Towards a Theory of Learning under Extreme Non-identifiability

Through the lens of causal learning and kernel clustering

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
M. Sc. Leena Chennuru Vankadara
aus Proddatur, India

Tübingen
2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	12.10.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr. Ulrike von Luxburg
2. Berichterstatter:	Prof. Dr. Arthur Gretton
3. Berichterstatter:	Prof. Dr. Ingo Steinwart

Contents

Acknowledgments

Thesis advisors greatly impact the quality of one's experience in graduate school. I was extremely fortunate to have had two incredibly supportive advisors: Ulrike von Luxburg and Debarghya Ghosdastidar. I feel very fortunate to have had Ulrike as a mentor. She has been incredibly supportive in every way imaginable. She allowed me absolute freedom in my research and always believed in me. She was always there when I needed her advice on both professional and personal fronts and motivated me during all my Ph.D. struggles. She always put my interests and goals above everything else, and I am deeply grateful for that. I know I can continue to count on her invaluable support and mentorship for the rest of my life, and I will always cherish this relationship.

I am very lucky to be Debarghya's first Ph.D. student. He was deeply invested in my success and growth as a graduate student. I could discuss ideas with him for hours together, no matter the time or day. His dedication to research and work ethic were truly inspiring. Our weekly meetings often lasted hours (even over zoom !) filled with insightful discussions and new ideas for research. He was extremely supportive of my ideas, allowed me complete research freedom, and placed a lot of trust in me. He was fully dedicated to his role as a supervisor, always offered valuable input in every aspect, and was constantly looking for ways to improve our relationship. I am deeply grateful for his unwavering support and guidance throughout my Ph.D.

I am also very fortunate to have had the pleasure of working with Dominik Janzing during an internship at Amazon Research. Dominik's enthusiasm for research and ideas was infectious. He thought deeply about ideas and was completely invested in them. Working with Dominik was truly stimulating, and this experience shaped the last two years of my Ph.D. I look forward to working on more exciting ideas with Dominik in my next endeavor.

I am truly grateful to all my collaborators — Siavash, Luca, Michael, Sebastian, Lenon, Maha, Pascal, Mila, and Philipp — without whom my Ph.D. journey would have been a lot less exciting. I am very thankful to Damien Garreau and Michael Perrot for plenty of useful advice at the beginning of my Ph.D. I am grateful to all my colleagues in Tuebingen — Moritz, Solveig, Luca, Sascha, David, Siavash, Michael, and Sebastian — for all the fun discussions as well as the open and honest discussions about work, Ph.D.

life, and mental health. It's not a luxury offered to everyone, and I am grateful for that. I am very grateful for my colleagues in Munich — Maha, Pascal, and Satyaki. Even though I only met them in person a couple of times, I was always made to feel welcome in the group.

I am thankful for the espresso machine in the building. I never found out who donated it or bought it, but thanks to whoever that is! In addition to providing fuel for deadlines, I fell in love with espresso and latte art because of this machine, and the passion for coffee is here to stay with me for a long time.

I would like to express my gratitude for the excellent research and collaborative environment in Tuebingen. I was able to attend reading groups and have discussions in other departments, collaborate, and build extremely useful relationships because of this. I would like to express my thanks to the Tuebingen AI center and the Baden-Wuerttemberg Stiftung for funding my Ph.D.

I am grateful for my family — my dad and Ravant — and my friends — Nikks, Tulasi, Vysh, Kusi, Akarsh — who I can always count on to be by my side in times of need.

To my mom, this would not have been possible without your unconditional support and love. You are my rock, and no amount of gratitude can ever pay you back for the support and inspiration you offer me.

There is no way I could have gotten through the last 5 years of my life without my incredible partner Faiz. He has been my constant source of support, bore me during the hardest of times, motivated me, and consoled me. Faiz, my love, there are no words to express how much your love, kindness, and unwavering support mean to me. I dedicate this thesis to you.

Abstract

Recent years have witnessed remarkable breakthroughs across various scientific and technological pursuits, including visual recognition and language understanding that are approaching human-level capabilities and protein folding. These breakthroughs have been driven in part by the availability of data and computational resources at unprecedented scales, as well as rapid progress in designing and optimizing highly complex learning algorithms.

Despite the impressive success of modern machine learning, the real-world applicability of these algorithms is hindered by two interrelated problems. First, a theoretical understanding of these complex methods is far from complete. Second, from a practical standpoint, these methods are brittle against adversarial examples and domain shifts and often do not comply with quality attributes such as fairness and privacy. Many of these problems can be characterized in the framework of *learning under extreme non-identifiability* and constitute the key challenges of contemporary research in machine learning. Unsurprisingly, the lack of a theoretical understanding of algorithms under this framework is particularly severe.

There are three key aspects any theory should consider in this framework:

(a) *Identification.* The problem of learning from finite samples fundamentally suffers from the issue of non-identifiability. The key to obtaining any theoretical justification is to define a model (through assumptions based on prior knowledge) and seek a *model-relative* justification. This is even more crucial for learning under extreme non-identifiability, where the issue of non-identifiability persists even in the infinite sample limit.

(b) *Estimation.* Once the parameters or quantities of interest are precisely identified, then the problem of learning is merely a problem of statistical estimation and can admit theoretical guarantees (for example, consistency or rates of convergence).

(c) *Computation.* Computation is the other key aspect of learning. While the focus of statistics is to provide guarantees under constraints on sample sizes, from a computational perspective, the focus is to obtain guarantees under constraints on the available computation.

In this thesis, we contribute to the theoretical foundations of the problem of learning under extreme non-identifiability with emphasis (to varying degrees) on each of the three aspects. In particular, we consider the problems of statistical clustering and causal learning and make the following contributions.

(a) *Theory of kernel clustering.* We provide recovery guarantees for kernel-based clustering under parametric and non-parametric assumptions on the data-generating process. We study phase transitions in the problem of high-dimensional Gaussian clustering and show that kernel-based clustering algorithms can be informational-theoretically optimal in their respective computational classes.

(b) *Theory of causal learning.* We introduce the framework of *causal learning theory* for forecasting and provide finite-sample uniform convergence guarantees for the causal risk of the class of vector autoregressive models. Under a linear causal model with potential latent confounders, we study the problem of causal generalization from the lens of interpolation and regularization.

Zusammenfassung

In den letzten Jahren wurden in verschiedenen wissenschaftlichen und technologischen Bereichen bemerkenswerte Durchbrüche erzielt. Dazu gehören die visuelle Erkennung und das Sprachverständnis, welche sich dem menschlichen Niveau nähern, sowie die Proteinfaltung. Diese Durchbrüche wurden teilweise vorangetrieben durch die Verfügbarkeit von Daten und Rechenleistung in noch nie dagewesenem Umfang sowie durch rasche Fortschritte bei der Entwicklung und Optimierung hochkomplexer Lernalgorithmen.

Trotz des beeindruckenden Erfolgs des modernen maschinellen Lernens wird die praktische Anwendbarkeit dieser Algorithmen durch zwei miteinander verknüpfte Probleme behindert. Erstens ist das theoretische Verständnis für diese komplexen Methoden noch lange nicht abgeschlossen. Zweitens sind diese Methoden von einem praktischen Standpunkt aus gesehen anfällig gegenüber feindlichen Beispielen und Domänenverschiebungen und erfüllen oft Qualitätsmerkmale wie Fairness und Datenschutz nicht. Viele dieser Probleme lassen sich im Rahmen des *Lernens unter extremer Nicht-Identifizierbarkeit* charakterisieren und stellen die zentralen Herausforderungen des zeitgenössischen maschinellen Lernens dar. Wenig überraschend ist der Mangel an theoretischem Verständnis von Algorithmen in diesem Rahmen besonders gravierend.

Es gibt drei Schlüsselaspekte, die jede Theorie in diesem Rahmen berücksichtigen sollte:

(a) *Identifikation.* Das Problem des Lernens von endlichen Stichproben leidet grundlegend unter dem Problem der Nicht- Identifizierbarkeit. Der Schlüssel zur Erlangung einer theoretischen Rechtfertigung ist ein Modell zu definieren (durch Annahmen, die auf Vorwissen beruhen) und eine *modellbezogene* Rechtfertigung zu suchen. Dies ist sogar noch entscheidender für das Lernen unter extremer Nicht-Identifizierbarkeit, wo das Problem der Nicht-Identifizierbarkeit auch bei unendlichen Stichproben fortbesteht.

(b) *Schätzung.* Sobald die gesuchten Parameter oder Größen genau bestimmt sind, ist das Problem des Lernens lediglich ein Problem der statistischen Schätzung und kann theoretische Garantien ermöglichen (zum Beispiel Konsistenz oder Konvergenzraten).

(c) *Berechnung.* Berechnung ist der dritte Schlüsselaspekt des Lernens. Während der Schwerpunkt der Statistik darin besteht, Garantien unter Beschränkungen des Stichprobenumfangs zu geben, liegt der Schwerpunkt aus der Sicht der Berechnung auf Garantien unter Beschränkung der verfügbaren Rechenleistung.

In dieser Arbeit leisten wir einen Beitrag zu den theoretischen Grundlagen des Lernens unter extremer Nicht-Identifizierbarkeit, wobei jeder dieser Aspekte (in unterschiedlichem Maße) im Vordergrund steht. Insbesondere betrachten wir die Probleme der statistischen Clusterbildung und des kausalen Lernens und leisten die folgenden Beiträge.

(a) *Theorie der kernelbasierten Clusterbildung.*

Wir bieten Wiederherstellungsgarantien für kernelbasierte Clusterbildung unter parametrischen und nicht-parametrischen Annahmen über den datenerzeugenden Prozess. Wir untersuchen Phasenübergänge beim Problem der hochdimensionalen Gausschen Clusterbildung und zeigen, dass kernelbasierte Cluster-Algorithmen in ihren jeweiligen Rechenklassen informationstheoretisch optimal sein können.

(b) *Theorie des kausalen Lernens.* Wir führen den Rahmen der *kausalen Lerntheorie* für Prognosen ein und bieten uniforme Konvergenzgarantien unter endlicher Stichprobe für die Klasse der vektorautoregressiven Modelle. Im Rahmen eines linearen Kausal modells mit potenziellen latenten Störfaktoren untersuchen wir das Problem der kausalen Generalisierung unter dem Aspekt der Interpolation und Regularisierung.

Part I

Introduction

A guide to the introduction

How to read the introduction? The introduction to this thesis is comprised of five sections. We begin by discussing the issue of non-identifiability in the problem of learning or inductive inference. We review the philosophical and theoretical perspectives on this problem and a means of reconciliation (Chapter 1). We then discuss the issue of theoretical justification of statistical learning methods in light of the issue of non-identifiability. We review the framework of statistical learning theory and highlight some landmark results (Chapter 2). In Chapter 3, we go beyond learning under the standard i.i.d. setting and introduce the issue of extreme non-identifiability. We discuss some of the challenging problems in modern machine learning in light of the extreme non-identifiability issue emphasizing the problem of statistical clustering and causal learning. In Chapter 4, we review existing work in the theory of statistical clustering and causal learning. In Chapter 5, we discuss some limitations of the current state of theory in clustering and causal learning and present our contributions to these topics.

1
Learning under non-identifiability.

> *"When the mind passes from the idea or impression of one object to the idea or belief of another, it is not determin'd by reason, but by certain principles, which associate together the ideas of these objects, and unite them in the imagination."*
>
> David Hume, A Treatise of Human Nature

In the summer of 1958, a press conference was held by the United States Office of Naval Research to unveil an astounding invention called the perceptron. The IBM 704 computer, which was large enough to fill an entire room, was fed 100 punch cards with squares placed on either the left or right side of a field. The perceptron was able to correctly distinguish between the two placements in 97 instances after only seeing 30 to 40 examples. The New York Times reported that the perceptron was expected to be "the first non-living organism to perceive, recognize, and identify its surroundings without any human training or control," according to its inventor Frank Rosenblatt [NYT, 1958].

Such intelligent machines — from the tale of a bronze man called Talos built by Hephaestus, the Greek god of invention, conceptualized as far back as 700BC, to modern science fiction references like Frankenstein's monster or our beloved star wars duo R2D2 and C3PO — have long captured the human imagination. While the perceptron fell short of its grand promise in the 1960s, fuelled by an abundance of data and enormous advances in computing power, we have come a very long way in building highly complex machines that achieve (or even surpass) human-level performance on a wide array of intelligent tasks like image recognition [Krizhevsky et al., 2017], speech recognition [Radford et al., 2022], or even discovering new algorithms [Fawzi et al., 2022]. Despite the enormous complexity of these modern intelligent machines, the conceptual goal of these machines is the same as it was for perception: *the problem of learning*.

1.1 *What is learning?*

WHAT IS LEARNING? On an abstract level, it refers to the problem of inferring *a law of nature* based on *experiences*. This is also referred to as the problem of inductive inference. Imagine being an explorer lost in a forest for weeks. Having not eaten in days, you come across a bunch of purple berries and are tempted to eat them. You decide to take a chance and find them delicious. For the next ten days, along your trail, you find more purple berries, and you find them delicious every time. Based on these experiences, you conclude a general law: all purple berries are edible and delicious. Similarly, much like how a child learns, after seeing some examples of squares placed on the right or left, the perception algorithm learns to distinguish between the two cases, even in unseen examples. Such inductive inferences, however, are merely probable. Assuming the truth of the premises, it is likely that the conclusion is true. This stands in contrast to the other common type of cogent reasoning — typically referred to as deductive inference — where the truth of the premises logically entails the truth of the conclusions. For example, if all berries are delicious and some berries are purple, you may conclude via deductive reasoning that all purple berries are delicious.

1.2 *Hume's induction dilemma.*

THE ISSUE OF NON-IDENTIFIABILITY, also known as the problem of induction, lies at the heart of the philosophical and theoretical foundations of the problem of learning. This problem was highlighted by the philosopher and empiricist David Hume in his book "A Treatise of Human Nature," published in 1739. Hume asks what the basis is for inductive inference and how we can justify inferring that the next purple berry we encounter will also be edible. By posing this question, Hume challenges the underlying assumption of inductive inference: the uniformity of the law of nature, which is the assumption that our past experiences will resemble those we have not yet had.

Hume then proceeds to present his famous two-horn dilemma of induction. According to Hume, all epistemological reasoning can be divided into two categories: deductive or demonstrative reasoning, and inductive or empirical reasoning. The first horn of the dilemma argues that it is impossible to justify the uniformity principle through purely deductive reasoning, as such reasoning always leads to conclusions that cannot be falsified. At the same time, we can find clear examples that contradict the uniformity principle - the next purple berry we encounter could be poisonous. The second horn of the dilemma argues that there can be no empirical or inductive basis for the uniformity principle either, as any such argument would be circular. For example, one might argue that inductive inferences have worked for us in the past, but this would again require the assumption of the uniformity of the law of nature, making the argument circular. Therefore, Hume concludes that there can be "no rational justification" for induction.

1.3 *No free lunch theorems.*

THE ESSENCE OF THE PROBLEM OF INDUCTION or the issue of non-identifiability is captured in the theory of learning via a series of mathematical impossibility results called the no-free-lunch (NFL) theorems [Wolpert, 1992, Schaffer, 1994, Wolpert, 1996, 2002].

To understand these results, let's introduce some notation to formally describe the learning problem. Suppose we have a dataset of training examples of the form $\{(x_i, y_i)_{i=1}^n\}$, where x_i's (e.g., one of the purple berries) belong to some input space $\mathcal{X}$ and y_i's (e.g., the deliciousness of the berry) belong to a target space $\mathcal{Y}$. The elements of the input space are called features or covariates, and the elements of the output space $\mathcal{Y}$ are called targets or labels. The goal of a learning algorithm is then to learn the underlying functional relationship between the features and the labels (e.g., the relationship between purple berries and deliciousness).

Now consider the problem of binary classification (i.e., where the labels y_i are assumed to be in $\{+1, -1\}$) on a discrete input space $\mathcal{X}$, and suppose that the training data is sampled from some unknown distribution $\mathbb{P}$. The NFL theorems show that, in expectation under a uniform distribution over all possible true labelings, no algorithm can achieve superior performance over another algorithm.[1] In other words, on average, no algorithm can do better than random guessing. Schaffer refers to this phenomenon as the "conservation of generalization performance" since it suggests that if an algorithm performs really well on some distributions, this must necessarily be offset by poor performance on others [Schaffer, 1994]. These results echo Hume's inductive skepticism.

DESPITE THE PHILOSOPHICAL and mathematical impossibility of induction, it is the foundation of not only computational learning but also the scientific method itself. So, how do we reconcile these impossibility results and Hume's dilemma of induction with the practicality of inductive inference in our everyday lives and in scientific research?[2]

1.4 *Reconciling theory and practice.*

While the NFL theorems appear rather strong at first glance, a path to reconciliation becomes clear after a moment of deliberation. The NFL results can be interpreted as saying that, without assuming any relationship between the experiences made so far

[1] In these results, performance is measured on a separate hold-out set also drawn from the same underlying distribution $\mathbb{P}$ but is disjoint from the training set.

[2] There has been extensive debate in philosophy around Hume's induction dilemma and various proposed resolutions. We refrain from diving into this literature and instead refer the interested reader to Sterkenburg and Grünwald [2021], Henderson [2022]

(i.e., the training data) and future experiences (i.e., test data) and if all functions between the features and the labels are equally likely, then learning is impossible. This is far from how we apply induction in practice. Wolpert and Schaffer were aware of this gap and believed that the "conservation law is theoretically sound but practically irrelevant". Nevertheless, these results highlight an idea fundamental to learning, i.e., learning is impossible if an algorithm does not incorporate any *prior knowledge* of the underlying task.

Analogously, in the philosophy literature, Sober and Okasha reconcile Hume's induction dilemma with the practicality of inductive inference similarly by taking a *local view of induction* [Sober, 1991, Okasha, 2005]. They argue that the induction dilemma can be sidestepped by questioning the presupposition of the uniformity of the law of nature. In other words, there is no *apriori* assumption, such as the uniformity of the law of nature, which can justify all inductive inferences. Instead, every inductive inference relies on some assumptions. This corresponds to challenging the apriori assumption of a uniform (or any) distribution over possible labelings in the NFL results. In the local view, this means that learning may be possible with additional *local assumptions* based on prior knowledge of the learning task.

1.5 *How do we incorporate prior knowledge?*

A standard assumption in the setting of statistical learning is that the training data and unseen test data are drawn from the same (unknown) probability distribution $\mathbb{P}$ (i.i.d.). Under this assumption, it is possible to find algorithms that are *universally consistent* [Stone, 1977, Devroye and Wagner, 1979b, Bartlett and Traskin, 2006, Collins et al., 2020]; that is, for any distribution $\mathbb{P}$, the performance of the algorithms gets closer to optimal with increasing size (n) of the training sample.[3] Such a result may elicit hope for a universal learning algorithm under this general i.i.d. assumption. However, for any fixed n, one can always find distributions on which the error incurred by these algorithms is arbitrarily close to random guessing. At the same time, another predictor is optimal and incurs zero error (see Devroye et al. [2013]). This clearly emphasizes the need for stronger assumptions that encode *local* prior knowledge to be incorporated into the learning process. There are several ways to achieve this, for example, by choosing an appropriate loss function or through the topology of the input space. In machine learning (ML), this is typically achieved by imposing some restrictions on the class of the functions from which one wishes to learn a good approximation to the optimal function. For example, in Bayesian approaches, via a prior distribution over the space of possible predictors, and in classical frequentist approaches, typically by choosing a predefined class of predictors. We will discuss this approach further in the next section.

[3] Note that even under the iid assumption, we don't observe the entire domain even as n approaches infinity. Universally consistent algorithms can (without additional assumptions) exist only due to the precise manner in which measurable spaces and functions are defined. See, for example, the discussion in Bousquet et al. [2003, Appendix B]

2
Theory of statistical learning

"Nothing is more practical than a good theory."

Kurt Lewin; Vladimir Vapnik.

First, let us recall the formal setup of binary classification. Given n samples of the form $\{(x_i, y_i)_{i=1}^n\}$ drawn i.i.d. from some unknown probability distribution $\mathbb{P}$ on $\mathbb{R}^d \times \{\pm 1\}$, and a non-negative loss function $l : \{\pm 1\} \times \{\pm 1\} \to \mathbb{R}^+$, the goal is to learn the function $f^* : \mathbb{R}^d \to \{\pm 1\}$ which minimizes the expected risk under $\mathbb{P}$

$$f^* := \underset{f:\mathbb{R}^d \to \{\pm 1\}}{\arg\inf} \; \mathbb{E}_{\mathbb{P}(x,y)} \, l(y, f(x)). \tag{2.1}$$

However, as discussed in Chapter 1, f^* is non-identifiable at finite n, and therefore the goal is to approximate f^* as well as possible under some assumptions.

2.1 *Parametric approach*

In classical statistics, a typical approach to solving this problem is to assume that the underlying distributions can be specified up to a small number of parameters independent of the sample size. Then, the maximum likelihood estimation (MLE) approach is typically used for parameter estimation. Theoretical analysis in this approach is specialized to the specified class of distributions and the corresponding effectiveness of MLE. This approach can yield good performance in simple settings and low dimensions where such strong priors may be justified. However, parametric approaches are generally not suitable for more real-world settings where it is typically not possible to model distributions using a small set of parameters.[1]

[1] Modern ML models are notoriously complex with number of parameters running into the billions.

2.2 *Non-parametric approach*

A MORE FLEXIBLE, NON-PARAMETRIC FRAMEWORK was pioneered by Vapnik and Chervonekis [Vapnik and Chervonenkis, 1971], building on the foundations of empirical process theory, specifically on uniform laws of large numbers such as the Glivenko-Cantelli-Kolmogorov theorem and its generalizations. In a non-parametric approach, the goal is to find and justify an inductive principle that can be applied in a general setting for any class of approximating functions. From an algorithmic point of view, a natural candidate for such a general inductive method is the principle of empirical risk minimization (ERM). The main idea is that one can approximate the expectation under the underlying distribution $\mathbb{P}$ in (2.1) by an expectation under the empirical distribution based on the training data. That is, to use the function $\hat{f}$, which minimizes the *empirical risk* over the space of all functions as an approximation to f^* in (2.1).

$$\hat{f} := \operatorname*{arg\,inf}_{f:\mathbb{R}^d \to \{\pm 1\}} \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)). \tag{2.2}$$

However, due to the issue of non-identifiability, the optimization problem in (2.2) is not well-specified. A typical approach to dealing with this issue is restricting the class of approximating functions to make the solution unique. For a fixed class of approximating functions $\mathcal{H}$ (for example, the class of neural networks with a fixed architecture), the ERM predictor f_{erm} is defined as the solution to the optimization problem in (2.3).

$$f_{\text{erm}} := \operatorname*{arg\,min}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)). \tag{2.3}$$

We are mainly interested in controlling the *excess risk* of the ERM predictor f_{erm} which is defined as the difference between the risk incurred by f_{erm} and the best possible risk under $\mathbb{P}$. Denoting the expected risk incurred by any predictor f under the distribution $\mathbb{P}$ as $\mathcal{R}(f) = \mathbb{E}_{\mathbb{P}(x,y)}\, l(y, f(x))$, the excess risk is given by

$$\mathcal{E}(f_{\text{erm}}) := \mathcal{R}(f_{\text{erm}}) - \mathcal{R}(f^*).$$

The excess risk can be decomposed into two components: estimation error and approximation error

$$\mathcal{R}(f_{\text{erm}}) - \mathcal{R}(f^*) = \underbrace{\mathcal{R}(f_{\text{erm}}) - \inf_{f\in\mathcal{H}} \mathcal{R}(f)}_{\text{estimation error}} + \underbrace{\inf_{f\in\mathcal{H}} \mathcal{R}(f) - \mathcal{R}(f^*)}_{\text{approximation error}}.$$

Observe that the two error terms are different in character. While the problem of controlling the approximation error is a conceptual one relating to the non-identifiability issue that was discussed in Section 1, controlling estimation error is a statistical problem that corresponds to seeking a *model-relative* justification. A majority of the vibrant field of statistical learning theory (SLT) focuses on the latter issue. Specifically, addressing the following questions has been the primary focus of this field (with answers to some questions more satisfactory than others).

1. What are the necessary and sufficient conditions under which the estimation error of f_{erm} vanishes as the number of training samples approaches infinity? This property is often referred to as statistical consistency.

2. How fast does the estimation error converge with increasing n? What is the optimal rate of convergence?

3. Can we find algorithms or learning principles with optimal predictive ability? What can we say about approximation error?

4. How do the computational properties of estimators interplay with their statistical properties? What are the fundamental limits of learning under computational constraints?

Here, we briefly review some standard approaches to addressing these questions and review some landmark results along the way. However, this area of research is incredibly rich, and this review should not, by any means, be considered exhaustive.

2.2.1 *Necessary and sufficient conditions for consistency of ERM*

A celebrated result in the field of statistical learning theory is a complete characterization of conditions that are both sufficient and necessary for the consistency of empirical risk minimization (ERM). Vapnik and Chervonenkis [1971] showed that in binary classification setting, consistency of empirical risk minimization is equivalent to stating that the hypothesis class $\mathcal{H}$ is a *uniform Glivenko Cantelli* class (uGC), that is, uniformly over all functions in $\mathcal{H}$ and uniformly over all probability distributions, empirical risk converges (in probability) to the expected risk. Furthermore, Vapnik and Chervonenkis [1971] showed that uniform Glivenko Cantelli hypothesis classes can be precisely characterized by the finiteness of a combinatorial measure of *complexity* or *capacity* of the function class called VC dimension. Alon et al. [1997] showed that such a characterization also holds for the setting of regression with a generalized notion of complexity measure called *fat-shattering-dimension*. $\mathcal{H}$ is uGC if and only if the fat-shattering dimension of $\mathcal{H}$ is finite at every finite scale. Algorithmic stability is another algorithm-dependent notion of complexity that can also be used to characterize the consistency of ERM [Kutin and Niyogi, 2002, Mukherjee et al., 2006]. Specifically, consistency of ERM holds if and only if ERM is stable on $\mathcal{H}$ (for some weak notion of stability).

Note that consistency is a property of the induction principle, i.e., ERM. However, consistency is closely related to the notion of statistical learnability, which is defined as a property of the hypothesis class $\mathcal{H}$. $\mathcal{H}$ is learnable if and only if there exists a learning mechanism for which consistency holds. In the setting of classification or regression, it has been shown that a necessary and sufficient condition for learnability is that $\mathcal{H}$ is uGC [Vapnik and Chervonenkis, 1971, Alon et al., 1997]. Since ERM is consistent under this condition, it holds that if a hypothesis class is learnable, then it is learnable by ERM. However, settings exist beyond supervised classification and regression where these notions diverge. See Shalev-Shwartz et al. [2010] for a thorough discussion on such settings and the interplay between uniform convergence, learnability, and stability. These results provide a complete answer to the first question concerning the consistency of ERM.

2.2.2 *Convergence rates for estimation error*

There has been intense work in the past two decades on the topic of deriving convergence rates for estimation error. Before we review them, let us first introduce the notion of *generalization gap*. Assume without loss of generality that $f_{\text{opt}} \in \arg\inf_{f\in\mathcal{H}} \mathcal{R}(f)$ exists. Notice that the estimation error admits the decomposition,

$$\mathcal{R}(f_{\text{erm}}) - \mathcal{R}(f_{\text{opt}}) = (\mathcal{R}(f_{\text{erm}}) - \hat{\mathcal{R}}(f_{\text{erm}}))+ \\ (\hat{\mathcal{R}}(f_{\text{erm}}) - \hat{\mathcal{R}}(f_{\text{opt}})) + (\hat{\mathcal{R}}(f_{\text{opt}}) - \mathcal{R}(f_{\text{opt}})).$$

The second expression in the decomposition cannot be strictly positive. Since f_{opt} is a fixed function, an application of the law of large numbers reveals that the third expression vanishes (under some mild assumptions as $\mathcal{O}(1/\sqrt{n})$). However, since f_{erm} is a function of the training data, a simple law of large numbers cannot be applied to the first expression. Consequently, the object of all analyses is typically the first term: the difference between an estimator's empirical risk and the expected risk. This term is typically referred to as the generalization error (or gap), and corresponding bounds for this expression are termed generalization bounds. There are several ways to derive generalization bounds (for example, uniform convergence, algorithmic stability, or margin-based bounds). Conceptually, we can broadly categorize them in two ways based on whether or not they depend on the underlying data distribution.

Data-independent generalization bounds

Obtaining bounds on the estimation error that holds independent of the data distribution can be of interest since they can be evaluated without the knowledge of the underlying probability distribution. To derive such bounds, two predominant approaches exist 1) uniform convergence and 2) algorithmic stability.[2] Both were already referenced in our discussion of conditions for the consistency of ERM.

Uniform convergence. A key observation that underlies this approach is the following. Since f_{erm} depends on the training data, it is not clear which model is chosen by the algorithm apriori. One way to overcome this is to apply a crude upper bound on generalization error as in (2.4) and treat the upper bound instead, which, incidentally, has been one of the most investigated objects in empirical process theory.

$$\mathcal{R}(f_{\text{erm}}) - \hat{\mathcal{R}}(f_{\text{erm}}) \leq \sup_{f\in\mathcal{H}}(\mathcal{R}(f) - \hat{\mathcal{R}}(f)). \tag{2.4}$$

Leveraging tools from empirical process theory, one can typically obtain high probability, *uniform convergence* bounds of the form[3]

$$\forall f \in \mathcal{H}, \;\; \mathcal{R}(f) - \hat{\mathcal{R}}(f) \leq \mathcal{O}\Big(\frac{cap(\mathcal{H})}{\sqrt{n}}\Big). \tag{2.5}$$

[2] There are many other approaches to deriving generalization bounds, for example, using information-theoretic properties. The two discussed here are widely entrenched.

[3] This bound should only serve as an exemplar. For example, there has been a lot of work on improving the convergence rate and obtaining minimax optimal convergence rates.

Here, $cap(\mathcal{H})$ refers to some measure of complexity or capacity of the hypothesis class $\mathcal{H}$, for example, VC-dimension for binary classification [Vapnik and Chervonenkis, 1971] or fat-shattering dimension in the context of regression with real-valued functions [Alon et al., 1997]. Many other complexity measures, such as the Natarajan dimension or pseudo-dimension, have been utilized to derive similar results for more general settings. For an overview of the different complexity measures, see Anthony et al. [1999]. A distinguishing feature of such bounds is that they are agnostic to the underlying probability distribution and hold uniformly over the space of all functions in $\mathcal{H}$. We now turn to the other prominent approach to deriving generalization bounds.

Algorithmic stability. Generalization bounds, such as the ones in (2.5), are typically used for model selection (for example, to select a hypothesis class with a suitable complexity measure). It is, therefore, crucial that the obtained bounds closely reflect the ground truth. However, due to the crude upper bound in (2.4), bounds in (2.5) can be weak since they hold uniformly over all the predictors in $\mathcal{H}$. The notion of algorithmic stability, which is inspired by Rogers and Wagner [1978], Devroye and Wagner [1979b,a], provides an alternative. Several notions of algorithmic stability exist, such as leave-one-out stability, cross-validation stability, and uniform stability, among many others.[4] Intuitively, they all capture the sensitivity of an algorithm's output to small perturbations of the training set. Uniform stability, proposed by Bousquet and Elisseeff [2002], is the strictest of these notions. It requires that for any two training sets from the input space that differ by one element, the algorithm's output (or loss) does not change much.[5] There is an entrenched line of work deriving generalization bounds for uniformly stable algorithms [Devroye and Wagner, 1979a, Bousquet and Elisseeff, 2002, Zhang, 2003, Maurer, 2017, Feldman and Vondrak, 2018, 2019, Bousquet et al., 2020]. Most notably, Feldman and Vondrak made significant breakthroughs by deriving nearly-tight high-probability generalization bounds for uniformly stable algorithms [Feldman and Vondrak, 2018, 2019]. The intuition here is that stability controls the estimator's variance, thereby controlling the generalization error.[6] Stability-based bounds can often be much tighter since they can take into account the properties of the specific model chosen by the algorithm. Here, the *size* of the hypothesis class does not matter but rather how the algorithm explores this space [Bousquet and Elisseeff, 2002]. A classical example is that of the k-nearest neighbor (kNN) classifier whose hypothesis class has an infinite VC dimension and, therefore, corresponding bounds are vacuous. However, Rogers and Wagner [1978] derived non-vacuous generalization bounds for the kNN classifier based on its stability. Another notable example is that of stochastic gradient descent for strongly convex losses for which uniform convergence bounds are

[4] See Kutin and Niyogi [2002] for an overview of different stability measures. Note that some stability notions are data-dependent (for example, average stability) and can also be utilized to derive generalization bounds [Shalev-Shwartz et al., 2010, Kuzborskij and Lampert, 2018]

[5] Many algorithms can be shown to satisfy this property

[6] Practical approaches, such as bagging, have been inspired by this intuition.

known to be at least $\Omega(\sqrt{d})$ times worse than those obtained via stability [Hardt et al., 2016, Feldman and Vondrak, 2018].

Data-dependent generalization bounds

In our discussion of algorithmic stability, we saw how sharper generalization bounds can be obtained in certain settings by considering the properties of the specific model chosen by the algorithm. Here, we discuss obtaining tighter bounds by considering complexity measures that depend on the data distribution. One of the widely employed data-dependent complexity measures, inspired by its utility in empirical process theory, is *Rademacher complexity* or the closely related *Gaussian complexity* [Koltchinskii, 2001, Mendelson, 2002, Bartlett et al., 2002, Bartlett and Mendelson, 2002]. Intuitively, they measure how well the functions in $\mathcal{H}$ *restricted to the training set* correlate with random noise. Since these measures capture properties of the data distribution, they often yield sharper uniform convergence bounds on generalization error. Crucially, bounds based on Gaussian or Rademacher complexity can be evaluated based on the training data and do not require knowledge of the underlying distribution.

One can further improve the estimation rates by considering bounds that do not hold uniformly over $\mathcal{H}$ but only a meaningful subset of $\mathcal{H}$ chosen in a data-dependent fashion. For example, one could only be concerned with the performance of predictors that achieve a small empirical risk on a given training set. The complexity of this subset could be considerably lower than that of $\mathcal{H}$, thereby improving the generalization bound. This intuition underlies the notion of *local* Rademacher complexity and *margin bounds*. In many other situations, one can obtain "fast rates" — convergence rates scaling at $\mathcal{O}(1/n)$ under low noise conditions [Tsybakov, 2004].

2.2.3 *Controlling approximation error. The framework of structural risk minimization.*

So far, we only discussed different approaches to controlling the estimation error of f_{erm}. Recall that excess risk is the sum of the estimation and approximation errors. Therefore, to control the excess risk, we need to obtain bounds on the approximation error. However, the NFL theorems discussed in Chapter 1 dictate that, without assumptions on the class of data distributions, one cannot provide finite-sample bounds simultaneously for the approximation error. Controlling the approximation error alone can be solved by simply choosing a sufficiently large $\mathcal{H}$. For example, if $\mathcal{H}$ is chosen to be the class of all possible functions, then the approximation error is 0. However, this is at odds with controlling the estimation error, which requires choosing $\mathcal{H}$ with low complexity. This dichotomy between the estimation and approximation errors is often termed as the bias-variance tradeoff in statistics.

The framework of *structural risk minimization* suggests choosing a complexity hierarchy of function classes $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_3 \subseteq \cdots$ and choose the class with the appropriate complexity based on the training data to achieve the correct tradeoff between the estimation and approximation errors. Many learning algorithms have been motivated by this approach, for example, support vector machines [Cortes and Vapnik, 1995] and ridge [Golub et al., 1999] or lasso regression [Tibshirani, 1996].

2.2.4 *Statistical-computational tradeoffs*

So far, we have discussed statistical aspects of learning problems. The main object of this study is to investigate the performance of algorithms under constraints on the number of observed samples. The other fundamental aspect of learning problems is a computational one. Here, the goal is to investigate the performance of algorithms under the additional constraints on the amount of available computation. There are numerous ways to investigate the interplay between the statistical and computational properties of estimators. Here, we restrict our attention to the two different questions commonly studied in learning theory. Naturally, one may expect some kind of a tradeoff between the statistical and computational properties of learning algorithms.

1. Phase transitions in learning. In the study of *phase transitions*, the primary concern is to obtain precise conditions under which it is possible to obtain a consistent estimator within a given *computational complexity class*, such as the class of estimators that can be obtained using computation that is polynomial in the number of samples. This, for instance, is a key question underlying the framework of *probably approximately correct (PAC)* learning proposed by Valiant [Valiant, 1984]. See Kearns and Vazirani [1994] for an introduction to this topic. In this setting, no emphasis is placed on ordering the different estimators within a computational class.

2. Trade-offs within a computational class. The other complementary goal is to establish the precise tradeoffs between statistics and computation *within a computational class.* For instance, one algorithmic question pursued in this setting is the following: Given an estimator that is statistically optimal (in some sense) within a computational class, is it possible to find an estimator in the class that uses less computation without relinquishing statistical optimality? Typical approaches to addressing this question involve using statistically efficient approximations of the optimal algorithm, for instance, via subsampling procedures [Rudi et al., 2015, Rudi and Rosasco, 2017, Calandriello et al., 2017].

2.3 *Failure of uniform convergence. Theory of deep learning.*

Recent breakthroughs in the field of machine learning have primarily been driven by the so-called deep-learning models fuelled by data and computation at unprecedented scales. However, the empirical success of these complex learning algorithms has been elusive to standard tools of statistical learning theory. These models exhibit a curious statistical behavior — heavily overparameterized models with the ability to fit random labels and often trained to interpolate the training data achieve impressive out-of-sample generalization performance even under the presence of large amounts of label noise [Zhang et al., 2021, Belkin et al., 2018]. This phenomenon is often referred to as *benign overfitting* in the literature. This is in seeming defiance of classical learning theory wisdom, which suggests choosing a hypothesis that balances data-fitting with some measure of complexity of the hypothesis.

Belkin [2021] argues that no general bounds (including data-dependent ones) on the difference between the training and the test risks of interpolating estimators can exist since such bounds would need to have the knowledge of the noise level of the problem apriori. Nagarajan and Kolter [2019] formally show that there exist settings where uniform convergence cannot explain generalization for interpolation. Bartlett and Long [2021] extend these results to show that any risk bounds for linear regression with the minimum norm interpolating solution must be loose for some probability distribution. In other words, meaningful risk bounds for interpolating estimators must rely on the specific properties of the data distribution and, therefore, cannot hold uniformly over every distribution. A considerable amount of work has now been done in analyzing the asymptotic and non-asymptotic generalization properties of interpolating estimators in classification and regression settings under different assumptions on the data distributions [Hastie et al., 2019, Mei and Montanari, 2019, Muthukumar et al., 2020, Montanari et al., 2019, Liang and Sur, 2020]. Another line of work studies the role of optimization and suggests that even if no explicit regularization is enforced during training, the choice of optimization enforces implicit regularization in training the models [Soudry et al., 2018, Gunasekar et al., 2018a,b, Ji and Telgarsky, 2019].

3
Extreme non-identifiability

3.1 From non-identifiability to extreme non-identifiability

In Chapter 1, we discussed the problem of non-identifiability in statistical learning under the i.i.d. assumption through the lens of the no-free-lunch theorems and Hume's induction dilemma, emphasizing the need for strong prior assumptions. In Chapter 2, we discussed how one can formally provide model-relative justification for statistical learning methods under prior assumptions, such as restrictions on the class of approximating functions. We also briefly reviewed common theoretical approaches toward this goal. The i.i.d. assumption in statistical learning seems reasonable, and significant methodological progress has been made across various applications via learning methods that operate under this assumption. However, the i.i.d. assumption is too restrictive in the context of modern machine learning, where we care about more than merely predicting statistical associations. Real-world deployment of machine learning systems, particularly in safety-critical applications, demands a wide array of attributes such as robustness against adversarial perturbations or interventions, privacy, fairness, or generalizability across domains and tasks. There has been a distressing amount of evidence suggesting that deep learning models are vulnerable to adversarial perturbations[Szegedy et al., 2014, Goodfellow et al., 2015], certain kinds of interventions [Quinonero-Candela et al., 2008, Torralba and Efros, 2011, Kuehlkamp et al., 2017, Csurka et al., 2017], and even spatial transformations [Biscione and Bowers, 2020]. Despite these limitations, there is increasing adoption of ML models even in safety-critical applications such as autonomous driving [Yurtsever et al., 2020], medical diagnostics [Richens et al., 2020], and criminal justice [Rudin, 2019]. There is clearly a pressing need for both developing new methods to this end and (arguably), more importantly, providing theoretical guarantees to justify their inference. This constitutes the focal challenge of current machine learning research [Carlini and Wagner, 2017, Madry et al., 2018, Bartlett et al., 2021, Bubeck and Sellke, 2021, Heinze-Deml and Meinshausen, 2021]. Many of the problems discussed above can be studied under the framework of learning under *extreme non-*

identifiability. Recall the issue of non-identifiability in the i.i.d. setting.

Non-identifiability in the i.i.d. setting. The underlying model cannot be uniquely specified given a finite training set. However, as the training set size goes to infinity, the underlying model can be uniquely identified.

In contrast, by extreme non-identifiability, we refer (informally) to the following issue.

Extreme non-identifiability. Even if one observes an infinite amount of data from the data distribution, the underlying model may not be uniquely specified. Indeed, even asymptotically, infinitely many models may exist that explain the observed data.

To further clarify, let us discuss some of the challenging problems in machine learning in light of the extreme non-identifiability issue.

3.2 *Learning under extreme non-identifiability*

3.2.1 *Statistical Clustering.*

Clustering has been one of the foundational learning problems and is studied under the framework of unsupervised learning. The problem of clustering is typically stated as follows: Given a finite sample $X = \{x_i\}_{i=1}^n$ drawn i.i.d. from some unknown probability distribution $\mathbb{P}$ over the domain $\mathbb{R}^d$, find a *meaningful* partition of X. This formulation naturally brings us to the question of what meaningful partitions are. In the long line of clustering literature, this has been addressed in many different ways. One natural approach to formalize the problem of clustering is to assume that the data is generated according to some (potentially non-parametric) mixture model, where the goal is to recover the underlying mixture components. Under this natural conceptualization, *without any additional assumptions*, it can be easily shown that even if we have full knowledge of the underlying distribution $\mathbb{P}$, the mixture components are still non-identifiable [Teicher, 1963, Holzmann et al., 2006, Vandermeulen and Scott, 2015, Miao et al., 2016, Aragam et al., 2020]. See Figure 3.1 for a simple illustration.

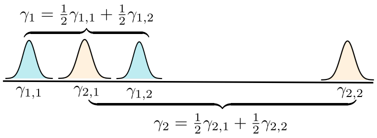

$\gamma_1 = \frac{1}{2}\gamma_{1,1} + \frac{1}{2}\gamma_{1,2}$
$\gamma_{1,1}$
$\gamma_{2,1}$
$\gamma_{1,2}$
$\gamma_{2,2}$
$\gamma_2 = \frac{1}{2}\gamma_{2,1} + \frac{1}{2}\gamma_{2,2}$

Figure 3.1: Example to show that even simple separation conditions do not suffice to overcome identifiability. As the distribution $\gamma_{2,2}$ moves arbitrarily far from the remaining distributions, the distance between γ_1 and γ_2 also increases arbitrarily. However, without additional assumptions, no clustering algorithm can recover the **desirable clusters** as defined by the **true components** γ_1 and γ_2. Figure from [Vankadara et al., 2021a].

Extreme non-identifiability in statistical clustering. Mixture models can be decomposed in infinitely many ways into their component distributions.

3.2.2 *Causal learning*

The problem of casual learning is one of the key challenges in modern machine learning, with close connections to many others such as robustness, fairness, transfer learning, and domain generalization [Peters et al., 2017, Schölkopf et al., 2021, 2011, Wang et al., 2022]. It can be viewed as an instance of the general problem of distributionally robust learning where the goal is to minimize the worst-case loss over a class of distributions defined by do-interventions [Meinshausen, 2018, Rothenhäusler et al., 2021].[1] Consider the "simple" problem of learning the *causal relationship* between two variables. x (e.g., taking a particular drug) and y (e.g., recovery of a disease). To learn such a relationship, we typically have access to *interventional* data, for example, through randomized control trials. However, access to interventional data may be prohibitive for a myriad of ethical, financial, or feasibility considerations, and one typically only has access to data from the statistical distribution. Assuming complete knowledge of the statistical distribution, even if one observes statistical dependence between x and y, it is unclear to what extent this can be attributed to one of the three possibilities: x influences y, y influences x or both x and y are influenced by some common cause z (as postulated in Reichenbach's common cause principle [Reichenbach, 1991]). See Figure 3.2.2 for a simple example.

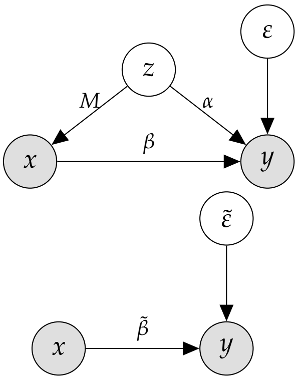


Figure 3.2: Graphical representation of the structural causal models given by $z \sim \mathcal{N}(0, I_l)\,, \varepsilon \sim \mathcal{N}(0, \sigma^2)\,, x = Mz\,, y = x^T\beta + z^T\alpha + \varepsilon$ **(top)** and $\tilde{\varepsilon} \sim \mathcal{N}(0, \tilde{\sigma}^2)\,, x \sim \mathcal{N}(0, MM^T)\,, y = x^T\tilde{\beta} + \tilde{\varepsilon}$ **(bottom)** with parameters $M \in \mathbb{R}^{d\times l}$, $\alpha \in \mathbb{R}^l, \beta \in \mathbb{R}^d$ and $\sigma^2 > 0$. Both linear causal models induce the same joint distribution over (x, y) for all parameters satisfying $\tilde{\beta} := \beta + \Gamma$ and $\tilde{\sigma}^2 := \sigma^2 + \|\alpha\|^2 - \|\Gamma\|^2_\Sigma$, where $\Gamma = M^{+T}\alpha$.

Extreme non-identifiability in causal learning. The underlying causal structure is non-identifiable even with full knowledge of the joint distribution. [Pearl, 2009, Peters et al., 2017]

In this manuscript, we will primarily focus on the problems of statistical clustering and causal learning; Below, we briefly mention a few other problems in the framework of learning under extreme non-identifiability.

3.2.3 *Domain Generalization*

One of the key limitations of the current state-of-the-art ML models is their inability to generalize to unseen *domains* (for example, see Heinze-Deml et al. [2018], Schölkopf et al. [2021] and references therein). For example, ML models trained on data collected on clear sunny days and deployed in the wild, for example, in autonomous driving systems, would be expected to make valid predictions even on foggy or rainy days. Other examples include using simulated data for training a model with the goal of generalization to real-world settings (see Figure 3.3 for an illustration).

This problem of generalization to unseen domains based on data from one or more related domains is called *domain generalization*. Note that this is different from *out-of-sample* generalization in the standard i.i.d. setting, where the test data is assumed to arise from the same distribution as the training data. Formally, given a collection of p datasets $D^j = \{(x_i^j, y_i^j)_{i=1}^n\}$ for $j \in [p]$, where $(x_i^j, y_i^j) \overset{i.i.d.}{\sim} \mathbb{P}^j_{xy}$, the goal is to learn a function $h : \mathbb{R}^d \times \mathbb{R}$

[1] In the language of distributionally robust optimization, the goal is to minimize the worst-case risk with respect to an *ambiguity set* of distributions which is typically defined as the class of distributions within a small radius of the training distribution with respect to some measure of discrepancy [Rahimian and Mehrotra, 2019].

such that the error incurred on some unseen domain $\mathbb{P}_{xy}^{test}$ is minimized, that is, h minimizes $\mathbb{E}_{\mathbb{P}^{test}(x,y)}\, l(y, g(x))$. Naturally, at this level of generality, if one makes no further assumptions about how the training and test distributions are related, the issue of non-identifiability persists even if we observe an infinite amount of data from the training distributions. Several other problems, such as transfer learning, domain adaptation, multi-task learning, and even causal learning, are closely related to the problem of domain generalization.

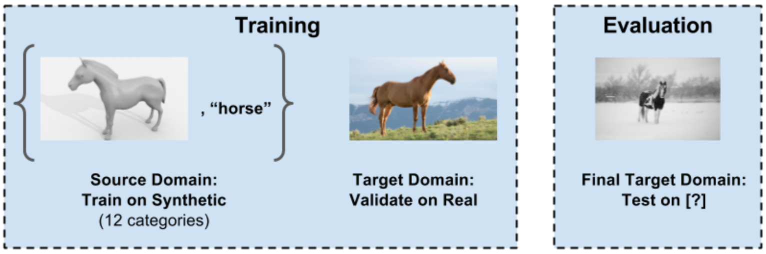


Figure 3.3: A sample image from a popular domain generalization dataset VISDA 2017 [Peng et al., 2017].

3.2.4 Transfer Learning

A humble but powerful tool that dictates the practical utility of deep learning models is the concept of *transfer learning*. For example, neural networks such as VGG16 [Simonyan and Zisserman, 2015] or Stable diffusion models [Rombach et al., 2022] in computer vision or the GPT-3 [Brown et al., 2020] and Whisper models [Radford et al., 2022] in natural language processing are trained on a huge corpus of data. These models are subsequently deployed in target domains or tasks that may differ from the training domain or task. In this setting, the models are assumed to have access to a small amount of labeled data from the target distribution. This problem is referred to as transfer learning and is closely related to the problem of domain generalization. Without additional assumptions on how the training and target domains or tasks are related, it suffers from the issue of extreme non-identifiability.

3.3 Reconciling theory and practice

Need for stronger assumptions. As discussed in Chapter 1, the problem of learning inherently suffers from the issue of non-identifiability and requires assumptions based on prior knowledge of the task. From the examples discussed so far, it is exceedingly clear that under extreme non-identifiability, it is imperative to impose arguably even stronger assumptions that render learning possible based on prior knowledge of the underlying problem setting. The principal challenge lies in finding a set of assumptions that satisfy some level of generality, are practically meaningful, and are amenable to theoretical analysis.

How do we make assumptions? Not surprisingly, the choice of assumptions highly depends on the underlying task, arguably even more so than in the case of learning in the i.i.d. setting. Numerous assumptions have been considered for various problem settings. For example, in the context of domain generalization or transfer learning, one may assume that the source and the target domains or tasks are "related" or close according to some discrepancy metric. In the context of clustering, one may assume that items within a cluster are closer to each other than those across clusters. For learning under extreme non-identifiability, specifying general yet practically relevant assumptions is a significant challenge both from a theoretical as well as a methodological point of view. In this manuscript, we focus on the problems of statistical clustering and causal learning. Below, we discuss some of the standard assumptions postulated in the corresponding literature. A high-level discussion along similar lines may be found in [Von Luxburg and Ben-David, 2005].

3.3.1 Assumptions for clustering

Unlike in the i.i.d. setting of classification and regression where non-identifiability is a finite sample issue, non-identifiability arises due to two factors in clustering:

1. ill-posedness at the population level,
2. finite-sample issues.

Accordingly, different kinds of assumptions are necessary to deal with them. Assumptions targeted toward addressing population-level non-identifiability specifically focus on addressing the following *conceptual* question:

What kind of clustering is desirable, given complete knowledge of the underlying data distribution?

Over decades of research on clustering, numerous assumptions have been considered to address this question. Dominant examples include assumptions based on objective functions, axiomatic approaches, mixture models, and density-based approaches.

In contrast, typical assumptions to deal with the issue of non-identifiability due to finite samples aim to address the following *statistical* question:

Given such a model of the desirable clustering, how can one find a clustering that is as close as possible if one only has access to a finite sample?

As one may expect, these assumptions are typically similar to those postulated for finite-sample non-identifiability in the i.i.d. setting, such as constraints on the complexity of the hypothesis class or data distributions. In Section 4.1, we will discuss both kinds of assumptions in further detail.

3.3.2 *Assumptions for causal learning*

Learning the causal relationships between a set of variables from observational data is one of the most challenging problems in machine learning today. A primary source of this challenge is the issue of (extreme) non-identifiability in causal learning. Accordingly, assumptions are made to address two key questions:

1. *Under what assumptions is it possible to identify causal relations at the population level?*

2. *Given an identifiable causal relationship, when can we estimate it from the data?*

We shall elaborate on the various approaches to learning causal relationships and their underlying assumptions in Section 4.2.

4
Theory of learning under extreme non-identifiability

In this section, we will discuss theoretical foundations for learning under extreme non-identifiability, emphasizing the problems of statistical clustering and causal learning.

4.1 Theory of statistical clustering

First, let us recall the problem of clustering in a statistical framework. Given a sample of n items $X = \{x_i\}_{i=1}^n$ drawn i.i.d. according to an unknown probability distribution $\mathbb{P}$ over $\mathbb{R}^d$, the goal is to find the *underlying cluster structure*. Clearly, one cannot formalize the clustering problem without further specifying the underlying or optimal clustering. In other words, one cannot formalize clustering without resolving the issue of population non-identifiability. Under a model of optimal clustering, one can ask for a *model-relative* justification of clustering methods. Conceptually, this is similar in spirit to controlling the estimation error in the setting of supervised learning. Therefore, the primary focus of learning theory for clustering rests on addressing similar questions to those posed in Chapter 2. Note, however, that unlike in the case of supervised learning where theoretical justification is sought for a general inductive principle, viz ERM, in clustering, an inductive principle of such generality does not exist beyond a small class of methods such as K-means and a myriad of its extensions. Therefore, guarantees are sought for individual clustering methods or a class of methods. See [Von Luxburg and Ben-David, 2005] for a discussion on some analogies between clustering and learning in the supervised setting.

The key goals of learning theory for clustering are to address the following questions:

1. What are the necessary and sufficient conditions under which a clustering method or a class of methods can consistently estimate the optimal clustering as the number of samples approaches infinity?

2. How fast does the clustering at finite samples converge with increasing n? What is the optimal rate of convergence?
3. Can we find algorithms with optimal predictive ability?
4. How do the computation properties of clustering methods interplay with their statistical properties?

We will now discuss a few models of the true underlying clustering considered in the literature and review some existing work that addresses these questions under the model assumptions.

4.1.1 *Objective functions*

In this framework, a model or optimal clustering is defined via a measure of the quality of a clustering. Examples in this framework include the problem of k-means clustering [MacQueen, 1967] or weighted k-means clustering, which are known to be closely related to certain graph-theoretic clustering objectives, such as the normalized cut problem [Shi and Malik, 2000, Dhillon et al., 2004]. In this section, we will use the k-means problem as an archetype of this framework.

Formally, for any distribution $\mathbb{P}$ on $\mathbb{R}^d$ and a parameter $k \in \mathbb{N}$, the goal of the k-means problem is to find k centers $C^* = \{c_1, c_2, \cdots, c_k\} \in \mathbb{R}^d$ such that the expected k-means cost under $\mathbb{P}$ is minimized, that is

$$C^* \in \operatorname*{arg\,min}_{C \in \mathbb{R}^d,\ |C|=k} \mathbb{E}_{\mathbb{P}(x)} \min_{c \in C} d(x,c), \tag{4.1}$$

where $d(\cdot,\cdot)$ is a measure of distance or dissimilarity. Under the standard squared l_2 distance on $\mathbb{R}^d$, this formulation corresponds to the standard K-means problem [MacQueen, 1967]. In this setting, the existence of C^* is guaranteed if the second moments of $\mathbb{P}$ are bounded [Linder, 2002]. However, the set of minimizers of (4.1) is not necessarily unique. Many generalizations of this problem, for instance, via other dissimilarity measures, have been studied in literature; notable among them is the problem of center-based clustering with Bergman divergences which has been extensively treated in the literature [Banerjee et al., 2005, Telgarsky and Dasgupta, 2012, Brécheteau et al., 2021, Paul et al., 2021a,b]. Under this framework, as in the supervised learning setting, a natural estimator for C^* can be obtained via *empirical cost/risk minimization*. Given a finite sample drawn according to $\mathbb{P}$, the ERM estimator is given by

$$\hat{C} \in \operatorname*{arg\,min}_{C \in \mathbb{R}^d,\ |C|=k} \frac{1}{n} \sum_{i=1}^{n} \min_{c \in C} d(x,c).$$

Consistency of ERM. Theoretical analysis for ERM estimators in this framework was initialized by the strong consistency results for the standard K-means problem with the l_2 distance. Pollard

[1981] and Abaya and Wise [1984] showed that under the mild assumption on finiteness of 2nd moment of $\mathbb{P}$, the cost of every empirical cost minimizer $\hat{C}$ converges to the cost of C^* almost surely under $\mathbb{P}$ as $n \to \infty$. More generally, Biau et al. [2008] and Levrard [2015] derived strong consistency results for the K-means problem over any separable Hilbert space. Such consistency results also exist for many related problems in this framework, for example, for sparse K-means, power K-means, and reduced K-means problems [Terada, 2014, 2015, Chakraborty and Das, 2020]. Consistency of ERM in this framework may also be characterized by the notion of algorithmic stability [Shalev-Shwartz et al., 2010].

Convergence rates for ERM In the framework of center-based clustering, there has been extensive work on deriving non-asymptotic guarantees for clustering methods [Pollard, 1981, Chou, 1994, Bartlett et al., 1998, Linder et al., 1994, Linder, 2000, 2002, Antos, 2005, Antos et al., 2005, Graf and Luschgy, 2007, Biau et al., 2008, Telgarsky and Dasgupta, 2013, Bachem et al., 2017, Brécheteau et al., 2021]. This line of work is strongly inspired by empirical process theory and uniform convergence guarantees for ERM in supervised learning. Accordingly, typical bounds for ERM in this framework take the following form. Given a set of k centers $C = \{c_1, c_2, \cdots, c_k\}$, let $R(C)$ denote the expected cost (or distortion) of C under $\mathbb{P}$ given by

$$R(C) = \mathbb{E}_{\mathbb{P}(x)} \min_{c \in C} d(x, c).$$

Analogously, let $\hat{R}(C)$ denote the empirical counterpart of $R(C)$ given by

$$\hat{R}(C) = \frac{1}{n} \sum_{i=1}^{n} \min_{c \in C} d(x, c).$$

For instance, Biau et al. [2008] showed that for any distribution $\mathbb{P}$ over a bounded domain in a separable Hilbert space such that $\|x\| \leq M$ for some $M < \infty$, there exists a constant C such that, for any $\delta \in (0, 1)$, with probability at least $(1 - \delta)$,

$$R(C^*) - R(\hat{C}) \leq CM^2 \sqrt{\frac{k^2 + \log \frac{1}{\delta}}{n}} \tag{4.2}$$

These bounds are optimal with respect to n in the minimax sense without additional assumptions [Antos, 2005]. More recently, for distributions with bounded support, nearly optimal upper bounds with respect to both n and k have also been derived in [Liu, 2021]. Uniform convergence bounds for ERM under less restrictive moment assumptions with a rate of $\mathcal{O}(1/\sqrt{n})$ have also been derived in this framework [Klochkov et al., 2021]. Better convergence rates (so-called "fast rates") scaling as $\mathcal{O}(1/n)$ can be achieved under additional assumptions, for example, for distributions with finite support or distributions with bounded support satisfying certain regularity conditions.

Improved estimation. There has been a multitude of methods that aimed to derive more robust estimators for center-based clustering. Most of these estimators are motivated from a practical standpoint, and only a few admit theoretical justification. Notable among them is the class of so-called *medians of means estimators* (MOM) provides an outlier-robust class of estimators for center-based clustering and admits strong theoretical guarantees similar to the ERM. For a nice overview of MOM estimation for center-based clustering with Bergman divergence, refer to Paul et al. [2021b]. For example, Klochkov et al. [2021] provides uniform convergence bounds scaling at a rate of $\mathcal{O}(\sqrt{k/n})$ for MOM estimators for center-based clustering in any separable Hilbert space under the mild assumption of the existence of 2nd order moments of the underlying distributions. These bounds are minimax optimal both with respect to k and n.

Statistical Computational Trade-offs. So far, we have only discussed the statistical properties of ERM and MOM estimators in the framework of center-based clustering. In the worst case, the ERM and the MOM estimators (which achieve optimal convergence rates) are NP-Hard. Numerous approximations have been proposed to address the computational complexity of ERM estimators. However, the theoretical understanding of these algorithms in a statistical framework is very sparse. Nystrom approximations [Calandriello and Rosasco, 2018] or Gaussian approximations [Biau et al., 2008] attempt to reduce the computational complexity of center-based clustering by constraining the underlying hypothesis class. However, in general, ERM over these classes is still NP-Hard. The most common approach to analyzing such approximations is by deriving guarantees on the approximation ratio, which is the ratio of the cost of the clustering obtained by the approximation to that obtained via the target estimator, such as ERM (see [Wang et al., 2019b] and references therein for a sample result of this kind).

4.1.2 *Density-based clustering*

The paradigm of density-based clustering is another widely popular approach to systematically resolve the ambiguity of "what is a good clustering?". In this framework, clusters are defined as *regions of high density separated by low-density regions* [Hartigan, 1975, 1981]. Assuming that n i.i.d. samples $X = \{x_1, x_2, \cdots, x_n\}$ from a distribution $\mathbb{P}$ with density f over $\mathbb{R}^d$ are observed, two distinct problems have been studied under this framework:

1. **Flat clustering.** Estimate the connected components of the upper-level set given by $\left\{x \in \mathbb{R}^d : f(x) \geq \lambda\right\}$, for a specified $\lambda \geq 0$.
2. **Cluster tree estimation.** Estimating the *cluster tree* of f, that is, a map $C : \mathbb{R} \to \mathbb{C}(\mathbb{R}^d)$, given by $C(\lambda) = \left\{x \in \mathbb{R}^d : f(x) \geq \lambda\right\}$,

where $\mathbb{C}(\mathbb{R}^d)$ denotes the set of all partitions of subsets of $\mathbb{R}^d$.

For both problems, a natural approach to estimating the level sets and their connected components based on X is to estimate the density f using some non-parametric approach such as kernel density estimators and then obtain the level sets (and their connected components) of the estimated density f_n. In a certain sense, this *plugin* estimator inherits the statistical properties of the density estimator f_n.

Consistency and convergence rates of the plugin estimator have been extensively investigated for the flat clustering problem [Tsybakov, 1997, Cuevas and Fraiman, 1997, Klemelä, 2004, Rigollet and Vert, 2009, Rinaldo et al., 2010]. However, this approach suffers from certain limitations. First, it is unclear how one may choose an appropriate λ since different choices of λ can create ambiguity in the notion of the *true clustering*. Second, even for simple univariate distributions, there may not exist a single λ for which the connected components of the level set do not correspond to visually distinct clusters (see Rinaldo et al. [2012, Figure 1] for an illustration). Furthermore, from an algorithmic point of view, estimating the level sets of typical densities used for estimation is extremely complicated, thereby limiting the practical utility of such estimators [Sriperumbudur and Steinwart, 2012, Chaudhuri et al., 2014].

The framework of cluster tree estimation is proposed as an alternative to mitigate some of these limitations of the flat clustering problem. Due to the algorithmic difficulties of the plugin approach discussed earlier, recent work focuses on algorithms that directly estimate the cluster tree. Hartigan [1981] showed that the single linkage algorithm satisfies a weaker notion of consistency called *fractional consistency*, which eliminates spurious clusters attained due to finite sample effects. Chaudhuri et al. [2014] show that a modified single linkage algorithm and an estimator based on the k-nearest neighbor graph also achieve consistency in the weak sense that depends on a separation criterion between the clusters under a mild assumption of uniform continuity of the density. They provide optimal convergence rates in the minimax sense with respect to the cluster separation criterion. These results were extended by Balakrishnan et al. [2013] to distributions over manifolds satisfying certain regularity conditions.

The DBSCAN algorithm [Ester et al., 1996] and its modifications are some of the most popular clustering methods for practitioners in the framework of density-based clustering. Despite being a rather popular practical approach, statistical properties of DBSCAN have only recently been analyzed in Sriperumbudur and Steinwart [2012]. For $\alpha-$Hölder continuous densities, Sriperumbudur et al. [2010] adaptively estimate the first split λ^* that yields two connected components in its density level set and provide consistency as well as optimal convergence rates under this setting. For smooth densities, Wang et al. [2019a] derived consistency and nearly optimal rates for cluster tree estimation. Further extensions of these analyses to a general class of estimators based on kernel density estimation can be found in Steinwart et al. [2019].

Statistical-computational tradeoffs. The computational complexity of DBSCAN is known to be $\mathcal{O}(n^2)$ in the worst case [Gan and Tao, 2015]. A lot of work has been done on finding approximations that run in sub-quadratic or almost linear time based, for example, on approximate nearest neighbors [Huang and Bian, 2009, Kumar and Reddy, 2016], subsampling [Liu, 2006, Viswanath and Babu, 2009], or parallel computation. However, most of these methods are based on heuristics and lack statistical guarantees. The only algorithms with statistical guarantees are provided in Jang and Jiang [2019] and Esfandiari et al. [2021]. Jang and Jiang [2019] provide a modification of the DBSCAN algorithm called DBSCAN++, which retains optimal statistical rates while achieving a sub-quadratic computational complexity. They also demonstrate that DBSCAN++ exhibits a general tradeoff between computational complexity and convergence rates. Esfandiari et al. [2021] improved these results to obtain an almost linear time approximation to DBSCAN while retaining the optimal convergence rates.

4.1.3 *Axiomatic approaches*

Another principled approach to evaluating clustering methods is the so-called axiomatic approach [Kleinberg, 2002, Ben-David and Ackerman, 2008, Zadeh and Ben-David, 2012]. Here, unlike the approaches discussed so far, an explicit target/population clustering is not specified. Instead, a list of axioms or desirable properties is postulated, and the extent to which different clustering methods satisfy them is investigated. Kleinberg [2002] postulated three conditions one may expect clustering functions to satisfy. Under these axioms, Kleinberg famously derived an impossibility result, proving that no clustering method can satisfy all three criteria simultaneously. However, it has been argued that the impossibility result is merely an artifact of the specific formalism invoked in Kleinberg's work [Ben-David and Ackerman, 2008]. For example, Zadeh and Ben-David [2012] modified the criterion and derived a positive result under the modified axioms. In particular, they showed that the single linkage algorithm is consistent with the three modified axioms simultaneously.

4.1.4 *Mixture models*

Another systematic approach to evaluate clustering algorithms can be found in the so-called planted models, where the goal is to obtain *recovery guarantees* under distributional assumptions [Dasgupta, 1999, Yan and Sarkar, 2016, Banks et al., 2018, Vankadara and Ghoshdastidar, 2020, Vankadara et al., 2021a]. For example, under the assumption that a mixture of distributions generates the data, the goal is to recover the membership of each item in a finite sample. As we already discussed in the previous chapter, population-level identifiability persists for general mixture models without further assumptions. Mixture models are identifiable under certain *parametric assumptions* such as Gaussiantiy of the components [Bruni and Koch, 1985, Teicher, 1963]. Accordingly, recovery guarantees under parametric assumptions (typically Gaussianity) have been extensively studied in the clustering literature. Dasgupta [1999] presented a clustering method that can provably learn a mixture of high-dimensional Gaussians. Recent work in this direction focuses on statistical-computational tradeoffs or *phase transitions* in clustering a mixture of high-dimensional Gaussians [Banks et al., 2018, Ashtiani et al., 2018]. More recently, the issue of population-level identifiability has also been addressed under non-parametric assumptions such as separability, independence of marginals, or by incorporating certain symmetries. The framework of obtaining recovery guarantees under parametric as well as non-parametric assumptions is a central focus of this thesis. We discuss this approach in further detail in Chapters 6 and 7.

Despite decades of research in the theory of clustering, existing guarantees for clustering are far from satisfactory. We will discuss some shortcomings of the current state of clustering theory, which partly motivates the contributions of this thesis, in Chapter 5.

4.2 *Theory of causal learning*

Causal learning is the process of using observations to infer cause-effect relationships. The study of causality has a rich history, dating back to ancient philosophers such as Aristotle, who made key contributions to our understanding of causality. In the centuries that followed, many other philosophers and scientists continued to develop and refine our understanding of causality, including figures such as John Stuart Mill and David Hume. However, the formal study of causal learning is relatively recent. In the 20th century, the field of causal learning underwent significant developments, with the advent of modern statistical methods and the development of new mathematical frameworks for representing and reasoning about causal relationships.

4.2.1 Modeling causal relationshps

There are several ways to model causal relationships among a set of variables. One popular approach is to use the framework of *causal graphical models (CGM)* [Pearl, 2009], which represent variables as nodes of a directed graph, and *direct cause-effect* relationships between the variables as directed edges. This powerful framework enables us to reason about the effects of interventions or manipulations on a subset of the variables. Here, we primarily focus on a subclass of causal graphical models called *structural causal models (SCM)* [Peters et al., 2017].

Given a set of variables $\{x_1, x_2, \cdots, x_n\}$ equipped with a directed acyclic graph G, SCMs define each variable x_i as a result of a deterministic function of its parents in G, denoted by $Pa(x_i) \subset \{x_1, x_2, \cdots, x_n\}$ and a noise variable u_i:

$$x_i := f_i(Pa(x_i), u_i), \quad (i = 1, 2, \cdots, n).$$

The (random) noise variables $\{u_1, \cdots, u_n\}$ are assumed to be jointly independent. An SCM induces a joint distribution over the variables, which satisfies the *causal Markov condition*. This condition states that each variable x_i is independent of its non-descendants in G conditioned on its parents. Equivalently, the joint distribution satisfies all the conditional independence relations that are implied by a (global) graphical criterion called d-separation, which is based on a notion of "path blocking". Note that different graphs can satisfy the same d-separation criteria and therefore enforce the same set of conditional independence relationships. The class of all such graphs is called the Markov equivalence class [Pearl, 2009].

This framework provides a natural means to formalize the notion of interventions or manipulations which are fundamental to the problem of causal learning. Many different kinds of interventions are considered in causal learning. Hard interventions on a variable x_i correspond to modifying the structural equation f_i and setting it to a fixed value c. From a graphical point of view, this corresponds to severing all the incoming edges into x_i [Peters et al., 2017]. Interventions induce a shift in the underlying distribution — which is the joint distribution induced by the corresponding set of structural equations. This is formally denoted using the *do* operator as $do(x := c)$. Many other kinds of interventions, such as soft or randomized interventions, are also considered in the literature [Vankadara et al., 2021b].

4.2.2 Causal discovery

A majority of research in causal learning is focused on the problem of *causal discovery*, which is the process of inferring the (qualitative) causal relationships between the variables. In the SCM or DCGM framework, the goal of causal discovery is to recover the underlying causal graph from observational data. This can be formalized as a problem of statistical estimation of parameters that describe the causal graph [Glymour et al., 2019]. Like any statistical estimation problem, there are two key aspects to consider: *identifiability* and *estimation*. Once the parameters of interest are correctly identified, the problem is reduced to a statistical one. Therefore, we can also ask the same theoretical questions about consistency or convergence rates that we discussed in the context of statistical learning and clustering.

While the causal graph is not identifiable from observational data in general, under suitable assumptions such as *faithfulness*[1], it is possible to recover the underlying causal graph up to its Markov equivalence class (MEC) [Pearl, 2009]. For example, under the assumption of faithfulness and absence of latent confounders, the PC algorithm can be shown to consistently recover the causal graph asymptotically with the number of observed samples [Spirtes et al., 2000]. Subject to faithfulness, the FCI algorithm is known to be consistent even in the presence of latent confounders [Spirtes et al., 2000]. These methods and many extensions rely on conditional independence tests to prune the edges of a fully connected causal DAG. They are typically referred to as *constraint-based methods*. In general, constraint-based approaches cannot provide finite sample bounds since they are only equipped with "pointwise" (and not uniform) consistency guarantees [Glymour et al., 2019].

While these approaches are rather powerful and make relatively few assumptions, there are subject to several limitations. Conditional independence testing is a notoriously difficult problem at finite samples and often requires additional assumptions [Shah and Peters, 2020]. Moreover, they cannot identify the causal graph within a Markov equivalence class. For example, when there are only two observable quantities[2], conditional independence tests cannot determine cause-effect relationships between them.

An interesting approach to dealing with these limitations is to impose constraints on the class of functions (f_i) in the SCM. The main intuition behind this idea is that these structural functions leave a footprint in the observational distribution, which can be exploited to recover the causal structure [Peters et al., 2017]. This bears a resemblance to the standard setting of statistical learning, where constraints on the complexity of the hypothesis class are crucial to derive finite-sample generalization bounds. For example, in the absence of latent confounders, if the structural equations are constrained to be linear, and the noise variables are assumed to be non-Gaussian, the causal structure can be shown to be identifiable [Shimizu et al., 2006]. These models, commonly known as LiNGAM models, can be consistently and (computationally) efficiently estimated from observational data even in the two variable settings [Shimizu et al., 2011]. Several generalizations of these results under relaxed assumptions have been obtained.

[1] While the causal DAG implies certain independence properties in the joint distribution, the distribution may satisfy additional independence properties not implied by the DAG. Faithfulness assumes that the distribution only satisfies independence properties implied by the graph.

[2] Learning cause-effect relationships between two variables is a rather important problem in causal learning from a practical point of view.

The post-non-linear causal models (PNL) have the most general form in the SCM framework, under which it can be shown that the causal directions are identifiable [Zhang and Hyvarinen, 2012]. These methods are often used in conjunction with constraint-based methods to refine the edges in the Markov equivalence class [Zhang and Hyvarinen, 2012]. Identifiability of the causal structure is a crucial problem in SCM-based learning, and most of the theory in this setting is focused on deriving conditions that guarantee identifiability. However, to the best of our knowledge, there is hardly any work on finite sample analysis (for example, sample complexity bounds) for causal discovery.

4.2.3 *Causal inference*

While the problem of causal discovery is concerned with recovering the underlying causal structure among a set of variables, causal inference or reasoning aims to quantify causal influences given the causal structure. Identification and estimation are again the central aspects to consider. The typical setup of causal inference considers a (potentially multi-dimensional) causal driver x influencing a target variable y, and the goal is to estimate the effect of interventions on x on the target variable or outcome y. In the SCM framework, this often amounts to estimating the structural functions. The potential outcomes (PO) framework [Rubin, 1974, Neyman, 1923] is another popular framework in the study of causal inference. In this framework, causal effects are described as *treatment effects*, which evaluates the differences in the outcome under two contrasting interventions [Rosenbaum and Rubin, 1983, Imbens and Rubin, 2015]. The causal driver also called the treatment variable, is often assumed to be binary, representing the treatment or control groups.

Identification and estimation are the two critical aspects of the theory of causal inference. In the theory of identification, the central question that is asked is: *Under what conditions can causal effects be precisely identified?* If causal effects can be accurately identified, the task of estimating causal effects becomes a statistical estimation problem. Randomized controlled trials (RCTs) are considered the gold standard for identifying and estimating causal effects. In an RCT, participants are randomly assigned to either the treatment or control groups. The treatment group receives the intervention/treatment, while the control group receives either a placebo or no intervention. By comparing the outcomes of the two groups, one can estimate the causal effect of the treatment on the outcome. When RCTs are infeasible, one seeks to assess causal effects from observational data. Without additional assumptions, causal effects are not identifiable purely from observational data.

A common assumption in the causal inference literature is the absence of latent confounding variables. Under this assumption, the *do-calculus* provides a complete set of rules that allow the identification of causal effects from the observational distribution [Robins, 1986, Pearl, 2009]. Estimating causal effects in this setting is feasible under the so-called *strong ignorability* condition. There are several ways to estimate the causal/treatment effects in this setting. Common approaches include regression-based adjustment, matching, and re-weighting methods [Rosenbaum, 2002, Abadie and Imbens, 2006, Van Der Laan and Rubin, 2006, Robins et al., 2017]. Typical guarantees for these approaches focus on asymptotic consistency and convergence rates in low-dimensional settings. However, these methods are known to be prone to biases in high-dimensional regimes or under strong confounding [Belloni et al., 2014, 2017, Chernozhukov et al., 2018].

Doubly robust methods typically use a combination of methods (for example, propensity-score methods and regression methods) to mitigate the bias of, say, regression-based adjustments [Belloni et al., 2014, 2017, Chernozhukov et al., 2018]. Assuming the existence of a consistent estimator for propensity score, these approaches derive $\sqrt{n}$ consistency guarantees for treatment effect estimation in high dimensions. There is little emphasis in this literature on obtaining small sample convergence rates. Notable exceptions include [Shalit et al., 2017], which provides finite-sample generalization error bounds for estimating individual treatment effects.

In the presence of latent confounders, the problem of estimating treatment effects is incredibly hard and requires additional information or stronger assumptions. Approaches based on *instrumental variables* are some of the most popular methods in this setting. Many algorithms have been proposed for estimating causal effects in the presence of instrumental variables [Newey and Powell, 2003, Chen and Christensen, 2018, Singh et al., 2019, Muandet et al., 2020]. These methods admit strong theoretical guarantees that hold at finite samples. For a review of these results, we refer the reader to Singh et al. [2019], Muandet et al. [2016].

5
Thesis contributions

5.1 Guarantees for kernel-based clustering

5.1.1 What is missing?

A majority of the theoretical analysis of clustering focuses on frameworks based on objective functions, such as center-based clustering. As discussed earlier, analyses in this framework define the underlying clustering as the one that consistently partitions the data domain. Such objective functions are typically ad hoc and have no generality beyond specific applications such as vector quantization. To illustrate, consider the following example. Kernel k-means is one of the most popular algorithms for clustering among practitioners. Consistency guarantees and corresponding convergence rates exist for kernel k-means (see, for example, [Biau et al., 2008]). However, the target clustering strongly depends on the choice of the kernel function. If one adopts a non-informative kernel, such as the identity kernel, arbitrary partitions of the data domain may be optimal. While statistical consistency and corresponding convergence guarantees are necessary, they are not sufficient. It is crucial to consider models of clustering that are practically relevant and sufficiently general while being amenable to theoretical analysis.

Density-based clustering provides one of the more promising formalisms of clustering. Algorithms motivated by this general framework, such as DBSCAN and its extensions, are practically popular and admit strong theoretical guarantees. However, they are still subject to several limitations. They are known to fail when clusters widely differ in their densities. Furthermore, similar to density estimation, they suffer from the curse of dimensionality. Generally, no single formalization of the true/underlying clustering suffices for every practical application. There has been a long discussion in the clustering literature on the need to develop a taxonomy of clustering frameworks and methods [Von Luxburg and Ben-David, 2005, Von Luxburg et al., 2012]. This clearly emphasizes the need for evaluating clustering algorithms under sufficient general and practically meaningful formalisms.

5.1.2 Contributions

Kernel-based clustering methods such as kernel k-means or spectral clustering are some of the most popular clustering methods among practitioners. However, theoretical guarantees for these methods under a general formalism of clustering are surprisingly limited. In this thesis, we take a step toward addressing this gap between theory and practice. We consider the fairly general framework of mixture models to define a notion of true clustering. Under this framework, we focus on addressing the following questions under parametric or non-parametric assumptions on the mixture distribution:

1. What are the necessary and sufficient conditions under which kernel clustering algorithms can consistently recover the underlying clustering?
2. How do the statistical properties interplay with computational properties?

To this end, our work results in the following contributions:

(a) We consider the problem of clustering a mixture of Gaussians in the high-dimensional regime. Under this formulation, we study *phase transitions* – sharp information-theoretic thresholds below which no algorithm can, provably, recover the true clustering. We show that kernel k-means and an efficient semi-definite relaxation are both information-theoretically (near) optimal in their respective computational classes (Chapter 6).

(b) We study the statistical performance of kernel clustering algorithms under general non-parametric conditions on the mixture models. We provide necessary and sufficient *separability* conditions under which these algorithms can consistently recover the true clustering (Chapter 7).

5.2 Contributions to the theory of causal learning

The investigation of how to identify causal relationships using only observational data is a relatively recent area of study. Guarantees that hold at finite samples are generally scarce, particularly in high-dimensional settings or in time series settings, and are limited to simple models. Although there has been a surge in research on developing theoretical foundations for causal inference, we are far from a complete understanding.

In this thesis, we attempt to improve our understanding of the problem of causal learning from both a theoretical and conceptual standpoint. Our results are motivated by Janzing [2019] which formally establishes a close analogy between "generalizing from empirical to observational distributions" and "generalizing from observational to interventional distributions" albeit for highly constructed statistical and causal models. Intuitively speaking, the analogy suggests that the bias due to observing small sample sizes is qualitatively similar to the bias due to shifts in distributions. Such observations have also been made in the distributionally robust learning literature [Zhu et al., 2020]. An immediate consequence of this observation is the following conjecture:

> Can techniques for learning models with good out-of-sample generalization performance (e.g., regularization) also help learn models with good out-of-distribution generalization and vice-versa?[1]

A preliminary and positive answer to this question is indicated in Janzing (2019), which suggests that standard norm-based regularization techniques typically recommended for better statistical learning may also help learn better causal models. Furthermore, distributionally robust optimization approaches are increasingly being utilized to learn predictive models with good out-of-sample generalization performance when few samples are available [Zhu et al., 2020].

More generally, similar to the setting of statistical learning where uniform convergence bounds are sought for general hypothesis classes, one can ask:

> When can we obtain guarantees of *causal generalization* — generalizing from the observational to interventional distributions — that hold uniformly over a hypothesis class under restrictions on the complexity of the class?

To this end, the following contributions are made by this thesis:

(a) We introduce the framework of *causal learning theory* for forecasting. Using this framework, we obtain a characterization of the difference between statistical and *causal risks*, which helps identify sources of divergence between them. When no hidden confounders are present, the problem of causal generalization amounts to learning under covariate shifts, albeit with additional structure (restriction to interventional distributions under the vector autoregressive models(VAR). This structure allows us to obtain uniform convergence bounds on causal generalizability for the class of vector autoregressive models. To the best of our knowledge, this is the first work that provides theoretical guarantees for causal generalization in the time-series setting (Chapter 8).

(b) A large volume of recent work shows that in complex model classes, interpolators can achieve statistical generalization and even be optimal for statistical learning. As discussed in Chapter 2, uniform convergence fails to explain the generalization

[1] The term out-of-distribution generalization is used a bit informally to refer to generalization from the observational to the class of interventional distributions

properties of interpolators in high-dimensional and overparameterized settings. Despite increasing interest in learning models with good causal properties, there is no understanding of whether such interpolators can also achieve *causal generalization*. To address this gap, we study causal learning from observational data through the lens of interpolation and its counterpart—regularization under a linear causal model (Chapter 9).

5.3 *Thesis overview*

This thesis is based on the following publications.

1. (Chapter 6) **Leena C Vankadara**, Debarghya Ghoshdastidar. On the optimality of kernels for high-dimensional clustering. *In Artificial Intelligence and Statistics (2020).*

2. (Chapter 7) **Leena C Vankadara**, Sebastian Brodt, Ulrike von Luxburg, Debarghya Ghoshdastidar. Recovery Guarantees for Kernel-based Clustering under Non-parametric Mixture Models. *In Artificial Intelligence and Statistics (2021).* **This paper received an oral presentation which is awarded to 3% of all the submissions).**

3. (Chapter 8) **Leena C Vankadara**, Philipp M Faller, Michaela Hardt, Lenon Minorics, Debarghya Ghoshdastidar, Dominik Janzing. Causal Forecasting: Generalization Bounds for Autoregressive Models. *In Uncertainity of Artificial Intelligence (2022).*

4. (Chapter 9) **Leena C Vankadara**, Luca Rendsburg, Ulrike von Luxburg, Debharghya Ghosdastidar. Interpolation and Regularization for Causal Learning. *In Neural Information Processing Systems (2022).*

In addition to these papers, I published the following papers during the course of my PhD.

While I am the leading author for this paper, I did not include it in my thesis since it does not fit the theme of this thesis.

5. **Leena C Vankadara**, Siavash Haghiri, Michael Lohaus, Faiz Ul Wahab, Ulrike von Luxburg, Insights into ordinal embedding algorithms: a systematic evaluation. *In Journal of Machine Learning Research (2023).*

I am one of the two primary contributors to the following paper.

6. Luca Rendsburg, **Leena C Vankadara**, Debharghya Ghosdastidar, Ulrike von Luxburg, A Consistent Estimator for Confounding Strength. *A preprint (2022).*

The following paper was published after starting my Ph.D. but was completed during my Master thesis and a research internship that followed it.

7. **Leena C Vankadara**, Ulrike von Luxburg. Measures of distortion for ML. *In Neural Information Processing Systems (2018).*

In addition, I contributed to the following publications.

8. Mahalakshmi Sabanayagam, **Leena C Vankadara**, Debarghya Ghoshdastidar. Consistency of Clustering and Two-sample Testing of Graphons. *In International Conference for Learning Representations (2022).*

9. Maximilian Fleissner, **Leena C Vankadara**, Debharghya Ghosdastidar, Explainability of Kernel Clustering. *A preprint (2022).*

10. Pascal Esser, **Leena C Vankadara**, Debarghya Ghoshdastidar. Learning Theory Can (Sometimes) Explain Generalisation in Graph Neural Networks. *In Neural Information Processing Systems (2021).*

Part II

Publications

6
Optimality of kernels for high dimensional clustering

On the optimality of kernels for high-dimensional clustering

Leena Chennuru Vankadara
University of Tübingen, IMPRS-IS

Debarghya Ghoshdastidar
Technical University of Munich

Abstract

This paper studies the optimality of kernel methods in high-dimensional data clustering. Recent works have studied the large sample performance of kernel clustering in the high-dimensional regime, where Euclidean distance becomes less informative. However, it is unknown whether popular methods, such as kernel k-means, are optimal in this regime. We consider the problem of high-dimensional Gaussian clustering and show that, for a class of dot-product kernels, the sufficient conditions for partial recovery of clusters using the NP-hard kernel k-means objective matches the known information-theoretic limit up to a factor of $\sqrt{2}$ for large k. It also exactly matches the known upper bounds for the non-kernel setting. We also show that a semi-definite relaxation of the kernel k-means procedure matches upto constant factors, the spectral threshold, below which no polynomial time algorithm is known to succeed. This is the first work that provides such optimality guarantees for the kernel k-means as well as its convex relaxation. Our proofs demonstrate the utility of the less known polynomial concentration results for random variables with exponentially decaying tails in higher-order analysis of kernel methods.

1 Introduction

Kernel methods are one of the most empirically successful class of machine learning techniques. While being easy to implement, kernel methods are well known to improve empirical performance of algorithms and are also related to other successful machine learning principles such as Gaussian process and neural networks (Kanagawa et al., 2018; Jacot, Gabriel, and Hongler, 2018). At the heart of kernel based learning lies the *kernel trick* which implicitly maps the data to a high, possibly infinite, dimensional *reproducing kernel Hilbert space* (RKHS), and hence, induces non-linearity into classical linear learning models such as support vector machines, principle component analysis or k-means. Kernel methods are based on a solid theoretical foundation, which makes them conducive to theoretical analysis. There has been considerable theoretical research on kernel based supervised learning from a statistical perspective (Steinwart and Christmann, 2008; Mendelson and Neeman, 2010), and to some extent, in the context of semi-supervised learning (Wasserman and Lafferty, 2008; Mai and Couillet, 2018). Perhaps surprisingly, much less is known about the statistical performance of kernel methods beyond such settings, for instance, kernel based clustering.

A long-standing issue in the theoretical study of clustering, and also kernel based clustering, has been the lack of a universally accepted notion of *goodness* of clustering. A popular definition of good clustering is one that consistently or near-optimally partitions the data domain. Based on this perspective, there exist *approximation guarantees* for solving kernel based cost functions (S. Wang, Gittens, and Mahoney, 2019) and *consistency results* showing that the clustering asymptotically approaches a limiting clustering (Luxburg, Belkin, and Bousquet, 2008). In such analyses, the optimal cost function is inherently tied to the chosen kernel and hence can be arbitrarily far from the "ground truth." For instance, even an arbitrary clustering can be optimal (can achieve maximal clustering objective) for trivial kernels such as constant or identity kernels. Another approach to measure the performance of a clustering algorithm is by establishing recovery guarantees under distributional assumptions, sometimes known as *planted models.* Distributional assumptions, or specifically (sub)-Gaussian mixture model assumption, is often considered in the theory of clustering. While learning a mixture of Gaussians has always been an important research problem, Dasgupta (1999), for the time, presented a provable clustering algorithm to

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

learn a mixture of *high-dimensional* Gaussians. Theoretical research on learning high-dimensional Gaussians have ever since been highly significant, owing to the ubiquity of high-dimensional data in practice. Recent works in this direction provide *phase transitions* for both clustering and parameter estimation of a mixture of high-dimensional Gaussians (Banks et al., 2018; Ashtiani et al., 2018).

Couillet and Benaych-Georges (2016) initiated the theoretical study of kernel methods for high-dimensional Gaussian clustering, and in particular, presented the large sample behaviour of kernel spectral clustering in the regime where number of samples grow linearly with the data dimension. The statistical difficulty in this regime stems from the fact the Euclidean distance tends to be less informative in high dimensions and intra-cluster distances could be systematically larger than inter-cluster distances. Yan and Sarkar (2016) generalised the problem setup to sub-Gaussian mixtures and derived sufficient conditions for achieving zero clustering error using convex relaxations of the kernel k-means objective. In both works, the analysis is restricted to computationally efficient clustering algorithms and the optimality of kernel methods, in terms of comparing necessary and sufficient conditions for clustering, is not addressed.

In this paper, we study the phase transitions — sharp information-theoretic thresholds below which no algorithm can, provably, recover the true clustering better than chance — of the high-dimensional Gaussian clustering problem. Our setting is similar to Banks et al. (2018), where the number of samples is linear in the problem dimension. However, we focus on the case where one has access to only a kernel matrix. In other words, while the information-theoretic thresholds inherent to the Gaussian clustering problem are expected to remain unchanged in the kernel setting with a non-trivial kernel, we prove that one can nearly achieve such thresholds using popular kernel methods. The **main contributions** in this paper are the following: **(1)** We identify the smallest separation between the means of latent clusters such that the clusters are statistically distinguishable under a kernel k-means objective in the sense of partial recovery, that is, error smaller than random guessing. Our result **matches the phase transition** for high-dimensional Gaussian clustering without kernels (Banks et al., 2018). **(2)** We analyse a common **semi-definite relaxation** of the kernel k-means objective and present sufficient conditions for partial recovery that **match, up to constant factors, the known spectral threshold** — akin to the Kesten-Stigum threshold in the community detection under stochastic block model literature (Baik, Arous, and Péché, 2005; Paul, 2007).

Our main results obtained from the analysis of the two kernel-based clustering algorithms and the best known results for the same problem in a non-kernel setup are summarized in the table below. k is the number of clusters, and α is the ratio of sample size to the data dimension which remains asymptotically finite in our setting.

The lower and the upper bounds are on the minimum separation of the clusters required to achieve partial recovery. The first column contains the bounds for the information-theoretic threshold. The second column contains the bounds corresponding to the computational class of poly-time algorithms.

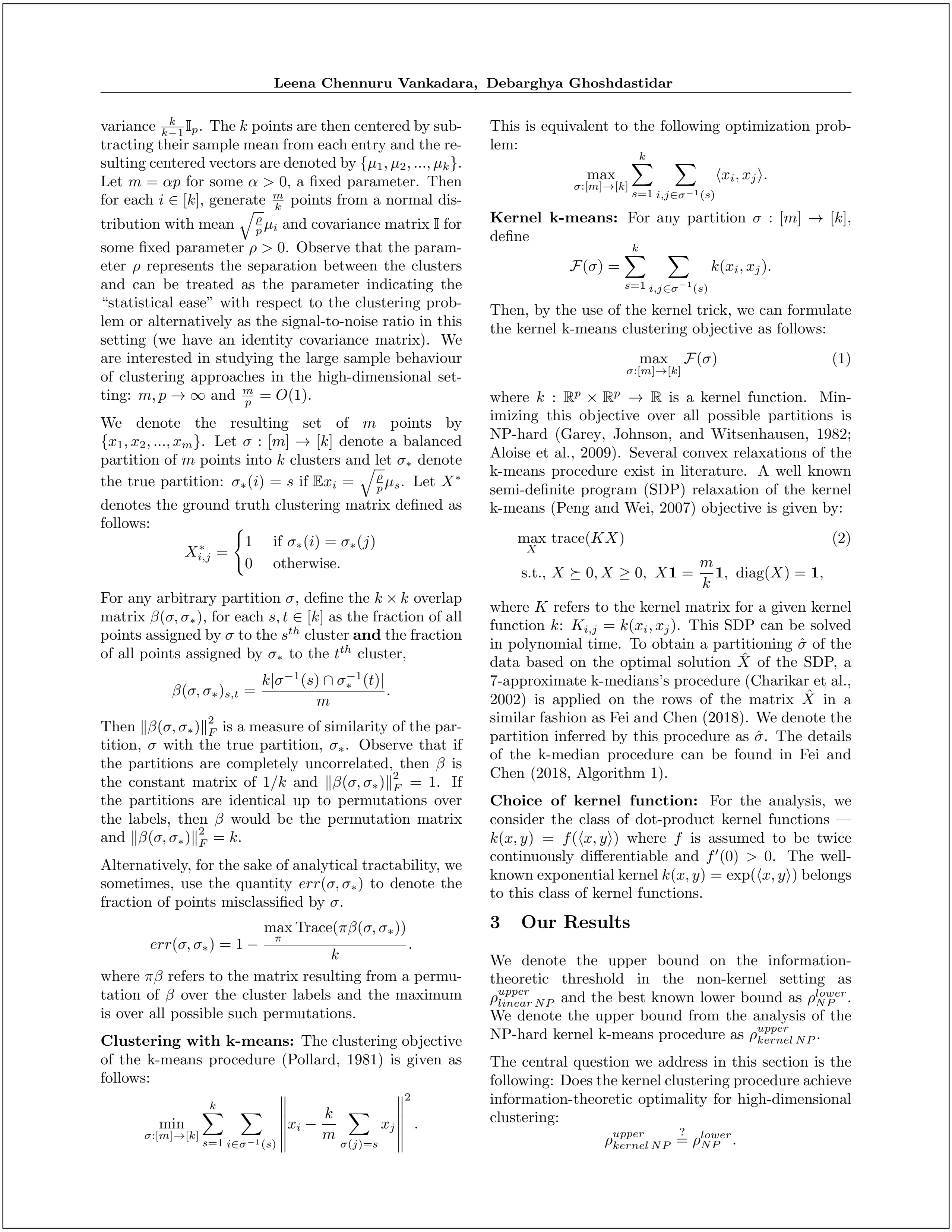
	Information-theoretic limit	**Poly-time solvable**
Lower bounds	$\sqrt{\frac{2(k-1)\log(k-1)}{\alpha}}$	$\frac{k-1}{\sqrt{\alpha}}$
Upper bounds (non-kernel)	$2\sqrt{\frac{k\log k}{\alpha}}+2\log k$	$O(k-1\vee\frac{k-1}{\sqrt{\alpha}})$
Upper bounds (kernel)	$2\sqrt{\frac{k\log k}{\alpha}}+2\log k$	$O(k\vee\frac{k}{\sqrt{\alpha}})$

As noted in Couillet and Benaych-Georges (2016), one requires a second-order analysis since first-order approximation of the kernel function does not suffice for the analysis in the high-dimensional setting. To this end, our proofs show that recent polynomial concentration inequalities (Götze, Sambale, and Sinulis, 2019) can be useful for second-order analysis of kernel methods.

2 Background and Setting

Notation: We denote the set of natural numbers $\{1, 2, \ldots, k\}$ by $[k]$. For any matrix A, $\|A\|_F$ refers to the Frobenius norm of the matrix. For any vector x, $\|x\|$ and $\|x\|_1$ refer to the Euclidean and l_1 norms of the vector. $\mathbb{I}$ denotes the identity matrix. For any $A \in \mathbb{R}^{m\times m}$, $\|A\|_{\infty\to 1}$ refers to the $\infty \to 1$ operator norm and defined as $\sup_{y,z\in\{\pm 1\}^m}(y^TAz)$. For any n real numbers $\{a_i\}_{i=1}^n$, $(a_1 \vee a_2 \ldots \vee a_n)$ refers to the maximum of the sequence:$\max_i a_i$. For any random variable x, $\mathbb{E}x$ denotes the expectation of x.

Setting: Our setting is akin to the one used in Banks et al. (2018), specifically due to the existence of a near-optimal phase transition for the information-theoretic threshold in the setting. We assume that the data is generated according to the following process. Let k be the number of clusters. Then, k points $\{\mu_1', \mu_2', ..., \mu_k'\} \in \mathbb{R}^p$ are generated independently according to a normal distribution with mean 0 and co-



variance $\frac{k}{k-1}\mathbb{I}_p$. The k points are then centered by subtracting their sample mean from each entry and the resulting centered vectors are denoted by $\{\mu_1, \mu_2, ..., \mu_k\}$. Let $m = \alpha p$ for some $\alpha > 0$, a fixed parameter. Then for each $i \in [k]$, generate $\frac{m}{k}$ points from a normal distribution with mean $\sqrt{\frac{\rho}{p}}\mu_i$ and covariance matrix $\mathbb{I}$ for some fixed parameter $\rho > 0$. Observe that the parameter ρ represents the separation between the clusters and can be treated as the parameter indicating the "statistical ease" with respect to the clustering problem or alternatively as the signal-to-noise ratio in this setting (we have an identity covariance matrix). We are interested in studying the large sample behaviour of clustering approaches in the high-dimensional setting: $m, p \to \infty$ and $\frac{m}{p} = O(1)$.

We denote the resulting set of m points by $\{x_1, x_2, ..., x_m\}$. Let $\sigma : [m] \to [k]$ denote a balanced partition of m points into k clusters and let σ_* denote the true partition: $\sigma_*(i) = s$ if $\mathbb{E}x_i = \sqrt{\frac{\rho}{p}}\mu_s$. Let X^* denotes the ground truth clustering matrix defined as follows:

$$X^*_{i,j} = \begin{cases} 1 & \text{if } \sigma_*(i) = \sigma_*(j) \\ 0 & \text{otherwise.} \end{cases}$$

For any arbitrary partition σ, define the $k \times k$ overlap matrix $\beta(\sigma, \sigma_*)$, for each $s, t \in [k]$ as the fraction of all points assigned by σ to the s^{th} cluster **and** the fraction of all points assigned by σ_* to the t^{th} cluster,

$$\beta(\sigma, \sigma_*)_{s,t} = \frac{k|\sigma^{-1}(s) \cap \sigma_*^{-1}(t)|}{m}.$$

Then $\|\beta(\sigma, \sigma_*)\|_F^2$ is a measure of similarity of the partition, σ with the true partition, σ_*. Observe that if the partitions are completely uncorrelated, then β is the constant matrix of $1/k$ and $\|\beta(\sigma, \sigma_*)\|_F^2 = 1$. If the partitions are identical up to permutations over the labels, then β would be the permutation matrix and $\|\beta(\sigma, \sigma_*)\|_F^2 = k$.

Alternatively, for the sake of analytical tractability, we sometimes, use the quantity $err(\sigma, \sigma_*)$ to denote the fraction of points misclassified by σ.

$$err(\sigma, \sigma_*) = 1 - \frac{\max_{\pi} \mathrm{Trace}(\pi\beta(\sigma, \sigma_*))}{k}.$$

where $\pi\beta$ refers to the matrix resulting from a permutation of β over the cluster labels and the maximum is over all possible such permutations.

Clustering with k-means: The clustering objective of the k-means procedure (Pollard, 1981) is given as follows:

$$\min_{\sigma:[m]\to[k]} \sum_{s=1}^{k} \sum_{i \in \sigma^{-1}(s)} \left\| x_i - \frac{k}{m} \sum_{\sigma(j)=s} x_j \right\|^2.$$

This is equivalent to the following optimization problem:

$$\max_{\sigma:[m]\to[k]} \sum_{s=1}^{k} \sum_{i,j \in \sigma^{-1}(s)} \langle x_i, x_j \rangle.$$

Kernel k-means: For any partition $\sigma : [m] \to [k]$, define

$$\mathcal{F}(\sigma) = \sum_{s=1}^{k} \sum_{i,j \in \sigma^{-1}(s)} k(x_i, x_j).$$

Then, by the use of the kernel trick, we can formulate the kernel k-means clustering objective as follows:

$$\max_{\sigma:[m]\to[k]} \mathcal{F}(\sigma) \tag{1}$$

where $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a kernel function. Minimizing this objective over all possible partitions is NP-hard (Garey, Johnson, and Witsenhausen, 1982; Aloise et al., 2009). Several convex relaxations of the k-means procedure exist in literature. A well known semi-definite program (SDP) relaxation of the kernel k-means (Peng and Wei, 2007) objective is given by:

$$\begin{aligned} &\max_X \ \mathrm{trace}(KX) && (2) \\ &\text{s.t., } X \succeq 0, X \geq 0, \ X\mathbf{1} = \frac{m}{k}\mathbf{1}, \ \mathrm{diag}(X) = \mathbf{1}, \end{aligned}$$

where K refers to the kernel matrix for a given kernel function k: $K_{i,j} = k(x_i, x_j)$. This SDP can be solved in polynomial time. To obtain a partitioning $\hat{\sigma}$ of the data based on the optimal solution $\hat{X}$ of the SDP, a 7-approximate k-medians's procedure (Charikar et al., 2002) is applied on the rows of the matrix $\hat{X}$ in a similar fashion as Fei and Chen (2018). We denote the partition inferred by this procedure as $\hat{\sigma}$. The details of the k-median procedure can be found in Fei and Chen (2018, Algorithm 1).

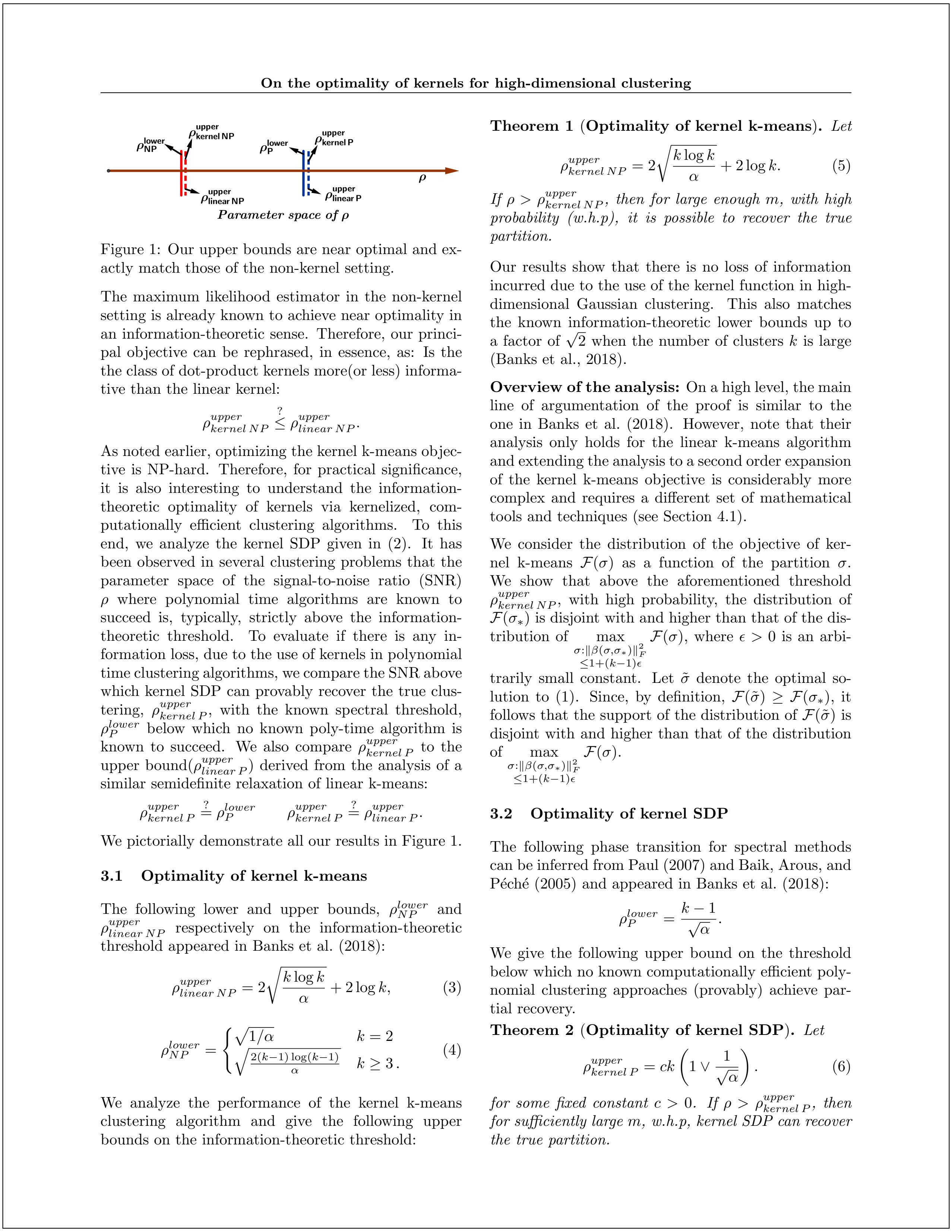
Choice of kernel function: For the analysis, we consider the class of dot-product kernel functions — $k(x, y) = f(\langle x, y \rangle)$ where f is assumed to be twice continuously differentiable and $f'(0) > 0$. The well-known exponential kernel $k(x, y) = \exp(\langle x, y \rangle)$ belongs to this class of kernel functions.

3 Our Results

We denote the upper bound on the information-theoretic threshold in the non-kernel setting as $\rho^{upper}_{linear\,NP}$ and the best known lower bound as ρ^{lower}_{NP}. We denote the upper bound from the analysis of the NP-hard kernel k-means procedure as $\rho^{upper}_{kernel\,NP}$.

The central question we address in this section is the following: Does the kernel clustering procedure achieve information-theoretic optimality for high-dimensional clustering:

$$\rho^{upper}_{kernel\,NP} \overset{?}{=} \rho^{lower}_{NP}.$$



On the optimality of kernels for high-dimensional clustering

$\rho^{\text{lower}}_{\text{NP}}$
$\rho^{\text{upper}}_{\text{kernel NP}}$
$\rho^{\text{upper}}_{\text{linear NP}}$
$\rho^{\text{lower}}_{\text{P}}$
$\rho^{\text{upper}}_{\text{kernel P}}$
$\rho^{\text{upper}}_{\text{linear P}}$
ρ
Parameter space of ρ

Figure 1: Our upper bounds are near optimal and exactly match those of the non-kernel setting.

The maximum likelihood estimator in the non-kernel setting is already known to achieve near optimality in an information-theoretic sense. Therefore, our principal objective can be rephrased, in essence, as: Is the the class of dot-product kernels more(or less) informative than the linear kernel:

$$\rho^{upper}_{kernel\,NP} \overset{?}{\leq} \rho^{upper}_{linear\,NP}.$$

As noted earlier, optimizing the kernel k-means objective is NP-hard. Therefore, for practical significance, it is also interesting to understand the information-theoretic optimality of kernels via kernelized, computationally efficient clustering algorithms. To this end, we analyze the kernel SDP given in (2). It has been observed in several clustering problems that the parameter space of the signal-to-noise ratio (SNR) ρ where polynomial time algorithms are known to succeed is, typically, strictly above the information-theoretic threshold. To evaluate if there is any information loss, due to the use of kernels in polynomial time clustering algorithms, we compare the SNR above which kernel SDP can provably recover the true clustering, $\rho^{upper}_{kernel\,P}$, with the known spectral threshold, ρ^{lower}_{P} below which no known poly-time algorithm is known to succeed. We also compare $\rho^{upper}_{kernel\,P}$ to the upper bound($\rho^{upper}_{linear\,P}$) derived from the analysis of a similar semidefinite relaxation of linear k-means:

$$\rho^{upper}_{kernel\,P} \overset{?}{=} \rho^{lower}_{P} \qquad \rho^{upper}_{kernel\,P} \overset{?}{=} \rho^{upper}_{linear\,P}.$$

We pictorially demonstrate all our results in Figure 1.

3.1 Optimality of kernel k-means

The following lower and upper bounds, ρ^{lower}_{NP} and $\rho^{upper}_{linear\,NP}$ respectively on the information-theoretic threshold appeared in Banks et al. (2018):

$$\rho^{upper}_{linear\,NP} = 2\sqrt{\frac{k\log k}{\alpha}} + 2\log k, \qquad (3)$$

$$\rho^{lower}_{NP} = \begin{cases} \sqrt{1/\alpha} & k = 2 \\ \sqrt{\frac{2(k-1)\log(k-1)}{\alpha}} & k \geq 3\,. \end{cases} \qquad (4)$$

We analyze the performance of the kernel k-means clustering algorithm and give the following upper bounds on the information-theoretic threshold:

Theorem 1 (**Optimality of kernel k-means**). *Let*

$$\rho^{upper}_{kernel\,NP} = 2\sqrt{\frac{k\log k}{\alpha}} + 2\log k. \qquad (5)$$

If $\rho > \rho^{upper}_{kernel\,NP}$, then for large enough m, with high probability (w.h.p), it is possible to recover the true partition.

Our results show that there is no loss of information incurred due to the use of the kernel function in high-dimensional Gaussian clustering. This also matches the known information-theoretic lower bounds up to a factor of $\sqrt{2}$ when the number of clusters k is large (Banks et al., 2018).

Overview of the analysis: On a high level, the main line of argumentation of the proof is similar to the one in Banks et al. (2018). However, note that their analysis only holds for the linear k-means algorithm and extending the analysis to a second order expansion of the kernel k-means objective is considerably more complex and requires a different set of mathematical tools and techniques (see Section 4.1).

We consider the distribution of the objective of kernel k-means $\mathcal{F}(\sigma)$ as a function of the partition σ. We show that above the aforementioned threshold $\rho^{upper}_{kernel\,NP}$, with high probability, the distribution of $\mathcal{F}(\sigma_*)$ is disjoint with and higher than that of the distribution of $\max\limits_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} \mathcal{F}(\sigma)$, where $\epsilon > 0$ is an arbitrarily small constant. Let $\tilde{\sigma}$ denote the optimal solution to (1). Since, by definition, $\mathcal{F}(\tilde{\sigma}) \geq \mathcal{F}(\sigma_*)$, it follows that the support of the distribution of $\mathcal{F}(\tilde{\sigma})$ is disjoint with and higher than that of the distribution of $\max\limits_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} \mathcal{F}(\sigma)$.

3.2 Optimality of kernel SDP

The following phase transition for spectral methods can be inferred from Paul (2007) and Baik, Arous, and Péché (2005) and appeared in Banks et al. (2018):

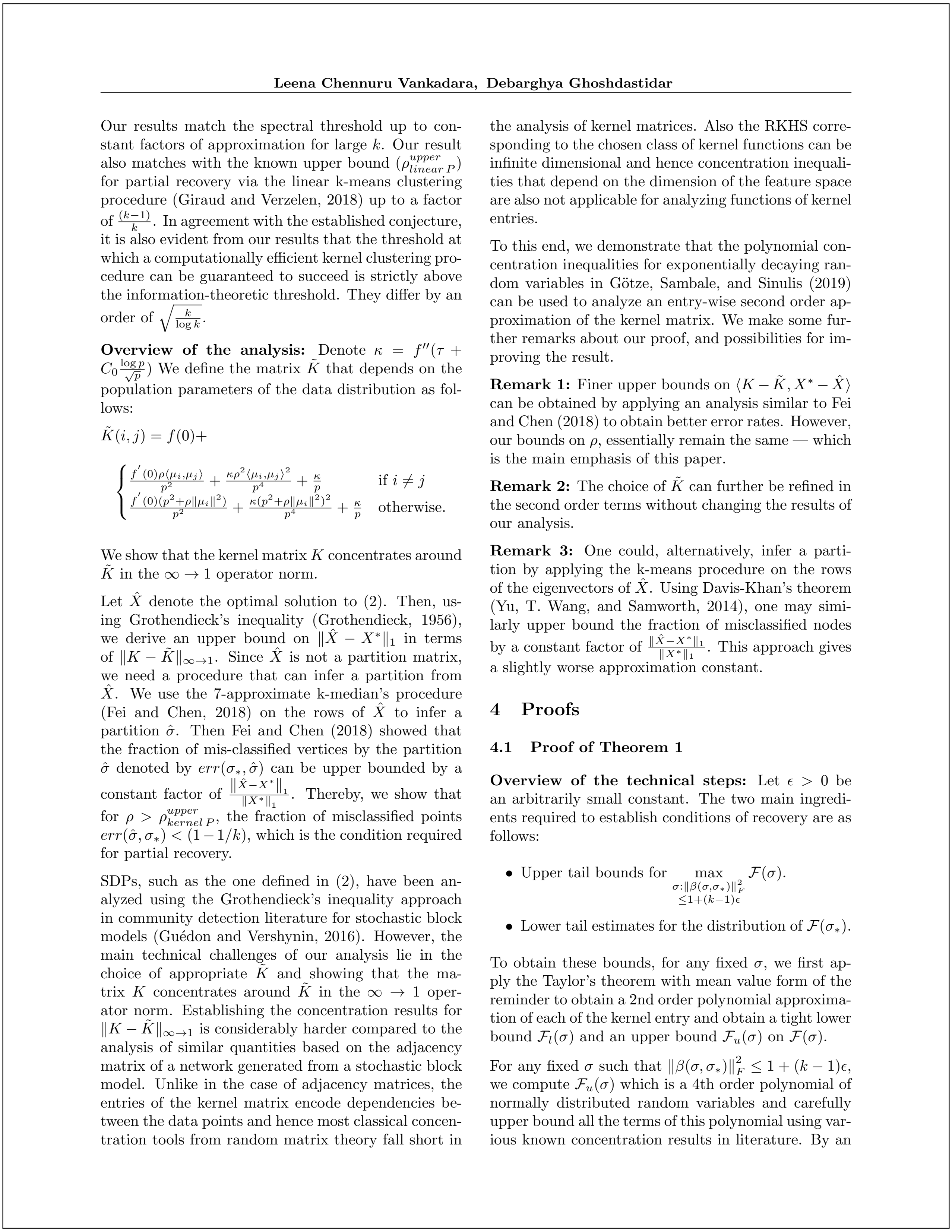
$$\rho^{lower}_{P} = \frac{k-1}{\sqrt{\alpha}}.$$

We give the following upper bound on the threshold below which no known computationally efficient polynomial clustering approaches (provably) achieve partial recovery.

Theorem 2 (**Optimality of kernel SDP**). *Let*

$$\rho^{upper}_{kernel\,P} = ck\left(1 \vee \frac{1}{\sqrt{\alpha}}\right). \qquad (6)$$

for some fixed constant $c > 0$. If $\rho > \rho^{upper}_{kernel\,P}$, then for sufficiently large m, w.h.p, kernel SDP can recover the true partition.



Our results match the spectral threshold up to constant factors of approximation for large k. Our result also matches with the known upper bound ($\rho^{upper}_{linear\,P}$) for partial recovery via the linear k-means clustering procedure (Giraud and Verzelen, 2018) up to a factor of $\frac{(k-1)}{k}$. In agreement with the established conjecture, it is also evident from our results that the threshold at which a computationally efficient kernel clustering procedure can be guaranteed to succeed is strictly above the information-theoretic threshold. They differ by an order of $\sqrt{\frac{k}{\log k}}$.

Overview of the analysis: Denote $\kappa = f''(\tau + C_0\frac{\log p}{\sqrt{p}})$ We define the matrix $\tilde{K}$ that depends on the population parameters of the data distribution as follows:

$$\tilde{K}(i,j) = f(0)+$$

$$\begin{cases} \frac{f^{'}(0)\rho\langle\mu_i,\mu_j\rangle}{p^2} + \frac{\kappa\rho^2\langle\mu_i,\mu_j\rangle^2}{p^4} + \frac{\kappa}{p} & \text{if } i \neq j \\ \frac{f^{'}(0)(p^2+\rho\|\mu_i\|^2)}{p^2} + \frac{\kappa(p^2+\rho\|\mu_i\|^2)^2}{p^4} + \frac{\kappa}{p} & \text{otherwise.} \end{cases}$$

We show that the kernel matrix K concentrates around $\tilde{K}$ in the $\infty \to 1$ operator norm.

Let $\hat{X}$ denote the optimal solution to (2). Then, using Grothendieck's inequality (Grothendieck, 1956), we derive an upper bound on $\|\hat{X} - X^*\|_1$ in terms of $\|K - \tilde{K}\|_{\infty\to 1}$. Since $\hat{X}$ is not a partition matrix, we need a procedure that can infer a partition from $\hat{X}$. We use the 7-approximate k-median's procedure (Fei and Chen, 2018) on the rows of $\hat{X}$ to infer a partition $\hat{\sigma}$. Then Fei and Chen (2018) showed that the fraction of mis-classified vertices by the partition $\hat{\sigma}$ denoted by $err(\sigma_*, \hat{\sigma})$ can be upper bounded by a constant factor of $\frac{\|\hat{X}-X^*\|_1}{\|X^*\|_1}$. Thereby, we show that for $\rho > \rho^{upper}_{kernel\,P}$, the fraction of misclassified points $err(\hat{\sigma}, \sigma_*) < (1 - 1/k)$, which is the condition required for partial recovery.

SDPs, such as the one defined in (2), have been analyzed using the Grothendieck's inequality approach in community detection literature for stochastic block models (Guédon and Vershynin, 2016). However, the main technical challenges of our analysis lie in the choice of appropriate $\tilde{K}$ and showing that the matrix K concentrates around $\tilde{K}$ in the $\infty \to 1$ operator norm. Establishing the concentration results for $\|K - \tilde{K}\|_{\infty\to 1}$ is considerably harder compared to the analysis of similar quantities based on the adjacency matrix of a network generated from a stochastic block model. Unlike in the case of adjacency matrices, the entries of the kernel matrix encode dependencies between the data points and hence most classical concentration tools from random matrix theory fall short in the analysis of kernel matrices. Also the RKHS corresponding to the chosen class of kernel functions can be infinite dimensional and hence concentration inequalities that depend on the dimension of the feature space are also not applicable for analyzing functions of kernel entries.

To this end, we demonstrate that the polynomial concentration inequalities for exponentially decaying random variables in Götze, Sambale, and Sinulis (2019) can be used to analyze an entry-wise second order approximation of the kernel matrix. We make some further remarks about our proof, and possibilities for improving the result.

Remark 1: Finer upper bounds on $\langle K - \tilde{K}, X^* - \hat{X}\rangle$ can be obtained by applying an analysis similar to Fei and Chen (2018) to obtain better error rates. However, our bounds on ρ, essentially remain the same — which is the main emphasis of this paper.

Remark 2: The choice of $\tilde{K}$ can further be refined in the second order terms without changing the results of our analysis.

Remark 3: One could, alternatively, infer a partition by applying the k-means procedure on the rows of the eigenvectors of $\hat{X}$. Using Davis-Khan's theorem (Yu, T. Wang, and Samworth, 2014), one may similarly upper bound the fraction of misclassified nodes by a constant factor of $\frac{\|\hat{X}-X^*\|_1}{\|X^*\|_1}$. This approach gives a slightly worse approximation constant.

4 Proofs

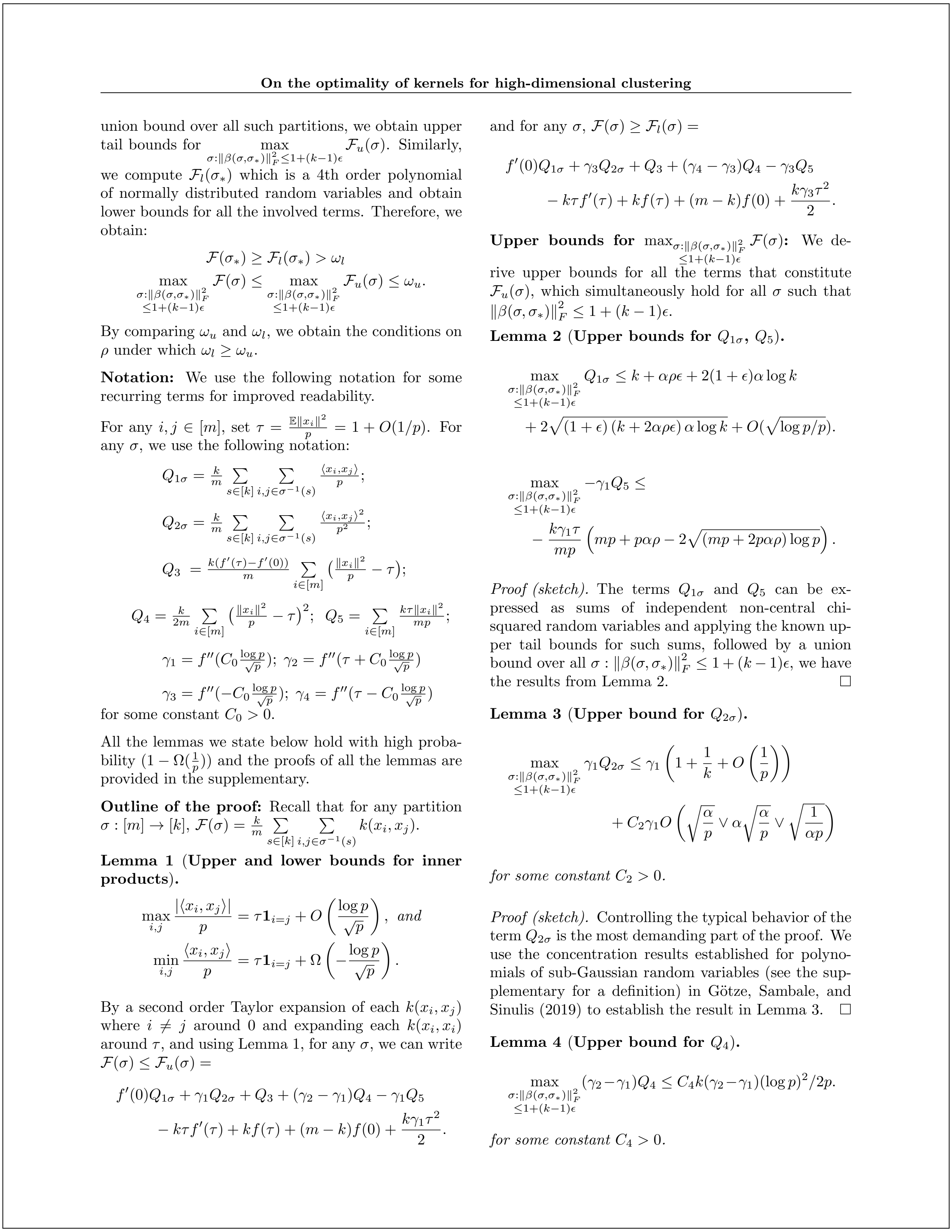
4.1 Proof of Theorem 1

Overview of the technical steps: Let $\epsilon > 0$ be an arbitrarily small constant. The two main ingredients required to establish conditions of recovery are as follows:

- Upper tail bounds for $\max\limits_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \leq 1+(k-1)\epsilon}} \mathcal{F}(\sigma)$.
- Lower tail estimates for the distribution of $\mathcal{F}(\sigma_*)$.

To obtain these bounds, for any fixed σ, we first apply the Taylor's theorem with mean value form of the reminder to obtain a 2nd order polynomial approximation of each of the kernel entry and obtain a tight lower bound $\mathcal{F}_l(\sigma)$ and an upper bound $\mathcal{F}_u(\sigma)$ on $\mathcal{F}(\sigma)$.

For any fixed σ such that $\|\beta(\sigma, \sigma_*)\|_F^2 \leq 1 + (k-1)\epsilon$, we compute $\mathcal{F}_u(\sigma)$ which is a 4th order polynomial of normally distributed random variables and carefully upper bound all the terms of this polynomial using various known concentration results in literature. By an



union bound over all such partitions, we obtain upper tail bounds for $\max_{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\leq 1+(k-1)\epsilon} \mathcal{F}_u(\sigma)$. Similarly, we compute $\mathcal{F}_l(\sigma_*)$ which is a 4th order polynomial of normally distributed random variables and obtain lower bounds for all the involved terms. Therefore, we obtain:

$$\mathcal{F}(\sigma_*) \geq \mathcal{F}_l(\sigma_*) > \omega_l$$
$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} \mathcal{F}(\sigma) \leq \max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} \mathcal{F}_u(\sigma) \leq \omega_u.$$

By comparing ω_u and ω_l, we obtain the conditions on ρ under which $\omega_l \geq \omega_u$.

Notation: We use the following notation for some recurring terms for improved readability.

For any $i, j \in [m]$, set $\tau = \frac{\mathbb{E}\|x_i\|^2}{p} = 1 + O(1/p)$. For any σ, we use the following notation:

$$Q_{1\sigma} = \tfrac{k}{m} \sum_{s\in[k]} \sum_{i,j\in\sigma^{-1}(s)} \tfrac{\langle x_i,x_j\rangle}{p};$$

$$Q_{2\sigma} = \tfrac{k}{m} \sum_{s\in[k]} \sum_{i,j\in\sigma^{-1}(s)} \tfrac{\langle x_i,x_j\rangle^2}{p^2};$$

$$Q_3 \ = \tfrac{k(f'(\tau)-f'(0))}{m} \sum_{i\in[m]} \big(\tfrac{\|x_i\|^2}{p} - \tau\big);$$

$$Q_4 = \tfrac{k}{2m} \sum_{i\in[m]} \big(\tfrac{\|x_i\|^2}{p} - \tau\big)^2; \ \ Q_5 = \sum_{i\in[m]} \tfrac{k\tau\|x_i\|^2}{mp};$$

$$\gamma_1 = f''(C_0 \tfrac{\log p}{\sqrt{p}}); \ \gamma_2 = f''(\tau + C_0 \tfrac{\log p}{\sqrt{p}})$$

$$\gamma_3 = f''(-C_0 \tfrac{\log p}{\sqrt{p}}); \ \gamma_4 = f''(\tau - C_0 \tfrac{\log p}{\sqrt{p}})$$

for some constant $C_0 > 0$.

All the lemmas we state below hold with high probability $(1 - \Omega(\frac{1}{p}))$ and the proofs of all the lemmas are provided in the supplementary.

Outline of the proof: Recall that for any partition $\sigma : [m] \to [k]$, $\mathcal{F}(\sigma) = \frac{k}{m} \sum_{s\in[k]} \sum_{i,j\in\sigma^{-1}(s)} k(x_i, x_j)$.

Lemma 1 (**Upper and lower bounds for inner products**)**.**

$$\max_{i,j} \frac{|\langle x_i, x_j\rangle|}{p} = \tau \mathbf{1}_{i=j} + O\left(\frac{\log p}{\sqrt{p}}\right), \ \textit{and}$$
$$\min_{i,j} \frac{\langle x_i, x_j\rangle}{p} = \tau \mathbf{1}_{i=j} + \Omega\left(-\frac{\log p}{\sqrt{p}}\right).$$

By a second order Taylor expansion of each $k(x_i, x_j)$ where $i \neq j$ around 0 and expanding each $k(x_i, x_i)$ around τ, and using Lemma 1, for any σ, we can write $\mathcal{F}(\sigma) \leq \mathcal{F}_u(\sigma) =$

$$f'(0)Q_{1\sigma} + \gamma_1 Q_{2\sigma} + Q_3 + (\gamma_2 - \gamma_1)Q_4 - \gamma_1 Q_5$$
$$- k\tau f'(\tau) + kf(\tau) + (m-k)f(0) + \frac{k\gamma_1\tau^2}{2}.$$

and for any σ, $\mathcal{F}(\sigma) \geq \mathcal{F}_l(\sigma) =$

$$f'(0)Q_{1\sigma} + \gamma_3 Q_{2\sigma} + Q_3 + (\gamma_4 - \gamma_3)Q_4 - \gamma_3 Q_5$$
$$- k\tau f'(\tau) + kf(\tau) + (m-k)f(0) + \frac{k\gamma_3\tau^2}{2}.$$

Upper bounds for $\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} \mathcal{F}(\sigma)$: We derive upper bounds for all the terms that constitute $\mathcal{F}_u(\sigma)$, which simultaneously hold for all σ such that $\|\beta(\sigma,\sigma_*)\|_F^2 \leq 1 + (k-1)\epsilon$.

Lemma 2 (**Upper bounds for $Q_{1\sigma}$, Q_5**)**.**

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} Q_{1\sigma} \leq k + \alpha\rho\epsilon + 2(1+\epsilon)\alpha\log k$$
$$+ 2\sqrt{(1+\epsilon)\,(k + 2\alpha\rho\epsilon)\,\alpha\log k} + O(\sqrt{\log p/p}).$$

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} -\gamma_1 Q_5 \leq$$
$$- \frac{k\gamma_1\tau}{mp}\left(mp + p\alpha\rho - 2\sqrt{(mp + 2p\alpha\rho)\log p}\right).$$

Proof (sketch). The terms $Q_{1\sigma}$ and Q_5 can be expressed as sums of independent non-central chi-squared random variables and applying the known upper tail bounds for such sums, followed by a union bound over all $\sigma : \|\beta(\sigma,\sigma_*)\|_F^2 \leq 1 + (k-1)\epsilon$, we have the results from Lemma 2. □

Lemma 3 (**Upper bound for $Q_{2\sigma}$**)**.**

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} \gamma_1 Q_{2\sigma} \leq \gamma_1\left(1 + \frac{1}{k} + O\left(\frac{1}{p}\right)\right)$$
$$+ C_2\gamma_1 O\left(\sqrt{\frac{\alpha}{p}} \vee \alpha\sqrt{\frac{\alpha}{p}} \vee \sqrt{\frac{1}{\alpha p}}\right)$$

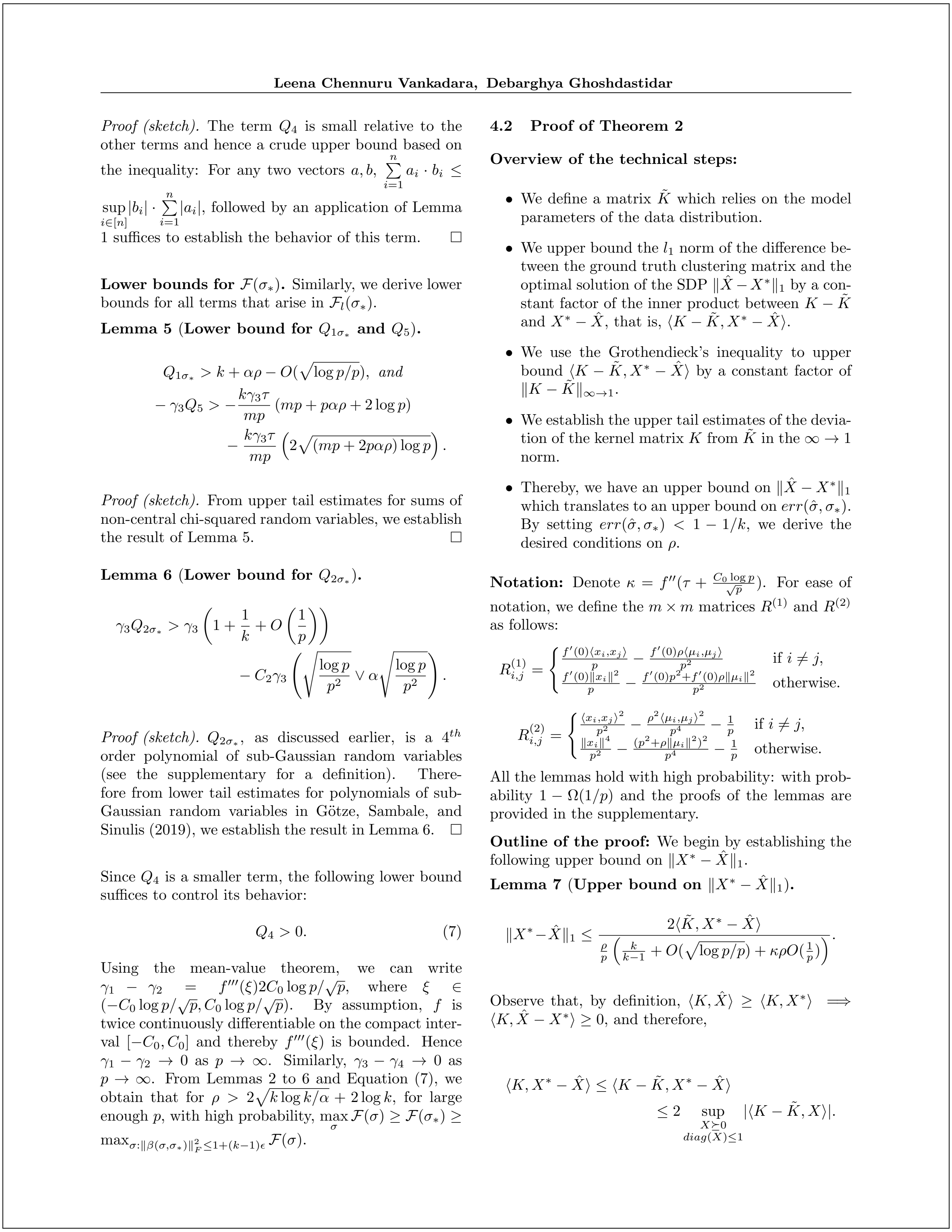
for some constant $C_2 > 0$.

Proof (sketch). Controlling the typical behavior of the term $Q_{2\sigma}$ is the most demanding part of the proof. We use the concentration results established for polynomials of sub-Gaussian random variables (see the supplementary for a definition) in Götze, Sambale, and Sinulis (2019) to establish the result in Lemma 3. □

Lemma 4 (**Upper bound for Q_4**)**.**

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} (\gamma_2 - \gamma_1)Q_4 \leq C_4 k(\gamma_2 - \gamma_1)(\log p)^2/2p.$$

for some constant $C_4 > 0$.



Proof (sketch). The term Q_4 is small relative to the other terms and hence a crude upper bound based on the inequality: For any two vectors a, b, $\sum_{i=1}^{n} a_i \cdot b_i \leq \sup_{i\in[n]} |b_i| \cdot \sum_{i=1}^{n} |a_i|$, followed by an application of Lemma 1 suffices to establish the behavior of this term. □

Lower bounds for $\mathcal{F}(\sigma_*)$. Similarly, we derive lower bounds for all terms that arise in $\mathcal{F}_l(\sigma_*)$.

Lemma 5 (**Lower bound for $Q_{1\sigma_*}$ and Q_5**).

$$\begin{aligned} Q_{1\sigma_*} &> k + \alpha\rho - O(\sqrt{\log p/p}), \ \textit{and} \\ -\gamma_3 Q_5 &> -\frac{k\gamma_3\tau}{mp}\left(mp + p\alpha\rho + 2\log p\right) \\ &\quad - \frac{k\gamma_3\tau}{mp}\left(2\sqrt{(mp + 2p\alpha\rho)\log p}\right). \end{aligned}$$

Proof (sketch). From upper tail estimates for sums of non-central chi-squared random variables, we establish the result of Lemma 5. □

Lemma 6 (**Lower bound for $Q_{2\sigma_*}$**).

$$\begin{aligned} \gamma_3 Q_{2\sigma_*} &> \gamma_3\left(1 + \frac{1}{k} + O\left(\frac{1}{p}\right)\right) \\ &\quad - C_2\gamma_3\left(\sqrt{\frac{\log p}{p^2}} \vee \alpha\sqrt{\frac{\log p}{p^2}}\right). \end{aligned}$$

Proof (sketch). $Q_{2\sigma_*}$, as discussed earlier, is a 4^{th} order polynomial of sub-Gaussian random variables (see the supplementary for a definition). Therefore from lower tail estimates for polynomials of sub-Gaussian random variables in Götze, Sambale, and Sinulis (2019), we establish the result in Lemma 6. □

Since Q_4 is a smaller term, the following lower bound suffices to control its behavior:

$$Q_4 > 0. \tag{7}$$

Using the mean-value theorem, we can write $\gamma_1 - \gamma_2 = f'''(\xi) 2C_0 \log p/\sqrt{p}$, where $\xi \in (-C_0 \log p/\sqrt{p}, C_0 \log p/\sqrt{p})$. By assumption, f is twice continuously differentiable on the compact interval $[-C_0, C_0]$ and thereby $f'''(\xi)$ is bounded. Hence $\gamma_1 - \gamma_2 \to 0$ as $p \to \infty$. Similarly, $\gamma_3 - \gamma_4 \to 0$ as $p \to \infty$. From Lemmas 2 to 6 and Equation (7), we obtain that for $\rho > 2\sqrt{k \log k/\alpha} + 2\log k$, for large enough p, with high probability, $\max_{\sigma} \mathcal{F}(\sigma) \geq \mathcal{F}(\sigma_*) \geq \max_{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \leq 1+(k-1)\epsilon} \mathcal{F}(\sigma)$.

4.2 Proof of Theorem 2

Overview of the technical steps:

- We define a matrix $\tilde{K}$ which relies on the model parameters of the data distribution.
- We upper bound the l_1 norm of the difference between the ground truth clustering matrix and the optimal solution of the SDP $\|\hat{X} - X^*\|_1$ by a constant factor of the inner product between $K - \tilde{K}$ and $X^* - \hat{X}$, that is, $\langle K - \tilde{K}, X^* - \hat{X}\rangle$.
- We use the Grothendieck's inequality to upper bound $\langle K - \tilde{K}, X^* - \hat{X}\rangle$ by a constant factor of $\|K - \tilde{K}\|_{\infty\to 1}$.
- We establish the upper tail estimates of the deviation of the kernel matrix K from $\tilde{K}$ in the $\infty \to 1$ norm.
- Thereby, we have an upper bound on $\|\hat{X} - X^*\|_1$ which translates to an upper bound on $err(\hat{\sigma}, \sigma_*)$. By setting $err(\hat{\sigma}, \sigma_*) < 1 - 1/k$, we derive the desired conditions on ρ.

Notation: Denote $\kappa = f''(\tau + \frac{C_0 \log p}{\sqrt{p}})$. For ease of notation, we define the $m \times m$ matrices $R^{(1)}$ and $R^{(2)}$ as follows:

$$R^{(1)}_{i,j} = \begin{cases} \frac{f'(0)\langle x_i, x_j\rangle}{p} - \frac{f'(0)\rho\langle \mu_i, \mu_j\rangle}{p^2} & \text{if } i \neq j, \\ \frac{f'(0)\|x_i\|^2}{p} - \frac{f'(0)p^2 + f'(0)\rho\|\mu_i\|^2}{p^2} & \text{otherwise.} \end{cases}$$

$$R^{(2)}_{i,j} = \begin{cases} \frac{\langle x_i, x_j\rangle^2}{p^2} - \frac{\rho^2\langle \mu_i, \mu_j\rangle^2}{p^4} - \frac{1}{p} & \text{if } i \neq j, \\ \frac{\|x_i\|^4}{p^2} - \frac{(p^2 + \rho\|\mu_i\|^2)^2}{p^4} - \frac{1}{p} & \text{otherwise.} \end{cases}$$

All the lemmas hold with high probability: with probability $1 - \Omega(1/p)$ and the proofs of the lemmas are provided in the supplementary.

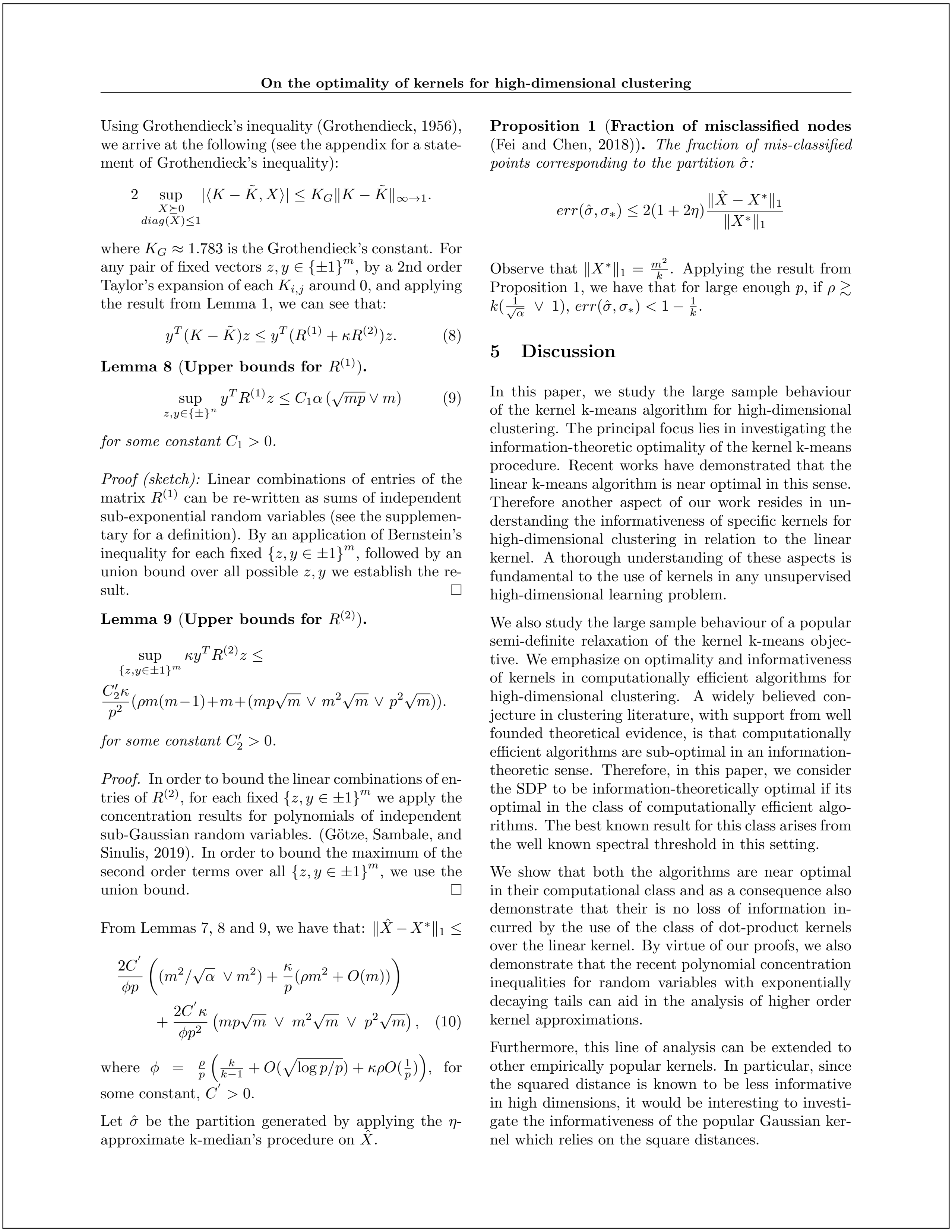
Outline of the proof: We begin by establishing the following upper bound on $\|X^* - \hat{X}\|_1$.

Lemma 7 (**Upper bound on $\|X^* - \hat{X}\|_1$**).

$$\|X^* - \hat{X}\|_1 \leq \frac{2\langle \tilde{K}, X^* - \hat{X}\rangle}{\frac{\rho}{p}\left(\frac{k}{k-1} + O(\sqrt{\log p/p}) + \kappa\rho O(\frac{1}{p})\right)}.$$

Observe that, by definition, $\langle K, \hat{X}\rangle \geq \langle K, X^*\rangle \implies \langle K, \hat{X} - X^*\rangle \geq 0$, and therefore,

$$\begin{aligned} \langle K, X^* - \hat{X}\rangle &\leq \langle K - \tilde{K}, X^* - \hat{X}\rangle \\ &\leq 2 \sup_{\substack{X \succeq 0 \\ diag(X) \leq 1}} |\langle K - \tilde{K}, X\rangle|. \end{aligned}$$



Using Grothendieck's inequality (Grothendieck, 1956), we arrive at the following (see the appendix for a statement of Grothendieck's inequality):

$$2 \sup_{\substack{X \succeq 0 \\ diag(X) \le 1}} |\langle K - \tilde{K}, X \rangle| \le K_G \|K - \tilde{K}\|_{\infty \to 1}.$$

where $K_G \approx 1.783$ is the Grothendieck's constant. For any pair of fixed vectors $z, y \in \{\pm 1\}^m$, by a 2nd order Taylor's expansion of each $K_{i,j}$ around 0, and applying the result from Lemma 1, we can see that:

$$y^T (K - \tilde{K}) z \le y^T (R^{(1)} + \kappa R^{(2)}) z. \tag{8}$$

Lemma 8 (**Upper bounds for** $R^{(1)}$).

$$\sup_{z, y \in \{\pm\}^n} y^T R^{(1)} z \le C_1 \alpha \left(\sqrt{mp} \vee m\right) \tag{9}$$

for some constant $C_1 > 0$.

Proof (sketch): Linear combinations of entries of the matrix $R^{(1)}$ can be re-written as sums of independent sub-exponential random variables (see the supplementary for a definition). By an application of Bernstein's inequality for each fixed $\{z, y \in \pm 1\}^m$, followed by an union bound over all possible z, y we establish the result. □

Lemma 9 (**Upper bounds for** $R^{(2)}$).

$$\sup_{\{z, y \in \pm 1\}^m} \kappa y^T R^{(2)} z \le$$
$$\frac{C_2' \kappa}{p^2} (\rho m(m-1) + m + (mp\sqrt{m} \vee m^2 \sqrt{m} \vee p^2 \sqrt{m})).$$

for some constant $C_2' > 0$.

Proof. In order to bound the linear combinations of entries of $R^{(2)}$, for each fixed $\{z, y \in \pm 1\}^m$ we apply the concentration results for polynomials of independent sub-Gaussian random variables. (Götze, Sambale, and Sinulis, 2019). In order to bound the maximum of the second order terms over all $\{z, y \in \pm 1\}^m$, we use the union bound. □

From Lemmas 7, 8 and 9, we have that: $\|\hat{X} - X^*\|_1 \le$

$$\frac{2C^{'}}{\phi p} \left((m^2/\sqrt{\alpha} \vee m^2) + \frac{\kappa}{p} (\rho m^2 + O(m)) \right) + \frac{2C^{'} \kappa}{\phi p^2} \left(mp\sqrt{m} \vee m^2 \sqrt{m} \vee p^2 \sqrt{m} \right), \tag{10}$$

where $\phi = \frac{\rho}{p} \left(\frac{k}{k-1} + O(\sqrt{\log p/p}) + \kappa \rho O(\frac{1}{p}) \right)$, for some constant, $C^{'} > 0$.

Let $\hat{\sigma}$ be the partition generated by applying the η-approximate k-median's procedure on $\hat{X}$.

Proposition 1 (**Fraction of misclassified nodes** (Fei and Chen, 2018)). *The fraction of mis-classified points corresponding to the partition* $\hat{\sigma}$*:*

$$err(\hat{\sigma}, \sigma_*) \le 2(1 + 2\eta) \frac{\|\hat{X} - X^*\|_1}{\|X^*\|_1}$$

Observe that $\|X^*\|_1 = \frac{m^2}{k}$. Applying the result from Proposition 1, we have that for large enough p, if $\rho \gtrsim k(\frac{1}{\sqrt{\alpha}} \vee 1)$, $err(\hat{\sigma}, \sigma_*) < 1 - \frac{1}{k}$.

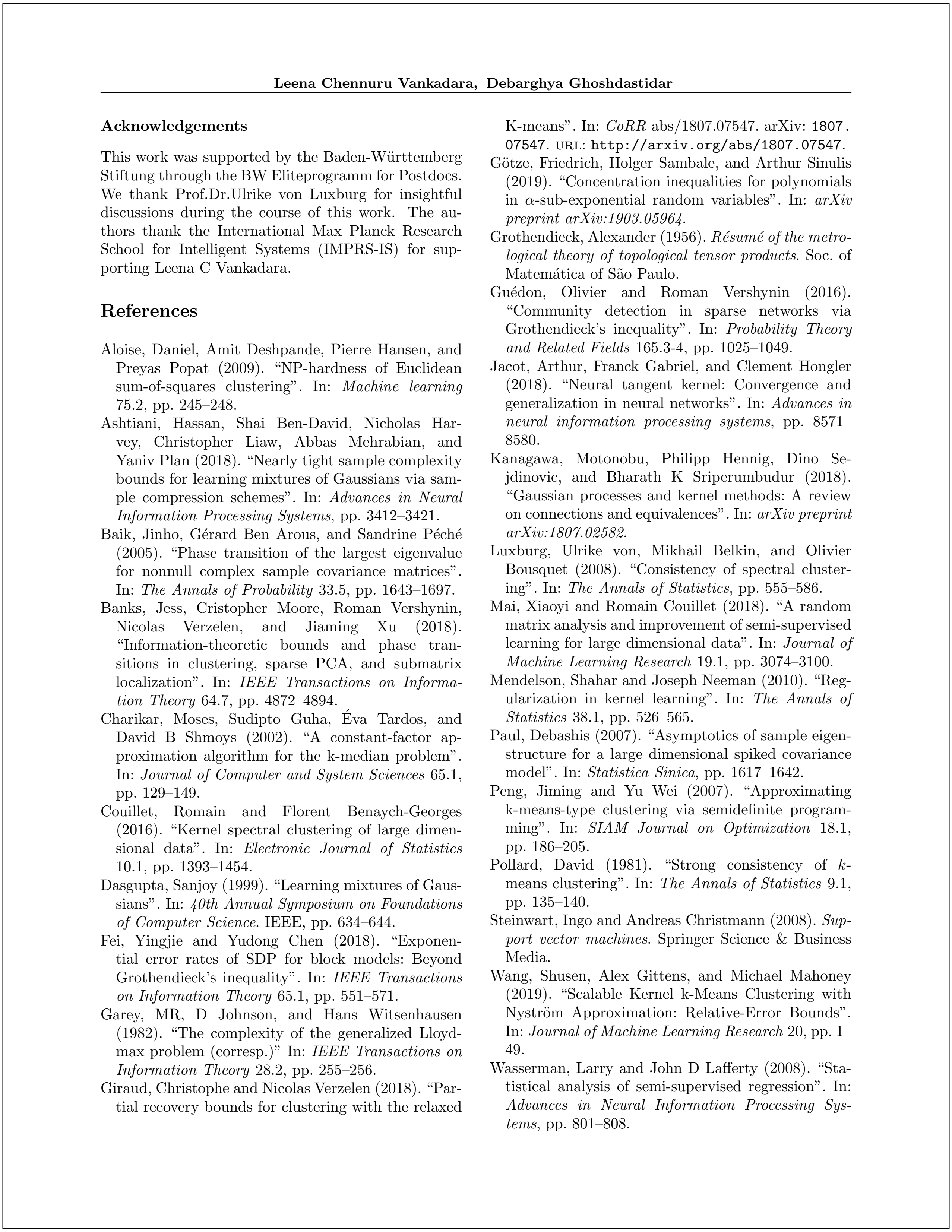
5 Discussion

In this paper, we study the large sample behaviour of the kernel k-means algorithm for high-dimensional clustering. The principal focus lies in investigating the information-theoretic optimality of the kernel k-means procedure. Recent works have demonstrated that the linear k-means algorithm is near optimal in this sense. Therefore another aspect of our work resides in understanding the informativeness of specific kernels for high-dimensional clustering in relation to the linear kernel. A thorough understanding of these aspects is fundamental to the use of kernels in any unsupervised high-dimensional learning problem.

We also study the large sample behaviour of a popular semi-definite relaxation of the kernel k-means objective. We emphasize on optimality and informativeness of kernels in computationally efficient algorithms for high-dimensional clustering. A widely believed conjecture in clustering literature, with support from well founded theoretical evidence, is that computationally efficient algorithms are sub-optimal in an information-theoretic sense. Therefore, in this paper, we consider the SDP to be information-theoretically optimal if its optimal in the class of computationally efficient algorithms. The best known result for this class arises from the well known spectral threshold in this setting.

We show that both the algorithms are near optimal in their computational class and as a consequence also demonstrate that their is no loss of information incurred by the use of the class of dot-product kernels over the linear kernel. By virtue of our proofs, we also demonstrate that the recent polynomial concentration inequalities for random variables with exponentially decaying tails can aid in the analysis of higher order kernel approximations.

Furthermore, this line of analysis can be extended to other empirically popular kernels. In particular, since the squared distance is known to be less informative in high dimensions, it would be interesting to investigate the informativeness of the popular Gaussian kernel which relies on the square distances.



Acknowledgements

This work was supported by the Baden-Württemberg Stiftung through the BW Eliteprogramm for Postdocs. We thank Prof.Dr.Ulrike von Luxburg for insightful discussions during the course of this work. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Leena C Vankadara.

References

Aloise, Daniel, Amit Deshpande, Pierre Hansen, and Preyas Popat (2009). “NP-hardness of Euclidean sum-of-squares clustering”. In: *Machine learning* 75.2, pp. 245–248.

Ashtiani, Hassan, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan (2018). “Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes”. In: *Advances in Neural Information Processing Systems*, pp. 3412–3421.

Baik, Jinho, Gérard Ben Arous, and Sandrine Péché (2005). “Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices”. In: *The Annals of Probability* 33.5, pp. 1643–1697.

Banks, Jess, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu (2018). “Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization”. In: *IEEE Transactions on Information Theory* 64.7, pp. 4872–4894.

Charikar, Moses, Sudipto Guha, Éva Tardos, and David B Shmoys (2002). “A constant-factor approximation algorithm for the k-median problem”. In: *Journal of Computer and System Sciences* 65.1, pp. 129–149.

Couillet, Romain and Florent Benaych-Georges (2016). “Kernel spectral clustering of large dimensional data”. In: *Electronic Journal of Statistics* 10.1, pp. 1393–1454.

Dasgupta, Sanjoy (1999). “Learning mixtures of Gaussians”. In: *40th Annual Symposium on Foundations of Computer Science*. IEEE, pp. 634–644.

Fei, Yingjie and Yudong Chen (2018). “Exponential error rates of SDP for block models: Beyond Grothendieck’s inequality”. In: *IEEE Transactions on Information Theory* 65.1, pp. 551–571.

Garey, MR, D Johnson, and Hans Witsenhausen (1982). “The complexity of the generalized Lloyd-max problem (corresp.)” In: *IEEE Transactions on Information Theory* 28.2, pp. 255–256.

Giraud, Christophe and Nicolas Verzelen (2018). “Partial recovery bounds for clustering with the relaxed K-means”. In: *CoRR* abs/1807.07547. arXiv: `1807.07547`. URL: `http://arxiv.org/abs/1807.07547`.

Götze, Friedrich, Holger Sambale, and Arthur Sinulis (2019). “Concentration inequalities for polynomials in α-sub-exponential random variables”. In: *arXiv preprint arXiv:1903.05964*.

Grothendieck, Alexander (1956). *Résumé of the metrological theory of topological tensor products*. Soc. of Matemática of São Paulo.

Guédon, Olivier and Roman Vershynin (2016). “Community detection in sparse networks via Grothendieck’s inequality”. In: *Probability Theory and Related Fields* 165.3-4, pp. 1025–1049.

Jacot, Arthur, Franck Gabriel, and Clement Hongler (2018). “Neural tangent kernel: Convergence and generalization in neural networks”. In: *Advances in neural information processing systems*, pp. 8571–8580.

Kanagawa, Motonobu, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur (2018). “Gaussian processes and kernel methods: A review on connections and equivalences”. In: *arXiv preprint arXiv:1807.02582*.

Luxburg, Ulrike von, Mikhail Belkin, and Olivier Bousquet (2008). “Consistency of spectral clustering”. In: *The Annals of Statistics*, pp. 555–586.

Mai, Xiaoyi and Romain Couillet (2018). “A random matrix analysis and improvement of semi-supervised learning for large dimensional data”. In: *Journal of Machine Learning Research* 19.1, pp. 3074–3100.

Mendelson, Shahar and Joseph Neeman (2010). “Regularization in kernel learning”. In: *The Annals of Statistics* 38.1, pp. 526–565.

Paul, Debashis (2007). “Asymptotics of sample eigenstructure for a large dimensional spiked covariance model”. In: *Statistica Sinica*, pp. 1617–1642.

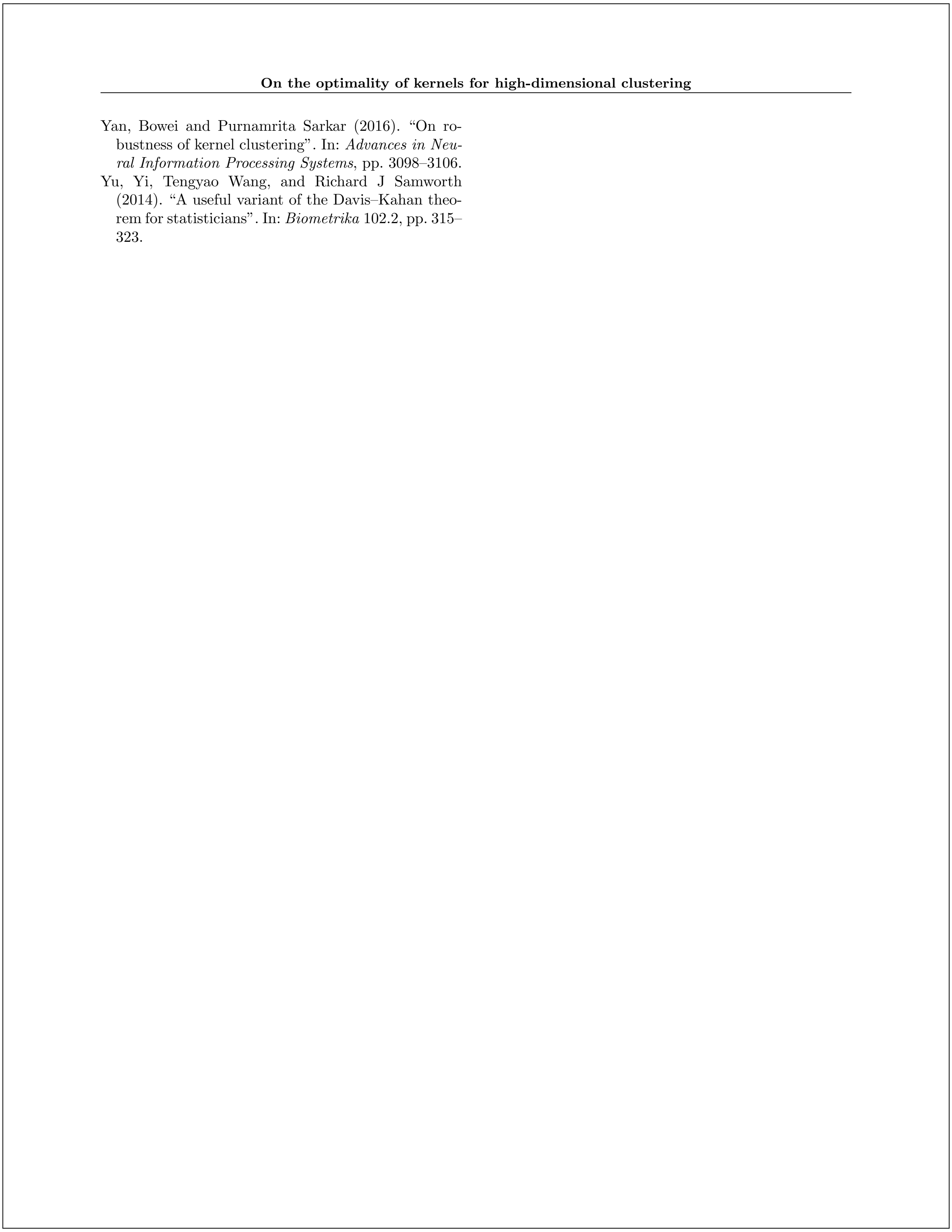
Peng, Jiming and Yu Wei (2007). “Approximating k-means-type clustering via semidefinite programming”. In: *SIAM Journal on Optimization* 18.1, pp. 186–205.

Pollard, David (1981). “Strong consistency of k-means clustering”. In: *The Annals of Statistics* 9.1, pp. 135–140.

Steinwart, Ingo and Andreas Christmann (2008). *Support vector machines*. Springer Science & Business Media.

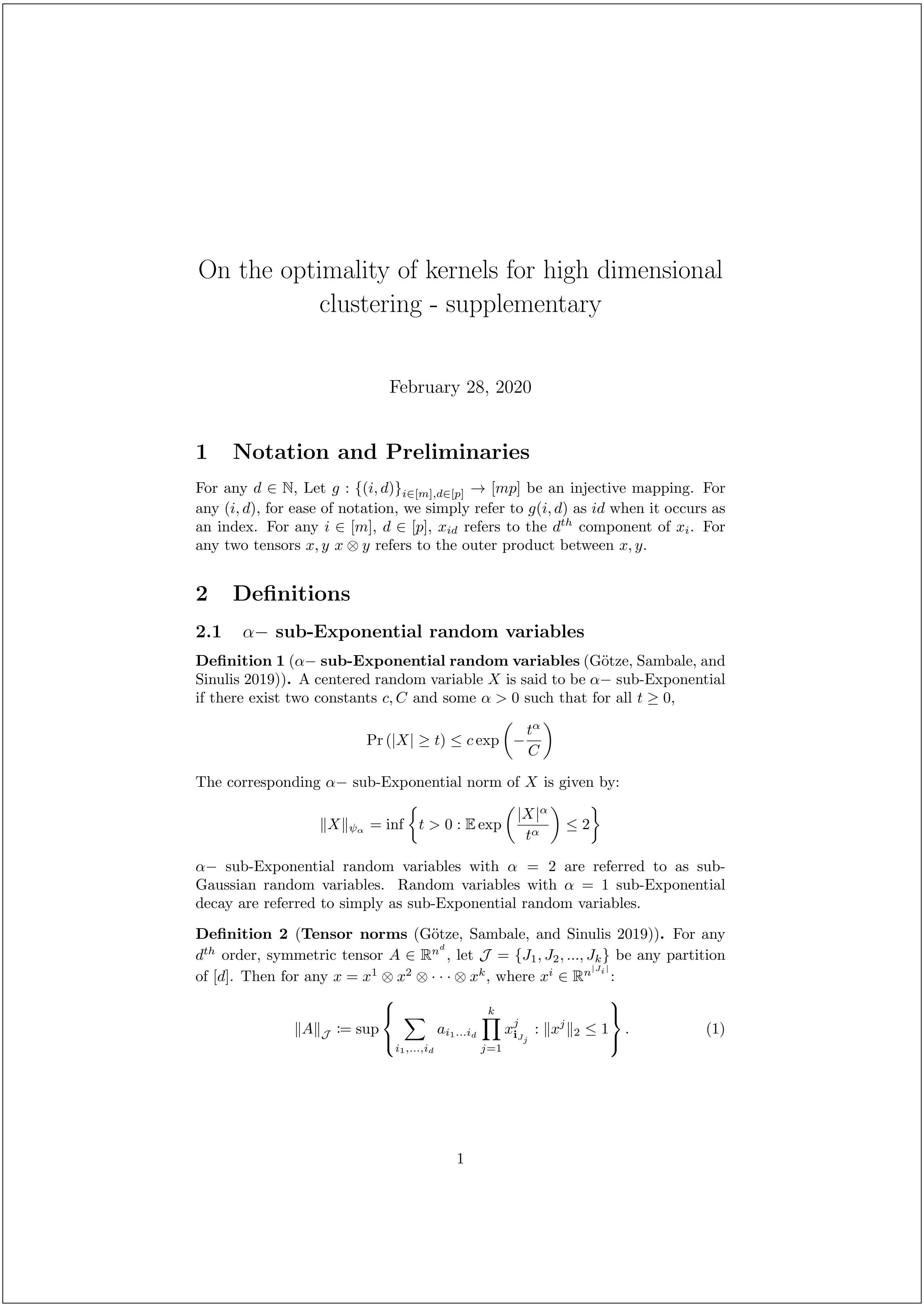
Wang, Shusen, Alex Gittens, and Michael Mahoney (2019). “Scalable Kernel k-Means Clustering with Nyström Approximation: Relative-Error Bounds”. In: *Journal of Machine Learning Research* 20, pp. 1–49.

Wasserman, Larry and John D Lafferty (2008). “Statistical analysis of semi-supervised regression”. In: *Advances in Neural Information Processing Systems*, pp. 801–808.



Yan, Bowei and Purnamrita Sarkar (2016). “On robustness of kernel clustering”. In: *Advances in Neural Information Processing Systems*, pp. 3098–3106.

Yu, Yi, Tengyao Wang, and Richard J Samworth (2014). “A useful variant of the Davis–Kahan theorem for statisticians”. In: *Biometrika* 102.2, pp. 315–323.



On the optimality of kernels for high dimensional clustering - supplementary

February 28, 2020

1 Notation and Preliminaries

For any $d \in \mathbb{N}$, Let $g : \{(i,d)\}_{i\in[m],d\in[p]} \to [mp]$ be an injective mapping. For any (i,d), for ease of notation, we simply refer to $g(i,d)$ as id when it occurs as an index. For any $i \in [m]$, $d \in [p]$, x_{id} refers to the d^{th} component of x_i. For any two tensors x, y $x \otimes y$ refers to the outer product between x, y.

2 Definitions

2.1 $\alpha-$ sub-Exponential random variables

Definition 1 ($\alpha-$ **sub-Exponential random variables** (Götze, Sambale, and Sinulis 2019))**.** A centered random variable X is said to be $\alpha-$ sub-Exponential if there exist two constants c, C and some $\alpha > 0$ such that for all $t \geq 0$,

$$\Pr\left(|X| \geq t\right) \leq c \exp\left(-\frac{t^\alpha}{C}\right)$$

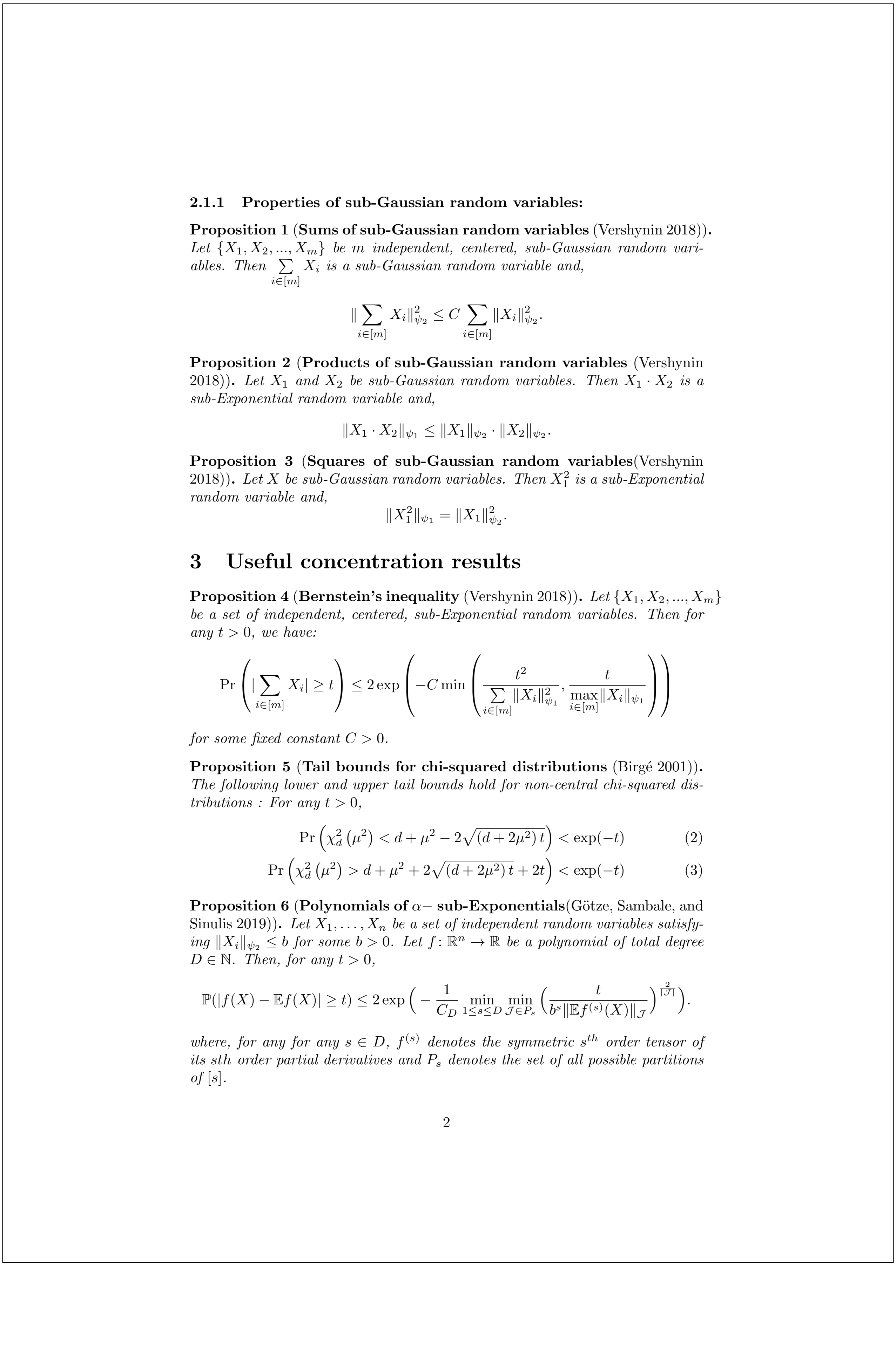
The corresponding $\alpha-$ sub-Exponential norm of X is given by:

$$\|X\|_{\psi_\alpha} = \inf\left\{t > 0 : \mathbb{E}\exp\left(\frac{|X|^\alpha}{t^\alpha}\right) \leq 2\right\}$$

$\alpha-$ sub-Exponential random variables with $\alpha = 2$ are referred to as sub-Gaussian random variables. Random variables with $\alpha = 1$ sub-Exponential decay are referred to simply as sub-Exponential random variables.

Definition 2 (**Tensor norms** (Götze, Sambale, and Sinulis 2019))**.** For any d^{th} order, symmetric tensor $A \in \mathbb{R}^{n^d}$, let $\mathcal{J} = \{J_1, J_2, ..., J_k\}$ be any partition of $[d]$. Then for any $x = x^1 \otimes x^2 \otimes \cdots \otimes x^k$, where $x^i \in \mathbb{R}^{n^{|J_i|}}$:

$$\|A\|_{\mathcal{J}} \coloneqq \sup\left\{\sum_{i_1,\dots,i_d} a_{i_1\dots i_d} \prod_{j=1}^{k} x^j_{\mathbf{i}_{J_j}} : \|x^j\|_2 \leq 1\right\}. \tag{1}$$



2.1.1 Properties of sub-Gaussian random variables:

Proposition 1 (**Sums of sub-Gaussian random variables** (Vershynin 2018))**.** *Let* $\{X_1, X_2, ..., X_m\}$ *be* m *independent, centered, sub-Gaussian random variables. Then* $\sum\limits_{i\in[m]} X_i$ *is a sub-Gaussian random variable and,*

$$\|\sum_{i\in[m]} X_i\|_{\psi_2}^2 \leq C \sum_{i\in[m]} \|X_i\|_{\psi_2}^2.$$

Proposition 2 (**Products of sub-Gaussian random variables** (Vershynin 2018))**.** *Let* X_1 *and* X_2 *be sub-Gaussian random variables. Then* $X_1 \cdot X_2$ *is a sub-Exponential random variable and,*

$$\|X_1 \cdot X_2\|_{\psi_1} \leq \|X_1\|_{\psi_2} \cdot \|X_2\|_{\psi_2}.$$

Proposition 3 (**Squares of sub-Gaussian random variables**(Vershynin 2018))**.** *Let* X *be sub-Gaussian random variables. Then* X_1^2 *is a sub-Exponential random variable and,*

$$\|X_1^2\|_{\psi_1} = \|X_1\|_{\psi_2}^2.$$

3 Useful concentration results

Proposition 4 (**Bernstein's inequality** (Vershynin 2018))**.** *Let* $\{X_1, X_2, ..., X_m\}$ *be a set of independent, centered, sub-Exponential random variables. Then for any* $t > 0$*, we have:*

$$\Pr\left(|\sum_{i\in[m]} X_i| \geq t\right) \leq 2\exp\left(-C\min\left(\frac{t^2}{\sum\limits_{i\in[m]}\|X_i\|_{\psi_1}^2}, \frac{t}{\max\limits_{i\in[m]}\|X_i\|_{\psi_1}}\right)\right)$$

for some fixed constant $C > 0$*.*

Proposition 5 (**Tail bounds for chi-squared distributions** (Birgé 2001))**.** *The following lower and upper tail bounds hold for non-central chi-squared distributions : For any* $t > 0$*,*

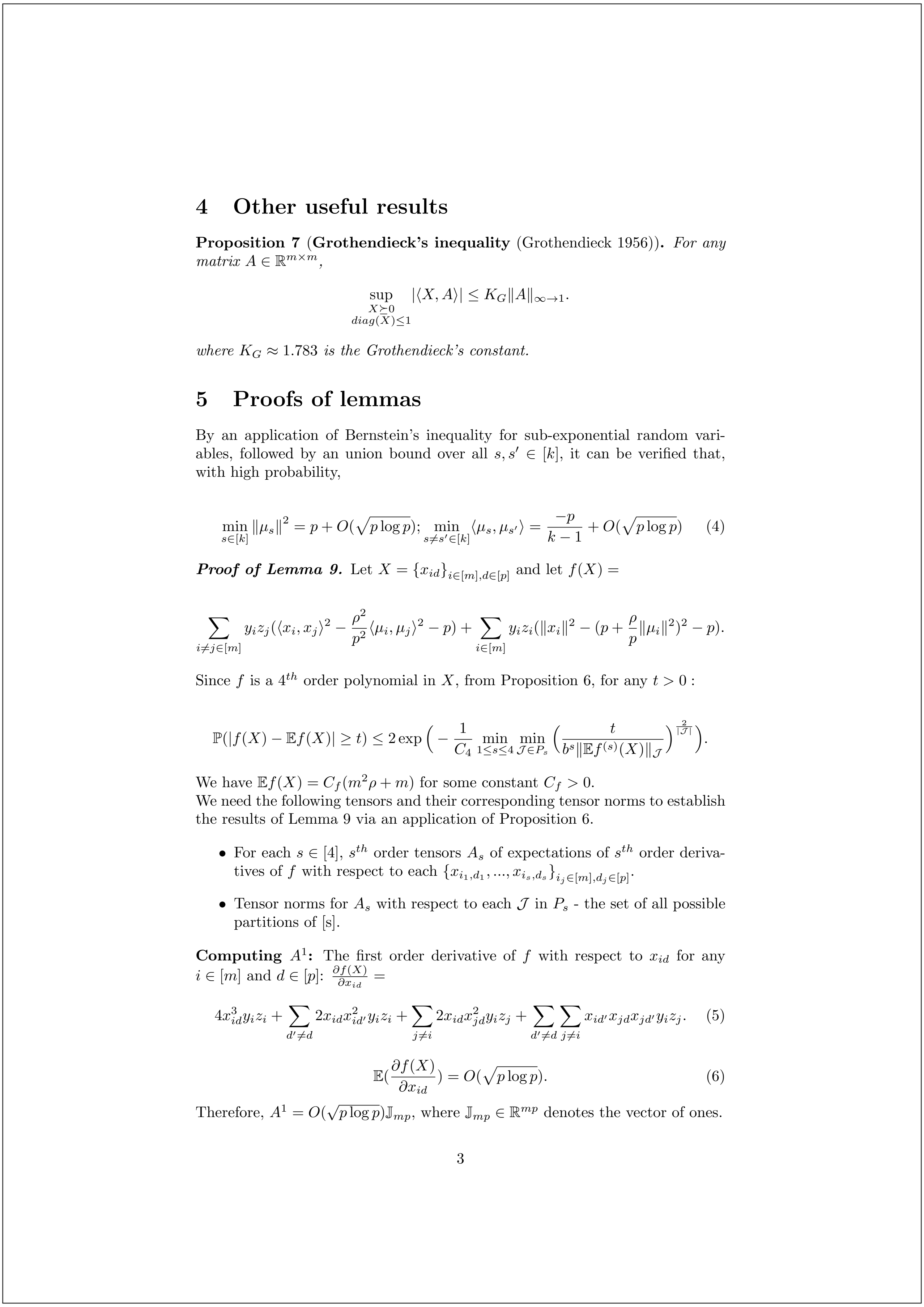
$$\Pr\left(\chi_d^2\left(\mu^2\right) < d + \mu^2 - 2\sqrt{\left(d + 2\mu^2\right)t}\right) < \exp(-t) \qquad (2)$$

$$\Pr\left(\chi_d^2\left(\mu^2\right) > d + \mu^2 + 2\sqrt{\left(d + 2\mu^2\right)t} + 2t\right) < \exp(-t) \qquad (3)$$

Proposition 6 (**Polynomials of** $\alpha-$ **sub-Exponentials**(Götze, Sambale, and Sinulis 2019))**.** *Let* $X_1, \ldots, X_n$ *be a set of independent random variables satisfying* $\|X_i\|_{\psi_2} \leq b$ *for some* $b > 0$*. Let* $f\colon \mathbb{R}^n \to \mathbb{R}$ *be a polynomial of total degree* $D \in \mathbb{N}$*. Then, for any* $t > 0$*,*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\Big(-\frac{1}{C_D}\min_{1\leq s\leq D}\min_{\mathcal{J}\in P_s}\Big(\frac{t}{b^s\|\mathbb{E}f^{(s)}(X)\|_{\mathcal{J}}}\Big)^{\frac{2}{|\mathcal{J}|}}\Big).$$

where, for any for any $s \in D$*,* $f^{(s)}$ *denotes the symmetric* s^{th} *order tensor of its sth order partial derivatives and* P_s *denotes the set of all possible partitions of* $[s]$*.*



4 Other useful results

Proposition 7 (**Grothendieck's inequality** (Grothendieck 1956)). *For any matrix $A \in \mathbb{R}^{m \times m}$,*

$$\sup_{\substack{X \succeq 0 \\ diag(X) \leq 1}} |\langle X, A\rangle| \leq K_G \|A\|_{\infty \to 1}.$$

where $K_G \approx 1.783$ is the Grothendieck's constant.

5 Proofs of lemmas

By an application of Bernstein's inequality for sub-exponential random variables, followed by an union bound over all $s, s' \in [k]$, it can be verified that, with high probability,

$$\min_{s \in [k]} \|\mu_s\|^2 = p + O(\sqrt{p \log p}); \min_{s \neq s' \in [k]} \langle \mu_s, \mu_{s'}\rangle = \frac{-p}{k-1} + O(\sqrt{p \log p}) \quad (4)$$

Proof of Lemma 9. Let $X = \{x_{id}\}_{i \in [m], d \in [p]}$ and let $f(X) =$

$$\sum_{i \neq j \in [m]} y_i z_j (\langle x_i, x_j\rangle^2 - \frac{\rho^2}{p^2}\langle \mu_i, \mu_j\rangle^2 - p) + \sum_{i \in [m]} y_i z_i (\|x_i\|^2 - (p + \frac{\rho}{p}\|\mu_i\|^2)^2 - p).$$

Since f is a 4^{th} order polynomial in X, from Proposition 6, for any $t > 0$:

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2 \exp\Big(-\frac{1}{C_4} \min_{1 \leq s \leq 4} \min_{\mathcal{J} \in P_s} \Big(\frac{t}{b^s \|\mathbb{E}f^{(s)}(X)\|_{\mathcal{J}}}\Big)^{\frac{2}{|\mathcal{J}|}}\Big).$$

We have $\mathbb{E}f(X) = C_f(m^2\rho + m)$ for some constant $C_f > 0$.
We need the following tensors and their corresponding tensor norms to establish the results of Lemma 9 via an application of Proposition 6.

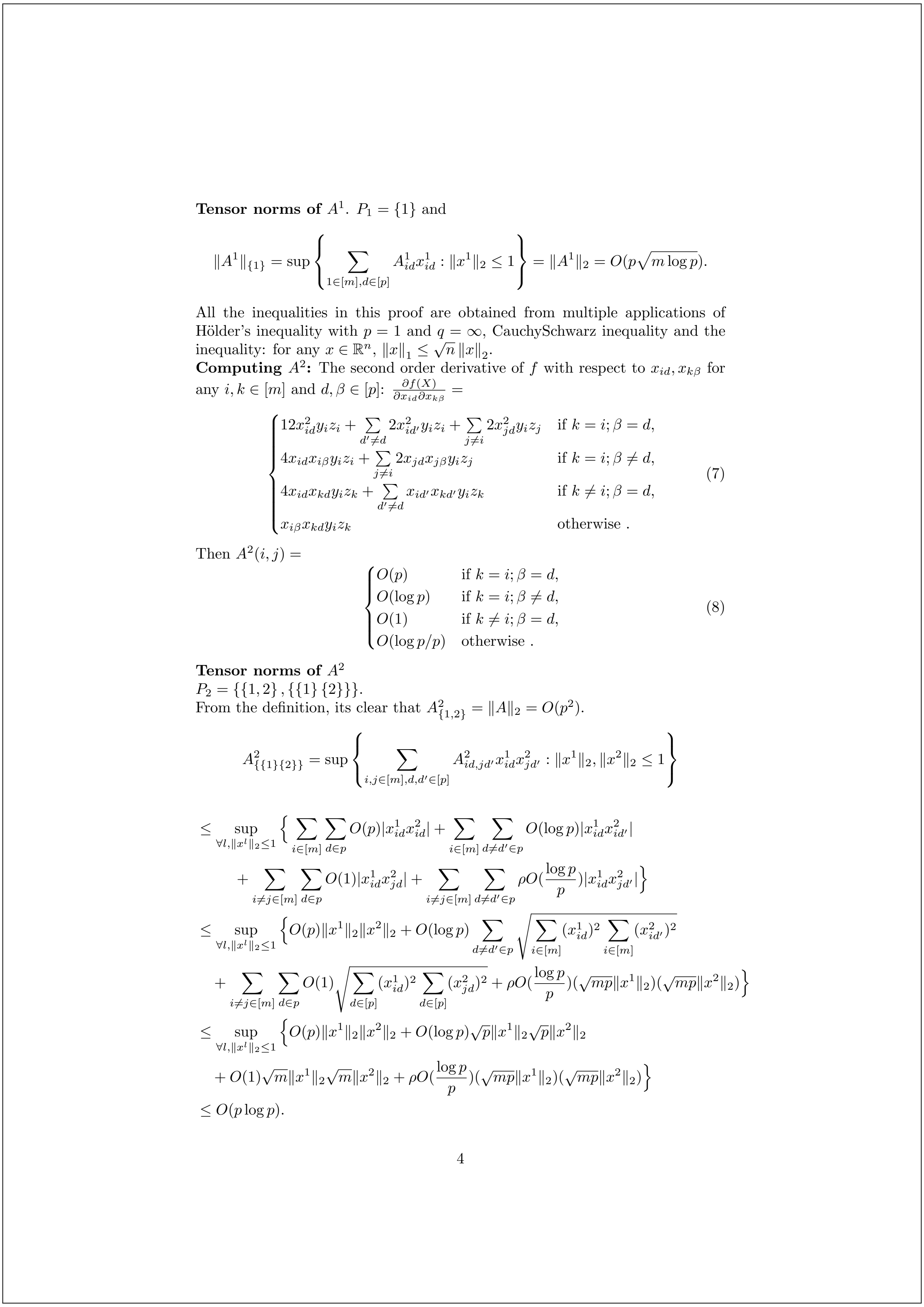
- For each $s \in [4]$, s^{th} order tensors A_s of expectations of s^{th} order derivatives of f with respect to each $\{x_{i_1,d_1}, ..., x_{i_s,d_s}\}_{i_j \in [m], d_j \in [p]}$.
- Tensor norms for A_s with respect to each $\mathcal{J}$ in P_s - the set of all possible partitions of [s].

Computing A^1: The first order derivative of f with respect to x_{id} for any $i \in [m]$ and $d \in [p]$: $\frac{\partial f(X)}{\partial x_{id}} =$

$$4x_{id}^3 y_i z_i + \sum_{d' \neq d} 2x_{id}x_{id'}^2 y_i z_i + \sum_{j \neq i} 2x_{id}x_{jd}^2 y_i z_j + \sum_{d' \neq d}\sum_{j \neq i} x_{id'}x_{jd}x_{jd'}y_i z_j. \quad (5)$$

$$\mathbb{E}(\frac{\partial f(X)}{\partial x_{id}}) = O(\sqrt{p \log p}). \quad (6)$$

Therefore, $A^1 = O(\sqrt{p \log p})\mathbb{J}_{mp}$, where $\mathbb{J}_{mp} \in \mathbb{R}^{mp}$ denotes the vector of ones.



Tensor norms of A^1. $P_1 = \{1\}$ and

$$\|A^1\|_{\{1\}} = \sup\left\{ \sum_{1\in[m],d\in[p]} A^1_{id}x^1_{id} : \|x^1\|_2 \le 1 \right\} = \|A^1\|_2 = O(p\sqrt{m\log p}).$$

All the inequalities in this proof are obtained from multiple applications of Hölder's inequality with $p = 1$ and $q = \infty$, CauchySchwarz inequality and the inequality: for any $x \in \mathbb{R}^n$, $\|x\|_1 \le \sqrt{n}\,\|x\|_2$.

Computing A^2: The second order derivative of f with respect to $x_{id}, x_{k\beta}$ for any $i, k \in [m]$ and $d, \beta \in [p]$: $\frac{\partial f(X)}{\partial x_{id}\partial x_{k\beta}} =$

$$\begin{cases} 12x^2_{id}y_iz_i + \sum_{d'\neq d} 2x^2_{id'}y_iz_i + \sum_{j\neq i} 2x^2_{jd}y_iz_j & \text{if } k = i; \beta = d, \\ 4x_{id}x_{i\beta}y_iz_i + \sum_{j\neq i} 2x_{jd}x_{j\beta}y_iz_j & \text{if } k = i; \beta \neq d, \\ 4x_{id}x_{kd}y_iz_k + \sum_{d'\neq d} x_{id'}x_{kd'}y_iz_k & \text{if } k \neq i; \beta = d, \\ x_{i\beta}x_{kd}y_iz_k & \text{otherwise .} \end{cases} \tag{7}$$

Then $A^2(i,j) =$

$$\begin{cases} O(p) & \text{if } k = i; \beta = d, \\ O(\log p) & \text{if } k = i; \beta \neq d, \\ O(1) & \text{if } k \neq i; \beta = d, \\ O(\log p/p) & \text{otherwise .} \end{cases} \tag{8}$$

Tensor norms of A^2

$P_2 = \{\{1,2\}, \{\{1\}\{2\}\}\}$.

From the definition, its clear that $A^2_{\{1,2\}} = \|A\|_2 = O(p^2)$.

$$A^2_{\{\{1\}\{2\}\}} = \sup\left\{ \sum_{i,j\in[m],d,d'\in[p]} A^2_{id,jd'}x^1_{id}x^2_{jd'} : \|x^1\|_2, \|x^2\|_2 \le 1 \right\}$$

$$\begin{aligned}
&\le \sup_{\forall l, \|x^l\|_2\le 1} \Big\{ \sum_{i\in[m]}\sum_{d\in p} O(p)|x^1_{id}x^2_{id}| + \sum_{i\in[m]}\sum_{d\neq d'\in p} O(\log p)|x^1_{id}x^2_{id'}| \\
&\quad + \sum_{i\neq j\in[m]}\sum_{d\in p} O(1)|x^1_{id}x^2_{jd}| + \sum_{i\neq j\in[m]}\sum_{d\neq d'\in p} \rho O(\frac{\log p}{p})|x^1_{id}x^2_{jd'}| \Big\} \\
&\le \sup_{\forall l, \|x^l\|_2\le 1} \Big\{ O(p)\|x^1\|_2\|x^2\|_2 + O(\log p)\sum_{d\neq d'\in p}\sqrt{\sum_{i\in[m]}(x^1_{id})^2\sum_{i\in[m]}(x^2_{id'})^2} \\
&\quad + \sum_{i\neq j\in[m]}\sum_{d\in p} O(1)\sqrt{\sum_{d\in[p]}(x^1_{id})^2\sum_{d\in[p]}(x^2_{jd})^2} + \rho O(\frac{\log p}{p})(\sqrt{mp}\|x^1\|_2)(\sqrt{mp}\|x^2\|_2) \Big\} \\
&\le \sup_{\forall l, \|x^l\|_2\le 1} \Big\{ O(p)\|x^1\|_2\|x^2\|_2 + O(\log p)\sqrt{p}\|x^1\|_2\sqrt{p}\|x^2\|_2 \\
&\quad + O(1)\sqrt{m}\|x^1\|_2\sqrt{m}\|x^2\|_2 + \rho O(\frac{\log p}{p})(\sqrt{mp}\|x^1\|_2)(\sqrt{mp}\|x^2\|_2) \Big\} \\
&\le O(p\log p).
\end{aligned}$$

Computing A^3: The third order derivative of f with respect to $x_{id}, x_{k\beta}, x_{\alpha l}$ for any $i, k \in [m]$ and $d, \beta \in [p]$: $\frac{\partial f(X)}{\partial x_{id} \partial x_{k\beta} \partial x_{\alpha l}} =$

$$\begin{cases} 24 x_{id} y_i z_i & \text{if } \alpha = k = i; l = \beta = d, \\ 4 x_{il} y_i z_i & \text{if } \alpha = k = i; l \neq \beta = d, \\ 4 x_{\alpha d} y_i z_\alpha & \text{if } \alpha \neq k = i; l = \beta = d, \\ x_{\alpha\beta} y_i z_\alpha & \text{if } \alpha \neq k = i; l = d \neq \beta, \\ 0 & \text{otherwise .} \end{cases} \tag{9}$$

Then $A^3_{id,k\beta,\alpha l} =$

$$\begin{cases} O(\frac{\log p}{p}) y_i z_i & \text{if } \alpha = k = i; l = \beta = d, \\ O(\frac{\log p}{p}) y_i z_i & \text{if } \alpha = k = i; l \neq \beta = d, \\ O(\frac{\log p}{p}) y_i z_\alpha & \text{if } \alpha \neq k = i; l = \beta = d, \\ O(\frac{\log p}{p}) y_i z_\alpha & \text{if } \alpha \neq k = i; l = d \neq \beta, \\ 0 & \text{otherwise .} \end{cases} \tag{10}$$

Tensor norms of A^3: The set of all possible partitions of [3] up to symmetries is $P_3 =$

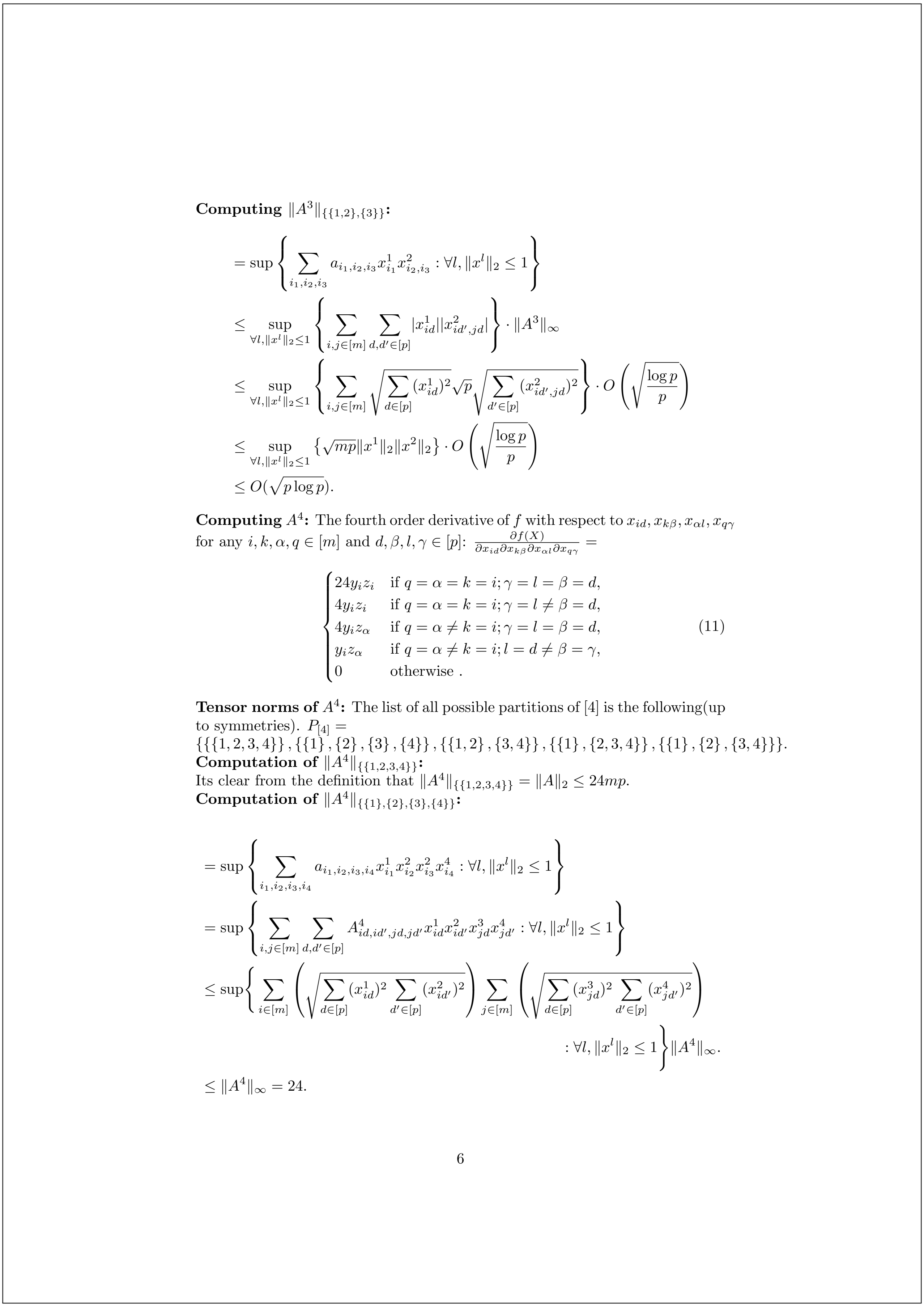
$$\{\{\{1,2,3\}\}, \{\{1\}, \{2\}, \{3\}\}, \{\{1\}, \{2,3\}\}\}$$

.

Computing $\|A^3\|_{\{\{1,2,3\}\}}$: It follows from the definition that $\|A^3\|_{\{\{1,2,3\}\}} = \|A^3\|_2 = O(m\sqrt{p \log p})$.

Computing $\|A^3\|_{\{\{1\},\{2\},\{3\}\}}$:

$$\begin{aligned} &= \sup \left\{ \sum_{i_1, i_2, i_3} a_{i_1, i_2, i_3} x^1_{i_1} x^2_{i_2} x^2_{i_3} : \forall l, \|x^l\|_2 \leq 1 \right\} \\ &\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sum_{i,j \in [m]} \sum_{d,d' \in [p]} |x^1_{id}||x^2_{id'}||x^3_{jd}| \right\} \cdot O\left(\sqrt{\frac{\log p}{p}} \right) \\ &\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sum_{i,j \in [m]} \sqrt{\sum_{d \in [p]} (x^1_{id})^2} \sqrt{\sum_{d \in [p]} (x^2_{jd})^2} \sqrt{p} \sqrt{\sum_{d \in [p]} (x^3_{id'})^2} \right\} \cdot O\left(\sqrt{\frac{\log p}{p}} \right) \\ &\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sqrt{mp} \|x^1\|_2 \|x^2\|_2 \|x^3\|_2 \right\} \cdot O\left(\sqrt{\frac{\log p}{p}} \right) \\ &\leq O(\sqrt{p \log p}). \end{aligned}$$



Computing $\|A^3\|_{\{\{1,2\},\{3\}\}}$:

$$
\begin{aligned}
&= \sup \left\{ \sum_{i_1,i_2,i_3} a_{i_1,i_2,i_3} x^1_{i_1} x^2_{i_2,i_3} : \forall l, \|x^l\|_2 \leq 1 \right\} \\
&\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sum_{i,j\in[m]} \sum_{d,d'\in[p]} |x^1_{id}||x^2_{id',jd}| \right\} \cdot \|A^3\|_\infty \\
&\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sum_{i,j\in[m]} \sqrt{\sum_{d\in[p]} (x^1_{id})^2} \sqrt{p} \sqrt{\sum_{d'\in[p]} (x^2_{id',jd})^2} \right\} \cdot O\left(\sqrt{\frac{\log p}{p}}\right) \\
&\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sqrt{mp} \|x^1\|_2 \|x^2\|_2 \right\} \cdot O\left(\sqrt{\frac{\log p}{p}}\right) \\
&\leq O(\sqrt{p \log p}).
\end{aligned}
$$

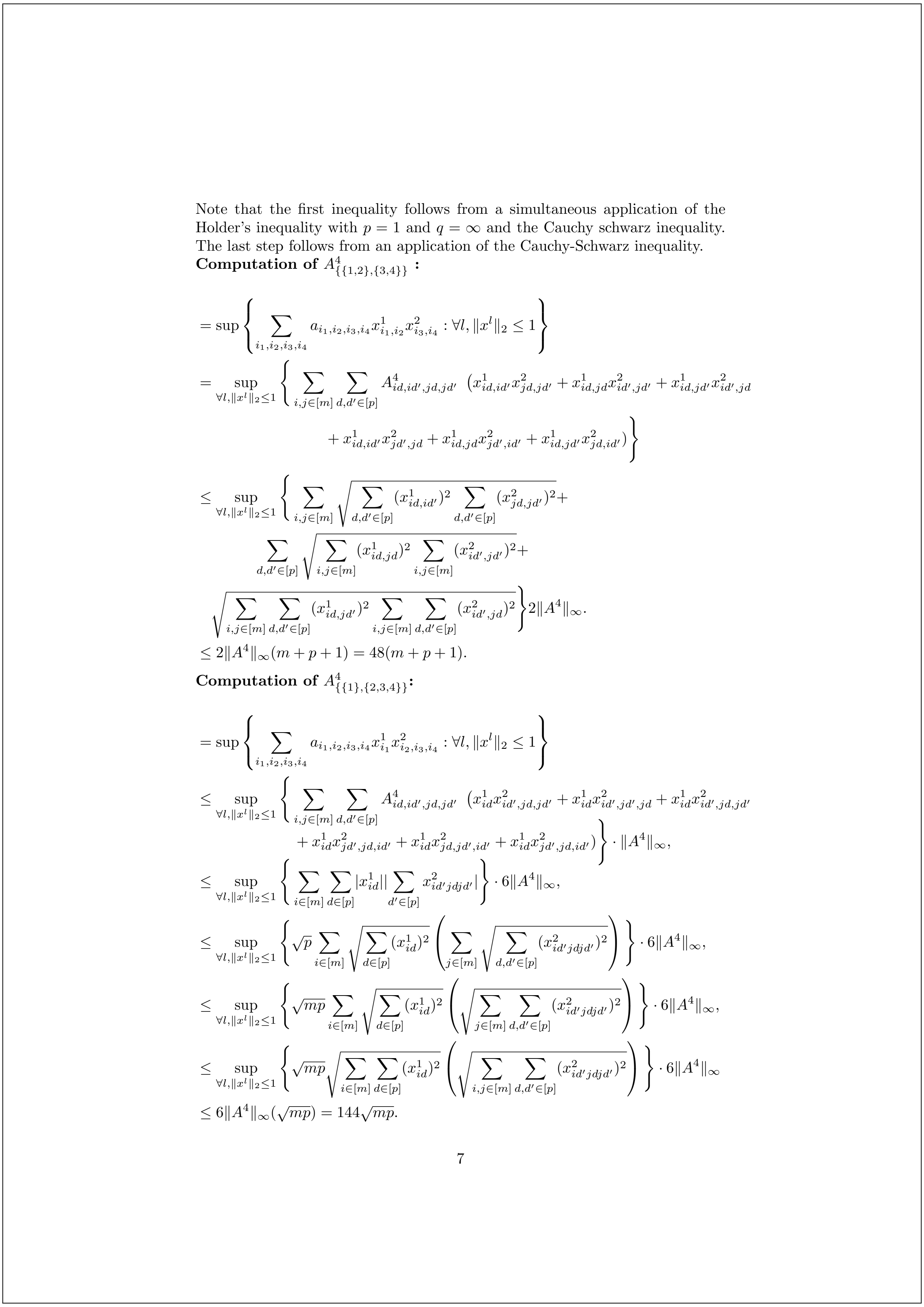
Computing A^4: The fourth order derivative of f with respect to $x_{id}, x_{k\beta}, x_{\alpha l}, x_{q\gamma}$ for any $i, k, \alpha, q \in [m]$ and $d, \beta, l, \gamma \in [p]$: $\frac{\partial f(X)}{\partial x_{id} \partial x_{k\beta} \partial x_{\alpha l} \partial x_{q\gamma}} =$

$$
\begin{cases}
24 y_i z_i & \text{if } q = \alpha = k = i; \gamma = l = \beta = d, \\
4 y_i z_i & \text{if } q = \alpha = k = i; \gamma = l \neq \beta = d, \\
4 y_i z_\alpha & \text{if } q = \alpha \neq k = i; \gamma = l = \beta = d, \\
y_i z_\alpha & \text{if } q = \alpha \neq k = i; l = d \neq \beta = \gamma, \\
0 & \text{otherwise .}
\end{cases}
\tag{11}
$$

Tensor norms of A^4: The list of all possible partitions of [4] is the following(up to symmetries). $P_{[4]} =$ $\{\{\{1,2,3,4\}\}, \{\{1\},\{2\},\{3\},\{4\}\}, \{\{1,2\},\{3,4\}\}, \{\{1\},\{2,3,4\}\}, \{\{1\},\{2\},\{3,4\}\}\}$.
Computation of $\|A^4\|_{\{\{1,2,3,4\}\}}$:
Its clear from the definition that $\|A^4\|_{\{\{1,2,3,4\}\}} = \|A\|_2 \leq 24mp$.
Computation of $\|A^4\|_{\{\{1\},\{2\},\{3\},\{4\}\}}$:

$$
\begin{aligned}
&= \sup \left\{ \sum_{i_1,i_2,i_3,i_4} a_{i_1,i_2,i_3,i_4} x^1_{i_1} x^2_{i_2} x^2_{i_3} x^4_{i_4} : \forall l, \|x^l\|_2 \leq 1 \right\} \\
&= \sup \left\{ \sum_{i,j\in[m]} \sum_{d,d'\in[p]} A^4_{id,id',jd,jd'} x^1_{id} x^2_{id'} x^3_{jd} x^4_{jd'} : \forall l, \|x^l\|_2 \leq 1 \right\} \\
&\leq \sup \left\{ \sum_{i\in[m]} \left(\sqrt{\sum_{d\in[p]} (x^1_{id})^2 \sum_{d'\in[p]} (x^2_{id'})^2} \right) \sum_{j\in[m]} \left(\sqrt{\sum_{d\in[p]} (x^3_{jd})^2 \sum_{d'\in[p]} (x^4_{jd'})^2} \right) \right. \\
&\qquad\qquad\qquad\qquad \left. : \forall l, \|x^l\|_2 \leq 1 \right\} \|A^4\|_\infty . \\
&\leq \|A^4\|_\infty = 24.
\end{aligned}
$$

6



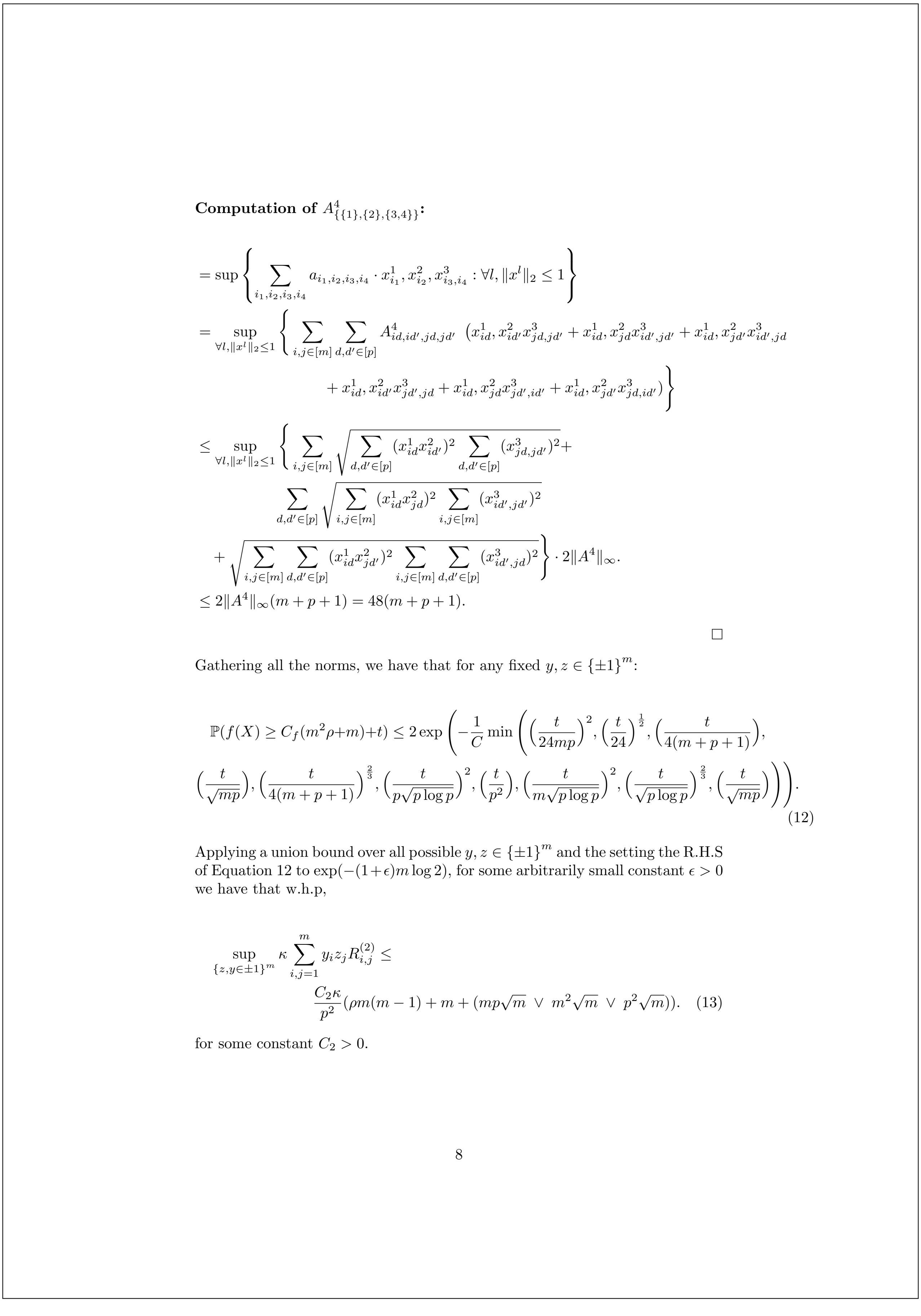
Note that the first inequality follows from a simultaneous application of the Holder's inequality with $p = 1$ and $q = \infty$ and the Cauchy schwarz inequality. The last step follows from an application of the Cauchy-Schwarz inequality.

Computation of $A^4_{\{\{1,2\},\{3,4\}\}}$ **:**

$$
\begin{aligned}
&= \sup \left\{ \sum_{i_1,i_2,i_3,i_4} a_{i_1,i_2,i_3,i_4} x^1_{i_1,i_2} x^2_{i_3,i_4} : \forall l, \|x^l\|_2 \leq 1 \right\} \\
&= \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sum_{i,j\in[m]} \sum_{d,d'\in[p]} A^4_{id,id',jd,jd'} \ \left(x^1_{id,id'} x^2_{jd,jd'} + x^1_{id,jd} x^2_{id',jd'} + x^1_{id,jd'} x^2_{id',jd} \right.\right. \\
&\qquad\qquad \left.\left. + x^1_{id,id'} x^2_{jd',jd} + x^1_{id,jd} x^2_{jd',id'} + x^1_{id,jd'} x^2_{jd,id'} \right) \right\} \\
&\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sum_{i,j\in[m]} \sqrt{\sum_{d,d'\in[p]} (x^1_{id,id'})^2 \sum_{d,d'\in[p]} (x^2_{jd,jd'})^2} + \right. \\
&\qquad \sum_{d,d'\in[p]} \sqrt{\sum_{i,j\in[m]} (x^1_{id,jd})^2 \sum_{i,j\in[m]} (x^2_{id',jd'})^2} + \\
&\left. \sqrt{\sum_{i,j\in[m]} \sum_{d,d'\in[p]} (x^1_{id,jd'})^2 \sum_{i,j\in[m]} \sum_{d,d'\in[p]} (x^2_{id',jd})^2} \right\} 2\|A^4\|_\infty. \\
&\leq 2\|A^4\|_\infty (m+p+1) = 48(m+p+1).
\end{aligned}
$$

Computation of $A^4_{\{\{1\},\{2,3,4\}\}}$**:**

$$
\begin{aligned}
&= \sup \left\{ \sum_{i_1,i_2,i_3,i_4} a_{i_1,i_2,i_3,i_4} x^1_{i_1} x^2_{i_2,i_3,i_4} : \forall l, \|x^l\|_2 \leq 1 \right\} \\
&\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sum_{i,j\in[m]} \sum_{d,d'\in[p]} A^4_{id,id',jd,jd'} \ \left(x^1_{id} x^2_{id',jd,jd'} + x^1_{id} x^2_{id',jd',jd} + x^1_{id} x^2_{id',jd,jd'} \right.\right. \\
&\qquad\qquad \left.\left. + x^1_{id} x^2_{jd',jd,id'} + x^1_{id} x^2_{jd,jd',id'} + x^1_{id} x^2_{jd',jd,id'} \right) \right\} \cdot \|A^4\|_\infty, \\
&\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sum_{i\in[m]} \sum_{d\in[p]} |x^1_{id}| |\sum_{d'\in[p]} x^2_{id'jdjd'}| \right\} \cdot 6\|A^4\|_\infty, \\
&\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sqrt{p} \sum_{i\in[m]} \sqrt{\sum_{d\in[p]} (x^1_{id})^2} \left(\sum_{j\in[m]} \sqrt{\sum_{d,d'\in[p]} (x^2_{id'jdjd'})^2} \right) \right\} \cdot 6\|A^4\|_\infty, \\
&\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sqrt{mp} \sum_{i\in[m]} \sqrt{\sum_{d\in[p]} (x^1_{id})^2} \left(\sqrt{\sum_{j\in[m]} \sum_{d,d'\in[p]} (x^2_{id'jdjd'})^2} \right) \right\} \cdot 6\|A^4\|_\infty, \\
&\leq \sup_{\forall l, \|x^l\|_2 \leq 1} \left\{ \sqrt{mp} \sqrt{\sum_{i\in[m]} \sum_{d\in[p]} (x^1_{id})^2} \left(\sqrt{\sum_{i,j\in[m]} \sum_{d,d'\in[p]} (x^2_{id'jdjd'})^2} \right) \right\} \cdot 6\|A^4\|_\infty \\
&\leq 6\|A^4\|_\infty (\sqrt{mp}) = 144\sqrt{mp}.
\end{aligned}
$$



Computation of $A^4_{\{\{1\},\{2\},\{3,4\}\}}$:

$$
\begin{aligned}
&= \sup\left\{\sum_{i_1,i_2,i_3,i_4} a_{i_1,i_2,i_3,i_4}\cdot x^1_{i_1}, x^2_{i_2}, x^3_{i_3,i_4} : \forall l, \|x^l\|_2 \le 1\right\}\\
&= \sup_{\forall l, \|x^l\|_2\le 1}\left\{\sum_{i,j\in[m]}\sum_{d,d'\in[p]} A^4_{id,id',jd,jd'}\ \left(x^1_{id}, x^2_{id'}x^3_{jd,jd'} + x^1_{id}, x^2_{jd}x^3_{id',jd'} + x^1_{id}, x^2_{jd'}x^3_{id',jd}\right.\right.\\
&\qquad\qquad \left.\left. + x^1_{id}, x^2_{id'}x^3_{jd',jd} + x^1_{id}, x^2_{jd}x^3_{jd',id'} + x^1_{id}, x^2_{jd'}x^3_{jd,id'}\right)\right\}\\
&\le \sup_{\forall l, \|x^l\|_2\le 1}\left\{\sum_{i,j\in[m]}\sqrt{\sum_{d,d'\in[p]}(x^1_{id}x^2_{id'})^2\sum_{d,d'\in[p]}(x^3_{jd,jd'})^2}+\right.\\
&\qquad \sum_{d,d'\in[p]}\sqrt{\sum_{i,j\in[m]}(x^1_{id}x^2_{jd})^2\sum_{i,j\in[m]}(x^3_{id',jd'})^2}\\
&\quad \left. + \sqrt{\sum_{i,j\in[m]}\sum_{d,d'\in[p]}(x^1_{id}x^2_{jd'})^2\sum_{i,j\in[m]}\sum_{d,d'\in[p]}(x^3_{id',jd})^2}\right\}\cdot 2\|A^4\|_\infty.\\
&\le 2\|A^4\|_\infty(m+p+1) = 48(m+p+1).
\end{aligned}
$$

□

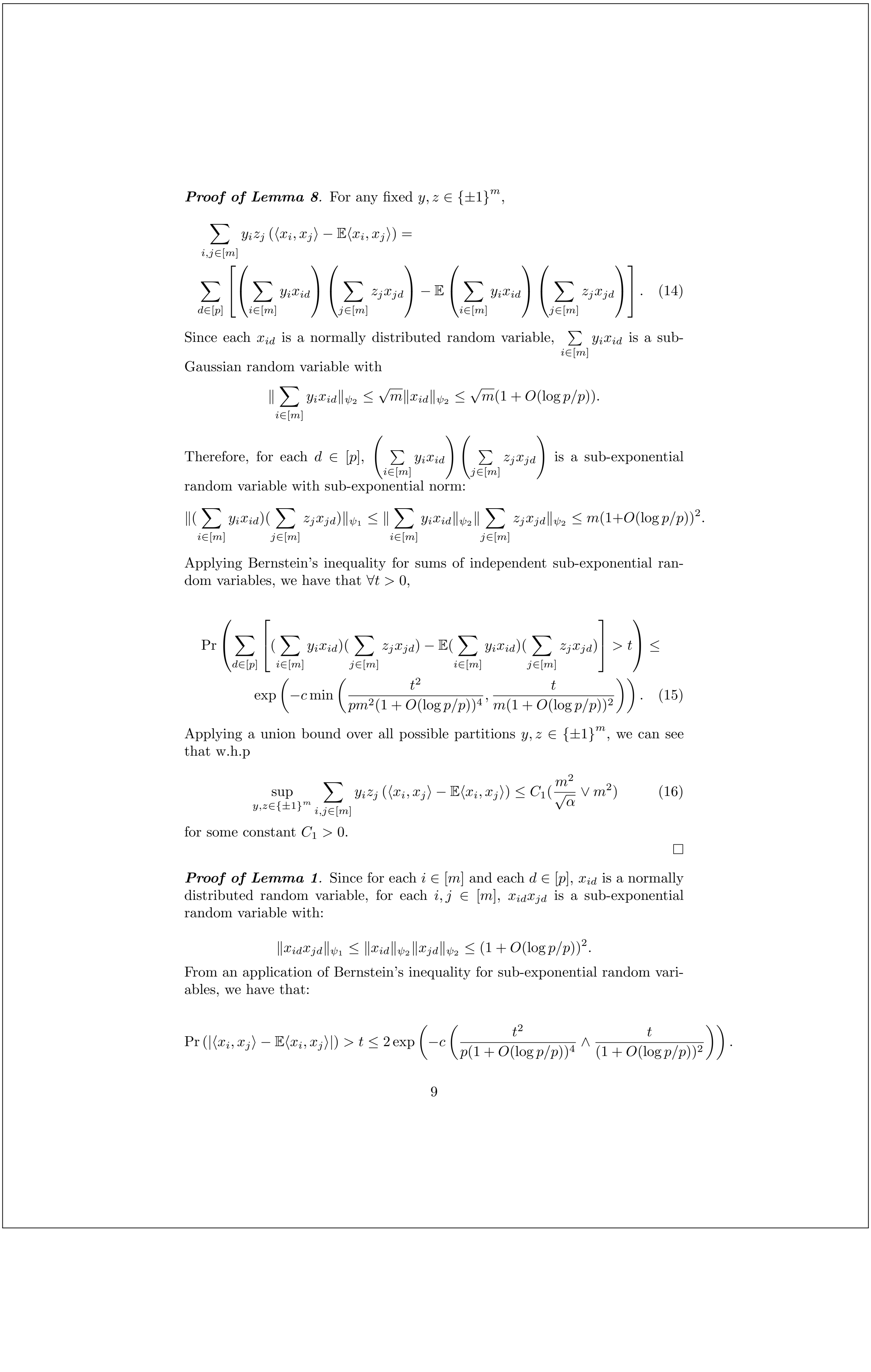
Gathering all the norms, we have that for any fixed $y, z \in \{\pm 1\}^m$:

$$
\mathbb{P}(f(X) \ge C_f(m^2\rho{+}m){+}t) \le 2\exp\left(-\frac{1}{C}\min\left(\left(\frac{t}{24mp}\right)^2, \left(\frac{t}{24}\right)^{\frac{1}{2}}, \left(\frac{t}{4(m+p+1)}\right),\right.\right.
$$
$$
\left.\left.\left(\frac{t}{\sqrt{mp}}\right), \left(\frac{t}{4(m+p+1)}\right)^{\frac{2}{3}}, \left(\frac{t}{p\sqrt{p\log p}}\right)^2, \left(\frac{t}{p^2}\right), \left(\frac{t}{m\sqrt{p\log p}}\right)^2, \left(\frac{t}{\sqrt{p\log p}}\right)^{\frac{2}{3}}, \left(\frac{t}{\sqrt{mp}}\right)\right)\right). \tag{12}
$$

Applying a union bound over all possible $y, z \in \{\pm 1\}^m$ and the setting the R.H.S of Equation 12 to $\exp(-(1{+}\epsilon)m\log 2)$, for some arbitrarily small constant $\epsilon > 0$ we have that w.h.p,

$$
\sup_{\{z,y\in\pm 1\}^m} \kappa\sum_{i,j=1}^m y_i z_j R^{(2)}_{i,j} \le
$$
$$
\frac{C_2\kappa}{p^2}(\rho m(m-1) + m + (mp\sqrt{m}\ \vee\ m^2\sqrt{m}\ \vee\ p^2\sqrt{m})). \tag{13}
$$

for some constant $C_2 > 0$.



Proof of Lemma 8. For any fixed $y, z \in \{\pm 1\}^m$,

$$\sum_{i,j\in[m]} y_i z_j \left(\langle x_i, x_j\rangle - \mathbb{E}\langle x_i, x_j\rangle\right) =$$
$$\sum_{d\in[p]} \left[\left(\sum_{i\in[m]} y_i x_{id}\right)\left(\sum_{j\in[m]} z_j x_{jd}\right) - \mathbb{E}\left(\sum_{i\in[m]} y_i x_{id}\right)\left(\sum_{j\in[m]} z_j x_{jd}\right)\right]. \quad (14)$$

Since each x_{id} is a normally distributed random variable, $\sum_{i\in[m]} y_i x_{id}$ is a sub-Gaussian random variable with

$$\|\sum_{i\in[m]} y_i x_{id}\|_{\psi_2} \le \sqrt{m}\|x_{id}\|_{\psi_2} \le \sqrt{m}(1 + O(\log p/p)).$$

Therefore, for each $d \in [p]$, $\left(\sum_{i\in[m]} y_i x_{id}\right)\left(\sum_{j\in[m]} z_j x_{jd}\right)$ is a sub-exponential random variable with sub-exponential norm:

$$\|(\sum_{i\in[m]} y_i x_{id})(\sum_{j\in[m]} z_j x_{jd})\|_{\psi_1} \le \|\sum_{i\in[m]} y_i x_{id}\|_{\psi_2}\|\sum_{j\in[m]} z_j x_{jd}\|_{\psi_2} \le m(1{+}O(\log p/p))^2.$$

Applying Bernstein's inequality for sums of independent sub-exponential random variables, we have that $\forall t > 0$,

$$\Pr\left(\sum_{d\in[p]}\left[(\sum_{i\in[m]} y_i x_{id})(\sum_{j\in[m]} z_j x_{jd}) - \mathbb{E}(\sum_{i\in[m]} y_i x_{id})(\sum_{j\in[m]} z_j x_{jd})\right] > t\right) \le$$
$$\exp\left(-c \min\left(\frac{t^2}{pm^2(1 + O(\log p/p))^4}, \frac{t}{m(1 + O(\log p/p))^2}\right)\right). \quad (15)$$

Applying a union bound over all possible partitions $y, z \in \{\pm 1\}^m$, we can see that w.h.p

$$\sup_{y,z\in\{\pm 1\}^m} \sum_{i,j\in[m]} y_i z_j \left(\langle x_i, x_j\rangle - \mathbb{E}\langle x_i, x_j\rangle\right) \le C_1(\frac{m^2}{\sqrt{\alpha}} \vee m^2) \quad (16)$$

for some constant $C_1 > 0$.

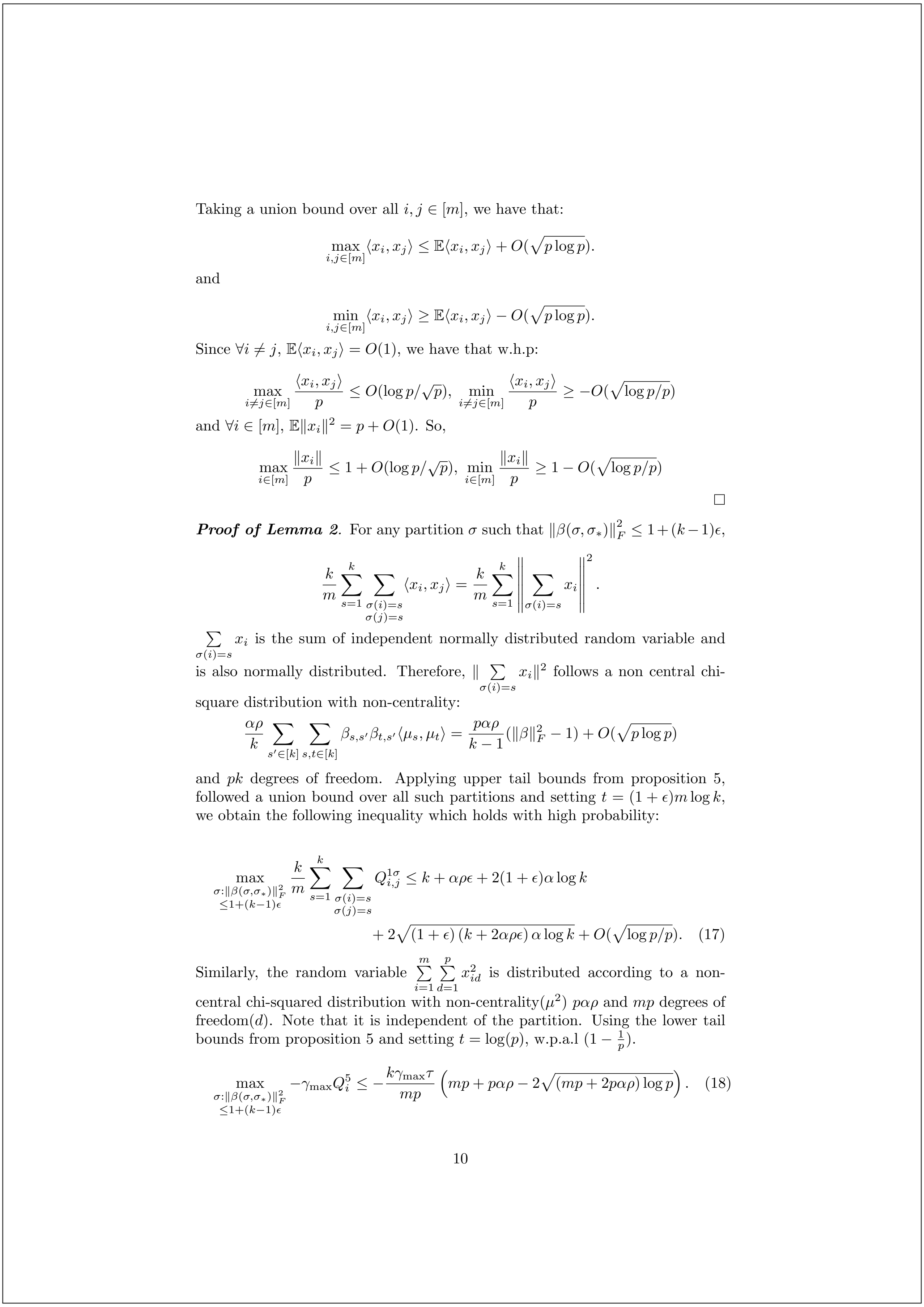
□

Proof of Lemma 1. Since for each $i \in [m]$ and each $d \in [p]$, x_{id} is a normally distributed random variable, for each $i, j \in [m]$, $x_{id}x_{jd}$ is a sub-exponential random variable with:

$$\|x_{id}x_{jd}\|_{\psi_1} \le \|x_{id}\|_{\psi_2}\|x_{jd}\|_{\psi_2} \le (1 + O(\log p/p))^2.$$

From an application of Bernstein's inequality for sub-exponential random variables, we have that:

$$\Pr\left(|\langle x_i, x_j\rangle - \mathbb{E}\langle x_i, x_j\rangle|\right) > t \le 2\exp\left(-c\left(\frac{t^2}{p(1 + O(\log p/p))^4} \wedge \frac{t}{(1 + O(\log p/p))^2}\right)\right).$$



Taking a union bound over all $i, j \in [m]$, we have that:

$$\max_{i,j\in[m]} \langle x_i, x_j\rangle \leq \mathbb{E}\langle x_i, x_j\rangle + O(\sqrt{p\log p}).$$

and

$$\min_{i,j\in[m]} \langle x_i, x_j\rangle \geq \mathbb{E}\langle x_i, x_j\rangle - O(\sqrt{p\log p}).$$

Since $\forall i \neq j$, $\mathbb{E}\langle x_i, x_j\rangle = O(1)$, we have that w.h.p:

$$\max_{i\neq j\in[m]} \frac{\langle x_i, x_j\rangle}{p} \leq O(\log p/\sqrt{p}),\ \min_{i\neq j\in[m]} \frac{\langle x_i, x_j\rangle}{p} \geq -O(\sqrt{\log p/p})$$

and $\forall i \in [m]$, $\mathbb{E}\|x_i\|^2 = p + O(1)$. So,

$$\max_{i\in[m]} \frac{\|x_i\|}{p} \leq 1 + O(\log p/\sqrt{p}),\ \min_{i\in[m]} \frac{\|x_i\|}{p} \geq 1 - O(\sqrt{\log p/p})$$

□

Proof of Lemma 2. For any partition σ such that $\|\beta(\sigma, \sigma_*)\|_F^2 \leq 1 + (k-1)\epsilon$,

$$\frac{k}{m}\sum_{s=1}^{k}\sum_{\substack{\sigma(i)=s\\ \sigma(j)=s}} \langle x_i, x_j\rangle = \frac{k}{m}\sum_{s=1}^{k}\left\|\sum_{\sigma(i)=s} x_i\right\|^2.$$

$\sum_{\sigma(i)=s} x_i$ is the sum of independent normally distributed random variable and is also normally distributed. Therefore, $\|\sum_{\sigma(i)=s} x_i\|^2$ follows a non central chi-square distribution with non-centrality:

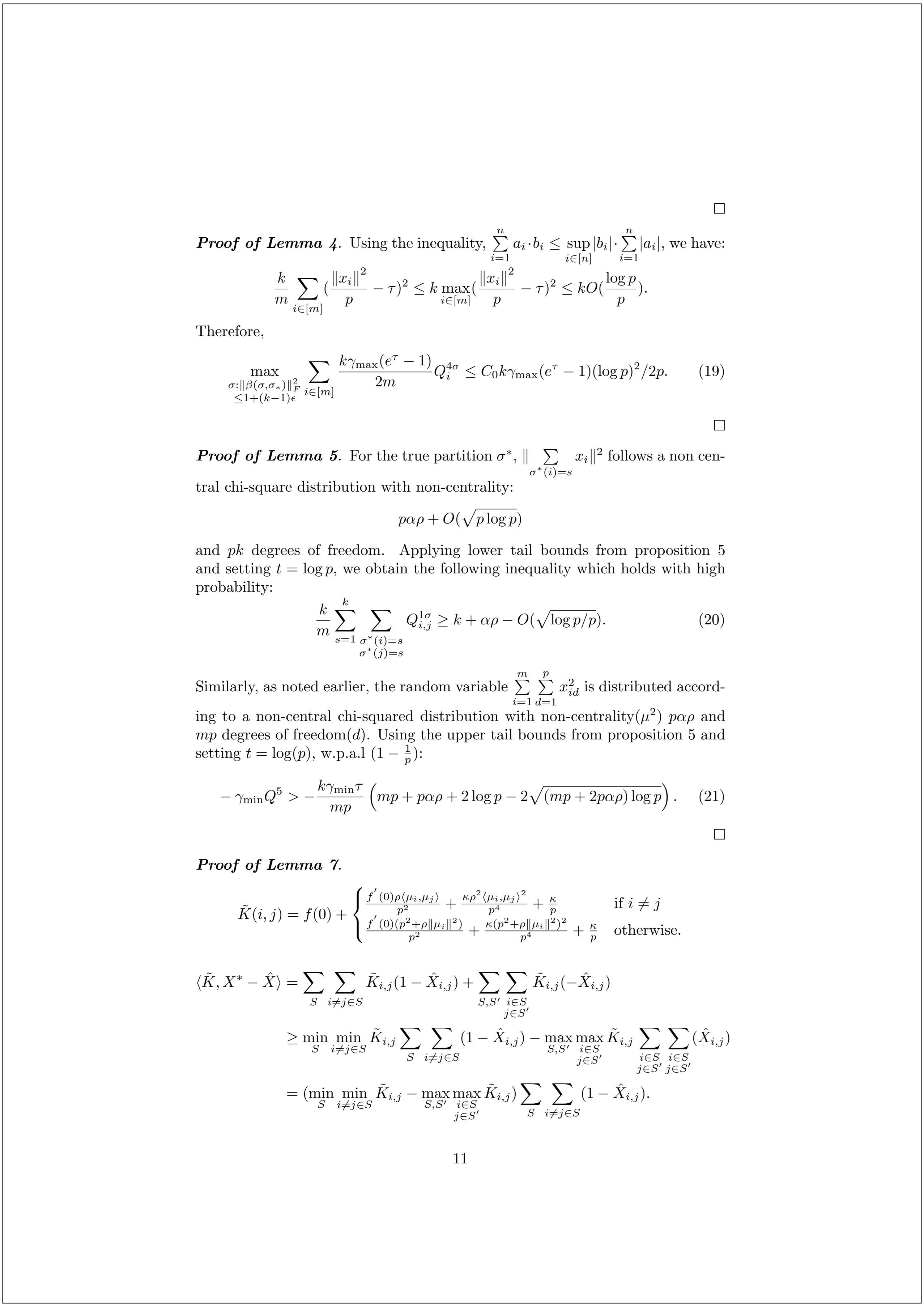
$$\frac{\alpha\rho}{k}\sum_{s'\in[k]}\sum_{s,t\in[k]} \beta_{s,s'}\beta_{t,s'}\langle \mu_s, \mu_t\rangle = \frac{p\alpha\rho}{k-1}(\|\beta\|_F^2 - 1) + O(\sqrt{p\log p})$$

and pk degrees of freedom. Applying upper tail bounds from proposition 5, followed a union bound over all such partitions and setting $t = (1+\epsilon)m\log k$, we obtain the following inequality which holds with high probability:

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} \frac{k}{m}\sum_{s=1}^{k}\sum_{\substack{\sigma(i)=s\\ \sigma(j)=s}} Q_{i,j}^{1\sigma} \leq k + \alpha\rho\epsilon + 2(1+\epsilon)\alpha\log k$$
$$+ 2\sqrt{(1+\epsilon)\,(k+2\alpha\rho\epsilon)\,\alpha\log k} + O(\sqrt{\log p/p}). \quad (17)$$

Similarly, the random variable $\sum_{i=1}^{m}\sum_{d=1}^{p} x_{id}^2$ is distributed according to a non-central chi-squared distribution with non-centrality(μ^2) $p\alpha\rho$ and mp degrees of freedom(d). Note that it is independent of the partition. Using the lower tail bounds from proposition 5 and setting $t = \log(p)$, w.p.a.l $(1 - \frac{1}{p})$.

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2\\ \leq 1+(k-1)\epsilon}} -\gamma_{\max} Q_i^5 \leq -\frac{k\gamma_{\max}\tau}{mp}\left(mp + p\alpha\rho - 2\sqrt{(mp+2p\alpha\rho)\log p}\right). \quad (18)$$



□

Proof of Lemma 4. Using the inequality, $\sum_{i=1}^{n} a_i \cdot b_i \leq \sup_{i\in[n]} |b_i| \cdot \sum_{i=1}^{n} |a_i|$, we have:

$$\frac{k}{m} \sum_{i\in[m]} (\frac{\|x_i\|^2}{p} - \tau)^2 \leq k \max_{i\in[m]} (\frac{\|x_i\|^2}{p} - \tau)^2 \leq kO(\frac{\log p}{p}).$$

Therefore,

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \leq 1+(k-1)\epsilon}} \sum_{i\in[m]} \frac{k\gamma_{\max}(e^{\tau}-1)}{2m} Q_i^{4\sigma} \leq C_0 k\gamma_{\max}(e^{\tau}-1)(\log p)^2/2p. \tag{19}$$

□

Proof of Lemma 5. For the true partition σ^*, $\| \sum_{\sigma^*(i)=s} x_i\|^2$ follows a non central chi-square distribution with non-centrality:

$$p\alpha\rho + O(\sqrt{p\log p})$$

and pk degrees of freedom. Applying lower tail bounds from proposition 5 and setting $t = \log p$, we obtain the following inequality which holds with high probability:

$$\frac{k}{m} \sum_{s=1}^{k} \sum_{\substack{\sigma^*(i)=s \\ \sigma^*(j)=s}} Q_{i,j}^{1\sigma} \geq k + \alpha\rho - O(\sqrt{\log p/p}). \tag{20}$$

Similarly, as noted earlier, the random variable $\sum_{i=1}^{m} \sum_{d=1}^{p} x_{id}^2$ is distributed according to a non-central chi-squared distribution with non-centrality(μ^2) $p\alpha\rho$ and mp degrees of freedom(d). Using the upper tail bounds from proposition 5 and setting $t = \log(p)$, w.p.a.l $(1 - \frac{1}{p})$:

$$-\gamma_{\min} Q^5 > -\frac{k\gamma_{\min}\tau}{mp} \left(mp + p\alpha\rho + 2\log p - 2\sqrt{(mp + 2p\alpha\rho)\log p} \right). \tag{21}$$

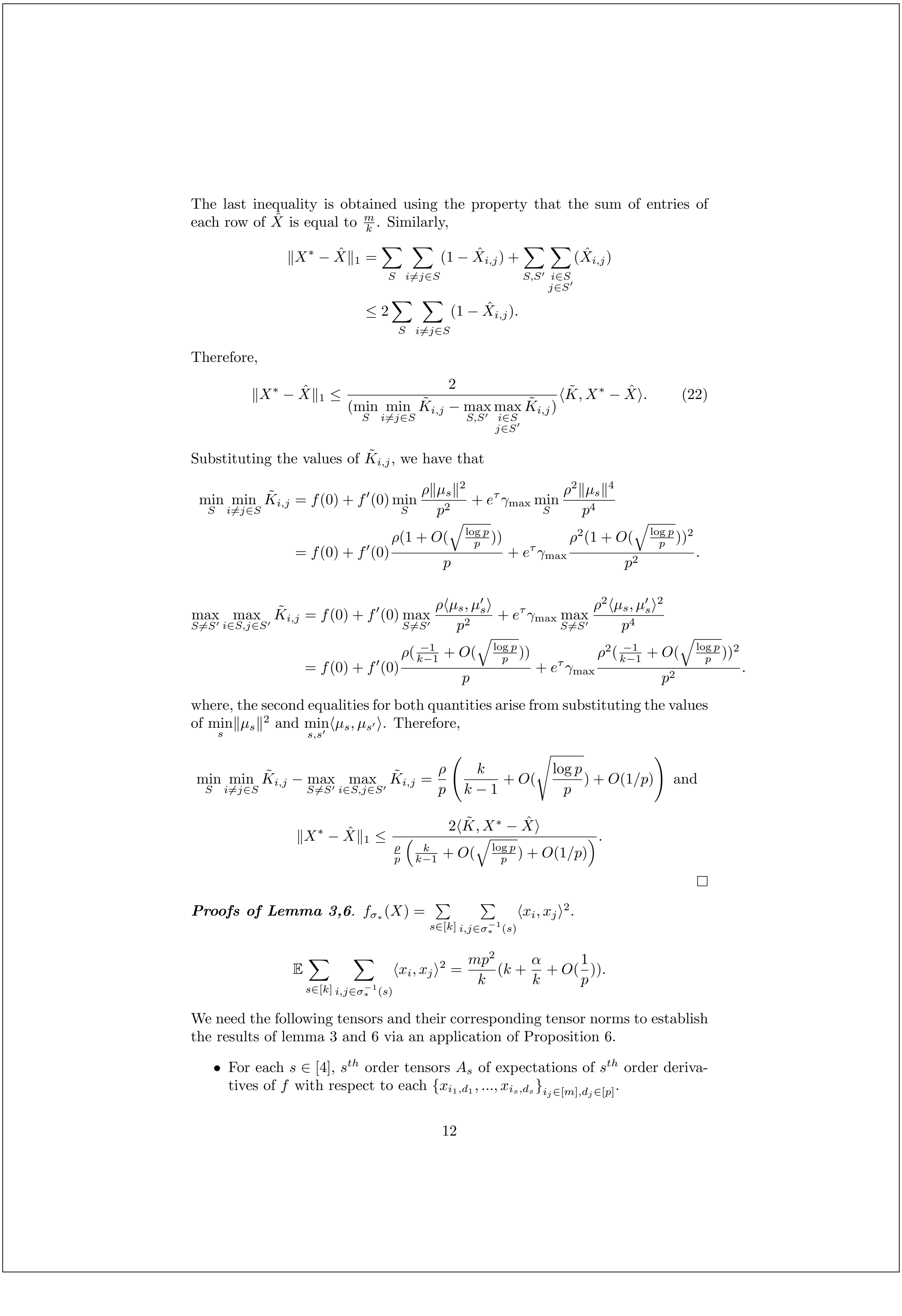
□

Proof of Lemma 7.

$$\tilde{K}(i,j) = f(0) + \begin{cases} \frac{f^{'}(0)\rho\langle\mu_i,\mu_j\rangle}{p^2} + \frac{\kappa\rho^2\langle\mu_i,\mu_j\rangle^2}{p^4} + \frac{\kappa}{p} & \text{if } i \neq j \\ \frac{f^{'}(0)(p^2+\rho\|\mu_i\|^2)}{p^2} + \frac{\kappa(p^2+\rho\|\mu_i\|^2)^2}{p^4} + \frac{\kappa}{p} & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \langle\tilde{K}, X^* - \hat{X}\rangle &= \sum_{S} \sum_{i\neq j\in S} \tilde{K}_{i,j}(1 - \hat{X}_{i,j}) + \sum_{S,S'} \sum_{\substack{i\in S \\ j\in S'}} \tilde{K}_{i,j}(-\hat{X}_{i,j}) \\ &\geq \min_{S} \min_{i\neq j\in S} \tilde{K}_{i,j} \sum_{S} \sum_{i\neq j\in S} (1 - \hat{X}_{i,j}) - \max_{S,S'} \max_{\substack{i\in S \\ j\in S'}} \tilde{K}_{i,j} \sum_{\substack{i\in S \\ j\in S'}} \sum_{\substack{i\in S \\ j\in S'}} (\hat{X}_{i,j}) \\ &= (\min_{S} \min_{i\neq j\in S} \tilde{K}_{i,j} - \max_{S,S'} \max_{\substack{i\in S \\ j\in S'}} \tilde{K}_{i,j}) \sum_{S} \sum_{i\neq j\in S} (1 - \hat{X}_{i,j}). \end{aligned}$$

11



The last inequality is obtained using the property that the sum of entries of each row of $\hat{X}$ is equal to $\frac{m}{k}$. Similarly,

$$\begin{aligned}\|X^* - \hat{X}\|_1 &= \sum_{S}\sum_{i\neq j\in S}(1-\hat{X}_{i,j}) + \sum_{S,S'}\sum_{\substack{i\in S\\ j\in S'}}(\hat{X}_{i,j})\\ &\leq 2\sum_{S}\sum_{i\neq j\in S}(1-\hat{X}_{i,j}).\end{aligned}$$

Therefore,

$$\|X^* - \hat{X}\|_1 \leq \frac{2}{(\min\limits_{S}\min\limits_{i\neq j\in S}\tilde{K}_{i,j} - \max\limits_{S,S'}\max\limits_{\substack{i\in S\\ j\in S'}}\tilde{K}_{i,j})}\langle \tilde{K}, X^* - \hat{X}\rangle. \qquad (22)$$

Substituting the values of $\tilde{K}_{i,j}$, we have that

$$\begin{aligned}\min_{S}\min_{i\neq j\in S}\tilde{K}_{i,j} &= f(0) + f'(0)\min_{S}\frac{\rho\|\mu_s\|^2}{p^2} + e^{\tau}\gamma_{\max}\min_{S}\frac{\rho^2\|\mu_s\|^4}{p^4}\\ &= f(0) + f'(0)\frac{\rho(1+O(\sqrt{\frac{\log p}{p}}))}{p} + e^{\tau}\gamma_{\max}\frac{\rho^2(1+O(\sqrt{\frac{\log p}{p}}))^2}{p^2}.\end{aligned}$$

$$\begin{aligned}\max_{S\neq S'}\max_{i\in S,j\in S'}\tilde{K}_{i,j} &= f(0) + f'(0)\max_{S\neq S'}\frac{\rho\langle\mu_s,\mu_s'\rangle}{p^2} + e^{\tau}\gamma_{\max}\max_{S\neq S'}\frac{\rho^2\langle\mu_s,\mu_s'\rangle^2}{p^4}\\ &= f(0) + f'(0)\frac{\rho(\frac{-1}{k-1}+O(\sqrt{\frac{\log p}{p}}))}{p} + e^{\tau}\gamma_{\max}\frac{\rho^2(\frac{-1}{k-1}+O(\sqrt{\frac{\log p}{p}}))^2}{p^2}.\end{aligned}$$

where, the second equalities for both quantities arise from substituting the values of $\min\limits_{s}\|\mu_s\|^2$ and $\min\limits_{s,s'}\langle\mu_s,\mu_{s'}\rangle$. Therefore,

$$\min_{S}\min_{i\neq j\in S}\tilde{K}_{i,j} - \max_{S\neq S'}\max_{i\in S,j\in S'}\tilde{K}_{i,j} = \frac{\rho}{p}\left(\frac{k}{k-1} + O(\sqrt{\frac{\log p}{p}}) + O(1/p)\right) \text{ and}$$

$$\|X^* - \hat{X}\|_1 \leq \frac{2\langle\tilde{K}, X^* - \hat{X}\rangle}{\frac{\rho}{p}\left(\frac{k}{k-1} + O(\sqrt{\frac{\log p}{p}}) + O(1/p)\right)}.$$

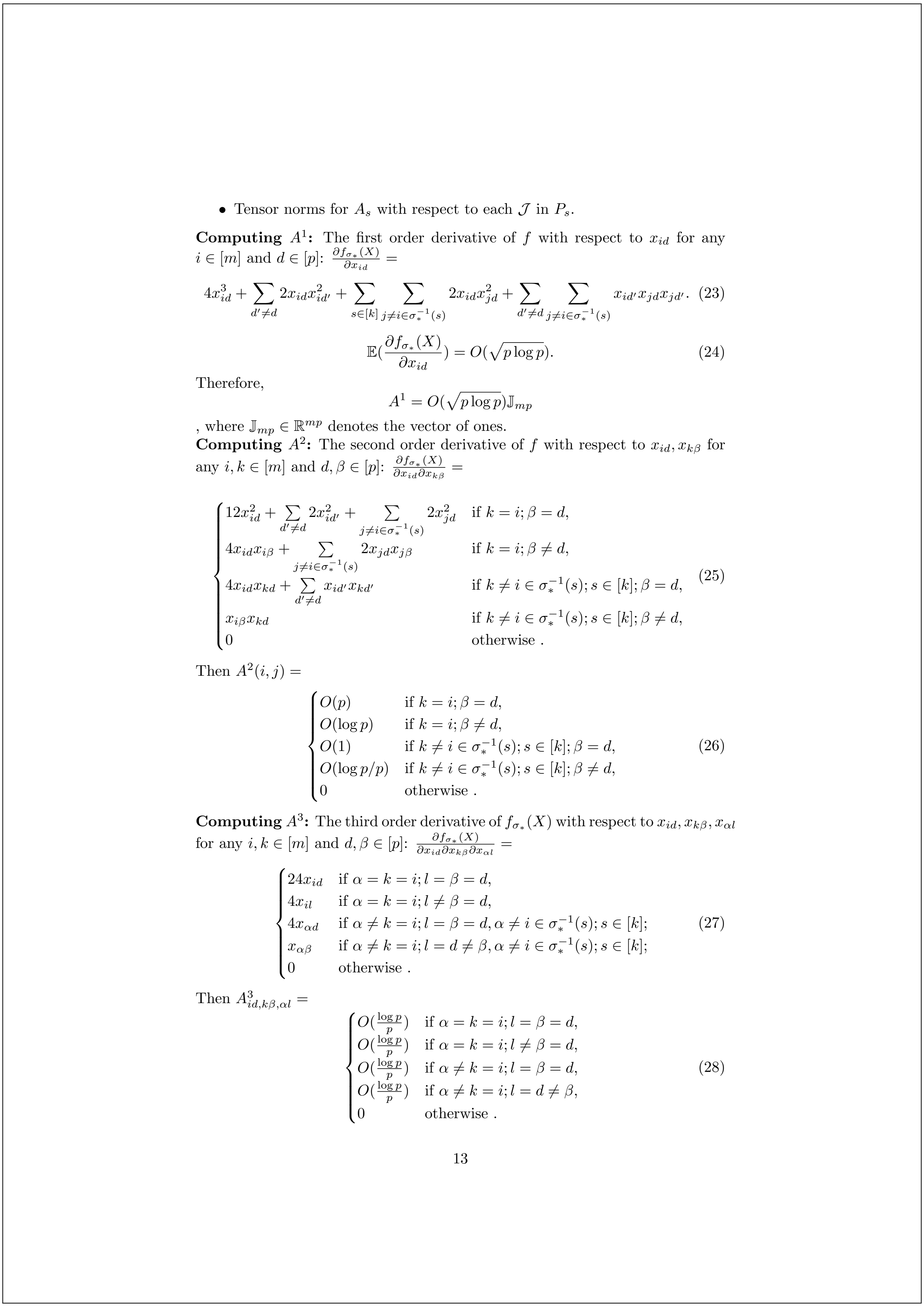
□

Proofs of Lemma 3,6. $f_{\sigma_*}(X) = \sum\limits_{s\in[k]}\sum\limits_{i,j\in\sigma_*^{-1}(s)}\langle x_i, x_j\rangle^2$.

$$\mathbb{E}\sum_{s\in[k]}\sum_{i,j\in\sigma_*^{-1}(s)}\langle x_i, x_j\rangle^2 = \frac{mp^2}{k}(k + \frac{\alpha}{k} + O(\frac{1}{p})).$$

We need the following tensors and their corresponding tensor norms to establish the results of lemma 3 and 6 via an application of Proposition 6.

- For each $s\in[4]$, s^{th} order tensors A_s of expectations of s^{th} order derivatives of f with respect to each $\{x_{i_1,d_1},...,x_{i_s,d_s}\}_{i_j\in[m],d_j\in[p]}$.



- Tensor norms for A_s with respect to each $\mathcal{J}$ in P_s.

Computing A^1: The first order derivative of f with respect to x_{id} for any $i \in [m]$ and $d \in [p]$: $\frac{\partial f_{\sigma_*}(X)}{\partial x_{id}} =$

$$4x_{id}^3 + \sum_{d'\neq d} 2x_{id}x_{id'}^2 + \sum_{s\in[k]}\sum_{j\neq i\in\sigma_*^{-1}(s)} 2x_{id}x_{jd}^2 + \sum_{d'\neq d}\sum_{j\neq i\in\sigma_*^{-1}(s)} x_{id'}x_{jd}x_{jd'}. \tag{23}$$

$$\mathbb{E}(\frac{\partial f_{\sigma_*}(X)}{\partial x_{id}}) = O(\sqrt{p\log p}). \tag{24}$$

Therefore,

$$A^1 = O(\sqrt{p\log p})\mathbb{J}_{mp}$$

, where $\mathbb{J}_{mp} \in \mathbb{R}^{mp}$ denotes the vector of ones.

Computing A^2: The second order derivative of f with respect to $x_{id}, x_{k\beta}$ for any $i, k \in [m]$ and $d, \beta \in [p]$: $\frac{\partial f_{\sigma_*}(X)}{\partial x_{id}\partial x_{k\beta}} =$

$$\begin{cases} 12x_{id}^2 + \sum\limits_{d'\neq d} 2x_{id'}^2 + \sum\limits_{j\neq i\in\sigma_*^{-1}(s)} 2x_{jd}^2 & \text{if } k = i; \beta = d, \\ 4x_{id}x_{i\beta} + \sum\limits_{j\neq i\in\sigma_*^{-1}(s)} 2x_{jd}x_{j\beta} & \text{if } k = i; \beta \neq d, \\ 4x_{id}x_{kd} + \sum\limits_{d'\neq d} x_{id'}x_{kd'} & \text{if } k \neq i \in \sigma_*^{-1}(s); s \in [k]; \beta = d, \\ x_{i\beta}x_{kd} & \text{if } k \neq i \in \sigma_*^{-1}(s); s \in [k]; \beta \neq d, \\ 0 & \text{otherwise .} \end{cases} \tag{25}$$

Then $A^2(i,j) =$

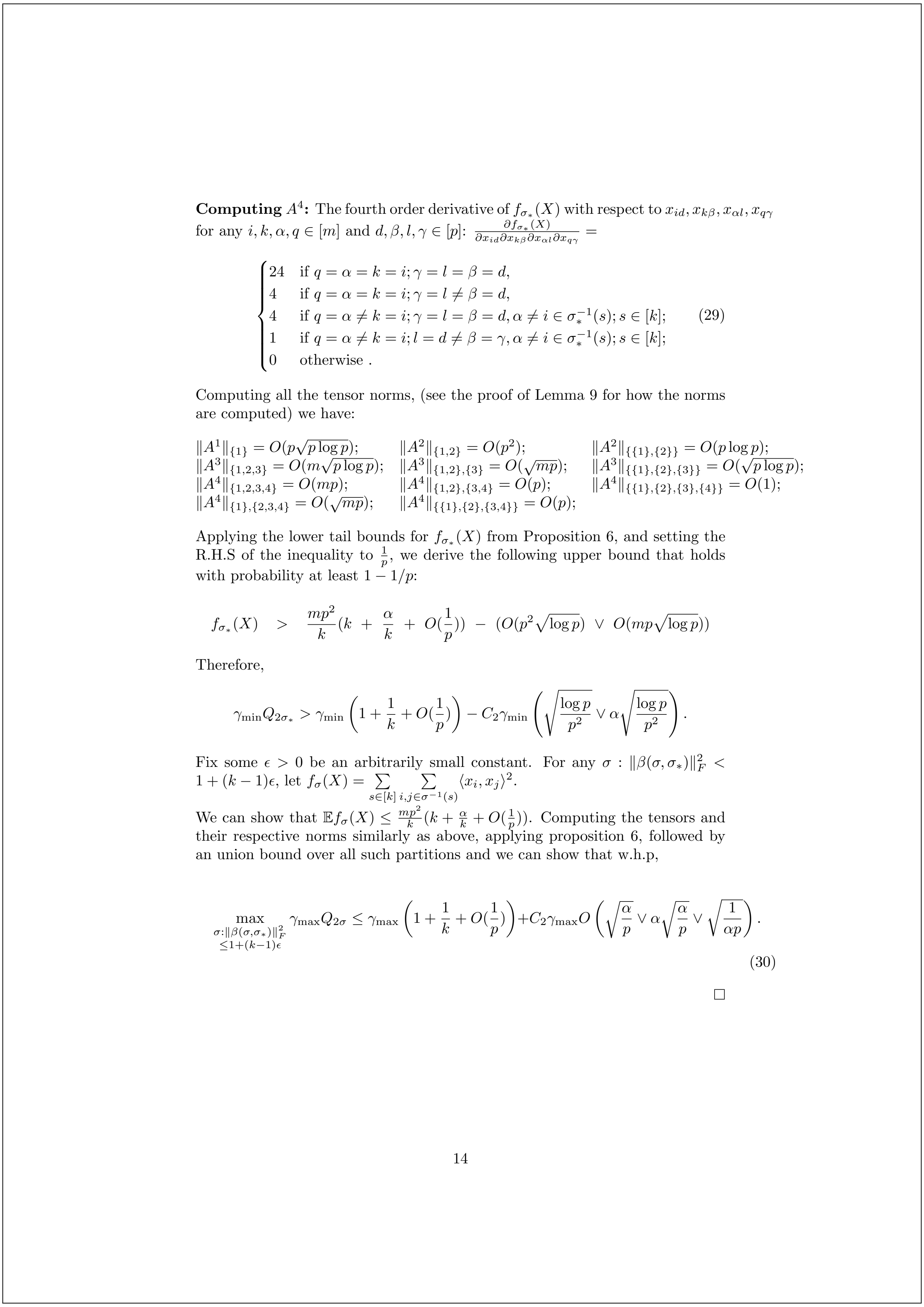
$$\begin{cases} O(p) & \text{if } k = i; \beta = d, \\ O(\log p) & \text{if } k = i; \beta \neq d, \\ O(1) & \text{if } k \neq i \in \sigma_*^{-1}(s); s \in [k]; \beta = d, \\ O(\log p/p) & \text{if } k \neq i \in \sigma_*^{-1}(s); s \in [k]; \beta \neq d, \\ 0 & \text{otherwise .} \end{cases} \tag{26}$$

Computing A^3: The third order derivative of $f_{\sigma_*}(X)$ with respect to $x_{id}, x_{k\beta}, x_{\alpha l}$ for any $i, k \in [m]$ and $d, \beta \in [p]$: $\frac{\partial f_{\sigma_*}(X)}{\partial x_{id}\partial x_{k\beta}\partial x_{\alpha l}} =$

$$\begin{cases} 24x_{id} & \text{if } \alpha = k = i; l = \beta = d, \\ 4x_{il} & \text{if } \alpha = k = i; l \neq \beta = d, \\ 4x_{\alpha d} & \text{if } \alpha \neq k = i; l = \beta = d, \alpha \neq i \in \sigma_*^{-1}(s); s \in [k]; \\ x_{\alpha\beta} & \text{if } \alpha \neq k = i; l = d \neq \beta, \alpha \neq i \in \sigma_*^{-1}(s); s \in [k]; \\ 0 & \text{otherwise .} \end{cases} \tag{27}$$

Then $A^3_{id,k\beta,\alpha l} =$

$$\begin{cases} O(\frac{\log p}{p}) & \text{if } \alpha = k = i; l = \beta = d, \\ O(\frac{\log p}{p}) & \text{if } \alpha = k = i; l \neq \beta = d, \\ O(\frac{\log p}{p}) & \text{if } \alpha \neq k = i; l = \beta = d, \\ O(\frac{\log p}{p}) & \text{if } \alpha \neq k = i; l = d \neq \beta, \\ 0 & \text{otherwise .} \end{cases} \tag{28}$$



Computing A^4: The fourth order derivative of $f_{\sigma_*}(X)$ with respect to $x_{id}, x_{k\beta}, x_{\alpha l}, x_{q\gamma}$ for any $i, k, \alpha, q \in [m]$ and $d, \beta, l, \gamma \in [p]$: $\frac{\partial f_{\sigma_*}(X)}{\partial x_{id} \partial x_{k\beta} \partial x_{\alpha l} \partial x_{q\gamma}} =$

$$\begin{cases} 24 & \text{if } q = \alpha = k = i; \gamma = l = \beta = d, \\ 4 & \text{if } q = \alpha = k = i; \gamma = l \neq \beta = d, \\ 4 & \text{if } q = \alpha \neq k = i; \gamma = l = \beta = d, \alpha \neq i \in \sigma_*^{-1}(s); s \in [k]; \\ 1 & \text{if } q = \alpha \neq k = i; l = d \neq \beta = \gamma, \alpha \neq i \in \sigma_*^{-1}(s); s \in [k]; \\ 0 & \text{otherwise .} \end{cases} \tag{29}$$

Computing all the tensor norms, (see the proof of Lemma 9 for how the norms are computed) we have:

$$\begin{array}{lll} \|A^1\|_{\{1\}} = O(p\sqrt{p\log p}); & \|A^2\|_{\{1,2\}} = O(p^2); & \|A^2\|_{\{\{1\},\{2\}\}} = O(p\log p); \\ \|A^3\|_{\{1,2,3\}} = O(m\sqrt{p\log p}); & \|A^3\|_{\{1,2\},\{3\}} = O(\sqrt{mp}); & \|A^3\|_{\{\{1\},\{2\},\{3\}\}} = O(\sqrt{p\log p}); \\ \|A^4\|_{\{1,2,3,4\}} = O(mp); & \|A^4\|_{\{1,2\},\{3,4\}} = O(p); & \|A^4\|_{\{\{1\},\{2\},\{3\},\{4\}\}} = O(1); \\ \|A^4\|_{\{1\},\{2,3,4\}} = O(\sqrt{mp}); & \|A^4\|_{\{\{1\},\{2\},\{3,4\}\}} = O(p); & \end{array}$$

Applying the lower tail bounds for $f_{\sigma_*}(X)$ from Proposition 6, and setting the R.H.S of the inequality to $\frac{1}{p}$, we derive the following upper bound that holds with probability at least $1 - 1/p$:

$$f_{\sigma_*}(X) \quad > \quad \frac{mp^2}{k}(k \; + \; \frac{\alpha}{k} \; + \; O(\frac{1}{p})) \; - \; (O(p^2\sqrt{\log p}) \;\vee\; O(mp\sqrt{\log p}))$$

Therefore,

$$\gamma_{\min} Q_{2\sigma_*} > \gamma_{\min}\left(1 + \frac{1}{k} + O(\frac{1}{p})\right) - C_2\gamma_{\min}\left(\sqrt{\frac{\log p}{p^2}} \vee \alpha\sqrt{\frac{\log p}{p^2}}\right).$$

Fix some $\epsilon > 0$ be an arbitrarily small constant. For any $\sigma : \|\beta(\sigma, \sigma_*)\|_F^2 < 1 + (k-1)\epsilon$, let $f_\sigma(X) = \sum_{s \in [k]} \sum_{i,j \in \sigma^{-1}(s)} \langle x_i, x_j \rangle^2$.

We can show that $\mathbb{E} f_\sigma(X) \leq \frac{mp^2}{k}(k + \frac{\alpha}{k} + O(\frac{1}{p}))$. Computing the tensors and their respective norms similarly as above, applying proposition 6, followed by an union bound over all such partitions and we can show that w.h.p,

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \leq 1+(k-1)\epsilon}} \gamma_{\max} Q_{2\sigma} \leq \gamma_{\max}\left(1 + \frac{1}{k} + O(\frac{1}{p})\right) + C_2\gamma_{\max} O\left(\sqrt{\frac{\alpha}{p}} \vee \alpha\sqrt{\frac{\alpha}{p}} \vee \sqrt{\frac{1}{\alpha p}}\right). \tag{30}$$

□

7
Kernel clustering under non-parametric mixture models

Recovery Guarantees for Kernel-based Clustering under Non-parametric Mixture Models

Leena C. Vankadara[1] **Sebastian Bordt**[1,2] **Ulrike von Luxburg**[1,2] **Debarghya Ghoshdastidar**[3]

University of Tübingen[1]
Max Planck Institute for Intelligent Systems, Tübingen[2]
Technical University of Munich[3]

Abstract

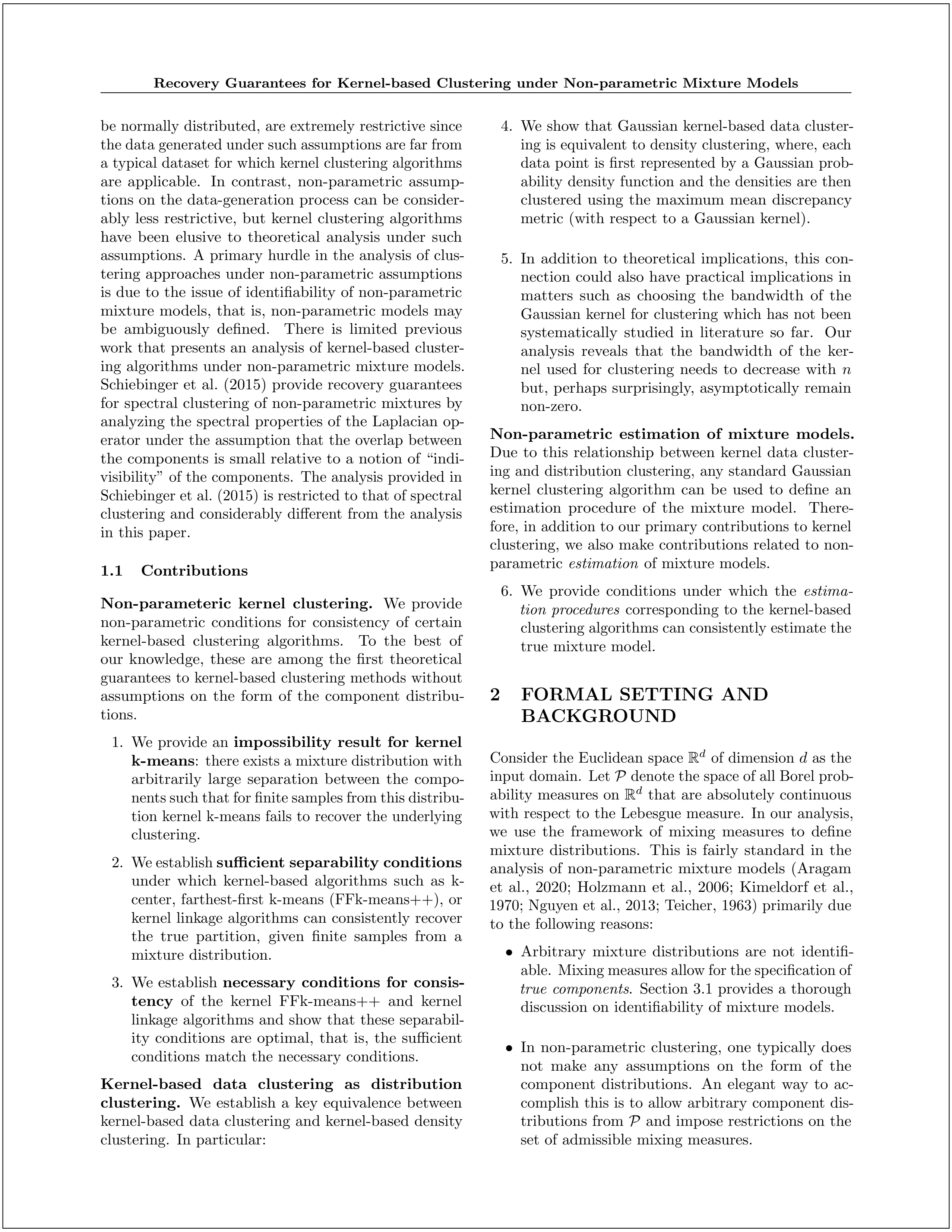
Despite the ubiquity of kernel-based clustering, surprisingly few statistical guarantees exist beyond settings that consider strong structural assumptions on the data generation process. In this work, we take a step towards bridging this gap by studying the statistical performance of kernel-based clustering algorithms under non-parametric mixture models. We provide necessary and sufficient separability conditions under which these algorithms can consistently recover the underlying *true clustering.* Our analysis provides guarantees for kernel clustering approaches without structural assumptions on the form of the component distributions. Additionally, we establish a key equivalence between kernel-based data-clustering and kernel density-based clustering. This enables us to provide consistency guarantees for kernel-based estimators of non-parametric mixture models. Along with theoretical implications, this connection could have practical implications, including in the systematic choice of the bandwidth of the Gaussian kernel in the context of clustering.

1 INTRODUCTION

Clustering refers to the unsupervised task of partitioning a given data sample or the input space into *meaningful* regions. Kernel clustering approaches such as kernel k-means (Dhillon et al., 2004) and kernel spectral clustering (Ng et al., 2002) are widely adopted by practitioners, particularly for partitioning non-spherical complex cluster structures. Beyond their good practical behavior, kernel methods are appealing due to their amenability to theoretical analysis. However, as an anomaly, kernel clustering has been elusive to theoretical analysis, in particular, under general non-parametric assumptions on the data generation process. One of the principle sources for this gap between theory and practice had been the lack of a universally accepted characterization of the quality of a clustering. One popular notion of the goodness of clustering is defined as the one that consistently partitions the data space. Consistency is, however, only a necessary condition for clustering algorithms. It simply checks if an algorithm asymptotically converges to a limiting partition. The optimality of this limiting partition is not studied under consistency. As an example, spectral clustering has been shown to be consistent (Luxburg et al., 2008) for any similarity function k. However, if one uses a similarity function based on an uninformative kernel such as the identity kernel, then the obtained limiting partition is clearly not guaranteed to be a desirable one. Density based clustering (Hartigan, 1975; Hartigan, 1981; Rinaldo et al., 2010) is another popular line of work with theoretical backing, where clusters are defined as connected components of high-density regions, referred to as density level sets. The imprecise notion of a high-density region is overcome using the so called cluster-tree approach (Chaudhuri et al., 2014; Sriperumbudur and Steinwart, 2012), where a continuum of all level sets is simultaneously considered.

Another systematic approach to overcome the ambiguity concerning the quality of clustering lies in the so called *model-based clustering*, which assumes that the data is generated from a mixture distribution and the goal is to partition the data in congruity with the components that generate the data. However, theoretical analysis of kernel clustering methods have been confined to settings with parametric distributions (Yan et al., 2016; Couillet et al., 2016; Vankadara et al., 2020). Parametric assumptions such as the Gaussian mixture setting, where the components are assumed to

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).



be normally distributed, are extremely restrictive since the data generated under such assumptions are far from a typical dataset for which kernel clustering algorithms are applicable. In contrast, non-parametric assumptions on the data-generation process can be considerably less restrictive, but kernel clustering algorithms have been elusive to theoretical analysis under such assumptions. A primary hurdle in the analysis of clustering approaches under non-parametric assumptions is due to the issue of identifiability of non-parametric mixture models, that is, non-parametric models may be ambiguously defined. There is limited previous work that presents an analysis of kernel-based clustering algorithms under non-parametric mixture models. Schiebinger et al. (2015) provide recovery guarantees for spectral clustering of non-parametric mixtures by analyzing the spectral properties of the Laplacian operator under the assumption that the overlap between the components is small relative to a notion of "indivisibility" of the components. The analysis provided in Schiebinger et al. (2015) is restricted to that of spectral clustering and considerably different from the analysis in this paper.

1.1 Contributions

Non-parameteric kernel clustering. We provide non-parametric conditions for consistency of certain kernel-based clustering algorithms. To the best of our knowledge, these are among the first theoretical guarantees to kernel-based clustering methods without assumptions on the form of the component distributions.

1. We provide an **impossibility result for kernel k-means**: there exists a mixture distribution with arbitrarily large separation between the components such that for finite samples from this distribution kernel k-means fails to recover the underlying clustering.
2. We establish **sufficient separability conditions** under which kernel-based algorithms such as k-center, farthest-first k-means (FFk-means++), or kernel linkage algorithms can consistently recover the true partition, given finite samples from a mixture distribution.
3. We establish **necessary conditions for consistency** of the kernel FFk-means++ and kernel linkage algorithms and show that these separability conditions are optimal, that is, the sufficient conditions match the necessary conditions.

Kernel-based data clustering as distribution clustering. We establish a key equivalence between kernel-based data clustering and kernel-based density clustering. In particular:

4. We show that Gaussian kernel-based data clustering is equivalent to density clustering, where, each data point is first represented by a Gaussian probability density function and the densities are then clustered using the maximum mean discrepancy metric (with respect to a Gaussian kernel).
5. In addition to theoretical implications, this connection could also have practical implications in matters such as choosing the bandwidth of the Gaussian kernel for clustering which has not been systematically studied in literature so far. Our analysis reveals that the bandwidth of the kernel used for clustering needs to decrease with n but, perhaps surprisingly, asymptotically remain non-zero.

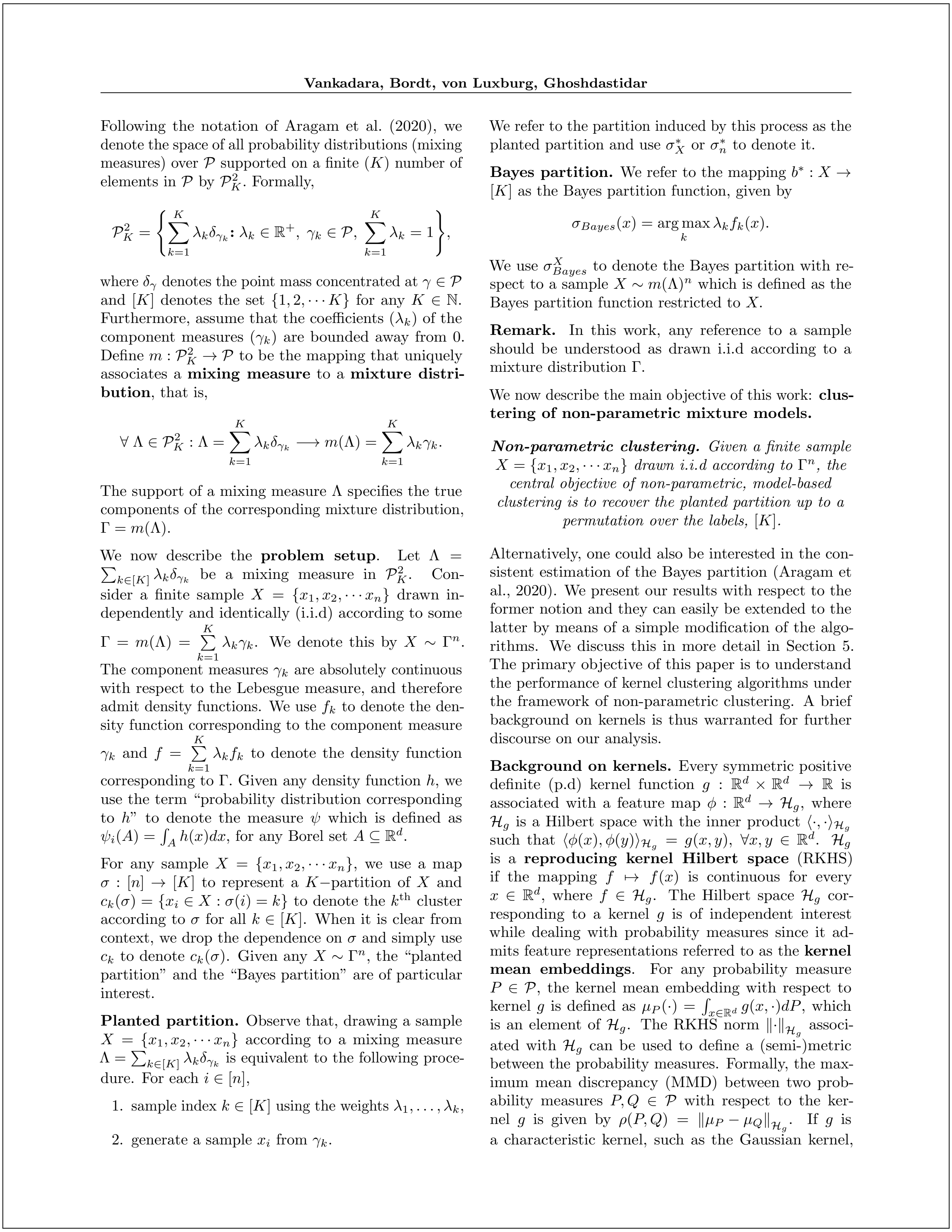
Non-parametric estimation of mixture models. Due to this relationship between kernel data clustering and distribution clustering, any standard Gaussian kernel clustering algorithm can be used to define an estimation procedure of the mixture model. Therefore, in addition to our primary contributions to kernel clustering, we also make contributions related to non-parametric *estimation* of mixture models.

6. We provide conditions under which the *estimation procedures* corresponding to the kernel-based clustering algorithms can consistently estimate the true mixture model.

2 FORMAL SETTING AND BACKGROUND

Consider the Euclidean space $\mathbb{R}^d$ of dimension d as the input domain. Let $\mathcal{P}$ denote the space of all Borel probability measures on $\mathbb{R}^d$ that are absolutely continuous with respect to the Lebesgue measure. In our analysis, we use the framework of mixing measures to define mixture distributions. This is fairly standard in the analysis of non-parametric mixture models (Aragam et al., 2020; Holzmann et al., 2006; Kimeldorf et al., 1970; Nguyen et al., 2013; Teicher, 1963) primarily due to the following reasons:

- Arbitrary mixture distributions are not identifiable. Mixing measures allow for the specification of *true components*. Section 3.1 provides a thorough discussion on identifiability of mixture models.
- In non-parametric clustering, one typically does not make any assumptions on the form of the component distributions. An elegant way to accomplish this is to allow arbitrary component distributions from $\mathcal{P}$ and impose restrictions on the set of admissible mixing measures.



Following the notation of Aragam et al. (2020), we denote the space of all probability distributions (mixing measures) over $\mathcal{P}$ supported on a finite (K) number of elements in $\mathcal{P}$ by $\mathcal{P}^2_K$. Formally,

$$\mathcal{P}^2_K = \left\{ \sum_{k=1}^{K} \lambda_k \delta_{\gamma_k} \colon \lambda_k \in \mathbb{R}^+,\ \gamma_k \in \mathcal{P},\ \sum_{k=1}^{K} \lambda_k = 1 \right\},$$

where δ_γ denotes the point mass concentrated at $\gamma \in \mathcal{P}$ and $[K]$ denotes the set $\{1, 2, \cdots K\}$ for any $K \in \mathbb{N}$. Furthermore, assume that the coefficients (λ_k) of the component measures (γ_k) are bounded away from 0. Define $m : \mathcal{P}^2_K \to \mathcal{P}$ to be the mapping that uniquely associates a **mixing measure** to a **mixture distribution**, that is,

$$\forall\ \Lambda \in \mathcal{P}^2_K : \Lambda = \sum_{k=1}^{K} \lambda_k \delta_{\gamma_k} \longrightarrow m(\Lambda) = \sum_{k=1}^{K} \lambda_k \gamma_k.$$

The support of a mixing measure Λ specifies the true components of the corresponding mixture distribution, $\Gamma = m(\Lambda)$.

We now describe the **problem setup**. Let $\Lambda = \sum_{k \in [K]} \lambda_k \delta_{\gamma_k}$ be a mixing measure in $\mathcal{P}^2_K$. Consider a finite sample $X = \{x_1, x_2, \cdots x_n\}$ drawn independently and identically (i.i.d) according to some $\Gamma = m(\Lambda) = \sum\limits_{k=1}^{K} \lambda_k \gamma_k$. We denote this by $X \sim \Gamma^n$. The component measures γ_k are absolutely continuous with respect to the Lebesgue measure, and therefore admit density functions. We use f_k to denote the density function corresponding to the component measure γ_k and $f = \sum\limits_{k=1}^{K} \lambda_k f_k$ to denote the density function corresponding to Γ. Given any density function h, we use the term "probability distribution corresponding to h" to denote the measure ψ which is defined as $\psi_i(A) = \int_A h(x)dx$, for any Borel set $A \subseteq \mathbb{R}^d$.

For any sample $X = \{x_1, x_2, \cdots x_n\}$, we use a map $\sigma : [n] \to [K]$ to represent a $K-$partition of X and $c_k(\sigma) = \{x_i \in X : \sigma(i) = k\}$ to denote the k^{th} cluster according to σ for all $k \in [K]$. When it is clear from context, we drop the dependence on σ and simply use c_k to denote $c_k(\sigma)$. Given any $X \sim \Gamma^n$, the "planted partition" and the "Bayes partition" are of particular interest.

Planted partition. Observe that, drawing a sample $X = \{x_1, x_2, \cdots x_n\}$ according to a mixing measure $\Lambda = \sum_{k \in [K]} \lambda_k \delta_{\gamma_k}$ is equivalent to the following procedure. For each $i \in [n]$,

1. sample index $k \in [K]$ using the weights $\lambda_1, \ldots, \lambda_k$,
2. generate a sample x_i from γ_k.

We refer to the partition induced by this process as the planted partition and use σ^*_X or σ^*_n to denote it.

Bayes partition. We refer to the mapping $b^* : X \to [K]$ as the Bayes partition function, given by

$$\sigma_{Bayes}(x) = \arg\max_k \lambda_k f_k(x).$$

We use σ^X_{Bayes} to denote the Bayes partition with respect to a sample $X \sim m(\Lambda)^n$ which is defined as the Bayes partition function restricted to X.

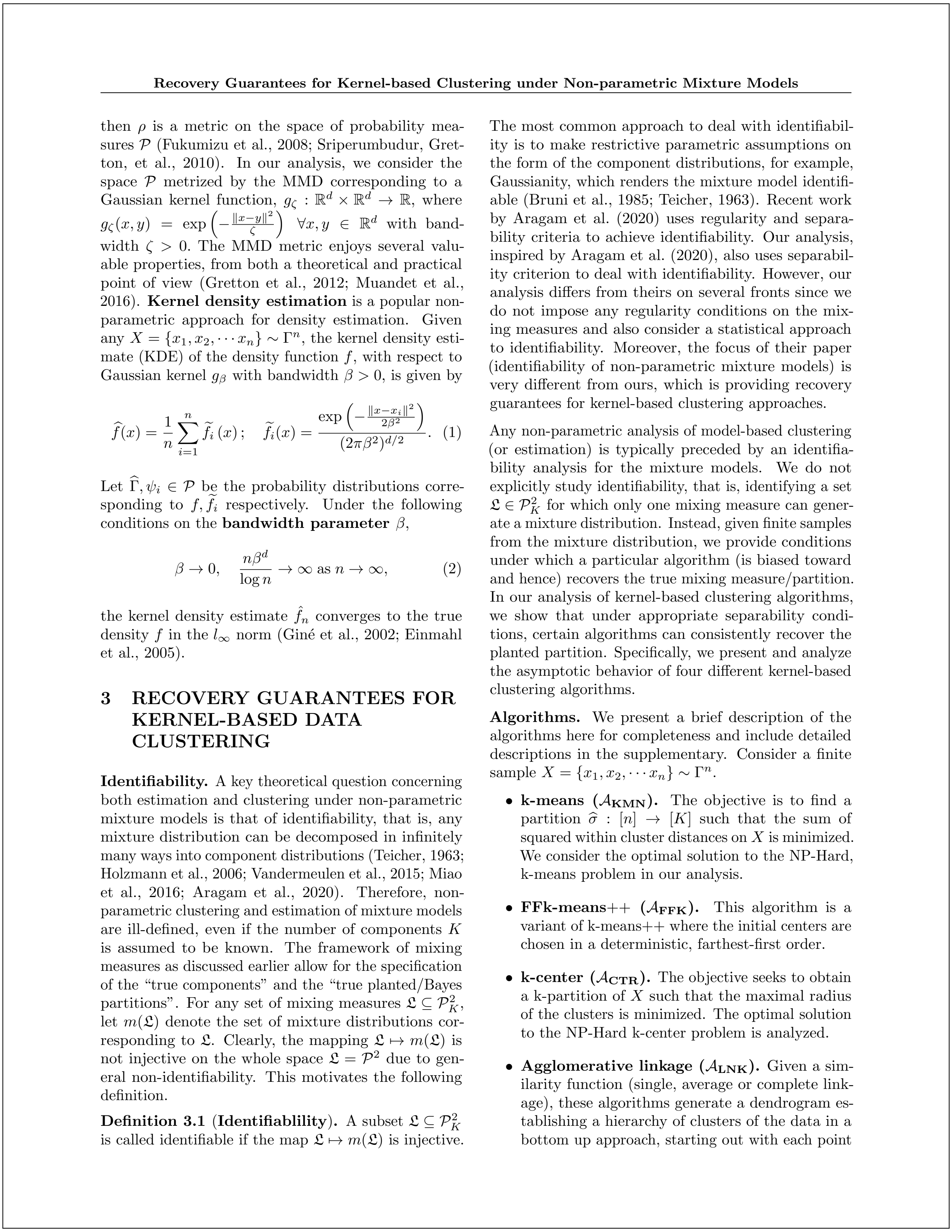
Remark. In this work, any reference to a sample should be understood as drawn i.i.d according to a mixture distribution Γ.

We now describe the main objective of this work: **clustering of non-parametric mixture models.**

Non-parametric clustering. *Given a finite sample $X = \{x_1, x_2, \cdots x_n\}$ drawn i.i.d according to Γ^n, the central objective of non-parametric, model-based clustering is to recover the planted partition up to a permutation over the labels, $[K]$.*

Alternatively, one could also be interested in the consistent estimation of the Bayes partition (Aragam et al., 2020). We present our results with respect to the former notion and they can easily be extended to the latter by means of a simple modification of the algorithms. We discuss this in more detail in Section 5. The primary objective of this paper is to understand the performance of kernel clustering algorithms under the framework of non-parametric clustering. A brief background on kernels is thus warranted for further discourse on our analysis.

Background on kernels. Every symmetric positive definite (p.d) kernel function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is associated with a feature map $\phi : \mathbb{R}^d \to \mathcal{H}_g$, where $\mathcal{H}_g$ is a Hilbert space with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_g}$ such that $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}_g} = g(x, y)$, $\forall x, y \in \mathbb{R}^d$. $\mathcal{H}_g$ is a **reproducing kernel Hilbert space** (RKHS) if the mapping $f \mapsto f(x)$ is continuous for every $x \in \mathbb{R}^d$, where $f \in \mathcal{H}_g$. The Hilbert space $\mathcal{H}_g$ corresponding to a kernel g is of independent interest while dealing with probability measures since it admits feature representations referred to as the **kernel mean embeddings**. For any probability measure $P \in \mathcal{P}$, the kernel mean embedding with respect to kernel g is defined as $\mu_P(\cdot) = \int_{x \in \mathbb{R}^d} g(x, \cdot) dP$, which is an element of $\mathcal{H}_g$. The RKHS norm $\|\cdot\|_{\mathcal{H}_g}$ associated with $\mathcal{H}_g$ can be used to define a (semi-)metric between the probability measures. Formally, the maximum mean discrepancy (MMD) between two probability measures $P, Q \in \mathcal{P}$ with respect to the kernel g is given by $\rho(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_g}$. If g is a characteristic kernel, such as the Gaussian kernel,



then ρ is a metric on the space of probability measures $\mathcal{P}$ (Fukumizu et al., 2008; Sriperumbudur, Gretton, et al., 2010). In our analysis, we consider the space $\mathcal{P}$ metrized by the MMD corresponding to a Gaussian kernel function, $g_\zeta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, where $g_\zeta(x, y) = \exp\left(-\frac{\|x-y\|^2}{\zeta}\right) \; \forall x, y \in \mathbb{R}^d$ with bandwidth $\zeta > 0$. The MMD metric enjoys several valuable properties, from both a theoretical and practical point of view (Gretton et al., 2012; Muandet et al., 2016). **Kernel density estimation** is a popular non-parametric approach for density estimation. Given any $X = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$, the kernel density estimate (KDE) of the density function f, with respect to Gaussian kernel g_β with bandwidth $\beta > 0$, is given by

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{f}_i\,(x)\,; \quad \widetilde{f}_i(x) = \frac{\exp\left(-\frac{\|x-x_i\|^2}{2\beta^2}\right)}{(2\pi\beta^2)^{d/2}}. \tag{1}$$

Let $\widehat{\Gamma}, \psi_i \in \mathcal{P}$ be the probability distributions corresponding to $f, \widetilde{f}_i$ respectively. Under the following conditions on the **bandwidth parameter** β,

$$\beta \to 0, \quad \frac{n\beta^d}{\log n} \to \infty \text{ as } n \to \infty, \tag{2}$$

the kernel density estimate $\hat{f}_n$ converges to the true density f in the l_∞ norm (Giné et al., 2002; Einmahl et al., 2005).

3 RECOVERY GUARANTEES FOR KERNEL-BASED DATA CLUSTERING

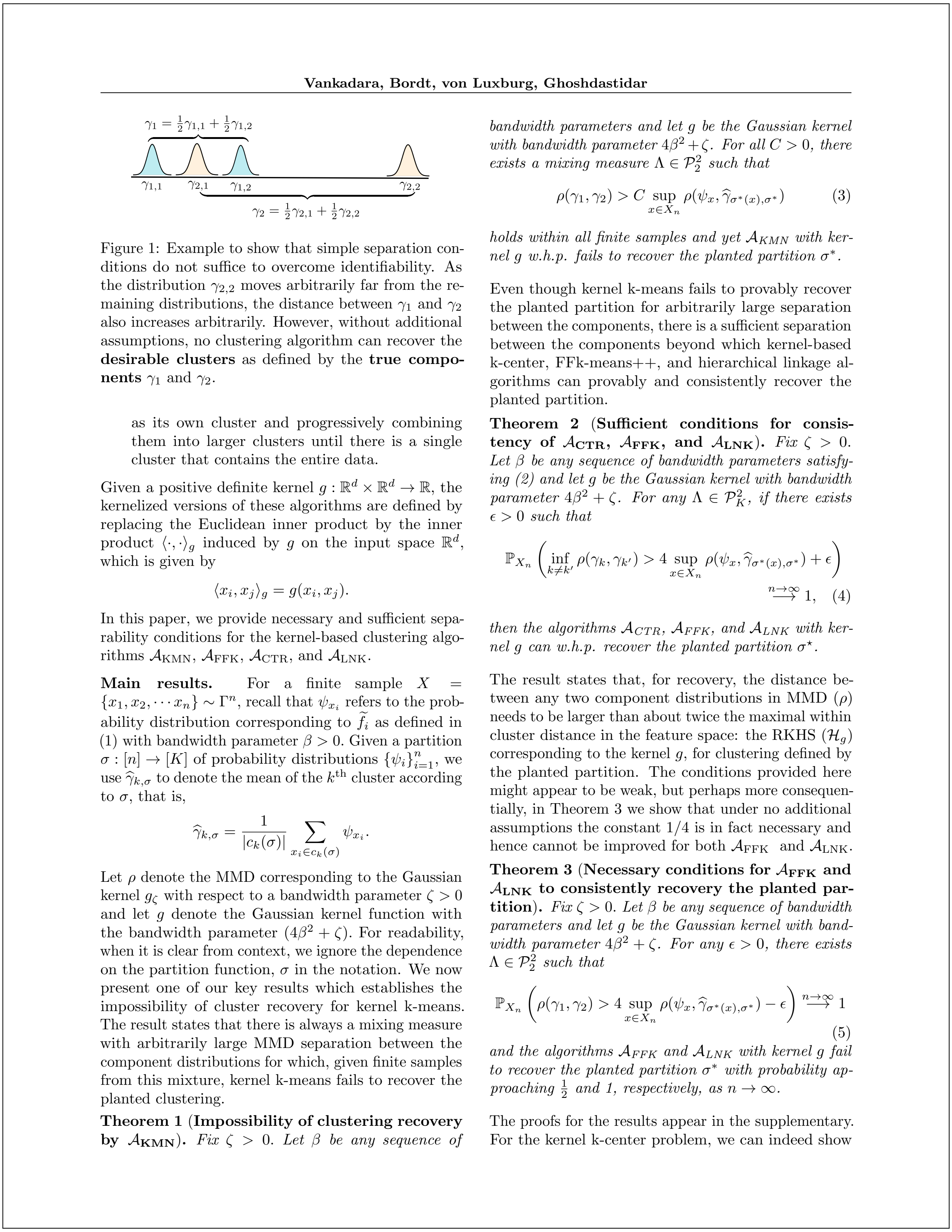
Identifiability. A key theoretical question concerning both estimation and clustering under non-parametric mixture models is that of identifiability, that is, any mixture distribution can be decomposed in infinitely many ways into component distributions (Teicher, 1963; Holzmann et al., 2006; Vandermeulen et al., 2015; Miao et al., 2016; Aragam et al., 2020). Therefore, non-parametric clustering and estimation of mixture models are ill-defined, even if the number of components K is assumed to be known. The framework of mixing measures as discussed earlier allow for the specification of the "true components" and the "true planted/Bayes partitions". For any set of mixing measures $\mathfrak{L} \subseteq \mathcal{P}^2_K$, let $m(\mathfrak{L})$ denote the set of mixture distributions corresponding to $\mathfrak{L}$. Clearly, the mapping $\mathfrak{L} \mapsto m(\mathfrak{L})$ is not injective on the whole space $\mathfrak{L} = \mathcal{P}^2$ due to general non-identifiability. This motivates the following definition.

Definition 3.1 (**Identifiablility**). A subset $\mathfrak{L} \subseteq \mathcal{P}^2_K$ is called identifiable if the map $\mathfrak{L} \mapsto m(\mathfrak{L})$ is injective. The most common approach to deal with identifiability is to make restrictive parametric assumptions on the form of the component distributions, for example, Gaussianity, which renders the mixture model identifiable (Bruni et al., 1985; Teicher, 1963). Recent work by Aragam et al. (2020) uses regularity and separability criteria to achieve identifiability. Our analysis, inspired by Aragam et al. (2020), also uses separability criterion to deal with identifiability. However, our analysis differs from theirs on several fronts since we do not impose any regularity conditions on the mixing measures and also consider a statistical approach to identifiability. Moreover, the focus of their paper (identifiability of non-parametric mixture models) is very different from ours, which is providing recovery guarantees for kernel-based clustering approaches.

Any non-parametric analysis of model-based clustering (or estimation) is typically preceded by an identifiability analysis for the mixture models. We do not explicitly study identifiability, that is, identifying a set $\mathfrak{L} \in \mathcal{P}^2_K$ for which only one mixing measure can generate a mixture distribution. Instead, given finite samples from the mixture distribution, we provide conditions under which a particular algorithm (is biased toward and hence) recovers the true mixing measure/partition. In our analysis of kernel-based clustering algorithms, we show that under appropriate separability conditions, certain algorithms can consistently recover the planted partition. Specifically, we present and analyze the asymptotic behavior of four different kernel-based clustering algorithms.

Algorithms. We present a brief description of the algorithms here for completeness and include detailed descriptions in the supplementary. Consider a finite sample $X = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$.

- **k-means** ($\mathcal{A}_{\mathbf{KMN}}$). The objective is to find a partition $\widehat{\sigma} : [n] \to [K]$ such that the sum of squared within cluster distances on X is minimized. We consider the optimal solution to the NP-Hard, k-means problem in our analysis.
- **FFk-means++** ($\mathcal{A}_{\mathbf{FFK}}$). This algorithm is a variant of k-means++ where the initial centers are chosen in a deterministic, farthest-first order.
- **k-center** ($\mathcal{A}_{\mathbf{CTR}}$). The objective seeks to obtain a k-partition of X such that the maximal radius of the clusters is minimized. The optimal solution to the NP-Hard k-center problem is analyzed.
- **Agglomerative linkage** ($\mathcal{A}_{\mathbf{LNK}}$). Given a similarity function (single, average or complete linkage), these algorithms generate a dendrogram establishing a hierarchy of clusters of the data in a bottom up approach, starting out with each point



$\gamma_1 = \frac{1}{2}\gamma_{1,1} + \frac{1}{2}\gamma_{1,2}$
$\gamma_{1,1}$
$\gamma_{2,1}$
$\gamma_{1,2}$
$\gamma_{2,2}$
$\gamma_2 = \frac{1}{2}\gamma_{2,1} + \frac{1}{2}\gamma_{2,2}$

Figure 1: Example to show that simple separation conditions do not suffice to overcome identifiability. As the distribution $\gamma_{2,2}$ moves arbitrarily far from the remaining distributions, the distance between γ_1 and γ_2 also increases arbitrarily. However, without additional assumptions, no clustering algorithm can recover the **desirable clusters** as defined by the **true components** γ_1 and γ_2.

as its own cluster and progressively combining them into larger clusters until there is a single cluster that contains the entire data.

Given a positive definite kernel $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the kernelized versions of these algorithms are defined by replacing the Euclidean inner product by the inner product $\langle \cdot, \cdot \rangle_g$ induced by g on the input space $\mathbb{R}^d$, which is given by

$$\langle x_i, x_j \rangle_g = g(x_i, x_j).$$

In this paper, we provide necessary and sufficient separability conditions for the kernel-based clustering algorithms $\mathcal{A}_{\text{KMN}}$, $\mathcal{A}_{\text{FFK}}$, $\mathcal{A}_{\text{CTR}}$, and $\mathcal{A}_{\text{LNK}}$.

Main results. For a finite sample $X = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$, recall that ψ_{x_i} refers to the probability distribution corresponding to $\widetilde{f}_i$ as defined in (1) with bandwidth parameter $\beta > 0$. Given a partition $\sigma : [n] \to [K]$ of probability distributions $\{\psi_i\}_{i=1}^n$, we use $\widehat{\gamma}_{k,\sigma}$ to denote the mean of the k^{th} cluster according to σ, that is,

$$\widehat{\gamma}_{k,\sigma} = \frac{1}{|c_k(\sigma)|} \sum_{x_i \in c_k(\sigma)} \psi_{x_i}.$$

Let ρ denote the MMD corresponding to the Gaussian kernel g_ζ with respect to a bandwidth parameter $\zeta > 0$ and let g denote the Gaussian kernel function with the bandwidth parameter $(4\beta^2 + \zeta)$. For readability, when it is clear from context, we ignore the dependence on the partition function, σ in the notation. We now present one of our key results which establishes the impossibility of cluster recovery for kernel k-means. The result states that there is always a mixing measure with arbitrarily large MMD separation between the component distributions for which, given finite samples from this mixture, kernel k-means fails to recover the planted clustering.

Theorem 1 (**Impossibility of clustering recovery by $\mathcal{A}_{\text{KMN}}$**). *Fix $\zeta > 0$. Let β be any sequence of bandwidth parameters and let g be the Gaussian kernel with bandwidth parameter $4\beta^2 + \zeta$. For all $C > 0$, there exists a mixing measure $\Lambda \in \mathcal{P}_2^2$ such that*

$$\rho(\gamma_1, \gamma_2) > C \sup_{x \in X_n} \rho(\psi_x, \widehat{\gamma}_{\sigma^*(x),\sigma^*}) \qquad (3)$$

holds within all finite samples and yet $\mathcal{A}_{KMN}$ with kernel g w.h.p. fails to recover the planted partition σ^*.

Even though kernel k-means fails to provably recover the planted partition for arbitrarily large separation between the components, there is a sufficient separation between the components beyond which kernel-based k-center, FFk-means++, and hierarchical linkage algorithms can provably and consistently recover the planted partition.

Theorem 2 (**Sufficient conditions for consistency of $\mathcal{A}_{\text{CTR}}$, $\mathcal{A}_{\text{FFK}}$, and $\mathcal{A}_{\text{LNK}}$**). *Fix $\zeta > 0$. Let β be any sequence of bandwidth parameters satisfying (2) and let g be the Gaussian kernel with bandwidth parameter $4\beta^2 + \zeta$. For any $\Lambda \in \mathcal{P}_K^2$, if there exists $\epsilon > 0$ such that*

$$\mathbb{P}_{X_n}\left(\inf_{k \neq k'} \rho(\gamma_k, \gamma_{k'}) > 4 \sup_{x \in X_n} \rho(\psi_x, \widehat{\gamma}_{\sigma^*(x),\sigma^*}) + \epsilon\right) \xrightarrow{n\to\infty} 1, \qquad (4)$$

then the algorithms $\mathcal{A}_{CTR}$, $\mathcal{A}_{FFK}$, and $\mathcal{A}_{LNK}$ with kernel g can w.h.p. recover the planted partition $\sigma^\star$.

The result states that, for recovery, the distance between any two component distributions in MMD (ρ) needs to be larger than about twice the maximal within cluster distance in the feature space: the RKHS ($\mathcal{H}_g$) corresponding to the kernel g, for clustering defined by the planted partition. The conditions provided here might appear to be weak, but perhaps more consequentially, in Theorem 3 we show that under no additional assumptions the constant 1/4 is in fact necessary and hence cannot be improved for both $\mathcal{A}_{\text{FFK}}$ and $\mathcal{A}_{\text{LNK}}$.

Theorem 3 (**Necessary conditions for $\mathcal{A}_{\text{FFK}}$ and $\mathcal{A}_{\text{LNK}}$ to consistently recovery the planted partition**). *Fix $\zeta > 0$. Let β be any sequence of bandwidth parameters and let g be the Gaussian kernel with bandwidth parameter $4\beta^2 + \zeta$. For any $\epsilon > 0$, there exists $\Lambda \in \mathcal{P}_2^2$ such that*

$$\mathbb{P}_{X_n}\left(\rho(\gamma_1, \gamma_2) > 4 \sup_{x \in X_n} \rho(\psi_x, \widehat{\gamma}_{\sigma^*(x),\sigma^*}) - \epsilon\right) \xrightarrow{n\to\infty} 1 \qquad (5)$$

and the algorithms $\mathcal{A}_{FFK}$ and $\mathcal{A}_{LNK}$ with kernel g fail to recover the planted partition σ^* with probability approaching $\frac{1}{2}$ and 1, respectively, as $n \to \infty$.

The proofs for the results appear in the supplementary. For the kernel k-center problem, we can indeed show

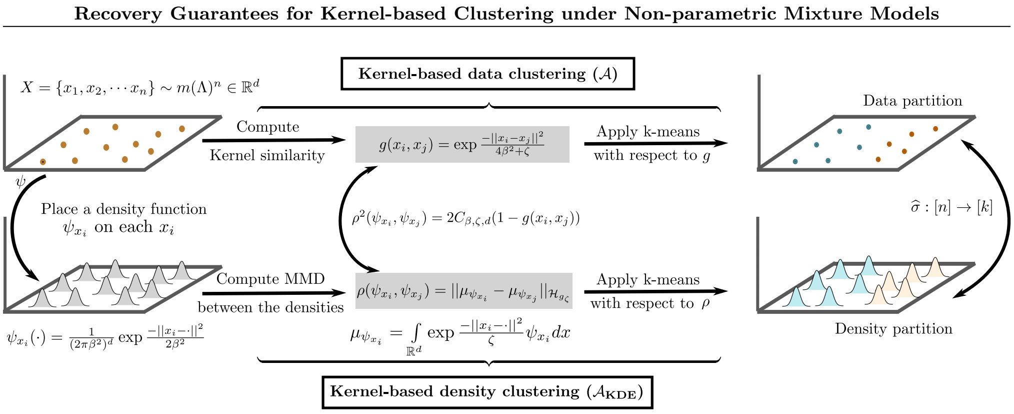

Recovery Guarantees for Kernel-based Clustering under Non-parametric Mixture Models
Kernel-based data clustering ($\mathcal{A}$)
$X = \{x_1, x_2, \cdots x_n\} \sim m(\Lambda)^n \in \mathbb{R}^d$
Compute
Kernel similarity
$g(x_i, x_j) = \exp \frac{-||x_i - x_j||^2}{4\beta^2 + \zeta}$
Apply k-means
with respect to g
Data partition
ψ
Place a density function
ψ_{x_i} on each x_i
$\rho^2(\psi_{x_i}, \psi_{x_j}) = 2C_{\beta,\zeta,d}(1 - g(x_i, x_j))$
$\widehat{\sigma} : [n] \to [k]$
Compute MMD
between the densities
$\rho(\psi_{x_i}, \psi_{x_j}) = ||\mu_{\psi_{x_i}} - \mu_{\psi_{x_j}}||_{\mathcal{H}_{g_\zeta}}$
Apply k-means
with respect to ρ
Density partition
$\psi_{x_i}(\cdot) = \frac{1}{(2\pi\beta^2)^d} \exp \frac{-||x_i - \cdot||^2}{2\beta^2}$
$\mu_{\psi_{x_i}} = \int_{\mathbb{R}^d} \exp \frac{-||x_i - \cdot||^2}{\zeta} \psi_{x_i} dx$
Kernel-based density clustering ($\mathcal{A}_{\mathbf{KDE}}$)

Figure 2: Illustration of the equivalence between kernel-based data clustering and distribution clustering. For Gaussian kernel clustering algorithm $\mathcal{A}$ using a bandwidth parameter $\eta > 0$, decompose η to obtain *any* $\beta > 0$ and $\zeta > 0$ satisfying $4\beta^2 + \zeta = \eta$. Then $\mathcal{A}$ can equivalently be reformulated as a kernel-based density clustering procedure as shown in the figure.

that the constant in the sufficient conditions (4) can further be improved to $1/3$ when $K = 2$. However, we believe that for any arbitrary K, the conditions provided in (4) cannot be further improved. This can be shown for a linear kernel and we leave the more general case of the Gaussian kernel as a conjecture. Our results not only show that certain kernel-based clustering algorithms can exploit separability to recover the planted clustering but also clearly show that under no additional assumptions very strong separability conditions are necessary to obtain recovery guarantees for kernel-based clustering. Furthermore, due to reasons of identifiability, simple separation conditions between the component distributions do not suffice to derive consistent recovery guarantees. For instance consider a simple example of a mixture distribution shown in Figure 1. As $\gamma_{2,2}$ moves arbitrarily far from the remaining distributions, the distance between the two component distributions, γ_1, γ_2 becomes arbitrarily far. However, without additional assumptions, it is not possible for a clustering algorithm to recover the desirable clustering even if we see infinite amount of data. Therefore, the separability conditions on the component distributions are necessarily dependent on the geometric properties of the distribution and not merely on the sample size or the dimension of the input space as it often is in the parametric setting. Our results, providing necessary and sufficient recovery conditions for kernel-based data clustering algorithms (Theorems 1, 2 , and 3), are obtained by analyzing an equivalent density/distribution clustering procedure which is considerably easier to analyze. Specifically, this equivalence allows us to exploit the metric geometry of the space of probability measures on the Euclidean space. Our proof techniques are motivated by the work of Aragam et al. (2020). We now describe this relationship between kernel-based data clustering and kernel-based density clustering.

4 EQUIVALENCE BETWEEN KERNEL-BASED DATA CLUSTERING AND DISTRIBUTION CLUSTERING

In this section, we present a density clustering procedure and describe its close relationship to kernel-based data clustering. Given a finite sample X, the density clustering procedure clusters the component probability distributions (ψ_i) of the kernel density estimate with respect to X using MMD as the metric between the distributions. This procedure is illustrated in Figure 2. As shown in Figure 2, the partition obtained by this density clustering procedure can be used to define a partition on the sample X. This partition can alternatively be obtained by using a simple kernel-based data clustering procedure. We now describe this density clustering procedure, which we denote by $\mathcal{A}_{\mathrm{KDE}}$.

Kernel-based density clustering $\mathcal{A}_{\mathbf{KDE}}$. Consider Gaussian kernel g_ζ for some $\zeta > 0$. Given sample $X \sim \Gamma^n$:

- Estimate the density of Γ by $\widehat{f} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{f}_i$ as in (1) with a bandwidth parameter $\beta > 0$.
- Consider MMD corresponding to the Gaussian kernel g_ζ as the metric between the distributions. Cluster the probability distributions $\{\psi_i\}_{i=1}^n$ corresponding to $\left\{\widetilde{f}_i\right\}_{i=1}^n$ by means of a distance based clustering algorithm (for example, k-means) to

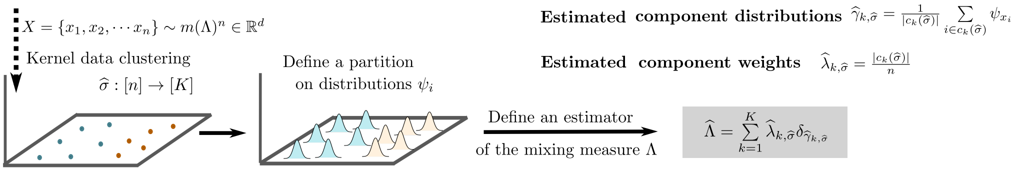

$X = \{x_1, x_2, \cdots x_n\} \sim m(\Lambda)^n \in \mathbb{R}^d$
Kernel data clustering
$\widehat{\sigma} : [n] \to [K]$
Define a partition
on distributions ψ_i
Estimated component distributions $\widehat{\gamma}_{k,\widehat{\sigma}} = \frac{1}{|c_k(\widehat{\sigma})|} \sum_{i \in c_k(\widehat{\sigma})} \psi_{x_i}$
Estimated component weights $\widehat{\lambda}_{k,\widehat{\sigma}} = \frac{|c_k(\widehat{\sigma})|}{n}$
Define an estimator
of the mixing measure Λ
$\widehat{\Lambda} = \sum_{k=1}^{K} \widehat{\lambda}_{k,\widehat{\sigma}} \delta_{\widehat{\gamma}_{k,\widehat{\sigma}}}$

Figure 3: Illustration of the estimation procedure defined with respect to a kernel clustering algorithm. Any Gaussian kernel clustering algorithm can be used to define a partition on component density functions $\{\psi_i\}_{i=1}^n$ which can in turn be used to define an estimator of the mixing measure Λ.

obtain a partition function $\widehat{\sigma}$.

This procedure is also illustrated in Figure 2. We show that for appropriately chosen bandwidth parameters, any kernel-based data clustering algorithm can be equivalently formulated as a density clustering procedure ($\mathcal{A}_{\text{KDE}}$). Recall that β and ζ are the bandwidth parameters of the Gaussian kernels used in $\mathcal{A}_{\text{KDE}}$ for kernel density estimation and for defining the MMD respectively. Then, let $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the Gaussian kernel with bandwidth parameter $4\beta^2 + \zeta$. The following lemma shows that the maximum mean discrepancy between the component distributions (ψ_i) is closely related to kernel evaluations on the input data.

Lemma 1 (**MMD between components is closely related to kernel evaluations between input data.**). *Given any sample $X \in \mathbb{R}^d$, let the component KDE distributions (ψ_i) be defined in the usual way. For all $x_i, x_j \in X$,*

$$\rho^2(\psi_i, \psi_j) = C_{\beta,\zeta,d}(1 - g(x_i, x_j))$$

where $C_{\beta,\zeta,d}$ is a constant dependent on the bandwidths β, ζ and the input dimension d.

We obtain this result by explicitly computing the MMD between the component distributions. Theorem 4 is then an immediate consequence of Lemma 1, which states that every kernel based data-clustering algorithm can equivalently be formulated as a kernel-based density clustering procedure (see Figure 2).

Theorem 4 (**Equivalence between kernel data–clustering and $\mathcal{A}_{\mathbf{KDE}}$**). *Any Gaussian kernel-based (data) clustering algorithm can equivalently be formulated as a clustering of the component KDE distributions with respect to the MMD metric corresponding to a Gaussian kernel for appropriately chosen bandwidth parameters.*

This simple result is consequential for practical considerations such as in the choice of bandwidth parameter for kernel data clustering (see Section 6) as well as for theoretical considerations. As it turns out, the density clustering procedure ($\mathcal{A}_{\text{KDE}}$) of the component KDE distributions can be used to define an estimator of the true mixing measure, that is, true component distributions and the corresponding weights. The equivalence between the two procedures, therefore, allows us to derive consistency guarantees for the estimators by analyzing the corresponding kernel-based clustering algorithms.

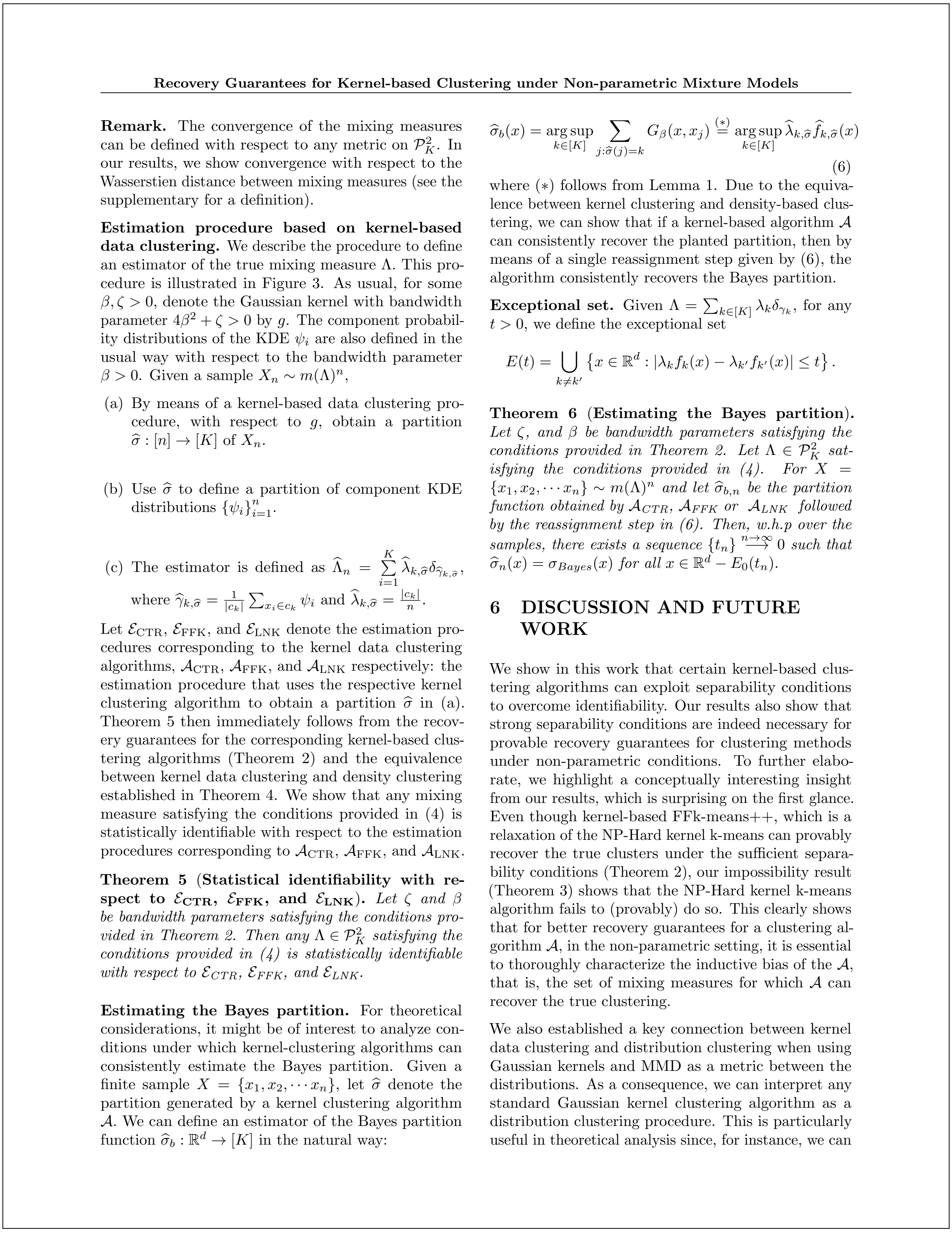
5 CONSISTENCY OF ESTIMATING MIXTURE MODELS

Estimation procedure. By an estimation procedure, we refer to any algorithm that takes a sample X drawn according to some mixing measure Λ, that is, $X \sim m(\Lambda)^n$ and provides an estimate $\widehat{\Lambda}$ of Λ.

Identifiability. Identifiability is also a key issue for estimation. Similar to our analysis of non-parametric clustering, we circumvent an *explicit* analysis of identifiability. Moreover, in the preceding discussion, identifiability is defined as a deterministic property of a set of mixing measures. We introduce a ***statistical*** notion of identifiability which can be defined as a property of either a mixing measure or a set of mixing measures. Additionally, in contrast to identifiability, statistical identifiability is defined with respect to an algorithm and therefore, it is a more intuitive and natural definition in the analysis of estimation procedures. Intuitively, the set of all mixing measures which are identifiable with respect to an estimation procedure $\mathcal{E}$ encodes the ***inductive bias*** of $\mathcal{E}$.

Definition 5.1 (**Statistical identifiability**). Let ϱ be some metric defined on the space of all mixing measures $\mathcal{P}_K^2$. A mixing measure Λ is statistically identifiable with respect to an estimation procedure $\mathcal{E}$ if the sequence of mixing measures $\left\{\widehat{\Lambda}_n = \mathcal{E}(X_n)\right\}$ converges in probability to Λ, where given $X_n \sim m(\Lambda)^n$.

Furthermore, a set of mixing measures $\mathfrak{L} \subset \mathcal{P}_K^2$ is said to be statistically identifiable with respect to estimation procedure $\mathcal{E}$ if every mixing measure $\Lambda \in \mathfrak{L}$ is statistically identifiable with respect to $\mathcal{E}$.



Remark. The convergence of the mixing measures can be defined with respect to any metric on $\mathcal{P}_K^2$. In our results, we show convergence with respect to the Wasserstien distance between mixing measures (see the supplementary for a definition).

Estimation procedure based on kernel-based data clustering. We describe the procedure to define an estimator of the true mixing measure Λ. This procedure is illustrated in Figure 3. As usual, for some $\beta, \zeta > 0$, denote the Gaussian kernel with bandwidth parameter $4\beta^2 + \zeta > 0$ by g. The component probability distributions of the KDE ψ_i are also defined in the usual way with respect to the bandwidth parameter $\beta > 0$. Given a sample $X_n \sim m(\Lambda)^n$,

(a) By means of a kernel-based data clustering procedure, with respect to g, obtain a partition $\widehat{\sigma} : [n] \to [K]$ of X_n.

(b) Use $\widehat{\sigma}$ to define a partition of component KDE distributions $\{\psi_i\}_{i=1}^n$.

(c) The estimator is defined as $\widehat{\Lambda}_n = \sum\limits_{i=1}^{K} \widehat{\lambda}_{k,\widehat{\sigma}} \delta_{\widehat{\gamma}_{k,\widehat{\sigma}}}$, where $\widehat{\gamma}_{k,\widehat{\sigma}} = \frac{1}{|c_k|} \sum_{x_i \in c_k} \psi_i$ and $\widehat{\lambda}_{k,\widehat{\sigma}} = \frac{|c_k|}{n}$.

Let $\mathcal{E}_{\mathrm{CTR}}$, $\mathcal{E}_{\mathrm{FFK}}$, and $\mathcal{E}_{\mathrm{LNK}}$ denote the estimation procedures corresponding to the kernel data clustering algorithms, $\mathcal{A}_{\mathrm{CTR}}$, $\mathcal{A}_{\mathrm{FFK}}$, and $\mathcal{A}_{\mathrm{LNK}}$ respectively: the estimation procedure that uses the respective kernel clustering algorithm to obtain a partition $\widehat{\sigma}$ in (a). Theorem 5 then immediately follows from the recovery guarantees for the corresponding kernel-based clustering algorithms (Theorem 2) and the equivalence between kernel data clustering and density clustering established in Theorem 4. We show that any mixing measure satisfying the conditions provided in (4) is statistically identifiable with respect to the estimation procedures corresponding to $\mathcal{A}_{\mathrm{CTR}}$, $\mathcal{A}_{\mathrm{FFK}}$, and $\mathcal{A}_{\mathrm{LNK}}$.

Theorem 5 (**Statistical identifiability with respect to $\mathcal{E}_{\mathrm{CTR}}$, $\mathcal{E}_{\mathrm{FFK}}$, and $\mathcal{E}_{\mathrm{LNK}}$**). *Let ζ and β be bandwidth parameters satisfying the conditions provided in Theorem 2. Then any $\Lambda \in \mathcal{P}_K^2$ satisfying the conditions provided in (4) is statistically identifiable with respect to $\mathcal{E}_{CTR}$, $\mathcal{E}_{FFK}$, and $\mathcal{E}_{LNK}$.*

Estimating the Bayes partition. For theoretical considerations, it might be of interest to analyze conditions under which kernel-clustering algorithms can consistently estimate the Bayes partition. Given a finite sample $X = \{x_1, x_2, \cdots x_n\}$, let $\widehat{\sigma}$ denote the partition generated by a kernel clustering algorithm $\mathcal{A}$. We can define an estimator of the Bayes partition function $\widehat{\sigma}_b : \mathbb{R}^d \to [K]$ in the natural way:

$$\widehat{\sigma}_b(x) = \underset{k\in[K]}{\arg\sup} \sum_{j:\widehat{\sigma}(j)=k} G_\beta(x, x_j) \overset{(*)}{=} \underset{k\in[K]}{\arg\sup}\, \widehat{\lambda}_{k,\widehat{\sigma}} \widehat{f}_{k,\widehat{\sigma}}(x) \tag{6}$$

where $(*)$ follows from Lemma 1. Due to the equivalence between kernel clustering and density-based clustering, we can show that if a kernel-based algorithm $\mathcal{A}$ can consistently recover the planted partition, then by means of a single reassignment step given by (6), the algorithm consistently recovers the Bayes partition.

Exceptional set. Given $\Lambda = \sum_{k\in[K]} \lambda_k \delta_{\gamma_k}$, for any $t > 0$, we define the exceptional set

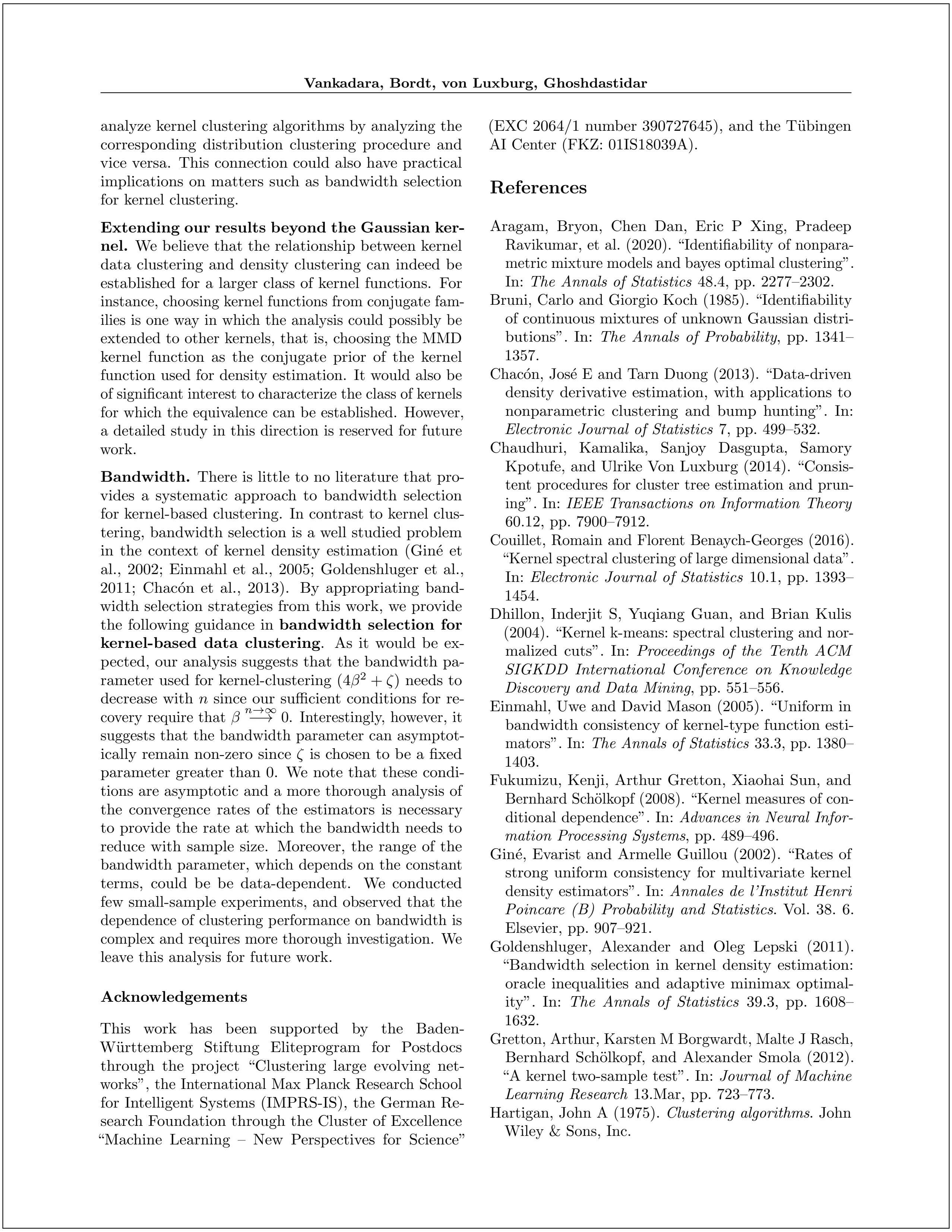
$$E(t) = \bigcup_{k \neq k'} \left\{ x \in \mathbb{R}^d : |\lambda_k f_k(x) - \lambda_{k'} f_{k'}(x)| \leq t \right\}.$$

Theorem 6 (**Estimating the Bayes partition**). *Let ζ, and β be bandwidth parameters satisfying the conditions provided in Theorem 2. Let $\Lambda \in \mathcal{P}_K^2$ satisfying the conditions provided in (4). For $X = \{x_1, x_2, \cdots x_n\} \sim m(\Lambda)^n$ and let $\widehat{\sigma}_{b,n}$ be the partition function obtained by $\mathcal{A}_{CTR}$, $\mathcal{A}_{FFK}$ or $\mathcal{A}_{LNK}$ followed by the reassignment step in (6). Then, w.h.p over the samples, there exists a sequence $\{t_n\} \overset{n\to\infty}{\longrightarrow} 0$ such that $\widehat{\sigma}_n(x) = \sigma_{Bayes}(x)$ for all $x \in \mathbb{R}^d - E_0(t_n)$.*

6 DISCUSSION AND FUTURE WORK

We show in this work that certain kernel-based clustering algorithms can exploit separability conditions to overcome identifiability. Our results also show that strong separability conditions are indeed necessary for provable recovery guarantees for clustering methods under non-parametric conditions. To further elaborate, we highlight a conceptually interesting insight from our results, which is surprising on the first glance. Even though kernel-based FFk-means++, which is a relaxation of the NP-Hard kernel k-means can provably recover the true clusters under the sufficient separability conditions (Theorem 2), our impossibility result (Theorem 3) shows that the NP-Hard kernel k-means algorithm fails to (provably) do so. This clearly shows that for better recovery guarantees for a clustering algorithm $\mathcal{A}$, in the non-parametric setting, it is essential to thoroughly characterize the inductive bias of the $\mathcal{A}$, that is, the set of mixing measures for which $\mathcal{A}$ can recover the true clustering.

We also established a key connection between kernel data clustering and distribution clustering when using Gaussian kernels and MMD as a metric between the distributions. As a consequence, we can interpret any standard Gaussian kernel clustering algorithm as a distribution clustering procedure. This is particularly useful in theoretical analysis since, for instance, we can



analyze kernel clustering algorithms by analyzing the corresponding distribution clustering procedure and vice versa. This connection could also have practical implications on matters such as bandwidth selection for kernel clustering.

Extending our results beyond the Gaussian kernel. We believe that the relationship between kernel data clustering and density clustering can indeed be established for a larger class of kernel functions. For instance, choosing kernel functions from conjugate families is one way in which the analysis could possibly be extended to other kernels, that is, choosing the MMD kernel function as the conjugate prior of the kernel function used for density estimation. It would also be of significant interest to characterize the class of kernels for which the equivalence can be established. However, a detailed study in this direction is reserved for future work.

Bandwidth. There is little to no literature that provides a systematic approach to bandwidth selection for kernel-based clustering. In contrast to kernel clustering, bandwidth selection is a well studied problem in the context of kernel density estimation (Giné et al., 2002; Einmahl et al., 2005; Goldenshluger et al., 2011; Chacón et al., 2013). By appropriating bandwidth selection strategies from this work, we provide the following guidance in **bandwidth selection for kernel-based data clustering**. As it would be expected, our analysis suggests that the bandwidth parameter used for kernel-clustering $(4\beta^2 + \zeta)$ needs to decrease with n since our sufficient conditions for recovery require that $\beta \overset{n\to\infty}{\longrightarrow} 0$. Interestingly, however, it suggests that the bandwidth parameter can asymptotically remain non-zero since ζ is chosen to be a fixed parameter greater than 0. We note that these conditions are asymptotic and a more thorough analysis of the convergence rates of the estimators is necessary to provide the rate at which the bandwidth needs to reduce with sample size. Moreover, the range of the bandwidth parameter, which depends on the constant terms, could be be data-dependent. We conducted few small-sample experiments, and observed that the dependence of clustering performance on bandwidth is complex and requires more thorough investigation. We leave this analysis for future work.

Acknowledgements

This work has been supported by the Baden-Württemberg Stiftung Eliteprogram for Postdocs through the project "Clustering large evolving networks", the International Max Planck Research School for Intelligent Systems (IMPRS-IS), the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1 number 390727645), and the Tübingen AI Center (FKZ: 01IS18039A).

References

Aragam, Bryon, Chen Dan, Eric P Xing, Pradeep Ravikumar, et al. (2020). "Identifiability of nonparametric mixture models and bayes optimal clustering". In: *The Annals of Statistics* 48.4, pp. 2277–2302.

Bruni, Carlo and Giorgio Koch (1985). "Identifiability of continuous mixtures of unknown Gaussian distributions". In: *The Annals of Probability*, pp. 1341–1357.

Chacón, José E and Tarn Duong (2013). "Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting". In: *Electronic Journal of Statistics* 7, pp. 499–532.

Chaudhuri, Kamalika, Sanjoy Dasgupta, Samory Kpotufe, and Ulrike Von Luxburg (2014). "Consistent procedures for cluster tree estimation and pruning". In: *IEEE Transactions on Information Theory* 60.12, pp. 7900–7912.

Couillet, Romain and Florent Benaych-Georges (2016). "Kernel spectral clustering of large dimensional data". In: *Electronic Journal of Statistics* 10.1, pp. 1393–1454.

Dhillon, Inderjit S, Yuqiang Guan, and Brian Kulis (2004). "Kernel k-means: spectral clustering and normalized cuts". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 551–556.

Einmahl, Uwe and David Mason (2005). "Uniform in bandwidth consistency of kernel-type function estimators". In: *The Annals of Statistics* 33.3, pp. 1380–1403.

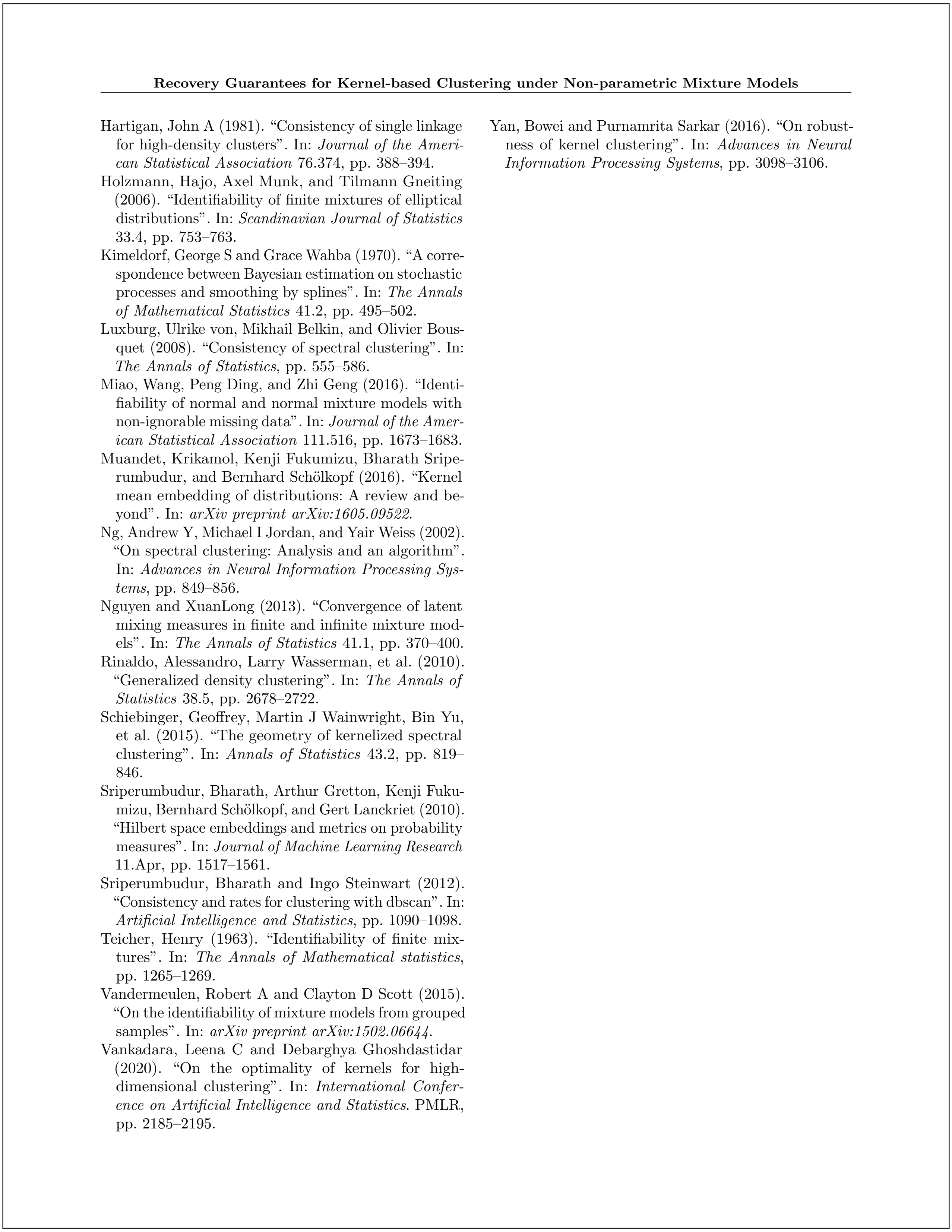
Fukumizu, Kenji, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf (2008). "Kernel measures of conditional dependence". In: *Advances in Neural Information Processing Systems*, pp. 489–496.

Giné, Evarist and Armelle Guillou (2002). "Rates of strong uniform consistency for multivariate kernel density estimators". In: *Annales de l'Institut Henri Poincare (B) Probability and Statistics*. Vol. 38. 6. Elsevier, pp. 907–921.

Goldenshluger, Alexander and Oleg Lepski (2011). "Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality". In: *The Annals of Statistics* 39.3, pp. 1608–1632.

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola (2012). "A kernel two-sample test". In: *Journal of Machine Learning Research* 13.Mar, pp. 723–773.

Hartigan, John A (1975). *Clustering algorithms.* John Wiley & Sons, Inc.



Hartigan, John A (1981). “Consistency of single linkage for high-density clusters”. In: *Journal of the American Statistical Association* 76.374, pp. 388–394.

Holzmann, Hajo, Axel Munk, and Tilmann Gneiting (2006). “Identifiability of finite mixtures of elliptical distributions”. In: *Scandinavian Journal of Statistics* 33.4, pp. 753–763.

Kimeldorf, George S and Grace Wahba (1970). “A correspondence between Bayesian estimation on stochastic processes and smoothing by splines”. In: *The Annals of Mathematical Statistics* 41.2, pp. 495–502.

Luxburg, Ulrike von, Mikhail Belkin, and Olivier Bousquet (2008). “Consistency of spectral clustering”. In: *The Annals of Statistics*, pp. 555–586.

Miao, Wang, Peng Ding, and Zhi Geng (2016). “Identifiability of normal and normal mixture models with non-ignorable missing data”. In: *Journal of the American Statistical Association* 111.516, pp. 1673–1683.

Muandet, Krikamol, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf (2016). “Kernel mean embedding of distributions: A review and beyond”. In: *arXiv preprint arXiv:1605.09522*.

Ng, Andrew Y, Michael I Jordan, and Yair Weiss (2002). “On spectral clustering: Analysis and an algorithm”. In: *Advances in Neural Information Processing Systems*, pp. 849–856.

Nguyen and XuanLong (2013). “Convergence of latent mixing measures in finite and infinite mixture models”. In: *The Annals of Statistics* 41.1, pp. 370–400.

Rinaldo, Alessandro, Larry Wasserman, et al. (2010). “Generalized density clustering”. In: *The Annals of Statistics* 38.5, pp. 2678–2722.

Schiebinger, Geoffrey, Martin J Wainwright, Bin Yu, et al. (2015). “The geometry of kernelized spectral clustering”. In: *Annals of Statistics* 43.2, pp. 819–846.

Sriperumbudur, Bharath, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert Lanckriet (2010). “Hilbert space embeddings and metrics on probability measures”. In: *Journal of Machine Learning Research* 11.Apr, pp. 1517–1561.

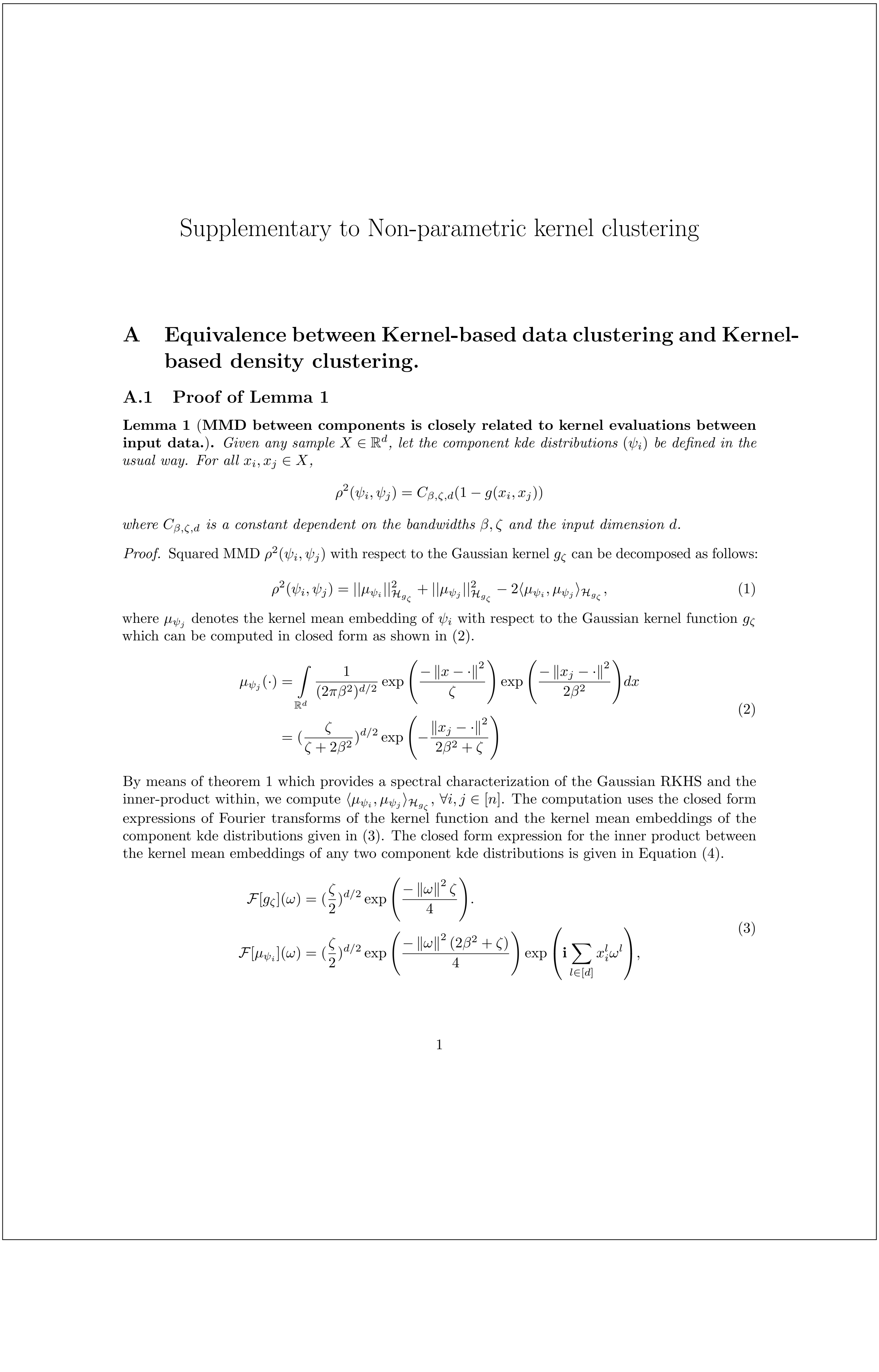
Sriperumbudur, Bharath and Ingo Steinwart (2012). “Consistency and rates for clustering with dbscan”. In: *Artificial Intelligence and Statistics*, pp. 1090–1098.

Teicher, Henry (1963). “Identifiability of finite mixtures”. In: *The Annals of Mathematical statistics*, pp. 1265–1269.

Vandermeulen, Robert A and Clayton D Scott (2015). “On the identifiability of mixture models from grouped samples”. In: *arXiv preprint arXiv:1502.06644*.

Vankadara, Leena C and Debarghya Ghoshdastidar (2020). “On the optimality of kernels for high-dimensional clustering”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2185–2195.

Yan, Bowei and Purnamrita Sarkar (2016). “On robustness of kernel clustering”. In: *Advances in Neural Information Processing Systems*, pp. 3098–3106.



Supplementary to Non-parametric kernel clustering

A Equivalence between Kernel-based data clustering and Kernel-based density clustering.

A.1 Proof of Lemma 1

Lemma 1 (MMD between components is closely related to kernel evaluations between input data.). *Given any sample $X \in \mathbb{R}^d$, let the component kde distributions (ψ_i) be defined in the usual way. For all $x_i, x_j \in X$,*

$$\rho^2(\psi_i, \psi_j) = C_{\beta,\zeta,d}(1 - g(x_i, x_j))$$

where $C_{\beta,\zeta,d}$ is a constant dependent on the bandwidths β, ζ and the input dimension d.

Proof. Squared MMD $\rho^2(\psi_i, \psi_j)$ with respect to the Gaussian kernel g_ζ can be decomposed as follows:

$$\rho^2(\psi_i, \psi_j) = ||\mu_{\psi_i}||^2_{\mathcal{H}_{g_\zeta}} + ||\mu_{\psi_j}||^2_{\mathcal{H}_{g_\zeta}} - 2\langle \mu_{\psi_i}, \mu_{\psi_j} \rangle_{\mathcal{H}_{g_\zeta}}, \tag{1}$$

where μ_{ψ_j} denotes the kernel mean embedding of ψ_i with respect to the Gaussian kernel function g_ζ which can be computed in closed form as shown in (2).

$$\begin{aligned}
\mu_{\psi_j}(\cdot) &= \int\limits_{\mathbb{R}^d} \frac{1}{(2\pi\beta^2)^{d/2}} \exp\left(\frac{-\|x - \cdot\|^2}{\zeta}\right) \exp\left(\frac{-\|x_j - \cdot\|^2}{2\beta^2}\right) dx \\
&= (\frac{\zeta}{\zeta + 2\beta^2})^{d/2} \exp\left(-\frac{\|x_j - \cdot\|^2}{2\beta^2 + \zeta}\right)
\end{aligned} \tag{2}$$

By means of theorem 1 which provides a spectral characterization of the Gaussian RKHS and the inner-product within, we compute $\langle \mu_{\psi_i}, \mu_{\psi_j} \rangle_{\mathcal{H}_{g_\zeta}}$, $\forall i, j \in [n]$. The computation uses the closed form expressions of Fourier transforms of the kernel function and the kernel mean embeddings of the component kde distributions given in (3). The closed form expression for the inner product between the kernel mean embeddings of any two component kde distributions is given in Equation (4).

$$\begin{aligned}
\mathcal{F}[g_\zeta](\omega) &= (\frac{\zeta}{2})^{d/2} \exp\left(\frac{-\|\omega\|^2 \zeta}{4}\right). \\
\mathcal{F}[\mu_{\psi_i}](\omega) &= (\frac{\zeta}{2})^{d/2} \exp\left(\frac{-\|\omega\|^2 (2\beta^2 + \zeta)}{4}\right) \exp\left(\mathbf{i} \sum_{l \in [d]} x_i^l \omega^l\right),
\end{aligned} \tag{3}$$

where $\mathbf{i}$ denotes the imaginary unit and satisfies $\mathbf{i}^2 = -1$.

$$\begin{aligned}\langle \mu_{\psi_i}, \mu_{\psi_j} \rangle_{\mathcal{H}_{g_\zeta}} &= \frac{1}{(2\pi)^{d/2}} \int \frac{\mathcal{F}[\mu_{\psi_i}](\omega)\overline{\mathcal{F}[\mu_{\psi_i}](\omega)}}{\mathcal{F}[g_\zeta](\omega)} d\omega \\ &= \left(\frac{\zeta}{4\beta^2+\zeta}\right)^{d/2} \exp\left(\frac{-\|x_i - x_j\|^2}{4\beta^2+\zeta}\right)\end{aligned} \tag{4}$$

Substituting the values of $\langle \mu_{\psi_i}, \mu_{\psi_j} \rangle_{\mathcal{H}_{g_\zeta}}$ for any $i, j \in [n]$ we obtain

$$\rho^2(\psi_i, \psi_j) = 2\left(\frac{\zeta}{4\beta^2+\zeta}\right)^{d/2} (1 - g(x_i, x_j)) \tag{5}$$

□

The following result given by Kimeldorf et al. (1970) and Wendland (2004) provides a spectral characterization of the RKHS corresponding to any translation-invariant kernel.

Theorem 1 (**Spectral characterization of RKHS. (Kimeldorf et al., 1970; Wendland, 2004)**). *Let k be a translation-invariant kernel on $\mathbb{R}^d$ such that $k(x, y) := \psi(x - y)$ where $\Phi \in C(\mathbb{R}^d) \cap L_1(R^d)$. Then the corresponding RKHS $\mathcal{H}$ is given by*

$$\mathcal{H} = \left\{ f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \ \|f\|^2_{\mathcal{H}_g} = \frac{1}{(2\pi)^{d/2}} \int \frac{|\mathcal{F}[f](\omega)|^2}{\mathcal{F}[\psi](\omega)} d\omega < \infty \right\}, \tag{6}$$

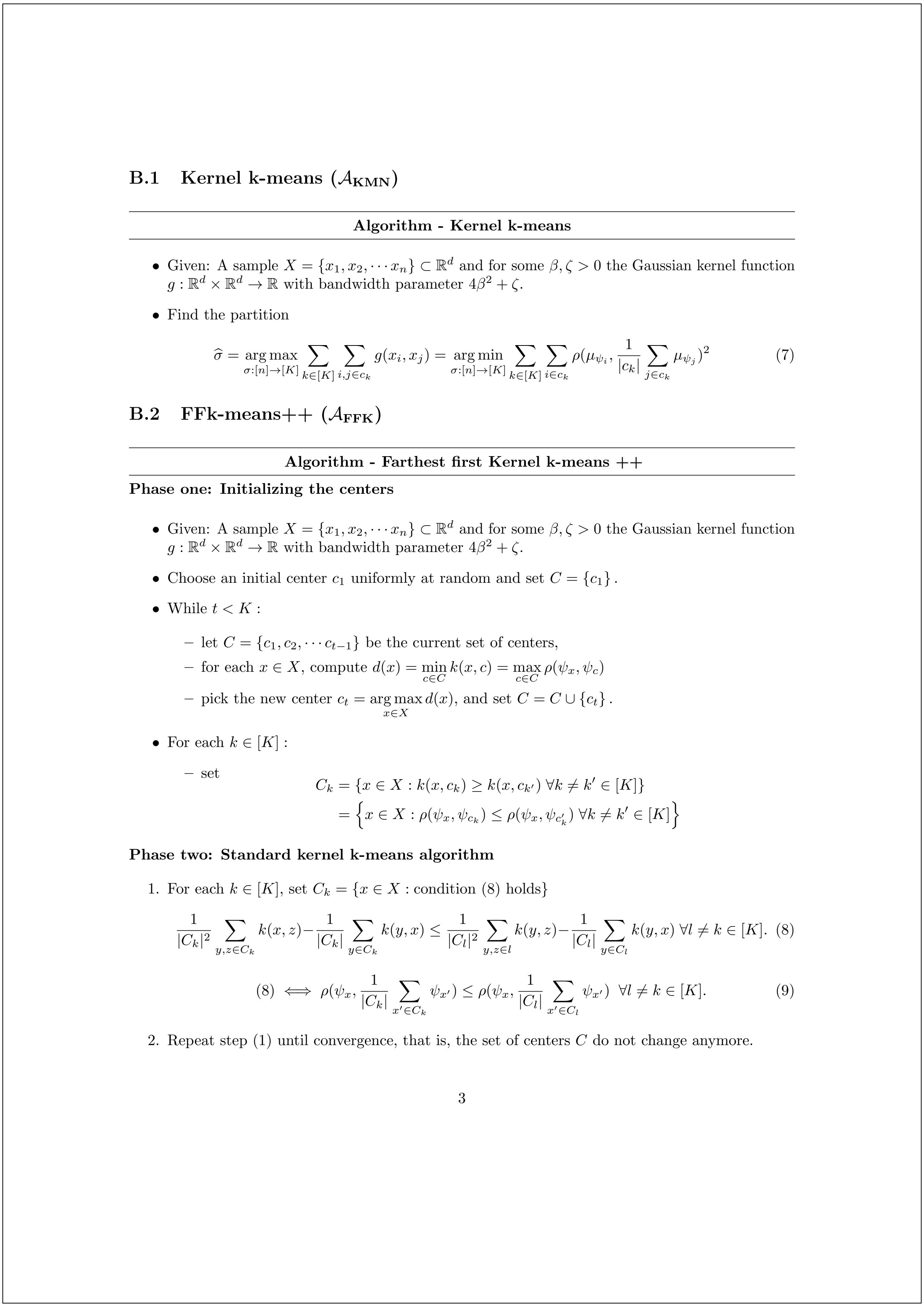
where $|\cdot|$ denotes the magnitude of the enclosed quantity and $\mathcal{F}[f](\omega)$ denotes the Fourier transform of the function f. The inner product on $\mathcal{H}$ is defined as $\langle f, g \rangle_{\mathcal{H}} = \frac{1}{(2\pi)^{d/2}} \int \frac{\mathcal{F}[f](\omega)\overline{\mathcal{F}[g](\omega)}}{\mathcal{F}[\psi](\omega)} d\omega, \quad f, g \in \mathcal{H}$, where $\overline{\mathcal{F}[g](\omega)}$ denotes the complex conjugate of $\mathcal{F}[g](\omega)$.

A.2 Proof of Theorem 4

Theorem 4 immediately follows from Lemma 1. For any data clustering algorithm with respect to the Gaussian kernel $\eta > 0$, decompose η into any two positive quantities $\beta, \zeta > 0$ satisfying $\eta = 4\beta^2 + \zeta$. Due to Lemma 1, the kernel clustering algorithm equivalently defines a clustering of the component kde distributions $\{\psi_i\}_{i=1}^n$.

B Algorithms

For completeness, we briefly describe the kernel-based clustering algorithms ($\mathcal{A}_{\text{KMN}}$, $\mathcal{A}_{\text{CTR}}$, $\mathcal{A}_{\text{FFK}}$, and $\mathcal{A}_{\text{LNK}}$) here. In each of the algorithms, we describe the standard kernel data clustering procedure as well as the equivalent kernel density clustering procedures (see Theorem 4). The component kde distributions are defined in the usual way with respect to the bandwidth parameter $\beta > 0$ and ρ is defined with respect to the Gaussian kernel with bandwidth parameter $\zeta > 0$.



B.1 Kernel k-means ($\mathcal{A}_{\mathbf{KMN}}$)

Algorithm - Kernel k-means

- Given: A sample $X = \{x_1, x_2, \cdots x_n\} \subset \mathbb{R}^d$ and for some $\beta, \zeta > 0$ the Gaussian kernel function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with bandwidth parameter $4\beta^2 + \zeta$.
- Find the partition

$$\widehat{\sigma} = \arg\max_{\sigma:[n]\to[K]} \sum_{k\in[K]} \sum_{i,j\in c_k} g(x_i, x_j) = \arg\min_{\sigma:[n]\to[K]} \sum_{k\in[K]} \sum_{i\in c_k} \rho(\mu_{\psi_i}, \frac{1}{|c_k|} \sum_{j\in c_k} \mu_{\psi_j})^2 \tag{7}$$

B.2 FFk-means++ ($\mathcal{A}_{\mathbf{FFK}}$)

Algorithm - Farthest first Kernel k-means ++

Phase one: Initializing the centers

- Given: A sample $X = \{x_1, x_2, \cdots x_n\} \subset \mathbb{R}^d$ and for some $\beta, \zeta > 0$ the Gaussian kernel function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with bandwidth parameter $4\beta^2 + \zeta$.
- Choose an initial center c_1 uniformly at random and set $C = \{c_1\}$.
- While $t < K$:
    - let $C = \{c_1, c_2, \cdots c_{t-1}\}$ be the current set of centers,
    - for each $x \in X$, compute $d(x) = \min_{c\in C} k(x, c) = \max_{c\in C} \rho(\psi_x, \psi_c)$
    - pick the new center $c_t = \arg\max_{x\in X} d(x)$, and set $C = C \cup \{c_t\}$.
- For each $k \in [K]$:
    - set

$$\begin{aligned} C_k &= \{x \in X : k(x, c_k) \geq k(x, c_{k'}) \ \forall k \neq k' \in [K]\} \\ &= \Big\{x \in X : \rho(\psi_x, \psi_{c_k}) \leq \rho(\psi_x, \psi_{c'_k}) \ \forall k \neq k' \in [K]\Big\} \end{aligned}$$

Phase two: Standard kernel k-means algorithm

1. For each $k \in [K]$, set $C_k = \{x \in X : \text{condition (8) holds}\}$

$$\frac{1}{|C_k|^2} \sum_{y,z\in C_k} k(x,z) - \frac{1}{|C_k|} \sum_{y\in C_k} k(y,x) \leq \frac{1}{|C_l|^2} \sum_{y,z\in l} k(y,z) - \frac{1}{|C_l|} \sum_{y\in C_l} k(y,x) \ \forall l \neq k \in [K]. \tag{8}$$

$$(8) \iff \rho(\psi_x, \frac{1}{|C_k|} \sum_{x'\in C_k} \psi_{x'}) \leq \rho(\psi_x, \frac{1}{|C_l|} \sum_{x'\in C_l} \psi_{x'}) \ \ \forall l \neq k \in [K]. \tag{9}$$

2. Repeat step (1) until convergence, that is, the set of centers C do not change anymore.

B.3 Kernel K-center($\mathcal{A}_{\mathbf{CTR}}$)

Algorithm - Kernel K-center

- Given: A sample $X = \{x_1, x_2, \cdots x_n\} \subset \mathbb{R}^d$ and for some $\beta, \zeta > 0$ the Gaussian kernel function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with bandwidth parameter $4\beta^2 + \zeta$.
- Find the partition

$$\begin{aligned}\widehat{\sigma} &= \operatorname*{arg\,max}_{\sigma:[n]\to[K]} \inf_{l\in[n]} \frac{-1}{|c_k^l|^2} \sum_{i,j\in c_k^l} k(x_i, x_j) + \frac{1}{|c_k^l|} \sum_{i\in c_k^l} k(x_i, x_l) \\ &= \operatorname*{arg\,min}_{\sigma:[n]\to[K]} \max_{i\in[n]} \rho(\psi_i, \widehat{\gamma}_{\sigma(i),\sigma})\end{aligned}$$

B.4 Agglomerative hierarchical clustering ($\mathcal{A}_{\mathbf{LNK}}$)

Given a sample $X = \{x_1, x_2, \cdots x_n\} \subset \mathbb{R}^d$ and a similarity function $S : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, hierarchical clustering algorithms seek to generate a cluster tree (dendrogram) establishing a hierarchy of relationships between the elements of the sample. Aggolomerative methods, in contrast to divisive methods, seek a bottom up approach, starting out with each point as its own cluster and progressively combining them into larger clusters until there is a single cluster that contains all the elements of the sample X. The criterion for merging hinges on the underlying similarity function, which in our case is the kernel matrix computed on the sample for a given kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. We discuss two of the popular hierarchical clustering algorithms that exist in literature: **single linkage** and **complete linkage** methods. The distinguishing factor across the two methods is the choice of the criterion C used to merge any two clusters $c, c' \subset X$ $(c \cap c' = \varnothing)$, which are given below in 10.

$$C(c, c') = \underbrace{\max_{x\in c, y\in c'} k(x, y) = \min_{x\in c, y\in c'} \rho(\psi_x, \psi_y)\,,}_{\textbf{Single linkage}} \quad \text{and} \quad \underbrace{\min_{x\in c, y\in c'} k(x, y) = \max_{x\in c, y\in c'} \rho(\psi_x, \psi_y)\,.}_{\textbf{Complete linkage}} \tag{10}$$

By substituting the different criterion $C(c, c')$ to merge any two clusters c, c' in Algorithm 1, we obtain variants of the corresponding algorithms.

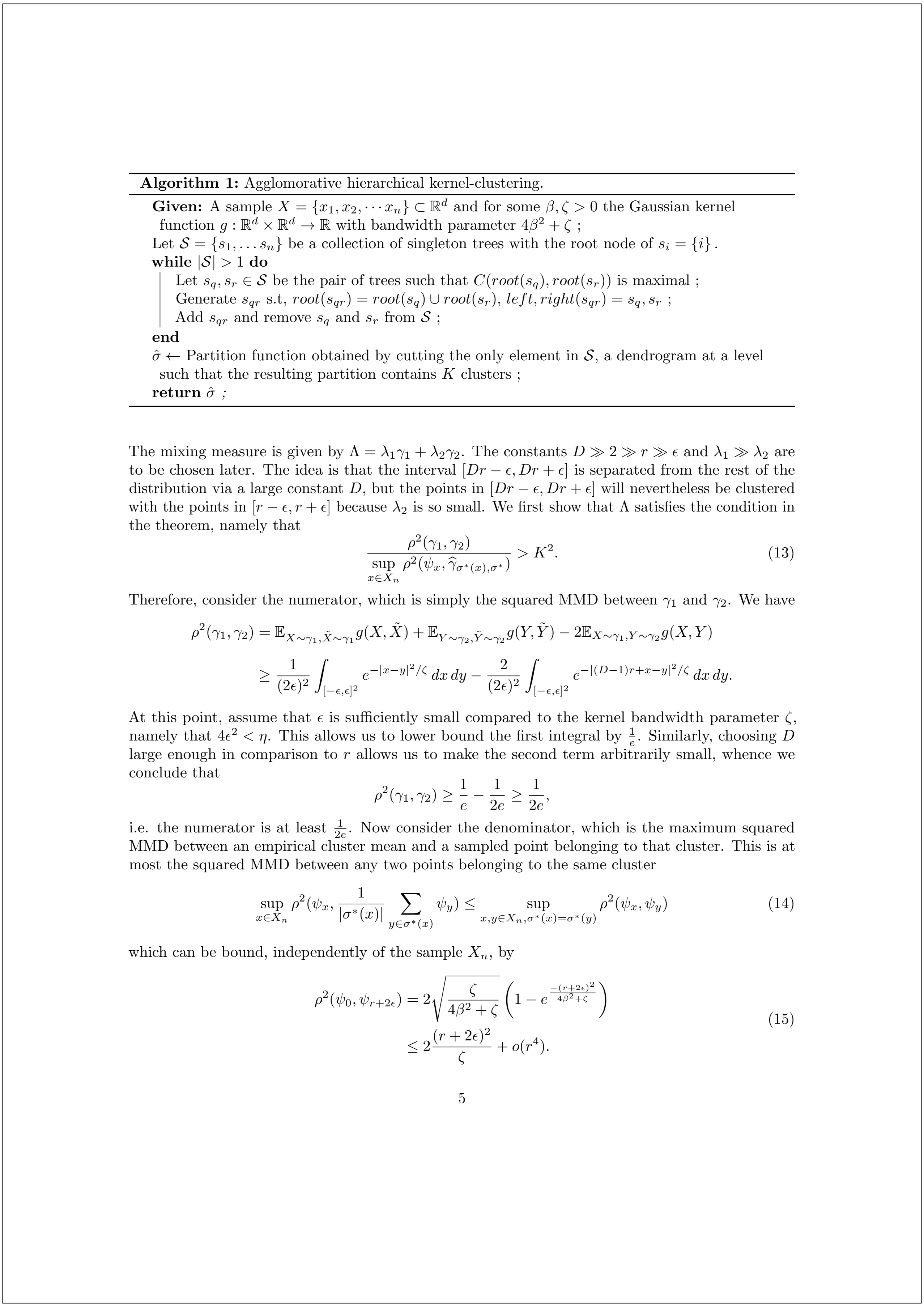
C Impossibility of recovery by kernel k-means(Proof of Theorem 1)

Proof. Fix the kernel bandwidth parameter $\zeta > 0$. Consider the following example in $\mathbb{R}$, where $\mathcal{U}([a, b])$ denotes the uniform distribution on the real interval $[a, b]$. Let

$$\gamma_1 = m\left(\frac{1}{2}\mathcal{U}([-\epsilon, \epsilon]) + \frac{1}{2}\mathcal{U}([r - \epsilon, r + \epsilon])\right) \tag{11}$$

and

$$\gamma_2 = \mathcal{U}([Dr - \epsilon, Dr + \epsilon]). \tag{12}$$



Algorithm 1: Agglomorative hierarchical kernel-clustering.

```
Given: A sample $X = \{x_1, x_2, \cdots x_n\} \subset \mathbb{R}^d$ and for some $\beta, \zeta > 0$ the Gaussian kernel
  function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with bandwidth parameter $4\beta^2 + \zeta$ ;
Let $\mathcal{S} = \{s_1, \ldots s_n\}$ be a collection of singleton trees with the root node of $s_i = \{i\}$ .
while $|\mathcal{S}| > 1$ do
    Let $s_q, s_r \in \mathcal{S}$ be the pair of trees such that $C(root(s_q), root(s_r))$ is maximal ;
    Generate $s_{qr}$ s.t, $root(s_{qr}) = root(s_q) \cup root(s_r)$, $left, right(s_{qr}) = s_q, s_r$ ;
    Add $s_{qr}$ and remove $s_q$ and $s_r$ from $\mathcal{S}$ ;
end
$\hat{\sigma} \leftarrow$ Partition function obtained by cutting the only element in $\mathcal{S}$, a dendrogram at a level
  such that the resulting partition contains $K$ clusters ;
return $\hat{\sigma}$ ;
```

The mixing measure is given by $\Lambda = \lambda_1\gamma_1 + \lambda_2\gamma_2$. The constants $D \gg 2 \gg r \gg \epsilon$ and $\lambda_1 \gg \lambda_2$ are to be chosen later. The idea is that the interval $[Dr - \epsilon, Dr + \epsilon]$ is separated from the rest of the distribution via a large constant D, but the points in $[Dr - \epsilon, Dr + \epsilon]$ will nevertheless be clustered with the points in $[r - \epsilon, r + \epsilon]$ because λ_2 is so small. We first show that Λ satisfies the condition in the theorem, namely that

$$\frac{\rho^2(\gamma_1, \gamma_2)}{\sup\limits_{x \in X_n} \rho^2(\psi_x, \widehat{\gamma}_{\sigma^*(x), \sigma^*})} > K^2. \tag{13}$$

Therefore, consider the numerator, which is simply the squared MMD between γ_1 and γ_2. We have

$$\begin{aligned} \rho^2(\gamma_1, \gamma_2) &= \mathbb{E}_{X \sim \gamma_1, \tilde{X} \sim \gamma_1} g(X, \tilde{X}) + \mathbb{E}_{Y \sim \gamma_2, \tilde{Y} \sim \gamma_2} g(Y, \tilde{Y}) - 2\mathbb{E}_{X \sim \gamma_1, Y \sim \gamma_2} g(X, Y) \\ &\geq \frac{1}{(2\epsilon)^2} \int_{[-\epsilon, \epsilon]^2} e^{-|x-y|^2/\zeta} \, dx \, dy - \frac{2}{(2\epsilon)^2} \int_{[-\epsilon, \epsilon]^2} e^{-|(D-1)r + x - y|^2/\zeta} \, dx \, dy. \end{aligned}$$

At this point, assume that ϵ is sufficiently small compared to the kernel bandwidth parameter ζ, namely that $4\epsilon^2 < \eta$. This allows us to lower bound the first integral by $\frac{1}{e}$. Similarly, choosing D large enough in comparison to r allows us to make the second term arbitrarily small, whence we conclude that

$$\rho^2(\gamma_1, \gamma_2) \geq \frac{1}{e} - \frac{1}{2e} \geq \frac{1}{2e},$$

i.e. the numerator is at least $\frac{1}{2e}$. Now consider the denominator, which is the maximum squared MMD between an empirical cluster mean and a sampled point belonging to that cluster. This is at most the squared MMD between any two points belonging to the same cluster

$$\sup_{x \in X_n} \rho^2(\psi_x, \frac{1}{|\sigma^*(x)|} \sum_{y \in \sigma^*(x)} \psi_y) \leq \sup_{x, y \in X_n, \sigma^*(x) = \sigma^*(y)} \rho^2(\psi_x, \psi_y) \tag{14}$$

which can be bound, independently of the sample X_n, by

$$\begin{aligned} \rho^2(\psi_0, \psi_{r+2\epsilon}) &= 2\sqrt{\frac{\zeta}{4\beta^2 + \zeta}} \left(1 - e^{\frac{-(r+2\epsilon)^2}{4\beta^2+\zeta}}\right) \\ &\leq 2\frac{(r + 2\epsilon)^2}{\zeta} + o(r^4). \end{aligned} \tag{15}$$

Here $r+2\epsilon$ is the maximum distance of any two points belonging to the same cluster and we used (5). Thus, choosing a small r allows us to make the denominator arbitrarily small, and the fraction in (13) can become larger than any fixed K^2.
Now, we show that k-means does w.h.p. not recover the planted partition. The idea is to choose $\lambda_1 \gg \lambda_2$. In our sample X_n from $m(\Lambda)$, denote the number of points within $[-\epsilon, \epsilon]$ by N_1, the number of points within within $[r-\epsilon, r+\epsilon]$ by N_2, and the number of points within $[Dr-\epsilon, Dr+\epsilon]$ by N_3. Assume that n is large enough s.t. $N_1, N_2, N_3 > 0$. We rely on the equivalence between kernel-based data clustering and kernel-based density clustering and directly consider the MMD between component distributions ψ_{x_i} (compare section B.1). That is we consider k-means w.r.t. the norm $\|\cdot\|^2 =<\cdot,\cdot>_{\mathcal{H}_{g_\zeta}}$. The k-means objective of the planted partition is at least

$$N_1\left\|\mu_{\psi_\epsilon} - \frac{N_1\,\mu_{\psi_{-\epsilon}} + N_2\,\mu_{\psi_{r-\epsilon}}}{N_1+N_2}\right\|^2 + N_2\left\|\mu_{\psi_{r-\epsilon}} - \frac{N_1\mu_{\psi_\epsilon} + N_2\mu_{\psi_{r+\epsilon}}}{N_1+N_2}\right\|^2 \geq \frac{N_1N_2}{N_1+N_2}\left\|\mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}}\right\|^2 + O(\epsilon).$$

Similarly, the k-means objective of the alternative partition where the points in $[r-\epsilon, r+\epsilon]$ and $[Dr-\epsilon, Dr+\epsilon]$ form a cluster is at most

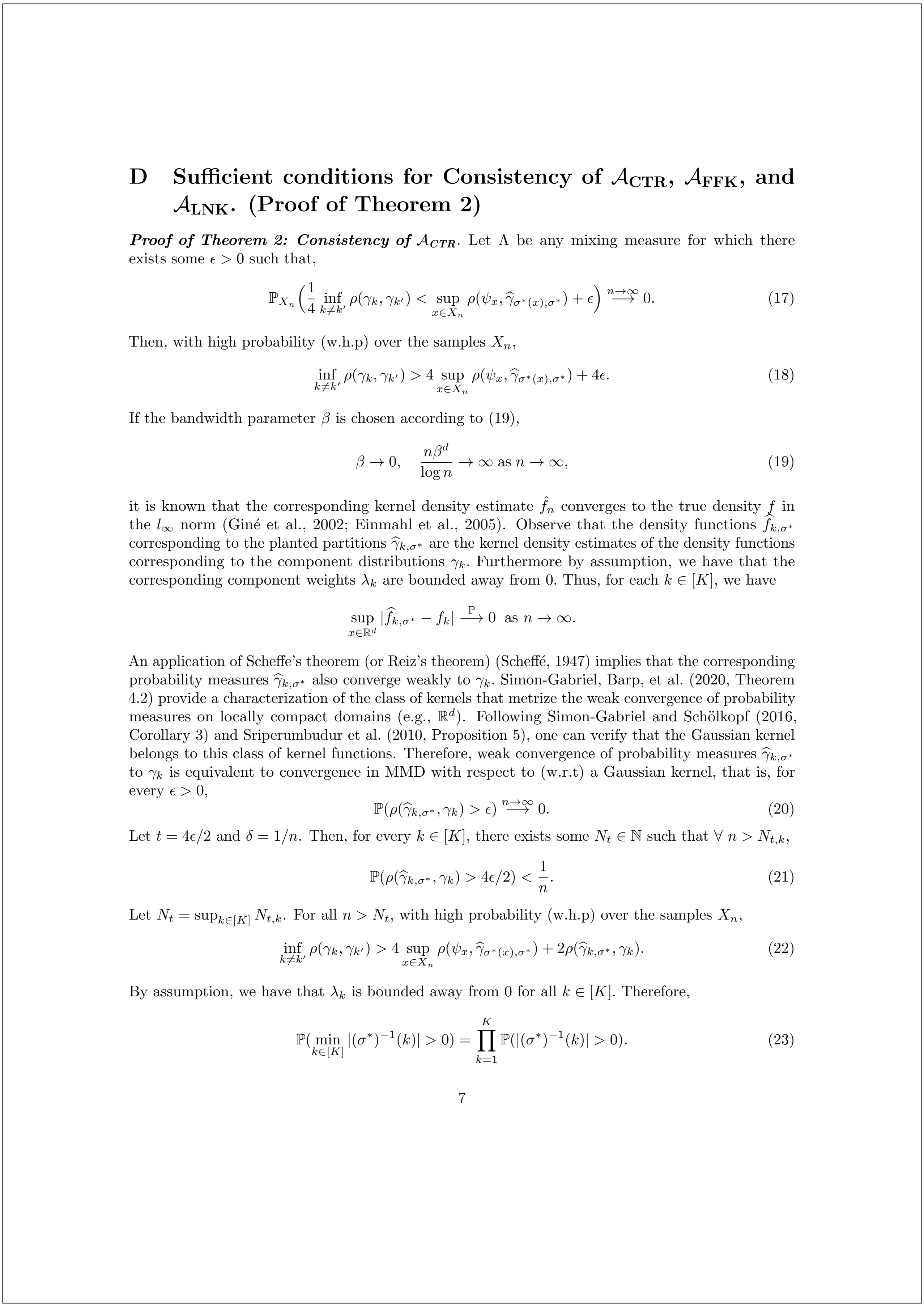
$$\begin{aligned} &N_1\left\|\mu_{\psi_0} - \mu_{\psi_{2\epsilon}}\right\|^2 + N_2\left\|\mu_{\psi_{r-\epsilon}} - \frac{N_2\mu_{\psi_{r+\epsilon}} + N_3\mu_{\psi_{Dr+\epsilon}}}{N_2+N_3}\right\|^2 + N_3\left\|\mu_{\psi_{Dr+\epsilon}} - \frac{N_2\mu_{\psi_{r-\epsilon}} + N_3\mu_{\psi_{Dr-\epsilon}}}{N_2+N_3}\right\|^2 \\ \leq &N_1\left\|\mu_{\psi_0} - \mu_{\psi_{2\epsilon}}\right\|^2 + \frac{N_2N_3}{N_2+N_3}\left\|\mu_{\psi_{r-\epsilon}} - \mu_{\psi_{Dr+\epsilon}}\right\|^2 + O(\epsilon). \end{aligned}$$

Thus, k-means will choose the alternative partition if

$$\begin{aligned} &N_1\left\|\mu_{\psi_0} - \mu_{\psi_{2\epsilon}}\right\|^2 + \frac{N_2N_3}{N_2+N_3}\left\|\mu_{\psi_{r-\epsilon}} - \mu_{\psi_{Dr+\epsilon}}\right\|^2 + O(\epsilon) \leq \frac{N_1N_2}{N_1+N_2}\left\|\mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}}\right\|^2 \\ &\iff \frac{\left\|\mu_{\psi_{r-\epsilon}} - \mu_{\psi_{Dr+\epsilon}}\right\|^2}{\left\|\mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}}\right\|^2} + O(\epsilon) \leq \frac{N_1}{N_3}\frac{N_2+N_3}{N_1+N_2} - \frac{N_1(N_2+N_3)}{N_2N_3}\frac{\left\|\mu_{\psi_0} - \mu_{\psi_{2\epsilon}}\right\|^2}{\left\|\mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}}\right\|^2} \\ &\iff \frac{\left\|\mu_{\psi_{r-\epsilon}} - \mu_{\psi_{Dr+\epsilon}}\right\|^2}{\left\|\mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}}\right\|^2} + O(\epsilon) \leq \frac{N_1}{N_3}\left(\frac{N_2+N_3}{N_1+N_2} - \left(1+\frac{N_3}{N_2}\right)\frac{\left\|\mu_{\psi_0} - \mu_{\psi_{2\epsilon}}\right\|^2}{\left\|\mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}}\right\|^2}\right). \end{aligned} \tag{16}$$

□

First note that the norms in equation (16) are deterministic quantities that depend on ϵ, r and D. The N_i are Binomial random variables parametrized by λ_1 and λ_2, i.e. $N_1 \sim \text{Binom}(n, \lambda_1/2)$, $N_2 \sim \text{Binom}(n, \lambda_1/2)$ and $N_3 \sim \text{Binom}(n, \lambda_2)$. All terms involving $N_i's$ w.h.p. concentrate around their expectation. Thus, choosing $\lambda_1 \gg \lambda_2$ allows us to make the fraction $\frac{N_1}{N_3}$ w.h.p. arbitrarily large. Choosing ϵ small enough (in comparison to r) ensures that the $O(\epsilon)$ term on the LHS is small enough, and that the bracketed term on the RHS is at least $\frac{1}{4}$.



D Sufficient conditions for Consistency of $\mathcal{A}_{\mathbf{CTR}}$, $\mathcal{A}_{\mathbf{FFK}}$, and $\mathcal{A}_{\mathbf{LNK}}$. (Proof of Theorem 2)

Proof of Theorem 2: Consistency of $\mathcal{A}_{CTR}$. Let Λ be any mixing measure for which there exists some $\epsilon > 0$ such that,

$$\mathbb{P}_{X_n}\Big(\frac{1}{4}\inf_{k\neq k'}\rho(\gamma_k,\gamma_{k'}) < \sup_{x\in X_n}\rho(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*}) + \epsilon\Big) \overset{n\to\infty}{\longrightarrow} 0. \tag{17}$$

Then, with high probability (w.h.p) over the samples X_n,

$$\inf_{k\neq k'}\rho(\gamma_k,\gamma_{k'}) > 4\sup_{x\in X_n}\rho(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*}) + 4\epsilon. \tag{18}$$

If the bandwidth parameter β is chosen according to (19),

$$\beta \to 0, \quad \frac{n\beta^d}{\log n} \to \infty \text{ as } n \to \infty, \tag{19}$$

it is known that the corresponding kernel density estimate $\hat{f}_n$ converges to the true density f in the l_∞ norm (Giné et al., 2002; Einmahl et al., 2005). Observe that the density functions $\widehat{f}_{k,\sigma^*}$ corresponding to the planted partitions $\widehat{\gamma}_{k,\sigma^*}$ are the kernel density estimates of the density functions corresponding to the component distributions γ_k. Furthermore by assumption, we have that the corresponding component weights λ_k are bounded away from 0. Thus, for each $k \in [K]$, we have

$$\sup_{x\in\mathbb{R}^d}|\widehat{f}_{k,\sigma^*} - f_k| \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty.$$

An application of Scheffe's theorem (or Reiz's theorem) (Scheffé, 1947) implies that the corresponding probability measures $\widehat{\gamma}_{k,\sigma^*}$ also converge weakly to γ_k. Simon-Gabriel, Barp, et al. (2020, Theorem 4.2) provide a characterization of the class of kernels that metrize the weak convergence of probability measures on locally compact domains (e.g., $\mathbb{R}^d$). Following Simon-Gabriel and Schölkopf (2016, Corollary 3) and Sriperumbudur et al. (2010, Proposition 5), one can verify that the Gaussian kernel belongs to this class of kernel functions. Therefore, weak convergence of probability measures $\widehat{\gamma}_{k,\sigma^*}$ to γ_k is equivalent to convergence in MMD with respect to (w.r.t) a Gaussian kernel, that is, for every $\epsilon > 0$,

$$\mathbb{P}(\rho(\widehat{\gamma}_{k,\sigma^*},\gamma_k) > \epsilon) \overset{n\to\infty}{\longrightarrow} 0. \tag{20}$$

Let $t = 4\epsilon/2$ and $\delta = 1/n$. Then, for every $k \in [K]$, there exists some $N_t \in \mathbb{N}$ such that $\forall\ n > N_{t,k}$,

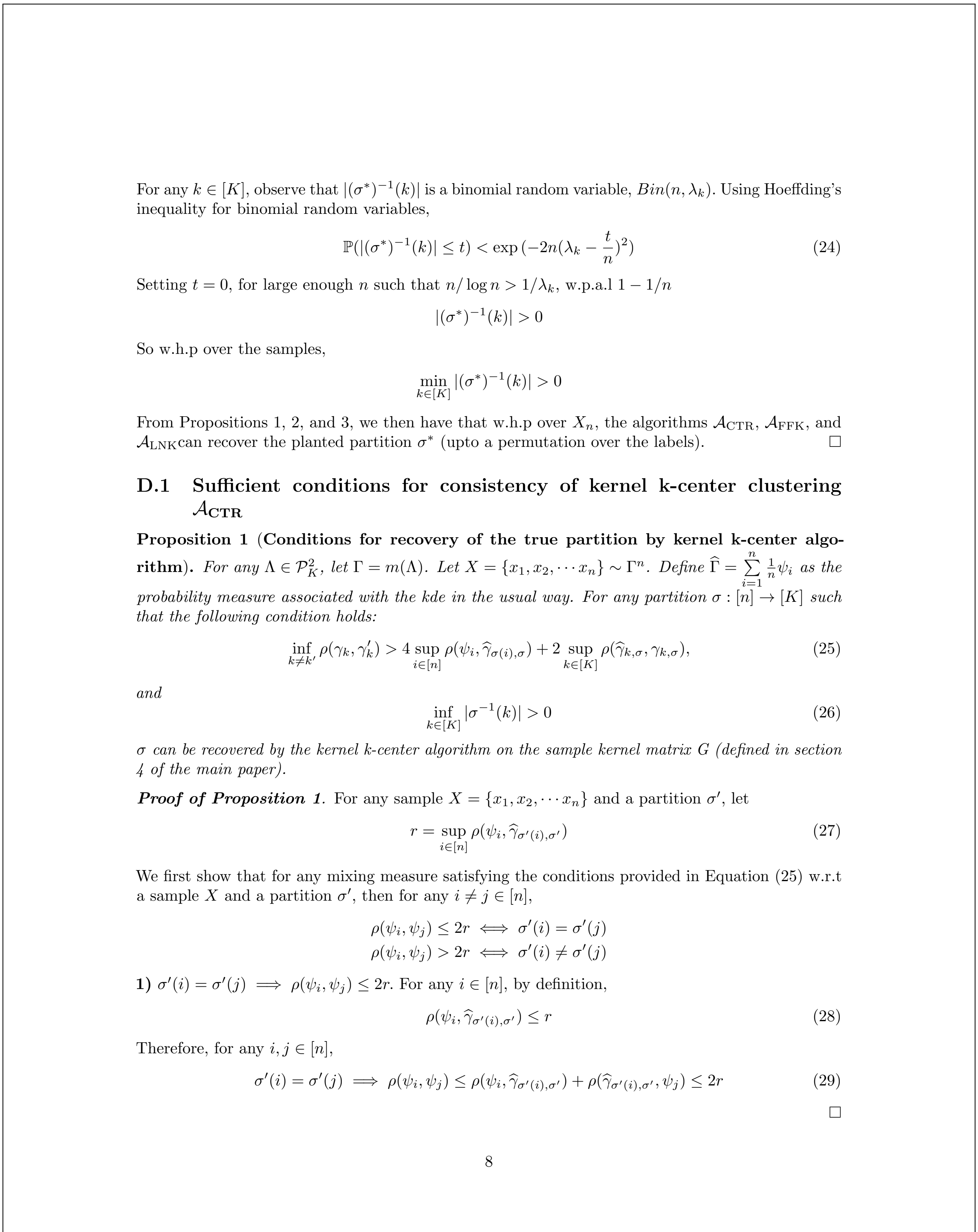
$$\mathbb{P}(\rho(\widehat{\gamma}_{k,\sigma^*},\gamma_k) > 4\epsilon/2) < \frac{1}{n}. \tag{21}$$

Let $N_t = \sup_{k\in[K]} N_{t,k}$. For all $n > N_t$, with high probability (w.h.p) over the samples X_n,

$$\inf_{k\neq k'}\rho(\gamma_k,\gamma_{k'}) > 4\sup_{x\in X_n}\rho(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*}) + 2\rho(\widehat{\gamma}_{k,\sigma^*},\gamma_k). \tag{22}$$

By assumption, we have that λ_k is bounded away from 0 for all $k \in [K]$. Therefore,

$$\mathbb{P}(\min_{k\in[K]}|(\sigma^*)^{-1}(k)| > 0) = \prod_{k=1}^{K}\mathbb{P}(|(\sigma^*)^{-1}(k)| > 0). \tag{23}$$



For any $k \in [K]$, observe that $|(\sigma^*)^{-1}(k)|$ is a binomial random variable, $Bin(n, \lambda_k)$. Using Hoeffding's inequality for binomial random variables,

$$\mathbb{P}(|(\sigma^*)^{-1}(k)| \leq t) < \exp\left(-2n(\lambda_k - \frac{t}{n})^2\right) \tag{24}$$

Setting $t = 0$, for large enough n such that $n/\log n > 1/\lambda_k$, w.p.a.l $1 - 1/n$

$$|(\sigma^*)^{-1}(k)| > 0$$

So w.h.p over the samples,

$$\min_{k \in [K]} |(\sigma^*)^{-1}(k)| > 0$$

From Propositions 1, 2, and 3, we then have that w.h.p over X_n, the algorithms $\mathcal{A}_{\text{CTR}}$, $\mathcal{A}_{\text{FFK}}$, and $\mathcal{A}_{\text{LNK}}$can recover the planted partition σ^* (upto a permutation over the labels). □

D.1 Sufficient conditions for consistency of kernel k-center clustering $\mathcal{A}_{\text{CTR}}$

Proposition 1 (Conditions for recovery of the true partition by kernel k-center algorithm). *For any $\Lambda \in \mathcal{P}_K^2$, let $\Gamma = m(\Lambda)$. Let $X = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$. Define $\widehat{\Gamma} = \sum_{i=1}^{n} \frac{1}{n}\psi_i$ as the probability measure associated with the kde in the usual way. For any partition $\sigma : [n] \to [K]$ such that the following condition holds:*

$$\inf_{k \neq k'} \rho(\gamma_k, \gamma'_k) > 4 \sup_{i \in [n]} \rho(\psi_i, \widehat{\gamma}_{\sigma(i),\sigma}) + 2 \sup_{k \in [K]} \rho(\widehat{\gamma}_{k,\sigma}, \gamma_{k,\sigma}), \tag{25}$$

and

$$\inf_{k \in [K]} |\sigma^{-1}(k)| > 0 \tag{26}$$

σ can be recovered by the kernel k-center algorithm on the sample kernel matrix G (defined in section 4 of the main paper).

Proof of Proposition 1. For any sample $X = \{x_1, x_2, \cdots x_n\}$ and a partition σ', let

$$r = \sup_{i \in [n]} \rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) \tag{27}$$

We first show that for any mixing measure satisfying the conditions provided in Equation (25) w.r.t a sample X and a partition σ', then for any $i \neq j \in [n]$,

$$\rho(\psi_i, \psi_j) \leq 2r \iff \sigma'(i) = \sigma'(j)$$
$$\rho(\psi_i, \psi_j) > 2r \iff \sigma'(i) \neq \sigma'(j)$$

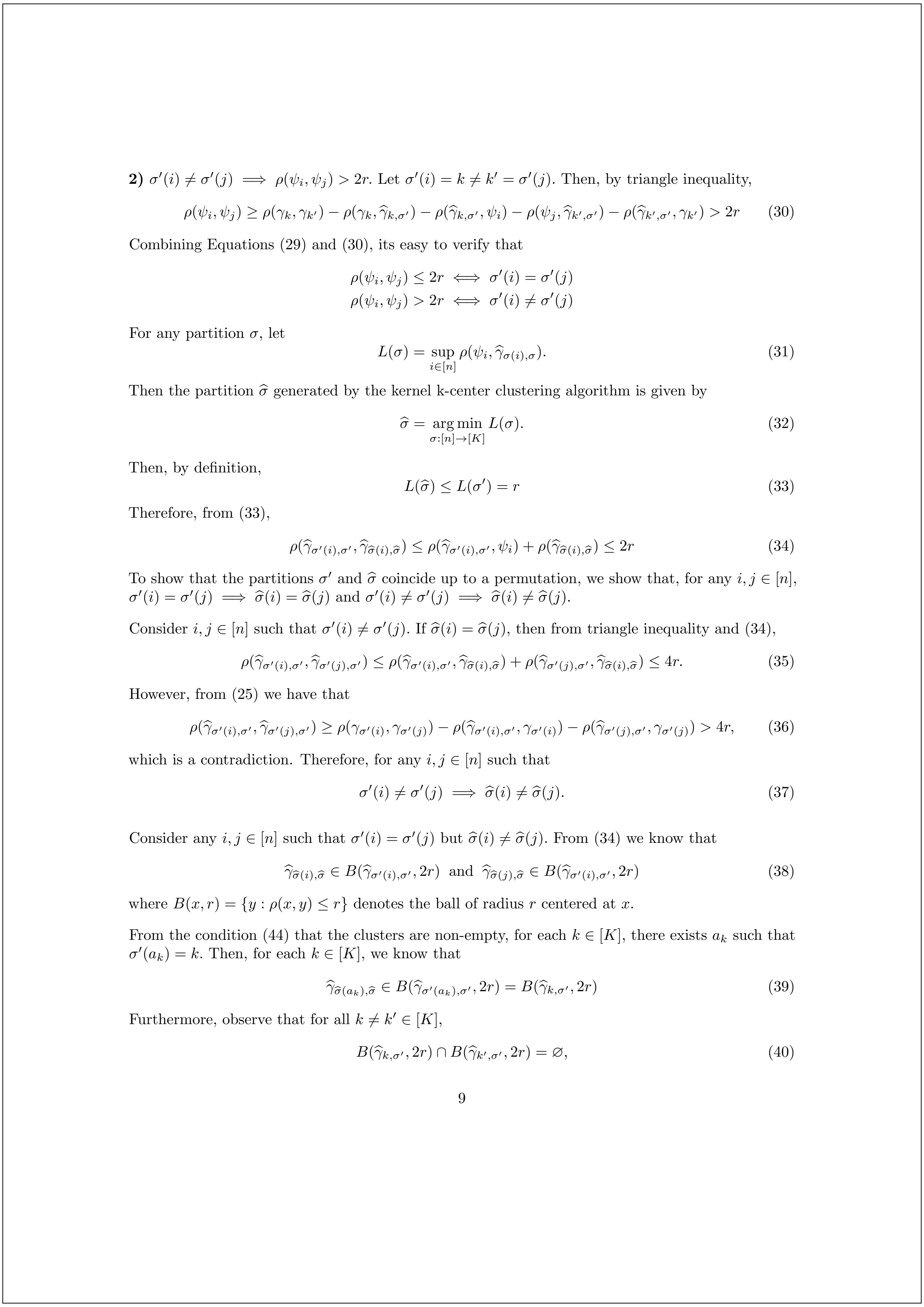
1) $\sigma'(i) = \sigma'(j) \implies \rho(\psi_i, \psi_j) \leq 2r$. For any $i \in [n]$, by definition,

$$\rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) \leq r \tag{28}$$

Therefore, for any $i, j \in [n]$,

$$\sigma'(i) = \sigma'(j) \implies \rho(\psi_i, \psi_j) \leq \rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) + \rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \psi_j) \leq 2r \tag{29}$$

□



2) $\sigma'(i) \neq \sigma'(j) \implies \rho(\psi_i, \psi_j) > 2r$. Let $\sigma'(i) = k \neq k' = \sigma'(j)$. Then, by triangle inequality,

$$\rho(\psi_i, \psi_j) \geq \rho(\gamma_k, \gamma_{k'}) - \rho(\gamma_k, \widehat{\gamma}_{k,\sigma'}) - \rho(\widehat{\gamma}_{k,\sigma'}, \psi_i) - \rho(\psi_j, \widehat{\gamma}_{k',\sigma'}) - \rho(\widehat{\gamma}_{k',\sigma'}, \gamma_{k'}) > 2r \tag{30}$$

Combining Equations (29) and (30), its easy to verify that

$$\begin{aligned} \rho(\psi_i, \psi_j) \leq 2r &\iff \sigma'(i) = \sigma'(j) \\ \rho(\psi_i, \psi_j) > 2r &\iff \sigma'(i) \neq \sigma'(j) \end{aligned}$$

For any partition σ, let

$$L(\sigma) = \sup_{i \in [n]} \rho(\psi_i, \widehat{\gamma}_{\sigma(i),\sigma}). \tag{31}$$

Then the partition $\widehat{\sigma}$ generated by the kernel k-center clustering algorithm is given by

$$\widehat{\sigma} = \operatorname*{arg\,min}_{\sigma:[n]\to[K]} L(\sigma). \tag{32}$$

Then, by definition,

$$L(\widehat{\sigma}) \leq L(\sigma') = r \tag{33}$$

Therefore, from (33),

$$\rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}}) \leq \rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \psi_i) + \rho(\widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}}) \leq 2r \tag{34}$$

To show that the partitions σ' and $\widehat{\sigma}$ coincide up to a permutation, we show that, for any $i, j \in [n]$, $\sigma'(i) = \sigma'(j) \implies \widehat{\sigma}(i) = \widehat{\sigma}(j)$ and $\sigma'(i) \neq \sigma'(j) \implies \widehat{\sigma}(i) \neq \widehat{\sigma}(j)$.

Consider $i, j \in [n]$ such that $\sigma'(i) \neq \sigma'(j)$. If $\widehat{\sigma}(i) = \widehat{\sigma}(j)$, then from triangle inequality and (34),

$$\rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \widehat{\gamma}_{\sigma'(j),\sigma'}) \leq \rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}}) + \rho(\widehat{\gamma}_{\sigma'(j),\sigma'}, \widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}}) \leq 4r. \tag{35}$$

However, from (25) we have that

$$\rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \widehat{\gamma}_{\sigma'(j),\sigma'}) \geq \rho(\gamma_{\sigma'(i)}, \gamma_{\sigma'(j)}) - \rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \gamma_{\sigma'(i)}) - \rho(\widehat{\gamma}_{\sigma'(j),\sigma'}, \gamma_{\sigma'(j)}) > 4r, \tag{36}$$

which is a contradiction. Therefore, for any $i, j \in [n]$ such that

$$\sigma'(i) \neq \sigma'(j) \implies \widehat{\sigma}(i) \neq \widehat{\sigma}(j). \tag{37}$$

Consider any $i, j \in [n]$ such that $\sigma'(i) = \sigma'(j)$ but $\widehat{\sigma}(i) \neq \widehat{\sigma}(j)$. From (34) we know that

$$\widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}} \in B(\widehat{\gamma}_{\sigma'(i),\sigma'}, 2r) \text{ and } \widehat{\gamma}_{\widehat{\sigma}(j),\widehat{\sigma}} \in B(\widehat{\gamma}_{\sigma'(i),\sigma'}, 2r) \tag{38}$$

where $B(x, r) = \{y : \rho(x, y) \leq r\}$ denotes the ball of radius r centered at x.

From the condition (44) that the clusters are non-empty, for each $k \in [K]$, there exists a_k such that $\sigma'(a_k) = k$. Then, for each $k \in [K]$, we know that

$$\widehat{\gamma}_{\widehat{\sigma}(a_k),\widehat{\sigma}} \in B(\widehat{\gamma}_{\sigma'(a_k),\sigma'}, 2r) = B(\widehat{\gamma}_{k,\sigma'}, 2r) \tag{39}$$

Furthermore, observe that for all $k \neq k' \in [K]$,

$$B(\widehat{\gamma}_{k,\sigma'}, 2r) \cap B(\widehat{\gamma}_{k',\sigma'}, 2r) = \varnothing, \tag{40}$$

since otherwise there exists some $x \in B(\widehat{\gamma}_{k,\sigma'}, 2r) \cap B(\widehat{\gamma}_{k',\sigma'}, 2r)$, i.e.,

$$\begin{aligned} &\rho(x, \widehat{\gamma}_{k,\sigma'}) \leq 2r \ \text{ and } \ \rho(x, \widehat{\gamma}_{k',\sigma'}) \leq 2r, \\ \implies &\rho(\widehat{\gamma}_{k,\sigma'}, \widehat{\gamma}_{k',\sigma'}) \leq \rho(x, \widehat{\gamma}_{k,\sigma'}) + \rho(x, \widehat{\gamma}_{k',\sigma'}) \leq 4r, \end{aligned}$$

which is a contradiction.

Moreover, by definition, $\sigma'(a_k) \neq \sigma'(a_{k'})$ for all $k, k' \in [K]$, from (37), we have

$$\widehat{\sigma}(a_1) \neq \widehat{\sigma}(a_2) \cdots \neq \widehat{\sigma}(a_K) \tag{41}$$

Since there are only K centers, (39), (40) and (41) imply that

- For any $i \in [n]$, there exists some $k \in [K]$ such that $\widehat{\sigma}(i) = \widehat{\sigma}(a_k)$, and
- $\widehat{\gamma}_{\widehat{\sigma}(a_k),\widehat{\sigma}} \in B(\widehat{\gamma}_{\sigma'(i),\sigma'}, 2r) \implies \widehat{\gamma}_{\widehat{\sigma}(a_{k'}),\widehat{\sigma}} \notin B(\widehat{\gamma}_{\sigma'(i),\sigma'}, 2r)$ for all $k' \neq k \in [K]$.

So, from (38),

$$\sigma'(i) = \sigma'(j) \implies \widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}} = \widehat{\gamma}_{\widehat{\sigma}(j),\widehat{\sigma}} \implies \widehat{\sigma}(i) = \widehat{\sigma}(j), \tag{42}$$

since, if $\widehat{\sigma}(i) \neq \widehat{\sigma}(j)$, then $\rho(\widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}}, \widehat{\gamma}_{\widehat{\sigma}(j),\widehat{\sigma}}) > 4r$.

Therefore, the partitions σ' and $\widehat{\sigma}$ coincide up to a permutation over the labels.

D.2 Sufficient conditions for kernel kmeans++ algorithm - proofs

Proposition 2 (Sufficient conditions for recovery by kernel k-means ++). *For any $\Lambda \in \mathcal{P}_K^2$, let $\Gamma = m(\Lambda)$. Let $X = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$. Define $\widehat{\Gamma} = \sum_{i=1}^{n} \frac{1}{n}\psi_i$ as the probability measure associated with the kde in the usual way. For any partition $\sigma' : [n] \to [K]$ such that the following condition holds:*

$$\inf_{k \neq k'} \rho(\gamma_k, \gamma'_k) > 4 \sup_{i \in [n]} \rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) + 2 \sup_{k \in [K]} \rho(\widehat{\gamma}_{k,\sigma'}, \gamma_{k,\sigma'}), \tag{43}$$

and

$$\inf_{k \in [K]} |(\sigma')^{-1}(k)| > 0 \tag{44}$$

σ can be recovered by a (deterministic) kernel k-means++ algorithm on the sample kernel matrix G.

Proof of Proposition 2. Let,

$$r = \sup_{i \in [n]} \rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}), \text{ and } B_k = B(\widehat{\gamma}_{k,\sigma'}, r) \ \ \forall k \in [K]. \tag{45}$$

Claim: Let C be the set of centers initialized in phase one of the k-means ++ algorithm as described. Then, for each $k \in [K]$,

$$c_k \in B_k \tag{46}$$

Proof: For every $i \in [n]$, by definition,

$$\rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) \leq r \implies \psi_i \in B_{\sigma'(i)}. \tag{47}$$

Therefore, without loss of generality (W.L.O.G), let $c_1 \in B_1$. For any $t < K$, assume that $C_t = \{c_1, c_2, \cdots c_t\}$ and $c_k \in B_k \;\; \forall k \in [t]$ (upto a permutation over the labels). Note that B_k is non-empty for every $k \in [K]$.

From the proof of Proposition 1, for any mixing measure satisfying the conditions provided in (43),

$$\rho(\psi_i, \psi_j) \leq 2r \iff \sigma'(i) = \sigma'(j) \tag{48}$$

$$\rho(\psi_i, \psi_j) > 2r \iff \sigma'(i) \neq \sigma'(j) \tag{49}$$

Therefore, since $c_k \in B_k$ for all $k \in [K]$, $d(\psi_i) = \rho^2(\psi_i, c_k) \leq 2r$ for all $\sigma'(i) = k$. Therefore,

$$d(\psi_i) \text{ is } \begin{cases} \leq 2r & \forall \psi_i \in B_k, \text{ and } k \leq t, \\ > 2r & \text{otherwise.} \end{cases} \tag{50}$$

Since $c_{t+1} = \arg\max\limits_{\psi_i} d(\psi_i)$, $c_{t+1} \in B_s$ for some $s \notin C_t$.

■

Claim: Kernel k-means algorithm does not affect the centers obtained in Phase one of the algorithm.
Proof: From claim 1, in phase one of the algorithm, the centers $C = \{c_1, c_2, \cdots c_K\}$ are obtained such that $c_k \in B_k$ for all $k \in [K]$. For each $k \in [K]$, clusters $\{C_1, C_2, \cdots C_K\}$ are then defined as follows.

$$C_k = \left\{ i \in [n] : \rho^2(c_k, \psi_i) \geq \rho^2(c_{k'}, \psi_i) \quad \forall k \neq k' \in [K] \right\} \tag{51}$$

From (48), we have that

$$\begin{aligned} \rho^2(\psi_i, c_k) &\leq 4r^2 && \text{if } \sigma'(i) = k \\ \rho^2(\psi_i, c_k) &> 4r^2 && \text{otherwise .} \end{aligned}$$

Therefore, the partition obtained in the Phase 1 of the algorithm coincides with σ' up to a permutation over the labels, that is,

$$C_k = \{\psi_i \in X : \sigma'(i) = k\}, \tag{52}$$

and

$$\sum_{i:\sigma'(i)=k} \psi_i = \widehat{\gamma}_{k,\sigma'} \in B_k. \tag{53}$$

Clearly,

$$\rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) \leq 2r \leq \rho(\psi_i, \widehat{\gamma}_{k,\sigma'}) > 2r \; \forall k \neq \sigma'(i).$$

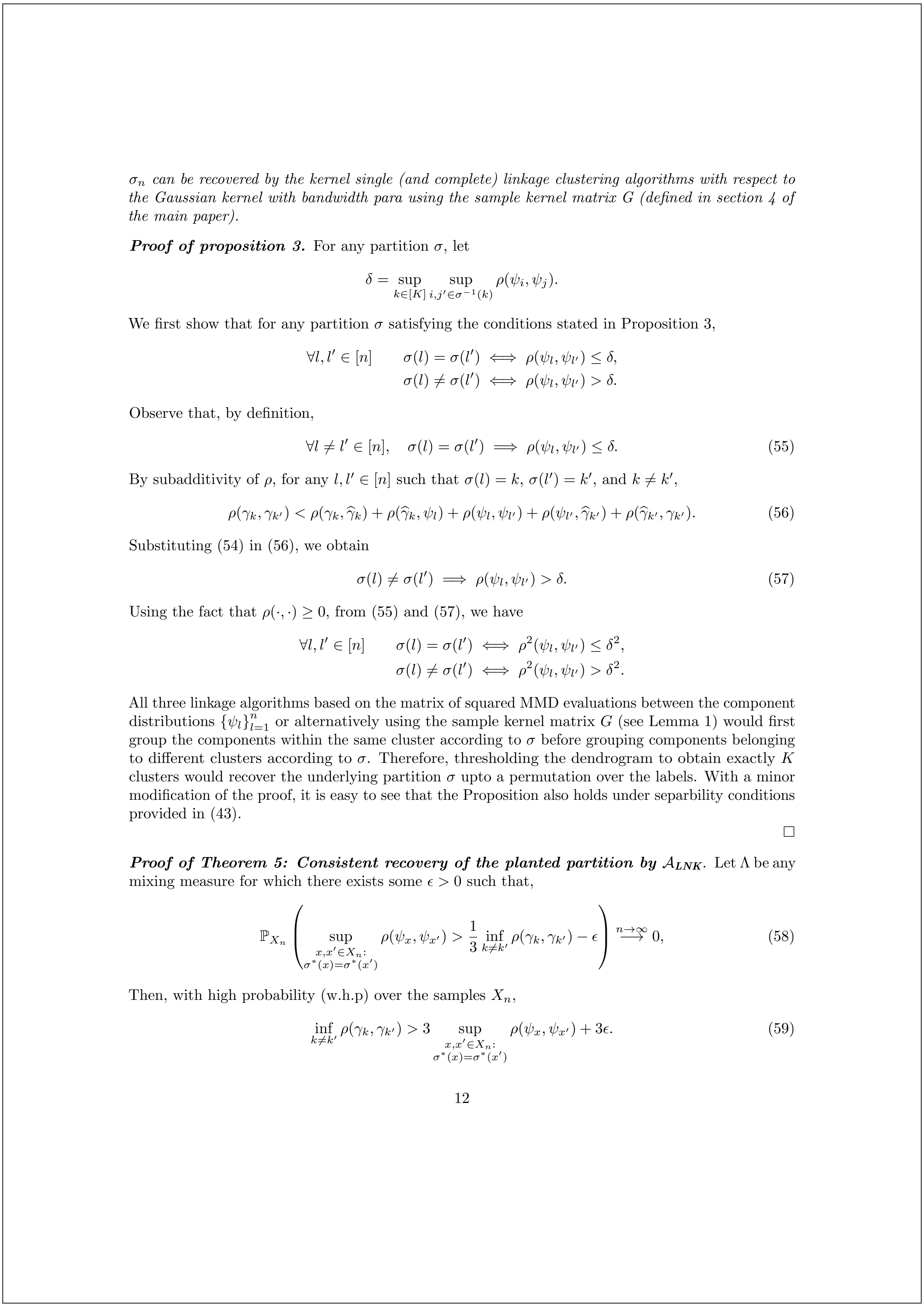
Therefore, the clusters obtained in the phase 1 of the algorithm do not change in the Phase 2 of the algorithm and the partition obtained by $\mathcal{A}_{\mathrm{FFK}}$coincides with that of σ' up to a permutation over the labels.

■ □

D.3 Sufficient conditions for kernel linkage clustering algorithms (Proof of Theorem 2 - Part III)

Proposition 3 (**Recovery by single linkage clustering**). *For any $\Lambda \in \mathcal{P}_K^2$, let $\Gamma = m(\Lambda)$. Let $X_n = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$ be a sample. Define $\widehat{\Gamma} = \sum\limits_{i=1}^{n} \frac{1}{n} \psi_i$ as the probability measure associated with the kde in the usual way. For any partition σ_n such that the following condition holds:*

$$\inf_{k \neq k'} \rho(\gamma_k, \gamma'_k) > 3 \sup_k \sup_{l \neq l' \in \sigma_n^{-1}(k)} \rho(\psi_l, \psi_{l'}) + 2 \sup_{k \in [K]} \rho(\widehat{\gamma}_{k,\sigma_n}, \gamma_{k,\sigma_n}), \tag{54}$$



σ_n can be recovered by the kernel single (and complete) linkage clustering algorithms with respect to the Gaussian kernel with bandwidth para using the sample kernel matrix G (defined in section 4 of the main paper).

Proof of proposition 3. For any partition σ, let

$$\delta = \sup_{k\in[K]} \sup_{i,j'\in\sigma^{-1}(k)} \rho(\psi_i, \psi_j).$$

We first show that for any partition σ satisfying the conditions stated in Proposition 3,

$$\begin{aligned} \forall l, l' \in [n] \qquad \sigma(l) = \sigma(l') &\iff \rho(\psi_l, \psi_{l'}) \leq \delta, \\ \sigma(l) \neq \sigma(l') &\iff \rho(\psi_l, \psi_{l'}) > \delta. \end{aligned}$$

Observe that, by definition,

$$\forall l \neq l' \in [n], \quad \sigma(l) = \sigma(l') \implies \rho(\psi_l, \psi_{l'}) \leq \delta. \tag{55}$$

By subadditivity of ρ, for any $l, l' \in [n]$ such that $\sigma(l) = k$, $\sigma(l') = k'$, and $k \neq k'$,

$$\rho(\gamma_k, \gamma_{k'}) < \rho(\gamma_k, \widehat{\gamma}_k) + \rho(\widehat{\gamma}_k, \psi_l) + \rho(\psi_l, \psi_{l'}) + \rho(\psi_{l'}, \widehat{\gamma}_{k'}) + \rho(\widehat{\gamma}_{k'}, \gamma_{k'}). \tag{56}$$

Substituting (54) in (56), we obtain

$$\sigma(l) \neq \sigma(l') \implies \rho(\psi_l, \psi_{l'}) > \delta. \tag{57}$$

Using the fact that $\rho(\cdot, \cdot) \geq 0$, from (55) and (57), we have

$$\begin{aligned} \forall l, l' \in [n] \qquad \sigma(l) = \sigma(l') &\iff \rho^2(\psi_l, \psi_{l'}) \leq \delta^2, \\ \sigma(l) \neq \sigma(l') &\iff \rho^2(\psi_l, \psi_{l'}) > \delta^2. \end{aligned}$$

All three linkage algorithms based on the matrix of squared MMD evaluations between the component distributions $\{\psi_l\}_{l=1}^n$ or alternatively using the sample kernel matrix G (see Lemma 1) would first group the components within the same cluster according to σ before grouping components belonging to different clusters according to σ. Therefore, thresholding the dendrogram to obtain exactly K clusters would recover the underlying partition σ upto a permutation over the labels. With a minor modification of the proof, it is easy to see that the Proposition also holds under separbility conditions provided in (43).

□

Proof of Theorem 5: Consistent recovery of the planted partition by $\mathcal{A}_{LNK}$. Let Λ be any mixing measure for which there exists some $\epsilon > 0$ such that,

$$\mathbb{P}_{X_n}\left(\sup_{\substack{x,x'\in X_n: \\ \sigma^*(x)=\sigma^*(x')}} \rho(\psi_x, \psi_{x'}) > \frac{1}{3} \inf_{k\neq k'} \rho(\gamma_k, \gamma_{k'}) - \epsilon \right) \xrightarrow{n\to\infty} 0, \tag{58}$$

Then, with high probability (w.h.p) over the samples X_n,

$$\inf_{k\neq k'} \rho(\gamma_k, \gamma_{k'}) > 3 \sup_{\substack{x,x'\in X_n: \\ \sigma^*(x)=\sigma^*(x')}} \rho(\psi_x, \psi_{x'}) + 3\epsilon. \tag{59}$$

Furthermore, we know that for every $\epsilon > 0$,

$$\mathbb{P}(\rho(\widehat{\gamma}_{k,\sigma^*}, \gamma_k) > \epsilon) \overset{n\to\infty}{\longrightarrow} 0. \tag{60}$$

Let $t = 3\epsilon/2$ and $\delta = 1/n$. Then, for every $k \in [K]$, there exists some $N_t \in \mathbb{N}$ such that $\forall\ n > N_{t,k}$,

$$\mathbb{P}(\rho(\widehat{\gamma}_{k,\sigma^*}, \gamma_k) > 3\epsilon/2) < \frac{1}{n}. \tag{61}$$

Let $N_t = \sup_{k\in[K]} N_{t,k}$. For all $n > N_t$, with high probability (w.h.p) over the samples X_n,

$$\inf_{k\neq k'} \rho(\gamma_k, \gamma_{k'}) > 3 \sup_{\substack{x,x'\in X_n:\\ \sigma^*(x)=\sigma^*(x')}} \rho(\psi_x, \psi_{x'}) + 2\rho(\widehat{\gamma}_{k,\sigma^*}, \gamma_k). \tag{62}$$

From Proposition 3, we have that w.h.p over X_n, kernel single linkage clustering algorithm recovers the true partition σ^* (upto a permutation over the labels).

□

E Necessary conditions for consistency of $\mathcal{A}_{\mathbf{FFK}}$ and $\mathcal{A}_{\mathbf{LNK}}$. (Proof of Theorem 3)

E.1 Proof for $\mathcal{A}_{\mathbf{FFK}}$

Fix the kernel bandwidth parameter $\zeta > 0$. Let r, ϵ and K be small constants that satisfy $1 > r > 2K > 16\epsilon$. Consider the following example in $\mathbb{R}$, where $\mathcal{U}([a, b])$ denotes the uniform distribution on the real interval $[a, b]$. Let

$$\gamma_1 = m\left(\frac{1}{2}\mathcal{U}([-\epsilon, \epsilon]) + \frac{1}{2}\mathcal{U}([r-\epsilon, r+\epsilon])\right) \tag{63}$$

and

$$\gamma_2 = m\left(\frac{1}{2}\mathcal{U}([2r-K-\epsilon, 2r-K+\epsilon]) + \frac{1}{2}\mathcal{U}([3r-K-\epsilon, 3r-K+\epsilon])\right). \tag{64}$$

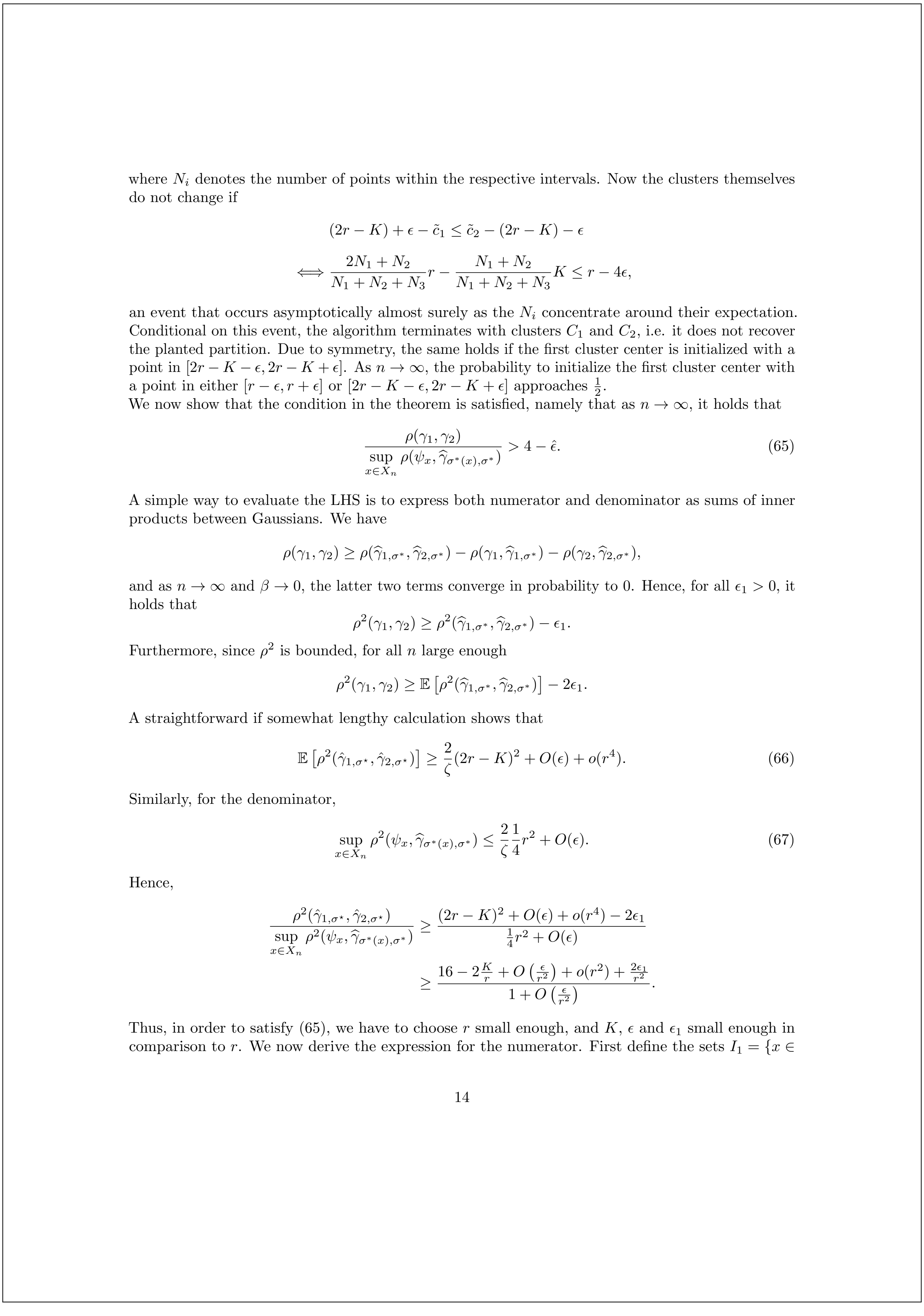
The mixing measure is given by $\Lambda = \frac{1}{2}\gamma_1 + \frac{1}{2}\gamma_2$. The idea is that because $K > 0$, the two clusters are just not separated enough.

To see that $\mathcal{A}_{\text{FFK}}$ fails to recover the planted partition with probability approaching $\frac{1}{2}$, consider the case where the first cluster center is initialized with a point $c_1 \in [r-\epsilon, r+\epsilon]$. The farthest first heuristic then chooses a second cluster center $c_2 \in [3r - K\epsilon, 3r - K + \epsilon]$. Since $K > 4\epsilon$, the initial clusters will be given by

$$C_1 = \{x : x \leq 2r - K + \epsilon\} \quad \text{and} \quad C_2 = \{x : x \geq 3r - K - \epsilon\}.$$

Consequently, in the first iteration of phase two of the algorithm (compare section B.2), the new cluster centers satisfy

$$\tilde{c}_1 \geq \frac{rN_2 + (2r-K)N_3}{N_1 + N_2 + N_3} - \epsilon \quad \text{and} \quad \tilde{c}_2 \geq 3r - K - \epsilon,$$



where N_i denotes the number of points within the respective intervals. Now the clusters themselves do not change if

$$(2r-K)+\epsilon-\tilde{c}_1 \leq \tilde{c}_2-(2r-K)-\epsilon$$

$$\iff \frac{2N_1+N_2}{N_1+N_2+N_3}r-\frac{N_1+N_2}{N_1+N_2+N_3}K \leq r-4\epsilon,$$

an event that occurs asymptotically almost surely as the N_i concentrate around their expectation. Conditional on this event, the algorithm terminates with clusters C_1 and C_2, i.e. it does not recover the planted partition. Due to symmetry, the same holds if the first cluster center is initialized with a point in $[2r-K-\epsilon, 2r-K+\epsilon]$. As $n \to \infty$, the probability to initialize the first cluster center with a point in either $[r-\epsilon, r+\epsilon]$ or $[2r-K-\epsilon, 2r-K+\epsilon]$ approaches $\frac{1}{2}$.
We now show that the condition in the theorem is satisfied, namely that as $n \to \infty$, it holds that

$$\frac{\rho(\gamma_1,\gamma_2)}{\sup\limits_{x\in X_n} \rho(\psi_x, \widehat{\gamma}_{\sigma^*(x),\sigma^*})} > 4-\hat{\epsilon}. \tag{65}$$

A simple way to evaluate the LHS is to express both numerator and denominator as sums of inner products between Gaussians. We have

$$\rho(\gamma_1,\gamma_2) \geq \rho(\widehat{\gamma}_{1,\sigma^*},\widehat{\gamma}_{2,\sigma^*})-\rho(\gamma_1,\widehat{\gamma}_{1,\sigma^*})-\rho(\gamma_2,\widehat{\gamma}_{2,\sigma^*}),$$

and as $n \to \infty$ and $\beta \to 0$, the latter two terms converge in probability to 0. Hence, for all $\epsilon_1 > 0$, it holds that

$$\rho^2(\gamma_1,\gamma_2) \geq \rho^2(\widehat{\gamma}_{1,\sigma^*},\widehat{\gamma}_{2,\sigma^*})-\epsilon_1.$$

Furthermore, since ρ^2 is bounded, for all n large enough

$$\rho^2(\gamma_1,\gamma_2) \geq \mathbb{E}\left[\rho^2(\widehat{\gamma}_{1,\sigma^*},\widehat{\gamma}_{2,\sigma^*})\right]-2\epsilon_1.$$

A straightforward if somewhat lengthy calculation shows that

$$\mathbb{E}\left[\rho^2(\hat{\gamma}_{1,\sigma^\star},\hat{\gamma}_{2,\sigma^\star})\right] \geq \frac{2}{\zeta}(2r-K)^2+O(\epsilon)+o(r^4). \tag{66}$$

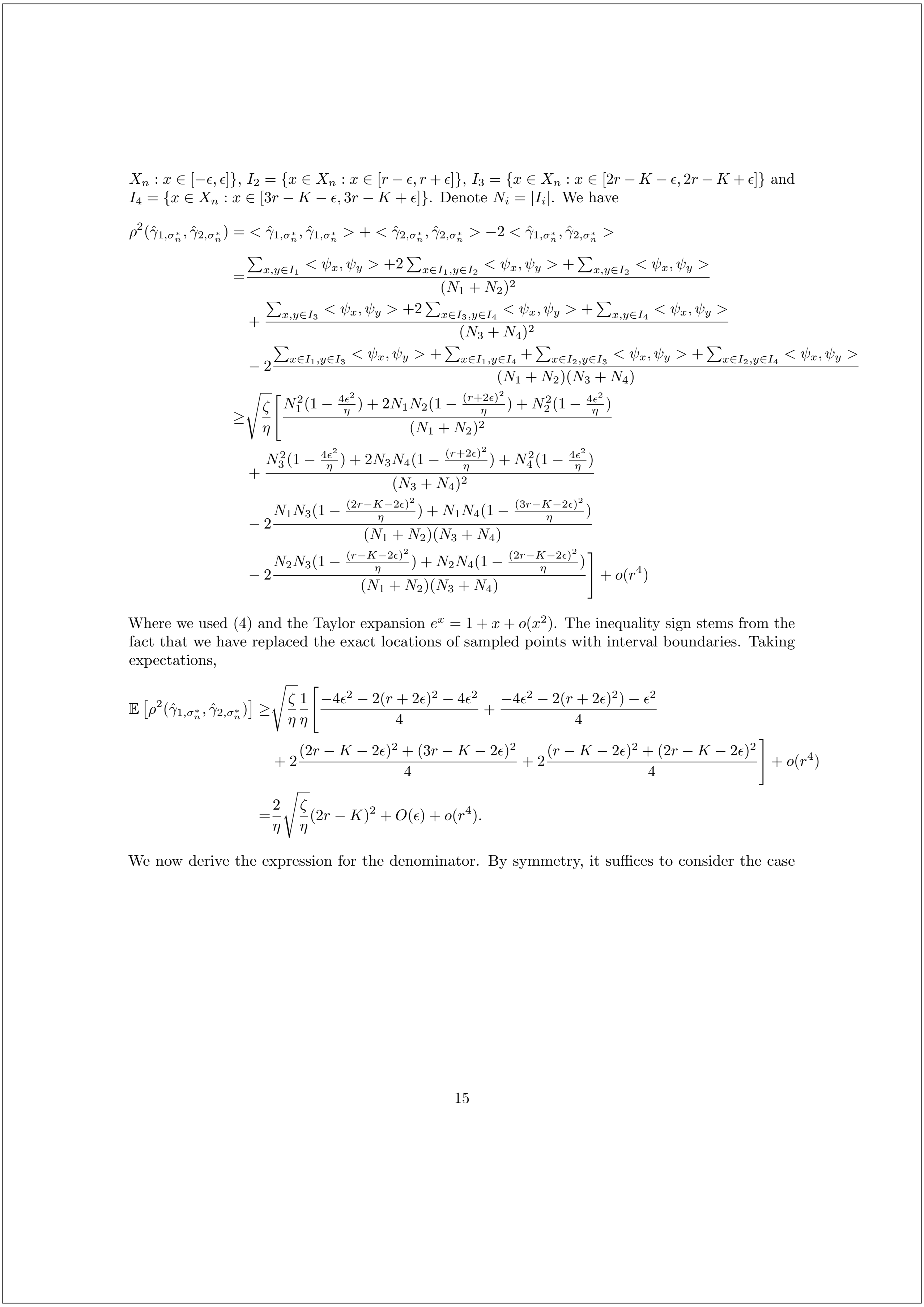
Similarly, for the denominator,

$$\sup_{x\in X_n} \rho^2(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*}) \leq \frac{2}{\zeta}\frac{1}{4}r^2+O(\epsilon). \tag{67}$$

Hence,

$$\begin{aligned}\frac{\rho^2(\hat{\gamma}_{1,\sigma^\star},\hat{\gamma}_{2,\sigma^\star})}{\sup\limits_{x\in X_n}\rho^2(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*})} &\geq \frac{(2r-K)^2+O(\epsilon)+o(r^4)-2\epsilon_1}{\frac{1}{4}r^2+O(\epsilon)}\\ &\geq \frac{16-2\frac{K}{r}+O\left(\frac{\epsilon}{r^2}\right)+o(r^2)+\frac{2\epsilon_1}{r^2}}{1+O\left(\frac{\epsilon}{r^2}\right)}.\end{aligned}$$

Thus, in order to satisfy (65), we have to choose r small enough, and K, ϵ and ϵ_1 small enough in comparison to r. We now derive the expression for the numerator. First define the sets $I_1 = \{x \in$



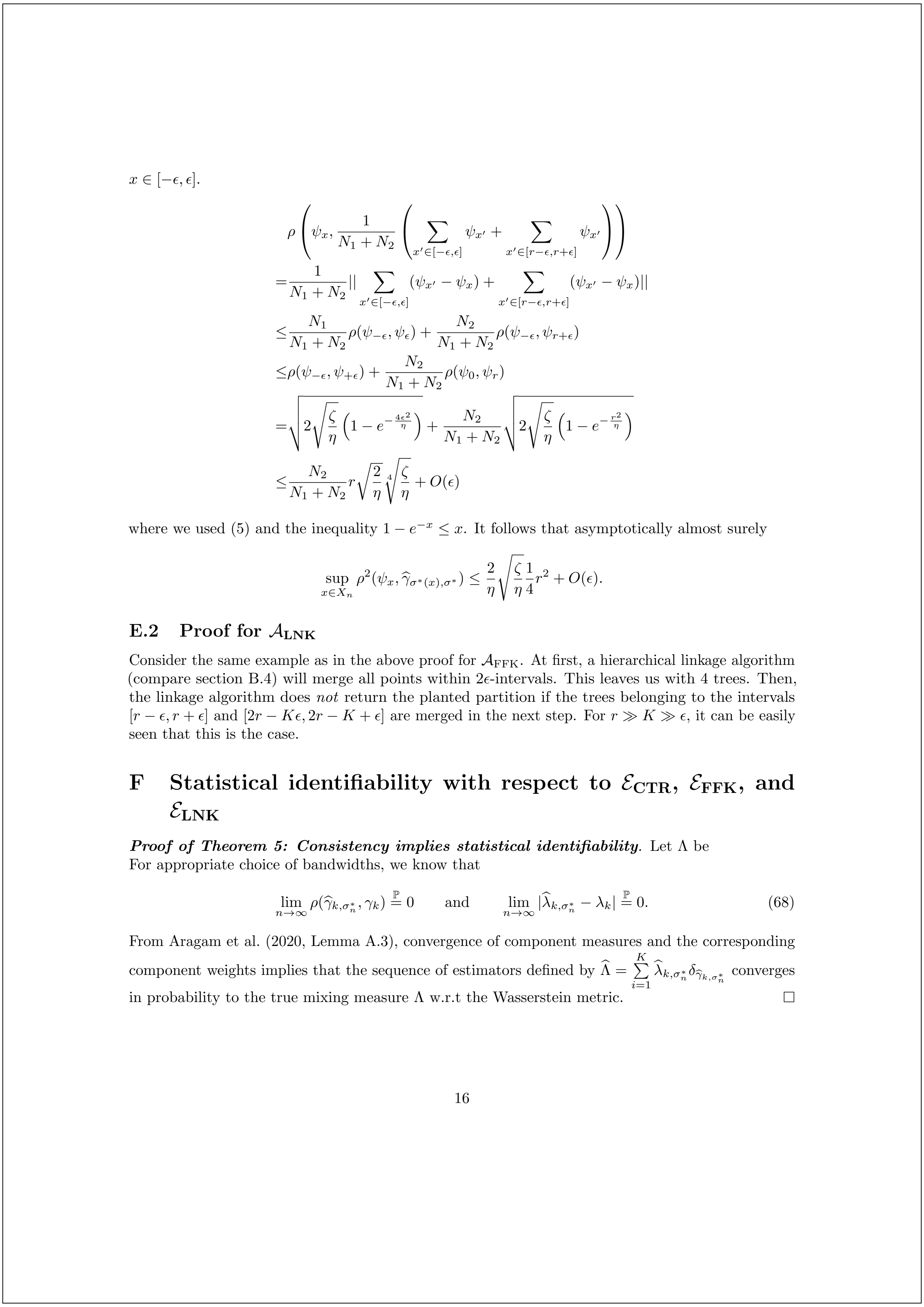
$X_n : x \in [-\epsilon, \epsilon]\}$, $I_2 = \{x \in X_n : x \in [r-\epsilon, r+\epsilon]\}$, $I_3 = \{x \in X_n : x \in [2r-K-\epsilon, 2r-K+\epsilon]\}$ and $I_4 = \{x \in X_n : x \in [3r-K-\epsilon, 3r-K+\epsilon]\}$. Denote $N_i = |I_i|$. We have

$$\begin{aligned}
\rho^2(\hat{\gamma}_{1,\sigma_n^*}, \hat{\gamma}_{2,\sigma_n^*}) = &< \hat{\gamma}_{1,\sigma_n^*}, \hat{\gamma}_{1,\sigma_n^*} > + < \hat{\gamma}_{2,\sigma_n^*}, \hat{\gamma}_{2,\sigma_n^*} > -2 < \hat{\gamma}_{1,\sigma_n^*}, \hat{\gamma}_{2,\sigma_n^*} > \\
=& \frac{\sum_{x,y\in I_1} < \psi_x, \psi_y > +2\sum_{x\in I_1, y\in I_2} < \psi_x, \psi_y > + \sum_{x,y\in I_2} < \psi_x, \psi_y >}{(N_1+N_2)^2} \\
&+ \frac{\sum_{x,y\in I_3} < \psi_x, \psi_y > +2\sum_{x\in I_3, y\in I_4} < \psi_x, \psi_y > + \sum_{x,y\in I_4} < \psi_x, \psi_y >}{(N_3+N_4)^2} \\
&- 2\frac{\sum_{x\in I_1, y\in I_3} < \psi_x, \psi_y > + \sum_{x\in I_1, y\in I_4} + \sum_{x\in I_2, y\in I_3} < \psi_x, \psi_y > + \sum_{x\in I_2, y\in I_4} < \psi_x, \psi_y >}{(N_1+N_2)(N_3+N_4)} \\
\geq& \sqrt{\frac{\zeta}{\eta}}\left[\frac{N_1^2(1-\frac{4\epsilon^2}{\eta}) + 2N_1N_2(1-\frac{(r+2\epsilon)^2}{\eta}) + N_2^2(1-\frac{4\epsilon^2}{\eta})}{(N_1+N_2)^2}\right. \\
&+ \frac{N_3^2(1-\frac{4\epsilon^2}{\eta}) + 2N_3N_4(1-\frac{(r+2\epsilon)^2}{\eta}) + N_4^2(1-\frac{4\epsilon^2}{\eta})}{(N_3+N_4)^2} \\
&- 2\frac{N_1N_3(1-\frac{(2r-K-2\epsilon)^2}{\eta}) + N_1N_4(1-\frac{(3r-K-2\epsilon)^2}{\eta})}{(N_1+N_2)(N_3+N_4)} \\
&\left.- 2\frac{N_2N_3(1-\frac{(r-K-2\epsilon)^2}{\eta}) + N_2N_4(1-\frac{(2r-K-2\epsilon)^2}{\eta})}{(N_1+N_2)(N_3+N_4)}\right] + o(r^4)
\end{aligned}$$

Where we used (4) and the Taylor expansion $e^x = 1 + x + o(x^2)$. The inequality sign stems from the fact that we have replaced the exact locations of sampled points with interval boundaries. Taking expectations,

$$\begin{aligned}
\mathbb{E}\left[\rho^2(\hat{\gamma}_{1,\sigma_n^*}, \hat{\gamma}_{2,\sigma_n^*})\right] \geq& \sqrt{\frac{\zeta}{\eta}}\frac{1}{\eta}\left[\frac{-4\epsilon^2 - 2(r+2\epsilon)^2 - 4\epsilon^2}{4} + \frac{-4\epsilon^2 - 2(r+2\epsilon)^2) - \epsilon^2}{4}\right. \\
&\left.+ 2\frac{(2r-K-2\epsilon)^2 + (3r-K-2\epsilon)^2}{4} + 2\frac{(r-K-2\epsilon)^2 + (2r-K-2\epsilon)^2}{4}\right] + o(r^4) \\
=& \frac{2}{\eta}\sqrt{\frac{\zeta}{\eta}}(2r-K)^2 + O(\epsilon) + o(r^4).
\end{aligned}$$

We now derive the expression for the denominator. By symmetry, it suffices to consider the case



$x \in [-\epsilon, \epsilon]$.

$$
\begin{aligned}
&\rho\left(\psi_x, \frac{1}{N_1+N_2}\left(\sum_{x'\in[-\epsilon,\epsilon]}\psi_{x'} + \sum_{x'\in[r-\epsilon,r+\epsilon]}\psi_{x'}\right)\right)\\
=&\frac{1}{N_1+N_2}|| \sum_{x'\in[-\epsilon,\epsilon]}(\psi_{x'}-\psi_x) + \sum_{x'\in[r-\epsilon,r+\epsilon]}(\psi_{x'}-\psi_x)||\\
\le&\frac{N_1}{N_1+N_2}\rho(\psi_{-\epsilon},\psi_\epsilon) + \frac{N_2}{N_1+N_2}\rho(\psi_{-\epsilon},\psi_{r+\epsilon})\\
\le&\rho(\psi_{-\epsilon},\psi_{+\epsilon}) + \frac{N_2}{N_1+N_2}\rho(\psi_0,\psi_r)\\
=&\sqrt{2\sqrt{\frac{\zeta}{\eta}}\left(1-e^{-\frac{4\epsilon^2}{\eta}}\right)} + \frac{N_2}{N_1+N_2}\sqrt{2\sqrt{\frac{\zeta}{\eta}}\left(1-e^{-\frac{r^2}{\eta}}\right)}\\
\le&\frac{N_2}{N_1+N_2}r\sqrt{\frac{2}{\eta}}\sqrt[4]{\frac{\zeta}{\eta}} + O(\epsilon)
\end{aligned}
$$

where we used (5) and the inequality $1 - e^{-x} \le x$. It follows that asymptotically almost surely

$$\sup_{x\in X_n} \rho^2(\psi_x, \widehat{\gamma}_{\sigma^*(x),\sigma^*}) \le \frac{2}{\eta}\sqrt{\frac{\zeta}{\eta}}\frac{1}{4}r^2 + O(\epsilon).$$

E.2 Proof for $\mathcal{A}_{\mathbf{LNK}}$

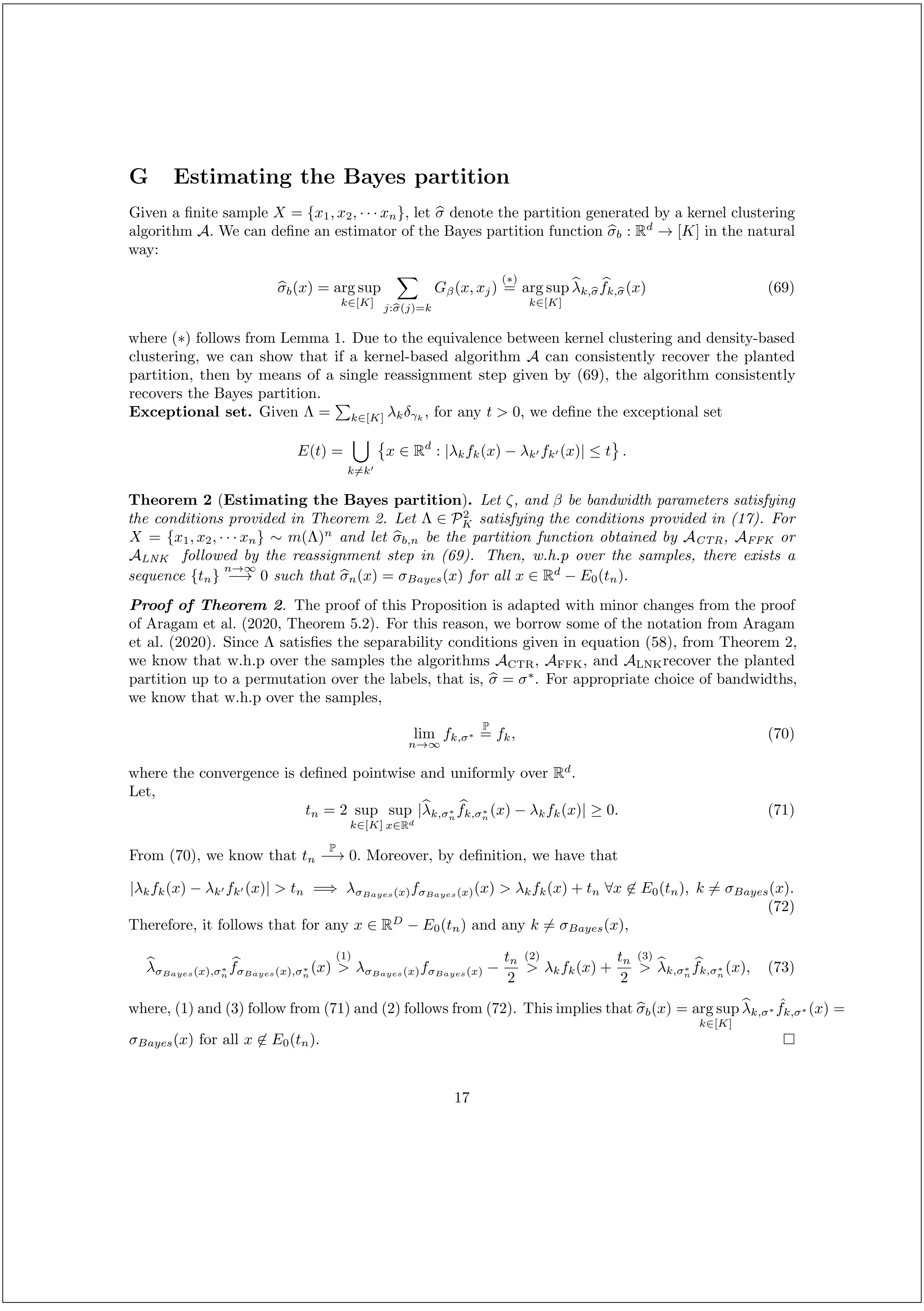
Consider the same example as in the above proof for $\mathcal{A}_{\text{FFK}}$. At first, a hierarchical linkage algorithm (compare section B.4) will merge all points within 2ϵ-intervals. This leaves us with 4 trees. Then, the linkage algorithm does *not* return the planted partition if the trees belonging to the intervals $[r-\epsilon, r+\epsilon]$ and $[2r - K\epsilon, 2r - K + \epsilon]$ are merged in the next step. For $r \gg K \gg \epsilon$, it can be easily seen that this is the case.

F Statistical identifiability with respect to $\mathcal{E}_{\mathbf{CTR}}$, $\mathcal{E}_{\mathbf{FFK}}$, and $\mathcal{E}_{\mathbf{LNK}}$

Proof of Theorem 5: Consistency implies statistical identifiability. Let Λ be
For appropriate choice of bandwidths, we know that

$$\lim_{n\to\infty}\rho(\widehat{\gamma}_{k,\sigma_n^*},\gamma_k) \overset{\mathbb{P}}{=} 0 \qquad \text{and} \qquad \lim_{n\to\infty}|\widehat{\lambda}_{k,\sigma_n^*} - \lambda_k| \overset{\mathbb{P}}{=} 0. \tag{68}$$

From Aragam et al. (2020, Lemma A.3), convergence of component measures and the corresponding component weights implies that the sequence of estimators defined by $\widehat{\Lambda} = \sum_{i=1}^{K}\widehat{\lambda}_{k,\sigma_n^*}\delta_{\widehat{\gamma}_{k,\sigma_n^*}}$ converges in probability to the true mixing measure Λ w.r.t the Wasserstein metric. □



G Estimating the Bayes partition

Given a finite sample $X = \{x_1, x_2, \cdots x_n\}$, let $\widehat{\sigma}$ denote the partition generated by a kernel clustering algorithm $\mathcal{A}$. We can define an estimator of the Bayes partition function $\widehat{\sigma}_b : \mathbb{R}^d \to [K]$ in the natural way:

$$\widehat{\sigma}_b(x) = \arg\sup_{k \in [K]} \sum_{j : \widehat{\sigma}(j) = k} G_\beta(x, x_j) \overset{(*)}{=} \arg\sup_{k \in [K]} \widehat{\lambda}_{k,\widehat{\sigma}} \widehat{f}_{k,\widehat{\sigma}}(x) \tag{69}$$

where $(*)$ follows from Lemma 1. Due to the equivalence between kernel clustering and density-based clustering, we can show that if a kernel-based algorithm $\mathcal{A}$ can consistently recover the planted partition, then by means of a single reassignment step given by (69), the algorithm consistently recovers the Bayes partition.

Exceptional set. Given $\Lambda = \sum_{k \in [K]} \lambda_k \delta_{\gamma_k}$, for any $t > 0$, we define the exceptional set

$$E(t) = \bigcup_{k \neq k'} \left\{ x \in \mathbb{R}^d : |\lambda_k f_k(x) - \lambda_{k'} f_{k'}(x)| \leq t \right\}.$$

Theorem 2 (**Estimating the Bayes partition**). *Let ζ, and β be bandwidth parameters satisfying the conditions provided in Theorem 2. Let $\Lambda \in \mathcal{P}_K^2$ satisfying the conditions provided in (17). For $X = \{x_1, x_2, \cdots x_n\} \sim m(\Lambda)^n$ and let $\widehat{\sigma}_{b,n}$ be the partition function obtained by $\mathcal{A}_{CTR}$, $\mathcal{A}_{FFK}$ or $\mathcal{A}_{LNK}$ followed by the reassignment step in (69). Then, w.h.p over the samples, there exists a sequence $\{t_n\} \overset{n \to \infty}{\longrightarrow} 0$ such that $\widehat{\sigma}_n(x) = \sigma_{Bayes}(x)$ for all $x \in \mathbb{R}^d - E_0(t_n)$.*

Proof of Theorem 2. The proof of this Proposition is adapted with minor changes from the proof of Aragam et al. (2020, Theorem 5.2). For this reason, we borrow some of the notation from Aragam et al. (2020). Since Λ satisfies the separability conditions given in equation (58), from Theorem 2, we know that w.h.p over the samples the algorithms $\mathcal{A}_{\text{CTR}}$, $\mathcal{A}_{\text{FFK}}$, and $\mathcal{A}_{\text{LNK}}$recover the planted partition up to a permutation over the labels, that is, $\widehat{\sigma} = \sigma^*$. For appropriate choice of bandwidths, we know that w.h.p over the samples,

$$\lim_{n \to \infty} f_{k,\sigma^*} \overset{\mathbb{P}}{=} f_k, \tag{70}$$

where the convergence is defined pointwise and uniformly over $\mathbb{R}^d$.

Let,

$$t_n = 2 \sup_{k \in [K]} \sup_{x \in \mathbb{R}^d} |\widehat{\lambda}_{k,\sigma_n^*} \widehat{f}_{k,\sigma_n^*}(x) - \lambda_k f_k(x)| \geq 0. \tag{71}$$

From (70), we know that $t_n \overset{\mathbb{P}}{\longrightarrow} 0$. Moreover, by definition, we have that

$$|\lambda_k f_k(x) - \lambda_{k'} f_{k'}(x)| > t_n \implies \lambda_{\sigma_{Bayes}(x)} f_{\sigma_{Bayes}(x)}(x) > \lambda_k f_k(x) + t_n \; \forall x \notin E_0(t_n), \; k \neq \sigma_{Bayes}(x). \tag{72}$$

Therefore, it follows that for any $x \in \mathbb{R}^D - E_0(t_n)$ and any $k \neq \sigma_{Bayes}(x)$,

$$\widehat{\lambda}_{\sigma_{Bayes}(x),\sigma_n^*} \widehat{f}_{\sigma_{Bayes}(x),\sigma_n^*}(x) \overset{(1)}{>} \lambda_{\sigma_{Bayes}(x)} f_{\sigma_{Bayes}(x)} - \frac{t_n}{2} \overset{(2)}{>} \lambda_k f_k(x) + \frac{t_n}{2} \overset{(3)}{>} \widehat{\lambda}_{k,\sigma_n^*} \widehat{f}_{k,\sigma_n^*}(x), \tag{73}$$

where, (1) and (3) follow from (71) and (2) follows from (72). This implies that $\widehat{\sigma}_b(x) = \arg\sup_{k \in [K]} \widehat{\lambda}_{k,\sigma^*} \hat{f}_{k,\sigma^*}(x) = \sigma_{Bayes}(x)$ for all $x \notin E_0(t_n)$. □

8
Causal generalization in autoregressive models

Causal Forecasting:
Generalization Bounds for Autoregressive Models

Leena Chennuru Vankadara*[1] Philipp Michael Faller[2] Michaela Hardt[2] Lenon Minorics[2]
Debarghya Ghoshdastidar[3] Dominik Janzing[2]

[1]University of Tübingen
[2]Amazon Research
[3]Technical University of Munich, Munich Data Science Institute.

Abstract

Despite the increasing relevance of forecasting methods, causal implications of these algorithms remain largely unexplored. This is concerning considering that, even under simplifying assumptions such as causal sufficiency, the statistical risk of a model can differ significantly from its *causal risk*. Here, we study the problem of *causal generalization*—generalizing from the observational to interventional distributions—in forecasting. Our goal is to find answers to the question: How does the efficacy of an autoregressive (VAR) model in predicting statistical associations compare with its ability to predict under interventions? To this end, we introduce the framework of *causal learning theory* for forecasting. Using this framework, we obtain a characterization of the difference between statistical and causal risks, which helps identify sources of divergence between them. Under causal sufficiency, the problem of causal generalization amounts to learning under covariate shifts albeit with additional structure (restriction to interventional distributions under the VAR model). This structure allows us to obtain uniform convergence bounds on causal generalizability for the class of VAR models. To the best of our knowledge, this is the first work that provides theoretical guarantees for causal generalization in the time-series setting.

1 INTRODUCTION

Forecasting algorithms are increasingly relevant in a variety of applications including meteorology, climatology, economics, and business. While traditional economic modelling relies on relatively simple time series models (Brockwell et al. 1991), e.g., autoregressive models, or methods like co-integration, modern business planning heavily uses neural networks for forecasting (Faloutsos et al. 2018; Januschowski et al. 2020; Salinas et al. 2020). Despite the advancements of forecast quality, causal implications are not yet well understood. There has been notable progress in 'explainable' models in the sense of feature relevance (Lundberg et al. 2017; Molnar 2019; Janzing et al. 2020; Wang et al. 2020) with potential applications in forecasting. Furthermore, specialized models (Hatt et al. 2021; Bica et al. 2020; Lim et al. 2018) have shown remarkable success for causal inference in forecasting.

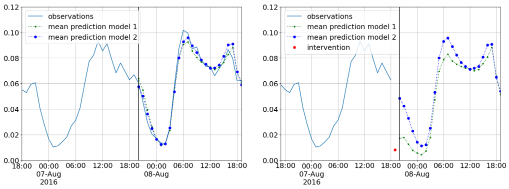


Figure 1: An example time series with predictions of two DeepAR models (top) under an intervention in red (bottom) on the Traffic dataset. While we do not know the ground-truth, we see that two models disagree when faced with an intervention more than on the in-distribution forecasting. Since at most one of them can be right, we conclude that at least the other one makes a notable forecasting error under the intervention.

It is common practice in business and econometrics to learn statistical forecasting models and interpret them causally. In practice, while forecasting models tend to agree on their statistical predictions, they can differ substantially on their causal predictions (see Figure 1 for an example). In particular, this practice is considered justified under simplifying assumptions such as causal sufficiency and the absence of contemporaneous effects (see for instance Hyvärinen et al. (2010, Section 1)). Here, we are interested in the funda-

*Part of this work was completed during an internship at Amazon Research.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

mental question: what is the relation between the statistical predictability of a forecasting model and its causal generalizability — ability to predict under interventions.

We argue that even for very simple models and even under simplifying assumptions such as causal sufficiency and absence of contemporaneous influence, causal interpretation of forecasting models is non-trivial. To appreciate the challenges, consider a simple example of a process with strongly correlated observations where $x_t \approx x_{t-1}$, and hence $x_t \approx x_{t-2}$. These observations can be explained either by a causal model with a strong influence of x_{t-1} on x_t or a causal model with a strong influence from x_{t-2} on x_t. The difference between the models gets apparent when an intervention randomizes x_{t-1} and x_{t-2} independently. Then, predictions become hard, particularly when x_{t-1} and x_{t-2} are set to significantly different values. While both models are similar in their statistical predictions, they differ substantially in their *causal predictions*. This example already shows that, even in a simple setting, causal and statistical predictability can differ significantly. The question of causal generalization is thus practically relevant and non-trivial and begs for a better theoretical understanding.

Specifically, we consider the simple class of vector autoregressive models (VAR) and ask the question

How does the efficacy of an autoregressive model in predicting statistical associations compare with its ability to predict under interventions?

These models are widely applied in domains ranging from econometrics (Lütkepohl 2009; Grabowski et al. 2020) and finance (Zivot et al. 2006) to neuroscience (Valdés-Sosa et al. 2005).

Connection to Covariate Shift. The problem of causal generalization is closely related to the problem of covariate shift. To see this, we first ignore the time series setting and consider the scenario where a variable Y should be predicted from a variable X, which is known not to be an effect of Y. If there is no common cause of X and Y, that is, we assume causal sufficiency (Spirtes et al. 1993), the statistical relation between X and Y is entirely due to the influence of X on Y. Therefore, the observational and interventional conditionals coincide ($P_{Y|x=x^*} = P_{Y|do(x=x^*)}$ in Pearl's language (Pearl 2009)) and the true parameters would be optimal both from a statistical and causal perspective. However, due to *estimation bias*, a prediction model learned using finite samples from P_x may perform poorly when randomized interventions draw x-values from a different distribution $\tilde{P}_X$, which is the usual covariate shift scenario (Sugiyama et al. 2012). In our setting, X and Y are represented by the past and the present values of a (possibly multivariate) time series, respectively. Accordingly, we focus on interventional distributions that are natural for this setting: independent interventions at different time points and components of the multivariate process. Hence, we have additional structure in comparison with the standard covariate shift problem. We are not aware of any theoretical work on covariate shift in the time-series setting. Nevertheless, we describe the connections to learning theory in the standard covariate shift setting and other related work in Section 6.

Our Contributions. Our central goal in this work is to develop a formal and thorough understanding of causal generalization for the class of VAR models.

a. To this end, we introduce a framework of causal learning theory for forecasting to analyze when forecasting models can generalize from the *observational* to the *interventional distributions* (Section 2). This is closely related to the setting of learning under domain adaptation.
b. Using this framework, we provide a characterization of the difference in the statistical and *causal* risks (Section 3). Such a characterization allows us to identify the sources of divergence between the two quantities. Our results show that the strength of correlation of the underlying process plays a key role in determining causal generalizability. They also highlight that already for simple models, causal and statistical errors can even diverge.
c. Further, we provide finite-sample, uniform convergence bounds on causal generalization for the class of VAR models (Section 3). Our simulations demonstrate that our bounds indeed capture the key drivers of causal generalization. To the best of our knowledge, this is the first work that provides theoretical guarantees for causal generalization of any kind in the time-series setting.
d. As a by-product of our analysis, we provide an explicit characterization of the powers of a companion matrix (see Section 2) using symmetric Schur polynomials (Macdonald 1998) of its eigenvalues (Lemma 2) which, to the best of our knowledge, has not been noted in the literature. This result could be of independent interest in theoretical endeavors that build upon companion matrices which, for instance, are ubiquitous in stochastic processes and in Linear-Time-Invariant dynamical systems (Davison 1976; Melnyk et al. 2016).
e. We conduct experiments with a variety of deep neural networks on real data. Our experiments approach causal risks in this setting and explore its relationship to uncertainty.

2 CAUSAL LEARNING THEORY FOR FORECASTING

In this section, we introduce a framework to formally evaluate the quality of a forecasting model with respect to prediction and the validity of its causal implications. We refer to this framework as causal learning theory for forecasting. First, we introduce some relevant notation.

Notation. For any stochastic process $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$, we use $\mathbf{x}^n_{t-\omega} = \{x_{t-\omega-n+1}, \cdots, x_{t-\omega-1}, x_{t-\omega}\}$ to denote the *set* of $x_{t-\omega}$ and the $n-1$ variables in the past of $x_{t-\omega}$. We distinguish this from y^n_t which denotes the *vector* $\left(x_t, x_{t-1}, \cdots, x_{t-n+1}\right)^T \in \mathbb{R}^{nd}$. When it is clear from context, to reduce cumbersome notation, we simply use y_t. For any random variable x, $\mathbb{E}[x]$ denotes its expectation. For any matrix A, we use $A_{i:}$ and $A_{:j}$ to denote the ith row and jth column of A respectively. We use A^j_{1k} to denote the $(1,k)$th element of A^j. For any vector x_t at time t, we use $x_{t,i}$ to denote the ith element of x_t. We use $\lambda_{\max}(A), \lambda_{\min}(A), \kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$ to denote the maximum and minimum eigenvalues and the condition number of A respectively. $\mathbb{I}_p$ denotes the identity matrix of size p, $\mathbb{N}, \mathbb{Z}$ denote the set of natural numbers and integers respectively and $[n]$ denotes the set $\{1, 2, \cdots n\}$.

To evaluate the statistical and causal efficacy of an estimator we introduce the notions of statistical and *causal* forecast risks. To define statistical forecast risk, we consider the setting of $\omega-$step forecasting where the goal is to predict x_t from observations $\mathbf{x}^n_{t-\omega}$ drawn from a stochastic process $\{x_t\}_{t\in\mathbb{Z}}$ for some $\omega \in \mathbb{N}$. To define the causal forecast risk, we consider interventions on $x_{t-\omega,i}$ for some $i \in [d]$.[1]

Definition 2.1 (Statistical forecast error). The statistical forecast error of an estimator $\hat{f}$ in the prediction of a target variable x_t from $\mathbf{x}^n_{t-\omega}$, drawn from the *observational distribution*, can be defined as

$$\mathcal{S}_\omega = \mathbb{E}_{\mathbb{P}(x_t, \mathbf{x}^n_{t-\omega})}\left[\left(x_t - \hat{f}(x^n_{t-\omega})\right)^2\right]. \tag{1}$$

The empirical counterpart ($\hat{\mathcal{S}_\omega}$), is defined naturally by replacing the expectation by the empirical mean.

For causal questions, we want to investigate the behavior of a model under interventions. Here, we consider atomic interventions. Using Pearl's do notation (Pearl 2009), an atomic intervention $do(x = x^*)$ refers to *setting* the variable x to some value x^*.

Definition 2.2 (Causal errors). The interventional forecast error of $\hat{f}$ in predicting the *effect of an intervention* $do(x_{t-\omega,i} = x^*_{t-\omega,i})$, on target variable x_t is defined as

$$\mathcal{G}_{do_{\omega,i}} = \mathbb{E}_{\mathbb{P}_{do_{\omega,i}}(x_t, \mathbf{x}^n_{t-\omega})}\left[\left(x_t - \hat{f}(x^n_{t-\omega})\right)^2\right], \tag{2}$$

where $do_{\omega,i}$ is shorthand for $do(x_{t-\omega,i} = x^*_{t-\omega,i})$ and $\mathbb{P}_{do_{\omega,i}}$ denotes the distribution induced by the intervention $do(x_{t-\omega,i} = x^*_{t-\omega,i})$. To isolate from the dependence on specific values that the intervened variables are set to, we present our results via the notion of *average causal error*. It is defined as the expected interventional error for interventions drawn from the marginal distribution of $x_{t-\omega,i}$ since it provides a natural scale at which the statistical and causal errors can be compared.

$$\mathcal{G}_{\omega,i} = \mathbb{E}_{x^*_{t-\omega,i}\sim\mathbb{P}(x_{t-\omega,i})}\left[\mathcal{G}_{do_{\omega,i}}\right]. \tag{3}$$

Statistical and Causal Learning Theory. Consider the standard framework of statistical learning in time-series prediction. For any stochastic process $\{x_t\}_{t\in\mathbb{Z}}$ taking values in $\mathcal{X}$, given a loss function $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, the goal of statistical learning is to learn a function $f^*_{\mathcal{S}}$ that achieves the optimal statistical risk $\mathcal{S}^\omega(f)$: Since the true process is unknown, the empirical average ($\widehat{\mathcal{S}}^\omega$) of generalization risk is used to estimate $\mathcal{S}^\omega$. Statistical generalization bounds of the form: $\mathcal{S}^\omega(f) < \widehat{\mathcal{S}}^\omega(f) + \mathcal{C}(\mathcal{F}, n)$ are then used to provide guarantees on the uniform deviation of empirical risk from expected risk given sufficiently many samples and when the "complexity" of the function class is small.

Analogously, the goal of *causal learning* is to find a function $f^*_{\mathcal{G}}$ that achieves the optimal *causal* risk $\mathcal{G}^\omega(f)$ In contrast to statistical learning, the empirical averages of the causal error cannot be utilized to estimate $\mathcal{G}_\omega$ since we often do not have access to data from the interventional distributions. Instead, we are only provided with data from the observational/statistical distribution of the stochastic process and the goal of causal learning theory is to understand, to what extent is it possible to provide *causal generalization* guarantees of the form: $\mathcal{G}^\omega(f) < \widehat{\mathcal{S}}^\omega(f) + \mathcal{C}(\mathcal{F}, n)$.

To summarize, we ask: Can the predictors in $\mathcal{F}$ generalize from the *empirical observational distribution* to the *true interventional distribution* assuming that we control the complexity of $\mathcal{F}$ and that we observe sufficiently many samples drawn from the observational distribution? One cannot address this question in a very general setting and would need model assumptions to make any meaningful statements. To this end, we now formally introduce our problem setup and some preliminaries. We provide additional relevant background in the Appendix A.

Statistical and Causal Models. We assume that the stochastic process $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$ follows a weakly stationary vector autoregressive model(VAR(p)) of order p for some $p, d \in \mathbb{N}$ which is defined as

$$x_t = A_1x_{t-1} + A_2x_{t-2} + \cdots A_Px_{t-p} + \epsilon_t, \tag{4}$$

where $x_t \in \mathbb{R}^d$ is a vector-valued time-series, for all $i \in [p]$, $A_i \in \mathbb{R}^{d\times d}$ are the coefficients of the VAR model, and $\epsilon_t \in \mathbb{R}^d$ denotes the noise vector such that $\mathbb{E}[\epsilon_t] = 0$ and $\mathbb{E}[\epsilon_t\epsilon^T_{t+h}] = \Sigma_\epsilon$ if $h = 0$ and 0 otherwise. For some $\sigma^2_\epsilon > 0$, we simply set $\Sigma_\epsilon = \sigma^2_\epsilon\mathbb{I}$ for enhanced readability. Our results can be easily generalized to arbitrary covariance matrices by means of the spectral properties ($\lambda_{\min}, \lambda_{\max}$) of Σ_ϵ. The autocovariance matrix of $\{x_t\}_{t\in\mathbb{Z}}$ plays a central role in our results and analysis. For any $n \in \mathbb{N}$, we use Σ_n to denote the autocovariance matrix of size n defined as $\mathbb{E}[(y^n_t - \mathbb{E}[y^n_t])(y^n_t - \mathbb{E}[y^n_t])^T]$. It is convenient to rewrite

[1] The results for simultaneous interventions are qualitatively similar to those of interventions on single variables, and for ease of exposition, we present our discussion in the latter case.

a VAR model of order p in Equation (4) as a VAR(1) model, $y_t = Ay_{t-1} + e_t$, where $y_t \in \mathbb{R}^{dp}, e_t \in \mathbb{R}^{dp}$ are defined as $y_t = \left(x_t, x_{t-i}, \cdots, x_{t-p+1}\right)^T$, $e_t = \left(\epsilon_t, 0, \cdots, 0\right)^T$, and $A \in \mathbb{R}^{dp \times dp}$ is a *(multi) companion matrix* defined as:

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{pmatrix}. \tag{5}$$

The eigenvalues of the multi-companion matrix A fully characterize the stability and stationarity of the VAR process. For a VAR(p) process to be weakly stationary, that is for the mean and the covariance of the process to not change over time, the eigenvalues of A, which satisfy

$$\det|\mathbb{I}_d \lambda^p - A_1 \lambda^{p-1} - A_2 \lambda^{p-2} - \cdots - A_p| = 0, \tag{6}$$

are constrained to not lie on the unit circle. If the magnitude of all the eigenvalues are $|\lambda_i| < 1$, then the process is stable, that is, its values do not diverge (Lütkepohl 2013).

Causal Models. Under the assumptions of causal sufficiency and absence of contemporaneous influences, a causal interpretation of the VAR model in (4) as structural equations naturally yields the corresponding causal model. We consider the family of all VAR models as our function class $\mathcal{F}$ of statistical and causal estimators.

3 CAUSAL GENERALIZATION FOR VAR

In this section, we present causal generalization bounds for the family of VAR models under atomic interventions. We first provide an overview of our results in the more general case of VAR(p) models and later provide a thorough interpretation of the results, often by deriving simplified versions of the results for AR(p) models. We begin by providing an exact characterization of the difference in statistical and causal errors in terms of the model and estimated parameters and the autocovariance matrix of the underlying process.

Lemma 1 (Difference in Causal and Statistical errors (VAR)). *Consider a vector-valued time series $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$, following a VAR(q) process parameterized by $\{A_1, A_2, \cdots A_q\}$. Let $\nu = \max\{p, q\}$. For any VAR(p) model f with parameters $\{\hat{A}_1, \hat{A}_2, \cdots \hat{A}_p\}$,*

$$|\mathcal{G}_{\omega,i} - \mathcal{S}_\omega| = 2\left|(A^\omega_{ii} - \hat{A}^\omega_{ii}) \sum_{k\neq i}^{d\nu} (A^\omega_{ik} - \hat{A}^\omega_{ik})\Sigma^\nu_{ik}\right|,$$

where Σ^ν denotes the autocovariance matrix of x_t of size ν, A is a multi-companion matrix of the form described in (5) with the first d rows populated by $\{A'_1, A'_2, \cdots A'_\nu\}$, with A'_l defined as A_l for all $l \leq p$ and as $\mathbf{0}_{d\times d}$ for all $l > p$. $\hat{A}$ is analogously defined.

Building on Lemma 1, we establish that the condition number of the autocovariance matrix of the underlying process controls causal generalizability from the *observational* to *interventional distributions.*

Proposition 1 (Stability Controls Causal Generalization (VAR)). *Let $\{x_t\}_{t\in\mathbb{Z}}$ follow a VAR(q) process for some $q \in \mathbb{N}$. For any VAR(p) model,*

$$|\mathcal{G}_{\omega,i} - \mathcal{S}_\omega| \leq (2\kappa(\Sigma^\nu) - 1)(\mathcal{S}_\omega - \sigma^2_\epsilon), \tag{7}$$

where $\kappa(\Sigma^\nu)$ denotes the condition number of the autocovariance matrix Σ^ν. Further, one can construct processes where equality holds upto a small constant factor.

The result states that the difference in expected causal and statistical errors is controlled by the *condition number* of the autocovaraince matrix of size $\max\{p, q\}$. It also states that without incorporating additional information, one cannot obtain a much tighter bound which is also verified by our experiments in Section 4. The condition number of the autocovariance matrix can get arbitrarily large as the process gets closer to the boundary of the stability domain. This result therefore shows that even for very simple classes of forecasting models, causal interpretations can get challenging. We later provide a detailed interpretation of this result and provide an explicit bound on $\kappa(\Sigma_\nu)$ in terms of the stability parameter for AR(p) models (Corollary 2).

Proposition 1 allows us to employ generalization bounds for time-series (Yu 1994; Meir 2000; Mohri et al. 2009; McDonald et al. 2017) to derive finite-sample *causal generalization bounds* for VAR models. In particular, we utilize Rademacher complexity bounds for generalization in time-series under mixing conditions (Mohri et al. 2009) to derive Theorem 1.

Theorem 1 (Finite sample bounds for VAR(p) models). *Let $\mathcal{F}$ denote the family of all VAR models of dimension d and order p. For any $n > \max\{p, q\} \in \mathbb{N}$, let $\mu, m > 0$ be integers such that $2\mu m = n$ and $\delta > 2(\mu - 1)\rho^m$ for a fixed constant $0 < \rho < 1$ determined by the underlying process. Let $\{x_1, x_2, \cdots x_n\} \in \mathbb{R}^d$ be a finite sample drawn from a VAR(q) process. Then, simultaneously for every $f \in \mathcal{F}$, under the square loss truncated at M, with probability at least $1 - \delta$,*

$$\mathcal{G}_{\omega,i} \leq \zeta\hat{\mathcal{S}}_\omega + \zeta\widehat{\mathfrak{R}}_\mu(\mathcal{F}) + 3\zeta M\sqrt{\frac{\log\frac{4}{\delta'}}{2\mu}} \tag{8}$$

where $\zeta = 2\kappa(\Sigma^\nu)$, $\delta' = \delta - 2(\mu - 1)\rho^m$, and $\widehat{\mathfrak{R}}_\mu(\mathcal{F})$ denotes the empirical Rademacher complexity of $\mathcal{F}$.

Our causal generalization bound in Theorem 1 suggests that, given sufficiently many samples, the true causal error can be guaranteed to be close to empirical statistical error if our VAR models come from a class with a small Rademacher

complexity, particularly when the process is associated with a small stability parameter.

We now focus on providing a detailed interpretation of our results. First, we take a minor detour to present a technical result (Lemma 2) which is useful both in deriving some of our main results as well as in interpreting them.

Lemma 2 (Expressing powers of a companion matrix using symmetric polynomials). *For a companion matrix A with distinct eigenvalues, for any $k \in [p]$, the $(1,k)$th element of A^j, can be expressed using Schur polynomials of the eigenvalues $\lambda = \{\lambda_1, \lambda_2, \cdots \lambda_p\}$ of A, that is, $A^j_{1,k} = S_{j,k}(\lambda)$, where $S_{j,k}(\lambda)$ refers to the Schur polynomial indexed by $K = \{j, 1, \cdots k-1 \text{ times} \cdots, 1, 0, \cdots, 0\}$.*

Lemma 2 shows that the coefficients of the powers of a companion matrix can be fully characterized using symmetric Schur polynomials of its eigenvalues. A good overview of these polynomials can be found in Chaugule et al. (2019). An advantage of expressing the coefficients using symmetric Schur polynomials is that these polynomials have been a subject of extensive research in combinatorics and an equivalence between several alternate definitions has been established. To name a few, Cauchy's bialternant expression, (Cauchy 1815; Jacobi 1841), the combinatorial formula (Macdonald 1998) or Jacobi–Trudi identity (Jacobi 1841) are all equivalent ways to define Schur polynomials. It is therefore possible and often beneficial to choose the definition that yields the most useful notion for the context. We utilize this connection to interpret our results. First, for easier interpretation, we simplify Lemma 1 to the following result for scalar AR models.

Corollary 1 (Difference in Causal and Statistical errors (AR)). *Let $\{x_t\}_{t\in\mathbb{Z}}$ follow an AR(q) process. Then, for any AR(p) model with parameters $\hat{A}$,*

$$|\mathcal{G}_{do_\omega} - \mathcal{S}_\omega| = 2\left|(A^\omega_{11} - \hat{A}^\omega_{11})\sum_{k=2}^{\nu}(A^\omega_{1k} - \hat{A}^\omega_{1k})\gamma_{k-1}\right|, \quad (9)$$

where, for any $k \in \mathbb{N}$, γ_k denotes the autocovariance of $\{x_t\}_{t\in\mathbb{Z}}$ with lag k. A and $\widehat{A}$ are the corresponding companion matrices of the model and estimated parameters as defined in Lemma 1.

Lemma 1 identifies factors that control causal generalizability. We now describe them.

Correlations control causal generalizability. Recall our motivating example of the two highly correlated time-series where the casual and statistical errors diverge. Intuitively, one would therefore expect that large correlations among time series potentially induce large differences between observational and interventional distributions. The quantitative dependence of causal generalizability on the correlation structure of the process is, however, less obvious. Lemma 1 confirms the intuition and shows that correlations between the intervened time-series $x_{t-\omega,i}$ across both the components and time instances in $\mathbf{x}_{t-\omega}$ control generalizability from observational to the interventional distributions.

High-dimensional and higher-order processes can hurt generalization. For high-dimensional processes it is not unlikely to have strong correlations across components, which may obscure causal relations in the same way as strong correlations across time does for univariate processes. Lemma 1 also supports this intuition and shows that strong correlations across components as well as time instances play a role. With increasing order or dimension of the processes, larger orders of covariances across time and dimensions could entail poor causal generalizability.

Dependence on ω. The dependence of the error on ω arises through the elements of the matrix power A^k. A simple computation shows that, even for an AR(2) model, the dependence of these coefficients on the model parameters is asymmetric and highly intricate. However, using the Cauchy's bialternant formulation of Schur polynomials, we have that for any AR(p) model, the coefficients A^ω_{1k} can be expressed as $A^\omega_{1k} = (-1)^{k+1}\dfrac{\sum_{i=1}^{p}\lambda_i^{p+\omega-1}e_k(\lambda_i)}{\det\left|\{\lambda_k^{p-k'}\}_{k,k'\in[p]}\right|}$, where $e_k(\lambda_i)$ refers to the elementary symmetric polynomial of order k and with variables $\{\lambda_1, \cdots \lambda_{i-1}, \lambda_{i+1}, \cdots, \lambda_p\}$. While this is not the most interpretable definition per se, the dependence of the coefficients on ω is easily understood and it is easy to verify that if the underlying model as well as the estimated model are both stable ($|\lambda| < 1$), the coefficients and hence the difference in errors exponentially decays with interventions arbitrarily in the past of the target variable and if either of the process is not stable ($|\lambda| > 1$), the difference can indeed diverge.

Proposition 1 allows us to obtain a high-level perspective on causal generalizability. It states that the condition number of the autocovariance matrix controls causal generalizability. Both the maximum and the minimum eigenvalue of the autocovariance matrix (and hence the condition number) can be used as a measure of stability and hence determine the strength of correlation of the underlying process (Basu et al. 2015; Melnyk et al. 2016). As the process gets closer to the boundary of stability domain, the autocovariance matrix gets singular and hence the condition number of the autocovariance matrix can get arbitrarily large. Proposition 1, therefore, can be interpreted as if the underlying process gets closer to the boundary of the stability domain the causal and statistical errors can diverge.

For intuition, let us revisit our motivating example from the introduction with strongly correlated observations in an $AR(p)$ process. Let, without loss of generality $p = q$. Introducing the vectors $a := (a_1, a_2, \ldots, a_p)$ and $\hat{a} := (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p)$ and the covariance matrix $\Sigma_p = \Sigma_{\max\{p,q\}}$. Then the quotient between causal and statistical error for predicting one time step ahead i.e. ($\omega = 1$) reads:

$$\frac{\mathcal{G}_{do_\omega}}{\mathcal{S}_\omega} = \frac{(\hat{a}-a)^T(\hat{a}-a)+\sigma_\epsilon^2}{(\hat{a}-a)^T\Sigma_p(\hat{a}-a)+\sigma_\epsilon^2}, \tag{10}$$

Where we have assumed X_t to have unit variance without loss of generality. The quotient is maximized if $(\hat{a}-a)$ is a multiple of the eigenvector to the smallest eigenvalue of Σ_p. This aligns with the intuition that causal loss diverges when the auto-covariance matrix gets singular. Moreover, we see that the vector $(\hat{a}-a)$ can be large with little observable effect when it mainly consists of eigenvectors with small eigenvalues of Σ_p. In the extreme case, if the minimum eigenvalue of the autocovariance matrix is 0, it is possible to arbitrarily deviate from the true model parameters along the direction of the corresponding eigenvector which can significantly affect the causal error without affecting the statistical error at all. For an $AR(2)$ process, for instance, we obtain $\Sigma_p = \begin{pmatrix} 1 & a_1/(1-a_2) \\ a_1/(1-a_2) & 1 \end{pmatrix}$, which becomes singular for $a_1 = \pm(1-a_2)$ which indeed is the boundary of the stability domain (see for example, Lütkepohl (2009)). This is the limit in Section 1 where $X_t = \pm X_{t-1}$. The eigenvector for eigenvalue 0 reads $(1, \mp 1)$. Accordingly, the quotient (10) diverges when $\hat{a}$ differs from a by $(1, \mp 1)$.

This further highlights that even for simple classes of forecasting models and with simplifying assumptions such as causal sufficiency, causal risks may even diverge from statistical risks. To show this formally, by means of Lemma 2, we can derive an explicit upper bound on the condition number of the autocovariance matrix $\kappa(\Sigma_{\max\{p,q\}})$ for AR(p) models and arrive at Corollary 2.

Corollary 2 (Stability Controls Causal Generalization (AR)). *Consider an AR(q) process, such that eigenvalues of its companion matrix satisfy $|\lambda| < \delta < 1$. For any AR(q) model f,*

$$|\mathcal{G}_{\omega,i} - \mathcal{S}_\omega| \leq K_p \mathcal{S}_\omega(f)\nu(1+\delta)^{2\nu}/(1-\delta^2), \tag{11}$$

where K_p is some finite constant that depends on the order p of the underlying process.

The bound in Corollary 2 is elegant due to its simplicity and generality. However, the cost of generality of the bound that relies only on the stability parameter is clearly that it cannot explain the variations in behavior exhibited by individual processes with the same stability parameter. For instance, consider an AR(2) model with parameters a_1 and a_2 with $a_2 \approx 0$ so that it is essentially an AR(1) model. Then, it is easy to verify that $\lambda_2 \approx 0$. The combinatorial definition of the Schur polynomials (Macdonald 1998) allows us to express the coefficients as follows: $A_{11}^\omega = \sum_{i=0}^{\omega} \lambda_1^{\omega-i}\lambda_2^i, \quad A_{12}^\omega = \sum_{i=1}^{\omega-1} \lambda_1^{\omega-i}\lambda_2^i$. Combining this with Corollary 1, it is easy to see that if the estimated model is also close to AR(1), then the coefficients A_{12}^ω and $\widehat{A}_{12}^\omega$ and hence the difference in statistical and causal errors is close to 0. The bound in (11) which relies on the stability parameter does not capture this. For tighter bounds that utilize additional information about the spectrum of the companion matrix, we can exploit the connection to Schur polynomials to arrive at the following bound.

$$|\mathcal{G}_{\omega,i} - \mathcal{S}_\omega| \leq K_{p,q} \max\left\{\delta, \widehat{\delta}\right\}^\omega \sum_{k=2}^{\nu} (S_{\omega k}^{\lambda} - S_{\omega k}^{\hat{\lambda}})\gamma_{k-1},$$

where $K_{p,q}$ is a constant that depends on p, q, δ and $\widehat{\delta}$ are the stability parameters of the true and estimated processes respectively and λ and $\widehat{\lambda}$ denote the set of eigenvalues of A and $\widehat{A}$ respectively.

4 SIMULATIONS

To verify the practical behavior of causal and statistical risks, we provide some simple simulations to study the errors of different estimators under AR processes. For each presented plot, we draw parameters for 10,000 stationary $AR(p)$ processes using rejection sampling. We draw the coefficients of each process independently and uniformly from $[-2, 2]$ and reject sets of parameters that yield a non-stationary process. For each process, we draw a training sample with 100 timesteps and a test sample with 1000 timesteps. For all figures in the main paper we set $\omega = 1$. To estimate the coefficients we use Ordinary Least Squares (OLS). In Appendix E we provide additional plots with hidden confounder, as well as varying order, sample size, ω and other estimators: Ridge, Lasso, and Elastic Net regressors. OLS minimizes the empirical statistical error, that is, $\sum_{y_i,\hat{y}_i}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ denotes the model prediction with estimated parameters $\hat{a}$.

In line with our theoretical results, we find that even for simple scalar AR processes of small orders, the causal error of the estimators is often several times larger than the statistical error (see Figure 2). In Figure 3 we sorted the randomly drawn datasets by their autocorrelation (measured by the condition number κ of the autocorrelation matrix) and split the sorted list into buckets of 500 dataset. For each we calculated the maximum, mean and 90% quantile of the difference in causal and statistical error for the OLS and Ridge estimators. The plots corresponding to the other estimators are provided in Appendix E We can see that upto constant factors, our theoretical finite sample causal generalization bound matches the difference in causal and statistical risks observed empirically.

5 EXPERIMENTS ON REAL DATA

Data. We conduct experiments on the m4 hourly dataset (Makridakis et al. 2018) that includes timeseries from a diverse set of sources. The datasets has a hourly frequency

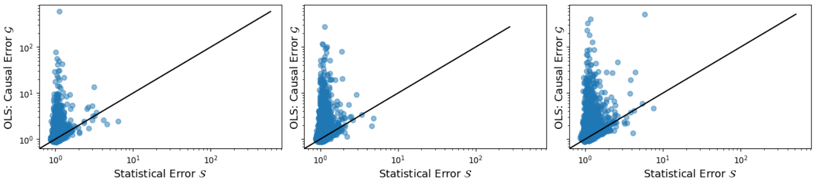


Figure 2: The causal error $\mathcal{G}$ versus the statistical error $\mathcal{S}$ for AR(p) processes with $p = 3, 5, 7$.

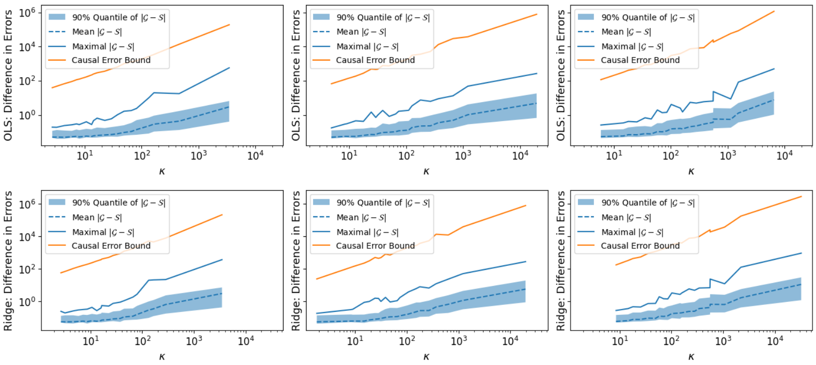


Figure 3: The maximal difference between statistical error $\mathcal{S}$ and causal error $\mathcal{G}$ as well as an estimate for the generalization bound in Theorem 1 for increasing condition number κ for process orders $p = 3, 5, 7$ (from left to right). The maximum is taken over 500 datasets with the closest κ. Our theoretical bounds (orange) closely match the empirical evaluations up to constant factors (blue).

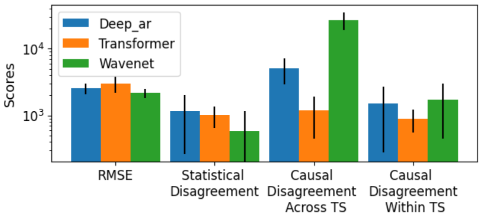


Figure 4: Results of the evaluation of three different deep neural network architectures on the m4hourly dataset. The "RMSE" is computed comparing prediction on the observational data against the ground truth. The disagreement from Def. 5.1 compares the root-mean-square deviation between the predictions of two models of the same architecture on the observational data ("Statistical Disagreement") and interventional distributions ("Causal Disagreement Across TS" sampling interventions from all of time-series and "Causal Disagreement Within TS" sampling interventions from prior points within the time series). The results are averaged over 5 runs of training and evaluation and include standard deviation in black.

and a prediction length of 48. To create an interventional distribution without a generative model, for each time series we replace the last time step prior to the evaluation window by sampling at random either from all time-series at that time step (referred to as *across-ts*) or from previous values of the same time series (referred to as *within-ts*). Appendix F records results for additional datasets.

Models. We include three popular deep neural network architectures in our evaluation. DeepAR consists of an RNN that takes the previous time steps as inputs and predicts the parameters of an auto-regressive model (Gasthaus et al. 2019). Wavenet is a hierarchical CNN developed for speech-to-text (Oord et al. 2016). Transformer is an attention-based deep neural network widely applied to NLP tasks including translation (Vaswani et al. 2017). For all these models we use AutoGluonTS's default hyperparameters.

Metrics. For the observational distribution, we compute the root-mean-square error (RMSE) comparing average prediction for each time point with the ground truth in the evaluation set. For the interventional distribution we are lacking ground truth. Therefore, we train two separate models and compute their disagreement.

Definition 5.1. The disagreement is the average root-mean-square deviation of the mean forecasts of two models. The average is taken over a set of time-series. If the time-series come from the original dataset, we call it the statistical disagreement. If they come from one of the interventional datasets, we call it causal disagreement and specify the type of intervention as across time-series or within time-series.

This disagreement is a measure of uncertainty introduced by the randomness in the training and evaluation procedure. Here, however, we use it to approach the causal risk, that we cannot compute directly. If the disagreement is high on the interventional distribution at least one of the models must have a high causal risk. For comparison, we also included this disagreement measure for examples from the observational distribution. Finally, to explore the relationship between causal forecasting error and uncertainty, we also compute the width of the 80% prediction interval for both the observational and interventional distribution.

Definition 5.2. The 80% prediction width of a forecasting model is the absolute distance between the 0.9 quantile and 0.1 quantile of the forecast distribution. It is averaged over a set of time-series that can come from the observational or the interventional distritibutions.

The experiments were conducted using GluonTS (Alexandrov et al. 2019) with default hyperparameters on instances with 4 virtual CPUs and a 2.9 GHz processor.

Limitations. The dataset and models have clear shortcomings. Likely, the dataset is not causally sufficient. Also, we did not tune the models. Moreover, we are lacking samples from the marginal distribution for the interventions and groundtruth on what happens under these interventions. Nevertheless, we hope to get a sense for how popular deep learning networks can behave on real data for relevant prediction tasks under interventions.

Results. Figure 4 shows the results of the metrics when we evaluate the models on the datsets for both observation and interventional distributions. We see that the causal disagreement between two models of the same architecture and hyper-parameters can be much higher than their disagreement on the observational distribution. While there are only smaller differences in the statistical risk between the model architectures, their causal disagreement differs more. Overall, the the causal disagreement can be high, which implies high causal risk, but it varies across datasets and model architectures. Wavenet's disagreement is an order of magnitude larger when sampling interventions from other time-series. For transformer models their interventional disagreement is close to the observational one.

Uncertainty.

When we compare the width of the 80% interval of predictions in Table 2 we see that this uncertainty measure is higher for the interventional distribution compared to the observational one. Moreover, directionally it relates to the causal disagreement across models. Unlike the disagreement that requires a second model to be trained, this uncertainty measure is readily available from the predicted forecasts.

Model	observ.	across-ts interv.	within-ts interv.
DeepAR	940.0 ± 126.2	1329.2 ± 187.5	953.1 ± 124.2
wavenet	1253.9 ± 96.6	3444.7 ± 649.4	1612.7 ± 257.7
transformer	1259.3 ± 139.3	1355.1 ± 129.6	1255.7 ± 139.3

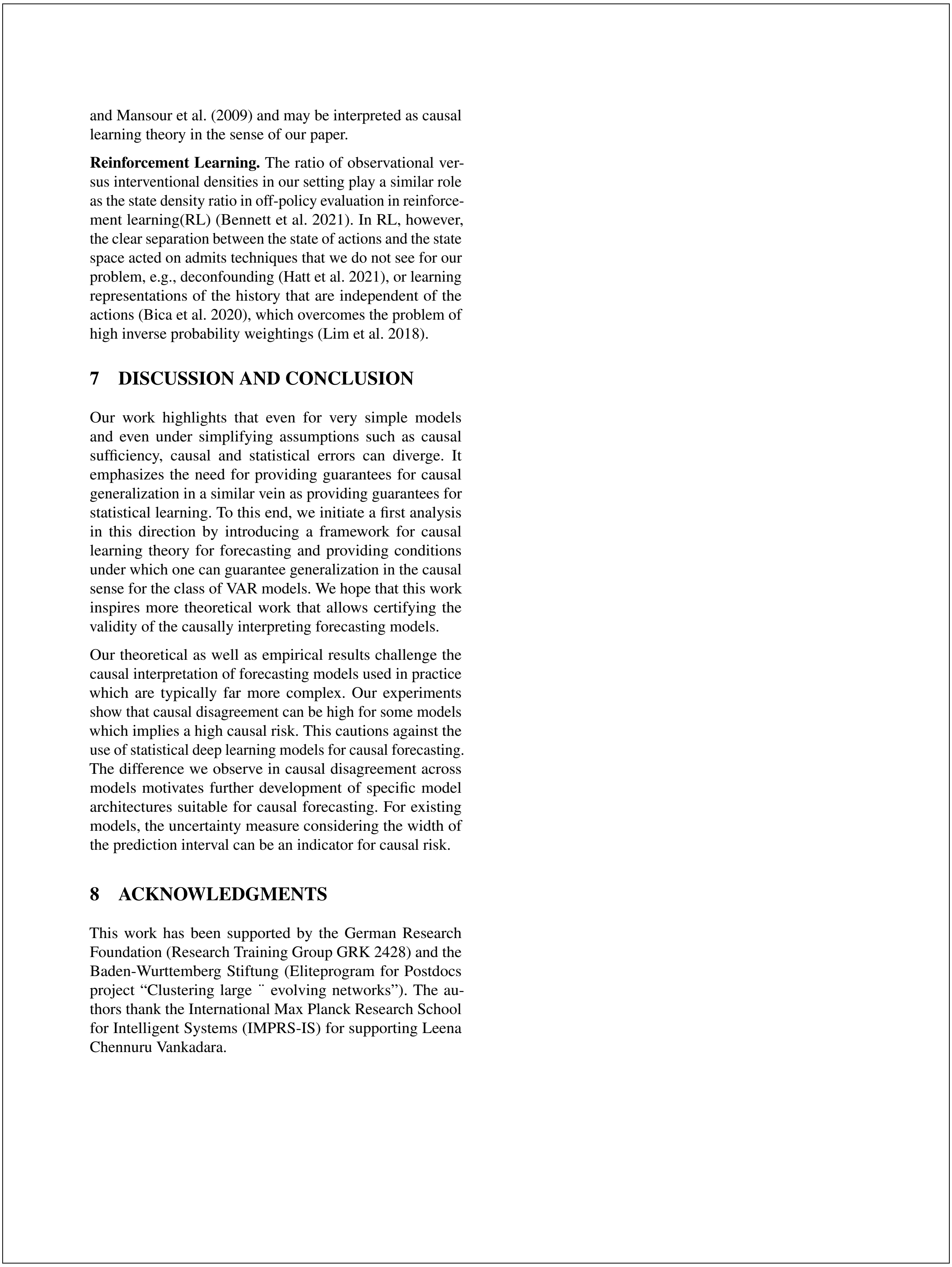
Table 1: 80% prediction width, see Def. 5.2, for observational and interventional forecasts. Averaged over 5 runs with std.

6 RELATED WORK

Our work intersects with domain adaption, RL, and treatment effect estimation, reviewed separately below.

Domain Adaptation. The literature that is perhaps most relevant to our context is that of learning theory for domain adaptation, in particular, for covariate shift. Theoretical analysis of domain adaptation when labelled samples from the source distribution and unlabelled samples from the target distribution are generated i.i.d was initiated by Ben-David et al. (2007), who provided VC bounds for binary classification under covariate shifts based on a *discrepancy measure* $d_{\mathcal{F}}$ between source and target distributions that depends on the hypothesis class $\mathcal{F}$ and is estimable from finite samples. Mansour et al. (2009) extended the work to the context of regression in the i.i.d setting by adapting the discrepancy measure for more general loss functions and by providing tighter, data-dependent Rademacher bounds. Despite the i.i.d assumption that is necessary to derive their finite-sample bounds, the results in Mansour et al. (2009) are perhaps the most relevant to our setting. We can utilize one of the main results from Mansour et al. (2009, Theorem 8) which does not rely on the i.i.d assumption to arrive at the following population-level bound for our setting: $|\mathcal{G}_{\omega,i}(f, f^*) - \mathcal{S}_{\omega}(f, f^*)| \leq \sup_{f,f' \in \mathcal{F}} |\mathcal{G}_{\omega,i}(f, f') - \mathcal{S}_{\omega}(f, f')|$. These bounds are clearly non-informative in our context since they do not incorporate structural knowledge of the class of interventional distributions under a VAR model.

Estimation of Treatment Effects. A related problem is that of estimating treatment effects in the potential outcomes framework (Hill et al. 2006; Shi et al. 2019), where the goal is to estimate the effects of binary-valued treatments from observational data under a multivariate confounding model. Our setting is more general in that variables in the multivariate process can take a continuum of interventions and play a multiplicity of roles — each variable plays the role of treatment, confounder, and the target variable. Of particular relevance is the work of Shalit et al. (2017) and Johansson et al. (2020), who prove generalization error bounds on estimating individual-level treatment effects in terms of standard generalization error and a distance measure between the treated and control distributions. This result is similar to domain adaptation bounds in Ben-David et al. (2007)



and Mansour et al. (2009) and may be interpreted as causal learning theory in the sense of our paper.

Reinforcement Learning. The ratio of observational versus interventional densities in our setting play a similar role as the state density ratio in off-policy evaluation in reinforcement learning(RL) (Bennett et al. 2021). In RL, however, the clear separation between the state of actions and the state space acted on admits techniques that we do not see for our problem, e.g., deconfounding (Hatt et al. 2021), or learning representations of the history that are independent of the actions (Bica et al. 2020), which overcomes the problem of high inverse probability weightings (Lim et al. 2018).

7 DISCUSSION AND CONCLUSION

Our work highlights that even for very simple models and even under simplifying assumptions such as causal sufficiency, causal and statistical errors can diverge. It emphasizes the need for providing guarantees for causal generalization in a similar vein as providing guarantees for statistical learning. To this end, we initiate a first analysis in this direction by introducing a framework for causal learning theory for forecasting and providing conditions under which one can guarantee generalization in the causal sense for the class of VAR models. We hope that this work inspires more theoretical work that allows certifying the validity of the causally interpreting forecasting models.

Our theoretical as well as empirical results challenge the causal interpretation of forecasting models used in practice which are typically far more complex. Our experiments show that causal disagreement can be high for some models which implies a high causal risk. This cautions against the use of statistical deep learning models for causal forecasting. The difference we observe in causal disagreement across models motivates further development of specific model architectures suitable for causal forecasting. For existing models, the uncertainty measure considering the width of the prediction interval can be an indicator for causal risk.

8 ACKNOWLEDGMENTS

This work has been supported by the German Research Foundation (Research Training Group GRK 2428) and the Baden-Wurttemberg Stiftung (Eliteprogram for Postdocs project “Clustering large ¨ evolving networks”). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Leena Chennuru Vankadara.

REFERENCES

Alexandrov, Alexander, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang (2019). *GluonTS: Probabilistic Time Series Models in Python*. arXiv: `1906.05264 [cs.LG]`.

Basu, Sumanta and George Michailidis (2015). “Regularized estimation in sparse high-dimensional time series models”. In: *The Annals of Statistics* 43.4, pp. 1535–1567.

Ben-David, Shai, John Blitzer, Koby Crammer, Fernando Pereira, et al. (2007). “Analysis of representations for domain adaptation”. In: *Advances in neural information processing systems* 19, p. 137.

Bennett, Andrew, Nathan Kallus, Lihong Li, and Ali Mousavi (2021). “Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1999–2007.

Bica, Ioana, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar (2020). “Estimating counterfactual treatment outcomes over time through adversarially balanced representations”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: `https://openreview.net/forum?id=BJg866NFvB`.

Brand, Louis (1964). “The companion matrix and its properties”. In: *The American Mathematical Monthly* 71.6, pp. 629–634.

Brockwell, Peter J, Richard A Davis, and Stephen E Fienberg (1991). *Time series: theory and methods: theory and methods*. Springer Science & Business Media.

Cauchy, Augustin Louis (1815). “M ’e memory on functions which can obtain only two equal values é and of opposite signs as a result of the transpositions op ’e r é es between the variables that they contain”. In: *Journal de l’Ecole polytechnique* 10.17, pp. 29–112.

Chaugule, Prasad, Mrinal Kumar, Nutan Limaye, Chandra Kanta Mohapatra, Adrian She, and Srikanth Srinivasan (2019). “Schur Polynomials do not have small formulas if the Determinant doesn’t!” In: *arXiv preprint arXiv:1911.12520*.

Davison, Edward (1976). “The robust control of a servomechanism problem for linear time-invariant multivariable systems”. In: *IEEE transactions on Automatic Control* 21.1, pp. 25–34.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. URL: `http://archive.ics.uci.edu/ml`.

Faloutsos, Christos, Jan Gasthaus, Tim Januschowski, and Yuyang Wang (2018). “Forecasting Big Time Series: Old and New”. In: 11.12. ISSN: 2150-8097.

Fisk, Steve (2005). “A very short proof of Cauchy’s interlace theorem for eigenvalues of Hermitian matrices”. In: *arXiv preprint math/0502408*.

Gasthaus, Jan, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski (Apr. 2019). “Probabilistic Forecasting with Spline Quantile Function RNNs”. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1901–1910.

Grabowski, Wojciech and Aleksander Welfe (2020). “The Tobit cointegrated vector autoregressive model: An application to the currency market”. In: *Economic Modelling* 89, pp. 88–100.

Hatt, Tobias and Stefan Feuerriegel (2021). *Sequential Deconfounding for Causal Inference with Unobserved Confounders*. arXiv: `2104.09323 [stat.ME]`.

Hill, Jennifer and Jerome P Reiter (2006). “Interval estimation for treatment effects using propensity score matching”. In: *Statistics in medicine* 25.13, pp. 2230–2256.

Horn, Roger A and Charles R Johnson (2012). *Matrix analysis*. Cambridge university press.

Hyvärinen, Aapo, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer (2010). “Estimation of a structural vector autoregression model using non-gaussianity.” In: *Journal of Machine Learning Research* 11.5.

Jacobi, Carl Gustav Jakob (1841). “De functionibus alternantibus earumque divisione per productum e differentiis elementorum conflatum.” In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 1841.22, pp. 360–371.

Januschowski, Tim, Jan Gasthaus, Yuyang Wang, David Salinas, Valentin Flunkert, Michael Bohlke-Schneider, and Laurent Callot (2020). “Criteria for classifying forecasting methods”. In: *International Journal of Forecasting* 36.1. M4 Competition, pp. 167–177. ISSN: 0169-2070.

Janzing, D., L. Minorics, and P. Bloebaum (Aug. 2020). “Feature relevance quantification in explainable AI: A causal problem”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. Online: PMLR, pp. 2907–2916.

Johansson, Fredrik D, Uri Shalit, Nathan Kallus, and David Sontag (2020). “Generalization bounds and representation learning for estimation of potential outcomes and causal effects”. In: *arXiv preprint arXiv:2001.07426*.

Lim, Bryan, Ahmed Alaa, and Mihaela van der Schaar (2018). “Forecasting Treatment Responses over Time Using Recurrent Marginal Structural Networks”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., pp. 7494–7504.

Lundberg, S. and S. Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 4765–4774.

Lütkepohl, Helmut (2009). "Econometric analysis with vector autoregressive models". In: *Handbook of computational econometrics*, pp. 281–319.

– (2013). "Vector autoregressive models". In: *Handbook of Research Methods and Applications in Empirical Macroeconomics*. Edward Elgar Publishing.

Macdonald, Ian Grant (1998). *Symmetric functions and Hall polynomials*. Oxford university press.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilis Assimakopoulos (June 2018). "The M4 Competition: Results, findings, conclusion and way forward". In: *International Journal of Forecasting* 34. DOI: 10.1016/j.ijforecast.2018.06.001.

Mansour, Yishay, Mehryar Mohri, and Afshin Rostamizadeh (2009). "Domain adaptation: Learning bounds and algorithms". In: *arXiv preprint arXiv:0902.3430*.

McDonald, Daniel J, Cosma Rohilla Shalizi, and Mark Schervish (2017). "Nonparametric risk bounds for time-series forecasting". In: *The Journal of Machine Learning Research* 18.1, pp. 1044–1083.

Meir, Ron (2000). "Nonparametric time series prediction through adaptive model selection". In: *Machine learning* 39.1, pp. 5–34.

Melnyk, Igor and Arindam Banerjee (2016). "Estimating structured vector autoregressive models". In: *International Conference on Machine Learning*. PMLR, pp. 830–839.

El-Mikkawy, Moawwad EA (2003). "Explicit inverse of a generalized Vandermonde matrix". In: *Applied mathematics and computation* 146.2-3, pp. 643–651.

Mohri, Mehryar and Afshin Rostamizadeh (2009). "Rademacher Complexity Bounds for Non-I.I.D. Processes". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc.

Mokkadem, Abdelkader (1988). "Mixing properties of ARMA processes". In: *Stochastic processes and their applications* 29.2, pp. 309–315.

Molnar, C. (2019). *Interpretable Machine Learning*. Molnar, C. URL: https://christophm.github.io/interpretable-ml-book/.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). *WaveNet: A Generative Model for Raw Audio*. arXiv: 1609.03499 [cs.SD].

Pearl, Judea (2009). *Causality*. Cambridge university press.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference*. The MIT Press.

Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski (2020). "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". In: *International Journal of Forecasting* 36.3, pp. 1181–1191. ISSN: 0169-2070.

Shalit, Uri, Fredrik D Johansson, and David Sontag (2017). "Estimating individual treatment effect: generalization bounds and algorithms". In: *International Conference on Machine Learning*. PMLR, pp. 3076–3085.

Shi, Claudia, David M Blei, and Victor Veitch (2019). "Adapting neural networks for the estimation of treatment effects". In: *arXiv preprint arXiv:1906.02120*.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction, and Search*. New York, NY: Springer-Verlag.

Sugiyama, Masashi and Motoaki Kawanabe (2012). *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press. ISBN: 0262017091.

Valdés-Sosa, Pedro A, Jose M Sánchez-Bornot, Agustín Lage-Castellanos, Mayrim Vega-Hernández, Jorge Bosch-Bayard, Lester Melie-García, and Erick Canales-Rodríguez (2005). "Estimating brain functional connectivity with sparse multivariate autoregression". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1457, pp. 969–981.

Varga, Richard S (2010). *Geršgorin and his circles*. Vol. 36. Springer Science & Business Media.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.

Wang, Jiaxuan, Jenna Wiens, and Scott Lundberg (2020). *Shapley Flow: A Graph-based Approach to Interpreting Model Predictions*. arxiv:2010.14592.

Yu, Bin (1994). "Rates of convergence for empirical processes of stationary mixing sequences". In: *The Annals of Probability*, pp. 94–116.

Zivot, Eric and Jiahui Wang (2006). "Vector autoregressive models for multivariate time series". In: *Modeling Financial Time Series with S-Plus®*, pp. 385–429.

A BACKGROUND

Notation. We recall the notation and some key definitions here for the reader's convenience. For any stochastic process $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$, we use $\mathbf{x}^n_{t-\omega} = \{x_{t-\omega-n+1}, \cdots, x_{t-\omega-1}, x_{t-\omega}\}$ to denote the *set* of $x_{t-\omega}$ and the $n-1$ variables in the past of $x_{t-\omega}$. We distinguish this from y^n_t which denotes the *vector* $\left(x_t, x_{t-1}, \cdots, x_{t-n+1}\right)^T \in \mathbb{R}^{nd}$. When it is clear from context, to reduce cumbersome notation, we simply use y_t. For any random variable x, $\mathbb{E}[x]$ denotes its expectation. For any matrix A, we use $A_{i:}$ and $A_{:j}$ to denote the ith row and jth column of A respectively. We use A^j_{1k} to denote the $(1,k)$th element of A^j. For any vector x_t at time t, we use $x_{t,i}$ to denote the ith element of x_t. We use $\lambda_{\max}(A), \lambda_{\min}(A), \kappa(A)$ to denote the maximum and minimum eigenvalues and the condition number of A respectively, where $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$. $\mathbb{I}_p$ denotes the identity matrix of size p, $\mathbb{N}, \mathbb{Z}$ denote the set of natural numbers and integers respectively and $[n]$ denotes the set $\{1, 2, \cdots n\}$.

Definition A.1 (Vector Autoregressive Model). A vector autoregressive model (VAR(p)) of dimension d and order p is defined as

$$x_t = A_1 x_{t-1} + A_2 x_{t-2} + \cdots A_P x_{t-p} + \epsilon_t, \tag{12}$$

where $x_t \in \mathbb{R}^d$ is a vector-valued time-series, for all $i \in [p]$, $A_i \in \mathbb{R}^{d\times d}$ are the coefficients of the VAR model, and $\epsilon_t \in \mathbb{R}^d$ denotes the noise vector such that $\mathbb{E}[\epsilon_t] = 0$ and $\mathbb{E}[\epsilon_t \epsilon^T_{t+h}] = \Sigma_\epsilon$ if $h = 0$ and 0 otherwise. For some $\sigma^2_\epsilon > 0$, we simply set $\Sigma_\epsilon = \sigma^2_\epsilon \mathbb{I}$ for enhanced readability. Our results can be easily generalized to arbitrary covariance matrices by means of the spectral properties $(\lambda_{\min}, \lambda_{\max})$ of Σ_ϵ.

Definition A.2 (Weak Stationarity). A stochastic process $\{x_t\}_{t\in\mathbb{Z}}$ is weakly stationary if the mean and the covariance of the process does not change over time, that is, for all $t, \tau \in \mathbb{Z}$

$$\mathbb{E}[x_t] = \mathbb{E}[x_{t+\tau}], \quad \mathbb{C}_x(t, t+\tau) = \mathbb{C}_x(0, \tau), \tag{13}$$

where $\mathbb{C}_x(t, t+\tau) = \mathbb{E}[(x_t - \mathbb{E}[x_t])(x_{t+\tau} - \mathbb{E}[x_{t+\tau}])]$ denotes the autocovariance function.

The autocovariance matrix of $\{x_t\}_{t\in\mathbb{Z}}$ plays a central role in our results and analysis. For any $n \in \mathbb{N}$, we use Σ_n to denote the autocovariance matrix of size n defined as $\mathbb{E}[(y^n_t - \mathbb{E}[y^n_t])(y^n_t - \mathbb{E}[y^n_t])^T]$.

It is often quite convenient to rewrite a VAR model of order p in Equation (12) as a VAR(1) model, $y_t = Ay_{t-1} + e_t$, where $y_t \in \mathbb{R}^{dp}, e_t \in \mathbb{R}^{dp}$ are defined as $y_t = \left(x_t, x_{t-i}, \cdots, x_{t-p+1}\right)^T$, $e_t = \left(\epsilon_t, 0, \cdots, 0\right)^T$, and $A \in \mathbb{R}^{dp\times dp}$ is a *(multi) companion matrix* defined as:

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{pmatrix}. \tag{14}$$

The eigenvalues of the multi-companion matrix A fully characterize the stability and stationarity of the VAR process. For a VAR(p) process to be weakly stationary, the eigenvalues of A, which satisfy

$$\det|\mathbb{I}_d\lambda^p - A_1\lambda^{p-1} - A_2\lambda^{p-2} - \cdots - A_p| = 0, \tag{15}$$

are constrained to not lie on the unit circle. If the magnitude of the eigenvalues are $|\lambda_i| < 1$ for all $i \in [dp]$, then the underlying process is stable, that is, its values do not diverge (Lütkepohl 2013).

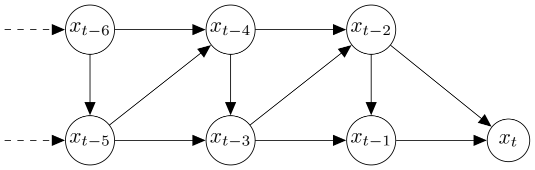


Figure 5: Causal DAG of an AR(2) model

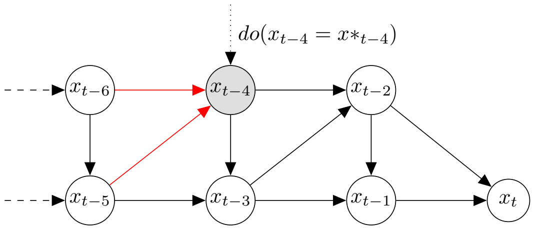


Figure 6: Graphical representation of the effect of an intervention $do(x_{t-4} = x*_{t-4})$ on an AR(2) model. Incoming edges into x_{t-4} are removed in the new DAG which are in red.

Definition A.3 (Empirical Rademacher Complexity). Given a finite sample $X = \{x_1, x_2, \cdots, x_n\} \in \mathbb{R}^d$, the empirical Rademacher complexity of a hypothesis class $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ is defined as:

$$\hat{\mathfrak{R}(\mathcal{F})} = \frac{2}{n}\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} \sigma_i f(x_i)| \right],$$

where $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_n)$ and for all $i \in [n]$, σ_i are independent random variables drawn from the Rademacher distribution, that is, a uniform distribution over $\{-1, +1\}$.

B PROOFS OF MAIN RESULTS

Lemma 3 (Expressing powers of a companion matrix using symmetric polynomials). *For a companion matrix A with distinct eigenvalues and for any $k \in [p]$, the $(1, k)$th element of A^ω, can be expressed as a Schur polynomial of the eigenvalues $\lambda = \{\lambda_1, \lambda_2, \cdots \lambda_p\}$ of A. in particular, $|A^\omega_{1,k}| = S_{\mu_{\omega,k},\lambda}$ where $S_{\mu_{\omega,k},\lambda}$ refers to the Schur polynomial over λ indexed by $\{\omega, 1, \cdots k-1 \text{ times} \cdots, 1, 0, \cdots, 0\}$.*

Proof. For convenience, we use the notation λ and λ/λ_i to denote the sets $\{\lambda_1, \lambda_2, \cdots, \lambda_p\}$ and $\{\lambda_1, \lambda_2, \cdots, \lambda_{i-1}, \lambda_{i+1}, \cdots, \lambda_p\}$ respectively.

Assuming that the eigenvalues $\lambda = \{\lambda_i\}_{i=1}^p$ of a companion matrix A are distinct, it can be diagonalized as $A = V\Lambda V^{-1}$, where $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_p)$ is the diagonal matrix of eigenvalues of A and V is a vandermonde matrix (Brand 1964) given by

$$V_\lambda = \begin{pmatrix} \lambda_1^{p-1} & \lambda_2^{p-1} & \cdots & \lambda_p^{p-1} \\ \lambda_1^{p-2} & \lambda_2^{p-2} & \cdots & \lambda_p^{p-2} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1 & \lambda_2 & \cdots & \lambda_p \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{16}$$

For any $i \in [p]$, let $e_k(\lambda/\lambda_i)$ denote the elementary symmetric polynomial of order k with variables in λ/λ_i and let

$$\alpha_i = \frac{1}{\prod\limits_{j \neq i} (\lambda_i - \lambda_j)}. \tag{17}$$

The inverse of the Vandermonde matrix V can then be explicitly computed (El-Mikkawy 2003) to obtain

$$V^{-1} = \begin{pmatrix} \alpha_1 & -\alpha_1 e_1(\lambda/\lambda_1) & \cdots & (-1)^{p-1}\alpha_1 e_{p-1}(\lambda/\lambda_1) \\ \alpha_2 & -\alpha_2 e_1(\lambda/\lambda_2) & \cdots & (-1)^{p-1}\alpha_2 e_{p-1}(\lambda/\lambda_2) \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_p & -\alpha_p e_1(\lambda/\lambda_p) & \cdots & (-1)^{p-1}\alpha_p e_{p-1}(\lambda/\lambda_p) \end{pmatrix}, \tag{18}$$

Using the diagonalization of A, we can compute its power A^ω as

$$A^\omega = V\Lambda^\omega V^{-1} \tag{19}$$

and the coefficients A^ω_{1k} can be computed as

$$(-1)^{k-1}\sum_{i=1}^{p}\alpha_i\lambda_i^{p+\omega-1}e_{k-1}(\lambda/\lambda_i)$$

Claim. $|A^\omega_{1k}|$ is the Schur polynomial $S_{\{\omega,1,1,\cdots\ k-1\text{times}\cdots\ 1,\ 0,0,\cdots,0\}}$

For any $\mu = \{\mu_1, \mu_2, \cdots, \mu_p\}$ such that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_p$ consider the generalized Vandermonde matrix $V_{\mu,\lambda}$ defined as

$$V_{\mu,\lambda} = \begin{pmatrix} \lambda_1^{p-1+\mu_1} & \lambda_2^{p-1+\mu_1} & \cdots & \lambda_p^{p-1+\mu_1} \\ \lambda_1^{p-2+\mu_2} & \lambda_2^{p-2+\mu_2} & \cdots & \lambda_p^{p-2+\mu_2} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1^{1+\mu_{p-1}} & \lambda_2^{1+\mu_{p-1}} & \cdots & \lambda_p^{1+\mu_{p-1}} \\ \lambda_1^{\mu_p} & \lambda_2^{\mu_p} & \cdots & \lambda_p^{\mu_p} \end{pmatrix}. \tag{20}$$

The Bilaternant formulation defines Schur polynomial $S_{\mu,\lambda}$ as

$$S_{\mu,\lambda} = \frac{\text{det}(V_{\mu,\lambda})}{\text{det}(V_\lambda)}. \tag{21}$$

It can be shown that the determinant of the vandemonde matrix V_λ can be given as

$$\text{det}(V_\lambda) = \prod_{1\leq i<j\leq n}(\lambda_i - \lambda_j). \tag{22}$$

A proof of this statement can be found in most standard texts on Matrix analysis, for example, see Horn et al. (2012).

For any $i, k \in [p]$, consider the generalized Vandermonde matrix $V_{\mu_k,\lambda/\lambda_i}$, where $\mu_k = \{1, 1, \cdots\ k-1\text{times}\cdots\ 1,\ 0, 0, \cdots, 0\}$. That is,

$$V_{\mu_k,\lambda/\lambda_i} = \begin{pmatrix} \lambda_1^{p-1} & \lambda_2^{p-1} & \cdots & \lambda_{i-1}^{p-1} & \lambda_{i+1}^{p-1} & \cdots & \lambda_p^{p-1} \\ \lambda_1^{p-2} & \lambda_2^{p-2} & \cdots & \lambda_{i-1}^{p-2} & \lambda_{i+1}^{p-2} & \cdots & \lambda_p^{p-2} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ \lambda_1^{p-(k-1)} & \lambda_2^{p-(k-1)} & \cdots & \lambda_{i-1}^{p-(k-1)} & \lambda_{i+1}^{p-(k-1)} & \cdots & \lambda_p^{p-(k-1)} \\ \lambda_1^{p-(k+1)} & \lambda_2^{p-(k+1)} & \cdots & \lambda_{i-1}^{p-(k+1)} & \lambda_{i+1}^{p-(k+1)} & \cdots & \lambda_p^{p-(k+1)} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{23}$$

From (21), we know that

$$\text{det}(V_{\mu_k,\lambda/\lambda_i}) = \text{det}(V_{\lambda/\lambda_i})S_{\mu_k,\lambda/\lambda_i},$$

where $S_{\mu_k,\lambda/\lambda_i}$ is the Schur polynomial of variables λ/λ_i indexed by $\mu_k = \{1, 1, \cdots\ k-1\text{times}\cdots\ 1,\ 0, 0, \cdots, 0\}$. Using a combinatorial definition of a Schur polynomial as a summation over semi-standard representations over a Young's Tableaux (see Macdonald (1998) for an exposition), it is easy to verify that

(24)

$$S_{\mu_k,\lambda/\lambda_i} = e_{k-1}(\lambda/\lambda_i).$$

Therefore, combining (22) and (24) we can write

$$\det(V_{\mu_k,\lambda/\lambda_i}) = \det(V_{\lambda/\lambda_i})e_{k-1}(\lambda/\lambda_i) = e_{k-1}(\lambda/\lambda_i) \prod_{\substack{1\le l<l'\le p\\ l,l'\neq i}} (\lambda_l - \lambda_{l'})$$

Now, observe that we can rewrite A^{ω}_{1k} as

$$\begin{aligned} A^{\omega}_{1k} &= (-1)^{k-1}\sum_{i=1}^{p} \alpha_i \lambda_i^{p+\omega-1} e_{k-1}(\lambda/\lambda_i), \\ &= (-1)^{k-1}\sum_{i=1}^{p} (-1)^{i+1} \lambda_i^{p+\omega-1} e_{k-1}(\lambda/\lambda_i) \prod_{\substack{1\le l<l'\le p\\ l,l'\neq i}} (\lambda_l - \lambda_{l'})/\det(V_\lambda), \\ &= (-1)^{k-1}\sum_{i=1}^{p} (-1)^{i+1} \lambda_i^{p+\omega-1} \det(V_{\mu_k,\lambda/\lambda_i})/\det(V_\lambda). \end{aligned}$$

Finally, letting $\mu_{\omega,k} = \{\omega, 1, 1, \cdots\ k-1 \text{times} \cdots\ 1,\ 0, 0, \cdots, 0\}$, consider the generalized Vandermonde matrix $V_{\mu_{\omega,k},\lambda}$ given by

$$V_{\mu_{\omega,k},\lambda} = \begin{pmatrix} \lambda_1^{p-1+\omega} & \lambda_2^{p-1+\omega} & \cdots & \lambda_p^{p-1+\omega} \\ \lambda_1^{p-1} & \lambda_2^{p-1} & \cdots & \lambda_p^{p-1} \\ \lambda_1^{p-2} & \lambda_2^{p-2} & \cdots & \lambda_p^{p-2} \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_1^{p-(k-1)} & \lambda_2^{p-(k-1)} & \cdots & \lambda_p^{p-(k-1)} \\ \lambda_1^{p-(k+1)} & \lambda_2^{p-(k+1)} & \cdots & \lambda_p^{p-(k+1)} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{25}$$

Using the Laplace expansion to compute the determinant along the first row of $V_{\mu_{\omega,k},\lambda}$ and observing that for any $i \in [p]$, the minor of $V_{\mu_{\omega,k},\lambda}(1,i)$ is given by $\det(V_{\mu_k,\lambda/\lambda_i})$, we have

$$\sum_{i=1}^{p} (-1)^{i+1} \lambda_i^{p+\omega-1} e_{k-1}(\lambda_i) \prod_{\substack{1\le l<l'\le p\\ l,l'\neq i}} (\lambda_l - \lambda_{l'}) = \det(V_{\mu_{\omega,k},\lambda})$$

and once again by invoking the bialternant formulation for Schur polynomials, we have

$$|A^{\omega}_{1k}| = \sum_{i=1}^{p} \alpha_i \lambda_i^{p+\omega-1} e_{k-1}(\lambda_i) = \frac{\det(V_{\mu_{\omega,k},\lambda})}{\det(V_\lambda)} = S_{\mu_{\omega,k},\lambda}.$$

□

Lemma 4 (Form of Interventional Autocovariance matrix). *Consider a vector-valued time series $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$, following a VAR(q) process with autocovariance matrix of size $nd \times nd$ denoted by Σ_n. Consider simultaneous atomic interventions on components $\{l_1, l_2, \cdots, l_r\} \subset [d]$ of $x_{t-\omega}$, that is, consider the intervention $do(x_{t-\omega,l_1} = x*_{t-\omega,l_1}, \cdots, x_{t-\omega,l_r} =$*

$x*_{t-\omega,l_r}$). Then, the autocovariance matrix of size $nd \times nd$ (Γ'_n) of the corresponding joint interventional distribution, denoted by $\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})$ is given by

$$\Gamma'_n(i,j) = \begin{cases} 0 & \text{if } i \neq j, i = l_m,\ j = l_m\ \forall m \in [r] \\ x*^2_{t-\omega,l_m} & \text{if } i = j = l_m\ \forall m \in [r] \\ \Sigma_n(i,j) & \text{otherwise} \end{cases}. \tag{26}$$

Moreover, let

$$\Gamma_n = \mathbb{E}_{\{x*_{t-\omega,l_m}\}_{m\in[r]} \sim \prod_{m\in[r]} \mathbb{P}(x_{t-\omega},l_m)}\Gamma'_n.$$

Then,

$$\Gamma_n(i,j) = \begin{cases} 0 & \text{if } i \neq j, i = l_m,\ j = l_m\ \forall m \in [r] \\ \Sigma_n(i,j) & \text{otherwise} \end{cases}. \tag{27}$$

The autocovariance matrix of the interventional distribution under simultaneous interventions on consecutive time-steps can be analogously obtained.

Proof of Lemma 4.. Note that due to time ordering and since instantaneous effects are not modelled by a VAR model, there is no directed path from any of the variables $x_{t-\omega,l_1}, x_{t-\omega,l_2}, \cdots, x_{t-\omega,l_r}$ to $\mathbf{x}^n_{t-\omega-1}$ as well as to variables in $\{x_{t-\omega,1}, x_{t-\omega,2}, \cdots, x_{t-\omega,d}\} / x_{t-\omega,l_1}, x_{t-\omega,l_2}, \cdots, x_{t-\omega,l_r}$. Peters et al. (2017, Proposition 6.14) provides graphical criterion for determining the existence of a total causal effect from a variable x to a variable y under interventions on x. Absence of a directed path from x to y implies there is no total causal effect from x to y and from Proposition 6.12 of Peters et al. (2017), we know that $x \perp\!\!\!\perp y$ under the corresponding interventional distribution. As a consequence of these Propositions, we have our desired result.

□

Lemma 5 (Difference in Causal and Statistical error (VAR(p))). *Consider a vector-valued time series $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$, following a VAR(q) process with model parameters $\{A_1, A_2, \cdots A_q\}$. Assuming $n > \max\{p,q\}$, for any VAR(p) model f with parameters $\{\widehat{A}_1, \widehat{A}_2, \cdots \widehat{A}_p\}$,*

$$|\mathcal{G}_{do_\omega}(f) - \mathcal{S}(f)| = \sum_{i=1}^{d} (A^\omega_{i:} - \widehat{A}^\omega_{i:})^T (\Gamma - \Sigma)(A^\omega_{i:} - \widehat{A}^\omega_{i:}), \tag{28}$$

Proof of Lemma 5. Let A denote the multi-companion matrix corresponding to the true VAR(q) process with model parameters $\{A_1, A_2, \cdots, A_q\}$ of the form described in (14) with the first d rows populated by $\{A'_1, A'_2, \cdots A'_{\max\{p,q\}}\}$, where A'_l is defined as A_l for all $l \leq q$ and as $\mathbf{0}_{d\times d}$ for all $l > q$. Define $\widehat{A}^{(\max\{p,q\})}$ analogously as the multi-companion matrix corresponding to parameters $\left\{\widehat{A}_1, \widehat{A}_2, \cdots, \widehat{A}_p\right\}$ of the estimated VAR(p) model f obtained independently from some statistical estimation procedure $\mathcal{E}$.

Using (12) recursively, we can write

$$y^{(\max\{p,q\})}_t = A^\omega y^{(\max\{p,q\})}_{t-\omega} + A^\omega e^{(\max\{p,q\})}_{t-\omega+1} + A^{\omega-1} e^{(\max\{p,q\})}_{t-\omega+2} + \cdots + A e^{(\max\{p,q\})}_{t-1} + e^{(\max\{p,q\})}_t \tag{29}$$

To reduce cumbersome notation, we let $\zeta_t = A^\omega e_{t-\omega+1} + A^{\omega-1} e_{t-\omega+2} + \cdots + A e_{t-1} + e_t \in \mathbb{R}^{dp}$ and write

$$Y_t = A^\omega y_{t-\omega} + \zeta_t. \tag{30}$$

Let $\hat{x}_t$ denote the prediction of the target variable x_t corresponding to the estimated model f. Then, Statistical error $\mathcal{O}_\omega$ defined with respect to the squared norm can be computed as follows:

$$\begin{aligned}
\mathcal{O}_\omega &= \mathbb{E}_{\mathbb{P}(\mathbf{x}^n_{t-\omega}, x_t)}[\|x_t - \hat{x}_t\|^2] \\
&= \sum_{i=1}^{d} \mathbb{E}[x_{t,i} - \hat{x}_{t,i}]^2 && \text{(Subscript omitted for convenience)} \\
&= \sum_{i=1}^{d} \mathbb{E}[A^\omega_{i,:} y_{t-\omega} + \zeta_{t,\omega,i} - \hat{A}^\omega_{i,:} y_{t-\omega}]^2 \\
&= \sum_{i=1}^{d} (A^\omega_{i:} - \hat{A}^\omega_{i:})^T \mathbb{E}[y_{t-\omega} y^T_{t-\omega}](A^\omega_{i:} - \widehat{A}^\omega_{i:}) + \mathbb{E}[\zeta^2_{t,\omega,i}] && (\mathbb{E}[x_{t-i}\epsilon^T_t] = 0,\ \forall i \in \mathbb{N}\ ) \\
&= \sum_{i=1}^{d} (A^\omega_{i:} - \widehat{A}^\omega_{i:})^T \Sigma_{\max\{p,q\}} (A^\omega_{i:} - \widehat{A}^\omega_{i:}) + \mathbb{E}[\zeta^2_{t,\omega,i}]
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathcal{G}_{do_\omega} &= \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega}, x_t)}(\|x_t - \hat{x}_t\|^2) && (31) \\
&= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}_{do_\omega}(x_t|\mathbf{x}^n_{t-\omega})} \left[x_{t,i} - \hat{x}_{t,i}\right]^2 && (32) \\
&= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}_{do_\omega}(x_t|\mathbf{x}^n_{t-\omega})} \left[x^2_{t,i} + \hat{x}^2_{t,i} - 2x_{t,i}\hat{x}_{t,i}\right]^2 && (33) \\
&= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}_{do_\omega}(x_t|\mathbf{x}^q_{t-\omega})} \left[x^2_{t,i} + \hat{x}^2_{t,i} - 2x_{t,i}\hat{x}_{t,i}\right]^2 && (34) \\
&= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}(X_t|\mathbf{x}^q_{t-\omega})} \left[x^2_{t,i} + \hat{x}^2_{t,i} - 2x_{t,i}\hat{x}_{t,i}\right]^2 && (35) \\
&= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^q_{t-\omega})} \left[\mathbb{E}_{\mathbb{P}(x_t|\mathbf{x}^q_{t-\omega})}[x^2_{t,i}] + ((\widehat{A}^\omega_{i:})^T y_{t-\omega})^2 - 2\mathbb{E}_{\mathbb{P}(x_t|\mathbf{x}^q_{t-\omega})}[x_{t,i}](\widehat{A}^\omega_{i:})^T y_{t-\omega}\right]^2 && (36) \\
&= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})} \left[((A^\omega_{i:})^T y_{t-\omega} + \zeta_t)^2 + ((\widehat{A}^\omega_{i:})^T y_{t-\omega})^2 - 2((A^\omega_{i:})^T y_{t-\omega})((\widehat{A}^\omega_{i:})^T y_{t-\omega})\right]^2 && (37) \\
&= \sum_{i=1}^{d} \left((A^\omega_{i:} - \widehat{A}^\omega_{i:})^T \mathbb{E}_{do_\omega}(y_{t-\omega} y^T_{t-\omega})(A^\omega_{i:} - \widehat{A}^\omega_{i:}) + \mathbb{E}(\zeta^2_{t,i})\right) \qquad (\mathbb{E}(x_{t-i}\epsilon^T_t) = 0,\ \forall i \in \mathbb{N}\ ) && (38) \\
&= \sum_{i=1}^{d} (A^\omega_{i:} - \widehat{A}^\omega_{i:})^T \Gamma'_{\max\{p,q\}} (A^\omega_{i:} - \widehat{A}^\omega_{i:}) + \mathbb{E}(\zeta^2_{t,i}) && (39)
\end{aligned}$$

To see why Equation (35) holds, note that the structural equations that specify the dependence of x_t on $\mathbf{x}^q_{t-\omega}$ remain unchanged under interventions on $x_{t-\omega}$ and therefore the conditional distributions remain unchanged under these interventions.

Therefore,

$$\mathbb{E}_{x*_{t-\omega} \sim \mathbb{P}(x_{t-\omega})} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega}, x_t)}(\|x_t - \hat{x}_t\|^2) = \sum_{i=1}^{d} (A^\omega_{i:} - \widehat{A}^\omega_{i:})^T \Gamma_{\max\{p,q\}} (A^\omega_{i:} - \widehat{A}^\omega_{i:}) + \mathbb{E}(\zeta^2_{t,i}),$$

where Γ can be obtained using Lemma 4.

□

Corollary 3 (Difference in Causal and Statistical errors (AR)). *Let $\{x_t\}$ follow an AR(q) process. Then, for any AR(p)* *model f with parameters $\{\widehat{a}_1, \widehat{a}_2, \cdots, \widehat{a}_p\}$,*

$$|\mathcal{G}_{do_\omega}(f) - \mathcal{S}_\omega(f)| = 2\left|(A^\omega_{1,1} - \widehat{A}^\omega_{1,1}) \sum_{k=2}^{\max\{p,q\}} (A^\omega_{1,k} - \widehat{A}^\omega_{1,k})\gamma_{k-1}\right|, \tag{40}$$

where, for any $k \in \mathbb{N}$, γ_k denotes the autocovariance of $\{x_t\}$ with lag k. A and $\widehat{A}$ are the corresponding companion matrices of the model and estimated parameters as defined in Lemma 5.

***Proof of Corollary 3**.* Corollary 1 directly follows from Lemmas 4 and 5. □

Proposition 2 (Stability Controls Causal Generalization (VAR)). *Consider a VAR(q) process. Assuming $n > \max\{p, q\}$, for any VAR(p) model f,*

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq 2\kappa(\Sigma_{\max\{p,q\}})(\mathcal{S}_\omega(f) - \sigma^2_\epsilon), \tag{41}$$

where $\kappa(\Sigma_{\max\{p,q\}})$ denotes the condition number of the autocovariance matrix $\Sigma_{\max\{p,q\}}$.

Proof. From Lemma 5, it remains to prove that

$$|\sum_{j=1}^{d}(A^\omega_{j:} - \widehat{A}^\omega_{j:})^T(\Gamma - \Sigma)(A^\omega_{j:} - \widehat{A}^\omega_{j:})| \leq (2\kappa(\Sigma) - 1)(\mathcal{S}_\omega(f) - \sigma^2_\epsilon).$$

First, we show that

$$|(A^\omega_{j:} - \widehat{A}^\omega_{j:})^T(\Gamma - \Sigma)(A^\omega_{j:} - \widehat{A}^\omega_{j:})| \leq (2\lambda_{\max}(\Sigma)) \left\|A^\omega_{j:} - \widehat{A}^\omega_{j:}\right\|^2. \tag{42}$$

Case 1. $(A^\omega_{j:} - \widehat{A}^\omega_{j:})^T(\Gamma - \Sigma)(A^\omega_{j:} - \widehat{A}^\omega_{j:}) \geq 0$.

$$|(A^\omega_{j:} - \widehat{A}^\omega_{j:})^T(\Gamma - \Sigma)(A^\omega_{j:} - \widehat{A}^\omega_{j:})| = (A^\omega_{j:} - \widehat{A}^\omega_{j:})^T(\Gamma - \Sigma)(A^\omega_{j:} - \widehat{A}^\omega_{j:}), \tag{43}$$

$$\leq (\lambda_{\max}(\Gamma) - \lambda_{min}(\Sigma)) \left\|A^\omega_{j:} - \widehat{A}^\omega_{j:}\right\|^2. \tag{44}$$

where (44) holds by an application of Rayleigh's principle. We still need to show that $\lambda_{\max}(\Gamma) \leq 2\lambda_{\max}(\Sigma)$.

Without loss of generality, assume that $i = 1$, that is the component of $x_{t-\omega}$ that is intervened upon is indexed by 1. Note that, this merely simplifies notation and the following steps also hold simultaneous interventions on multiple components and consecutive time instances without any additional steps.

Representing Σ and Γ in block matrix form, we have

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix}. \tag{45}$$

From Lemma 4, we have

$$\Gamma_{11} \in \mathbb{R}^{1\times 1} = \sigma^2 = \mathbb{E}(X_t^2),\ \Gamma_{12}^T = \Gamma_{21} \in \mathbb{R}^{1\times d\max\{p,q\}-1} = 0,\ \text{and}\ \Gamma_{22} = \Sigma_{22}.$$

We can write Γ as follows:

$$\Gamma = \Gamma_1' + \Gamma_2', \tag{46}$$

where

$$\Gamma_1' = \begin{pmatrix} \sigma^2 & \mathbf{0}_{1\times d\max\{p,q\}-1} \\ \mathbf{0}_{d\max\{p,q\}-1\times 1} & \mathbf{0}_{d\max\{p,q\}-1\times d\max\{p,q\}-1} \end{pmatrix}, \tag{47}$$

and

$$\Gamma_2' = \begin{pmatrix} \mathbf{0}_{1\times 1} & \mathbf{0}_{1\times d\max\{p,q\}-1} \\ \mathbf{0}_{d\max\{p,q\}-1\times 1} & \Sigma_{22} \end{pmatrix}. \tag{48}$$

Since Γ_1' and Γ_2' are Hermitian matrices, $\lambda_{\max}(\Gamma) \leq \lambda_{\max}(\Gamma_1') + \lambda_{\max}(\Gamma_2')$.

Observe that Γ_2 is a principal sub-matrix of Γ obtained by deleting the first row and column, by Cauchy's interlacing theorem (Fisk 2005), we have

$$\lambda_{\max}(\Gamma_2') \leq \lambda_{\max}(\Sigma). \tag{49}$$

Note that, when we intervene simultaneously on multiple components and time instances, instead of setting the first row to 0, the covariance matrix of the corresponding interventional distribution Γ can be obtained by deleting the off-diagonal elements of the corresponding rows and columns. It remains to show that $\sigma^2 \leq \lambda_{\max}(\Sigma)$. Note that

$$\lambda_{\max}(\Gamma_2') = \sigma^2 = \Sigma_{11} = e_1^T \Sigma e_1 \leq \lambda_{\max}(\Sigma), \tag{50}$$

where e_i denotes the ith standard basis vector. Combining (49) and (50) we have

$$\lambda_{\max}(\Gamma) \leq 2\lambda_{\max}(\Sigma) \tag{51}$$

and

$$|(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T(\Gamma - \Sigma)(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})| \leq (2\lambda_{\max}(\Sigma) - \lambda_{\min}(\Sigma)) \left\| A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:} \right\|^2 . \tag{52}$$

Case 2. $(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T(\Gamma - \Sigma)(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}) \leq 0$.

$$\begin{aligned} |(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T(\Gamma - \Sigma)(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})| &= (A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T(\Sigma - \Gamma)(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}), && (53)\\ &\leq (\lambda_{\max}(\Sigma) - \lambda_{\min}(\Gamma)) \left\| A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:} \right\|^2 . && (54) \end{aligned}$$

Using the same arguments used in deriving upper bounds for $\lambda_{\max}(\Gamma)$, we can show that $\lambda_{\min}(\Gamma) \geq \lambda_{\min}(\Sigma)$. Therefore, we have

$$\begin{aligned} |\mathcal{G}_{do_{\omega,i}}(f) - \mathcal{S}_{\omega}(f)| &\leq \sum_{j \in [d]} (2\lambda_{\max}(\Sigma) - \lambda_{\min(\Sigma)}) \left\| A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:} \right\|^2 && (55)\\ &\leq (2\lambda_{\max}(\Sigma) - \lambda_{\min(\Sigma)}) \sum_{j \in [d]} \left\| A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:} \right\|^2 && (56)\\ &\leq (2\kappa(\Sigma) - 1)(\mathcal{S}_{\omega}(f) - \sigma_{\epsilon}^2). && (57) \end{aligned}$$

To see why (57) holds, observe that

$$\begin{aligned} \mathcal{S}_{\omega}(f) - \sigma_{\epsilon}^2 &= \sum_{j=1}^{d} (A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T \Sigma (A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}) && (58)\\ &\geq \sum_{j=1}^{d} (A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T \Sigma (A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}) && (59)\\ &\geq \sum_{j=1}^{d} \lambda_{\min}(\Sigma) \left\| A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:} \right\|^2 . && (60) \end{aligned}$$

We now show that we can construct AR(2) processes such that the bound in Proposition 1 is tight upto a small constant factor. Consider an AR(2) process with true model parameters a_1 and a_2. The autocorrelation matrix Σ_2 of this process is given by $\Sigma_p = \begin{pmatrix} 1, \gamma \\ \gamma, 1 \end{pmatrix}$ where $\gamma = \frac{a_1}{1 - a_2}$. The eigenvalues of Σ_2 are given by $\lambda_1 = 1 + \gamma$ and $\lambda_2 = 1 - \gamma$ corresponding to eigenvectors u_1 and u_2 respectively. Without loss of generality assume $\gamma > 0$ which yields $\lambda_1 \geq \lambda_2$. Denote vectors $a = (a_1, a_2)$ and $\hat{a} = (\hat{a}_1, \hat{a}_2)$. Consider an AR(2) process with parameters $\hat{a}_1, \hat{a_2}$ such that $(a - \hat{a}) = u_2$. Then assuming $\omega = 1$, we have that

$$\frac{\mathcal{G}_{do_1} - \mathcal{S}_1}{\mathcal{S}_1 - \sigma_{\epsilon}^2} = \frac{\|a - \hat{a}\|^2 - (a - \hat{a})^T \Sigma (a - \hat{a})}{(a - \hat{a})^T \Sigma (a - \hat{a})} = \frac{\gamma}{1 - \gamma} = (\kappa(\Sigma) - 1)/2.$$

As a approaches the boundary of the stability domain, the process gets more strongly correlated and $\lambda_{\min}$ approaches 0 and the relative difference in causal and statistical errors diverges. □

Lemma 6 (**Bounds on** a_k)**.** *For any AR(p) model such that the non-zero eigenvalues of the companion matrix are distinct and satisfy* $|\lambda| \leq \delta < 1$,

$$|a_k| \leq \binom{p}{k}\delta^k. \tag{61}$$

Proof of Lemma 6. From Lemma 3, we know that

$$\begin{aligned}
|a_k| &= |S_{\{1,1,\cdots k\ times\ 1,0,\cdots,0\}}(\{\lambda_1, \lambda_2, \cdots, \lambda_p\})| \\
&= |\sum_{\{i_1<i_2<\cdots<i_k\}\in[p]} \lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_k}| \\
&\leq \sum_{\{i_1<i_2<\cdots<i_k\}\in[p]} |\lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_k}| && (|x+y| \leq |x|+|y|) \\
&\leq \sum_{\{i_1<i_2<\cdots<i_k\}\in[p]} \delta^k && (|\lambda_i| \leq \delta) \\
&= \binom{p}{k}\delta^k.
\end{aligned}$$

□

Lemma 7 (**Bounds on** γ_k)**.** *For any stochastic process* $\{x_t\}_{t\in\mathbb{Z}}$ *following an AR(p) model the non-zero eigenvalues of the companion matrix are distinct and satisfy* $|\lambda| \leq \delta < 1$

$$|\gamma_k| \leq \frac{C\sigma_\epsilon^2\delta^k}{1-\delta^2}$$

Proof of Lemma 7. Using the infinite-moving average representation of X_t (See Brockwell et al. (1991)), we have

$$x_t = \sum_{i=0}^{\infty} A_{11}^i \epsilon_{t-i} \tag{62}$$

$$\begin{aligned}
|\mathbb{E}[x_l, x_r]| &= |\mathbb{E}[(\sum_{i_1=0}^{\infty} A_{11}^{i_1}\epsilon_{l-i_1})(\sum_{i_2=0}^{\infty} A_{11}^{i_2}\epsilon_{r-i_2})]| && (63) \\
&= |\sum_{i=0}^{\infty} A_{11}^i A_{11}^{i+|l-r|}\mathbb{E}[\epsilon_t\epsilon_t^T]| && (64) \\
&= |\sigma_\epsilon^2 \sum_{i=0}^{\infty} A_{11}^i A_{11}^{i+|l-r|}| && (65) \\
&\leq K_p\delta^{|l-r|}\sigma_\epsilon^2 \sum_{i=0}^{\infty}\delta^{2i} && (66) \\
&\leq K_p\sigma_\epsilon^2\frac{\delta^{|l-r|}}{1-\delta^2} && (67)
\end{aligned}$$

To see why (66) holds observe that, from Lemma 3,

$$A_{11}^i = S_{\{i,0,\cdots,0\}} \leq \sum_{\{i_1\leq i_2\leq\cdots\leq i_k\}\in[p]} |\lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_k}| \leq p^p\delta^i$$

□

Lemma 8 (**Lower Bounds on** $\lambda_{\min}(\Sigma)$)**.** *For any stochastic process* $\{x_t\}_{t\in\mathbb{Z}}$ *following an AR(p) model the non-zero eigenvalues of the companion matrix are distinct and satisfy* $|\lambda| \leq \delta < 1$

$$\lambda_{\min}(\Sigma) \geq \frac{\sigma_\epsilon^2}{(1+\delta)^{2p}}$$

Proof. First, note that

$$(1 + \sum_{k=1}^{p} |a_k|) \leq \sum_{k=0}^{p} \binom{p}{k} \delta^k = (1+\delta)^p (\text{Binomial Theorem}).$$

Combining this with the results from Lemma 11 and Proposition 3, we have

$$\lambda_{\min}(\Sigma) \geq 2\pi \inf_{\omega} f(\omega) \geq \frac{\sigma_\epsilon^2}{\nu_{\max}(\mathcal{A})} \geq \frac{\sigma_\epsilon^2}{(1 + \sum\limits_{k=1}^{p} |a_k|)^2}$$

$$\lambda_{\min}(\Sigma) \geq \frac{\sigma_\epsilon^2}{(1 + \sum\limits_{k=1}^{p} |a_k|)^2} \geq \frac{\sigma_\epsilon^2}{(1+\delta)^{2p}}.$$

□

Lemma 9 (**Upper Bounds on** $\lambda_{\max}(\Sigma)$)**.** *For any stochastic process* $\{x_t\}_{t\in\mathbb{Z}}$ *following an AR(p) model the non-zero eigenvalues of the companion matrix are distinct and satisfy* $|\lambda| \leq \delta < 1$

$$\lambda_{\max}(\Sigma) \leq 2K_p \sigma_\epsilon^2 n \frac{1}{1-\delta^2}$$

Proof. By Gershgorin's theorem (Varga 2010), we can derive an upper bound on the maximum eigenvalue of Σ_n as follows:

$$\lambda_{\max}(\Sigma_n) \leq \max_{i\in[n]} (\Sigma_{ii} + \sum_{j\neq i} |\Sigma_{ij}|).$$

Note that the autocovariance matrix of an AR process which is defined as $\Sigma_{i,j} = \gamma_{|i-j|}$ (the autocovariance of lag $|i-j|$) has a Toeplitz structure. Due to this Toeplitz structure of the autocovariance matrix, we can see that

$$\lambda_{\max}(\Sigma_n) < 2\sum_{i=1}^{n} |\gamma_{i-1}| < 2K_p \sigma_\epsilon^2 \sum_{i=1}^{n} \frac{\delta^{i-1}}{1-\delta^2} \leq 2K_p n \sigma_\epsilon^2 \frac{1}{1-\delta^2}$$

□

Corollary 4 (**Stability Controls Causal Generalization (AR(p))**)**.** *Consider an AR(q) process, such that eigenvalues of its companion matrix satisfy* $|\lambda| < \delta < 1$. *For any AR(q) model* f,

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq K_p \mathcal{S}_\omega(f) \frac{\max\{p,q\}\,(1+\delta)^{2\max\{p,q\}}}{(1-\delta^2)}, \tag{68}$$

where K_p *is some finite constant that depends on the order p of the underlying process.*

Proof of Corollary 4. From Proposition 2, we already know that

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq 2\kappa(\Sigma_{\max\{p,q\}})(\mathcal{S}_\omega(f) - \sigma_\epsilon^2), \tag{69}$$

From Lemma 8 and Lemma 9, we have that

$$\lambda_{\min}(\Sigma_{\max\{p,q\}}) \geq \frac{\sigma_\epsilon^2}{(1+\delta)^{2p}}$$

and

$$\lambda_{\max}(\Sigma_{\max\{p,q\}}) \leq 2K_p \max\{p,q\}\,\sigma_\epsilon^2 \frac{1}{1-\delta^2}$$

.

Combining these results, we have the desired result. □

Theorem 2 (Finite sample bounds for VAR(p) models). *Let $\mathcal{F}$ denote the family of all VAR models of dimension* d *and order* p*. For any $n > \max\{p, q\} \in \mathbb{N}$, let $\mu, m > 0$ be integers such that $2\mu m = n$ and $\delta > 2(\mu - 1)\rho^m$ for a fixed constant $0 < \rho < 1$ determined by the underlying process. Let $\{x_1, x_2, \cdots x_n\} \in \mathbb{R}^d$ be a finite sample drawn from a VAR(q) process. Then, simultaneously for every $f \in \mathcal{F}$, under the square loss truncated at M, with probability at least $1 - \delta$,*

$$\mathcal{G}_{\omega,i} \leq \zeta \hat{\mathcal{S}}_\omega + \zeta \widehat{\mathfrak{R}}_\mu(\mathcal{F}) + 3\zeta M \sqrt{\frac{\log \frac{4}{\delta'}}{2\mu}} \tag{70}$$

where $\zeta = 2\kappa(\Sigma^\nu)$, $\delta' = \delta - 2(\mu - 1)\rho^m$, and $\widehat{\mathfrak{R}}_\mu(\mathcal{F})$ denotes the empirical Rademacher complexity of $\mathcal{F}$.

Proof of Theorem 2. From Proposition , we already have that

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq (2\kappa(\Sigma_{\max\{p,q\}}) - 1)(\mathcal{S}_\omega(f) - \sigma_\epsilon^2). \tag{71}$$

Additionally, processes that follow VAR models are known to be β mixing and in particular, they are geometrically completely regular, that is, there exists some $0 < \rho < 1$ such that $\beta(k) = C\rho^k$ for some constant C, where $\beta(k)$ denotes the β mixing coefficient of the process (Mokkadem 1988). Theorem 2 then follows by applying Rademacher bounds (Mohri et al. 2009, Theorem 1) for generalization in time-series under mixing conditions. □

C RELATIVE INTERVENTIONS

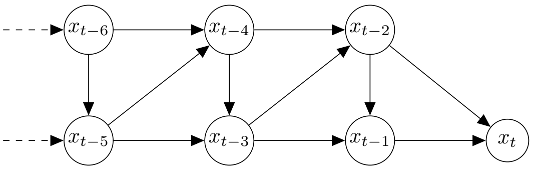


Figure 7: Causal DAG of an AR(2) model

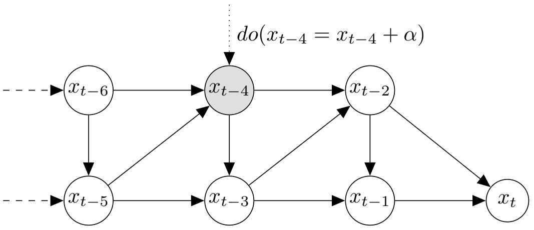


Figure 8: Graphical representation of the effect of an intervention $do(x_{t-4} = x_{t-4} + \alpha)$ on an AR(2) model. Dependencies are retained.

Assume for simplicity $p = q$ and $d = 1$. Let A and $\widehat{A}$ denote the companion matrices corresponding to the true and estimated parameters respectively. Then, rewriting the VAR(p) model as a VAR(1) model, we have

$$x_t = A_{11}^\omega x_{t-\omega} + A_{12}^\omega x_{t-\omega-1} + \cdots + A_{1p}^\omega x_{t-\omega-p+1} + A_{11}^{\omega-1} \epsilon_{t-\omega+1} + \cdots + A_{11}\epsilon_{t-1} + \epsilon_t. \tag{72}$$

Let $\zeta_t = A_{11}^{\omega-1}\epsilon_{t-\omega+1} + \cdots + A_{11}\epsilon_{t-1} + \epsilon_t$. Then, Statistical error S_ω can be computed as

$$\mathbb{E}[x_t - \hat{x}_t]^2 = \mathbb{E}[\sum_{i=1}^{p}(A_{1i}^\omega - \widehat{A}_{1i}^\omega)x_{t-\omega-i+1} + \zeta_t^2] \tag{73}$$

$$= \sum_{ij=1}^{p}(A_{1i}^\omega - \widehat{A}_{1i}^\omega)(A_{1j}^\omega - \widehat{A}_{1j}^\omega)\Sigma_{ij} + \mathbb{E}[\zeta_t^2] \tag{74}$$

The causal error $\mathcal{G}_{do_\omega}$ due to the effect of an intervention $do(x_{t-\omega} = x_{t-\omega} + \alpha)$ can be computed as

$$\mathbb{E}_{do_\omega}[x_t - \hat{x}_t]^2 = \mathbb{E}[\sum_{i=1}^{p}(A^\omega_{1i} - \widehat{A}^\omega_{1i})x_{t-\omega-i+1} + (A^\omega_{11} - \widehat{A}^\omega_{11})\alpha + \zeta^2_t] \tag{75}$$

$$= \sum_{ij=1}^{p}(A^\omega_{1i} - \widehat{A}^\omega_{1i})(A^\omega_{1j} - \widehat{A}^\omega_{1j})\Sigma_{ij} + (A^\omega_{11} - \widehat{A}^\omega_{11})^2\alpha^2 + \mathbb{E}[\zeta^2_t] \tag{76}$$

To see why (76) holds, recall that $\mathbb{E}[x_t] = 0, \mathbb{E}[\epsilon_t] = 0, \mathbb{E}[x_{t-i}\epsilon_t] = 0 \; \forall i \in \mathbb{N}$.

Lemma 10 (Difference in Causal and Statistical errors (AR) under Relative Interventions). *Let $\{X_t\}$ follow an AR(q) process. Then, for any AR(p) model f with parameters $\{\widehat{a}_1, \widehat{a}_2, \cdots, \widehat{a}_p\}$,*

$$\mathcal{G}_{do_\omega}(f) - \mathcal{S}_\omega(f) = (A^\omega_{1,1} - \widehat{A}^\omega_{1,1})^2\alpha^2, \tag{77}$$

where, A and $\widehat{A}$ are the corresponding companion matrices of the model and estimated parameters.

D OTHER RESULTS

Proposition 3. *(Basu et al. 2015, Proposition 2.2) Consider a (matrix-valued) polynomial $\mathcal{A}(z) = I_d - \sum_{k=1}^{p} A_k z^k, x \in \mathbb{C},\; p \in \mathbb{N}$, satisfying $det(\mathcal{A}(z)) \neq 0$ for all $|z| < 1$, $\mu_{\max}(\mathcal{A}) \leq (1 + (\nu_{row} + \nu_{col})/2)^2$, where*

$$\nu_{row} = \sum_{k=1}^{p} \max_{1\leq i\leq d} \sum_{j=1}^{d} |A_k(i,j)|, \quad \nu_{col} = \sum_{k=1}^{p} \max_{1\leq i\leq d} \sum_{i=1}^{d} |A_k(i,j)|.$$

Lemma 11 (Bounds on spectrum of Σ). *Let $\{X_t\}$ be a second-order stationary time series with spectral density $f(\omega)$ and let Σ_n denote the autocorrelation matrix of size $n \times n$ given by $\Sigma_n(i,j) = \gamma_{|i-j|} = \mathbb{E}(x_{t+i}, x_{t+j})$ for any $i, j \in \mathbb{Z}$. Then the extremal eigenvalues of Σ are bounded as follows.*

$$\lambda_{\min}(\Sigma_n) \geq 2\pi \inf_\omega f(\omega) \quad \text{and} \quad \lambda_{\max}(\Sigma_n) \leq 2\pi \sup_\omega f(\omega) \; \forall n \in \mathbb{N}$$

Furthermore, the bound holds uniformly for all $n \in \mathbb{N}$. See Brockwell et al. (1991, Proposition 4.5.3) for a proof of the Lemma.

E ADDITIONAL EXPERIMENTAL RESULTS

In section 5 we described experiments with simulated autoregressive processes. Here, we provide additional plots from these experiments.

E.1 STATISTICAL AND CAUSAL ERRORS

In the main paper we have seen that even in very simple AR models the causal error of an OLS regression estimator can be several times larger than its statistical error. In Figures 9, 10 and 11 we can see that this is also the case for OLS, Lasso and ElasticNet regression and different process orders. All methods can be seen as the solution to an optimization problem, minimizing the empirical statistical error plus some penalty term $\Omega(\hat{a})$, that is, $\sum_{y_i,\hat{y}_i}(y_i - \hat{y}_i)^2 + \lambda\Omega(\hat{a})$, where $\hat{y}_i$ denotes the model prediction with estimated parameters $\hat{a}$ and $\lambda > 0$ the strength of the regularization. For OLS, the penalty term is zero. For Ridge and Lasso the penalty is the l_2 and l_1 norm respectively, i.e. $\Omega(\hat{a}) = \|\hat{a}\|_2$ for Ridge and $\Omega(\hat{a}) = \|\hat{a}\|_1$ for Lasso. For ElasticNet we have $\Omega(\hat{a}) = \mu \cdot \|\hat{a}\|_1 + (1-\mu) \cdot \|\hat{a}\|_2$, where μ is a parameter balancing the l_1 and l_2 penalty.

We used standard grid-search and 5-fold cross-validation to find the optimal regularization strength. For ElasticNet, we additional optimized μ with the grid search. Except for Figures 12, we use 100 training and 1000 test samples. For all experiments, we simulate our processes with noise variance $\sigma^2 = 1$.

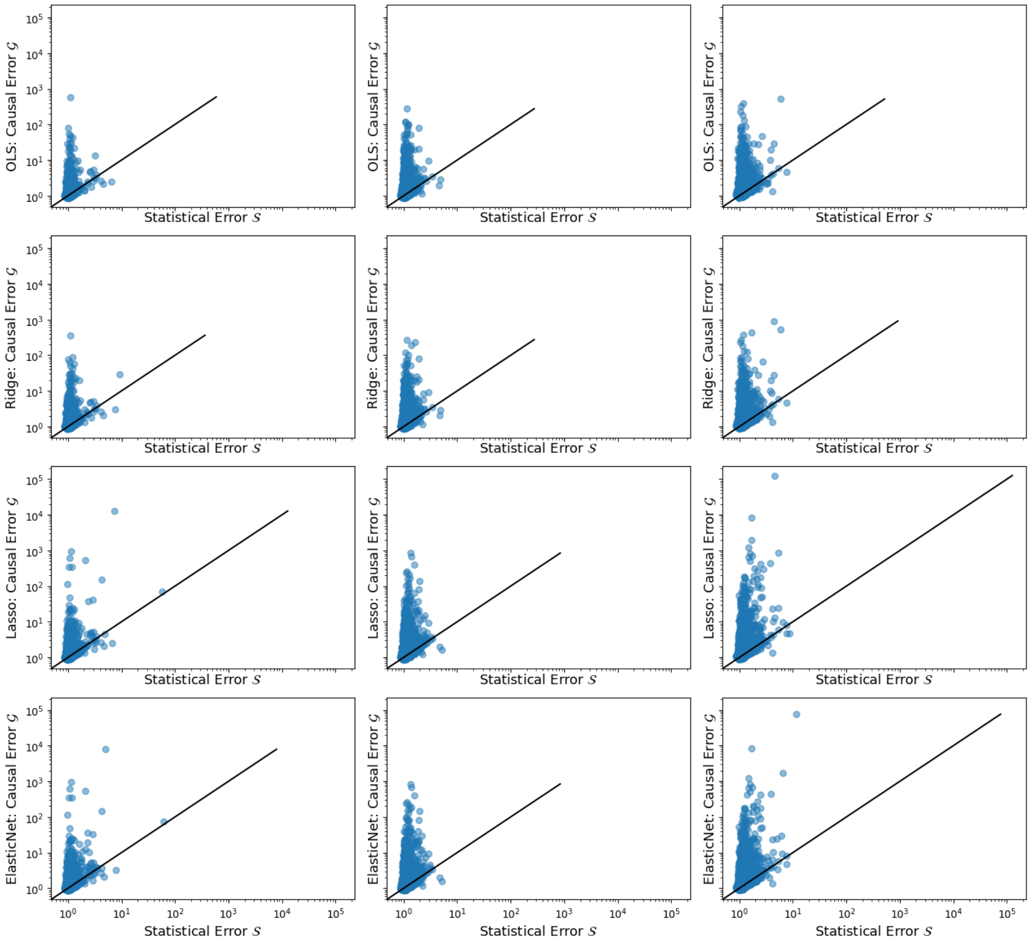


Figure 9: The causal error $\mathcal{G}$ plotted against the statistical error $\mathcal{S}$ for process orders $p = 3, 5, 7$ (from left to right) and estimators OLS, Lasso and ElasticNet (from top to bottom).

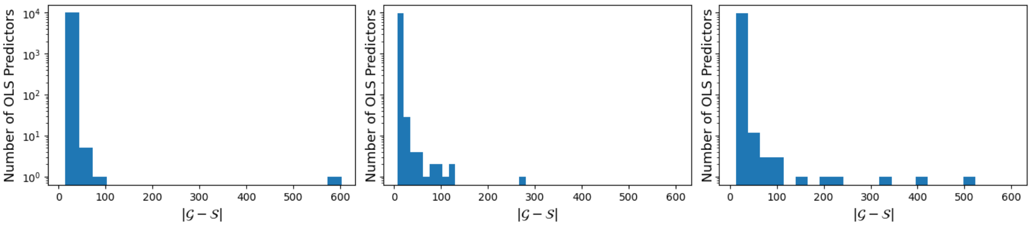


Figure 10: Histogram of the difference $|\mathcal{G} - \mathcal{S}|$ for orders $p = 3, 5, 7$.

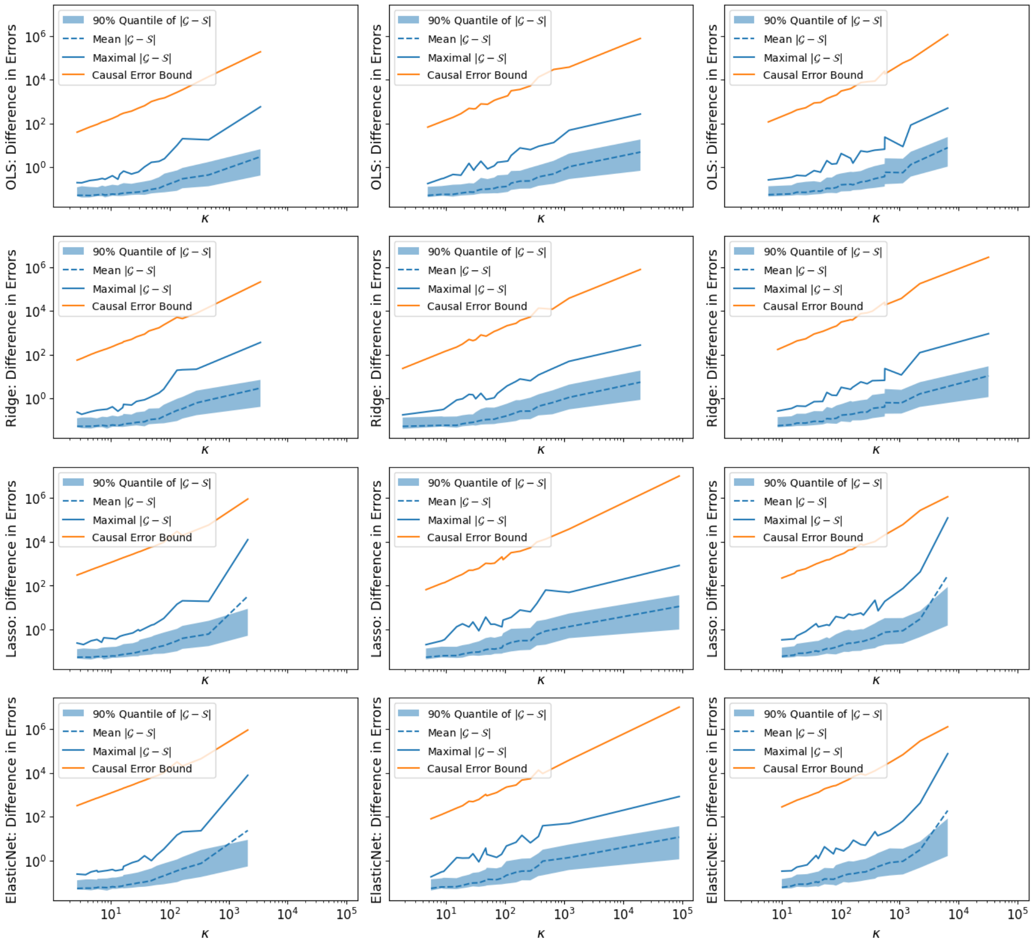


Figure 11: The maximal difference of statistical and causal error $|\mathcal{G} - \mathcal{S}|$ plotted against the condition number of the autocorrelation matrix κ for process orders $p = 3, 5, 7$ (from left to right) and estimators OLS, Lasso and ElasticNet (from top to bottom).

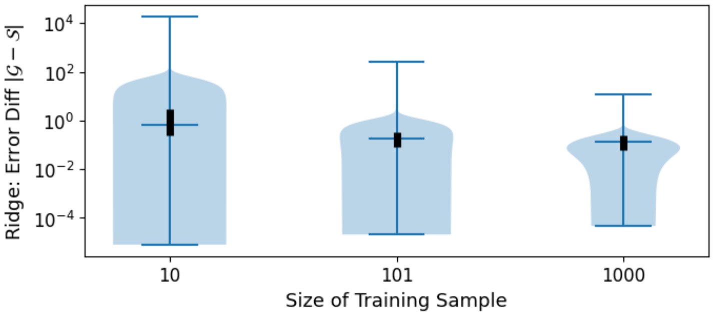


Figure 12: The absolute difference $|\mathcal{G} - \mathcal{S}|$ of causal and statistical error plotted against the sample size for process orders $p = 5$, sample sizes 10, 100, 1000 using Ridge regression. The blue bars mark the 0, 0.5 and 1 quantile and the black block goes from the 0.25 to the 0.75 quantile.

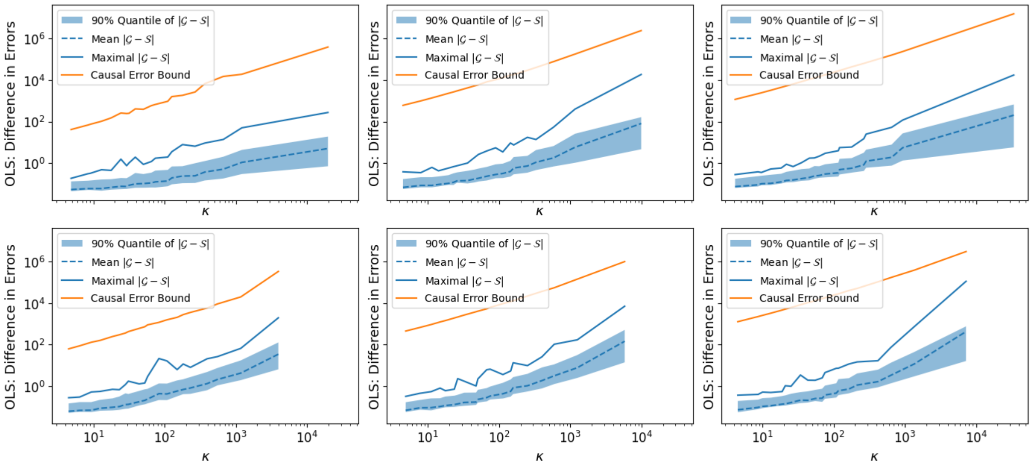


Figure 13: The maximal difference of errors $|\mathcal{G} - \mathcal{S}|$ as well as the generalization bound from Theorem 2 plotted against condition number of the autocorrelation matrix for process order $p = 5$, steps predicted ahead $\omega = 1, 5, 7$ (from left to right). The top row show interventions only on the most recent timestep x_{t-1} where the bottom row shows interventions on all previous timesteps before the prediction.

Dataset	electricity			traffic		
Model	observ.	across-ts interv.	within-ts interv.	observ.	across-ts interv.	within-ts interv.
DeepAR	381.550 ± 21.647	449.781 ± 27.536	375.632 ± 20.851	0.0282 ± 0.0015	0.0288 ± 0.0017	0.0294 ± 0.0018
wavenet	470.691 ± 15.886	799.307 ± 65.722	588.469 ± 39.911	0.0246 ± 0.0003	0.0279 ± 0.0003	0.0299 ± 0.0003
transformer	413.174 ± 31.243	575.946 ± 35.456	407.372 ± 29.073	0.0282 ± 0.0023	0.0312 ± 0.0031	0.0328 ± 0.0033

Table 2: 80% prediction width for observational and interventional forecasts. Averaged over 5 runs with std.

Increasing sample size. As one would expect, in Figure 12, we can see that the absolute difference of the errors decreases for larger training samples. We show this result for the ridge regression estimator. Results for other estimators are similar. The respective means are 13.28, 0.48 and 0.18 from left to right and the standard deviations are 264.54, 4.35 and 0.27, which is hard to read from the plot due to the scale of the outliers.

Violations of causal sufficiency. In Figure 14 we violated the causal sufficiency assumption by introducing a hidden confounder. To this end we draw a two-dimensional AR(1) process by drawing each entry of the parameter matrix A independently and uniformly from $[-2, 2]$ and reject matrices that yield non-stationary processes. We then only use one of the two dimensions as training and test sample. The other one acts as hidden confounder. We also use only the sample of the observed dimension to estimate the autocorrelation of the process, which is the x-axis of the plots in Figure 14.

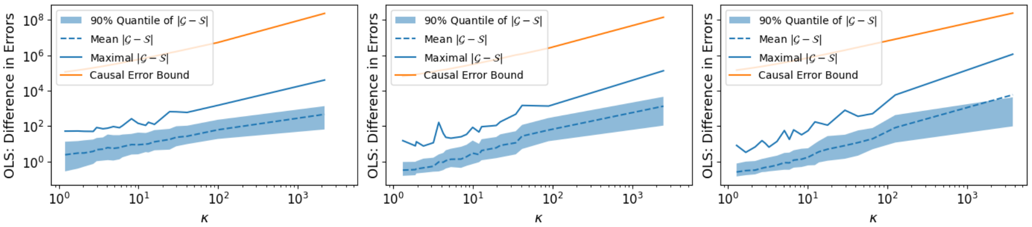


Figure 14: The maximal difference of errors $|\mathcal{G} - \mathcal{S}|$ as well as the generalization bound from Theorem 2 plotted against condition number of the autocorrelation matrix for process order $p = 5$, steps predicted ahead $\omega = 1, 5, 7$ (from left to right).

F ADDITIONAL EXPERIMENTS ON REAL DATA

Following the setup from Section 5, we added two datasets. **Data.** We conduct additional experiments on the traffic dataset that records the occupancy rates of car lanes on freeways in the San Francisco Bay Area (Dua et al. 2017) and the electricity dataset (Dua et al. 2017) that records the electricity consumption of 370 customers hourly.

Results. The additional results on these two datasets in Figure 15 confirm our previous discussion, that the causal disagreement between two models of the same architecture and hyper-parameters can be much higher than their disagreement on the observational distribution. While there are only smaller differences in the statistical risk between the model architectures, their causal disagreement differs more. Wavenet continues to have a high causal disagreement. The disagreement can be viewed as an uncertainty measure over the model training. An additional uncertainty measure can be derived from the forecasts themselves which represent a distribution over future time-series continuations. Table 2 reports the average width that captures 80% of the samples drawn from the forecast distribution. We see that this it is yields similar results to those of Figure 15: The prediction width is wider for the interventional distributions and varies across datasets and model architectures.

The causal disagreement can be high for some models which implies a high causal risk. This cautions against the use of statistical deep learning models to forecast what will happen under interventions. The difference we observe in causal disagreement across models motivates further development of specific model architectures suitable for causal forecasting. For existing models, the uncertainty measure considering the width of the prediction interval can be an indicator for causal risk.

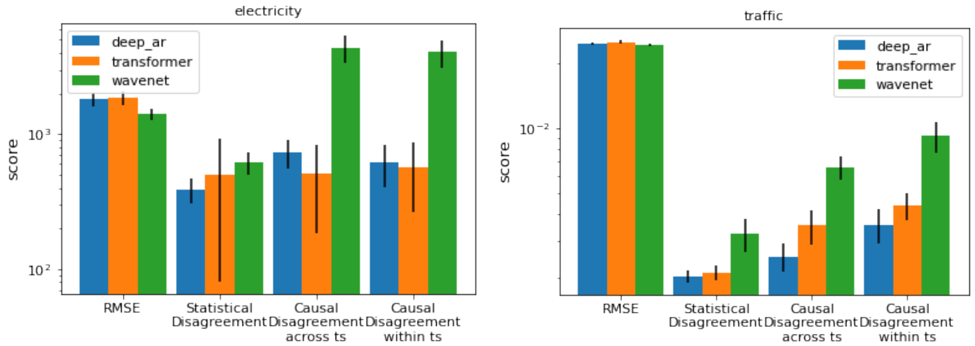


Figure 15: Results of the evaluation of three different deep neural network architectures on the electricity dataset on the top row and the traffic dataset on the bottom row. The RMSE is computed comparing prediction on the observational data against the ground truth and comparing the predictions across two models of the same architecture on the observational and interventional distributions. The results are averaged over 5 runs of training and evaluation and include standard deviation.

REFERENCES

Alexandrov, Alexander, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang (2019). *GluonTS: Probabilistic Time Series Models in Python*. arXiv: `1906.05264 [cs.LG]`.

Basu, Sumanta and George Michailidis (2015). “Regularized estimation in sparse high-dimensional time series models”. In: *The Annals of Statistics* 43.4, pp. 1535–1567.

Ben-David, Shai, John Blitzer, Koby Crammer, Fernando Pereira, et al. (2007). “Analysis of representations for domain adaptation”. In: *Advances in neural information processing systems* 19, p. 137.

Bennett, Andrew, Nathan Kallus, Lihong Li, and Ali Mousavi (2021). “Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1999–2007.

Bica, Ioana, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar (2020). “Estimating counterfactual treatment outcomes over time through adversarially balanced representations”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: `https://openreview.net/forum?id=BJg866NFvB`.

Brand, Louis (1964). “The companion matrix and its properties”. In: *The American Mathematical Monthly* 71.6, pp. 629–634.

Brockwell, Peter J, Richard A Davis, and Stephen E Fienberg (1991). *Time series: theory and methods: theory and methods*. Springer Science & Business Media.

Cauchy, Augustin Louis (1815). “M ’e memory on functions which can obtain only two equal values é and of opposite signs as a result of the transpositions op ’e r é es between the variables that they contain”. In: *Journal de l’Ecole polytechnique* 10.17, pp. 29–112.

Chaugule, Prasad, Mrinal Kumar, Nutan Limaye, Chandra Kanta Mohapatra, Adrian She, and Srikanth Srinivasan (2019). “Schur Polynomials do not have small formulas if the Determinant doesn’t!” In: *arXiv preprint arXiv:1911.12520*.

Davison, Edward (1976). “The robust control of a servomechanism problem for linear time-invariant multivariable systems”. In: *IEEE transactions on Automatic Control* 21.1, pp. 25–34.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. URL: `http://archive.ics.uci.edu/ml`.

Faloutsos, Christos, Jan Gasthaus, Tim Januschowski, and Yuyang Wang (2018). “Forecasting Big Time Series: Old and New”. In: 11.12. ISSN: 2150-8097.

Fisk, Steve (2005). “A very short proof of Cauchy’s interlace theorem for eigenvalues of Hermitian matrices”. In: *arXiv preprint math/0502408*.

Gasthaus, Jan, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski (Apr. 2019). “Probabilistic Forecasting with Spline Quantile Function RNNs”. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1901–1910.

Grabowski, Wojciech and Aleksander Welfe (2020). “The Tobit cointegrated vector autoregressive model: An application to the currency market”. In: *Economic Modelling* 89, pp. 88–100.

Hatt, Tobias and Stefan Feuerriegel (2021). *Sequential Deconfounding for Causal Inference with Unobserved Confounders*. arXiv: `2104.09323 [stat.ME]`.

Hill, Jennifer and Jerome P Reiter (2006). “Interval estimation for treatment effects using propensity score matching”. In: *Statistics in medicine* 25.13, pp. 2230–2256.

Horn, Roger A and Charles R Johnson (2012). *Matrix analysis*. Cambridge university press.

Hyvärinen, Aapo, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer (2010). “Estimation of a structural vector autoregression model using non-gaussianity.” In: *Journal of Machine Learning Research* 11.5.

Jacobi, Carl Gustav Jakob (1841). “De functionibus alternantibus earumque divisione per productum e differentiis elementorum conflatum.” In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 1841.22, pp. 360–371.

Januschowski, Tim, Jan Gasthaus, Yuyang Wang, David Salinas, Valentin Flunkert, Michael Bohlke-Schneider, and Laurent Callot (2020). “Criteria for classifying forecasting methods”. In: *International Journal of Forecasting* 36.1. M4 Competition, pp. 167–177. ISSN: 0169-2070.

Janzing, D., L. Minorics, and P. Bloebaum (Aug. 2020). “Feature relevance quantification in explainable AI: A causal problem”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. Online: PMLR, pp. 2907–2916.

Johansson, Fredrik D, Uri Shalit, Nathan Kallus, and David Sontag (2020). “Generalization bounds and representation learning for estimation of potential outcomes and causal effects”. In: *arXiv preprint arXiv:2001.07426*.

Lim, Bryan, Ahmed Alaa, and Mihaela van der Schaar (2018). “Forecasting Treatment Responses over Time Using Recurrent Marginal Structural Networks”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., pp. 7494–7504.

Lundberg, S. and S. Lee (2017). “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 4765–4774.

Lütkepohl, Helmut (2009). “Econometric analysis with vector autoregressive models”. In: *Handbook of computational econometrics*, pp. 281–319.

– (2013). “Vector autoregressive models”. In: *Handbook of Research Methods and Applications in Empirical Macroeconomics*. Edward Elgar Publishing.

Macdonald, Ian Grant (1998). *Symmetric functions and Hall polynomials*. Oxford university press.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilis Assimakopoulos (June 2018). “The M4 Competition: Results, findings, conclusion and way forward”. In: *International Journal of Forecasting* 34. DOI: `10.1016/j.ijforecast.2018.06.001`.

Mansour, Yishay, Mehryar Mohri, and Afshin Rostamizadeh (2009). “Domain adaptation: Learning bounds and algorithms”. In: *arXiv preprint arXiv:0902.3430*.

McDonald, Daniel J, Cosma Rohilla Shalizi, and Mark Schervish (2017). “Nonparametric risk bounds for time-series forecasting”. In: *The Journal of Machine Learning Research* 18.1, pp. 1044–1083.

Meir, Ron (2000). “Nonparametric time series prediction through adaptive model selection”. In: *Machine learning* 39.1, pp. 5–34.

Melnyk, Igor and Arindam Banerjee (2016). “Estimating structured vector autoregressive models”. In: *International Conference on Machine Learning*. PMLR, pp. 830–839.

El-Mikkawy, Moawwad EA (2003). “Explicit inverse of a generalized Vandermonde matrix”. In: *Applied mathematics and computation* 146.2-3, pp. 643–651.

Mohri, Mehryar and Afshin Rostamizadeh (2009). “Rademacher Complexity Bounds for Non-I.I.D. Processes”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc.

Mokkadem, Abdelkader (1988). “Mixing properties of ARMA processes”. In: *Stochastic processes and their applications* 29.2, pp. 309–315.

Molnar, C. (2019). *Interpretable Machine Learning*. Molnar, C. URL: `https://christophm.github.io/interpretable-ml-book/`.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). *WaveNet: A Generative Model for Raw Audio*. arXiv: `1609.03499 [cs.SD]`.

Pearl, Judea (2009). *Causality*. Cambridge university press.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference*. The MIT Press.

Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski (2020). "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". In: *International Journal of Forecasting* 36.3, pp. 1181–1191. ISSN: 0169-2070.

Shalit, Uri, Fredrik D Johansson, and David Sontag (2017). "Estimating individual treatment effect: generalization bounds and algorithms". In: *International Conference on Machine Learning*. PMLR, pp. 3076–3085.

Shi, Claudia, David M Blei, and Victor Veitch (2019). "Adapting neural networks for the estimation of treatment effects". In: *arXiv preprint arXiv:1906.02120*.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction, and Search*. New York, NY: Springer-Verlag.

Sugiyama, Masashi and Motoaki Kawanabe (2012). *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press. ISBN: 0262017091.

Valdés-Sosa, Pedro A, Jose M Sánchez-Bornot, Agustín Lage-Castellanos, Mayrim Vega-Hernández, Jorge Bosch-Bayard, Lester Melie-García, and Erick Canales-Rodríguez (2005). "Estimating brain functional connectivity with sparse multivariate autoregression". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1457, pp. 969–981.

Varga, Richard S (2010). *Geršgorin and his circles*. Vol. 36. Springer Science & Business Media.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.

Wang, Jiaxuan, Jenna Wiens, and Scott Lundberg (2020). *Shapley Flow: A Graph-based Approach to Interpreting Model Predictions*. arxiv:2010.14592.

Yu, Bin (1994). "Rates of convergence for empirical processes of stationary mixing sequences". In: *The Annals of Probability*, pp. 94–116.

Zivot, Eric and Jiahui Wang (2006). "Vector autoregressive models for multivariate time series". In: *Modeling Financial Time Series with S-Plus®*, pp. 385–429.

9
Interpolation and regularization for causal learning

Interpolation and Regularization for Causal Learning

Leena Chennuru Vankadara*
University of Tübingen

Luca Rendsburg*
University of Tübingen

Ulrike von Luxburg
University of Tübingen

Debarghya Ghoshdastidar
Technical University of Munich

Abstract

Recent work shows that in complex model classes, interpolators can achieve statistical generalization and even be optimal for statistical learning. However, despite increasing interest in learning models with good causal properties, there is no understanding of whether such interpolators can also achieve *causal generalization*. To address this gap, we study causal learning from observational data through the lens of interpolation and its counterpart—regularization. Under a simple linear causal model, we derive precise asymptotics for the causal risk of the min-norm interpolator and ridge regressors in the high-dimensional regime. We find a large range of behavior that can be precisely characterized by a new measure of *confounding strength*. When confounding strength is positive, which holds under independent causal mechanisms—a standard assumption in causal learning—we find that interpolators cannot be optimal. Indeed, causal learning requires stronger regularization than statistical learning. Beyond this assumption, when confounding is negative, we observe a phenomenon of self-induced regularization due to positive alignment between statistical and causal signals. Here, causal learning requires weaker regularization than statistical learning, interpolators can be optimal, and optimal regularization can even be negative.

1 Introduction

We consider the problem of learning the causal influence of multivariate covariates $x \in \mathbb{R}^d$ on a scalar target variable $y \in \mathbb{R}$ purely from observational data and under the presence of hidden confounders. Formally, given finite samples $\{(x_i, y_i)\}_{i=1}^n$ drawn independently and identically (i.i.d) from the joint *observational distribution* $p(x, y) = p(x)p(y|x)$, the goal of causal learning is to predict the effects on the target variable y under interventions on the covariates x. In other words, the goal is to learn a predictive model that minimizes the expected loss on a random draw from the *interventional distribution* $p_{do}(x, y) = p(x)p(y|do(x))$, which can be different from the observational distribution.

Recently, Janzing (2019) established a close analogy between statistical and causal learning (albeit under a highly constructed confounded model). As a consequence, Janzing (2019) suggested that standard statistical learning-theoretic techniques (such as norm-based regularization) may also help learn good causal models. However, the classical statistical principles of bias-variance trade-off have been challenged in recent years by highly complex classes of models that are trained to interpolate the data and yet achieve remarkable generalization properties across a broad range of problem domains (Zhang et al., 2021). A large volume of recent work suggests that interpolation can be compatible with and may even be necessary to achieve optimal statistical generalization in the high-dimensional regime (Belkin et al., 2018; Belkin et al., 2019a; Liang et al., 2020; Feldman, 2020). Despite the surge in interest, causal properties of such interpolators have not yet been explored. In this work, we

[*] denotes equal contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

consider a simple linear causal model in the high-dimensional regime ($n, d \to \infty, d/n \in \mathcal{O}(1)$) and ask: can interpolators achieve good causal generalization?

1.1 Motivation and Related Work

Resemblance between statistical and causal generalization Causal learning can be regarded as an instance of the general problem of learning under distribution shifts, where the training (observational) distribution is shifted from the test (interventional) distribution. In the framework of out-of-distribution generalization, an interesting proposition for causal learning arises from the following high-level idea. Observing small sample sizes may induce a similar bias as distribution shifts. Therefore, techniques for learning models with good *out-of-sample* generalization (such as regularization) may also help to learn models with good *out-of-distribution* generalization and vice-versa. The literature provides plentiful evidence to support this general principle for different classes of distribution shifts. For instance, under a broad class of distribution shifts, distributionally robust optimization is equivalent to norm-based regularization (Xu et al., 2009; Shafieezadeh Abadeh et al., 2015; Gao et al., 2017; Shafieezadeh-Abadeh et al., 2019; Blanchet et al., 2019; Kuhn et al., 2019). Analogously, distributionally robust optimization techniques are also employed for statistical learning under limited samples (Zhu et al., 2020). Particularly relevant to our work is Janzing (2019), which formally establishes a close analogy between "generalizing from *empirical to observational distributions*" and "generalizing from *observational to interventional distributions*" under a highly constructed confounding model. This analogy suggests that standard norm-based regularization such as lasso or ridge, typically used for statistical learning, may also help learn better causal models.

Interpolation can be compatible with statistical learning Explicit norm-based regularization techniques were initially motivated by the classical learning theory principle of bias-variance trade-off, which is characterized by a U-shaped generalization curve. This principle recommends to avoid interpolation and instead suggests to balance data fitting with the complexity of the hypothesis class. Recently, however, these classical principles have been challenged by deep learning models. Despite being highly complex with the ability to fit even random labels and often trained to interpolate the training data, they achieve state-of-the-art out-of-sample generalization across many domains (Zhang et al., 2021). A partial explanation is provided by the *double-descent* phenomenon (Belkin et al., 2019b; Belkin, 2021). Extending the generalization curve beyond the interpolation threshold reveals two regimes: the classical U-curve in the *underparameterized* regime and a decreasing curve in the *overparameterized* regime. This behaviour is not limited to deep neural networks, but extends to other settings such as random feature models and random forests (Belkin et al., 2019b; Hastie et al., 2022; Mei et al., 2021). Follow-up work suggests that in the overparameterized regime, interpolators can indeed achieve low statistical risk (Belkin et al., 2019a; Liang et al., 2020; Bartlett et al., 2020; Tsigler et al., 2020; Muthukumar et al., 2020).

Is interpolation compatible with causal learning? On account of the parallels between statistical (out-of-sample) and causal (out-of-distribution) learning, it is therefore natural to ask: *can interpolators also learn good causal models?* One line of empirical work suggests that naively applying distributionally robust learning techniques such as importance reweighting or distributionally robust optimization approaches (which are equivalent to certain forms of regularization) offers vanishing benefits over empirical risk minimization in overparameterized model classes (Byrd et al., 2019; Sagawa et al., 2020; Gulrajani et al., 2021). However, there is also empirical evidence that suggests that augmenting such techniques with additional explicit norm-based regularization may help to learn distributionally robust models in the overparameterized regime (Sagawa et al., 2020; Donhauser et al., 2021). In the context of causal learning, it has been suggested that explicit regularization can be beneficial and might even need to be stronger than for statistical learning (Janzing, 2019; Vankadara et al., 2021). Existing work, therefore, remains unclear about the role of explicit regularization in causal learning, or correspondingly, whether interpolation is compatible with causal learning. In this work, we take a theoretical approach to systematically address these questions.

1.2 Our Contributions

We provide a first analysis of causal generalization from observational data in the modern, overparameterized and interpolating regime under a simple linear causal model. Specifically, we consider the interpolating minimum l_2 norm least-squares estimator and the family of regularized ridge regression estimators in the proportional asymptotic regime. We seek answers to the following questions: is there a regime where the optimal causal regularization parameter is 0, that is, can we observe *benign causal overfitting*? Furthermore, if the optimal causal regularization parameter is positive, how strongly do we need to regularize? How does the optimal causal regularization compare to the optimal statistical regularization? While our analysis is exhaustive, we emphasize the results under the assumption of independent causal mechanisms (Janzing et al., 2010), a standard assumption in causal learning.

- **Precise asymptotics of the causal risk (Section 3).** We provide precise asymptotics of the *causal risk* of the ridge regression estimator as well as the minimum l_2 norm interpolator in the high-dimensional setting: $n, d \to \infty, d/n \to \gamma \in (0, \infty)$. Our results confirm that, similar to the statistical setting, the causal generalization curve of the min-norm estimator exhibits the double-descent phenomenon. This is because the variance term diverges at the interpolation threshold and is decreasing in the overparameterized regime ($\gamma > 1$).
- **A measure of confounding strength ζ (Section 2.1).** We introduce a new measure of *confounding strength* ζ that quantifies the relative contribution of the "confounding signal" to the "causal signal". It can be interpreted as the strength of the distribution shift between the observational and interventional distributions. While ζ can take any real value in general, it is restricted to $[0, 1]$ under the assumption of independent causal mechanisms. There, it induces a strict, model-independent ordering of all causal models that entail the same observational distribution.
- **Benign causal overfitting (Section 4).** When the causal signal dominates the statistical signal ($\zeta < 0$), we observe a phenomenon of self-induced regularization due to the confounding signal. As a consequence, the optimal causal regularization can be 0 or negative even if the optimal statistical regularization is strictly positive. Under the assumption of independent causal mechanisms, however, we show that there is no benign causal overfitting. This is in contrast to benign statistical overfitting, which can occur in the highly underparameterized regime ($\gamma \to 0$).
- **Optimal causal vs. statistical regularization (Section 5).** We show that causal learning requires weaker regularization than statistical learning when the confounding strength ζ is negative. However, when $\zeta > 0$ and in particular under the principle of independent causal mechanisms, we show that causal learning requires stronger regularization than statistical learning. More specifically, the optimal causal regularization is strictly increasing in confounding strength.

2 Problem Setup

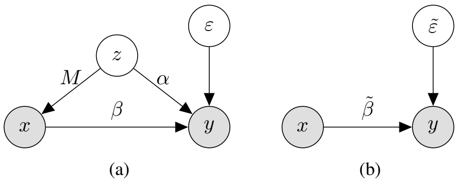


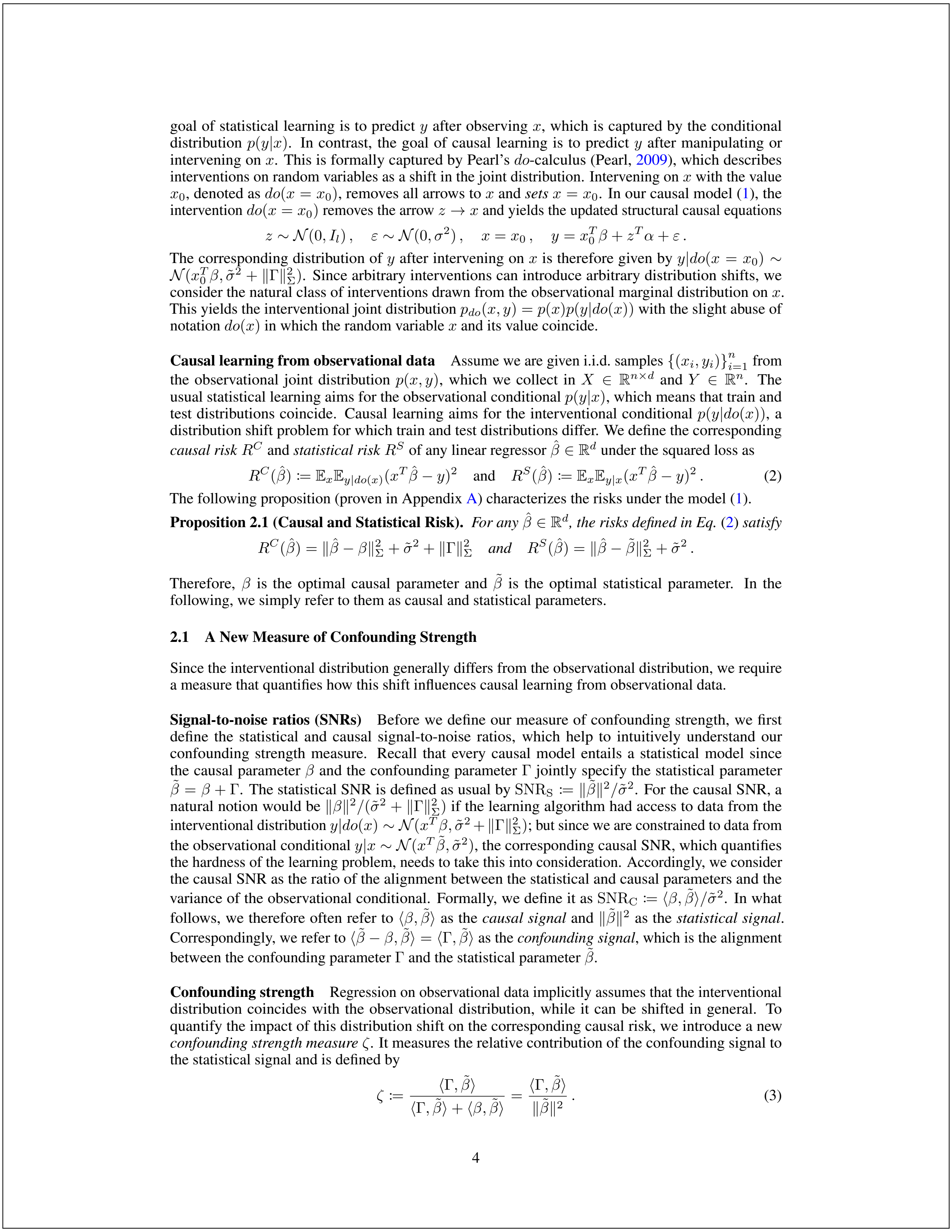
Figure 1: (*a*) Graphical model of the causal model defined in (1). (*b*) The usual statistical model. In both figures, observed random variables are shaded and unobserved variables are white.

We consider a linear causal model with parameters $M \in \mathbb{R}^{d \times l}$, $\alpha \in \mathbb{R}^l, \beta \in \mathbb{R}^d$ with $l \geq d$ and $\sigma^2 > 0$ described via the *structural equations*

$$z \sim \mathcal{N}(0, I_l)\,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,, \quad x = Mz\,, \quad y = x^T\beta + z^T\alpha + \varepsilon\,. \tag{1}$$

The covariates $x \in \mathbb{R}^d$ and the target $y \in \mathbb{R}$ are *confounded* through $z \in \mathbb{R}^l$, which follows a standard normal distribution. This structure implies that $\mathbb{E}x = 0$ and the covariance of x is $\Sigma := \operatorname{Cov} x = MM^T$. A graphical representation of this causal model is given in Figure 1a. The observational joint distribution of this causal model is given by $p(x, y) = p(x)p(y|x)$, where $x \sim \mathcal{N}(0, \Sigma)$ and $y|x \sim \mathcal{N}(x^T\tilde{\beta}, \tilde{\sigma}^2)$. Here, the statistical parameter $\tilde{\beta} := \beta + \Gamma$ consists of the causal parameter β and a confounding parameter $\Gamma := M^{+T}\alpha$, where M^{+T} is shorthand for $(M^+)^T$ and M^+ denotes the Moore-Penrose inverse of M. The statistical noise is given by $\tilde{\sigma}^2 := \sigma^2 + \|\alpha\|^2 - \|\Gamma\|_\Sigma^2$, where $\|x\|_\Sigma^2 := x^T\Sigma x$ denotes the generalized norm. [1] Note that the observational distribution alone cannot distinguish the causal model from the one in Figure 1b. The

[1] Note that $\|\alpha\|^2 - \|\Gamma\|_\Sigma^2 = \|\alpha\|_{I-M^+M}^2 \geq 0$, where $I - M^+M$ is the orthogonal projection onto $\ker M$.



goal of statistical learning is to predict y after observing x, which is captured by the conditional distribution $p(y|x)$. In contrast, the goal of causal learning is to predict y after manipulating or intervening on x. This is formally captured by Pearl's do-calculus (Pearl, 2009), which describes interventions on random variables as a shift in the joint distribution. Intervening on x with the value x_0, denoted as $do(x = x_0)$, removes all arrows to x and *sets* $x = x_0$. In our causal model (1), the intervention $do(x = x_0)$ removes the arrow $z \to x$ and yields the updated structural causal equations

$$z \sim \mathcal{N}(0, I_l)\,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,, \quad x = x_0\,, \quad y = x_0^T\beta + z^T\alpha + \varepsilon\,.$$

The corresponding distribution of y after intervening on x is therefore given by $y|do(x = x_0) \sim \mathcal{N}(x_0^T\beta, \tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2)$. Since arbitrary interventions can introduce arbitrary distribution shifts, we consider the natural class of interventions drawn from the observational marginal distribution on x. This yields the interventional joint distribution $p_{do}(x, y) = p(x)p(y|do(x))$ with the slight abuse of notation $do(x)$ in which the random variable x and its value coincide.

Causal learning from observational data Assume we are given i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ from the observational joint distribution $p(x, y)$, which we collect in $X \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^n$. The usual statistical learning aims for the observational conditional $p(y|x)$, which means that train and test distributions coincide. Causal learning aims for the interventional conditional $p(y|do(x))$, a distribution shift problem for which train and test distributions differ. We define the corresponding *causal risk* R^C and *statistical risk* R^S of any linear regressor $\hat{\beta} \in \mathbb{R}^d$ under the squared loss as

$$R^C(\hat{\beta}) := \mathbb{E}_x\mathbb{E}_{y|do(x)}(x^T\hat{\beta} - y)^2 \quad \text{and} \quad R^S(\hat{\beta}) := \mathbb{E}_x\mathbb{E}_{y|x}(x^T\hat{\beta} - y)^2\,. \tag{2}$$

The following proposition (proven in Appendix A) characterizes the risks under the model (1).

Proposition 2.1 (Causal and Statistical Risk). *For any $\hat{\beta} \in \mathbb{R}^d$, the risks defined in Eq. (2) satisfy*

$$R^C(\hat{\beta}) = \|\hat{\beta} - \beta\|_\Sigma^2 + \tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2 \quad \textit{and} \quad R^S(\hat{\beta}) = \|\hat{\beta} - \tilde{\beta}\|_\Sigma^2 + \tilde{\sigma}^2\,.$$

Therefore, β is the optimal causal parameter and $\tilde{\beta}$ is the optimal statistical parameter. In the following, we simply refer to them as causal and statistical parameters.

2.1 A New Measure of Confounding Strength

Since the interventional distribution generally differs from the observational distribution, we require a measure that quantifies how this shift influences causal learning from observational data.

Signal-to-noise ratios (SNRs) Before we define our measure of confounding strength, we first define the statistical and causal signal-to-noise ratios, which help to intuitively understand our confounding strength measure. Recall that every causal model entails a statistical model since the causal parameter β and the confounding parameter Γ jointly specify the statistical parameter $\tilde{\beta} = \beta + \Gamma$. The statistical SNR is defined as usual by $\mathrm{SNR_S} := \|\tilde{\beta}\|^2/\tilde{\sigma}^2$. For the causal SNR, a natural notion would be $\|\beta\|^2/(\tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2)$ if the learning algorithm had access to data from the interventional distribution $y|do(x) \sim \mathcal{N}(x^T\beta, \tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2)$; but since we are constrained to data from the observational conditional $y|x \sim \mathcal{N}(x^T\tilde{\beta}, \tilde{\sigma}^2)$, the corresponding causal SNR, which quantifies the hardness of the learning problem, needs to take this into consideration. Accordingly, we consider the causal SNR as the ratio of the alignment between the statistical and causal parameters and the variance of the observational conditional. Formally, we define it as $\mathrm{SNR_C} := \langle\beta, \tilde{\beta}\rangle/\tilde{\sigma}^2$. In what follows, we therefore often refer to $\langle\beta, \tilde{\beta}\rangle$ as the *causal signal* and $\|\tilde{\beta}\|^2$ as the *statistical signal*. Correspondingly, we refer to $\langle\tilde{\beta} - \beta, \tilde{\beta}\rangle = \langle\Gamma, \tilde{\beta}\rangle$ as the *confounding signal*, which is the alignment between the confounding parameter Γ and the statistical parameter $\tilde{\beta}$.

Confounding strength Regression on observational data implicitly assumes that the interventional distribution coincides with the observational distribution, while it can be shifted in general. To quantify the impact of this distribution shift on the corresponding causal risk, we introduce a new *confounding strength measure* ζ. It measures the relative contribution of the confounding signal to the statistical signal and is defined by

$$\zeta := \frac{\langle\Gamma, \tilde{\beta}\rangle}{\langle\Gamma, \tilde{\beta}\rangle + \langle\beta, \tilde{\beta}\rangle} = \frac{\langle\Gamma, \tilde{\beta}\rangle}{\|\tilde{\beta}\|^2}\,. \tag{3}$$

Other notions of confounding strength are possible, but we will see later that this definition is well-suited to capture the shift strength for causal learning from observational data. Without further restrictions, ζ can take any value in $\mathbb{R}$. This measure divides the causal models into the following three regimes, depending on the relationship between causal and statistical signal:

- $\zeta \geq 1$: the causal signal $\langle \beta, \tilde{\beta} \rangle$ is non-positive, which implies that causal and statistical parameters are orthogonal or negatively aligned. Statistical learning is adversarial to causal learning.
- $0 < \zeta < 1$: causal and statistical parameters are positively aligned but the causal signal is weaker than the statistical signal $\|\tilde{\beta}\|^2$, for example $\beta = \tilde{\beta}/2$.
- $\zeta \leq 0$: the causal signal dominates the statistical signal, for example $\beta = 2\tilde{\beta}$.

The SNRs are related to the confounding strength measure via $\mathrm{SNR_C} = (1 - \zeta)\, \mathrm{SNR_S}$. In particular, the causal signal decreases as the confounding strength increases.

The regime $0 \leq \zeta \leq 1$ is practically most relevant Causal learning often requires strong assumptions because causal models cannot be uniquely identified by their observational distribution. A standard assumption is the principle of independent causal mechanisms (ICM) (Janzing et al., 2010; Lemeire et al., 2013; Peters et al., 2017), which informally asserts that causal mechanisms share no information. In our causal model (1), a corresponding assumption could be that the causal mechanisms α and β are drawn from rotationally invariant distributions. This implies that $\langle \beta, \Gamma \rangle \to 0$ as $d \to \infty$, which in turn falls in the regime $0 \leq \zeta \leq 1$. While our following analysis covers all possible causal models, we pay special attention to this regime because it might be of highest practical relevance. Note that for $\langle \beta, \Gamma \rangle = 0$, our measure of confounding strength coincides with the measure $\zeta' = \|\Gamma\|^2/(\|\Gamma\|^2 + \|\beta\|^2)$ introduced by Janzing et al. (2017). It measures the relative contribution of causal and confounding signal in terms of lengths rather than inner products.

3 Causal and Statistical Risk of High-Dimensional Regression Models

Causal learning is extremely challenging, because it requires scarcely available interventional data or has to rely on other information such as exogenous (Rothenhäusler et al., 2021) or instrumental variables (Angrist et al., 1991). In our setting where only observational data are available, causal learning requires additional model assumptions. One such approach has been followed by the Concorr method (Janzing, 2019) which leverages the ICM assumption to make an improved choice of regularization parameter under a linear regression model. To fully characterize the effect of regularization on causal generalization, we consider two estimators for learning causal models from observational data $(X, Y) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$: the min-norm interpolator and ridge regressors. The *min-norm interpolator* is the minimum l_2 norm solution to the least squares regression problem

$$\hat{\beta}_0(X, Y) := \arg\min\{\|\hat{\beta}\|_2 : \hat{\beta} \in \arg\min_{\hat{\beta} \in \mathbb{R}^d} \|Y - X\hat{\beta}\|^2\}. \tag{4}$$

A closed form is given by $\hat{\beta}_0(X, Y) = (X^T X)^+ X^T Y$. For $\lambda > 0$, the *ridge regressor* solves

$$\hat{\beta}_\lambda(X, Y) := \arg\min_{\hat{\beta} \in \mathbb{R}^d} \frac{1}{n}\|Y - X\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|^2\,, \tag{5}$$

which has the explicit solution $\hat{\beta}_\lambda(X, Y) = (X^T X + n\lambda I_d)^{-1} X^T Y$. The min-norm interpolator can be obtained as a limiting case from the ridge regression solution via $\hat{\beta}_0(X, Y) = \lim_{\lambda \to 0^+} \hat{\beta}_\lambda(X, Y)$. Whenever it is clear from the context, we drop the dependence of the predictors on X and Y.

Before proceeding with the analysis, we motivate the idea that appropriate regularization can help to learn causal models from purely observational data. To this end, we compare regularization chosen by statistical cross validation to regularization based on an *interventional validation set* in Figure 2. Since cross validation implicitly assumes that there is no confounding, it is close to Bayes optimal for $\zeta = 0$ when $n \gg d$. However, as confounding increases, it falls behind regularization based on the interventional validation set. The latter even yields Bayes optimal risk again in the purely confounded setting $\zeta = 1$, where the lack of causal signal ($\beta = 0$) is encoded by infinite regularization. While we might not have access to an interventional validation set in practice, our theory will show that knowledge of confounding strength is sufficient for choosing appropriate regularization. Finally, we want to caution that even though regularization can help, it does not remove the hardness of causal learning. Reliable causal inference still requires stronger assumptions or additional data.

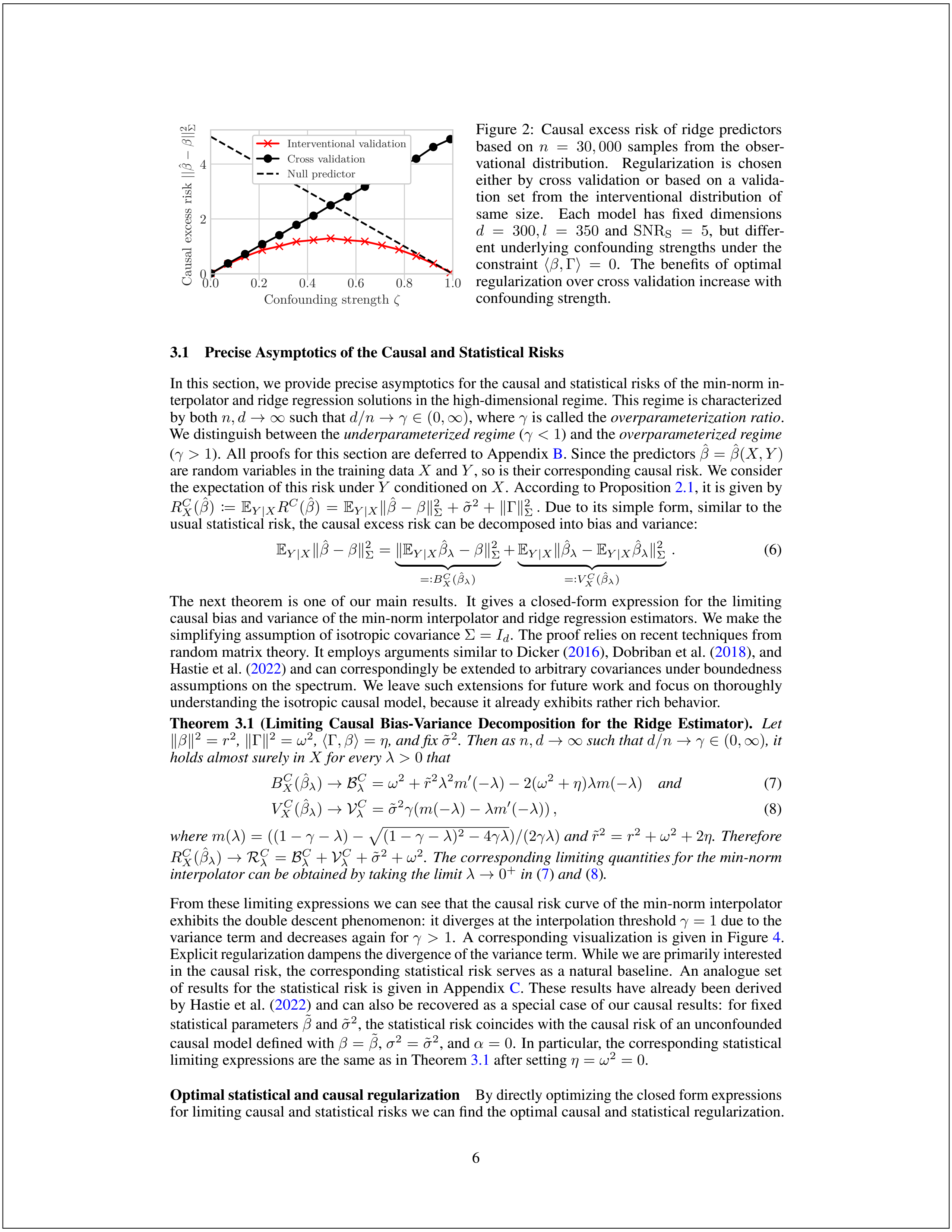

Causal excess risk $\|\hat{\beta}-\beta\|_\Sigma^2$
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0
Confounding strength ζ
Interventional validation
Cross validation
Null predictor

Figure 2: Causal excess risk of ridge predictors based on $n = 30,000$ samples from the observational distribution. Regularization is chosen either by cross validation or based on a validation set from the interventional distribution of same size. Each model has fixed dimensions $d = 300, l = 350$ and $\mathrm{SNR}_\mathrm{S} = 5$, but different underlying confounding strengths under the constraint $\langle \beta, \Gamma \rangle = 0$. The benefits of optimal regularization over cross validation increase with confounding strength.

3.1 Precise Asymptotics of the Causal and Statistical Risks

In this section, we provide precise asymptotics for the causal and statistical risks of the min-norm interpolator and ridge regression solutions in the high-dimensional regime. This regime is characterized by both $n, d \to \infty$ such that $d/n \to \gamma \in (0, \infty)$, where γ is called the *overparameterization ratio*. We distinguish between the *underparameterized regime* ($\gamma < 1$) and the *overparameterized regime* ($\gamma > 1$). All proofs for this section are deferred to Appendix B. Since the predictors $\hat{\beta} = \hat{\beta}(X, Y)$ are random variables in the training data X and Y, so is their corresponding causal risk. We consider the expectation of this risk under Y conditioned on X. According to Proposition 2.1, it is given by $R_X^C(\hat{\beta}) := \mathbb{E}_{Y|X} R^C(\hat{\beta}) = \mathbb{E}_{Y|X} \|\hat{\beta} - \beta\|_\Sigma^2 + \tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2$. Due to its simple form, similar to the usual statistical risk, the causal excess risk can be decomposed into bias and variance:

$$\mathbb{E}_{Y|X} \|\hat{\beta} - \beta\|_\Sigma^2 = \underbrace{\|\mathbb{E}_{Y|X} \hat{\beta}_\lambda - \beta\|_\Sigma^2}_{=: B_X^C(\hat{\beta}_\lambda)} + \underbrace{\mathbb{E}_{Y|X} \|\hat{\beta}_\lambda - \mathbb{E}_{Y|X} \hat{\beta}_\lambda\|_\Sigma^2}_{=: V_X^C(\hat{\beta}_\lambda)} . \tag{6}$$

The next theorem is one of our main results. It gives a closed-form expression for the limiting causal bias and variance of the min-norm interpolator and ridge regression estimators. We make the simplifying assumption of isotropic covariance $\Sigma = I_d$. The proof relies on recent techniques from random matrix theory. It employs arguments similar to Dicker (2016), Dobriban et al. (2018), and Hastie et al. (2022) and can correspondingly be extended to arbitrary covariances under boundedness assumptions on the spectrum. We leave such extensions for future work and focus on thoroughly understanding the isotropic causal model, because it already exhibits rather rich behavior.

Theorem 3.1 (Limiting Causal Bias-Variance Decomposition for the Ridge Estimator). *Let $\|\beta\|^2 = r^2$, $\|\Gamma\|^2 = \omega^2$, $\langle \Gamma, \beta \rangle = \eta$, and fix $\tilde{\sigma}^2$. Then as $n, d \to \infty$ such that $d/n \to \gamma \in (0, \infty)$, it holds almost surely in X for every $\lambda > 0$ that*

$$B_X^C(\hat{\beta}_\lambda) \to \mathcal{B}_\lambda^C = \omega^2 + \tilde{r}^2 \lambda^2 m'(-\lambda) - 2(\omega^2 + \eta)\lambda m(-\lambda) \quad \textit{and} \tag{7}$$

$$V_X^C(\hat{\beta}_\lambda) \to \mathcal{V}_\lambda^C = \tilde{\sigma}^2 \gamma (m(-\lambda) - \lambda m'(-\lambda)) , \tag{8}$$

where $m(\lambda) = ((1 - \gamma - \lambda) - \sqrt{(1 - \gamma - \lambda)^2 - 4\gamma\lambda})/(2\gamma\lambda)$ and $\tilde{r}^2 = r^2 + \omega^2 + 2\eta$. Therefore $R_X^C(\hat{\beta}_\lambda) \to \mathcal{R}_\lambda^C = \mathcal{B}_\lambda^C + \mathcal{V}_\lambda^C + \tilde{\sigma}^2 + \omega^2$. The corresponding limiting quantities for the min-norm interpolator can be obtained by taking the limit $\lambda \to 0^+$ in (7) and (8).

From these limiting expressions we can see that the causal risk curve of the min-norm interpolator exhibits the double descent phenomenon: it diverges at the interpolation threshold $\gamma = 1$ due to the variance term and decreases again for $\gamma > 1$. A corresponding visualization is given in Figure 4. Explicit regularization dampens the divergence of the variance term. While we are primarily interested in the causal risk, the corresponding statistical risk serves as a natural baseline. An analogue set of results for the statistical risk is given in Appendix C. These results have already been derived by Hastie et al. (2022) and can also be recovered as a special case of our causal results: for fixed statistical parameters $\tilde{\beta}$ and $\tilde{\sigma}^2$, the statistical risk coincides with the causal risk of an unconfounded causal model defined with $\beta = \tilde{\beta}$, $\sigma^2 = \tilde{\sigma}^2$, and $\alpha = 0$. In particular, the corresponding statistical limiting expressions are the same as in Theorem 3.1 after setting $\eta = \omega^2 = 0$.

Optimal statistical and causal regularization By directly optimizing the closed form expressions for limiting causal and statistical risks we can find the optimal causal and statistical regularization.

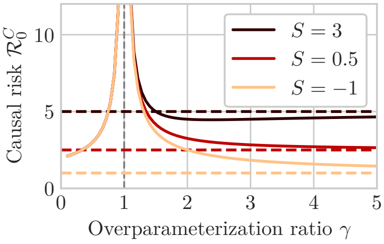


Figure 3: Limiting causal excess risk $\mathcal{R}_0^C$ (without the constant $\tilde{\sigma}^2 + \omega^2$) of the min-norm interpolator for different causal signal strengths S. Dashed lines are the corresponding null-risks ω^2, which are outperformed more often as S increases. For $\gamma < 1$, all three curves coincide.

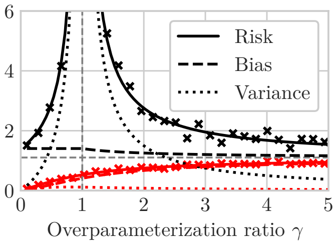


Figure 4: Limiting bias-variance decomposition and causal excess risk of the min-norm interpolator (black) and optimally regularized ridge regression (red). Crosses indicate finite-sample risks of $n = d/\gamma$ samples with $d = 300$. The finite risks are well-predicted by their theoretical limit.

For any $\gamma \in (0, \infty)$, the optimal statistical regularization $\lambda_S^*(\gamma) := \arg\inf_{\lambda \in (0,\infty)} \mathcal{R}_\lambda^S$ can be expressed in closed-form as $\lambda_S^*(\gamma) = \mathrm{SNR_S}^{-1} \gamma$. The closed-form expression for the optimal causal regularization parameter $\lambda_C^*(\gamma) := \arg\inf_{\lambda \in (0,\infty)} \mathcal{R}_\lambda^C$ is a root of a 4th order polynomial and as such considerably intricate. For readability, we do not include it here. We investigate the behavior of the optimal causal and statistical regularization in Section 4 and 5.

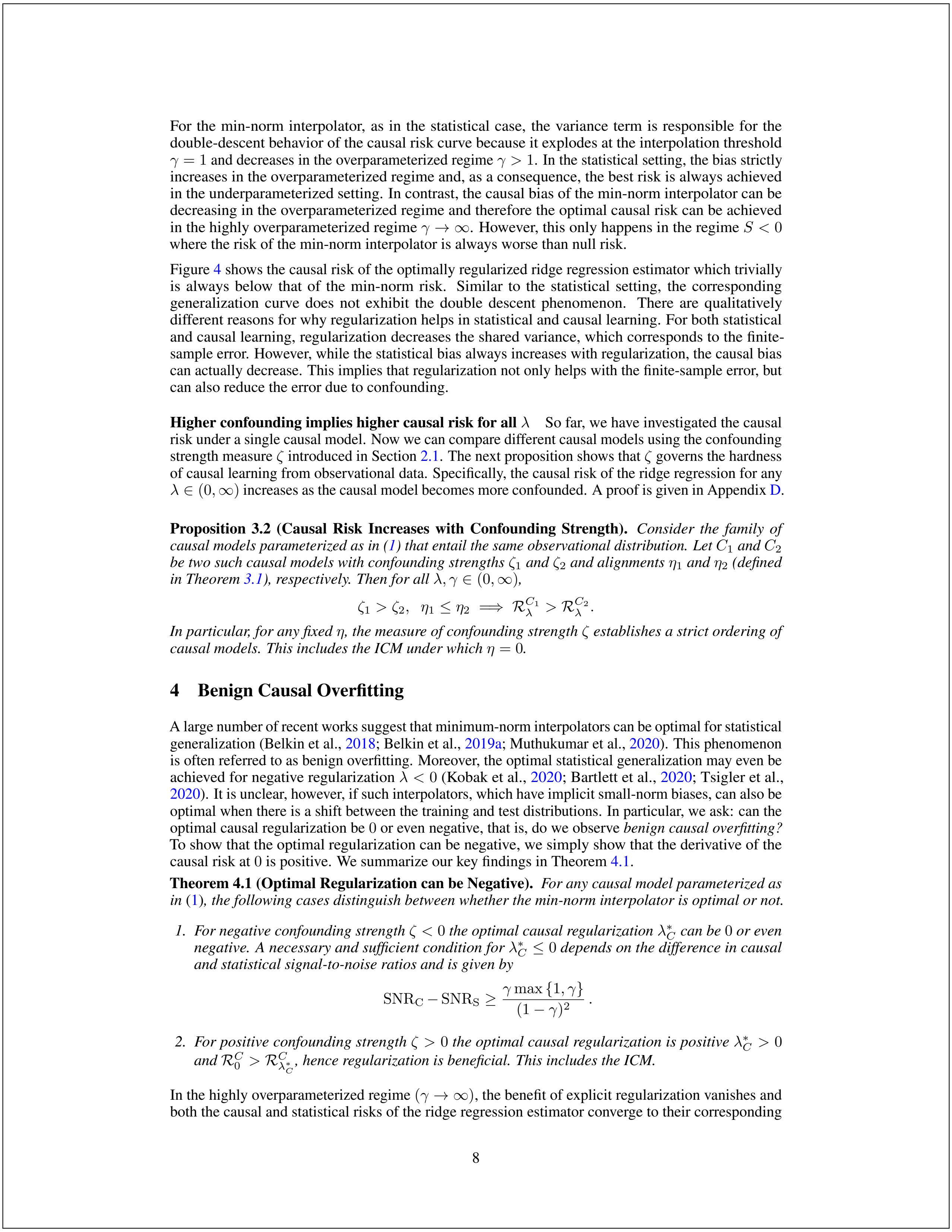
3.2 Basic Behavior of the Limiting Risk

We start to analyze the results by assessing the basic behavior of the limiting causal risk. The causal risk of the null estimator $\hat{\beta} = 0$ serves as a natural baseline to evaluate the performance of the the min-norm interpolator and the ridge regression estimators.

Regimes of the min-norm interpolator Theorem 3.1 characterizes the limiting causal risk of the min-norm interpolator. Its behavior is controlled by the causal signal-to-noise ratio, which we defined as $\mathrm{SNR_C} = (1-\zeta)\,\mathrm{SNR_S}$. However, as we will later see, the causal risk of the min-norm interpolator can be lower than null risk when $\zeta < 0.5$. To distinguish the regimes of the min-norm interpolator, it is therefore convenient to consider the closely related quantity $S = (1 - 2\zeta)\,\mathrm{SNR_S}$. It distinguishes between three different regimes (visualized in Figure 3).

- For $S > 1$, the causal signal dominates the noise and the min-norm interpolator can perform better than null risk in both under- and overparameterized regime.
- For $0 \leq S \leq 1$, the causal signal is weaker than the noise. Only the underparameterized regime can beat the null risk, whereas the overparameterized regime is always worse.
- The previous two cases resemble the behavior of the statistical risk in the corresponding regimes of the statistical SNR. Contrary to the statistical risk, however, the causal risk admits a third regime $S < 0$. In this case, the min-norm interpolator always performs worse than null risk. Here, the causal signal $\langle \beta, \tilde{\beta} \rangle$ is dominated by the confounding signal $\langle \Gamma, \tilde{\beta} \rangle$, and interpolating the observational data overfits to the confounding.

Bias and variance The bias-variance decomposition of the causal risk given in Theorem 3.1 is visualized in Figure 4 for the min-norm interpolator and the optimally ridge-regularized regressor. The figure also shows the causal risk based on finite samples from the model, which is in high agreement with our asymptotic results. We compare the causal risk to the corresponding statistical risk. First note that the causal and statistical variance terms coincide exactly for both the min-norm interpolator and ridge regressors. This is because the variance term of the squared loss depends only on the variance in the training data, but not on the target parameter β or $\tilde{\beta}$. Since the training data are the same for both causal and statistical learning, the variance terms trivially coincide.



For the min-norm interpolator, as in the statistical case, the variance term is responsible for the double-descent behavior of the causal risk curve because it explodes at the interpolation threshold $\gamma = 1$ and decreases in the overparameterized regime $\gamma > 1$. In the statistical setting, the bias strictly increases in the overparameterized regime and, as a consequence, the best risk is always achieved in the underparameterized setting. In contrast, the causal bias of the min-norm interpolator can be decreasing in the overparameterized regime and therefore the optimal causal risk can be achieved in the highly overparameterized regime $\gamma \to \infty$. However, this only happens in the regime $S < 0$ where the risk of the min-norm interpolator is always worse than null risk.

Figure 4 shows the causal risk of the optimally regularized ridge regression estimator which trivially is always below that of the min-norm risk. Similar to the statistical setting, the corresponding generalization curve does not exhibit the double descent phenomenon. There are qualitatively different reasons for why regularization helps in statistical and causal learning. For both statistical and causal learning, regularization decreases the shared variance, which corresponds to the finite-sample error. However, while the statistical bias always increases with regularization, the causal bias can actually decrease. This implies that regularization not only helps with the finite-sample error, but can also reduce the error due to confounding.

Higher confounding implies higher causal risk for all λ So far, we have investigated the causal risk under a single causal model. Now we can compare different causal models using the confounding strength measure ζ introduced in Section 2.1. The next proposition shows that ζ governs the hardness of causal learning from observational data. Specifically, the causal risk of the ridge regression for any $\lambda \in (0, \infty)$ increases as the causal model becomes more confounded. A proof is given in Appendix D.

Proposition 3.2 (Causal Risk Increases with Confounding Strength). *Consider the family of causal models parameterized as in (1) that entail the same observational distribution. Let C_1 and C_2 be two such causal models with confounding strengths ζ_1 and ζ_2 and alignments η_1 and η_2 (defined in Theorem 3.1), respectively. Then for all $\lambda, \gamma \in (0, \infty)$,*

$$\zeta_1 > \zeta_2, \quad \eta_1 \leq \eta_2 \implies \mathcal{R}_\lambda^{C_1} > \mathcal{R}_\lambda^{C_2}.$$

In particular, for any fixed η, the measure of confounding strength ζ establishes a strict ordering of causal models. This includes the ICM under which $\eta = 0$.

4 Benign Causal Overfitting

A large number of recent works suggest that minimum-norm interpolators can be optimal for statistical generalization (Belkin et al., 2018; Belkin et al., 2019a; Muthukumar et al., 2020). This phenomenon is often referred to as benign overfitting. Moreover, the optimal statistical generalization may even be achieved for negative regularization $\lambda < 0$ (Kobak et al., 2020; Bartlett et al., 2020; Tsigler et al., 2020). It is unclear, however, if such interpolators, which have implicit small-norm biases, can also be optimal when there is a shift between the training and test distributions. In particular, we ask: can the optimal causal regularization be 0 or even negative, that is, do we observe *benign causal overfitting?* To show that the optimal regularization can be negative, we simply show that the derivative of the causal risk at 0 is positive. We summarize our key findings in Theorem 4.1.

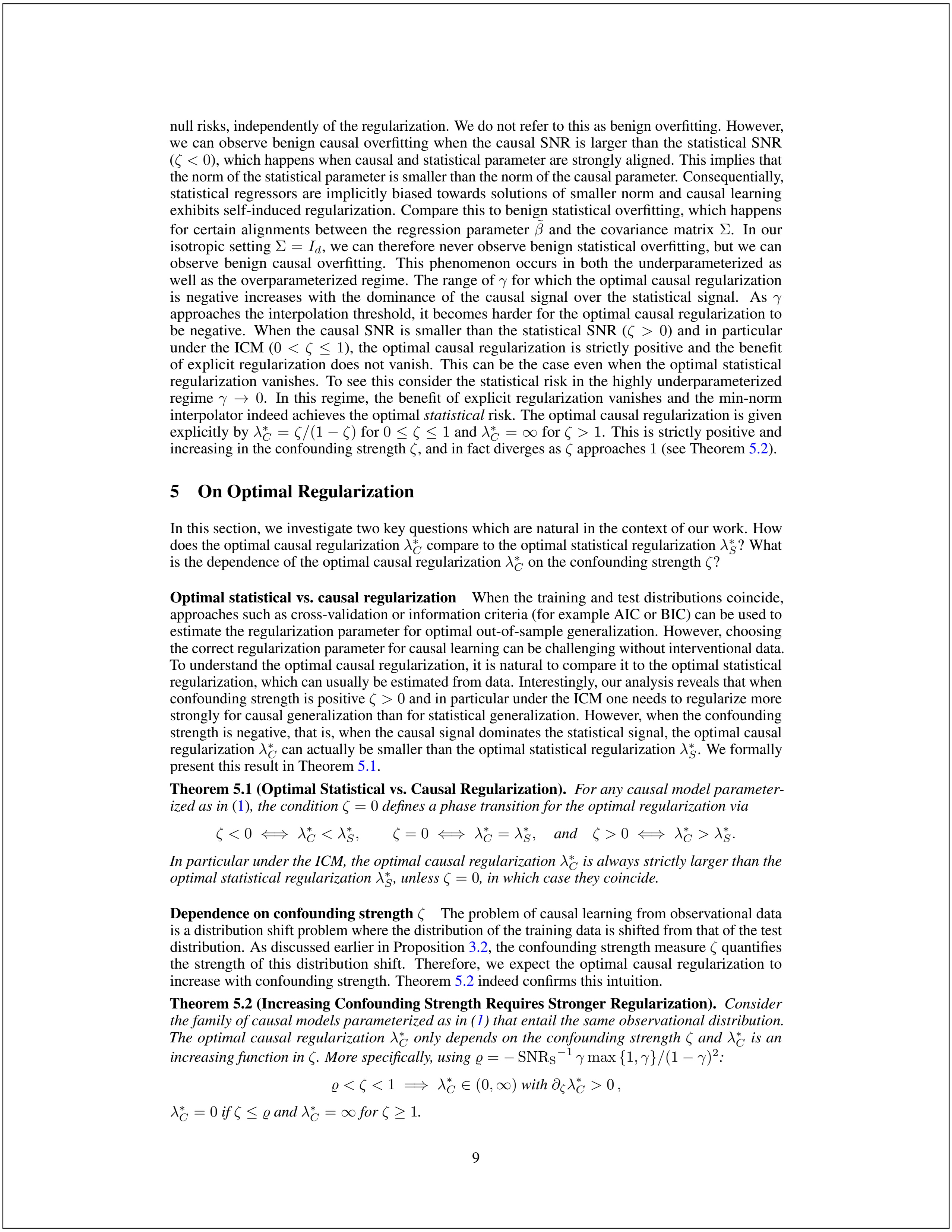
Theorem 4.1 (Optimal Regularization can be Negative). *For any causal model parameterized as in* (1)*, the following cases distinguish between whether the min-norm interpolator is optimal or not.*

1. *For negative confounding strength $\zeta < 0$ the optimal causal regularization λ_C^* can be 0 or even negative. A necessary and sufficient condition for $\lambda_C^* \leq 0$ depends on the difference in causal and statistical signal-to-noise ratios and is given by*

$$\mathrm{SNR_C} - \mathrm{SNR_S} \geq \frac{\gamma \max\{1, \gamma\}}{(1-\gamma)^2}\,.$$

2. *For positive confounding strength $\zeta > 0$ the optimal causal regularization is positive $\lambda_C^* > 0$ and $\mathcal{R}_0^C > \mathcal{R}_{\lambda_C^*}^C$, hence regularization is beneficial. This includes the ICM.*

In the highly overparameterized regime ($\gamma \to \infty$), the benefit of explicit regularization vanishes and both the causal and statistical risks of the ridge regression estimator converge to their corresponding



null risks, independently of the regularization. We do not refer to this as benign overfitting. However, we can observe benign causal overfitting when the causal SNR is larger than the statistical SNR ($\zeta < 0$), which happens when causal and statistical parameter are strongly aligned. This implies that the norm of the statistical parameter is smaller than the norm of the causal parameter. Consequentially, statistical regressors are implicitly biased towards solutions of smaller norm and causal learning exhibits self-induced regularization. Compare this to benign statistical overfitting, which happens for certain alignments between the regression parameter $\tilde{\beta}$ and the covariance matrix Σ. In our isotropic setting $\Sigma = I_d$, we can therefore never observe benign statistical overfitting, but we can observe benign causal overfitting. This phenomenon occurs in both the underparameterized as well as the overparameterized regime. The range of γ for which the optimal causal regularization is negative increases with the dominance of the causal signal over the statistical signal. As γ approaches the interpolation threshold, it becomes harder for the optimal causal regularization to be negative. When the causal SNR is smaller than the statistical SNR ($\zeta > 0$) and in particular under the ICM ($0 < \zeta \leq 1$), the optimal causal regularization is strictly positive and the benefit of explicit regularization does not vanish. This can be the case even when the optimal statistical regularization vanishes. To see this consider the statistical risk in the highly underparameterized regime $\gamma \to 0$. In this regime, the benefit of explicit regularization vanishes and the min-norm interpolator indeed achieves the optimal *statistical* risk. The optimal causal regularization is given explicitly by $\lambda_C^* = \zeta/(1-\zeta)$ for $0 \leq \zeta \leq 1$ and $\lambda_C^* = \infty$ for $\zeta > 1$. This is strictly positive and increasing in the confounding strength ζ, and in fact diverges as ζ approaches 1 (see Theorem 5.2).

5 On Optimal Regularization

In this section, we investigate two key questions which are natural in the context of our work. How does the optimal causal regularization λ_C^* compare to the optimal statistical regularization λ_S^*? What is the dependence of the optimal causal regularization λ_C^* on the confounding strength ζ?

Optimal statistical vs. causal regularization When the training and test distributions coincide, approaches such as cross-validation or information criteria (for example AIC or BIC) can be used to estimate the regularization parameter for optimal out-of-sample generalization. However, choosing the correct regularization parameter for causal learning can be challenging without interventional data. To understand the optimal causal regularization, it is natural to compare it to the optimal statistical regularization, which can usually be estimated from data. Interestingly, our analysis reveals that when confounding strength is positive $\zeta > 0$ and in particular under the ICM one needs to regularize more strongly for causal generalization than for statistical generalization. However, when the confounding strength is negative, that is, when the causal signal dominates the statistical signal, the optimal causal regularization λ_C^* can actually be smaller than the optimal statistical regularization λ_S^*. We formally present this result in Theorem 5.1.

Theorem 5.1 (Optimal Statistical vs. Causal Regularization). *For any causal model parameterized as in (1), the condition $\zeta = 0$ defines a phase transition for the optimal regularization via*

$$\zeta < 0 \iff \lambda_C^* < \lambda_S^*, \qquad \zeta = 0 \iff \lambda_C^* = \lambda_S^*, \quad \textit{and} \quad \zeta > 0 \iff \lambda_C^* > \lambda_S^*.$$

In particular under the ICM, the optimal causal regularization λ_C^* is always strictly larger than the optimal statistical regularization λ_S^*, unless $\zeta = 0$, in which case they coincide.

Dependence on confounding strength ζ The problem of causal learning from observational data is a distribution shift problem where the distribution of the training data is shifted from that of the test distribution. As discussed earlier in Proposition 3.2, the confounding strength measure ζ quantifies the strength of this distribution shift. Therefore, we expect the optimal causal regularization to increase with confounding strength. Theorem 5.2 indeed confirms this intuition.

Theorem 5.2 (Increasing Confounding Strength Requires Stronger Regularization). *Consider the family of causal models parameterized as in (1) that entail the same observational distribution. The optimal causal regularization λ_C^* only depends on the confounding strength ζ and λ_C^* is an increasing function in ζ. More specifically, using $\varrho = -\,\mathrm{SNR_S}^{-1}\gamma \max\{1,\gamma\}/(1-\gamma)^2$:*

$$\varrho < \zeta < 1 \implies \lambda_C^* \in (0,\infty) \textit{ with } \partial_\zeta \lambda_C^* > 0\,,$$

$\lambda_C^* = 0$ *if* $\zeta \leq \varrho$ *and* $\lambda_C^* = \infty$ *for* $\zeta \geq 1$.

6 Summary and Extensions

We characterize the role of explicit regularization for causal learning from observational data by computing the asymptotic risk of ridge-regularized regressors and the min-norm interpolator (Theorem 3.1). Under the principle of independent causal mechanisms (ICM), we find that causal learning requires stronger regularization than statistical learning (Theorem 5.1). A practical implication is that the regularization parameter for causal learning should be chosen larger than what is suggested by cross-validation. We can precisely state how much larger based on an estimate of confounding strength (Janzing et al., 2017; Janzing et al., 2018). Beyond ICM, we show that strong alignments between causal and statistical parameters can cause self-induced regularization and lead to benign causal overfitting (Theorem 4.1). One could consider generalizing our assumptions: arbitrary covariances, shifts in the marginal distributions of covariates, soft interventions, more complex hypothesis classes, or non-linear causal relationships. Since the linear model already exhibits rich behavior, we focus in this paper on understanding the simple setting. Below, we briefly discuss extensions of our analysis to causal learning under soft interventions, non-linearity, and non-Gaussianity.

Soft interventions It is not always appropriate to consider causal learning under hard interventions. Instead, it is often of interest to consider *soft interventions*. In these settings, the qualitative statements derived from our analysis still hold. To illustrate this, we consider the class of shift interventions where the structural dependence of the covariates x is not destroyed as in the case of hard interventions but the observed covariates are merely perturbed (i.e., interventions of the form $do(x := x + \nu)$). Then it turns out that $\text{Causal risk}_{\text{soft}} = \text{Causal risk}_{\text{hard}} + \text{Statistical risk}$. From our results, it then follows that under ICM, $\lambda^{\text{statistical}} \leq \lambda^{\text{causal}}_{\text{soft}} \leq \lambda^{\text{causal}}_{\text{hard}}$ This also supports our intuition since under soft interventions, we typically aim to achieve a tradeoff between statistical and causal predictability. We include a complete analysis under shift interventions in Appendix F.

Extensions to non-linear models It is feasible to extend the analysis to structural causal models that arise in a reproducing kernel Hilbert space corresponding to a positive definite kernel (i.e, where the best statistical model $\tilde{f}$ and the best causal model f are functions in some RKHS). There are two major technical challenges to deriving the theoretical analysis in such non-linear settings. Both are beyond what can be done in this paper and are left for future work, but we briefly outline them below.

1. **Extend the definition of confounding strength ζ beyond the linear setting.** Since such a definition is non-trivial already in the linear setting, it is challenging to meaningfully generalize this to the non-linear setting. However, under non-linear causal models in the RKHS, we can naturally extend this definition by replacing the Euclidean norms with functional norms in the RKHS. Generalizing the analysis beyond this setting would require further careful consideration.
2. **Derive limiting expressions for causal risk of regularized regressors in a non-linear hypothesis class.** In the case of kernel regression, this would still be feasible via recent random matrix theory results [27]. By optimizing the limiting expressions with respect to the regularization parameter, one can obtain the parameter that achieves the optimal causal risk and subsequently identify the relationship between optimal causal regularization and confounding strength.

Beyond Gaussianity The analysis can be extended beyond the Gaussian setting by considering random variables generated by finite mixtures of Gaussians. Due to the universality phenomenon in the high-dimensional limit, we believe that our limiting expressions (and the qualitative messages derived henceforth) would be rather robust to shifts in the marginal distribution as long as moments of order $(4 + \delta)$ for some $\delta > 0$ are bounded. We conducted experiments to verify this claim and the corresponding results can be found in Appendix G. They show that for distributions with finite 4th moments, the finite-sample risks of the min-norm interpolator and causally optimally regularized ridge regressor closely match the theoretically derived asymptotic risks.

Acknowledgments and Disclosure of Funding

This work has been supported by the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1:390727645) and the BMBF Tübingen AI Center (FKZ: 01IS18039A). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Leena C Vankadara and Luca Rendsburg.

References

Angrist, Joshua D and Alan B Keueger (1991). "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*.

Bai, Zhidong and Jack W Silverstein (2010). *Spectral analysis of large dimensional random matrices*. Springer.

Bartlett, Peter L et al. (2020). "Benign overfitting in linear regression". *Proceedings of the National Academy of Sciences*.

Belkin, Mikhail (2021). "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation". *Acta Numerica*.

Belkin, Mikhail, Siyuan Ma, and Soumik Mandal (2018). "To understand deep learning we need to understand kernel learning". *International Conference on Machine Learning (ICML)*.

Belkin, Mikhail, Alexander Rakhlin, and Alexandre B Tsybakov (2019a). "Does data interpolation contradict statistical optimality?" *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Belkin, Mikhail et al. (2019b). "Reconciling modern machine-learning practice and the classical bias–variance trade-off". *Proceedings of the National Academy of Sciences*.

Blanchet, Jose, Yang Kang, and Karthyek Murthy (2019). "Robust Wasserstein profile inference and applications to machine learning". *Journal of Applied Probability*.

Byrd, Jonathon and Zachary Lipton (2019). "What is the effect of importance weighting in deep learning?" *International Conference on Machine Learning (ICML)*.

Dicker, Lee H (2016). "Ridge regression and asymptotic minimax estimation over spheres of growing dimension". *Bernoulli*.

Dobriban, Edgar and Stefan Wager (2018). "High-dimensional asymptotics of prediction: Ridge regression and classification". *The Annals of Statistics*.

Donhauser, Konstantin et al. (2021). "Interpolation can hurt robust generalization even when there is no noise". *Advances in Neural Information Processing Systems (NeurIPS)*.

Feldman, Vitaly (2020). "Does learning require memorization? a short tale about a long tail". *Symposium on Theory of Computing (STOC)*.

Gao, Rui, Xi Chen, and Anton J Kleywegt (2017). "Distributional robustness and regularization in statistical learning". *arXiv preprint arXiv:1712.06050*.

Gulrajani, Ishaan and David Lopez-Paz (2021). "In Search of Lost Domain Generalization". *International Conference on Learning Representations (ICLR)*.

Hachem, Walid, Philippe Loubaton, and Jamal Najim (2007). "Deterministic equivalents for certain functionals of large random matrices". *The Annals of Applied Probability*.

Hastie, Trevor et al. (2022). "Surprises in high-dimensional ridgeless least squares interpolation". *The Annals of Statistics*.

Janzing, Dominik (2019). "Causal Regularization". *Advances in Neural Information Processing Systems (NeurIPS)*.

Janzing, Dominik and Bernhard Schölkopf (2010). "Causal Inference Using the Algorithmic Markov Condition". *IEEE Transactions on Information Theory*.

– (2017). "Detecting Confounding in Multivariate Linear Models via Spectral Analysis". *Journal of Causal Inference*.

– (2018). "Detecting non-causal artifacts in multivariate linear regression models". *International Conference on Machine Learning (ICML)*.

Kobak, Dmitry, Jonathan Lomond, and Benoit Sanchez (2020). "The Optimal Ridge Penalty for Real-world High-dimensional Data Can Be Zero or Negative due to the Implicit Ridge Regularization." *Journal of Machine Learning Research (JMLR)*.

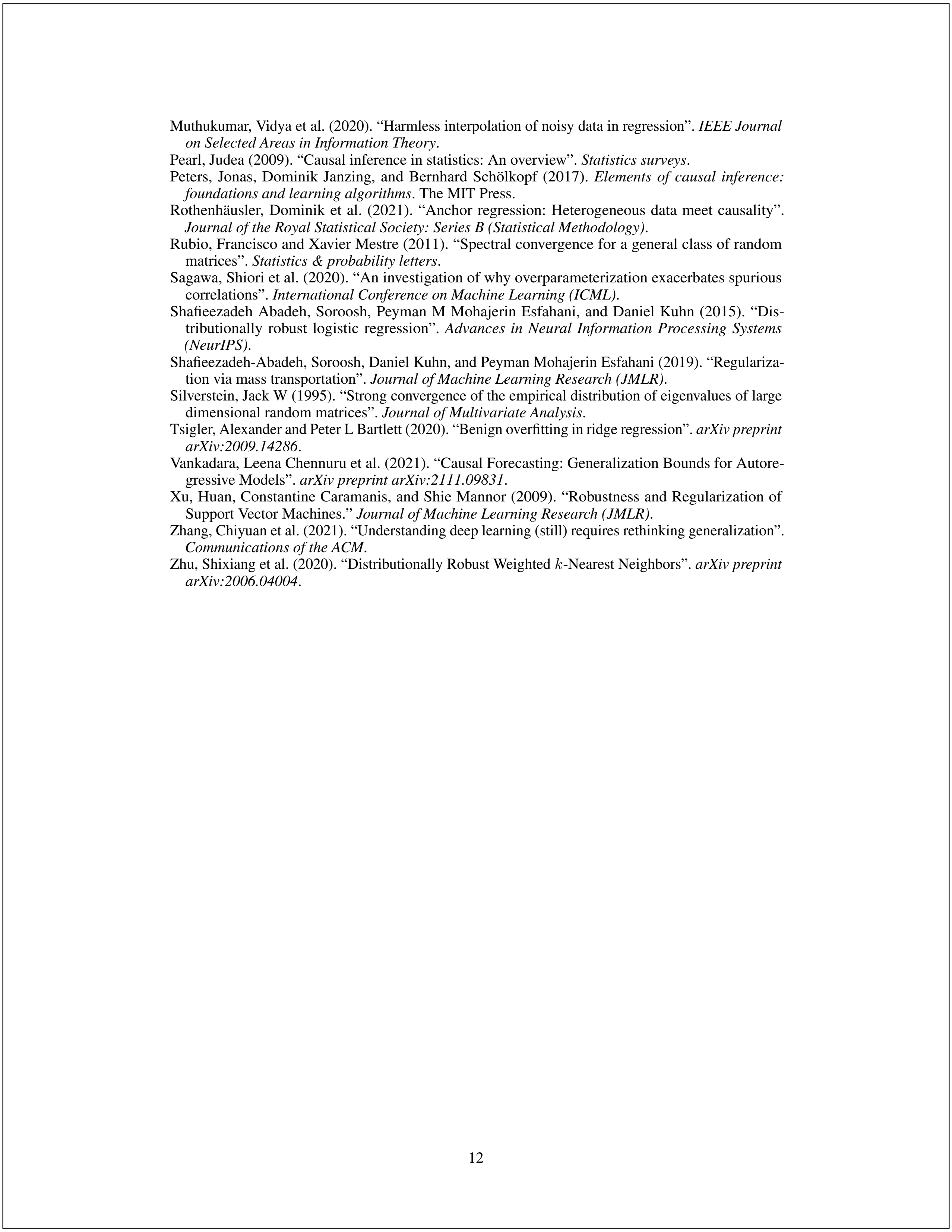
Kuhn, Daniel et al. (2019). "Wasserstein distributionally robust optimization: Theory and applications in machine learning". *Operations research & management science in the age of analytics*.

Lemeire, Jan and Dominik Janzing (2013). "Replacing Causal Faithfulness with Algorithmic Independence of Conditionals". *Minds and Machines*.

Liang, Tengyuan and Alexander Rakhlin (2020). "Just interpolate: Kernel "ridgeless" regression can generalize". *The Annals of Statistics*.

Marčenko, Vladimir A and Leonid Andreevich Pastur (1967). "Distribution of eigenvalues for some sets of random matrices". *Mathematics of the USSR-Sbornik*.

Mei, Song and Andrea Montanari (2021). "The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve". *Communications on Pure and Applied Mathematics*.



Muthukumar, Vidya et al. (2020). “Harmless interpolation of noisy data in regression”. *IEEE Journal on Selected Areas in Information Theory*.

Pearl, Judea (2009). “Causal inference in statistics: An overview”. *Statistics surveys*.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Rothenhäusler, Dominik et al. (2021). “Anchor regression: Heterogeneous data meet causality”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Rubio, Francisco and Xavier Mestre (2011). “Spectral convergence for a general class of random matrices”. *Statistics & probability letters*.

Sagawa, Shiori et al. (2020). “An investigation of why overparameterization exacerbates spurious correlations”. *International Conference on Machine Learning (ICML)*.

Shafieezadeh Abadeh, Soroosh, Peyman M Mohajerin Esfahani, and Daniel Kuhn (2015). “Distributionally robust logistic regression”. *Advances in Neural Information Processing Systems (NeurIPS)*.

Shafieezadeh-Abadeh, Soroosh, Daniel Kuhn, and Peyman Mohajerin Esfahani (2019). “Regularization via mass transportation”. *Journal of Machine Learning Research (JMLR)*.

Silverstein, Jack W (1995). “Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices”. *Journal of Multivariate Analysis*.

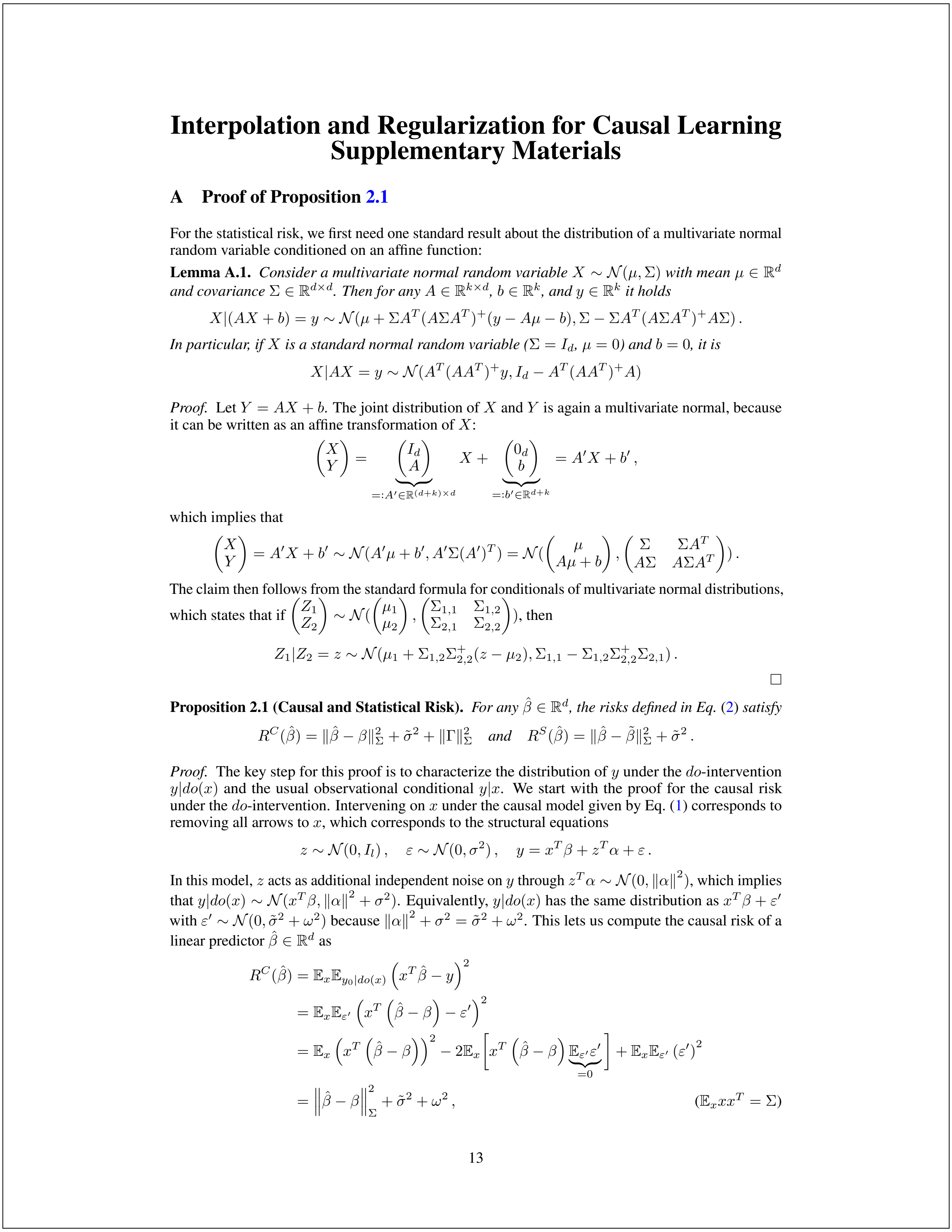
Tsigler, Alexander and Peter L Bartlett (2020). “Benign overfitting in ridge regression”. *arXiv preprint arXiv:2009.14286*.

Vankadara, Leena Chennuru et al. (2021). “Causal Forecasting: Generalization Bounds for Autoregressive Models”. *arXiv preprint arXiv:2111.09831*.

Xu, Huan, Constantine Caramanis, and Shie Mannor (2009). “Robustness and Regularization of Support Vector Machines.” *Journal of Machine Learning Research (JMLR)*.

Zhang, Chiyuan et al. (2021). “Understanding deep learning (still) requires rethinking generalization”. *Communications of the ACM*.

Zhu, Shixiang et al. (2020). “Distributionally Robust Weighted k-Nearest Neighbors”. *arXiv preprint arXiv:2006.04004*.



Interpolation and Regularization for Causal Learning Supplementary Materials

A Proof of Proposition 2.1

For the statistical risk, we first need one standard result about the distribution of a multivariate normal random variable conditioned on an affine function:

Lemma A.1. *Consider a multivariate normal random variable $X \sim \mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$. Then for any $A \in \mathbb{R}^{k\times d}$, $b \in \mathbb{R}^k$, and $y \in \mathbb{R}^k$ it holds*

$$X|(AX+b) = y \sim \mathcal{N}(\mu + \Sigma A^T(A\Sigma A^T)^+(y - A\mu - b), \Sigma - \Sigma A^T(A\Sigma A^T)^+ A\Sigma)\,.$$

In particular, if X is a standard normal random variable ($\Sigma = I_d$, $\mu = 0$) and $b = 0$, it is

$$X|AX = y \sim \mathcal{N}(A^T(AA^T)^+ y, I_d - A^T(AA^T)^+ A)$$

Proof. Let $Y = AX + b$. The joint distribution of X and Y is again a multivariate normal, because it can be written as an affine transformation of X:

$$\begin{pmatrix}X\\Y\end{pmatrix} = \underbrace{\begin{pmatrix}I_d\\A\end{pmatrix}}_{=:A'\in\mathbb{R}^{(d+k)\times d}} X + \underbrace{\begin{pmatrix}0_d\\b\end{pmatrix}}_{=:b'\in\mathbb{R}^{d+k}} = A'X + b'\,,$$

which implies that

$$\begin{pmatrix}X\\Y\end{pmatrix} = A'X + b' \sim \mathcal{N}(A'\mu + b', A'\Sigma(A')^T) = \mathcal{N}(\begin{pmatrix}\mu\\A\mu + b\end{pmatrix}, \begin{pmatrix}\Sigma & \Sigma A^T\\ A\Sigma & A\Sigma A^T\end{pmatrix})\,.$$

The claim then follows from the standard formula for conditionals of multivariate normal distributions, which states that if $\begin{pmatrix}Z_1\\Z_2\end{pmatrix} \sim \mathcal{N}(\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}, \begin{pmatrix}\Sigma_{1,1} & \Sigma_{1,2}\\ \Sigma_{2,1} & \Sigma_{2,2}\end{pmatrix})$, then

$$Z_1|Z_2 = z \sim \mathcal{N}(\mu_1 + \Sigma_{1,2}\Sigma_{2,2}^+(z - \mu_2), \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^+\Sigma_{2,1})\,.$$

□

Proposition 2.1 (Causal and Statistical Risk). *For any $\hat\beta \in \mathbb{R}^d$, the risks defined in Eq. (2) satisfy*

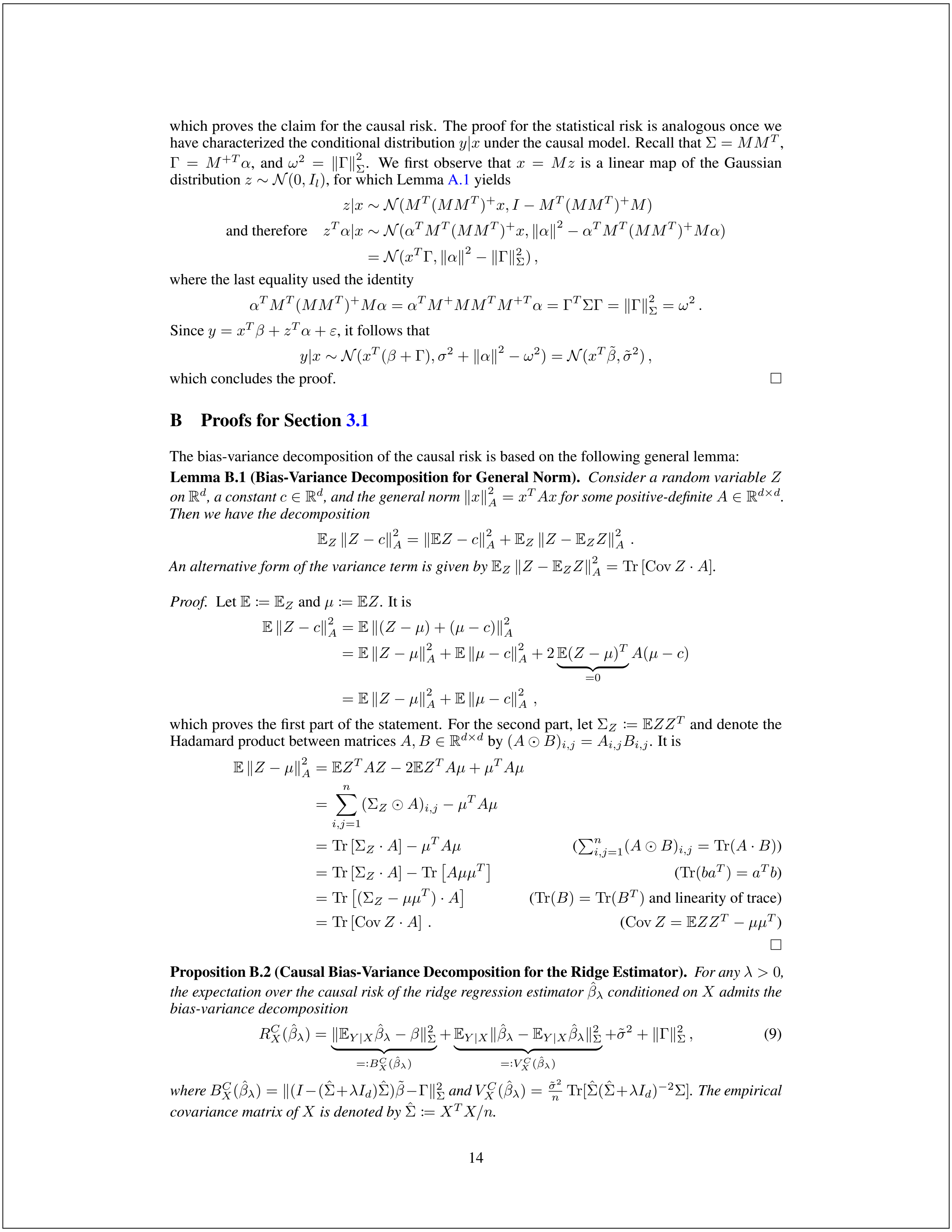
$$R^C(\hat\beta) = \|\hat\beta - \beta\|_\Sigma^2 + \tilde\sigma^2 + \|\Gamma\|_\Sigma^2 \quad \textit{and} \quad R^S(\hat\beta) = \|\hat\beta - \tilde\beta\|_\Sigma^2 + \tilde\sigma^2\,.$$

Proof. The key step for this proof is to characterize the distribution of y under the do-intervention $y|do(x)$ and the usual observational conditional $y|x$. We start with the proof for the causal risk under the do-intervention. Intervening on x under the causal model given by Eq. (1) corresponds to removing all arrows to x, which corresponds to the structural equations

$$z \sim \mathcal{N}(0, I_l)\,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,, \quad y = x^T\beta + z^T\alpha + \varepsilon\,.$$

In this model, z acts as additional independent noise on y through $z^T\alpha \sim \mathcal{N}(0, \|\alpha\|^2)$, which implies that $y|do(x) \sim \mathcal{N}(x^T\beta, \|\alpha\|^2 + \sigma^2)$. Equivalently, $y|do(x)$ has the same distribution as $x^T\beta + \varepsilon'$ with $\varepsilon' \sim \mathcal{N}(0, \tilde\sigma^2 + \omega^2)$ because $\|\alpha\|^2 + \sigma^2 = \tilde\sigma^2 + \omega^2$. This lets us compute the causal risk of a linear predictor $\hat\beta \in \mathbb{R}^d$ as

$$\begin{aligned}
R^C(\hat\beta) &= \mathbb{E}_x\mathbb{E}_{y_0|do(x)}\left(x^T\hat\beta - y\right)^2\\
&= \mathbb{E}_x\mathbb{E}_{\varepsilon'}\left(x^T\left(\hat\beta - \beta\right) - \varepsilon'\right)^2\\
&= \mathbb{E}_x\left(x^T\left(\hat\beta - \beta\right)\right)^2 - 2\mathbb{E}_x\left[x^T\left(\hat\beta - \beta\right)\underbrace{\mathbb{E}_{\varepsilon'}\varepsilon'}_{=0}\right] + \mathbb{E}_x\mathbb{E}_{\varepsilon'}\left(\varepsilon'\right)^2\\
&= \left\|\hat\beta - \beta\right\|_\Sigma^2 + \tilde\sigma^2 + \omega^2\,, && (\mathbb{E}_x xx^T = \Sigma)
\end{aligned}$$



which proves the claim for the causal risk. The proof for the statistical risk is analogous once we have characterized the conditional distribution $y|x$ under the causal model. Recall that $\Sigma = MM^T$, $\Gamma = M^{+T}\alpha$, and $\omega^2 = \|\Gamma\|_\Sigma^2$. We first observe that $x = Mz$ is a linear map of the Gaussian distribution $z \sim \mathcal{N}(0, I_l)$, for which Lemma A.1 yields

$$\begin{aligned} z|x &\sim \mathcal{N}(M^T(MM^T)^+x, I - M^T(MM^T)^+M) \\ \text{and therefore} \quad z^T\alpha|x &\sim \mathcal{N}(\alpha^T M^T(MM^T)^+x, \|\alpha\|^2 - \alpha^T M^T(MM^T)^+M\alpha) \\ &= \mathcal{N}(x^T\Gamma, \|\alpha\|^2 - \|\Gamma\|_\Sigma^2)\,, \end{aligned}$$

where the last equality used the identity

$$\alpha^T M^T(MM^T)^+M\alpha = \alpha^T M^+ MM^T M^{+T}\alpha = \Gamma^T\Sigma\Gamma = \|\Gamma\|_\Sigma^2 = \omega^2\,.$$

Since $y = x^T\beta + z^T\alpha + \varepsilon$, it follows that

$$y|x \sim \mathcal{N}(x^T(\beta+\Gamma), \sigma^2 + \|\alpha\|^2 - \omega^2) = \mathcal{N}(x^T\tilde{\beta}, \tilde{\sigma}^2)\,,$$

which concludes the proof. □

B Proofs for Section 3.1

The bias-variance decomposition of the causal risk is based on the following general lemma:

Lemma B.1 (Bias-Variance Decomposition for General Norm). *Consider a random variable Z on $\mathbb{R}^d$, a constant $c \in \mathbb{R}^d$, and the general norm $\|x\|_A^2 = x^T Ax$ for some positive-definite $A \in \mathbb{R}^{d\times d}$. Then we have the decomposition*

$$\mathbb{E}_Z \|Z - c\|_A^2 = \|\mathbb{E}Z - c\|_A^2 + \mathbb{E}_Z \|Z - \mathbb{E}_Z Z\|_A^2\;.$$

An alternative form of the variance term is given by $\mathbb{E}_Z \|Z - \mathbb{E}_Z Z\|_A^2 = \mathrm{Tr}\,[\mathrm{Cov}\, Z \cdot A]$.

Proof. Let $\mathbb{E} := \mathbb{E}_Z$ and $\mu := \mathbb{E}Z$. It is

$$\begin{aligned} \mathbb{E}\|Z - c\|_A^2 &= \mathbb{E}\|(Z-\mu) + (\mu - c)\|_A^2 \\ &= \mathbb{E}\|Z-\mu\|_A^2 + \mathbb{E}\|\mu - c\|_A^2 + 2\underbrace{\mathbb{E}(Z-\mu)^T}_{=0} A(\mu - c) \\ &= \mathbb{E}\|Z-\mu\|_A^2 + \mathbb{E}\|\mu - c\|_A^2\;, \end{aligned}$$

which proves the first part of the statement. For the second part, let $\Sigma_Z := \mathbb{E}ZZ^T$ and denote the Hadamard product between matrices $A, B \in \mathbb{R}^{d\times d}$ by $(A \odot B)_{i,j} = A_{i,j}B_{i,j}$. It is

$$\begin{aligned} \mathbb{E}\|Z-\mu\|_A^2 &= \mathbb{E}Z^T AZ - 2\mathbb{E}Z^T A\mu + \mu^T A\mu \\ &= \sum_{i,j=1}^n (\Sigma_Z \odot A)_{i,j} - \mu^T A\mu \\ &= \mathrm{Tr}\,[\Sigma_Z \cdot A] - \mu^T A\mu && (\textstyle\sum_{i,j=1}^n (A\odot B)_{i,j} = \mathrm{Tr}(A\cdot B)) \\ &= \mathrm{Tr}\,[\Sigma_Z \cdot A] - \mathrm{Tr}\left[A\mu\mu^T\right] && (\mathrm{Tr}(ba^T) = a^T b) \\ &= \mathrm{Tr}\left[(\Sigma_Z - \mu\mu^T)\cdot A\right] && (\mathrm{Tr}(B) = \mathrm{Tr}(B^T) \text{ and linearity of trace}) \\ &= \mathrm{Tr}\,[\mathrm{Cov}\, Z \cdot A]\;. && (\mathrm{Cov}\, Z = \mathbb{E}ZZ^T - \mu\mu^T) \end{aligned}$$

□

Proposition B.2 (Causal Bias-Variance Decomposition for the Ridge Estimator). *For any $\lambda > 0$, the expectation over the causal risk of the ridge regression estimator $\hat{\beta}_\lambda$ conditioned on X admits the bias-variance decomposition*

$$R_X^C(\hat{\beta}_\lambda) = \underbrace{\|\mathbb{E}_{Y|X}\hat{\beta}_\lambda - \beta\|_\Sigma^2}_{=:B_X^C(\hat{\beta}_\lambda)} + \underbrace{\mathbb{E}_{Y|X}\|\hat{\beta}_\lambda - \mathbb{E}_{Y|X}\hat{\beta}_\lambda\|_\Sigma^2}_{=:V_X^C(\hat{\beta}_\lambda)} + \tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2\,, \tag{9}$$

where $B_X^C(\hat{\beta}_\lambda) = \|(I - (\hat{\Sigma} + \lambda I_d)\hat{\Sigma})\tilde{\beta} - \Gamma\|_\Sigma^2$ and $V_X^C(\hat{\beta}_\lambda) = \frac{\tilde{\sigma}^2}{n}\,\mathrm{Tr}[\hat{\Sigma}(\hat{\Sigma} + \lambda I_d)^{-2}\Sigma]$. The empirical covariance matrix of X is denoted by $\hat{\Sigma} := X^T X/n$.

Proof. Recall that $R_X^C(\hat{\beta}_\lambda) = \mathbb{E}_{Y|X} \left\| \hat{\beta}_\lambda - \beta \right\|_\Sigma^2$. The first part of the statement follows directly from Lemma B.1 with $\hat{\beta}_\lambda$ as a random variable in $Y|X$ and β. The remainder of the proof consists of computing expectation and covariance of the ridge regression solution $\hat{\beta}_\lambda = \hat{\beta}_\lambda(X, Y)$ under the distribution $Y|X$. The samples (X, Y) are drawn from the observational distribution of the causal model defined in Eq. (1). As shown in the proof of Proposition 2.1, the corresponding conditional distribution is $y|x \sim \mathcal{N}(x^T\tilde{\beta}, \tilde{\sigma}^2)$. Since (X, Y) consist of independent draws, this implies $Y|X \sim \mathcal{N}(X\tilde{\beta}, \tilde{\sigma}^2 I_n)$. Together with $\hat{\beta}_\lambda = (X^TX + n\lambda I)^{-1}X^TY$ this yields

$$\begin{aligned}\hat{\beta}_\lambda|X &\sim \mathcal{N}((X^TX + n\lambda I)^{-1}X^TX\tilde{\beta}, (X^TX + n\lambda I)^{-1}X^T\tilde{\sigma}^2 I_n X(X^TX + n\lambda I)^{-1}) \\ &= \mathcal{N}(\left(\hat{\Sigma} + \lambda I_d\right)^{-1} \hat{\Sigma}\tilde{\beta}, \frac{\tilde{\sigma}^2}{n} \left(\hat{\Sigma} + \lambda I_d\right)^{-1} \hat{\Sigma} \left(\hat{\Sigma} + \lambda I_d\right)^{-1}).\end{aligned}$$

The characterizations of $B_X^C(\hat{\beta}_\lambda)$ and $V_X^C(\hat{\beta}_\lambda)$ then simply follow from plugging in expectation and covariance of $\hat{\beta}_\lambda$:

$$\begin{aligned}B_X^C(\hat{\beta}_\lambda) = \left\| \mathbb{E}_{Y|X}\hat{\beta}_\lambda - \beta \right\|_\Sigma^2 = \left\| \left(\hat{\Sigma} + \lambda I_d\right)^{-1} \hat{\Sigma}\tilde{\beta} - \beta \right\|_\Sigma^2 &= \|(I - \Pi_\lambda)(\beta + \Gamma) - \beta\|_\Sigma^2 \\ &= \|\Pi_\lambda\beta - (I - \Pi_\lambda)\Gamma\|_\Sigma^2\end{aligned}$$

and, using the alternate form of the variance term from Lemma B.1,

$$\begin{aligned}V_X^C(\hat{\beta}_\lambda) = \operatorname{Tr}\left[\operatorname{Cov}_{Y|X} \hat{\beta}_\lambda \cdot \Sigma\right] &= \operatorname{Tr}\left[\frac{\tilde{\sigma}^2}{n} \left(\hat{\Sigma} + \lambda I_d\right)^{-1} \hat{\Sigma} \left(\hat{\Sigma} + \lambda I_d\right)^{-1} \cdot \Sigma\right] \\ &= \frac{\tilde{\sigma}^2}{n} \operatorname{Tr}\left[\hat{\Sigma} \left(\hat{\Sigma} + \lambda I_d\right)^{-2} \Sigma\right],\end{aligned}$$

where the last equality used that $\left(\hat{\Sigma} + \lambda I_d\right)^{-1}$ commutes with $\hat{\Sigma}$. □

Theorem 2 (Limiting Causal Bias-Variance Decomposition for the Ridge Estimator). *Let $\|\beta\|^2 = r^2$, $\|\Gamma\|^2 = \omega^2$, $\langle\Gamma, \beta\rangle = \eta$, and $\sigma_\epsilon^2 = \tilde{\sigma}^2$. Then as $n, d \to \infty$ such that $d/n \to \gamma \in (0, \infty)$, it holds almost surely in X for every $\lambda > 0$ that*

$$B_X^C(\hat{\beta}_\lambda) \to \mathcal{B}_\lambda^C := \omega^2 + \tilde{r}^2\lambda^2 m'(-\lambda) - 2(\omega^2 + \eta)\lambda m(-\lambda) \quad \textit{and} \tag{7}$$

$$V_X^C(\hat{\beta}_\lambda) \to \mathcal{V}_\lambda^C := \tilde{\sigma}^2\gamma(m(-\lambda) - \lambda m'(-\lambda)), \tag{8}$$

where $m(\lambda) = ((1 - \gamma - \lambda) - \sqrt{(1 - \gamma - \lambda)^2 - 4\gamma\lambda})/(2\gamma\lambda)$ and $\tilde{r}^2 = r^2 + \omega^2 + 2\eta$. Therefore $R_X^C(\hat{\beta}_\lambda) \to \mathcal{R}_\lambda^C := \mathcal{B}_\lambda^C + \mathcal{V}_\lambda^C + \tilde{\sigma}^2 + \omega^2$. The corresponding limiting quantities for the min-norm interpolator can be obtained by taking the limit $\lambda \to 0^+$ in equations (7) and (8), which yields

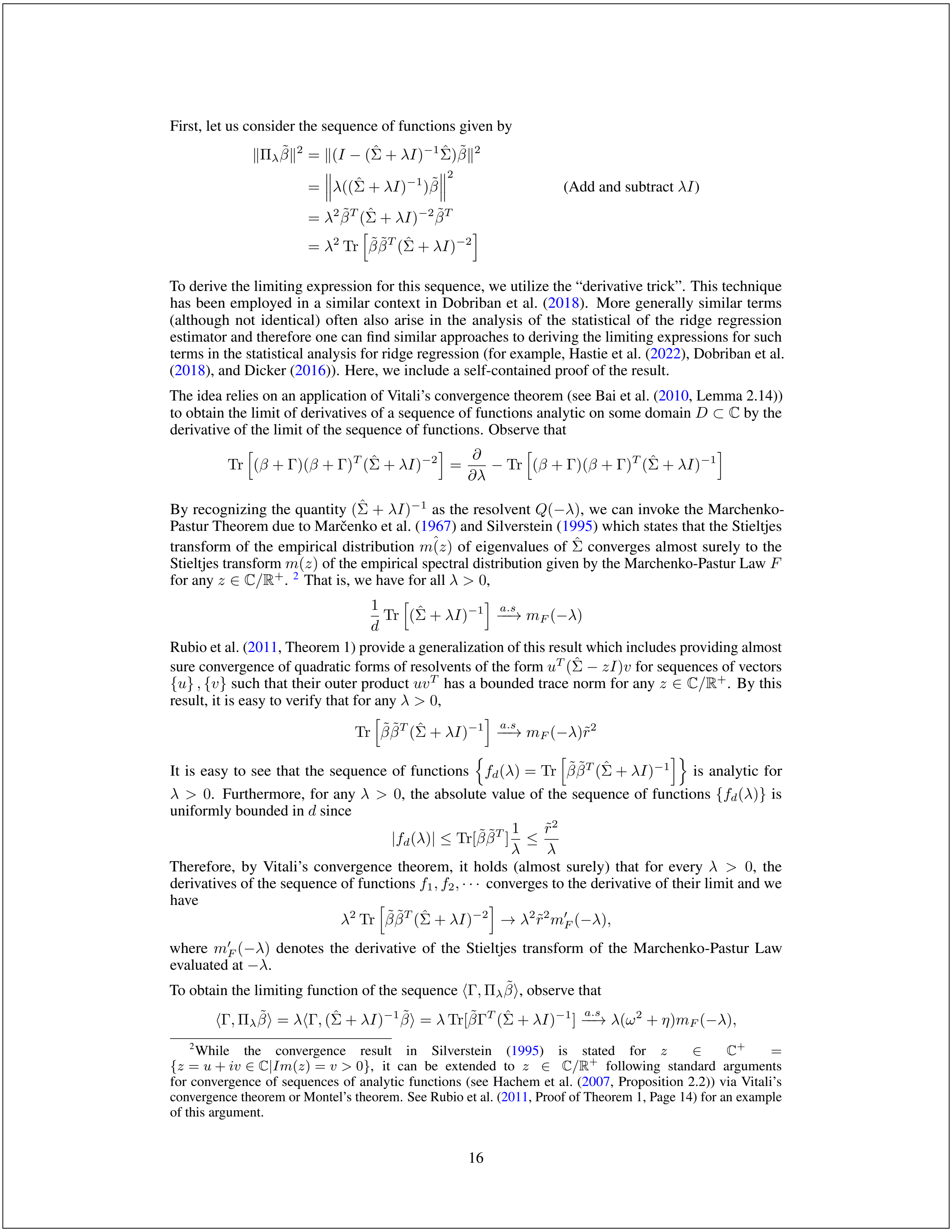
$$B_X^C(\hat{\beta}_0) \to \mathcal{B}_0^C = \begin{cases} \omega^2, & \gamma < 1 \\ \omega^2 + (r^2 - \omega^2)(1 - \frac{1}{\gamma}), & \gamma > 1 \end{cases}, \quad V_X^C(\hat{\beta}_0) \to \mathcal{V}_0^C = \begin{cases} \tilde{\sigma}^2 \frac{\gamma}{1-\gamma}, & \gamma < 1 \\ \tilde{\sigma}^2 \frac{1}{\gamma - 1}, & \gamma > 1 \end{cases}.$$

Therefore $R_X^C(\hat{\beta}_0) \to \mathcal{R}_0^C = \mathcal{B}_0^C + \mathcal{V}_0^C + \tilde{\sigma}^2 + \omega^2$.

Proof. From Proposition B.2, the causal risk $R_X^C(\hat{\beta}_\lambda)$ can be decomposed as a sum of the causal bias $B_X^C(\hat{\beta}_\lambda)$, and causal variance $V_X^C(\hat{\beta}_\lambda)$. In what follows, we derive the limiting expressions for $B_X^C(\hat{\beta}_\lambda)$ and $V_X^C(\hat{\beta}_\lambda)$ to obtain the limiting causal risk for any $\gamma \in (0, \infty)$.

Limiting expressions for causal bias

$$\begin{aligned}B_X^C(\hat{\beta}_\lambda) &= \|\beta - \mathbb{E}_{|X}\hat{\beta}_\lambda\|_\Sigma^2 = \|\Pi_\lambda\beta - (I - \Pi_\lambda)\Gamma\|^2 \qquad (\Sigma = I) \\ &= \|\Pi_\lambda(\beta + \Gamma) - \Gamma\|^2 \\ &= \|\Pi_\lambda\tilde{\beta}\|^2 + \|\Gamma\|^2 - 2\langle\Gamma, \Pi_\lambda(\tilde{\beta})\rangle\end{aligned}$$



First, let us consider the sequence of functions given by

$$\begin{aligned}\|\Pi_\lambda \tilde{\beta}\|^2 &= \|(I-(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma})\tilde{\beta}\|^2 \\ &= \left\|\lambda((\hat{\Sigma}+\lambda I)^{-1})\tilde{\beta}\right\|^2 && \text{(Add and subtract } \lambda I\text{)} \\ &= \lambda^2 \tilde{\beta}^T(\hat{\Sigma}+\lambda I)^{-2}\tilde{\beta}^T \\ &= \lambda^2 \operatorname{Tr}\left[\tilde{\beta}\tilde{\beta}^T(\hat{\Sigma}+\lambda I)^{-2}\right]\end{aligned}$$

To derive the limiting expression for this sequence, we utilize the “derivative trick”. This technique has been employed in a similar context in Dobriban et al. (2018). More generally similar terms (although not identical) often also arise in the analysis of the statistical of the ridge regression estimator and therefore one can find similar approaches to deriving the limiting expressions for such terms in the statistical analysis for ridge regression (for example, Hastie et al. (2022), Dobriban et al. (2018), and Dicker (2016)). Here, we include a self-contained proof of the result.

The idea relies on an application of Vitali's convergence theorem (see Bai et al. (2010, Lemma 2.14)) to obtain the limit of derivatives of a sequence of functions analytic on some domain $D \subset \mathbb{C}$ by the derivative of the limit of the sequence of functions. Observe that

$$\operatorname{Tr}\left[(\beta+\Gamma)(\beta+\Gamma)^T(\hat{\Sigma}+\lambda I)^{-2}\right] = \frac{\partial}{\partial\lambda} - \operatorname{Tr}\left[(\beta+\Gamma)(\beta+\Gamma)^T(\hat{\Sigma}+\lambda I)^{-1}\right]$$

By recognizing the quantity $(\hat{\Sigma}+\lambda I)^{-1}$ as the resolvent $Q(-\lambda)$, we can invoke the Marchenko-Pastur Theorem due to Marčenko et al. (1967) and Silverstein (1995) which states that the Stieltjes transform of the empirical distribution $\hat{m(z)}$ of eigenvalues of $\hat{\Sigma}$ converges almost surely to the Stieltjes transform $m(z)$ of the empirical spectral distribution given by the Marchenko-Pastur Law F for any $z \in \mathbb{C}/\mathbb{R}^+$. [2] That is, we have for all $\lambda > 0$,

$$\frac{1}{d}\operatorname{Tr}\left[(\hat{\Sigma}+\lambda I)^{-1}\right] \xrightarrow{a.s} m_F(-\lambda)$$

Rubio et al. (2011, Theorem 1) provide a generalization of this result which includes providing almost sure convergence of quadratic forms of resolvents of the form $u^T(\hat{\Sigma}-zI)v$ for sequences of vectors $\{u\},\{v\}$ such that their outer product uv^T has a bounded trace norm for any $z \in \mathbb{C}/\mathbb{R}^+$. By this result, it is easy to verify that for any $\lambda > 0$,

$$\operatorname{Tr}\left[\tilde{\beta}\tilde{\beta}^T(\hat{\Sigma}+\lambda I)^{-1}\right] \xrightarrow{a.s} m_F(-\lambda)\tilde{r}^2$$

It is easy to see that the sequence of functions $\left\{f_d(\lambda) = \operatorname{Tr}\left[\tilde{\beta}\tilde{\beta}^T(\hat{\Sigma}+\lambda I)^{-1}\right]\right\}$ is analytic for $\lambda > 0$. Furthermore, for any $\lambda > 0$, the absolute value of the sequence of functions $\{f_d(\lambda)\}$ is uniformly bounded in d since

$$|f_d(\lambda)| \leq \operatorname{Tr}[\tilde{\beta}\tilde{\beta}^T]\frac{1}{\lambda} \leq \frac{\tilde{r}^2}{\lambda}$$

Therefore, by Vitali's convergence theorem, it holds (almost surely) that for every $\lambda > 0$, the derivatives of the sequence of functions $f_1, f_2, \cdots$ converges to the derivative of their limit and we have

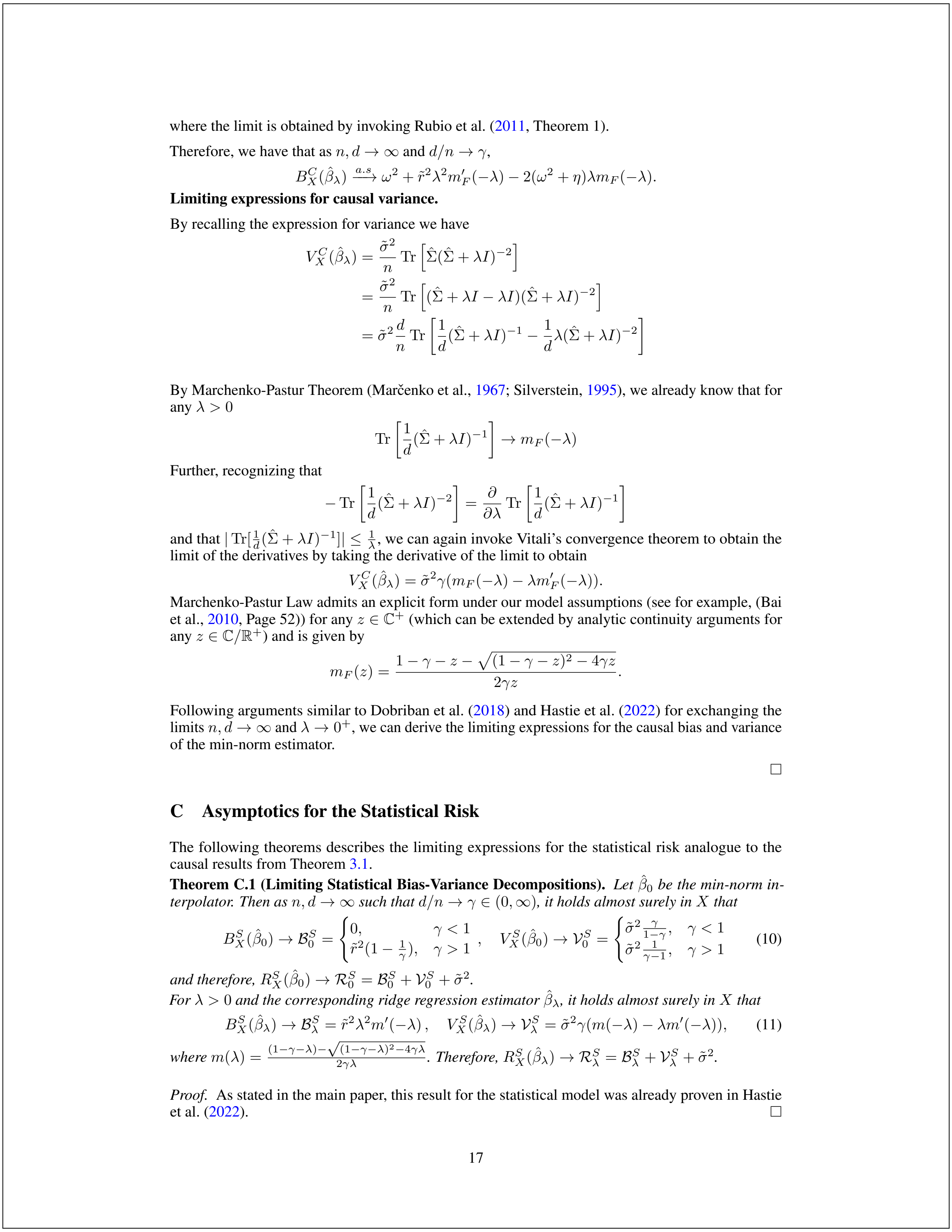
$$\lambda^2 \operatorname{Tr}\left[\tilde{\beta}\tilde{\beta}^T(\hat{\Sigma}+\lambda I)^{-2}\right] \to \lambda^2\tilde{r}^2 m_F'(-\lambda),$$

where $m_F'(-\lambda)$ denotes the derivative of the Stieltjes transform of the Marchenko-Pastur Law evaluated at $-\lambda$.

To obtain the limiting function of the sequence $\langle\Gamma, \Pi_\lambda\tilde{\beta}\rangle$, observe that

$$\langle\Gamma, \Pi_\lambda\tilde{\beta}\rangle = \lambda\langle\Gamma, (\hat{\Sigma}+\lambda I)^{-1}\tilde{\beta}\rangle = \lambda \operatorname{Tr}[\tilde{\beta}\Gamma^T(\hat{\Sigma}+\lambda I)^{-1}] \xrightarrow{a.s} \lambda(\omega^2+\eta)m_F(-\lambda),$$

[2] While the convergence result in Silverstein (1995) is stated for $z \in \mathbb{C}^+ = \{z = u+iv \in \mathbb{C}|Im(z) = v > 0\}$, it can be extended to $z \in \mathbb{C}/\mathbb{R}^+$ following standard arguments for convergence of sequences of analytic functions (see Hachem et al. (2007, Proposition 2.2)) via Vitali's convergence theorem or Montel's theorem. See Rubio et al. (2011, Proof of Theorem 1, Page 14) for an example of this argument.



where the limit is obtained by invoking Rubio et al. (2011, Theorem 1).

Therefore, we have that as $n, d \to \infty$ and $d/n \to \gamma$,

$$B_X^C(\hat{\beta}_\lambda) \xrightarrow{a.s} \omega^2 + \tilde{r}^2\lambda^2 m_F'(-\lambda) - 2(\omega^2 + \eta)\lambda m_F(-\lambda).$$

Limiting expressions for causal variance.

By recalling the expression for variance we have

$$\begin{aligned} V_X^C(\hat{\beta}_\lambda) &= \frac{\tilde{\sigma}^2}{n} \operatorname{Tr}\left[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-2}\right] \\ &= \frac{\tilde{\sigma}^2}{n} \operatorname{Tr}\left[(\hat{\Sigma} + \lambda I - \lambda I)(\hat{\Sigma} + \lambda I)^{-2}\right] \\ &= \tilde{\sigma}^2 \frac{d}{n} \operatorname{Tr}\left[\frac{1}{d}(\hat{\Sigma} + \lambda I)^{-1} - \frac{1}{d}\lambda(\hat{\Sigma} + \lambda I)^{-2}\right] \end{aligned}$$

By Marchenko-Pastur Theorem (Marčenko et al., 1967; Silverstein, 1995), we already know that for any $\lambda > 0$

$$\operatorname{Tr}\left[\frac{1}{d}(\hat{\Sigma} + \lambda I)^{-1}\right] \to m_F(-\lambda)$$

Further, recognizing that

$$-\operatorname{Tr}\left[\frac{1}{d}(\hat{\Sigma} + \lambda I)^{-2}\right] = \frac{\partial}{\partial \lambda} \operatorname{Tr}\left[\frac{1}{d}(\hat{\Sigma} + \lambda I)^{-1}\right]$$

and that $|\operatorname{Tr}[\frac{1}{d}(\hat{\Sigma} + \lambda I)^{-1}]| \leq \frac{1}{\lambda}$, we can again invoke Vitali's convergence theorem to obtain the limit of the derivatives by taking the derivative of the limit to obtain

$$V_X^C(\hat{\beta}_\lambda) = \tilde{\sigma}^2\gamma(m_F(-\lambda) - \lambda m_F'(-\lambda)).$$

Marchenko-Pastur Law admits an explicit form under our model assumptions (see for example, (Bai et al., 2010, Page 52)) for any $z \in \mathbb{C}^+$ (which can be extended by analytic continuity arguments for any $z \in \mathbb{C}/\mathbb{R}^+$) and is given by

$$m_F(z) = \frac{1 - \gamma - z - \sqrt{(1 - \gamma - z)^2 - 4\gamma z}}{2\gamma z}.$$

Following arguments similar to Dobriban et al. (2018) and Hastie et al. (2022) for exchanging the limits $n, d \to \infty$ and $\lambda \to 0^+$, we can derive the limiting expressions for the causal bias and variance of the min-norm estimator.

□

C Asymptotics for the Statistical Risk

The following theorems describes the limiting expressions for the statistical risk analogue to the causal results from Theorem 3.1.

Theorem C.1 (Limiting Statistical Bias-Variance Decompositions). *Let $\hat{\beta}_0$ be the min-norm interpolator. Then as $n, d \to \infty$ such that $d/n \to \gamma \in (0, \infty)$, it holds almost surely in X that*

$$B_X^S(\hat{\beta}_0) \to \mathcal{B}_0^S = \begin{cases} 0, & \gamma < 1 \\ \tilde{r}^2(1 - \frac{1}{\gamma}), & \gamma > 1 \end{cases}, \quad V_X^S(\hat{\beta}_0) \to \mathcal{V}_0^S = \begin{cases} \tilde{\sigma}^2\frac{\gamma}{1-\gamma}, & \gamma < 1 \\ \tilde{\sigma}^2\frac{1}{\gamma-1}, & \gamma > 1 \end{cases} \tag{10}$$

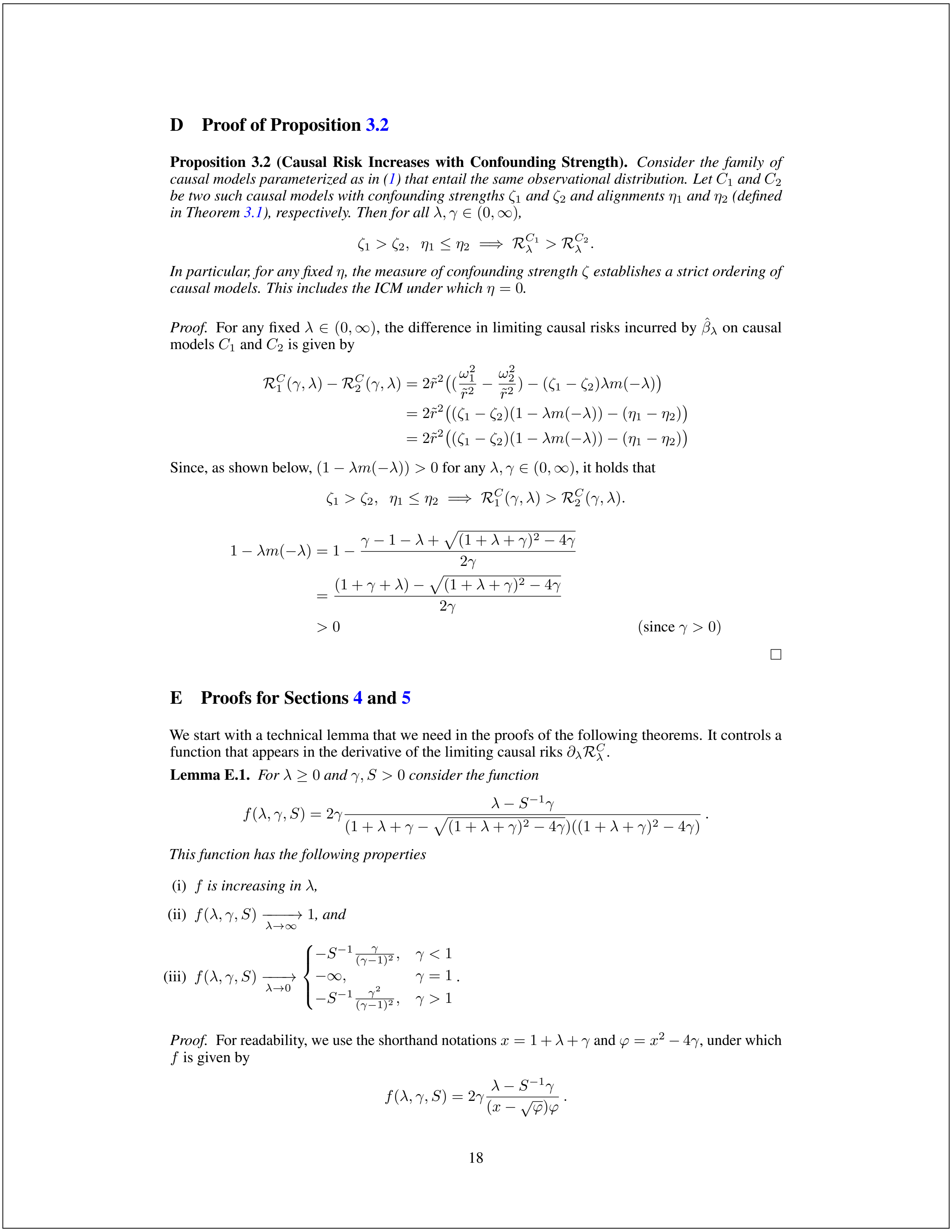
and therefore, $R_X^S(\hat{\beta}_0) \to \mathcal{R}_0^S = \mathcal{B}_0^S + \mathcal{V}_0^S + \tilde{\sigma}^2$.

For $\lambda > 0$ and the corresponding ridge regression estimator $\hat{\beta}_\lambda$, it holds almost surely in X that

$$B_X^S(\hat{\beta}_\lambda) \to \mathcal{B}_\lambda^S = \tilde{r}^2\lambda^2 m'(-\lambda)\,, \quad V_X^S(\hat{\beta}_\lambda) \to \mathcal{V}_\lambda^S = \tilde{\sigma}^2\gamma(m(-\lambda) - \lambda m'(-\lambda)), \tag{11}$$

where $m(\lambda) = \frac{(1-\gamma-\lambda)-\sqrt{(1-\gamma-\lambda)^2-4\gamma\lambda}}{2\gamma\lambda}$. Therefore, $R_X^S(\hat{\beta}_\lambda) \to \mathcal{R}_\lambda^S = \mathcal{B}_\lambda^S + \mathcal{V}_\lambda^S + \tilde{\sigma}^2$.

Proof. As stated in the main paper, this result for the statistical model was already proven in Hastie et al. (2022). □



D Proof of Proposition 3.2

Proposition 3.2 (Causal Risk Increases with Confounding Strength). *Consider the family of causal models parameterized as in (1) that entail the same observational distribution. Let C_1 and C_2 be two such causal models with confounding strengths ζ_1 and ζ_2 and alignments η_1 and η_2 (defined in Theorem 3.1), respectively. Then for all $\lambda, \gamma \in (0, \infty)$,*

$$\zeta_1 > \zeta_2, \quad \eta_1 \leq \eta_2 \implies \mathcal{R}_\lambda^{C_1} > \mathcal{R}_\lambda^{C_2}.$$

In particular, for any fixed η, the measure of confounding strength ζ establishes a strict ordering of causal models. This includes the ICM under which $\eta = 0$.

Proof. For any fixed $\lambda \in (0, \infty)$, the difference in limiting causal risks incurred by $\hat{\beta}_\lambda$ on causal models C_1 and C_2 is given by

$$\begin{aligned}
\mathcal{R}_1^C(\gamma,\lambda) - \mathcal{R}_2^C(\gamma,\lambda) &= 2\tilde{r}^2\big((\frac{\omega_1^2}{\tilde{r}^2} - \frac{\omega_2^2}{\tilde{r}^2}) - (\zeta_1 - \zeta_2)\lambda m(-\lambda)\big) \\
&= 2\tilde{r}^2\big((\zeta_1 - \zeta_2)(1 - \lambda m(-\lambda)) - (\eta_1 - \eta_2)\big) \\
&= 2\tilde{r}^2\big((\zeta_1 - \zeta_2)(1 - \lambda m(-\lambda)) - (\eta_1 - \eta_2)\big)
\end{aligned}$$

Since, as shown below, $(1 - \lambda m(-\lambda)) > 0$ for any $\lambda, \gamma \in (0, \infty)$, it holds that

$$\zeta_1 > \zeta_2, \quad \eta_1 \leq \eta_2 \implies \mathcal{R}_1^C(\gamma,\lambda) > \mathcal{R}_2^C(\gamma,\lambda).$$

$$\begin{aligned}
1 - \lambda m(-\lambda) &= 1 - \frac{\gamma - 1 - \lambda + \sqrt{(1+\lambda+\gamma)^2 - 4\gamma}}{2\gamma} \\
&= \frac{(1+\gamma+\lambda) - \sqrt{(1+\lambda+\gamma)^2 - 4\gamma}}{2\gamma} \\
&> 0 && \text{(since } \gamma > 0)
\end{aligned}$$

□

E Proofs for Sections 4 and 5

We start with a technical lemma that we need in the proofs of the following theorems. It controls a function that appears in the derivative of the limiting causal riks $\partial_\lambda \mathcal{R}_\lambda^C$.

Lemma E.1. *For $\lambda \geq 0$ and $\gamma, S > 0$ consider the function*

$$f(\lambda,\gamma,S) = 2\gamma \frac{\lambda - S^{-1}\gamma}{(1+\lambda+\gamma - \sqrt{(1+\lambda+\gamma)^2 - 4\gamma})((1+\lambda+\gamma)^2 - 4\gamma)}.$$

This function has the following properties

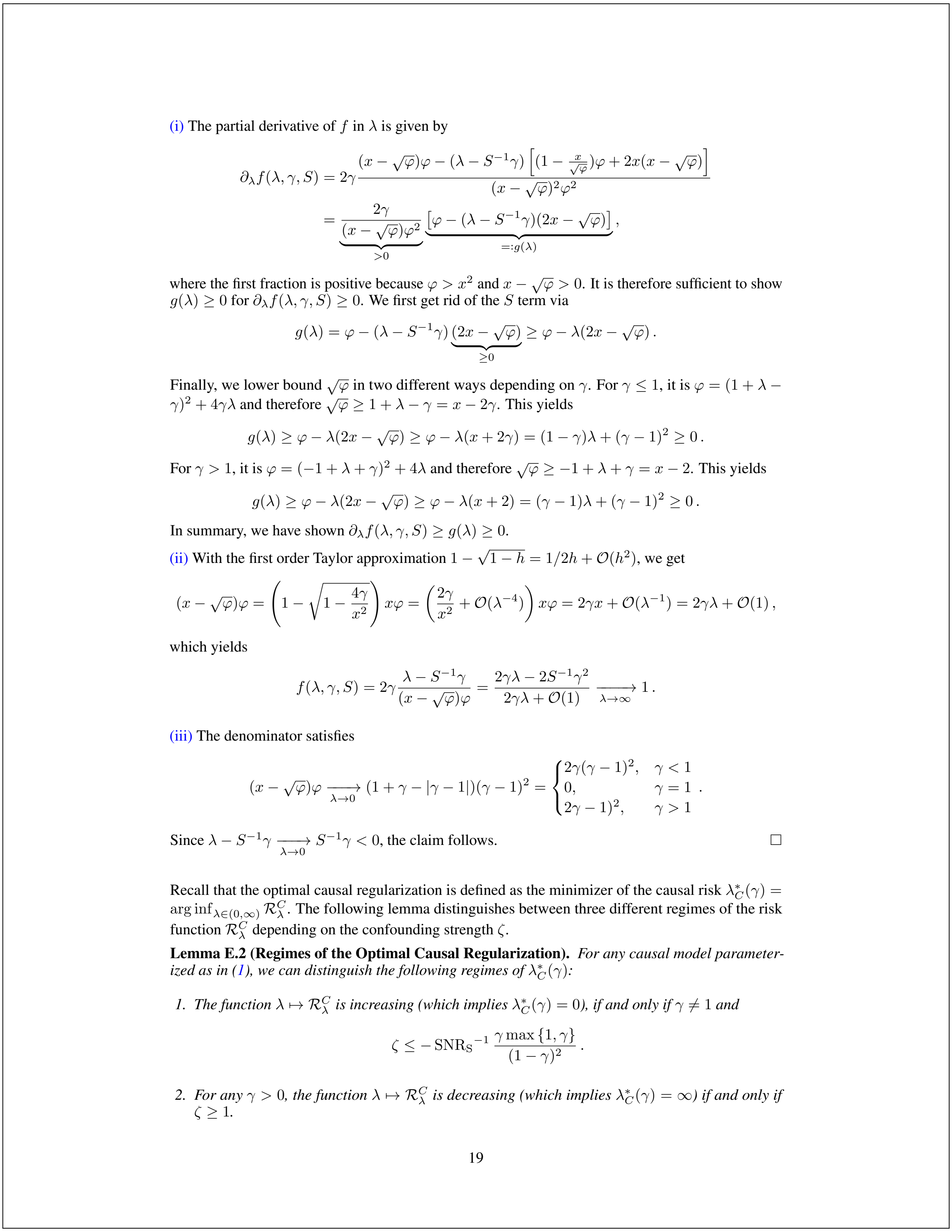
(i) *f is increasing in λ,*

(ii) *$f(\lambda,\gamma,S) \xrightarrow[\lambda\to\infty]{} 1$, and*

(iii) $f(\lambda,\gamma,S) \xrightarrow[\lambda\to 0]{} \begin{cases} -S^{-1}\frac{\gamma}{(\gamma-1)^2}, & \gamma < 1 \\ -\infty, & \gamma = 1 \\ -S^{-1}\frac{\gamma^2}{(\gamma-1)^2}, & \gamma > 1 \end{cases}.$

Proof. For readability, we use the shorthand notations $x = 1 + \lambda + \gamma$ and $\varphi = x^2 - 4\gamma$, under which f is given by

$$f(\lambda,\gamma,S) = 2\gamma \frac{\lambda - S^{-1}\gamma}{(x - \sqrt{\varphi})\varphi}.$$



(i) The partial derivative of f in λ is given by

$$\begin{aligned}\partial_\lambda f(\lambda,\gamma,S) &= 2\gamma\frac{(x-\sqrt{\varphi})\varphi-(\lambda-S^{-1}\gamma)\left[(1-\frac{x}{\sqrt{\varphi}})\varphi+2x(x-\sqrt{\varphi})\right]}{(x-\sqrt{\varphi})^2\varphi^2}\\ &= \underbrace{\frac{2\gamma}{(x-\sqrt{\varphi})\varphi^2}}_{>0}\underbrace{\left[\varphi-(\lambda-S^{-1}\gamma)(2x-\sqrt{\varphi})\right]}_{=:g(\lambda)},\end{aligned}$$

where the first fraction is positive because $\varphi > x^2$ and $x-\sqrt{\varphi}>0$. It is therefore sufficient to show $g(\lambda)\geq 0$ for $\partial_\lambda f(\lambda,\gamma,S)\geq 0$. We first get rid of the S term via

$$g(\lambda)=\varphi-(\lambda-S^{-1}\gamma)\underbrace{(2x-\sqrt{\varphi})}_{\geq 0}\geq\varphi-\lambda(2x-\sqrt{\varphi})\,.$$

Finally, we lower bound $\sqrt{\varphi}$ in two different ways depending on γ. For $\gamma\leq 1$, it is $\varphi=(1+\lambda-\gamma)^2+4\gamma\lambda$ and therefore $\sqrt{\varphi}\geq 1+\lambda-\gamma=x-2\gamma$. This yields

$$g(\lambda)\geq\varphi-\lambda(2x-\sqrt{\varphi})\geq\varphi-\lambda(x+2\gamma)=(1-\gamma)\lambda+(\gamma-1)^2\geq 0\,.$$

For $\gamma>1$, it is $\varphi=(-1+\lambda+\gamma)^2+4\lambda$ and therefore $\sqrt{\varphi}\geq -1+\lambda+\gamma=x-2$. This yields

$$g(\lambda)\geq\varphi-\lambda(2x-\sqrt{\varphi})\geq\varphi-\lambda(x+2)=(\gamma-1)\lambda+(\gamma-1)^2\geq 0\,.$$

In summary, we have shown $\partial_\lambda f(\lambda,\gamma,S)\geq g(\lambda)\geq 0$.

(ii) With the first order Taylor approximation $1-\sqrt{1-h}=1/2h+\mathcal{O}(h^2)$, we get

$$(x-\sqrt{\varphi})\varphi=\left(1-\sqrt{1-\frac{4\gamma}{x^2}}\right)x\varphi=\left(\frac{2\gamma}{x^2}+\mathcal{O}(\lambda^{-4})\right)x\varphi=2\gamma x+\mathcal{O}(\lambda^{-1})=2\gamma\lambda+\mathcal{O}(1)\,,$$

which yields

$$f(\lambda,\gamma,S)=2\gamma\frac{\lambda-S^{-1}\gamma}{(x-\sqrt{\varphi})\varphi}=\frac{2\gamma\lambda-2S^{-1}\gamma^2}{2\gamma\lambda+\mathcal{O}(1)}\xrightarrow[\lambda\to\infty]{}1\,.$$

(iii) The denominator satisfies

$$(x-\sqrt{\varphi})\varphi\xrightarrow[\lambda\to 0]{}(1+\gamma-|\gamma-1|)(\gamma-1)^2=\begin{cases}2\gamma(\gamma-1)^2, & \gamma<1\\ 0, & \gamma=1\\ 2\gamma-1)^2, & \gamma>1\end{cases}.$$

Since $\lambda-S^{-1}\gamma\xrightarrow[\lambda\to 0]{}S^{-1}\gamma<0$, the claim follows. □

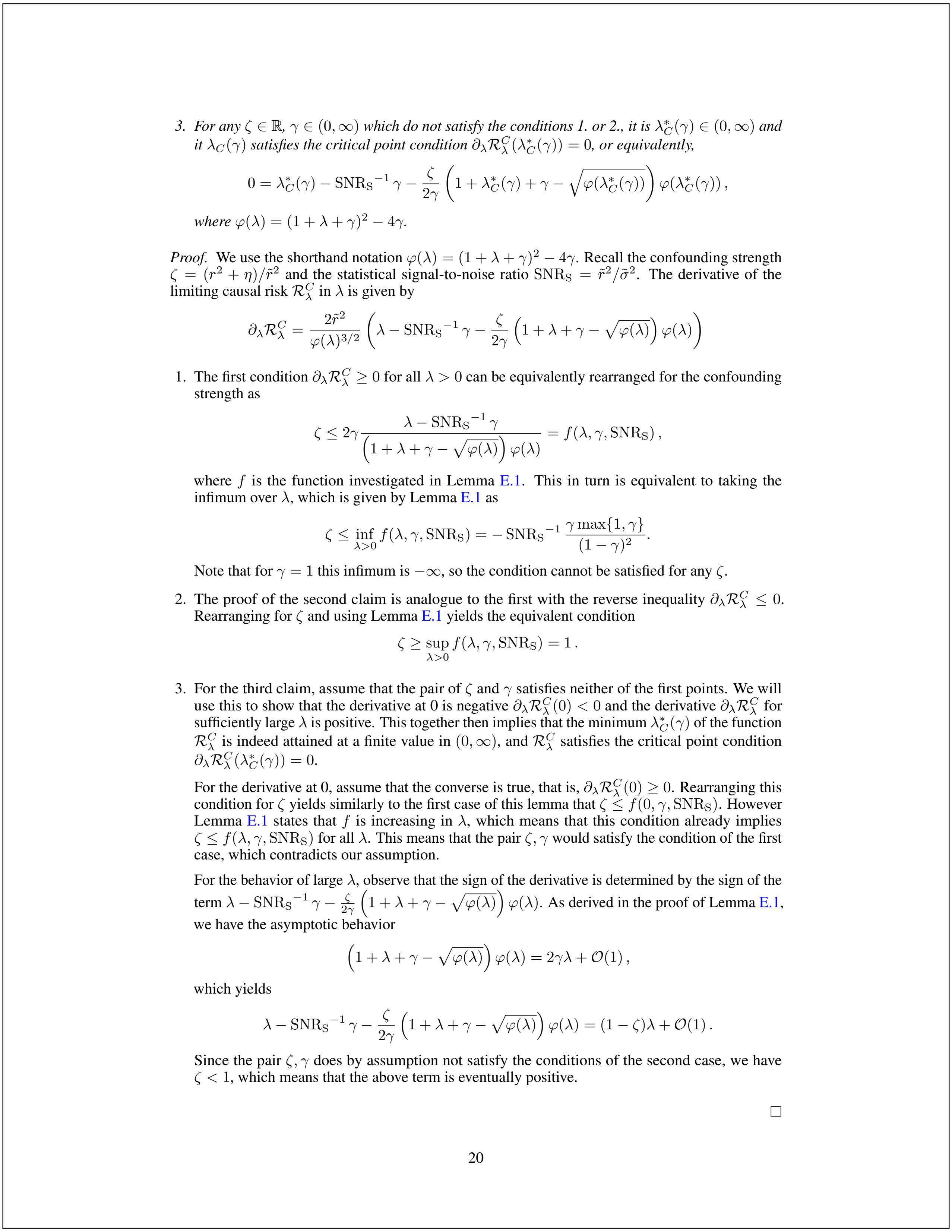
Recall that the optimal causal regularization is defined as the minimizer of the causal risk $\lambda^*_C(\gamma)=\arg\inf_{\lambda\in(0,\infty)}\mathcal{R}^C_\lambda$. The following lemma distinguishes between three different regimes of the risk function $\mathcal{R}^C_\lambda$ depending on the confounding strength ζ.

Lemma E.2 (Regimes of the Optimal Causal Regularization). *For any causal model parameterized as in (1), we can distinguish the following regimes of $\lambda^*_C(\gamma)$:*

1. *The function $\lambda\mapsto\mathcal{R}^C_\lambda$ is increasing (which implies $\lambda^*_C(\gamma)=0$), if and only if $\gamma\neq 1$ and*

$$\zeta\leq -\mathrm{SNR_S}^{-1}\frac{\gamma\max\{1,\gamma\}}{(1-\gamma)^2}\,.$$

2. *For any $\gamma>0$, the function $\lambda\mapsto\mathcal{R}^C_\lambda$ is decreasing (which implies $\lambda^*_C(\gamma)=\infty$) if and only if $\zeta\geq 1$.*



3. *For any $\zeta \in \mathbb{R}$, $\gamma \in (0, \infty)$ which do not satisfy the conditions 1. or 2., it is $\lambda^*_C(\gamma) \in (0, \infty)$ and it $\lambda_C(\gamma)$ satisfies the critical point condition $\partial_\lambda \mathcal{R}^C_\lambda(\lambda^*_C(\gamma)) = 0$, or equivalently,*

$$0 = \lambda^*_C(\gamma) - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda^*_C(\gamma) + \gamma - \sqrt{\varphi(\lambda^*_C(\gamma))}\right)\varphi(\lambda^*_C(\gamma))\,,$$

where $\varphi(\lambda) = (1 + \lambda + \gamma)^2 - 4\gamma$.

Proof. We use the shorthand notation $\varphi(\lambda) = (1 + \lambda + \gamma)^2 - 4\gamma$. Recall the confounding strength $\zeta = (r^2 + \eta)/\tilde{r}^2$ and the statistical signal-to-noise ratio $\mathrm{SNR_S} = \tilde{r}^2/\tilde{\sigma}^2$. The derivative of the limiting causal risk $\mathcal{R}^C_\lambda$ in λ is given by

$$\partial_\lambda \mathcal{R}^C_\lambda = \frac{2\tilde{r}^2}{\varphi(\lambda)^{3/2}}\left(\lambda - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda)\right)$$

1. The first condition $\partial_\lambda \mathcal{R}^C_\lambda \geq 0$ for all $\lambda > 0$ can be equivalently rearranged for the confounding strength as

$$\zeta \leq 2\gamma \frac{\lambda - \mathrm{SNR_S}^{-1}\gamma}{\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda)} = f(\lambda, \gamma, \mathrm{SNR_S})\,,$$

where f is the function investigated in Lemma E.1. This in turn is equivalent to taking the infimum over λ, which is given by Lemma E.1 as

$$\zeta \leq \inf_{\lambda > 0} f(\lambda, \gamma, \mathrm{SNR_S}) = -\,\mathrm{SNR_S}^{-1}\frac{\gamma \max\{1, \gamma\}}{(1 - \gamma)^2}.$$

Note that for $\gamma = 1$ this infimum is $-\infty$, so the condition cannot be satisfied for any ζ.

2. The proof of the second claim is analogue to the first with the reverse inequality $\partial_\lambda \mathcal{R}^C_\lambda \leq 0$. Rearranging for ζ and using Lemma E.1 yields the equivalent condition

$$\zeta \geq \sup_{\lambda > 0} f(\lambda, \gamma, \mathrm{SNR_S}) = 1\,.$$

3. For the third claim, assume that the pair of ζ and γ satisfies neither of the first points. We will use this to show that the derivative at 0 is negative $\partial_\lambda \mathcal{R}^C_\lambda(0) < 0$ and the derivative $\partial_\lambda \mathcal{R}^C_\lambda$ for sufficiently large λ is positive. This together then implies that the minimum $\lambda^*_C(\gamma)$ of the function $\mathcal{R}^C_\lambda$ is indeed attained at a finite value in $(0, \infty)$, and $\mathcal{R}^C_\lambda$ satisfies the critical point condition $\partial_\lambda \mathcal{R}^C_\lambda(\lambda^*_C(\gamma)) = 0$.

   For the derivative at 0, assume that the converse is true, that is, $\partial_\lambda \mathcal{R}^C_\lambda(0) \geq 0$. Rearranging this condition for ζ yields similarly to the first case of this lemma that $\zeta \leq f(0, \gamma, \mathrm{SNR_S})$. However Lemma E.1 states that f is increasing in λ, which means that this condition already implies $\zeta \leq f(\lambda, \gamma, \mathrm{SNR_S})$ for all λ. This means that the pair ζ, γ would satisfy the condition of the first case, which contradicts our assumption.

   For the behavior of large λ, observe that the sign of the derivative is determined by the sign of the term $\lambda - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda)$. As derived in the proof of Lemma E.1, we have the asymptotic behavior

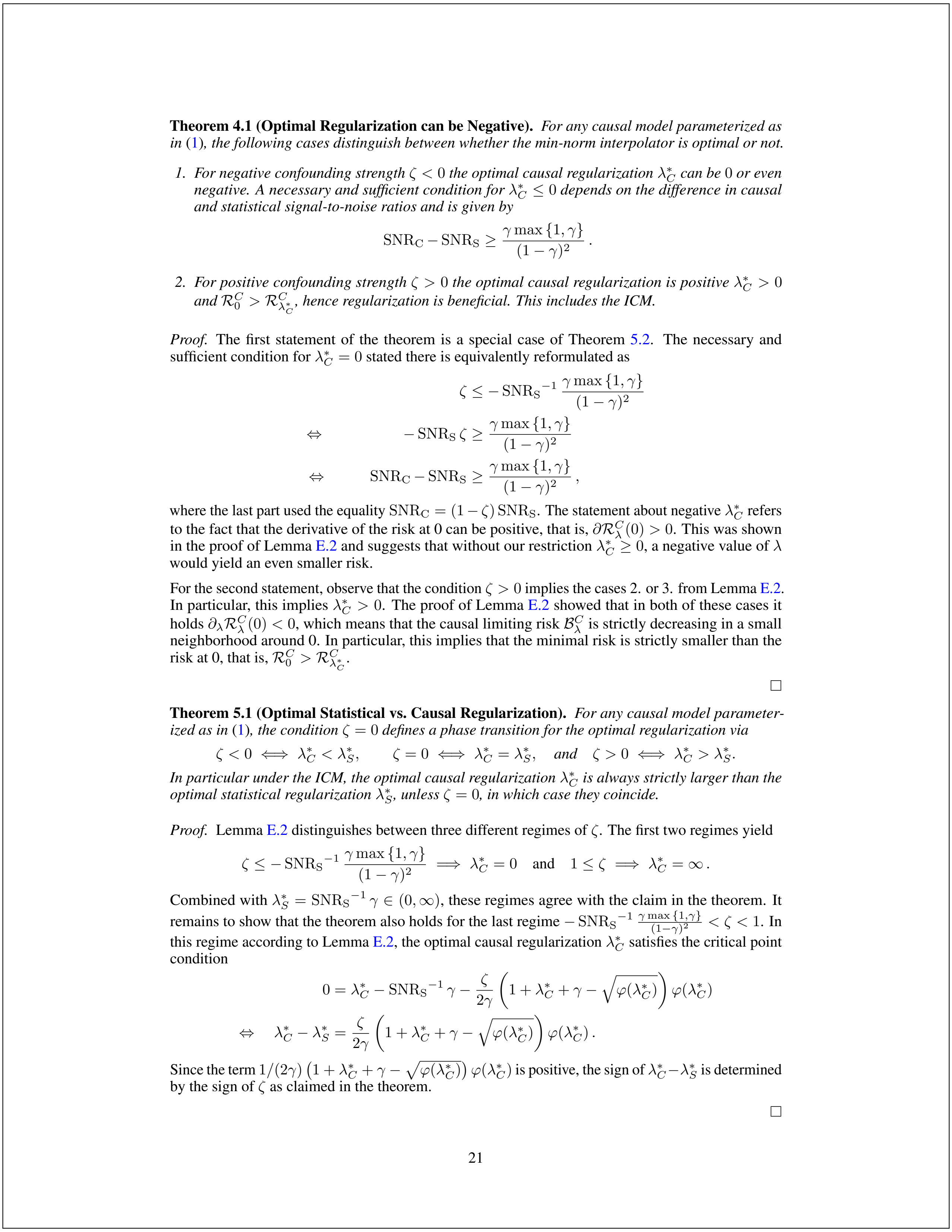
$$\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda) = 2\gamma\lambda + \mathcal{O}(1)\,,$$

   which yields

$$\lambda - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda) = (1 - \zeta)\lambda + \mathcal{O}(1)\,.$$

   Since the pair ζ, γ does by assumption not satisfy the conditions of the second case, we have $\zeta < 1$, which means that the above term is eventually positive.

□



Theorem 4.1 (Optimal Regularization can be Negative). *For any causal model parameterized as in* (1)*, the following cases distinguish between whether the min-norm interpolator is optimal or not.*

1. *For negative confounding strength $\zeta < 0$ the optimal causal regularization λ_C^* can be 0 or even negative. A necessary and sufficient condition for $\lambda_C^* \leq 0$ depends on the difference in causal and statistical signal-to-noise ratios and is given by*

$$\mathrm{SNR_C} - \mathrm{SNR_S} \geq \frac{\gamma \max\{1, \gamma\}}{(1-\gamma)^2}\,.$$

2. *For positive confounding strength $\zeta > 0$ the optimal causal regularization is positive $\lambda_C^* > 0$ and $\mathcal{R}_0^C > \mathcal{R}_{\lambda_C^*}^C$, hence regularization is beneficial. This includes the ICM.*

Proof. The first statement of the theorem is a special case of Theorem 5.2. The necessary and sufficient condition for $\lambda_C^* = 0$ stated there is equivalently reformulated as

$$\begin{aligned} && \zeta &\leq -\,\mathrm{SNR_S}^{-1} \frac{\gamma \max\{1, \gamma\}}{(1-\gamma)^2} \\ \Leftrightarrow && -\,\mathrm{SNR_S}\,\zeta &\geq \frac{\gamma \max\{1, \gamma\}}{(1-\gamma)^2} \\ \Leftrightarrow && \mathrm{SNR_C} - \mathrm{SNR_S} &\geq \frac{\gamma \max\{1, \gamma\}}{(1-\gamma)^2}\,, \end{aligned}$$

where the last part used the equality $\mathrm{SNR_C} = (1-\zeta)\,\mathrm{SNR_S}$. The statement about negative λ_C^* refers to the fact that the derivative of the risk at 0 can be positive, that is, $\partial\mathcal{R}_\lambda^C(0) > 0$. This was shown in the proof of Lemma E.2 and suggests that without our restriction $\lambda_C^* \geq 0$, a negative value of λ would yield an even smaller risk.

For the second statement, observe that the condition $\zeta > 0$ implies the cases 2. or 3. from Lemma E.2. In particular, this implies $\lambda_C^* > 0$. The proof of Lemma E.2 showed that in both of these cases it holds $\partial_\lambda \mathcal{R}_\lambda^C(0) < 0$, which means that the causal limiting risk $\mathcal{B}_\lambda^C$ is strictly decreasing in a small neighborhood around 0. In particular, this implies that the minimal risk is strictly smaller than the risk at 0, that is, $\mathcal{R}_0^C > \mathcal{R}_{\lambda_C^*}^C$.

□

Theorem 5.1 (Optimal Statistical vs. Causal Regularization). *For any causal model parameterized as in* (1)*, the condition $\zeta = 0$ defines a phase transition for the optimal regularization via*

$$\zeta < 0 \iff \lambda_C^* < \lambda_S^*, \qquad \zeta = 0 \iff \lambda_C^* = \lambda_S^*, \quad \textit{and} \quad \zeta > 0 \iff \lambda_C^* > \lambda_S^*.$$

In particular under the ICM, the optimal causal regularization λ_C^* is always strictly larger than the optimal statistical regularization λ_S^*, unless $\zeta = 0$, in which case they coincide.

Proof. Lemma E.2 distinguishes between three different regimes of ζ. The first two regimes yield

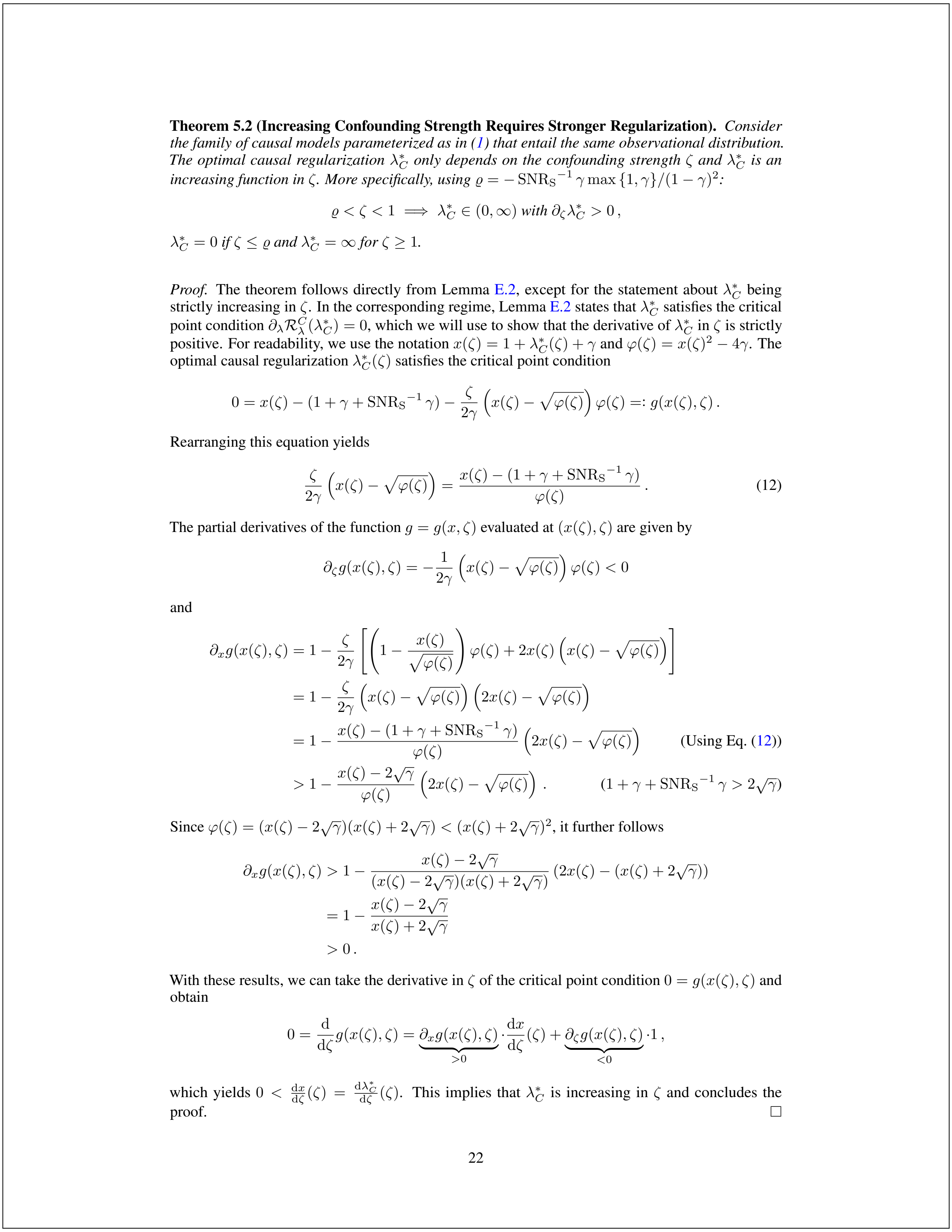
$$\zeta \leq -\,\mathrm{SNR_S}^{-1} \frac{\gamma \max\{1, \gamma\}}{(1-\gamma)^2} \implies \lambda_C^* = 0 \quad \text{and} \quad 1 \leq \zeta \implies \lambda_C^* = \infty\,.$$

Combined with $\lambda_S^* = \mathrm{SNR_S}^{-1}\gamma \in (0, \infty)$, these regimes agree with the claim in the theorem. It remains to show that the theorem also holds for the last regime $-\,\mathrm{SNR_S}^{-1} \frac{\gamma \max\{1,\gamma\}}{(1-\gamma)^2} < \zeta < 1$. In this regime according to Lemma E.2, the optimal causal regularization λ_C^* satisfies the critical point condition

$$\begin{aligned} && 0 &= \lambda_C^* - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda_C^* + \gamma - \sqrt{\varphi(\lambda_C^*)}\right)\varphi(\lambda_C^*) \\ \Leftrightarrow && \lambda_C^* - \lambda_S^* &= \frac{\zeta}{2\gamma}\left(1 + \lambda_C^* + \gamma - \sqrt{\varphi(\lambda_C^*)}\right)\varphi(\lambda_C^*)\,. \end{aligned}$$

Since the term $1/(2\gamma)\left(1 + \lambda_C^* + \gamma - \sqrt{\varphi(\lambda_C^*)}\right)\varphi(\lambda_C^*)$ is positive, the sign of $\lambda_C^* - \lambda_S^*$ is determined by the sign of ζ as claimed in the theorem.

□



Theorem 5.2 (Increasing Confounding Strength Requires Stronger Regularization). *Consider the family of causal models parameterized as in (1) that entail the same observational distribution. The optimal causal regularization λ_C^* only depends on the confounding strength ζ and λ_C^* is an increasing function in ζ. More specifically, using $\varrho = -\,\mathrm{SNR_S}^{-1}\,\gamma \max\{1,\gamma\}/(1-\gamma)^2$:*

$$\varrho < \zeta < 1 \implies \lambda_C^* \in (0,\infty) \textit{ with } \partial_\zeta \lambda_C^* > 0\,,$$

$\lambda_C^* = 0$ if $\zeta \leq \varrho$ and $\lambda_C^* = \infty$ for $\zeta \geq 1$.

Proof. The theorem follows directly from Lemma E.2, except for the statement about λ_C^* being strictly increasing in ζ. In the corresponding regime, Lemma E.2 states that λ_C^* satisfies the critical point condition $\partial_\lambda \mathcal{R}_\lambda^C(\lambda_C^*) = 0$, which we will use to show that the derivative of λ_C^* in ζ is strictly positive. For readability, we use the notation $x(\zeta) = 1 + \lambda_C^*(\zeta) + \gamma$ and $\varphi(\zeta) = x(\zeta)^2 - 4\gamma$. The optimal causal regularization $\lambda_C^*(\zeta)$ satisfies the critical point condition

$$0 = x(\zeta) - (1 + \gamma + \mathrm{SNR_S}^{-1}\,\gamma) - \frac{\zeta}{2\gamma}\left(x(\zeta) - \sqrt{\varphi(\zeta)}\right)\varphi(\zeta) =: g(x(\zeta),\zeta)\,.$$

Rearranging this equation yields

$$\frac{\zeta}{2\gamma}\left(x(\zeta) - \sqrt{\varphi(\zeta)}\right) = \frac{x(\zeta) - (1 + \gamma + \mathrm{SNR_S}^{-1}\,\gamma)}{\varphi(\zeta)}\,. \tag{12}$$

The partial derivatives of the function $g = g(x,\zeta)$ evaluated at $(x(\zeta),\zeta)$ are given by

$$\partial_\zeta g(x(\zeta),\zeta) = -\frac{1}{2\gamma}\left(x(\zeta) - \sqrt{\varphi(\zeta)}\right)\varphi(\zeta) < 0$$

and

$$\begin{aligned}
\partial_x g(x(\zeta),\zeta) &= 1 - \frac{\zeta}{2\gamma}\left[\left(1 - \frac{x(\zeta)}{\sqrt{\varphi(\zeta)}}\right)\varphi(\zeta) + 2x(\zeta)\left(x(\zeta) - \sqrt{\varphi(\zeta)}\right)\right] \\
&= 1 - \frac{\zeta}{2\gamma}\left(x(\zeta) - \sqrt{\varphi(\zeta)}\right)\left(2x(\zeta) - \sqrt{\varphi(\zeta)}\right) \\
&= 1 - \frac{x(\zeta) - (1 + \gamma + \mathrm{SNR_S}^{-1}\,\gamma)}{\varphi(\zeta)}\left(2x(\zeta) - \sqrt{\varphi(\zeta)}\right) && \text{(Using Eq. (12))} \\
&> 1 - \frac{x(\zeta) - 2\sqrt{\gamma}}{\varphi(\zeta)}\left(2x(\zeta) - \sqrt{\varphi(\zeta)}\right)\,. && (1 + \gamma + \mathrm{SNR_S}^{-1}\,\gamma > 2\sqrt{\gamma})
\end{aligned}$$

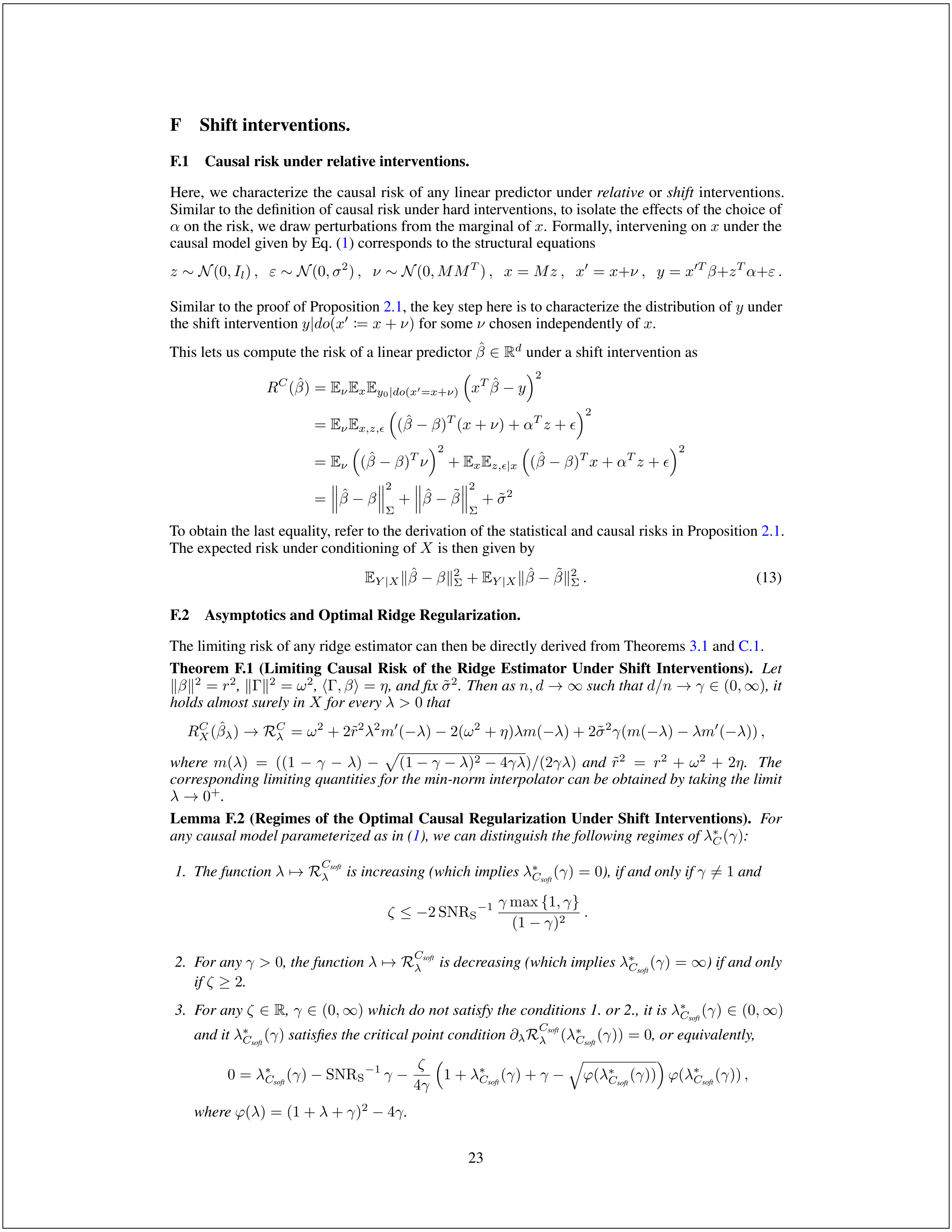
Since $\varphi(\zeta) = (x(\zeta) - 2\sqrt{\gamma})(x(\zeta) + 2\sqrt{\gamma}) < (x(\zeta) + 2\sqrt{\gamma})^2$, it further follows

$$\begin{aligned}
\partial_x g(x(\zeta),\zeta) &> 1 - \frac{x(\zeta) - 2\sqrt{\gamma}}{(x(\zeta) - 2\sqrt{\gamma})(x(\zeta) + 2\sqrt{\gamma})}\left(2x(\zeta) - (x(\zeta) + 2\sqrt{\gamma})\right) \\
&= 1 - \frac{x(\zeta) - 2\sqrt{\gamma}}{x(\zeta) + 2\sqrt{\gamma}} \\
&> 0\,.
\end{aligned}$$

With these results, we can take the derivative in ζ of the critical point condition $0 = g(x(\zeta),\zeta)$ and obtain

$$0 = \frac{\mathrm{d}}{\mathrm{d}\zeta} g(x(\zeta),\zeta) = \underbrace{\partial_x g(x(\zeta),\zeta)}_{>0} \cdot \frac{\mathrm{d}x}{\mathrm{d}\zeta}(\zeta) + \underbrace{\partial_\zeta g(x(\zeta),\zeta)}_{<0} \cdot 1\,,$$

which yields $0 < \frac{\mathrm{d}x}{\mathrm{d}\zeta}(\zeta) = \frac{\mathrm{d}\lambda_C^*}{\mathrm{d}\zeta}(\zeta)$. This implies that λ_C^* is increasing in ζ and concludes the proof. □



F Shift interventions.

F.1 Causal risk under relative interventions.

Here, we characterize the causal risk of any linear predictor under *relative* or *shift* interventions. Similar to the definition of causal risk under hard interventions, to isolate the effects of the choice of α on the risk, we draw perturbations from the marginal of x. Formally, intervening on x under the causal model given by Eq. (1) corresponds to the structural equations

$$z \sim \mathcal{N}(0, I_l)\,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,, \quad \nu \sim \mathcal{N}(0, MM^T)\,, \quad x = Mz\,, \quad x' = x + \nu\,, \quad y = x'^T\beta + z^T\alpha + \varepsilon\,.$$

Similar to the proof of Proposition 2.1, the key step here is to characterize the distribution of y under the shift intervention $y|do(x' := x + \nu)$ for some ν chosen independently of x.

This lets us compute the risk of a linear predictor $\hat{\beta} \in \mathbb{R}^d$ under a shift intervention as

$$\begin{aligned} R^C(\hat{\beta}) &= \mathbb{E}_\nu \mathbb{E}_x \mathbb{E}_{y_0|do(x'=x+\nu)} \left(x^T\hat{\beta} - y\right)^2 \\ &= \mathbb{E}_\nu \mathbb{E}_{x,z,\epsilon} \left((\hat{\beta} - \beta)^T(x + \nu) + \alpha^T z + \epsilon\right)^2 \\ &= \mathbb{E}_\nu \left((\hat{\beta} - \beta)^T \nu\right)^2 + \mathbb{E}_x \mathbb{E}_{z,\epsilon|x} \left((\hat{\beta} - \beta)^T x + \alpha^T z + \epsilon\right)^2 \\ &= \left\|\hat{\beta} - \beta\right\|_\Sigma^2 + \left\|\hat{\beta} - \tilde{\beta}\right\|_\Sigma^2 + \tilde{\sigma}^2 \end{aligned}$$

To obtain the last equality, refer to the derivation of the statistical and causal risks in Proposition 2.1. The expected risk under conditioning of X is then given by

$$\mathbb{E}_{Y|X}\|\hat{\beta} - \beta\|_\Sigma^2 + \mathbb{E}_{Y|X}\|\hat{\beta} - \tilde{\beta}\|_\Sigma^2\,. \tag{13}$$

F.2 Asymptotics and Optimal Ridge Regularization.

The limiting risk of any ridge estimator can then be directly derived from Theorems 3.1 and C.1.

Theorem F.1 (Limiting Causal Risk of the Ridge Estimator Under Shift Interventions). *Let $\|\beta\|^2 = r^2$, $\|\Gamma\|^2 = \omega^2$, $\langle\Gamma, \beta\rangle = \eta$, and fix $\tilde{\sigma}^2$. Then as $n, d \to \infty$ such that $d/n \to \gamma \in (0, \infty)$, it holds almost surely in X for every $\lambda > 0$ that*

$$R_X^C(\hat{\beta}_\lambda) \to \mathcal{R}_\lambda^C = \omega^2 + 2\tilde{r}^2\lambda^2 m'(-\lambda) - 2(\omega^2 + \eta)\lambda m(-\lambda) + 2\tilde{\sigma}^2\gamma(m(-\lambda) - \lambda m'(-\lambda))\,,$$

where $m(\lambda) = ((1 - \gamma - \lambda) - \sqrt{(1 - \gamma - \lambda)^2 - 4\gamma\lambda})/(2\gamma\lambda)$ and $\tilde{r}^2 = r^2 + \omega^2 + 2\eta$. The corresponding limiting quantities for the min-norm interpolator can be obtained by taking the limit $\lambda \to 0^+$.

Lemma F.2 (Regimes of the Optimal Causal Regularization Under Shift Interventions). *For any causal model parameterized as in (1), we can distinguish the following regimes of $\lambda_C^*(\gamma)$:*

1. *The function $\lambda \mapsto \mathcal{R}_\lambda^{C_{soft}}$ is increasing (which implies $\lambda_{C_{soft}}^*(\gamma) = 0$), if and only if $\gamma \neq 1$ and*

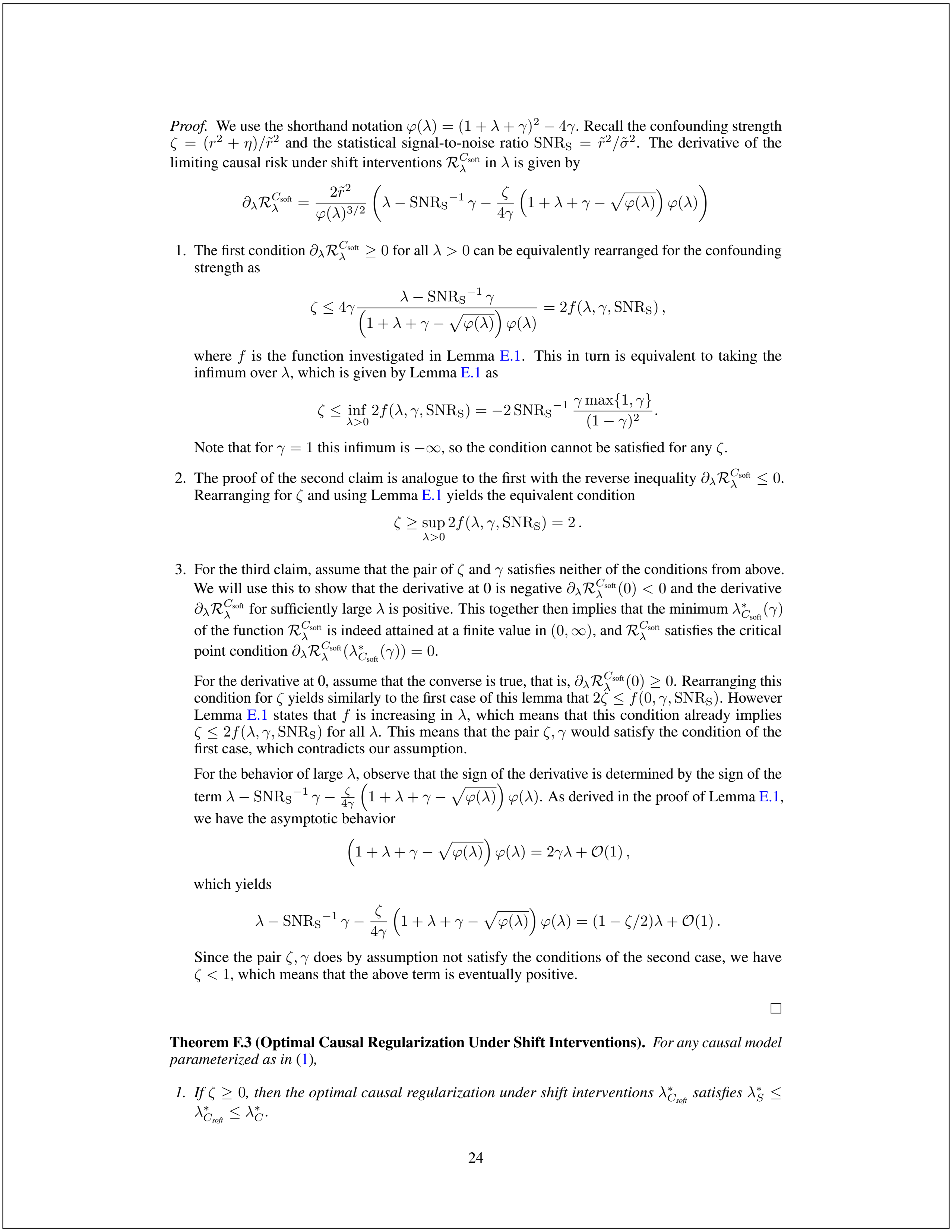
   $$\zeta \leq -2\,\mathrm{SNR_S}^{-1}\,\frac{\gamma \max\{1, \gamma\}}{(1 - \gamma)^2}\,.$$

2. *For any $\gamma > 0$, the function $\lambda \mapsto \mathcal{R}_\lambda^{C_{soft}}$ is decreasing (which implies $\lambda_{C_{soft}}^*(\gamma) = \infty$) if and only if $\zeta \geq 2$.*

3. *For any $\zeta \in \mathbb{R}$, $\gamma \in (0, \infty)$ which do not satisfy the conditions 1. or 2., it is $\lambda_{C_{soft}}^*(\gamma) \in (0, \infty)$ and it $\lambda_{C_{soft}}^*(\gamma)$ satisfies the critical point condition $\partial_\lambda \mathcal{R}_\lambda^{C_{soft}}(\lambda_{C_{soft}}^*(\gamma)) = 0$, or equivalently,*

   $$0 = \lambda_{C_{soft}}^*(\gamma) - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{4\gamma}\left(1 + \lambda_{C_{soft}}^*(\gamma) + \gamma - \sqrt{\varphi(\lambda_{C_{soft}}^*(\gamma))}\right)\varphi(\lambda_{C_{soft}}^*(\gamma))\,,$$

   where $\varphi(\lambda) = (1 + \lambda + \gamma)^2 - 4\gamma$.



Proof. We use the shorthand notation $\varphi(\lambda) = (1+\lambda+\gamma)^2 - 4\gamma$. Recall the confounding strength $\zeta = (r^2+\eta)/\tilde{r}^2$ and the statistical signal-to-noise ratio $\mathrm{SNR_S} = \tilde{r}^2/\tilde{\sigma}^2$. The derivative of the limiting causal risk under shift interventions $\mathcal{R}_\lambda^{C_{\text{soft}}}$ in λ is given by

$$\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}} = \frac{2\tilde{r}^2}{\varphi(\lambda)^{3/2}} \left(\lambda - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{4\gamma}\left(1+\lambda+\gamma-\sqrt{\varphi(\lambda)}\right)\varphi(\lambda) \right)$$

1. The first condition $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}} \geq 0$ for all $\lambda > 0$ can be equivalently rearranged for the confounding strength as

$$\zeta \leq 4\gamma \frac{\lambda - \mathrm{SNR_S}^{-1}\gamma}{\left(1+\lambda+\gamma-\sqrt{\varphi(\lambda)}\right)\varphi(\lambda)} = 2f(\lambda,\gamma,\mathrm{SNR_S}),$$

where f is the function investigated in Lemma E.1. This in turn is equivalent to taking the infimum over λ, which is given by Lemma E.1 as

$$\zeta \leq \inf_{\lambda>0} 2f(\lambda,\gamma,\mathrm{SNR_S}) = -2\,\mathrm{SNR_S}^{-1}\frac{\gamma\max\{1,\gamma\}}{(1-\gamma)^2}.$$

Note that for $\gamma = 1$ this infimum is $-\infty$, so the condition cannot be satisfied for any ζ.

2. The proof of the second claim is analogue to the first with the reverse inequality $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}} \leq 0$. Rearranging for ζ and using Lemma E.1 yields the equivalent condition

$$\zeta \geq \sup_{\lambda>0} 2f(\lambda,\gamma,\mathrm{SNR_S}) = 2\,.$$

3. For the third claim, assume that the pair of ζ and γ satisfies neither of the conditions from above. We will use this to show that the derivative at 0 is negative $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}}(0) < 0$ and the derivative $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}}$ for sufficiently large λ is positive. This together then implies that the minimum $\lambda^*_{C_{\text{soft}}}(\gamma)$ of the function $\mathcal{R}_\lambda^{C_{\text{soft}}}$ is indeed attained at a finite value in $(0,\infty)$, and $\mathcal{R}_\lambda^{C_{\text{soft}}}$ satisfies the critical point condition $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}}(\lambda^*_{C_{\text{soft}}}(\gamma)) = 0$.

   For the derivative at 0, assume that the converse is true, that is, $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}}(0) \geq 0$. Rearranging this condition for ζ yields similarly to the first case of this lemma that $2\zeta \leq f(0,\gamma,\mathrm{SNR_S})$. However Lemma E.1 states that f is increasing in λ, which means that this condition already implies $\zeta \leq 2f(\lambda,\gamma,\mathrm{SNR_S})$ for all λ. This means that the pair ζ,γ would satisfy the condition of the first case, which contradicts our assumption.

   For the behavior of large λ, observe that the sign of the derivative is determined by the sign of the term $\lambda - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{4\gamma}\left(1+\lambda+\gamma-\sqrt{\varphi(\lambda)}\right)\varphi(\lambda)$. As derived in the proof of Lemma E.1, we have the asymptotic behavior

$$\left(1+\lambda+\gamma-\sqrt{\varphi(\lambda)}\right)\varphi(\lambda) = 2\gamma\lambda + \mathcal{O}(1)\,,$$

   which yields

$$\lambda - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{4\gamma}\left(1+\lambda+\gamma-\sqrt{\varphi(\lambda)}\right)\varphi(\lambda) = (1-\zeta/2)\lambda + \mathcal{O}(1)\,.$$

   Since the pair ζ,γ does by assumption not satisfy the conditions of the second case, we have $\zeta < 1$, which means that the above term is eventually positive.

□

Theorem F.3 (Optimal Causal Regularization Under Shift Interventions). *For any causal model parameterized as in* (1),

1. *If* $\zeta \geq 0$, *then the optimal causal regularization under shift interventions* $\lambda^*_{C_{soft}}$ *satisfies* $\lambda^*_S \leq \lambda^*_{C_{soft}} \leq \lambda^*_C$.

2. *If* $\zeta < 0$, *then* $\lambda_C^* \leq \lambda_{C_{soft}}^* \leq \lambda_S^*$.

Indeed, the optimal causal regularization under shift interventions satisfies $\lambda_{C_{soft}}^* = \lambda_S^* + (\lambda_C^* - \lambda_S^*)/2$.

Proof. Lemma E.2 distinguishes between three different regimes of ζ. The first two regimes yield

$$\zeta \leq -2\,\mathrm{SNR_S}^{-1}\,\frac{\gamma \max\{1,\gamma\}}{(1-\gamma)^2} \implies \lambda_C^* = 0 \quad \text{and} \quad 2 \leq \zeta \implies \lambda_C^* = \infty\,.$$

Combined with $\lambda_S^* = \mathrm{SNR_S}^{-1}\gamma \in (0,\infty)$, these regimes agree with the claim in the theorem. It remains to show that the theorem also holds for the last regime $-2\,\mathrm{SNR_S}^{-1}\frac{\gamma\max\{1,\gamma\}}{(1-\gamma)^2} < \zeta < 2$. In this regime according to Lemma E.2, the optimal causal regularization λ_C^* satisfies the critical point condition

$$\begin{aligned} 0 &= \lambda_{C_\mathrm{soft}}^* - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{4\gamma}\left(1 + \lambda_{C_\mathrm{soft}}^* + \gamma - \sqrt{\varphi(\lambda_{C_\mathrm{soft}}^*)}\right)\varphi(\lambda_{C_\mathrm{soft}}^*) \\ \Leftrightarrow \quad \lambda_{C_\mathrm{soft}}^* - \lambda_S^* &= \frac{\zeta}{4\gamma}\left(1 + \lambda_{C_\mathrm{soft}}^* + \gamma - \sqrt{\varphi(\lambda_{C_\mathrm{soft}}^*)}\right)\varphi(\lambda_{C_\mathrm{soft}}^*)\,. \end{aligned}$$

Similarly, we know from the proof of Theorem 5.1 λ_C^* satisfies

$$\begin{aligned} 0 &= \lambda_C^* - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda_C^* + \gamma - \sqrt{\varphi(\lambda_C^*)}\right)\varphi(\lambda_C^*) \\ \Leftrightarrow \quad \lambda_C^* - \lambda_{C_\mathrm{soft}}^* &= \frac{\zeta}{4\gamma}\left(1 + \lambda_{C_\mathrm{soft}}^* + \gamma - \sqrt{\varphi(\lambda_{C_\mathrm{soft}}^*)}\right)\varphi(\lambda_{C_\mathrm{soft}}^*)\,. \end{aligned}$$

Since the term $1/(2\gamma)\left(1 + \lambda_{C_\mathrm{soft}}^* + \gamma - \sqrt{\varphi(\lambda_{C_\mathrm{soft}}^*)}\right)\varphi(\lambda_{C_\mathrm{soft}}^*)$ is positive, the sign of $\lambda_{C_\mathrm{soft}}^* - \lambda_S^*$ and $\lambda_C^* - \lambda_{C_\mathrm{soft}}^*$ is determined by the sign of ζ as claimed in the theorem. □

G Beyond Gaussianity

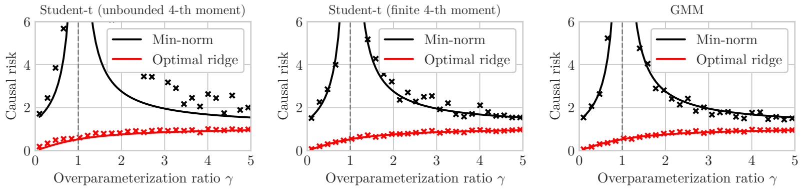


Figure 5: Causal risk of the minimum norm l_2 interpolator and the (causally)optimally regularized ridge regressor under a student-t distribution with unbounded 4th moments (3 degrees of freedom, left), a student-t distribution with bounded 4th moments (10 degrees of freedom, middle), a mixture of Gaussians (right). We choose the parameters $d = 300, l = 350$, statistical signal $\tilde{r}^2 = 5$, statistical noise $\tilde{\sigma}^2 = 1$, causal noise $\sigma^2 = .5$ and confounding strength $\zeta = 0.5$. For Gaussian mixtures, we consider a (centered and normalized) mixture of $k = 5$ Gaussians. Each individual mixture component has mean $\mu_i \sim \mathcal{N}(0_l, \frac{k^2}{(k-1)l} I_l)$ and identity covariance $\mathrm{Cov}_i = I_l$.

The analysis of this paper can be extended beyond the Gaussian setting by considering random variables generated by finite mixtures of Gaussians. The analysis can get considerably more technical and is left as future work, but we include a brief discussion here. Due to the Universality phenomenon in the high-dimensional limit, we believe that our limiting expressions (and the qualitative messages derived henceforth) would be rather robust to shifts in the marginal distribution as long as moments of order $(4+\delta)$ for some $\delta > 0$ are bounded. We conducted experiments to verify this claim and the corresponding results can be found in Figure 5. These experiments compare our theoretically derived asymptotic risks with finite-sample risks of the min-norm interpolator and causally optimally regularized ridge regressor. Instead of Gaussian confounders $z \sim \mathcal{N}(0, I_l)$, we only fix the first two moments 0 and I_l and generate z such that $\mathbb{E}[z] = 0$, $\mathrm{Cov}[z] = I$ from heavy-tailed multivariate t-distribution with different degrees of freedom, and finite mixture of Gaussians. Each plot shows the causal risk of min-norm interpolator and optimally regularized ridge regressor based on finite samples along with our theoretical asymptotic predictions. Our experiments show that, for distributions with finite 4th moments, the finite-sample risks closely match the theoretical results.

Part III

Discussion

10
Discussion

Many problems in modern machine learning, including fairness and robustness to domain shifts, interventions, and adversarial examples, fall under the umbrella of learning under extreme non-identifiability.[1] The central challenge to *theoretically justified* learning under extreme non-identifiability lies in finding a set of assumptions that is sufficiently general, practically meaningful, and amenable to theoretical analysis. This is an extremely challenging task.

For example, in the theory of clustering under a mixture model, formulating such a set of assumptions is incredibly hard. While parametric assumptions such as the Gaussianity of the components yield strong theoretical guarantees, the generality and practical relevance of such assumptions to real-world settings are limited. On the other hand, while non-parametric assumptions are sufficiently general, the resulting bounds are weak, thereby limiting the practical relevance of such assumptions. Other assumptions in the theory of kernel clustering include assuming a mixture model on the reproducing kernel Hilbert space (RKHS) of the kernel [Chen and Yang, 2021]. However, it is unclear how such assumptions translate into assumptions on the class of data distributions, thereby hindering the practical applicability of such results. Assumptions such as the stability of the clusters have been thought to offer certain useful generality. However, the practical utility of such measures has been shown to be rather limited [Ben-David et al., 2006]. These examples emphasize the difficulty of finding a set of meaningful assumptions in clustering. This is the primary hurdle to developing a taxonomy of clustering methods long advocated by many experts on clustering [Von Luxburg et al., 2012]. While understanding the theoretical properties of clustering methods under these assumptions is essential, there is certainly room for understanding the performance of these methods under a general and practically relevant set of assumptions.[2]

Similar challenges are encountered in the theory of causal learning. On a positive note, the principle of independence of causal mechanisms offers a powerful bias for learning causal models [Peters et al., 2017]. For example, in one of our papers [not included in this thesis], we show that a mathematical instantiation of the ICM

[1] A lot of these problems have indeed been known to be closely related to the problem of learning causal relationships [Schölkopf et al., 2021].

[2] The framework of density-based clustering[Hartigan, 1981, Chaudhuri et al., 2014] does offer such a meaningful formalism. However, the framework suffers from limitations, as we discussed in Chapter 4. Moreover, there is no single framework for clustering that can work for all practical applications, and it is necessary to study the performance of clustering methods under more general formalisms.

principle allows for partial identification of the causal parameter purely from observational data in high dimensions, even in the presence of latent confounding [Rendsburg et al., 2022]. However, the space of possible assumptions that enable causal learning from observational data is still very much open. Under such assumptions, it would be very interesting to develop the foundations for a *learning theory of causality* (akin to statistical learning theory) that holds for rather general hypothesis classes.

Bibliography

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Efren F Abaya and Gary L Wise. Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal on Applied Mathematics*, 44(1):183–189, 1984.

Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

Martin Anthony, Peter L Bartlett, Peter L Bartlett, et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

András Antos. Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Transactions on Information Theory*, 51(11):4022–4032, 2005.

András Antos, László Gyorfi, and Andras Gyorgy. Individual convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory*, 51(11):4013–4022, 2005.

Bryon Aragam, Chen Dan, Eric P Xing, Pradeep Ravikumar, et al. Identifiability of nonparametric mixture models and bayes optimal clustering. *The Annals of Statistics*, 48(4):2277–2302, 2020.

Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In *Advances in Neural Information Processing Systems*, pages 3412–3421, 2018.

Olivier Bachem, Mario Lucic, S Hamed Hassani, and Andreas Krause. Uniform deviation bounds for k-means clustering. In *International Conference on Machine Learning*, pages 283–291. PMLR, 2017.

Sivaraman Balakrishnan, Srivatsan Narayanan, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Cluster trees on manifolds. *Advances in Neural Information Processing Systems*, 26, 2013.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, Joydeep Ghosh, and John Lafferty. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005.

Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894, 2018.

Peter Bartlett and Mikhail Traskin. Adaboost is consistent. *Advances in Neural Information Processing Systems*, 19, 2006.

Peter Bartlett, Sébastien Bubeck, and Yeshwanth Cherapanamjeri. Adversarial examples in multi-layer random relu networks. *Advances in Neural Information Processing Systems*, 34:9241–9252, 2021.

Peter L Bartlett and Philip M Long. Failures of model-dependent generalization bounds for least-norm interpolation. *J. Mach. Learn. Res.*, 22:204–1, 2021.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Tamás Linder, and Gábor Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information theory*, 44(5):1802–1813, 1998.

Peter L Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 2021.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning (ICML)*, 2018.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Alexandre Belloni, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Shai Ben-David and Margareta Ackerman. Measures of clustering quality: A working set of axioms for clustering. *Advances in neural information processing systems*, 21, 2008.

Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In *International conference on computational learning theory*, pages 5–19. Springer, 2006.

Gérard Biau, Luc Devroye, and Gábor Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

Valerio Biscione and Jeffrey S Bowers. Learning translation invariance in cnns. In *Neural Information Processing Systems 2020: Shared Visual Representations in Human and Machine Intelligence*. 2020.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer, 2003.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

Claire Brécheteau, Aurélie Fischer, and Clément Levrard. Robust bregman clustering. *The Annals of Statistics*, 49(3):1679–1701, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Carlo Bruni and Giorgio Koch. Identifiability of continuous mixtures of unknown gaussian distributions. *The Annals of Probability*, pages 1341–1357, 1985.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822, 2021.

Daniele Calandriello and Lorenzo Rosasco. Statistical and computational trade-offs in kernel k-means. *Advances in neural information processing systems*, 31, 2018.

Daniele Calandriello, Alessandro Lazaric, and Michal Valko. Efficient second-order online kernel learning with adaptive embedding. *Advances in Neural Information Processing Systems*, 30, 2017.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

Saptarshi Chakraborty and Swagatam Das. Detecting meaningful clusters from high-dimensional data: A strongly consistent sparse center-based clustering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Kamalika Chaudhuri, Sanjoy Dasgupta, Samory Kpotufe, and Ulrike Von Luxburg. Consistent procedures for cluster tree estimation and pruning. *IEEE Transactions on Information Theory*, 60 (12):7900–7912, 2014.

Xiaohong Chen and Timothy M Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018.

Xiaohui Chen and Yun Yang. Hanson–wright inequality in hilbert spaces with application to k-means clustering for non-euclidean data. *Bernoulli*, 27(1):586–614, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. DOI: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Philip A Chou. The distortion of vector quantizers trained on n vectors decreases to the optimum as o/sub p/(1/n). In *Proceedings of 1994 IEEE International Symposium on Information Theory*, page 457. IEEE, 1994.

Benoît Collins, Sushma Kumari, and Vladimir G Pestov. Universal consistency of the k-nn rule in metric spaces and nagata dimension. *ESAIM: Probability and Statistics*, 24:914–934, 2020.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Gabriela Csurka et al. *Domain adaptation in computer vision applications*. Springer, 2017.

Antonio Cuevas and Ricardo Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, pages 2300–2312, 1997.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science*, pages 634–644. IEEE, 1999.

Luc Devroye and T Wagner. Distribution-free performance bounds with the resubstitution error estimate (corresp.). *IEEE Transactions on Information Theory*, 25(2):208–210, 1979a.

Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979b.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 551–556, 2004.

Hossein Esfandiari, Vahab Mirrokni, and Peilin Zhong. Almost linear time density level set estimation via dbscan. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7349–7357, 2021.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930): 47–53, 2022.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1270–1279. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/feldman19a.html.

Junhao Gan and Yufei Tao. Dbscan revisited: Mis-claim, unfixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 519–530, 2015.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, 21(1):185–194, 1999.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Siegfried Graf and Harald Luschgy. *Foundations of quantization for probability distributions*. Springer, 2007.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. *arXiv preprint arXiv:1806.00468*, 2018a.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018b.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

John A Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374): 388–394, 1981.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.

Leah Henderson. The Problem of Induction. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition, 2022.

Hajo Holzmann, Axel Munk, and Tilmann Gneiting. Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics*, 33(4):753–763, 2006.

Ming Huang and Fuling Bian. A grid and density based fast spatial clustering algorithm. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, volume 4, pages 260–263. IEEE, 2009.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Jennifer Jang and Heinrich Jiang. Dbscan++: Towards fast and scalable density clustering. In *International conference on machine learning*, pages 3019–3029. PMLR, 2019.

D. Janzing. Causal regularization. In *Advances in Neural Information Processing Systems 33*, 2019.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

Michael J Kearns and Umesh Vazirani. *An introduction to computational learning theory*. 1994.

Jon Kleinberg. An impossibility theorem for clustering. *Advances in neural information processing systems*, 15, 2002.

Jussi Klemelä. Complexity penalized support estimation. *Journal of multivariate analysis*, 88(2):274–297, 2004.

Yegor Klochkov, Alexey Kroshnin, and Nikita Zhivotovskiy. Robust k-means clustering for distributions with two moments. *The Annals of Statistics*, 49(4):2206–2230, 2021.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5): 1902–1914, 2001.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Andrey Kuehlkamp, Benedict Becker, and Kevin Bowyer. Gender-from-iris or gender-from-mascara? In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 1151–1159. IEEE, 2017.

K Mahesh Kumar and A Rama Mohan Reddy. A fast dbscan clustering algorithm by accelerating neighbor searching using groups method. *Pattern Recognition*, 58:39–48, 2016.

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 275–282, 2002.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.

Clément Levrard. Nonasymptotic bounds for vector quantization in hilbert spaces. *The Annals of Statistics*, 43(2):592–619, 2015.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and minimum-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

Tamás Linder. On the training distortion of vector quantizers. *IEEE Transactions on Information Theory*, 46(4):1617–1623, 2000.

Tamás Linder. Learning-theoretic methods in vector quantization. In *Principles of nonparametric learning*, pages 163–210. Springer, 2002.

Tamás Linder, Gábor Lugosi, and Kenneth Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, 40(6):1728–1740, 1994.

Bing Liu. A fast density-based clustering algorithm for large databases. In *2006 International Conference on Machine Learning and Cybernetics*, pages 996–1000. IEEE, 2006.

Yong Liu. Refined learning bounds for kernel and approximate *k*-means. *Advances in Neural Information Processing Systems*, 34: 6142–6154, 2021.

J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Andreas Maurer. A second-order look at stability and generalization. In *Conference on learning theory*, pages 1461–1475. PMLR, 2017.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

Shahar Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE transactions on Information Theory*, 48(1):251–263, 2002.

Wang Miao, Peng Ding, and Zhi Geng. Identifiability of normal and normal mixture models with non-ignorable missing data. *Journal of the American Statistical Association*, 111(516):1673–1683, 2016.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016.

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33:2710–2721, 2020.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1): 161–193, 2006.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (translated). *Statistical Science*, 5:465–480, 1923.

NYT. Electronic 'brain' teaches itslef, 1958. URL `https://www.nytimes.com/1958/07/13/archives/electronic-brain-teaches-itself.html?smid=url-share`.

Samir Okasha. Does hume's argument against induction rest on a quantifier-shift fallacy? In *Proceedings of the Aristotelian Society*, volume 105, 2005.

Debolina Paul, Saptarshi Chakraborty, and Swagatam Das. On the uniform concentration bounds and large sample properties of clustering with bregman divergences. *Stat*, 10(1):e360, 2021a.

Debolina Paul, Saptarshi Chakraborty, Swagatam Das, and Jason Xu. Uniform concentration bounds toward a unified framework for robust clustering. *Advances in Neural Information Processing Systems*, 34:8307–8319, 2021b.

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 2009.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge, 2017.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

David Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140, 1981.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *OpenAI Blog*, 2022.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1991.

Luca Rendsburg, Leena Chennuru Vankadara, Debarghya Ghoshdastidar, and Ulrike von Luxburg. A consistent estimator for confounding strength. *arXiv preprint arXiv:2211.01903*, 2022.

Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):1–9, 2020.

Philippe Rigollet and Régis Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, pages 1154–1178, 2009.

Alessandro Rinaldo, Larry Wasserman, et al. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010.

Alessandro Rinaldo, Aarti Singh, Rebecca Nugent, and Larry Wasserman. Stability of density-based clustering. *The Journal of Machine Learning Research*, 13(1):905–948, 2012.

J. M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods — applications to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

P.R. Rosenbaum. *Observational Studies*. Springer Series in Statistics. Springer, 2002. ISBN 9780387989679. URL https://books.google.de/books?id=K0OglGXtpGMC.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28, 2015.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Cullen Schaffer. A conservation law for generalization performance. In *Machine Learning Proceedings 1994*, pages 259–265. Elsevier, 1994.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, and Kun Zhang. Robust learning via cause-effect models. *arXiv preprint arXiv:1112.2738*, 2011.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Elliott Sober. *Reconstructing the past: Parsimony, evolution, and inference*. MIT press, 1991.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Bharath Sriperumbudur and Ingo Steinwart. Consistency and rates for clustering with dbscan. In *Artificial Intelligence and Statistics*, pages 1090–1098, 2012.

Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

Ingo Steinwart, Bharath K Sriperumbudur, Philipp Thomann, and D ONE Solutions AG. Adaptive clustering using kernel density estimators. *stat*, 1050:17, 2019.

Tom F Sterkenburg and Peter D Grünwald. The no-free-lunch theorems of supervised learning. *Synthese*, 199(3):9979–10015, 2021.

Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical statistics*, pages 1265–1269, 1963.

Matus J Telgarsky and Sanjoy Dasgupta. Moment-based uniform deviation bounds for k-means and friends. *Advances in Neural Information Processing Systems*, 26, 2013.

Matus Jan Telgarsky and Sanjoy Dasgupta. Agglomerative bregman clustering. In *29th International Conference on Machine Learning, ICML 2012*, pages 1527–1534, 2012.

Yoshikazu Terada. Strong consistency of reduced k-means clustering. *Scandinavian Journal of Statistics*, 41(4):913–931, 2014.

Yoshikazu Terada. Strong consistency of factorial k-means clustering. *Annals of the Institute of Statistical Mathematics*, 67(2):335–357, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Alexandre B Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

Robert A Vandermeulen and Clayton D Scott. On the identifiability of mixture models from grouped samples. *arXiv preprint arXiv:1502.06644*, 2015.

Leena C Vankadara and Debarghya Ghoshdastidar. On the optimality of kernels for high-dimensional clustering. In *International Conference on Artificial Intelligence and Statistics*, pages 2185–2195. PMLR, 2020.

Leena C Vankadara, Sebastian Bordt, Ulrike von Luxburg, and Debarghya Ghoshdastidar. Recovery guarantees for kernel-based clustering under non-parametric mixture models. In *International Conference on Artificial Intelligence and Statistics*, pages 3817–3825. PMLR, 2021a.

Leena Chennuru Vankadara, Philipp Michael Faller, Lenon Minorics, Debarghya Ghoshdastidar, and Dominik Janzing. Causal forecasting: Generalization bounds for autoregressive models. *arXiv preprint arXiv:2111.09831*, 2021b.

VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

P Viswanath and V Suresh Babu. Rough-dbscan: A fast hybrid density based clustering method for large data sets. *Pattern Recognition Letters*, 30(16):1477–1488, 2009.

Ulrike Von Luxburg and Shai Ben-David. Towards a statistical theory of clustering. In *Pascal workshop on statistics and optimization of clustering*, pages 20–26. London, UK, 2005.

Ulrike Von Luxburg, Robert C Williamson, and Isabelle Guyon. Clustering: Science or art? In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 65–79. JMLR Workshop and Conference Proceedings, 2012.

Daren Wang, Xinyang Lu, and Alessandro Rinaldo. Dbscan: Optimal rates for density-based cluster estimation. *Journal of machine learning research*, 2019a.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Shusen Wang, Alex Gittens, and Michael Mahoney. Scalable kernel k-means clustering with nyström approximation: Relative-error bounds. *Journal of Machine Learning Research*, 20:1–49, 2019b.

David H Wolpert. On the connection between in-sample testing and generalization error. *Complex Systems*, 6(1):47, 1992.

David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

David H Wolpert. The supervised learning no-free-lunch theorems. *Soft computing and industry*, pages 25–42, 2002.

Bowei Yan and Purnamrita Sarkar. On robustness of kernel clustering. In *Advances in Neural Information Processing Systems*, pages 3098–3106, 2016.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

Reza Bosagh Zadeh and Shai Ben-David. A uniqueness theorem for clustering. *arXiv preprint arXiv:1205.2600*, 2012.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Tong Zhang. Leave-one-out bounds for kernel methods. *Neural computation*, 15(6):1397–1437, 2003.

Shixiang Zhu, Liyan Xie, Minghe Zhang, Rui Gao, and Yao Xie. Distributionally robust weighted *k*-nearest neighbors. *arXiv preprint arXiv:2006.04004*, 2020.