Changelog

Major changes will be described here. I will probably not include minor changes here.

- June 18, incorporate Hu's comments.
- June 12, Initial version.

Optimal Algorithms for Experts and Mixtures of Gaussians

Version of Monday 22nd June, 2020 at 10:13

by

Christopher Vui Seng Liaw

Bachelor of Applied Science, University of British Columbia, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF SCIENCE
(Computer Science)

The University of British Columbia
(Vancouver)

August 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Optimal Algorithms for Experts and Mixtures of Gaussians

submitted by **Christopher Vui Seng Liaw** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Computer Science**.

Examining Committee:

Nicholas J. A. Harvey, Computer Science
Supervisor

Hu Fu, Computer Science
Supervisory Committee Member

Bruce Shepherd, Computer Science
Supervisory Committee Member

Abstract

This thesis makes contributions to two problems in learning theory: prediction with expert advice and learning mixtures of Gaussians.

The problem of prediction with expert advice can be cast as a sequential game between an algorithm and an adversary as follows. At each time step, an algorithm chooses one of n options (or experts) and the adversary sets a cost for each expert. The algorithm's goal is to minimize its *regret*, i.e. its cost relative to the best expert in hindsight. The celebrated multiplicative weights algorithm is known to be optimal if the the game is terminated at a fixed, *known* time and the number of experts is large. Optimal algorithms are also known when the number of experts is 2, 3, or 4.

If the game does not terminate at a *known* time or is run indefinitely, the optimal algorithm is not known for any number of experts. We contribute to this problem by giving the optimal algorithm when there are 2 experts. Our algorithm is designed by considering a continuous analogue of the problem, which is solved using ideas from stochastic calculus.

In the second part of the thesis, we look at distribution learning, which is a fundamental task in statistics that has been studied for over a century. We consider such a problem where the distribution is a mixture of k Gaussians in $\mathbb{R}^d$. The objective is density estimation: given i.i.d. samples from the unknown distribution, produce a distribution whose total variation from the unknown distribution is at most ε. We prove that $\tilde{\Theta}(kd^2/\varepsilon^2)$ are sufficient and necessary for this task, suppressing logarithmic factors. This improves both the known upper bound and lower bound for this problem.

Lay Summary

Machine learning is now ubiquitous in our daily lives. Despite its success, many basic and fundamental questions remain unanswered.

One such example is in decision making. Suppose one makes a decision every day, say by deciding between several choices. Can one make decisions in such a way as to have no regret in not knowing the best choice beforehand? Perhaps surprisingly, this problem has important applications in finance, machine learning, and algorithm design. In this thesis, we design an algorithm for making decisions with strong theoretical guarantees.

The second problem is in statistics: if we want to model some phenomenon, how much data do we need to collect? For example, this may model how different groups of people react to different vaccines. Collecting data is often an expensive endeavour. In this thesis, we design an algorithm that optimally trades off between accuracy and the amount of data collected.

Preface

The results presented in this thesis are the result of two major research projects that the authour has had the privilege to partake in.

- The material in Chapter 2 is on the classical problem of prediction with expert advice. It is based on the paper in [83] and is joint work with my supervisor Nick Harvey, as well as Sikander Randhawa and Edwin Perkins. In [83], we give a new algorithm when there are two experts and prove that the algorithm is optimal; this also answers an open question posed by [106] on the minimax strategy for two experts. The problem was proposed by Nick Harvey. All the authours, including myself, were heavily involved in coming up with the ideas for the solution, its execution, and the writing of the manuscript.

- In Chapter 4, we study the problem of learning mixtures of Gaussians. The material there is based on joint work with Hassan Ashtiani, Shai Ben-David, Nick Harvey, Abbas Mehrabian, and Yaniv Plan which appeared in the 2018 Conference on Neural Information Processing Systems [14]. In [14], we give a new algorithm for learning mixtures of Gaussians and show that, up to minor polylogarithmic factors, the algorithm uses the minimum number of samples. The problem was introduced to me by Abbas. The concept of compression was conceived by Hassan, Shai, and Abbas (see also [16, §7.2]). I was involved with proving the lower bound for mixtures of Gaussians as well as using the compression ideas to develop an upper bound with near-optimal sample complexity.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to begin by thanking my supervisor, Nick Harvey. I remember first meeting Nick at the SFU summer school on randomized algorithms prior to my final year of undergrad, although, at the time I was far too shy to introduce myself to him. My memories of the summer school have since faded but I do recall that his lectures were one of the main reasons I applied to do a PhD at UBC CS. The next five years of working with Nick have been an absolute blast. Nick is an absolutely brilliant researcher. Whenever I felt stuck on a problem, Nick would always have an insightful new direction to try. More importantly, he has always been extremely caring and patient. I am forever grateful to Nick and there is no doubt I would not have been able to get this far without him.

Next, I would like to thank the other members of my thesis committee: Hu Fu and Bruce Shepherd. Interacting with Hu throughout my PhD was a great pleasure and I have thoroughly enjoyed the projects that we worked on together. Although I have not worked with Bruce on any projects, I will forever be grateful for the insightful advice he has given me.

A grateful thank you to everybody I have collaborated with during my PhD: Abbas Mehrabian, Hassan Ashtiani, Shai Ben-David, Yaniv Plan, Edwin Perkins, Sikander Randhawa, Paul Liu, Tasuku Soma, Hu Fu, Nick Harvey, Roman Vershynin, Peter Kling, Petra Berenbrink, Zhihao Tang, Pinyan Lu, and Robert Reiss. A big thank you to Yaniv Plan and Edwin Perkins for teaching me everything that I know about probability. I would also like to thank the many faculty members in theory group that I have taken courses with and learned a lot from: David Kirkpatrick, Anne Condon, Joel Friedman, and Will Evans.

During my PhD, I had the chance to do an internship with Aranyak and Siva at Google Research. I thank them and the other interns for making the internship such an enjoyable experience.

My PhD would not have been possible without the many lab-mates and friends that surrounded me: Sikander, Taylor, Mehrdad, Coulter, Richard, Joey, Paul, Ron, Victor, Jason, Reza, Heddy, Sharan, Neil, Chris, David, Greg, Amit, Raunak, Yifan, Aaron, Bader, Kyle. Talking about research was also a blast but so was playing basketball or disc golf, hanging out at the beach at midnight, or cheering on the Raptors.

A big shoutout to all the administrative staff at UBC for helping with the paperwork and making sure I get paid. Thanks to UBC and NSERC for providing me with funding during my PhD.

Lastly, and most importantly, I would like to thank my family. Without their constant support, this thesis would be no more than a fantasy.

Chapter 1

Introduction

Machine learning has become a ubiquitous tool in many different fields, ranging from medical applications [41, 97, 141] to playing video games [108, 139] to advertising [21, 115, 123] and much, much more. The field of *learning theory* aims to provide a theoretical and mathematical foundation to machine learning.

In this thesis, we focus on two basic and classical problems in learning theory. The first problem we study is the problem of prediction with expert advice. This is a fundamental problem whose origin dates back to the 1950s with the work of Hannan [82]. The second problem that we study is the problem of learning mixtures of Gaussians which dates back to the late 19th century when Karl Pearson was developing his mathematical theory of evolution [112].

This thesis makes contributions to both of these problems. For the problem of prediction with expert advice, we give an optimal algorithm for an important special case of the problem. More importantly, we introduce a new technique, based on Brownian Motion and stochastic analysis, to analyze regret in online learning. For the problem of learning mixtures of Gaussians, we introduce the concept of sample compression within the context of distribution learning and use this to give an algorithm that provably uses the least number of samples.

In the next two sections of this chapter, we give some background and an overview of these two problems.

1.1 Prediction with Expert Advice

The problem of prediction with expert advice (also referred to as the experts problem) can be cast as the following sequential game between an adversary and an algorithm. At each time step t, the algorithm must pick one of n choices (perhaps randomly). Then an adversary, knowing what the algorithm's strategy is but not the outcome of its randomness, assigns a cost (or a loss) to each choice. The algorithm's goal is to develop a strategy so that its accumulated cost is almost as good as the best single choice in hindsight.

As an example of this problem, each of the choices could correspond to a different route that a person can take to go to work each day. Then the *cost* of each route corresponds to the time it takes to get to work. The problem for the commuter is to design a strategy of picking each day's route to

be almost as good as having stuck with the best single route in hindsight.

Although the problem itself appears to be simple, solutions to this problem have been a key component in a number of different areas. Here, we list a small number of applications. We refer the reader to the survey of Arora et al. [13] for many more examples.

- It has been used as a core subroutine in designing fast algorithms for approximately solving linear programs [118, 144]. This, in turn, has made it useful for solving a variety of combinatorial optimization problems. For example, the greedy algorithm for set cover can be seen as an instantiation of this approach [144].
- In combinatorial optimization, algorithms for prediction with expert advice have formed an important component in designing algorithms for a number of problems. Examples include computing maximum flows in graphs [38], computing multicommodity flows [71], and computing sparsest cut [12].
- In learning theory, *boosting*, where one combines many weak learning algorithms into a very effective learning algorithm, can also be seen as an application of the learning with experts paradigm [70].
- In complexity theory, this problem has been used to construct Boolean functions which are inapproximable by circuits of bounded size [90].

The most well-known algorithm for the experts problem is the celebrated multiplicative weights update algorithm, introduced independently by Littlestone and Warmuth [103] as the weighted majority algorithm and by Vovk [140] as the aggregating strategies algorithm. Here, we will give the algorithm but the analysis can be found in a number of different sources including [26, Theorem 2.1],[73, Theorem 2.1], or [31, Theorem 2.2]. At a high-level, the algorithm maintains the total loss incurred by each expert. (Initially each expert has incurred total loss of 0). At each time step t, the algorithm chooses each expert with some probability which depends on the expert's total loss thus far; the more loss an expert has incurred relative to the other experts, the less likely the algorithm will choose that expert. The pseudocode is given in Algorithm 1. The parameter $\eta_t > 0$ is a step-size which is allowed to depend on all events *up to* time $t-1$.

Algorithm 1 The multiplicative weights update algorithm.

1: Initialize $L_0 \leftarrow (0, \ldots, 0)$.
2: **for** $t = 1, 2, \ldots$ **do**
3: Set probability vector p_t so that $p_{t,i} \propto \exp(-\eta_t L_{t-1,i})$, i.e.

$$p_{t,i} = \frac{\exp(-\eta_t L_{t-1,i})}{\sum_{j \in [n]} \exp(-\eta_t L_{t-1,j})}.$$

4: Choose expert $i \in [n]$ according to p_t.
5: Receive cost vector c_t, incur cost $c_{t,i}$ and update $L_t \leftarrow L_{t-1} + c_t$.
6: **end for**

To discuss the merits of Algorithm 1, we will need to formalize one small piece of notation. Fix an algorithm and let its strategy at time t be p_t, where p_t is an n-dimensional probability vector and is allowed to depend on all information up to time $t-1$. Let $c_1, \ldots$ be the sequence of cost vectors. The

expected loss of the algorithm at time T is $\sum_{t=1}^{T} c_t^\top p_t$. Let $L_{T,i} = \sum_{t=1}^{T} c_{t,i}$ be the total cost of expert i up until time T. The algorithm's *regret* is defined as

$$\text{Regret}(T) \coloneqq \sum_{t=1}^{T} c_t^\top p_t - \min_{i \in [n]} L_{t,i}. \tag{1.1}$$

Eq. (1.1) is exactly the gap between the expected cost of the algorithm and the single best expert in hindsight, i.e. the algorithm's regret for not having the foresight to choose expert i.

Theorem 1.1. *Fix $T > 0$ and assume $c_t \in [0,1]$ for all $t \in \{1, \ldots, T\}$. Setting $\eta_t = 2\sqrt{2T \ln n}$, Algorithm 1 guarantees*

$$\text{Regret}(T) \leq \sqrt{T \ln(n)/2}.$$

References. [26, Theorem 2.1],[73, Theorem 2.1], or [31, Theorem 2.2]

The theorem asserts that for *any* sequence of cost vectors $c_1, \ldots, c_T$, the multiplicative weights algorithm (with the correct tuning *depending only* on T) has regret bounded by $\sqrt{T \ln(n)/2}$. It is important to note the order of quantifiers in the statement of Theorem 1.1. For any time T which is *fixed in advance*, Algorithm 1 with the correct tuning, *depending on* T, achieves a regret of at most $\sqrt{T \ln(n)/2}$ at time T (and in fact, for all times up to T). Guarantees of this type are often referred to as *fixed-time* guarantees. We also refer to the setting where T is known to the algorithm a priori as the *fixed-time* setting.

Theorem 1.1 is known to be tight in the sense that for *any* algorithm

$$\limsup_{n \to \infty} \lim_{T \to \infty} \frac{\text{Regret}(T)}{\sqrt{T \ln(n)/2}} \geq 1.$$

This means that for *any* algorithm, for any $\varepsilon > 0$, there exists n, T and a sequence of cost vectors $c_1, \ldots, c_T \in [0,1]^n$ such that $\text{Regret}(T) \geq (1-\varepsilon)\sqrt{T \ln(n)/2}$. Note that the algorithm is allowed to know T in advance. It is known that Algorithm 1 is *not* tight when the number of experts is small. For $n = 2$, the optimal algorithm was given by Cover [39]. He used a natural dynamic programming formulation for the problem to give an algorithm that achieves $\text{Regret}(T) \leq \sqrt{T/2\pi} + O(1)$ for every sequence of loss vectors. For $n = 3$, Abbasi-Yadkori et al. [2] showed that one can obtain $\text{Regret}(T) \leq \sqrt{8T/9\pi} + o(T)$ and that this is optimal. For $n = 4$, Bayraktar et al. [19] showed that the optimal regret is $\sqrt{\pi T/8}$. For any fixed $n \geq 5$, the optimal algorithm is unknown.

For some applications, the time horizon T may not be known in advance or may be extremely large. When this is the case, we would much prefer an *anytime* guarantee, which controls the regret at *all* points in time and not just at a fixed time. One possible way to achieve this is via the well-known "doubling trick" [32, §4.6]. The doubling trick works as follows. Suppose that we had an algorithm with a fixed-time guarantee (for example, Algorithm 1 tuned appropriately). We then initialize the algorithm with a time horizon of T_0 and run the algorithm for T_0 steps. Then, we reset the algorithm by discarding its current state and re-initialize it with a time horizon with $2T_0$. After running the algorithm for $2T_0$ additional time steps, we reset the algorithm again and re-initialize it with a time horizon $4T_0$. This process continues indefinitely and every time we reset the algorithm, we double the

time horizon that the algorithm receives as input.

It is a short calculation to show that the doubling trick transforms any algorithm that achieves a sublinear regret for a fixed time T, say $O(T^\alpha)$ where $\alpha \in (1/2, 1)$, into another algorithm which achieves a regret of $O(t^\alpha)$ for *all* $t > 0$.[1] We note that the constant hidden by the $O(\cdot)$ may be larger in the latter setting. Although the doubling trick is theoretically simple, it is very wasteful from a practical point of view as it throws away all progress that the algorithm has made whenever it resets the algorithm. A much more elegant approach is to use a dynamic step size as shown by the following algorithm.

Theorem 1.2. *Assume $c_t \in [0,1]^n$ for all $t \geq 1$. Setting $\eta_t = 2\sqrt{t \ln n}$, Algorithm 1 guarantees*

$$\text{Regret}(t) \leq \sqrt{t \ln(n)} \quad \forall t \geq 1.$$

References. [26, Theorem 2.4] or [73, Proposition 2.1]

Note that the guarantee in Theorem 1.2 holds for *any* time t. It is unknown whether Theorem 1.2 is tight but by Theorem 1.1, it can by off by at most $\sqrt{2}$.

Unlike in the fixed-time setting, the optimal algorithm for few experts is unknown, even for $n = 2$. In Chapter 2, we design an algorithm for the anytime setting with $n = 2$ experts and show that it is optimal. To state our result, we define the function $M(x) \coloneqq \sqrt{\pi x}\,\text{erfi}(\sqrt{x}) + e^x$.

Theorem 1.3. *Assume $c_t \in [0,1]^2$ for all $t \geq 1$. There is an algorithm which achieves*

$$\text{Regret}(t) \leq \frac{\gamma}{2}\sqrt{t} \quad \forall t \geq 1,$$

where $\gamma \approx 1.30693$ is the unique positive root of $M(x^2/2)$. Moreover, the constant $\gamma/2$ cannot be improved.

The function M is an example of a confluent hypergeometric function; this is a broad class of functions which include many well-known classes of functions, such as Laguerre polynomials and Bessel functions. Interestingly, the roots of confluent hypergeometric functions have played an important role in the study of fundamental properties of Brownian Motion and random walks [23, 48, 79, 114]. It is natural to wonder whether there are connections between regret minimization and probability. In Chapter 2, we give one example by showing how Theorem 1.3 can be used to recover a probabilistic statement of random walks due to Davis [48]. Other connections can also be found in the work of Rakhlin and Sridharan [119].

The design of the algorithm has a number of key features which we point out here. First, we design the algorithm by considering a continuous analogue version of regret. Next, we use stochastic calculus to find an elegant relationship between (continuous) regret and any (continuous-time) algorithm. This relationship exposes a partial differential equation which we can then solve to obtain an optimal continuous-time algorithm.

[1] To briefly sketch this, suppose that $t = T_0 + 2T_0 + \cdots + 2^k T_0 = \Theta(2^k T_0)$ for some $k \geq 1$ and $T_0 \geq 1$. If an algorithm can be tuned to give regret $C \cdot T^\alpha$ ($C > 0$ is some constant) for all $T > 0$ then the transformed algorithm would have regret $C \cdot T_0^\alpha \sum_{i=0}^k 2^{\alpha i}$. The sum itself is $O(2^{\alpha k})$ so the regret at time t is $O(2^{\alpha k} T_0^\alpha) = O(t^\alpha)$.

One question remains: how can we transform the continuous-time algorithm into a discrete-time algorithm? It turns out that the optimal continuous-time algorithm is actually a derivative of a certain potential function. To obtain a discrete-time algorithm, we look at the *discrete* derivative of the *same* potential function.

1.2 Learning Mixtures of Gaussians

The problem of distribution learning is a classical and fundamental problem in statistics that dates back to the work of Karl Pearson in the late 19th century [112]. At the time, Karl Pearson was developing a mathematical theory of evolution. The problem itself is quite simple: if one has a dataset, can one understand the underlying distribution from which the data originated? It is typical to make some assumptions on the underlying distribution; for example, the data coming from a single source may be coming from a Gaussian (or Normal) distribution. A single dataset may actually consist of data from multiple sources. In this case, the dataset itself may be comprised of a combination, or *mixture*, of Gaussians.

As an example of this, we can consider an experiment that Karl Pearson performed with some crab data that he received from Prof. Weldon. The dataset that Karl Pearson possessed contained the ratio between the length of the forehead and the length of the body for some of the crabs in Naples. If the dataset contained data for a *single* species of crabs then this ratio ought to form a Normal distribution. Karl Pearson observed that the dataset did not appear to be symmetric, let alone to follow a Normal distribution. However, he observed that the dataset was very well approximated as a mixture of two Gaussian components which provided evidence that the dataset actually contained at least two distinct species.

The problem of distribution learning and, in particular the problem of learning mixtures of Gaussians, continued to garner a tremendous amount of study over the past century. In modern data science, practitioners often try to model their data using a Gaussian mixture model and many software packages have implemented algorithms to perform this task [1, 113]. In contrast to the crab example above, Gaussian mixture models are often used in settings where the data is very high-dimensional [78, 116, 125].

The most common heuristic used in practice to fit a Gaussian mixture model to data is the expectation-maximization (EM) heuristic [50]. However, the EM heuristic for fitting Gaussian mixture models is not very well understood and it is unknown whether EM (or some variant of it) will always converge to the true Gaussian mixture. Nonetheless, there is a growing body of work which aims to understand EM in the context of learning Gaussian mixtures [43, 44, 46, 98].

The problem of learning mixtures of Gaussians was introduced to the theoretical computer science community by Dasgupta [42]. Since then, there was been a flurry of work on the subject [8, 9, 15, 34, 45, 58, 68, 72, 85, 86, 88, 89, 91, 101, 109, 121, 129, 133, 136]. Typically one of two models are considered: the parameter estimation model and the density estimation model. Here, we will give a brief, informal discussion of what mixture of Gaussians are, the parameter estimation model, and the density estimation model. A more formal background for mixtures of Gaussians and density estimation is given in Chapter 3.

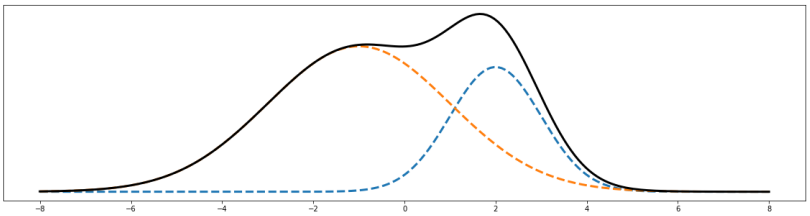


Figure 1.1: An example of a mixture of two Gaussians in $\mathbb{R}$. The dashed coloured lines correspond to the p.d.f. of individual Gaussian distributions. The black solid line is a weighted average of the two Gaussians.

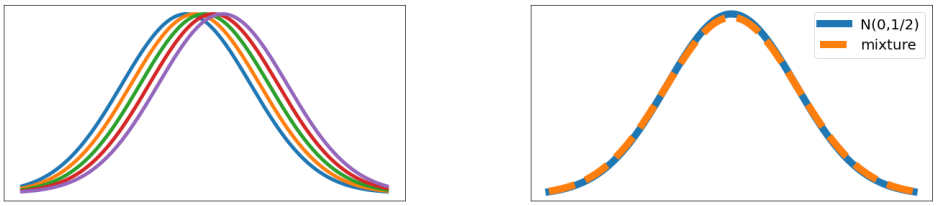
Mixtures of Gaussians. A Gaussian distribution in $\mathbb{R}^d$ is a distribution which is specified by a mean vector $\mu \in \mathbb{R}^d$ and a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$.[2] The vector μ determines where the center of Gaussian is and the covariance matrix Σ determines how skewed and spread out the distribution is along certain directions. A mixture of k Gaussians in $\mathbb{R}^d$ consists of k components where each component is itself a Gaussian distribution. Each component $i \in [k]$ has a weight $w_i \geq 0$ (called the *mixing weights*) with $\sum_{i \in [k]} w_i = 1$. To sample from a mixture of k Gaussians, one first samples a component with respect to the probability w and then samples from the corresponding Gaussian distribution. Figure 1.1 gives an example of a mixture of two Gaussians.

Parameter Estimation. In parameter estimation, the goal is to estimate, for an unknown mixture of Gaussians, the mixing weights of each of the components, the mean vector of each of the components, and the covariance matrix of each of the components. This model was first introduced by Dasgupta [42] and culminated in the work of Moitra and Valiant [109] who gave algorithms with running time that was polynomial in the number of samples and with provably minimal assumptions.

In general, the parameter estimation problem can be quite difficult. Indeed, Moitra and Valiant [109] showed that, while the running time of their algorithm is polynomial in the number of samples that the algorithm uses, the *sample complexity* (i.e. the number of samples that the algorithm uses) is exponential in the number of components that appear in the mixture, even for $d = 1$. The issue lies in the fact that it is possible to construct two different mixtures of Gaussians where all the components have distinct parameters while the distribution themselves are nearly identical. Figure 1.2 shows an example of the construction in [109] and illustrates how a single Gaussian can be almost indistinguishable from a mixture of Gaussians.

Density Estimation. In density estimation, the goal is only to return a distribution which is close (according to some notion of "distance") to the distribution that we are trying to learn. This alleviates the issue present in the previous discussion where it is possible to have mixtures of Gaussians with different parameters whose distributions are nearly identical. Indeed, in the density estimation model, returning any of these nearly identical distributions would be considered a success. For example, going back to Figure 1.2, if the true distribution was indeed the mixture of Gaussians, the algorithm would

[2] There are certain restrictions on what Σ is; we relegate these details to Chapter 3.



(a) 5 different Gaussian components. **(b)** A Gaussian and mixture of Gaussians.

Figure 1.2: Figure 1.2a shows five different Gaussian components. We note that the Gaussian components are reasonably separated. In Figure 1.2b, two distributions are plotted. The blue curve is a standard Gaussian with mean 0 and variance 1/2 and the orange dashed curve is a mixture of Gaussians whose components are exactly those in Figure 1.2a. We note that the mixture of Gaussians is nearly indistinguishable from just a single Gaussian. This example was taken from [109] and shows that it is very difficult to distinguish between a single Gaussian and a mixture of Gaussians. In the density estimation model, the problem is relaxed and we say that an algorithm is successful if it outputs a distribution which is nearly indistinguishable from the original distribution.

still be considered correct if it had output a single Gaussian.

It turns out that considering this particular model for learning mixture of Gaussians is sufficient to circumvent the exponential lower bound of Moitra and Valiant [109]. Indeed, a number of previous works have shown that, in the density estimation model, there are algorithms which require only a polynomial number of samples [15, 68, 133]. However, it remained an open problem to determine *exactly* how many samples are needed in order to learn a mixture of Gaussian in this model. In Chapter 4, we design a new algorithm for learning mixtures of Gaussians and prove that the algorithm provably uses the minimum number of samples, up to logarithmic factors. Informally, we will prove the following theorem.

Theorem 1.4 (Informal)**.** *There is an algorithm which, given $\tilde{O}(kd^2/\varepsilon^2)$ i.i.d. samples from an unknown mixture of Gaussians in d dimensions with k components, returns a distribution whose density is "ε-close" to the unknown distribution. Moreover, any such algorithm requires $\tilde{\Omega}(kd^2/\varepsilon^2)$.*

In Theorem 1.4, the $\tilde{O}$ and $\tilde{\Omega}$ notation suppresses polylog(kd/ε) factors.

The algorithm is based on the idea of *sample compression*. We show that if one can *compress* the identity of the distribution, using samples from that distribution, then there exists a sample-efficient algorithm for learning the distribution as well. In fact, the reduction transforms the sample compression scheme into a sample-efficient algorithm.

1.3 Thesis Organization

Chapter 2 describes the optimal anytime algorithm for two experts. In Chapter 3, we give some necessary probability background to understand the work on learning mixtures of Gaussians. A reader with a basic background on distribution learning is encouraged to skip Chapter 3 and directly read Chapter 4. We note that Chapter 2 is completely disjoint from Chapter 3 and Chapter 4.

Chapter 2

Optimal anytime regret with two experts

Chapter Summary. In this chapter, we study the problem of prediction with expert advice. We show that, for two experts and costs in $[0, 1]$, there is an algorithm achieving $\text{Regret}(t) \leq \frac{\gamma}{2}\sqrt{t}$ for all $t > 0$, where $\gamma \approx 1.30693$ is the root of a confluent hypergeometric function with certain parameters. Furthermore, this is optimal, in the sense that the constant $\gamma/2$ is best possible. Prior to this work, there were no known optimal algorithms for any number of experts, in the anytime setting.

2.1 Introduction

In this chapter, we study the problem of prediction with expert advice, whose origin can be traced back to the 1950s [82]. The problem is a sequential game between an adversary and an algorithm as follows. There are n actions, which are called "experts". At each time step, the algorithm computes a distribution over the experts, then randomly chooses an expert according to that distribution; concurrently, the adversary chooses a cost for each expert, with knowledge of the algorithm's distribution but not its random choice. The cost of each expert is then revealed to the algorithm, and the algorithm incurs the cost that its chosen expert incurred. The goal is to design an algorithm whose expected *regret* is small. That is, the goal is to minimize the difference between the algorithm's expected total cost and the total cost of the best expert. This problem and its variants have been a key component in many results in TCS and machine learning; some examples were discussed in Chapter 1 and we refer the reader to [13] for more applications.

The most well-known algorithm for the experts problem is the celebrated multiplicative weights update algorithm (MWU), as discussed in Chapter 1 [103, 140]. In the fixed-time setting (where a time horizon T is known in advance), MWU (with the optimal tuning of its step size) suffers a regret of $\sqrt{(T/2)\ln n}$ at time T, where n is the number of experts [30, 32]. This bound on the regret of MWU is known to be tight for any even n [77]. It is also known [32] that $\sqrt{(T/2)\ln n}$ is asymptotically optimal for any algorithm as $n, T \to \infty$. Hence, MWU is a minimax optimal[1] algorithm as $n, T \to \infty$.

[1] This means that the algorithm minimizes the maximum, over all adversaries, of the regret.

Interestingly, MWU is *not* optimal for small values of n. For $n = 2$, Cover [39] observed decades earlier that a natural dynamic programming formulation of the problem leads to a simple analysis showing that the minimax optimal regret is $\sqrt{T/2\pi}$.

For some applications, the time horizon T is not known in advance; examples include any sort of online tasks (e.g., online learning), or tasks requiring convergence over time (e.g., convergence to equilibria). An alternative model, more suited to those scenarios, is the *anytime setting*[2], in which algorithms are not given T but must bound the regret *for all* T. Yet another model is to assume that T is random with a known distribution [106]. For example, the *geometric horizon setting* of Gravin, Peres, and Sivan [76] assumes that T is a geometric random variable. In this setting, they gave the optimal algorithm for two and three experts. Moreover, they propose a conjecture on the relationship between the fixed-time and the geometric horizon settings that could lead to optimal bounds for all n.

Our focus is the anytime setting. One can convert algorithms for the fixed-time setting to the anytime setting by the well-known "doubling trick" [32, §4.6]. This involves restarting the fixed-time horizon algorithm every power-of-two steps with new parameters. If the fixed-time algorithm has regret $O(T^c)$ at time T for some $c \in (0, 1)$ then the doubling trick yields an algorithm with regret $O(t^c)$ at time t for every $t \geq 1$.[3] On the one hand, this is a conceptually simple and generic reduction. On the other hand, restarting the algorithm and discarding its state is clearly wasteful and probably not very practical.

Instead of using the doubling trick, one can use variants of MWU with a dynamic step size; see, e.g., [31, §2.3], [111, Theorem 1], [26, §2.5]. This is a much more elegant and practical approach and is even simpler to implement. However, the analysis is more difficult than for MWU, and is rarely taught. It is known that, with an appropriate choice of step sizes, MWU can guarantee[4] a regret of $\sqrt{t \ln n}$ for all $t \geq 1$ and all $n \geq 2$ (see [26, Theorem 2.4] or [73, Proposition 2.1]). However, it is unknown whether $\sqrt{t \ln n}$ is the minimax optimal anytime regret, for any value of n.

Results and techniques. This work considers the anytime setting with $n = 2$ experts. We show that the optimal regret is $\frac{\gamma}{2}\sqrt{t}$, where $\gamma \approx 1.30693$ is a fundamental constant that arises in the study of Brownian motion [114]. (Note that $\gamma/2 \approx 0.653 < 0.833 \approx \sqrt{\ln 2}$.) It is not a priori obvious why this fundamental constant should play a role in both Brownian motion and regret. Nevertheless, some connections are known. For example, in the fixed-time setting, the optimal algorithms for $n \in \{2, 3, 4\}$ (see [76]) and the optimal lower bound for $n \to \infty$ all involve properties of random walks. Since Brownian motion is a continuous limit of random walks, a connection between anytime regret and Brownian motion is plausible.

Our techniques to analyze the optimal anytime regret are a significant departure from previous work on regret minimization. First, we define a continuous-time analogue of the problem which expresses the regret as a stochastic integral. This allows us to utilize tools from stochastic calculus to arrive at

[2]Other authors have referred to this setting as an "unknown time horizon" or "bounds that hold uniformly over time".

[3] Note that the constant hidden inside the $O(\cdot)$ for the anytime setting is larger than the constant hidden inside the $O(\cdot)$ for the fixed time setting.

[4]It can be shown, by modifying arguments of [77], that this is the optimal anytime analysis for MWU with step sizes $c/\sqrt{t}$.

a potential function whose derivative gives the optimal *continuous*-time algorithm. Remarkably, the optimal *discrete*-time algorithm is the *discrete* derivative of the same potential function. Thus our work can be seen as bridging continuous and discrete optimization in a novel way.

The potential function that we derive involves a "confluent hypergeometric function". Such functions often arise in solutions to differential equations, and are useful in discrete mathematics [75, §5.5]. Our use of these functions may seem exotic, but they appear to be inherent to our problem since they also arise in the matching lower bound.

Motivation. The main purpose of this chapter is to determine optimal constants in regret bounds. The quest for optimal constants has been a perennial activity in TCS, particularly for fundamental problems in approximation algorithms. Often completely new techniques are required to obtain the optimal constants, such as for Max Cut [74], bounded-degree spanning trees [132], or constrained submodular maximization [29]. The techniques developed to find these optimal constants have had considerable longevity.

The experts problem is also a fundamental problem in TCS, as shown by the 2019 FOCS Test of Time Award for [102]. Finding the optimal constant in regret is one of the most basic and natural questions regarding the experts problem. Although the initial work on the experts problem focused on the simpler fixed-time setting, the anytime setting is arguably more natural since the algorithm requires fewer assumptions. For example, even in the original motivating example of predicting the weather [32], it makes more sense for the prediction task to continue indefinitely than to end at a fixed time. Thus, finding the optimal constant for anytime regret is a fundamental and compelling research question.

Why study the case $n=2$ when there are more applications for the case where n is large? Determining the minimax anytime regret for large n is an enticing question, but at present all algorithmic techniques seem unable to improve the $\sqrt{t \ln n}$ upper bound. This leads to the question: is there an inherent difference between the fixed-time and anytime settings, or is our pool of algorithmic techniques simply too limited? To address this question, we focus on the simple (yet still non-trivial) setting of $n=2$ in order to develop a range of new techniques (use of stochastic calculus, PDEs, stopping times and confluent hypergeometric functions). We hope that these techniques will play a role in resolving the question for large n.

There is some technical evidence that confluent hypergeometric functions have key properties not just for $n = 2$ but for large n as well. For example, the constant $\gamma/2$ appearing in the minimax regret for $n = 2$ is defined by $\gamma = \alpha(1/2)$, where α is a function giving the root of a confluent hypergeometric function with certain parameters. (See Claim 2.12 and [114, Prop. 1(b)].) It is known that $\alpha(1/n)/2 \xrightarrow{n\to\infty} \sqrt{\ln(n)/2}$, which gives some reassurance that these techniques could yield a $\sqrt{t \ln(n)/2}$ bound for large n.

Even if we disregard future generalizations, the anytime regret for two experts is interesting to study. It is one of the simplest questions about regret that remained unanswered, highlighting our incomplete understanding of the area. It is surprising that it took 55 years from Cover's two expert result [39] to perform an optimal analysis when reversing the quantifiers from "for all T there exists

an algorithm" to "there exists an algorithm such that for all T".

Application. An interesting application of our results is to a problem in probability theory that does not involve regret at all. Let $(X_t)_{t\geq 0}$ be a standard random walk. Then $\mathbb{E}[|X_\tau|] \leq \gamma \mathbb{E}[\sqrt{\tau}]$ for every stopping time τ; moreover, the constant γ cannot be improved.[5] This result is originally due to Davis [48, Eq. (3.8)], who proved it first for Brownian motion and later derived the result for random walks (via the Skorokhod embedding). We give a new derivation of Davis' result from our results in Subsection 2.2.4.

Related work. The minimax regret for the experts problem has been well-studied in the fixed-time horizon setting. For two experts the minimax regret was shown to be $\sqrt{T/2\pi}$ by Cover in 1965 [39]. It has been known for twenty years that $\sqrt{T\ln(n)/2}$ is the minimax regret as $n \to \infty$ [30, 32]. Building on the work of Gravin et al. [76], it has recently been shown that the minimax regret is $\sqrt{8T/9\pi}$ for three experts [2] and $\sqrt{\pi T/8}$ for four experts [19]. The anytime setting is not as well understood. In the two-experts setting, Luo and Schapire [106] demonstrate that, if the time horizon T is chosen by an adversary and unknown to the algorithm then the algorithm may be forced to incur regret at least $\sqrt{T/\pi}$. This exceeds the minimax regret of $\sqrt{T/2\pi}$ if T is known to the algorithm a priori, which indicates that the adversary has more power to force regret when it is allowed to select the time horizon.

Recently, interactions between algorithms in discrete and continuous *time* have been fruitful in other lines of work, e.g., [3, 27–29, 36, 61, 66, 93, 100, 142]. There is also a line of work that makes connections between the experts problem (in the finite-time horizon and geometric-time horizon setting) and PDEs [18, 19, 63, 64, 95, 96]. There is also work connecting regret minimization to option pricing [49] and to the Black-Scholes formula [4], which is based on Brownian motion and stochastic calculus. Intuitively, stochastic calculus is a crucial tool to optimally hedge against future costs, which we exploit too.

Our work crucially uses stopping times for Brownian motion hitting a time-dependent boundary. Such techniques have also been used for non-adversarial bandits to approximate Gittins indices (see, e.g., [25]).

2.2 Discussion of results and techniques

2.2.1 Formal problem statement

The problem may be stated formally as follows. Let n denote the number of experts. There is a deterministic algorithm $\mathcal{A}$, and a deterministic adversary $\mathcal{B}$ that knows $\mathcal{A}$. For each integer $t \geq 1$, there is a prediction task that is said to occur at time t. In this task, $\mathcal{A}$ picks a probability distribution $x_t \in [0,1]^n$, and $\mathcal{B}$ picks a cost vector $\ell_t \in [0,1]^n$. The coordinate $\ell_{t,j}$ denotes the cost of the j^{th} expert

[5] At first glance, the inequality may seem to contradict the Law of the Iterated Logarithm. However, we remark that if $\tau := \inf\{t > 0 : |X_t| \geq c\sqrt{t \ln \ln t}\}$ for some $c \in (0, \sqrt{2})$ then $\mathbb{E}[\sqrt{\tau}] = \infty$ (despite τ being a.s. finite) and the inequality is trivial.

at time t.[6]

After x_t is chosen the vector ℓ_t is revealed, so x_t depends on $\ell_1, \ldots, \ell_{t-1}$ (and implicitly $x_1, \ldots, x_{t-1}$). The vector ℓ_t depends on $\mathcal{A}$ and on $\ell_1, \ldots, \ell_{t-1}$ (and implicitly $x_1, \ldots, x_t$, since $\mathcal{A}$ is deterministic and known to $\mathcal{B}$). The game can end whenever $\mathcal{B}$ wishes, or continue forever. Since $\mathcal{A}$ is deterministic and known to $\mathcal{B}$, the entire sequence of interactions, including the ending time, can be predetermined by $\mathcal{B}$.

The cost incurred by the algorithm at time t is the inner product $\langle x_t, \ell_t \rangle$. This may be thought of as the "expected cost" of the algorithm, although the algorithm is actually deterministic. The total expected cost of the algorithm up to time t is $\sum_{i=1}^{t} \langle x_i, \ell_i \rangle$. For $j \in [n]$, the total cost of the j^{th} expert up to time t is $L_{t,j} = \sum_{i=1}^{t} \ell_{i,j}$. The regret at time t of algorithm $\mathcal{A}$ against adversary $\mathcal{B}$ is the difference between the algorithm's total expected cost and the total cost of the best expert, i.e.,

$$\text{Regret}(n, t, \mathcal{A}, \mathcal{B}) \;=\; \sum_{i=1}^{t} \langle x_i, \ell_i \rangle \;-\; \min_{j \in [n]} L_{t,j}.$$

Anytime setting. This work focuses on the anytime setting. In this setting, one may view the algorithm as running forever, with the goal of minimizing, for *all* t, the regret normalized by $\sqrt{t}$. Alternatively, one may view the game as ending at a time chosen by the adversary, and the algorithm must minimize the regret at that ending time. (It does not matter whether the adversary chooses the ending time in advance or dynamically, since $\mathcal{A}$ and $\mathcal{B}$ are deterministic so all interactions are predetermined.) These two views are equivalent because the algorithm cannot distinguish between them.

Formally, we will design an algorithm which achieves the infimum in the following expression.

$$\text{AnytimeNormRegret}(n) \;\coloneqq\; \inf_{\mathcal{A}} \sup_{\mathcal{B}} \sup_{t \geq 1} \frac{\text{Regret}(n, t, \mathcal{A}, \mathcal{B})}{\sqrt{t}}. \tag{2.1}$$

The minimax anytime regret is unknown even in the case of $n = 2$. The best known bounds at present are

$$0.564 \;\approx\; \sqrt{1/\pi} \;\leq\; \text{AnytimeNormRegret}(2) \;\leq\; \sqrt{\ln 2} \;\approx\; 0.833. \tag{2.2}$$

The lower bound, due to [106], demonstrates a gap between the anytime setting and the fixed-time setting, where the optimal normalized regret is $\sqrt{1/2\pi}$ [39]. Our main result in this chapter is that $\text{AnytimeNormRegret}(2) = \gamma/2 \approx 0.653$ and consequently neither inequality in Eq. (2.2) is tight.

[6] Alternatively, we may view $\mathcal{A}$ is a *randomized* algorithm which picks expert i with probability $x_{t,i}$. In this setting, x_t is known to the adversary but the realization of which expert is chosen is only revealed at the end of the round. We note that the vector x_t is still deterministic even if we consider the $\mathcal{A}$ to be a randomized algorithm.

2.2.2 Statement of results

To state our results, we require two definitions.

$$\begin{aligned} \operatorname{erfi}(x) &= \frac{2}{\sqrt{\pi}} \int_0^x e^{z^2} \, \mathrm{d}z \\ M_0(x) &= e^x - \sqrt{\pi x} \operatorname{erfi}(\sqrt{x}) \end{aligned} \tag{2.3}$$

The first is the imaginary error function, a well-known special function that relates to the Gaussian error function. The second is an example of a confluent hypergeometric function, a very broad class of special functions that includes, e.g., Bessel functions and Laguerre polynomials. (See Subsection 2.2.6 for formal definitions.) Our analysis makes use of a few elementary properties of these functions. A key constant used in this chapter is γ, which is defined to be the smallest[7] positive root[8] of $M_0(x^2/2)$, i.e.,

$$\gamma := \min \left\{ x > 0 : M_0(x^2/2) = 0 \right\} \approx 1.3069... \tag{2.4}$$

Theorem 2.1 (Main result)**.** *In the anytime setting with two experts, the minimax optimal normalized regret (over deterministic algorithms $\mathcal{A}$ and adversaries $\mathcal{B}$) is*

$$\text{AnytimeNormRegret}(2) = \inf_{\mathcal{A}} \sup_{\mathcal{B}} \sup_{t \geq 1} \frac{\text{Regret}(2, t, \mathcal{A}, \mathcal{B})}{\sqrt{t}} = \frac{\gamma}{2}. \tag{2.5}$$

The proof of this theorem has two parts: an upper bound, in Section 2.3, which exhibits an optimal algorithm, and a lower bound, in Section 2.4, which exhibits an optimal randomized adversary. The algorithm is very short, and it appears below in Algorithm 2.

One might imagine that some form of duality theory is involved in our matching upper and lower bounds. Indeed, if the costs are in $\{0, 1\}$ one may write AnytimeNormRegret(2) as the value of an infinite-dimensional linear program. In this case, our algorithm can be seen as a feasible solution to the primal linear program with value $\gamma/2$ and our random walk construction can be seen as a feasible solution to the dual linear program with value $\gamma/2 - \varepsilon$ for any $\varepsilon > 0$. In this chapter, we do not explicitly adopt this viewpoint. Instead, γ arises in our lower bound as the maximizer in Eq. (2.27), whereas γ arises in our upper bound as the minimizer in Eq. (2.48). We are not aware of any direct relationship between those two equations, although we speculate that a relationship might exist via some sort of unexplored duality theory for PDEs.

Comparison to existing techniques. A duality viewpoint is adopted by Gravin et al. [76] in the fixed-time and geometric horizon settings using von Neumann's minimax theorem. Their dual problem is characterized by properties of random walks, which allows one to determine the optimal dual value directly without reference to the primal. It is conceivable that some form of von Neumann's minimax theorem can be applied for the anytime setting, although it is unclear due to the appearance of the supremum and $1/\sqrt{t}$ in (2.5). Our results of Section 2.4 may be viewed as using random walks to

[7] In fact, γ is the *unique* positive root. See Fact 2.11.

[8] The *roots* of certain confluent hypergeometric functions have appeared in studying some natural phenomena of Brownian motion; for some examples see [23, 48, 79, 114].

construct feasible dual solutions of value $\gamma/2 - \varepsilon \;\; \forall \varepsilon > 0$, but it is *not obvious* that these solutions converge to the optimal dual value.

The only way we know of to prove optimality of those dual solutions is to construct an algorithm whose regret is $\gamma\sqrt{t}/2$. This is the more challenging part of this chapter, which we discuss in Sections 2.3 and 2.5. Interestingly, unlike some previous work, we explicitly obtain an optimal algorithm for costs in $[0, 1]$, not just for costs in $\{0, 1\}$.

Remark 2.2. *Our lower bound can be strengthened to show that, for any algorithm $\mathcal{A}$,*

$$\sup_{\mathcal{B}} \limsup_{t \geq 1} \frac{\text{Regret}(2, t, \mathcal{A}, \mathcal{B})}{\sqrt{t}} \;\geq\; \frac{\gamma}{2}.$$

In particular, even if $\mathcal{A}$ is granted a "warm-up" period during which its regret is ignored, an adversary can still force it to incur large regret afterwards.

The algorithm's description and analysis relies heavily on a function $R\colon \mathbb{R}_{\geq 0} \times \mathbb{R} \to \mathbb{R}$ defined by

$$R(t, g) = \begin{cases} 0 & (t = 0) \\ \frac{g}{2} + \kappa\sqrt{t} \cdot M_0\left(g^2/2t\right) & (t > 0 \text{ and } g \leq \gamma\sqrt{t}) \\ \frac{\gamma\sqrt{t}}{2} & (t > 0 \text{ and } g \geq \gamma\sqrt{t}) \end{cases} \qquad \text{where} \qquad \kappa = \frac{1}{\sqrt{2\pi}\,\text{erfi}(\gamma/\sqrt{2})} \tag{2.6}$$

and M_0 as defined in (2.3). The function R may seem mysterious at first, but in fact arises naturally from the solution to a stochastic calculus problem[9] in Section 2.5. In our usage of this function, t will correspond to the time and g will correspond to the *gap* between (i.e., absolute difference of) the total loss for the two experts. One may verify that R is continuous on $\mathbb{R}_{>0} \times \mathbb{R}$ because the second and third cases agree on the curve $\left\{ \; (t, \gamma\sqrt{t}) \; : \; t > 0 \; \right\}$ since γ satisfies $M_0(\gamma^2/2) = 0$. We next define a function p to be

$$p(t, g) = \tfrac{1}{2}\big(R(t, g+1) - R(t, g-1)\big). \tag{2.7}$$

This is the discrete derivative of R at time t and gap g. The algorithm constructs its distribution x_t so that $p(t, g)$ is the probability mass assigned to the expert with the greatest accumulated loss so far. It is shown later that $p(t, g) \in [0, 1/2]$ whenever $t \geq 1$ and $g \geq 0$ so that p is indeed a probability and the algorithm is well defined. We remark that $p(t, 0) = 1/2$ (Lemma 2.17) for all $t \geq 1$ so the algorithm places equal mass on both experts when their cumulative losses are equal.

2.2.3 Techniques

Lower Bound. The common approach to prove lower bounds in the experts problem is to consider a random adversary that changes the gap by ± 1 at each step. In the fixed-time setting, the adversary has no control over the time horizon; it is known to both the adversary and the algorithm beforehand. The adversary in the anytime setting has the additional power to choose the time horizon, without

[9] As we describe below, the regret against a random adversary is a stochastic integral. Viewing this problem in continuous time, then designing a function to minimize the integral leads to a PDE which R solves.

Algorithm 2 The algorithm achieving the minimax anytime regret for two experts. At each time step, each expert incurs a cost in the interval $[0, 1]$, so the cost vector ℓ_t lies in $[0, 1]^2$.

Initialize $L_0 \leftarrow \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$.
for $t = 1, 2, \ldots$ **do**
    If necessary, swap indices so that $L_{t-1,1} \geq L_{t-1,2}$.
    The current gap is $g_{t-1} \leftarrow L_{t-1,1} - L_{t-1,2}$.
    Set $x_t \leftarrow \big[\, p(t, g_{t-1}),\ 1 - p(t, g_{t-1}) \,\big]$, where p is the function defined by (2.7).
    ▷ Observe cost vector ℓ_t and incur cost $\langle x_t, \ell_t \rangle$.
    $L_t \leftarrow L_{t-1} + \ell_t$
end for

informing the algorithm, and therefore it is not surprising that an adversary using a fixed time horizon does not provide a good anytime lower bound.

To obtain the optimal lower bound, we allow the adversary to select a *random time*, τ, as the time horizon. In general, a random adversary in the anytime setting has the ability to generate an infinitely long sequence of random bits and then select a time horizon as a function of the *entire* sequence. The optimal adversary, on the other hand, comes from a significantly *weaker* class of random adversaries that *do not* examine their sequence of random bits ahead of time. Instead it accesses its random bits one at a time and declares that it is time to stop once the current gap reaches some time-dependent threshold. It is not a priori obvious that the optimal adversary comes from this weaker class. We describe the adversary and its connection to the lower bound below.

First, let us view the regret as a discrete stochastic process. To analyze this stochastic process, we use an elementary identity known as Tanaka's Formula for random walks, which allows us to write the regret process as $\mathrm{Regret}(t) = Z_t + g_t/2$ where Z_t is a martingale with $Z_0 = 0$ and g_t is the current gap at time t. When τ is a sufficiently "nice" *stopping time*[10], the *Optional Stopping Theorem* (O.S.T.) yields $\mathbb{E}\,[Z_\tau] = Z_0 = 0$. (This step is trivial in the fixed-time and geometric horizon settings since they involve stopping times that are always nice.) To use the O.S.T., we restrict ourselves to consider adversaries which select the time horizon to be a nice stopping time.

In particular, we consider adversaries that select τ as the first time that the gap g_t exceeds some time dependent boundary $f(t)$[11]. This approach follows an established doctrine that connects optimal stopping and stochastic control problems to free-boundary problems [37, 117]. Applying the O.S.T., one might expect that $\mathbb{E}\,[\mathrm{Regret}(\tau)] = \mathbb{E}\,[g_\tau]\,/2 \geq \mathbb{E}\,[f(\tau)]\,/2$. Unfortunately, such an argument must involve additional assumptions; otherwise the adversary could just select the boundary $f(t)$ to be arbitrarily large, and the resulting regret lower bound would violate known upper bounds.

The issue lies in the fact that the O.S.T. requires certain conditions on the martingale and stopping time. First observe that it is *not* sufficient for the stopping time to be *almost surely finite*. (Otherwise, one could use a boundary $f(t) = \Theta(\sqrt{t \ln \ln t})$ and the Law of the Iterated Logarithm [65] to prove lower bounds that contradict the $O(\sqrt{t})$ upper bound of Cover or MWU.) Therefore, it is tempting to remove from consideration all adversaries that use a τ with $\mathbb{E}\,[\tau] = \infty$. However, imposing the

[10] Intuitively, a stopping time must make the decision that now is the time to stop without knowledge of future random bits.

[11] Note that $\tau = \min\{\, t \geq 0 : g_t \geq f(t) \,\}$ is a stopping time.

restriction $\mathbb{E}[\tau] < \infty$ makes the adversaries much too weak, and some delicate care is required here, as we will see shortly. We have yet to nail down the boundary, and at this point a lucky guess is required. We will consider boundaries of the form $f(t) = c\sqrt{t}$ since this would be in harmony with the known $\Theta(\sqrt{t})$ regret bounds. To ensure that $\mathbb{E}[\tau] < \infty$, it is known [23, 130] that choosing $c < 1$ is necessary and sufficient. Unfortunately this would yield a regret lower bound of $\sqrt{t}/2$, which is trivial since the algorithm can easily be forced to have regret $1/2$ at time $t = 1$. Therefore, we must relax the restriction that $\mathbb{E}[\tau] < \infty$.

Fortunately there is a strengthening of the O.S.T. with a weaker and somewhat surprising hypothesis that leads to optimal results in our setting. We show that the optimal adversary chooses a stopping time to satisfy this weak hypothesis. This strengthened O.S.T. states: if Z_t is a martingale with bounded increments (i.e. $\sup_{t\geq 0} |Z_{t+1} - Z_t| \leq K$ for some $K > 0$) and τ is a stopping time satisfying $\mathbb{E}\left[\sqrt{\tau}\right] < \infty$, then $\mathbb{E}[Z_\tau] = 0$. The crucial detail is to bound the *expected square root* of τ. This result is stated formally in Theorem 2.29. It remains to choose as large a boundary as possible such that the associated stopping time of hitting the boundary satisfies $\mathbb{E}[\sqrt{\tau}] < \infty$. Using classical results of Breiman [23] and Greenwood and Perkins [79], we show that the optimal choice of c is γ.

Upper Bound. Our analysis of Algorithm 2, to prove the upper bound in Theorem 2.1, uses a deceptively simple argument where R defined in Eq. (2.6) acts as a potential function. Specifically, we show that the change in regret from time $t-1$ with gap g_{t-1} to time t with gap g_t is at most $R(t, g_t) - R(t-1, g_{t-1})$. This implies that $\max_g R(t, g)$ is an upper bound on the regret at time t. The analysis has a number of key features. First, note that the potential function R is bivariate; it depends on both the *time* t as well as the *state* g_t. To deal with this bivariate potential, we use a tool known as the discrete Itô formula. This formula allows us to relate the regret to the potential R, while elegantly handling changes to both time and state. In fact, the potential R turns out to be an extremely tight approximation to the actual regret. Previously, there have been several works that make use of bivariate potentials (e.g. [35, 107]). However, to the best of our knowledge, our work is the first to use the discrete Itô formula in the setting of regret minimization.

The function R and the use of discrete Itô do not come "out of thin air"; they come from considering a continuous-time analogue of the problem. This continuous viewpoint brings a wealth of analytical tools that do not exist (or are more cumbersome) in the discrete setting. As discussed in the lower bound section above, in discrete-time it is natural to assume the gap process evolves as a reflected random walk. In order to formulate the continuous-time problem, we assume that the continuous adversary evolves the gap between the best and worst expert as a reflected Brownian motion (the continuous-time analogue of a random walk). Using this adversary, the continuous-time regret becomes a stochastic integral.

The most natural way to analyze an integral is to use the fundamental theorem of calculus (FTC). However, the continuous-time regret is defined by a stochastic integral so the FTC cannot be applied[12]. However there is a stochastic analog of the FTC, namely the (continuous) Itô formula, which we state in Theorem 2.39. We use it to provide an insightful decomposition of the continuous-time regret. In

[12] The integrator is reflected Brownian motion, which is not of bounded variation.

particular, this decomposition suggests that the algorithm should satisfy an analytic condition known as the *backwards heat equation*. A key resulting idea is: if the algorithm satisfies the backward heat equation, then there is a natural potential function that upper bounds the regret of the algorithm. This enables a systematic approach to obtain an explicit continuous-time algorithm and a potential function that bounds the continuous algorithm's regret. To go back to the discrete setting, using the *same* potential function, we replace applications of Itô's formula with the discrete Itô formula. Remarkably, this leads to *exactly* the same regret bound as the continuous setting.

2.2.4 Application

As mentioned in Section 2.1, the following theorem of Davis can be proven as a corollary of our techniques. Intriguingly, the proof involves regret, despite the fact that regret does not appear in the theorem statement.

Theorem 2.3 (Davis [48])**.** *Let $(X_t)_{t\geq 0}$ be a standard random walk. Then $\mathbb{E}\left[|X_\tau|\right] \leq \gamma \mathbb{E}\left[\sqrt{\tau}\right]$ for every stopping time τ; moreover, the constant γ cannot be improved.*

Proof. We begin by proving the first assertion. Suppose that $\mathrm{Regret}(T)$ is the regret process when Algorithm 2 is used against a random adversary. As discussed in Subsection 2.2.3, we can write the regret process as $\mathrm{Regret}(T) = Z_T + g_T/2$ where Z_T is a martingale and g_T evolves as a reflected random walk. Moreover, if τ is a stopping time satisfying $\mathbb{E}\left[\sqrt{\tau}\right] < \infty$, then $\mathbb{E}\left[Z_\tau\right] = 0$ (see Theorem 2.29).

The upper bound in Theorem 2.1 asserts that $\gamma\sqrt{T}/2 \geq \mathrm{Regret}(T) = Z_T + g_T/2$ for any fixed $T \geq 0$. Hence, $\gamma\mathbb{E}\left[\sqrt{\tau}\right]/2 \geq \mathbb{E}\left[g_\tau\right]/2$. Replacing g_τ with $|X_\tau|$ (since both g_t and $|X_t|$ are reflected random walks), the proof of the first assertion is complete.

The fact that no constant smaller than γ is possible is a direct consequence of the results of Breiman [23] and Greenwood and Perkins [79] as mentioned in Subsection 2.2.3 (see also Section 2.4 or [48]). □

Remark 2.4. *Davis [48] proved Theorem 2.3 for both random walks and Brownian motion. We are also able to recover the result for Brownian motion as a corollary of our continuous-time result (Theorem 2.37). The proof is very similar to that above.*

2.2.5 An expression for the regret involving the gap

In our two-expert prediction problem, the most important scenario restricts each cost vector ℓ_t to be either $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ or $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$. That is, at each time step, some expert incurs cost 1 and the other expert incurs no cost. This restricted scenario is equivalent to the condition $g_t - g_{t-1} \in \{\pm 1\}\ \forall t \geq 1$, where $g_t \coloneqq |L_{t,1} - L_{t,2}|$ is the gap at time t. To prove the optimal lower bound it suffices to consider this restricted scenario. The optimal upper bound is first proven in the restricted scenario, then extended to general cost vectors in Subsection 2.3.3. With the sole exception of Subsection 2.3.3, we assume the restricted scenario.

We now present an expression, valid for any algorithm, that emphasizes how the regret depends on the *change* in the gap. This expression will be useful in proving both the upper and lower bounds. Henceforth we write $\mathrm{Regret}(t) \coloneqq \mathrm{Regret}(2, t, \mathcal{A}, \mathcal{B})$ where $\mathcal{A}$ and $\mathcal{B}$ are usually implicit from the context.

Proposition 2.5. *Assume the restricted setting in which $g_t - g_{t-1} \in \{\pm 1\}$ for every $t \geq 1$. When $g_{t-1} \neq 0$, let p_t denote the probability mass assigned by the algorithm to the "worst expert", i.e., if $L_{t-1,1} \geq L_{t-1,2}$ then $p_t = x_{t,1}$ and otherwise $p_t = x_{t,2}$. The quantity p_t may depend arbitrarily on $\ell_1, \ldots, \ell_{t-1}$. Then*

$$\mathrm{Regret}(T) \;=\; \sum_{t=1}^{T} p_t \cdot (g_t - g_{t-1}) \cdot \mathbb{I}[g_{t-1} \neq 0] \;+\; \sum_{t=1}^{T} \langle x_t, \ell_t \rangle \cdot \mathbb{I}[g_{t-1} = 0]. \tag{2.8}$$

Furthermore, assume that if $g_{t-1} = 0$, then $p_t = x_{t,1} = x_{t,2} = 1/2$. In this case

$$\mathrm{Regret}(T) \;=\; \sum_{t=1}^{T} p_t \cdot (g_t - g_{t-1}). \tag{2.9}$$

Remark 2.6. *Observe that* (2.9) *is a discrete analog of a Riemann-Stieltjes integral of p with respect to g. If $(g_t)_{t\geq 0}$ is a random process, then* (2.9) *is called a discrete stochastic integral. In the specific case that $(g_t)_{t\geq 0}$ is a reflected random walk (the absolute value of a standard random walk), then Eq.* (2.8) *is the Doob decomposition [94, Theorem 10.1] of the regret process $\big(\mathrm{Regret}(t)\big)_{t\geq 0}$, i.e., the first sum is a martingale and the second sum is an increasing predictable process.*

Proof. Define $\Delta_{\mathrm{R}}(t) = \mathrm{Regret}(t) - \mathrm{Regret}(t-1)$. The total cost of the best expert at time t is $L_t^* \coloneqq \min\{L_{t,1}, L_{t,2}\}$. The change in regret at time t is the cost incurred by the algorithm minus the change in the total cost of the best expert, so $\Delta_{\mathrm{R}}(t) = \langle x_t, \ell_t \rangle - (L_t^* - L_{t-1}^*)$.

Case 1: $g_{t-1} \neq 0$**.** In this case, the best expert at time $t-1$ remains a best expert at time t. If the worst expert incurs cost 1, then the algorithm incurs cost p_t and the best expert incurs cost 0, so $\Delta_{\mathrm{R}}(t) = p_t$ and $g_t - g_{t-1} = 1$. Otherwise, the best expert incurs cost 1 and the algorithm incurs cost $1 - p_t$, so $\Delta_{\mathrm{R}}(t) = -p_t$ and $g_t - g_{t-1} = -1$. For either choice of cost, we have $\Delta_{\mathrm{R}}(t) = p_t \cdot (g_t - g_{t-1})$.

Case 2: $g_{t-1} = 0$. Both experts are best, but one incurs no cost, so $L_t^* = L_{t-1}^*$ and $\Delta_{\mathrm{R}}(t) = \langle x_t, \ell_t \rangle$.

The above two cases prove Eq. (2.8). For the last assertion, we have that $\langle x_t, \ell_t \rangle = 1/2 = p_t \cdot (g_t - g_{t-1})$ whenever $g_{t-1} = 0$. Hence, we can collapse the two sums in Eq. (2.8) into one to get Eq. (2.9). □

2.2.6 Background and basic facts on confluent hypergeometric functions

In this subsection, we collect some basic facts about confluent hypergeometric functions which will be useful in the proof of Theorem 2.1 (for both the upper bound and the lower bound).

For any $a, b \in \mathbb{R}$ with $b \notin \mathbb{Z}_{\leq 0}$, the confluent hypergeometric function of the first kind is defined as

$$M(a, b, z) = \sum_{n=0}^{\infty} \frac{(a)_n z^n}{(b)_n n!}, \tag{2.10}$$

where $(x)_n \coloneqq \prod_{i=0}^{n-1}(x+i)$ is the Pochhammer symbol. See, e.g., Abramowitz and Stegun [5, Eq. (13.1.2)].

For notational convenience, for $i \in \{0, 1, 2, \dots, \}$, we write

$$M_i(x) = M(i - 1/2, i + 1/2, x). \tag{2.11}$$

Fact 2.7 ([5, Eq. (13.4.9)])**.** *If $b \notin \mathbb{Z}_{\leq 0}$ then $\frac{\mathrm{d}}{\mathrm{d}x} M(a, b, x) = \frac{a}{b} \cdot M(a + 1, b + 1, x)$. Consequently,*

(1) $M_0'(x) = -M_1(x)$*; and*

(2) $M_1'(x) = \frac{1}{3} \cdot M_2(x)$.

Fact 2.8. *The following identities hold:*

(1) $M_0(x) = -\sqrt{\pi x}\,\mathrm{erfi}(\sqrt{x}) + e^x$.

(2) $M_1(x) = \frac{\sqrt{\pi}\,\mathrm{erfi}(\sqrt{x})}{2\sqrt{x}}$.

(3) $M_2(x) = \frac{3(2e^x\sqrt{x} - \sqrt{\pi}\,\mathrm{erfi}(\sqrt{x}))}{4x^{3/2}}$.

(4) $\frac{2}{3} \cdot M_2(x) \cdot x + M_1(x) = e^x$.

Proof.

(2): See [5], equations (7.1.21) or (13.6.19), and use that $\mathrm{erfi}(x) = -i\,\mathrm{erf}(ix)$, where $i = \sqrt{-1}$.

(1): Differentiating the right-hand side (using the definition of erfi in (2.3)) yields $-\frac{\sqrt{\pi}\,\mathrm{erfi}(\sqrt{x})}{2\sqrt{x}}$. So the right-hand side is an anti-derivative of $-M_1(x)$, by part (2). Thus, the identity (1) follows from Fact 2.7(1) and the initial condition $M_0(0) = 1$.

(3): This follows directly by differentiating (2) and Fact 2.7(2).

(4): Immediate from (2) and (3). □

Fact 2.9. *The function $M_0(x)$ is decreasing and concave on $[0, \infty)$.*

Remark 2.10. *In fact, $M_0(x)$ is decreasing and concave on $\mathbb{R}$ but we do not require this fact.*

Proof. By Fact 2.7, we have $M_0'(x) = -M_1(x)$ and $M_0''(x) = -\frac{1}{3} \cdot M_2(x)$. Note that the coefficients of $M_1(x), M_2(x)$ in their Taylor series are all non-negative. As $x \geq 0$, we have that $M_0'(x), M_0''(x) \leq 0$ as desired. □

Fact 2.11. *The function $x \mapsto M_0(x^2/2)$ has a unique positive root at $x = \gamma$. Moreover $M_0(x^2/2) > 0$ for $x \in (0, \gamma)$ and $M_0(x^2/2) < 0$ for $x \in (\gamma, \infty)$.*

Proof. The Maclaurin expansion of $M_0(x^2/2)$ is given by

$$M_0\left(\frac{x^2}{2}\right) = 1 - \sum_{k=1}^{\infty} \frac{1}{(2k-1)k!} \frac{x^{2k}}{2^k}.$$

Note that $M_0(0) = 1$. It is clear, from the series expansion above (and Fact 2.9), that $M_0(x^2/2)$ is strictly decreasing in x on $(0, \infty)$ and $\lim_{x\to\infty} M_0(x^2/2) = -\infty$. Hence, $M_0(x^2/2)$ contains a positive root γ and it is unique. Finally, it is clear that $M_0(x^2/2)$ is positive on $(0, \gamma)$ and negative on (γ, ∞). □

Claim 2.12. *For any $\varepsilon > 0$, there exists $a_\varepsilon \in (-1, -1/2)$ such that the smallest*[13] *positive root c_ε of $z \mapsto M(a_\varepsilon, 1/2, z^2/2)$ satisfies $c_\varepsilon \geq \gamma - \varepsilon$.*

[13] In fact, there is a unique positive root.

Proof. Following Perkins' notation [114], let $\lambda_0(-c, c)$ be such that c is the smallest positive root of $x \mapsto M(-\lambda_0(-c, c), 1/2, x^2/2)$. By [114, Proposition 1], the map $c \mapsto \lambda_0(-c, c)$ is strictly decreasing and continuous on $\mathbb{R}_{>0}$, so it has a continuous inverse α. From (2.4) and Fact 2.8(1), we see that $\lambda_0(-\gamma, \gamma) = 1/2$, hence $\alpha(1/2) = \gamma$. By continuity, for all $\varepsilon > 0$, there exists $\delta \in (0, 1/2)$ such that $\alpha(1/2 + \delta) > \gamma - \varepsilon$. Then we may take $a_\varepsilon = -(1/2 + \delta)$ and $c_\varepsilon = \alpha(1/2 + \delta)$. □

2.3 Upper bound

In this section, we prove the upper bound in Theorem 2.1 via a sequence of simple steps. We remind the reader that for simplicity, we will assume that the gap changes by ± 1 at each step, which corresponds to each loss vector ℓ_t being either $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ or $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$. The analysis can be extended to general loss vectors in $[0, 1]^2$ through the use of concavity arguments. The details can be found in Subsection 2.3.3.

The proof in this section uses the potential function R which, as explained in Subsection 2.2.3, is defined via continuous-time arguments in Section 2.5. Moreover, the structure of the proof is heavily inspired by the proof in the continuous setting. Finally, we remark that the analysis of this section uses the potential function in a modular way[14], and could conceivably be used to analyze other algorithms (e.g., MWU).

Moving forward, we will need a few observations about the functions R and p, which were defined in equations (2.6) and (2.7).

Lemma 2.13. *For any $t > 0$, $R(t, g)$ is concave and non-decreasing in g.*

The proof of Lemma 2.13 can be verified by "inspecting the curve" and is nothing more than a calculus exercise. First, we have the following two calculations.

Lemma 2.14. *Consider the function $\tilde{R}(t, g) = \frac{g}{2} + \kappa\sqrt{t} M_0\left(\frac{g^2}{2t}\right)$. Then $\frac{\partial}{\partial g}\tilde{R}(t, g) = \frac{1}{2}\left(1 - \frac{\operatorname{erfi}(g/\sqrt{2t})}{\operatorname{erfi}(\gamma/\sqrt{2})}\right)$.*

Lemma 2.15. $\frac{\partial}{\partial g} R(t, g) = \frac{1}{2}\left(1 - \frac{\operatorname{erfi}(g/\sqrt{2t})}{\operatorname{erfi}(\gamma/\sqrt{2})}\right)_+$.

The previous two lemmata are essentially special cases of Lemma 2.45.

Proof of Lemma 2.13. The fact that $R(t, g)$ is non-decreasing in g follows from Lemma 2.15 because its derivative in g is non-negative. The concavity of $R(t, g)$ (in g) follows from the fact that erfi is non-decreasing, so $\frac{\partial}{\partial g} R(t, g)$ is non-increasing in g. □

As a consequence of Lemma 2.13, we can easily obtain the maximum value of $R(t, g)$ for any t.

Lemma 2.16. *For any $t > 0$, we have $R(t, g) \le \gamma\sqrt{t}/2$.*

Proof. Lemma 2.13 shows that $R(t, g)$ is non-decreasing in g. By definition, $R(t, g)$ is constant for $g \ge \gamma\sqrt{t}$. It follows that $\max_g R(t, g) \le R(t, \gamma\sqrt{t}) = \gamma\sqrt{t}/2$. □

[14] Our analysis may also be viewed as an amortized analysis. With this viewpoint, the algorithm incurs amortized regret at most $\frac{\gamma}{2}(\sqrt{t} - \sqrt{t-1}) \approx \gamma/4\sqrt{t}$ at each time step t.

In the definition of the prediction task, the algorithm must produce a probability vector x_t. Recalling the definition of x_t in Algorithm 2, it is not a priori clear whether x_t is indeed a probability vector. We now verify that it is, since Lemma 2.17 implies that $p(t,g) \in [0,1/2]$ for all t, g.

Lemma 2.17. *Fix $t \geq 1$. Then*
(1) $p(t,0) = 1/2$;
(2) $p(t,g)$ is non-increasing in g; and
(3) $p(t,g) \geq 0$.

Proof. For the first assertion, we have

$$p(t,0) = \frac{1}{2}(R(t,1) - R(t,-1)) = \frac{1}{2}\left(\frac{1}{2} + \kappa\sqrt{t}M_0(1/2t) + \frac{1}{2} - \kappa\sqrt{t}M_0(1/2t)\right) = \frac{1}{2}.$$

For the second equality, we used that $1 \leq \gamma \leq \gamma\sqrt{t}$ for all $t \geq 1$. The second assertion follows from concavity of R, which was shown in Lemma 2.13, and an elementary property of concave functions (Fact A.1). The final assertion holds because R is non-decreasing in g, which is also shown in Lemma 2.13. □

2.3.1 Analysis when gap increments are ± 1

In this subsection we prove the upper bound of Theorem 2.1 for a restricted class of adversaries (that nevertheless capture the core of the problem). The analysis is extended to all adversaries in Subsection 2.3.3.

Theorem 2.18. *Let $\mathcal{A}$ be the algorithm described in Algorithm 2. For any adversary $\mathcal{B}$ such that each cost vector ℓ_t is either $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ or $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$, we have*

$$\sup_{t \geq 1} \frac{\mathrm{Regret}(2, t, \mathcal{A}, \mathcal{B})}{\sqrt{t}} \;\leq\; \frac{\gamma}{2}.$$

Our analysis will rely on an identity known as the discrete Itô formula, which is the discrete analogue of Itô's formula from stochastic analysis (see Theorem 2.39). To make this connection (in addition to future connections) more apparent, we define the discrete derivatives of a function f to be

$$\begin{aligned} f_g(t,g) &= \frac{f(t,g+1) - f(t,g-1)}{2}, \\ f_t(t,g) &= f(t,g) - f(t-1,g), \\ f_{gg}(t,g) &= \big(f(t,g+1) + f(t,g-1)\big) - 2f(t,g). \end{aligned}$$

It was remarked earlier that $p(t,g)$ is the discrete derivative of R, and this is because

$$p(t,g) \;=\; R_g(t,g). \tag{2.12}$$

Lemma 2.19 (Discrete Itô formula)**.** *Let $g_0, g_1, \ldots$ be any sequence of real numbers (not necessarily random) satisfying $|g_t - g_{t-1}| = 1$. Then for any function f and any fixed time $T \geq 1$, we have*

$$f(T, g_T) - f(0, g_0) \;=\; \sum_{t=1}^{T} f_g(t, g_{t-1}) \cdot (g_t - g_{t-1}) \;+\; \sum_{t=1}^{T} \left(\frac{1}{2} f_{gg}(t, g_{t-1}) + f_t(t, g_{t-1}) \right). \tag{2.13}$$

This lemma is a small generalization of [94, Example 10.9] to accommodate a bivariate function f that depends on t. The proof is essentially identical and is provided here for completeness.

Proof. By telescoping, $f(T, g_T) - f(0, g_0) = \sum_{t=1}^{T} \big(f(t, g_t) - f(t-1, g_{t-1}) \big)$. Consider a fixed $t \in [T]$. We can write

$$\begin{aligned} f(t, g_t) - f(t-1, g_{t-1}) = &\left(f(t, g_t) - \frac{f(t, g_{t-1}+1) + f(t, g_{t-1}-1)}{2} \right) \\ &+ \left(\frac{f(t, g_{t-1}+1) + f(t, g_{t-1}-1)}{2} - f(t-1, g_{t-1}) \right). \end{aligned} \tag{2.14}$$

For the first bracketed term, by considering the cases $g_t = g_{t-1} + 1$ and $g_t = g_{t-1} - 1$, we have

$$\begin{aligned} f(t, g_t) - \frac{f(t, g_{t-1}+1) + f(t, g_{t-1}-1)}{2} &= \frac{f(t, g_{t-1}+1) - f(t, g_{t-1}-1)}{2} \cdot (g_t - g_{t-1}) \\ &= f_g(t, g_{t-1}) \cdot (g_t - g_{t-1}). \end{aligned} \tag{2.15}$$

Note that the above step is the only place where the assumption that $|g_t - g_{t-1}| = 1$ is used. For the second bracketed term, we have

$$\begin{aligned} \frac{f(t, g_{t-1}+1) + f(t, g_{t-1}-1)}{2} - f(t-1, g_{t-1}) &= \frac{f(t, g_{t-1}+1) + f(t, g_{t-1}-1) - 2f(t, g_{t-1})}{2} \\ &\quad + \big(f(t, g_{t-1}) - f(t-1, g_{t-1}) \big) \\ &= \frac{1}{2} f_{gg}(t, g_{t-1}) + f_t(t, g_{t-1}). \end{aligned}$$

This gives the desired formula. □

Now we show how the regret has a formula similar to (2.13). Recall that Lemma 2.17(1) guarantees $p(t, 0) = 1/2$, i.e., $x_t = [1/2, 1/2]$. Hence, (2.9) gives

$$\text{Regret}(T) = \sum_{t=1}^{T} p(t, g_{t-1}) \cdot (g_t - g_{t-1}) \tag{2.16}$$

where $g_0 = 0$ and $g_t \geq 0$ for all $t \geq 1$.

Key technical step. The following is the most non-obvious step of the proof. We will apply discrete Itô to Eq. (2.16), taking $f = R$. Since $p = R_g = f_g$, observe that the main difference between Eq. (2.13) and Eq. (2.16) is the absence of $\frac{1}{2} f_{gg}(t, g_{t-1}) + f_t(t, g_{t-1})$ in Eq. (2.16). In the continuous setting, we will see that a key idea is to try to obtain a solution satisfying $(\frac{1}{2}\partial_{gg} + \partial_t) f = 0$; this is the well-known

backwards heat equation. In the discrete setting, by a remarkable stroke of luck, we have the following analogous property.

Lemma 2.20 (Discrete backwards heat inequality)**.** $\frac{1}{2}R_{gg}(t,g) + R_t(t,g) \geq 0$ *for all* $t \in \mathbb{R}_{\geq 1}$ *and* $g \in \mathbb{R}_{\geq 0}$.

This lemma is the most technical part of the discrete analysis and we dedicate Subsection 2.3.2 to its proof. We now have all the ingredients needed to prove our main theorem (in the present special case).

Proof of Theorem 2.18. Apply Lemma 2.19 to the function R and the sequence $g_0, g_1, \ldots$ of (integer) gaps produced by the adversary $\mathcal{B}$. Then, for any time $T \geq 0$,

$$\begin{aligned}
&R(T, g_T) - R(0, g_0) \\
&= \sum_{t=1}^{T} R_g(t, g_{t-1}) \cdot (g_t - g_{t-1}) + \sum_{t=1}^{T} \Big(\frac{1}{2} R_{gg}(t, g_{t-1}) + R_t(t, g_{t-1})\Big) && \text{(by Lemma 2.19)} \\
&\geq \sum_{t=1}^{T} p(t, g_{t-1}) \cdot (g_t - g_{t-1}) && \text{(by (2.12) and Lemma 2.20)} \\
&= \text{Regret}(T) && \text{(by (2.16))}.
\end{aligned}$$

Since $g_0 = 0$ and $R(0,0) = 0$, applying Lemma 2.16 shows that $\text{Regret}(T) \leq R(T, g_T) \leq \gamma\sqrt{T}/2$. □

The reader at this point may be wondering why γ is the right constant to appear in the analysis. In Section 2.5, we will define the function R specifically to obtain γ in the preceding analysis. In the next section, our matching lower bound will prove that γ is indeed the right constant.

2.3.2 Proof of Lemma 2.20

In this subsection, we prove the discrete backwards heat inequality (Lemma 2.20). We begin with a few basic facts.

Lemma 2.21. *For all* $u \in [0, 1/2]$*, we have* $M_0(u) \geq \sqrt{1-2u}$.

Proof. The Maclaurin expansion of $M_0(u)$ is given by

$$M_0(u) = 1 - \sum_{k=1}^{\infty} \frac{1}{(2k-1)k!} u^k.$$

Note that $\frac{\mathrm{d}^k}{\mathrm{d}x^k}\sqrt{1-2x} = -\frac{(2k-3)!!}{(1-2x)^{(2k-1)/2}}$, where $(n)!!$ denotes the double factorial (note that $(-1)!! = 1$).[15] Hence, the Maclaurin expansion of $\sqrt{1-2u}$ is

$$\sqrt{1-2u} = 1 - \sum_{k=1}^{\infty} \frac{(2k-3)!!}{k!} u^k.$$

[15] If $n \in \mathbb{Z}_{\geq 0}$, we define $(n)!! = \prod_{k=0}^{\lceil n/2 \rceil - 1}(n-2k)$. If $n \in \mathbb{Z}_{<0}$, we define $(n)!!$ via the recursive relation $(n)!! = \frac{(n+2)!!}{n+2}$ so that $(-1)!! = \frac{(1)!!}{1} = 1$.

It is not hard to verify that $(2k-3)!! \geq \frac{1}{2k-1}$. This implies that $M_0(u) \geq \sqrt{1-2u}$. □

Lemma 2.22. *For all $z \in [0,1)$ and $x \in \mathbb{R}$, we have*

$$M_0\left(\frac{(x+z)^2}{2}\right) + M_0\left(\frac{(x-z)^2}{2}\right) \geq 2\sqrt{1-z^2} M_0\left(\frac{x^2}{2(1-z^2)}\right).$$

Proof. Fix $z \in [0,1)$ and consider the function

$$h_z(x) = M_0\left(\frac{(x+z)^2}{2}\right) + M_0\left(\frac{(x-z)^2}{2}\right) - 2\sqrt{1-z^2} M_0\left(\frac{x^2}{2(1-z^2)}\right).$$

Note that $h_z(0) \geq 0$ by applying Lemma 2.21 with $u = z^2/2$. We will show that $x = 0$ is the minimizer of h_z which implies the lemma.

Indeed, computing derivatives, we have

$$h_z'(x) = -M_1\left(\frac{(x+z)^2}{2}\right) \cdot (x+z) - M_1\left(\frac{(x-z)^2}{2}\right) \cdot (x-z) + 2M_1\left(\frac{x^2}{2(1-z^2)}\right) \cdot \frac{x}{\sqrt{1-z^2}}.$$

As $h_z'(0) = 0$, $x = 0$ is a critical point of h_z. We will now show that h_z is convex which certifies that $x = 0$ is indeed a minimizer.

To obtain h_z'', we differentiate term-by-term. Let $u = \frac{(x+z)^2}{2}$. Then

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x} M_1\left(\frac{(x+z)^2}{2}\right) \cdot (x+z) &= \frac{M_2\left(\frac{(x+z)^2}{2}\right) \cdot (x+z)^2}{3} + M_1\left(\frac{(x+z)^2}{2}\right) \\
&= \frac{2M_2(u) \cdot u}{3} + M_1(u) \\
&= \frac{2u(2e^u\sqrt{u} - \sqrt{\pi}\,\mathrm{erfi}(\sqrt{u}))}{4u^{3/2}} + \frac{\sqrt{\pi}\,\mathrm{erfi}(\sqrt{u})}{2\sqrt{u}} \\
&= e^u = \exp\left(\frac{(x+z)^2}{2}\right).
\end{aligned}$$

The first equality is by Fact 2.7 and the third equality is by identities (2) and (3) in Fact 2.8. We can similarly show that

$$\frac{\mathrm{d}}{\mathrm{d}x} M_1\left(\frac{(x-z)^2}{2}\right) \cdot (x-z) = \exp\left(\frac{(x-z)^2}{2}\right).$$

Finally, for the last term, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x} M_1\left(\frac{x^2}{2(1-z^2)}\right) \cdot \frac{x}{\sqrt{1-z^2}} &= M_2\left(\frac{x^2}{2(1-z^2)}\right) \cdot \frac{x^2}{(1-z^2)^{3/2}} + M_1\left(\frac{x^2}{2(1-z^2)}\right) \cdot \frac{1}{\sqrt{1-z^2}} \\
&= \frac{1}{\sqrt{1-z^2}}\left(M_2\left(\frac{x^2}{2(1-z^2)}\right) \cdot \frac{x^2}{(1-z^2)} + M_1\left(\frac{x^2}{2(1-z^2)}\right)\right) \\
&= \frac{\exp\left(\frac{x^2}{2(1-z^2)}\right)}{\sqrt{1-z^2}},
\end{aligned}$$

where the first equality uses Fact 2.7 and the last equality is by identity (4) in Fact 2.8.

Hence, we have

$$h_z''(x) = \frac{2e^{x^2/2(1-z^2)} - (e^{(x+z)^2/2} + e^{(x-z)^2/2})\sqrt{1-z^2}}{\sqrt{1-z^2}}.$$

So to check that $h_z''(x) \geq 0$ for all $x \in \mathbb{R}$, it suffices to check that

$$\frac{(e^{(x+z)^2/2} + e^{(x-z)^2/2})\sqrt{1-z^2}}{2} \leq e^{x^2/2(1-z^2)}.$$

Indeed, we have

$$\begin{aligned}
\frac{(e^{(x+z)^2/2} + e^{(x-z)^2/2})\sqrt{1-z^2}}{2} &\leq \frac{(e^{(x+z)^2/2} + e^{(x-z)^2/2})e^{-z^2/2}}{2} \\
&= e^{x^2/2}\frac{(e^{xz} + e^{-xz})}{2} \\
&\leq e^{x^2/2}e^{x^2z^2/2} \\
&= e^{x^2(1+z^2)/2} \\
&\leq e^{x^2/2(1-z^2)},
\end{aligned}$$

where the first inequality is because $1 - a \leq e^{-a}$ for all $a \in \mathbb{R}$, the second inequality is because $(e^a + e^{-a})/2 = \cosh(a) \leq e^{a^2/2}$ for all $a \in \mathbb{R}$, and the last inequality is because $1 + a \leq 1/(1-a)$ for all $a < 1$. This proves that h_z is convex which concludes the proof that $x = 0$ is a minimizer for h_z and hence, completes the proof of the lemma. □

We are now ready to prove the discrete backwards heat inequality.

Proof of Lemma 2.20. The inequality $R_t(t,g) + \frac{1}{2}R_{gg}(t,g) \geq 0$ is equivalent to

$$R(t, g+1) + R(t, g-1) \geq 2R(t-1, g). \tag{2.17}$$

We first prove the claim for $t = 1$. In this case, the RHS of Eq. (2.17) is identically 0. On the other hand, the LHS of Eq. (2.17) is non-decreasing in g by Lemma 2.13. Hence, it suffices to prove the inequality for $g = 0$. With $t = 1$ and $g = 0$, we have

$$R(1,1) + R(1,-1) = 2\kappa M_0(1/2).$$

As M_0 is decreasing (Fact 2.9) and $1/2 \leq \gamma^2/2$, we have $M_0(1/2) \geq M_0(\gamma^2/2) = 0$. So Eq. (2.17) holds for $t = 1$ and $g \geq 0$.

For the remainder of the proof, we assume that $t > 1$. Observe that $\gamma\sqrt{t} - 1 \leq \gamma\sqrt{t-1} \leq \gamma\sqrt{t} + 1$ (since $t \geq 1$).[16] We will consider a few cases depending on the value of g.

[16]The inequality $\gamma\sqrt{t} - 1 \leq \gamma\sqrt{t-1}$ is equivalent to $\sqrt{t} - \sqrt{t-1} \leq 1/\gamma$. As $t \mapsto \sqrt{t}$ is concave and $t \geq 1$, the LHS is maximized at $t = 1$ (Fact A.1). Hence, the inequality is true provided $\sqrt{2} \leq 1 + 1/\gamma$. One can check numerically that this last inequality is true as $\gamma \leq 2$.

Case 1: $g \le \gamma\sqrt{t} - 1$**.** In this case, $g+1 \le \gamma\sqrt{t}$, $g \le \gamma\sqrt{t-1}$, and $g-1 \le \gamma\sqrt{t}$. Hence,

$$\begin{aligned} R(t, g+1) &= \frac{g+1}{2} + \kappa\sqrt{t} \cdot M_0\left(\frac{(g+1)^2}{2t}\right) \\ R(t, g-1) &= \frac{g-1}{2} + \kappa\sqrt{t} \cdot M_0\left(\frac{(g-1)^2}{2t}\right) \\ R(t-1, g) &= \frac{g}{2} + \kappa\sqrt{t} \cdot M_0\left(\frac{g^2}{2(t-1)}\right). \end{aligned}$$

So Eq. (2.17) is equivalent to

$$\sqrt{t} \cdot M_0\left(\frac{(g+1)^2}{2t}\right) + \sqrt{t} \cdot M_0\left(\frac{(g-1)^2}{2t}\right) \ge 2\sqrt{t-1} \cdot M_0\left(\frac{g^2}{2(t-1)}\right), \tag{2.18}$$

or rearranging, is equivalent to

$$M_0\left(\frac{(g+1)^2}{2t}\right) + M_0\left(\frac{(g-1)^2}{2t}\right) \ge 2\sqrt{1-1/t} \cdot M_0\left(\frac{g^2}{2(t-1)}\right).$$

The latter inequality is true by Lemma 2.22 using $x = g/\sqrt{t}$ and $z = 1/\sqrt{t} \in (0,1)$.

Case 2: $\gamma\sqrt{t} - 1 \le g \le \gamma\sqrt{t-1}$**.** Let $\tilde{R}$ be the function defined in Lemma 2.14. In this case, we have

$$R(t, g+1) = \gamma\sqrt{t} = \tilde{R}(t, \gamma\sqrt{t}) \ge \tilde{R}(t, g+1) = \frac{g+1}{2} + \kappa\sqrt{t} \cdot M_0\left(\frac{(g+1)^2}{2t}\right).$$

The inequality is by Lemma 2.14 which implies that $\tilde{R}(t, g+1)$ is *non-increasing* for $g \in (\gamma\sqrt{t} - 1, \infty)$. Using the lower bound on $R(t, g+1)$, Eq. (2.17) is again implied by Eq. (2.18) and we have already verified that Eq. (2.18) is true.

Case 3: $\gamma\sqrt{t-1} \le g$**.** Note that for $g \ge \gamma\sqrt{t-1}$, the functions $R(t-1, g)$ and $R(t, g+1)$ are constant in g but $R(t, g-1)$ is non-decreasing in g. Hence, it suffices to check Eq. (2.17) for $g = \gamma\sqrt{t-1}$ which holds by case 2. □

2.3.3 Analysis of Algorithm 2 for general cost vectors

In this section, we prove the upper bound of Theorem 2.1 in full generality.

Theorem 2.23. *Let $\mathcal{A}$ be the algorithm described in Algorithm 2. For any adversary $\mathcal{B}$ (allowing any cost vectors $\ell_t \in [0,1]^2$), we have*

$$\sup_{t \ge 1} \frac{\mathrm{Regret}(2, t, \mathcal{A}, \mathcal{B})}{\sqrt{t}} \le \frac{\gamma}{2}.$$

In Subsection 2.3.1, since the gap was integer-valued, the identity of the best expert could only change when the gap is exactly 0 (at which time there are two best experts). In general, the gap can

be real-valued, so the best expert can switch abruptly, which affects our formula for the regret. We will need to generalize Proposition 2.5 to deal with this possibility. Let $\Delta_{\mathrm{R}}(t) = \mathrm{Regret}(t) - \mathrm{Regret}(t-1)$.

Proposition 2.24. *Let g_{t-1} be the gap after time $t-1$ but before playing an action at time t. Let g_t be the gap after time t. Let $p(t, g_{t-1})$ denote the probability mass assigned to the worst expert at time t. Suppose that $p(t, 0) = 1/2$ for all $t \geq 1$.*

1. *If a best expert at time $t-1$ remains a best expert at time t then*

$$\Delta_{\mathrm{R}}(t) = (g_t - g_{t-1})p(t, g_{t-1}).$$

2. *If a best expert at time $t-1$ is no longer a best expert at time t then*

$$\Delta_{\mathrm{R}}(t) = g_t - (g_t + g_{t-1})p(t, g_{t-1}).$$

Moreover, $g_t + g_{t-1} \leq 1$.

The proof of this is very similar to that of Proposition 2.5

Remark 2.25. *Note that, at any specific time, the set of best experts may have size either one or two so the choice of the best expert in Proposition 2.24 may be ambiguous. However, note that if $g_{t-1} = 0$ (i.e., there are two best experts at time $t-1$) then $p(t, g_{t-1}) = 1/2$ so both formulas give $\Delta_{\mathrm{R}}(t) = \frac{1}{2}g_t$. On the other hand, if $g_t = 0$ (i.e., there are two best experts at time t) then both formulas give $\Delta_{\mathrm{R}}(t) = -g_{t-1}p(t, g_{t-1})$. Hence there is no issue with the ambiguity.*

Proof of Proposition 2.24. Fix t and for notational convenience, let $p = p(t, g_{t-1})$ throughout the proof. In addition, throughout the proof, we use expert 1 to refer to the worst expert at time $t-1$ (chosen arbitrarily if the choice of worst expert is not unique) and use expert 2 to refer to the other expert. Let $\ell_{t,1}, \ell_{t,2} \in [0,1]$ be the respective losses at time t and $L_{t,1}, L_{t,2}$ be the respective *cumulative* losses up to time t. Note that $g_{t-1} = L_{t-1,1} - L_{t-1,2}$. Finally, we set $L_t^* = \min_{i\in[2]} L_{t,i}$. By assumption, $L_{t-1}^* = L_{t-1,2}$.

For the first assertion we have $L_t^* = L_{t,2}$ (because a best expert remains a best expert). Note that $\ell_{t,1} + \ell_{t,2} = (L_{t,1} - L_{t,2}) - (L_{t-1,1} - L_{t-1,2}) = g_t - g_{t-1}$. So the cost of the algorithm can be can be written as

$$p\ell_{t,1} + (1-p)\ell_{t,2} = p(g_t - g_{t-1}) + \ell_{t,2}.$$

On the other hand, $L_t^* - L_{t-1}^* = L_{t,2} - L_{t-1,2} = \ell_{t,2}$. Subtracting this from the above display equation gives $\Delta_{\mathrm{R}}(t) = (g_t - g_{t-1})p$.

In the second assertion, we have $L_t^* = L_{t,1}$. Again, the algorithm incurs cost $p\ell_{t,1} + (1-p)\ell_{t,2}$. This time, note that $\ell_{t,1} - \ell_{t,2} = (L_{t,1} - L_{t,2}) - (L_{t-1,1} - L_{t-1,2}) = -g_t - g_{t-1}$. So the algorithm incurs cost $-p(g_t + g_{t-1}) + \ell_{t,2}$. On the other hand,

$$L_t^* - L_{t-1}^* = L_{t,1} - L_{t-1,2} = L_{t,1} - L_{t-1,1} + L_{t-1,1} - L_{t-1,2} = \ell_{t,1} + g_{t-1} = \ell_{t,2} - g_{t-1},$$

where the last equality uses the identity $\ell_{t,1} - \ell_{t,2} = -g_t - g_{t-1}$. Subtracting this last quantity with the change in the algorithm's cost gives $\Delta_{\mathrm{R}}(t) = g_{t-1} - p(g_t + g_{t-1})$.

To complete the proof for the second assertion, it remains to check that $g_t + g_{t-1} \le 1$. From above, we have the identity, $g_t + g_{t-1} = \ell_{t,2} - \ell_{t,1} \le \ell_{t,2} \le 1$, as desired. □

We will need the following identity which is essentially the same as Lemma 2.19 but without specializing to the case where $|g_t - g_{t-1}| = 1$.

Lemma 2.26. *Let $g_0, g_1, \ldots$ be a sequence of real numbers. Then for any function f and any fixed time $T \ge 1$, we have*

$$\begin{aligned} f(T, g_T) - f(0, g_0) &= \sum_{t=1}^{T} f(t, g_t) - \frac{f(t, g_{t-1}+1) + f(t, g_{t-1}-1)}{2} \\ &\quad + \sum_{t=1}^{T} \left(\frac{1}{2} f_{gg}(t, g_{t-1}) + f_t(t, g_{t-1}) \right). \end{aligned} \tag{2.19}$$

Proof. The proof is identical to the proof of Lemma 2.19 except that we do not perform the simplification in Eq. (2.15). □

When we assumed the gaps were integer-valued, we had

$$\Delta_{\mathrm{R}}(t) = R(t, g_t) - \frac{R(t, g_{t-1}+1) + R(t, g_{t-1}-1)}{2}$$

because both sides were equal to $R_g(t, g_{t-1}) \cdot (g_t - g_{t-1})$. This does not hold in the general setting, but we will be able to prove the following inequality.

Lemma 2.27. *For all $t \ge 1$,*

$$\Delta_{\mathrm{R}}(t) \le R(t, g_t) - \frac{R(t, g_{t-1}+1) + R(t, g_{t-1}-1)}{2}.$$

The proof of Lemma 2.27 appears in Subsection 2.3.4. Given Lemma 2.27, we can now prove our upper bound in general.

Proof of Theorem 2.23. Fix any $T \ge 1$. Then

$$\begin{aligned} R(T, g_T) - R(0, g_0) &= \sum_{t=1}^{T} R(t, g_t) - \frac{R(t, g_{t-1}+1) + R(t, g_{t-1}-1)}{2} \\ &\quad + \sum_{t=1}^{T} \left(\frac{1}{2} R_{gg}(t, g_{t-1}) + R_t(t, g_{t-1}) \right) \quad \text{(Lemma 2.26)} \\ &\ge \sum_{t=1}^{T} \Delta_{\mathrm{R}}(t) \quad \text{(Lemma 2.27 and Lemma 2.20)} \\ &= \mathrm{Regret}(T). \end{aligned}$$

As $g_0 = 0$ and $R(0, 0) = 0$, we have $\mathrm{Regret}(T) \le R(T, g_T) \le \gamma\sqrt{T}/2$, where the last inequality is by Lemma 2.16. □

2.3.4 Proof of Lemma 2.27

Proof of Lemma 2.27. Fix $t \geq 1$. We will consider the two cases corresponding to the two cases in Proposition 2.24.

Case 1: A best expert at time $t-1$ remains a best expert at time t. In this case, $\Delta_{\mathrm{R}}(t) = (g_t - g_{t-1})p(t, g_{t-1})$, so it suffices to check that

$$p(t, g_{t-1}) \cdot (g_t - g_{t-1}) \leq R(t, g_t) - \frac{R(t, g_{t-1}+1) + R(t, g_{t-1}-1)}{2}. \tag{2.20}$$

Rearranging, the above inequality is equivalent to

$$R(t, g_t) - \frac{R(t, g_{t-1}+1) + R(t, g_{t-1}-1)}{2} - p(t, g_{t-1}) \cdot (g_t - g_{t-1}) \geq 0.$$

If g_{t-1} is fixed then notice that the LHS of the above expression is concave in g_t. To see this, Lemma 2.13 implies that $R(t, g_t)$ is concave in g_t, the second term is constant in g_t, and the last term is linear in g_t. Hence, it suffices to verify the inequality when $g_t = g_{t-1} \pm 1$ (Fact A.2). Indeed, if $|g_t - g_{t-1}| = 1$ then

$$\begin{aligned} R(t, g_t) - \frac{R(t, g_{t-1}+1) + R(t, g_{t-1}-1)}{2} &= \frac{R(t, g_{t-1}+1) - R(t, g_{t-1}-1)}{2} \cdot (g_t - g_{t-1}) \\ &= p(t, g_{t-1}) \cdot (g_t - g_{t-1}), \end{aligned}$$

where the second equality used the definition of p.

Case 2: A best expert at time $t-1$ is no longer a best expert at time t. This case is nearly identical to the previous case but in this case $\Delta_{\mathrm{R}}(t) = g_t - (g_t + g_{t-1})p(t, g_{t-1})$ with the promise that $g_t + g_{t-1} \leq 1$. Hence, the inequality we need to verify is that

$$g_t - (g_t + g_{t-1})p(t, g_{t-1}) \leq R(t, g_t) - \frac{R(t, g_{t-1}+1) + R(t, g_{t-1}-1)}{2}. \tag{2.21}$$

Once again, we do this via a concavity argument. Fix $g_{t-1} \in [0, 1]$. Since $g_t + g_{t-1} \leq 1$, we have $g_t \in [0, 1 - g_{t-1}]$. Notice that the LHS of Eq. (2.21) is linear in g_t and the RHS of Eq. (2.21) is concave in g_t (by Lemma 2.13). Hence, it suffices to check the inequality assuming $g_t \in \{0, 1 - g_{t-1}\}$. Note that the case $g_t = 0$ is handled by case 1 since the LHS of Eq. (2.20) and Eq. (2.21) are identical (see also the remark after Proposition 2.24).

Now assume that $g_t = 1 - g_{t-1}$. Then Eq. (2.21) becomes

$$1 - g_{t-1} - p(t, g_{t-1}) \leq R(t, 1 - g_{t-1}) - \frac{R(t, g_{t-1}+1) + R(t, g_{t-1}-1)}{2}$$

Recall that $p(t,g) = \frac{R(t,g+1)-R(t,g-1)}{2}$ so that the above inequality is equivalent to

$$1 - g_{t-1} - \frac{R(t, g_{t-1}+1) + R(t, g_{t-1}-1)}{2} \le R(t, 1-g_{t-1}) - \frac{R(t, g_{t-1}+1) + R(t, g_{t-1}-1)}{2}.$$

Rearranging the inequality becomes

$$1 \le g_{t-1} + R(t, 1-g_{t-1}) - R(t, g_{t-1}-1).$$

Note that $g_{t-1} \le 1 \le \gamma\sqrt{t}$ (since $t \ge 1$ and $\gamma \ge 1$). Hence, by definition of R, the RHS of the above inequality is

$$\begin{aligned} g_{t-1} + R(t, 1-g_{t-1}) - R(t, g_{t-1}-1) &= g_{t-1} + \frac{1-g_{t-1}}{2} + \kappa\sqrt{t} M_0\left(\frac{(1-g_{t-1})^2}{2}\right) \\ &\quad - \frac{g_{t-1}-1}{2} - \kappa\sqrt{t} M_0\left(\frac{(g_{t-1}-1)^2}{2}\right) \\ &= 1, \end{aligned}$$

and obviously, $1 \le 1$. □

2.4 Lower bound

The main result of this section is the following theorem, which implies the lower bound in Theorem 2.1.

Theorem 2.28. *For any algorithm $\mathcal{A}$ and any $\epsilon > 0$, there exists an adversary $\mathcal{B}_\epsilon$ such that*

$$\sup_{t \ge 1} \frac{\mathrm{Regret}(2, t, \mathcal{A}, \mathcal{B}_\epsilon)}{\sqrt{t}} \ \ge \ \frac{\gamma - \varepsilon}{2}. \tag{2.22}$$

In the statement of Theorem 2.28, the sup can be replaced by a lim sup; see Subsection 2.4.1.

It is common in the literature for regret lower bounds to be proven by random adversaries; see, e.g., [31, Theorem 3.7]. We will also consider a random adversary, but the novelty is the use of a non-trivial stopping time at which it can be shown that the regret is large.

A random adversary. Suppose an adversary produces a sequence of cost vectors $\ell_1, \ell_2, \ldots \in \{0,1\}^2$ as follows. For all $t \ge 1$,

- If $g_{t-1} > 0$ then ℓ_t is randomly chosen to be one of the vectors $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ or $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$, uniformly and independent of $\ell_1, \ldots, \ell_{t-1}$. Thus $g_t - g_{t-1}$ is uniform in $\{\pm 1\}$.
- If $g_{t-1} = 0$ then $\ell_t = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ if $x_{t,1} \ge 1/2$, and $\ell_t = \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ if $x_{t,2} > 1/2$. In both cases $g_t = 1$.

As remarked above, the process $(g_t)_{t \ge 0}$ has the same distribution as the absolute value of a standard random walk (which is also known as a reflected random walk).

We now obtain from (2.8) a lower bound on the regret of any algorithm against this adversary. The

adversary's behavior when $g_{t-1} = 0$ ensures that $\langle x_t, \ell_t \rangle \geq 1/2$, showing that

$$\text{Regret}(T) \;\geq\; \underbrace{\sum_{t=1}^{T} p_t \left(g_t - g_{t-1}\right) \cdot \mathbb{I}[g_{t-1} \neq 0]}_{\text{martingale}} \;+\; \underbrace{\frac{1}{2} \sum_{t=1}^{T} \mathbb{I}[g_{t-1} = 0]}_{\text{local time}} \qquad \forall T \in \mathbb{N}.$$

(Equality holds if the algorithm sets $x_t = [1/2, 1/2]$ whenever $g_{t-1} = 0$.) The first sum is a martingale indexed by t. (This holds because $g_t - g_{t-1}$ has conditional expectation 0 when $g_{t-1} \neq 0$, and $\mathbb{I}[g_{t-1} \neq 0] = 0$ when $g_{t-1} = 0$.) The second sum is called the local time of the random walk. Using Tanaka's formula [94, Ex. 10.8], the local time can be written as $\sum_{t=1}^{T} \mathbb{I}[g_{t-1} = 0] = g_t - Z'_t$ where Z'_t is a martingale with uniformly bounded increments and $Z'_0 = 0$. Thus, combining the two martingales, we have

$$\text{Regret}(t) \;\geq\; Z_t + \frac{g_t}{2} \qquad \forall t \in \mathbb{Z}_{\geq 0}, \tag{2.23}$$

where Z_t is a martingale with uniformly bounded increments and $Z_0 = 0$.

Intuition for a stopping time. Optional stopping theorems assert that, under some hypotheses, the expected value of a martingale at a stopping time equals the value at the start. Using such a theorem, at a stopping time τ it would hold that $\mathbb{E}\left[\text{Regret}(\tau)\right] \geq \mathbb{E}\left[g_\tau\right]/2$ (under some hypotheses on τ and Z). Thus it is natural to design a stopping time τ that maximizes $\mathbb{E}\left[g_\tau\right]$ and satisfies the hypotheses. We know from (2.2) that the optimal anytime regret at time t is $\Theta(\sqrt{t})$, so one reasonable stopping time would be

$$\tau(c) \;:=\; \min\left\{\, t > 0 \,:\, g_t \geq c\sqrt{t} \,\right\}$$

for some constant c yet to be determined. If $\tau(c)$ and Z satisfy the hypotheses of the optional stopping theorem, then it will hold that $\mathbb{E}\left[\text{Regret}(\tau(c))\right] \geq \frac{c}{2}\,\mathrm{E}[\,\sqrt{\tau(c)}\,]$. From this, it follows, fairly easily, that $\text{AnytimeNormRegret}(2) \geq c/2$; this will be argued more carefully later.

An optional stopping theorem. The optional stopping theorems appearing in standard references require one of the following hypotheses: (i) τ is almost surely bounded, or (ii) $\mathbb{E}\left[\tau\right]$ is bounded and the martingale has bounded increments, or (iii) the martingale is almost surely bounded and τ is almost surely finite. See, e.g., [24, Theorem 5.33], [65, Theorem 4.8.5], [94, Theorem 10.11], [80, Theorem 12.5.1], [126, Theorem II.57.4], or [143, Theorem 10.10]. These will not suffice for our purposes. For example, condition (ii) is the only useful hypothesis for our setting. It is known [23, 130] that $\mathbb{E}\left[\tau(c)\right] < \infty$, with $\tau(c)$ as above, if and only if $c < 1$; this yields a weak lower bound on the regret. Instead, we will require the following theorem, which has a weaker hypothesis (due to the square root). We are unable to find a reference for this theorem, although it is presumably folklore, so we provide a proof of this result below.

Theorem 2.29. *Let Z_t be a martingale and $K > 0$ a constant such that $|Z_t - Z_{t-1}| \leq K$ almost surely for all t. Let τ be a stopping time. If $\mathbb{E}\left[\sqrt{\tau}\right] < \infty$ then $\mathbb{E}\left[Z_\tau\right] = \mathbb{E}\left[Z_0\right]$.*

Before we prove Theorem 2.29, some preliminary definitions are required. For a martingale

$(X_t)_{t\in\mathbb{N}}$, define its maximum process $X_t^* = \max_{0\le s\le t}|X_s|$ and its quadratic variation process $[X]_t = \sum_{1\le s\le t}(X_s - X_{s-1})^2$.

Theorem 2.30 (Davis [47])**.** *There exists a constant C such that for any martingale $(X_t)_{t\in\mathbb{N}}$ with $X_0 = 0$, $\mathbb{E}[X_\infty^*] \le C\mathbb{E}\left[[X]_\infty^{1/2}\right]$.*

Here, we prove a slightly more general variant of Theorem 2.29. To recover Theorem 2.29, we apply the following theorem with $\sigma = 0$ and then take expectations to get that $\mathbb{E}[Z_\tau] = Z_0$. The more general version will be useful to prove Theorem 2.28 with the sup replaced by a lim sup; see Subsection 2.4.1.

Theorem 2.31. *Let $(Z_t)_{t\in\mathbb{Z}_{\ge 0}}$ be a martingale with respect to a filtration $\{\mathcal{F}_t\}$ and $K > 0$ a constant such that $|Z_t - Z_{t-1}| \le K$ almost surely for all t. Let $\sigma \le \tau$ be stopping times and suppose that $\mathbb{E}[\sqrt{\tau}] < \infty$. Then the random variables Z_σ, Z_τ are almost surely well-defined and $\mathbb{E}[Z_\tau \mid \mathcal{F}_\sigma] = Z_\sigma$.*

Proof. Define the stopped process $Z_{t\wedge\tau}$, which is also a martingale [94, Theorem 10.15]. Since $\mathbb{E}[\sqrt{\tau}] < \infty$ we have $\mathbb{P}[\tau < \infty] = 1$. On the event $\{\tau < \infty\}$, $(Z_{t\wedge\tau})_{t\ge 0}$ has a well-defined limit, which is used as the almost sure definition of Z_τ. As $\{\tau < \infty\} \subseteq \{\sigma < \infty\}$, the same argument shows that $(Z_{t\wedge\sigma})_{\ge 0}$ has a well-defined limit, and we use this as the almost sure definition of Z_σ.

We claim that also $Z_{t\wedge\tau} \xrightarrow{L_1} Z_\tau \in L_1$ and $Z_{t\wedge\sigma} \xrightarrow{L_1} Z_\sigma \in L_1$, from which the theorem concludes as follows. By the definition of conditional expectation, we need to check that $\mathbb{E}[Z_\tau \mathbb{I}_A] = \mathbb{E}[Z_\sigma \mathbb{I}_A]$ for all $A \in \mathcal{F}_\sigma$. To that end, fix $A \in \mathcal{F}_\sigma$ and note that $A \cap \{\sigma \le t\} \in \mathcal{F}_{\sigma\wedge t}$. For any fixed t, $t\wedge\sigma \le t\wedge\tau \le t$, so the optional sampling theorem [94, Theorem 10.11] applied to the stopped process yields $\mathbb{E}[Z_{t\wedge\tau} \mid \mathcal{F}_{t\wedge\sigma}] = Z_{t\wedge\sigma}$. Hence,

$$\mathbb{E}\left[Z_{\tau\wedge t}\mathbb{I}_A\mathbb{I}_{\{\sigma\le t\}}\right] = \mathbb{E}\left[Z_{\sigma\wedge t}\mathbb{I}_A\mathbb{I}_{\{\sigma\le t\}}\right]. \tag{2.24}$$

Since $Z_{\tau\wedge t} \xrightarrow{L_1} Z_\tau \in L_1$, it follows that $Z_{\tau\wedge t}\mathbb{I}_A\mathbb{I}_{\{\sigma\le t\}} \xrightarrow{L_1} Z_\tau\mathbb{I}_A\mathbb{I}_{\{\sigma<\infty\}}$. This is because

$$\begin{aligned}\mathrm{E}\left[|Z_{\tau\wedge t}\mathbb{I}_A\mathbb{I}_{\sigma\le t} - Z_\tau\mathbb{I}_A\mathbb{I}_{\sigma<\infty}|\right] &\le \mathbb{E}\left[|Z_{\tau\wedge t}\mathbb{I}_A\mathbb{I}_{\sigma\le t} - Z_\tau\mathbb{I}_A\mathbb{I}_{\sigma\le t}|\right] + \mathbb{E}\left[|Z_\tau\mathbb{I}_A\mathbb{I}_{\sigma<\infty} - Z_\tau\mathbb{I}_A\mathbb{I}_{\sigma\le t}|\right] \\ &\le \mathbb{E}\left[|Z_{t\wedge\tau} - Z_\tau|\right] + \mathbb{E}\left[|Z_\tau|\,\mathbb{I}_{t<\sigma<\infty}\right].\end{aligned}$$

The quantity $\mathbb{E}[|Z_{t\wedge\tau} - Z_\tau|] \to 0$ because $Z_{t\wedge\tau} \xrightarrow{L_1} Z_\tau$. Next, $Z_\tau \in L_1$ and $\mathbb{I}_{t<\sigma<\infty} \to 0$ a.s. so $\mathbb{E}[|Z_\tau|\,\mathbb{I}_{t<\sigma<\infty}] \to 0$ by dominated convergence. Finally, note that $Z_\tau\mathbb{I}_A\mathbb{I}_{\sigma<\infty} = Z_\tau\mathbb{I}_A$ as $\mathbb{I}_{\sigma<\infty} = 1$ a.s. Hence,

$$\mathbb{E}\left[Z_{\tau\wedge t}\mathbb{I}_A\mathbb{I}_{\{\sigma\le t\}}\right] \xrightarrow{t\to\infty} \mathbb{E}\left[Z_\tau\mathbb{I}_A\right]. \tag{2.25}$$

Similarly,

$$\mathbb{E}\left[Z_{\sigma\wedge t}\mathbb{I}_A\mathbb{I}_{\{\sigma\le t\}}\right] \xrightarrow{t\to\infty} \mathbb{E}\left[Z_\sigma\mathbb{I}_A\right]. \tag{2.26}$$

Combining Eq. (2.24), Eq. (2.25), and Eq. (2.26) gives $\mathbb{E}[Z_\tau\mathbb{I}_A] = \mathbb{E}[Z_\sigma\mathbb{I}_A]$ as desired.

It remains to show that $Z_{\tau\wedge t} \xrightarrow{L_1} Z_\tau \in L_1$ and $Z_{\sigma\wedge t} \xrightarrow{L_1} Z_\sigma \in L_1$. We will only prove the convergence for $Z_{\tau\wedge t}$ as the two arguments are identical. The L_1 convergence is proven using the dominated convergence theorem [94, Corollary 6.26], which requires exhibiting a random variable that bounds $|Z_{t\wedge\tau}|$ for all t and has finite expectation. For notational convenience, let $X_t = Z_{t\wedge\tau}$. Clearly

$|X_t| \le X_t^* \le X_\infty^*$, so it remains to show that $\mathbb{E}[X_\infty^*] < \infty$. Using Theorem 2.30 and that Z has increments bounded by K,

$$\mathbb{E}[X_\infty^*] \;\le\; C\mathbb{E}\left[[X]_\infty^{1/2}\right] \;=\; C\mathbb{E}\left[\Big(\sum_{1\le s\le\tau}(Z_s - Z_{s-1})^2\Big)^{1/2}\right] \;\le\; CK\mathbb{E}\left[\tau^{1/2}\right] \;<\; \infty.$$

The dominated convergence theorem states that $Z_{t\wedge\tau} \xrightarrow{L_1} Z_\tau \in L_1$, as required. □

Optimizing the stopping time. Since the martingale Z_t defined above has bounded increments, Theorem 2.29 may be applied so long as $\mathrm{E}[\sqrt{\tau(c)}] < \infty$, in which case the preceding discussion yields AnytimeNormRegret(2) $\ge c/2$. We reiterate that the condition $\mathrm{E}[\sqrt{\tau(c)}] < \infty$ is a stronger assumption than $\tau(c)$ being almost surely finite. So it remains to determine

$$\sup\{\, c \ge 0 \,:\, \mathrm{E}[\sqrt{\tau(c)}] < \infty \,\}, \tag{2.27}$$

where $\tau(c)$ is the first time at which a standard random walk crosses the two-sided boundary $\pm c\sqrt{t}$. We will use the following result, in which M is the confluent hypergeometric function defined in Subsection 2.2.6.

Theorem 2.32 (Breiman [23], Theorem 2)**.** *Let $c > 1$ and $a < 0$ be such that c is the smallest positive root of the function $x \mapsto M(a, 1/2, x^2/2)$. Then there exists a constant K such that $\mathbb{P}[\tau(c) > u] \sim Ku^a$.*

Remark 2.33. *Breiman's result is not stated in exactly this form because he focused on the case $a \in \mathbb{Z}_{<0}$, in which case M degenerates to a polynomial. One can show by direct calculation that the function $\theta(a)$ in his equation (2.6) is identical to our function $M(a, 1/2, c^2/2)$ for all $a \in \mathbb{R}$.*

An alternative approach is to use a result of Greenwood and Perkins [79, Theorem 5], which shows in a more general context that $\mathbb{P}[\tau(c) > u] = u^{-\lambda_0(-c,c)}\pi(u)$ where $-\lambda_0(-c, c)$ is the largest non-positive eigenvalue of a certain Sturm-Liouville equation and $\pi(u)$ is a "slowly-varying function". It is shown by Perkins [114, Proposition 1] that c is the smallest positive root of $x \mapsto M(-\lambda_0(-c, c), 1/2, x^2/2)$. A standard result [69, Lemma VIII.8.2] states that any slowly-varying function π satisfies $\pi(u) = O(u^\varepsilon)$ for every $\varepsilon > 0$. This alternative approach suffices to prove Theorem 2.28 since (2.28) *is unaffected by the slowly-varying function.*

Recall the definition of γ in (2.4). For intuition, let us apply Theorem 2.32 with $c = \gamma$, which is defined so that it is the root for $a = -1/2$ (see Eq. (2.11) and Fact 2.8). It then follows that

$$\mathbb{E}\left[\sqrt{\tau(\gamma)}\right] \;=\; \int_0^\infty \mathbb{P}\left[\sqrt{\tau(\gamma)} > s\right] \mathrm{d}s \;=\; \int_0^\infty \mathbb{P}\left[\tau(\gamma) > s^2\right] \mathrm{d}s \;\sim\; K\int_0^\infty s^{-1}\mathrm{d}s,$$

by Theorem 2.32. This integral is infinite, so Theorem 2.29 cannot be applied to $\tau(\gamma)$. However, the integral is on the cusp of being finite. By slightly decreasing a below $-1/2$, and slightly modifying c to be the new root, we should obtain a finite integral, showing that $\mathrm{E}[\sqrt{\tau(c)}]$ is finite. The following proof uses analytic properties of M to show that this is possible.

Proof of Theorem 2.28. Fix any $\varepsilon > 0$ that is sufficiently small. Consider the random adversary and the stopping times $\tau(c)$ described above. By Claim 2.12, there exists $a_\varepsilon \in (-1, -1/2)$ and $c_\varepsilon \geq \gamma - \varepsilon$ such that c_ε is the unique positive root of $z \mapsto M(a_\varepsilon, 1/2, z^2/2)$. As in the above calculations, Theorem 2.32 shows that

$$\mathbb{E}\left[\sqrt{\tau(c_\varepsilon)}\right] \;=\; \int_0^\infty \mathbb{P}\left[\tau(c_\varepsilon) > s^2\right] \mathrm{d}s \;\sim\; K \int_0^\infty s^{2a_\varepsilon} \mathrm{d}s \;<\; \infty, \tag{2.28}$$

since $a_\varepsilon < -1/2$. It follows that $\tau(c_\varepsilon)$ is almost surely finite, and therefore $\mathrm{Regret}(\tau(c_\varepsilon))$ and $g_{\tau(c_\varepsilon)}$ are almost surely well defined. Applying Theorem 2.29 to the martingale Z_t appearing in Eq. (2.23), we obtain that

$$\mathbb{E}\left[\mathrm{Regret}(\tau(c_\varepsilon))\right] \;\geq\; \frac{1}{2}\mathbb{E}\left[g_{\tau(c_\varepsilon)}\right] \;=\; \frac{1}{2}\mathbb{E}\left[c_\varepsilon\sqrt{\tau(c_\varepsilon)}\right].$$

By the probabilistic method, there exists a finite sequence of cost vectors $\ell_1, \ldots, \ell_t$ (depending on $\mathcal{A}$ and ε) for which the regret of $\mathcal{A}$ at time t is at least $c_\varepsilon\sqrt{t}/2$. The adversary $\mathcal{B}_\varepsilon$ (which knows $\mathcal{A}$) provides this sequence of cost vectors to algorithm $\mathcal{A}$, thereby proving (2.22). □

2.4.1 Large regret infinitely often

In this subsection, we sketch the following theorem.

Theorem 2.34. *For any algorithm $\mathcal{A}$ and any $\varepsilon > 0$, there exists an adversary $\mathcal{B}_\varepsilon$ such that*

$$\limsup_{t\geq 1} \frac{\mathrm{Regret}(2, t, \mathcal{A}, \mathcal{B}_\epsilon)}{\sqrt{t}} \;\geq\; \frac{\gamma - \varepsilon}{2}. \tag{2.29}$$

Sketch. We use the same adversary as in Theorem 2.28 so that

$$\mathrm{Regret}(t) \geq Z_t + \frac{g_t}{2},$$

where Z_t is a martingale with $Z_0 = 0$ and g_t evolves as a reflected random walk. Let $\mathcal{F}_t \coloneqq \sigma(g_0, \ldots, g_t)$ be the natural filtration. Finally, let $c_\varepsilon \geq \gamma - \varepsilon$ be as in the proof of Theorem 2.28.

Define the stopping times $\tau_0 \coloneqq 0$ and $\tau_i \coloneqq \inf\left\{\, t > \tau_{i-1} \,:\, g_t \geq c_\varepsilon\sqrt{t} \,\right\}$ for $i \geq 1$. Note that, by the strong Markov property, for each $i \geq 1$, the process $\{g_{\tau_{i-1}+t}\}_{t\geq 0}$ is a reflected random walk started at position $g_{\tau_{i-1}} > 0$. Moreover, observe that τ_i is similar to the stopping time used in Theorem 2.28 in that the asymptotics of the boundary are the same but the starting point is perturbed by a (random) additive constant. It is not hard to show (via [79, Theorem 5]) that $\mathbb{E}\left[\sqrt{\tau_i}\right] < \infty$.[17] Hence, we can apply Theorem 2.31 to obtain that $\mathbb{E}\left[Z_{\tau_i} \mid \mathcal{F}_{\tau_{i-1}}\right] = Z_{\tau_{i-1}}$ for all $i \geq 1$.

We will now inductively construct a sequence of events which satisfy the conclusions of the theorem. To that end, define the events

$$A_i = \left\{\tau_i < \infty, Z_{\tau_i} \geq \ldots \geq Z_{\tau_1} \geq 0\right\}.$$

For the base case, we have $A_1 = \{\tau_1 < \infty, Z_{\tau_1} \geq 0\}$. In the proof of Theorem 2.28, we have already verified that $\mathbb{P}[A_1] > 0$ (this also follows from the previous paragraph). For the inductive step, suppose

[17] Verifying that $\mathbb{E}[\sqrt{\tau_i}] < \infty$ is the only non-rigorous portion of the proof.

that $\mathbb{P}[A_{i-1}] > 0$. The condition that $\mathbb{E}\left[Z_{\tau_i} \mid \mathcal{F}_{\tau_{i-1}}\right] = Z_{\tau_{i-1}}$ implies that, for any $B \in \mathcal{F}_{\tau_{i-1}}$ with $\mathbb{P}[B] > 0$, the event $B \cap \left\{\tau_i < \infty, Z_{\tau_i} \geq Z_{\tau_{i-1}}\right\}$ has positive probability. Taking $B = A_{i-1}$ implies that $\mathbb{P}[A_i] > 0$.

To conclude, for any $n \geq 1$, the event A_n has positive probability. Hence, there exists a sequence of times $T_1, \ldots, T_n < \infty$ and loss vectors up to time T_n that guarantee $g_{T_i} \geq c_\varepsilon \sqrt{T_i}$ for all $i \in [n]$ and $Z_{T_n} \geq \ldots \geq Z_{T_1} \geq 0$. In particular, for all $i \in [n]$,

$$\text{Regret}(T_i) \geq Z_{T_i} + \frac{g_{T_i}}{2} \geq \frac{c_\varepsilon}{2}\sqrt{T_i}.$$

As $n \geq 1$ was arbitrary, the theorem follows. □

2.5 Derivation of a continuous-time analogue of Algorithm 2

The purpose of this section is to show how the potential function R defined in Eq. (2.6) arises naturally as the solution of a stochastic calculus problem. The derivation of that function is accomplished by defining, then solving, an analogue of the regret minimization problem in continuous time. The main advantage of considering this continuous setting is the wealth of analytic methods available, such as stochastic calculus.

2.5.1 Defining the continuous regret problem

Continuous time regret problem. The continuous regret problem is inspired by Eq. (2.9). Notice that, when the adversary chooses cost vectors in $\{\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right], \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]\}$, the sequence of gaps $g_0, g_1, g_2, \ldots$ live in the support of a reflected random walk. The goal in the discrete case is to find an algorithm p that bounds the regret over all possible sample paths of a reflected random walk. In continuous time it is natural to consider a stochastic integral with respect to reflected Brownian motion, denoted $|B_t|$, instead. Our goal now is to find a continuous-time algorithm whose regret is small for almost all reflected Brownian motion paths.

Definition 2.35 (Continuous Regret)**.** Let $p : \mathbb{R}_{>0} \times \mathbb{R}_{\geq 0} \to [0, 1]$ be a continuous function that satisfies $p(t, 0) = 1/2$ for every $t > 0$. Let B_t be a standard one-dimensional Brownian motion. Then, the *continuous regret* of p with respect to B is the stochastic integral

$$\text{ContRegret}(T, p, B) \;=\; \int_0^T p(t, |B_t|)\mathrm{d}\,|B_t|\,. \tag{2.30}$$

Remark 2.36. *The condition $p(t, 0) = 1/2$ is due to Eq.* (2.30) *being inspired by Eq.* (2.9)*, which requires this condition.*

In this definition we may think of p as a continuous-time algorithm and B as a continuous-time adversary. The goal for the remainder of this section is to prove the following result.

Theorem 2.37. *There exists a continuous-time algorithm p^* such that*

$$\text{ContRegret}(T, p^*, B) \;\le\; \frac{\gamma\sqrt{T}}{2} \qquad \forall T \in \mathbb{R}_{\ge 0},\ \textit{almost surely.} \tag{2.31}$$

Remark 2.38. *A natural question arises upon reviewing the definition of continuous regret: What role does Brownian motion play in Definition 2.35 and is it the "correct" stochastic process to consider in order to uncover the optimal algorithm? In the analysis that follows, the only properties of reflected Brownian motion that we use are its non-negativity and that its* quadratic variation *is t. It turns out that one can generalize Theorem 2.37 by allowing any non-negative, continuous semi-martingale X to control the gap process, and by letting time grow at the rate of the* quadratic variation *of X. See Subsection 2.5.4 for more details.*

2.5.2 Connections to stochastic calculus and the backward heat equation

Since $\text{ContRegret}(T)$ evolves as a stochastic integral with respect to a semi-martingale[18] (namely reflected Brownian motion), Itô's lemma provides an insightful decomposition. The following statement of Itô's lemma is a specialization of [124, Theorem IV.3.3] for the special case of reflected Brownian motion.[19]

Notation. Up to now, we have used the symbol g as the second parameter to the bivariate functions p and R. Henceforth, it will be more consistent with the usual notation in the literature to use x to denote g. We will also use the notation $C^{1,2}$ to denote the class of bivariate functions that are continuously differentiable in their first argument and twice continuously differentiable in their second argument.

Theorem 2.39 (Itô's formula)**.** *Let $f\colon \mathbb{R}_{\ge 0} \times \mathbb{R} \to \mathbb{R}$ be $C^{1,2}$. Then, almost surely,*

$$f(T, |B_T|) - f(0, |B_0|) \;=\; \int_0^T \partial_x f(t, |B_t|)\mathrm{d}\,|B_t| + \int_0^T \Big[\underbrace{\partial_t f(t, |B_t|) + \tfrac{1}{2}\partial_{xx} f(t, |B_t|)}_{=:\overset{*}{\Delta} f(t, |B_t|)}\Big]\mathrm{d}t. \tag{2.32}$$

The integrand of the second integral is an important quantity arising in PDEs and stochastic processes (see, e.g., [62, pp. 263]). We will denote it by $\overset{*}{\Delta} f(t,x) \coloneqq \partial_t f(t,x) + \frac{1}{2}\partial_{xx} f(t,x)$. Some discussion about the statement of Theorem 2.39 appears in Appendix A.2.2.

Applying Itô's formula to the continuous regret. The continuous regret is defined by a stochastic integral. In standard calculus, when presented with an integral to evaluate, we usually turn to the Fundamental Theorem of Calculus (FTC) for intuition and insight. The analogue of the FTC for stochastic calculus is Itô's formula. In order to apply Itô's formula to the continuous regret, we pattern match Eq. (2.30) and Eq. (2.32). Comparing these equations, it is natural to assume that $p = \partial_x f$

[18] A semi-martingale is a stochastic process that can written as the sum of a local martingale and a process of finite variation.

[19] Specifically, we are using the statement of Itô's formula that appears in Remark 1 after Theorem IV.3.3 in [124] with $X_t = |B_t|$ and $A_t = t$. Note that y in their notation is t in ours and $\langle |B|, |B| \rangle_t = t$.

for a function f that is $C^{1,2}$ with $f(0,0) = 0$, $\partial_x f \in [0,1]$, and $\partial_x f(t,0) = 1/2$; the latter two conditions are needed for Definition 2.35 to be applicable. Itô's formula then yields

$$\text{ContRegret}(T, p = \partial_x f, B) \;=\; \int_0^T \partial_x f(t, |B_t|) \mathrm{d}\,|B_t| \;=\; f(T, |B_T|) - \int_0^T \overset{*}{\Delta} f(t, |B_t|) \mathrm{d}t. \tag{2.33}$$

Path independence and the backward heat equation. At this point a useful idea arises: as a thought experiment, suppose that $\overset{*}{\Delta} f = 0$. Then the second integral would vanish, and we would have the appealing expression $\text{ContRegret}(T, p, B) = f(T, |B_T|)$. Moreover, since f is a deterministic function, the right-hand side depends only on $|B_T|$ rather than the entire Brownian path $B|_{[0,T]}$. Thus, the same must be true of the left-hand side: at time T, the continuous regret of the algorithm p depends only on T and $|B_T|$ (the gap). We say that say that such an algorithm has *path independent regret*. Our supposition that led to these attractive consequences was only that $\overset{*}{\Delta} f = 0$, which turns out to be a well studied condition.

Definition 2.40. Let $f \colon \mathbb{R}_{>0} \times \mathbb{R} \to \mathbb{R}$ be a $C^{1,2}$ function. If $\overset{*}{\Delta} f(t,x) = 0$ for all $(t,x) \in \mathbb{R}_{>0} \times \mathbb{R}$ then we say that f satisfies the *backward heat equation*. A synonymous statement is that f is *space-time harmonic*.

We may summarize the preceding discussion with the following proposition.

Proposition 2.41. *Let $f : \mathbb{R}_{>0} \times \mathbb{R} \to \mathbb{R}$ be a $C^{1,2}$ function that satisfies $\overset{*}{\Delta} f = 0$ everywhere with $f(0,0) = 0$. Let $p = \partial_x f$. Then,*

$$\int_0^T p(t, |B_t|) \mathrm{d}\,|B_t| \;=\; f(T, |B_T|). \tag{2.34}$$

Suppose that a function f satisfies the hypothesis of Proposition 2.41 *and in addition* $p = \partial_x f \in [0,1]$ with $p(t,0) = 1/2$. Then, we would have

$$\text{ContRegret}(T, p, B) = f(T, |B_T|). \tag{2.35}$$

We are unable to derive a function that satisfies the properties required for Eq. (2.35) to hold along with $\max_{x \geq 0} f(T,x) \leq \gamma\sqrt{T}/2$. Instead, we will begin by relaxing the constraint that $p(t,x) \in [0,1]$ and allow $p(t,x)$ to be negative. We will overload the notation $\text{ContRegret}(\cdot)$ to include such functions. In the next section, we will derive a family of such functions that all achieve $\text{ContRegret}(T, p, |B_T|) = f(T, |B_T|) = O(\sqrt{T})$. This is done by setting up and solving the backwards heat equation. Next, we use a "smoothing" argument to obtain a family of functions that all achieve $\text{ContRegret}(T, p, |B_T|) = O(\sqrt{T})$, and that *do* satisfy $p(t,x) \in [0,1]$. Finally, we will optimize $\text{ContRegret}(T, \cdot, |B_T|)$ over this family of functions to prove Theorem 2.37. The constant γ will appear as a consequnce of this optimization problem.

Satisfying the backward heat equation

The main result of this section is the derivation of a family of functions $\tilde{p} : \mathbb{R}_{>0} \times \mathbb{R} \to \mathbb{R}$ that satisfy $\tilde{p}(t,x) \leq 1$, $\tilde{p}(t,0) = 1/2$ and

$$\text{ContRegret}(T, \tilde{p}, B) \;=\; f(T, |B_T|) \;=\; O(\sqrt{T}), \tag{2.36}$$

but do not necessarily satisfy $\tilde{p}(t,x) \geq 0$.

The first step is to find a function f which satisfies the partial differential equation $\overset{*}{\Delta} f = 0$. Since the boundary condition $\tilde{p}(t,0) = 1/2$ is a condition on $\tilde{p} = \partial_x f$, not on f itself, it will be convenient to solve a PDE for $\tilde{p}$ instead, and then derive f by integrating. However, some care is needed since not all antiderivates of $\tilde{p}$ (in x) will satisfy the backwards heat equation. Fortunately, we have a useful lemma showing that if $\tilde{p}$ satisfies the backward heat equation, then we can construct an f that also does.

Lemma 2.42. *Suppose that $h : \mathbb{R}_{>0} \times \mathbb{R} \to \mathbb{R}$ is a $C^{1,2}$ function. Define*

$$f(t,x) := \int_0^x h(t,y)\mathrm{d}y - \frac{1}{2}\int_0^t \partial_x h(s,0)\mathrm{d}s.$$

Then,

(1) $f \in C^{1,2}$,
(2) If $\overset{*}{\Delta} h = 0$ over $\mathbb{R}_{>0} \times \mathbb{R}$ then $\overset{*}{\Delta} f = 0$ over $\mathbb{R}_{>0} \times \mathbb{R}$,
(3) $h = \partial_x f$.

Proof. First, we check that $f \in C^{1,2}$. Let $(t,x) \in \mathbb{R}_{>0} \times \mathbb{R}$. It is easy to check via standard applications of the Dominated Convergence Theorem (DCT) and the Fundamental Theorem of Calculus (FTC) that

(1) $\partial_t f(t,x) = \int_0^x \partial_t h(t,y)\mathrm{d}y - \frac{1}{2}\partial_x h(s,0)$,
(2) $\partial_x f(t,x) = h(t,x)$, and
(3) $\partial_{xx} f(t,x) = \partial_x h(t,x)$.

All of the above partial derivatives are clearly continuous since h is $C^{1,2}$.

Next, we show that if $\overset{*}{\Delta} h(t,x) = 0$ for all $(t,x) \in \mathbb{R}_{>0} \times \mathbb{R}$, then $\overset{*}{\Delta} f(t,x) = 0$ for all $\mathbb{R}_{>0} \times \mathbb{R}$. By DCT and FTC,

$$\begin{aligned}
\overset{*}{\Delta} f(t,x) \;&=\; \left(\partial_t + \frac{1}{2}\partial_{xx}\right)\left(\int_0^x h(t,y)\mathrm{d}y - \int_0^t \frac{1}{2}\partial_x h(s,0)\mathrm{d}s\right) \\
&=\; \int_0^x \partial_t h(t,y)\mathrm{d}y + \frac{1}{2}\partial_{xx}\int_0^x h(t,y)\mathrm{d}y - \left(\partial_t + \frac{1}{2}\partial_{xx}\right)\int_0^t \frac{1}{2}\partial_x h(s,0)\mathrm{d}s \qquad \text{(by DCT)} \\
&=\; \int_0^x \partial_t h(t,y)\mathrm{d}y + \frac{1}{2}\partial_x h(t,x) - \frac{1}{2}\partial_x h(t,0) \qquad \text{(by FTC)} \\
&=\; \int_0^x \underbrace{\left(\partial_t h(t,y) + \frac{1}{2}\partial_{xx} h(t,y)\right)}_{=0}\mathrm{d}y \qquad \text{(by FTC)} \\
&=\; 0.
\end{aligned}$$

An application of FTC shows that $\partial_x f(t,x) = h(t,x)$ for every (t,x) as $y \mapsto h(t,y)$ is continuous. $\square$

Defining boundary conditions for p. Obtaining a particular solution to the backward heat equation requires sufficient boundary conditions in order to uniquely identify $\tilde{p}$. The boundary condition mentioned above is that $\tilde{p}(t,0) = 1/2$ for all t. This condition together with the backward heat equation clearly do not suffice to uniquely determine $\tilde{p}$. Therefore, we impose some reasonable boundary conditions on $\tilde{p}$.

What should the value be at the boundary? Intuitively, $x \mapsto \tilde{p}(t,x)$ should be a decreasing function because $\tilde{p}$ represents the weight placed on the worst expert as a function of the gap. Therefore, it is natural to consider an "upper boundary" which specifies the point at which the difference in experts' total costs is so great that the algorithm places zero weight on the worst expert. The upper boundary can be specified by a curve, $\{\ (t,\phi(t))\ :\ t>0\ \}$ for some continuous function $\phi : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$. We will incorporate this idea by requiring $\tilde{p}(t,\phi(t)) = 0$ for all $t > 0$.

Where should the boundary be? One reasonable choice for the boundary is to use $\phi_\alpha(t) = \alpha\sqrt{t}$ for some constant $\alpha > 0$, as this is similar to the boundary used by the random adversary in the lower bound of Section 2.4. These conditions are combined into the following partial differential equation:

$$\begin{aligned}
&\text{(backward heat equation)} && \partial_t u(t,x) + \tfrac{1}{2}\partial_{xx} u(t,x) \;=\; 0 && \text{for all } (t,x) \in \mathbb{R}_{>0} \times \mathbb{R} && (2.37)\\
&\text{(upper boundary)} && u(t,\alpha\sqrt{t}) \;=\; 0 && \text{for all } t > 0 && (2.38)\\
&\text{(lower boundary)} && u(t,0) \;=\; \tfrac{1}{2} && \text{for all } t > 0. && (2.39)
\end{aligned}$$

Next we show that the following function solves this PDE. Define $\tilde{p}_\alpha : \mathbb{R}_{>0} \times \mathbb{R} \to \mathbb{R}$ by

$$\tilde{p}_\alpha(t,x) \;\coloneqq\; \frac{1}{2}\left(1 - \frac{\operatorname{erfi}\left(x/\sqrt{2t}\right)}{\operatorname{erfi}\left(\alpha/\sqrt{2}\right)}\right). \tag{2.40}$$

Lemma 2.43. *$\tilde{p}_\alpha$ satisfies the following properties:*

(1) $\tilde{p}_\alpha$ is $C^{1,2}$ over $\mathbb{R}_{>0} \times \mathbb{R}$,

(2) $\tilde{p}_\alpha$ satisfies the constraints in Eq. (2.37), Eq. (2.38) and Eq. (2.39), and

(3) For all $t > 0$ and all $x \geq 0$, $\tilde{p}_\alpha(t,x) \leq 1/2$.

Proof. Let us assume that we can write $u(t,x) = v(x/\sqrt{t})$. Then, we have $\partial_t u(t,x) = -\frac{x}{2t^{3/2}} v'(x/\sqrt{t})$, and $\frac{1}{2}\partial_{xx} u(t,x) = \frac{1}{2t} v''(x/\sqrt{t})$. The backward heat equation enforces that $v''(x/\sqrt{t}) = \frac{x}{\sqrt{t}} v'(x/\sqrt{t})$. By a change of variables ($z = x/\sqrt{t}$), we obtain the following ordinary differential equation

$$v''(z) \;=\; z \cdot v'(z). \tag{2.41}$$

Hence, $v'(z) = C \cdot e^{\frac{z^2}{2}}$ for some constant C. We can then integrate to obtain $v(z) = \int_0^z Ce^{y^2/2}\,\mathrm{d}y + D = \int_0^{z/\sqrt{2}} \sqrt{2}Ce^{u^2}\,\mathrm{d}u + D$, for some constant D. For the last equality, we made the change of variables $u = y/\sqrt{2}$ in the integral. Therefore, by the definition of erfi (and replacing $C\sqrt{2}$ with C), we have

$v(z) = C\,\mathrm{erfi}(z/\sqrt{2}) + D$. Hence, for some constants $C, D \in \mathbb{R}$, we have

$$u(t,x) \;=\; C\,\mathrm{erfi}(x/\sqrt{2t}) + D.$$

Plugging in the boundary condition at $x = 0$ and recalling that $\mathrm{erfi}(0) = 0$ we see that $D = 1/2$. Plugging in the boundary condition that $u(t, \alpha\sqrt{t}) = 0$ and using that $D = 1/2$ we see that $C = -\frac{1}{2\,\mathrm{erfi}(\alpha/\sqrt{2})}$. Therefore, we have that the following function

$$q(t,x) \;=\; \frac{1}{2}\left(1 - \frac{\mathrm{erfi}\left(x/\sqrt{2t}\right)}{\mathrm{erfi}\left(\alpha/\sqrt{2}\right)}\right)$$

satisfies the backwards heat equation and the boundary conditions. Moreover, $q \in C^{1,2}$ on $\mathbb{R}_{>0} \times \mathbb{R}$. □

Lemma 2.43 shows that $\tilde{p}_\alpha(t,x)$ nearly defines a valid continuous time algorithm, in that it satisfies the conditions of Definition 2.35 except for non-negativity. Next, we will integrate $\tilde{p}_\alpha$ as described in Lemma 2.42. Define the function $\tilde{R}_\alpha \colon \mathbb{R}_{>0} \times \mathbb{R} \to \mathbb{R}$ as

$$\tilde{R}_\alpha(t,x) = \frac{x}{2} + \kappa_\alpha\sqrt{t} \cdot M_0\left(\frac{x^2}{2t}\right) \qquad \text{where} \qquad \kappa_\alpha = \frac{1}{\sqrt{2\pi}\,\mathrm{erfi}(\alpha/\sqrt{2})}. \tag{2.42}$$

Lemma 2.44. $\tilde{R}_\alpha(t,x) = \int_0^x \tilde{p}_\alpha(t,y)\,\mathrm{d}y - \frac{1}{2}\int_0^t \partial_x \tilde{p}_\alpha(s,0)\,\mathrm{d}s$.

First we need to compute some derivatives.

Lemma 2.45. *The following identities hold for every $\alpha > 0$.*

1. $\partial_x \tilde{R}_\alpha(t,x) = \tilde{p}_\alpha(t,x) = \frac{1}{2}\left(1 - \frac{\mathrm{erfi}(x/\sqrt{2t})}{\mathrm{erfi}(\alpha/\sqrt{2t})}\right)$.
2. $\partial_{xx} \tilde{R}_\alpha(t,x) = \partial_x \tilde{p}_\alpha(t,x) = -\kappa_\alpha \cdot \frac{\exp(x^2/2t)}{\sqrt{2t}}$.

Proof. The proof is a straightforward calculation. We have

$$\begin{aligned}
\partial_x \tilde{R}_\alpha(t,x) &= \frac{1}{2} - \kappa_\alpha \frac{x}{\sqrt{t}} \cdot M_1\left(\frac{x^2}{2t}\right) \\
&= \frac{1}{2} - \frac{1}{\sqrt{2\pi}\,\mathrm{erfi}(\alpha/\sqrt{2})} \cdot \frac{x}{\sqrt{t}} \cdot \frac{\sqrt{\pi}\,\mathrm{erfi}(x/\sqrt{2t})}{2 \cdot x/\sqrt{2t}} \\
&= \frac{1}{2}\left(1 - \frac{\mathrm{erfi}(x/\sqrt{2t})}{\mathrm{erfi}(\alpha/\sqrt{2})}\right),
\end{aligned}$$

where the first equality uses Fact 2.7 and the second equality uses the identity (2) in Fact 2.8. This proves the first identity.

For the second identity, using the definition of $\mathrm{erfi}(\cdot)$, we have

$$\partial_{xx} \tilde{R}_\alpha = \partial_x \tilde{p}_\alpha(t,x) = -\frac{\exp(x^2/2t)}{\sqrt{2\pi}\,\mathrm{erfi}(\alpha/\sqrt{2})\sqrt{s}} = -\kappa_\alpha \cdot \frac{\exp(x^2/2t)}{\sqrt{2t}}.$$ □

Proof of Lemma 2.44. By the first identity in Lemma 2.45, we have

$$\int_0^x \tilde{p}_\alpha(t,y)\,\mathrm{d}y = \tilde{R}_\alpha(t,x) - \tilde{R}_\alpha(t,0) \tag{2.43}$$

Note that $\tilde{R}_\alpha(t,0) = \kappa_\alpha\sqrt{t}$. Next, the second identity of Lemma 2.45 implies that $-\partial_x\tilde{p}_\alpha(s,0) = \frac{\kappa_\alpha}{2\sqrt{s}}$. Hence,

$$-\frac{1}{2}\int_0^t \partial_x\tilde{p}_\alpha(s,0)\,\mathrm{d}s = \kappa_\alpha\sqrt{t} = \tilde{R}_\alpha(t,0). \tag{2.44}$$

Summing Eq. (2.43) and Eq. (2.44) gives

$$\int_0^x \tilde{p}_\alpha(t,y)\,\mathrm{d}y - \frac{1}{2}\int_0^t \partial_x\tilde{p}_\alpha(s,0)\,\mathrm{d}s = \tilde{R}_\alpha(t,x) - \tilde{R}_\alpha(t,0) + \tilde{R}_\alpha(t,0) = \tilde{R}_\alpha(t,x). \qquad \square$$

By Lemma 2.43, the function $\tilde{p}_\alpha$ satisfies the hypothesis of the function h in Lemma 2.42. Hence, we can apply Lemma 2.42 with $h = \tilde{p}_\alpha$ and $f = \tilde{R}_\alpha$ to assert the following properties on $\tilde{R}_\alpha$.

Lemma 2.46. *$\tilde{R}_\alpha$ satisfies the following properties:*
(1) $\tilde{R}_\alpha$ is $C^{1,2}$,
(2) $\tilde{R}_\alpha$ satisfies $\overset{*}{\Delta}\tilde{R}_\alpha = 0$ over $\mathbb{R}_{>0}\times\mathbb{R}$,
(3) $\partial_x\tilde{R}_\alpha(t,x) = \tilde{p}_\alpha(t,x)$.

Since $\mathrm{erfi}(\cdot)$ is a strictly increasing function with $\mathrm{erfi}(0) = 0$, observe that $\tilde{p}_\alpha$ has exactly one root at $\alpha\sqrt{t}$. Therefore, for every T, we have

$$\mathrm{ContRegret}(T,\tilde{p}_\alpha,B) \;=\; \tilde{R}_\alpha(T,|B_T|) \;\le\; \max_{x\ge 0}\tilde{R}_\alpha(T,x) \;\le\; \left(\frac{\alpha}{2} + \kappa_\alpha M_0\left(\frac{\alpha^2}{2}\right)\right)\sqrt{T}. \tag{2.45}$$

This establishes (2.36), as desired.

Resolving the non-negativity issue

The only remaining step is to modify $\tilde{p}_\alpha$ so that it lies in the interval $[0,1/2]$. We will modify $\tilde{p}_\alpha$ in the most natural way: by modifying all negative values to be zero. Specifically, we set

$$p_\alpha(t,x) \;\coloneqq\; \begin{cases} 0 & (t=0) \\ (\tilde{p}_\alpha(t,x))_+ & (t>0) \end{cases} \;=\; \begin{cases} 0 & (t=0) \\ \frac{1}{2}\left(1 - \frac{\mathrm{erfi}(x/\sqrt{2t})}{\mathrm{erfi}(\alpha/\sqrt{2})}\right)_+ & (t>0) \end{cases}. \tag{2.46}$$

Here, we use the notation $(x)_+ = \max\{0,x\}$. Note that $p_\alpha(t,0) = 1/2$ for all $t>0$ and $p_\alpha(t,x)\in[0,1/2]$ for all $t,x\ge 0$. So p_α defines a valid continuous-time algorithm. From Eq. (2.46), we obtain a truncated version of $\tilde{R}_\alpha$ as

$$R_\alpha(t,x) \;\coloneqq\; \begin{cases} 0 & (t=0) \\ \tilde{R}_\alpha(t,x) & (t>0\wedge x\le\alpha\sqrt{t}) \\ \tilde{R}_\alpha(t,\alpha\sqrt{t}) & (t>0\wedge x\ge\alpha\sqrt{t}) \end{cases}. \tag{2.47}$$

It is straightforward to verify that $\partial_x R_\alpha = p_\alpha$. This is because for $x \le \alpha\sqrt{t}$, $p_\alpha(t,x) = \tilde{p}_\alpha(t,x)$ and $R_\alpha(t,x) = \tilde{R}_\alpha(t,x)$ (we have computed the derivatives in Lemma 2.46). In addition, $R_\alpha(t,x)$ is constant (in x) for $x \ge \alpha\sqrt{t}$ so its derivative (in x) is 0.

If R_α were sufficiently smooth then we could immediately apply Eq. (2.35) (or Theorem 2.39) to obtain a formula for the regret of p_α. The only flaw is that $\partial_{xx} R_\alpha$ is not well-defined on the curve $\left\{ (t, \alpha\sqrt{t}) : t > 0 \right\}$ so R_α is not in $C^{1,2}$ and Theorem 2.39 cannot be applied directly. The reader who believes that this issue is unlikely to be problematic may wish to take Lemma 2.47 on faith and skip ahead to Subsection 2.5.3.

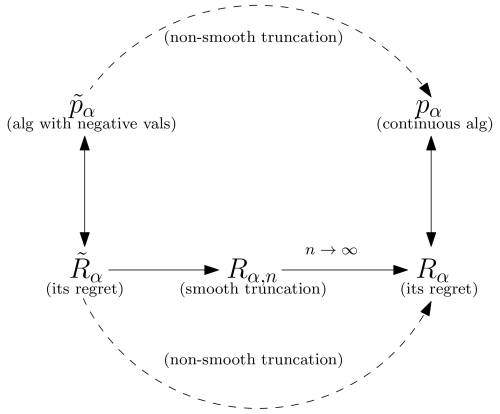


Figure 2.1: The relationships between $\tilde{p}_\alpha$, $\tilde{R}_\alpha$, $R_{\alpha,n}$, p_α, and R_α

Lemma 2.47. *Fix $\alpha > 0$. Then, almost surely, for all $T \ge 0$,* $\mathrm{ContRegret}(T, p_\alpha, B) \le R_\alpha(T, |B_T|)$.

Here, we will present a high-level overview of the proof of this lemma; the details can be found in Appendix A.2. Let $\phi(x)$ be a smooth function satisfying $\phi(x) = 1$ for $x \le 0$ and $\phi(x) = 0$ for $x \ge 1$. For $n \in \mathbb{N}$, define $\phi_n(x) = \phi(nx)$ and the approximations

$$R_{\alpha,n}(t,x) \coloneqq \tilde{R}_\alpha(t,x)\phi_n(x - \alpha\sqrt{t}) + \tilde{R}_\alpha(t,\alpha\sqrt{t})(1 - \phi_n(x - \alpha\sqrt{t})).$$

It is relatively straightforward to check that $R_{\alpha,n}(t,x) \xrightarrow{n\to\infty} R_\alpha(t,x)$ pointwise and similarly for the derivatives. The important property is that $R_{\alpha,n}$ is smooth so Itô's formula may be applied. Lemma 2.47 is then proved by taking limits and controlling the error terms.

The remainder of this section proves Theorem 2.37 by setting $p^* = p_\alpha$ for the optimal α.

2.5.3 Optimizing the boundary to minimize the continuous regret problem

By Lemma 2.47, $\mathrm{ContRegret}(T, \partial_x R_\alpha, B) \le R_\alpha(T, |B_T|) \le R_\alpha(T, \alpha\sqrt{T})$, where the last inequality is because $\partial_x R_\alpha(t,x) = p_\alpha(t,x)$ is positive for $x \in [0, \alpha\sqrt{t})$ and 0 for $x \ge \alpha\sqrt{t}$. Define

$$h(\alpha) \coloneqq R_\alpha(1,\alpha) = \frac{\alpha}{2} + \kappa_\alpha M_0(\alpha^2/2)$$

and note that $R_\alpha(T, \alpha\sqrt{T}) = \sqrt{T} \cdot h(\alpha)$. Thus, the only remaining task is now to solve the following optimization problem.

$$\min_{\alpha>0} h(\alpha) = \min_{\alpha>0} \left\{ \frac{\alpha}{2} + \kappa_\alpha \cdot M_0\left(\frac{\alpha^2}{2}\right) \right\} \tag{2.48}$$

The following lemma verifies that there exists some α for which $\mathrm{ContRegret}(T, \partial_x R_\alpha, B) \le \frac{\gamma\sqrt{T}}{2}$, completing the proof of Theorem 2.37

Lemma 2.48. *Fix $T > 0$. Then $\min_\alpha R_\alpha(T, \alpha\sqrt{T}) = R_\gamma(T, \gamma\sqrt{T}) = \frac{\gamma\sqrt{T}}{2}$.*

Lemma 2.48 follows easily from the following claim

Claim 2.49. *$h'(\alpha) = -\frac{\exp(\alpha^2/2)}{\pi \operatorname{erfi}(\alpha/\sqrt{2})} \cdot M_0(\alpha^2/2)$. In particular, $h'(\alpha) < 0$ for $\alpha \in (0, \gamma)$, $h'(\gamma) = 0$, and $h'(\alpha) > 0$ for $\alpha \in (\gamma, \infty)$.*

Proof. Recall that $h(\alpha) = \frac{\alpha}{2} + \frac{M_0(\alpha^2/2)}{\sqrt{2\pi}\operatorname{erfi}(\alpha/\sqrt{2})}$. Hence,

$$\begin{aligned} h'(\alpha) &= \frac{1}{2} - \frac{\alpha \cdot M_1(\alpha^2/2)}{\sqrt{2\pi}\operatorname{erfi}(\alpha/\sqrt{2})} - \frac{\exp(\alpha^2/2) \cdot M_0(\alpha^2/2)}{\pi \operatorname{erfi}(\alpha/\sqrt{2})} && \text{(by Fact 2.7)} \\ &= -\frac{\exp(\alpha^2/2) \cdot M_0(\alpha^2/2)}{\pi \operatorname{erfi}(\alpha/\sqrt{2})} && \text{(by Fact 2.8(2)).} \end{aligned}$$

This proves the first assertion.

Next, observe that $\frac{\exp(\alpha^2/2)}{\operatorname{erfi}(\alpha/\sqrt{2})}$ is positive for all $\alpha \in \mathbb{R}$. Hence, by Fact 2.11, we have that $h'(\alpha) < 0$ for $\alpha \in (0, \gamma)$, $h'(\gamma) = 0$, and $h'(\alpha) > 0$ for $\alpha \in (\gamma, \infty)$. □

Proof of Lemma 2.48. Claim 2.49 implies that γ is the global minimizer for $h(\alpha)$. Therefore, for every $\alpha > 0$, we have $R_\alpha(T, \alpha\sqrt{T}) = \sqrt{T} \cdot h(\alpha) \ge \sqrt{T} \cdot h(\gamma) = R_\gamma(T, \gamma\sqrt{T})$. This proves the first equality. The second equality is because $M_0(\gamma^2/2) = 0$ by definition of γ. □

2.5.4 Continuous regret against any continuous semi-martingale

Recall that the continuous regret upper bound (Theorem 2.37) involved the adversary evolving the gap process as a reflected Brownian motion, which is a continuous semi-martingale. In this section, we generalize the definition of continuous regret to allow arbitrary, non-negative, continuous semi-martingales to control the gap process, and derive an analogue of Theorem 2.37 in this generalized setting. We use the notation $[X]_t$ to denote the quadratic variation process of X (see Appendix A.2.2 for definitions).

We begin with a generalized definition of continuous regret.

Definition 2.50 (Continuous Regret)**.** Let $p : \mathbb{R}_{>0} \times \mathbb{R}_{\ge 0} \to [0, 1]$ be a continuous function that satisfies $p(t, 0) = 1/2$ for every $t > 0$. Let X_t be a continuous, non-negative, semi-martingale. Then, the *continuous regret* of p with respect to X is the stochastic integral

$$\mathrm{ContRegret}(T, p, X) \;=\; \int_0^T p(t, X_t) \mathrm{d}X_t. \tag{2.49}$$

The main result for this generalized setting is as follows.

Theorem 2.51. *There exists a continuous-time algorithm p^* such that for any continuous, non-negative, semi-martingale X,*

$$\text{ContRegret}(T, p^*, X) \;\leq\; \frac{\gamma}{2}\sqrt{[X]_T} \quad \forall T \in \mathbb{R}_{\geq 0}, \text{ almost surely.} \tag{2.50}$$

We provide an overview of the proof of this result below. For the sake of exposition, we sketch the proof of Theorem 2.51 in the setting where we allow p^* to take values in $(-\infty, 1]$. Truncating p^* as was done in Subsection 2.5.2 yields Theorem 2.51 as stated.

Sketch. Let $p^*(t,x) \coloneqq \tilde{p}_\gamma([X]_t, x)$ and $R(t,x) \coloneqq \tilde{R}_\gamma(t,x)$. (See Eq. (2.40) and Eq. (2.42) for definitions of $\tilde{p}_\gamma$ and $\tilde{R}_\gamma$). Recall the following three important properties of R from Lemma 2.46:

(1) R is $C^{1,2}$,
(2) R satisfies $\overset{*}{\Delta} R = 0$ over $\mathbb{R}_{>0} \times \mathbb{R}$,
(3) $\partial_x R(t,x) = \tilde{p}_\gamma(t,x)$.

Since R is $C^{1,2}$, we may apply Itô's formula (specifically Eq. (A.15) with $A_t = [X]_t$, which is a bounded variation process since it is increasing) to obtain

$$\begin{aligned}
R([X]_T, X_T) \;&=\; \int_0^T \partial_x R([X]_t, X_t)\mathrm{d}X_t + \int_0^T \partial_t R([X]_t, X_t) + \frac{1}{2}\partial_{xx} R([X]_t, X_t)\mathrm{d}[X]_t \\
&=\; \int_0^T p^*(t, X_t)\mathrm{d}X_t + \int_0^T \underbrace{\partial_t R([X]_t, X_t) + \frac{1}{2}\partial_{xx} R([X]_t, X_t)}_{=\overset{*}{\Delta} R([X]_t, X_t)}\,\mathrm{d}[X]_t \quad (\partial_x R = \tilde{p}_\gamma) \\
&=\; \int_0^T p^*(t, X_t)\mathrm{d}X_t \qquad (\overset{*}{\Delta} R = 0) \\
&=\; \text{ContRegret}(T, p^*, X).
\end{aligned}$$

Next, recall the upper bound on R from Eq. (2.45):

$$R(t,x) \;=\; R_\gamma(t,x) \;\leq\; \left(\frac{\gamma}{2} + \kappa_\gamma M_0\left(\frac{\gamma^2}{2}\right)\right)\sqrt{t} \;=\; \frac{\gamma}{2}\sqrt{t},$$

where the final equality is because γ is a root of $M_0\left(\frac{x^2}{2}\right)$. Putting everything together, we have

$$\text{ContRegret}(T, p^*, X) \;=\; R([X]_T, X_T) \;\leq\; \frac{\gamma}{2}\sqrt{[X]_T},$$

as desired. □

Chapter 3

Background on Mixtures of Gaussians

In this chapter, we provide some background on the problem of learning mixtures of Gaussians in the density estimation.

3.1 Notation and Definitions

We use $\mathbb{R}$ to denote the set of real numbers and $\mathbb{R}_{\geq 0}$ to denote the set of non-negative real numbers. For an integer n, we may also write $[n] = \{1, \ldots, n\}$.

Definition 3.1 (Gaussian distribution)**.** A d-dimensional Gaussian is a distribution which is specified by a mean $\mu \in \mathbb{R}^d$ and a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Here, Σ must be a positive semidefinite matrix.[1] The probability density function (p.d.f.) is given by

$$\mathcal{N}(\mu, \Sigma)(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma (x-\mu)\right) \tag{3.1}$$

We will often denote a Gaussian with mean μ and covariance matrix Σ by $\mathcal{N}(\mu, \Sigma)$.

Let $\Delta_k = \{w \in \mathbb{R}^k : w_i \geq 0, \sum_{i \in [k]} w_i = 1\}$ denote the k-dimensional probability simplex.

Definition 3.2 (Mixture distributions)**.** Let $\mathcal{F}$ be a class of distributions. The class of k-mixtures of $\mathcal{F}$, denoted by k-mix$(\mathcal{F})$ is defined as

$$k\text{-mix}(\mathcal{F}) \coloneqq \left\{ \sum_{i=1}^{k} w_i f_i : (w_1, \ldots, w_k) \in \Delta_k, f_i \in \mathcal{F} \right\}. \tag{3.2}$$

In this thesis, the most important special case of this definition is when $\mathcal{F}$ is the class of Gaussian distributions. In this case, a mixture of k Gaussians is a distribution specified by k mean vectors $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$, k covariance matrices $\Sigma_1, \ldots, \Sigma_k \in \mathbb{R}^{d \times d}$, and a probability vector $w \in \Delta_k$. The p.d.f. is then given by $\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$. One way to sample from this distribution is to first sample a coordinate i according to the vector w then to sample a point from $\mathcal{N}(\mu_i, \Sigma_i)$.

[1] We say a matrix $\Sigma \in \mathbb{R}^{d \times d}$ is positive semidefinite if Σ is symmetric and all its eigenvalues are non-negative. An alternative definition is that Σ is symmetric and $v^\top \Sigma v \geq 0$ for all $v \in \mathbb{R}^d$.

3.2 Probability Background

In Chapter 4, we will deal exclusively with continuous distributions so all results stated in this chapter will assume continuous distributions. Nonetheless, all the preliminary probability background presented here hold for discrete distributions as well.

In this section, let f, g denote probability density functions (p.d.f.) on $\mathbb{R}^d$. We will abuse notation and often refer to a distribution via its p.d.f. In this section, $\mathcal{B}$ denotes the Borel σ-algebra on $\mathbb{R}^d$. For an event $A \in \mathcal{B}$ and a distribution f, we write $\mathbb{P}_f[A] = \int_A f(x)\,\mathrm{d}x$ as the probability that the event A happens under f.

Definition 3.3 (Total Variation Distance)**.** The *total variation distance* between f, g, which we denote by $d_{\mathrm{TV}}(f, g)$ is defined as

$$d_{\mathrm{TV}}(f,g) = \sup_{A\in\mathcal{B}} \int_A f(x) - g(x)\,\mathrm{d}x = \sup_{A\in\mathcal{B}} \{\mathbb{P}_f[A] - \mathbb{P}_g[A]\}$$

If $X \sim f$ and $Y \sim g$, we may abuse notation and write $d_{\mathrm{TV}}(X, Y) = d_{\mathrm{TV}}(f, g)$.

Fact 3.4. *Let $A^+ = \{x \in \Omega : f(x) > g(x)\}$ and $A^- = \{x : f(x) \le g(x)\}$. Then*

$$d_{\mathrm{TV}}(f,g) = \mathbb{P}_f[A^+] - \mathbb{P}_g[A^+] = \mathbb{P}_g[A^-] - \mathbb{P}_f[A^+] = \frac{1}{2}\left\|f - g\right\|_1.$$

Here, $\|f\|_1 = \int_{\mathbb{R}^d} f\,\mathrm{d}x$.

Remark 3.5. *Fact 3.4 implies that d_{TV} is a metric.*

Proof. We first show that $\mathbb{P}_f[A^+] - \mathbb{P}_g[A^+] = \mathbb{P}_g[A^-] - \mathbb{P}_f[A^+]$. To see this, observe that $A^+ \cup A^- = \mathbb{R}^d$ and that the union is disjoint. Hence, $\mathbb{P}_f[A^+] + \mathbb{P}_f[A^-] = 1 = \mathbb{P}_g[A^+] + \mathbb{P}_g[A^-]$. Rearranging gives the claim.

Next we show that $\mathbb{P}_f[A^+] - \mathbb{P}_g[A^+] = \frac{1}{2}\left\|f - g\right\|_1$. Indeed, by definition of A^+, A^-,

$$\begin{aligned}\left\|f - g\right\|_1 &= \int_{A^+} f(x) - g(x)\,\mathrm{d}x + \int_{A^-} g(x) - f(x)\,\mathrm{d}x \\ &= (\mathbb{P}_f[A^+] - \mathbb{P}_g[A^+]) + (\mathbb{P}_g[A^-] - \mathbb{P}_f[A^-]) \\ &2(\mathbb{P}_f[A^+] - \mathbb{P}_g[A^+]),\end{aligned}$$

so $\mathbb{P}_f[A^+] - \mathbb{P}_g[A^+] = \frac{1}{2}\left\|f - g\right\|_1$.

It remains to show that $d_{\mathrm{TV}}(f, g) = \mathbb{P}_f[A^+] - \mathbb{P}_g[A^+]$. Clearly, $d_{\mathrm{TV}}(f, g) \ge \mathbb{P}_f[A^+] - \mathbb{P}_g[A^+]$ by definition of total variation distance. We now prove the reverse inequality. Let $A \in \mathcal{B}$ be an arbitrary event. Then

$$\begin{aligned}\mathbb{P}_f[A] - \mathbb{P}_g[A] &= (\mathbb{P}_f[A \cap A^+] - \mathbb{P}_g[A \cap A^+]) + (\mathbb{P}_f[A \cap A^-] - \mathbb{P}_g[A \cap A^-]) \\ &\le \mathbb{P}_f[A \cap A^+] - \mathbb{P}_g[A \cap A^+] \\ &\le \mathbb{P}_f[A^+] - \mathbb{P}_g[A^+]\end{aligned}$$

Taking the supremum over all events A gives that $d_{\mathrm{TV}}(f, g) \le \mathbb{P}_f[A^+] - \mathbb{P}_g[A^+]$ as desired. □

Fact 3.6. *Let X and Y be arbitrary random variables. For any function F, we have*

$$d_{\mathrm{TV}}\left(F(X), F(Y)\right) \;\leq\; d_{\mathrm{TV}}\left(X, Y\right).$$

Proof. This follows from the observation that

$$\mathbb{P}\left[F(X) \in A\right] - \mathbb{P}\left[F(Y) \in A\right] = \mathbb{P}\left[X \in F^{-1}(A)\right] - \mathbb{P}\left[Y \in F^{-1}(A)\right] \leq d_{\mathrm{TV}}\left(X, Y\right),$$

so taking supremum of the left-hand side gives the result. □

Definition 3.7 (KL Divergence)**.** Let f, g be distributions. The *Kullback-Leibler diversion* (or KL divergence) between f and g, denoted $D_{\mathrm{KL}}(f \parallel g)$, is defined as

$$D_{\mathrm{KL}}(f \parallel g) = \int_{\mathbb{R}^d} \ln\left(\frac{f(x)}{g(x)}\right) \mathrm{d}x.$$

It is important to note that, in general, the KL divergence is *not* symmetric, i.e. it is not necessarily true that $D_{\mathrm{KL}}(f \parallel g) = D_{\mathrm{KL}}(g \parallel f)$. In particular, the KL divergence is not a metric.

Fact 3.8 ([135, p. 85])**.** $D_{\mathrm{KL}}(f \parallel g) \geq 0$.

There is an important relationship between the total variation distance and the KL divergence; this is given by the Pinsker's Inequality.

Lemma 3.9 (Pinsker's Inequality [135, Lemma 2.5])**.** $2d_{\mathrm{TV}}(f, g)^2 \leq D_{\mathrm{KL}}(f \parallel g)$.

3.3 Density Estimation

As usual, we let $\mathcal{F}$ denote a class of distributions. At a high-level, a *density estimation algorithm* is an algorithm that takes in i.i.d. samples from some unknown $f \in \mathcal{F}$ and outputs a distribution $g \in \mathcal{F}$ such that $d_{\mathrm{TV}}(f, g)$ is small. Of course, this may not always be possible. For example, if f is just a single Gaussian then, with very small probability, all the samples could come from its tails. For this reason, we will be happy if the algorithm returns a distribution $g \in \mathcal{F}$ for which $d_{\mathrm{TV}}(f, g)$ is small *most of the time*.

Let us now define the problem more formally. We begin by defining the notion of probably approximately correct (PAC) learning of distributions.

Definition 3.10 (PAC learning of distributions, realizable case)**.** Let $\mathcal{F}$ be a class of densities. We say that $\mathcal{F}$ is *PAC-learnable* with sample complexity $m\colon (0,1) \times (0,1) \to \mathbb{N}$ if there exists an algorithm $\mathcal{A}$ such that for all $f \in \mathcal{F}$ the following holds: given $\varepsilon, \delta > 0$ and $m(\varepsilon, \delta)$ i.i.d. samples from f as input to $\mathcal{A}$, with probability $1 - \delta$ (over the samples), $\mathcal{A}$ outputs a density $\hat{g}$ such that $d_{\mathrm{TV}}(f, \hat{g}) \leq \varepsilon$.

As an example of Definition 3.10, let $\mathcal{F}$ denote the class of Gaussian distributions, i.e. distributions of the form $\mathcal{N}(\mu, \Sigma)$ where μ, Σ are unknown. An intuitive algorithm for learning this distribution is to obtain a collection of m i.i.d. samples from $\mathcal{N}(\mu, \Sigma)$, compute its empirical mean, $\hat{\mu}$, and its empirical covariance matrix, $\hat{\Sigma}$. We then output the distribution $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$. We will be able to show that if

$m = O\left(\frac{d^2}{\varepsilon^2} \cdot \log(1/\delta)\right)$ then, with probability $1-\delta$ over the samples, $d_{\mathrm{TV}}(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\hat{\mu},\hat{\Sigma})) \leq \varepsilon$. As the argument does require some work, the details are relegated to Subsection 3.3.3. This shows that Gaussian distributions are PAC-learnable with sample complexity $O\left(\frac{d^2}{\varepsilon^2} \cdot \log(1/\delta)\right)$.

Definition 3.10 assumes that the true distribution is itself a member of $\mathcal{F}$. We can further generalize the definition to assume that the true distribution is *not* a member of $\mathcal{F}$.

Definition 3.11 (PAC learning of distributions, agnostic case)**.** Let $\mathcal{F}$ be a class of densities. We say that $\mathcal{F}$ is C-agnostic *PAC-learnable* with sample complexity $m\colon (0,1)\times(0,1) \to \mathbb{N}$ if there exists an algorithm $\mathcal{A}$ such that for all distributions f (not necessarily in $\mathcal{F}$) the following holds: given $\varepsilon, \delta > 0$ and $m(\varepsilon,\delta)$ i.i.d. samples from f as input to $\mathcal{A}$, with probability $1-\delta$ (over the samples), $\mathcal{A}$ outputs a density $\hat{g}$ such that

$$d_{\mathrm{TV}}(f,\hat{g}) \leq C \cdot \inf_{g\in\mathcal{F}} d_{\mathrm{TV}}(f,g) + \varepsilon.$$

We can understand this definition as follows. Suppose that f is the true distribution and that it is close (but not exactly equal to) a member in $\mathcal{F}$. Then an algorithm is an agnostic PAC-learner if it outputs a distribution in $\mathcal{F}$ which approximates f only slightly worse than the best member in $\mathcal{F}$ would approximate f.

Definition 3.11 is also a bit more useful than Definition 3.10 in practice. For example, it is common to assume that $\mathcal{F}$ is a mixture of high-dimensional Gaussians with a moderate number of components. On the other hand, it is unlikely that the real data is a exactly a mixture of Gaussians, in which wase Definition 3.10 would not be applicable. However, it is likely that the real data is very well-approximated by a mixture of Gaussians with not too many components in which case Definition 3.11 would be useful.

3.3.1 Learning Finite Hypothesis Classes

Suppose that $\mathcal{F}$ is a *finite* hypothesis class. Consider the following problem. We are given i.i.d. access to an unknown distribution f and our goal is to output a distribution $g \in \mathcal{F}$ such that $d_{\mathrm{TV}}(f,g)$ is small.

Theorem 3.12 ([53, Theorem 6.3])**.** *Let $\mathcal{F}$ be a finite class of distributions and $M = \log|\mathcal{F}|$. There is an algorithm $\mathcal{A}$ such that for any distribution f, if $\mathcal{A}$ is given $O(\log(M/\delta)/\varepsilon^2)$ i.i.d. samples from f then with probability at least $1-\delta$ (over the samples), $\mathcal{A}$ outputs a distribution $\hat{g} \in \mathcal{F}$ satisfying*

$$d_{\mathrm{TV}}(f,\hat{g}) \leq 3 \cdot \min_{g\in\mathcal{F}} d_{\mathrm{TV}}(f,g) + \varepsilon.$$

Remark 3.13. *Bousquet et al. [22] showed that the constant 3 is tight if $\hat{g} \in \mathcal{F}$. However, if one removes the restriction that $\hat{g} \in \mathcal{F}$ then Bousquet et al. [22] showed that the constant 3 can be improved to 2. For example, it is sufficient to allow $\hat{g}$ to be a mixture of densities in $\mathcal{F}$.*

The idea is as follows. Recall from Fact 3.4 that if we define the event $E_{g,g'} = \{g > g'\}$ then $d_{\mathrm{TV}}(g,g') = \mathbb{P}_g[E_{g,g'}] - \mathbb{P}_{g'}[E_{g,g'}]$. Let $\mathcal{E} = \{E_{g,g'} : g,g' \in \mathcal{F}, g \neq g'\}$ be a set of "candidate events". For each g, we define $\Delta_g = \max_{E\in\mathcal{E}} |\mathbb{P}_g[E] - \mathbb{P}_f[E]|$. We can understand Δ_g as a proxy for $d_{\mathrm{TV}}(g,f)$.

The algorithm then returns $\arg\min_{g\in\mathcal{F}} \Delta_g$, i.e. the distribution g which has the smallest estimate for the total variation distance. Of course, one cannot compute Δ_g because one cannot compute $\mathbb{P}_f[E]$. However, one can obtain a very good estimate of $\mathbb{P}_f[E]$ by sampling from f. The formal algorithm is given in Algorithm 3.

Algorithm 3 An algorithm for learning with respect to a finite class of distributions

Input: Sample access to unknown distribution f, finite family $\mathcal{F}$ of distributions, and number of samples n.

Output: A distribution $\hat{g} \in \mathcal{F}$.

1: Let $\mathcal{E} = \{E_{g,g'} : g \in \mathcal{F}\}$ where $E_{g,g'} = \{g > g'\}$.
2: Draw n samples from f, call these $X_1, \ldots, X_n$.
3: For each $E \in \mathcal{E}$, compute $\overline{p}_E = \frac{1}{n}\sum_{i\in[n]} \mathbb{I}[X_i \in E]$.
4: For each $g \in \mathcal{F}$, compute its discrepancy $\Delta_g = \sup_{E\in\mathcal{E}} |\mathbb{P}_g[E] - \overline{p}_E|$.
5: Return $\hat{g} \in \arg\min_{g\in\mathcal{F}} \Delta_g$.

Proof of Theorem 3.12. Fix an event $E \in \mathcal{E}$. Then $X_1, \ldots, X_n$ are i.i.d. random variables and $\mathbb{P}_f[E] = \mathbb{E}[\mathbb{I}[X_i \in E]]$. Hence, by Hoeffding's Inequality, we have

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[X_i \in E] - \mathbb{P}_f[E]\right| \geq \varepsilon/2\right] \leq 2\exp\left(-n\varepsilon^2/2\right)$$

Choosing $n \geq 2\log\left(2M^2/\delta\right)/\varepsilon^2$, we have that for the fixed event E,

$$\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[X_i \in E] - \mathbb{P}_f[E]\right| \leq \varepsilon/2 \tag{3.3}$$

with probability at least $1 - \delta/M^2$. Since $|\mathcal{E}| \leq M^2$, we can take a union bound to get that Eq. (3.3) holds for all $E \in \mathcal{E}$ with probabiltiy at least $1 - \delta$. For the remainder of the proof, we condition on this event and we write $\overline{p}_E = \frac{1}{n}\sum_{i\in[n]} \mathbb{I}[X_i \in E]$.

Let $\hat{g}$ be the output of the algorithm and let $g^* \in \arg\min_{g\in\mathcal{E}} d_{\mathrm{TV}}(f, g)$. By the triangle inequality, we have $d_{\mathrm{TV}}(f, \hat{g}) \leq d_{\mathrm{TV}}(\hat{g}, g^*) + d_{\mathrm{TV}}(f, g^*)$. We now bound $d_{\mathrm{TV}}(\hat{g}, g^*)$. Writing $E^* = E_{\hat{g},g^*}$, we have

$$\begin{aligned}
d_{\mathrm{TV}}(\hat{g}, g^*) &= |\mathbb{P}_{\hat{g}}[E^*] - \mathbb{P}_{g^*}[E^*]| \\
&\leq |\mathbb{P}_{\hat{g}}[E^*] - \mathbb{P}_f[E^*]| + |\mathbb{P}_f[E^*] - \mathbb{P}_{g^*}[E^*]| && \text{(triangle inequality)} \\
&\leq |\mathbb{P}_{\hat{g}}[E^*] - \mathbb{P}_f[E^*]| + d_{\mathrm{TV}}(f, g^*) && \text{(definition of } d_{\mathrm{TV}}(f, g^*)) \\
&\leq |\mathbb{P}_{\hat{g}}[E^*] - \overline{p}_{E^*}| + |\overline{p}_{E^*} - \mathbb{P}_f[E^*]| + d_{\mathrm{TV}}(f, g^*) && \text{(triangle inequality)} \\
&\leq |\mathbb{P}_{\hat{g}}[E^*] - \overline{p}_{E^*}| + \varepsilon/2 + d_{\mathrm{TV}}(f, g^*) && \text{(Eq. (3.3) holds)} \\
&\leq \Delta_{\hat{g}} + \varepsilon/2 + d_{\mathrm{TV}}(f, g^*) && \text{(definition of } \Delta_{\hat{g}}) \\
&\leq \Delta_{g^*} + \varepsilon/2 + d_{\mathrm{TV}}(f, g^*) && \text{(Line 5)} \\
&\leq \max_{E\in\mathcal{E}} |\mathbb{P}_{g^*}[E] - \overline{p}_E| + \varepsilon/2 + d_{\mathrm{TV}}(f, g^*) && \text{(definition of } \Delta_{g^*}) \\
&\leq \max_{E\in\mathcal{E}} |\mathbb{P}_{g^*}[E] - \mathbb{P}_f[E]| + |\mathbb{P}_f[E] - \overline{p}_E| + \varepsilon/2 + d_{\mathrm{TV}}(f, g^*) && \text{(triangle inequality)}
\end{aligned}$$

$$\leq \max_{E \in \mathcal{E}} |\mathbb{P}_{g^*}[E] - \mathbb{P}_f[E]| + \varepsilon + d_{\mathrm{TV}}(f, g^*) \qquad \text{(Eq. (3.3) holds)}$$

$$\leq d_{\mathrm{TV}}(f, g^*) + \varepsilon + d_{\mathrm{TV}}(f, g^*) \qquad \text{(definition of } d_{\mathrm{TV}}(f, g^*)\text{)}.$$

We conclude that $d_{\mathrm{TV}}(f, \hat{g}) \leq 3 d_{\mathrm{TV}}(f, g^*) + \varepsilon$. □

3.3.2 Covering Arguments

Now suppose that $\mathcal{F}$ is an *infinite* hypothesis class. It turns out that the results in the last section can be used to obtain an algorithm for this setting as long as $\mathcal{F}$ is somewhat structured.

Definition 3.14 (ε-cover)**.** A hypothesis class $\hat{\mathcal{F}}$ is an ε-cover of $\mathcal{F}$ if for every $f \in \mathcal{F}$, there exists $\hat{f} \in \hat{\mathcal{F}}$ such that $d_{\mathrm{TV}}(f, \hat{f}) \leq \varepsilon$.

Note that in the definition of ε-cover, we did not require that $\hat{\mathcal{F}} \subseteq \mathcal{F}$.

Corollary 3.15. *Let $\mathcal{F}$ be a hypothesis class (possibly infinite) and f be an arbitrary distribution. Let $\hat{\mathcal{F}}$ be an ε_1-cover for $\mathcal{F}$ and set $M = \log |\hat{\mathcal{F}}|$. There is an algorithm, which given $\hat{\mathcal{F}}$, and $n \geq \log\left(2M^2/\delta\right)/\varepsilon_2^2$ i.i.d. samples from f, with probability at least $1 - \delta$ (over the samples) outputs $\hat{g} \in \hat{\mathcal{F}}$ satisfying*

$$d_{\mathrm{TV}}(f, \hat{g}) \leq 3 \cdot \inf_{g \in \mathcal{F}} d_{\mathrm{TV}}(f, g) + 3\varepsilon_1 + \varepsilon_2$$

Proof. We use Algorithm 3 with $\hat{\mathcal{F}}$ as the input hypothesis set. In this case, Theorem 3.12 guarantees that with probability $1 - \delta$, the output will be a distribution $\hat{g} \in \hat{\mathcal{F}}$ satisfying

$$d_{\mathrm{TV}}(f, \hat{g}) \leq 3 \cdot \min_{g \in \hat{\mathcal{F}}} d_{\mathrm{TV}}(f, g) + \varepsilon_2.$$

Now let $g \in \mathcal{F}$ be arbitrary and let $g' \in \hat{\mathcal{F}}$ be such that $d_{\mathrm{TV}}(g, g') \leq \varepsilon_2$. In this case, we have

$$d_{\mathrm{TV}}(f, \hat{g}) \leq 3 d_{\mathrm{TV}}(f, g') + \varepsilon \leq 3 d_{\mathrm{TV}}(f, g) + 3\varepsilon_1 + \varepsilon_2.$$

Taking the infimum over all $g \in \mathcal{F}$ on the RHS gives the claim. □

Covering arguments are not always useful. In particular, if $\hat{\mathcal{F}}$ is an infinite-sized ε-cover for $\mathcal{F}$ then Corollary 3.15 only gives a vacuous statement. For example, if $\mathcal{F}$ is the class of Gaussian distributions on $\mathbb{R}$ with unit variance, it is impossible to find a finite ε-cover for any value of $\varepsilon \in (0, 1)$. The issue lies in the fact that an ε-cover is constructed *before* seeing any samples from f. In Chapter 4, we will introduce the compression framework which will allow us to construct a "data-dependent ε-cover".

3.3.3 An algorithm for learning a single Gaussian

Theorem 3.16. *Let $\mathcal{F}$ be the class of Gaussian distributions in $\mathbb{R}^d$. Then $\mathcal{F}$ is PAC-learnable with sample complexity $O\left(\frac{d^2}{\varepsilon^2} \cdot \log(1/\delta)\right)$.*

To prove this theorem, we need to show that there is an algorithm $\mathcal{A}$ with the following property: if it takes as input an accuracy parameter $\varepsilon > 0$ and a confidence parameter $\delta > 0$, then it requires

only $O\left(\frac{d^2}{\varepsilon^2}\cdot\log(1/\delta)\right)$ samples from an unknown Gaussian distribution to produce an ε-approximation of it (in total variation distance).

The algorithm $\mathcal{A}$ is as simple as it can get. Suppose the unknown Gaussian is $\mathcal{N}(\mu,\Sigma)$. First, the algorithm draws $2m$ samples from $\mathcal{N}(\mu,\Sigma)$ where $m=O\left(\frac{d^2}{\varepsilon^2}\cdot\log(1/\delta)\right)$. Let us call these samples $X_1,\dots,X_{2m}$. The algorithm then computes the empirical mean $\hat{\mu}=\frac{1}{m}\sum_{i=1}^m X_i$ and the empirical covariance matrix $\hat{\Sigma}=\frac{1}{2m}\sum_{i=1}^m(X_{2i}-X_{2i-1})(X_{2i}-X_{2i-1})^\top$. The output is the distribution $\mathcal{N}(\hat{\mu},\hat{\Sigma})$.

We now analyze this algorithm. From Lemma B.5 in Appendix B.1, we have that

$$2d_{\mathrm{TV}}(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\hat{\mu},\hat{\Sigma}))^2\le\frac{1}{2}\Big(\mathrm{LD}\left(\hat{\Sigma},\Sigma\right)+(\hat{\mu}-\mu)^\top\Sigma^{-1}(\hat{\mu}-\mu)\Big),$$

where $\mathrm{LD}\,(A,B)\coloneqq\mathrm{Tr}(B^{-1}A-I)-\log\det(B^{-1}A)$ is the log-det divergence between PSD matrices A,B (see Definition B.1 in Appendix B.1). Hence, it suffices to show that, with the above algorithm, we have

$$\mathrm{LD}\left(\hat{\Sigma},\Sigma\right)\le O(\varepsilon^2)\quad\text{and}\quad(\hat{\mu}-\mu)^\top\Sigma^{-1}(\hat{\mu}-\mu)\le O(\varepsilon^2)$$

with probability at least $1-\delta$. This will follow from the following two claims.

Claim 3.17. *Suppose $m\ge\frac{d+2\sqrt{d\log(1/\delta)}+2\log(1/\delta)}{\varepsilon^2}$ and let $\hat{\mu}=\frac{1}{m}\sum_{i=1}^m X_i$ where X_i are drawn independently from $\mathcal{N}(\mu,\Sigma)$. Then*

$$\mathbb{P}\left[(\hat{\mu}-\mu)^\top\Sigma^{-1}(\hat{\mu}-\mu)\ge\varepsilon^2\right]\le\delta$$

Proof. Note that $\Sigma^{-1/2}\cdot(X_i-\mu)\sim N(0,I)$ and so $\frac{1}{m}\sum_{i=1}^m\Sigma^{-1/2}(X_i-\mu)\sim N(0,\frac{1}{m}I)$. Let $g=\frac{1}{\sqrt{m}}\sum_{i=1}^m\Sigma^{-1/2}(X_i-\mu)\sim N(0,1)$ so that $(\hat{\mu}-\mu)^\top\Sigma^{-1}(\hat{\mu}-\mu)=\frac{1}{m}g^\top g$. The claim is thus equivalent to proving that

$$\mathbb{P}\left[\|g\|_2^2\ge m\varepsilon^2\right]\le\delta. \tag{3.4}$$

Note that $\|g\|_2^2$ has a chi-squared distribution with parameter d since $\|g\|_2^2=\sum_{j=1}^d g_j^2$ where $g_j\sim N(0,1)$. By Lemma B.13, we have, for any $t>0$,

$$\mathbb{P}\left[\|g\|_2^2\ge d+2\sqrt{dt}+2t\right]\le\exp(-t) \tag{3.5}$$

Choosing $t=\log(1/\delta)$ gives that Eq. (3.5) is at most δ. By our choice of m, we have that $m\varepsilon^2\ge d+2\sqrt{d\log(1/\delta)}+2\log(1/\delta)$ so Eq. (3.4) holds. □

Claim 3.18. *If $m\ge C(d^2+d\log(1/\delta))/\varepsilon^2$ for some sufficiently large constant $C>0$ then with probability at least $1-\delta$,*

$$(1-\varepsilon/\sqrt{d})\Sigma\preceq\hat{\Sigma}\preceq(1+\varepsilon/\sqrt{d})\Sigma.$$

Proof. Set $\alpha=\varepsilon/\sqrt{d}$. First observe that $\frac{X_{2i}-X_{2i-1}}{\sqrt{2}}\sim\mathcal{N}(0,\Sigma)$. Let $g_i=\Sigma^{-1/2}\frac{X_{2i}-X_{2i-1}}{\sqrt{2}}\sim\mathcal{N}(0,I_d)$. Thus, we have $\hat{\Sigma}=\frac{1}{m}\Sigma^{1/2}\left(\sum_{i=1}^m g_ig_i^\top\right)\Sigma^{1/2}$. By Fact 3.19, it suffices to show that

$$(1-\alpha)I_d\preceq\frac{1}{m}\sum_{i=1}^m g_ig_i^\top\preceq(1+\alpha)I_d. \tag{3.6}$$

Applying Lemma B.14 with $t = \sqrt{1 + \log(1/\delta)/d}$ gives that, as long as $m \geq C(d + \log(1/\delta))/\alpha^2 = C(d^2 + d\log(1/\delta))/\varepsilon^2$ then Eq. (3.6) holds with probability at least $1 - \delta$. □

Fact 3.19. *Let A, B be $n \times n$ PSD matrices and suppose that $A \preceq B$. Let X be any $k \times n$ matrix. Then $XAX^\top \preceq XBX^\top$.*

Chapter 4

Near-optimal sample complexity bounds for learning mixtures of Gaussians

Chapter Summary. In this chapter, we give an algorithm for learning mixtures of k Gaussians in $\mathbb{R}^d$ with sample complexity $\tilde{O}(kd^2/\varepsilon^2)$.[1] We will also show a matching lower bound and show that any algorithm for learning mixtures of k Gaussians in $\mathbb{R}^d$ requires $\tilde{\Omega}(kd^2/\varepsilon^2)$ samples.

4.1 Introduction

Estimating distributions from observed data is a fundamental task in statistics that has been studied for over a century. This task frequently arises in applied machine learning where it is commonly assumed that the distribution can be modeled approximately with a mixture of Gaussians. There are many popular software packages which have implemented heuristics for learning mixtures of Gaussians; the most common heuristic is the expectation maximization (EM) algorithm. The theoretical machine learning community also has a lengthy history on distribution learning; we refer the reader to [56] for a survey on learning structured distributions.

The purpose of this chapter is to develop a general and generic technique for distribution learning. We will then apply this technique to the fundamental setting of learning mixtures of Gaussians. The learning model we adapt is *density estimation*, which is described in detail in Section 3.3. To summarize, in this model, we are given i.i.d. samples from the unknown target distribution and our goal is to find a distribution that is close to the target in *total variation (TV) distance*. This chapter will focus on *sample complexity*, i.e. the number of samples for which it is sufficient for some algorithm to obtain a close estimate.

The technique for proving upper bounds on the sample complexity utilizes a novel notion of *sample compression*. More specifically, we show that if it is possible to "encode" members of a class of

[1] We write $\tilde{O}(f(n))$ to mean $O(f(n)\operatorname{polylog}(f(n)))$. So $\tilde{O}(kd^2/\varepsilon^2)$ means $O(kd^2/\varepsilon^2 \cdot \operatorname{polylog}(kd/\varepsilon))$. Similarly, $\tilde{\Omega}(f(n))$ means $\Omega(f(n)/\operatorname{polylog}(f(n)))$. So $\tilde{\Omega}(kd^2/\varepsilon^2)$ means $\Omega(kd^2/\varepsilon^2 \operatorname{polylog}(kd/\varepsilon))$.

distributions with a carefully chosen subset of samples *drawn* from the distribution then this yields an upper bound on the sample complexity for learning with respect to that class. In fact, given an efficient sample compression scheme, we will show how to transform it into a sample-efficient learning algorithm. Hence, by constructing sample compression schemes for mixtures of Gaussians, we will obtain new upper bounds on the sample complexity of learning with respect to these classes.

4.1.1 Main results for mixtures of Gaussians

In this section, we go over our main results for learning mixtures of multivariate Gaussians. Let k denote the number of mixture components and d denote the dimension. Henceforth, the notations $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ suppress $\mathrm{polylog}(kd/\varepsilon\delta)$ factors. We stress that the results that we present here do not have any dependencies on any other parameters (such as the condition number or the minimum mixing weight).

Theorem 4.1. *The class of k-mixtures of d-dimensional Gaussians can be learned in the realizable setting, and can be 12-learned in the agnostic setting, using $\widetilde{O}(kd^2/\varepsilon^2)$ samples.*

Prior to this result, the best known upper bounds on the sample complexity of this problem were $\widetilde{O}(kd^2/\varepsilon^4)$, due to [15], and $O(k^4d^4/\varepsilon^2)$, based on a VC-dimension bound. For the case of a single Gaussian (i.e., $k = 1$), a sample complexity bound of $O(d^2/\varepsilon^2)$ is known as we described in Section 3.3.3.

Our second main result is a lower bound matching Theorem 4.1 up to logarithmic factors.

Theorem 4.2. *Any method for learning the class of k-mixtures of d-dimensional Gaussians in the realizable setting has sample complexity $\widetilde{\Omega}(kd^2/\varepsilon^2)$.*

Note that this is a worst-case (i.e., minimax) lower bound: for any learning algorithm, there always exists at least one distribution that requires that many samples. Previously, the best known lower bound on the sample complexity was $\widetilde{\Omega}(kd/\varepsilon^2)$ [133]. Even for a single Gaussian (i.e., $k = 1$), an $\widetilde{\Omega}(d^2/\varepsilon^2)$ lower bound was not known prior to this work.

Our third main result is an upper bound for learning mixtures of *axis-aligned* Gaussians, i.e., Gaussians with diagonal covariance matrices. This bound is also near-optimal as it matches the $\Omega(kd/\varepsilon^2)$ bound found in [133].

Theorem 4.3. *The class of k-mixtures of axis-aligned d-dimensional Gaussians can be learned in the realizable setting, and can be 12-learned in the agnostic setting, using $\widetilde{O}(kd/\varepsilon^2)$ samples.*

Theorem 4.3 also appears in Ashtiani's thesis [16, §7.6]. His proof is also via compression schemes; however the compression we present here differs significantly from the one found in [16].

Our techniques. The upper bounds are proved using a novel compression framework. In particular, we show that distributions which can be "compressed" by representing it using a small number of its own samples then it can also be *learned* with a small number of samples. In fact, given any compression scheme, there is a black-box reduction which transforms the compression scheme into a sample-efficient algorithm. We then show that mixtures of Gaussians have an efficient compression scheme.

Next we discuss the main ideas used in the proof of our lower bound, Theorem 4.2. In order to prove our lower bound for mixtures of Gaussians, we first prove a lower bound of $\widetilde{\Omega}(d^2/\varepsilon^2)$ for learning a single Gaussian. Although the approach is quite intuitive, the details are intricate and much care is required to make a formal proof. The main step is to construct a large family (of size $2^{\Omega(d^2)}$) of covariance matrices such that the associated Gaussian distributions are well-separated in terms of their total variation distance, while simultaneously ensuring that their Kullback-Leibler divergences are small. Once this is established, we can then apply a generalized version of Fano's inequality to complete the proof.

To construct this family of covariance matrices, we sample $2^{\Omega(d^2)}$ matrices from the following probabilistic process: start with an identity covariance matrix; then choose a random subspace of dimension $d/9$ and slightly increase the eigenvalues corresponding to this eigenspace. It is easy to bound the KL divergences between the constructed Gaussians. To lower bound the TV distance, we show that for every pair of these distributions, there is some subspace for which a vector drawn from one Gaussian will have slightly larger projection than a vector drawn from the other Gaussian. Quantifying this gap will then give us the desired lower bound on the total variation distance.

Computational efficiency. Although our approach for proving sample complexity upper bounds is algorithmic, our focus is not on computational efficiency. The resulting algorithms have nearly optimal sample complexities, but their running times are exponential in the dimension d and the number of mixture components k. More precisely, the running time is $2^{kd^2\,\mathrm{polylog}(d,k,1/\varepsilon,1/\delta)}$ for mixtures of general Gaussians, and $2^{kd\,\mathrm{polylog}(d,k,1/\varepsilon,1/\delta)}$ for mixtures of axis-aligned Gaussians. The existence of an algorithm for density estimation that runs in time $\mathrm{poly}(k,d)$ is unknown even for the class of mixtures of axis-aligned Gaussians, see [59, Question 1.1].

Chapter outline. Next, we review some related work. In Section 4.2, we provide justification for our learning model. In Section 4.3, we formally define compression schemes for distributions, prove their closure properties, and show their connection with density estimation. Theorems 4.1 and 4.3 are proved in Section 4.4. Theorem 4.2 is proven in Section 4.5. A collection of standard facts can be found in the appendices.

4.1.2 Related work

Distribution learning is a vast topic and many approaches have been considered in the literature. This section reviews the approaches that are particularly relevant to our work.

For parametric families of distributions, a common approach is to use the samples to estimate the parameters of the distribution, possibly in a maximum likelihood sense, or possibly aiming to approximate the true parameters. For the specific case of mixtures of Gaussians, there is a substantial theoretical literature on algorithms that approximate the mixing weights, means and covariances (e.g., [11, 20, 42, 110]); see [87] for a survey. The strictness of this objective cuts both ways. On the one hand, a successful learner uncovers substantial structure of the target distribution. On the other hand, this objective is impossible when the means and covariances are extremely close. Thus, algorithms

for parameter estimation of mixtures necessarily require some separation assumptions on the target parameters.

Density estimation has a long history in the statistics literature, where the focus is on the sample complexity question; see [51, 52, 131] for general background. It was first studied in the computational learning theory community under the name *PAC learning of distributions* by [92], whose focus is on the computational complexity of the learning problem.

Various measures of dissimilarity between distributions have been considered in existing density estimation schemes. For example, one natural measure is the TV distance [52, Chapter 5]; this has been used by several existing algorithms for mixtures of Gaussians [15, 33, 45]. Another natural measure is the KL divergence, which has also been used for mixtures of Gaussians [67]. Yet another natural measure is the L^p distance for $p > 1$; for example, some prior work has used the L^2 distance for density estimation [7, 57]. (The L^p distance between densities f and g is defined as $\|f-g\|_p \coloneqq \left(\int_{\mathbb{R}^d} |f(x)-g(x)|^p \,\mathrm{d}x\right)^{1/p}$.) This chapter focuses on the TV distance (i.e., the L^1 distance), and we provide justification for this choice in Section 4.2.

The *minimum distance estimate* [52, Section 6.8] is one possible approach for deriving sample complexity upper bounds for distribution learning. This approach is based on uniform convergence theory. In particular, an upper bound for any class of distributions can be achieved by bounding the VC-dimension of an associated set system, called the Yatracos class (see [52, page 58] for the definition). For example, [60] used this approach to bound the sample complexity of learning high-dimensional log-concave distributions. For the class of single Gaussians in d dimensions, this approach leads to the optimal sample complexity upper bound of $O(d^2/\varepsilon^2)$. However, for mixtures of Gaussians and axis-aligned Gaussians in $\mathbb{R}^d$, the best known VC-dimension bounds (see [52, Section 8.5] and [10, Theorem 8.14]) result in loose upper bounds of $O(k^4d^4/\varepsilon^2)$ and $O((k^4d^2 + k^3d^3)/\varepsilon^2)$, respectively.

Another approach is to first approximate the mixture class using a more manageable class such as piecewise polynomials, and then study the associated Yatracos class; see, e.g., [33]. However, piecewise polynomials do a poor job in approximating d-dimensional Gaussians, resulting in an exponential dependence on d.

For density estimation of mixtures of Gaussians using the TV distance, the best known sample complexity upper bounds (in terms of k and d) are $\widetilde{O}(kd^2/\varepsilon^4)$ for general Gaussians and $\widetilde{O}(kd/\varepsilon^4)$ for axis-aligned Gaussians, both due to [15]. For the general Gaussian case, their method takes an i.i.d. sample of size $\widetilde{O}(kd^2/\varepsilon^2)$ and partitions this sample in every possible way into k subsets. Based on those partitions, $k^{\widetilde{O}(kd^2/\varepsilon^2)}$ "candidate distributions" are generated. The problem is then reduced to learning with respect to this finite class of candidates. Their sample complexity has a suboptimal factor of $1/\varepsilon^4$, of which $1/\varepsilon^2$ arises in their approach for choosing the best candidate, and another factor $1/\varepsilon^2$ is due to the exponent in the number of candidates.

Our approach via compression schemes also ultimately reduces the problem to learning with respect to finite classes, although yielding a more refined bound than previous work. In the case of mixtures of Gaussians, one factor of $1/\varepsilon^2$ is again incurred due to learning with respect to finite classes. The key is that the number of compressed samples is only $\widetilde{O}_d(1)$, so the overall sample complexity bound has only an $\widetilde{O}(1/\varepsilon^2)$ dependence on ε.

	Number of Gaussians	Dimension	Axis-aligned	Sample complexity	Reference
upper bounds	1	d	no	$O(d^2/\varepsilon^2)$	standard
	1	d	yes	$O(d/\varepsilon^2)$	standard
	k	1	n/a	$\widetilde{O}(k/\varepsilon^2)$	[33]
	k	d	no	$\widetilde{O}(kd^2/\varepsilon^2)$	this chapter
	k	d	yes	$\widetilde{O}(kd/\varepsilon^2)$	this chapter
lower bounds	1	d	no	$\widetilde{\Omega}(d^2/\varepsilon^2)$	this chapter
	1	d	yes	$\widetilde{\Omega}(d/\varepsilon^2)$	[134]
	k	1	n/a	$\widetilde{\Omega}(k/\varepsilon^2)$	[134]
	k	d	no	$\widetilde{\Omega}(kd^2/\varepsilon^2)$	this chapter
	k	d	yes	$\widetilde{\Omega}(kd/\varepsilon^2)$	[134]

Table 4.1: Bounds on the sample complexities of learning Gaussian mixtures and their subclasses. The lower bounds are minimax (i.e., worst-case). The bounds in the first two rows are well known; proofs can be found in [15].

As for lower bounds on the sample complexity for learning mixtures of Gaussians under the TV distance, much fewer results are known. The only lower bound prior to this work is due to [134], which shows a bound of $\widetilde{\Omega}(kd/\varepsilon^2)$ for learning mixtures of axis-aligned Gaussians (and hence for general Gaussians as well). This bound is tight for the axis-aligned case, as we show in Theorem 4.3, but loose in the general case, as we show in Theorem 4.2. We note that an alternative construction was provided in [54] giving the same lower bound as ours using a deterministic construction.

A summary of bounds on the sample complexity for learning Gaussian mixtures and their subclasses is presented in Table 4.1.

4.2 Justification of our model

Several of the existing models for learning mixtures of Gaussians need some structural assumption on the distribution. For example, learning under the parameter estimation model requires that the means are sufficiently separated and that the mixing weights are not too small, see the discussion after [87, Definition 1].

A key motivation for our work is to study a model for learning mixtures of Gaussians that requires no structural assumptions at all. Specifically, we would like to identify a model in which Gaussians can be learned up to error ϵ with sample complexity depending only on k, d and ϵ, then derive optimal sample complexity bounds in that model. Density estimation under the TV distance is one such model: Ashtiani et al. [15, Theorem 14] and Theorem 4.1 in this paper show that mixtures of Gaussians can be learned up to error ε with sample complexity depending on k, d, and ε only. In this section we provide further justification for using this particular model.

In Section 4.2.1 we argue that the TV distance is not an arbitrary choice. If instead we had used

the KL divergence or any L^p distance, with $p > 1$, then the sample complexity must necessarily depend on the structural properties of the distribution. Thus, TV distance is a natural choice.

4.2.1 Comparison to KL divergence and L^p distances

In this section we consider the problem of density estimation for a mixture of Gaussians, using a distance measure that is either the KL divergence[2] or an L^p distance with $p > 1$. Under these distance measures, we show that the sample complexity of this problem must necessarily depend on structural properties of the distribution — that is, it cannot be bounded purely as a function of k, d and ϵ.

First we consider using the KL divergence. We show that no algorithm can guarantee that the KL divergence between the true distribution and the output distribution is smaller than any finite number with a uniformly bounded number of samples. In fact, this even holds for mixtures of two one-dimensional Gaussians with unit variances.

Theorem 4.4. *Let $\mathcal{F}$ be the class of mixtures of two Gaussians in $\mathbb{R}$, both of which have unit variance. Let $\mathcal{A}$ be any algorithm (possibly randomized) whose input is a finite-length sequence of real numbers and whose output is a (Lebesgue) measurable density function. Then for every $m \in \mathbb{N}$ and every $\tau > 0$, there exists a density $f \in \mathcal{F}$ such that if $X'_1, \dots, X'_m \sim f$ then $D_{\mathrm{KL}}(f \parallel \mathcal{A}(X'_1, \dots, X'_m)) \geq \tau$ with probability at least 0.98.*

Remark 4.5. *We note that Feldman et al. [68] consider learning mixtures of axis-aligned Gaussians under KL divergence. However, Theorem 4.4 does not contradict the results in [68] because they assume that the means and variances are bounded.*

The intuition behind the theorem is as follows. Let $a \in \mathbb{N}$ and consider the set of distributions $(1-\delta) \cdot \mathcal{N}(0,1) + \delta \cdot \mathcal{N}(a,1)$ where $\delta \ll 1/m$. Any algorithm that draws m samples from such a distribution will likely have all of its samples come from $\mathcal{N}(0,1)$. However, the only way for the KL divergence to be small is if the distribution returned by $\mathcal{A}$ has non-negligible mass near the $\mathcal{N}(a,1)$ distribution, which is impossible since the samples provide no information about a.

Let ν be the Lebesgue measure on $\mathbb{R}$. We begin with a simple calculation that will be useful later.

Claim 4.6. *Suppose $I \subseteq \mathbb{R}$ satisfies $\nu(I) \geq \gamma$. Moreover, let $f, h \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ be measurable density functions such that $f(x) \geq \beta, h(x) \leq \alpha$ for all $x \in I$ and $f(x) > 0$ for all $x \in \mathbb{R}$. Then $D_{\mathrm{KL}}(f \parallel h) \geq \gamma\beta \log(\beta/\alpha) - 1/e$.*

Proof. First, let us write

$$D_{\mathrm{KL}}(f \parallel h) = \int_I f(x) \log \frac{f(x)}{h(x)} \, \mathrm{d}x + \int_{I^c} f(x) \log \frac{f(x)}{h(x)} \, \mathrm{d}x.$$

For the first integral, we have

$$\int_I f(x) \log \frac{f(x)}{h(x)} \, \mathrm{d}x \geq \int_I \beta \log \frac{\beta}{\alpha} \, \mathrm{d}x \geq \gamma\beta \log(\beta/\alpha).$$

[2] Recall that KL divergence is not symmetric; we only consider using KL divergence in one direction.

Next, we bound the second integral and show that it has value at least $-1/e$ which completes the proof. Let $F = \int_{I^c} f(x)\,\mathrm{d}x$ and $H = \int_{I^c} h(x)\,\mathrm{d}x$. Note that $F > 0$ as $f(x) > 0$ for all $x \in \mathbb{R}$. If $H = 0$ then $h(x) = 0$ almost everywhere on I^c so the second integral is $+\infty$.

So assume that $H > 0$. Then f/F and h/H are densities on I^c. Hence, we have

$$\int_{I^c} f(x) \log \frac{f(x)}{h(x)}\,\mathrm{d}x = \underbrace{F \int_{I^c} \frac{f(x)}{F} \log \frac{f(x)/F}{h(x)/H}\,\mathrm{d}x}_{\geq 0} + F \int_{I^c} \frac{f(x)}{F} \log \frac{F}{H}\,\mathrm{d}x \geq F \log(F/H),$$

where the inequality is because the KL divergence of two densities is always non-negative. Since $H \leq 1$, we have $-\log(H) \geq 0$ so $F \log(F/H) = F \log F - F \log H \geq F \log F \geq -1/e$.

Combining the two bounds gives the claim. □

Proof of Theorem 4.4. We allow the algorithm $\mathcal{A}$ to be randomized. Denote by $\mathcal{A}(X_1, \ldots, X_m, R)$ the output of $\mathcal{A}$ given input $X_1, \ldots, X_m$ (the sampled data from the "true" distribution) and an independent source of randomness R. We will first analyze the behavior of the algorithm when the true distribution is $\mathcal{N}(0,1)$ and show that there exists some $a' \in \mathbb{R}$ for which the algorithm's output puts almost no probability mass on around a'. We then show that if the true distribution is a carefully chosen mixture of $\mathcal{N}(0,1)$ and $\mathcal{N}(a',1)$, then the algorithm's output does not change with high probability, so it still puts almost no mass on $\mathcal{N}(a',1)$; hence the KL divergence of the output and the true distribution is large.

Define the parameters $\delta = \frac{0.01}{m}$, $\beta = \frac{\delta}{\sqrt{2\pi}} \exp(-1/32)$, and $\alpha = \beta \exp\left(\frac{-4\tau - 4/e}{\beta}\right)$.

Let $X_1, \ldots, X_m \sim \mathcal{N}(0,1)$ and set $h = \mathcal{A}(X_1, \ldots, X_m, R)$. Note that h is random. Define the (random) set $H = \{\, x \in \mathbb{R} : h(x) \geq \alpha \,\}$. Then $\nu(H) \leq 1/\alpha$. For $a \in \mathbb{Z}$, define $I_a = [a - 1/4, a + 1/4]$. Note that the I_a are disjoint intervals. Hence $\sum_{a \in \mathbb{Z}} \nu(I_a \cap H) \leq 1/\alpha$ deterministically so $\mathbb{E}\left[\sum_{a \in \mathbb{Z}} \nu(I_a \cap H)\right] \leq 1/\alpha$. Note that the left hand side of the inequality is an infinite sum while the right hand side is a finite number. Since expectation is linear, we can find $a' \in Z$ such that $\mathbb{E}\left[\nu(I_{a'} \cap H)\right] \leq 1/400$. By Markov's Inequality, $\nu(I_{a'} \cap H) \leq 1/4$ with probability at least 0.99. We condition on this event.

Define $f = (1-\delta)\mathcal{N}(0,1) + \delta \cdot \mathcal{N}(a',1)$ and note that for all $x \in I_{a'}$ we have $f(x) \geq \frac{\delta}{\sqrt{2\pi}} \exp(-1/32) = \beta$, and f is positive everywhere. Let $J_{a'} = I_{a'} \setminus H$. Then $\nu(J_{a'}) \geq \nu(I_{a'}) - \nu(I_{a'} \cap H) \geq 1/4$, and for all $x \in J_{a'}$ we have $f(x) \geq \beta$ and $h(x) < \alpha$. So

$$D_{\mathrm{KL}}(f \parallel h) \geq \beta \log(\beta/\alpha)/4 - 1/e = \tau,$$

where the inequality is by Claim 4.6 and the equality is by definition of α. Hence, $D_{\mathrm{KL}}(f \parallel h) \geq \tau$ with probability at least 0.99.

Note that $d_{\mathrm{TV}}(f, \mathcal{N}(0,1)) \leq \delta$. If $S = (X_1, \ldots, X_m)$ and $S' = (X'_1, \ldots, X'_m)$ where $X_i \sim \mathcal{N}(0,1)$ and $X'_i \sim f$ then $d_{\mathrm{TV}}(S, S') \leq m\delta = 0.01$. Hence, if $h = \mathcal{A}(S, R)$ and $h' = \mathcal{A}(S', R)$ then $d_{\mathrm{TV}}(h, h') \leq 0.01$ so $\mathbb{P}[D_{\mathrm{KL}}(f \parallel h') \geq \tau] \geq \mathbb{P}[D_{\mathrm{KL}}(f \parallel h) \geq \tau] - 0.01 \geq 0.98$, completing the proof. □

Next we consider L^p distances, and prove a result analogous to Theorem 4.4. The main difference is that the argument uses Gaussians with non-unit variance, which can strongly influence the L^p distance.

Theorem 4.7. *Let $\mathcal{F}$ be the class of mixtures of two Gaussians in $\mathbb{R}$. Let $\mathcal{A}$ be any algorithm (possibly randomized) whose input is a finite-length sequence of real numbers and whose output is a (Lebesgue) measurable density function. Then for every $p > 1$, every $m \in \mathbb{N}$, and every $\tau > 0$, there exists a density $f \in \mathcal{F}$ such that if $X'_1, \ldots, X'_m \sim f$ then $\|f - \mathcal{A}(X'_1, \ldots, X'_m)\|_p \geq \tau$ with probability at least* 0.98.

Proof of Theorem 4.7. First, define the parameters $\delta = \frac{0.01}{m}$, $\sigma^{p-1} = \frac{\delta^p}{\tau^p 6^p}\sqrt{\ln(9/2\pi)}$, and $M = 4\sigma\sqrt{\ln(9/2\pi)}$.

Let $X_1, \ldots, X_m \sim \mathcal{N}(0,1)$ and set $h = \mathcal{A}(X_1, \ldots, X_m, R)$, where, as in the proof of Theorem 4.4, R is the algorithm's independent source of randomness. Note that h is random. Define $H = \{\, x \in \mathbb{R} : h(x) \geq \delta/6\sigma \,\}$. Then $\nu(H) \leq 6\sigma/\delta$. For $a \in \mathbb{Z}$, define the intervals $I_a = [aM - M/4, aM + M/4]$ and note that I_a are disjoint intervals. Hence, $\sum_{a\in\mathbb{Z}} \nu(I_a \cap H) \leq 6\sigma/\delta$ deterministically so $\mathbb{E}\left[\sum_{a\in\mathbb{Z}} \nu(I_a \cap H)\right] \leq 6\sigma/\delta$. Note that the left hand side of the inequality is an infinite sum while the right hand side of the inequality is a finite number. Since expectation is linear, we can find $a' \in Z$ such that $\mathbb{E}\left[\nu(I_{a'} \cap H)\right] \leq M/400$. By Markov's Inequality, $\nu(I_{a'} \cap H) \leq M/4$ with probability at least 0.99. We condition on this event.

Define $f = (1-\delta)\mathcal{N}(0,1) + \delta \cdot \mathcal{N}(a', \sigma^2)$. Now, note that for $x \in I_{a'}$, we have

$$f(x) \geq \delta \frac{1}{\sqrt{2\pi}\sigma} \exp(-(M/4)^2/2\sigma^2) = \delta/3\sigma.$$

Let $J_{a'} = I_{a'} \setminus H$. Then $\nu(J_{a'}) \geq M/2 - M/4 = M/4 = \sigma\sqrt{\ln(9/2\pi)}$ and for all $x \in J_{a'}$, we have $f(x) \geq \delta/3\sigma$ and $h(x) \leq \delta/6\sigma$. So

$$\|f - h\|_p^p \;\geq\; \int_{J_a} |f(x) - h(x)|^p \,\mathrm{d}x \;\geq\; \frac{\delta^p}{(6\sigma)^p}\sigma\sqrt{\ln(9/2\pi)} \;=\; \frac{\delta^p}{6^p\sigma^{p-1}}\sigma\sqrt{\ln(9/2\pi)} \;=\; \tau^p,$$

where the last equality is by definition of σ. Hence, $\|f - h\|_p \geq \tau$ with probability at least 0.99.

Note that $d_{\mathrm{TV}}(f, \mathcal{N}(0,1)) \leq \delta$. If $S = (X_1, \ldots, X_m)$ and $S' = (X'_1, \ldots, X'_m)$ where $X_i \sim \mathcal{N}(0,1)$ and $X'_i \sim f$ then $d_{\mathrm{TV}}(S, S') \leq m\delta = 0.01$. Hence, if $h = \mathcal{A}(S, R)$ and $h' = \mathcal{A}(S', R)$ then $d_{\mathrm{TV}}(h, h') \leq 0.01$ so $\mathbb{P}[\|f - h'\|_p \geq c] \geq \mathbb{P}[\|f - h\|_p \geq c] - 0.01 \geq 0.98$. □

4.3 Compression

In this section, we will precisely define a notion of sample compression and show how it can be used for distribution learning. Much of the material presented in this section can also be found in Ashtiani's thesis [16, §7.2].

4.3.1 Definition of compression

Let $\mathcal{F}$ be a distribution over a domain Z. A compression scheme for $\mathcal{F}$ involves two parties: an *encoder* and a *decoder*. Intuitively, one can think of the encoder and decoder as follows.

- The *encoder* knows a distribution $g \in \mathcal{F}$. Its goal is to communicate this to the decoder. However, the encoder must do this as follows. First, the encoder draws m i.i.d. samples from g.

Having drawn these m samples, the encoder selects a list of length τ from the set of samples. Furthermore, the encoder may construct a bit sequence of length t. The encoder then sends a message consisting of the list of samples and the bit sequence. Ideally, m, τ, t should all be relatively small.

- The *decoder* receives the message from the encoder consisting of the list of samples and the bit sequence. Using this, it then outputs a distribution $\hat{g}$. The decoder is successful if $\hat{g}$ is close to g (in total variation distance).

Let us make two remarks at this point. First, it could be possible that the samples that the encoder draws is not at all representative of g at all. For example, if g is a single Gaussian distribution, with very small probability, all its samples could come from one of its tails. Thus, we will only ensure that the compression scheme is successful with some constant probability, say $2/3$.

Second, requiring that $g \in \mathcal{F}$ is a condition which will only be useful for the realizable setting. To handle the agnostic setting, we will modify the encoder where it knows a distribution $g \in \mathcal{F}$ but it is only allowed to draw samples from a *different* distribution q (not necessarily in $\mathcal{F}$).

Third, the covering arguments that we described in Subsection 3.3.2 is actually a special case of compression by setting $m = \tau = 0$. In this case, the encoder only sends bits to the decoder and the bit sequence that it sends would correspond to an element in the cover.

Formal definitions. We now make the above definitions of compression schemes more precise.

Definition 4.8. A decoder for a distribution class $\mathcal{F}$ on a domain Z is a deterministic function $\mathcal{J} \colon \cup_{n=0}^{\infty} Z^n \times \cup_{n=0}^{\infty} \{0,1\}^n \to \mathcal{F}$. In other words, $\mathcal{J}$ takes as input a finite sequence of elements from Z and a finite bit-string and outputs an element in $\mathcal{F}$.

Definition 4.9 (robust compression schemes)**.** Let $\tau, t, m \colon (0,1) \to \mathbb{Z}$ be decreasing functions and let $r \in [0,2]$. We say $\mathcal{F}$ admits (τ, t, m) r-robust compression if there exists a decoder $\mathcal{J}$ for $\mathcal{F}$ such that for any distribution $g \in \mathcal{F}$, and for any distribution q on Z with $\|g - q\|_1 \leq r$, the following holds.

> For any $\varepsilon \in (0,1)$, if a set of $m(\varepsilon)$ samples S is drawn from $q^{m(\varepsilon)}$ then with probability $2/3$ (over the samples), there exists a sequence L of at most $\tau(\varepsilon)$ elements of S, and a bit-sequence B of length at most $t(\varepsilon)$, such that $\|\mathcal{J}(L, B) - g\|_1 \leq \varepsilon$.

We note here that L is an *ordered sequence* and is allowed to contain duplicates. Thus, it is conceivable that the *order* of the sequence may be used by the decoder.

Remark 4.10. *In the special case where $r = 0$, we will refer to the compression scheme as being a* non-robust *compression scheme. In the non-robust case, the samples come from the distribution g itself. However, in the robust case, the samples come from a distribution q which is only close to g (in total variation distance).*

Remark 4.11. *In Definition 4.9, we required that L and B exist with probability $2/3$. If one wishes to boost this probability, one could draw k sets of samples $S_1, \ldots, S_k$ where each $S_i \sim q^{m(\varepsilon)}$. For each S_i, a suitable L and B will fail to exist with probability at most $1/3$, so the probability that a suitable L and B do not exist in $S_1, \ldots, S_k$ is at most 3^{-k}. Thus to get a compression scheme with success probability $1 - \delta$ we can set $k = \lceil \log_3(1/\delta) \rceil$. We will often make use of this calculation in our proofs.*

4.3.2 Connection between compression and learning

We now show that if a class of distributions has a robust compression scheme then it can be learned in the agnostic density estimation model. In fact, our proof will show how to utilize a robust compression scheme as a blackbox to obtain a sample-efficient algorithm for density estimation.

The main idea is as follows. Note that the encoder cannot be implemented in the density estimation model because the encoder requires prior knowledge of the distribution g. However, the interaction between the encoder and decoder is only via a short message. Thus, we make a collection of all possible messages that may have been sent from the encoder to the decoder. The assertion that a sample compression scheme exists implies that for at least one of these messages, the output of the decoder will be a distribution which is close to g.

At this point, the only remaining task is to select a distribution, from the finite collection that we constructed, that is close to g. In Subsection 3.3.1, we described an algorithm that does precisely this and we will now use the algorithm as a blackbox. For convenience, we restate the theorem here.

Theorem 3.12 ([53, Theorem 6.3])**.** *Let $\mathcal{F}$ be a finite class of distributions and $M = \log|\mathcal{F}|$. There is an algorithm $\mathcal{A}$ such that for any distribution f, if $\mathcal{A}$ is given $O(\log(M/\delta)/\varepsilon^2)$ i.i.d. samples from f then with probability at least $1-\delta$ (over the samples), $\mathcal{A}$ outputs a distribution $\hat{g} \in \mathcal{F}$ satisfying*

$$d_{\mathrm{TV}}(f, \hat{g}) \leq 3 \cdot \min_{g \in \mathcal{F}} d_{\mathrm{TV}}(f, g) + \varepsilon.$$

Our approach for relating compression schemes and density estimation, described informally above, is made formal by the following theorem. It uses Theorem 3.12 to select a good distribution that the decoder can output. Note that we assume the learner knows all the problem parameters, such as $k, d, \varepsilon, \delta, \tau, t, m$, and r, but is oblivious to the target distribution.

We will begin with the realizable setting.

Theorem 4.12 (compression implies learning, realizable setting)**.** *Suppose $\mathcal{F}$ admits (τ, t, m) compression. Let $\tau'(\varepsilon) \coloneqq \tau(\varepsilon) + t(\varepsilon)$. Then $\mathcal{F}$ can be PAC-learned with sample complexity*

$$O\left(m(\varepsilon/4)\log(1/\delta) + \frac{\tau'(\varepsilon/4)\log(m(\varepsilon/4)\log(1/\delta)) + \log(1/\delta)}{\varepsilon^2}\right) = \widetilde{O}\left(m\Big(\frac{\varepsilon}{4}\Big) + \frac{\tau'(\varepsilon/4)}{\varepsilon^2}\right).$$

In other words, there is an algorithm which receives the above number of samples from an unknown distribution $g \in \mathcal{F}$ and outputs $\hat{g}$ such that $\|g - \hat{g}\|_1 \leq \varepsilon$.

Proof. Let $\mathcal{J}$ be the decoder which guarantees the existence of a (τ, t, m) compression scheme for $\mathcal{F}$. Let $g \in \mathcal{F}$ denote the unknown distribution.

By assumption, with probability $2/3$, if S is a set of $m(\varepsilon)$ i.i.d. samples from g then there exists a sequence L of at most $\tau(\varepsilon)$ elements from S and a bit-sequence B of length at most $t(\varepsilon)$ such that

$$\|\mathcal{J}(L, B) - g\|_1 \leq \varepsilon. \tag{4.1}$$

As in Remark 4.11, we can boost the success probability that there exists L and B satisfying Eq. (4.1) from $2/3$ to $1-\delta$ by drawing $\lceil \log_3(1/\delta) \rceil$ sets of $m(\varepsilon)$ i.i.d. samples from g. For the rest of the argument,

we condition on the event that we have already drawn $N = m(\varepsilon)\lceil \log_3(1/\delta)\rceil$ samples and that there exists L and B satisfying Eq. (4.1).

Of course, the learner is unaware of L and B but it can generate all possible sequences of length $\tau(\varepsilon)$ and all possible bit-sequences of length $t(\varepsilon)$. For each of these inputs, it obtains a distribution from $\mathcal{J}$ and at least one of these sequences will satisfy Eq. (4.1).

The total number of sequences of length at most $\tau(\varepsilon)$ is bounded above by $(N+1)^{\tau(\varepsilon)}$ (the additional $+1$ is to account for the fact that the sequence L may be less than $\tau(\varepsilon)$). Similarly, the number of bit sequences of length at most $t(\varepsilon)$ is bounded above by $2^{t(\varepsilon)+1}$. Hence, the learner can generate a finite hypothesis class of size

$$M \le (N+1)^{\tau(\varepsilon)} \cdot 2^{t(\varepsilon)+1} \le (N+1)^{\tau'(\varepsilon)+1}$$

with the guarantee that at least one hypothesis is within total variation distance ε of g.

We can now appeal to Theorem 3.12 which asserts that there is an algorithm such that, given $O(\log(M/\delta)/\varepsilon^2)$ i.i.d. samples from g, outputs $\hat{g}$ with $\|g - \hat{g}\|_1 \le 4\varepsilon$ with probability at least $1-\delta$.

In total, the algorithm has drawn at most

$$O\left(m(\varepsilon)\log(1/\delta) + \frac{\tau'(\varepsilon)\log(m(\varepsilon)\log(1/\delta)) + \log(1/\delta)}{\varepsilon^2}\right)$$

samples from g and its success probability is at least $1-2\delta$ (because the compression scheme has probability at most δ of failure and the algorithm of Theorem 3.12 has probability at most δ of failure). Replacing δ with $\delta/2$ and ε with $\varepsilon/4$ proves the theorem. □

The following theorem relates robust compression to agnostic learning. Its proof is a slightly more technical version of the proof of Theorem 4.12.

Theorem 4.13 (compression implies learning, agnostic setting)**.** *Suppose $\mathcal{F}$ admits (τ, t, m) r-robust compression. Let $\tau'(\varepsilon) \coloneqq \tau(\varepsilon) + t(\varepsilon)$. Then $\mathcal{F}$ can be $\max\{3, 2/r\}$-learned in the agnostic setting using*

$$O\left(m\Big(\frac{\varepsilon}{6}\Big)\log\Big(\frac{1}{\delta}\Big) + \frac{\tau'(\varepsilon/6)\log(m(\frac{\varepsilon}{6})\log_3(1/\delta)) + \log(1/\delta)}{\varepsilon^2}\right) \;=\; \widetilde{O}\left(m\Big(\frac{\varepsilon}{6}\Big) + \frac{\tau'(\varepsilon/6)}{\varepsilon^2}\right) \text{ samples.}$$

In other words, there is an algorithm which receives the above number of samples from an unknown distribution q and outputs $\hat{g}$ such that

$$\|\hat{g} - q\|_1 \le \max\{3, 2/r\} \cdot \inf_{f\in\mathcal{F}} \|f - q\|_1 + \varepsilon.$$

Proof. The proof of this theorem is similar to that of Theorem 4.12 but some care is needed to deal with the possibility that the unknown distribution q may not be in $\mathcal{F}$.

Let $\alpha = \inf_{f\in\mathcal{F}} \|f - q\|_1$ be the approximation error of q with respect to $\mathcal{F}$. The goal of the learner is to find a distribution $\hat{h}$ such that $\|\hat{h} - q\|_1 \le \max\{3, 2/r\}\cdot\alpha + \varepsilon$.

First, consider the case $\alpha \le r$. In this case, we develop a learner that finds a distribution $\hat{h}$ such

that $\|\hat{h} - q\|_1 \le 3\alpha + \varepsilon$. Let $g \in \mathcal{F}$ be a distribution such that

$$\|g - q\|_1 \;\le\; \alpha + \frac{\varepsilon}{12}. \tag{4.2}$$

Such a g exists by the definition of α. By assumption, $\mathcal{F}$ admits (τ, t, m) compression. Let $\mathcal{J}$ denote the corresponding decoder. Given ε, the learner first asks for an i.i.d. sample $S \sim q^{m(\varepsilon/6)\cdot\log_3(2/\delta)}$. Recall the definition of robust compression and Remark 4.11, which allows us to amplify the success probability of the decoder. Then, with probability at least $1 - \delta/2$, there exist $L \in S^{\tau(\varepsilon/6)}$ and $B \in \{0,1\}^{t(\varepsilon/6)}$ satisfying the following guarantee: letting $h^* \coloneqq \mathcal{J}(L, B)$, we have

$$\|h^* - g\|_1 \;\le\; \frac{\varepsilon}{6}. \tag{4.3}$$

The learner is of course unaware of L and B. However, given the sample S, it can try all of the possibilities for L and B and create a candidate set of distributions. More concretely, let

$$H \;=\; \{\, \mathcal{J}(L, B) \;:\; L \in S^{\tau(\varepsilon/6)},\, B \in \{0,1\}^{t(\varepsilon/6)} \,\}.$$

Note that

$$|H| \;\le\; \big(m(\varepsilon/6)\log_3(2/\delta)\big)^{\tau(\varepsilon/6)} 2^{t(\varepsilon/6)} \;\le\; \big(m(\varepsilon/6)\log_3(2/\delta)\big)^{\tau'(\varepsilon/6)}.$$

Since H is finite, we can use the algorithm of Theorem 3.12 to find a good candidate $\hat{h}$ from H. In particular, we set the accuracy parameter in Theorem 3.12 to be $\varepsilon/16$ and the confidence parameter to be $\delta/2$. In this case, Theorem 3.12 requires

$$\frac{\log(6|H|^2/\delta)}{2(\varepsilon/16)^2} \;=\; O\left(\frac{\tau'(\varepsilon/6)\log(m(\frac{\varepsilon}{6})\log_3(\frac{1}{\delta})) + \log(\frac{1}{\delta})}{\varepsilon^2}\right) \;=\; \widetilde{O}(\tau'(\varepsilon/6)/\varepsilon^2)$$

additional samples, and its output $\hat{h}$ satisfies the following guarantee:

$$\begin{aligned}
\|\hat{h} - q\|_1 \;&\le\; 3\|h^* - q\|_1 + 4\frac{\varepsilon}{16} && \text{(by Theorem 3.12)} \\
&\le\; 3(\|h^* - g\|_1 + \|g - q\|_1) + \frac{\varepsilon}{4} \\
&\le\; 3\left(\frac{\varepsilon}{6} + \left(\alpha + \frac{\varepsilon}{12}\right)\right) + \frac{\varepsilon}{4} && \text{(by (4.2) and (4.3))} \\
&=\; 3\alpha + \varepsilon.
\end{aligned}$$

Note that the above procedure uses $\widetilde{O}\big(m(\varepsilon/6) + \tau'(\varepsilon/6)/\varepsilon^2\big)$ samples, and the probability of failure is at most δ. That is, the probability of either H not containing a good h^*, or the failure of Theorem 3.12 in choosing a good candidate among H, is bounded by $\delta/2 + \delta/2 = \delta$.

The other case, $\alpha > r$, is trivial: the learner outputs some distribution $\widehat{h}$. Since $\widehat{h}$ and q are density functions, we have $\|\widehat{h} - q\|_1 \le 2 < \frac{2}{r}\cdot\alpha < \max\{3, 2/r\}\cdot\alpha + \varepsilon$. □

4.3.3 Combining compression schemes

To conclude this section, we state a few results showing that compression schemes can be combined in useful ways. These results concern product distributions (which will be useful for axis-aligned Gaussians) and mixture distributions (which will be useful for mixtures of Gaussians).

First, Lemma 4.14 below states that if a class $\mathcal{F}$ of distributions can be robustly compressed, then the class of distributions that are formed by taking products of members of $\mathcal{F}$ can also be robustly compressed. If $p_1, \ldots, p_d$ are distributions over domains $Z_1, \ldots, Z_d$, then $\prod_{i=1}^d p_i$ denotes the standard product distribution over $\prod_{i=1}^d Z_i$. For a class $\mathcal{F}$ of distributions, define

$$\mathcal{F}^d \coloneqq \left\{ \prod_{i=1}^d p_i \,:\, p_1, \ldots, p_d \in \mathcal{F} \right\}.$$

Lemma 4.14 (compressing product distributions)**.** *For any τ, t, m, r, d,*

$$\begin{aligned} &\text{if } \mathcal{F} \text{ admits } \big(\quad \tau(\varepsilon), \qquad t(\varepsilon), \qquad\quad m(\varepsilon) \qquad \big) \ r\text{-robust compression,} \\ &\text{then } \mathcal{F}^d \text{ admits } \big(d \cdot \tau(\varepsilon/d),\ d \cdot t(\varepsilon/d),\ \log_3(3d) \cdot m(\varepsilon/d) \big) \ r\text{-robust compression.} \end{aligned}$$

The proof of Lemma 4.14 is fairly intuitive: if one has a compression scheme for $\mathcal{F}$, one can imagine running d copies of the compression scheme, one for each coordinate.

For the proof of Lemma 4.14, we need the following standard proposition which can be proved, e.g., using the coupling characterization of the total variation distance.

Proposition 4.15 (Lemma 3.3.7 in [122])**.** *For $i \in [d]$, let p_i and q_i be probability distributions over the same domain Z. Then $\|\Pi_{i=1}^d p_i - \Pi_{i=1}^d q_i\|_1 \le \sum_{i=1}^d \|p_i - q_i\|_1$.*

Proof of Lemma 4.14. Let $G = \Pi_{i=1}^d g_i$ be an arbitrary element of $\mathcal{F}^d$, with all $g_i \in \mathcal{F}$. Let Q be an arbitrary distribution over Z^d, subject to $\|G - Q\|_1 \le r$. Let $q_1, \ldots, q_d$ be the marginal distributions of Q on the d components. Observe that $\|q_j - g_j\|_1 \le r$ for each $j \in [d]$, since Fact 3.6 implies that projection onto a coordinate cannot increase the total variation distance.

The lemma's hypothesis is that $\mathcal{F}$ admits (τ, t, m) r-robust compression. Let $\mathcal{J}$ denote the corresponding decoder, let $m_0 \coloneqq m(\varepsilon/d)\log_3(3d)$, and $S \sim Q^{m_0}$. To prove the lemma we must encode an ε-approximation of G using $d \cdot \tau(\varepsilon/d)$ elements of S and $d \cdot t(\varepsilon/d)$ bits.

Since S contains m_0 samples, each of which is a d-dimensional vector, we may think of S as a $d \times m_0$ matrix over Z. Let S_i denote the ith row of this matrix. That is, for $i \in [d]$, let $S_i \in Z^{m_0}$ be the vector of the ith components of all elements of S. By definition of q_i, we have $S_i \sim q_i^{m_0}$ for each i. As observed above, we have $\|q_i - g_i\| \le r$.

Apply Remark 4.11 with parameters ϵ/d and $\delta = 1/3d$ for each $i \in [d]$. Then, for each i, the following statement holds with probability at least $1 - 1/3d$: there exists a sequence L_i of at most $\tau(\varepsilon/d)$ elements of S_i, and a sequence B_i of at most $t(\varepsilon/d)$ bits, such that $\|\mathcal{J}(L_i, B_i) - g_i\|_1 \le \varepsilon/d$. By the union bound, this statement holds simultaneously for all $i \in [d]$ with probability at least $2/3$. We may encode these $L_1, \ldots, L_d, B_1, \ldots, B_d$ using $d \cdot \tau(\varepsilon/d)$ samples from S and $d \cdot t(\varepsilon/d)$ bits. Our decoder for $\mathcal{F}^d$ then

extracts $L_1, \ldots, L_d, B_1, \ldots, B_d$ from these samples and bits, and then outputs $\prod_{i=1}^{d} \mathcal{J}(L_i, B_i) \in \mathcal{F}^d$. Finally, Proposition 4.15 gives $\|\Pi_{i=1}^{d} \mathcal{J}(L_i, B_i) - G\|_1 \leq \sum_{i=1}^{d} \|\mathcal{J}(L_i, B_i) - g_i\|_1 \leq d \cdot \varepsilon/d \leq \varepsilon$, completing the proof. □

Our next lemma states that if a class $\mathcal{F}$ of distributions can be compressed, then the class of distributions that are formed by taking mixtures of members of $\mathcal{F}$ can also be compressed.

Lemma 4.16 (compressing mixtures, non-robustly)**.** *For any τ, t, m, r, d, suppose*

$$\textit{If } \mathcal{F} \textit{ admits } \big(\quad \tau(\varepsilon), \qquad\qquad t(\varepsilon), \qquad\qquad m(\varepsilon) \qquad \big) \textit{ (non-robust) compression}$$

$$\textit{then } k\textit{-mix}(\mathcal{F}) \textit{ admits } \big(\; k\tau(\varepsilon/3), \; kt(\varepsilon/3) + k\log_2(3k/\varepsilon), \; \tfrac{48k\log(6k)}{\varepsilon} m(\varepsilon/3) \; \big) \textit{ (non-robust) compression.}$$

We begin with a high-level overview of the proof and the proof itself will make this discussion formal. Suppose first, for the sake of simplicity, that one had a uniform mixture, i.e. all the mixing weights are equal to $1/k$. In this case, if one had a compression scheme for $\mathcal{F}$, one could imagine running k copies of the compression scheme for each of the different components in the mixture.

Now, what if the mixing weights were not all equal? In this case, one could still a compression scheme similar to that above but the encoder will tell the decoder what the mixing weights are (up to a very small discretization error).

Finally, what if a component has an extremely small mixing weight? In this case, it suffices to not bother to encode it at all since it has a negligible impact on the total variation distance.

Proof of Lemma 4.16. Consider any $g \in k\text{-mix}(\mathcal{F})$, so $g = \sum_{i \in [k]} w_i f_i$ for some distributions $f_1, \ldots, f_k \in \mathcal{F}$ and mixing weights $w_1, \ldots, w_k$. Define $m_0 \coloneqq 48m(\varepsilon/3)k\log(6k)/\varepsilon$, and draw $S \sim g^{m_0}$. Then S has the same distribution as the process that performs m_0 independent trials as follows: select a component i according to the weights w, then draw a sample from f_i. In the latter process, we may define S_i to be the sequence of samples that were generated using f_i. Our encoder for g will discretize the mixing weights, then use the compression scheme for $\mathcal{F}$ to separately encode each S_i.

Encoding the mixing weights. We encode $w_1, \ldots, w_k$ using bits as follows. Consider an $(\varepsilon/3k)$-net in ℓ_∞ for Δ_k of size $(3k/\varepsilon)^k$ (see Lemma B.17). Let $(\widehat{w}_1, \ldots, \widehat{w}_k)$ be an element in the net that has

$$\|(\widehat{w}_1, \ldots, \widehat{w}_k) - (w_1, \ldots, w_k)\|_\infty \;\leq\; \varepsilon/3k. \tag{4.4}$$

Encoding the element $(\widehat{w}_1, \ldots, \widehat{w}_k)$ from the net requires only $k\log_2(3k/\varepsilon)$ bits.

Encoding S_i. For any $i \in [k]$, we say that index i is *negligible* if $w_i \leq \varepsilon/(6k)$. For any negligible index we will approximate f_i by an arbitrary distribution $\widehat{f}_i$. For any non-negligible index we will likely have enough samples from f_i to use the compression scheme for $\mathcal{F}$ to encode a distribution $\widehat{f}_i$ that approximates f_i.

Define $m_1 = m(\varepsilon/3)\log(6k)$. For each non-negligible index i, a standard Chernoff bound shows that, with probability at least $1 - 1/6k$, we have $|S_i| \geq m_1$. By a union bound, this statement holds simultaneously for all non-negligible $i \in [k]$ with probability at least $5/6$.

Apply Remark 4.11 with parameters $\epsilon/3$ and $\delta = 1/6k$ for each non-negligible index i. Then, for each such i, the following statement holds with probability at least $1-1/6k$: there exist $\tau(\varepsilon/3)$ samples from S_i and $t(\varepsilon/3)$ bits from which the decoder can construct a distribution $\widehat{f_i}$ with

$$\|f_i - \widehat{f_i}\|_1 \;\leq\; \varepsilon/3. \tag{4.5}$$

By the union bound, this statement holds simultaneously for all non-negligible indices with probability at least $5/6$. The encoding consists of these samples and bits for each non-negligible i, whereas for negligible i we use the same number of samples and bits, chosen arbitrarily.

By a union bound, the failure probability of the encoding is at most $2 \cdot (1 - 5/6) = 1/3$.

Complexity of the encoding. The discretized weights require $k\log_2(3k/\varepsilon)$ bits. For each index $i \in [k]$, we use at most $\tau(\varepsilon/3)$ samples and $t(\varepsilon/3)$ bits. Thus, the total number of bits is $k \cdot t(\varepsilon/3) + k\log_2(3k/\varepsilon)$, and the total number of samples is $k \cdot \tau(\varepsilon/3)$.

Decoding. The decoder for k-mix$(\mathcal{F})$ is explicitly given the discretized weights $\widehat{w}_1, \ldots, \widehat{w}_k$. It is also given, for each index i, $\tau(\varepsilon/3)$ samples and $t(\varepsilon/3)$ bits, which it provides to the decoder for $\mathcal{F}$, yielding the distribution $\widehat{f_i}$. (Recall that, for a negligible index i, the distribution $\widehat{f_i}$ can be arbitrary.) The decoder outputs the distribution $\sum_i \widehat{w}_i \widehat{f_i}$.

To complete the proof of the lemma, we will show that $\|\sum_i w_i f_i - \sum_i \widehat{w}_i \widehat{f_i}\|_1 \leq \varepsilon$ with probability at least $2/3$. Let $N \subseteq [k]$ denote the set of negligible components. Recall that the encoder succeeds with probability at least $2/3$, in which case the decoded distributions $\widehat{f_i}$ will satisfy (4.5) for each $i \notin N$. We then have

$$\begin{aligned}
\left\|\sum_{i\in[k]} (\widehat{w}_i\widehat{f_i} - w_i f_i)\right\|_1 \;&\leq\; \left\|\sum_{i\in[k]} w_i(\widehat{f_i} - f_i)\right\|_1 + \left\|\sum_{i\in[k]} (\widehat{w}_i - w_i)\widehat{f_i}\right\|_1 \\
&\leq\; \left\|\sum_{i\in N} w_i(\widehat{f_i} - f_i)\right\|_1 + \left\|\sum_{i\notin N} w_i(\widehat{f_i} - f_i)\right\|_1 + \sum_{i\in[k]} |\widehat{w}_i - w_i| \cdot \left\|\widehat{f_i}\right\|_1 \\
&\leq\; \sum_{i\in N} w_i \cdot 2 \;+\; \sum_{i\notin N} w_i \cdot \frac{\varepsilon}{3} \;+\; \sum_{i\in[k]} \frac{\varepsilon}{3k} \cdot 1 \quad \text{(by (4.4) and (4.5))} \\
&\leq\; k \cdot \frac{\varepsilon}{6k} \cdot 2 \;+\; \frac{\varepsilon}{3} \;+\; \frac{\varepsilon}{3} \;=\; \varepsilon \quad \text{(by definition of } N\text{)}.
\end{aligned}$$

This completes the analysis of the compression scheme for k-mix$(\mathcal{F})$. □

Lemma 4.16 shows that non-robust compression of $\mathcal{F}$ implies non-robust compression of k-mix$(\mathcal{F})$. We do not know whether an analogous statement holds for robust compression. That is, does robust compression of $\mathcal{F}$ imply robust compression of k-mix$(\mathcal{F})$, for a general class $\mathcal{F}$? Nevertheless, in the

next lemma we show that if $\mathcal{F}$ can be robustly compressed, then k-mix($\mathcal{F}$) can be *learned in the agnostic setting.*

Lemma 4.17 (learning mixtures, robustly)**.** *Suppose $\mathcal{F}$ admits $(\tau(\varepsilon), t(\varepsilon), m(\varepsilon))$ r-robust compression, and let $\tau'(\varepsilon) \coloneqq \tau(\varepsilon) + t(\varepsilon)$. Then k-mix($\mathcal{F}$) admits $3(1 + 2/r)$-agnostic learning with sample complexity*

$$\widetilde{O}\left(\frac{km(\varepsilon/10)}{\varepsilon} + \frac{k\tau'(\varepsilon/10)\log m(\varepsilon/10)}{\varepsilon^2}\right).$$

The proof of Lemma 4.17 is quite technical and we relegate it to Appendix B.4.

4.4 Upper bound: learning mixtures of Gaussians by compression schemes

The main positive results of this paper are sample complexity bounds for learning mixtures of Gaussians (Theorems 4.1 and 4.3). In this section we prove these results by describing compression schemes for a single Gaussian, then applying the techniques of the previous section. To begin, we illustrate the techniques by considering the simple setting of non-robust compression for a single-dimensional Gaussian.

4.4.1 A simple example: mixtures of axis-aligned Gaussians, non-robustly

In this short section, we give an illustrative use of our compression framework to prove an upper bound of $\widetilde{O}(kd/\varepsilon^2)$ for the sample complexity of learning mixtures of k axis-aligned Gaussians in the realizable setting. The next section gives a much more general argument that works for general Gaussians in the agnostic setting.

Lemma 4.18. *The class of single-dimensional Gaussians admits a $(3, O(\log(1/\varepsilon)), 3)$ non-robust compression scheme.*

Before we prove Lemma 4.18, let us give a high-level intuition about the compression scheme. By Lemma 4.14 and Lemma 4.16, it suffices to find a compression scheme for a single, one-dimensional Gaussian. Indeed, Lemma 4.16 asserts that finding a compression scheme for a mixture of axis-aligned Gaussians reduces to finding a compression scheme for a single axis-aligned Gaussian. Next, note that an axis-aligned Gaussian is just a vector of independent one-dimensional Gaussians. So, Lemma 4.14 asserts that it suffices to compress a single one-dimensional Gaussian.

How can we compress a single one-dimensional Gaussian? Here is the idea for first encoding σ: imagine that we drew two samples $g_1, g_2 \sim \mathcal{N}(\mu, \sigma^2)$ where μ, σ^2 are unknown. Note that $g = \frac{1}{\sqrt{2}}(g_1 - g_2) \sim \mathcal{N}(0, \sigma^2)$. Now, with a fairly large constant probability, $|g| \in [0.01\sigma, 100\sigma]$. So $\lambda|g| = \sigma$ for some $\lambda \in [0.01, 100]$. In other words, there is some scaling of g that recovers σ (up to a sign). Moreover, the scaling is in some bounded interval so we can discretize it and the encoder will sending the scaling as well as g_1, g_2 to the decoder. This allows us to encode σ.

What about encoding μ? Note that if $g \in \mathcal{N}(\mu, \sigma^2)$ then g is usually within, say 100 standard deviations away from μ, i.e. $|g - \mu| \leq 100\sigma$. So given g, we will discretize an interval of length 100σ around g, and at least one of this points will be sufficiently close to μ.

Remark 4.19. *It is also possible to encode a Gaussian as follows. Suppose that we draw $O(1/\varepsilon)$ samples from $\mathcal{N}(\mu, \sigma^2)$. Then with good probability, there exists samples $X_1 \approx \mu + \sigma$ and another sample $X_2 \approx \mu - \sigma$. In this case $\frac{X_1+X_2}{2} \approx \mu$ and $\frac{X_1-X_2}{2} \approx \sigma$. This idea is carried out formally by Ashtiani [16, §7.6].*

Proof. Let $c < 1 < C$ be such that $\mathbb{P}_{X\sim\mathcal{N}(0,1)}[c < |X| < C] \geq 0.99$. Let $\mathcal{N}(\mu, \sigma^2)$ be the target distribution. We first show how to encode σ. Let $g_1, g_2 \sim \mathcal{N}(\mu, \sigma^2)$. Then $g = \frac{1}{\sqrt{2}}(g_1 - g_2) \sim \mathcal{N}(0, \sigma^2)$. So with probability at least 0.99, we have $\sigma c < |g| < \sigma C$. Conditioned on this event, we have $\lambda \coloneqq \sigma/g \in [-1/c, 1/c]$. We now choose $\hat{\lambda} \in \{0, \pm\varepsilon/2C^2, \pm 2\varepsilon/2C^2, \pm 3\varepsilon/2C^2 \ldots, \pm 1/c\}$ satisfying $|\hat{\lambda} - \lambda| \leq \varepsilon/(4C^2)$, and encode the standard deviation by $(g_1, g_2, \hat{\lambda})$. The decoder then estimates $\hat{\sigma} \coloneqq \hat{\lambda}(g_1 - g_2)/\sqrt{2}$. Note that $|\hat{\sigma} - \sigma| \leq |\hat{\lambda} - \lambda||g| \leq \sigma\varepsilon/(4C)$ and that the encoding requires two sample points and $O(\log(C^2/c\varepsilon)) = O(\log(1/\varepsilon))$ bits (for encoding $\hat{\lambda}$).

Now we turn to encoding μ. Let $g_3 \sim \mathcal{N}(\mu, \sigma^2)$. Then $|g_3 - \mu| \leq C\sigma$ with probability at least 0.99. We will condition on this event, which implies existence of some $\eta \in [-C, C]$ such that $g_3 + \sigma\eta = \mu$. We choose $\hat{\eta} \in \{0, \pm\varepsilon/2, \pm 2\varepsilon/2, \pm 3\varepsilon/2 \ldots, \pm C\}$ such that $|\hat{\eta} - \eta| \leq \varepsilon/4$, and encode the mean by $(g_3, \hat{\eta})$. The decoder estimates $\hat{\mu} \coloneqq g_3 + \hat{\sigma}\hat{\eta}$. Again, note that $|\hat{\mu} - \mu| = |\sigma\eta - \hat{\sigma}\hat{\eta}| \leq |\sigma\eta - \sigma\hat{\eta}| + |\sigma\hat{\eta} - \hat{\sigma}\hat{\eta}| \leq \sigma\varepsilon/2$. Moreover, encoding the mean requires one sample point and $O(\log(1/\varepsilon))$ bits.

To summarize, the decoder has $|\hat{\mu} - \mu| \leq \sigma\varepsilon/2$ and $|\hat{\sigma} - \sigma| \leq \sigma\varepsilon/2$. Plugging these bounds into Lemma B.6 gives $\|\mathcal{N}(\mu, \sigma^2) - \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)\|_1 \leq \varepsilon$, as required. □

Remark 4.20. *Note that in the above result, the samples are not "compressed" in the usual sense of this verb. Nevertheless, our formal definition of compression, Definition 4.9, allows $m = \tau$.*

To complete the proof of Theorem 4.3 in the realizable setting, we note that Lemma 4.18 combined with Lemma 4.14 implies that the class of axis-aligned Gaussians in $\mathbb{R}^d$ admits an

$$\big(\, O(d),\, O(d\log(d/\varepsilon)),\, O(\log(3d))\,\big)$$

non-robust compression scheme. (Note that any axis-aligned Gaussian is a product of one-dimensional Gaussians.) Then, by Lemma 4.16, the class of mixtures of k axis-aligned Gaussians admits an

$$\big(\, O(kd),\, O(kd\log(d/\varepsilon) + k\log(k/\varepsilon)),\, O(k\log(k)\log(d)/\varepsilon)\,\big)$$

non-robust compression scheme. Theorem 4.13 now implies that the class of k-mixtures of axis-aligned Gaussians in $\mathbb{R}^d$ can be learned using $\widetilde{O}(kd/\varepsilon^2)$ samples in the realizable setting.

4.4.2 Learning axis-aligned and general Gaussians in the agnostic setting

We now turn to the general case and prove an upper bound of $\widetilde{O}(kd^2/\varepsilon^2)$ for the sample complexity of learning mixtures of k Gaussians in d dimensions, and an upper bound of $\widetilde{O}(kd/\varepsilon^2)$ for the sample complexity of learning mixtures of k axis-aligned Gaussians, both in the agnostic sense. The heart of the proof is to show that Gaussians have robust compression schemes in any dimension.

Lemma 4.21. *For any positive integer d, the class $\mathcal{G}^d$ of d-dimensional Gaussians admits an*

$$\big(\ O(d\log(2d)),\ O(d^2\log(2d)\log(d/\varepsilon)),\ O(d\log(2d))\ \big)$$

2/3-robust compression scheme.

Remark 4.22. *The proof of Lemma 4.21 can be amended to give an r-robust compression schemes for any $r < 1$, which will change the constant 12 in the agnostic results of Theorem 4.1 and Theorem 4.3 to any constant larger than 9, at the expense of worse constants for τ, t and m. This is straightforward but creates additional cumbersome notation, hence we omit the details.*

Before proving Lemma 4.21, we show how can it be combined with the previous lemmata to prove our main upper bounds.

Proof of Theorem 4.1. Combining Lemma 4.21 with Lemma 4.17 shows that the class of k-mixtures of d-dimensional Gaussians is 12-agnostically learnable with sample complexity $\widetilde{O}(kd^2/\varepsilon^2)$. □

Proof of Theorem 4.3. Recall that $\mathcal{G}^d$ denote the class of d-dimensional Gaussian distributions. Applying Lemma 4.21 with $d = 1$, Lemma 4.14 shows that $\mathcal{G}^d$ admits $\big(O(d), O(d\log(d/\varepsilon)), O(\log(3d))\big)$ 2/3-robust compression. Lemma 4.17 then implies that the class k-mix$(\mathcal{G}^d)$ is 12-agnostically learnable with sample complexity $\widetilde{O}(kd/\varepsilon^2)$, completing the proof. □

4.4.3 Proof of Lemma 4.21

We first provide a high-level overview of the proof. For simplicity, let us assume that we would like to encode the distribution $\mathcal{N}(0,\Sigma)$, where $\Sigma \in \mathbb{R}^{d\times d}$ has rank d. Let $v_1,\dots,v_d$ be an orthogonal set of vectors which satisfy $\Sigma = \sum_{i=1}^d v_i v_i^\top$ (note that the vectors v_i are not normalized to have norm 1). Now let $g_1,\dots,g_d$ be a collection of d samples from $\mathcal{N}(0,\Sigma)$. As $\mathrm{span}\{g_1,\dots,g_d\} = \mathbb{R}^d$ with probability 1, a natural idea is the following: for each i, find real coefficients $\lambda_{i,1},\dots,\lambda_{i,d}$ such that $v_i = \sum_{j=1}^d \lambda_{i,j} g_j$. We could then have the encoder send $g_1,\dots,g_d$ and (a discretization of) the set $\{\lambda_{i,j}\}_{i,j\in[d]}$; the decoder would then be able to recover an approximation of Σ. If the discretization of each $\lambda_{i,j}$ can be accomplished with at most b bits each, then this would give a compression scheme where the encoder draws d points and sends d points along with $d^2 b$ bits to the decoder.

However, there is a difficulty: we must control the bit complexity of a suitable discretization of $\lambda_{i,j}$—to achieve the optimal sample complexity bound, the bit complexity must be bounded by $\mathrm{polylog}(d, 1/\varepsilon)$. The key to achieving a suitable discretization is the following fact from geometric functional analysis (Lemma 4.23, cf. [105, Corollary 4.1]): given a set $T = \{g_1,\dots,g_m\}$ of $m = O(d)$ i.i.d. samples from $\mathcal{N}(0,\Sigma)$, we have $\frac{1}{20}\cdot\mathcal{E} \subseteq \mathrm{conv}(T)$, where $\mathcal{E}$ is the ellipsoid whose principle axes are $v_1,\dots,v_d$. This enables us to express each v_i as $\sum_{j=1}^m \lambda_{i,j} g_j$, where each $\lambda_{i,j}$ is guaranteed to lie in a bounded interval; we can then discretize by building an ε-net of size $\mathrm{poly}(d, 1/\varepsilon)$ on this interval. The bit complexity would then be $\mathrm{polylog}(d, 1/\varepsilon)$, as desired.

Now suppose we would like to encode the distribution $\mathcal{N}(\mu,\Sigma)$ (again, assuming Σ has full-rank). Note that if $g_1, g_2 \sim \mathcal{N}(\mu,\Sigma)$ then $\frac{g_1-g_2}{\sqrt{2}} \sim \mathcal{N}(0,\Sigma)$, and thus we can reduce to the compression scheme

idea discussed above. To encode μ, the idea is that a *single* sample $g \sim \mathcal{N}(0, \Sigma)$ is unlikely to be too far from μ. More specifically, if $\mathcal{E}$ is the ellipsoid defined by Σ centered at 0, then, with very high probability, $\mu \in g + O(\sqrt{d}) \cdot \mathcal{E}$. Thus, we can essentially build an ε-net of the set $g + O(\sqrt{d}) \cdot \mathcal{E}$ and the encoder can send g as well as the identity of the point in the ε-net closest to μ.

We now proceed to the formal proof. The goal is to obtain a 2/3-robust compression scheme for $\mathcal{G}^d$. Accordingly, we consider any target distribution Q for which there exists a Gaussian $\mathcal{N}(\mu, \Sigma)$ satisfying $\|Q - \mathcal{N}(\mu, \Sigma)\|_1 \leq 2/3$. Recall, from Fact 3.4, that this implies that $d_{\mathrm{TV}}(Q, \mathcal{N}(\mu, \Sigma)) \leq 1/3$.

We may assume that Σ has full rank, since there is a reduction from the case of rank-deficient Σ. If the rank of Σ is $\rho < d$, then any $X \sim \mathcal{N}(\mu, \Sigma)$ lies in some affine subspace $\mathcal{S}$ of dimension ρ. Thus, by Fact 3.4, any $X \sim Q$ lies in $\mathcal{S}$ with probability at least 2/3. With high probability, after seeing $10d$ samples from Q, at least $\rho + 1$ points from $\mathcal{S}$ will appear in the sample. We encode $\mathcal{S}$ using these samples, and for the rest of the process we work in this affine subspace, and discard outside points.

Definition of $v_1, \ldots, v_d, \Psi$. Since Σ has full rank, we may find an orthogonal set of vectors $v_1, \ldots, v_d$ satisfying $\Sigma = \sum_{i=1}^{d} v_i v_i^\top$. For convenience, let $\Psi = \Sigma^{1/2}$ be the unique positive definite square root of Σ. Observe that

$$\Psi \;=\; \sum_{i=1}^{d} \frac{v_i v_i^\top}{\|v_i\|}, \qquad \Sigma^{-1} \;=\; \sum_{i=1}^{d} \frac{v_i v_i^\top}{\|v_i\|^4}, \qquad \text{and} \qquad \Psi^{-1} \;=\; \sum_{i=1}^{d} \frac{v_i v_i^\top}{\|v_i\|^3}. \tag{4.6}$$

We first prove a lemma that is similar to known results in random matrix theory [104, Corollary 4.1], but is tailored for our purposes. The notation $a \cdot B_2^d$ denotes $\{\, y \in \mathbb{R}^d \,:\, \|y\| \leq a \,\}$. The convex hull of a set T is denoted by $\mathrm{conv}(T)$.

Lemma 4.23. *Let $q_1, \ldots, q_m \in \mathbb{R}^d$ be i.i.d. samples from a distribution Q where $d_{\mathrm{TV}}(Q, \mathcal{N}(0, I_d)) \leq 2/3$. Let*

$$T \;\coloneqq\; \{\, \pm q_i \;:\; \|q_i\| \leq 4\sqrt{d} \,\}.$$

Then for a large enough absolute constant C, if $m \geq Cd(1 + \log d)$ then

$$\mathbb{P}\left[\, \frac{1}{20} B_2^d \not\subseteq \mathrm{conv}(T) \,\right] \;\leq\; 1/6.$$

Proof. Let $S^{d-1} \coloneqq \{\, y \in \mathbb{R}^d \,:\, \|y\| = 1 \,\}$. Consider the following statement:

$$\max_{q \in T} |\langle y, q \rangle| \;\geq\; \frac{1}{20} \qquad \forall y \in S^{d-1}. \tag{4.7}$$

We first show that (4.7) implies that $\frac{1}{20} B_2^d \subseteq \mathrm{conv}(T)$, which is the event that we wish to analyze. Let $P \coloneqq \mathrm{conv}(T)$. Its polar is $P^\circ = \{\, y \in \mathbb{R}^d \,:\, |\langle y, q \rangle| \leq 1 \;\forall q \in T \,\}$. So (4.7) implies $P^\circ \subseteq 20 B_2^d$. As polarity reverses containment and the polar of B_2^d is itself, we obtain $P \supseteq (20 B_2^d)^\circ = (\frac{1}{20}) B_2^d$.

We now bound the probability that (4.7) fails using an ε-net argument. For this, fix some $y \in S^{d-1}$ and let $g \sim \mathcal{N}(0, I_d)$ and let $X_y \coloneqq \langle y, g \rangle$. Notice that $X_y \sim \mathcal{N}(0, 1)$. Since the pdf of X_y is bounded above by $\frac{1}{\sqrt{2\pi}} < 1$, we have $\mathbb{P}\left[|X_y| \leq \frac{1}{10}\right] \leq 1/5$. Moreover, by Lemma B.13, $\mathbb{P}\left[\|g\|_2 > 4\sqrt{d}\right] \leq$

$\exp(-3)$. Hence by the union bound,

$$\mathbb{P}\left[|X_y| \le \frac{1}{10} \vee \|g\|_2 > 4\sqrt{d}\right] \le 1/5 + \exp(-3) < 0.25.$$

Now, let

$$Y_{y,i} \coloneqq \langle y, q_i \rangle \qquad \forall y \in S^{d-1},\, i \in [m],$$

and let $E_{y,i}$ be the event $\{|Y_{y,i}| \le \frac{1}{10} \vee \|q_i\| > 4\sqrt{d}\}$. As $d_{\mathrm{TV}}\left(Q, \mathcal{N}(0, I_d)\right) \le 2/3$, we have $\mathbb{P}\left[[]\, E_{y,i}\right] \le 0.25 + 2/3 < 0.92$. Thus

$$\mathbb{P}\left[\bigwedge_{i \in [m]} E_{y,i}\right] < (0.92)^m.$$

Let N be an $(1/80\sqrt{d})$-net of S^{d-1} in ℓ_2 with $|N| \le (240\sqrt{d})^d$ (see Lemma B.16). By the union bound, since $m \ge Cd(1 + \log d)$ for C large enough, with probability at least $1 - (240\sqrt{d})^d (0.92)^m \ge 5/6$, for all $y \in N$ there exists $i \in [m]$ such that $|Y_{y,i}| \ge \frac{1}{10}$ and $\|q_i\| \le 4\sqrt{d}$.

To complete the proof, we suppose that this event holds, and show that (4.7) also holds. Consider any $y \in S^{d-1}$, and let $y' \in N$ satisfy $\|y - y'\|_2 \le 1/80\sqrt{d}$. Let q_i be such that $\|q_i\| \le 4\sqrt{d}$ and $|Y_{y',i}| \ge \frac{1}{10}$. These imply that $\pm q_i \in T$ and that

$$|Y_{y,i}| \;\ge\; \left|Y_{y',i}\right| - \frac{|q_i|}{80\sqrt{d}} \;\ge\; \frac{1}{10} - \frac{1}{20} \;=\; \frac{1}{20}.$$

Thus $|\langle y, q_i \rangle| \ge 1/20$, as required. □

We next show how to encode the mean and the eigenvectors.

Lemma 4.24. *Let C be a sufficiently large absolute constant. Suppose that S contains $2m = 2Cd(1 + \log d)$ samples drawn from Q, where $d_{\mathrm{TV}}\left(Q, \mathcal{N}(\mu, \Sigma)\right) \le 1/3$. Then, with probability at least $2/3$, one can encode vectors $\hat{v}, \ldots, \hat{v}_d, \hat{\mu} \in \mathbb{R}^d$ satisfying*

$$\|\Psi^{-1}(\hat{v}_j - v_j)\| \;\le\; \varepsilon/24d^2 \qquad \forall j \in [d], \tag{4.8}$$

$$\|\Psi^{-1}(\hat{\mu} - \mu)\| \;\le\; \varepsilon/2, \tag{4.9}$$

using $O(d^2 \log(2d) \log(2d/\varepsilon))$ bits and the points in S.

Proof. The samples in S will be denoted $X_1, \ldots, X_{2m}$.

Encoding $\widehat{v}_j$. We define "normalized" samples

$$Y_i \;\coloneqq\; \frac{1}{\sqrt{2}} \Psi^{-1}(X_{2i} - X_{2i-1}) \qquad \forall i \in [m].$$

If we were in the non-robust case, then X_{2i} and X_{2i-1} would both have distribution $\mathcal{N}(\mu, \Sigma)$, so Y_i would have distribution $\mathcal{N}(0, I)$. Instead, both X_{2i} and X_{2i-1} have TV distance at most $1/3$ from $\mathcal{N}(\mu, \Sigma)$. It follows that Y_i has TV distance at most $2/3$ from $\mathcal{N}(0, I)$. (This may be seen, for example,

by the coupling definition of TV distance; see [94, Eq. (18.10)].) Define the event

$$\mathcal{E} \coloneqq \left\{ \frac{1}{C} B_2^d \subseteq \operatorname{conv}\{\pm Y_i : i \in \mathcal{I}\} \right\} \quad \text{where} \quad \mathcal{I} \coloneqq \{i \in [m] : \|Y_i\| \le 4\sqrt{d}\}.$$

Since C is large, and in particular $C \ge 20$, by Lemma 4.23 we have $\mathbb{P}[\mathcal{E}] \ge 5/6$. Our encoding will assume that the event $\mathcal{E}$ occurs.

Fix some $j \in [d]$. Referring to (4.6), we see that $\Psi^{-1} v_j = v_j / \|v_j\|$ has unit norm. Since $\mathcal{E}$ occurs, we can write

$$\frac{\Psi^{-1} v_j}{C} = \sum_{i \in [m]} \theta_{j,i} Y_i$$

for some vector $\theta_j \in [-1,1]^m$ supported on $\mathcal{I}$. Applying Ψ to both sides, we obtain

$$v_j = \frac{C}{\sqrt{2}} \sum_{i \in \mathcal{I}} \theta_{j,i} (X_{2i} - X_{2i-1}).$$

Consider the natural $(\varepsilon/96Cmd^3)$-net for $[-1,1]^m$ in the ℓ_∞ norm, formed by the Cartesian product of 1-dimensional nets (see Lemma B.17). This net has size at most $(96Cmd^3/\varepsilon)^m$. Recalling that $m = O(d(1+\log d))$, it follows that any element of the net can be described using $O(m \log(2d/\varepsilon))$ bits. Let $\widehat{\theta}_j$ be an element in the net that is closest to θ_j. Since each θ_j is supported on $\mathcal{I}$, and the net has the Cartesian product structure, we may choose $\widehat{\theta}_j$ also to be supported on $\mathcal{I}$. Define

$$\hat{v}_j \coloneqq \frac{C}{\sqrt{2}} \sum_{i \in \mathcal{I}} \widehat{\theta}_{j,i} (X_{2i} - X_{2i-1}).$$

The error of this encoding is

$$\begin{aligned} \left\| \Psi^{-1}(\hat{v}_j - v_j) \right\| &= \frac{C}{\sqrt{2}} \left\| \sum_{i \in \mathcal{I}} (\theta_{j,i} - \widehat{\theta}_{j,i}) \Psi^{-1} (X_{2i} - X_{2i-1}) \right\| \\ &\le \frac{C}{\sqrt{2}} |\mathcal{I}| \big(\max_{i \in \mathcal{I}} |\theta_{j,i} - \widehat{\theta}_{j,i}| \big) \big(\max_{i \in \mathcal{I}} \sqrt{2} \|Y_i\| \big) \end{aligned}$$

By definition of $\widehat{\theta}_j$, we have $\left\| \widehat{\theta}_j - \theta_j \right\|_\infty \le \varepsilon/96Cmd^3$. By definition of $\mathcal{I}$, we have $\|Y_i\| \le 4\sqrt{d}$, leading to the bound

$$\left\| \Psi^{-1}(\hat{v}_j - v_j) \right\| \le \frac{C}{\sqrt{2}} m \Big(\frac{\varepsilon}{96Cmd^3} \Big) \big(4\sqrt{2}\sqrt{d} \big) \le \frac{\varepsilon}{24d^2}, \tag{4.10}$$

establishing (4.8). The vectors $\widehat{v}_1, \dots, \widehat{v}_d$ are encoded simply using $\widehat{\theta}_1, \dots, \widehat{\theta}_d$. Each $\widehat{\theta}_i$ requires $O(m \log(2d/\epsilon))$ bits. Recall that $m = O(d \log(2d))$. Hence, the total number of bits required is $O(d^2 \log(2d) \log(2d/\varepsilon))$.

Encoding $\hat{\mu}$. Let $Z_i \coloneqq \Psi^{-1}(X_i - \mu)$ and observe that Z_i has a distribution with TV distance at most $1/3$ to $\mathcal{N}(0, I)$. Define the event

$$\mathcal{E}' \coloneqq \{\min\{\|Z_1\|, \|Z_2\|\} \le 4\sqrt{d}\}.$$

Lemma B.13 implies that

$$\mathbb{P}[\|Z_i\| \ge 4\sqrt{d}] \le \exp(-3) + 1/3 < \sqrt{1/6}.$$

Thus $\mathbb{P}[\mathcal{E}'] \ge 5/6$. Our encoding will assume that the event $\mathcal{E}'$ occurs.

By symmetry assume $\|Z_1\| \le 4\sqrt{d}$, and suppose $Z_1 = \sum_{j\in[d]} \lambda_j v_j / \|v_j\|$. Thus $\sum \lambda_j^2 \le 16d^2$. Furthermore, from the definitions of Z_1 and Ψ we have

$$\mu = X_1 - \Psi Z_1 = X_1 - \sum_{j\in[d]} \lambda_j v_j.$$

Consider an $(\varepsilon/3d)$-net for $4\sqrt{d}B_2^d$ of size $O(d^{1.5}/\varepsilon)^d$ (see Lemma B.16). Observe that $\lambda \in 4\sqrt{d}B_2^d$, and let $\widehat{\lambda}$ be the closest element to λ in this net. The encoding is

$$\hat{\mu} \coloneqq X_1 - \sum_{j\in[d]} \widehat{\lambda}_j \widehat{v}_j.$$

The error of this encoding is

$$\begin{aligned}
\left\|\Psi^{-1}(\mu - \hat{\mu})\right\| &= \left\|\sum_{j\in[d]} \Psi^{-1}(\lambda_j v_j - \widehat{\lambda}_j \widehat{v}_j)\right\| \\
&\le \sum_{j\in[d]} \left\|\widehat{\lambda}_j(\Psi^{-1}v_j - \Psi^{-1}\widehat{v}_j) + (\lambda_j - \widehat{\lambda}_j)\Psi^{-1}v_j\right\| \\
&\le d \cdot \max_{j\in[d]} \left\{\left|\widehat{\lambda}_j\right| \cdot \left\|\Psi^{-1}v_j - \Psi^{-1}\widehat{v}_j\right\| + \left|\lambda_j - \widehat{\lambda}_j\right| \cdot \left\|\Psi^{-1}v_j\right\|\right\}.
\end{aligned}$$

By definition of $\widehat{\lambda}$, we have $\left\|\widehat{\lambda}\right\|_\infty \le 4\sqrt{d}$ and $\left\|\lambda - \widehat{\lambda}\right\|_\infty \le \varepsilon/3d$. From (4.6) we have $\left\|\Psi^{-1}v_j\right\| \le 1$. Lastly, using (4.10) we have $\left\|\Psi^{-1}(\hat{v}_j - v_j)\right\| \le \varepsilon/24d^2$, leading to the bound

$$\left\|\Psi^{-1}(\mu - \hat{\mu})\right\| \le d \cdot \left(4\sqrt{d} \cdot \frac{\varepsilon}{24d^2} + \frac{\varepsilon}{3d} \cdot 1\right) \le \varepsilon/2,$$

establishing (4.9). The encoding for $\hat{\mu}$ consists only of $\widehat{\lambda}$. Since $\widehat{\lambda}$ comes from a net of size $O(d^{1.5}/\varepsilon)^d$, the number of bits required for the encoding is $O(d\log(d/\varepsilon))$.

All encodings will succeed so long as both $\mathcal{E}$ and $\mathcal{E}'$ occur, which happens with probability at least $2/3$. □

Lemma 4.21 now follows immediately from the following lemma.

Lemma 4.25. *Suppose that the vectors $\hat{v}_1, \dots, \hat{v}_d, \hat{\mu} \in \mathbb{R}^d$ satisfy*

$$\|\Psi^{-1}(\hat{v}_j - v_j)\| \;\le\; \rho \le 1/6d \qquad \forall j \in [d] \tag{4.11}$$

$$\|\Psi^{-1}(\hat{\mu} - \mu)\| \;\le\; \zeta. \tag{4.12}$$

Then

$$d_{\mathrm{TV}}\left(\mathcal{N}\big(\mu, \sum_{i\in[d]} v_i v_i^\top\big), \mathcal{N}\big(\hat{\mu}, \sum_{i\in[d]} \hat{v}_i \hat{v}_i^\top\big)\right) \;\le\; \frac{\sqrt{9d^3\rho^2 + \zeta^2}}{2}.$$

Proof. In this proof, we will use the log-det divergence, which is defined in Definition B.1. Define $\widehat{\Sigma} \coloneqq \sum_i \hat{v}_i \hat{v}_i^\top$. We will show that

$$\mathrm{LD}\left(\widehat{\Sigma}, \Sigma\right) \;\le\; 9d^3\rho^2. \tag{4.13}$$

If this is true, then Lemma B.5 and (4.12) yield

$$d_{\mathrm{TV}}\left(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\hat{\mu},\widehat{\Sigma})\right)^2 \;\le\; \frac{1}{4}\Big(\mathrm{LD}\left(\widehat{\Sigma}, \Sigma\right) + (\mu-\hat{\mu})^\top \Sigma^{-1}(\mu-\hat{\mu})\Big) \;\le\; \frac{1}{4}(9d^3\rho^2 + \zeta^2),$$

which completes the proof of the lemma.

Thus, we focus on (4.13). Recalling from (4.6) that $\Psi = \Sigma^{1/2}$, from Claim B.2 we have

$$\begin{aligned} \mathrm{LD}\left(\widehat{\Sigma}, \Sigma\right) \;&=\; \mathrm{LD}\left(\Psi^{-1}\widehat{\Sigma}\Psi^{-1}, \Psi^{-1}\Sigma\Psi^{-1}\right) \;=\; \mathrm{LD}\left(B, I\right) \\ \text{where} \quad B \;&\coloneqq\; \sum_{i=1}^d \Psi^{-1}\hat{v}_i\hat{v}_i^\top\Psi^{-1}. \end{aligned}$$

We will show that $\|B - I\| \le 3d\rho$, or equivalently $-3d\rho I \preccurlyeq B - I \preccurlyeq 3d\rho I$. Then Lemma B.4 will imply that $\mathrm{LD}\left(\widehat{\Sigma}, \Sigma\right) = \mathrm{LD}\left(B, I\right) \le 9d^3\rho^2$, which establishes (4.13).

To complete the proof, we have

$$\begin{aligned} \|B - I\| \;&=\; \left\| \sum_{i=1}^d (\Psi^{-1}\hat{v}_i\hat{v}_i^\top\Psi^{-1} - \Psi^{-1}v_i v_i^\top\Psi^{-1}) \right\| \\ &\le\; \sum_{i=1}^d \left\| \Psi^{-1}\hat{v}_i\hat{v}_i^\top\Psi^{-1} - \Psi^{-1}v_i v_i^\top\Psi^{-1} \right\| \\ &=\; \sum_{i=1}^d \|x_i x_i^\top - y_i y_i^\top\|, \end{aligned}$$

with $x_i \coloneqq \Psi^{-1}\hat{v}_i$ and $y_i \coloneqq \Psi^{-1}v_i$. Referring to (4.6) we see that $\|y_i\| = \|\Psi^{-1}v_i\| = 1$. By the lemma's hypothesis, $\|x_i - y_i\| \le \rho$. By applying the following simple lemma, we conclude that $\|B - I\| \le 3d\rho$. □

Lemma 4.26. *Suppose x, y satisfy $\|y\| = 1$ and $\|x - y\| \le \varepsilon \le 1$. Then we have $\|xx^\top - yy^\top\| \le 3\varepsilon$.*

Proof. Suppose $x = y + z$ with $\|z\| \le \varepsilon$. Then,

$$\|xx^\top - yy^\top\| = \|yz^\top + zy^\top + zz^\top\| \le \|yz^\top\| + \|zy^\top\| + \|zz^\top\| \le \varepsilon + \varepsilon + \varepsilon^2 \le 3\varepsilon,$$

where we have used the facts that $\|AB\| \le \|A\| \cdot \|B\|$ for any two size-compatible matrices A and B, and that for any column or row vector v, the operator norm of v as a matrix coincides with its Euclidean norm as a vector. □

4.5 The lower bound for Gaussians and their mixtures

In this section, we establish a lower bound of $\widetilde{\Omega}(d^2/\varepsilon^2)$ for learning a single Gaussian, and then lift it to obtain a lower bound of $\widetilde{\Omega}(kd^2/\varepsilon^2)$ for learning mixtures of k Gaussians in d dimensions. The lower bound holds for the realizable setting, and therefore also holds in the agnostic setting.

The high-level strategy for our lower bound follows a strategy adopted in earlier work for mixtures of spherical Gaussians [134]. The idea is to create a large number of distributions that are pairwise close in KL divergence (roughly ε^2) but pairwise far in TV distance (roughly ε). An application of the following lemma will then yield the desired sample complexity bound.

Lemma 4.27. *Let $\kappa\colon \mathbb{R} \to \mathbb{R}$ be a function and let $\mathcal{F}$ be a class of distributions such that, for all small enough $\varepsilon > 0$, there exist distributions $f_1, \dots, f_M \in \mathcal{F}$ with*

$$D_{\mathrm{KL}}(f_i \parallel f_j) \le \kappa(\varepsilon) \quad \text{and} \quad d_{\mathrm{TV}}(f_i, f_j) > 2\varepsilon \qquad \forall i \ne j \in [M].$$

Then any method that learns $\mathcal{F}$ to within total variation distance ε with success probability at least $2/3$ has sample complexity $\Omega\left(\frac{\log M}{\kappa(\varepsilon)\log(1/\varepsilon)}\right)$.

The preceding lemma is a straightforward consequence of the following result, which is a generalized form of Fano's inequality. It may be found in [145, Lemma 3].

Lemma 4.28 (Generalized Fano inequality)**.** *Let the distributions $f_1, \dots, f_M$ satisfy*

$$D_{\mathrm{KL}}(f_i \parallel f_j) \le \beta \quad \text{and} \quad \|f_i - f_j\|_1 > \alpha \qquad \forall i \ne j \in [M].$$

Consider any density estimation method that has an explicit description of $f_1, \dots, f_M$, receives n i.i.d. samples from some f_i without knowing i, then outputs an estimate $\widehat{f}$ for f_i. For each i, define $e_i \coloneqq \mathbf{E}\|f_i - \widehat{f}\|_1$ for the case in which the method receives samples from f_i. Then

$$\max_i e_i \ge \alpha \frac{\log M - n\beta + \log 2}{2 \log M}.$$

Proof of Lemma 4.27. Consider a distribution learning method for learning $\mathcal{F}$ with sample complexity $m(\varepsilon)$, and consider M distributions $f_1, \dots, f_M$ satisfying the hypotheses. The method will receive samples from f_j, where $j \in [M]$ is unknown. We will amplify its success probability by running it k times, then apply the generalized Fano inequality.

Perform a sequence of k trials as follows. In each trial, the method receives $m(\varepsilon)$ samples from (the same) f_j. The trial is a success if the method outputs some density g whose TV distance from f_j is at most ε. Since the method's sample complexity is $m(\varepsilon)$, each trial is a success with probability at least $2/3$. After performing the k trials, there have been k densities $g_1, \dots, g_k$ produced as output. If some f_i is within TV distance ε from at least $k/2$ of these outputs, then this f_i is used as the overall output $\widehat{f}$ of this amplified method; otherwise, $\widehat{f} = f_1$ is the overall output.

Let $\mathcal{E}$ be the event that at least $k/2$ of the trials were a success. By a standard Chernoff bound, $\mathbb{P}[\mathcal{E}] \geq 1 - \exp(-\Omega(k))$. When event $\mathcal{E}$ occurs, then at least $k/2$ of $g_1, \dots, g_k$ have TV distance at most ε from the target density f_j, so the overall output must be $\widehat{f} = f_j$, so $\|f_j - \widehat{f}\|_1 = 0$. Thus, the expected error is

$$e_j \;=\; \mathbf{E}\|f_j - \widehat{f}\|_1 \;\leq\; \mathbb{P}[\mathcal{E}^c] \cdot 2 \;\leq\; \exp(-\Omega(k)) \qquad \forall j \in [M].$$

The total number of samples is $n = km(\varepsilon)$, so Lemma 4.28 gives

$$\alpha \frac{\log M - (km(\varepsilon))\kappa(\varepsilon) + \log 2}{2 \log M} \;\leq\; \exp(-\Omega(k)).$$

Choose $k = \Theta(\log(1/\varepsilon))$ to be sufficiently large. Rearranging gives $m(\varepsilon) = \Omega(\log M / \kappa(\varepsilon) \log(1/\varepsilon))$, as required. □

Our main lower bound for learning a single Gaussian is the following result.

Theorem 4.29. *Any algorithm that learns the class of d-dimensional Gaussians in $\mathbb{R}^d$ in the realizable setting within total variation distance ε and with success probability $2/3$ has sample complexity $\Omega\left(\frac{d^2}{\varepsilon^2 \log(1/\varepsilon)}\right)$.*

Proof. In order to apply Lemma 4.27, we must create a large number M of Gaussian distributions whose pairwise KL divergence is at most κ, and whose pairwise TV distance is at least 2ε. We will accomplish this with parameters $M = 2^{\Omega(d^2)}$ and $\kappa = O(\varepsilon^2)$, so Lemma 4.27 will yield the desired lower bound.

The existence of these M distributions will be shown using the probabilistic method. Specifically, let us fix parameters $r = 9$ and $\lambda = \Theta(\varepsilon d^{-1/2})$. For each $a \in [M]$, we pick U_a to be a random matrix of size $d \times d/r$ with orthonormal columns (the columns of U_a are chosen to be the first d/r vectors of a uniformly random orthonormal basis of $\mathbb{R}^d$.). From this, we create the distribution

$$f_a \;\coloneqq\; \mathcal{N}(0, \Sigma_a) \qquad \text{where} \qquad \Sigma_a = I_d + \lambda U_a U_a^\top \qquad \forall a \in [M].$$

To apply Lemma 4.27, we must analyze the pairwise KL divergences and TV distances between $f_1, \dots, f_M$.

Bound on KL divergences. This analysis is straightforward since there is a closed-form expression for the KL divergence between any two Gaussians. First, observe that any two Σ_a and Σ_b have the same

spectrum: there are d/r eigenvalues equal to $1+\lambda$ and the remaining eigenvalues equal 1. Consequently,

$$\log\det(\Sigma_b\Sigma_a^{-1}) \;=\; \log(\det\Sigma_b\cdot\det\Sigma_a^{-1}) \;=\; 0. \tag{4.14}$$

Next observe that

$$\Sigma_a^{-1} \;=\; I - \frac{\lambda}{1+\lambda}U_aU_a^\top; \tag{4.15}$$

this may be verified simply by multiplying by Σ_a. Thus

$$\begin{aligned}
2\cdot D_{\mathrm{KL}}(f_a \parallel f_b) &= \mathrm{Tr}(\Sigma_a^{-1}\Sigma_b - I) \qquad \text{(by (4.14) and Lemma B.3)}\\
&= \mathrm{Tr}\left(\left(I - \frac{\lambda}{1+\lambda}U_aU_a^\top\right)(I+\lambda U_bU_b^\top) - I\right) \qquad \text{(by (4.15))}\\
&= \mathrm{Tr}\left(\lambda U_bU_b^\top - \frac{\lambda}{1+\lambda}U_aU_a^\top - \frac{\lambda^2}{1+\lambda}U_aU_a^\top U_bU_b^\top\right) \qquad (4.16)\\
&= \lambda\cdot\frac{d}{r} - \frac{\lambda}{1+\lambda}\cdot\frac{d}{r} - \frac{\lambda^2}{1+\lambda}\cdot\|U_a^\top U_b\|_F^2\\
&\le \frac{\lambda^2 d}{(1+\lambda)r} \;\le\; \frac{\lambda^2 d}{r} \;=\; O(\varepsilon^2).
\end{aligned}$$

This bound holds with probability 1.

Bound on TV distances. The remaining step is to show that $d_{\mathrm{TV}}(f_a, f_b) = \Omega(\varepsilon)$ for all $a \neq b$. Then, by scaling ε by a constant factor, we may apply Lemma 4.27 and complete the proof.

First we provide some intuition on why such an inequality should hold. Let S_a be the subspace spanned by the columns of U_a. One would expect that a vector drawn from $\mathcal{N}(0,\Sigma_a)$ should have a slightly larger projection onto S_a then a vector drawn from $\mathcal{N}(0,\Sigma_b)$. This would reveal an event that has slightly higher probability under the former distribution than under the latter. Recalling the definition of the TV distance as a supremum over events (see Fact 3.4), such an argument would give the desired lower bound on the TV distance.

Here we use a simpler argument, formulated as Lemma 4.31, which shows that a lower bound on $d_{\mathrm{TV}}(f_a, f_b)$ can be obtained if $\|U_a^\top U_b\|_F^2$ is small. This would hold if the columns of U_a are nearly pairwise orthogonal to the columns of U_b, which intuitively should hold since U_a and U_b are chosen randomly. This is formalized in Lemma 4.30 below, which shows that, with positive probability, $\|U_a^\top U_b\|_F^2 \le d/2r$ for all $a \neq b$. Then Lemma 4.31 implies that, for all $a \neq b$, $d_{\mathrm{TV}}(f_a, f_b) = \Omega\left(\min\{1, \lambda\sqrt{d/r}\}\right) = \Omega(\varepsilon)$, by our choice of parameters. □

The main technical lemma underlying our lower bound is Lemma 4.30.

Lemma 4.30. *Suppose $d \ge r \ge 9$. There exists $M = 2^{\Omega(d^2/r)}$ such that the following holds. Let the matrices U_a, for $a \in [M]$, be independently chosen with size $d \times d/r$ and with orthonormal columns. Then, with positive probability, we have $\|U_a^\top U_b\|_F^2 \le d/2r$ for all $a \neq b$.*

Proof. The columns of each matrix U_a are chosen to be the first d/r vectors of a uniformly random

orthonormal basis of $\mathbb{R}^d$. We will show that, for any two such matrices U_a and U_b, with probability $1 - 2^{-\Omega(d^2/r)}$ we have $\|U_a^\top U_b\|_F^2 \le d/2r$. The lemma then follows by a union bound.

Fix $a, b \in [M]$ with $a \neq b$. By rotational invariance, we may assume without loss of generality that $U_a = \left(\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right)$. Thus $\|U_a^\top U_b\|_F^2 \stackrel{d}{=} \|U_{d/r}\|_F^2$, where $U_{d/r}$ is a $d/r \times d/r$ principal submatrix of a uniformly random orthogonal matrix U. (Alternatively, the columns of $U_{d/r}$ are the first d/r coordinates of d/r orthonormal vectors in $\mathbb{R}^d$ chosen uniformly at random.) Hence, it suffices to show that $\|U_{d/r}\|_F^2 \le d/2r$ with probability at least $1 - 2^{-\Omega(d^2/r)}$. The main difficulty is that $U_{d/r}$ does not have independent entries, due to the orthonormality, but intuitively it should behave very similarly to a matrix with independent Gaussian entries.

Relating to a Gaussian matrix. The matrix U is naturally related to the Gaussian matrix $G \in \mathbb{R}^{d \times d/r}$ with i.i.d. $\mathcal{N}(0, 1/d)$ entries. Similarly, the matrix $U_{d/r}$ is naturally related to the Gaussian matrix $G_{d/r} \in \mathbb{R}^{d/r \times d/r}$ comprised of the first d/r rows of G. To see this, let $G = U_G \Sigma_G V_G^\top$ be the singular value decomposition of G, where $U_G \in \mathbb{R}^{d \times d/r}$ and $\Sigma_G, V_G \in \mathbb{R}^{d/r \times d/r}$. Observe that, by rotational invariance, the columns of U_G are d/r uniformly random orthonormal vectors, and therefore the top d/r rows of U_G (which we denote, slightly awkwardly, by $(U_G)_{d/r}$) have the same distribution as $U_{d/r}$. More precisely, since U_G is independent of Σ_G, V_G, we have

$$G_{d/r} \;=\; (U_G)_{d/r} \Sigma_G V_G^\top \;\stackrel{d}{=}\; U_{d/r} \Sigma_G V_G^\top. \tag{4.17}$$

Observe that $\mathbb{E}\|G_{d/r}\|_F^2 = (d/r)^2 \cdot (1/d) = d/r^2$, so it remains to show that $\|U_{d/r}\|_F^2$ is unlikely to exceed this by a factor $r/2$.

The Frobenius norms $\|G_{d/r}\|_F$ and $\|U_{d/r}\|_F$ can be related as follows. By (4.17),

$$\begin{aligned}
\|G_{d/r}\|_F &\stackrel{d}{=} \|U_{d/r} \Sigma_G V_G^\top\|_F \\
&= \sqrt{\mathrm{Tr}(U_{d/r} \Sigma_G V_G^\top \cdot V_G \Sigma_G U_{d/r}^\top)} \\
&= \sqrt{\mathrm{Tr}(U_{d/r} \Sigma_G \cdot \Sigma_G U_{d/r}^\top)} \qquad \text{(since } V_G \text{ is orthogonal)} \\
&= \|U_{d/r} \Sigma_G\|_F \\
&\ge \sigma_{\min}(\Sigma_G) \|U_{d/r}\|_F,
\end{aligned} \tag{4.18}$$

where $\sigma_{\min}(\Sigma_G)$ denotes the smallest singular value of Σ_G.

Moments of $\|U_{d/r}\|_F$. Intuitively, (4.18) should show that $\|U_{d/r}\|_F$ is unlikely to deviate significantly above $\mathbb{E}\|G_{d/r}\|_F$, since $\mathbb{E}\sigma_{\min}(\Sigma_G) \ge 1 - 1/\sqrt{r}$ (by Theorem B.12) and since $\|G_{d/r}\|_F^2$ concentrates sharply around its mean (as it is a sum of i.i.d. random variables). To make this precise, we will bound the (suitably modified) pth moment of (4.18), for any $p \ge 1$.

Since an upper bound on $\|U_{d/r}\|_F$ is desired, it will be convenient, and sufficient, to consider only the moments of positive deviations. To formalize this idea, recall the notation $(x)_+ \coloneqq \max\{0, x\}$, and observe that the map $x \mapsto (x)_+^p$ is monotone and convex on $\mathbb{R}$ for $p \ge 1$. The bound on the (modified)

moments proceeds as follows:

$$\begin{aligned}\mathbb{E}\big[(\|G_{d/r}\|_F - \sqrt{d}/r)_+^p\big] &\geq \mathbb{E}\big[(\sigma_{\min}(\Sigma_G)\cdot\|U_{d/r}\|_F - \sqrt{d}/r)_+^p\big] \qquad \text{(by (4.18) and monotonicity of } (\cdot)_+^p\text{)}\\ &= \mathbb{E}\Big[\,\mathbb{E}\big[\,(\sigma_{\min}(\Sigma_G)\cdot\|U_{d/r}\|_F - \sqrt{d}/r)_+^p \mid U_{d/r}\,\big]\,\Big]\end{aligned}$$

The next step uses Jensen's inequality for the conditional expectation $\mathbb{E}[\cdot|U_{d/r}]$, and convexity of $x \mapsto (x)_+^p$ to obtain

$$\begin{aligned}&\geq \mathbb{E}\Big[\,\big(\,\mathbb{E}[\,\sigma_{\min}(\Sigma_G)\cdot\|U_{d/r}\|_F - \sqrt{d}/r \mid U_{d/r}\,]\,\big)_+^p\,\Big]\\ &= \mathbb{E}\Big[\,\big(\,\mathbb{E}[\,\sigma_{\min}(\Sigma_G)\,]\cdot\|U_{d/r}\|_F - \sqrt{d}/r\,\big)_+^p\,\Big] \qquad \text{(independence of } \Sigma_G \text{ and } U_{d/r}\text{)}\\ &\geq \mathbb{E}\Big[\big((1-1/\sqrt{r})\cdot\|U_{d/r}\|_F - \sqrt{d}/r\big)_+^p\Big], \qquad (4.19)\end{aligned}$$

by monotonicity of $(\cdot)_+^p$ again, and by applying Theorem B.12 to the matrix $\sqrt{d}G$ (whose entries are i.i.d. $\mathcal{N}(0,1)$), which yields $\mathbb{E}\sigma_{\min}(\sqrt{d}G) \geq \sqrt{d} - \sqrt{d/r}$ and therefore $\mathbb{E}\sigma_{\min}(G) \geq 1 - 1/\sqrt{r}$.

High-probability bound on $\|U_{d/r}\|_F$. All that remains is the routine task of deriving a high-probability bound from moment bounds. Observe that $\|G_{d/r}\|_F \stackrel{d}{=} \|g\|_2/\sqrt{d}$, where $g \sim \mathcal{N}(0, I_{(d/r)^2})$. Lemma B.8 states that $(\|g\|_2 - d/r)_+$ is $O(1)$-subgaussian; by scaling, $(\|G_{d/r}\|_F - \sqrt{d}/r)_+$ is $O(1/\sqrt{d})$-subgaussian. Since the property of being $O(\sigma)$-subgaussian can be characterized via moments (see Lemma B.9), and since inequality (4.19) shows that the pth moments of $((1-1/\sqrt{r})\cdot\|U_{d/r}\|_F - \sqrt{d}/r)_+$ are bounded by the moments of $(\|G_{d/r}\|_F - \sqrt{d}/r)_+$, for all $p \geq 1$, we conclude that $((1-1/\sqrt{r})\cdot\|U_{d/r}\|_F - \sqrt{d}/r)_+$ is also $O(1/\sqrt{d})$-subgaussian. This allows us to bound the right tail of $(1-1/\sqrt{r})\cdot\|U_{d/r}\|_F - \sqrt{d}/r$ (but not the left tail, due to the $(\cdot)_+$). Recalling the definition of a subgaussian random variable (Definition B.7), we have

$$\mathbb{P}\Big[\,(1-\sqrt{1/r})\cdot\|U_{d/r}\|_F - \sqrt{d}/r \;\leq\; t\,\Big] \;\geq\; 1 - 2^{-\Omega(t^2 d)} \qquad \forall t > 0. \qquad (4.20)$$

Fix $t = \sqrt{d}/(12\sqrt{r})$. By simple manipulations, the event in (4.20) is equivalent to

$$\|U_{d/r}\|_F^2 \;\leq\; \frac{d}{r}\cdot\left(\frac{\frac{1}{\sqrt{r}} + \frac{1}{12}}{1 - \frac{1}{\sqrt{r}}}\right)^2.$$

This right-hand side is at most $d/2r$ for all $r \geq 9$. It follows that $\|U_{d/r}\|_F^2 \leq d/2r$ with probability at least $1 - 2^{-\Omega(d^2/r)}$, completing the proof. □

Lemma 4.31. *Suppose that $\lambda \leq 1/4$. If $\|U_a^\top U_b\|_F^2 \leq d/2r$, then $d_{\mathrm{TV}}(f_a, f_b) = \Omega\left(\min\{1, \lambda\sqrt{d/r}\}\right)$.*

The proof will make use of the following fact.

Fact 4.32 ([84, Fact 7(c) in Section 24.4])**.** *Let A, B be size-compatible matrices. Then*

$$\max\{\sigma_{\min}(A)\|B\|_F, \sigma_{\min}(B)\|A\|_F\} \leq \|AB\|_F \leq \min\{\sigma_{\max}(A)\|B\|_F, \sigma_{\max}(B)\|A\|_F\}. \qquad (4.21)$$

Proof of Lemma 4.31. The proof relies on the following approximate characterization of the TV distance between two zero-mean Gaussians. For any two symmetric positive definite matrices Σ_a and Σ_b of the same size,

$$d_{\mathrm{TV}}\left(\mathcal{N}(0,\Sigma_a),\mathcal{N}(0,\Sigma_b)\right) \;=\; \Theta\left(\min\{1, \|\Sigma_a^{-1/2}\Sigma_b\Sigma_a^{-1/2} - I\|_F\}\right).$$

This result appears in [55, Theorem 1.1]; see also [17, Corollary 2]. Hence to complete the proof it suffices to show that $\|\Sigma_a^{-1/2}\Sigma_b\Sigma_a^{-1/2} - I\|_F \geq \frac{4}{5}\lambda\sqrt{d/r}$. Observe that $\Sigma_a^{-1/2}\Sigma_b\Sigma_a^{-1/2} - I = \Sigma_a^{-1/2}(\Sigma_b - \Sigma_a)\Sigma_a^{-1/2}$. Applying the left inequality in (4.21) twice gives

$$\|\Sigma_a^{-1/2}\Sigma_b\Sigma_a^{-1/2} - I\|_F \geq \sigma_{\min}(\Sigma_a^{-1/2})^2\|\Sigma_b - \Sigma_a\|_F.$$

Recall that the eigenvalues of Σ_a are 1 and $1+\lambda$, hence $\sigma_{\min}(\Sigma_a^{-1/2}) = (1+\lambda)^{-1/2} \geq \sqrt{4/5}$ since $\lambda \leq 1/4$. Moreover, $\|\Sigma_b - \Sigma_a\|_F = \lambda\|U_bU_b^\top - U_aU_a^\top\|_F$, and since $U_bU_b^\top - U_aU_a^\top$ is symmetric, we have

$$\begin{aligned}
\|U_bU_b^\top - U_aU_a^\top\|_F^2 &= \mathrm{Tr}((U_bU_b^\top - U_aU_a^\top)(U_bU_b^\top - U_aU_a^\top)) \\
&= \mathrm{Tr}(U_bU_b^\top U_bU_b^\top) + \mathrm{Tr}(U_aU_a^\top U_aU_a^\top) - \mathrm{Tr}(U_bU_b^\top U_aU_a^\top) - \mathrm{Tr}(U_aU_a^\top U_bU_b^\top) \\
&= \mathrm{Tr}(U_bU_b^\top) + \mathrm{Tr}(U_aU_a^\top) - \mathrm{Tr}(U_b^\top U_aU_a^\top U_b) - \mathrm{Tr}(U_a^\top U_bU_b^\top U_a) \\
&= d/r + d/r - \|U_a^\top U_b\|_F^2 - \|U_a^\top U_b\|_F^2 \\
&\geq 2d/r - 2d/2r = d/r,
\end{aligned}$$

hence $\|\Sigma_a^{-1/2}\Sigma_b\Sigma_a^{-1/2} - I\|_F \geq \frac{4}{5}\lambda\sqrt{d/r}$, completing the proof of the lemma. □

Finally, we prove our lower bound for mixtures, Theorem 4.34, for which we will need a standard result.

Lemma 4.33. *Let $T \geq 4$ and $k \in \mathbb{N}$. There exists a set of tuples $\mathcal{X} \subseteq [T]^k$ such that $|\mathcal{X}| \geq T^{\Omega(k)}$ and every pair of distinct $x, y \in \mathcal{X}$ have Hamming distance $|\{i \in [k] : x_i \neq y_i\}| \geq k/4$.*

Proof. This can be proven in several different ways. The conclusion of the lemma states that $\mathcal{X}$ is a code over the alphabet $[T]$ of rate $\Omega(1)$ and relative distance at least $1/4$. By standard results [81, Proposition 3.3.2], the T-ary entropy function H_T satisfies $1 - H_T(1/4) \geq 1/4$ as $T \geq 4$. By the Gilbert-Varshamov bound [81, Theorem 4.2.1], there exists such a code of rate $1/8$. □

Theorem 4.34. *Any algorithm that learns the class of mixtures of k Gaussians in $\mathbb{R}^d$ in the realizable setting within total variation distance ε and with success probability at least $2/3$ has sample complexity $\Omega\left(\frac{kd^2}{\varepsilon^2\log(1/\varepsilon)}\right)$.*

Proof. This proof builds on the lower bound construction for learning a single Gaussian (Theorem 4.29), and extends it to a lower bound for learning a mixture of Gaussians. The high-level idea is simple: create a family of distributions in k-mix$(\mathcal{G}^d)$ such that each Gaussian uses a covariance matrix as constructed in Theorem 4.29. As we will use Lemma 4.27 again to obtain the sample complexity lower bound, it suffices to construct $2^{\Omega(kd^2)}$ distributions in k-mix$(\mathcal{G}^d)$ with pairwise KL divergence $O(\varepsilon^2)$

and pairwise TV distance $\Omega(\varepsilon)$. Some care is required to ensure that the TV distance is large, and we will adopt some ideas used in earlier work for mixtures of spherical Gaussians [134, Appendix C.2]. In more detail, the construction proceeds as follows.

First, we construct a family of covariance matrices. The proof of Theorem 4.29 shows that there exists a family of symmetric, positive definite matrices $\Sigma_1, \dots, \Sigma_T$ with $T = 2^{\Omega(d^2)}$ satisfying

$$D_{\mathrm{KL}}(\mathcal{N}(0, \Sigma_i) \parallel \mathcal{N}(0, \Sigma_j)) \;\le\; O(\varepsilon^2) \qquad \forall i \neq j \tag{4.22a}$$

$$d_{\mathrm{TV}}\left(\mathcal{N}(0, \Sigma_i), \mathcal{N}(0, \Sigma_j)\right) \;\ge\; \Omega(\varepsilon) \qquad \forall i \neq j \tag{4.22b}$$

$$\Sigma_i \;\prec\; 2I \qquad \forall i. \tag{4.22c}$$

Next we will create a family of distributions in k-mix$(\mathcal{G}^d)$ for which each Gaussian in the mixture uses one of these Σ_i matrices as its covariance matrix. However, there is a tension. On the one hand, we'd like any two of these mixture distributions to use disjoint sets of covariance matrices, so that the TV distance between the mixtures is large. On the other hand, that constraint would greatly reduce the number of mixture distributions we can create, and we want many distributions in order to maximize the lower bound. This tension is resolved by a compromise obtained via error-correcting codes.

The formal construction proceeds as follows. First, we pick $\mu_1, \dots, \mu_k \in \mathbb{R}^d$, which will serve as the means for the Gaussians. The only constraint is that they should be far apart: for some Δ, to be chosen later, we have $\|\mu_i - \mu_j\|_2 \ge \Delta$ for all $i \neq j$. Each mixture distribution will be a uniform mixture of k Gaussians, for which the ith Gaussian has mean μ_i. The choice of covariance matrices is determined using the error-correcting code. Specifically, let $\mathcal{X} \subset [T]^k$ be a set as in Lemma 4.33 above. The family of mixture distributions is

$$\mathcal{F} \coloneqq \{\, f_x \,:\, x \in \mathcal{X} \,\} \qquad \text{where} \qquad f_x \;\coloneqq\; \frac{1}{k}\Big(\mathcal{N}(\mu_1, \Sigma_{x_1}) + \cdots + \mathcal{N}(\mu_k, \Sigma_{x_k})\Big).$$

As desired, we have $|\mathcal{F}| = T^{\Omega(k)} = 2^{\Omega(kd^2)}$.

To analyze $\mathcal{F}$, the first task is to prove the pairwise KL divergence upper bound. This is straightforward. Fix distinct $x, y \in \mathcal{X}$. For each i, (4.22a) shows that

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_i, \Sigma_{x_i}) \parallel \mathcal{N}(\mu_i, \Sigma_{y_i})) \;=\; D_{\mathrm{KL}}(\mathcal{N}(0, \Sigma_{x_i}) \parallel \mathcal{N}(0, \Sigma_{y_i})) \;\le\; O(\varepsilon^2).$$

Convexity of KL divergence [40, Theorem 2.7.2] then shows that $D_{\mathrm{KL}}(f_x \parallel f_y) \le O(\varepsilon^2)$.

The remaining task is to prove $d_{\mathrm{TV}}(f_x, f_y) \ge \Omega(\varepsilon)$ for all distinct $f_x, f_y \in \mathcal{F}$. The intuition is as follows. Say that index $i \in [k]$ *disagrees* if $x_i \neq y_i$. Whenever i disagrees, the ith Gaussian in f_x and ith Gaussian in f_y have TV distance $\Omega(\varepsilon)$ by (4.22b). Moreover, the total mixture weight apportioned to disagreeing indices is at least $1/4$, since the code ensures that the number of disagreements is at least $k/4$, and each mixture uses uniform weights on its components. Thus, the disagreeing coordinates should suffice to show that the TV distance is $\Omega(\varepsilon)$. Proving this formally requires somewhat more care because each Gaussian is supported on all of $\mathbb{R}^d$, so there is interaction between all Gaussians involved in the mixtures. However, the parameter Δ ensures that the means are far apart, so the interaction is

negligible.

More formally, let $A'_j \subseteq \mathbb{R}^d$ be such that

$$\mathbb{P}_{g\sim\mathcal{N}(\mu_j,\Sigma_{x_j})}[g \in A'_j] - \mathbb{P}_{g\sim\mathcal{N}(\mu_j,\Sigma_{y_j})}[g \in A'_j] \;=\; d_{\mathrm{TV}}\left(\mathcal{N}(\mu_j,\Sigma_{x_j}),\mathcal{N}(\mu_j,\Sigma_{y_j})\right). \tag{4.23}$$

Define

$$A_j \;=\; A'_j \cap B_j \qquad \text{where} \qquad B_j \;=\; \left\{\, x \in \mathbb{R}^d \,:\, \|x-\mu_j\|_2 < \Delta/2 \,\right\}.$$

Note that the separation of $\mu_1,\ldots,\mu_k$ implies that the balls $B_1,\ldots,B_k$ are pairwise disjoint. Consequently, the sets $A_1,\ldots,A_k$ are also pairwise disjoint.

Several preliminary inequalities are required concerning these events. First,

$$\begin{aligned}
\mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{x_i})}[g \notin B_i] \;&=\; \mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{x_i})}[\|g-\mu_i\|_2^2 \geq (\Delta/2)^2] \\
&=\; \mathbb{P}_{g\sim\mathcal{N}(0,\Sigma_{x_i})}[\|g\|_2^2 \geq (\Delta/2)^2] && \text{(translating to zero-mean)} \\
&\leq\; \mathbb{P}_{g\sim\mathcal{N}(0,I_d)}[\|g\|_2^2 \geq \Delta^2/8] && \text{(by (4.22c))} \\
&\leq\; \varepsilon^2/k^2,
\end{aligned} \tag{4.24}$$

by applying Lemma B.13 with $t = 2\ln(k/\varepsilon)$ and choosing Δ to satisfy $\Delta^2/8 = d+2\sqrt{dt}+2t$. Inequality (4.24) also holds replacing x_i with y_i. Since $A'_i \setminus A_i \subseteq B_i^c$, (4.24) shows that

$$\left|\; \mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{x_i})}[g \in A_i] - \mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{x_i})}[g \in A'_i] \;\right| \;\leq\; \mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{x_i})}[g \notin B_i] \;\leq\; \varepsilon^2/k^2. \tag{4.25}$$

This inequality also holds using y_i instead of x_i. For $i \neq j$, we have $A_j \subseteq B_i^c$, so

$$\mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{y_i})}[g \in A_j] \;\leq\; \mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{y_i})}[g \notin B_i] \;\leq\; \varepsilon^2/k^2. \tag{4.26}$$

Finally, by (4.23), (4.25) and the triangle inequality,

$$\mathbb{P}_{g\sim\mathcal{N}(\mu_j,\Sigma_{x_j})}[g \in A_j] - \mathbb{P}_{g\sim\mathcal{N}(\mu_j,\Sigma_{y_j})}[g \in A_j] \;\geq\; d_{\mathrm{TV}}\left(\mathcal{N}(\mu_j,\Sigma_{x_j}),\mathcal{N}(\mu_j,\Sigma_{y_j})\right) - 2\varepsilon^2/k^2. \tag{4.27}$$

The total variation distance is lower bounded as follows. Let $A \coloneqq A_1 \cup \cdots \cup A_k$. Then

$$\begin{aligned}
&d_{\mathrm{TV}}\left(f_x, f_y\right) \\
&\geq\; \mathbb{P}_{g\sim f_x}[g \in A] - \mathbb{P}_{g\sim f_y}[g \in A] \\
&=\; \sum_{j=1}^{k} \left(\mathbb{P}_{g\sim f_x}[g \in A_j] - \mathbb{P}_{g\sim f_y}[g \in A_j]\right) && \text{(by disjointness of the } A_j\text{)} \\
&=\; \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{k} \left(\mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{x_i})}[g \in A_j] - \mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{y_i})}[g \in A_j]\right) && \text{(expanding } f_x, f_y \text{ as } k\text{-mixtures)} \\
&=\; \frac{1}{k}\sum_{j=1}^{k} \left(\mathbb{P}_{g\sim\mathcal{N}(\mu_j,\Sigma_{x_j})}[g \in A_j] - \mathbb{P}_{g\sim\mathcal{N}(\mu_j,\Sigma_{y_j})}[g \in A_j]\right) && \text{(summands with } i = j\text{)}
\end{aligned}$$

$$
\begin{aligned}
&+ \ \frac{1}{k}\sum_{j=1}^{k}\sum_{i\neq j}\Big(\underbrace{\mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{x_i})}[g\in A_j]}_{\geq 0} - \underbrace{\mathbb{P}_{g\sim\mathcal{N}(\mu_i,\Sigma_{y_i})}[g\in A_j]}_{\leq \varepsilon^2/k^2 \text{ by (4.26)}}\Big) \qquad \text{(summands with } i\neq j\text{)} \\
&\geq \ \frac{1}{k}\sum_{j=1}^{k}\Big(\mathbb{P}_{g\sim\mathcal{N}(\mu_j,\Sigma_{x_j})}[g\in A_j] - \mathbb{P}_{g\sim\mathcal{N}(\mu_j,\Sigma_{y_j})}[g\in A_j]\Big) - \varepsilon^2 \\
&\geq \ \frac{1}{k}\sum_{j=1}^{k}\Big(d_{\mathrm{TV}}\left(\mathcal{N}(\mu_j,\Sigma_{x_j}),\mathcal{N}(\mu_j,\Sigma_{y_j})\right) - 2\varepsilon^2/k^2\Big) - \varepsilon^2 \qquad \text{(by (4.27))} \\
&\geq \ \frac{1}{k}(k/4)\Omega(\varepsilon) - 3\varepsilon^2 \ = \ \Omega(\varepsilon),
\end{aligned}
$$

where the last inequality is because $d_{\mathrm{TV}}\left(\mathcal{N}(\mu_j,\Sigma_{x_j}),\mathcal{N}(\mu_j,\Sigma_{y_j})\right) \geq \Omega(\varepsilon)$ whenever $x_j \neq y_j$, which is the case for at least $k/4$ of the indices j. □

Appendix A

Appendix for Chapter 2

A.1 Standard facts

Fact A.1. *Suppose $f\colon \mathbb{R} \to \mathbb{R}$ is concave. Then for any $\alpha < \beta$, the function $g(t) = f(t+\beta) - f(t+\alpha)$ is non-increasing.*

Fact A.2. *Suppose that $f\colon \mathbb{R} \to \mathbb{R}$ is concave. Let $\alpha < \beta$. Then $f(x) \geq \min\{f(\alpha), f(\beta)\}$ for all $x \in [\alpha, \beta]$.*

A.2 Proof of Lemma 2.47

The main idea of the proof is that we will approximate R_α by a sequence of smooth functions (i.e. functions in $C^{2,2}$).

Fix $\alpha > 0$. Recall that $\tilde{R}_\alpha(t,x) = \frac{x}{2} + \kappa_\alpha \sqrt{t} \cdot M_0\left(\frac{x^2}{2t}\right)$ for $t > 0, x \in \mathbb{R}$, where $\kappa_\alpha = \frac{1}{\sqrt{2\pi}\,\mathrm{erfi}(\alpha/\sqrt{2})}$. (For $t = 0$, it suffices to define $\tilde{R}_\alpha(t,x) = 0$.) We also have the truncated version, R_α, defined as

$$R_\alpha(t,x) = \begin{cases} \tilde{R}_\alpha(t,x) & t > 0 \wedge x \leq \alpha\sqrt{t} \\ \tilde{R}_\alpha(t, \alpha\sqrt{t}) & t > 0 \wedge x \geq \alpha\sqrt{t} \\ 0 & t = 0 \end{cases}.$$

Recall also that $p_\alpha = \partial_x R_\alpha$. For convenience, we restate the lemma.

Lemma 2.47. *Fix $\alpha > 0$. Then, almost surely, for all $T \geq 0$, $\mathrm{ContRegret}(T, p_\alpha, B) \leq R_\alpha(T, |B_T|)$.*

For the remainder of this section, we will write $\tilde{f} = \tilde{R}_\alpha$ and $f = R_\alpha$. Let $\phi(x)$ be any non-increasing C^2 function satisfying $\phi(x) = 1$ for $x \leq 0$ and $\phi(x) = 0$ for $x \geq 1$. For concreteness, we may take

$$\phi(x) = \begin{cases} 1 & x \leq 0 \\ (1-x) + \frac{1}{2\pi}\sin(2\pi x) & x \in [0,1] \\ 0 & x \geq 1 \end{cases}. \tag{A.1}$$

We leave it as an easy calculus exercise to verify that ϕ is indeed a non-increasing C^2 function.

Next, define $\phi_n(x) = \phi(nx)$ and

$$f_n(t,x) = \tilde{f}(t,x) \cdot \phi_n(x - \alpha\sqrt{t}) + f(t, \alpha\sqrt{t}) \cdot \left(1 - \phi_n(x - \alpha\sqrt{t})\right).$$

Note that $f_n \in C^{2,2}$ on $\mathbb{R}_{>0} \times \mathbb{R}$ for all n. The function f_n is a smooth approximation to f and its limit is exactly $f\,(= R_\alpha)$.

Claim A.3. *For every $t > 0, x \in \mathbb{R}$, $\lim_{n\to\infty} f_n(t,x) = f(t,x)$.*

Proof. If $x \le \alpha\sqrt{t}$ then $\phi_n(x - \alpha\sqrt{t}) = 1$ so $f_n(t,x) = \tilde{f}(t,x) = f(t,x)$. In particular, this also holds for the limit. Next, suppose that $a = x - \alpha\sqrt{t} > 0$. If $n > 1/a$ then $\phi_n(x - \alpha\sqrt{t}) = 0$ so $f_n(t,x) = \tilde{f}(t, \alpha\sqrt{t}) = f(t,x)$. □

Recall that our goal is to relate $f(T, |B_T|)$ and $\int_0^T \partial_x f(t, |B_t|)\,\mathrm{d}|B_t|$. However, one cannot apply Itô's formula to f directly as it is not in $C^{1,2}$. Instead, we will apply Itô's formula to the smoothed version of f, namely f_n, and then take limits. The remainder of this section does this limiting argument carefully.

For technical reasons (namely that $\tilde{f}(t,x)$ has a pole when $t \to 0$ and $x \neq 0$), we will not be able to start the stochastic integral at 0. Hence, we will fix $\varepsilon > 0$ and, at the end of the proof, we will allow $\varepsilon \to 0$.

The following lemma bounds the stochastic integral of $\partial_x f_n$ with respect to $|B_t|$.

Lemma A.4. *Almost surely, for every $T \ge \varepsilon$*

$$\begin{aligned}\int_\varepsilon^T \partial_x f_n(t, |B_t|)\,\mathrm{d}|B_t| \le{}& f_n(T, |B_T|) - f_n(\varepsilon, |B_\varepsilon|) \\ &- \int_\varepsilon^T \frac{\alpha}{2\sqrt{t}} \cdot \phi_n'(|B_t| - \alpha\sqrt{t}) \cdot \left(f(t, \alpha\sqrt{t}) - \tilde{f}(t, |B_t|)\right) \mathrm{d}t \\ &- \frac{1}{2}\int_\varepsilon^T \phi_n''(|B_t| - \alpha\sqrt{t}) \cdot \left(f(t, \alpha\sqrt{t}) - \tilde{f}(t, |B_t|)\right) \mathrm{d}t.\end{aligned} \tag{A.2}$$

Proof. The proof is by Itô's formula (Theorem 2.39) applied to f_n. We have, for all $T \ge \varepsilon$,

$$f_n(T, |B_T|) - f_n(\varepsilon, |B_\varepsilon|) = \int_\varepsilon^T \partial_x f_n(t, |B_t|)\,\mathrm{d}|B_t| + \int_\varepsilon^T \partial_t f_n(t, |B_t|) + \frac{1}{2}\partial_{xx} f_n(t, |B_t|)\,\mathrm{d}t. \tag{A.3}$$

Computing derivatives of f_n, we have

$$\begin{aligned}\partial_t f_n(t,x) = {}& (\partial_t \tilde{f}(t,x)) \cdot \phi_n(x - \alpha\sqrt{t}) - \frac{\alpha}{2\sqrt{t}} \tilde{f}(t,x) \phi_n'(x - \alpha\sqrt{t}) \\ &+ \partial_t (f(t, \alpha\sqrt{t})) \cdot (1 - \phi_n(x - \alpha\sqrt{t})) + \frac{\alpha}{2\sqrt{t}} f(t, \alpha\sqrt{t}) \cdot \phi_n'(x - \alpha\sqrt{t})\end{aligned} \tag{A.4}$$

$$\partial_x f_n(t,x) = (\partial_x \tilde{f}(t,x)) \cdot \phi_n(x - \alpha\sqrt{t}) + \tilde{f}(t,x)\phi_n'(x - \alpha\sqrt{t}) - f(t, \alpha\sqrt{t})\phi_n'(x - \alpha\sqrt{t}) \tag{A.5}$$

$$\begin{aligned}\partial_{xx} f_n(t,x) = {}& (\partial_{xx} \tilde{f}(t,x)) \cdot \phi_n(x - \alpha\sqrt{t}) + 2(\partial_x \tilde{f}(t,x))\phi_n'(x - \alpha\sqrt{t}) \\ &+ \left(\tilde{f}(t,x) - f(t, \alpha\sqrt{t})\right) \phi_n''(x - \alpha\sqrt{t}).\end{aligned} \tag{A.6}$$

Recalling the notation $\overset{*}{\Delta} = \partial_t + \frac{1}{2}\partial_{xx}$, we have

$$\begin{aligned}\overset{*}{\Delta} f_n(t,x) &= \left(\overset{*}{\Delta}\tilde{f}(t,x)\right)\cdot\phi_n(x-\alpha\sqrt{t}) + \partial_t(f(t,\alpha\sqrt{t}))\cdot(1-\phi_n(x-\alpha\sqrt{t}))\\ &\quad + (\partial_x\tilde{f}(t,x))\phi_n'(x-\alpha\sqrt{t})\\ &\quad + \frac{\alpha}{2\sqrt{t}}\cdot(f(t,\alpha\sqrt{t})-\tilde{f}(t,x))\cdot\phi_n'(x-\alpha\sqrt{t}) + \frac{1}{2}\left(\tilde{f}(t,x)-f(t,\alpha\sqrt{t})\right)\phi_n''(x-\alpha\sqrt{t}).\end{aligned} \tag{A.7}$$

By Lemma 2.46, $\overset{*}{\Delta}\tilde{f} = 0$. By Claim A.5 below, $\partial_t(f(t,\alpha\sqrt{t})) > 0$. Next, observe that $(\partial_x\tilde{f}(t,x))\cdot\phi_n'(x-\alpha\sqrt{t}) \geq 0$. To see this, if $x \leq \alpha\sqrt{t}$ then $\phi_n'(x-\alpha\sqrt{t}) = 0$. On the other hand, if $x > \alpha\sqrt{t}$ then $\phi_n'(x-\alpha\sqrt{t}) \leq 0$ because ϕ_n is non-increasing and $\partial_x\tilde{f}(t,x) \leq 0$ by Lemma 2.46 and Eq. (2.40). Hence, we can lower bound Eq. (A.7) by

$$\overset{*}{\Delta} f_n(t,x) \geq \frac{\alpha}{2\sqrt{t}}\cdot(f(t,\alpha\sqrt{t})-\tilde{f}(t,x))\cdot\phi_n'(x-\alpha\sqrt{t}) + \frac{1}{2}\left(\tilde{f}(t,x)-f(t,\alpha\sqrt{t})\right)\phi_n''(x-\alpha\sqrt{t}). \tag{A.8}$$

Plugging Eq. (A.8) into Eq. (A.3) gives

$$\begin{aligned}f_n(T,|B_T|) - f_n(\varepsilon,|B_\varepsilon|) &\geq \int_\varepsilon^T \partial_x f_n(t,|B_t|)\,\mathrm{d}|B_t|\\ &\quad + \int_\varepsilon^T \frac{\alpha}{2\sqrt{t}}\cdot\phi_n'(|B_t|-\alpha\sqrt{t})\cdot\left(f(t,\alpha\sqrt{t})-\tilde{f}(t,|B_t|)\right)\mathrm{d}t\\ &\quad + \frac{1}{2}\int_\varepsilon^T \phi_n''(|B_t|-\alpha\sqrt{t})\cdot\left(f(t,\alpha\sqrt{t})-\tilde{f}(t,|B_t|)\right)\mathrm{d}t.\end{aligned} \tag{A.9}$$

Rearranging Eq. (A.9) gives the lemma. □

Claim A.5. *If $t > 0$ then $\partial_t(\tilde{f}(t,\alpha\sqrt{t})) > 0$.*

Proof. Note that

$$\tilde{f}(t,\alpha\sqrt{t}) = \sqrt{t}\cdot\left(\frac{\alpha}{2} + \frac{M_0(\alpha^2/2)}{\sqrt{2\pi}\,\mathrm{erfi}(\alpha/\sqrt{2})}\right) = \sqrt{t}\cdot f(1,\alpha).$$

So it suffices to check that $\tilde{f}(1,\alpha) > 0$. To see this, note that $\tilde{f}(1,0) = \kappa_\alpha > 0$ and $\partial_x\tilde{f}(1,x) \geq 0$ as long as $x \leq \alpha$ (by the first identity of Lemma 2.45). Hence, $\tilde{f}(1,\alpha) > 0$. □

At this point, we would like to take limits on both sides of Eq. (A.2). This is achieved by the following two lemmas.

Lemma A.6. *Almost surely, for every $T \geq \varepsilon$,*

1. $\lim_{n\to\infty}\int_\varepsilon^T \frac{\alpha}{2\sqrt{t}}\cdot\phi_n'(|B_t|-\alpha\sqrt{t})\cdot\left(f(t,\alpha\sqrt{t})-\tilde{f}(t,|B_t|)\right)\mathrm{d}t = 0$*; and*
2. $\lim_{n\to\infty}\int_\varepsilon^T \phi_n''(|B_t|-\alpha\sqrt{t})\cdot\left(f(t,\alpha\sqrt{t})-\tilde{f}(t,|B_t|)\right)\mathrm{d}t = 0$.

Lemma A.7. *For every $T \geq \varepsilon$,*

$$\int_\varepsilon^T \partial_x f_n(t,|B_t|)\,\mathrm{d}|B_t| \xrightarrow{L^2} \int_\varepsilon^T \partial_x f(t,|B_t)\,\mathrm{d}|B_t|$$

as $n \to \infty$.

Within this section, $X_n \xrightarrow{L^2} X$ means that $\mathbb{E}\left[(X_n - X)^2\right] \to 0$ as $n \to \infty$. We relegate the proofs of Lemma A.6 and Lemma A.7 to Appendix A.2.1. We now take limits on both sides of Eq. (A.2) to obtain the following bound on the stochastic integral of $\partial_x f$.

Lemma A.8. *Almost surely, for every* $T \geq \varepsilon$,

$$\int_\varepsilon^T \partial_x f(t, |B_t|) \, \mathrm{d}|B_t| \leq f(T, |B_T|) - f(\varepsilon, |B_\varepsilon|). \tag{A.10}$$

Proof. By Lemma A.7, for every $T \geq \varepsilon$,

$$\int_\varepsilon^T \partial_x f_n(t, |B_t|) \, \mathrm{d}|B_t| \xrightarrow{L^2} \int_\varepsilon^T \partial_x f(t, |B_t) \, \mathrm{d}|B_t|.$$

Hence, there exists a subsequence n_k such that

$$\int_\varepsilon^T \partial_x f_{n_k}(t, |B_t|) \, \mathrm{d}|B_t| \xrightarrow{\text{a.s.}} \int_\varepsilon^T \partial_x f(t, |B_t) \, \mathrm{d}|B_t|.$$

Using Lemma A.4 to bound the left-hand-side and then Lemma A.6 to take limits gives that Eq. (A.10) holds for any fixed $T \geq \varepsilon$. Hence, almost surely, Eq. (A.10) holds for all rational $T \geq \varepsilon$. As both sides of Eq. (A.10) are continuous as a function of T, Eq. (A.10) holds for all $T \geq \varepsilon$. □

Proof (of Lemma 2.47). We will work in the probability 1 set where Lemma A.8 holds (for every rational $\varepsilon > 0$) and $t \mapsto B_t$ is continuous.

Fix $T > 0$. Note that $\mathrm{ContRegret}(T, \partial_x f, B)$ is defined because $\partial_x f \in [0, 1/2]$ and $\partial_x f(t, 0) = 1/2$ for all $t > 0$ (see Eq. (2.46)). Recalling Definition 2.35, we have, for $\varepsilon \leq T$,

$$\begin{aligned}
\mathrm{ContRegret}(T, \partial_x f, B) &= \int_0^T \partial_x f(t, |B_t|) \, \mathrm{d}|B_t| \\
&= \int_\varepsilon^T \partial_x f(t, |B_t|) \, \mathrm{d}|B_t| + \int_0^\varepsilon \partial_x f(t, |B_t|) \, \mathrm{d}|B_t| \\
&\leq f(T, |B_T|) - f(\varepsilon, |B_\varepsilon|) + \int_0^\varepsilon \partial_x f(t, |B_t|) \, \mathrm{d}|B_t| \qquad \text{(Lemma A.8)}.
\end{aligned}$$

The right-hand-side is continuous in ε so taking $\varepsilon \to 0$ (and recalling that $f(0, 0) = 0$), gives

$$\mathrm{ContRegret}(T, \partial_x f, B) \leq f(T, |B_T|).$$ □

A.2.1 Additional proofs from Appendix A.2

Before we prove Lemma A.6, we will need one key observation.

Lemma A.9. *Fix* $\varepsilon > 0$. *Then there is a constant* $C_\varepsilon > 0$ *(depending also on* α*) such that for* $t > 0$ *and* x *satisfying* $\left|x - \alpha\sqrt{t}\right| \leq 1$,

1. $\left|\tilde{f}(t,x) - f(t,\alpha\sqrt{t})\right| \le C_\varepsilon \cdot (x-\alpha\sqrt{t})^2$; *and*
2. $\left|\partial_x \tilde{f}(t,x)\right| \le C_\varepsilon \cdot \left|x-\alpha\sqrt{t}\right|$.

Proof. The key observation is that $f(t,\alpha\sqrt{t})$ is already a first-order Taylor expansion of $\tilde{f}(t,x)$ (in x) about the point $\gamma\sqrt{t}$. Indeed, $\tilde{f}(t,\alpha\sqrt{t}) = f(t,\alpha\sqrt{t})$ and $(\partial_x\tilde{f})(t,\alpha,\sqrt{t}) = 0$. Hence, by Taylor's Theorem (see e.g. [128, Theorem 5.15])

$$\left|\tilde{f}(t,x) - f(t,\alpha\sqrt{t})\right| \le \frac{1}{2}\cdot(x-\alpha\sqrt{t})^2 \cdot \sup_{t\ge\varepsilon,\left|x-\alpha\sqrt{t}\right|\le 1}\left|\partial_{xx}\tilde{f}(t,x)\right|$$

By the second identity in Lemma 2.45, we have

$$\left|\partial_{xx}\tilde{f}(t,x)\right| = \frac{\kappa_\alpha \exp\left(x^2/2t\right)}{\sqrt{2t}}.$$

Since $t \ge \varepsilon$ and $x \le 1+\alpha\sqrt{t}$, we have

$$\begin{aligned}\left|\partial_{xx}\tilde{f}(t,x)\right| &\le \frac{\kappa_\alpha \exp\left((1+\alpha\sqrt{t})^2/2t\right)}{\sqrt{2\varepsilon}} \\ &= \frac{\kappa_\alpha \exp(\alpha^2/2 + \alpha/\sqrt{t} + 1/t)}{\sqrt{2\varepsilon}} \\ &\le \frac{\kappa_\alpha \exp(\alpha^2/2 + \alpha/\sqrt{\varepsilon} + 1/\varepsilon)}{\sqrt{2\varepsilon}}.\end{aligned}$$

So one can take $C_\varepsilon = \frac{\kappa_\alpha \exp(\alpha^2/2+\alpha/\sqrt{\varepsilon}+1/\varepsilon)}{\sqrt{2\varepsilon}}$. This gives the first assertion.

The second assertion is similar. Indeed, since $(\partial_x\tilde{f})(t,\alpha\sqrt{t}) = 0$, we have

$$\begin{aligned}\left|(\partial_x\tilde{f})(t,x)\right| &= \left|(\partial_x\tilde{f})(t,x) - (\partial_x\tilde{f})(t,\alpha\sqrt{t})\right| \\ &\le \left|x-\alpha\sqrt{t}\right| \cdot \sup_{t\ge\varepsilon,|x-\alpha\sqrt{t}|\le 1}|\partial_{xx}\tilde{f}(t,x)| \\ &\le C_\varepsilon \cdot \left|x-\alpha\sqrt{t}\right|.\end{aligned}$$

□

We also need a simple claim which bounds the value of $|\phi_n'(x)|$ and $|\phi_n''(x)|$.

Claim A.10. *There is an absolute constant $C > 0$ such that $|\phi_n'(x)| \le Cn$ and $|\phi_n''(x)| \le Cn^2$.*

Proof. Note that $\phi_n'(x) = n\cdot\phi'(x)$ and $n^2\cdot\phi''(x)$. It is easy to see, from differentiating Eq. (A.1) or by continuity and compact arguments, that there exists $C > 0$ such that $|\phi'(x)|, |\phi''(x)| \le C$ for all $x \in \mathbb{R}$. □

Proof (of Lemma A.6). We start with the second assertion. The first assertion is similar but simpler. We claim that there exists a constant C' (depending on ε and α) such that

$$\left|\phi_n''(|B_t|-\alpha\sqrt{t})\cdot\left(f(t,\alpha\sqrt{t}) - \tilde{f}(t,|B_t|)\right)\right| \le C'\mathbb{I}[|B_t|-\alpha\sqrt{t}\in[0,1/n]] \tag{A.11}$$

Indeed, if $|B_t| - \alpha\sqrt{t} \notin [0, 1/n]$ then $\phi_n''(|B_t| - \alpha\sqrt{t}) = 0$ so both sides of Eq. (A.11) are equal to 0. On the other hand, if $|B_t| - \alpha\sqrt{t} \in [0, 1/n]$ then Lemma A.9 shows that $|f(t, \alpha\sqrt{t}) - \tilde{f}(t, |B_t|)| \le C_\varepsilon/n^2$ where C_ε is the constant from Lemma A.9. Next, Claim A.10 gives $|\phi_n''(|B_t| - \alpha\sqrt{t})| \le Cn^2$. So taking $C' = C_\varepsilon \cdot C$ gives Eq. (A.11). Hence,

$$\begin{aligned}\left|\int_\varepsilon^T \phi_n''(|B_t| - \alpha\sqrt{t}) \cdot \left(f(t, \alpha\sqrt{t}) - \tilde{f}(t, |B_t|)\right) \mathrm{d}t\right| &\le \int_\varepsilon^T C' \cdot \mathbb{I}[|B_t| - \alpha\sqrt{t} \in [0, 1/n]] \, \mathrm{d}t \\ &= C' \cdot m\left(\left\{\ t \in [\varepsilon, T]\ :\ |B_t| - \alpha\sqrt{t} \in [0, 1/n]\ \right\}\right),\end{aligned}$$

where m denotes the Lebesgue measure. By continuity of measure, we have

$$\lim_n m\left(\left\{\ t \in [\varepsilon, T]\ :\ |B_t| - \alpha\sqrt{t} \in [0, 1/n]\ \right\}\right) = \int_\varepsilon^T \mathbb{I}\left[|B_t| = \alpha\sqrt{t}\right] \mathrm{d}t = 0 \quad \text{a.s.}$$

This proves the second assertion.

For the first assertion, we can use the bound (from Lemma A.9 and Claim A.10)

$$\left|\phi_n'(x - \alpha\sqrt{t}) \cdot \left(f(t, \alpha\sqrt{t}) - \tilde{f}(t, x)\right)\right| \le \frac{C'}{n}\mathbb{I}[x - \alpha\sqrt{t} \in [0, 1/n]] \le \frac{C'}{n}. \tag{A.12}$$

Hence,

$$\begin{aligned}\left|\int_\varepsilon^T \frac{\alpha}{2\sqrt{t}} \cdot \phi_n'(|B_t| - \alpha\sqrt{t}) \cdot \left(f(t, \alpha\sqrt{t}) - \tilde{f}(t, |B_t|)\right) \mathrm{d}t\right| &\le \int_\varepsilon^T \frac{\alpha}{2\sqrt{t}} \frac{C'}{n} \, \mathrm{d}t \\ &\le C'\alpha\sqrt{T}/n \to 0. \qquad \square\end{aligned}$$

Proof (of Lemma A.7). By Eq. (A.5), we have

$$\begin{aligned}\partial_x f_n(t, x) - \partial_x f(t, x) = &\left(\partial_x \tilde{f}(t, x)\phi_n(x - \alpha\sqrt{t}) - \partial_x f(t, x)\right) \\ &+ \left(\phi_n'(x - \alpha\sqrt{t}) \cdot \left(\tilde{f}(t, x) - f(t, \alpha\sqrt{t})\right)\right).\end{aligned} \tag{A.13}$$

For the first bracketed term, since $\partial_x \tilde{f}(t, x) = \partial_x f(t, x)$ when $x \le \alpha\sqrt{t}$ and $\partial_x f(t, x) = 0$ when $x \ge \alpha\sqrt{t}$, we have

$$\begin{aligned}\left|\partial_x \tilde{f}(t, x)\phi_n(x - \alpha\sqrt{t})\right| &= \left|\partial_x \tilde{f}(t, x)\phi_n(x - \alpha\sqrt{t})\right| \mathbb{I}[x - \alpha\sqrt{t} \in [0, 1/n]] \\ &\le \frac{C'}{n},\end{aligned}$$

where the final inequality is by the second assertion in Lemma A.9. The second bracketed term has been bounded in Eq. (A.12), and so we have proved

$$\left|\partial_x f_n(t, x) - \partial_x f(t, x)\right| \le \frac{C''}{n} \text{ for all } t \ge \varepsilon \text{ and all } x. \tag{A.14}$$

Tanaka's formula (see [127, Theorem IV.43.3]) states that

$$|B_t| = \int_0^t \operatorname{sign}(B_s)\,\mathrm{d}B_s + L_t =: W_t + L_t,$$

where L is the local time at zero of B and W is a Brownian motion. Recall that $t \mapsto L_t$ is a continuous non-decreasing random process which increases only on the set $\{\, t : B_t = 0 \,\}$. Therefore by the Itô isometry property, for any $T \geq \varepsilon$,

$$\begin{aligned}
&\mathrm{E}\Big[\Big(\int_\varepsilon^T \partial_x f_n(t,|B_t|)\,\mathrm{d}|B|_t - \int_\varepsilon^T \partial_x f(t,|B_t|)\,\mathrm{d}|B|_t\Big)^2\Big] \\
&\quad \leq 2\mathbb{E}\left[\Big(\int_\varepsilon^T (\partial_x f_n - \partial_x f)(t,|B_t|)\,\mathrm{d}W_t\Big)^2\right] + 2\mathbb{E}\left[\Big(\int_\varepsilon^T (\partial_x f_n - \partial_x f)(t,|B_t|))\,\mathrm{d}L_t\Big)^2\right] \\
&\quad = 2\mathbb{E}\left[\int_\varepsilon^T (\partial_x f_n - \partial_x f)(t,|B_t|)^2\,\mathrm{d}t\right] + 2\mathbb{E}\left[\Big(\int_\varepsilon^T (\partial_x f_n - \partial_x f)(t,0)\,\mathrm{d}L_t\Big)^2\right].
\end{aligned}$$

Now use (A.14) to bound the right-hand side by

$$2(C''/n)^2 T + 2(C''/n)^2 \mathbb{E}\left[L_T^2\right] \leq C'''n^{-2}T,$$

where the last inequality uses Tanaka's formula (and the fact that W_t is also a standard Brownian motion) to bound

$$\mathbb{E}\left[L_T^2\right] = \mathbb{E}\left[(|B_T| - W_T)^2\right] \leq 2\mathbb{E}\left[|B_T|^2\right] + 2\mathbb{E}\left[|W_T|^2\right] = 4\mathbb{E}\left[|B_T|^2\right] = O(T).$$

The result follows. □

A.2.2 Discussion on the statement of Theorem 2.39

In this paper, we use the version of Itô's formula that appears in Remark 1 after Theorem IV.3.3 in [124]. It states that if $f \in C^{1,2}$, X is a continuous semimartingale[1] and A is a process with bounded variation then

$$\begin{aligned}
f(A_T, X_T) - f(A_0, X_0) = &\int_0^T \partial_x f(A_t, X_t)\,\mathrm{d}X_t + \int_0^T \partial_t f(A_t, X_t)\,\mathrm{d}A_t \\
&+ \frac{1}{2}\int_0^T \partial_{xx} f(A_t, X_t)\,\mathrm{d}\langle X, X\rangle_t.
\end{aligned} \tag{A.15}$$

In our setting, we take $X_t = |B_t|$ and $A_t = t$. We now explain the notation $\langle X, X\rangle$.

(1) For a continuous local martingale M, $\langle M, M\rangle$ is the unique increasing continuous process vanishing at 0 such that $M^2 - \langle M, M\rangle$ is a martingale [124, Theorem IV.1.8].

(2) If X is a continuous semimartingale with M being the (continuous) local martingale part then $\langle X, X\rangle = \langle M, M\rangle$ [124, Definition IV.1.20].

[1] A continuous semimartingale X is a process that can be written as $X = M + N$ where M is a continuous local martingale and N is a continuous adapted process of finite variation.

Tanaka's formula [127, Theorem IV.43.3] asserts that $|B_t| = W_t + L_t$ where W_t is a Brownian Motion and L_t is the local time of B_t at 0, which is an increasing, continuous, adapated process. Hence, $|B_t|$ is a semimartingale with $\langle |B|, |B| \rangle_t = \langle W, W \rangle_t = t$. Plugging these into Eq. (A.15) gives

$$f(T, |B_T|) - f(0, |B_0|) \;=\; \int_0^T \partial_x f(t, |B_t|) \mathrm{d}\,|B_t| + \int_0^T \Big[\partial_t f(t, |B_t|) + \tfrac{1}{2}\partial_{xx} f(t, |B_t|)\Big] \mathrm{d}t,$$

which is what appears in Theorem 2.39.

Appendix B

Appendix for Chapter 4

B.1 Standard facts

Definition B.1. Let A and B be symmetric positive definite matrices of the same size. The *log-det divergence* of A and B is defined as $\mathrm{LD}(A,B) \coloneqq \mathrm{Tr}(B^{-1}A - I) - \log\det(B^{-1}A)$.

The log-det divergence is an asymmetric measure of distance between matrices and is closely related to the KL divergence between their corresponding Gaussian distributions, as can be seen from Lemma B.3.

Claim B.2. *Let A, B and C be square matrices of the same size. Suppose that A and B are symmetric, positive definite and C is invertible. Then* $\mathrm{LD}(A,B) = \mathrm{LD}(CAC, CBC)$.

Proof. From the definition it is apparent that $\mathrm{LD}(A,B)$ only depends on the spectrum of $B^{-1}A$. So the claim follows from the fact that $B^{-1}A$ and $(CBC)^{-1}CAC$ have the same spectrum. This fact holds because v is an eigenvector for $B^{-1}A$ of eigenvalue λ if and only if $C^{-1}v$ is an eigenvector for $(CBC)^{-1}CAC$ of eigenvalue λ. □

Lemma B.3 (Rasmussen and Williams [120, Equation A.23])**.** *For two full-rank Gaussians $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$, their KL divergence is*

$$\begin{aligned} & D_{\mathrm{KL}}(\mathcal{N}(\mu_0,\Sigma_0) \parallel \mathcal{N}(\mu_1,\Sigma_1)) \\ &= \frac{1}{2}\left(\mathrm{Tr}(\Sigma_1^{-1}\Sigma_0 - I) + (\mu_0-\mu_1)^\top \Sigma_1^{-1}(\mu_0-\mu_1) - \log\det(\Sigma_0\Sigma_1^{-1})\right) \\ &= \frac{1}{2}\left(\mathrm{LD}(\Sigma_0,\Sigma_1) + (\mu_0-\mu_1)^\top \Sigma_1^{-1}(\mu_0-\mu_1)\right). \end{aligned}$$

Lemma B.4. *Let A, B be symmetric, positive definite matrices, satisfying $(1-\alpha)B \preceq A \preceq (1+\alpha)B$ for some $\alpha \in [0, 1/2]$. Then* $\mathrm{LD}(A,B) \le d\alpha^2$.

Proof. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of $B^{-1}A$. By the hypothesis, each $\lambda_i \in [1-\alpha, 1+\alpha]$. So,

$$\mathrm{LD}(A,B) \;=\; \mathrm{Tr}(B^{-1}A - I) - \log\det(B^{-1}A) \;=\; \sum_{i=1}^{d}(\lambda_i - 1) \;-\; \log\prod_{i=1}^{d}\lambda_i$$

$$= \sum_{i=1}^{d}(\lambda_i - 1 - \log(\lambda_i)) \;\le\; \sum_{i=1}^{d}(\lambda_i - 1)^2 \;\le\; d\alpha^2.$$

The first inequality follows from $x - 1 - \log x \le (x-1)^2$, valid for any $x \ge 1/2$. □

Lemma B.5. *For two full-rank Gaussians $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu', \Sigma')$, their total variation distance is bounded by*

$$\begin{aligned} 2d_{\mathrm{TV}}(\mathcal{N}(\mu_0,\Sigma_0),\mathcal{N}(\mu_1,\Sigma_1))^2 &\le D_{\mathrm{KL}}(\mathcal{N}(\mu_0,\Sigma_0 \parallel \mathcal{N}(\mu_1,\Sigma_1)) \\ &= \frac{1}{2}\Big(\mathrm{LD}\,(\Sigma_0,\Sigma_1) + (\mu_0-\mu_1)^\top \Sigma_1^{-1}(\mu_0-\mu_1)\Big). \end{aligned}$$

Proof. Follows from Lemma B.3 and Lemma 3.9. □

Lemma B.6. *For any $\mu, \sigma, \widehat{\mu}, \widehat{\sigma} \in \mathbb{R}$ with $|\widehat{\mu} - \mu| \le \varepsilon\sigma$ and $|\widehat{\sigma} - \sigma| \le \varepsilon\sigma$ and $\varepsilon \in [0, 2/3]$ we have*

$$\|\mathcal{N}(\mu,\sigma^2) - \mathcal{N}(\widehat{\mu},\widehat{\sigma}^2)\|_1 \;\le\; 2\varepsilon.$$

Proof. By Lemma B.5,

$$4d_{\mathrm{TV}}(\mathcal{N}(\hat{\mu},\hat{\sigma}^2),\mathcal{N}(\mu,\sigma^2))^2 \;\le\; \frac{\hat{\sigma}^2}{\sigma^2} - 1 - \log\Big(\frac{\hat{\sigma}^2}{\sigma^2}\Big) + \frac{|\mu-\hat{\mu}|^2}{\sigma^2} \;\le\; \Big(\frac{\hat{\sigma}}{\sigma}\Big)^2 - 1 - \log\left(\Big(\frac{\hat{\sigma}}{\sigma}\Big)^2\right) + \varepsilon^2.$$

Since $z := \hat{\sigma}/\sigma \in [1-\varepsilon, 1+\varepsilon]$ and $\varepsilon \le 2/3$, using the inequality $x^2 - 1 - \log(x^2) \le 3(x-1)^2$ valid for all $|x-1| \le 2/3$, we find

$$d_{\mathrm{TV}}(\mathcal{N}(\hat{\mu},\hat{\sigma}^2),\mathcal{N}(\mu,\sigma^2))^2 \;\le\; \frac{1}{4}(3(z-1)^2 + \varepsilon^2) \;\le\; \frac{1}{4}(4\varepsilon^2) \;=\; \varepsilon^2.$$

The lemma follows since the L^1 distance is symmetric and is equal to twice the TV distance. □

Definition B.7. A random variable X is said to be σ-subgaussian if $\mathbb{P}[|X| \ge t] \le 2\exp(-t^2/\sigma^2)$ for all $t > 0$.

For instance if $X \sim \mathcal{N}(0,1)$ then X is $\sqrt{2}$-subgaussian, see, e.g., Abramowitz and Stegun [6, formula (7.1.13)].

Lemma B.8 (Theorem 3.1.1 in [138])**.** *Let $g \sim \mathcal{N}(0, I_d)$. Then $(\|g\|_2 - \sqrt{d})$ is $O(1)$-subgaussian. Consequently, $(\|g\|_2 - \sqrt{d})_+$ is also $O(1)$-subgaussian.*

Lemma B.9 (Proposition 2.5.2 in [138])**.** *There exist absolute positive constants C_1, C_2 with the following properties. A random variable X is σ-subgaussian if $\sup_{p\ge1} p^{-1/2}\,(\mathbb{E}|X|^p)^{1/p} \le C_1\sigma$. Conversely, if $\sup_{p\ge1} p^{-1/2}\,(\mathbb{E}|X|^p)^{1/p} \le C_2\sigma$ then X is σ-subgaussian.*

Lemma B.10 (Hoeffding's Inequality, Proposition 2.6.1 in [138])**.** *Let $X_1, \dots, X_n$ be independent, mean-zero random variables and suppose X_i is σ_i-subgaussian. Then, for some global constant $c > 0$ and any $t \ge 0$,*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i\right| > t\right] \;\le\; 2\exp\left(\frac{-ct^2}{\sum_{i=1}^{n}\sigma_i^2}\right).$$

Lemma B.11. *Let $g_1, \ldots, g_n \sim \mathcal{N}(0,1)$ and $a_1, \ldots, a_n > 0$. Then, there is a global constant $c > 0$ such that for every $t \geq 0$,*

$$\mathbb{P}\left[\left|\sum_{i=1}^n a_i g_i^2 - \mathbb{E}\left(\sum_{i=1}^n a_i g_i^2\right)\right| \geq t\right] \;\leq\; 2\exp\left(-c\min\left\{\frac{t^2}{\sum_{i=1}^n a_i^2}, \frac{t}{\max_i a_i}\right\}\right).$$

Proof. This statement follows from Bernstein's inequality for subexponential random variables (Theorem 2.8.1 in [138]). □

Theorem B.12 (Gordon's Theorem, Theorem 5.32 in [137])**.** *For a matrix A, let $\sigma_{\min}(A)$ denote the smallest positive singular value of A. Let G be an $m \times n$ matrix with entries independently drawn from $\mathcal{N}(0,1)$. Then $\mathbb{E}\sigma_{\min}(G) \geq \sqrt{m} - \sqrt{n}$.*

B.2 Concentration Inequalities

Lemma B.13 ([99, Lemma 1])**.** *Let X have the chi-squared distribution with parameter d; that is, $X = \sum_{i=1}^d X_i^2$ where the X_i are i.i.d. standard normal. Then,*

$$\begin{aligned}\mathbb{P}[X - d \geq 2\sqrt{dt} + 2t] &\leq \exp(-t) \text{ and}\\ \mathbb{P}[d - X \geq 2\sqrt{dt}] &\leq \exp(-t).\end{aligned}$$

Lemma B.14 (Corollary 5.50 in [137])**.** *There exist an absolute constant C with the following property. Let $X_1, \ldots, X_m \sim \mathcal{N}(0, I_d)$ and let $\varepsilon \in (0,1)$. Suppose that $t \geq 1$. If $m \geq Ct^2 d/\varepsilon^2$ then*

$$\mathbb{P}\left[\left\|\frac{1}{m}\sum_{i=1}^m X_i X_i^\top - I_d\right\| > \varepsilon\right] \;<\; 2\exp(-t^2 d).$$

B.3 Other standard facts

Definition B.15 (ε-net)**.** Let $\varepsilon \geq 0$. We say $N \subseteq X$ is an ε-net for X in metric d if for each $x \in X$ there exists some $y \in N$ such that $d(x,y) \leq \varepsilon$.

Lemma B.16 (Corollary 4.2.13 in [138])**.** *For any $\varepsilon \in (0,1]$, there exists an ε-net for B_2^d in ℓ_2 metric of size $(3/\varepsilon)^d$.*

Recall the ℓ_∞ metric between $(x_1, \ldots, x_d)$ and $(y_1, \ldots, y_d)$ is defined as $\max_i |x_i - y_i|$.

Lemma B.17. *For any $\varepsilon \in (0,1]$ there exists an ε-net for $[-1,1]^d$ in ℓ_∞ metric of size ε^{-d}.*

Proof. Partition $[-1,1]^d$ into ε^{-d} cubes of side-length 2ε. The cube centers form an ε-net for $[-1,1]^d$ in ℓ_∞. □

B.4 Proof of Lemma 4.17

We first give a high-level idea of the proof. Let g be the target distribution and suppose there exists $\rho \geq 0$ and $f \in k\text{-mix}(\mathcal{F})$ such that $\|g - f\|_1 \leq \rho$. Since $f \in k\text{-mix}(\mathcal{F})$, we can write $f = \sum_{i \in [k]} w_i f_i$,

where $f_i \in \mathcal{F}$, $w_i \geq 0$, and $\sum_{i\in[k]} w_i = 1$. A first attempt would be to try to write $g = \sum_{i\in[k]} w_i g_i$ such that each $\|g_i - f_i\| \leq r$; if this were true, then given a sufficient number of samples from f, we would have sufficient samples from each f_i, and then we could use an r-robust compression scheme for each f_i to output some $\hat{g}_i$ satisfying $\|g_i - \hat{g}_i\|_1 \leq \varepsilon$. Alas, it is not clear whether we can ensure that $\|g_i - f_i\|_1 \leq r$ for all i. However, Lemma B.18 below asserts that we can write $g = \sum_{i\in[k]} w_i g_i$ in such a way that for each i, either $\|g_i - f_i\| \leq r$ or w_i is small (in fact, the *sum* of all such weights is small) and, hence, their contribution to the TV distance is small. Thus, we will only need to deal with the case where $\|g_i - f_i\| \leq r$, a task for which r-robust compression is well-suited.

Lemma B.18. *Let g be a density and suppose there exists $f = \sum_{i\in[k]} w_i f_i$ with $(w_1, \ldots, w_k) \in \Delta_k$ and each $f_i \in \mathcal{F}$ such that $\|g - f\|_1 \leq \rho$ for some $\rho \geq 0$. Then we can write $g = \sum_{i\in[k]} w_i g_i$ such that, for any $r > 0$,*

$$\sum_{i\,:\,\|g_i - f_i\|_1 > r} w_i \;<\; \rho/r.$$

The proof of this lemma appears below in Appendix B.4.1. We now turn to proving Lemma 4.17.

Proof of Lemma 4.17. Let g be the target distribution. Let $f \in k\text{-mix}(\mathcal{F})$ such that $\|f - g\|_1 \leq \rho$ for some $\rho \geq 0$. Let $g = \sum_{i\in[k]} w_i g_i$ be the representation given by Lemma B.18. The learner first takes $M = 160m(\varepsilon/10)\log(3k/\delta)k/\varepsilon$ samples from g. Let S be the set of these samples. We view g as a mixture of the g_i, so S can be partitioned into k subsets such that the ith subset has distribution g_i. We learn each of the components individually. The learner does not know which sample point comes from which component, but it can try all possible ways of partitioning S into k subsets, hence generating several 'candidate distributions', such that at least one of them is close to g. Moreover, the learner also 'guesses' the weights w_i as follows: let W be an $(\varepsilon/10k)$-net in ℓ_∞ for Δ_k of size $(10k/\varepsilon)^k$ (see Lemma B.17). So there exists some point $(\widehat{w}_1, \ldots, \widehat{w}_k) \in W$ such that $\max_i |w_i - \widehat{w}_i| \leq \varepsilon/10k$.

We say component i is *tiny* if $w_i < \varepsilon/20k$, and we say component i is *far* if $\|g_i - f_i\|_1 > r$. We say a component is *nice* if it is neither far nor tiny. The sum of weights of tiny components is at most $\varepsilon/20$, and the sum of weights of far components is at most ρ/r by Lemma B.18.

The number of samples from component i is binomial with mean Mw_i. By a Chernoff bound and a union bound over nice components, with probability at least $1-\delta/3$, there are at least $m(\varepsilon/10)\log(3k/\delta)$ points from each nice component. If this is the case, then the definition of robust compression implies that for each such component g_i, with probability at least $1-\delta/3k$ there exists a sequence $L_i \in S^{\tau(\varepsilon/10)}$ and a sequence $B_i \in \{0,1\}^{t(\varepsilon/10)}$ such that $\|\mathcal{J}(L_i, B_i) - f_i\|_1 \leq \varepsilon/10$, where $\mathcal{J}$ is the corresponding decoder. By a union bound over nice components, this is simultaneously true for all nice components, with probability at least $1 - \delta/3$.

Thus far we have proved that with probability at least $1 - 2\delta/3$ there exist sequences $L_1, \ldots, L_k \in S^{\tau(\varepsilon/10)}$ and $B_1, \ldots, B_k \in \{0,1\}^{t(\varepsilon/10)}$ such that $\|\mathcal{J}(L_i, B_i) - f_i\|_1 \leq \varepsilon/10$ for each nice component i. The learner builds the following set of candidate distributions:

$$\mathcal{C} \;\coloneqq\; \left\{ \sum_{i=1}^{k} \widehat{w}_i \mathcal{J}(L_i, B_i) \;:\; L_1, \ldots, L_k \in S^{\tau(\varepsilon/10)},\, B_1, \ldots, B_k \in \{0,1\}^{t(\varepsilon/10)},\, (\widehat{w}_1, \ldots, \widehat{w}_k) \in W \right\}.$$

We claim that with probability at least $1-2\delta/3$ at least one of the distributions in $\mathcal{C}$ is $(3\varepsilon/10+2\rho/r+\rho)$-close to g. This corresponds to the 'correct' sequences L_i, B_i, and $\widehat{w}_i$. To show this, let T denote the set of tiny components, let F denote the set of far components, and let N denote the nice components. Then we have

$$\begin{aligned}
\left\|\sum_{i\in[k]} \widehat{w}_i \mathcal{J}(L_i,B_i) - w_i g_i\right\|_1 &\leq \left\|\sum_{i\in[k]} w_i(\mathcal{J}(L_i,B_i)-f_i)\right\|_1 + \left\|\sum_{i\in[k]} (\widehat{w}_i - w_i)\mathcal{J}(L_i,B_i)\right\|_1 + \|f-g\|_1 \\
&\leq \sum_{i\in T\cup F} w_i \left\|\mathcal{J}(L_i,B_i)-f_i\right\|_1 + \sum_{i\in N} w_i \left\|\mathcal{J}(L_i,B_i)-f_i\right\|_1 \\
&+ \sum_{i\in[k]} |\widehat{w}_i - w_i| \cdot \left\|\mathcal{J}(L_i,B_i)\right\|_1 + \rho \\
&\leq \sum_{i\in T\cup F} w_i \cdot 2 + \sum_{i\in N} w_i \cdot (\varepsilon/10) + \sum_{i\in[k]} (\varepsilon/10k) + \rho \\
&\leq (\varepsilon/10 + 2\rho/r) + \varepsilon/10 + \varepsilon/10 + \rho,
\end{aligned}$$

where the first two inequalities are by the triangle inequality and the last inequality is by definition of tiny and far. This proves the claim.

Next the learner applies the algorithm of Theorem 3.12 (with error parameter $\varepsilon/40$) to obtain a member of $\mathcal{C}$ whose distance from g is bounded by $3\cdot(3\varepsilon/10+2\rho/r+\rho)+4(\varepsilon/40) \leq \varepsilon + 3\rho(1+2/r)$, as required. The overall failure probability is bounded by $2\delta/3$ (probability of the claim failing) plus $\delta/3$ (the probability that algorithm of Theorem 3.12 fails).

The sample complexity of the algorithm is bounded as follows. The number of candidate distributions can be bounded by

$$|\mathcal{C}| \;\leq\; \left(M^{\tau(\varepsilon/10)} 2^{t(\varepsilon/10)}\right)^k \cdot (10k/\varepsilon)^k \leq M^{k\tau'(\varepsilon/10)} \cdot (10k/\varepsilon)^k,$$

whence the total sample complexity can be bounded by

$$\begin{aligned}
&M + \frac{\log(3|\mathcal{C}|^2/\delta)}{2\varepsilon^2} \\
&\quad= O\left(m\Big(\frac{\varepsilon}{10}\Big)\log\Big(\frac{k}{\delta}\Big)\frac{k}{\varepsilon} + \frac{\log(1/\delta) + k\log(k/\varepsilon) + k\tau'(\varepsilon/10)\log\left(m(\frac{\varepsilon}{10})\log(k/\varepsilon)k/\varepsilon\right)}{\varepsilon^2}\right) \\
&\quad= \widetilde{O}\left(\frac{km(\varepsilon/10)}{\varepsilon} + \frac{k\tau'(\varepsilon/10)\log m(\varepsilon/10)}{\varepsilon^2}\right),
\end{aligned}$$

completing the proof. □

B.4.1 Proof of Lemma B.18

Let $\mathcal{X} \coloneqq \{x : g(x) < f(x)\}$. Our goal is to "transform" each f_i into another density g_i so that $g = \sum_{i\in[k]} w_i g_i$. Note that $\mathcal{X}$ consists of the domain points on which f exceeds g. Hence, to transform each f_i into g_i, we would "scale it down multiplicatively" on points in $\mathcal{X}$, and "scale it up additively"

on points not in $\mathcal{X}$. These transformations need to be done carefully for each function g_i to end up being non-negative and integrate to 1.

To that end, we define

$$g_i(x) \;\coloneqq\; \begin{cases} f_i(x)g(x)/f(x) & \text{for } x \in \mathcal{X}, \\ f_i(x) + \Delta_i(x) & \text{for } x \notin \mathcal{X}, \end{cases}$$

where

$$\Delta_i(x) \;\coloneqq\; \big(g(x) - f(x)\big) \left(\int_{\mathcal{X}} f_i(y) \cdot \frac{f(y) - g(y)}{f(y)} \,\mathrm{d}y \right) \Big/ \int_{\mathcal{X}} \big(f(y) - g(y)\big) \,\mathrm{d}y.$$

Recall that Z is the domain of g and the densities in $\mathcal{F}$. We now check that each g_i is a density and that $g = \sum_{i \in [k]} w_i g_i$.

Claim B.19. *For all $i \in [k]$, g_i is a density on Z.*

Proof. We first check that $g_i(x) \geq 0$ for all x. If $x \in \mathcal{X}$, then $g_i(x) \geq 0$ because f_i, g, f are all densities and hence non-negative. If $x \notin \mathcal{X}$, then $\Delta_i(x) \geq 0$ because $g(x) - f(x) \geq 0$ on $\mathcal{X}^c$ and $f(x) - g(x) \geq 0$ on $\mathcal{X}$. We now check that $\int_Z g_i(x) \,\mathrm{d}x = 1$. Since both g and f are densities, both integrate to 1 over Z, and therefore

$$\int_{\mathcal{X}^c} (g(x) - f(x)) \,\mathrm{d}x \;=\; \int_{\mathcal{X}} (f(x) - g(x)) \,\mathrm{d}x. \tag{B.1}$$

The following calculation completes the proof.

$$\begin{aligned}
\int_{\mathcal{X}^c} g_i(x) \,\mathrm{d}x \;&=\; \int_{\mathcal{X}^c} (\Delta_i(x) + f_i(x)) \,\mathrm{d}x \\
&=\; \frac{\int_{\mathcal{X}^c} (g(x) - f(x)) \,\mathrm{d}x}{\int_{\mathcal{X}} (f(y) - g(y)) \,\mathrm{d}y} \cdot \int_{\mathcal{X}} \left(f_i(y) \cdot \frac{f(y) - g(y)}{f(y)} \right) \mathrm{d}y + \int_{\mathcal{X}^c} f_i(x) \,\mathrm{d}x \\
&=\; \int_{\mathcal{X}} \left(f_i(y) \cdot \frac{f(y) - g(y)}{f(y)} \right) \mathrm{d}y + \int_{\mathcal{X}^c} f_i(x) \,\mathrm{d}x \\
&=\; \int_{\mathcal{X}} f_i(y) \cdot \left(1 - \frac{g(y)}{f(y)} \right) \mathrm{d}y + \int_{\mathcal{X}^c} f_i(y) \,\mathrm{d}y \\
&=\; \int_{\mathcal{X}} f_i(y) \,\mathrm{d}y - \int_{\mathcal{X}} f_i(y) \cdot \left(\frac{g(y)}{f(y)} \right) \mathrm{d}y + \int_{\mathcal{X}^c} f_i(y) \,\mathrm{d}y \\
&=\; 1 - \int_{\mathcal{X}} f_i(y) \cdot \frac{g(y)}{f(y)} \,\mathrm{d}y \\
&=\; 1 - \int_{\mathcal{X}} g_i(y) \,\mathrm{d}y
\end{aligned}$$

In the above calculations, the second equality is by definition of Δ_i, the third equality is by (B.1), the sixth equality is because f_i is a density, and the last equality is by definition of g_i. □

Claim B.20. $g = \sum_{i \in [k]} w_i g_i$.

Proof. First suppose $x \in \mathcal{X}$. Since $\sum_{i\in[k]} w_i f_i = f$, we have

$$\sum_{i\in[k]} w_i g_i(x) \;=\; \sum_{i\in[k]} w_i f_i(x)\frac{g(x)}{f(x)} \;=\; g(x).$$

On the other hand, for $x \notin \mathcal{X}$ we have

$$\begin{aligned}
\sum_{i\in[k]} w_i g_i(x) &= \sum_{i\in[k]} w_i \Delta_i(x) + w_i f_i(x) \\
&= \sum_{i\in[k]} w_i \left(\frac{(g(x)-f(x))}{\int_{\mathcal{X}} (f(y)-g(y))\,\mathrm{d}y} \cdot \int_{\mathcal{X}} \left(f_i(y) \cdot \frac{f(y)-g(y)}{f(y)} \right) \mathrm{d}y \right) + \sum_{i\in[k]} w_i f_i(x) \\
&= \frac{(g(x)-f(x))}{\int_{\mathcal{X}} (f(y)-g(y))\,\mathrm{d}y} \cdot \int_{\mathcal{X}} \left(\sum_{i\in[k]} w_i f_i(y) \cdot \frac{f(y)-g(y)}{f(y)} \right) \mathrm{d}y + f(x) \\
&= \frac{(g(x)-f(x))}{\int_{\mathcal{X}} (f(y)-g(y))\,\mathrm{d}y} \cdot \int_{\mathcal{X}} (f(y)-g(y))\,\mathrm{d}y + f(x) \\
&= g(x) - f(x) + f(x) \;=\; g(x),
\end{aligned}$$

where the first equality is by definition of g_i, the second equality is by definition of Δ_i, and second last equality is by because $\sum_{i\in[k]} w_i f_i = f$. □

Let $I \coloneqq \{\, i \in [k] \,:\, \|f_i - g_i\|_1 > r \,\}$. It remains to show that $\sum_{i\in I} w_i < \rho/r$. Observe from the definition of the g_i that we also have $\mathcal{X} = \{\, x \,:\, g_i(x) < f_i(x) \,\}$ for each $i \in [k]$. Thus, using Claim B.20,

$$\|f-g\|_1 \;=\; 2\int_{\mathcal{X}} (f(x)-g(x))\,\mathrm{d}x \;=\; 2\sum_{i\in[k]} w_i \int_{\mathcal{X}} (f_i(x)-g_i(x))\,\mathrm{d}x \;=\; \sum_{i\in[k]} w_i \|f_i - g_i\|_1.$$

Thus, from the hypothesis of the lemma,

$$\rho \;\geq\; \|f-g\|_1 \;=\; \sum_{i\in[k]} w_i \|f_i-g_i\|_1 \;\geq\; \sum_{i\in I} w_i \|f_i-g_i\|_1 \;>\; \sum_{i\in I} w_i r,$$

by definition of I. This gives $\sum_{i\in I} w_i < \rho/r$, as required.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org. → page 5

[2] Yasin Abbasi-Yadkori, Peter L. Bartlett, and Victor Gabillon. Near minimax optimal players for the finite-time 3-expert prediction problem. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3033–3042, 2017. → pages 3, 11

[3] Sepehr Abbasi-Zadeh, Nikhil Bansal, Guru Guruganesh, Aleksandar Nikolov, Roy Schwartz, and Mohit Singh. Sticky Brownian rounding and its applications to constraint satisfaction problems. *arXiv preprint arXiv:1812.07769*, 2018. → page 11

[4] Jacob D. Abernethy, Rafael M. Frongillo, and Andre Wibisono. Minimax option pricing meets black-scholes in the limit. In *Proceedings of the 44th Symposium on Theory of Computing Conference*, pages 1029–1040. ACM, 2012. → page 11

[5] Milton Abramowitz and Irene A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965. → pages 18, 19

[6] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York; National Bureau of Standards, Washington, DC, 1984. ISBN 0-471-80007-4. Reprint of the 1972 edition, Selected Government Publications, available at http://people.math.sfu.ca/~cbm/aands/. → page 94

[7] Jayadev Acharya, Ilias Diakonikolas, Chinmay Hegde, Jerry Li, and Ludwig Schmidt. Fast and near-optimal algorithms for approximating distributions by histograms. In *Proceedings of PODS*, 2015. → page 56

[8] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017. → page 5

[9] Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James Voss. The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. In *Conference on Learning Theory*, pages 1135–1164, 2014. → page 5

[10] Martin Anthony and Peter Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, 1999. → page 56

[11] Sanjeev Arora and Ravi Kannan. Learning mixtures of separated nonspherical gaussians. *Ann. Appl. Probab.*, 15(1A):69–92, 02 2005. doi:10.1214/105051604000000512. URL https://doi.org/10.1214/105051604000000512. → page 55

[12] Sanjeev Arora, Elad Hazan, and Satyen Kale. $o(\sqrt{\log n})$ approximation to sparsest cut in $\tilde{O}(n^2)$ time. *SIAM Journal on Computing*, 39(5):1748–1771, 2010. → page 2

[13] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012. → pages 2, 8

[14] 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. → page v

[15] Hassan Ashtiani, Shai Ben-David, and Abbas Mehrabian. Sample-efficient learning of mixtures. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. → pages ix, 5, 7, 54, 56, 57

[16] Mohammad Zokaei Ashtiani. *A PAC-Theory of Clustering with Advice*. PhD thesis, University of Waterloo, 2018. → pages v, 54, 60, 69

[17] S. S. Barsov and V. V. Ul'yanov. Estimates of the proximity of Gaussian measures. *Sov. Math., Dokl.*, 34:462–466, 1987. ISSN 0197-6788. → page 81

[18] Erhan Bayraktar, Ibrahim Ekren, and Yili Zhang. On the asymptotic optimality of the comb strategy for prediction with expert advice. *arXiv preprint arXiv:1902.02368*, 2019. → page 11

[19] Erhan Bayraktar, Ibrahim Ekren, and Xin Zhang. Finite-time 4-expert prediction problem. *Communications in Partial Differential Equations*, pages 1–44, 2020. → pages 3, 11

[20] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 103–112, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4244-7. doi:10.1109/FOCS.2010.16. → page 55

[21] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013. → page 1

[22] Olivier Bousquet, Daniel Kane, and Shay Moran. The optimal approximation factor in density estimation. In *Conference on Learning Theory*, pages 318–341, 2019. → page 48

[23] Leo Breiman. First exit times for a square root boundary. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Contributions to Probability Theory, Part 2*, pages 9–16. University of California Press, 1967. → pages 4, 13, 16, 17, 31, 33

[24] Leo Breiman. *Probability*. SIAM, 1992. → page 31

[25] Monica Brezzi and Tze Leung Lai. Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control*, 27(1):87–108, 2002. ISSN 0165-1889. → page 11

[26] Sébastien Bubeck. Introduction to online optimization, December 2011. unpublished. → pages 2, 3, 4, 9

[27] Sébastien Bubeck, Michael B. Cohen, Yin Tat Lee, James R. Lee, and Aleksander Madry. k-server via multiscale entropic regularization. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 3–16. ACM, 2018. → page 11

[28] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discrete Comput. Geom.*, 59(4):757–783, June 2018. ISSN 0179-5376. doi:10.1007/s00454-018-9992-1. URL https://doi.org/10.1007/s00454-018-9992-1. → page 11

[29] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6): 1740–1766, 2011. → pages 10, 11

[30] Nicolò Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59(3):392–411, 1999. → pages 8, 11

[31] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006. → pages 2, 3, 9, 30

[32] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3): 427–485, 1997. → pages 3, 8, 9, 10, 11

[33] Siu-On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pages 604–613, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2710-7. doi:10.1145/2591796.2591848. → pages 56, 57

[34] Kamalika Chaudhuri, Sanjoy Dasgupta, and Andrea Vattani. Learning mixtures of gaussians using the k-means algorithm. *arXiv preprint arXiv:0912.0086*, 2009. → page 5

[35] Kamalika Chaudhuri, Yoav Freund, and Daniel J. Hsu. A parameter-free hedging algorithm. In *Advances in Neural Information Processing Systems 22*, pages 297–305, 2009. → page 16

[36] Chandra Chekuri, TS Jayram, and Jan Vondrák. On multiplicative weight updates for concave and submodular function maximization. In *Proceedings of the Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 201–210, 2015. → page 11

[37] Herman Chernoff. Optimal stochastic control. *Sankhyā: The Indian Journal of Statistics, Series A*, 30:221–252, 1968. → page 15

[38] Paul Christiano, Jonathan A Kelner, Aleksander Madry, Daniel A Spielman, and Shang-Hua Teng. Electrical flows, laplacian systems, and faster approximation of maximum flow in undirected graphs. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 273–282, 2011. → page 2

[39] Thomas M. Cover. Behavior of sequential predictors of binary sequences. In *Proceedings of the 4th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*. Publishing House of the Czechoslovak Academy of Sciences, Prague, 1965. → pages 3, 9, 10, 11, 12

[40] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006. ISBN 978-0-471-24195-9; 0-471-24195-4. → page 82

[41] Joseph A Cruz and David S Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2:117693510600200030, 2006. → page 1

[42] Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999. → pages 5, 6, 55

[43] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007. → page 5

[44] Sanjoy Dasgupta and Leonard J Schulman. A two-round variant of em for gaussian mixtures. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 152–159, 2000. → page 5

[45] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Conference on Learning Theory*, pages 1183–1213, 2014. → pages 5, 56

[46] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. In *Conference on Learning Theory*, pages 704–710, 2017. → page 5

[47] Burgess Davis. On the intergrability[1] of the martingale square function. *Israel Journal of Mathematics*, 8:187–190, 1970. → page 32

[48] Burgess Davis. On the L_p norms of stochastic integrals and other martingales. *Duke Math. J*, 43(4):697–704, 1976. → pages 4, 11, 13, 17

[49] Peter M. DeMarzo, Ilan Kremer, and Yishay Mansour. Online trading algorithms and robust option pricing. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 477–486. ACM, 2006. → page 11

[50] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. → page 5

[51] Luc Devroye. *A course in density estimation*, volume 14 of *Progress in Probability and Statistics*. Birkhäuser Boston, Inc., Boston, MA, 1987. ISBN 0-8176-3365-0. → page 56

[52] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001. ISBN 0-387-95117-2. doi:10.1007/978-1-4613-0125-7. → page 56

[53] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012. → pages 48, 62

[54] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The minimax learning rate of normal and Ising undirected graphical models. *arXiv preprint arXiv:1806.06887*, 2018. → page 57

[1] This appears to be a typographical error in the title of the paper.

[55] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*, 2018. → page 81

[56] Ilias Diakonikolas. Learning Structured Distributions. In Peter Bühlmann, Petros Drineas, Michael Kane, and Mark van der Laan, editors, *Handbook of Big Data*, chapter 15, pages 267–283. Chapman and Hall/CRC, 2016. → page 53

[57] Ilias Diakonikolas, Elena Grigorescu, Jerry Li, Abhiram Natarajan, Krzysztof Onak, and Ludwig Schmidt. Communication-efficient distributed learning of discrete probability distributions. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. → page 56

[58] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017. → page 5

[59] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84, Oct 2017. doi:10.1109/FOCS.2017.16. Available on arXiv:1611.03473 [cs.LG]. → page 55

[60] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Learning multivariate log-concave distributions. In *Proceedings of Machine Learning Research*, volume 65 of *COLT'17*, pages 1–17, 2017. ISBN 3-540-35294-5, 978-3-540-35294-5. URL http://proceedings.mlr.press/v65/diakonikolas17a/diakonikolas17a.pdf. → page 56

[61] Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019. → page 11

[62] J. L. Doob. *Classical Potential Theory and Its Probabilistic Counterparts*. Springer-Verlag, 1984. → page 36

[63] Nadeja Drenska. *A PDE approach to a Prediction Problem Involving Randomized Strategies*. PhD thesis, New York University, 2017. → page 11

[64] Nadejda Drenska and Robert V Kohn. Prediction with expert advice: A PDE perspective. *Journal of Nonlinear Science*, 30(1):137–173, 2020. → page 11

[65] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, fifth edition, 2019. → pages 15, 31

[66] Ronen Eldan and Assaf Naor. Krivine diffusions attain the Goemans–Williamson approximation ratio. *arXiv preprint arXiv:1906.10615*, 2019. → page 11

[67] Jon Feldman, Rocco A. Servedio, and Ryan O'Donnell. PAC learning axis-aligned mixtures of Gaussians with no separation assumption. In *Proceedings of the 19th Annual Conference on Learning Theory*, COLT'06, pages 20–34, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-35294-5, 978-3-540-35294-5. doi:10.1007/11776420_5. → page 56

[68] Jon Feldman, Rocco A Servedio, and Ryan ODonnell. Pac learning axis-aligned mixtures of gaussians with no separation assumption. In *International Conference on Computational Learning Theory*, pages 20–34. Springer, 2006. → pages 5, 7, 58

[69] William Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, second edition, 1971. → page 33

[70] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995. → page 2

[71] Naveen Garg and Jochen Koenemann. Faster and simpler algorithms for multicommodity flow and other fractional packing problems. *SIAM Journal on Computing*, 37(2):630–652, 2007. → page 2

[72] Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 761–770, 2015. → page 5

[73] Sébastien Gerchinovitz. *Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation techniques*. PhD thesis, Université Paris-Sud, 2011. → pages 2, 3, 4, 9

[74] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995. → page 10

[75] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics*. Addison-Wesley, 1994. → page 10

[76] Nick Gravin, Yuval Peres, and Balasubramanian Sivan. Towards optimal algorithms for prediction with expert advice. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 528–547. SIAM, 2016. → pages 9, 11, 13

[77] Nick Gravin, Yuval Peres, and Balasubramanian Sivan. Tight Lower Bounds for Multiplicative Weights Algorithmic Families. In Ioannis Chatzigiannakis, Piotr Indyk, Fabian Kuhn, and Anca Muscholl, editors, *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*, volume 80 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 48:1–48:14, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-041-5. doi:10.4230/LIPIcs.ICALP.2017.48. URL http://drops.dagstuhl.de/opus/volltexte/2017/7499. → pages 8, 9

[78] Hayit Greenspan, Amit Ruf, and Jacob Goldberger. Constrained gaussian mixture model framework for automatic segmentation of mr brain images. *IEEE transactions on medical imaging*, 25(9):1233–1245, 2006. → page 5

[79] Priscilla Greenwood and Edwin Perkins. A conditioned limit theorem for random walk and brownian local time on square root boundaries. *Annals of Probability*, 11:227–261, 1983. → pages 4, 13, 16, 17, 33, 34

[80] Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes*. Oxford University Press, third edition, 2001. → page 31

[81] Venkatesan Guruswami, Atri Rudra, and Madhu Sudan. Essential coding theory, 2019. available at https://cse.buffalo.edu/faculty/atri/courses/coding-theory/book/. → page 81

[82] James Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957. → pages 1, 8

[83] Nicholas J. A. Harvey, Christopher Liaw, Edwin Perkins, and Sikander Randhawa. Optimal anytime regret with two experts. *arXiv preprint arXiv:2002.08994*, 2020. → page v

[84] Leslie Hogben, editor. *Handbook of linear algebra*. Discrete Mathematics and its Applications (Boca Raton). CRC Press, Boca Raton, FL, second edition, 2014. ISBN 978-1-4665-0728-9. → page 80

[85] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018. → page 5

[86] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20, 2013. → page 5

[87] Adam Kalai, Ankur Moitra, and Gregory Valiant. Disentangling Gaussians. *Communications of the ACM*, 55(2), 2012. → pages 55, 57

[88] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010. → page 5

[89] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Disentangling gaussians. *Communications of the ACM*, 55(2):113–120, 2012. → page 5

[90] Satyen Kale. Boosting and hard-core set constructions: a simplified approach. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 131. Citeseer, 2007. → page 2

[91] Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan Ullman. Differentially private algorithms for learning mixtures of separated gaussians. In *Advances in Neural Information Processing Systems*, pages 168–180, 2019. → page 5

[92] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the Twenty-sixth Annual ACM Symposium on Theory of Computing*, STOC '94, pages 273–282, New York, NY, USA, 1994. ACM. ISBN 0-89791-663-8. doi:10.1145/195058.195155. URL http://doi.acm.org/10.1145/195058.195155. → page 56

[93] Robert Kleinberg, Georgios Piliouras, and Éva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 533–542, 2009. → page 11

[94] Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer, 2008. → pages 18, 22, 31, 32, 73

[95] Vladimir A. Kobzar, Robert V. Kohn, and Zhilei Wang. New potential-based bounds for prediction with expert advice. *arXiv preprint arXiv:1911.01641*, 2019. → page 11

[96] Vladimir A. Kobzar, Robert V. Kohn, and Zhilei Wang. New potential-based bounds for the geometric-stopping version of prediction with expert advice. *arXiv preprint arXiv:1912.03132*, 2019. → page 11

[97] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015. → page 1

[98] Jeongyeol Kwon and Constantine Caramanis. Em algorithm is sample-optimal for learning mixtures of well-separated gaussians. *arXiv preprint arXiv:2002.00329*, 2020. → page 5

[99] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5), 2000. → page 95

[100] Yin Tat Lee and Santosh S. Vempala. Geodesic walks in polytopes. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 927–940. ACM, 2017. → page 11

[101] Jerry Li and Ludwig Schmidt. Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In *Conference on Learning Theory*, pages 1302–1382, 2017. → page 5

[102] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. In *Proceedings of FOCS*, pages 256–261, 1989. → page 10

[103] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994. → pages 2, 8

[104] Alexander E. Litvak, Alain Pajor, Mark Rudelson, and Nicole Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Adv. Math.*, 195(2): 491–523, 2005. ISSN 0001-8708. URL https://doi.org/10.1016/j.aim.2004.08.004. → page 71

[105] Alexander E. Litvak, Alain Pajor, Mark Rudelson, and Nicole Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Adv. Math.*, 195(2): 491–523, 2005. ISSN 0001-8708. URL https://doi.org/10.1016/j.aim.2004.08.004. → page 70

[106] Haipeng Luo and Robert E. Schapire. Towards minimax online learning with unknown time horizon. In *Proceedings of ICML*, 2014. → pages v, 9, 11, 12

[107] Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: AdaNormalHedge. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1286–1304, 2015. → page 16

[108] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. → page 1

[109] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010. → pages x, 5, 6, 7

[110] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 93–102, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4244-7. doi:10.1109/FOCS.2010.15. → page 55

[111] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009. → page 9

[112] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894. → pages 1, 5

[113] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. → page 5

[114] Edwin Perkins. On the Hausdorff dimension of the Brownian slow points. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 64:369–399, 1983. → pages 4, 9, 10, 13, 20, 33

[115] Claudia Perlich, Brian Dalessandro, Troy Raeder, Ori Stitelman, and Foster Provost. Machine learning for targeted display advertising: Transfer learning in action. *Machine learning*, 95(1): 103–127, 2014. → page 1

[116] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4): 695–706, 2006. → page 5

[117] Goran Peskir and Albert Shiryaev. *Optimal Stopping and Free-Boundary Problems*. Birkhäuser Verlag, 2006. → page 15

[118] Serge A Plotkin, David B Shmoys, and Éva Tardos. Fast approximation algorithms for fractional packing and covering problems. *Mathematics of Operations Research*, 20(2):257–301, 1995. → page 2

[119] Alexander Rakhlin and Karthik Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. In *Conference on Learning Theory*, pages 1704–1722, 2017. → page 4

[120] Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. → page 93

[121] Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017. → page 5

[122] Rolf-Dieter Reiss. *Approximate distributions of order statistics with applications to nonparametric statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1989. ISBN 0-387-96851-2. URL https://doi.org/10.1007/978-1-4613-9620-8. → page 65

[123] Kan Ren, Weinan Zhang, Ke Chang, Yifei Rong, Yong Yu, and Jun Wang. Bidding machine: Learning to bid for directly optimizing profits in display advertising. *IEEE Transactions on Knowledge and Data Engineering*, 30(4):645–659, 2017. → page 1

[124] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013. → pages 36, 91

[125] Eitan Richardson and Yair Weiss. On gans and gmms. In *Advances in Neural Information Processing Systems*, pages 5847–5858, 2018. → page 5

[126] L. C. G. Rogers and David Williams. *Diffusions, Markov Processes and Martingales. Volume 1: Foundations*. Cambridge University Press, second edition, 2000. → page 31

[127] L. C. G. Rogers and David Williams. *Diffusions, Markov Processes and Martingales. Volume 2: Itô Calculus*, volume 2. Cambridge University Press, second edition, 2000. → pages 91, 92

[128] Walter Rudin. *Principles of Mathematical Analysis*. John Wiley & Sons, third edition, 1976. → page 89

[129] Arora Sanjeev and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257, 2001. → page 5

[130] L. A. Shepp. A first passage problem for the Wiener process. *The Annals of Mathematical Statistics*, 38(6):1912–1914, 1967. → pages 16, 31

[131] Bernard W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986. ISBN 0-412-24620-1. → page 56

[132] Mohit Singh and Lap Chi Lau. Approximating minimum bounded degree spanning trees to within one of optimal. *J. ACM*, 62(1):1:1–1:19, 2015. → page 10

[133] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 1395–1403, 2014. → pages 5, 7, 54

[134] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical Gaussian mixtures. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1395–1403. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5251-near-optimal-sample-estimators-for-spherical-gaussian-mixtures.pdf. → pages 57, 76, 82

[135] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. → page 47

[136] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004. → page 5

[137] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012. → page 95

[138] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.html. → pages 94, 95

[139] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017. → page 1

[140] Volodimir G. Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990. → pages 2, 8

[141] Yu Wang, Igor V Tetko, Mark A Hall, Eibe Frank, Axel Facius, Klaus FX Mayer, and Hans W Mewes. Gene selection from microarray data for cancer classificationa machine learning approach. *Computational biology and chemistry*, 29(1):37–46, 2005. → page 1

[142] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47): E7351–E7358, 2016. → page 11

[143] David Williams. *Probability with Martingales*. Cambridge University Press, 1991. → page 31

[144] Neal E Young. Randomized rounding without solving the linear program. In *Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 170–178. Society for industrial and applied mathematics, 1995. → page 2

[145] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997. → page 76