Generalizing Complex Hypotheses on Product Distributions: Auctions, Prophet Inequalities, and Pandora's Problem

Chenghao Guo CHENGHAO@MIT.EDU
Massachusetts Institute of Technology.

Zhiyi Huang ZHIYI@CS.HKU.HK
The Univeristy of Hong Kong.

Zhihao Gavin Tang TANG.ZHIHAO@MAIL.SHUFE.EDU.CN
ITCS, Shanghai University of Finance and Economics.

Xinzhi Zhang XINZHI20@CS.WASHINGTON.EDU
University of Washington.

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

This paper explores a theory of generalization for learning problems on product distributions, complementing the existing learning theories in the sense that it does not rely on any complexity measures of the hypothesis classes. The main contributions are two general sample complexity bounds: (1) $\tilde{O}\left(\frac{nk}{\epsilon^2}\right)$ samples are sufficient and necessary for learning an ϵ-optimal hypothesis in *any problem* on an n-dimensional product distribution, whose marginals have finite supports of sizes at most k; (2) $\tilde{O}\left(\frac{n}{\epsilon^2}\right)$ samples are sufficient and necessary for any problem on n-dimensional product distributions if it satisfies a notion of strong monotonicity from the algorithmic game theory literature. As applications of these theories, we match the optimal sample complexity for single-parameter revenue maximization (Guo et al., STOC 2019), improve the state-of-the-art for multi-parameter revenue maximization (Gonczarowski and Weinberg, FOCS 2018) and prophet inequality (Correa et al., EC 2019; Rubinstein et al., ITCS 2020), and provide the first and tight sample complexity bound for Pandora's problem.

Keywords: Generalization, Sample Complexity, Auctions, Prophet Inequalities, Pandora's Problem

1. Introduction

> *The learning process is a process of choosing an appropriate function from a given set of functions.* — *Vapnik (1998)*

Generalization is widely recognized as one of the fundamental pillars of learning theory. A general learning problem asks whether we can select a function, often referred to as a hypothesis, from a hypothesis class to maximize or minimize the expectation w.r.t. an underlying distribution over the data domain, based on samples from the distribution. While it may be easy to select a hypothesis that maximizes or minimizes the average over the samples, how can we ensure that it *generalizes* and gets a similar performance on the underlying distribution? More quantitatively, we may ask about its sample complexity: *how many samples are sufficient and necessary for choosing a hypothesis that is optimal on the true distribution up to an ϵ error?*

A widely studied example is the classification problem in supervised learning. In this problem, each data point is a feature-label pair $(x, y) \in X \times Y$, where X is the feature domain, e.g., $\mathbb{R}^n$,

and Y is the label domain, e.g., $\{0, 1\}$. Each hypothesis corresponds to a classifier, i.e., a feature-to-label mapping $f : X \mapsto Y$; its value on a data point (x, y) is $L\big(f(x), y\big)$ for some loss function L, e.g., $\big|f(x) - y\big|$. The goal is to learn a classifier from samples to minimize the expected loss on the underlying distribution.

Meanwhile, the general learning problem also captures a wide range of optimization problems in the Bayesian model. The problem of learning revenue-maximizing auctions from data is a recent example, which has received a lot of attention in algorithmic game theory and more generally in theoretical computer science. In this example, each data point comprises the valuations of the bidders; each hypothesis corresponds to an auction and its value is defined to be the revenue of the auction on the given valuations. We aim to learn an auction from sample valuations to maximize the expected revenue on the underlying value distributions.

1.1. Generalization from Complexity Measures of the Hypothesis Class

Most sample complexity bounds in learning theory rely on detailed structures of the hypothesis class $\mathcal{H}$, and they hold for arbitrary distributions over the data domain. In particular, they build on various complexity measures of the hypothesis class, including the covering number Anthony and Bartlett (2009), Vapnik-Chervonenkis (VC) dimension Vapnik and Chervonenkis (2015), Natarajan dimension Natarajan (1989), pseudo-dimension Pollard (1990), fat-shattering dimension Bartlett et al. (1996), Rademacher complexity Bartlett and Mendelson (2002); Koltchinskii and Panchenko (2000), local Rademacher complexity Bartlett et al. (2002), etc. Informally, each complexity measure provides a parameter d which represents the "degrees-of-freedom" of the hypothesis class $\mathcal{H}$, and the corresponding sample complexity upper bound has the form $\tilde{O}\big(\frac{d}{\epsilon^2}\big)$.

For example, the VC dimension characterizes the sample complexity of binary classification problems, and the Natarajan dimesnion captures that of multiclass classification problems (see, e.g., Shalev-Shwartz and Ben-David (2014)), *if the underlying distribution could be arbitrary*.

Further, the example of learning revenue-optimal auctions from data, in particular, the special case of selling a single item to n bidders, has been investigated using the covering number (e.g., Devanur et al. (2016), Gonczarowski and Nisan (2017)), pseudo-dimension (e.g., Morgenstern and Roughgarden (2015)), and Rademacher complexity (e.g., Syrgkanis (2017)). The "degrees-of-freedom" bounds in these works are all $\tilde{O}\big(\frac{n}{\epsilon}\big)$ and thus, lead to the same $\tilde{O}\big(\frac{n}{\epsilon^3}\big)$ sample complexity upper bound.[1] The bound once again holds for *arbitrary distributions of the valuations, even correlated ones*, although the problem of revenue-optimal auction design often considers product value distributions.

1.2. Our Contributions: A Theory of Generalization on Product Distributions

While the above theories suggest that simpler hypotheses generalize better, it has been increasingly important to consider complex ones. On the one hand, deep neural networks generalize surprisingly well on classification problems on real-world data despite their complexity (e.g., Zhang et al. (2017)). On the other hand, the revenue-optimal auction for selling even two heterogeneous items could have an infinite menu complexity (e.g., Daskalakis et al. (2013)). To this end, this paper asks:

1. Nonetheless, these works are different in that they either prove sample complexity bounds for different families of distributions beyond the $[0, 1]$-bounded ones, and/or provide slightly different bounds in the logarithmic factors.

Is there a complementary theory of generalization building on the simplicity of data instead of the hypothesis class?

In particular, it is standard to assume that the data is drawn from a product distribution in optimization problems in the Bayesian model, including the aforementioned revenue-maximization problem, prophet inequality and Pandora's problem. Can we get sample complexity bounds from the independence of data dimensions, and only minimum knowledge about the hypotheses?

Implicit Attempts in Previous Works. We first review several recent sample complexity bounds for specific optimization problems that implicitly explore the power of independent data dimensions. Cole and Roughgarden (2014) and Roughgarden and Schrijvers (2016) used independence in the single-parameter revenue maximization problems to analyze coordinate-wise the convergence of the empirical distribution to the true distribution. Correa et al. (2019) employed a similar approach on prophet inequality. Their analyses of convergence, however, are problem dependent.

Cai and Daskalakis (2017) proposed a hybrid argument that used the independence of data dimensions in multi-item auctions to derive sample complexity bounds from complexity measures of the hypothesis class w.r.t. each coordinate of the data domain. Their hybrid approach benefits from independence, yet still relies on complexity measures of the hypothesis class.

Gonczarowski and Weinberg (2018) and Guo et al. (2019) are the closest to this paper. Gonczarowski and Weinberg (2018) exploited independence in multi-parameter revenue maximization to construct an improved covering number that holds specifically on product distributions. Although they did not explicitly ask the above conceptual question, their techniques implicitly showed generalization of complex hypotheses on product distributions; the resulting bound is inferior in the logarithmic factor compared to Theorem 5 in this paper. Guo et al. (2019) used independent data dimensions and a notion of monotonicity to derive optimal sample complexity bounds for single-parameter revenue maximization problems in the matroid setting. It is the closest to this paper. Part of our results can be viewed as generalizing theirs to all problems with the same notion of monotonicity.

Our Results. The contributions of the paper are two general sample complexity bounds from the independence of data dimensions, unrelated to any complexity measures of the hypothesis class. Both results use the same algorithm which we call the *product empirical reward maximizer/risk minimize* (PERM). It selects the best hypothesis w.r.t. a product empirical distribution such that each coordinate is a uniform distribution over the corresponding coordinate of the samples. This is different from the usual notion of empirical distribution, i.e., the uniform distribution over sample vectors, which is not a product distribution in general. The first result considers finite data domains.

Informal Theorem *Suppose the data has n independent dimensions, each of which takes up to k possible values. Then, $O\left(\frac{nk}{\epsilon^2}\log\frac{1}{\delta}\right)$ samples are sufficient for a uniform convergence of any hypothesis class up to ϵ-optimality with probability at least $1-\delta$. In other words, simultaneously and uniformly for every hypothesis, its expectation under E is ϵ-close to its expectation under D. Further, $\Omega\left(\frac{nk}{\epsilon^2}\right)$ samples are necessary.*

The proof of the upper bound is simple in hindsight. We bound the total variation distance between the product empirical distribution and the true distribution, using the connection between the total variation distance and the Hellinger distance, and a vector concentration inequality.

Despite its simplicity, the above theorem gives strong sample complexity upper bounds for the aforementioned optimization problems. In single-parameter revenue maximization, e.g., single-item

	This paper		Previous results
	Finite domain (§3)	Strong monotonicity (§4)	
General bound	$\boldsymbol{O\left(\frac{nk}{\epsilon^2}\log\frac{1}{\delta}\right)}$	$\boldsymbol{\tilde{O}\left(\frac{n}{\epsilon^2}\right)}$	-
Single-parameter	$O\left(\frac{n}{\epsilon^3}\log\frac{1}{\delta}\right)$	$\boldsymbol{\tilde{O}\left(\frac{n}{\epsilon^2}\right)}$	$\tilde{O}\left(\frac{n}{\epsilon^2}\right)$
Multi-parameter	$\boldsymbol{O\left(\frac{n}{\epsilon^4}\log\frac{1}{\delta}\right)}$	-	$O\left(\frac{n}{\epsilon^4}\log\frac{n}{\epsilon\delta}\right)$
Prophet inequality	$O\left(\frac{n}{\epsilon^3}\log\frac{1}{\delta}\right)$	$\boldsymbol{\tilde{O}\left(\frac{n}{\epsilon^2}\right)}$	$\tilde{O}\left(\frac{n^2}{\epsilon^2}\right)$, $\tilde{O}\left(\frac{n}{\epsilon^6}\right)$
Pandora's problem	$O\left(\frac{n^3}{\epsilon^3}\log\frac{1}{\delta}\right)$	$\boldsymbol{\tilde{O}\left(\frac{n}{\epsilon^2}\right)}$	-

Table 1: Summary of sample complexity upper bounds, in comparisons with the state-of-the-art. The results in bold are the best upper bounds in different settings. We use single-item auctions, and n-item auctions with a unit-demand bidder as the running examples of single- and multi-parameter revenue maximization. The bounds may vary in other settings; see Sec. 3 and 4. The previous results for the three settings are from Guo et al. (2019), Gonczarowski and Weinberg (2018), and Correa et al. (2019), Rubinstein et al. (2020) respectively.

auctions, the value can be discretized to multiples of ϵ Devanur et al. (2016); replacing $k = \frac{1}{\epsilon}$ gives an $O\left(\frac{n}{\epsilon^3}\log\frac{1}{\delta}\right)$ upper bound. Hence, using only that the value domain could be discretized, the theorem matches the sample complexity upper bounds mentioned in Section 1.1, which relied on detailed knowledge of single-parameter auctions such as its various complexity measures. In fact, the above theorem improves in the log factor.

In multi-parameter revenue maximization, e.g., with one unit-demand bidder and n items, the value domain can be discretized to multiples of ϵ^2 Balcan et al. (2008); hence, letting $k = \frac{1}{\epsilon^2}$ gives an $O\left(\frac{n}{\epsilon^4}\log\frac{1}{\delta}\right)$ upper bound, improving the state-of-the-art by Gonczarowski and Weinberg (2018) in the log factor.[2]

Further, we show that the type domain in prophet inequality can be discretized to multiples of ϵ, leading to an $O\left(\frac{n}{\epsilon^3}\log\frac{1}{\delta}\right)$ upper bound. It improves the previous bound by Correa et al. (2019) in the dependence in n, and the bound by Rubinstein et al. (2020) in the dependence in ϵ. In prophet inequality with i.i.d. rewards in particular, it implies that $\tilde{O}(n)$ samples are sufficient to learn a 0.745-competitive algorithm. Concurrent to Rubinstein et al. (2020), this answers an open question by Correa et al. (2019).[3]

Finally, we show that the type domain of Pandora's problem can also be discretized to multiples of ϵ, giving the first polynomial sample complexity bound for the problem.

Our second result revisits the notion of strong monotonicity, a key ingredient of the optimal sample complexity bounds for single-parameter revenue maximization by Guo et al. (2019). Strong monotonicity means that the expected value of the optimal hypothesis w.r.t. a distribution $\tilde{\mathbf{D}}$ does not decrease when it is applied to another distribution $\mathbf{D}$ that stochastically dominates $\tilde{\mathbf{D}}$. We generalize the analysis by Guo et al. (2019) to any strongly monotone problem.

2. Gonczarowski and Weinberg (2018) only claim polynomial sample complexity; the stated bound is derived using their techniques to the best of our efforts.
3. The best algorithm, with full knowledge of the distribution, is strictly better than 0.745-competitive. Hence, we may consider ϵ a constant in this result. Further, unlike our model, Correa et al. (2019) consider unbounded distributions and multiplicative approximation; nonetheless, Appendix C.4 shows how to get the stated bounds in their model.

Informal Theorem *Suppose the data has n independent dimensions. Then, $\tilde{\Theta}\left(\frac{n}{\epsilon^2}\right)$ samples are sufficient and necessary for learning an arbitrary strongly monotone hypothesis class.*

Further, we show that prophet inequality and Pandora's problem are both strongly monotone. Using this theorem, we get an $\tilde{O}\left(\frac{n}{\epsilon^2}\right)$ upper bound for single-parameter revenue maximization, prophet inequality, and Pandora's problem.[4] The linear dependence on the data dimension n is tight for all three problems. We remark that while the bound for single-parameter revenue maximization is the same as that by Guo et al. (2019), ours directly uses the PERM, which corresponds to the empirical Myerson auction, while that by Guo et al. (2019) needs an appropriate regularization to the product empirical distributions and uses the regularized empirical Myerson auction.

2. Preliminaries

2.1. Model

A *general learning problem* (e.g., Chapter 1.4 of Vapnik (2013)) is defined by a *hypothesis class* denoted as $\mathcal{H}$. We will abuse notation and refer to the problem defined by a hypothesis class $\mathcal{H}$ as problem $\mathcal{H}$. Each *hypothesis* $h \in \mathcal{H}$ is a mapping from $\mathbf{T} = T_1 \times T_2 \times \cdots \times T_n$ to $[0, 1]$, where $T_i \subseteq \mathbb{R}$ is the domain of the i-th coordinate of the data type. We will refer to n as the *data dimension* of the problem, to make a distinction with various learning dimensions in the literature which measure the complexity of the hypothesis class. For a concrete running example, readers may think of $T_1 = T_2 = \cdots = T_n = [0, 1]$. For any data type $\mathbf{t} \in \mathbf{T}$, and any hypothesis $h \in \mathcal{H}$, $h(\mathbf{t})$ is the reward obtained by hypothesis h on a data point of type $\mathbf{t}$.

Given a distribution $\mathbf{D}$ over $\mathbf{T}$, we seek to pick a hypothesis $h \in \mathcal{H}$ to maximize the expected reward: $h\big(\mathbf{D}\big) \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{t}\sim\mathbf{D}}\big[h\big(\mathbf{t}\big)\big]$.

Further, let $\textsc{Opt}_{\mathcal{H}}\big(\mathbf{D}\big)$ denote the optimal expected reward. We will omit the subscript $\mathcal{H}$ for brevity when the hypothesis class is clear from the context. $\textsc{Opt}_{\mathcal{H}}\big(\mathbf{D}\big) \stackrel{\text{def}}{=} \sup_{h\in\mathcal{H}} h\big(\mathbf{D}\big)$.

Throughout this paper, we will assume that $\mathbf{D} = D_1 \times D_2 \times \cdots \times D_n$ is a product distribution, as it is a standard assumption in all examples considered in the paper. See Section 2.2 for details.

A *learning algorithm* for a problem $\mathcal{H}$ takes N i.i.d. samples from the underlying distribution $\mathbf{D}$ as input and returns a hypothesis $h \in \mathcal{H}$. Let E_i denote the uniform distribution over the i-th coordinate of the samples. We call $\mathbf{E} = E_1 \times E_2 \times \cdots \times E_n$ the *product empirical distribution*, and its optimal hypothesis $h_{\mathbf{E}}$ the *product empirical reward maximizer* (PERM).

For any $0 \leq \epsilon \leq 1$, a hypothesis h is an *ϵ-additive approximation* if: $h\big(\mathbf{D}\big) \geq \textsc{Opt}\big(\mathbf{D}\big) - \epsilon$.

The *sample complexity* of a problem $\mathcal{H}$ is the minimum number of samples N for which there is a learning algorithm so that, *for any distribution* $\mathbf{D}$, it takes N i.i.d. samples and returns an ϵ-additive approximation with probability at least $1 - \delta$. The sample complexity bounds in this paper depend on the data dimension n, the approximation parameter ϵ, and the confidence parameter δ . In our first set of results, they further depend on the sizes of the data type domain T_i's. Importantly, they are independent of any complexity measure of the hypothesis class $\mathcal{H}$. We remark that when all hypotheses are $[0, H]$-bounded, our sample complexity bounds imply $\epsilon \cdot H$-additive approximations.

4. Concurrently and independently, Fu and Lin (2020) also proved an $\tilde{O}\left(\frac{n}{\epsilon^2}\right)$ upper bound for Pandora's problem.

2.2. Examples

Next, we define several example problems for which the theory developed in this paper improves or matches the state-of-the-art sample complexity bounds. We define each problem only with the minimum detail necessary for verifying that it is a special case of the above model. In particular, we intentionally do not characterize the optimal hypothesis to stress the main feature of our theory: *it requires almost no knowledge of the hypothesis class*; instead, it only needs that (1) $\mathbf{D}$ is a product distribution, and (2) some generic structural property, e.g., the data type domain can be discretized, or the problem satisfies strong monotonicity, which will be discussed in more details in Section 4.

Single-parameter Revenue Maximization. For simplicity of exposition, we use single-item auctions with n bidders as the running example. Each bidder i has a type $t_i \in [0, 1]$ that represents its value for the item, and is drawn independently from D_i. An auction A maps any type profile $\mathbf{t}$ to an allocation $\mathbf{x} \in [0, 1]^n$, $\|\mathbf{x}\|_1 \leq 1$, and a payment vector $\mathbf{p} \in [0, 1]^n$. For any bidder i, x_i is the probability that the bidder gets the item, and p_i is its payment. Its utility equals $x_i t_i - p_i$.

A bidder's type is private information known only to itself; therefore, the auctioneer must ask the bidders to report the values. Hence, the literature focuses on dominant-strategy incentive compatible (DSIC) auctions, which ensure that for any bidder, reporting the value truthfully always maximizes its utility. The goal is to pick a DSIC auction to maximize the expected revenue. Readers are referred to Myerson (1981) for a characterization of the revenue optimal auction.

To place it in our framework, define the hypothesis class $\mathcal{H}$ by having a hypothesis h_A for every DSIC auction A, such that $h_A(\mathbf{t})$ equals the revenue of auction A on a type profile $\mathbf{t}$.

Readers who are familiar with auction theory may verify that the techniques in Section 3 apply to arbitrary single-parameter problems, and those in Section 4 apply to the matroid setting.[5]

Multi-parameter Revenue Maximization. For simplicity of exposition, we use single-bidder auctions with n items as the running example. Section 3 will discuss the extension to multi-bidder multi-item auctions. The bidder has a type $\mathbf{t} \in [0, 1]^n$ such that t_i is its value for item i, and is drawn independently from D_i. There are various settings with different definitions of the bidder's value for subsets of items. The most-studied ones are the *unit-demand* bidder, whose value for a subset of items is the maximum value for a single item in the subset, and the *additive* bidder, whose value is the sum of item values it gets. An auction A maps any type $\mathbf{t}$ to a subset of items to be allocated to the bidder and its payment. The bidder's utility is equal to its value for the allocated subset minus the payment. The goal is to pick a DSIC auction to maximize the expected revenue. Readers are referred to Cai et al. (2012) for an LP-based characterization of the optimal auction.

Similar to the single-parameter setting, define the hypothesis class $\mathcal{H}$ by having a hypothesis h_A for every DSIC auction A, such that $h_A(\mathbf{t})$ equals the revenue of auction A on type $\mathbf{t}$.

Prophet Inequality. Consider n rewards which arrive one at a time; each reward $t_i \in [0, 1]$ is drawn independently from D_i. On observing each reward, the algorithm must immediately decide whether to take it or not; it can take at most one reward. The goal is to maximize the expected reward. Readers are referred to Samuel-Cahn (1984) for an algorithm that gets at least a half of the expected max reward, and Correa et al. (2017) for an improved algorithm in the case of i.i.d. rewards that gets a 0.745 fraction of the expected max reward. Some readers may know it as the optimal stopping problem, while it is often known as prophet inequality in theoretical computer science.

5. This part of our results rely on a notion called strong revenue monotonicity, which we will discuss in more details in Section 4. It is only known to hold in the matroid setting. Whether it further generalizes is an open problem.

To put it in our framework, define the hypothesis class $\mathcal{H}$ by having a hypothesis h_A for every algorithm A, such that $h_A(\mathbf{t})$ equals what the algorithm gets when the reward sequence is $\mathbf{t}$.

Pandora's Problem. Consider n boxes; each box i has a reward $t_i \in [0,1]$ drawn from D_i and a fixed cost $c_i \in [0,1]$ for opening it. In each round, the algorithm decides whether to take the best observed reward, or to open a new box. The goal is to maximize the reward it gets minus the total cost. To interpret it in our framework, define the hypothesis class $\mathcal{H}$ by having a hypothesis h_A for every algorithm A. If we let $h_A(\mathbf{t})$ equals what the algorithm gets minus the cost when the reward sequence is $\mathbf{t}$, its range would be $[-n,1]$ instead of $[0,1]$. In the main text, we use a simple normalization, which let h_A be the above value plus n and scaled by $\frac{1}{n+1}$. Appendix D.4 presents a more specialized method that gives the tight sample complexity bound.

2.3. Metrics for Probability Distributions

Consider two distributions P, Q over a sample domain T. For concreteness, think of T as a cube in the Euclidean space, e.g., $[0,1]$ or $[0,1]^n$.

Total Variation Distance. The *total variation distance* is a half of the L_1 distance: $\delta\big(P,Q\big) = \frac{1}{2}\|P-Q\|_1$.

The following useful fact about total variation distance follows by its definition.

Lemma 1 *For any distributions P, Q over a sample domain T, and any function $h : T \mapsto [0,1]$:* $\big|h\big(P\big) - h\big(Q\big)\big| \le \delta\big(P,Q\big)$.

Recall that we are interested in multi-dimensional product distributions. It is hard to directly measure the total variation distance among such distributions. The standard method is to instead consider either the Kullback-Leibler divergence or the Hellinger distance; they are both additive in that we can account for the distance in each coordinate separately, and both can be related to the total variation distance. This paper uses the latter because it has better properties.

Hellinger Distance. The *Hellinger distance* between P and Q, denoted as $H\big(P,Q\big)$, is given by: $H^2\big(P,Q\big) = \frac{1}{2}\int_T \big(\sqrt{dP} - \sqrt{dQ}\big)^2$.

More formally, for any measure λ over T so that both P and Q are absolutely continuous w.r.t. λ, let $\frac{dP}{d\lambda}$ and $\frac{dQ}{d\lambda}$ be the Radon-Nikodym derivatives. We have: $H^2\big(P,Q\big) = \frac{1}{2}\int_T \left(\sqrt{\frac{dP}{d\lambda}} - \sqrt{\frac{dQ}{d\lambda}}\right)^2 d\lambda$.

For example, if P and Q are continuous over $[0,1]$ with density functions p and q, or if P and Q are distributions over a discrete set T with probability mass functions p and q, we have:

$$H^2\big(P,Q\big) = \frac{1}{2}\int_0^1 \big(\sqrt{p(t)} - \sqrt{q(t)}\big)^2 dt \quad \text{or} \quad H^2\big(P,Q\big) = \frac{1}{2}\sum_{t\in T}\big(\sqrt{p(t)} - \sqrt{q(t)}\big)^2 .$$

The next two lemmas relate the Hellinger distance with the total variation distance, and formalize its additivity. Readers are referred to Gibbs and Su (2002) for details of these properties and a comprehensive discussion on different metrics for probability distributions.

Lemma 2 *Suppose P and Q are distributions over a sample domain T. Then, we have:* $H^2\big(P,Q\big) \le \delta\big(P,Q\big) \le \sqrt{2}H\big(P,Q\big)$.

Lemma 3 *Suppose $\mathbf{P}$ and $\mathbf{Q}$ are product distributions over $\mathbf{T} = T_1 \times T_2 \times \cdots \times T_n$. Then,* $1 - H^2\big(\mathbf{P},\mathbf{Q}\big) = \prod_{i=1}^n \big(1 - H^2\big(P_i,Q_i\big)\big)$.

2.4. Vector Concentration Inequality

We will use the following Bernstein-style concentration inequality that bounds the ℓ_2-norm of the sum of independent random vectors.

Lemma 4 (Equation 6.12 of Ledoux and Talagrand (1991)) *Let $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_N}$ be N i.i.d. random vectors in $\mathbb{R}^d$ such that for all $i \in [N]$, $\mathbb{E}\left[\|\mathbf{X_i}\|_2^2\right] \leq \sigma^2$, and $\|\mathbf{X_i}\|_2 \leq M$ for some constant $M > 0$. Then, for any positive Δ:*

$$\mathbf{Pr}\left[\left|\left\|\sum_{i=1}^N \mathbf{X_i}\right\|_2 - \mathbb{E}\left[\left\|\sum_{i=1}^N \mathbf{X_i}\right\|_2\right]\right| > \Delta\right] \leq 2\exp\left(-\frac{\Delta^2}{2N\sigma^2}\left(2 - \exp(\frac{2M\Delta}{N\sigma^2})\right)\right).$$

3. Problems with Finite (Discretized) Domain

3.1. General Sample Complexity Bounds

This section offers a theory of generalization for arbitrary hypotheses with finite domain.

Theorem 5 *For any product distribution $\mathbf{D}$ over $\mathbf{T}$ such that $|T_i| \leq k$ for all $1 \leq i \leq n$, suppose for some sufficiently large constant $C > 0$, the number of samples is at least: $C \cdot \frac{nk}{\epsilon^2} \log \frac{1}{\delta}$. Then, with probability at least $1 - \delta$, for any $h : \mathbf{T} \to [0, 1]$, we have $\left|h(\mathbf{D}) - h(\mathbf{E})\right| \leq \epsilon$. In particular, the PERM is an ϵ-additive approximation.*

The proof is simple in hindsight. The key is bounding the convergence of the product empirical distribution to the true distribution in terms of the Hellinger distance, as in the next lemma. We remark two natural attempts to prove Theorem 5 and Lemma 3 which would lead inferior bounds. First, one may consider a hybrid argument that switches D_i to E_i one coordinate at a time and bounds the resulting generalization error; see, e.g., Cai and Daskalakis (2017) for applications of this strategy on related but different results. It would lead to an additional factor n and a worse log factor which depends on n. Second, one may apply standard concentration inequalities and the union bound to cap the deviation for each data type in each coordinate, and calculate the resulting Hellinger distance. See the first version of this paper for this strategy. It would lead to a worse log factor, which depends on both k and n.

Lemma 6 *When the number of samples is N, with probability at least $1-\delta$, we have: $H^2(\mathbf{D}, \mathbf{E}) = O\left(\frac{nk}{N} \log \frac{1}{\delta}\right)$.*

We proceed with the proof of Theorem 5 assuming the correctness of the lemma, whose proof is deferred to Appendix A.1.

Proof of Theorem 5: By Lemma 6 and the stated number of samples, with probability at least $1 - \delta$, $H^2(\mathbf{D}, \mathbf{E}) \leq \frac{\epsilon^2}{2}$. Further by Lemma 2, we have $\delta(\mathbf{D}, \mathbf{E}) \leq \sqrt{2} \cdot H(\mathbf{D}, \mathbf{E}) \leq \epsilon$. Then, the theorem follows by Lemma 1. ■

We complement Theorem 5 with a matching lower bound up to a logarithmic factor. The proof is deferred to Appendix A.2.

Theorem 7 *There exists a sufficiently small constant $c > 0$, such that for all $\epsilon \in [0, 1]$, there exists a problem $\mathcal{H}$ on a finite domain $\mathbf{T} = T_1 \times \cdots \times T_n$ with $|T_i| = k$ for all $1 \leq i \leq n$, such that no algorithm gives an expected ϵ-additive approximation with probability larger than $\frac{2}{3}$ if the number of input samples of the algorithm is less than: $c \cdot \frac{nk}{\epsilon^2}$.*

3.2. Applications

Although some problems are defined on continuous domains, most can be discretized, including all examples defined in Section 2.2. Observe that by rounding each sample to the closest discretized value, we effectively sample from the discretized distribution. Next we discuss the applications of Theorem 5 on the examples; the discretization arguments are either from previous works, or deferred to the appendix because they do not provide much insight. While some results will be subsumed by those in the next section, they are already comparable with the state-of-the-art in meaningful ways. We restress that our bounds are obtained knowing effectively nothing about the problems other than that the domains can be discretized, while previous works generally rely on detailed problem structures. Since it only rely on an appropriate discretization of the date domain, we believe the same approach can be applied to other problems not covered in this paper.

Single-parameter Revenue Maximization. Devanur et al. (2016) showed that we may w.l.o.g. round values down to multiples of ϵ in single-parameter revenue maximization if the target is an $O(\epsilon)$-additive approximation. Hence, with $k = \frac{1}{\epsilon}$, Theorem 5 matches the previous results by Morgenstern and Roughgarden (2015), Devanur et al. (2016), and Syrgkanis (2017) discussed in Section 1.1, i.e., the best bounds before the recent work of Guo et al. (2019). In fact, our bound is better in the logarithmic factors.

Theorem 8 *In a single-item auction with n bidders whose values are bounded in $[0, 1]$, the sample complexity is at most $O\big(\frac{n}{\epsilon^3}\log\frac{1}{\delta}\big)$.*

Multi-parameter Revenue Maximization. We consider the case for selling n-items to a unit-demand (resp., additive) bidder. It is known that rounding the item values down to multiples of ϵ^2 (resp., ϵ^2/n) for a unit-demand (resp. additive) bidder is w.l.o.g. due to a reduction from approximate DSIC auctions to DSIC auctions, which states any ϵ^2-DSIC mechanism can be transformed into a DSIC mechanism with at most ϵ loss in revenue (see, e.g., Balcan et al. (2008), attributed to Nisan). Therefore, Theorem 5 gives the following bounds that improve the best known results by Gonczarowski and Weinberg (2018) in the log factor.

Theorem 9 *In a multi-item auction with n items and a unit-demand bidder whose values are bounded in $[0, 1]$, the sample complexity is at most $O\big(\frac{n}{\epsilon^4}\log\frac{1}{\delta}\big)$.*

Theorem 10 *In a multi-item auction with n items and an additive bidder whose values are bounded in $[0, 1]$, the sample complexity is at most $O\big(\frac{n^2}{\epsilon^4}\log\frac{1}{\delta}\big)$.*[6]

Multiple Bidders. For n-bidder m-item auctions, Theorem 5 gives an $\tilde{O}\big(\frac{mnk}{\epsilon^2}\big)$ bound if the buyers are unit-demand and their value domains are finite with size k, improving those by Gonczarowski and Weinberg (2018) in the log factor. For continuous value domains, however, there is no existing transformation from approximate DSIC to DSIC auctions in this more general setting; as a result, the aforementioned discretization no longer works. Either we settle with approximate DSIC auctions, or need to know more about the relation between approximate and exact DSIC auctions, which is a fundamental question on its own in auction theory. Readers are referred to Gonczarowski and Weinberg (2018) for an extensive discussion on this topic.

6. In the additive case, the optimal revenue is bounded by $[0, n]$. The stated bound is an ϵ-approximation w.r.t. the normalized revenue divided by a factor $\frac{1}{n}$. The bound would be $\tilde{O}\big(\frac{n^4}{\epsilon^4}\big)$ for an ϵ-approximation without normalization.

Prophet Inequality. We show that the rewards in prophet inequality can be discretized w.l.o.g. to multiples of ϵ. Hence, letting $k = \frac{1}{\epsilon}$, we get a sample complexity bound that improves the recent work of Correa et al. (2019) in the dependence in n, and the concurrent result by Rubinstein et al. (2020) in the dependence in ϵ, with bounded rewards in $[0, 1]$ and additive approximation. The discretization argument and an extension to the original setting of Correa et al. (2019) with unbounded rewards and multiplicative approximation is deferred to Appendix C.

Theorem 11 *In the prophet inequality setting with n items whose rewards are bounded in $[0, 1]$, the sample complexity is at most $O\big(\frac{n}{\epsilon^3}\log\frac{1}{\delta}\big)$.*

Pandora's Problem. Recall that the main text uses a simple normalization by a factor $\frac{1}{n+1}$ to ensure that the hypotheses in Pandora's problem has range $[0, 1]$. Hence, to get an ϵ-approximation w.r.t. Pandora's problem, we need a $\frac{\epsilon}{n+1}$-approximately optimal hypothesis. Further, we show that the rewards can be w.l.o.g. discretized to multiples of ϵ. Putting together, we get the first polynomial sample complexity for Pandora's problem. The discretization argument and a more specialized method that gives the optimal sample complexity are deferred to Appendix D.

Theorem 12 *In the Pandora's problem with n boxes whose rewards and costs are bounded in $[0, 1]$, the sample complexity is at most $O\big(\frac{n^3}{\epsilon^3}\log\frac{1}{\delta}\big)$.*

4. Strongly Monotone Problems

This section considers the sample complexity of a subset of problems which satisfy a structural property called strong monotonicity, without any restrictions on the supports of the data domain.

Definition 13 (Stochastic Dominance) *Suppose $\mathbf{D}$ and $\tilde{\mathbf{D}}$ are two product distributions distributions over $\mathbb{R}^n$, $\mathbf{D}$ stochastically dominates $\tilde{\mathbf{D}}$ if for any $i \in [n]$ and any $t \in \mathbb{R}$, the CDFs of D_i and $\tilde{D}_i$ satisfy $F_{D_i}(t) \le F_{\tilde{D}_i}(t)$.*

Definition 14 (Strong Monotonicity) *A problem $\mathcal{H}$ is* strongly monotone *if for any $\mathbf{D}$, any $\tilde{\mathbf{D}}$ that is stochastically dominated by $\mathbf{D}$, and the optimal hypothesis $h_{\tilde{\mathbf{D}}}$ of $\tilde{\mathbf{D}}$: $h_{\tilde{\mathbf{D}}}(\mathbf{D}) \ge h_{\tilde{\mathbf{D}}}(\tilde{\mathbf{D}}) = \textsc{Opt}(\tilde{\mathbf{D}})$.*

The name is inherited from the context of single-parameter revenue maximization, where each hypothesis h is a DSIC auction, $\mathbf{v} \sim \mathbf{D}$ is the value profile, and $h(\mathbf{D})$ is the expected revenue of the auction over the random realization of a value profile drawn from $\mathbf{D}$. Then, the above inequality states that running the optimal auction w.r.t. $\tilde{\mathbf{D}}$, a.k.a., Myerson's auction, on a distribution $\mathbf{D}$ that stochastically dominates $\tilde{\mathbf{D}}$, gets at least the optimal revenue w.r.t. the dominated distribution $\tilde{\mathbf{D}}$. This is precisely the notion of *strong revenue monotonicity* introduced by Devanur et al. (2016). The naming is to make a distinction with the existing weaker notion of revenue monotonicity, which only requires that optimal revenue w.r.t. $\mathbf{D}$ to be weakly larger than that w.r.t. $\tilde{\mathbf{D}}$. We restate below the weaker notion in the more general context in this paper.

Definition 15 (Weak Monotonicity) *A problem $\mathcal{H}$ is* weakly monotone *if for any $\mathbf{D}$, and any $\tilde{\mathbf{D}}$ that is stochastically dominated by $\mathbf{D}$: $\textsc{Opt}\big(\mathbf{D}\big) \ge \textsc{Opt}\big(\tilde{\mathbf{D}}\big)$.*

Finally, we remark that there is an even stronger notion of monotonicity which we call hypothesis-wise monotonicity.

Definition 16 (Hypothesis-wise Monotonicity) *A problem $\mathcal{H}$ is* hypothesis-wise monotone *if for any $\mathbf{D} \in \mathbb{R}^n$, any $\tilde{\mathbf{D}} \in \mathbb{R}^n$ that is stochastically dominated by $\mathbf{D}$, and any hypothesis $h \in \mathcal{H}$: $h(\mathbf{D}) \geq h(\tilde{\mathbf{D}})$.*

Clearly, hypothesis-wise monotonicity implies strong monotonicity, which in turns implies weak monotonicity. Weak monotonicity is insufficient for deriving the improved sample complexity bound with the techniques in this section. Hypothesis-wise monotonicity is too restrictive on the other hand; in fact, it fails to hold on any example considered in this paper.

The rest of the section argues that (1) strong monotonicity leads to an improved sample complexity bound, and (2) strong monotonicity holds in all but one examples considered in this paper, improving or matching the state-of-the-art sample complexity bounds. We stress that we did not make any simplifying assumption compared to previous works.

4.1. Sample Complexity Bounds for Strongly Monotone Problems

The main result for strongly monotone problems is the following improved sample complexity upper bound. In particular, the bound is independent of the support size of the distributions and, in fact, applies to continuous distributions.

Theorem 17 *For any strongly monotone problem $\mathcal{H}$, suppose the number of samples is at least $C \cdot \frac{n}{\epsilon^2} \log\left(\frac{n}{\epsilon}\right) \log\left(\frac{n}{\epsilon\delta}\right)$, where $C > 0$ is a sufficiently large constant independent of the problem $\mathcal{H}$. Then, the PERM is an ϵ-additive approximation with probability at least $1 - \delta$.*
Remark: *If the distributions are i.i.d., i.e., $D_i = D^*$ for any $i \in [n]$, it suffices to have the above number of sample values from D^* (rather than vectors from $\mathbf{D}$). Given the samples from D^*, we construct E^* to be the uniform distribution over the samples and let $\mathbf{E} = E^* \times E^* \times \cdots \times E^*$.*

The above upper bound is identical to that in the special case of single-parameter revenue maximization by Guo et al. (2019); the proof is also similar. The contributions of this paper are two-folds. First, we generalize it to arbitrary strongly monotone problems so that it can be further applied to a broader scope of problems, including the prophet inequalities and the Pandora's problem considered in this paper. Second, we show that the empirical maximizer itself achieves the optimal sample complexity bound when the reward function is bounded in $[0, 1]$; in contrast, Guo et al. (2019) need a regularized version of the empirical distributions called the dominated empirical distributions, and uses the corresponding regularized maximizer.

The proof and the applications of Theorem 17 are deferred to Appendix B. We remark that the above upper bound is tight up to a poly-logarithmic factor, due to an existing lower bound in the special case of single-parameter revenue maximization.

Theorem 18 *There is a strongly monotone problem $\mathcal{H}$ so that if the number of samples is less than $c \cdot \frac{n}{\epsilon^2}$, where $c > 0$ is a sufficiently small constant, no algorithm gets an expected ϵ-additive approximation.*

Proof Let $\mathcal{H}$ be the set of DSIC single-item auctions with n bidders. Restrict the bidders' valuations to be bounded in $[0, 1]$ so that the value/revenue of any hypothesis/auction $h \in \mathcal{H}$ on any value

profile is bounded in $[0, 1]$. By Devanur et al. (2016), the single-item revenue maximization problem is strongly monotone. Further by Guo et al. (2019), the sample complexity of $[0, 1]$-bounded valuations and ϵ-additive approximation is at least $\Omega(\frac{n}{\epsilon^2})$. ■

5. Future Directions and Other Related Works

Sample Complexity of Simple Auctions/Hypotheses. A branch of the literature of sample complexity of auctions considers simpler auction formats such as the second-price auction with reserve prices. Readers are referred to Balcan et al. (2018), Cai and Daskalakis (2017), and Morgenstern and Roughgarden (2016) for some examples. We restate that the theories developed in this paper are complementary to the existing ones; they are more suitable for problems with complex hypothesis classes (on product distributions). Hence, this paper does not try to apply the theories to these simpler families of auctions. That said, there are relatively few natural hypothesis classes whose "degrees-of-freedom" are smaller than the data dimensions. Hence, for strongly monotone problems, the sample complexity bound in Theorem 17 is competitive. Finally, we leave as a future question whether there are natural learning problems whose tight sample complexity bounds need both complexity measures of the hypotheses and the independence of data dimensions.

Beyond Product Distributions. Although arbitrarily correlated distributions seem intractable, it may be possible to generalize the theories in this paper to structured corrected distributions, which we leave as another future direction. Concretely, if we can learn from samples an appropriate representation of the data under which different dimensions are independent, we shall be able to combine it with the theories in this paper to get generalization bounds. To this end, the vast literature on principle component analysis (PCA) is related. See, e.g., Pearson (1901) and Jolliffe (2011). Independently, Brustle et al. (2019) made progress on this direction showing how to learn multi-item auctions when the value distribution is correlated yet admit special structures.

Classification Problems. We present a a preliminary result in Appendix E under a strong assumption that the feature distributions are independent *conditioned on any given label*. To further extend the theories in this paper to obtain useful generalization bounds for natural classification problems, we need to relax the assumption of having product conditional feature distributions, which is related to the last research direction. Moreover, although the algorithmic question of finding the optimal hypothesis w.r.t. a product distribution is well-studied for optimization problems in the Bayesian model, little is known about its counterpart for classification problems. In particular, it is unclear whether finding the best hypothesis w.r.t. the product empirical distribution is harder or easier than doing so w.r.t. the original notion of empirical distribution. On the one hand, the product empirical distribution is more structured; on the other hand, its support size is exponential in general, while the support size of the original empirical distribution is upper bounded by the number of samples.

Multi-parameter Auctions and Other Structural Properties. Multi-parameter revenue maximization is the only example in this paper that does not benefit from the improved sample complexity bound in Theorem 17 because it is not strong monotonicity. In fact, Hart and Reny (2015) showed that it is not even weakly monotone. Nonetheless, the hypotheses corresponding to multi-parameter auctions are very different from those used in the proof of the lower bound (Theorem 7). We consider it an interesting open question if there is another structural property (unrelated to the

complexity measures of the hypotheses) which applies to multi-parameter revenue maximization and, ideally, to a large family of problems, which lead to improved sample complexity bounds.

Acknowledgments

This work is supported by Science and Technology Innovation 2030 New Generation of Artificial Intelligence Major Project No.(2018AAA0100903), Program for Innovative Research Team of Shanghai University of Finance and Economics (IRTSHUFE) and the Fundamental Research Funds for the Central Universities. Zhiyi Huang is supported by the Hong Kong Research Grants Council (RGC) under Grant No. 17203717E.

The authors thank anonymous reviewers for pointing out an incorrect interpretation of the sample complexity upper bounds by Gonczarowski and Weinberg (2018) in the first version of the paper. ZH further thanks Rong Ge for helpful discussions on concentration inequalities for vectors.

References

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

Maria-Florina Balcan, Avrim Blum, Jason D Hartline, and Yishay Mansour. Reducing mechanism design to algorithm design via machine learning. *Journal of Computer and System Sciences*, 74 (8):1245–1270, 2008.

Maria-Florina Balcan, Tuomas Sandholm, and Ellen Vitercik. A general theory of sample complexity for multi-item profit maximization. In *Proceedings of the 19th ACM Conference on Economics and Computation*, pages 173–174. ACM, 2018.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized rademacher complexities. In *International Conference on Computational Learning Theory*, pages 44–58. Springer, 2002.

Johaness Brustle, Yang Cai, and Constantinos Daskalakis. Multi-item mechanisms without item-independence: Learnability via robustness. In *21st ACM Conference on Economics and Computation*, 2019.

Yang Cai and Constantinos Daskalakis. Learning multi-item auctions with (or without) samples. In *Proceedings of the 58th IEEE Annual Symposium on Foundations of Computer Science*, pages 516–527. IEEE, 2017.

Yang Cai, Constantinos Daskalakis, and S Matthew Weinberg. Optimal multi-dimensional mechanism design: Reducing revenue to welfare maximization. In *Proceedings of the 53rd IEEE Annual Symposium on Foundations of Computer Science*, pages 130–139. IEEE, 2012.

Richard Cole and Tim Roughgarden. The sample complexity of revenue maximization. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 243–252. ACM, 2014.

José Correa, Patricio Foncea, Ruben Hoeksma, Tim Oosterwijk, and Tjark Vredeveld. Posted price mechanisms for a random stream of customers. In *Proceedings of the 18th ACM Conference on Economics and Computation*, pages 169–186. ACM, 2017.

José Correa, Paul Dütting, Felix Fischer, and Kevin Schewior. Prophet inequalities for i.i.i. random variables from an unknown distribution. In *Proceedings of the 20th ACM Conference on Economics and Computation*, pages 3–17, New York, NY, USA, 2019. ACM.

Constantinos Daskalakis, Alan Deckelbaum, and Christos Tzamos. Mechanism design via optimal transport. In *Proceedings of the 14th ACM Conference on Electronic Commerce*, pages 269–286. ACM, 2013.

Nikhil R Devanur, Zhiyi Huang, and Christos-Alexandros Psomas. The sample complexity of auctions with side information. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pages 426–439. ACM, 2016.

Hu Fu and Tao Lin. Learning utilities and equilibria in non-truthful auctions. *arXiv preprint arXiv:2007.01722*, 2020.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

Yannai A Gonczarowski and Noam Nisan. Efficient empirical revenue maximization in single-parameter auction environments. In *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*, pages 856–868. ACM, 2017.

Yannai A. Gonczarowski and S. Matthew Weinberg. The sample complexity of up-to-ϵ multi-dimensional revenue maximization. In *Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science*, pages 416–426. IEEE, 2018.

Chenghao Guo, Zhiyi Huang, and Xinzhi Zhang. Settling the sample complexity of single-parameter revenue maximization. In *Proceedings of the 51st ACM Symposium on Theory of Computing*. ACM, 2019.

Sergiu Hart and Philip J Reny. Maximal revenue with multiple goods: Nonmonotonicity and other observations. *Theoretical Economics*, 10(3):893–922, 2015.

Ian Jolliffe. *Principal component analysis*. Springer, 2011.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–457. Springer, 2000.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer, 1991.

Jamie Morgenstern and Tim Roughgarden. Learning simple auctions. In *Conference on Learning Theory*, pages 1298–1318, 2016.

Jamie H Morgenstern and Tim Roughgarden. On the pseudo-dimension of nearly optimal auctions. In *Advances in Neural Information Processing Systems*, pages 136–144, 2015.

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

David Pollard. Empirical processes: theory and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–86. JSTOR, 1990.

Tim Roughgarden and Okke Schrijvers. Ironing in the dark. In *Proceedings of the 17th ACM Conference on Economics and Computation*, pages 1–18. ACM, 2016.

Aviad Rubinstein, Jack Z Wang, and S Matthew Weinberg. Optimal single-choice prophet inequalities from samples. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

Ester Samuel-Cahn. Comparison of threshold stop rules and maximum for independent nonnegative random variables. *The Annals of Probability*, pages 1213–1216, 1984.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Vasilis Syrgkanis. A sample complexity measure with applications to learning optimal auctions. In *Advances in Neural Information Processing Systems*, pages 5352–5359, 2017.

Vladimir Vapnik. Statistical learning theory. *New York*, pages 156–160, 1998.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer, 2013.

Vladimir Vapnik and Alexey Y Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of Complexity*, pages 11–30. Springer, 2015.

Martin L Weitzman. Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, pages 641–654, 1979.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *Proceedings of 6th International Conference on Learning Representations*, 2017.

Appendix A. Finite Domain

A.1. Convergence of Product Empirical Distribution: Proof of Lemma 6

We will reduce Lemma 6 to a vector concentration inequality. First:

$$\begin{aligned} 1 - H^2(\mathbf{D}, \mathbf{E}) &= \prod_{i=1}^{n} \left(1 - H^2(D_i, E_i)\right) && \text{(Lemma 3)} \\ &\geq 1 - \sum_{i=1}^{n} H^2(D_i, E_i)\ . \end{aligned}$$

Hence, it suffices to show that with probability at least $1 - \delta$:

$$\sum_{i=1}^{n} H^2(D_i, E_i) \leq O\left(\frac{nk}{N} \log \frac{1}{\delta}\right). \tag{1}$$

By definition of Hellinger distance, for any $1 \leq i \leq n$:

$$H^2(D_i, E_i) = \frac{1}{2} \sum_{t \in T_i} \left(\sqrt{f_{D_i}(t)} - \sqrt{f_{E_i}(t)}\right)^2 = \frac{1}{2} \sum_{t \in T_i} \left(\frac{f_{D_i}(t) - f_{E_i}(t)}{\sqrt{f_{D_i}(t)} + \sqrt{f_{E_i}(t)}}\right)^2.$$

Next, we bound the right-hand-side by the following lemma, which can be viewed a smoothed variant of the connection between the Hellinger distance and χ-square distance.

Lemma 19 *For any $f_D, f_E \in [0,1]$:*

$$\left(\frac{f_D - f_E}{\sqrt{f_D} + \sqrt{f_E}}\right)^2 \leq \frac{\left(f_D - f_E\right)^2}{\max\left\{f_D, \frac{1}{N}\log\frac{1}{\delta}\right\}} + \frac{1}{N}\log\frac{1}{\delta}\,.$$

Proof If $f_E > \frac{1}{N}\log\frac{1}{\delta}$ or $f_D > \frac{1}{N}\log\frac{1}{\delta}$, the left-hand-side is at most the first term on the right-hand-side. Otherwise, the left-hand-side is at most $\max\{f_D, f_E\} \leq \frac{1}{N}\log(\frac{1}{\delta})$. ■

Sum Eqn. (1) over $1 \leq i \leq n$, and apply Lemma 19 to the right-hand-side:

$$\sum_{i=1}^{n} H^2(D_i, E_i) \leq \frac{1}{2}\sum_{i=1}^{n}\sum_{t\in T_i}\frac{\left(f_{D_i}(t) - f_{E_i}(t)\right)^2}{\max\left\{f_{D_i}(t), \frac{1}{N}\log\frac{1}{\delta}\right\}} + \frac{nk}{N}\log\frac{1}{\delta}\,.$$

Therefore, it suffices to show that with probability at least $1-\delta$:

$$\sum_{i=1}^{n}\sum_{t\in T_i}\frac{\left(f_{D_i}(t) - f_{E_i}(t)\right)^2}{\max\left\{f_{D_i}(t), \frac{1}{N}\log\frac{1}{\delta}\right\}} \leq O\left(\frac{nk}{N}\log\frac{1}{\delta}\right)\,. \tag{2}$$

To interpret the left-hand-side as the squared ℓ_2-norm of the sum of i.i.d. vectors, we associate each sample $\mathbf{s_j} \sim \mathbf{D}$, $1 \leq j \leq N$, with a $\sum_{i=1}^{n}|T_i|$-dimensional random vector $\mathbf{X_j}$. Concretely, for any $j \in [N], i \in [n]$ and for any $t \in T_i$:

$$X_{jit} = \frac{\mathbf{1}[s_{ji} = t] - f_{D_i}(t)}{\sqrt{\max\left\{f_{D_i}(t), \frac{1}{N}\log\frac{1}{\delta}\right\}}}\,. \tag{3}$$

Then, Eqn. (2) can be restated as:

$$\left\|\frac{1}{N}\sum_{j=1}^{N}\mathbf{X_j}\right\|_2^2 \leq O\left(\frac{nk}{N}\log\frac{1}{\delta}\right) \quad \text{or equivalently} \quad \left\|\sum_{j=1}^{N}\mathbf{X_j}\right\|_2^2 \leq O\left(Nnk\log\frac{1}{\delta}\right)\,.$$

We establish the following properties of the random vectors.

Lemma 20 *For any $j \in [N]$, the random vector $\mathbf{X_j}$ defined in Eqn. (3) satisfies: (1) $\mathbb{E}\left[\mathbf{X_j}\right] = \mathbf{0}$, (2) $\mathbb{E}\left[\|\mathbf{X_j}\|_2^2\right] \leq nk$, and (3) $\|\mathbf{X_j}\|_2 \leq \sqrt{\frac{Nnk}{\log\frac{1}{\delta}}}$.*

Proof The first property follows by definition. The second one is true because:

$$\mathbb{E}\left[\|\mathbf{X_j}\|_2^2\right] \leq \sum_{i=1}^{n}\sum_{t\in T_i}\frac{\mathbb{E}\left[(\mathbf{1}[\mathbf{s_j}_i = t] - f_{D_i}(t))^2\right]}{f_{D_i}(t)} = \sum_{i=1}^{n}\sum_{t\in T_i}(1 - f_{D_i}(t)) \leq nk\,.$$

The last property follows by $X_{ijk}^2 \leq N/\log\frac{1}{\delta}$. This is why we need the smoothed variant in Lemma 19 instead of the original inequality between the Hellinger and χ-square distances. ■

By Lemma 20, and Lemma 4 with $\sigma^2 = 320nk$, $M = \sqrt{Nnk/\log\frac{1}{\delta}}$, and $\Delta = 100\sqrt{Nnk\log\frac{1}{\delta}}$, we get

$$\begin{aligned}\mathbf{Pr}\left[\left|\left\|\sum_{j=1}^{N}\mathbf{X_j}\right\|_2 - \mathbb{E}\left[\left\|\sum_{j=1}^{N}\mathbf{X_j}\right\|_2\right]\right| > \Delta\right] &\leq 2\exp\Big(-\frac{\Delta^2}{2N\sigma^2}\big(2-\exp(\frac{2M\Delta}{N\sigma^2})\big)\Big)\\ &\leq 2\delta^{2.05}\,,\end{aligned}$$

so for $0 < \delta < 1/2$, with probability at least $1-\delta$:

$$\left|\left\|\sum_{j=1}^{N}\mathbf{X_j}\right\|_2 - \mathbb{E}\left[\left\|\sum_{j=1}^{N}\mathbf{X_j}\right\|_2\right]\right| \leq O\left(\sqrt{Nnk\log\frac{1}{\delta}}\right).$$

Finally, it remains to bound the expected ℓ_2-norm of $\sum_{j=1}^{N}\mathbf{X_j}$:

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{j=1}^{N}\mathbf{X_j}\right\|_2\right]^2 &\leq \mathbb{E}\left[\left\|\sum_{j=1}^{N}\mathbf{X_j}\right\|_2^2\right] && \text{(Cauchy-Schwarz)}\\ &= \sum_{j=1}^{N}\mathbb{E}\left[\left\|\mathbf{X_j}\right\|_2^2\right] && \text{(independence of } \mathbf{X_j}\text{'s, and } \mathbb{E}\left[\mathbf{X_j}\right] = \mathbf{0} \text{ by Lemma 20)}\\ &\leq Nnk\,. && (\mathbb{E}\left[\|\mathbf{X_j}\|_2^2\right] \leq nk \text{ by Lemma 20)}\end{aligned}$$

A.2. Lower Bound for Finite-domain Problems: Proof of Theorem 7

For notation simplicity, we consider $T = \{0, \pm1, \ldots, \pm k\}$ with support size $2k+1$. We first define the hypothesis class $\mathcal{H}$. Each hypothesis is specified by a binary nk-dimensional vector $\mathbf{v} \in \{\pm1\}^{n\times k}$. Specifically, $\mathcal{H} = \{h^{\mathbf{v}}\}$ where $h^{\mathbf{v}} : T^n \to [0,1]$ is defined as

$$h^{\mathbf{v}}(\mathbf{t}) := \mathbb{1}\left[\exists i \in [n], j \in [k], \mathbf{t} = \Big(0,\ldots,0,\underset{i\text{'th}}{v_{i,j}\cdot j},0,\ldots,0\Big)\right].$$

Next, we consider a family of distributions $\mathcal{D} = \{\mathbf{D}^{\mathbf{v}}\}$ that are also indexed by $\mathbf{v}$. For each dimension i of $\mathbf{D}^{\mathbf{v}}$, the probability density function is defined as the following:

$$f_{D_i^{\mathbf{v}}}(t_i) = \begin{cases} 1-\frac{1}{n}, & \text{if } t_i = 0\\ \frac{1}{2nk}(1-\epsilon), & \text{if } t_i = -v_{i,j}\cdot j \text{ for some } j \in [k]\\ \frac{1}{2nk}(1+\epsilon), & \text{if } t_i = v_{i,j}\cdot j \text{ for some } j \in [k]\end{cases}$$

Our plan is to show that any algorithm that gets a ϵ-approximation on all distributions in $\mathcal{D}$ must take $\Omega(\frac{nk}{\epsilon^2})$ number of samples.

When the underlying distribution is $\mathbf{D}^{\mathbf{v}}$, the corresponding optimal hypothesis is $h^{\mathbf{v}}$. Intuitively, in order to achieve a good approximation to $h^{\mathbf{v}}$, an algorithm has to specify a vector $\mathbf{v}'$ close enough to $\mathbf{v}$ based on the samples. We formalize the intuition by calculating the loss of choosing $h^{\mathbf{v}'}$.

Lemma 21 *For all* $\mathbf{v}, \mathbf{v}'$*, we have*

$$h^{\mathbf{v}}(\mathbf{D}^{\mathbf{v}}) - h^{\mathbf{v}'}(\mathbf{D}^{\mathbf{v}}) = \Omega(\frac{\epsilon}{nk} \cdot d(\mathbf{v}, \mathbf{v}')),$$

where $d(\mathbf{v}, \mathbf{v}')$ *is the hamming distance between* $\mathbf{v}$ *and* $\mathbf{v}'$*.*

Proof For all $\mathbf{v}, \mathbf{v}'$, we have that

$$\begin{aligned} h^{\mathbf{v}}(\mathbf{D}^{\mathbf{v}'}) &= \sum_{i=1}^{n}\sum_{j=1}^{k} \mathbf{Pr}_{t_i \sim D_i^{\mathbf{v}}}[t_i = v'_{i,j} \cdot j] \cdot \prod_{\ell \neq i} \mathbf{Pr}_{t_\ell \sim D_\ell^{\mathbf{v}}}[t_\ell = 0] \\ &= (1 - \frac{1}{n})^{n-1} \sum_{i=1}^{n}\sum_{j=1}^{k} \mathbf{Pr}_{t_i \sim D_i^{\mathbf{v}}}[t_i = v'_{i,j} \cdot j] \\ &= (1 - \frac{1}{n})^{n-1} \sum_{i=1}^{n}\sum_{j=1}^{k} \left[\mathbb{1}[v_{i,j} = v'_{i,j}] \cdot \frac{1+\epsilon}{2nk} + \mathbb{1}[v_{i,j} \neq v'_{i,j}] \cdot \frac{1-\epsilon}{2nk} \right] . \end{aligned}$$

Next, we bound the lose of choosing $h^{\mathbf{v}'}$ by the hamming distance between $\mathbf{v}$ and $\mathbf{v}'$.

$$h^{\mathbf{v}}(\mathbf{D}^{\mathbf{v}}) - h^{\mathbf{v}'}(\mathbf{D}^{\mathbf{v}}) = (1 - \frac{1}{n})^{n-1} \sum_{i=1}^{n}\sum_{j=1}^{k} \mathbb{1}[v_{i,j} \neq v'_{i,j}] \cdot \frac{\epsilon}{nk} = \Omega(\frac{\epsilon}{nk} \cdot d(\mathbf{v}, \mathbf{v}'))$$

■

Let $\mathbf{s}$ be the samples and A be any (randomized) algorithm that takes samples $\mathbf{s}$ as inputs and outputs a vector $A(\mathbf{s}) \in \{\pm 1\}^{n \times k}$. The next lemma states that if two distributions differ in only one dimension, then the total probability of A guessing wrongly for the two distributions is at least $\Omega(1)$ if the number of samples is $O(\frac{nk}{\epsilon^2})$.

Lemma 22 *For any* $\mathbf{D}^{\bar{\mathbf{v}}}$ *and* $\mathbf{D}^{\underline{\mathbf{v}}}$ *where* $\mathbf{D}^{\bar{\mathbf{v}}}$ *and* $\mathbf{D}^{\underline{\mathbf{v}}}$ *only differ in one dimension* (i, j)*, i.e.,* $\bar{v}_{i,j} = 1$*,* $\underline{v}_{i,j} = -1$*, and for any algorithm* A*, when* $N = O(\frac{nk}{\epsilon^2})$*,*

$$\mathbf{Pr}_{\mathbf{s} \sim (\mathbf{D}^{\bar{\mathbf{v}}})^N}[A(\mathbf{s})_{i,j} \neq \bar{v}_{i,j}] + \mathbf{Pr}_{\mathbf{s} \sim (\mathbf{D}^{\underline{\mathbf{v}}})^N}[A(\mathbf{s})_{i,j} \neq \underline{v}_{i,j}] \geq \Omega(1).$$

Proof Since the n dimensions of the distribution are independent, to guess the (i, j)-th dimension of the underlying $\mathbf{v}$, the only useful samples are those $\mathbf{s}$ with $s_i = \pm j$, which happens with probability $\frac{1}{nk}$. Thus, when the number of samples is $O(\frac{nk}{\epsilon^2})$, with high probability, we have at most $O(\frac{1}{\epsilon^2})$ number of useful samples. Note that these samples are either from

$$\bar{f}(\pm j) = \frac{1 \pm \epsilon}{2} \text{ or } \underline{f}(\pm j) = \frac{1 \mp \epsilon}{2}.$$

whose total variation distance is $\Theta(\epsilon)$, then if we only have $O(\frac{1}{\epsilon^2})$ samples, with constant probability we cannot distinguish distinguish whether the samples are from $\mathbf{D}^{\bar{\mathbf{v}}}$ or $\mathbf{D}^{\underline{\mathbf{v}}}$. In other words, for any algorithm A, $A(\mathbf{s})_{i,j}$ must be inconsistent with the underlying distribution with constant probability, which concludes the proof. ■

To finish the proof, we consider the performance of A on a uniform distribution over all distributions in $\mathcal{D}$. Formally, let U be a uniform distribution on $\{\pm 1\}^{n\times k}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{v}\sim U}\mathbb{E}_{\mathbf{s}\sim(\mathbf{D}^{\mathbf{v}})^N}[d(A(\mathbf{s}),\mathbf{v})] &= \sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{E}_{\mathbf{v}\sim U}\mathbb{E}_{\mathbf{s}\sim(\mathbf{D}^{\mathbf{v}})^N}[\mathbb{1}(A(\mathbf{s})_{i,j}\neq v_{i,j})] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{E}_{\mathbf{v}\sim U}\mathbf{Pr}_{\mathbf{s}\sim(\mathbf{D}^{\mathbf{v}})^N}[A(\mathbf{s})_{i,j})\neq v_{i,j}] \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k}\Big(\mathbb{E}_{\bar{\mathbf{v}}\sim U}\mathbf{Pr}_{\mathbf{s}\sim(\mathbf{D}^{\bar{\mathbf{v}}})^N}[A(\mathbf{s})_{i,j})\neq \bar{v}_{i,j}] \\
&\qquad + \mathbb{E}_{\underline{\mathbf{v}}\sim U}\mathbf{Pr}_{\mathbf{s}\sim(\mathbf{D}^{\underline{\mathbf{v}}})^N}[A(\mathbf{s})_{i,j})\neq \underline{v}_{i,j}]\Big) \\
&\geq \Omega(nk), \qquad \text{(by Lemma 22)}
\end{aligned}$$

where $d(\cdot,\cdot)$ denotes the hamming distance. By Lemma 21, this implies a $\Omega(\epsilon)$ error of the output. Therefore, there exists a distribution $\mathbf{D}^{\mathbf{v}}$ that cannot be learned by A with $O(\epsilon)$ additive error.

Appendix B. Strong Monotonicity

B.1. Proof of Theorem 17

By the Bernstein inequality and union bound, we can relate the CDFs of underlying distribution $\mathbf{D}$ and the empirical distribution $\mathbf{E}$ as follows.

Lemma 23 (e.g., Lemma 5 of Guo et al. (2019)) *With probability at least $1-\delta$, we have that for any $i\in[n]$, and any $t_i\in[0,1]$:*

$$\left|F_{D_i}(t_i)-F_{E_i}(t_i)\right| \leq \sqrt{\frac{2F_{D_i}(t_i)\big(1-F_{D_i}(t_i)\big)\ln(2Nn\delta^{-1})}{N}}+\frac{\ln(2Nn\delta^{-1})}{N}\,.$$

The rest of the subsection shows the stated additive approximation factor under the assumption that the inequality in Lemma 23 holds.

We introduce two auxiliary distribution $\hat{\mathbf{D}}$ and $\check{\mathbf{D}}$, where the former serves as an upper bound of $\mathbf{E}$ and the latter as a lower bound. Concretely, for any $i\in[n]$, define the CDF of $\hat{D}_i$ as follows:

$$F_{\hat{D}_i}(t_i) = \begin{cases} 1 & t_i = 1\,; \\ \max\left\{0, F_{D_i}(t_i) - \sqrt{\frac{2F_{D_i}(t_i)(1-F_{D_i}(t_i))\ln(2Nn\delta^{-1})}{N}} - \frac{\ln(2Nn\delta^{-1})}{N}\right\} & 0\leq t_i<1\,. \end{cases} \tag{4}$$

The case of $t_i=1$ is defined separately because its CDF must be 1 for any distribution with support bounded in $[0,1]$. The other cases are defined to be the smallest possible value of $F_{E_i}(t_i)$ according to Lemma 23 and the trivial lower bound of $F_{E_i}(t_i)\geq 0$.

Similarly, for any $i \in [n]$, define the CDF of $\check{D}_i$ as follows:

$$F_{\check{D}_i}(t_i) = \begin{cases} 0 & t_i = 0\,; \\ \min\left\{1, F_{D_i}(t_i) + \sqrt{\frac{2F_{D_i}(t_i)(1-F_{D_i}(t_i))\ln(2Nn\delta^{-1})}{N}} + \frac{\ln(2Nn\delta^{-1})}{N}\right\} & 0 < t_i \le 1\,. \end{cases} \tag{5}$$

Then, the empirical distribution is sandwiched between the auxiliary distributions by definition.

Lemma 24 *Assuming the inequality in Lemma 23, we have:*

$$\hat{\mathbf{D}} \succeq \mathbf{E} \succeq \check{\mathbf{D}}\,.$$

Therefore, we can lower bound the performance of the empirical maximizer, i.e., $h_{\mathbf{E}}$, on the underlying distribution $\mathbf{D}$ through a sequence of inequalities below:

$$\begin{aligned}
h_{\mathbf{E}}(\mathbf{D}) &\ge h_{\mathbf{E}}(\hat{\mathbf{D}}) - \delta(\hat{\mathbf{D}}, \mathbf{D}) && (h_{\mathbf{E}} \text{ bounded in } [0,1], \text{Lemma 1}) \\
&\ge h_{\mathbf{E}}(\mathbf{E}) - \delta(\hat{\mathbf{D}}, \mathbf{D}) && (\text{strong monotonicity, } \hat{\mathbf{D}} \succeq \mathbf{E}) \\
&= \textsc{Opt}(\mathbf{E}) - \delta(\hat{\mathbf{D}}, \mathbf{D}) && (\text{definition of } \textsc{Opt}(\mathbf{E})) \\
&\ge \textsc{Opt}(\check{\mathbf{D}}) - \delta(\hat{\mathbf{D}}, \mathbf{D}) && (\text{weak monotonicity, } \mathbf{E} \succeq \check{\mathbf{D}}) \\
&\ge h_{\mathbf{D}}(\check{\mathbf{D}}) - \delta(\hat{\mathbf{D}}, \mathbf{D}) && (\text{definition of } \textsc{Opt}(\check{\mathbf{D}})) \\
&\ge h_{\mathbf{D}}(\mathbf{D}) - \delta(\check{\mathbf{D}}, \mathbf{D}) - \delta(\hat{\mathbf{D}}, \mathbf{D}) && (h_{\mathbf{D}} \text{ bounded in } [0,1], \text{Lemma 1}) \\
&= \textsc{Opt}(\mathbf{D}) - \delta(\check{\mathbf{D}}, \mathbf{D}) - \delta(\hat{\mathbf{D}}, \mathbf{D})\,. && (\text{definition of } \textsc{Opt}(\mathbf{D}))
\end{aligned}$$

Therefore, it remains to show that with the number of samples stated in the theorem:

$$\delta(\check{\mathbf{D}}, \mathbf{D}) \le \frac{\epsilon}{2} \quad , \quad \delta(\hat{\mathbf{D}}, \mathbf{D}) \le \frac{\epsilon}{2}\,. \tag{6}$$

By Lemma 2, it suffices to upper bound the Hellinger distances, as in the following lemmas.

Lemma 25 *For any distribution* $\mathbf{D}$ *and the corresponding* $\hat{\mathbf{D}}$ *defined in Eqn. (4), we have:*

$$H^2(\hat{\mathbf{D}}, \mathbf{D}) \le O\left(\frac{n}{N}\log\left(\frac{Nn}{\delta}\right)\log\left(\frac{N}{\log(Nn\delta^{-1})}\right)\right)\,.$$

Lemma 26 *For any distribution* $\mathbf{D}$ *and the corresponding* $\check{\mathbf{D}}$ *defined in Eqn. (5), we have:*

$$H^2(\check{\mathbf{D}}, \mathbf{D}) \le O\left(\frac{n}{N}\log\left(\frac{Nn}{\delta}\right)\log\left(\frac{N}{\log(Nn\delta^{-1})}\right)\right)\,.$$

The proofs of the above lemmas, which we include at the end of the subsection for completeness, are analogous to the proof of a similar lemma w.r.t. Kullback-Leibler divergence by Guo et al. (2019). The main difference is that the lemma by Guo et al. (2019) requires additional conditions that lower bound the probability masses of the two endpoints of the support, while ours do not.

As corollaries of the lemma, with a number of samples stated in the lemma, we have:

$$H^2(\hat{\mathbf{D}}, \mathbf{D}) \le \frac{\epsilon^2}{8} \quad , \quad H^2(\check{\mathbf{D}}, \mathbf{D}) \le \frac{\epsilon^2}{8}\,.$$

Putting together with Lemma 2 proves Eqn. (6), which finishes the proof of Theorem 17.

Proof [Proof of Lemma 25 and Lemma 26] By symmetry, it suffices to prove one of them. Below we prove Lemma 25.

For simplicity of notations in this proof, let $\Gamma = \frac{\ln(2Nn\delta^{-1})}{N}$ be the coefficient that appears in the definition of $\hat{\mathbf{D}}$, i.e., Eqn. (4). Further define:

$$g(y) = y - \sqrt{2\Gamma \cdot y(1-y)} - \Gamma \,.$$

Then, Eqn. (4) can be written as:

$$F_{\hat{D}_i}(t) = \begin{cases} 1 & t = 1 \,; \\ \max\left\{0, g\big(F_{D_i}(t)\big)\right\} & 0 \le t < 1 \,. \end{cases} \tag{7}$$

Further, the inequality in the lemma can be written as:

$$H^2\big(\hat{\mathbf{D}}, \mathbf{D}\big) \le O\left(n\Gamma \log \frac{1}{\Gamma}\right) \,,$$

or equivalently:

$$1 - H^2\big(\hat{\mathbf{D}}, \mathbf{D}\big) \ge 1 - O\left(n\Gamma \log \frac{1}{\Gamma}\right) \,.$$

By Lemma 3, we have:

$$1 - H^2\big(\hat{\mathbf{D}}, \mathbf{D}\big) = \prod_{i=1}^{n} \left(1 - H^2\big(\hat{D}_i, D_i\big)\right) \,.$$

Hence, it suffices to show that for any $i \in [n]$:

$$1 - H^2\big(\hat{D}_i, D_i\big) \ge 1 - O\left(\Gamma \log \frac{1}{\Gamma}\right) \,,$$

or equivalently:

$$H^2\big(\hat{D}_i, D_i\big) \le O\left(\Gamma \log \frac{1}{\Gamma}\right) \,.$$

By definition, the squared Hellinger distance is:

$$H^2\big(\hat{D}_i, D_i\big) = \frac{1}{2} \int_{x \in [0,1]} \left(\sqrt{dF_{D_i}(x)} - \sqrt{dF_{\hat{D}_i}(x)}\right)^2 \,. \tag{8}$$

We shall partition $[0, 1]$ into three subsets based on how the CDF of $\hat{D}_i$ is defined in Eqn. (7): (a) the values whose $F_{\hat{D}_i}(t) = 0$, (b) $t = 1$ whose $F_{\hat{D}_i}(t) = 1$, and (c) the rest of the values whose $0 < F_{\hat{D}_i}(t) < 1$. Then, we account for their contributions to Eqn. (8) separately.

Part (a). Consider the values whose CDF is 0 w.r.t. $\hat{D}_i$. To formally define this subset of values, recall that $g(y) = y - \sqrt{2\Gamma \cdot y(1-y)} - \Gamma$. Let $F_\ell \in [0,1]$ be the unique solution for:

$$g(F_\ell) = 0 \,.$$

The value of $g(F_\ell)$ is strictly less than 0 when $F_\ell = \Gamma$, and is strictly greater than 0 when $F_\ell = 4\Gamma$. Hence, we have:

$$\Gamma < F_\ell < 4\Gamma \,. \tag{9}$$

Let ℓ be the minimum value whose CDF is at least F_ℓ, i.e.:

$$\ell = \inf\left\{x : F_{D_i}(t) \geq F_\ell\right\} .$$

Then, for values in $[0, \ell)$, we have $F_{\hat{D}_i}(t) = 0$ and therefore:

$$\begin{aligned}
\frac{1}{2}\int_{t\in[0,\ell)} \left(\sqrt{dF_{D_i}(t)} - \sqrt{dF_{\hat{D}_i}(t)}\right)^2 &= \lim_{t\to\ell-} \frac{1}{2} F_{D_i}(t) \\
&\leq \frac{1}{2} F_\ell && \text{(definition of } \ell\text{)} \\
&< 2\Gamma \,. && \text{(Eqn. (9))}
\end{aligned}$$

Part (b). For simplicity of notations, let $f(1) = f_{D_i}(1)$ and $\hat{f}(1) = f_{\hat{D}_i}(1)$ be the probability that $x = 1$ in D_i and $\hat{D}_i$ respectively. We have:

$$\begin{aligned}
\hat{f}(1) &= 1 - \lim_{t\to 1-} F_{\hat{D}_i}(t) \\
&= 1 - \lim_{t\to 1-} \left(F_{D_i}(t) - \sqrt{2\Gamma \cdot F_{D_i}(t)\big(1 - F_{D_i}(t)\big)} - \Gamma\right) && \text{(Eqn. (7))} \\
&= f(1) + \sqrt{2\Gamma \cdot f(1)\big(1 - f(1)\big)} + \Gamma \,.
\end{aligned}$$

As corollaries, we have:

$$\hat{f}(1) \geq \max\left\{f(1), \Gamma\right\} ,$$

and:

$$\begin{aligned}
\big(\hat{f}(1) - f(1)\big)^2 &= \Gamma \cdot \big(\sqrt{2f(1)\big(1 - f(1)\big)} + \sqrt{\Gamma}\big)^2 \\
&\leq \Gamma \cdot \big(\sqrt{2f(1)} + \sqrt{\Gamma}\big)^2 \\
&\leq \Gamma \cdot \max\left\{f(1), \Gamma\right\} .
\end{aligned}$$

Using the above two inequalities, the contribution from $x = 1$ is at most:

$$\begin{aligned}
\frac{1}{2}\big(\sqrt{f(1)} - \sqrt{\hat{f}(1)}\big)^2 &= \frac{1}{2} \frac{\big(\hat{f}(1) - f(1)\big)^2}{\big(\sqrt{f(1)} + \sqrt{\hat{f}(1)}\big)^2} \\
&\leq \frac{\big(\hat{f}(1) - f(1)\big)^2}{2\hat{f}(1)} \\
&\leq \frac{\Gamma}{2} \,.
\end{aligned}$$

Part (c). It remains to bound the contribution from values $t \in [\ell, 1)$. By Eqn. (7) and the definition of ℓ, the CDF w.r.t. $\hat{D}_i$ of any value in this range is defined by a continuous mapping:

$$F_{\hat{D}_i}(t) = g\big(F_{D_i}(t)\big) \ .$$

Therefore, the CDFs w.r.t. D_i and $\hat{D}_i$ have the same set of discontinuities in this range, i.e., the same set of point masses. We will first bound the contribution of values in $[\ell, 1)$ under the assumption that both D_i and $\hat{D}_i$ are continuous in this range. Then, we will demonstrate how to generalize the result to arbitrary distributions by handling the common point masses appropriately.

Under the assumption of continuity, the contribution to the Hellinger distance by this part is:

$$\begin{aligned}\frac{1}{2}\int_{t\in[\ell,1)}\left(\sqrt{dF_{D_i}(t)}-\sqrt{dF_{\hat{D}_i}(t)}\right)^2 &= \frac{1}{2}\int_{t\in[\ell,1)}\left(\sqrt{\frac{dF_{\hat{D}_i(t)}}{dF_{D_i}(t)}}-1\right)^2 dF_{D_i}(t)\\ &= \frac{1}{2}\int_{t\in[\ell,1)}\left(\sqrt{g'\big(F_{D_i}(t)\big)}-1\right)^2 dF_{D_i}(t)\ .\end{aligned}$$

By the definition of g, we have:

$$g'(y) = 1 + \frac{(2y-1)\sqrt{\Gamma}}{\sqrt{2y\big(1-y\big)}}\ .$$

Therefore, it suffices to upper bound the following integral:

$$\int_{F_\ell}^{1}\left(\sqrt{1+\tfrac{(2y-1)\sqrt{\Gamma}}{\sqrt{2y(1-y)}}}-1\right)^2 dy$$

We will bound it in $[F_\ell, 1-\Gamma)$ and $[1-\Gamma, 1]$ separately. The former is at most:

$$\begin{aligned}\int_{F_\ell}^{1-\Gamma}\left(\sqrt{1+\tfrac{(2y-1)\sqrt{\Gamma}}{\sqrt{2y\big(1-y\big)}}}-1\right)^2 dy &\le \int_{F_\ell}^{1-\Gamma}\frac{(2y-1)^2\Gamma}{2y\big(1-y\big)}dy && \big(\big|\sqrt{1+x}-1\big|\le|x|\big)\\ &\le \int_{F_\ell}^{1-\Gamma}\frac{\Gamma}{2y\big(1-y\big)}dy && (0\le y\le 1)\\ &= \frac{\Gamma}{2}\left(\ln\frac{1-\Gamma}{F_\ell}+\ln\frac{1-F_\ell}{\Gamma}\right)\\ &\le \frac{\Gamma}{2}\left(\ln\frac{1}{F_\ell}+\ln\frac{1}{\Gamma}\right)\\ &< \Gamma\ln\frac{1}{\Gamma}\ . && \text{(Eqn. (9))}\ .\end{aligned}$$

For the latter, it is upper bounded by:

$$\begin{aligned}\int_{1-\Gamma}^{1}\left(\sqrt{1+\tfrac{(2y-1)\sqrt{\Gamma}}{\sqrt{2y\big(1-y\big)}}}-1\right)^2 dy &\le \int_{1-\Gamma}^{1}\frac{(2y-1)\sqrt{\Gamma}}{\sqrt{2y\big(1-y\big)}}dy && (\sqrt{1+x}-1\le\sqrt{x}\text{ for }x>0)\\ &= \sqrt{2(1-\Gamma)}\Gamma\\ &\le \sqrt{2}\Gamma\ .\end{aligned}$$

Combining the upper bounds of the integrals over the two intervals, the contribution from part (c) under the assumption of continuity is at most $O\big(\Gamma \log \frac{1}{\Gamma}\big)$.

Finally, consider any point mass t^* in the two distributions D_i and $\hat{D}_i$. Let $\bar{y} = F_{D_i}(t^*)$ and $\underline{y} = \lim_{t \to t^*-} F_{D_i}(t)$. Then, the probability mass of t^* w.r.t. D_i is $\bar{y} - \underline{y}$, and that w.r.t. $\hat{D}_i$ is $g(\bar{y}) - g(\underline{y})$. Hence, the contribution of t^* to the Hellinger distance is:

$$\begin{aligned}
\frac{1}{2}\left(\sqrt{\bar{y}-\underline{y}} - \sqrt{g(\bar{y})-g(\underline{y})}\right)^2 &= \frac{1}{2}\left(\sqrt{\frac{g(\bar{y})-g(\underline{y})}{\bar{y}-\underline{y}}} - 1\right)^2 (\bar{y}-\underline{y}) \\
&= \frac{1}{2}\left(\sqrt{\frac{1}{\bar{y}-\underline{y}}\int_{\underline{y}}^{\bar{y}} g'(y)dy} - 1\right)^2 (\bar{y}-\underline{y}) \\
&\le \frac{1}{2}\int_{\underline{y}}^{\bar{y}} \left(\sqrt{g'(y)} - 1\right)^2 dy\,.
\end{aligned}$$

The last inequality follows by the convexity of $(\sqrt{x}-1)^2$ and Jensen's inequality. The RHS is precisely the contribution by values with CDF in $(\underline{y}, \bar{y}]$ in the continuous case. Therefore, by applying this argument to all point masses, we reduce the problem to the continuous case. ■

B.2. Applications of Theorem 17

Single-parameter Revenue Maximization. As we have discussed at the beginning of the section, strong monotonicity corresponds to strong revenue monotonicity in the context of single-parameter revenue maximization, which is shown by Devanur et al. (2016). In particular, for single-item auction, it follows from Theorem 17 that $\tilde{O}(n\epsilon^{-2})$ samples are sufficient for getting an ϵ-additive approximation when the bidders' valuations are bounded in $[0,1]$, matching the optimal bound by Guo et al. (2019). The main difference compared with Guo et al. (2019) lies in that we achieve the optimal upper bound using the empirical maximizer, which corresponds to Myerson's optimal auction w.r.t. the empirical distributions, while Guo et al. (2019) needs to apply appropriate regularization to the empirical distribution and uses the corresponding regularized empirical Myerson's auction.

Theorem 27 *In a single-item auction with n bidders whose values are bounded in $[0,1]$, suppose the number of samples is at least $C \cdot \frac{n}{\epsilon^2} \log\left(\frac{n}{\epsilon}\right) \log\left(\frac{n}{\epsilon\delta}\right)$ for some sufficiently large constant $C > 0$. Then, the empirical Myerson's auction is an ϵ-additive approximation with probability at least $1-\delta$.*

Prophet Inequality. In the context of prophet inequality, each hypothesis corresponds to a sequence of thresholds, one for each round, such that the algorithm accepts the first reward that is greater than or equal to the corresponding threshold. We show that this problem satisfies strong monotonicity; the proof is deferred to Appendix C.3.

Lemma 28 *The problem of prophet inequality is strongly monotone.*

As a corollary of Lemma 28, Theorem 17, and the fact that the optimal thresholds achieve at least one half of the expected max reward (e.g., Samuel-Cahn (1984)),[7] we get an $\tilde{O}\big(n\epsilon^{-2}\big)$ sample complexity upper bound.

7. In fact, is it known that an appropriate fixed threshold can achieve at least one half of the expected max.

Theorem 29 *For any instance of prophet inequality in which the rewards are bounded in* $[0,1]$*, suppose the number of samples is at least:* $C \cdot \frac{n}{\epsilon^2} \log\left(\frac{n}{\epsilon}\right) \log\left(\frac{n}{\epsilon\delta}\right)$ *for some sufficiently large constant* $C > 0$*. Then, the expected reward by the empirically optimal thresholds is an* ϵ*-additive approximation compared to the optimal thresholds and thus, is at least half of the expected max reward minus* ϵ*.*

Prophet Inequality for i.i.d. Rewards. If the rewards are i.i.d., Correa et al. (2017) prove an improved prophet inequality that achieves at least a 0.745 factor of the expected max reward. The strong monotonicity of prophet inequality for i.i.d. rewards follows as a special case of Lemma 28. Therefore, we get the same $\tilde{O}\left(n\epsilon^{-2}\right)$ sample complexity upper bound, matching the lower bound by Correa et al. (2019).

Theorem 30 *For any instance of prophet inequality with i.i.d. rewards bounded in* $[0,1]$*, suppose the number of sample rewards (rather than reward vectors) is at least* $C \cdot \frac{n}{\epsilon^2} \log\left(\frac{n}{\epsilon}\right) \log\left(\frac{n}{\epsilon\delta}\right)$ *for some sufficiently large constant* $C > 0$*. Then, the expected reward by the empirically optimal thresholds is an* ϵ*-additive approximation compared to the optimal thresholds and thus, is at least a* 0.745 *factor of the expected max reward minus* ϵ*.*

As mentioned in Section 1, the setting of Correa et al. (2019) is different from ours in that they consider unbounded distributions and multiplicative approximation. Indeed, we focus on bounded-support distributions and additive approximation in the main text of the paper in order to develop a generalization theory that requires minimum knowledge of the structure of the problems. Nonetheless, Appendix C.4 demonstrates how to combine the techniques in this paper and the special structures of the prophet inequality with i.i.d. rewards to get the same $\tilde{O}(n\epsilon^{-2})$ sample complexity upper bound in the setting of Correa et al. (2019), addressing an open problem therein.[8]

Pandora's Problem. An algorithm for the Pandora's problem is a mapping from the history of observed rewards to either one of the unopened boxes, or the decision to stop and take the best observed reward. Since the former has exponentially many possibilities even after discretization, the naïve upper bound on the size of the hypothesis class is doubly exponential. Nonetheless, we show that the problem is strongly monotone, highlighting that strong monotonicity is a structural property without any obvious connection to the complexity/simplicity of the hypothesis class. The proof is deferred to Appendix D.3.

Lemma 31 *Pandora's problem is strongly monotone.*

Recall that in Section 2 we use the simple treatment of defining the value of a hypothesis to be the value of the corresponding algorithm plus n and then scaled by $\frac{1}{n+1}$ to normalize its range to be $[0,1]$. Therefore, to get an ϵ-additive approximation in Pandora's problem, we need a $\frac{\epsilon}{n+1}$-additive approximation w.r.t. $\mathcal{H}$. As a corollary of Lemma 31 and Theorem 17, we get an $\tilde{O}(n^3\epsilon^{-2})$ sample complexity bound. See Appendix D.4 for an analysis tailored for Pandora's problem to get the following optimal bound.

Theorem 32 *For any instance of Pandora's problem in which the rewards are bounded in* $[0,1]$*, suppose the number of samples is at least* $C \cdot \frac{n}{\epsilon^2} \log^2\left(\frac{1}{\epsilon}\right) \log\left(\frac{n}{\epsilon}\right) \log\left(\frac{n}{\epsilon\delta}\right)$ *for some sufficiently large*

8. It is explicitly stated as an open question in the talk at EC 2019.

constant $C > 0$. Then, we can learn an ϵ-additive approximate algorithm from the samples. Further, to learn such an algorithm, the number of samples must be at least: $c \cdot \frac{n}{\epsilon^2}$ for some sufficiently small constant $c > 0$.

Appendix C. Missing Proofs about Prophet Inequality

C.1. Optimal Hypothesis

An optimal strategy of prophet inequality when $\mathbf{D}$ has bounded support, denote as $S_{\mathbf{D}}$, is called *backward induction*, where we recursively compute the optimal reward for items appear behind i and set the thresholds θ_i for item i. The algorithm for setting the strategy is as follows:

$\theta_n \leftarrow 0$ $\textsc{Opt}(D_n) = \mathbb{E}_{t_n \sim D_n}[t_n]$ **for** *i from $n-1$ to 1* **do**
    $\theta_i \leftarrow \textsc{Opt}(\mathbf{D}_{\geq i+1})$ $\textsc{Opt}(\mathbf{D}_{\geq i}) = \mathbb{E}_{t_i \geq \theta_i}[t_i] + \mathbf{Pr}[t_i < \theta_i]\textsc{Opt}(\mathbf{D}_{\geq i+1})$
end
// online strategy
$i \leftarrow 1$ **while** $i \leq n$ **do**
    if $t_i \geq \theta_i$ **then**
        Accept t_i and stop
    else
        $i \leftarrow i + 1$ *// observe the next reward*
    end
    if *i=n* **then**
        Accept t_n and stop *// if no item has been accepted, accept the last one*
    end
end

Algorithm 1: Optimal Strategy for Prophet Inequality in Bounded-support Case

One particular note for this strategy is that the thresholds for the last $n - i + 1$ dimension of $\mathbf{D}$ is independent of the arrivals there are before t_i. Therefore, in further discussion we can abuse $h_{\geq i}(\mathbf{D})$ to denote the expected reward of running the last $n - i + 1$ dimension of an backward induction strategy S on $\mathbf{D}_{\geq i} = \prod_{j=i}^{n} D_j$.

C.2. Discretization and Sample Complexity: Proof of Theorem 11

Let $\mathbf{D}_{\epsilon/2}$ be the discretized version of $\mathbf{D}$ obtained from rounding the values of each marginal distribution D_i down to the nearest multiples of ϵ. For all type $\mathbf{t} \sim \mathbf{D}$, define its downward discretization

$$\mathbf{t}_{\epsilon/2} = \lfloor \frac{2\mathbf{t}}{\epsilon} \rfloor \cdot \frac{\epsilon}{2} ,$$

also, for the optimal strategy $S_{\mathbf{D}}$ define a coupling optimal strategy $S'_{\mathbf{D}}$ for $\mathbf{t}_{\epsilon/2} \sim \mathbf{D}_{\epsilon/2}$: First re-sample

$$\mathbf{r} \sim \prod_{i=1}^{n} D_i(t \mid t \in [(t_i)_{\epsilon/2}, (t_i)_{\epsilon/2} + \frac{\epsilon}{2})) ,$$

then perform the original $S_{\mathbf{D}}$ on $\mathbf{t}' = \mathbf{t}_{\epsilon/2} + \mathbf{r}$ and return the accepted item. We introduce the re-sample step because $\mathbf{t}'$ and $\mathbf{t} \sim \mathbf{D}$ have the same distribution. Hence at any step $i \in [n]$, the

probability that $S'_{\mathbf{D}}$ accepts $\mathbf{t}_{\epsilon/2} \sim \mathbf{D}_{\epsilon/2}$ equals to that of $S_{\mathbf{D}}$ accepts $\mathbf{t} \sim \mathbf{D}$. We further show that the expected reward of the coupling optimal strategy is an ϵ-additive approximation of the original one:

Lemma 33 *Under the above definition, we have*

$$\mathbb{E}_{S'_{\mathbf{D}}}(h_{S'_{\mathbf{D}}}(\mathbf{D}_{\epsilon/2})) \geq h_{S_{\mathbf{D}}}(\mathbf{D}) - \frac{\epsilon}{2}\,.$$

Proof

$$\begin{aligned}
\mathbb{E}_{S'_{\mathbf{D}}}(h_{S'_{\mathbf{D}}}(\mathbf{D}_{\epsilon/2})) &= \int_{\mathbf{t}\in[0,1]^n} \mathbf{Pr}(\mathbf{t}' = \mathbf{t}) \cdot h_{S'_{\mathbf{D}}}(\mathbf{t}) d\mathbf{t} \\
&\leq \int_{\mathbf{t}\in[0,1]^n} f_{\mathbf{D}}(\mathbf{t}) \cdot \left\lfloor \frac{2h_{S'_{\mathbf{D}}}(\mathbf{t})}{\epsilon} \right\rfloor \cdot \frac{\epsilon}{2} d\mathbf{t} \\
&\leq \int_{\mathbf{t}\in[0,1]^n} f_{\mathbf{D}}(\mathbf{t}) \cdot h_{S_{\mathbf{D}}}(\mathbf{t}) d\mathbf{t} - \frac{\epsilon}{2} \\
&= h_{S_{\mathbf{D}}}(\mathbf{D}) - \frac{\epsilon}{2}
\end{aligned}$$

■

Now we go to the proof of Theorem 11. Let $\mathbf{E}_{\epsilon/2}$ be the distribution obtained from rounding down the values of each dimension of $\mathbf{E}$ to the nearest supports of $\frac{\epsilon}{2}$. We want to show that, when $N \geq C \cdot \frac{n}{\epsilon^3}\log(\frac{n}{\epsilon\delta})$, $h_{S_{\mathbf{E}_{\epsilon/2}}}$ will become the near-optimal hypothesis of $\mathbf{D}$.

Since $\mathbf{E}_{\epsilon/2}$ is also the empirical distribution of $\mathbf{D}_{\epsilon/2}$, and has finite support with size $\frac{1}{\epsilon/2}$ in each dimension, a corollary of Theorem 5 shows that when $N \geq C \cdot \frac{n}{\epsilon^3}\log(\frac{n}{\epsilon\delta})$,

$$h_{S_{\mathbf{D}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) - h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) \leq \frac{\epsilon}{2} \tag{10}$$

Then

$$\begin{aligned}
h_{S_{\mathbf{D}}}(\mathbf{D}) &\leq \mathbb{E}_{S'_{\mathbf{D}}}(h_{S'_{\mathbf{D}}}(\mathbf{D}_{\epsilon/2})) + \frac{\epsilon}{2} && (\text{Lemma 33}) \\
&\leq h_{S_{\mathbf{D}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) + \frac{\epsilon}{2} && (\text{Optimality of } S_{\mathbf{D}_{\epsilon/2}}) \\
&\leq h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) + \epsilon && (\text{From (10)})
\end{aligned}$$

It remains to show $h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2})$ approximates $h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D})$, i.e. the actual expected reward from the learned hypothesis. We elaborate it as follows:

Lemma 34

$$h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) \leq h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D})\,.$$

Proof Suppose $\mathbf{E}_{\epsilon/2}$ is specified by thresholds $\theta = (\theta_1, \cdots, \theta_n)$. We can assume without loss of generality that each θ_i is the multiple of $\frac{\epsilon}{2}$, since rounding the thresholds up does not affect the behavior of the strategy on $\mathbf{E}_{\epsilon/2}$.

Therefore, for any type $\mathbf{t}$ and its downward discretization $\mathbf{t}_{\epsilon/2}$, $S_{\mathbf{E}_{\epsilon/2}}$ accepts $\mathbf{t}$ at the i^{th} step if and only if $S_{\mathbf{E}_{\epsilon/2}}$ accepts $\mathbf{t}_{\epsilon/2}$ at the i^{th} step. This gives that

$$\begin{aligned} h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}) &= \int_{\mathbf{t}\in[0,1]^n} f_{\mathbf{D}}(\mathbf{t}) h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{t}) d\mathbf{t} \\ &\geq \int_{\mathbf{t}\in[0,1]^n} f_{\mathbf{D}}(\mathbf{t}) h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{t}_{\epsilon/2}) d\mathbf{t} \\ &= \int_{\mathbf{t}\in[0,1]^n} f_{\mathbf{D}_{\epsilon/2}}(\mathbf{t}) h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{t}_{\epsilon/2}) d\mathbf{t} \\ &= h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) \,. \end{aligned}$$

■

With this lemma in hand, we can complete the proof of Theorem 11.

C.3. Strong Monotonicity: Proof of Lemma 28

From now on, we abuse $h_{\geq i}(\mathbf{D})$ to denote the expected revenue of running the last $n-i+1$ dimension of $S_{\tilde{\mathbf{D}}_{\geq \mathbf{i}}}$ on $\mathbf{D}_{\geq i}$. We want to show by backward induction on i that $\forall$ product distributions $\mathbf{D}, \tilde{\mathbf{D}}$ such that $\mathbf{D} \succeq \tilde{\mathbf{D}}$, the expected reward of performing $S_{\tilde{\mathbf{D}}}$ on the last $n-i+1$ dimension of $\mathbf{D}$ is at least that of performing $S_{\tilde{\mathbf{D}}}$ on the last $n-i+1$ dimension of $\tilde{\mathbf{D}}$, i.e.

$$h_{\geq i}(\mathbf{D}_{\geq i}) \geq h_{\geq i}(\tilde{\mathbf{D}}_{\geq i}) \,.$$

Base case: When $i = n$, there is only one item with value t_n, and $S_{\tilde{\mathbf{D}}_{\geq \mathbf{n}}}$ accepts it with probability 1 and obtains the reward t_n. Therefore, we have

$$\begin{aligned} h_{\geq n}(\mathbf{D}) = \mathbb{E}_{D_n}[t_n] &= \int_{t=0}^{\infty} q^{D_n}(t) dt \\ &\geq \int_{t=0}^{\infty} q^{\tilde{D}_n}(t) dt && (D_n \succeq \tilde{D}_n) \\ &= \mathbb{E}_{\tilde{D}_n}[t_n] = h_{\geq n}(\tilde{\mathbf{D}}) \,. \end{aligned}$$

Inductive step: Assume the induction hypothesis holds for all $j > i$, i.e. $\forall j > i \; h_{\geq j}(\mathbf{D}) \geq h_{\geq j}(\tilde{\mathbf{D}})$. Then since $h_{\geq i}(\mathbf{D})$ satisfies the following recursion:

$$h_{\geq i}(\mathbf{D}) = \mathbf{Pr}_{D_i}[t_i \geq \theta_i] \cdot \mathbb{E}_{D_i}[t_i \mid t_i \geq \theta_i] + \mathbf{Pr}_{D_i}[t_i < \theta_i] \cdot h_{\geq i+1}(\mathbf{D})$$

where the first term on the right-hand-side is the expected reward when the i^{th} item is accepted, while the second one is the expected reward when the strategy accepts subsequent item. A similar recursion holds for $h_{\geq i}(\tilde{\mathbf{D}})$:

$$h_{\geq i}(\tilde{\mathbf{D}}) = \mathbf{Pr}_{D_i}[t_i \geq \theta_i] \cdot \mathbb{E}_{D_i}[t_i \mid t_i \geq \theta_i] + \mathbf{Pr}_{D_i}[t_i < \theta_i] \cdot h_{\geq i+1}(\mathbf{D})$$

We then compare the first and the second term of $h_{\geq i}(\mathbf{D})$ and $h_{\geq i}(\tilde{\mathbf{D}})$ respectively.

$$\begin{aligned}
&h_{\geq i}(\mathbf{D}) \\
=&\mathbf{Pr}_{D_i}[t_i \geq \theta_i] \cdot \mathbb{E}_{D_i}[t_i \mid t_i \geq \theta_i] + \mathbf{Pr}_{D_i}[t_i < \theta_i] \cdot h_{\geq i+1}(\mathbf{D}) \\
\geq&\mathbf{Pr}_{D_i}[t_i \geq \theta_i] \cdot \mathbb{E}_{\tilde{D}_i}[t_i \mid t_i \geq \theta_i] + (1 - \mathbf{Pr}_{D_i}[t_i \geq \theta_i]) \cdot h_{\geq i+1}(\tilde{\mathbf{D}}) \\
=&\mathbf{Pr}_{\tilde{D}_i}[t_i \geq \theta_i] \cdot \mathbb{E}_{\tilde{D}_i}[t_i \mid t_i \geq \theta_i] + (\mathbf{Pr}_{D_i}[t_i \geq \theta_i] - \mathbf{Pr}_{\tilde{D}_i}[t_i \geq \theta_i]) \cdot \mathbb{E}_{\tilde{D}_i}[t_i \mid t_i \geq \theta_i] \\
&+ (1 - \mathbf{Pr}_{D_i}[t_i \geq \theta_i]) \cdot h_{\geq i+1}(\tilde{\mathbf{D}}) \qquad (11)
\end{aligned}$$

where the first inequality comes from $D_i \succeq \tilde{D}_i$ and the induction hypothesis. Furthermore, from the optimality of $S_{\tilde{\mathbf{D}}}$ on $\tilde{\mathbf{D}}$,we must have

$$\mathbb{E}_{\tilde{D}_i}[t_i \mid t_i \geq \theta_i] \geq h_{\geq i+1}(\tilde{\mathbf{D}}) \ ,$$

Otherwise $S_{\tilde{\mathbf{D}}}$ could discard t_i unconditionally and achieve higher expected revenue. Therefore,

$$\begin{aligned}
(11) &\geq \mathbf{Pr}_{\tilde{D}_i}[t_i \geq \theta_i] \cdot \mathbb{E}_{\tilde{D}_i}[t_i \mid t_i \geq \theta_i] + (\mathbf{Pr}_{D_i}[t_i \geq \theta_i] - \mathbf{Pr}_{\tilde{D}_i}[t_i \geq \theta_i]) \cdot h_{\geq i+1}(\tilde{\mathbf{D}}) \\
&\quad + (1 - \mathbf{Pr}_{D_i}[t_i \geq \theta_i]) \cdot h_{\geq i+1}(\tilde{\mathbf{D}}) \\
&= \mathbf{Pr}_{\tilde{D}_i}[t_i \geq \theta_i] \cdot \mathbb{E}_{\tilde{D}_i}[t_i \mid t_i \geq \theta_i] + \mathbf{Pr}_{\tilde{D}_i}[t_i < \theta_i] \cdot h_{\geq i+1}(\tilde{\mathbf{D}}) \\
&= h_{\geq i}(\tilde{\mathbf{D}}) \ .
\end{aligned}$$

C.4. Prophet Inequality with i.i.d. Unbounded Rewards

In Correa et al. (2019), an ϵ-approximately optimal strategy for known distribution in the unbounded support and i.i.d. case has been introduced. Denote the strategy for distribution D as $R_{\mathbf{D}}$. We restate the algorithm to generate $R_{\mathbf{D}}$ is as follows:

Solve differential equation $y' = y(\log(y) - 1) - (\beta - 1)$ and $y(0) = 1$
where $\beta \approx 1/0.745$ **for** *i from* 1 *to* n **do**
  $\epsilon_i \leftarrow 1 - y(i/n)^{1/(n-1)}$
end
// online strategy $i \leftarrow 1$ **while** $i \leq n$ **do**
  if $\epsilon_i < \frac{\epsilon}{n}$ **then**
    $\epsilon_i \leftarrow 0$ *// Skip when acceptance probability* $< \frac{\epsilon}{n}$
  end
  if $q^{D_i}(t_i) \leq \epsilon_i$ **then**
    Accept t_i and stop
  else
    $i \leftarrow i + 1$
  end
end

Algorithm 2: Approximately Optimal Strategy for Unbounded Support Case

In the following discussion, we will show a new sample complexity bound of achieving ϵ-multiplicative approximation in the unbounded optimal stopping game.

Lemma 35 *For arbitrary distribution D, the sample complexity required for Algorithm 2 is at most $\tilde{O}(\frac{n}{\epsilon^2})$.*

The algorithm is to run the strategy R on a *dominated empirical distribution* $\tilde{E}$, which is defined below:

$$F_{\tilde{E}}(x) = \min\{1, F_E(x) - \sqrt{\frac{2F_E(x)(1-F_E(x))\ln(2Nn\delta^{-1})}{N}} - \frac{4\ln(2Nn\delta^{-1})}{N}\} .$$

In Lemma 5 of Guo et al. (2019), it is shown that with high probability $D \succeq \tilde{E}$ via a standard concentration bound.

In the following discussion, denote the stopping time of running strategy R on input $\mathbf{t}$ as $\tau(R, \mathbf{t})$

Lemma 36 *With high probability over samples for the algorithm,*

$$h_{R_{\mathbf{D}}}(\mathbf{D}) - h_{R_{\tilde{\mathbf{E}}}}(\mathbf{D}) < Pr_{\mathbf{t}\sim D^n}[\tau(R_{\tilde{\mathbf{E}}}, \mathbf{t}) < \tau(R_{\mathbf{D}}, \mathbf{t})] \cdot \textsc{Opt}(\mathbf{D}) .$$

Proof

Since $D \succeq \tilde{E}$ with high probability, $\forall t \in [n]$ the value threshold in $R_{\tilde{\mathbf{E}}}$ is lower than that of $R_{\mathbf{D}}$, i.e.

$$F_{\tilde{E}}((F_D)^{-1}(1-\epsilon_t)) > F_D((F_D)^{-1}(1-\epsilon_t)) ,$$

Therefore, fix an input value configuration $\mathbf{t}$, $\tau(R_{\tilde{\mathbf{E}}}, \mathbf{t}) \leq \tau(R_{\mathbf{D}}, \mathbf{t})$, and the only case where the revenue obtained from $R_{\tilde{\mathbf{E}}}$ is smaller than that from R_D should be $\tau(R_{\tilde{\mathbf{E}}}, \mathbf{t}) < \tau(R_{\mathbf{D}}, \mathbf{t})$.

Now it suffices to show that

$$\mathbb{E}[h_{R_{\mathbf{D}}}(\mathbf{D}) - h_{R_{\tilde{\mathbf{E}}}}(\mathbf{D}) \mid \tau(R_{\tilde{\mathbf{E}}}, \mathbf{t}) < \tau(R_{\mathbf{D}}, \mathbf{t})] = O(\textsc{Opt}(\mathbf{D})) .$$

For all $t \in [n]$, define A_t as the set of input such that $R_{\tilde{\mathbf{E}}}$ accepts at time t but $R_{\mathbf{D}}$ does not accept:

Then we can rewrite the above conditioned expected difference of revenue as follows:

$$\begin{aligned}
&\mathbb{E}[h_{R_{\mathbf{D}}}(\mathbf{D}) - h_{R_{\tilde{\mathbf{E}}}}(\mathbf{D}) \mid \tau(R_{\tilde{\mathbf{E}}}, \mathbf{t}) < \tau(R_{\mathbf{D}}, \mathbf{t})] \\
=\ &\mathbb{E}[h_{R_{\mathbf{D}}}(\mathbf{D}) - h_{R_{\tilde{\mathbf{E}}}}(\mathbf{D}) \mid \mathbf{t} \in \cup_{t=1}^{n-1} A_t] \\
\leq\ &\mathbb{E}[h_{R_{\mathbf{D}}}(\mathbf{D}) \mid \mathbf{t} \in \cup_{t=1}^{n-1} A_t] && (h_{R_{\tilde{\mathbf{E}}}}(\mathbf{D}) > 0) \\
=\ &\max_{t\in[n-1]} \mathbb{E}[h_{R_{\mathbf{D}}}(\mathbf{D}) \mid \mathbf{t} \in A_t] \\
=\ &\max_{t\in[n-1]} \mathbb{E}[h_{R_{\mathbf{D}}}(\mathbf{D}) \mid R_D \text{ does not accept before } t] \\
\leq\ &\textsc{Opt}(\mathbf{D})
\end{aligned}$$

■

Lemma 37 *When $m \geq \tilde{O}(n\epsilon^{-2})$, with high probability over samples for the algorithm,*

$$Pr_{\mathbf{t}\sim D^n}[\tau(R_{\tilde{\mathbf{E}}}, \mathbf{t}) < \tau(R_{\mathbf{D}}, \mathbf{t})] < O(\epsilon) .$$

Proof

$$
\begin{aligned}
& Pr_{\mathbf{t}\sim D^n}[\tau(R_{\tilde{\mathbf{E}}}, \mathbf{t}) < \tau(R_{\mathbf{D}}, \mathbf{t})] && (12)\\
\leq & Pr_{\mathbf{t}\sim D^n}[\mathbf{t} \in \sum_{t=1}^{n-1} A_t] \\
\leq & \sum_{t=1}^{n-1} Pr_{\mathbf{t}\sim D^n}[\epsilon_t \leq q^{\tilde{E}}(X_i) \leq \epsilon_t + \sqrt{\frac{8\epsilon_t(1-\epsilon_t)\ln(2Nn\delta^{-1})}{N}} + \frac{7\ln(2Nn\delta^{-1})}{N}] \\
\leq & \sum_{t=1}^{n-1} (\sqrt{\frac{8\epsilon_t(1-\epsilon_t)\ln(2Nn\delta^{-1})}{N}} + \frac{7\ln(2Nn\delta^{-1})}{N}) && (13)\\
= & \sum_{t=1}^{n-1} \sqrt{\frac{8\epsilon_t(1-\epsilon_t)\ln(2Nn\delta^{-1})}{N}} + O(\epsilon^2)\,, && (14)
\end{aligned}
$$

the second inequality is also shown in Lemma 7 of Guo et al. (2019) to be hold with high probability. Recall from Algorithm 2 that $\epsilon_t = 1 - y(\frac{t}{n})^{1/(n-1)}$. Now we bound (14) for $y(\frac{t}{n}) < \frac{1}{n}$ or $y(\frac{t}{n}) > \frac{1}{n}$:

Case 1: $y(\frac{t}{n}) \geq \frac{1}{n}$**.** In this case,

$$
\epsilon_t = 1 - y^{\frac{1}{n-1}} \leq 1 - e^{\frac{\log y}{n-1}} \leq 1 - e^{-\frac{\log n}{n-1}} \leq \frac{\log n}{n}\,,
$$

Therefore when $m \geq n\epsilon^{-2}\log n$

$$
\sum_{y(\frac{t}{n})\geq\frac{1}{n}} \sqrt{\frac{\epsilon_t(1-\epsilon_t)\ln(2Nn\delta^{-1})}{N}} \leq n \cdot \sqrt{\frac{\log n/n \cdot 1 \cdot \ln(2Nn\delta^{-1})}{N}} = O(\epsilon)\,.
$$

Case 2: $y(\frac{t}{n}) < \frac{1}{n}$**.** Since $y(x) \in [0,1]$ when $x \in [0,1]$, we have $\forall x \in [0,1]$,

$$
y'(x) = y(\log y - 1) - (\beta - 1) \leq -(\beta - 1) \leq -0.3414
$$

Therefore

$$
|\{t \in [n-1], \text{s.t. } y(\frac{t}{n}) < \frac{1}{n}\}| \leq \frac{1}{n} \cdot \frac{1}{0.3414} \cdot n \leq 3\,,
$$

and when $N \geq n\epsilon^{-2}\log n$,

$$
\sum_{y(\frac{t}{n})<\frac{1}{n}} \sqrt{\frac{\epsilon_t(1-\epsilon_t)\ln(2Nn\delta^{-1})}{N}} \leq 3 \cdot \sqrt{\frac{\epsilon_t(1-\epsilon_t)}{N}\ln(2Nn\delta^{-1})} = O(\frac{\epsilon}{\sqrt{n}})\,.
$$

Combining the two cases, we have

$$
\begin{aligned}
& Pr_{\mathbf{t}\sim D^n}[\tau(R_{\tilde{\mathbf{E}}}, \mathbf{t}) < \tau(R_{\mathbf{D}}, \mathbf{t})] \\
& \leq 4 \sum_{y(\frac{t}{n})<\frac{1}{n}} \sqrt{\frac{\epsilon_t(1-\epsilon_t)\ln(2Nn\delta^{-1})}{N}} + 4 \sum_{y(\frac{t}{n})\geq\frac{1}{n}} \sqrt{\frac{\epsilon_t(1-\epsilon_t)\ln(2Nn\delta^{-1})}{N}} + O(\epsilon^2) \\
& = O(\epsilon)\,.
\end{aligned}
$$

■

$i \leftarrow 1$ **while** $i \leq n$ **do**
  Open the i^{th} box and set $U_i \leftarrow \max_{j \leq i} t_j$ **if** $U_i \geq \sigma_i$ **then**
    Accept U_i and stop *// accept the highest opened box so far*
  else
    $i \leftarrow i + 1$
  end
end
if *i=n* **then**
  Accept U_n and stop *// if no item has been accepted, accept the box with highest reward*
end

Algorithm 3: Frequency Number Computation

Appendix D. Missing Proofs about Pandora's Problem

D.1. Optimal Hypothesis

Optimal strategy of Pandora's problem An optimal strategy $S_{\mathbf{D}}$ for $\mathbf{D}$, introduced by Weitzman (1979), opens the boxes sequentially according to its *reservation value* σ_i, the threshold of the maximum realized values below which opening the $i + 1^{th}$ box will give rise to a higher expected reward. A formal definition of σ_i is as follows:

$$\sigma_i \stackrel{\text{def}}{=} \inf_{\sigma}(\mathbb{E}_{t_i \sim D_i}[(t_i - \sigma)^+] = c_i) \ .$$

we assume without loss of generality that the reserve value is non-increasing with the index of each box, i.e. $\sigma_1 \geq \sigma_2 \geq \sigma_n$.

Also, for convenience we use U_i to denote the maximum value among the first i boxes:

$$U_i \stackrel{\text{def}}{=} \max_{j \leq i} t_j \ ,$$

and let $U_0 \stackrel{\text{def}}{=} 0$.

We restate the optimal strategy $S_{\mathbf{D}}$ in Weitzman (1979) as Algorithm 3:

D.2. Discretization and Sample Complexity: Proof of Theorem 12

The proof of Theorem 12 is almost the same as that of Theorem 11. We include it here only for completeness.

Let $\mathbf{D}_{\epsilon/2}$ be the discretized version of $\mathbf{D}$ obtained from rounding the values of each marginal distribution D_i down to the nearest multiples of $\epsilon/2$. Also, for all type $\mathbf{t} \sim \mathbf{D}$, let $\mathbf{t}_{\epsilon/2}$ be its downward discretization to the multiples of $\epsilon/2$. For the optimal strategy $S_{\mathbf{D}}$ define a coupling optimal strategy $S'_{\mathbf{D}}$ for discretized type $\mathbf{t}_{\epsilon/2}$: First re-sample

$$\mathbf{r} \sim \prod_{i=1}^{n} D_i(t \mid t \in [(t_i)_{\epsilon/2}, (t_i)_{\epsilon/2} + \frac{\epsilon}{2})) \ ,$$

then perform the original $S_{\mathbf{D}}$ on $\mathbf{t}' = \mathbf{t}_{\epsilon/2} + \mathbf{r}$ and return the accepted item. It is easy to see that after the re-sample step $\mathbf{t}'$ has the same distribution as $\mathbf{t} \sim \mathbf{D}$. Hence at any step $i \in [n]$ and for any

$j \in [i]$, the probability that $S'_{\mathbf{D}}$ accepts the j^{th} box of $\mathbf{t}_{\epsilon/2} \sim \mathbf{D}_{\epsilon/2}$ equals to that of $S_{\mathbf{D}}$ accepts the j^{th} box of $\mathbf{t} \sim \mathbf{D}$. We further show that the expected reward of the coupling optimal strategy is an ϵ-additive approximation of the original one:

Lemma 38 *Under the above definition, we have*

$$\mathbb{E}_{S'_{\mathbf{D}}}(h_{S'_{\mathbf{D}}}(\mathbf{D}_{\epsilon/2})) \geq h_{S_{\mathbf{D}}}(\mathbf{D}) - \frac{\epsilon}{2} .$$

Proof Same as the proof of Lemma 33. ■

Now we go to the proof of Theorem 12. Let $\mathbf{E}_{\epsilon/2}$ be the distribution obtained from rounding down the values of each dimension of $\mathbf{E}$ to the nearest supports of $\frac{\epsilon}{2}$. We want to show that, when $N \geq C \cdot \frac{n^3}{\epsilon^3} \log(\frac{n}{\epsilon\delta})$, $h_{S_{\mathbf{E}_{\epsilon/2}}}$ will become the near-optimal hypothesis of $\mathbf{D}$.

Since $\mathbf{E}_{\epsilon/2}$ is also the empirical distribution of $\mathbf{D}_{\epsilon/2}$, and has finite support with size $\frac{1}{\epsilon/2}$ in each dimension. Because the value value is bounded in $[-n, 1]$, a corollary of Theorem 5 shows that when $N \geq C \cdot \frac{n^3}{\epsilon^3} \log(\frac{n}{\epsilon\delta})$ for a large enough constant C,

$$\frac{h_{S_{\mathbf{D}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) + n}{n+1} - \frac{h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) + n}{n+1} \leq \frac{\epsilon}{2(n+1)} \tag{15}$$

Then

$$\begin{aligned}
h_{S_{\mathbf{D}}}(\mathbf{D}) &\leq \mathbb{E}_{S'_{\mathbf{D}}}(h_{S'_{\mathbf{D}}}(\mathbf{D}_{\epsilon/2})) + \frac{\epsilon}{2} && (\text{ Lemma 38}) \\
&\leq h_{S_{\mathbf{D}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) + \frac{\epsilon}{2} && (\text{Optimality of } S_{\mathbf{D}_{\epsilon/2}}) \\
&\leq h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) + \epsilon && (\text{ (15)})
\end{aligned}$$

It remains to show $h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2})$ approximates $h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D})$, i.e. the actual expected reward from the learned hypothesis. We elaborate it as follows:

Lemma 39

$$h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}_{\epsilon/2}) \leq h_{S_{\mathbf{E}_{\epsilon/2}}}(\mathbf{D}) .$$

Proof Same as the proof of Lemma 34. ■

With this lemma in hand, we can complete the proof of Theorem 12.

D.3. Strong Monotonicity: Proof of Lemma 31

It suffices to show that for any $\mathbf{D} \succeq \tilde{\mathbf{D}}$,

$$h_{S_{\tilde{\mathbf{D}}}}(\mathbf{D}) \geq h_{S_{\tilde{\mathbf{D}}}}(\tilde{\mathbf{D}}) .$$

Lemma 40 *For any $H \leq \sigma_{i+1}$,*

$$h_{S_{\tilde{\mathbf{D}}}}(\tilde{\mathbf{D}}|U_i = H) \leq \sigma_{i+1} - \sum_{j=1}^{i} c_j$$

Proof Since $S_{\tilde{\mathbf{D}}}$ is the optimal strategy, $h_{S_{\tilde{\mathbf{D}}}}(\tilde{\mathbf{D}}|U_i = H)$ is monotone in H, so it is only necessary to show the case when $U_i = H = \sigma_{i+1}$. But in this case, simply choosing the largest among first i boxes would give a revenue of $\sigma_{i+1} - \sum_{j=1}^{i} c_j$, so

$$h_{S_{\tilde{\mathbf{D}}}}(\tilde{\mathbf{D}}|U_i = \sigma_{i+1}) \leq \sigma_{i+1} - \sum_{j=1}^{i} c_j$$

is true due to the optimality of the mechanism. ■

We will use backward induction from n to 0 to prove the following statement.

Lemma 41 *For any $0 \leq i \leq n$ and any $u_i' \leq u_i \leq \sigma_{i+1}$,*

$$h_{S_{\tilde{\mathbf{D}}}}(\tilde{\mathbf{D}}|U_i = u_i) > h_{S_{\tilde{\mathbf{D}}}}(\tilde{\mathbf{D}}|U_i = u_i') \,.$$

Proof This holds trivially when $i = n$. In the following discussion, assume $i < n$ and the lemma holds for $i + 1$.

If $U_i \leq \sigma_{i+1}$, the mechanism will choose to open the next box. In this case, because $D_{i+1} \succeq \tilde{D_{i+1}}$, it suffices to show that for any $t_{i+1} \geq t_{i+1}'$,

$$h_{S_{\tilde{\mathbf{D}}}}(\mathbf{D}|U_i = u_i, X_{i+1} = t_{i+1}) \geq h_{S_{\tilde{\mathbf{D}}}}(\tilde{\mathbf{D}}|U_i = u_i', X_{i+1} = t_{i+1}') \,.$$

So it is enough to show that for any $u_{i+1} \geq u_{i+1}'$,

$$h_{S_{\tilde{\mathbf{D}}}}(\mathbf{D}|U_{i+1} = u_{i+1}) \geq h_{S_{\tilde{\mathbf{D}}}}(\tilde{\mathbf{D}}|U_{i+1} = u_{i+1}') \,.$$

We will consider three cases. In the first case, $u_{i+1} \geq u_{i+1}' > \sigma_{i+2}$. Then

$$h_{S_{\tilde{\mathbf{D}}}}(\mathbf{D}) = u_{i+1} - \sum_{j=1}^{i} c_i \geq u_{i+1}' - \sum_{j=1}^{i} c_j = h_{S_{\tilde{\mathbf{D}}}}(\tilde{\mathbf{D}}) \,.$$

In the second case, $u_{i+1} > \sigma_{i+2} \geq u_{i+1}'$, then we can apply Lemma 40 and have

$$h_{S_{\tilde{\mathbf{D}}}}(\tilde{\mathbf{D}}) \leq \sigma_{i+2} - \sum_{j=1}^{i+1} c_i < u_{i+1} - \sum_{j=1}^{i+1} c_k = h_{S_{\tilde{\mathbf{D}}}}(\mathbf{D})$$

In third case, $\sigma_{i+2} \geq u_{i+1} \geq u_{i+1}'$, the inequality follows directly from induction assumption. ■

D.4. Tight Bounds: Proof of Theorem 32

D.4.1. UPPER BOUND

We start by recalling the main obstacle for getting an $\tilde{O}\left(\frac{n}{\epsilon^2}\right)$ sample complexity upper bound as a direct corollary of Theorem 17. In Pandora's problem, an algorithm may pay a cost up to 1 to open each box and thus, the range of the realized objective is $[-n, 1]$ instead of $[0, 1]$. In the main text, we use a simple hypothesis class $\mathcal{H}$, which has a hypothesis h_A for each algorithm A, normalizing its

value to be in $[0,1]$ by letting it be the realized objective of A plus n and scaled by $\frac{1}{n+1}$. Therefore, to get an ϵ-additive approximation in Pandora's problem, we need a $\frac{\epsilon}{n+1}$-additive approximation w.r.t. the general learning problem $\mathcal{H}$. Therefore, applying Theorem 17 to this hypothesis class $\mathcal{H}$ gives only an $\tilde{O}\big(\frac{n^3}{\epsilon^2}\big)$ sample complexity bound.

Although the objective could be as small as $-n$ in the worst cases, intuitively the chance of getting such a bad objective shall be negligible if the algorithm is reasonable w.r.t. the underlying distribution. Indeed, we will reason that it is without loss of generality to consider algorithms that stop whenever the cost exceeds $\log\frac{1}{\epsilon}$. As a result, we avoid scaling the value of the hypotheses by a $n+1$ factor.

In particular, we consider the following notion of rational algorithms w.r.t. a given distribution.

Definition 42 (Rational Algorithms) *For any distribution $\mathbf{D}$ and any cost vector $\mathbf{c}$, an algorithm A for the Pandora's problem is* rational *w.r.t. $\mathbf{D}$ and $\mathbf{c}$ if whenever A opens a box i, the expected increase in the best observed reward is greater than or equal to the cost c_i.*

The next lemma follows by the definition of the optimal algorithm.

Lemma 43 *Suppose A is the optimal algorithm w.r.t. a distribution $\tilde{\mathbf{D}}$ and a cost vector $\mathbf{c}$. Then, A is rational w.r.t. any distribution $\mathbf{D}$ that stochastically dominates (including $\tilde{\mathbf{D}}$ itself), and $\mathbf{c}$.*

We will need a standard Bernstein type concentration bound for submartingales.

Lemma 44 *Let $S_0, S_1 \cdots, S_n$ be a submartingale with respect to filtration $\mathcal{F}_0, \mathcal{F}_1, \cdots, \mathcal{F}_k$. Suppose $S_0 = 0$, $|S_i - S_{i-1}| \le M$, $\sum_{i=1}^n \mathbb{E}[(S_i - S_{i-1})^2|\mathcal{F}_{i-1}] \le L$, then for any positive Δ,*

$$\mathbf{Pr}[S_n < -\Delta] \le \exp\left(\frac{\Delta^2}{2L + (2/3)M\Delta}\right).$$

Lemma 45 *Suppose an algorithm A is rational w.r.t. a distribution $\mathbf{D}$ and a cost vector $\mathbf{c}$. Then, the probability that A pays a cost more than $\Omega\big(\log\frac{1}{\epsilon}\big)$ is at most ϵ.*

Proof For $1 \le i \le n$, let X_i be objective after round i; if the algorithm stops before round i, let $X_i = X_{i-1}$. Let $X_0 = 0$. Then, by that A is rational, we have:

$$\mathbb{E}\big[X_i | X_1, X_2, \ldots, X_{i-1}\big] \ge X_{i-1}\,.$$

That is, X_i's form a discrete-time submartingale.

We have $-1 \le X_i - X_{i-1} \le 1$ by definition. Further, for any round $1 \le i \le n$, $X_i - X_{i-1}$ is upper bounded by the increment in the best observed reward in the round. Therefore, $\sum_{i:X_i \ge X_{i-1}}(X_i - X_{i-1})$ is at most the best observed reward at the end, which is upper bounded by 1. Hence, we have:

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}\big[(X_i - X_{i-1})^2\big] &\le \sum_{i=1}^n \mathbb{E}\big[\big|X_i - X_{i-1}\big|\big] && (-1 \le X_i - X_{i-1} \le 1)\\
&= \mathbb{E}\Big[\sum_{i=1}^n (X_{i-1} - X_i) + 2\sum_{i:X_i \ge X_{i-1}} (X_i - X_{i-1})\Big]\\
&= -\mathbb{E}\big[X_n\big] + 2\cdot\mathbb{E}\Big[\sum_{i:X_i \ge X_{i-1}} (X_i - X_{i-1})\Big]\\
&\le 2\,. && (\textstyle\sum_{i:X_i \ge X_{i-1}}(X_i - X_{i-1}) \le 1)
\end{aligned}$$

Since the cost is at most $-X_n$ by definition, it suffices to upper bound the probability that $X_n \le -\Omega\big(\log \frac{1}{\epsilon}\big)$. Then, the lemma follows by Bernstein's inequality for submartingales. ■

In the following arguments, consider a hypothesis class $\mathcal{H}$, which has a hypothesis for any algorithm A such that its value equals the objective of A, *without scaling*. Further, fixed the cost vector $\mathbf{c}$, let $h_{\mathbf{D}}$ be the hypothesis that corresponds to the optimal algorithm for $\mathbf{D}$ and $\mathbf{c}$. Finally, let $\bar{h}_{\mathbf{D}}$ be the hypothesis that corresponds to a truncated version of the optimal algorithm for $\mathbf{D}$, which stops whenever the cost exceeds $\Omega\big(\log \frac{1}{\epsilon}\big)$.

Lemma 46 *Fixed any cost vector* $\mathbf{c}$*. For any* $\mathbf{D} \succeq \tilde{\mathbf{D}}$*, the truncated version of optimal algorithm w.r.t.* $\tilde{\mathbf{D}}$ *gets an expected value greater than or equal to that of the untruncated version minus* ϵ*:*

$$\bar{h}_{\tilde{\mathbf{D}}}(\mathbf{D}) \ge h_{\tilde{\mathbf{D}}}(\mathbf{D}) - \epsilon\,.$$

Proof By Lemma 43, $h_{\tilde{\mathbf{D}}}$ is rational w.r.t. $\mathbf{D}$ and $\mathbf{c}$. Hence, by Lemma 45, the probability that the truncated version $\bar{h}_{\tilde{\mathbf{D}}}$ and the original version $h_{\tilde{\mathbf{D}}}$ give different outcomes is at most ϵ. Finally, whenever they are different, $h_{\tilde{\mathbf{D}}}$ gets at most 1 extra reward in subsequent rounds. Putting together proves the lemma. ■

We now prove the stated sample complexity upper bound.

Proof [Proof of Theorem 32 (Upper Bound)] We show that the truncated version of PERM gets the stated sample complexity bound. We prove an $O(\epsilon)$-additive approximation with the understanding that changing ϵ by a constant factor does not affect the stated sample complexity bound asymptotically.

It follows from a sequence of inequalities below, similar to those in Section 4:

$$\begin{aligned}
\bar{h}_{\mathbf{E}}(\mathbf{D}) &\ge \bar{h}_{\mathbf{E}}(\hat{\mathbf{D}}) - O\big(\log \tfrac{1}{\epsilon}\big)\cdot\delta(\hat{\mathbf{D}},\mathbf{D}) && (\bar{h}_{\mathbf{E}} \text{ bounded in } [-O\big(\log \tfrac{1}{\epsilon}\big), 1])\\
&\ge h_{\mathbf{E}}(\hat{\mathbf{D}}) - \epsilon - O\big(\log \tfrac{1}{\epsilon}\big)\cdot\delta(\hat{\mathbf{D}},\mathbf{D}) && (\text{Lemma 46})\\
&\ge h_{\mathbf{E}}(\mathbf{E}) - \epsilon - O\big(\log \tfrac{1}{\epsilon}\big)\cdot\delta(\hat{\mathbf{D}},\mathbf{D}) && (\text{strong monotonicity, } \hat{\mathbf{D}} \succeq \mathbf{E})\\
&= \textsc{Opt}(\mathbf{E}) - \epsilon - O\big(\log \tfrac{1}{\epsilon}\big)\cdot\delta(\hat{\mathbf{D}},\mathbf{D}) && (\text{definition of } \textsc{Opt}(\mathbf{E}))\\
&\ge \textsc{Opt}(\check{\mathbf{D}}) - \epsilon - O\big(\log \tfrac{1}{\epsilon}\big)\cdot\delta(\hat{\mathbf{D}},\mathbf{D}) && (\text{weak monotonicity, } \mathbf{E} \succeq \check{\mathbf{D}})\\
&\ge \bar{h}_{\mathbf{D}}(\check{\mathbf{D}}) - \epsilon - O\big(\log \tfrac{1}{\epsilon}\big)\cdot\delta(\hat{\mathbf{D}},\mathbf{D}) && (\text{definition of } \textsc{Opt}(\check{\mathbf{D}}))\\
&\ge \bar{h}_{\mathbf{D}}(\mathbf{D}) - \epsilon - O\big(\log \tfrac{1}{\epsilon}\big)\cdot\big(\delta(\hat{\mathbf{D}},\mathbf{D}) + \delta(\check{\mathbf{D}},\mathbf{D})\big) && (\bar{h}_{\mathbf{D}} \text{ bounded in } [-O\big(\log \tfrac{1}{\epsilon}\big), 1])\\
&\ge h_{\mathbf{D}}(\mathbf{D}) - 2\epsilon - O\big(\log \tfrac{1}{\epsilon}\big)\cdot\big(\delta(\hat{\mathbf{D}},\mathbf{D}) + \delta(\check{\mathbf{D}},\mathbf{D})\big) && (\text{Lemma 46})\\
&= \textsc{Opt}(\mathbf{D}) - 2\epsilon - O\big(\log \tfrac{1}{\epsilon}\big)\cdot\big(\delta(\hat{\mathbf{D}},\mathbf{D}) + \delta(\check{\mathbf{D}},\mathbf{D})\big)\,. && (\text{definition of } \textsc{Opt}(\mathbf{D}))
\end{aligned}$$

By Lemma 2, Lemma 25 and Lemma 26, we get an $O(\epsilon)$-additive approximation. ■

D.4.2. LOWER BOUND

Consider n boxes with cost $\frac{1}{n}$ each. Consider 2^n potential instances, in which the reward distribution of each box is either D^+ or D^-, defined by the following probability mass functions respectively:

$$f_{D^+}(x) = \begin{cases} \frac{1+\epsilon}{n} & x = 1\,; \\ 1 - \frac{1+\epsilon}{n} & x = 0\,. \end{cases} \qquad f_{D^-}(x) = \begin{cases} \frac{1-\epsilon}{n} & x = 1\,; \\ 1 - \frac{1-\epsilon}{n} & x = 0\,. \end{cases}$$

We will refer to each of these 2^n instances by the distribution $\mathbf{D}$, since the cost vector $\mathbf{c}$ is fixed.

The next lemma follows by the above definition and and simple calculations which we omit.

Lemma 47 *The squared Hellinger distance between D^+ and D^- is bounded by:*

$$H^2\big(D^+, D^-\big) = O\left(\frac{n}{\epsilon^2}\right).$$

To distinguish the algorithm for Pandora's problem and the learning algorithm, we will refer to the former as a hypothesis.

Since the rewards are either 0 or 1, any hypothesis is characterized by an ordered subsequence $i_1, i_2, \ldots, i_k$ of the boxes such that it opens the boxes one by one until it gets a reward 1; if all k rewards are 0, it stops and leaves the remaining $n-k$ boxes unopened. The optimal hypothesis chooses a box into the subsequence if and only if its distribution equals D^+ (order is irrelevant since they are identical). Therefore, for any instance $\mathbf{D}$ defined above, any hypothesis h, and any box $1 \le i \le n$, we say that h makes a mistake on box i w.r.t. $\mathbf{D}$ if either $D_i = D^+$ but i isn't in the subsequence chosen by h, or $D_i = D^-$ but i is in the subsequence. We simply say that the algorithm makes a mistake on box i w.r.t $\mathbf{D}$ if it selects a hypothesis that makes such a mistake. Whether a given learning algorithm makes a mistake might be a random event if it is randomized.

In the rest of the argument, we first argue that the additive approximation error scales linearly with number of mistakes made by the chosen hypothesis. Then, we argue through a sequence of lemmas that for any algorithm that takes less than $c \cdot \frac{n}{\epsilon^2}$ samples for some sufficiently small constant $c > 0$, there is an instance $\mathbf{D}$ for which it picks a hypothesis that makes $\Omega\big(n\big)$ mistakes with at least constant probability.

Lemma 48 *For any instance $\mathbf{D}$, if a hypothesis h makes k mistakes, then we have:*

$$h\big(\mathbf{D}\big) \le \textsc{Opt}\big(\mathbf{D}\big) - \Omega\left(\frac{k\epsilon}{n}\right).$$

Proof Suppose the instance have n^+ and n^- boxes with reward distributions equal to D^+ and D^- respectively. Further suppose h makes k^+ and k^- mistakes on the two types of boxes. Hence, h includes $n^+ - k^+$ boxes with distributions equal to D^+ and k^- boxes with distributions equal to D^- in its subsequence.

The expected reward minus cost for opening a box with distribution D^+ is $\frac{\epsilon}{n}$; opening a box with distribution D^- gives $-\frac{\epsilon}{n}$. Further, the probability of opening the i-th box in the sequence is equal to the probability that the first $i-1$ rewards are all 0.

Hence, the optimal is:

$$\textsc{Opt}\big(\mathbf{D}\big) = \frac{\epsilon}{n}\left(1 + \left(1 - \frac{1+\epsilon}{n}\right) + \cdots + \left(1 - \frac{1+\epsilon}{n}\right)^{n^+-1}\right).$$

The expected value of the hypothesis is at most (when it opens the $n^+ - k^+$ boxes with reward distributions equal to D^+ first):

$$h\big(\mathbf{D}\big) \le \frac{\epsilon}{n}\left(\sum_{i=1}^{n^+-k^+}\left(1 - \frac{1+\epsilon}{n}\right)^{i-1} - \left(1 - \frac{1+\epsilon}{n}\right)^{n^+-k^+}\sum_{i=1}^{k^-}\left(1 - \frac{1-\epsilon}{n}\right)^{i-1}\right)$$

Therefore, we have:

$$\begin{aligned}
\textsc{Opt}(\mathbf{D}) - h(\mathbf{D}) &\geq \frac{\epsilon}{n}\left(1-\frac{1+\epsilon}{n}\right)^{n^+-k^+}\left(\sum_{i=1}^{k^+}\left(1-\frac{1+\epsilon}{n}\right)^{i-1}+\sum_{i=1}^{k^-}\left(1-\frac{1-\epsilon}{n}\right)^{i-1}\right)\\
&\geq \frac{\epsilon}{n}\left(\sum_{i=1}^{k^+}\left(1-\frac{1+\epsilon}{n}\right)^{n}+\sum_{i=1}^{k^-}\left(1-\frac{1+\epsilon}{n}\right)^{n}\right)\\
&= \frac{\epsilon k}{n}\left(1-\frac{1+\epsilon}{n}\right)^{n}\\
&\geq \frac{\epsilon k}{n}\exp\left(-2-2\epsilon\right).
\end{aligned}$$

The last inequality is due to $1-x > e^{-2x}$ for $0 < x < \frac{1}{2}$. ■

Lemma 49 *For any algorithm A, any box $1 \leq i \leq n$, and any two neighboring instances $\mathbf{D}^+$ and $\mathbf{D}^-$ that differ only in the i-th coordinate, we have:*

$$\mathbf{Pr}\left[A \text{ makes a mistake on box } i \text{ w.r.t. } \mathbf{D}^+\right] + \mathbf{Pr}\left[A \text{ makes a mistake on box } i \text{ w.r.t. } \mathbf{D}^-\right] \geq \Omega(1)\,.$$

Proof By definition, any hypothesis h makes a mistake on box i w.r.t. either $\mathbf{D}^+$ or $\mathbf{D}^-$. Let $\mathcal{H}^+$ and $\mathcal{H}^-$ denote the two subsets of hypotheses respectively. On the one hand, we have:

$$\mathbf{Pr}\left[A \text{ picks } h \in \mathcal{H}^+ \text{ given samples from } \mathbf{D}^+\right] + \mathbf{Pr}\left[A \text{ picks } h \in \mathcal{H}^- \text{ given samples from } \mathbf{D}^+\right] = 1\,.$$

On the other hand, with less than $c \cdot \frac{n}{\epsilon^2}$ samples for some sufficiently constant $c > 0$, and by Lemma 47, we have:

$$\begin{aligned}
&\mathbf{Pr}\left[A \text{ picks } h \in \mathcal{H}^- \text{ given samples from } \mathbf{D}^+\right]\\
\geq &\mathbf{Pr}\left[A \text{ picks } h \in \mathcal{H}^- \text{ given samples from } \mathbf{D}^-\right] - O(1)\,,
\end{aligned}$$

for a sufficiently small constant inside the big-O notation. Putting together proves the lemma. ■

As a direct corollary, we have the following via a simple counting argument.

Lemma 50 *There is an instance $\mathbf{D}$ for which the algorithm makes $\Omega(n)$ mistakes in expectation.*

Proof By Lemma 49, if $\mathbf{D}$ is chosen from the 2^n possible instances uniformly at random, the algorithm makes a mistake on each box i with constant probability. So the lemma follows. ■

Proof [Proof of Theorem 32 (Lower Bound)] Consider the instance in Lemma 50. Suppose the algorithm makes αn mistakes in expectation where $\alpha > 0$ is a constant. Then, by a standard probability argument, the probability that it makes at least $\frac{\alpha n}{2}$ mistakes is at least $\frac{1}{2}$. Hence, by Lemma 48, the expected additive error is at least $\Omega(\epsilon)$. ■

Appendix E. Classification Problems: A Preliminary Discussion

In classification problems, there is a special data dimension which corresponds to the labels; the rest of the data dimensions correspond to the features. In particular, it is crucial that the labels are correlated with the features. Therefore, the assumption of independent data dimensions fail to hold. Nevertheless, below we present a straightforward extension of Theorem 5 under the assumption that the distribution of features *conditioned on any given label* is a product distribution. Although this preliminary result still relies on too strong an assumption to be useful in natural classification problems, we hope that it will serve as a stepping stone for follow-up works. See Section 5 for some related research directions.

The rest of the section follows the notations in classification problems and denotes each data point as a feature-label pair $(\mathbf{x}, y)$, where $\mathbf{x}$ is the feature vector and y is the label. We assume that there are ℓ labels $[\ell] = \{1, 2, \ldots, \ell\}$. Let $\mathbf{T} = \prod_{i=1}^{n} T_i$ denote an n-dimensional feature domain. Hence, the data domain under the model in Section 2 is $\mathbf{T} \times [\ell]$. Let D_Y denote the distribution of labels. Further, for any label $y \in [\ell]$, let $\mathbf{D}_{\mathbf{X}|y}$ denote a *product* distribution of features conditioned on having label y. For simplicity of notation, let $\mathbf{D}_{\mathbf{X}}$ denote the collection of conditional product feature distributions, and write $\mathbf{D} = \mathbf{D}_{\mathbf{X}} \circ D_Y$ be the joint distribution of feature-label pairs. By definition, the probability mass function of the joint distribution is:

$$f_{\mathbf{D}}(\mathbf{x}, y) = f_{D_Y}(y) \cdot f_{\mathbf{D}_{\mathbf{X}|y}}(\mathbf{x}) \,. \tag{16}$$

We say that such a distribution has product conditional feature distributions.

Generalized Product Empirical Distribution. We now generalize the definition of product empirical distribution to classification problems that have product conditional feature distributions. Let the empirical distribution of labels E_Y be the uniform distribution over sample labels. Further, for any label $y \in [\ell]$, let $\mathbf{E}_{\mathbf{X}|y}$ be the product empirical distribution w.r.t. the samples with label y. Concretely, for any $i \in [n]$, let the i-th coordinate of $\mathbf{E}_{\mathbf{X}|y}$ be a uniform distribution over the i-th coordinate of the samples with label y. As before, let $\mathbf{E}_{\mathbf{X}}$ denote the collection of product empirical feature distributions conditioned on the labels. Finally, let $\mathbf{E} = \mathbf{E}_{\mathbf{X}} \circ E_Y$.

By definition, the probability mass function of the joint distribution is:

$$f_{\mathbf{E}}(\mathbf{x}, y) = f_{E_Y}(y) \cdot f_{\mathbf{E}_{\mathbf{X}|y}}(\mathbf{x}) \,. \tag{17}$$

Finally, we define the *product empirical risk minimizer* (PERM) to be the best hypothesis w.r.t. $\mathbf{E}$. Here, note that we seek to minimize the objective.

Theorem 51 *Let $\mathbf{D} = \mathbf{D}_{\mathbf{X}} \circ D_Y$ be any distribution with product conditional feature distributions, over $\mathbf{T} \times [\ell]$ such that $|T_i| \leq k$ for any $1 \leq i \leq n$. For a sufficiently large constant $C > 0$, suppose the number of samples is at least:*

$$C \cdot \frac{nk\ell}{\epsilon^2} \log\left(\frac{\ell}{\delta}\right)$$

Then, with probability at least $1 - \delta$, for any $h : \mathbf{T} \times [\ell] \mapsto [0, 1]$, we have:

$$\left|h(\mathbf{D}) - h(\mathbf{E})\right| \leq \epsilon \,.$$

In particular, the PERM is an ϵ-additive approximation.

Proof Similar to the proof of Theorem 5, we rely on Lemma 1. It suffices to show that:

$$\delta(\mathbf{D}, \mathbf{E}) \leq \epsilon \,.$$

To do so, we first decompose it into two parts, the error due to the estimation of the label distribution, and that due to the conditional feature distributions. By Eqn. (16) and Eqn. (17):

$$\begin{aligned}
\delta(\mathbf{D}, \mathbf{E}) &= \frac{1}{2} \sum_{y \in [\ell]} \sum_{\mathbf{x} \in \mathbf{T}} \left| f_{D_Y}(y) \cdot f_{\mathbf{D}_{\mathbf{X}|y}}(\mathbf{x}) - f_{E_y}(y) \cdot f_{\mathbf{E}_{\mathbf{X}|y}}(\mathbf{x}) \right| \\
&\leq \frac{1}{2} \sum_{y \in [\ell]} \sum_{\mathbf{x} \in \mathbf{T}} \left(\left| f_{D_y}(y) - f_{E_y}(y) \right| \cdot f_{\mathbf{D}_{\mathbf{X}|y}}(\mathbf{x}) + f_{E_y}(y) \cdot \left| f_{\mathbf{D}_{\mathbf{X}|y}}(\mathbf{x}) - f_{\mathbf{E}_{\mathbf{X}|y}}(\mathbf{x}) \right| \right) \\
&= \delta(D_y, E_y) + \sum_{y \in [\ell]} f_{E_y}(y) \cdot \delta(\mathbf{D}_{\mathbf{X}|y}, \mathbf{E}_{\mathbf{X}|y}) \,.
\end{aligned}$$

By Lemma 6 and the stated number of samples, the squared Hellinger distance between the label distributions D_y and E_y is less than $\frac{\epsilon^2}{8}$. Further by the relation between the total variation and Hellinger distances, i.e., Lemma 2, we get that the first term on the RHS above is at most $\frac{\epsilon}{2}$.

It remains to bound the second term, i.e., the error due to the estimation of the conditional feature distributions. Fix any label $y \in [\ell]$. By definition, the number of samples with label y is $f_{E_Y}(y)N$. Therefore, by Lemma 6 and the stated number of samples, the squared Hellinger distance between the feature distributions conditioned on y is at most:

$$H^2(\mathbf{D}_{\mathbf{X}|y}, \mathbf{E}_{\mathbf{X}|y}) \leq \frac{\epsilon^2}{8\ell f_{E_Y}(y)} \,.$$

Further by Lemma 2, their total variation distance is at most:

$$\delta(\mathbf{D}_{\mathbf{X}|y}, \mathbf{E}_{\mathbf{X}|y}) \leq \frac{\epsilon}{2} \cdot \frac{1}{\sqrt{\ell f_{E_Y}(y)}} \,.$$

Hence, the second term is at most:

$$\begin{aligned}
\sum_{y \in [\ell]} f_{E_Y}(y) \cdot \delta(\mathbf{D}_{\mathbf{X}|y}, \mathbf{E}_{\mathbf{X}|y}) &\leq \sum_{y \in [\ell]} \frac{\epsilon}{2} \cdot \sqrt{\frac{f_{E_Y}(y)}{\ell}} \\
&\leq \frac{\epsilon}{2} \,.
\end{aligned}$$

The second inequality follows by $\sum_{y \in [\ell]} f_{E_Y}(y) = 1$ and the Cauchy-Schwartz inequality. ■