Dualizing Le Cam's method for functional estimation, with applications to estimating the unseens

Yury Polyanskiy and Yihong Wu*

January 13, 2022

Abstract

Le Cam's method (or the two-point method) is a commonly used tool for obtaining statistical lower bound and especially popular for functional estimation problems. This work aims to explain and give conditions for the tightness of Le Cam's lower bound in functional estimation from the perspective of convex duality. Under a variety of settings it is shown that the maximization problem that searches for the best two-point lower bound, upon dualizing, becomes a minimization problem that optimizes the bias-variance tradeoff among a family of estimators. For estimating linear functionals of a distribution our work strengthens prior results of Donoho-Liu [DL91] (for quadratic loss) by dropping the Hölderian assumption on the modulus of continuity. For exponential families our results extend those of Juditsky-Nemirovski [JN09] by characterizing the minimax risk for the quadratic loss under weaker assumptions on the exponential family.

We also provide an extension to the high-dimensional setting for estimating separable functionals. Notably, coupled with tools from complex analysis, this method is particularly effective for characterizing the "elbow effect" – the phase transition from parametric to nonparametric rates. As the main application of our methodology, we consider three problems in the area of "estimating the unseens", recovering the prior result of [PSW17] on population recovery and, in addition, obtaining two new ones:

- *Distinct elements problem*: Randomly sampling a fraction p of colored balls from an urn containing d balls in total, the optimal normalized estimation error of the number of distinct colors in the urn is within logarithmic factors of $d^{-\frac{1}{2}\min\{\frac{p}{1-p},1\}}$, exhibiting an elbow at $p=\frac{1}{2}$;
- *Fisher's species problem*: Given n independent observations drawn from an unknown distribution, the optimal normalized prediction error of the number of unseen symbols in the next (unobserved) $r \cdot n$ observations is within logarithmic factors of $n^{-\min\{\frac{1}{r+1},\frac{1}{2}\}}$, exhibiting an elbow at $r=1$.

*Y.P. is with the Department of EECS, MIT, Cambridge, MA, email: yp@mit.edu. Y.W. is with the Department of Statistics and Data Science, Yale University, New Haven, CT, email: yihong.wu@yale.edu.

Contents

1 Introduction

Le Cam's method (or the two-point method) is a commonly used tool for obtaining statistical lower bound [LC73, Yu97]. The rationale is that if two hypotheses are statistically indistinguishable, then the difference of their parameters presents a lower bound to the accuracy of any estimator. Although Le Cam's method can be loose in estimating high-dimensional parameters as it may not capture the correct dependency on the dimension, for estimating scalar-valued functionals it often yields the tight minimax rate even when the underlying parameter is high-dimensional. This work aims to explain and give conditions for the tightness of Le Cam's lower bound in functional estimation from the perspective of convex duality.

Let Θ and $\mathcal{X}$ be measurable spaces and P a transition probability kernel from Θ to $\mathcal{X}$. Then $\{P_\theta = P(\cdot|\theta) \colon \theta \in \Theta\}$ is a parametric family of distributions on $\mathcal{X}$. Let Π be a given subset of $\mathcal{P}(\Theta)$, the set of all probability measures on Θ. Let $T(\pi)$ be a real-valued affine functional of $\pi \in \Pi$. Denote the observations by $\boldsymbol{X} = (X_1, \ldots, X_n)$ and the latent parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, where conditioned on $\boldsymbol{\theta}$, X_i's are independent and distributed as

$$X_i \overset{\text{ind.}}{\sim} P_{\theta_i}, \quad i = 1, \ldots, n \tag{1}$$

Furthermore, we shall focus on the following settings, which are commonly assumed in empirical Bayes and compound estimation problems, respectively [Rob51, Zha97, Zha03].

- $\theta_i \overset{\text{iid}}{\sim} \pi$ for some $\pi \in \Pi$. In this case, $X_i \overset{\text{iid}}{\sim} \pi P$, where $\pi P \triangleq \int_\Theta P_\theta \pi(d\theta)$ denotes the mixture distribution induced by the mixing distribution (prior) π. Given $X_1, \ldots, X_n$, the goal is to estimate the functional $T(\pi)$ or π itself.
- θ_i's are deterministic whose empirical distribution $\pi_{\boldsymbol{\theta}} \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$ belongs to Π. Given $X_1, \ldots, X_n$, the goal is to estimate the functional $T(\pi_{\boldsymbol{\theta}})$ or the empirical distribution $\pi_{\boldsymbol{\theta}}$ itself. For affine functional $T(\pi) = \int h(\theta)\pi(d\theta)$, $T(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1} h(\theta_i)$ has a separable form and can be non-linear in the parameter $\boldsymbol{\theta}$.

The minimax quadratic risks of estimating the functional T in the iid and deterministic setting are defined respectively as:

$$R^*_{\text{iid}}(n) \triangleq \inf_{\hat{T}} \sup_{\pi \in \Pi} \mathbb{E}[|\hat{T}(X_1, \ldots, X_n) - T(\pi)|^2], \tag{2}$$

$$R^*_{\text{det}}(n) \triangleq \inf_{\hat{T}} \sup_{\boldsymbol{\theta}: \pi_{\boldsymbol{\theta}} \in \Pi} \mathbb{E}[|\hat{T}(X_1, \ldots, X_n) - T(\pi_{\boldsymbol{\theta}})|^2]. \tag{3}$$

The main result of this paper is the following: Under appropriate technical conditions such as the convexity and compactness of the space Π, for affine functional T, Le Cam's lower bound is tight up to universal constant factors for both iid and deterministic settings. More precisely, we have

$$R^*_{\text{iid}}(n) \asymp R^*_{\text{det}}(n) \asymp \max_{\theta, \theta' \in \Theta} \left\{ |T(\pi) - T(\pi')|^2 : \chi^2(\pi P \| \pi' P) \le \frac{1}{n} \right\}, \tag{4}$$

where $\chi^2(\cdot\|\cdot)$ denotes the χ^2-divergence. In the iid setting, this result strengthens the celebrated result of Donoho-Liu [DL91] for linear functionals. In addition, we show a counterpart of this characterization also holds for exponential families for estimating functionals linear in the mean parameters, where the χ^2-divergence in (4) is replaced by the squared Hellinger distance, extending the result of Juditsky-Nemirovski [JN09] to the quadratic risk and relaxing the assumptions. See Section 1.1 for more discussion. Throughout the paper we focus on the expected risk under the quadratic loss (for which the χ^2-divergence is a natural choice). As will be shown later, these results can be easily extended to high-probability risk bounds.

We now explain the intuition behind the main result (4). In the iid setting where $X_i \overset{\text{iid}}{\sim} \pi P$, the lower bound is a straightforward application of Le Cam's two point method, since $\chi^2(\pi P \| \pi' P) \leq \frac{1}{n}$ implies that the total variation of the product distributions $(\pi P)^{\otimes n}$ and $(\pi' P)^{\otimes n}$ is bounded away from one and thus impossible to be tested reliably given the sample $X_1, \ldots, X_n$, leading to a lower bound on the order of $|T(\pi) - T(\pi')|^2$. In the deterministic setting, the lower bound is shown by a generalized version of Le Cam's method using two priors (also known as fuzzy hypotheses testing [Tsy09, Sec. 2.7.4]), where we consider θ_i's drawn from the product distribution $\pi^{\otimes n}$ or $\pi'^{\otimes n}$ with appropriate truncation.

What is more surprising is perhaps the upper bound in (4), which in fact holds *without additional constant factors* in both the iid and deterministic settings. The key observation is that the maximization in (4) is a convex optimization problem, whose dual corresponds to a minimization problem that optimizes the bias-variance tradeoff of estimators of the following type:

$$\hat{T} = \frac{1}{n} \sum_{i=1}^{n} g(X_i). \tag{5}$$

In the absence of a duality gap, this shows the achievability of Le Cam's lower bound by optimizing the choice of g.

The duality view underlying the main result (4) is in fact natural. Indeed, the classical minimax theorem in decision theory states that, under regularity assumptions (cf. e.g. [Str85, Theorem 46.5]) the minimax risk and the least favorable Bayes risk coincide, namely

$$\inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[(\hat{T} - T(\theta))^2] = \sup_{\pi} \inf_{\hat{T}} \mathbb{E}_{\theta \sim \pi}[(\hat{T} - T(\theta))^2], \tag{6}$$

where the supremum on the right is taken over all priors on Θ. This can also be interpreted from the duality perspective,[1] where the primal variables corresponds to (randomized) estimators and the dual variables correspond to priors. However, the duality view of (6) is unwieldy except in special cases, due to the difficulty of finding the least favorable prior that maximizes the Bayes risk for large sample size. In this vein, the more effective result of (4) can be viewed as approximate version of the general minimax theorem for functional estimation.

To produce concrete results of minimax rate for specific applications, one needs to evaluate the value of the maximum in (4). Using tools from complex analysis, we do so for a number of problems and obtain new results on the optimal rate of convergence, characterizing, in particular, the "elbow effect", that is, the phase transition from parametric to nonparametric rates. As the main application of our methodology, we consider three problems in the area of "estimating the unseens", namely, *population recovery*, *distinct elements problem*, and *Fisher's species problem*. In addition to recovering the prior result of [PSW17] on the optimal rate of population recovery, we establish the following new results:

- Distinct elements problem: Randomly sampling a fraction p of colored balls from an urn containing n balls in total, the goal is to estimate the number of distinct colors in the urn [RRSS09, Val11, WY18]. We show that, as $n \to \infty$, the optimal normalized estimation error is within logarithmic factors of $n^{-\frac{1}{2}\min\{\frac{p}{1-p}, 1\}}$, exhibiting an elbow at $p = \frac{1}{2}$;

[1] This follows from standard arguments in optimization by rewriting the left-hand side as $\inf_{\hat{T}}\{t : \mathbb{E}_\theta[(\hat{T} - T(\theta))^2] \leq t, \forall \theta \in \Theta\}$ and the Lagrange multipliers correspond to priors. When both X and θ are finitely-valued, (6) is simply the duality of linear programming (LP).

- Fisher's species problem: Given n independent observations drawn from an unknown distribution, the goal is to predict the number of unseen symbols in the next (unobserved) $r \cdot n$ samples [FCW43, ET76, OSW16]. We show that, as $n \to \infty$, the optimal normalized prediction error is within logarithmic factors of $n^{-\min\{\frac{1}{r+1},\frac{1}{2}\}}$, exhibiting an elbow at $r = 1$.

We emphasize that the main focus of this paper to determine the minimax rate by means of convex duality without demonstrating an explicit choice of the optimal estimator. This is conceptually distinct from existing explicit construction of estimators such as smoothed estimators in the context of the species problem [OSW16] (which do not attain the optimal rate). Nevertheless, since the minimax rate in both iid and deterministic settings can be achieved by the estimator (5) parameterized by g, the optimal choice of g corresponds to the optimizer of certain convex optimization problem (cf. (24)), which can be solved efficiently in the case of finite Θ and $\mathcal{X}$ (e.g. the population recovery problem). Even for continuous models, this infinite-dimensional problem can often be effectively discretized leading to computational efficient construction of optimal estimators. For example, for the distinct elements and species problems, the estimators can be constructed in polynomial time as solutions to certain linear programs (see (84) and (98) respectively).

Before discussing the related literature, let us mention that the duality-based method in this paper need not be limited to functional estimation. In a companion paper [JPW20] we extend the methods to estimating the distribution itself (with respect to the total variation loss) in the context of the distinct elements problem. The connection to functional estimation is that estimating the distribution in total variation is equivalent to simultaneously estimating all bounded linear functionals; this view enables us to analyze Wolfowitz's *minimum-distance estimators* [Wol57] in the duality framework.

1.1 Related work

A celebrated result of Donoho-Liu [DL91] relates the minimax rate of estimating linear functionals to the Hellinger modulus of continuity. For the density estimation models, under certain assumptions, it is shown that the minimax rate coincides with the right-hand side of (4) with H^2 in place of the χ^2-divergence.[2] However, the constant factors may not be universal and depend on the problem or its hyper-parameters, thus precluding the application to high-dimensional problems. More importantly, the proof (of the upper bound) in [DL91] is based on constructing an estimator via pairwise hypotheses tests, by means of a binary search on the functional value. While this method can deal with general loss function, the limitation is that it assumes the Hölderianity of the modulus of continuity in order to show tightness. We refer the readers to Section 2.1 for a detailed comparison of the results.

The prior work that is closest to ours in spirit is that of Juditsky-Nemirovski [JN09] (cf. also the recent monograph [JN20]), where the main technology was also convex optimization and the minimax theorem. As opposed to the squared loss, they considered the ϵ-quantile loss, namely, an upper bound on the estimation accuracy that holds with probability at least $1-\epsilon$ for all parameters. For exponential families, under certain convexity assumptions, it is shown that the minimax ϵ-quantile risk is determined within absolute constant factors by the Hellinger modulus of continuity provided that ϵ is not too small. We extend this result to quadratic risk under more relaxed assumptions (see Section 4.1 for details). Note that the quadratic risk result cannot be obtained through the usual route of integrating the high-probability risk bound, since the optimal estimator for an ϵ-quantile loss potentially depends on ϵ. On the other hand, results on ϵ-quantile loss can be obtained from the quadratic risk by sampling splitting and applying the median (although the more

[2] The resulting moduli of continuity are in fact the same up to constant factors, as we show in Proposition 3.

direct argument in [JN09] achieves better constant factors). Nevertheless, the main advantage of our approach lies in its versatility, as witnessed, e.g., by the treatment of the deterministic setting.

Finally, let us mention that while the main objective of [JN09] was to obtain rate-optimal estimators by means of convex programming, in this paper and similar to the program in [DL91], the characterization of the minimax risk by convex optimization is mostly used as a mathematical tool for determining the minimax rates, although it also leads to efficient construction of estimations for specific problems.

1.2 Organization

The rest of the paper is organized as follows. Section 2 presents the main result for the iid setting. We provide two examples: population recovery (Section 2.2) and interval censoring (Section 2.3), which are finite-dimensional and infinite-dimensional application of the main theorem respectively. Section 3 extend the result to the deterministic setting for estimating separable functionals. The methods are then applied to the distinct elements problem (Section 3.1) and Fisher's species extrapolation problem (Section 3.2) to determine the minimax rates of convergence up to logarithmic factors. Finally, in Section 4 we extend the result to exponential families.

Section 5 contains the proofs of Theorems 10–12; further technical results and proofs are collected in Appendices B and C. Appendix A discusses applications to classical problems of density estimation and the Gaussian white noise model, the latter constituting a simple self-contained example where the dualization of Le Cam's lower bound can be carried out explicitly.

2 Iid setting

Recall that in the iid setting of (1), the sample consists of $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \pi P$. When P is the identity kernel, X_i's are directly drawn from π; otherwise, they are indirect observations. The minimax quadratic risk of estimating $T(\pi)$ over $\pi \in \Pi$ is denoted by $R^*_{\text{iid}}(n)$ in (2).

Define the modulus of continuity of functional T with respect to various distances (and quasi-distances) between distributions πP:

$$\delta_{\chi^2}(t) = \sup\{T(\pi') - T(\pi) : \chi^2(\pi' P \| \pi P) \le t^2, \pi, \pi' \in \Pi\} \tag{7}$$

$$\delta_{H^2}(t) = \sup\{T(\pi') - T(\pi) : H^2(\pi' P, \pi P) \le t^2, \pi, \pi' \in \Pi\} \tag{8}$$

$$\delta_{\text{TV}}(t) = \sup\{T(\pi') - T(\pi) : \text{TV}(\pi' P, \pi P) \le t, \pi, \pi' \in \Pi\} \tag{9}$$

where $\text{TV}(F, G) = \sup_E |F(E) - G(E)|$ is the total variation, $H^2(F, G) = \int d\nu \left(\sqrt{\frac{dF}{d\nu}} - \sqrt{\frac{dG}{d\nu}}\right)^2$ is the squared Hellinger distance (with ν being any dominating measure s.t. $F \ll \nu$ and $G \ll \nu$, e.g. $\nu = F + G$). Finally, the χ^2-divergence is defined as $\chi^2(F\|G) = \infty$ if $F \not\ll G$ and otherwise $\chi^2(F\|G) = \int dG \left(\frac{dF}{dG}\right)^2 - 1$. We note that $\text{TV}(F, G)$ and $H(F, G)$ are distances on $\mathcal{P}$. For a signed measure μ its total variation norm is denoted $\|\mu\|_{\text{TV}}$, so that $\text{TV}(F, G) = \|F - G\|_{\text{TV}}$.

Our main result is the following:

Theorem 1. *Suppose that $(\Theta, \mathcal{X}, P, T, \Pi)$ satisfy the following assumptions:*

A1 The functional $\pi \mapsto T(\pi)$ is affine;

A2 The set Π is convex;

A3 There exists a vector space of functions $\mathcal{F}$ on $\mathcal{X}$ such that $\mathcal{F}$ contains constants and is dense in $L_2(\mathcal{X}, \pi P)$ for every $\pi \in \Pi$;

A4 There exists a topology on Π such that:

A4a It is coarse enough that Π is compact;

A4b It is fine enough that $T(\pi)$, $\pi P f$ and $\pi P(f^2)$ are continuous in $\pi \in \Pi$ for all $f \in \mathcal{F}$.

Then

$$\frac{1}{(1+\sqrt{e})^2}\delta_{\chi^2}(\tfrac{1}{\sqrt{n}})^2 \le R^*(n) \le \delta_{\chi^2}(\tfrac{1}{\sqrt{n}})^2, \tag{10}$$

where the upper bound can be achieved by an estimator of the form $\hat{T} = \frac{1}{n}\sum_{i=1}^n g(X_i)$ for some $g \in \mathcal{F}$.

Corollary 2. *There exist an absolute constant $c > 0$ with the following property. In the setting of the previous Theorem for all $e^{-2n} \le \epsilon \le \frac{1}{16}$ we have*

$$\frac{1}{c}\delta_{\chi^2}\left(\sqrt{\frac{1}{n}\ln\frac{1}{\epsilon}}\right) \le R^*_{\text{iid}}(n,\epsilon) \le c\delta_{\chi^2}\left(\sqrt{\frac{1}{n}\ln\frac{1}{\epsilon}}\right), \tag{11}$$

where the ϵ-risk (confidence interval) is defined as

$$R^*_{\text{iid}}(n,\epsilon) \triangleq \inf_{\hat{T}} \sup_{\pi\in\Pi} \inf\{\rho : \mathbb{P}[|\hat{T}(X_1,\ldots,X_n) - T(\pi)| > \rho] \le \epsilon\}.$$

Some remarks are in order:

1. If Θ and $\mathcal{X}$ are finite, then $\mathcal{F}$ can be taken to be all functions on $\mathcal{X}$ and assumptions A3 and A4 are automatic.

2. If $\mathcal{X}$ is a normal topological space, then every probability measure ν is regular [DS58, IV.6.2] and the set $\mathcal{F}$ of all bounded continuous functions is dense in $L_2(\mathcal{X}, \nu)$, cf. [DS58, IV.8.19]. Other convenient choices of $\mathcal{F}$ are all Lipschitz functions (and Wasserstein W_1-convergence), all polynomials, trigonometric polynomials or sums of exponentials.

3. The continuity of $\pi P f$ under the weak topology on Π can be assured by demanding the following (strong Feller) property for kernel P: For any bounded measurable f, Pf is bounded continuous.

To prove Theorem 1, we start with several general properties and comparisons of various moduli of continuity.

Proposition 3. *Let $T(\pi)$ be affine in π. Assume that Π is convex. Then*

1. *(Concavity) $\delta_{H^2}(\sqrt{\cdot})$, $\delta_{\text{TV}}(\cdot)$ and $\delta_{\chi^2}(\sqrt{\cdot})$ are concave.*

2. *(Subadditivity) For any $c \in [0,1]$ and $t \ge 0$ we have:*

$$\delta_{\text{TV}}(ct) \ge c\delta_{\text{TV}}(t) \tag{12}$$

$$\delta_{H^2}(ct) \ge c^2\delta_{H^2}(t) \tag{13}$$

$$\delta_{\chi^2}(ct) \ge c^2\delta_{\chi^2}(t) \tag{14}$$

3. *(Comparison of various δ's) For all* $t \geq 0$ *we have*

$$\frac{1}{2}\delta_{H^2}(t) \leq \delta_{\chi^2}(t) \leq \delta_{H^2}(t) \leq \delta_{\mathrm{TV}}(t) \leq \delta_{H^2}(\sqrt{2t})\,. \tag{15}$$

4. *(Superlinearity) Let* $\Delta_{\max} \triangleq \sup\{T(\pi') - T(\pi) : \pi, \pi' \in \Pi\}$, *then*

$$\delta_{H^2}(t) \geq \delta_{\chi^2}(t) \geq \Delta_{\max}\frac{t}{2} \qquad \forall 0 < t \leq 1\,. \tag{16}$$

Proof. The first property follows from the convexity of $\mathrm{TV}(P,Q)$, $H^2(P,Q)$ and $\chi^2(P\|Q)$ in the pair (P,Q). The second one follows from the first and the fact that $\delta(0) = 0$. For the third, we recall standard bounds (cf. e.g. [Tsy09, Sec. 2.4.1]): For any pair of distributions P, Q we have

$$H^2(P,Q)/2 \leq \mathrm{TV}(P,Q) \leq H(P,Q)\,, \tag{17}$$

and

$$H^2(P,Q) \leq 2 - \frac{2}{\sqrt{1+\chi^2(P\|Q)}} \leq \chi^2(P\|Q)\,. \tag{18}$$

Together (17) and (18) establish all inequalities in (15) except the left-most one. For the latter we recall from [LC86, p. 48]:

$$\frac{1}{2}H^2(P,Q) \leq \chi^2\left(P\middle\|\frac{P+Q}{2}\right) \leq H^2(P,Q)\,. \tag{19}$$

Thus, for any (π, π') that are feasible for the $\delta_{H^2}(t)$ problem, $\pi_0 \triangleq \frac{\pi+\pi'}{2}$ and $\pi_0' \triangleq \pi'$ are feasible for the $\delta_{\chi^2}(t)$ problem, since $\chi^2(\pi_0' P \| \pi_0 P) \leq t^2$ according to (19), and satisfy $|T(\pi_0) - T(\pi_0')| = \frac{1}{2}|T(\pi) - T(\pi')|$.

Finally, for (16), consider any pair of distributions π_1, π_0 such that $T(\pi_0) - T(\pi_1) = \Delta > 0$. From the data-processing inequality we have for all $0 < \lambda < 1$:

$$\chi^2(\lambda\pi_1 P + (1-\lambda)\pi_0 P \| \frac{1}{2}\pi_1 P + \frac{1}{2}\pi_0 P) \leq \chi^2(\mathrm{Bern}(\lambda)\|\mathrm{Bern}(1/2)) = (1-2\lambda)^2\,.$$

Therefore, setting $\pi' = \lambda\pi_1 + (1-\lambda)\pi_0$ and $\pi = \frac{1}{2}\pi_1 + \frac{1}{2}\pi_0$ with $1 - 2\lambda = t \leq 1$ we obtain

$$\delta_{\chi^2}(t) \geq T(\pi') - T(\pi) = \frac{\Delta}{2}(1-2\lambda) = \frac{\Delta}{2}t\,.$$

Optimizing over π_0, π_1 yields (16).

□

Proof of Theorem 1. The lower bound simply follows from the χ^2-version of Le Cam's method. Consider a pair of distributions π, π' such that $\chi^2(\pi' P \| \pi P) \leq \frac{1}{n}$ to be optimized. From the tensorization property of χ^2-divergence we have

$$\chi^2((\pi' P)^{\otimes n} \| (\pi P)^{\otimes n}) = (1 + \chi^2(\pi' P \| (\pi P)))^n - 1 \leq e - 1. \tag{20}$$

Using Brown-Low's two-point lower bound [BL96] and optimizing over the pair π, π', we have

$$R^*_{\mathrm{iid}}(n) \geq \sup_{\pi,\pi' \in \Pi : \chi^2(\pi' P\|\pi P) \leq \frac{1}{n}} \frac{(T(\pi) - T(\pi'))^2}{\left(1 + \sqrt{1 + \chi^2((\pi' P)^{\otimes n} \| (\pi P)^{\otimes n})}\right)^2} \geq \frac{\delta_{\chi^2}\left(\frac{1}{\sqrt{n}}\right)^2}{(1+\sqrt{e})^2}. \tag{21}$$

To prove an upper bound we consider estimators of the form

$$\hat{T}_g = \frac{1}{n}\sum_{i=1}^n g(X_i)\,, \tag{22}$$

where $g \in \mathcal{F}$. We analyze the quadratic risk of this estimator by decomposing it into bias and variance part:

$$\mathbb{E}_{X_i \overset{\text{iid}}{\sim} \pi P}[|\hat{T}_g - T(\pi)|^2] \le \frac{1}{n}\mathrm{Var}_{\pi P}[g] + |T(\pi) - \pi P g|^2. \tag{23}$$

Taking worst-case π and optimizing over g we get

$$\sqrt{R^*_{\text{iid}}(n)} \le \inf_{g\in\mathcal{F}}\sup_{\pi\in\Pi}\left\{\frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\pi P}[g]} + |T(\pi) - \pi P g|\right\} = \delta_{\text{bv}}(\tfrac{1}{\sqrt{n}})\,,$$

where

$$\delta_{\text{bv}}(t) \triangleq \inf_{g\in\mathcal{F}}\sup_{\pi\in\Pi}\left\{t\sqrt{\mathrm{Var}_{\pi P}[g]} + |T(\pi) - \pi P g|\right\}. \tag{24}$$

The proof is completed by applying the next proposition. □

Proposition 4. *Under the conditions of Theorem 1, we have*

$$\delta_{\text{bv}}(t) \le \delta_{\chi^2}(t) \qquad \forall t \ge 0. \tag{25}$$

Furthermore, the supremum over π, π' in the definition of δ_{χ^2} is achieved: There exist $\pi_*, \pi'_* \in \Pi$ s.t. $\delta_{\chi^2}(t) = T(\pi'_*) - T(\pi_*)$ and $\chi^2(\pi'_* P \| \pi_* P) \le t^2$.

Before proving the proposition, we recall the minimax theorem due to Ky Fan [Fan53, Theorem 2]:[3]

Theorem 5 (Ky Fan)**.** *Let X be a compact space and Y an arbitrary set (not topologized). Let $f : X \times Y \to \mathbb{R}$ be such that for every $y \in Y$, $x \mapsto f(x, y)$ is upper semicontinuous on X. If f is concave-convex-like on $X \times Y$, then*

$$\max_{x\in X}\inf_{y\in Y} f(x,y) = \inf_{y\in Y}\max_{x\in X} f(x,y).$$

We recall that the function f is concave-convex-like on $X \times Y$ if a) for any two $x_1, x_2 \in X$ and $\lambda \in [0, 1]$ there exists $x_3 \in X$ such that for all $y \in Y$:

$$\lambda f(x_1, y) + (1-\lambda) f(x_2, y) \le f(x_3, y) \tag{26}$$

and b) for any two $y_1, y_2 \in Y$ and $\lambda \in [0, 1]$ there exists $y_3 \in Y$ such that for all $x \in X$:

$$\lambda f(x, y_1) + (1-\lambda) f(x, y_2) \ge f(x, y_3).$$

Proof of Proposition 4. We aim to apply the minimax theorem in order to get a more convenient expression for $\delta_{\text{bv}}(t)$. The function

$$(\pi, g) \mapsto \sqrt{\mathrm{Var}_{\pi P}[g]} + |T(\pi) - \pi P g|$$

[3] There it is stated for Hausdorff X, but this condition is not necessary, e.g., [BZ86]. Note that in defining convex-concave-like property we mandate it hold for all $0 \le t \le 1$ in (26), but it is also known that minimax theorem holds for functions that only satisfy, e.g., $t = 1/2$, see [Kön68].

satisfies all the conditions for applying Theorem 5 except for the concavity in π due to the last term (it is convex instead of concave). To mend this consider the following upper bound

$$|T(\pi) - \pi Pg| \le \sup_{\xi\in[0,2],\pi'\in\Pi} T(\pi) - \pi Pg - \xi(T(\pi') - \pi' Pg)\,.$$

Indeed, if $T(\pi) - \pi Pg > 0$, take $\xi = 0$; otherwise, take $\pi' = \pi, \xi = 2$.

So letting $u = (\pi, \pi', \xi) \in U \triangleq \Pi \times \Pi \times [0, 2]$ we consider the following function on $U \times \mathcal{F}$:

$$F_t(u, g) \triangleq T(\pi) - \pi Pg - \xi(T(\pi') - \pi' Pg) + t\sqrt{\mathrm{Var}_{\pi P}[g]}$$

We claim it is concave-convex-like. Convexity in g is easy: the term $|T(\pi) - \pi Pg|$ is clearly convex, whereas the convexity of $g \mapsto \sqrt{\mathrm{Var}_\mu[g]}$ follows from observation that without loss of generality we may assume $\mathbb{E}_\mu[g] = 0$ and then $\sqrt{\mathrm{Var}_\mu[g]} = \sqrt{\int g^2 d\mu} \triangleq \|g\|_{L_2(\mu)}$ is a norm (hence convex).

We proceed to checking the concave-like property of $F_t(u, g)$ in u. Define for convenience,

$$a(\pi) \triangleq T(\pi) - \pi Pg, \qquad b(\pi) = t\sqrt{\mathrm{Var}_{\pi P}[g]}$$

It is clear that $a(\pi)$ is affine, whereas $b(\pi)$ is concave. Indeed, $\sqrt{\cdot}$ is a concave and increasing scalar function, whereas $\mathrm{Var}_\mu[g] = \mu(g^2) - (\mu g)^2$ is concave in μ. So for $u = (\pi, \pi', \xi)$ we have

$$F_t(u, g) = a(\pi) - \xi a(\pi') + b(\pi)\,. \tag{27}$$

Consider $u_1 = (\pi_1, \pi_1', \xi_1)$ and $u_2 = (\pi_2, \pi_2', \xi_2)$ and $\lambda \in [0, 1]$. First, suppose that $\xi_1 = \xi_2 = 0$. We see that in this case

$$\lambda F_t(u_1, g) + (1 - \lambda)F_t(u_2, g) \le F_t(\lambda u_1 + (1 - \lambda)u_2, g)$$

since from (27) we see that F_t is concave in π. Then, taking $u_3 = \lambda u_1 + (1 - \lambda)u_2$ satisfies (26). Next, assume that either $\xi_1 > 0$ or $\xi_2 > 0$. Then define

$$\pi_3 \triangleq \lambda\pi_1 + (1 - \lambda)\pi_2\,, \quad \pi_3' \triangleq \frac{\lambda\xi_1}{\xi_3}\pi_1' + \frac{(1 - \lambda)\xi_2}{\xi_3}\pi_2', \quad \xi_3 \triangleq \lambda\xi_1 + (1 - \lambda)\xi_2\,.$$

And set $u_3 = (\pi_3, \pi_3', \xi_3)$. We claim that

$$\lambda F_t(u_1, g) + (1 - \lambda)F_t(u_2, g) \le F_t(u_3, g)\,. \tag{28}$$

Indeed, we have from affinity of $a(\cdot)$:

$$a\left(\frac{\lambda\xi_1}{\xi_3}\pi_1' + \frac{(1 - \lambda)\xi_2}{\xi_3}\pi_2'\right) = \frac{\lambda\xi_1}{\xi_3}a(\pi_1') + \frac{(1 - \lambda)\xi_2}{\xi_3}a(\pi_2')\,.$$

Therefore, we have

$$\begin{aligned}
\lambda a(\pi_1) + (1 - \lambda)a(\pi_2) &= a(\pi_3) \\
\lambda\xi_1 a(\pi_1') + (1 - \lambda)\xi_2 a(\pi_2') &= \xi_3 a(\pi_3') \\
\lambda b(\pi_1) + (1 - \lambda)b(\pi_2) &\le b(\pi_3)\,.
\end{aligned}$$

These three statements together with (27) prove (28).

Knowing that F_t is concave-convex-like, for applying the minimax theorem we only need to check that $u \mapsto F_t(u,g)$ is continuous for all g and that U is compact. This is satisfied by the assumption $A4$ of Theorem 1. Applying Theorem 5, we have

$$\delta_{\rm bv}(t) \le \inf_{g\in\mathcal{F}} \sup_{u\in U} F_t(u,g) = \inf_{g\in\mathcal{F}} \max_{u\in U} F_t(u,g) = \max_{u\in U} \inf_{g\in\mathcal{F}} F_t(u,g)\,. \tag{29}$$

(Note that the rightmost maximum exists thanks to Theorem 5.)

Next, to evaluate the rightmost term, fix $u = (\pi, \pi', \xi) \in U$ and consider the optimization

$$\psi_t(u) = \inf_{g\in\mathcal{F}} (\xi\pi' - \pi)Pg + t\sqrt{\mathrm{Var}_{\pi P}[g]}\,. \tag{30}$$

We claim that

$$\psi_t(u) = \begin{cases} -\infty, & \xi \neq 1 \\ -\infty, & \xi = 1, \chi^2(\pi' P \| \pi P) > t^2 \\ 0, & \text{otherwise} \end{cases} \tag{31}$$

which implies the desired (25) by continuing (29):

$$\max_{u\in U} \inf_{g\in\mathcal{F}} F_t(u,g) = \max\{T(\pi') - T(\pi) : \chi^2(\pi' P\|\pi P) \le t^2, \pi \in \Pi, \pi' \in \Pi\}\,.$$

To prove (31), we first recall that $\mathcal{F}$ contains constants. Thus if $\xi \neq 1$, we have that the first term in (30) can be driven to $-\infty$, while keeping the second term zero, by taking $g = c1$ and $c \to \pm\infty$. So fix $\xi = 1$. Recall a variational characterization of the χ^2-divergence:[4]

$$\chi^2(\mu\|\nu) = \sup_{g\in\mathcal{G}} \{|\mathbb{E}_\mu[g] - \mathbb{E}_\nu[g]|^2 : \mathrm{Var}_\nu[g] \le 1\}\,, \tag{32}$$

where $\mathcal{G}$ is any subset that is dense in $L_2(\nu)$. Thus, if $\chi^2(\pi' P\|\pi P) > t^2$ (in particular, if $\pi' P \not\ll \pi P$) there must exists $g_0 \in \mathcal{F}$ such that

$$\pi' P g_0 - \pi P g_0 < -t \qquad \mathrm{Var}_{\pi P}[g_0] \le 1$$

Thus taking $g = cg_0$ and $c \to \infty$ in (30) we again obtain that $\psi_t(u) = -\infty$. In the remaining case, $\chi^2(\pi' P\|\pi P) \le t^2$ and again from (32) we have that for any $g \in \mathcal{F}$

$$(\pi' - \pi)Pg \ge -t\sqrt{\mathrm{Var}_{\pi P}[g]}\,,$$

and thus $\psi_t(u) \ge 0$, while 0 is achievable by taking $g = 0$. $\square$

Proof of Corollary 2. Consider two distributions π and π' such that $H^2((\pi P)^{\otimes n}, (\pi' P)^{\otimes n}) = 2 - 2\beta$ then from [Tsy09, Theorem 2.2] we have that

$$R^*_{\rm iid}(n,\epsilon) \ge \frac{1}{2}|T(\pi) - T(\pi')| \tag{33}$$

provided that $\beta^2 > 1 - (1-2\epsilon)^2$. Thus, taking $\beta > \sqrt{4\epsilon}$ suffices. Recall the tensorization identity for H^2:

$$1 - \frac{1}{2}H^2((\pi P)^{\otimes n}, (\pi' P)^{\otimes n}) = \left(1 - \frac{1}{2}H^2(\pi P, \pi' P)\right)^n\,.$$

[4] For completeness, here is a short proof of (32). First, assume $\chi^2(\mu\|\nu) < \infty$. Denoting $f = \frac{d\mu}{d\nu}$ and assuming without loss of generality that $\mathbb{E}_\nu g = 0$ we have $|\mathbb{E}_\mu[g] - \mathbb{E}_\nu[g]|^2 = (\mathbb{E}_\nu[fg])^2 \le \mathrm{Var}_\nu g \mathrm{Var}_\nu f$, which completes the proof since $\mathrm{Var}_\nu f = \chi^2(\mu\|\nu)$. For the other direction, simply approximate f by elements of $\mathcal{G}$. If $\chi^2(\mu\|\nu) = \infty$, set $f_n = \min(f, n)$ and let $n \to \infty$.

Consequently, the bound (33) holds whenever $H^2(\pi P, \pi' P) \le t_n^2$ with $t_n^2 = 2 - 2(4\epsilon)^{\frac{1}{2n}}$. Note that for $e^{-2n} \le \epsilon \le \frac{1}{16}$ we always have $2 - 2(4\epsilon)^{\frac{1}{2n}} \ge \frac{1}{4n} \ln \frac{1}{\epsilon}$, implying

$$R^*_{\rm iid}(n,\epsilon) \ge \frac{1}{2}\delta_{H^2}\Big(\sqrt{\frac{1}{4n}\ln\frac{1}{\epsilon}}\Big) \ge \frac{1}{8}\delta_{\chi^2}\Big(\sqrt{\frac{1}{4n}\ln\frac{1}{\epsilon}}\Big)\,,$$

where in the last step we applied (15) and (14).

The right-hand bound in (11) follows from applying Theorem 1 to the following generic observation:

$$R^*_{\rm iid}(2nL,\epsilon) \le 2\sqrt{R^*_{\rm iid}(n)}\,, \qquad \forall L \ge 8 \ln\frac{2}{\epsilon}\,. \tag{34}$$

This follows from the standard "median trick": Consider a sample of size $n_1 = 2nL$ and denote by $\hat{T}_1, \ldots, \hat{T}_{2L}$ the result of evaluating best quadratic-risk estimator on $2L$ independent subsamples, each of size n. Let $\rho = 2\sqrt{R^*_{\rm iid}(n)}$. Then from Chebyshev's inequality we have for each $1 \le j \le 2L$

$$p \triangleq \mathbb{P}[|T(\pi) - \hat{T}_j| \ge \rho] \le \frac{1}{4}\,.$$

Define the estimator $\hat{T}$ to be the median of $(\hat{T}_1, \ldots, \hat{T}_{2L})$. Then from Hoeffding's inequality we have

$$\mathbb{P}[|T(\pi) - \hat{T}| > \rho] \le \sum_{k \ge L} \binom{2L}{k} p^k (1-p)^{2L-k} \le e^{-L/8}\,.$$

From the last statement, we conclude that (34) must hold. $\square$

2.1 Comparison to Donoho-Liu [DL91]

Theorem 1 is very similar to a celebrated result of Donoho-Liu [DL91], who showed that in the same setting, as $n \to \infty$, one has

$$C_0 \delta_{H^2}(\tfrac{1}{\sqrt{n}})^2 \le R^*_{\rm iid}(n) \le C_1 \delta_{H^2}(\tfrac{1}{\sqrt{n}})^2\,, \tag{35}$$

for some constants C_0, C_1, i.e. that the minimax rate for estimating the linear functionals T coincides with *modulus of continuity* of T with respect to Hellinger distance. In view of (15), $\delta_{H^2} \asymp \delta_{\chi^2}$ and thus (35) seems like exactly what Theorem 1 claims.

The differences, however, are three-fold. First, the technical assumptions required in [DL91] are: A1, A2 (from Theorem 1), boundedness $\sup_{\pi \in \Pi} |T(\pi)| < \infty$ and *Hölderianity* of δ_{H^2}:

$$\delta_{H^2}(t) = Ct^r + o(t^r)$$

for some $C, r > 0$ as $t \to 0$. Barring the latter, the assumptions are weaker than in Theorem 1.

The second, and crucial, difference is the fact that (35) only holds for a fixed statistical problem $(\Theta, \mathcal{X}, P, T, \Pi)$ and as $n \to \infty$, i.e. the proportionality constants in (35) are not uniform and can be *problem dependent*. This precludes one to analyze questions where the problem size (e.g. dimension) varies with the sample size n, etc. For example, in the population recovery problem considered in Section 2.2 for any fixed d and $n \to \infty$ we get parametric rate $R^*_{\rm iid}(n) \asymp \frac{1}{n}$. To get interesting phase-transitions one needs to let d slowly grow – and this cannot be handled in the setup of [DL91] where the problem is first fixed and then analyzed in the large-sample asymptotics of $n \to \infty$.

The third difference is the method of proof. While we (indirectly, via duality) show the existence of a good linear estimator, Donoho and Liu construct an estimator via binary search, which entails decomposing the problem into a dyadic sequence of testing problems between two composite hypotheses of the form $\{\pi : T(\pi) < a\}$ vs $\{\pi : T(\pi) > b\}$. The advantage of their method is that it can handle loss functions other than the quadratic loss. The advantage of our method is that our estimator is simply an empirical average of a certain function, that, in discrete cases, can be efficiently pre-computed by convex or linear programming. Furthermore, even for continuous models, the infinite-dimensional LP can be effectively "finite-dimensionalized" leading to computational efficient construction of optimal estimators (see Theorems 10 and 11 for examples).

To sum up, in the iid setting, the chief advantage of our method is in explicit universal constants comparing $R^*_{\text{iid}}(n)$ and $\delta_{\chi^2}(\frac{1}{\sqrt{n}})$. However, perhaps, the main advantage is (as we show in Section 2) that our methods extend to the deterministic setting, when one's goal is to estimate a functional of high-dimensional parameters.

2.2 Application: Population recovery

In the problem of lossy population recovery [DRWY12,WY12], let μ denote an unknown distribution on the d-dimensional Hamming space $\{0,1\}^d$. For n iid random binary strings $A_1,\dots,A_n$ drawn from μ, we observe their erased version $B_1,\dots,B_n \in \{0,1,?\}^d$., where each bit is erased with probability ϵ, and the goal is to estimate the distribution μ. It has been shown in [DRWY12] (cf. [PSW17, Appendix A]) estimating the entire distribution μ in the sup norm can be reduced to estimating the weight of the all-zero string $\mu(\mathbf{0})$ in terms of both sample and time complexity. Furthermore, for large d it is shown in [PSW17] that summarizing each string into its number of 1's is "almost sufficient" (in the sense that it changes the minimax rate by no more than logarithmic factors), as the number of 0's provides negligible information for estimating $\mu(\mathbf{0})$.

After these reductions, we arrive at a specialization of the setup in Theorem 1. Let θ_i denote the number of 1's in the unerased string A_i. Then $\theta_i \overset{\text{iid}}{\sim} \pi$ for some distribution π on $\mathcal{X} = \Theta = \{0,\dots,d\}$, where $T(\pi) \triangleq \pi(0) = \mu(\mathbf{0})$ is the quantity to be estimated on the basis of $X_1,\dots,X_n$, where X_i denotes the number of 1's in the erased string B_i. Note that conditioned on θ, we have

$$X_i \overset{\text{iid}}{\sim} P_\theta = \text{Binom}(\theta, 1-\epsilon)\,. \tag{36}$$

The minimax risk of population recovery, denoted by $R^*(n,d)$, has been characterized within logarithmic factors in [PSW17]. Next we deduce this result from the general Theorem 1, which boils down to characterizing the corresponding χ^2-modulus of continuity $\delta_{\chi^2}(t) = \delta_{\chi^2}(t,d)$. The following result can be distilled from [PSW17] (a proof is given in Appendix B for completeness):

Lemma 6. *For any $t \ge 0, d \ge 1$ we have*

$$\delta_{\chi^2}(t,d) \le t^{\min(1,\frac{1-\epsilon}{\epsilon})}\,. \tag{37}$$

Conversely, for $\epsilon \le \frac{1}{2}$, $\delta_{\chi^2}(t,d) \ge \frac{\min(t,1)}{4}$; for $\epsilon > 1/2$ there exists $t_0 = t_0(\epsilon)$ and $C = C(\epsilon)$ such that

$$\delta_{\chi^2}(t,d) \ge C\left(\frac{t}{\ln\frac{1}{t}}\right)^{\frac{1-\epsilon}{\epsilon}}, \tag{38}$$

provided that $t \le t_0$ and $d \ge C\ln^2\frac{1}{t}$.

Applying the general Theorem 1 together with Lemma 6, we obtain the following characterization of the minimax risks, where the rate of convergence exhibits an elbow effect at erasure probability $\epsilon = \frac{1}{2}$:

Corollary 7 ([PSW17]).

- *If $\epsilon \in (0, \frac{1}{2}]$, then for any $d \geq 1$,*

$$\frac{1}{8(1+\sqrt{e})^2 n} \leq R^*(n,d) \leq \frac{1}{n}.$$

- *If $\epsilon \in (\frac{1}{2}, 1)$, then there exists a constant $C = C(\epsilon) > 0$ such that we have*

$$\frac{1}{C}(n \log^2 n)^{-\frac{1-\epsilon}{\epsilon}} \leq R^*(n,d) \leq n^{-\frac{1-\epsilon}{\epsilon}},$$

where the lower bound holds provided that $d \geq Cn \log^4 n$.

2.3 Application: Interval censoring

Consider the following setup, commonly occurring in biostatistics. Subjects are arriving according to a Poisson process starting from time t_0. Upon arrival a treatment is administered to every subject. Following the treatment after a random delay $\theta_i \overset{\text{iid}}{\sim} \pi$ the ith subject experiences an event. Statistician terminates the experiment at time t_1 and counts all subjects for which the event has occurred by time t_1. The goal is to estimate the CDF or the median of π. (Note that statistician does not observe the time of event, only whether it happened or not, which makes it different from the right-censoring model of Kaplan-Meier [KM58]. It can be seen that conditioned on the number n of total subjects arrived between t_0 and t_1 the random time $A_{(i)}$ passed between the ith subject's arrival and experiment termination t_1 is simply an i-th order statistic of the iid uniform sample $A_i \overset{\text{iid}}{\sim} \text{Unif}[0, t_1 - t_0]$. This motivates the formal setting of "Case 1" below. See [HW97] for a survey and more details.

Fix $s_0 \in (0,1)$ and class of distributions Π on $\Theta = [0,1]$. For any distribution π we denote by F_π its CDF. We consider two functionals

$$T_c(\pi) \triangleq F_\pi(s_0) = \pi([0, s_0]), \qquad T_m(\pi) \triangleq \int_{[0,1]} \theta \pi(d\theta)\,.$$

For a given $\pi \in \Pi$ we generate $\theta_i \overset{\text{iid}}{\sim} \pi$, $i \in [n]$ and consider two cases of observations:

1. "Interval censoring, Case 1" (also known as "current status model"), see [GJ14, Section 2.3]. The observations are $X_1, \ldots, X_n$, where $X_i = (A_i, \Delta_i)$, given by $A_i \overset{\text{iid}}{\sim} G_1$ and $\Delta_i = 1\{\theta_i \leq A_i\}$. That is, we only see whether θ_i has occurred before an independent A_i. Here G_1 is a fixed (known) distribution on $[0,1]$. We denote the corresponding Markov kernel acting from $[0,1]$ to $\mathcal{X} = [0,1] \times \{0,1\}$ by $P^{(1)} = \{P_\theta^{(1)}(\cdot) : \theta \in [0,1]\}$.

2. "Interval censoring, Case 2", see [GJ14, Section 4.7]. This time observations are given by $X_i = (A_i, B_i, \Delta, \tilde{\Delta})$ with $(A_i, B_i) \overset{\text{iid}}{\sim} G_2$, where G_2 is some fixed distribution on $[0,1]^2$, and $\Delta_i = 1\{\theta_i \leq A_i\}$, $\tilde{\Delta}_i = 1\{\theta_i \leq B_i\}$. The Markov kernel for this case is denoted by $P^{(2)}$.

We denote the minimax risks for estimating $T_j, j \in \{c, m\}$ in case i, $i \in \{1,2\}$ by $R_n^{(i,j)}(\Pi)$. Under appropriate conditions on G_1, G_2 and Π, the following was shown in a series of works, cf. [GL95, Section 5] and [GW92, Example 3.1]:

$$R_n^{(1,m)}(\Pi) \asymp R_n^{(2,m)} \asymp \frac{1}{n}\,, \tag{39}$$

$$R_n^{(1,c)}(\Pi) \asymp \frac{1}{n^{2/3}}\,, \tag{40}$$

$$R_n^{(2,c)}(\Pi) \asymp \frac{1}{(n \log n)^{2/3}}\,. \tag{41}$$

We denote the χ^2-moduli of continuity for functionals T_c and T_m under the two models as follows:

$$\delta_{\chi^2}^{(i,j)}(t;\Pi) \triangleq \sup\{T_j(\pi) - T_j(\pi') : \pi, \pi' \in \Pi, \chi^2(\pi P^{(i)} \| \pi' P^{(i)}) \le t^2\}, \quad i = 1, 2, \quad j = c, m.$$

The next result shows that the minimax rates are determined by these moduli of continuity. The proof is given in Appendix B, which amounts to verifying the assumptions of Theorem 1.

Proposition 8. *Suppose that the set Π is convex and weakly closed. For the case of estimating T_c, in addition, we assume that s_0 is a point of continuity of $F_\pi(\cdot)$ for every $\pi \in \Pi$. Suppose that distributions G_1 and G_2 both have densities g_1 and g_2.*[5] *Then we have for all $i \in \{1, 2\}, j \in \{c, m\}$*

$$R_n^{(i,j)} \asymp \delta_{\chi^2}^{(i,j)}\left(\frac{1}{\sqrt{n}}; \Pi\right)^2$$

Furthermore, in each case the upper bound is attained by an estimator of the form $\sum_{k=1}^n \phi(X_k)$ for some continuous function ϕ on $\mathcal{X}$.

The above general characterization of the minimax risk by Proposition 8 can be converted into explicit rate of convergence. We give two examples:

1. Suppose that the distribution G_1 has density g_1 such that $g_1(a) \ge \epsilon_1 > 0$ for all $a \in (0, 1)$. Then we have for any weakly closed and convex Π:

$$R_n^{(1,m)} \asymp \delta_{\chi^2}^{(1,m)}\left(\frac{1}{\sqrt{n}}; \Pi\right)^2 \asymp \frac{1}{n}.$$

   Indeed, consider any two distribution π_1 and π_2 with corresponding CDFs given by F_1 and F_2. A simple calculation shows

$$\chi^2(\pi P^{(1)} \| \pi' P^{(1)}) = \int_0^1 g_1(a) \frac{(F_1(a) - F_2(a))^2}{F_2(a)(1 - F_2(a))}. \tag{42}$$

   Upper-bounding the denominator by $\frac{1}{4}$ and lower-bounding g_1 by ϵ_0 we obtain

$$\begin{aligned} \chi^2(\pi P^{(i)} \| \pi' P^{(i)}) &\ge 4\epsilon_0 \int_0^1 da (F_1(a) - F_2(a))^2 \qquad (43) \\ &\ge 4\epsilon_0 \left(\int_0^1 da (F_1(a) - F_2(a))\right)^2, \qquad (44) \end{aligned}$$

   where the last step is via Jensen's inequality. Since the integral equals $T_m(\pi_2) - T_m(\pi_1)$ we get that

$$\delta_{\chi^2}^{(1,m)}(t) \le \frac{t}{2\sqrt{\epsilon_0}}.$$

   Note also that from (16) we conclude that $\delta_{\chi^2}^{(1,m)}(t) \asymp t$. This recovers (39) under most general conditions. (Note that if density g_1 is zero on some interval $[a, b]$ then the model becomes unidentifiable.)

[5] For estimating T_c we could also demand only the existence of density in small interval around s_0 or (s_0, s_0).

2. Now suppose that G_1 has density g_1 that is continuous and positive at s_0. Consider the class $\Pi(\gamma,\epsilon) \triangleq \{\pi : F_\pi(s) \text{ is } \gamma\text{-Lipschitz for } s \in (s_1, s_2)\}$, where $s_1 = s_0 - \epsilon, s_2 = s_0 + \epsilon$. Then we claim
$$\delta^{(1,c)}_{\chi^2}(t) \asymp t^{1/3}\,. \tag{45}$$
Without loss of generality, we will assume that $\epsilon_0 \le g_1(s) \le \frac{1}{\epsilon_0}$ for all $s \in (s_1, s_2)$. (Otherwise, we simply reduce ϵ.) Notice that Π is indeed convex and weakly closed. For the lower bound consider any pair of CDFs F_1, F_2 such that (a) they both belong to Π, (b) $1/4 < F_2(s_1) < F_2(s_2) < 3/4$, and (c)
$$F_1(s) - F_2(s) = \begin{cases} 0, & s < s_3 \text{ or } s > s_4 \\ s_5 + \frac{\gamma}{2}|s - s_0|, & s_3 < s < s_4 \end{cases},$$
where $s_3 = s_0 - \tau_1, s_4 = s_0 + \tau_1$ and $s_5 = -\gamma\tau_1/2$. From (42) we obtain:
$$\chi^2(\pi_1 P^{(1)} \| \pi_2 P^{(1)}) = \int_{s_3}^{s_4} ds g_1(s) \frac{(F_1(s) - F_2(s))^2}{F_2(s)(1 - F_2(s))} \lesssim \tau_1^3\,.$$
Thus, this demonstrates $\delta_{\chi^2}(\tau_1^3) \ge \tau_1$, as claimed by (45).
For the upper bound, arguing as in (44) we get for any $\tau_1 < \epsilon$ that
$$t^2 \ge \chi^2(\pi_1 P^{(1)} \| \pi_2 P^{(1)}) \gtrsim \int_{s_0-\tau_1}^{s_0+\tau_1} |F_1(s_0 + x) - F_2(s_0 + x)|^2 ds\,.$$
Now, if we set $\delta = F_1(s_0) - F_2(s_0) > 0$ then from Lipschitzness we get that for $\tau_1 = \frac{\delta}{2\gamma}$ we have that $|F_1(s_0 + x) - F_2(s_0 + x)| \ge \frac{\delta}{2}$ and thus
$$t^2 \gtrsim \delta^3\,,$$
implying that $\delta_{\chi^2}(t) \lesssim t^{1/3}$, finishing the proof of (45).

These applications hopefully demonstrate the utility of Theorem 1. Indeed, the minimax rates (39) and (40) were obtained with a lot less effort compared to the existing literature [GJ14].[6] In particular, the previous upper bounds were derived by a lengthy analysis of the nonparametric maximum likelihood [GW92], or certain ad hoc histogram estimator [GL95]. On the other hand, establishing (41) by computing δ_{χ^2} is more involved and will be presented elsewhere.

3 Deterministic setting

In this section we consider the deterministic setting as described in Section 1. Namely, the observations are $\boldsymbol{X} = (X_1, \ldots, X_n)$, where $X_i \overset{\text{ind.}}{\sim} P_{\theta_i}$. Here the unknown parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is deterministic and belongs to the following constraint set
$$\boldsymbol{\Theta}_c = \left\{ \boldsymbol{\theta} \in \Theta^{\otimes n} : \frac{1}{n}\sum_{i=1}^n c(\theta_i) \le 1 \right\},$$

[6] Results in [GJ14] are stated with the extra assumption on the lower bound on the density of π at s_0, but this constraint seems not necessary for establishing rates.

for some cost function $c : \Theta \to \mathbb{R}$. Equivalently, $\boldsymbol{\Theta}_c$ consists of those $\boldsymbol{\theta}$ whose empirical distribution $\pi_{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$ belongs to the convex set

$$\Pi = \left\{ \pi \in \mathcal{P}(\Theta) : \int c(\theta)\pi(d\theta) \le 1 \right\}. \tag{46}$$

Let $h : \Theta \to \mathbb{R}$ and define the following affine functional

$$T(\pi) = \int h(\theta)\pi(d\theta). \tag{47}$$

Given $\boldsymbol{X} = (X_1, \ldots, X_n)$, the goal is to estimate $T(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1} h(\theta_i)$, which is a symmetric separable function of the parameter $\boldsymbol{\theta}$. The minimax quadratic risk $R^*_{\mathrm{det}}(n)$ is defined in (3), namely,

$$R^*_{\mathrm{det}}(n) = \inf_{\hat{T}} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_c} \mathbb{E}_{\boldsymbol{\theta}}[|\hat{T}(\boldsymbol{X}) - T(\pi_{\boldsymbol{\theta}})|^2] \tag{48}$$

Many problems studied in the high-dimensional functional estimation literature are of or can be reduced to questions of the above type. For example, in the Gaussian model where $X_i \sim N(\theta_i, 1)$, estimation of linear ($h(\theta) = \theta$) and quadratic functional ($h(\theta) = \theta^2$) has been well-studied and more recently under sparsity assumptions which correspond to adding further constraints with $c(\theta) = \mathbf{1}_{\{|\theta|>0\}}$ or $c(\theta) = |\theta|^q$ [CCTV16, CCT17]. Estimation of non-smooth functional such as the ℓ_1-norm ($h(\theta) = |\theta|$) has been studied in [LNS99, CL11].

The main idea of this section is that the minimax problem in the deterministic setting is similar to the iid setting studied in Section 2, where, instead of adversarially selecting a vector $\boldsymbol{\theta}$ from $\boldsymbol{\Theta}_c$, one generates each coordinate θ_i independently from some prior $\pi \in \Pi$ such that $\mathbb{E}_{\theta\sim\pi}[c(\theta)] \le 1$. By concentration, we expect the constraint $\frac{1}{n}\sum_{i=1}^n c(\theta_i) \le 1$ to be fulfilled approximately and indeed this product prior can be made valid with appropriate truncation. Furthermore, we expect $T(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n h(\theta_i)$ to be concentrated near its mean $T(\pi) = \int h(\theta)\pi(d\theta)$, which is a linear functional of π. Estimating the latter falls under the purview of Section 2 and hence its minimax rate is given by $\delta_{\chi^2}(\frac{1}{\sqrt{n}})$. Thus, it seems natural to expect that

$$R^*_{\mathrm{det}}(n) \asymp \delta_{\chi^2}\left(\frac{1}{\sqrt{n}}\right)^2 \tag{49}$$

up to universal constants. Alas, such statement cannot hold without conditions, as the next example demonstrates. However, the good news is that such counterexamples only occur in the "uninteresting" case of $R^*_{\mathrm{det}}(n) = 0$ or $R^*_{\mathrm{det}}(n) \asymp \frac{1}{n}$ (parametric rate).

Example 1. Let $\Theta = \mathcal{X} = \{0, 1\}$, $c(\theta) = 0$, so that $\Pi = \{\mathrm{Bern}(p) : 0 \le p \le 1\}$. Let $h(\theta) = \theta$ and consider the observation model $\mathbb{P}[X = \theta] = 1 - \mathbb{P}[X = 1 - \theta] = \tau$ (the binary symmetric channel). Note that the smallest value of δ_{χ^2} will occur at $\tau = 0$, and even then from $\chi^2(\mathrm{Bern}(p)\|\mathrm{Bern}(1/2)) = (1-2p)^2$ we obtain: $\delta_{\chi^2}(t) \ge t/2$ for $t \le 1$. At the same time, a simple unbiased estimator $\hat{T}(X_1, \ldots, X_n) = \frac{1}{n(1-2\tau)}\sum_{i=1}^n (1\{X_i = 1\} - \tau)$ achieves

$$R^*_{\mathrm{det}}(n) \le \frac{\tau(1-\tau)}{(1-2\tau)^2}\frac{1}{n}.$$

One immediate conclusion is that at $\tau = 0$ we have $R^*_{\mathrm{det}}(n) = 0$ while $\delta_{\chi^2}(t) > 0$ for all $t > 0$. Furthermore, even when $\tau > 0$ and $R^*_{\mathrm{det}}(n) \asymp \delta_{\chi^2}(1/\sqrt{n})^2 \asymp \frac{1}{n}$, the proportionality constant in the first relation is not uniform in τ, as $\lim_{\tau\to 0} \frac{R^*_{\mathrm{det}}(n)}{\delta_{\chi^2}(1/\sqrt{n})^2} = 0$. Therefore, we cannot expect the relation (49) to hold with universal (problem-independent) constants.

Remark 1 (Parametric lower bound)**.** Consider the setting where the constraint function c and the function h are both fixed and the sample size n grows. There is a general *dichotomy*: either risk $R^*_{\rm det}(n) = 0$ or $R^*_{\rm det}(n) = \Omega(\frac{1}{\sqrt{n}})$. Indeed, either there exists a pair $\theta_a, \theta_b \in \Theta$ s.t. $h(\theta_a) \neq h(\theta_b)$ and $\mathrm{TV}(P_{\theta_a}, P_{\theta_b}) < 1$, or there is no such pair. In the latter case, we have $h(\theta) = g(X_1)$ (i.e. $h(\theta)$ is a deterministic function of a single sample), and thus $R^*_{\rm det}(n) = 0$ for any $n \geq 1$. In the former case, we can lower bound $R^*_{\rm det}(n)$ by the Bayes risk when $\boldsymbol{\theta}$ has iid components with $\mathbb{P}[\theta_i = \theta_a] = \mathbb{P}[\theta_i = \theta_b] = \frac{1}{2}$.[7] Clearly, the corresponding Bayesian risk is $\Omega(1/\sqrt{n})$.

The main result of this section is:

Theorem 9. *Suppose that $(\Theta, \mathcal{X}, P, T, \Pi)$, with Π and T given in (46) and (47) respectively, satisfy conditions A1-A4 of Theorem 1. Then*

$$R^*_{\rm det}(n) \leq \delta_{\chi^2}\left(\frac{1}{\sqrt{n}}\right)^2, \tag{50}$$

achieved by an estimator of the form $\hat{T} = \frac{1}{n}\sum_{i=1}^n g(X_i)$ for some $g \in \mathcal{F}$. Furthermore, suppose the following extra conditions are satisfied

A5 $K_V = \sup_{\pi\in\Pi} \mathrm{Var}_{\theta\sim\pi}[T(\theta)] < \infty$;

A6 Cost function $c \geq 0$ and there exists $\theta_0 \in \Theta$ with $c(\theta_0) = 0$.

Then

$$R^*_{\rm det}(n) \geq \frac{1}{2400}\delta_{\chi^2}\left(\frac{1}{\sqrt{n}}\right)^2 - \frac{K_V}{2n}. \tag{51}$$

Proof. Recall that $T(\pi) = \int h(\theta)\pi(d\theta)$ and $T(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n h(\theta_i)$. To prove (50), consider an estimator $\hat{T}_g$ of the form (22) and, similarly to (23), let us analyze its risk by decomposing into bias and variance parts:

$$\sqrt{\mathbb{E}_{\boldsymbol{\theta}}[|\hat{T}_g(\boldsymbol{X}) - T(\pi_{\boldsymbol{\theta}})|^2]} \leq \frac{1}{n}\sqrt{\sum_{i=1}^n \mathrm{Var}_{P_{\theta_i}}[g]} + \left|\frac{1}{n}\sum_{i=1}^n (P_{\theta_i}g - h(\theta_i))\right| \tag{52}$$

Recall that the empirical distribution of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is denoted by $\pi_{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$, so that $\pi_{\boldsymbol{\theta}}P = \frac{1}{n}\sum_{i=1}^n P_{\theta_i}$. By the concavity of $\mu \mapsto \mathrm{Var}_\mu[g]$, upper-bounding

$$\sum_{i=1}^n \mathrm{Var}_{P_{\theta_i}}[g] \leq n \cdot \mathrm{Var}_{\pi_{\boldsymbol{\theta}}P}[g],$$

we continue (52) to get

$$\sqrt{R^*(n)} \leq \inf_g \sup_{\boldsymbol{\theta}} \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\pi_{\boldsymbol{\theta}}P}[g]} + |T(\pi_{\boldsymbol{\theta}}) - \pi_{\boldsymbol{\theta}}Pg|, \tag{53}$$

where the supremum is taken over all $\boldsymbol{\theta}$ whose empirical measure $\pi_{\boldsymbol{\theta}}$ belongs to Π. Thus we can extend the inner supremum to $\hat{\pi}$ ranging over all of Π, concluding

$$\sqrt{R^*(n)} \leq \delta_{\rm bv}(\tfrac{1}{\sqrt{n}})$$

[7] This prior needs to be modified if $c(\theta_a) > 1$ or $c(\theta_b) > 1$. Specifically, choose an arbitrary θ_0 such that $c(\theta_0) < 1$. Then we can choose $\boldsymbol{\theta}$ iid from $\pi = (1-\epsilon)\delta_{\theta_0} + \frac{\epsilon}{2}(\delta_{\theta_a} + \delta_{\theta_b})$ for sufficiently small constant ϵ.

with $\delta_{\rm bv}$ defined in (24). Applying Proposition 4 we get (50).

To prove (51), fix $\gamma \in (0,1)$ (to be specified later) and consider $\pi_0, \pi_0' \in \Pi$ such that $\chi^2(\pi_0' P \| \pi_0 P) \le \frac{1}{n}$ and $T(\pi_0') - T(\pi_0) = \delta$. Next define distributions

$$\pi_1 = \gamma \pi_0 + (1-\gamma)\delta_{\theta_0}, \quad \pi_1' = \gamma \pi_0' + (1-\gamma)\delta_{\theta_0}\,,$$

where θ_0 is from Assumption A6 such that $c(\theta_0) = 0$. From the convexity of $\chi^2(\cdot\|\cdot)$, we get

$$\chi^2(\pi_1' P \| \pi_1 P) \le \frac{\gamma}{d}, \qquad T(\pi_1') - T(\pi_1) = \gamma\delta\,.$$

Denote $\mu' = T(\pi_1'), \mu = T(\pi_1)$. Define distributions $\nu = \pi_1^{\otimes n}, \nu' = \pi_1'^{\otimes n}$ and note that $(\pi_1 P)^{\otimes n} = \nu P^{\otimes n}$. Then

$$\begin{aligned}
\mathrm{TV}(\nu P^{\otimes n}, \nu' P^{\otimes n}) = \mathrm{TV}((\pi_1 P)^{\otimes n}, (\pi_1' P)^{\otimes n})) &\overset{(a)}{\le} \frac{1}{2}\sqrt{\chi^2((\pi_1 P)^{\otimes n} \| (\pi_1' P)^{\otimes n})} \\
&\overset{(b)}{=} \frac{1}{2}\sqrt{(1+\chi^2(\pi_1 P\|\pi_1' P))^n - 1} \\
&\overset{(c)}{\le} \frac{1}{2}\sqrt{(1+\gamma/n)^n - 1} \le \frac{1}{2}\sqrt{e^\gamma - 1},
\end{aligned}$$

where (a) follows from the fact that $\mathrm{TV} \le \frac{1}{2}\sqrt{\chi^2}$ [GS02, Section 3]; (b) is from the tensorization identity in (20); (c) is from the convexity $\chi^2(\pi_1 P\|\pi_1' P) \le \gamma\chi^2(\pi_0 P\|\pi_0' P)$.

Next define sets $A, A' \subset \mathbf{\Theta}_c$:

$$A = \left\{\boldsymbol{\theta} \in \Theta^{\otimes n} : \frac{1}{n}\sum_{i=1} c(\theta_i) \le 1, T(\pi_{\boldsymbol{\theta}}) \le \mu + \frac{\gamma\delta}{3}\right\} \tag{54}$$

$$A' = \left\{\boldsymbol{\theta} \in \Theta^{\otimes n} : \frac{1}{n}\sum_{i=1} c(\theta_i) \le 1, T(\pi_{\boldsymbol{\theta}}) \ge \mu' - \frac{\gamma\delta}{3}\right\}\,. \tag{55}$$

From the Chebyshev and Markov inequalities we have

$$\nu(A^c), \nu'(A'^c) \le \gamma + \frac{9K_V}{n\gamma^2\delta^2}$$

Next, decompose distributions ν, ν' as convex combinations:

$$\nu = \nu(A)\nu_{|A} + \nu(A^c)\nu_{|A^c}\,, \nu' = \nu'(A')\nu'_{|A'} + \nu'(A'^c)\nu'_{|A'^c}\,,$$

where $\nu_{|B}(\cdot) \triangleq \nu(\cdot \cap B)/\nu(B)$ is the conditional version of the distribution ν.

By the triangle inequality and the data processing inequality of total variation, we get

$$\mathrm{TV}(\nu_{|A} P^{\otimes n}, \nu'_{|A'} P^{\otimes n}) \le \nu(A^c) + \nu'(A^c) + \mathrm{TV}(\nu P^{\otimes n}, \nu' P^{\otimes n})\,.$$

Altogether, we have a pair of distributions $\nu_1 \triangleq \nu_{|A}$ and $\nu_1' \triangleq \nu'_{|A'}$ both supported on $\mathbf{\Theta}_c$ such that $T(\pi_{\boldsymbol{\theta}}) \le \mu - \frac{\gamma\delta}{3}$ for ν_1-a.e. $\boldsymbol{\theta}$ and $T(\pi_{\boldsymbol{\theta}}) \ge \mu + \frac{\gamma\delta}{3}$ for ν_1'-a.e. $\boldsymbol{\theta}$. Applying the TV version of Le Cam's method for quadratic risk (see [Yu97, Lemma 1]) yields the following minimax lower bound:

$$R^*(n) \ge \frac{1}{4}\left(\frac{\gamma\delta}{3}\right)^2 (1-t)\,,$$

where $t \triangleq 2\gamma + \frac{18K_V}{n\gamma^2\delta^2} + \sqrt{e^\gamma - 1}/2$. Choosing $\gamma \in (0,1)$ to maximize the function $\gamma^2 - 2\gamma^3 - \sqrt{e^\gamma - 1}/2$, we obtain

$$R^*(n) \ge \frac{1}{2400}\delta^2 - \frac{K_V}{2n}\,.$$

Optimizing over the choice of π_0, π_0' thus yields (51). $\square$

Remark 2. Before presenting new results obtained from Theorem 9, as a quick application, consider the problem of estimating the ℓ_1-norm of a vector in the Gaussian location model [LNS99, CL11], where $X_i \sim N(\theta_i, 1)$, $h(\theta) = |\theta|$ and $T(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\|\boldsymbol{\theta}\|_1$, and $\Theta = [-1, 1]$. Using the method of polynomial approximation and moment matching, it was shown in [CL11] that $R^*_{\text{det}}(n) = \Theta((\frac{\log\log n}{\log n})^2)$. (In fact, the sharp constant as $n \to \infty$ was also found). To see how this result follows from Theorem 9, note that $K_V = 1$, we have $c\delta^2_{\chi^2}(\frac{1}{\sqrt{n}}) - \frac{1}{4n} \le R^*_{\text{det}}(n) \le \delta^2_{\chi^2}(\frac{1}{\sqrt{n}})$ for constant c, where

$$\delta_{\chi^2}(t) = \sup\left\{\int |\theta|\pi(d\theta) - \int |\theta|\pi'(d\theta) : \chi^2(\pi' * N(0,1)\|\pi * N(0,1)) \le t^2\right\}. \tag{56}$$

Here $*$ denotes convolution, and the supremum is taken over $\pi, \pi' \in \mathcal{P}([-1,1])$. The speed of convergence of $\delta_{\chi^2}(t)$ when $t \to 0$ is extremely slow and thus its behavior governs the minimax rate. Indeed, one can show that (see Appendix B)

$$\delta_{\chi^2}(t) = \Theta\left(\frac{\log\log\frac{1}{t}}{\log\frac{1}{t}}\right), \tag{57}$$

recovering the result of [CL11].

However, if the parameter space is unbounded with $\Theta = \mathbb{R}$ we have $K_V = \infty$ and lower bound in Theorem 9 is not applicable. (In fact it is easy to see that $\delta_{\chi^2}(t) = \infty$ for any t.) Nevertheless, applying a truncation argument, it was shown in [CL11] that $R^*_{\text{det}}(n) \asymp \frac{1}{\log n}$.

3.1 Application: Distinct Elements problem

In the *distinct elements* problem, given a sample randomly drawn from an urn containing multiple colored balls, the goal is to estimate the total number of distinct colors in the urn. This problem has been thoroughly investigated in both statistics and computer science under various formulations and sampling models. We refer the readers to the comprehensive survey [BF93], [CCMN00, RRSS09, Val11, Val12, WY18] for more recent work, and [WY18, Table 1] for a summary of the state of the art. In this section, we consider the following version of the distinct elements problem, where the number of balls in the urn is at most n and unknown a priori. We shall work with the so-called *Bernoulli sampling model* with sampling ratio p, a specific version of sampling without replacement, where the color of each ball is observed independently with probability p; see [WY18, Appendix A] for connections and near equivalence to other sampling models.

Most of the recent theoretical results aim at the sublinear regime of $p = o(1)$. In particular, it is known that the optimal sample complexity for consistency (in normalized error) is $p = \Theta(\frac{1}{\log n})$. For the linear regime, say, 1% of the balls are observed, existing results do not yield tight characterization of the optimal estimation accuracy. Next, we will apply the general Theorem 9 to determine the minimax risk up to logarithmic factors in the linear regime, and reveal an elbow effect in the optimal rate of convergence that precisely occurs at sampling ratio $\frac{1}{2}$.

Without loss of generality, assume that the number of colors in the universe (not necessarily in the urn) is n, and indexed by $[n] = \{1, \dots, n\}$. Let $\theta_i \in \mathbb{Z}_+$ be the number of balls of the ith color, $i = 1, \dots, n$. Thus, the parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ is constrained to belong to the set

$$\boldsymbol{\Theta}_c = \left\{\boldsymbol{\theta} \in \mathbb{Z}^n_+ : \frac{1}{n}\sum_{i=1}^n \theta_i \le 1\right\}.$$

We shall work with the Bernoulli sampling model with sampling ratio p, where the color of each ball is observed independently with probability p. Denote by N_i the number of observed balls of the ith color. Then we have $X_i \overset{\text{ind.}}{\sim} \text{Binom}(\theta_i)$. Given $(N_1, \ldots, N_n)$, the goal is to estimate the (normalized) number of distinct colors:

$$T(\pi_{\boldsymbol{\theta}}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\theta_i \geq 1\}}.$$

This problem is exactly in the deterministic setting of Theorem 9 and the minimax quadratic risk $R^*(n) \equiv R^*_{\text{det}}(n)$ is defined as in (3).

The following theorem (proved in Section 5.1) determines the sharp minimax risk up to logarithmic factors in the linear sampling regime (p being a constant). Note that the upper bound is explicit and non-asymptotic, which allows us to recover the prior result on the optimal sampling complexity $\Omega(\frac{n}{\log n})$, i.e. $p = \Omega(\frac{1}{\log n})$, for consistent estimation.

Theorem 10. *Fix $p \in (0,1)$. There exists a constant $c = c(p) > 0$ such that*

- *if $p \geq \frac{1}{2}$, then*
$$\frac{c}{n} \leq R^*(n) \leq \frac{1}{n} \tag{58}$$
- *if $p < \frac{1}{2}$, then*
$$\frac{c}{\log^2 n} n^{-\frac{p}{1-p}} \leq R^*(n) \leq n^{-\frac{p}{1-p}}, \tag{59}$$

where the upper bound holds for all n and the the lower bound holds for all $n \geq n_0 = n_0(p)$. Furthermore, this upper bound can be achieved within a constant factor by an estimator of the form

$$\hat{T} = \sum_{i \in [n]} g(N_i) \tag{60}$$

with $g(0) = 0$, where the coefficient g can be found by solving an LP of $O(n)$ variables and $O(n)$ constraints.

Remark 3 (Linear estimator)**.** Estimators of the form (60) are commonly known as *linear estimators*, since they can be equivalently expressed as linear combinations of *profiles* (also known as fingerprints) [OSW16, VV11]:

$$\hat{T} = \sum_{j \geq 0} g(j) \Phi_j,$$

where

$$\Phi_j \triangleq \sum_{i} \mathbf{1}_{\{N_i = j\}}, \tag{61}$$

called the jth profile, denotes the number of colors that occurred exactly j times in the sample. Since Theorem 10 guarantees we can choose $g(0) = 0$ in (60), the resulting estimator $\hat{T} = \sum_{j \geq 1} g(j) \Phi_j$ is fully data-driven and oblivious to the total number of possible colors, a desirable property in practice.

3.2 Application: Fisher's species problem

Dating back to Fisher [FCW43], *predicting the unseen species* is a classical question in statistics, where given a sample of n iid observations $X_1, \ldots, X_n$ drawn from an unknown probability discrete distribution $P = (p_x)$ on some countable alphabet $\mathcal{X}$, the goal is to estimate the number of hitherto unobserved symbols that would be observed if a new sample of $X'_1, \ldots, X'_m$ were collected, i.e.,

$$U = U_{n,m} \triangleq |\{X'_1, \ldots, X'_m\} \backslash \{X_1, \ldots, X_n\}|.$$

In particular, the sequence $m \mapsto U_{n,m}$ is called the species discovery curve, which provides guidance on how many new species would be observed were m more data points to be collected. For this reason, extrapolating the species discovery curve is of significant interest in various fields such as ecology [FCW43, CL92], computational linguistics [ET76], genomics [ILLL09], etc. Clearly, the more future data we want to extrapolate, the more difficult it is to obtain a reliable prediction.

In order to frame the problem in the deterministic estimation setting, we consider the Poissonized version of the problem as studied in [FCW43, ET76, OSW16], where the sizes of the available and future (unobserved) samples are $N \sim \mathrm{Poi}(n)$ and $M \sim \mathrm{Poi}(m)$, respectively. Due to the concentration of the Poisson distribution, standard arguments (see Appendix C) show that the minimax risk bounds proved next apply to the model with fixed sample sizes with little change. Denote the histogram in the observed and unobserved sample by $N_x = \sum_{i\in[N]} \mathbf{1}_{\{X_i=x\}}$ and $N'_x = \sum_{i\in[M]} \mathbf{1}_{\{X'_i=x\}}$, respectively. Then $\{N_x\} \overset{\text{ind.}}{\sim} \mathrm{Poi}(np_x)$ and $\{N'_x\} \overset{\text{ind.}}{\sim} \mathrm{Poi}(mp_x)$ are independent of each. In terms of histograms, the number of unseen species can be expressed as

$$U = \sum_x \mathbf{1}_{\{N_x=0, N'_x>0\}}. \tag{62}$$

Let $r \triangleq \frac{m}{n}$ denote the extrapolation ratio. Denote the normalized minimax mean squared error of estimating U by

$$\mathcal{E}_n(r) \triangleq \inf_{\hat{U}} \sup_P \frac{1}{m^2} \mathbb{E}_P[(\hat{U} - U)^2],$$

where the expectation is with respect to both the original and the future samples. We emphasize that this problem is fully non-parametric and no assumptions are imposed on the distribution P.

It is known since Good and Toulmin [GT56] that an unbiased estimator for U is

$$\hat{U}_{\mathrm{GT}} = -\sum_x (-1)^{N_x} \mathbf{1}_{\{N_x>0\}} = \sum_{j\geq 1} -(-r)^j \Phi_j,$$

where Φ_j is the jth profile defined in (61). If $r \leq 1$, that is, we extrapolate no more than what have been observed, this unbiased estimator achieves the (optimal) parametric rate

$$\frac{1}{m^2} \mathbb{E}[(U - \hat{U}_{\mathrm{GT}})^2] \lesssim \frac{1}{n}. \tag{63}$$

However, for $r > 1$, the variance of $\hat{U}$ is unbounded due to the exponential growth of the coefficients. Based on a technique called *smoothing* that modifies the unbiased estimator to obtain a good bias-variance tradeoff, Orlitsky et al [OSW16] constructed a family of estimators that encompass previous heuristics of Efron and Thisted [ET76] and provably achieve the following prediction risk:

$$\mathcal{E}_n(r) \lesssim n^{-\log_3(1+\frac{2}{r})}. \tag{64}$$

Conversely, the following lower bound is also shown in [OSW16]:

$$\mathcal{E}_n(r) \gtrsim n^{-C/r}.$$

for some absolute constant C. Thus, one can extrapolate with a vanishing risk provided that $r = o(\log n)$, and this condition is the best possible. However, for fixed r, the optimal rate remains open. In particular, the above achievable results (63) and (64) seem to suggest an "elbow effect" in the optimal convergence rate, which transitions from parametric rate to nonparametric rate when the extrapolation ratio r exceeds 1. The following result resolves this question in the positive:

Theorem 11 (Optimal rate for predicting the unseen)**.** *Let $r > 0$ be a constant. There exist constants c_0, c_1 that depend only on r, such that the following holds.*

- *If $r \leq 1$, then*
$$\frac{c_0}{n} \leq \mathcal{E}_n(r) \leq \frac{c_1}{n}; \tag{65}$$
- *If $r > 1$, then*
$$\frac{c_0 n^{-\frac{2}{r+1}}}{\log^2 n} \leq \mathcal{E}_n(r) \leq c_1 n^{-\frac{2}{r+1}} \log^4 n. \tag{66}$$

Furthermore, an estimator achieving the upper bound can be constructed and evaluated in time $O(n^a)$ for some absolute constant a.

It is worth mentioning that, unlike Theorem 10, Theorem 11 does not directly follow from the general result in Theorem 9 because of the infinite-dimensional nature of the species problem (the number of distinct species is potentially unbounded), which requires extra reduction argument. Furthermore, analyzing the behavior of the modulus of continuity (as a linear program) relies on delicate complex analysis, in particular, Hadamard's three-lines theorem and the Paley-Wiener theorem. The proof of Theorem 11 is provided in Section 5.2.

Remark 4 (Species versus distinct elements problem)**.** There is an obvious connection between the species problem considered here and the distinct elements problem considered in Section 3.1: Treating the union of observed and unobserved samples $\{X_1, \ldots, X_n, X'_1, \ldots, X'_m\}$ as the content of an urn, the former can be viewed as a special case of the latter with the urn size being $n+m$ and the fraction of observation being $p = \frac{n}{m+n} = \frac{1}{1+r}$. Thus, for the interesting case of $r > 1$, applying Theorem 10 yields the upper bound $\mathcal{E}_n(r) \leq O(n^{-\frac{1}{r}})$. Perhaps surprisingly, this strategy turns out to be suboptimal in view of Theorem 11. This suggests that the optimal estimator for the species problem is able to exploit the special structure in the color configuration arising from iid sampling.

4 Exponential families

Let us revisit the setting of Theorem 1 in the special case when Θ and $\mathcal{X}$ are both finite. Given a closed convex set $\Pi \subset \Theta$ we define

$$M_0 \triangleq \{\mu : \mu = \pi P, \pi \in \Pi\},$$

which again is closed and convex. Any (identifiable) linear functional $T(\pi)$ can in turn be represented as a linear functional of μ. Hence, the statistical problem at hand becomes: Given $X_i \overset{\text{iid}}{\sim} \mu \in M_0$ estimate $T(\mu) \triangleq \sum_{x \in \mathcal{X}} \mu(x) h(x) = \langle \mu, h \rangle$, where $h : \mathcal{X} \to \mathbb{R}$ is a fixed function.

Let us restate the problem in the language of exponential families. Without loss of generality, let $\mathcal{X} = [d] \triangleq \{1, \ldots, d\}$. Then for $j \in [d]$ let $\phi_j(x) = \mathbf{1}\{x = j\}$, and denote $\phi(x) = (\phi_1(x) \ldots, \phi_m(x))$. For every $\gamma = (\gamma_1, \ldots, \gamma_d) \in \mathbb{R}^d$ we define a distribution on $\mathcal{X}$

$$P_\gamma(x) = \exp\{\langle \gamma, \phi(x) \rangle - C(\gamma)\},$$

where $C(\gamma)$ is chosen from normalization; more explicitly, $P_\gamma(x) = \frac{\exp(\gamma_x)}{\sum_{x \in \mathcal{X}} \exp(\gamma_x)}$. We can see that P_γ forms an exponential family with *natural parameters* γ and *mean parameters* $\mu_f(\gamma) = \mathbb{E}_{X \sim P_\gamma}[\phi(X)]$. Theorem 1 then shows that there exists g such that the empirical-mean estimator

$$\hat{T}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^n g(X_i) \tag{67}$$

that achieves the minimax rate for estimating $T(P_\gamma) = \langle \mu_f(\gamma), h\rangle$ from $X_i \overset{\text{iid}}{\sim} P_\gamma$ over the class $\{P_\gamma : \mu_f(\gamma) \in M_0\}$.

It turns out that this result can be extended: many other exponential families (i.e. different choices of $\phi : \mathcal{X} \to \mathbb{R}^d$) still enjoy the same property of (near) optimality of empirical-mean estimators. This extends the result of [JN09] to square loss and a wider class of exponential families (see discussion in Section 4.1). We proceed to formal definitions.

A d-dimensional exponential family $\{P_\gamma\}_{\gamma\in\Gamma}$ of probability distributions on a measurable space $\mathcal{X}$ is given by a triplet (ν, ϕ, Γ), where ν is a reference measure on $\mathcal{X}$, $\phi : \mathcal{X} \to \mathbb{R}^d$ is a measurable map, $\Gamma \subset \mathbb{R}^d$ and

$$P_\gamma(dx) = \exp\{\langle \gamma, \phi(x)\rangle - C(\gamma)\}\nu(dx)\,,$$

with $\gamma \in \mathbb{R}^d$ called the *natural parameter*. Let $\mathcal{F}$ be the finite-dimensional linear space spanned by basis functions ϕ_i, i.e., $\mathcal{F} = \{\langle h, \phi\rangle : h \in \mathbb{R}^d\}$. We make two standing assumptions on the exponential family:[8]

1. The set Γ is open and convex; $C(\gamma) < \infty$ for all $\gamma \in \Gamma$.
2. For some $\gamma_0 \in \Gamma$ (and hence for all γ by absolute continuity $P_\gamma \ll P_{\gamma_0}$), the functions $\phi_1, \dots, \phi_d$ are linearly independent, i.e.

$$\mathrm{Var}_{X\sim P_{\gamma_0}}(\langle \phi(X), h\rangle) > 0 \qquad \forall h \in \mathbb{R}^d \setminus \{0\}\,. \tag{68}$$

In addition to the natural parameter γ, we define the *mean parameter* μ via the forward map

$$\mu_f(\gamma) \triangleq \mathbb{E}_{X\sim P_\gamma}[\phi(X)]\,.$$

It is well known (see e.g. [Bro86]) that inside Γ the function $\gamma \mapsto C(\gamma)$ is infinitely differentiable, whose first two derivatives give the mean and covariance of $\phi(X)$:

$$\mu_f(\gamma) = \nabla C(\gamma), \qquad \frac{\partial \mu_f}{\partial \gamma} = \mathrm{Hess}\, C(\gamma) = \mathrm{Cov}_{P_\gamma}[\phi(X)] \triangleq \Sigma(\gamma)\,. \tag{69}$$

The non-degeneracy assumption (68) implies

$$\Sigma(\gamma) \succ 0 \qquad \forall \gamma \in \Gamma\,. \tag{70}$$

Since $C(\gamma)$ is, thus, strictly convex on Γ, the map $\gamma \mapsto \mu_f = \nabla C(\gamma)$ is one-to-one. Since the Jacobian of this map is non-zero everywhere on Γ, by the inverse function theorem the image $M \triangleq \mu_f(\Gamma)$ is an open set in $\mathbb{R}^d$ and, furthermore, there is an infinitely-differentiable inverse map γ_r such that

$$\mu_f(\gamma_r(\mu)) = \mu \qquad \forall \mu \in M\,.$$

It is also known that Jacobian of γ_r can be computed as

$$\frac{\partial \gamma_r(\mu)}{\partial \mu} = \Sigma^{-1}(\gamma_r(\mu))\,. \tag{71}$$

For convenience we denote $\tilde{P}_\mu = P_{\gamma_r(\mu)}$ and $\tilde{\Sigma}(\mu) = \Sigma(\gamma_r(\mu))$.

[8] Note that the second assumption is without loss of generality: if there is a linear relation between coordinates of ϕ, then by reducing the dimension d we eventually will make the second assumption hold.

For a given constraint set $\Gamma_0 \subset \Gamma$ and a functional $T(\gamma)$, we define the minimax square-loss as usual

$$R_n^*(\Gamma_0) = \inf_{\hat{T}} \sup_{\gamma \in \Gamma_0} \mathbb{E}_{X_i \overset{\text{iid}}{\sim} P_\gamma} [|\hat{T}(X_1, \ldots, X_n) - T(\gamma)|^2] \,. \tag{72}$$

The main finding in this section is that for estimating linear functionals of the mean parameter μ, under certain convexity assumptions (that are strictly weaker than those in [JN09]), the minimax quadratic risk is characterized by certain moduli of continuity within universal constant factors. To this end, let ω_H denote the modulus of continuity of T on M_0 with respect to the Hellinger distance, i.e.

$$\omega_H(t) \triangleq \sup_{\gamma, \gamma' \in \Gamma_0} \{T(\gamma) - T(\gamma') : H(P_\gamma, P_{\gamma'}) \le t\}, \tag{73}$$

Theorem 12. *There exist absolute constants $c_0 > 0$ and $c_1 > 0$ with the following property. Fix any dimension $d \ge 1$ and any exponential family (ν, ϕ, Γ) satisfying regularity assumptions 1 and 2 above. Consider a subfamily of an exponential family corresponding to mean parameters $\mu \in M_0 \subset M \subset \mathbb{R}^d$, where M_0 is compact and convex. Assume that the subfamily M_0 satisfies the key condition*

$$\mu \mapsto \sqrt{\mathrm{Var}_{P_\mu}[\phi]} \text{ is concave in } \mu \in M_0 \text{ for all } \phi \in \mathcal{F}. \tag{74}$$

Let the functional $T(\gamma)$ be linear in the mean parameter*, i.e.,*

$$T(\gamma) = \langle h, \mu_f(\gamma) \rangle \tag{75}$$

for some $h \in \mathbb{R}^d$, and define the constraint set $\Gamma_0 = \gamma_r(M_0)$. Then we have

$$c_0 \omega_H(1/\sqrt{n}) \le \sqrt{R_n^*(\Gamma_0)} \le c_1 \omega_H(1/\sqrt{n}), , \tag{76}$$

and this rate is achieved by the estimator of the type (67) with $g \in \mathcal{F}$.

The proof is given in Section 5.3. We stress that constants c_0, c_1 in (76) do not depend on dimension of the exponential family, and thus as in [JN09] we can think of the above result as essentially non-parametric.

Remark 5. Note that in the setting of the preceding Theorem 12, we have

$$T(\gamma) = \mathbb{E}_{X \sim P_\gamma}[\phi_0(X)] \,,$$

for some $\phi_0 \in \mathcal{F}$. Thus, it may appear that a good estimator would arise from taking $g = \phi_0$ in (67). Indeed, it gives an unbiased estimator by design. However, the subtlety here is that $\mathrm{Var}_{P_\gamma}[\phi_0(X)]$ might be prohibitively large (such as in population recovery in Section 2.2). The main discovery here is that the concavity condition (74) guarantees existence of some other $g \ne \phi_0$ such that the empirical average of g is minimax rate-optimal.

Remark 6. To shed some light on how assumption (74) relates to the tightness of empirical-mean estimators, we observe that the Fisher information matrix for parameter γ is given by $I_F(\gamma) = \Sigma(\gamma)$, while for parameter μ we get $I_F(\mu) = \tilde{\Sigma}^{-1}(\mu)$. In one dimension $d = 1$, we see that (158) shows that $R_n^*(M_0) \le \frac{1}{n \min_\mu I_F(\mu)}$. From the Bayesian Cramér-Rao lower bound (van Trees inequality) [GL95], we expect a similar lower bound to hold, unless $I_F(\mu)$ grows very rapidly around its minimum. The latter situation is prohibited by the assumption (74), as shown by the key inequality (162). Thus, assumption (74) enters our proof in two crucial ways: for the applicability of the minimax theorem and for taming the behavior of Fisher information. Because of the latter, it is unclear whether (74) can be extended from concavity to, say, quasi-concavity.

4.1 Comparison to Juditsky-Nemirovski [JN09]

As opposed to the squared loss (72), Juditsky-Nemirovski [JN09] considered the ϵ-quantile loss and the corresponding minimax risk:

$$R^*_{n,\epsilon}(\Gamma_0) \triangleq \inf_{\hat{T}} \sup_{\gamma\in\Gamma_0} \inf\left\{r : P_\gamma[|\hat{T}(X_1,\dots,X_n) - T(\gamma)| > r] \le \epsilon\right\}.$$

Nevertheless, Theorem 12 proved for the quadratic risk can be translated to the ϵ-quantile loss similarly as done in Corollary 2.

Corollary 13. *In the setting of Theorem 12, whenever $e^{-2n} \le \epsilon \le 2^{-8}$ we have up to absolute constants of proportionality*

$$R^*_{n,\epsilon}(\Gamma_0) \asymp \omega_H\left(\sqrt{\frac{1}{n}\ln\frac{1}{\epsilon}}\right). \tag{77}$$

Proof. The lower bound is proved by a Hellinger-based two-point argument as in Corollary 2. The upper bound follows from the same median trick as in Corollary 2. Bounding $\omega_H(ct)$ by $\omega_H(t)$ from above and below is done exactly as in the proof of Theorem 12. □

We now discuss results of [JN09]. The following assumptions are made in [JN09] (later called *a simple observation schemes* in [JN20, Section 2.4.2])

1. The exponential family (ν, ϕ, Γ) has $\Gamma = \mathbb{R}^d$, i.e. the natural parameters γ can range over the entire space $\mathbb{R}^d$.
2. The functional $T(\gamma) = T(A(\xi))$ is affine in ξ, where $\gamma = A(\xi)$ is a reparametrization such that the map
$$\xi \mapsto C(A(\xi)+a) - C(A(\xi)) \text{ is concave for every } a \in \mathbb{R}^d. \tag{78}$$

Under these assumptions, it is shown that (cf. [JN09, Theorem 3.1 and Proposition 3.1])

$$\frac{1}{2}\omega_H\left(\sqrt{2\left(1-e^{-\frac{1}{2n}\log\frac{1}{4\epsilon}}\right)}\right) \le R^*_{n,\epsilon} \le \frac{1}{2}\omega_H\left(\sqrt{2\left(1-e^{-\frac{1}{n}\log\frac{2}{\epsilon}}\right)}\right); \tag{79}$$

in particular, whenever $\exp(-2n) \le \epsilon < \frac{1}{5}$, we have (77) (also within absolute constants). Thus, to compare our results with [JN09] we need to compare the assumptions. It turns out (see Section 5.3 for a proof) that (78) is equivalent to the following requirement

$$\xi \mapsto \mu_f(A(\xi)) \text{ is affine and } \mu \mapsto \mathrm{Var}_{P_\mu}[\phi(X)] \text{ is concave in } \mu \in M_0 \text{ for all } \phi \in \mathcal{F}. \tag{80}$$

This equivalence shows that our condition (74) is strictly weaker than (80). Let us consider a simple example showing difference between (74) and (80).

Example 2 (Exponential distribution)**.** Let $\mathcal{X} = \mathbb{R}^d_+$, $\gamma \in \mathbb{R}^d_+$ and take $P_\gamma(dx) = \prod_{i=1}^d e^{-\gamma_i x_i}\mathbf{1}_{\{x_i>0\}}dx_i$, i.e. $X \sim P_\gamma$ has d independent components, each exponentially distributed. In this case $\phi(x) = x$. The mean parameters are $\mu = (\gamma_1^{-1}, \dots, \gamma_d^{-1})$. Our goal is to estimate $T(\gamma) = \sum_{i=1}^d \mu_i$ over the ℓ_p-ball in $\mathbb{R}^d$: $M_0 = \{\mu : \sum_{i=1}^d \mu_i^p \le 1\}$, where $p \ge 1$. A simple calculation shows

$$\omega_H(t) \asymp t d^{\max(\frac{1}{2}-\frac{1}{p},0)} \tag{81}$$

up to absolute (i.e. p-independent) constants. (For $p \le 2$ the worst pair (μ, μ') are 1-sparse (with a single nonzero), whereas for $p > 2$ they are scaled constant vectors.) From Theorem 12 we conclude that the minimax quadratic risk is $\Theta(\frac{1}{n} d^{\max(1-\frac{2}{p},0)})$. In this simple case the empirical mean $\hat{T} = \frac{1}{n}\sum_{t=1}^n \sum_{i=1}^d (X_t)_i$ achieves the optimal rate for all p, d, suggesting that the problem is rather simple. However, while our condition (74) holds, the condition (80) imposed by [JN09] does not. In addition, the natural parameter ranges over a subset of $\mathbb{R}^n$, not all of $\mathbb{R}^n$ (again in violation of [JN09]).

The above example thus shows that the extension from (80) to (74) is not vacuous. Another example is the normal scale model $X \sim \mathcal{N}(0, \sigma^2), \sigma^2 > 0$ with $\phi(x) = x^2$. For this family, again (74) holds but not (80). For larger dimension $d > 1$, the family $X \sim \mathcal{N}(0, \Sigma)$ with Σ a $d \times d$ positive definite matrix, does not satisfy either (80) or (74). (However, in this case a partial remedy is possible – see [JN20, Section 3.4.1].)

We point out, however, an important case of where our methods fail, but methods of [JN09,JN20] succeed. Namely, in [JN20, Section 3.1.4] it is shown that (in the notation of Theorem 12) the minimax risk of estimating a linear (in $\mu = \mu_f(\gamma)$) functional satisfies (77) also when M_0 is a finite union of convex sets (and in fact, the linear functional $T(\gamma)$ is allowed to be different on different convex sets). Unfortunatelly, this elegant result does not extend to the quadratic risk as the following example demonstrates.

Example 3. Consider the goal of estimating the bias p of $X_i \overset{\text{iid}}{\sim} \text{Bern}(p)$ where $p \in \{1/4, 1/3\}$. Then the modulus of continuity $\omega_H(t) = 0$ for $t < H(\text{Bern}(1/4), \text{Bern}(1/3))$, but this does not contradict (77) since $R^*_{n,\epsilon} = 0$ for all sufficiently large n. At the same time, it is clear that quadratic risk $R^*_n \ge c_1 e^{-c_2 n}$ for some constants $c_1, c_2 > 0$. Consequently, in the setting when M_0 is a union of convex sets, characterization $R^*_n \asymp \omega_H(1/\sqrt{n})$ is not possible.

5 Additional proofs

5.1 Proof of Theorem 10

Proof. Clearly the sufficient statistic is the histogram of the observed colors, that is, $\{N_i : i \in [n]\}$, where N_i is the number of observed balls of the ith color. Thus we have $N_i \overset{\text{ind.}}{\sim} \text{Binom}(\theta_i, p)$. Therefore, the setting of Theorem 10 is a particularization of the general Theorem 9, with $\Theta = \mathcal{X} = \mathbb{Z}_+$, $P_\theta = \text{Binom}(\theta, p)$, $c(\theta) = \theta$, $\Pi = \{\pi \in \mathcal{P}(\mathbb{Z}_+) : \int \theta \pi(d\theta) \le 1\}$ (which is weakly compact), and $h(\theta) = \mathbf{1}_{\{\theta \ge 1\}}$ so that $T(\pi) = 1 - \pi_0$, where we identify π with its PMF $\pi_k \equiv \pi(\{k\})$. Furthermore, the assumptions of Theorem 9 are fulfilled (with $K_V \le \frac{1}{4}$ and $\theta_0 = 0$). Applying Theorem 9, it remains to characterize the behavior of $\delta_{\chi^2}(t)$. Note that $\delta_{\chi^2}(t)$ is closely related to $\delta_{\chi^2}(t, n)$ previously studied for the population recovery problem in Section 2.2 (with $\epsilon = 1 - p$). Both dealing with the binomial model, the only difference is the additional moment constraint in δ_{χ^2} and the difference in the domain ($\mathbb{Z}_+$ versus $\{0, \dots, n\}$). Indeed, we have

$$\begin{aligned}
\delta_{\chi^2}(t) &= \sup\{\pi_0 - \pi'_0 : \chi^2(\pi P \| \pi' P) \le t^2, \pi, \pi' \in \Pi\} \\
&\le \sup\{\pi_0 - \pi'_0 : \chi^2(\pi P \| \pi' P) \le t^2, \pi, \pi' \in \mathcal{P}(\mathbb{Z}_+)\} \triangleq \delta'_{\chi^2}(t) \\
&\le \sup\{\pi_0 - \pi'_0 : \text{TV}(\pi P, \pi' P) \le t, \pi, \pi' \in \mathcal{P}(\mathbb{Z}_+)\} \triangleq \delta'_{\text{TV}}(t) \qquad (82) \\
&\le t^{\min(1, \frac{p}{1-p})}, \qquad (83)
\end{aligned}$$

where the last inequality follows from Lemma 6 (in particular (179) for $d = \infty$). Substituting $t = 1/\sqrt{n}$, this completes the proof of the upper bound $\sqrt{R^*(n)} \le n^{-\frac{1}{2}\min(1,\frac{p}{1-p})} \triangleq \epsilon_n$ as in (58) and (59).

For the constructive part, consider an estimator of the form (60), namely $\hat{T} = \frac{1}{n}\sum_{i=1}^n g(N_i)$. Choose g to be the solution g^* to the following LP (below $h = (0,1,\dots,1)$):

$$\min_{g\in\mathbb{R}^{n+1}} \|Pg - h\|_\infty + \frac{1}{\sqrt{n}}\|g\|_\infty, \tag{84}$$

which is equal to the dual LP

$$\max_{\Delta\in\mathbb{R}^{n+1}} \{\langle \Delta, h\rangle : \|\Delta P\|_1 \le t, \|\Delta\|_1 \le 1\}.$$

By [PSW17, Lemma 7], this LP is upper bounded by twice the value of the (82) with $t = 1/\sqrt{n}$, which shows the choice of g^* achieves the quadratic risk $4\epsilon_n^2$. The LP (84) (with $O(n)$ variables and $O(n)$ constraints) can be solved in time that is polynomial in n.

To finish the proof of the upper bound, we show that we can impose the constraint that $g(0) = 0$ and still achieve the upper bounds (58)–(59) within a constant factor. Indeed, from (84) we conclude that for all $\theta = 0,\dots,n$, $|(Pg^*)(\theta) - h(\theta)| = |\mathbb{E}_{N\sim\mathrm{Binom}(\theta,p)}[g^*(N)] - h(\theta)| \le 2\epsilon_n$. Particularizing to $\theta = 0$, we have $|g^*(0)| \le 2\epsilon_n$. Consider the modified estimator $\tilde{g}$ given by $\tilde{g}(0) = 0$ and $\tilde{g}(j) = g^*(j)$ for all $j \ge 1$. We have $\|\tilde{g}\|_\infty \le \|g^*\|_\infty$ and $\|P\tilde{g} - h\|_\infty \le \|Pg^* - h\|_\infty + \|P(\tilde{g} - g)\|_\infty \le \|Pg^* - h\|_\infty + 2\epsilon_n$. This shows $\tilde{g}$ achieve a quadratic risk of at most $16\epsilon_n^2$.

Next we proceed to the lower bound. The parametric lower bound in (58) follows from Remark 1. To complete the proof of (59), it remains to show the lower bound: for any $p \le \frac{1}{2}$ and all $t \le t_0(p)$ we have

$$\delta_{\chi^2}(t) \ge c t^{\frac{p}{1-p}} (\log t)^{-2} \tag{85}$$

for some constant $c = c(p) > 0$. To this end, we demonstrate a pair of feasible $\tilde{\pi}, \tilde{\pi}' \in \Pi$ by modifying the construction in the proof of [PSW17, Lemma 12] to satisfy the additional moment constraints. Therein,[9] it was shown that there exist probability distributions π, π' on $\mathbb{Z}_+$, such that $|\pi(0) - \pi'(0)| \ge \delta$ and

$$H^2(\pi P, \pi' P) \le 4\left(e^2 \delta_1 \log\frac{1}{\delta_1}\right)^{\frac{2(1-p)}{p}}, \tag{86}$$

whenever $\delta_1 = \frac{\delta}{p}$ satisfies $\delta_1 < e^{-1}$. More precisely, π and π' are obtained as follows: Let $\alpha = 1 - \frac{p}{\log\frac{1}{\delta_1}}$, $\beta = \delta_1 \log\frac{1}{\delta_1}$. Define $g : \mathbb{C} \to \mathbb{C}$ by $g(z) = \beta^{\frac{1+z}{1-z}}$. Set

$$f(z) = (1-\alpha) g(\alpha z) - (1-\alpha) g(\alpha)\,.$$

Define a sequence $\{\Delta_k : k \in \mathbb{Z}_+\}$ via the coefficients of the Taylor expansion of f, i.e., $\Delta_k \triangleq [z^k] f(z)$. Then $\Delta_k = (1-\alpha)\alpha^k [z^k] g(z)$ for $k \ge 1$. Define the following geometric distribution μ on $\mathbb{Z}_+$ by $\mu_k \triangleq \bar{\alpha}\alpha^k$. Define now π and π' via

$$\pi_k \triangleq \mu_k + \Delta_k\,, \quad \pi'_k \triangleq \mu_k - \Delta_k\,.$$

As shown in [PSW17, Lemma 12] we have $\pi_0 - \pi'_0 = 2\Delta_0 \ge \delta$.

[9] Original version of [PSW17], as published in the proceedings, contained an error in this derivation, see *arXiv:1702.05574v3* for correction.

Now we estimate the mean of π, π'. Note that the mean of the geometric distribution μ is $\sum_{k\geq 0} k\mu_k = \frac{1}{1-\alpha}$. Furthermore, since the generating function of Δ is f, using the facts that $f'(z) = \alpha(1-\alpha)g'(\alpha z)$ and $g'(z) = \frac{2\log\beta}{(1-z)^2}\beta^{\frac{1+z}{1-z}}$, we have

$$\sum_{k\geq 0} k\Delta_k = f'(1) = \alpha(1-\alpha)g'(\alpha) = \frac{2\alpha\log\beta}{1-\alpha}\beta^{\frac{1+\alpha}{1-\alpha}}$$

Since $\delta_1 \leq e^{-1}$ we have $\delta_1 \leq \beta \leq e^{-1}$ and $\bar{\alpha} < 1/2$, implying that

$$\left|\frac{2\alpha\log\beta}{1-\alpha}\beta^{\frac{1+\alpha}{1-\alpha}}\right| \leq \frac{2}{\bar{\alpha}} e^{1-\frac{2}{\bar{\alpha}}} \log\frac{1}{\delta_1}.$$

Since $xe^{-x} \leq e^{-1}$ on $x \geq 1$ we conclude

$$\left|\sum_{k\geq 0} \Delta_k\right| \leq \log\frac{1}{\delta_1}$$

and therefore, the first moments of π, π' are both bounded by $1/\eta$, where we set $\eta \triangleq \frac{\bar{\alpha}}{p+1} \leq 1/2$. Finally, define

$$\tilde{\pi} = (1-\eta)\delta_0 + \eta\pi, \quad \tilde{\pi}' = (1-\eta)\delta_0 + \eta\pi',$$

From previous estimates we have $|\sum_k k\tilde{\pi}_k| \leq 1$ and $|\sum_k k\tilde{\pi}'_k| \leq 1$ whenever $\delta < \frac{p}{e}$. By convexity, we have $H^2(\tilde{\pi}P, \tilde{\pi}'P) \leq \eta H^2(\pi P, \pi' P) \leq \frac{1}{2}H^2(\pi P, \pi' P)$. In summary, we have constructed $\tilde{\pi}, \tilde{\pi}' \in \Pi$ such that $|\tilde{\pi}(0) - \tilde{\pi}'(0)| \geq \eta\delta = \frac{1}{2}\delta\bar{\alpha}$ and

$$H^2(\tilde{\pi}P, \tilde{\pi}'P) \leq \frac{C}{2}\left(e^2\delta_1\log\frac{1}{\delta_1}\right)^{\frac{2(1-p)}{p}}.$$

Finally, choosing δ_1 so that the RHS of the previous display is t^2, i.e., $\delta_1 = \Theta(t^{\frac{p}{1-p}}/\log\frac{1}{t})$, we have $|\tilde{\pi}(0) - \tilde{\pi}'(0)| \geq \Omega((t)^{\frac{p}{1-p}}/(\log\frac{1}{t})^2)$. This completes the proof of (85) and the theorem. □

5.2 Proof of Theorem 11

We first present a key lemma, the proof of which requires delicate complex analysis and is postponed till the end of this subsection.

Lemma 14. *Consider the Poisson kernel $P(\cdot|\theta) = \mathrm{Poi}(\theta)$. For $s, t > 0$, define*

$$\delta(s,t) \triangleq \sup_{\Delta}\left\{\int e^{-s\theta}\Delta(d\theta) : \|\Delta P\|_{\mathrm{TV}} \leq t, \|\Delta\|_{\mathrm{TV}} \leq 1\right\}. \tag{87}$$

where the supremum is taken over all finite signed measure Δ on $\mathbb{R}_+$. Then for any $s > 0$ and $0 \leq t \leq 1$,

$$\delta(s,t) \leq t^{\min\{1,\frac{2}{s}\}}. \tag{88}$$

Furthermore, fix $s \geq 2$ and consider $\delta_{\chi^2}(t)$ in (7) with $\Theta = \mathbb{R}_+, \mathcal{X} = \mathbb{Z}_+$, $P(\cdot|\theta) = \mathrm{Poi}(\theta)$, $\Pi = \{\pi \in \mathcal{P}(\mathbb{R}_+) : \int \theta\pi(d\theta) \leq 1\}$ and $T(\pi) = \int e^{-s\theta}\pi(d\theta)$. There exist positive constants $c = c(s), t_1 = t(s)$ such that for all $t \leq t_1$,

$$ct^{\frac{2}{s}}\log^{-2}\frac{1}{t} \leq \delta_{\chi^2}(t) \leq 2t^{\frac{2}{s}}. \tag{89}$$

Before proving Theorem 11, we note that the species problem does not completely fall within the purview of Theorem 9, because the number of distinct species can be infinite. However, if the total number of species is restricted to $O(n)$, then the minimax rate readily follows from the general Theorem 9 coupled with the characterization of the modulus of continuity in (89), cf. (91)-(92) below. To deal with the full species problem without restriction, some extra argument is needed, which involves the auxiliary LP (87) and introduces extra logarithmic factors in the upper bound of (66).

Proof. The result (65) for $r \le 1$ simply follows from using Good-Toulmin's unbiased estimator and a parametric lower bound (cf. [GT56, OSW16]). Next we focus on proving (66) for $r > 1$.

Lower bound. We begin with some easy reductions. By (62), $U = \sum_x \mathbf{1}_{\{N_x=0\}} - V$, where $V \triangleq \sum_x \mathbf{1}_{\{N_x=0, N'_x=0\}}$, and hence estimating U and V are equivalent. Next, since V is concentrated near its mean, estimating V and $\mathbb{E}[V]$ are essentially equivalent. Indeed, by (62) and independence, we have

$$\mathrm{Var}(U) = \sum_x \mathrm{Var}(\mathbf{1}_{\{N_x=0, N'_x>0\}}) \le \mathbb{E}[U] \le rn.$$

Therefore for any estimator $\hat{V}$,

$$\mathbb{E}[(\hat{V} - V)^2] \ge \frac{1}{2}\mathbb{E}[(\hat{V} - \mathbb{E}[V])^2] - \frac{1}{2}\mathrm{Var}(V) \ge \frac{1}{2}\mathbb{E}[(\hat{V} - \mathbb{E}[V])^2] - rn. \tag{90}$$

Define $\theta_x = np_x$ and $h(\theta) = e^{-(r+1)\theta}$. Then $\mathbb{E}[V] = \sum_x h(\theta_x)$.

In order to apply the general result of Theorem 9, we introduce a restricted version of the species problem, where the number of distinct species is at most n. Thus any lower bound for the restricted species problem also holds for the original species problem. Denote the parameters by $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n) \in \boldsymbol{\Theta}_c \triangleq \{\theta \in \mathbb{R}_+^n : \sum_{i=1}^n \theta_i \le n\}$. Let the optimal risk for the restrictive problem be defined as usual:

$$\mathcal{E}_n^{(res)}(r) \triangleq \inf_{\hat{V}} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_c} \frac{1}{n^2}\mathbb{E}[(\hat{V} - \mathbb{E}[V])^2]\,. \tag{91}$$

Applying Theorem 9 with $c(\theta) = \theta$, $P = \mathrm{Poi}(\cdot)$, $\Pi = \{\pi \in \mathcal{P}(\mathbb{R}_+) : \int \theta\pi(d\theta) \le 1\}$ (which is weakly compact), and $h(\theta) = e^{-(r+1)\theta}$ (which is bounded), we obtain

$$\delta_{\chi^2}\left(\frac{1}{\sqrt{n}}\right)^2 \ge \mathcal{E}_n^{(res)}(r) \ge c\left(\delta_{\chi^2}\left(\frac{1}{\sqrt{n}}\right)^2 - \frac{1}{n}\right), \tag{92}$$

for some absolute constant c. Applying (89) in Lemma 14 with $t = \frac{1}{\sqrt{n}}$ and $s = r+1$, we obtained the lower bound $\delta_{\chi^2}(\frac{1}{\sqrt{n}}) \gtrsim n^{-\frac{1}{r+1}} \log^{-2}(n)$. The desired lower bound in (66) then follows from $\mathcal{E}_n^{(res)}(r) \le \mathcal{E}_n(r)$, (90), and (92).

Upper bound. We start with the construction of the estimator. Let $n_0 = \frac{n}{\log n}$ and $n_1 = n + n_0$. For notational convenience, we work with Poisson sampling parameter $n_1 = n(1 + \frac{1}{\log n})$ in place of n. Thus given observations $\{N_x\} \overset{\text{ind.}}{\sim} \mathrm{Poi}(n_1 p_x)$, the goal is to estimate $U = \sum_x \mathbf{1}_{\{N_x=0, N'_x>0\}}$ in (62), the number of unseen symbols that would be present in the next rn_1 observations, where $N'_x \sim \mathrm{Poi}(rn_1 p_x)$ By Poisson splitting, we have access to two independent sets of Poisson observations $\{\tilde{N}_x\} \overset{\text{ind.}}{\sim} \mathrm{Poi}(\lambda_x)$ and $\{\tilde{N}'_x\} \overset{\text{ind.}}{\sim} \mathrm{Poi}(\lambda'_x)$, where $\lambda_x \triangleq np_x$, $\lambda'_x \triangleq n_0 p_x = \frac{\lambda_x}{\log n}$ and $\tilde{N}_x + \tilde{N}'_x = N_x$. Let

$\tilde{U} \triangleq \sum_x \mathbf{1}_{\{\tilde{N}_x=0\}} - \mathbf{1}_{\{N_x=0,N'_x=0\}}$. Since $U = \tilde{U} - \sum_x \mathbf{1}_{\{\tilde{N}_x=0,\tilde{N}'_x>0\}}$, where the last sum is observed, thus estimating U is equivalent to estimating $\tilde{U}$.

To this end, fix a bounded sequence $f : \mathbb{Z}_+ \to \mathbb{R}$ to be optimized later. Fix a large constant C_0 and set a threshold $b' = C_0 \log n$. Consider an estimator of the following form

$$\hat{U} = \sum_x \hat{T}_x \tag{93}$$

where

$$\hat{T}_x = \begin{cases} 0 & \tilde{N}'_x \geq b' \\ f(\tilde{N}_x) & \tilde{N}'_x < b'. \end{cases} \tag{94}$$

Define

$$h(\lambda) \triangleq e^{-\lambda} - e^{-(1+\gamma)\lambda}, \quad \gamma \triangleq (1+r)\left(1 + \frac{n_0}{n}\right) - 1. \tag{95}$$

Note that

$$\mathbb{E}[\tilde{U}] = \sum_x (e^{-np_x} - e^{-(1+r)(n+n_0)p_x}) = \sum_x h(\lambda_x).$$

Then

$$\mathbb{E}[(\hat{U} - \tilde{U})^2] = \left(\sum_x (\mathbb{E}[\hat{T}_x] - h(\lambda_x))\right)^2 + \mathrm{Var}(\hat{U} - \tilde{U}).$$

A simple calculation shows that (cf. [OSW16, Lemma 3])

$$\mathrm{Var}(\hat{U} - \tilde{U}) \leq n(\|f\|_\infty^2 + \gamma). \tag{96}$$

To bound the bias, let $\epsilon = \frac{n_0}{n} = \frac{1}{\log n}$ and note that $\lambda'_x = \epsilon \lambda_x$. Set $b = b'/\epsilon = C_0 \log^2 n$. Using the definition of $\hat{T}_x$ and the independence of $\{\tilde{N}_x\}$ and $\{\tilde{N}'_x\}$, we have:

$$\begin{aligned}
&|\mathbb{E}[\hat{T}_x - h(\lambda_x)]| \\
= &\left|\mathbb{E}\left[(\hat{T}_x - h(\lambda_x))\left(\mathbf{1}_{\left\{\tilde{N}'_x \geq b', \lambda'_x \geq \frac{b'}{2}\right\}} + \mathbf{1}_{\left\{\tilde{N}'_x \geq b', \lambda'_x \leq \frac{b'}{2}\right\}} + \mathbf{1}_{\{\tilde{N}'_x \leq b', \lambda'_x \leq 2b'\}} + \mathbf{1}_{\{\tilde{N}'_x \leq b', \lambda'_x \geq 2b'\}}\right)\right]\right| \\
\leq\ & h(\lambda_x)\mathbf{1}_{\left\{\lambda'_x \geq \frac{b'}{2}\right\}} + h(\lambda_x)\mathbb{P}[\tilde{N}'_x \geq b']\mathbf{1}_{\left\{\lambda'_x \leq \frac{b'}{2}\right\}} \\
&+ |\mathbb{E}[f(\tilde{N}_x)] - h(\lambda_x)|\mathbf{1}_{\{\lambda'_x \leq 2b'\}} + (\|h\|_\infty + 1)\mathbb{P}[\tilde{N}'_x \leq b']\mathbf{1}_{\{\lambda'_x \geq 2b'\}} \\
\leq\ & \underbrace{h(\lambda_x)\mathbf{1}_{\left\{\lambda_x \geq \frac{b}{2}\right\}}}_{\text{(I)}} + \underbrace{h(\lambda_x)\exp(-b\kappa)\mathbf{1}_{\left\{\lambda_x \leq \frac{b}{2}\right\}}}_{\text{(II)}} \\
&+ \underbrace{|\mathbb{E}[f(\tilde{N}_x)] - h(\lambda_x)|\mathbf{1}_{\{\lambda_x \leq 2b\}}}_{\text{(III)}} + \underbrace{(\|f\|_\infty + 1)\exp(-b\kappa)\mathbf{1}_{\{\lambda_x \geq 2b\}}}_{\text{(IV)}},
\end{aligned}$$

where we used the Chernoff bound for Poisson distributions [MU05, Theorem 4.4]: for any $\lambda > 0$, $\mathbb{P}[\mathrm{Poi}(\lambda/2) \geq \lambda] \leq \exp(-\kappa\lambda)$ and $\mathbb{P}[\mathrm{Poi}(2\lambda) \leq \lambda] \leq \exp(-\kappa\lambda)$, with $\kappa \triangleq \log 2 - \frac{1}{2}$. Note that

$$\sum_x \lambda_x = n. \tag{97}$$

So

$$\sum_x \text{(I)} \leq \sum_x e^{-\lambda_x}(1 - e^{-r\lambda_x})\mathbf{1}_{\left\{\lambda_x \geq \frac{b}{2}\right\}} \leq \sum_x e^{-\lambda_x}\gamma\lambda_x\mathbf{1}_{\left\{\lambda_x \geq \frac{b}{2}\right\}} \leq \gamma n n^{-\frac{C_0}{2}}.$$

and

$$\sum_x (\mathrm{II}) \leq \sum_x e^{-\lambda_x}(1-e^{-r\lambda_x})\exp(-b\kappa) \leq \gamma n n^{-C_0\kappa},$$

and

$$\sum_x (\mathrm{IV}) \leq (\|f\|_\infty + 1) n^{-C_0\kappa}\frac{n}{2b}.$$

By choosing C_0 to be large constant, we have

$$\sum_x (\mathrm{I}) + (\mathrm{II}) + (\mathrm{IV}) \leq \gamma n^{-10}(\|h\|_\infty + 1).$$

Next to bound the main term (III), we choose the coefficient f by solving an LP, which is directly related to the LP (87) in Lemma 14. Let $f(k) = kg(k-1)$, where $g : \mathbb{Z}_+ \to \mathbb{R}$ is some sequence to be optimized later. Then by Stein's identity for Poisson distributions, we have $\mathbb{E}[f(\tilde{N}_x)] = \lambda_x \mathbb{E}[g(\tilde{N}_x)]$. Put

$$S(\lambda) \triangleq \frac{h(\lambda)}{\lambda} = \frac{e^{-\lambda} - e^{-(\gamma+1)\lambda}}{\lambda}.$$

Then we have $\mathbb{E}[f(\tilde{N}_x)] - h(\lambda_x) = \lambda_x(\mathbb{E}[f(\tilde{N}_x)] - S(\lambda_x))$. Recall that the Poisson kernel P acts as follows:

- For any sequence $g : \mathbb{Z}_+ \to \mathbb{R}$, $Pg : \mathbb{R}_+ \to \mathbb{R}$ is a function defined via $(Pg)(\lambda) \triangleq \mathbb{E}[g(\mathrm{Poi}(\lambda))]$;
- For any distribution π on $\mathbb{R}_+$, πP denotes the Poisson mixture whose probability mass function is given by $(\pi P)(k) = \int e^{-\lambda}\frac{\lambda^k}{k!}\pi(d\lambda), k \geq 0$.

For any $t > 0$, define the following bias-variance tradeoff LP:

$$\delta(t) \triangleq \inf_g \|S - Pg\|_{L_\infty(\mathbb{R}_+)} + t\|g\|_{\ell_\infty(\mathbb{Z}_+)}. \tag{98}$$

Next we bound $\delta(t)$ by the dual LP:

$$\begin{aligned}
\delta(t) &\overset{\text{(a)}}{=} \inf_{g \in \ell_\infty(\mathbb{Z}_+)} \sup_{\|\Delta\|_{\mathrm{TV}} \leq 1, \|\nu\|_{\mathrm{TV}} \leq 1} \int (S - Pg)d\Delta + t\int g d\nu \\
&\overset{\text{(b)}}{=} \sup_{\|\Delta\|_{\mathrm{TV}} \leq 1, \|\nu\|_{\mathrm{TV}} \leq 1} \inf_{g \in \ell_\infty(\mathbb{Z}_+)} \int (S - Pg)d\Delta + t\int g d\nu \\
&\overset{\text{(c)}}{=} \sup_{\|\Delta\|_{\mathrm{TV}} \leq 1, \|\nu\|_{\mathrm{TV}} \leq 1} \inf_{g \in \ell_\infty(\mathbb{Z}_+)} \int S d\Delta + \int g d(t\nu - \Delta P) \\
&\overset{\text{(d)}}{=} \sup_\Delta \left\{ \int S d\Delta : \|\Delta\|_{\mathrm{TV}} \leq 1, \|\Delta P\|_{\mathrm{TV}} \leq t \right\},
\end{aligned} \tag{99}$$

where in (a) Δ and ν are finite signed measures on $\mathbb{R}_+$ and $\mathbb{Z}_+$, respectively; (b) follows from Ky Fan's minimax theorem (Theorem 5), since $\{\Delta : \|\Delta\|_{\mathrm{TV}} \leq 1\}$ and $\{\nu : \|\nu\|_{\mathrm{TV}} \leq 1\}$ are compact in their respective weak topology, and for every bounded g, $\nu \mapsto \int g d\nu$ and $\Delta \mapsto \int (S - Pg)d\Delta$ are both weakly continuous since both S and Pg are bounded; (c) follows from Fubini's theorem: $\int Pg d\Delta = \int g d(\Delta P)$; (d) is because

$$\inf_{g \in \ell_\infty(\mathbb{Z}_+)} \int S d\Delta + \int g d(t\nu - \Delta P) = \begin{cases} -\infty & t\nu \neq \Delta P \\ 0 & t\nu = \Delta P \end{cases}$$

To relate the LP (99) to the LP (87) considered in Lemma 14, the key observation is the following integral representation:

$$S(\lambda) = \int_1^{\gamma+1} e^{-\lambda s} ds.$$

Interchanging the integral with the supremum in (99), we obtain the following upper bound

$$\delta(t) \leq \int_1^{\gamma+1} \delta(s,t) ds \tag{100}$$

where $\delta(s,t)$ is defined in (87). In view of (88) and (100), we have

$$\delta(t) \leq \gamma t^{\frac{2\gamma}{1+\gamma}}. \tag{101}$$

Thus, for the specific value of $t = \frac{1}{\sqrt{n}}$, there exists $g^* : \mathbb{Z}_+ \to \mathbb{R}$, such that

$$\sup_{\lambda \geq 0} |\mathbb{E}[g^*(\mathrm{Poi}(\lambda))] - S(\lambda)| \leq \gamma n^{-\frac{1}{1+\gamma}}, \qquad \|g^*\|_\infty \leq \gamma n^{\frac{1}{2}-\frac{1}{1+\gamma}}. \tag{102}$$

Next, we truncate g^*. Set $\lambda_0 = 2b$ and $L = 2\lambda_0 = 4C_0 \log^2 n$ and define g by

$$g(k) = g^*(k)\mathbf{1}_{\{k \leq L\}}. \tag{103}$$

Since $f(k) = kg(k-1)$, we have $\|f\|_\infty \leq L\|g^*\|_\infty$. In view of (96) and (102), we have the variance bound

$$\mathrm{Var}(\hat{U} - U) \leq 4\gamma L^2 n^{\frac{2\gamma}{1+\gamma}} = O(\gamma n^{\frac{2\gamma}{1+\gamma}} \log^4 n).$$

Furthermore, truncation incurs a small bias since

$$\left|\mathbb{E}\left[g^*(\tilde{N}_x)\mathbf{1}_{\{\tilde{N}_x > L\}}\right]\right| \leq \|g^*\|_\infty \mathbb{P}\left[\tilde{N}_x > L\right].$$

Note that $\mathbb{P}[\mathrm{Poi}(\lambda) > L] \leq \lambda \sum_{i \geq L} \frac{\lambda^{i-1}}{(i-1)!} e^{-\lambda} = \lambda \mathbb{P}[\mathrm{Poi}(\lambda) > L - 1]$. Thus

$$\begin{aligned}\sum_x |\mathbb{E}[g^*(\tilde{N}_x)\mathbf{1}_{\{\tilde{N}_x > L\}}]|\mathbf{1}_{\{\lambda_x \leq \lambda_0\}} &\leq n\|g^*\|_\infty \mathbb{P}[\mathrm{Poi}(\lambda_0) > 2\lambda_0 - 1] \\ &\overset{(102)}{\leq} \gamma n^{\frac{3}{2}-\frac{2\gamma}{1+\gamma}} \exp(-\kappa\lambda_0/2) \leq n^{-5}. \end{aligned} \tag{104}$$

Thus

$$\begin{aligned}\sum_x (\mathrm{III}) &= \sum_x |\mathbb{E}[f(\tilde{N}_x)] - h(\lambda_x)|\mathbf{1}_{\{\lambda_x \leq \lambda_0\}} \\ &= \sum_x \lambda_x |\mathbb{E}[g(\tilde{N}_x)] - S(\lambda_x)|\mathbf{1}_{\{\lambda_x \leq \lambda_0\}} \qquad (105) \\ &\overset{(103)}{\leq} \sum_x \lambda_x |\mathbb{E}[g^*(\tilde{N}_x)] - S(\lambda_x)|\mathbf{1}_{\{\lambda_x \leq \lambda_0\}} + \sum_x |\mathbb{E}[g^*(\tilde{N}_x)\mathbf{1}_{\{\tilde{N}_x > L\}}]|\mathbf{1}_{\{\lambda_x \leq \lambda_0\}} \qquad (106) \\ &\leq \gamma n^{\frac{r}{1+\gamma}} + n^{-5}, \qquad (107)\end{aligned}$$

where the last step follows from (97), (102) and (104).

Putting everything together, we have

$$\mathbb{E}[(\hat{U}-U)^2] \le \left(\sum_x \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)}\right)^2 + \mathrm{Var}(\hat{U}-U)$$
$$= O(\gamma^2 n^{\frac{2\gamma}{1+\gamma}} \log^4 n) = O(n^{\frac{2r}{1+r}} \log^4 n),$$

where the last step follows from the definition of γ in (95) and that $\frac{2\gamma}{1+\gamma} = \frac{2r}{1+r} + \frac{2\epsilon}{1+\gamma}$ with $\epsilon = \frac{n_0}{n} = \frac{1}{\log n}$. Recall that the above is proved for sample size $n_1 = n(1+1/\log n) \asymp n$. Dividing both sides by n^2 yields the upper bound in (66).

Finally, we address the construction of the estimator and its computational complexity. From the above proof, combining (93), (94), (98), (103) and (104), we see that it suffices to choose an estimator of the following form

$$\hat{U} = \sum_x \tilde{N}_x \cdot g^*(\tilde{N}_x - 1)\mathbf{1}_{\{\tilde{N}_x < L\}}\mathbf{1}_{\{\tilde{N}'_x < b'\}} \tag{108}$$

where g^* is the solution of the following infinite-dimensional LP:

$$\inf_g \|S - Pg\|_{L_\infty([0,\lambda_0])} + \frac{1}{\sqrt{n}}\|g\|_{\ell_\infty}, \tag{109}$$

with $(Pg)(\lambda) = \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}\left[g(N)\mathbf{1}_{\{N\le L\}}\right]$. Recall that λ_0, L and b are all $\Theta(\log^2 n)$. Here the decision variable $g : \{0,\ldots,L\} \to \mathbb{R}$ is finite-dimensional; however the objective function involves the L_∞-norm and is equivalent to setting a continuum of constraints. It remains to show that one can find a finite-dimensional LP whose solution is as good as (109), statistically speaking. We do so by means of discretization. From (102) we see that it suffices to consider $\|g\|_\infty \le \gamma n^{\frac{1}{2}-\frac{1}{1+\gamma}}$. For some small ε to be specified, let $m = \lfloor \lambda_0/\varepsilon \rfloor$ and $M \triangleq \varepsilon\{1,\ldots,m\}$. Consider the following discretized version of (109),

$$\inf_g \|S - Pg\|_{L_\infty(M)} + \frac{1}{\sqrt{n}}\|g\|_{\ell_\infty}, \tag{110}$$

To compare (109) and (110), note that for any $\lambda \in [0,\lambda_0]$, there exists $\lambda' \in M$ such that $|\lambda-\lambda'| \le \varepsilon$. Note that $S(\lambda) = \frac{h(\lambda)}{\lambda} = \frac{e^{-\lambda}-e^{-(\gamma+1)\lambda}}{\lambda}$ is L-Lipschitz in λ for some L depending only on r. Therefore $|S(\lambda)-S(\lambda')| \le L\varepsilon$. Furthermore, since $D(\mathrm{Poi}(\lambda)\|\mathrm{Poi}(\lambda')) = \lambda \log\frac{\lambda}{\lambda'} + \lambda' - \lambda \le \frac{(\lambda-\lambda')^2}{\lambda'} \le \varepsilon$, by Pinsker's inequality, we have $|(Pg)(\lambda) - (Pg)(\lambda')| \le \|g\|_\infty \mathrm{TV}(\mathrm{Poi}(\lambda), \mathrm{Poi}(\lambda')) \le \gamma\sqrt{n\varepsilon}$. Choosing $\varepsilon = \frac{1}{n^2}$, we conclude that the value of (109) and (110) only differs by $O(n^{-1/2})$, and solving which is an LP with $O(\log^2 n)$ variables and $O(n^2)$ constraints, achieves the upper bound in (66). $\square$

To close this section, we prove Lemma 14. The proof relies on two key results from complex analysis: Hadamard's three-lines theorem and the Paley-Wiener theorem.

Proof. We follow the same program of H^∞-relaxation as in the proof of Theorem 6 in [PSW17]. For a complex valued function on $U \subset \mathbb{C}$ we define $\|f\|_{H^\infty(U)} = \sup_{z\in U}|f(z)|$. If f is holomorphic on a domain U then $\|f\|_{H^\infty(U)} = \|f\|_{H^\infty(\partial U)}$ by the maximum principle. The open unit disk is denoted below as D and the unit circle as ∂D. To each finite signed measure Δ on $\mathbb{R}_+$ we associate its Laplace transform:

$$f_\Delta(z) \triangleq \int_{\mathbb{R}_+} e^{az}\Delta(da),$$

which is a holomorphic function on $\{\Re \le 0\}$ and

$$\|f_\Delta\|_{H^\infty(\Re\le 0)} = \|f_\Delta\|_{H^\infty(\Re=0)} \le \|\Delta\|_{\mathrm{TV}} \triangleq \int_{\mathbb{R}} |\Delta|(da). \tag{111}$$

Similarly, to each finite signed measure ν on $\mathbb{Z}_+$ we associate its z-transform

$$f_\nu(z) \triangleq \sum_{m\in\mathbb{Z}_+} \nu(m) z^m\,.$$

Again, f_ν is holomorphic on a D with

$$\|f_\nu\|_{H^\infty(D)} = \|f_\nu\|_{H^\infty(\partial D)} \le \|\nu\|_{\mathrm{TV}} \triangleq \sum_{m\in\mathbb{Z}_+} |\nu(m)|. \tag{112}$$

Furthermore, if f_ν happens to be holomorphic on rD for $r > 1$, then we have from Cauchy integral formula

$$|\nu(m)| \le r^{-m} \|f\|_{H^\infty(rD)} \tag{113}$$

The important observation for this proof is the following identity:

$$f_{\Delta P}(z) = f_\Delta(z-1)\,, \tag{114}$$

where Δ and ΔP are measures on $\mathbb{R}_+$ and $\mathbb{Z}_+$, with the latter obtained by applying the Poisson kernel P to Δ, to wit, $\Delta P(m) = \int \frac{e^{-a}a^m}{m!}\Delta(da)$. Indeed, (114) simply follows from Fubini's theorem: $f_{\Delta P}(z) = \int \sum_{m\ge 0} \frac{e^{-a}a^m}{m!}\Delta(da) = \int e^{a(z-1)}\Delta(da) = f_\Delta(z-1)$.

We now proceed to proving (88):

$$\begin{aligned}
\delta(s,t) &= \sup_\Delta \left\{ \int e^{-s\theta}\Delta(d\theta) : \|\Delta P\|_{\mathrm{TV}} \le t, \|\Delta\|_{\mathrm{TV}} \le 1 \right\} \\
&= \sup_\Delta \{ f_\Delta(-s) : \|f_\Delta\|_{H^\infty(D-1)} \le t, \|f_\Delta\|_{H^\infty(\Re<0)} \le 1\} &\quad (115)\\
&\le \sup_f \{ f(-s) : \|f\|_{H^\infty(D-1)} \le t, \|f\|_{H^\infty(\Re<0)} \le 1\} \triangleq \delta_{H^\infty}(t) &\quad (116)
\end{aligned}$$

where (115) is by expressing the objective function in terms of Laplace transform of Δ, and relaxing the total variation constraint on ΔP by the H^∞-norm constraint, in view of (111), (112) and (114); (116) is by extending the optimization from Laplace transforms f_Δ to all holomorphic functions on $\{\Re < 0\}$.

To solve the optimization problem (116) we first notice that for $s \le 2$, we have $-s \in D-1$ and thus $\delta_{H^\infty}(t) = t$ (achieved by taking $f(z) = t$). Next consider $s > 2$. Let us reparameterize $f(z) = g(1 + \frac{s}{z})$. Note that (cf. Fig. 1)

$$\|f\|_{H^\infty(\Re<0)} = \sup_{\Re(z)<0} |f(z)| = \sup_{\Re(z)<0} \left| g\left(1+\frac{s}{z}\right)\right| = \sup_{\Re(w)<1} |g(w)| = \|g\|_{H^\infty(\Re<1)}.$$

Furthermore, since

$$1 + \frac{1}{w} \in D \iff \Re(w) \le -\frac{1}{2}, \tag{117}$$

we have

$$\|f\|_{H^\infty(D-1)} = \sup_{z\in D-1} \left| g\left(1+\frac{s}{z}\right)\right| = \sup_{1+\frac{x}{w-1}\in D} |g(w)| = \sup_{\Re(w)<1-\frac{s}{2}} |g(w)| = \|g\|_{H^\infty(\Re<1-\frac{s}{2})}.$$

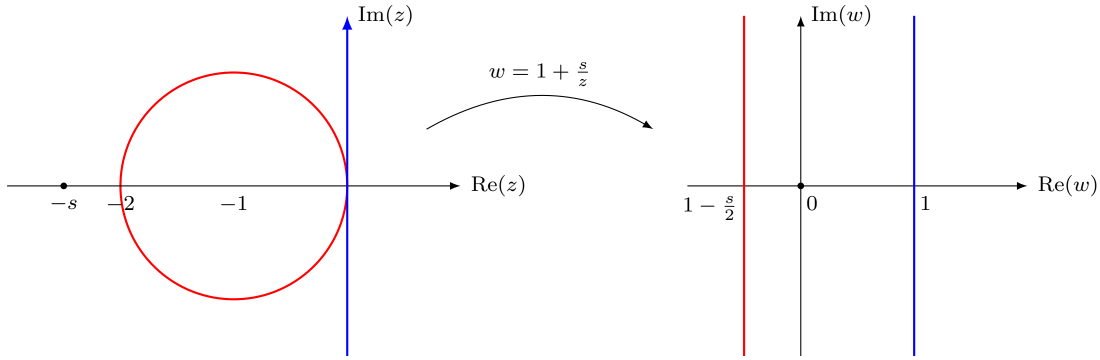

Figure 1: The function $w = 1 + \frac{s}{z}$ maps the circle $-1 + \partial D$ to the line $\Re = 1 - \frac{s}{2}$, $\Re = 0$ to $\Re = 1$, and the point $z = -s$ to $w = 0$.

Hence, we have

$$\delta_{H^\infty}(t) = \sup_g \{g(0) : \|g\|_{H^\infty(\Re<1-\frac{s}{2})} \le t, \|g\|_{H^\infty(\Re<1)} \le 1\}. \tag{118}$$

For any g feasible to (118), which is bounded on the strip $\{z : 1-\frac{s}{2} \le \Re(z) \le 1\}$, by Hadamard's three-lines theorem (see, e.g., [Sim11, Theorem 12.3]), $x \mapsto \log \|g\|_{H^\infty(\Re<x)}$ is convex. Since $(1 - \frac{s}{2})\frac{2}{s} + (1 - \frac{2}{s}) = 0$, we get

$$|g(0)| \le \|g\|_{H^\infty(\Re<0)} \le \left(\|g\|_{H^\infty(\Re<1-\frac{s}{2})}\right)^{\frac{2}{s}} \left(\|g\|_{H^\infty(\Re<1)}\right)^{1-\frac{2}{s}} \le t^{\frac{2}{s}},$$

for any g feasible for (118). Furthermore, this is achieved by taking $g(z) = t^{\frac{2}{s}(1-z)}$. So we have proved

$$\delta_{H^\infty}(t) = t^{\frac{2}{s}},$$

and the optimizer in (116) is

$$f_*(z) = t^{-\frac{2}{z}}, \tag{119}$$

which turns out to not depend on s. This completes the proof of (88).

Next we prove (89) for $s \ge 2$. The upper bound is clear:

$$\begin{aligned} \delta_{\chi^2}(t) &\le \delta_{\mathrm{TV}}(t) && (120)\\ &\le \sup_\Delta \left\{ \int \Delta(d\theta) e^{-s\theta} : \|\Delta P\|_{\mathrm{TV}} \le 2t, \|\Delta\|_{\mathrm{TV}} \le 2 \right\} && (121)\\ &= 2\delta(s,t) \le 2t^{\frac{2}{s}} && (122) \end{aligned}$$

where (120) is from (15), (121) is by dropping the constraint $\pi, \pi' \in \Pi$ and taking $\Delta = \pi' - \pi$, and (122) is by (88).

Finally, we prove the lower bound part of (89). To this end we need to produce a pair of distributions π, π' that are feasible for $\delta_{\chi^2}(t)$. We could try to take them to be positive and negative part of the measure Δ that whose Laplace transform coincides with (119), i.e., $f_\Delta = f_*$; however, this approach does not directly work (for example, if Δ were a finite measure, its characteristic function would have been given by $e^{\frac{ic_t}{\omega}} 1\{\omega \ne 0\}$, which is discontinuous at $\omega = 0$ and thus not the characteristic function of any finite measure on $\mathbb{R}$). Instead, below we construct a sequence of measures approximating Δ.

For each $0 < \alpha < 1$ (in the end we will take $\alpha \sim \frac{1}{\log \frac{1}{t}}$) define

$$f_\alpha(z) = \frac{1}{(z-1)^2} t^{-\frac{2}{z-\alpha}} = \frac{1}{(z-1)^2} e^{c_t/(z-\alpha)}, \quad c_t \triangleq 2 \log \frac{1}{t}.$$

Let G_α be a real-valued function on $\mathbb{R}$ (whose existence is to be established), such that its Laplace transform is given by f_α, i.e.

$$\int_{\mathbb{R}} G_\alpha(a) e^{az} da = f_\alpha(z) \qquad \forall z : \Re(z) \leq 0\,.$$

Let H_0 be the following probability distribution on $\mathbb{R}_+$

$$H_0(dx) = (1-\lambda)\delta_0(dx) + \lambda\gamma e^{-\gamma x} 1\{x \geq 0\}\, dx\,,$$

which is a mixture of a point mass at zero and an exponential distribution. We then take

$$\pi = H_0, \quad \pi' = (1-\tau_0)H_0 + \xi G_\alpha\,,$$

where

$$\tau_0 = \xi \int_{\mathbb{R}} G_\alpha(x) dx = \xi f_\alpha(0) = \xi e^{-\frac{c_t}{\alpha}} \tag{123}$$

so that π' is normalized. To complete the proof we have to prove that a certain choice of $(\alpha, \xi, \gamma, \lambda)$ achieves the following six goals for all sufficiently small t:

1. G_α is a real-valued density[10] supported on $\mathbb{R}_+$;
2. π' is a probability measure (i.e. it is a positive measure);
3. $\mathbb{E}_\pi[\theta] \leq 1$;
4. $\mathbb{E}_{\pi'}[\theta] \leq 1$;
5. The separation of means satisfies:

$$T(\pi') - T(\pi) \geq \frac{K}{(1+s)^2 \log^2 \frac{1}{t}} t^{\frac{2}{s}}\,,$$

   for some constant K (here and below, K denotes an absolute constant, possibly different on different lines), where recall that $T(\pi) = \mathbb{E}_\pi[e^{-s\theta}]$;
6. The χ^2-divergence satisfies:

$$\chi^2(\pi' P \| \pi P) \leq t^2.$$

We make the following choices of parameters:

$$\gamma = \frac{\alpha}{2}, \lambda = \frac{\alpha}{4}, \xi = \frac{\alpha^2}{16}, \alpha = \frac{1}{c_t} \tag{124}$$

Note that as $t \to 0$, all of the above vanish with polylog$(\frac{1}{t})$ speed.

[10] Although not directly needed for the statistical lower bound, we require G_α to have a density in order to apply the Paley-Wiener theorem which ensures it is supported on $\mathbb{R}_+$ and hence can be used as a valid prior.

We start with item 1. To get a formula for G_α we notice that the inverse Fourier transform is well-define. Indeed, since $|f_\alpha(i\omega)| = \frac{1}{1+\omega^2}\exp(-\frac{c_t\alpha}{\omega^2+\alpha^2})$, we have $\omega \mapsto f_\alpha(i\omega)$ is in $L_1(\mathbb{R})$. Hence there exists a continuous bounded function G_α on $\mathbb{R}$ whose Fourier transform is given by $f_\alpha(i\omega)$. Moreover, G_α is real-valued since $f_\alpha(-i\omega) = (f_\alpha(i\omega))^*$, where $*$ denotes the complex conjugation. To ensure that G_α is supported on $\mathbb{R}_+$, note that f_α is holomorphic in $\{\Re \le 0\}$ and, furthermore,

$$|f_\alpha(x+iy)| = \frac{1}{(1-x)^2+y^2}\exp\left(\frac{-c(x-\alpha)}{(x-\alpha)^2+y^2}\right),$$

thus

$$\sup_{x<0}\int_{\mathbb{R}} |f_\alpha(x+iy)|^2 dy \le \int_{\mathbb{R}} \frac{1}{1+y^2}dy \exp\left(\frac{c}{\alpha}\right) < \infty.$$

Then the Paley-Wiener theorem (cf. [Rud87, Theorem 19.2]) implies that G_α is supported on $\mathbb{R}_+$. We also get an estimate on the tail of $G_\alpha(a)$ for $a > 0$ as follows: By the inverse Fourier transform,

$$\begin{aligned} G_\alpha(a) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\frac{c_t}{i\omega-\alpha}}\frac{1}{(i\omega-1)^2}e^{-i\omega a}d\omega \\ &= \frac{1}{2\pi i}\int_{0-i\infty}^{0+i\infty} e^{\frac{c_t}{z-\alpha}}\frac{1}{(z-1)^2}e^{-za}dz \\ &= \frac{1}{2\pi i}\int_{\frac{\alpha}{2}-i\infty}^{\frac{\alpha}{2}+i\infty} e^{\frac{c_t}{z-\alpha}}\frac{1}{(z-1)^2}e^{-za}dz \qquad (125) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\frac{c_t}{i\omega-\frac{\alpha}{2}}}\frac{1}{(i\omega+\frac{\alpha}{2}-1)^2}e^{-(i\omega+\frac{\alpha}{2})a}d\omega, \end{aligned}$$

where in (125) we shifted the contour of integration since the integrand is holomorphic in the strip $\{0 \le \Re \le \frac{\alpha}{2}\}$. Thus

$$\begin{aligned} |G_\alpha(a)| &\le \frac{e^{-a\frac{\alpha}{2}}}{2\pi}\int_{-\infty}^{\infty}\frac{1}{\omega^2+(1-\frac{\alpha}{2})^2}d\omega \\ &= \frac{1}{2(1-\frac{\alpha}{2})}e^{-a\frac{\alpha}{2}} \le e^{-a\frac{\alpha}{2}}, \qquad (126) \end{aligned}$$

where the last step follows from $\int_{-\infty}^{\infty}\frac{1}{K^2+x^2}dx = \frac{\pi}{K}$ and the assumption that $\alpha \le 1$.

We proceed to item 2. In view of (126), to ensure the positivity of π' we only need to verify

$$(1-\tau_0)\lambda\gamma e^{-a\gamma} \ge \xi e^{-\frac{a\alpha}{2}}$$

Due to the choices in (124) this is equivalent to $1-\tau_0 \ge \frac{1}{2}$ which is satisfied for sufficiently small t.

For item 3, we have $\mathbb{E}_\pi[\theta] = \lambda\frac{1}{\gamma} = \frac{1}{2}$.

For item 4, we can compute the first moment of G_α from its Laplace transform as follows:

$$\int_0^\infty G_\alpha(a)ada = \frac{d}{dz}\bigg|_{z=0} f_\alpha(z) = e^{-\frac{c_t}{\alpha}}(2-\tfrac{c_t}{\alpha^2}) = e^{-\frac{1}{\alpha^2}}(2-\tfrac{1}{\alpha^3}) \to 0,$$

since $\alpha \to 0$ as $t \to 0$. Thus, we have $\mathbb{E}_{\pi'}[\theta] = (1-\tau_0)\frac{1}{2} + \xi\int aG_\alpha \to \frac{1}{2}$ as $t \to 0$.

For item 5, note that

$$T(G_\alpha) = \int e^{-sa}G_\alpha(a)da = f_\alpha(-s) = \frac{1}{(s+1)^2}t^{\frac{2}{s+\alpha}} \ge \frac{1}{(s+1)^2}t^{\frac{2}{s}}, \qquad (127)$$

Since $T(H_0) = 1 - \frac{s\lambda}{s+\gamma} \in [0,1]$, by linearity, we have from (127)

$$T(\pi') - T(\pi) = \; -\tau_0\left(1 - \frac{s\lambda}{s+\gamma}\right) + \xi\int_0^\infty e^{-a}G_\alpha(a)da \geq -\tau_0 + \frac{\xi}{(s+1)^2}t^{\frac{2}{s}}$$
$$\overset{(123)}{=} \xi\left(\frac{1}{(s+1)^2}t^{\frac{2}{s}} - e^{-4\log^2\frac{1}{t}}\right) \geq \frac{\xi}{2(s+1)^2}t^{\frac{2}{s}},$$

where the last step holds for all sufficiently small t.

Finally, for item 6, we have

$$\chi^2((1-\tau_0)H_0P + \xi G_\alpha P \| H_0P) = \sum_{m\geq 0}\frac{(\xi G_\alpha P(m) - \tau_0 H_0P(m))^2}{H_0P(m)}$$
$$= \xi^2\sum_{m\geq 0}\frac{G_\alpha P(m)^2}{H_0P(m)} - \tau_0^2 \leq \xi^2\sum_{m\geq 0}\frac{G_\alpha P(m)^2}{H_0P(m)}. \tag{128}$$

For the denominator we have

$$H_0P(m) = (1-\lambda)\mathbf{1}_{\{m=0\}} + \lambda(1-\beta)\beta^m, \quad \beta = \frac{1}{\gamma+1}. \tag{129}$$

To bound the numerator, by (114) the z-transform of $G_\alpha P$ is given by

$$f_{G_\alpha P}(z) = f_\alpha(z-1) = \frac{1}{(z-2)^2}e^{\frac{c_t}{z-1-\alpha}} \tag{130}$$

Our goal is to show that, for $r = 1 + \frac{\alpha}{2}$, we have $\|f_{G_\alpha P}\|_{H^\infty(rD)} \leq Kt$ for some constant K. Indeed, the first factor in (130) is bounded by $\|\frac{1}{(z-2)^2}\|_{H^\infty(rD)} \leq \frac{1}{(1-\alpha/2)^2} \leq 4$ for all sufficiently small t. For the second factor, in view of (117), for any $\rho > 0$ we have

$$\|e^{\rho/z}\|_{H^\infty(D-1)} = e^{-\rho/2}. \tag{131}$$

Set $\rho = 1 + \frac{3\alpha}{4}$, we have

$$\|f_{G_\alpha P}\|_{H^\infty(rD)} \overset{(a)}{\leq} 4\|e^{c_t/z}\|_{H^\infty(rD-1-\alpha)} \overset{(b)}{\leq} 4\|e^{c_t/z}\|_{H^\infty(\rho(D-1)} \overset{(c)}{=} 4e^{-\frac{c_t}{2\rho}} \overset{(d)}{\leq} 10t,$$

where (a) is by (130); (b) is because $rD - 1 - \alpha \subset \rho(D-1)$; (c) is by (131); (d) is by the choices in (124).

From Cauchy's integral formula (113) we obtain the estimate of the coefficients:

$$G_\alpha P(m) \leq Kr^{-m}t. \tag{132}$$

Using (129) and (132) we continue (128) to get

$$\chi^2((1-\tau_0)H_0P + \xi G_\alpha P \| H_0P) \leq Kt^2\frac{\xi^2}{\lambda\gamma}\sum_{m\geq 0}(r^2\beta)^{-m}.$$

Since $r^2\beta = 1 + \frac{\alpha}{2\epsilon} + o(\alpha)$ we conclude

$$\chi^2((1-\tau_0)H_0P + \xi G_\alpha P \| H_0P) \leq Kt^2\frac{\xi^2}{\lambda\gamma\alpha} \leq t^2$$

for all sufficiently small t due to (124). This completes the proof of (89). □

5.3 Proof of Theorem 12

Lemma 15 (Auxiliary convex analysis). *Let X and Y be a dual pair of finite-dimensional vector spaces and Π a compact convex subset of X. Let $f(x,y)$ be a function on $\Pi \times Y$ concave in x and convex in y. Assume in addition:*

1. *There exists $e_0 \in Y$ such that $\langle x, e_0 \rangle = 1$ for any $x \in \Pi$.*
2. *We have $f(x, y + ce_0) = f(x,y)$ for any $c \in \mathbb{R}$.*
3. *For any $c \in \mathbb{R}$ we have*[11]
$$f(x, cy) = |c| f(x,y)\,.$$

Fix $g \in Y$ and define the following quantities

$$d(x' \| x) \triangleq \sup\{\big\langle x - x', y\big\rangle : f(x,y) \le 1\}, \tag{133}$$
$$d_S(x', x) \triangleq d(x', (x+x')/2), \tag{134}$$
$$\delta_0(t) \triangleq \inf_y \sup_{x \in \Pi} t f(x,y) + |\langle x, g - y\rangle|, \tag{135}$$
$$\delta_1(t) \triangleq \sup_{x,x' \in \Pi} \{\big\langle x - x', g\big\rangle : d(x'\|x) \le t\}, \tag{136}$$
$$\delta_2(t) \triangleq \sup_{x,x' \in \Pi} \{\big\langle x - x', g\big\rangle : d_S(x', x) \le t\}. \tag{137}$$

We claim the following:

1. *$d_S(x', x) = d_S(x, x')$*
2. *$d_S(x', x) \le d(x\|x')$*
3. *$\frac{1}{2}\delta_2(t) \le \delta_1(t) \le \delta_2(t)$*
4. *And the key result:*
$$\frac{1}{2}\delta_1(t) \le \delta_0(t) \le \delta_1(t). \tag{138}$$

Proof. 1. This is clear.

2. To prove $d(x'\|(x+x')/2) \le d(x\|x')$ just notice that $f((x+x')/2, y) \le 1$ implies $f(x,y) \le 2$ by concavity and positivity.

3. Implied from above.

4. For the lower bound notice

$$\delta_0(t) = \inf_y \sup_{x,x' \in \Pi} t\frac{f(x,y) + f(x',y)}{2} + \frac{|\langle x, g-y\rangle| + |-\langle x', g-y\rangle|}{2}\,.$$

In the inner supremum we set x, x' to be the ones achieving $\delta_1(t)$. Then we have $\langle x - x', g\rangle \ge \delta_1$ and for any y we have
$$\big\langle x - x', y\big\rangle \le t f(x,y)\,. \tag{139}$$

[11] In particular, this implies that $f(x,y) = f(x,-y)$, $f(x,0) = 0$ and $f \ge 0$.

We further lower bound

$$\delta_0(t) \geq \frac{1}{2}\inf_y t(f(x,y)+f(x',y)) + |\langle x-x', g-y\rangle| \tag{140}$$

$$\geq \frac{1}{2}\inf_y tf(x,y) + |\langle x-x', g-y\rangle| \tag{141}$$

$$\geq \frac{1}{2}\langle x-x', g\rangle + \frac{1}{2}\inf_y tf(x,y) - \langle x-x', y\rangle\,, \tag{142}$$

where in the first step we used convexity of $|\cdot|$, in the second positivity of f and in the last step $|a| \geq a$. From (139) we conclude that $\delta_0 \geq \frac{1}{2}\delta_1$.

To prove an upper bound we denote the convex hull $\Pi_2 = \mathrm{co}\{0, 2\Pi\} = \{\mu x : x \in \Pi, \mu \in [0,2]\}$ and notice

$$\delta_0(t) \leq \inf_y \sup_{x\in\Pi, x'\in\Pi_2} tf(x,y) + \langle x-x', g-y\rangle$$

We now apply minimax theorem to get

$$\delta_0(t) \leq \sup_{x\in\Pi, x'\in\Pi_2} \inf_y tf(x,y) + \langle x-x', g-y\rangle$$

We notice that the inner infimum is $-\infty$ unless $\langle x-x', h\rangle = 0$, i.e. that $x' \in \Pi$, and thus

$$\delta_0(t) \leq \sup_{x\in\Pi, x'\in\Pi} \inf_y tf(x,y) + \langle x-x', g-y\rangle \tag{143}$$

Due to the homogeneity of $f(x,\cdot)$ we see that further

$$\inf_y tf(x,y) - \langle x-x', y\rangle = \begin{cases} -\infty, & d(x'\|x) > t\,, \\ 0, d(x'\|x) \leq t \end{cases}.$$

Consequently, the right-hand side of (143) evaluates to exactly $\delta_1(t)$.

□

Proof of Theorem 12. Recall that $D(P\|Q) = \int dP \log \frac{dP}{dQ}$ denote the Kullback-Leibler (KL) divergence. We need to introduce two other divergence-like quantities before proceeding.

$$d_J(P,Q) \triangleq D(P\|Q) + D(Q\|P) = \int dP \log\frac{dP}{dQ} + dQ\log\frac{dQ}{dP}$$

$$d(P_{\gamma'}\|P_\gamma) \triangleq \sup_{\phi\in\mathcal{F}}\{\mathbb{E}_{P_\gamma}[\phi] - \mathbb{E}_{P_{\gamma'}}[\phi] : \mathrm{Var}_{P_\gamma}[\phi] \leq 1\}.$$

We notice that d_J is known as the Jeffreys divergence, while $d(P_\gamma\|P_{\gamma'})$ describes the dissimilarity between distributions $P_{\gamma'}$ and P_γ in terms of the expectations of unit-variance functions in $\mathcal{F}$;[12] an explicit expression for d is given in (156) below. The modulus of continuity of T with respect to d_J and d will also play a role:

$$\omega_J(t) \triangleq \sup_{\gamma,\gamma'\in\Gamma_0}\{T(\gamma) - T(\gamma') : d_J(P_\gamma, P_{\gamma'}) \leq t^2\}\,,$$

$$\omega_d(t) \triangleq \sup_{\gamma,\gamma'\in\Gamma_0}\{T(\gamma) - T(\gamma') : d(P_{\gamma'}\|P_\gamma) \leq t\}$$

[12] Note that without the restriction $\phi \in \mathcal{F}$, the supremum coincides with $\chi(P_{\gamma'}\|P_\gamma)$; see (32).

We start by establishing the following comparison

$$\omega_J(t) \le \omega_H(t)\,. \tag{144}$$

Indeed, an application of Jensen inequality shows

$$-2\log(1-\tfrac{1}{2}H^2(P,Q)) = -2\log\int\sqrt{dPdQ} \le D(P||Q)$$

and from symmetry we, thus, have

$$-2\log(1-\tfrac{1}{2}H^2(P,Q)) \le \tfrac{1}{2}d_J(P,Q)\,.$$

Lower bounding the left-hand side we get $H^2(P,Q) \le \frac{1}{2}d_J(P,Q) \le d_J(P,Q)$ completing (144).

A routine two-point argument yields the lower bound

$$\omega_H(c_3/\sqrt{n}) \le \sqrt{R_n^*(\Gamma_0)} \tag{145}$$

for some absolute constant c_3.

Our proof will be completed in the following steps:

- First, we show by appealing to the minimax theorem the constructive part:

$$\sqrt{R_n^*(\Gamma_0)} \le \omega_d(1/\sqrt{n})\,. \tag{146}$$

- Next we will show that for some $c_2 > 0$ and all $t > 0$ we have

$$\omega_J(t) \ge \omega_d(c_2 t)\,. \tag{147}$$

- Subadditivity property of ω_d:

$$\omega_d(ct) \ge c\omega_d(t), \quad \forall 0 \le c \le 1\,. \tag{148}$$

- Together (144), (145) and the three steps above imply

$$\omega_H(\frac{c_0}{\sqrt{n}}) \le \sqrt{R_n^*} \le \omega_d(\frac{c_2}{c_2\sqrt{n}}) \le \frac{1}{c_2}\omega_H(\frac{1}{\sqrt{n}})\,. \tag{149}$$

  The proof will conclude by showing that for all $n \ge \frac{2}{c_3^2}$ we have (for some constant $c_0 > 0$):

$$\omega_H(\frac{c_3}{\sqrt{n}}) \ge c_0\omega_H(\frac{1}{\sqrt{n}}) \tag{150}$$

We proceed to proving the above claims. Let us extend the family $\mathcal{F}$ to $\mathcal{F}^* = \mathrm{span}\{\mathcal{F}, 1\}$ by adding constants. Similarly, we extend ϕ to $\phi^*(x) = (1, \phi(x)) \in \mathbb{R}^{d+1}$ by adding a constant coordinate. (Note that as exponential family ϕ^* no longer satisfies non-degeneracy condition (68)). We show an upper bound by considering estimators of the form

$$\hat{T} = \frac{1}{n}\sum_i g(X_i) = \frac{1}{n}\sum_i \langle \gamma^*, \phi^*(X_i)\rangle\,,$$

where g is an arbitrary (to be selected) element of $\mathcal{F}^*$, which we represented (here and below) as $g(x) = \langle \gamma^*, \phi^*(x)\rangle$ for some $\gamma^* \in \mathbb{R}^{d+1}$. (So everywhere above g and γ^* are coupled by this relation.) Recall that $\mathbb{E}_\gamma[\phi(X)] = \mu_f(\gamma) = \mu$ and the functional to be estimated is $T(\gamma) = \langle h, \mu\rangle = \mathbb{E}_\gamma[\langle h, \phi(X)\rangle]$. Define $h^* = (0, h)$ and $\mu^* = (1, \mu) = \mathbb{E}[\phi^*(X)]$. Then we have $T(\gamma) = \langle h^*, \mu^*\rangle = \mathbb{E}_\gamma[\langle h^*, \phi^*(X)\rangle]$. We have:

$$\inf_{g\in\mathcal{F}^*} \sup_{\gamma\in\Gamma_0} \sqrt{\mathbb{E}_\gamma[(T-\hat{T})^2]} \le \inf_{g\in\mathcal{F}^*} \sup_{\gamma\in\Gamma_0} \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{P_\gamma}[g(X)]} + |\mathbb{E}_{P_\gamma}[g(X)] - T(\gamma)| \tag{151}$$

$$= \inf_{\phi\in\mathcal{F}^*} \sup_{\mu\in M_0} \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\tilde{P}_\mu}[g(X)]} + |\langle \gamma^* - h^*, \mu^*\rangle| \tag{152}$$

$$= \delta_0(1/\sqrt{n}), \tag{153}$$

where δ_0 is defined as (similar to δ_{bv} previously defined in (24))

$$\delta_0(t) \triangleq \inf_{g\in\mathcal{F}^*} \sup_{\mu\in M_0} t\sqrt{\mathrm{Var}_{\tilde{P}_\mu}[g(X)]} + |\langle \gamma^* - h^*, \mu^*\rangle|.$$

This definition coincides with δ_0 defined in Lemma 15 if we set:

- $X = \mathbb{R}^{d+1}$, $\Pi = \{1\} \times M_0$, $Y = \mathcal{F}^*$
- Each element $g \in \mathcal{F}^*$ can be written as $g(x) = y_0 + \sum y_i\phi_i(x) = \langle y, \phi^*(x)\rangle$, this identifies $Y = \mathcal{F}^*$ with $\mathbb{R}^{d+1}$.
- We establish the dual pairing between X and Y as usual $\langle x, y\rangle = \sum_{i=0}^n x_iy_i$. Note that when $x = (1, \mu) \in \Pi$ and y is identified with g, we have $\langle x, y\rangle = \mathbb{E}_{\tilde{P}_\mu}[g(X)]$.
- For $x = (1, \mu) \in \Pi$ and y identified with g, we set $f(x, y) = \sqrt{\mathrm{Var}_{\tilde{P}_\mu}[g(X)]}$
- $e_0 = (1, 0, \ldots, 0)$ corresponds to the constant function 1 in $\mathcal{F}^*$.

Clearly $\langle x, e_0\rangle = \mathbb{E}_{P_\mu}[1] = 1$ for any $x \in \Pi$. Note also that in the definition of $d(P_\gamma \| P_{\gamma'})$ we may extend the supremum from $\mathcal{F}$ to $\mathcal{F}^*$ without change. With these settings, Lemma 15 shows

$$\frac{1}{2}\omega_d(t) \le \delta_0(t) \le \omega_d(t).$$

This completes the proof of (146).

We proceed to proving (147). We start with some preparatory remarks. A simple calculation reveals that

$$d_J(P_{\gamma_1}, P_{\gamma_2}) = \langle \gamma_1 - \gamma_2, \mu_f(\gamma_1) - \mu_f(\gamma_2)\rangle. \tag{154}$$

Similarly, we have the following expression for d:

$$d(\tilde{P}_{\mu'} \| \tilde{P}_\mu) = \sup_{a\in\mathbb{R}^n} \{\langle \mu - \mu', a\rangle : \left\langle \tilde{\Sigma}(\mu)a, a\right\rangle \le 1\} \tag{155}$$

$$= \sqrt{\left\langle \tilde{\Sigma}^{-1}(\mu)\Delta, \Delta\right\rangle}, \quad \Delta = \mu - \mu', \tag{156}$$

where we used the identity

$$\sup_{y\in\mathbb{R}^n} \{\langle y, b\rangle : \langle Ay, y\rangle \le 1\} = \sqrt{\langle A^{-1}b, b\rangle}, \tag{157}$$

which follows from the Cauchy-Schwarz inequality: $\langle y, b\rangle^2 = \langle A^{\frac{1}{2}}y, A^{-\frac{1}{2}}b\rangle^2 \le \left\langle A^{-1}b, b\right\rangle \langle Ay, y\rangle$.

Thus, we get a more explicit formula for ω_d:

$$\omega_d(t) = \sup_{\mu_1,\mu_2\in M_0} \left\{\langle \Delta, h\rangle : \left\langle \tilde{\Sigma}^{-1}(\mu_2)\Delta, \Delta \right\rangle \le t^2\,,\ \Delta = \mu_1 - \mu_2 \right\}\,. \tag{158}$$

This expression clearly shows (148).

We next establish a key inequality connecting the behavior of $\tilde{\Sigma}^{-1}(\lambda\mu_1 + \bar{\lambda}\mu_0)$ with the assumption (74). Consider the following chain of inequalities: for any $a \in \mathbb{R}^n$,

$$\left\langle \tilde{\Sigma}^{-1}(\lambda\mu_1 + \bar{\lambda}\mu_0)a, a\right\rangle^{\frac{1}{2}} = \sup_y \left\{ \langle y, a\rangle : \left\langle \tilde{\Sigma}(\lambda\mu_1 + \bar{\lambda}\mu_0)y, y\right\rangle^{\frac{1}{2}} \le 1 \right\} \tag{159}$$

$$\le \sup_y \left\{ \langle y, a\rangle : \lambda \left\langle \tilde{\Sigma}(\mu_1)y, y\right\rangle^{\frac{1}{2}} + \bar{\lambda}\left\langle \tilde{\Sigma}(\mu_2)y, y\right\rangle^{\frac{1}{2}} \le 1 \right\} \tag{160}$$

$$\le \sup_y \left\{ \langle y, a\rangle : \lambda \left\langle \tilde{\Sigma}(\mu_1)y, y\right\rangle^{\frac{1}{2}} \le 1 \right\} \tag{161}$$

$$= \frac{1}{\lambda}\left\langle \tilde{\Sigma}^{-1}(\mu_1)a, a\right\rangle^{\frac{1}{2}}\,, \tag{162}$$

where in (159) we used (157), in (160) we applied (74), in (161) we omitted the second term, which is non-negative by (70), and in (162) we used (157) again.

Next, we obtain an upper bound on $d_J(P_{\gamma_1}, P_{\gamma_2})$ by continuing from (154). We denote $\mu_i = \mu_f(\gamma_i), i = 1, 2$ and $\Delta = \mu_1 - \mu_2$. Notice

$$\gamma_1 - \gamma_2 = \int_0^1 \dot{\gamma}_\lambda d\lambda\,,$$

where with a slight abuse of notation we define $\gamma_\lambda \triangleq \gamma_r(\lambda\mu_1 + \bar{\lambda}\mu_2)$ and

$$\dot{\gamma}_\lambda = \frac{d}{d\lambda}\gamma_\lambda = \sum_{j=1}^n \frac{\partial\gamma_r}{\partial\mu_j}(\mu_{1,j} - \mu_{2,j}) \overset{(71)}{=} \Sigma(\gamma_\lambda)^{-1}\Delta\,. \tag{163}$$

Then we have

$$d_J(P_{\gamma_1}, P_{\gamma_2}) = \int_0^1 d\lambda \left\langle \tilde{\Sigma}^{-1}(\lambda\mu_1 + \bar{\lambda}\mu_2)\Delta, \Delta\right\rangle \tag{164}$$

$$\le \int_0^{1/2} d\lambda \frac{1}{\bar{\lambda}}\left\langle \tilde{\Sigma}^{-1}(\mu_2)\Delta, \Delta\right\rangle + \int_{1/2}^1 d\lambda \frac{1}{\lambda}\left\langle \tilde{\Sigma}^{-1}(\mu_1)\Delta, \Delta\right\rangle \tag{165}$$

$$= \ln 2 \cdot \left\langle (\tilde{\Sigma}^{-1}(\mu_2) + \tilde{\Sigma}^{-1}(\mu_1))\Delta, \Delta\right\rangle\,, \tag{166}$$

where (164) is from (163), (165) is from (162) and (166) is by computing the integrals.

Finally, consider a pair $\mu_1, \mu_2 \in M_0$ in the optimization (158), i.e. such that

$$\left\langle \tilde{\Sigma}(\mu_2)\Delta, \Delta\right\rangle \le t^2\,, \tag{167}$$

where as usual $\Delta = \mu_1 - \mu_2$. We set

$$\mu_1' = \frac{2}{3}\mu_1 + \frac{1}{3}\mu_2\,, \quad \mu_2' = \frac{1}{3}\mu_1 + \frac{2}{3}\mu_2\,, \tag{168}$$

From convexity we have $\mu_1', \mu_2' \in M_0$ and also

$$\langle \mu_1' - \mu_2', g \rangle = \frac{1}{3} \langle \Delta, g \rangle \,. \tag{169}$$

We claim that for some constant $c' > 0$ we have

$$d_J(\tilde{P}_{\mu_1'}, \tilde{P}_{\mu_2'}) \le c' t^2 \,, \tag{170}$$

which, together with (170) would clearly establish (147). Notice that from (168) and (162) we have

$$\left\langle \tilde{\Sigma}^{-1}(\mu_1')\Delta, \Delta \right\rangle \le 3 \left\langle \tilde{\Sigma}^{-1}(\mu_2)\Delta, \Delta \right\rangle \tag{171}$$

$$\left\langle \tilde{\Sigma}^{-1}(\mu_2')\Delta, \Delta \right\rangle \le \frac{3}{2} \left\langle \tilde{\Sigma}^{-1}(\mu_2)\Delta, \Delta \right\rangle \,. \tag{172}$$

Hence, the left-hand side in (166) is upper-bounded by a constant multiple of $\langle \tilde{\Sigma}(\mu_2)\Delta, \Delta \rangle$, which, in view of (167), shows (170) and, hence, (147).

We complete the proof by showing (150). Notice that $\omega_H(1/\sqrt{n}) \le \omega_H(c_0/\sqrt{\lfloor c_0^2 n \rfloor})$ and then from (76) we have for all $n \ge 2/c_0^2$ and $c_4 = \frac{\sqrt{2}}{c_0}$:

$$\omega_H(1/\sqrt{n}) \le \omega_d(\frac{1}{\sqrt{\lfloor c_0^2 n \rfloor}}) \le \omega_d(c_4/\sqrt{n}) \overset{(a)}{\le} \frac{c_4}{c_2 c_3} \omega_d(c_2 c_3/\sqrt{n}) \overset{(b)}{\le} \frac{c_4}{c_2 c_3} \omega_H(c_3/\sqrt{n}) \,,$$

where (a) follows from (148) with $c = \frac{c_2 c_3}{c_4}$; (b) follows from (147) and (144). This completes the proof of (150). $\square$

Proof of (78) $\iff$ (80). To show this equivalence, first notice the representation

$$C(\gamma + a) - C(\gamma) = \langle \mu_f(\gamma), a \rangle + \int_0^1 (1-s) a^T \Sigma(\gamma + sa) a \, ds \tag{173}$$

since $\nabla C(\gamma) = \mu_f(\gamma)$ and $\operatorname{Hess} C(\gamma) = \Sigma(\gamma)$ as in (69). Thus, from here (80) clearly imply (78) by virtue of

$$a^T \Sigma(\gamma) a = \operatorname{Var}_{P_\gamma}[\langle \phi(X), a \rangle] \,. \tag{174}$$

Conversely, (78) implies that the function

$$\xi \mapsto f_\epsilon(\xi) \triangleq \frac{1}{\epsilon} \{ C(A(\xi) + \epsilon a) - C(A(\xi)) \}$$

is concave for all $\epsilon > 0$. Taking the limit $\epsilon \to 0+$, cf. (173), we conclude that $\xi \mapsto \langle \mu_f(A(\xi)), a \rangle$ is concave for any a (in particular, for $-a$ as well), and hence $\xi \mapsto \mu_f(A(\xi))$ must be affine. Continuing, again from (78) we must have that

$$\xi \mapsto g_\epsilon(\xi) \triangleq \frac{1}{\epsilon} (f_\epsilon(\xi) - \langle \mu_f(A(\xi)), a \rangle)$$

is concave for any $\epsilon \ne 0$. Taking the limit as $\epsilon \to 0$, cf. (173), we conclude that $\xi \mapsto a^T \Sigma(A(\xi)) a$ must be concave, which implies the second claim in (80) in view of (174). $\square$

Acknowledgment

Y. Wu is supported in part by the NSF Grant CCF-1749241, an NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship. The research of Y. Polyanskiy was supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370 and by the MIT-IBM Watson AI Lab.

We thank Prof. A. Rakhlin and Prof. A. Tsybakov for pointing out [DL91] and [JN09], respectively. We thank Prof. A. Juditsky for many discussions and suggestions. We also thank the Associate Editor and the anonymous referees for comments that helped rewrite our paper.

A Classical applications

A.1 Density estimation

As an application of Theorem 1, we consider the classical problem of density estimation under smoothness conditions. For simplicity, we focus on the one-dimensional setting where π is a distribution on $[-1,1]$ with density ρ belonging to the Hölder class $\mathcal{P}(\beta, L)$ (with $0 < \beta \le 1$), namely, $|\rho(x) - \rho(y)| \le L|x-y|^\beta$ for any $x, y \in [-1,1]$. Given n iid observations drawn from ρ, the goal is to estimate the value of the density at point zero $\rho(0)$.

We now verify that this setting fulfills the assumptions of Theorem 1. First, we have $\Theta = \mathcal{X} = [-1,1]$ and P is the identity kernel: $P(x,E) = 1\{x \in E\}$. We take $\mathcal{F} = C[-1,1]$ to be all continuous functions on $[-1,1]$. Note that by identifying a measure π on $[-1,1]$ with its density ρ, we can set $T(\pi) = \rho(0)$ and view Π as a subset of $C[-1,1]$:

$$\Pi = \{\rho \in C[-1,1] : |\rho(x) - \rho(y)| \le L|x-y|^\beta\}.$$

If we endow Π and $C[-1,1]$ with the topology of uniform convergence, then Π becomes a closed convex subset of $C[-1,1]$ and the Arzela-Ascoli theorem [DS58, IV.6.7] implies that Π is in fact compact. Finally, it is clear that $\rho \mapsto \rho(0)$, $\rho \mapsto \int_{[-1,1]} \rho(x) f(x) dx$ and $\rho \mapsto \int_{[-1,1]} \rho(x) f^2(x) dx$ are all continuous on Π for any $f \in C[-1,1]$.

So all assumptions A1-A4 of the theorem are satisfied and the minimax quadratic risk is determined within absolute constant factors by $\delta_{\chi^2}(\frac{1}{\sqrt{n}})^2$. It is well-known that the modulus continuity here satisfies the following:

Lemma 16. *There exist constants c_0, c_1 depending on β and L, such that for all $t > 0$,*

$$c_0 t^{\frac{2\beta}{2\beta+1}} \le \delta_{\chi^2}(t) \le c_1 t^{\frac{2\beta}{2\beta+1}}.$$

Proof. For the upper bound, note that any $f \in \mathcal{P}(\beta, L)$ is everywhere bounded from above by some constant $C = C(\alpha, L)$, thanks to the fact that $f \ge 0$ and $\int f = 1$. Thus, for any $f, g \in \mathcal{P}(\beta, L)$ such that $|f(0) - g(0)| = \epsilon$ and $\chi^2(f\|g) \le t^2$, we have $\|f-g\|_2^2 \le Ct^2$. Let $p = |f-g|$. Then $p \ge 0$ and p is $(\beta, 2L)$-Hölder continuous. For sufficiently small ϵ, define $h : [-1,1] \to \mathbb{R}_+$ by $h(x) = \max\{\epsilon - 2L|x|^\beta, 0\}$. Then $p \ge h$ on $[-1,1]$ pointwise and hence

$$Ct^2 \ge \|f-g\|_2^2 \ge \|h\|_2^2 = C'\epsilon^{2+\frac{1}{\beta}}$$

for some constant C' depending on (β, L). This shows the upper bound. The lower bound follows from choosing f to be the uniform density, and $g(x) = f(x) + c|x|^\beta \operatorname{sign}(x) \mathbf{1}_{\{|x|^\beta \le \epsilon\}}$, for some small constant c depending on (β, L) and $\epsilon = t^{\frac{2\beta}{2\beta+1}}$. □

Applying Theorem 1, we recover the classical result:

$$\inf_{\hat{T}} \sup_{\rho \in \mathcal{P}(\beta, L)} \mathbb{E}_{X_1,\ldots,X_n \overset{\text{iid}}{\sim} f} |\hat{T}(X_1, \ldots, X_n) - \rho(0)|^2 \asymp n^{-\frac{2\beta}{2\beta+1}}. \tag{175}$$

Furthermore, Theorem 1 ensures that empirical-mean estimators (5) of the form $\hat{T} = \frac{1}{n}\sum_{i=1}^n g(X_i)$ are rate optimal for some appropriately chosen function g. Indeed, kernel density estimates are of this form, which achieve the minimax rate for suitably chosen kernel and bandwidth (cf. e.g. [Tsy09, Section 1.2]).

A.2 White Gaussian noise model

In this section we revisit the Gaussian white noise model and re-derive the classical result of [IH84, Don94] on the rate optimality (within constant factors) of linear estimators from convex duality. Let

$$dX_t = f(t)dt + \sigma dB_t, \quad t \in [0,1], \tag{176}$$

where the unknown function f belong to some *convex* set $\mathcal{F}$. Given $X = \{X_t : t \in [0,1]\}$, the goal is to estimate some affine functional $T(f)$ (such as $T(f) = f(1/2)$). Define the minimax risk as

$$R^*(\sigma) \triangleq \inf_{\hat{T}} \sup_{f \in \mathcal{F}} \mathbb{E}_f[(\hat{T}(X) - T(f))^2].$$

Although this is a special case of the n-sample exponential family model considered in Section 4 (with $\sigma = \frac{1}{\sqrt{n}}$), Theorem 12 proved for finite dimensions cannot be directly applied. Nevertheless, due to the simple structure of the Gaussian model, Le Cam's lower bound can be dualized explicitly, leading to the rate-optimality of linear estimators. Next we carry out this calculation as a self-contained example.

Consider a linear estimator of the form

$$\hat{T} = \int_0^1 g(t) dX_t,$$

where g is some continuous compactly-supported function to be optimized. (Denote all such functions by C_c.) Then the bias and variance are given respectively by

$$\begin{aligned} \mathbb{E}\hat{T} - T &= \langle f, g \rangle - T(f) \\ \mathrm{Var}(\hat{T}) &= \sigma^2 \|g\|_2^2. \end{aligned}$$

To bound the bias, note that, trivially,

$$\inf_{f \in \mathcal{F}} \langle f, g \rangle - T(f) \le \mathbb{E}\hat{T} - T \le \sup_{f \in \mathcal{F}} \langle f, g \rangle - T(f).$$

Without loss of generality, we can assume that $\sup_{f \in \mathcal{F}} \langle f, g \rangle - T(f) \ge 0 \ge \inf_{f \in \mathcal{F}} \langle f, g \rangle - T(f)$.[13] Therefore, we have

$$|\mathbb{E}\hat{T} - T| \le \sup_{f \in \mathcal{F}} \langle f, g \rangle - T(f) + \sup_{f \in \mathcal{F}} T(f) - \langle f, g \rangle = \sup_{f, f' \in \mathcal{F}} \langle f - f', g \rangle + T(f') - T(f).$$

[13] Suppose $\sup_{f \in \mathcal{F}} \langle f, g - h \rangle = \epsilon < 0$, i.e., the estimator is always negatively biased, then replacing g by $g - \epsilon$ improves the bias and retains the same variance.

Optimizing the bias-variance tradeoff over g leads to the following convex optimization problem:

$$\begin{aligned}
\sqrt{R^*(\sigma)} &\leq \inf_{g\in C_c} \sup_{f,f'\in\mathcal{F}} \langle f-f', g\rangle + T(f') - T(f) + \sigma\|g\|_2 \\
&= \inf_{g\in C_c} \sup_{f,f'\in\mathcal{F}, \|z\|_2\leq 1} \langle f-f', g\rangle + T(f') - T(f) + \sigma\langle g, z\rangle \\
&\stackrel{(a)}{=} \sup_{f,f'\in\mathcal{F}, \|z\|_2\leq 1} \left\{ T(f') - T(f) + \inf_{g\in C_c} \langle f-f'+\sigma z, g\rangle \right\} \\
&\stackrel{(b)}{=} \sup_{f,f'\in\mathcal{F}, \|f-f'\|_2\leq \sigma} T(f') - T(f) \\
&\stackrel{(c)}{\leq} C\sqrt{R^*(\sigma)},
\end{aligned}$$

where (a) follows from the minimax theorem (see, e.g., Theorem 5 in Section 2); (b) is simply because $\inf_{g\in C_c}\langle f, g\rangle = -\infty$ if $f\neq 0$ and 0 if $f=0$; finally, (c) follows from Le Cam's two-point lower bound since the KL divergence in the white noise model is given by

$$D(P_f\|P_{f'}) = \frac{1}{2\sigma^2}\|f-f'\|_2^2, \tag{177}$$

where P_f denotes the law of $\{X_t : t\in[0,1]\}$ as in (176), and C is an absolute constant. Thus we have shown that

$$\frac{\omega(\sigma)}{C} \leq \sqrt{R^*(\sigma)} \leq \omega(\sigma) \tag{178}$$

where $\omega(\sigma) \triangleq \sup_{f,f'\in\mathcal{F}}\{T(f') - T(f) : \|f-f'\|_2\leq\sigma\}$ is the modulus of continuity.

B Proof of technical results

Proof of Lemma 6. In [PSW17, Proposition 9] it is shown for any $d\in\mathbb{N}\cup\{\infty\}$,

$$\delta_{\mathrm{TV}}(t,d) \leq t^{\min(1,\frac{1-\epsilon}{\epsilon})}. \tag{179}$$

where $\delta_{\mathrm{TV}}(t,d)$ is defined for the same problem as δ_{χ^2} but with TV-distance in place of χ^2, cf. (9). From the general relation $\delta_{\chi^2}\leq\delta_{\mathrm{TV}}$ in (15) we get (37). Furthermore, due to (16) and (15), for $\epsilon\leq\frac{1}{2}$ we conclude

$$\frac{\min(t,1)}{4} \leq \delta_{\chi^2}(t,d) \leq t.$$

Next, we consider the case of $\epsilon > 1/2$. The following was shown in [PSW17, Lemma 12]: For every $\delta < \frac{1}{2e}$ and $d\geq\frac{2\epsilon}{1-\epsilon}\ln^2\frac{1}{\delta}$ there exists a pair of probability distributions π and π' on $\{0,\ldots,d\}$ such that $|\pi(0)-\pi'(0)|\geq\delta$ and

$$H^2(\pi P, \pi' P) \leq 36\left(e\delta\ln\frac{1}{\delta}\right)^{\frac{2\epsilon}{1-\epsilon}}. \tag{180}$$

Setting the RHS to t^2, we conclude that there exist $t_0 = t_0(\epsilon)$ and $C = C(\epsilon)$ such that for all $t\leq t_0$ and $d\geq C\ln^2\frac{1}{t}$, we have

$$\delta_{H^2}(t,d) \geq C\left(\frac{t}{\ln\frac{1}{t}}\right)^{\frac{1-\epsilon}{\epsilon}}.$$

This implies the desired (38) in view of the general inequality $\delta_{\chi^2}\geq\frac{1}{2}\delta_{H^2}$ in (15). $\square$

Proof of Proposition 8. We aim to apply Theorem 1. Assumptions A1 and A2 are verified. For A3 we take $\mathcal{F}$ to be the set of all continuous functions on $\mathcal{X}$ (cf. the second remark after Theorem 1). For A4 we endow Π with the weak topology. Since Π is a set of probability measures on the compact set $[0,1]$, it is tight and hence weakly compact, establishing A4a. For A4b we always have that $\pi \mapsto T_m(\pi)$ is weakly continuous, while for $\pi \mapsto T_c(\pi)$ weak continuity is implied by the assumption (indeed, if $\pi_n \stackrel{w}{\to} \pi$ then $F_{\pi_n}(s_0) \to F_\pi(s_0)$ due to assumption on s_0 being a point of continuity of F_π). To complete the verification of A4b, we need to verify that for any continuous ϕ on $\mathcal{X}$ the functional $\pi P \phi$ is weakly continuous. For example, consider the case of $i = 1$, in which case we have $\phi : [0,1] \times \{0,1\} \to \mathbb{R}$ and we can represent

$$\pi P \phi = \int_{[0,1]} \pi(d\theta) f(\theta), \quad f(\theta) \triangleq \int_{[0,1]} d a g_1(a) \phi(a,1) 1\{\theta \le a\} + \phi(a,0) 1\{\theta > a\} \tag{181}$$

We claim that f is continuous. Indeed, if $\theta_n \to \theta$ then $1\{\theta_n > a\} \to 1\{\theta > a\}$ for almost every $a \in [0,1]$. Hence, from the dominated convergence theorem we also have $f(\theta_n) \to f(\theta)$. Thus, the functional in (181) is weakly continuous as integral of a continuous function f. $\square$

Proof of (57). Let $H_k(x)$ denote the degree-k Hermite polynomial and note the fact that for $X \sim N(a,1)$, we have $\mathbb{E}[H_k(X)] = a^k$ and $\mathrm{Var}(H_k(X)) = k! \sum_{j=0}^{k-1} \binom{k}{j} \frac{a^{2j}}{j!}$. Thus $\mathrm{Var}(H_k(X)) \le k! 2^k$ provided $|a| \le 1$. Using the variational representation of the χ^2-divergence (32), for any feasible solution π, π' of (56), we have $|m_k(\pi) - m_k(\pi')| \le \sqrt{k! 2^k t}$, where $m_k(\pi) = \int \theta^k \pi(d\theta)$ denotes the kth moment of π. By existing results in approximation theory (see [CL11]), there exists a degree-k polynomial $p(x) = \sum_{i=0}^k a_i x^i$ and a constant C, such that $|a_i| \le C^k$ and $\sup_{|a|\le 1} ||a| - p(a)| \le \frac{C}{k}$. Therefore by the triangle inequality, we have $|\int |\theta| \pi'(d\theta) - \int |\theta| \pi'(d\theta)| \le \frac{C}{k} + \sqrt{t k! C^k}$. Choosing $k = c \frac{\log \frac{1}{t}}{\log \log \frac{1}{t}}$ for some small constant c proves the upper bound of (57).

To show the lower bound part, by the duality between best polynomial approximation and moment matching (see e.g. [WY16, Appendix E]), there exist $\pi, \pi' \in \mathcal{P}([-1,1])$ such that $m_i(\pi) = m_i(\pi')$ for $i = 1, \ldots, k$, and $\int |\theta| \pi'(d\theta) - \int |\theta| \pi'(d\theta) = 2 \inf_{\deg(p)=k} \sup_{|a|\le 1} ||a| - p(a)| \ge \frac{c}{k}$, where the last inequality is well-known in the approximation theory literature [CL11]. Furthermore, matching first k moments implies that the corresponding Gaussian mixture are close in χ^2-divergence [CL11]: $\chi^2(\pi' * N(0,1) \| \pi * N(0,1)) \le \frac{C^k}{k!}$. Choosing $k = c \frac{\log \frac{1}{t}}{\log \log \frac{1}{t}}$ for some large constant c proves the desired lower bound. $\square$

C Risks for Fisher's species problem with or without Poissonization

For fixed sample sizes (n,m), define the minimax quadratic risk for estimating $U = U_{n,m}$ as

$$\tilde{R}^*(n,m) \triangleq \inf_{\hat{U}} \sup_P \mathbb{E}_P[(\hat{U} - U)^2].$$

and $R^*(n,m)$ for the Poissonized model with sample sizes (N,M) distributed independently as $\mathrm{Poi}(n)$ and $\mathrm{Poi}(m)$). Also denote their normalized version by $\tilde{\mathcal{E}}_n(r) = \tilde{R}^*(n,m)/m^2$ and $\mathcal{E}_n(r) = R^*(n,m)/m^2$ (with $r = m/n$), the latter of which is addressed by Theorem 11. Nevertheless, the next result shows that $\tilde{\mathcal{E}}_n(r)$ satisfies the same upper and lower bounds Theorem 11 up to an additional $O\left(\frac{\log n}{n}\right)$ term. The proof of this lemma is standard (cf. [WY16, Appendix A]).

Lemma 17. *Let $r = m/n$ be a constant. Let $\alpha = 1/\log n$. Then for large n,*

$$\frac{1}{2}\tilde{R}^*(n(1+\alpha), m) - O(n\log n) \le R^*(n,m) \le 2\tilde{R}^*(n(1-\alpha), m) + O(n\log n). \tag{182}$$

Proof. Note the following facts about the risk $R^*(n,m)$:

1. $n \mapsto R^*(n,m)$ is non-increasing.
2. $0 \le R^*(n,m) \le R^*(0,m) \le m^2$.
3. $|\sqrt{R^*(n,m)} - \sqrt{R^*(n,m')}| \le |m-m'|$, due to the fact that $|U_{n,m} - U_{n,m'}| \le |m-m'|$.

For the left inequality, let $N \sim \mathrm{Poi}((1+\alpha)n)$ and $M \sim \mathrm{Poi}(m)$. Set $\Delta = C\sqrt{n\log n}$ for some large constant C. Using the Chernoff bound for Poisson, we have

$$\begin{aligned}\tilde{R}^*((1+\alpha)n, m) &\le \mathbb{E}[R^*(N,M)] \le \mathbb{E}[R^*(n,M)] + O(m^2)\mathbb{P}[N<n]\\ &\le \sum_{|m'-m|\le\Delta} R^*(n,m')\mathbb{P}\left[M=m'\right] + O(m^2)(\mathbb{P}[N>n] + \mathbb{P}[|M-m|>\Delta])\\ &\le (\sqrt{R^*(n,m)}+\Delta)^2 + O(m^2(e^{-\Omega(\alpha^2 n)} + e^{-\Omega(\Delta^2/n)})).\end{aligned}$$

For the right inequality, it suffices to consider the Bayes risk. Note that for any fixed prior π, the Bayes risks with fixed or Poissonized sample size, denoted by $\tilde{R}^*_\pi(n,m)$ and $R^*_\pi(n,m)$, is related by the identity $R^*_\pi(n,m) = \mathbb{E}[R^*_\pi(\mathrm{Poi}(n), \mathrm{Poi}(m))]$. Let $N \sim \mathrm{Poi}((1-\alpha)n)$ and $M \sim \mathrm{Poi}(m))$ be independent. Then

$$\begin{aligned}R^*_\pi((1-\alpha)n, m) &\ge \mathbb{E}[R^*_\pi(N,M)\mathbf{1}_{\{N'\le n, |M-m|\le\Delta\}}]\\ &\ge \mathbb{E}[R^*_\pi(n,M)\mathbf{1}_{\{|M-m|\le\Delta\}}]\mathbb{P}\left[N'\le n\right]\\ &\ge \max\{\sqrt{R^*_\pi(n,m)}-\Delta, 0\}^2(1-e^{-\Omega(\alpha^2 n)} - e^{-\Omega(\Delta^2/n)})).\end{aligned}$$

This completes the proof. ☐

References

[BF93] John Bunge and M Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[BL96] L. D. Brown and M. G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics*, 24:2524–2535, 1996.

[Bro86] L. D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. In S. S. Gupta, editor, *Lecture Notes-Monograph Series*, volume 9. Institute of Mathematical Statistics, Hayward, CA, 1986.

[BZ86] Jonathan M Borwein and D Zhuang. On Fan's minimax theorem. *Mathematical Programming*, 34(2):232–234, 1986.

[CCMN00] Moses Charikar, Surajit Chaudhuri, Rajeev Motwani, and Vivek Narasayya. Towards estimation error guarantees for distinct values. In *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 268–279. ACM, 2000.

[CCT17] Olivier Collier, Laëtitia Comminges, and Alexandre B Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958, 2017.

[CCTV16] Olivier Collier, Laëtitia Comminges, Alexandre B Tsybakov, and Nicolas Verzelen. Optimal adaptive estimation of linear functionals under sparsity. *arXiv preprint arXiv:1611.09744*, 2016.

[CL92] Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, 87(417):210–217, 1992.

[CL11] T. T. Cai and M. G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[DL91] David L. Donoho and Richard C. Liu. Geometrizing rates of convergence, II. *The Annals of Statistics*, 19:668–701, 1991.

[Don94] David L Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22:238–270, 1994.

[DRWY12] Zeev Dvir, Anup Rao, Avi Wigderson, and Amir Yehudayoff. Restriction access. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 19–33. ACM, 2012.

[DS58] N. Dunford and J.T. Schwartz. *Linear Operators: General theory*, volume 1. Interscience Publishers, New York, 1958.

[ET76] B. Efron and R. Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[Fan53] Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences*, 39(1):42–47, 1953.

[FCW43] Ronald Aylmer Fisher, A Steven Corbet, and Carrington B Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

[GJ14] Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*, volume 38. Cambridge University Press, 2014.

[GL95] R. D. Gill and B. Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1–2):59–79, 1995.

[GS02] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[GT56] I.J. Good and G.H. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

[GW92] Piet Groeneboom and Jon A Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media, 1992.

[HW97] Jian Huang and Jon A Wellner. Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 123–169. Springer, 1997.

[IH84] I.A. Ibragimov and R.Z. Has'minskii. On the nonparametric estimation of a value of a linear functional in the Gaussian white noise. *Theory of Probability & Its Applications*, 29(1):19–32, 1984.

[ILLL09] Iuliana Ionita-Laza, Christoph Lange, and Nan M Laird. Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, 106(13):5008–5013, 2009.

[JN09] Anatoli B Juditsky and Arkadi S Nemirovski. Nonparametric estimation by convex programming. *The Annals of Statistics*, 37(5A):2278–2300, 2009.

[JN20] Anatoli Juditsky and Arkadi Nemirovski. *Statistical Inference via Convex Optimization*, volume 69. Princeton University Press, 2020.

[JPW20] Soham Jana, Yury Polyanskiy, and Yihong Wu. Extrapolating the profile of a finite population. In *Proceedings of Conference on Learning Theory (COLT)*, Jul 2020. arXiv:2005.10561.

[KM58] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[Kön68] Heinz König. Über das von Neumannsche minimax-theorem. *Archiv der Mathematik*, 19(5):482–487, 1968.

[LC73] L. Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38 – 53, 1973.

[LC86] Lucien Le Cam. *Asymptotic methods in statistical decision theory.* Springer-Verlag, New York, NY, 1986.

[LNS99] Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the L_r norm of a regression function. *Probability Theory and Related Fields*, 113(2):221–253, 1999.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis.* Cambridge University Press, 2005.

[OSW16] Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences (PNAS)*, 113(47):13283–13288, 2016.

[PSW17] Y. Polyanskiy, A. T. Suresh, and Y. Wu. Sample complexity of population recovery. In *Proceedings of Conference on Learning Theory (COLT)*, Amsterdam, Netherland, Jul 2017. arXiv:1702.05574v3.

[Rob51] Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California, 1951.

[RRSS09] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

[Rud87] Walter Rudin. *Real and complex analysis*. McGraw-Hill, 1987.

[Sim11] Barry Simon. *Convexity: An analytic viewpoint*. Cambridge University Press, 2011.

[Str85] Helmut Strasser. *Mathematical theory of statistics: Statistical experiments and asymptotic decision theory*. Walter de Gruyter, Berlin, Germany, 1985.

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

[Val11] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

[Val12] Gregory Valiant. *Algorithmic Approaches to Statistical Questions*. PhD thesis, EECS Department, University of California, Berkeley, Sep 2012.

[VV11] Gregory Valiant and Paul Valiant. The power of linear estimators. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 403–412. IEEE, 2011.

[Wol57] J. Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, pages 75–88, 1957.

[WY12] Avi Wigderson and Amir Yehudayoff. Population recovery and partial identification. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 390–399. IEEE, 2012.

[WY16] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[WY18] Yihong Wu and Pengkun Yang. Sample complexity of the distinct element problem. *Mathematical Statistics and Learning*, 1(1):37–72, 2018.

[Yu97] Bin Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435, 1997.

[Zha97] Cun-Hui Zhang. Empirical bayes and compound estimation of normal means. *Statistica Sinica*, 7(1):181–193, 1997.

[Zha03] Cun-Hui Zhang. Compound decision theory and empirical Bayes methods. *The Annals of Statistics*, 31(2):379–390, 2003.