*Submitted to the Annals of Statistics*

# DUALIZING LE CAM'S METHOD FOR FUNCTIONAL ESTIMATION I: GENERAL THEORY

BY YURY POLYANSKIY* AND YIHONG WU†

*Massachusetts Institute of Technology* and Yale University†*

*Department of Statistics and Data Science*
*Yale University*
*New Haven, CT 06511*
*yihong.wu@yale.edu*

*Department of Electrical Engineering and Computer Science*
*Massachusetts Institute of Technology*
*Cambridge, MA 02139*
*yp@mit.edu*

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 minimizing an upper-bound on the quadratic risk over a family of estimators. Since by the minimax theorem two problems have the same value, this value also characterizes (up to a universal factor) the estimation rate. For estimating linear functionals of a distribution our work strengthens prior results of Donoho-Liu [9] (for quadratic loss) by dropping the Hölderian assumption on the modulus of continuity. For exponential families our results extend those of Juditsky-Nemirovski [23] 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.

An application of our methodology to the area of "estimating the unseens" is provided in the companion paper [32], resolving the optimal rates (within logarithmic factors) of the distinct elements problem and Fisher's species problem.

**1. Introduction.** *Le Cam's method* (or the two-point method) is a commonly used tool for obtaining statistical lower bound [25, 44]. 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 $\Theta$ and $\mathcal{X}$ be measurable spaces and $P$ a transition probability kernel from $\Theta$ to $\mathcal{X}$. Then $\{P_\theta = P(\cdot|\theta) \colon \theta \in \Theta\}$ is a parametric family of distributions on $\mathcal{X}$. Let $\Pi$ be a given subset of $\mathcal{P}(\Theta)$, the set of all probability measures on $\Theta$. Let $T(\pi)$ be a real-valued linear functional

*MSC2020 subject classifications:* Primary 62G05; secondary 62C20,62G86,62G07.
*Keywords and phrases:* Functional estimation, minimax optimality, convex optimization and duality, modulus of continuity, linear programming..

of $\pi \in \Pi$, i.e., for any $\pi, \pi' \in \Pi$ and $0 < \alpha < 1$, $T((1-\alpha)\pi' + \alpha\pi) = (1-\alpha)T(\pi') + \alpha T(\pi)$. 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 [35, 45, 46].

- $\theta_i \overset{\text{i.i.d.}}{\sim} \pi$ for some $\pi \in \Pi$. In this case, $X_i \overset{\text{i.i.d.}}{\sim} \pi P$, where $\pi P \triangleq \int_\Theta P_\theta \pi(d\theta)$ denotes the mixture distribution induced by the mixing distribution (prior) $\pi$. Given $X_1, \ldots, X_n$, the goal is to estimate the functional $T(\pi)$.
- $\theta_i$'s are deterministic whose empirical distribution $\pi_{\boldsymbol{\theta}} \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$ belongs to $\Pi$. Given $X_1, \ldots, X_n$, the goal is to estimate the functional $T(\pi_{\boldsymbol{\theta}})$. In the special case of linear functional $T(\pi) = \int h(\theta)\pi(d\theta)$ for some $h : \Theta \to \mathbb{R}$, $T(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1} h(\theta_i)$ 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 $\Pi$, for linear functional $T$, Le Cam's lower bound is tight up to universal constant factors for both iid and deterministic settings. More precisely, we have[1]

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

where $\chi^2(\cdot\|\cdot)$ denotes the $\chi^2$-divergence. In the iid setting, this result strengthens the celebrated result of Donoho-Liu [9] 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 $\chi^2$-divergence in (4) is replaced by the squared Hellinger distance, extending the result of Juditsky-Nemirovski [23] 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 $\chi^2$-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 $X_i \overset{\text{i.i.d.}}{\sim} \pi P$, the lower bound is a straightforward application of Le Cam's two point method, since $\chi^2(\pi P \| \pi' P) \le \frac{1}{n}$ 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 [37, Sec. 2.7.4]), where we consider $\theta_i$'s drawn from the product distribution $\pi^{\otimes n}$ or $\pi'^{\otimes n}$ with appropriate truncation.

[1] Throughout the paper we use standard asymptotic notation: for two positive sequences $\{x_n\}$ and $\{y_n\}$, we write $x_n = O(y_n)$ or $x_n \lesssim y_n$, if $x_n \le C y_n$ for an absolute constant $C$ and for all $n$; $x_n = \Omega(y_n)$ or $x_n \gtrsim y_n$, if $y_n = O(x_n)$; $x_n = \Theta(y_n)$ or $x_n \asymp y_n$, if $x_n = O(y_n)$ and $x_n = \Omega(y_n)$; $x_n = o(y_n)$ if $x_n/y_n \to 0$ as $n \to \infty$.

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 rooted in the classical minimax theorem in decision theory, which states that, under regularity assumptions (cf. e.g. [36, Theorem 46.5]) the minimax risk and the least favorable Bayes risk for estimating a real-valued functional (say, $S(\theta)$) coincide, namely

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

where the supremum on the right is taken over all priors on $\Theta$. This can also be interpreted from the duality perspective,[2] 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.

In addition to various examples in the present paper, as a chief application of our methodology, in the companion paper [32] we consider three representative problems in the area of "estimating the unseens", namely, *population recovery* [11, 39, 30, 7], *distinct elements problem* [34, 38, 42], and *Fisher's species problem* [14, 12, 29]. In addition to recovering the prior result of [30] on the optimal rate of population recovery, we resolve the optimal rates (up to logarithmic factors) for the distinct elements and the species problem, characterizing, in particular, the "elbow effect", that is, the phase transition from parametric to nonparametric rates. These results are outside the reach of existing approaches [42, 29], accomplished by first applying the abstract results in Sections 2 and 3 to reduce to computing the maximization in (4), which is then evaluated using tools from complex analysis.

We emphasize that the main result of this paper is a method for *determining the minimax rate* by means of convex duality *without* demonstrating an explicit choice of the optimal estimator. Conceptually this is a departure from existing approaches relying on ad hoc construction of estimators (e.g., those based on polynomial approximation for distinct elements [42] and those based on smoothing for the species problem [29], which do not attain the optimal rates). Nevertheless, our result shows that the minimax rate in both iid and deterministic settings can be achieved by the estimator (5) parametrized by $g$, the optimal $g$ corresponds to the optimizer of certain convex optimization problem (cf. (24)), which can be solved efficiently in the case of finite $\Theta$ and $\mathcal{X}$ (e.g. the population recovery problem). For certain continuous models, this infinite-dimensional problem can be effectively discretized leading to computationally efficient construction of optimal estimators. For example, for the distinct

[2] This follows from standard arguments in optimization by rewriting the left-hand side as $\inf_{\hat{S}}\{t : \mathbb{E}_\theta[(\hat{S} - S(\theta))^2] \le t, \forall\theta\in\Theta\}$ and the Lagrange multipliers associated with the constraints indexed by $\theta$ correspond to priors; see, e.g., [33, Sec. 28.3.3] for details. When both $X$ and $\theta$ are finitely-valued, (6) is simply the duality of linear programming (LP).

elements and species problems considered in the companion paper [32], the estimators can be constructed in polynomial time as solutions to certain linear programs.

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 follow-up paper [21] 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* [40] in the duality framework.

1.1. *Related work.* A celebrated result of Donoho-Liu [9] 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 $\chi^2$-divergence.[3] 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 [9] 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 functions, 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 [23] (cf. also the recent monograph [22]), where the main technology was also convex optimization and the minimax theorem. As opposed to the squared loss, they considered the $\epsilon$-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 $\epsilon$-quantile risk is determined within absolute constant factors by the Hellinger modulus of continuity provided that $\epsilon$ 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 $\epsilon$-quantile loss potentially depends on $\epsilon$. On the other hand, results on $\epsilon$-quantile loss can be obtained from the quadratic risk by sampling splitting and applying the median (although the more direct argument in [23] 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 [23] was to obtain rate-optimal estimators by means of convex programming, in this paper and similar to the program in [9], the characterization of the minimax risk by convex optimization is mostly used as a mathematical tool for determining the minimax rates, although it can also lead to efficient construction of estimations for certain specific problems.

1.2. *Organization.* The rest of the paper is organized as follows. Section 2 presents the main result for the iid setting and revisits the problem of interval censoring [17] in Section 2.2 as an application. Section 3 addresses the deterministic setting of estimating separable functionals, where we consider the problem of estimating the $\ell_1$-norm of the Gaussian mean [27, 4] as an example. Finally, in Section 4 we extend the result to exponential families. Due to space constraints, the proof of the main result of Section 4 is provided in Section 5 of the supplementary material [31]. Further technical proofs are collected in Section 6 of [31]. As further examples, in Section 7 of [31] we discuss applications to classical problems of density estimation and the Gaussian white noise model, the latter constituting a simple example where the dualization can be carried out explicitly.

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

**2. Iid setting.** Recall that in the iid setting of (1), the sample consists of $X_1, \dots, X_n \overset{iid}{\sim} \pi P$. When $P$ is the identity kernel, $X_i$'s are directly drawn from $\pi$; 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 $\pi P$:

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

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

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

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 $\nu$ being any dominating measure s.t. $F \ll \nu$ and $G \ll \nu$, e.g. $\nu = F + G$). Finally, the $\chi^2$-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 $\mathcal{P}$. In the present paper we will be mainly working with $\delta_{\chi^2}$ and $\delta_{H^2}$ (which are equivalent up to constants as we show next in Proposition 3); the modulus $\delta_{\text{TV}}$ plays an important role in the companion paper [32] thanks to the amenability of TV distance to complex-analytic techniques.

Recall the following notations: Given a kernel $P$ from $\Theta$ to $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$, $Pf$ is a function on $\Theta$ given by $(Pf)(\theta) = \int_{\mathcal{X}} P(dx|\theta) f(x)$. Thus for any prior $\pi$ on $\Theta$, $\pi P f = \int_\Theta \int_{\mathcal{X}} P(dx|\theta) \pi(d\theta) f(x)$.

The main result of this section 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 linear;*
A2 *The set $\Pi$ 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 $\Pi$ such that:*
  A4a *It is coarse enough that $\Pi$ 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*

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

*where the upper bound can be approached arbitrarily well by estimators of the form $\hat{T} = \frac{1}{n} \sum_{i=1}^n g(X_i)$ for $g \in \mathcal{F}$.*

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

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

*where the $\epsilon$-risk (confidence interval) is defined as*

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

Some remarks are in order:

1. If $\Theta$ 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 $\nu$ is regular [10, IV.6.2] and the set $\mathcal{F}$ of all bounded continuous functions is dense in $L_2(\mathcal{X}, \nu)$, cf. [10, IV.8.19]. Other possible choices of $\mathcal{F}$ are Lipschitz functions, polynomials, trigonometric polynomials or sums of exponentials.
3. The continuity of $\pi P f$ under the weak topology on $\Pi$ 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 linear in $\pi$. Assume that $\Pi$ is convex. Then*

1. *(Concavity) $\delta_{H^2}(\sqrt{\cdot})$, $\delta_{\mathrm{TV}}(\cdot)$ and $\delta_{\chi^2}(\sqrt{\cdot})$ are concave.*
2. *(Subadditivity) For any $c \in [0,1]$ and $t \geq 0$ we have:*

$$\delta_{\mathrm{TV}}(ct) \geq c\delta_{\mathrm{TV}}(t) \tag{12}$$

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

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

3. *(Comparison of various $\delta$'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. [37, 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

$$H^2(P,Q) \leq \chi^2(P\|Q)\,. \tag{18}$$

Together (17) and (18) establish all inequalities in (15) except the left-most one. For the latter we recall from [26, 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 $(\pi, \pi')$ 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 $\pi_1, \pi_0$ such that $T(\pi_0) - T(\pi_1) = \Delta > 0$. From the data-processing inequality we have for all $0 < \lambda < 1$:

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

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

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

Optimizing over $\pi_0, \pi_1$ yields (16).

□

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

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

Using Brown-Low's two-point lower bound [3] and optimizing over the pair $\pi, \pi'$, we have

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

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:

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

Taking the worst-case $\pi$ and optimizing over $g$ we get

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

where

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

The proof is completed by applying the next proposition. □

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

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

*Furthermore, the supremum over $\pi, \pi'$ in the definition of $\delta_{\chi^2}$ 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 Ky Fan's minimax theorem [13]. The following version is from [1, Theorem A].

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 $\max_{x\in X} \inf_{y\in Y} f(x,y)$ exists and equals*

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

We recall the definition that a 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 $\delta_{\mathrm{bv}}(t)$. The function

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

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

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

Indeed, if $T(\pi) - \pi P g < 0$, 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 $U \times \mathcal{F}$:

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

We claim it is concave-convex-like. Convexity in $g$ is easy: the first two terms are affine in $g$ and $g \mapsto \sqrt{\mathrm{Var}_\mu[g]}$ is convex.[4]

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

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

So for $u = (\pi, \pi', \xi)$ we have

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

It is clear that $a(\pi)$ is affine in $\pi$, whereas $b(\pi)$ is concave.[5]

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 $\pi$. 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

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

[4]Indeed, defining the norm $\|g\|_{L_2(\mu)} \triangleq \sqrt{\int g^2 d\mu}$, we have for any $a_0, a_1 > 0$ and $g_0, g_1$, $\sqrt{\mathrm{Var}_\mu[a_0 g_0 + a_1 g_1]} = \|a_0\bar{g}_0 + a_1\bar{g}_1\|_{L_2(\mu)} \le a_0\|\bar{g}_0\|_{L_2(\mu)} + a_1\|\bar{g}_1\|_{L_2(\mu)} = a_0\sqrt{\mathrm{Var}_\mu[g_0]} + a_1\sqrt{\mathrm{Var}_\mu[g_1]}$, where $\bar{g}_i \triangleq g_i - \int g_i d\mu$ for $i = 0, 1$.

[5]To see this, note that $\sqrt{\cdot}$ is a concave and increasing scalar function, whereas $\mathrm{Var}_\mu[g] = \mu(g^2) - (\mu g)^2$ is concave in $\mu$.

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 the 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

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

(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

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

We claim that

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

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 $\xi \ne 1$, the first term in (30) can be driven to $-\infty$, while keeping the second term zero, by taking $g \equiv c$ and $c \to \pm\infty$. So fix $\xi = 1$. Recall a variational characterization of the $\chi^2$-divergence [33, Section 7.13]:[6]

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

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, $\chi^2(\pi' P\|\pi P) \le t^2$. Again from (32), by scaling 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 $g = 0$. □

[6] 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 $n \to \infty$.

PROOF OF COROLLARY 2. Consider two distributions $\pi$ and $\pi'$ such that $H^2((\pi P)^{\otimes n}, (\pi' P)^{\otimes n}) = 2 - 2\beta$ then from [37, Theorem 2.2] we have that

$$\rho^*_{\text{iid}}(n, \epsilon) \geq \frac{1}{2}|T(\pi) - T(\pi')| \tag{33}$$

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

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

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

$$\rho^*_{\text{iid}}(n, \epsilon) \geq \frac{1}{2}\delta_{H^2}\left(\sqrt{\frac{1}{4n}\ln\frac{1}{\epsilon}}\right) \geq \frac{1}{8}\delta_{\chi^2}\left(\sqrt{\frac{1}{4n}\ln\frac{1}{\epsilon}}\right) ,$$

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:

$$\rho^*_{\text{iid}}(2nL, \epsilon) \leq 2\sqrt{R^*_{\text{iid}}(n)} , \qquad \forall L \geq 8 \ln \frac{2}{\epsilon} . \tag{34}$$

This follows from the standard "median trick" dating back to the 1980s (see discussion in [28] and [43, p. 1995]): 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 $\rho = 2\sqrt{R^*_{\text{iid}}(n)}$. Then from Chebyshev's inequality we have for each $1 \leq j \leq 2L$

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

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

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

From the last statement, we conclude that (34) must hold. □

2.1. *Comparison to Donoho-Liu [9].* Theorem 1 is very similar to a celebrated result of Donoho-Liu [9], who showed that in the same setting,[7] as $n \to \infty$, one has

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

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 [9] are: A1, A2 (from Theorem 1), boundedness $\sup_{\pi \in \Pi} |T(\pi)| < \infty$ and *Hölderianity* of $\delta_{H^2}$:

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

[7] The results of Donoho and Liu [9] are stated for direct observation ($P$ being identity). Nevertheless, it also holds for indirect observation. Indeed, consider the identifiable case, that is, $\pi \mapsto \pi P$ is injective. Then one can identify $T(\pi) = \tilde{T}(\mu)$ as a function of the data-generating distribution $\mu = \pi P$, which is linear (with all sub- and super-level sets convex).

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 the companion paper [32], for any fixed $d$ and $n \to \infty$ we get parametric rate $R^*_{\text{iid}}(n) \asymp \frac{1}{n}$. To get interesting phase transitions one needs to let $d$ slowly grow – and this cannot be handled in the setup of [9] 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 $g$ (see (5)). As mentioned in Section 1, for discrete problems with finite $\Theta$ and $\mathcal{X}$, the optimal $g$ can be efficiently pre-computed by convex or linear programming. Furthermore, for certain continuous models, the infinite-dimensional LP can be effectively "finite-dimensionalized" leading to computationally efficient construction of optimal estimators (see the distinct element and species problems in [32] 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 3) that our methods extend to the deterministic setting, when one's goal is to estimate a functional of high-dimensional parameters.

2.2. *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 $\theta_i \overset{iid}{\sim} \pi$ the $i$th 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 $\pi$. (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 [24]. 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 $i$th subject's arrival and experiment termination $t_1$ is simply an $i$-th order statistic of the iid uniform sample $A_i \overset{iid}{\sim} \text{Unif}[0, t_1 - t_0]$. This motivates the formal setting of "Case 1" below. See [19] for a survey and more details.

Fix $s_0 \in (0,1)$ and a class $\Pi$ of distributions on $\Theta = [0,1]$. For any distribution $\pi$ we denote by $F_\pi$ 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 $\theta_i \overset{iid}{\sim} \pi$, $i \in [n]$ and consider two cases of observations:

1. "Interval censoring, Case 1" (also known as "current status model"), see [17, Section 2.3]. The observations are $X_1, \ldots, X_n$, where $X_i = (A_i, \Delta_i)$, given by $A_i \overset{iid}{\sim} G_1$ and $\Delta_i = 1\{\theta_i \le A_i\}$. That is, we only see whether $\theta_i$ 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 $P^{(1)} = \{P^{(1)}_\theta(\cdot) : \theta \in [0,1]\}$.

2. "Interval censoring, Case 2", see [17, Section 4.7]. This time observations are given by $X_i = (A_i, B_i, \Delta, \tilde{\Delta})$ with $(A_i, B_i) \overset{iid}{\sim} G_2$, where $G_2$ is some fixed distribution on $[0,1]^2$, and $\Delta_i = 1\{\theta_i \le A_i\}$, $\tilde{\Delta}_i = 1\{\theta_i \le B_i\}$. 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 $\Pi$, the following was shown in a series of works, cf. [16, Section 5] and [18, Example 3.1]:

$$R_n^{(1,m)}(\Pi) \asymp R_n^{(2,m)} \asymp \frac{1}{n}\,, \tag{36}$$

$$R_n^{(1,c)}(\Pi) \asymp \frac{1}{n^{2/3}}\,, \tag{37}$$

$$R_n^{(2,c)}(\Pi) \asymp \frac{1}{(n\log n)^{2/3}}\,. \tag{38}$$

We denote the $\chi^2$-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 the supplementary material [31, Section 6], which amounts to verifying the assumptions of Theorem 1.

PROPOSITION 6. *Suppose that the set $\Pi$ 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 $g_2$.*[8] *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 $\phi$ on $\mathcal{X}$.*

The above general characterization of the minimax risk by Proposition 6 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 $a \in (0,1)$. (Note that if density $g_1$ is zero on some interval $[a,b]$ then the model becomes unidentifiable.) Then we have for any weakly closed and convex $\Pi$:

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

Indeed, consider $\pi_1, \pi_2 \in \Pi$ with corresponding CDFs given by $F_1$ and $F_2$. A simple calculation shows

$$\chi^2(\pi P^{(1)} \| \pi' P^{(1)}) = \int_0^1 g_1(a) \frac{(F_1(a) - F_2(a))^2}{F_2(a)(1 - F_2(a))}\,. \tag{39}$$

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

Upper-bounding the denominator by $\frac{1}{4}$ and lower-bounding $g_1$ by $\epsilon_1$ we obtain

$$\chi^2(\pi P^{(i)}\|\pi' P^{(i)}) \geq 4\epsilon_1 \int_0^1 da(F_1(a) - F_2(a))^2 \tag{40}$$

$$\geq 4\epsilon_1 \left(\int_0^1 da(F_1(a) - F_2(a))\right)^2, \tag{41}$$

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

$$\delta_{\chi^2}^{(1,m)}(t) \leq \frac{t}{2\sqrt{\epsilon_1}}.$$

Note also that from (16) we conclude that $\delta_{\chi^2}^{(1,m)}(t) \asymp t$. This recovers (36) under most general conditions.

2. Now suppose that $G_1$ has density $g_1$ that is continuous and positive at $s_0$. Fix $\gamma, \epsilon > 0$ and consider the class $\Pi = \Pi(\gamma, \epsilon) \triangleq \{\pi : F_\pi(s) \text{ is } \gamma\text{-Lipschitz for } s \in (s_1, s_2)\}$, where $s_1 = s_0 - \epsilon, s_2 = s_0 + \epsilon$. Then we claim

$$\delta_{\chi^2}^{(1,c)}(t) \asymp t^{1/3} \tag{42}$$

where the implied constants may depend upon $\gamma, \epsilon$. Without loss of generality, we will assume that $\epsilon_0 \leq g_1(s) \leq \frac{1}{\epsilon_0}$ for all $s \in (s_1, s_2)$. (Otherwise, we simply reduce $\epsilon$.) Notice that $\Pi$ is indeed convex and weakly closed.[9] For the lower bound consider any pair of CDFs $F_1, F_2$ such that (a) they both belong to $\Pi$, (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 (39) 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 $\delta_{\chi^2}(\tau_1^3) \geq \tau_1$, as claimed by (42).

For the upper bound, arguing as in (41) we get for any $\tau_1 < \epsilon$ that

$$t^2 \geq \chi^2(\pi_1 P^{(1)}\|\pi_2 P^{(1)}) \gtrsim \int_{s_0 - \tau_1}^{s_0 + \tau_1} |F_1(s_0 + x) - F_2(s_0 + x)|^2 ds.$$

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 $|F_1(s_0 + x) - F_2(s_0 + x)| \geq \frac{\delta}{2}$ and thus

$$t^2 \gtrsim \delta^3,$$

implying that $\delta_{\chi^2}(t) \lesssim t^{1/3}$, finishing the proof of (42).

[9] Indeed, consider any sequence $\pi_n \in \Pi$ weakly converging to $\pi$. Let $S \subset [0, 1]$ be the set of points of continuity of the CDF $F_\pi$. Since the complement of $S$ is countable, $S$ is dense. For any $r, s \in S \cap (s_1, s_2)$ we have $|F_{\pi_n}(r) - F_{\pi_n}(s)| \leq \gamma|r - s|$. By taking the limit $n \to \infty$ we obtain $|F_\pi(r) - F_\pi(s)| \leq \gamma|r - s|$. To extend this estimate to all $r, s$ we just need to find $r_n \downarrow r$ and $s_n \downarrow s$ with $r_n, s_n \in S$ and recall that $F_\pi(r_n) \to F_\pi(r)$ and $F_\pi(s_n) \to F_\pi(s)$.

These applications clearly demonstrate the utility of Theorem 1 for determining the minimax rate without demonstrating an explicit estimator. Indeed, the minimax rates (36) and (37) were obtained with a lot less effort compared to the existing literature [17].[10] In particular, the previous upper bounds were derived by a lengthy analysis of the nonparametric maximum likelihood [18], or certain ad hoc histogram estimator [16]. On the other hand, establishing (38) by computing $\delta_{\chi^2}$ 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

$$\boldsymbol{\Theta}_c = \left\{ \boldsymbol{\theta} \in \Theta^{\otimes n} : \frac{1}{n} \sum_{i=1}^{n} c(\theta_i) \le 1 \right\},$$

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

$$\Pi = \left\{ \pi \in \mathcal{P}(\Theta) : \int c(\theta) \pi(d\theta) \le 1 \right\}. \tag{43}$$

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

$$T(\pi) = \int h(\theta) \pi(d\theta). \tag{44}$$

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 $R^*_{\text{det}}(n)$ is defined in (3), namely,

$$R^*_{\text{det}}(n) = \inf_{\hat{T}} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_c} \mathbb{E}_{\boldsymbol{\theta}} [|\hat{T}(\boldsymbol{X}) - T(\pi_{\boldsymbol{\theta}})|^2] \tag{45}$$

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 $c(\theta) = |\theta|^q$ [6, 5]. Estimation of non-smooth functional such as the $\ell_1$-norm ($h(\theta) = |\theta|$) has been studied in [27, 4].

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 $\theta_i$ 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 $\pi$. 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

$$R^*_{\text{det}}(n) \asymp \delta_{\chi^2}\left(\frac{1}{\sqrt{n}}\right)^2 \tag{46}$$

[10] Results in [17] are stated with the extra assumption on the lower bound on the density of $\pi$ at $s_0$, but this constraint seems not necessary for establishing rates.

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 $R^*_{\rm det}(n) = 0$ or $R^*_{\rm det}(n) \asymp \frac{1}{n}$ (parametric rate).

EXAMPLE 1. Let $\Theta = \mathcal{X} = \{0,1\}$, $c(\theta) = 0$, so that $\Pi = \{\text{Bern}(p) : 0 \le p \le 1\}$. Let $h(\theta) = \theta$ and consider the observation model $\mathbb{P}[X = \theta] = 1 - \mathbb{P}[X = 1-\theta] = \tau$ (the binary symmetric channel). In other words, the goal is to estimate $T(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n \theta_i$ based on the noisy observations $\boldsymbol{X}$. Note that the smallest value of $\delta_{\chi^2}$ will occur at $\tau = 0$, and even then from $\chi^2(\text{Bern}(p)\|\text{Bern}(1/2)) = (1-2p)^2$ we obtain: $\delta_{\chi^2}(t) \ge t/2$ for $t \le 1$. At the same time, a simple unbiased estimator $\hat{T}(X_1,\dots,X_n) = \frac{1}{n(1-2\tau)}\sum_{i=1}^n (1\{X_i = 1\} - \tau)$ achieves

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

One immediate conclusion is that at $\tau = 0$ we have $R^*_{\rm det}(n) = 0$ while $\delta_{\chi^2}(t) > 0$ for all $t > 0$. Furthermore, even when $\tau > 0$ and $R^*_{\rm det}(n) \asymp \delta_{\chi^2}(1/\sqrt{n})^2 \asymp \frac{1}{n}$, the proportionality constant in the first relation is not uniform in $\tau$, as $\lim_{\tau\to 0}\frac{R^*_{\rm det}(n)}{\delta_{\chi^2}(1/\sqrt{n})^2} = 0$. Therefore, we cannot expect the relation (46) 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. $h(\theta_a) \ne h(\theta_b)$ and $\text{TV}(P_{\theta_a}, P_{\theta_b}) < 1$, 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 $n \ge 1$. In the former case, we can lower bound $R^*_{\rm det}(n)$ by the Bayes risk when $\boldsymbol{\theta}$ has iid components with $\mathbb{P}[\theta_i = \theta_a] = \mathbb{P}[\theta_i = \theta_b] = \frac{1}{2}$.[11] Clearly, the corresponding Bayesian risk is $\Omega(1/\sqrt{n})$.

The main result of this section is:

THEOREM 7. *Suppose that $(\Theta, \mathcal{X}, P, T, \Pi)$, with $\Pi$ and $T$ given in (43) and (44) respectively, satisfy conditions A1-A4 of Theorem 1. Then*

$$R^*_{\rm det}(n) \le \delta_{\chi^2}\left(\frac{1}{\sqrt{n}}\right)^2\,, \tag{47}$$

*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 $K_V = \sup_{\pi\in\Pi}\text{Var}_{\theta\sim\pi}[h(\theta)] < \infty$;
A6 *Cost function $c \ge 0$ and there exists $\theta_0 \in \Theta$ with $c(\theta_0) = 0$.*

*Then*

$$R^*_{\rm det}(n) \ge \frac{1}{2400}\delta_{\chi^2}\left(\frac{1}{\sqrt{n}}\right)^2 - \frac{K_V}{2n}\,. \tag{48}$$

[11] This prior needs to be modified if $c(\theta_a) > 1$ or $c(\theta_b) > 1$. Specifically, choose an arbitrary $\theta_0$ 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 $\epsilon$.

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 (47), 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:

$$\sqrt{\mathbb{E}_{\boldsymbol{\theta}}[|\hat{T}_g(\boldsymbol{X}) - T(\pi_{\boldsymbol{\theta}})|^2]} \le \frac{1}{n}\sqrt{\sum_{i=1}^n \mathrm{Var}_{P_{\theta_i}}[g]} + \left|\frac{1}{n}\sum_{i=1}^n (P_{\theta_i}g - h(\theta_i))\right| \tag{49}$$

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

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

we continue (49) to get

$$\sqrt{R^*_{\mathrm{det}}(n)} \le \inf_g \sup_{\boldsymbol{\theta}} \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\pi_{\boldsymbol{\theta}}P}[g]} + |T(\pi_{\boldsymbol{\theta}}) - \pi_{\boldsymbol{\theta}}Pg|\,, \tag{50}$$

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

$$\sqrt{R^*_{\mathrm{det}}(n)} \le \delta_{\mathrm{bv}}(\tfrac{1}{\sqrt{n}})$$

with $\delta_{\mathrm{bv}}$ defined in (24). Applying Proposition 4 we get (47).

To prove (48), fix $\gamma \in (0,1)$ (to be specified later) and consider $\pi_0, \pi_0' \in \Pi$ such that $\chi^2(\pi_0'P\|\pi_0P) \le \frac{1}{n}$ 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 $\theta_0$ is from Assumption A6 such that $c(\theta_0) = 0$. From the convexity of $\chi^2(\cdot\|\cdot)$, we get

$$\chi^2(\pi_1'P\|\pi_1P) \le \frac{\gamma}{d}, \qquad T(\pi_1') - T(\pi_1) = \gamma\delta\,.$$

Denote $\mu' = T(\pi_1'), \mu = T(\pi_1)$. Define distributions $\nu = \pi_1^{\otimes n}, \nu' = \pi_1'^{\otimes n}$ and note that $(\pi_1P)^{\otimes n} = \nu P^{\otimes n}$. Then

$$\begin{aligned}
\mathrm{TV}(\nu P^{\otimes n}, \nu' P^{\otimes n}) = \mathrm{TV}((\pi_1P)^{\otimes n}, (\pi_1'P)^{\otimes n}) &\overset{\text{(a)}}{\le} \frac{1}{2}\sqrt{\chi^2((\pi_1P)^{\otimes n}\|(\pi_1'P)^{\otimes n})} \\
&\overset{\text{(b)}}{=} \frac{1}{2}\sqrt{(1+\chi^2(\pi_1P\|\pi_1'P))^n - 1} \\
&\overset{\text{(c)}}{\le} \frac{1}{2}\sqrt{(1+\gamma/n)^n - 1} \le \frac{1}{2}\sqrt{e^\gamma - 1},
\end{aligned}$$

where (a) follows from the fact that $\mathrm{TV} \le \frac{1}{2}\sqrt{\chi^2}$ [15, Section 3]; (b) is from the tensorization identity in (20); (c) is from the convexity $\chi^2(\pi_1P\|\pi_1'P) \le \gamma\chi^2(\pi_0P\|\pi_0'P)$.

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

$$A = \left\{\boldsymbol{\theta} \in \Theta^{\otimes n} : \frac{1}{n}\sum_{i=1} c(\theta_i) \le 1, T(\pi_{\boldsymbol{\theta}}) \le \mu + \frac{\gamma\delta}{3}\right\} \tag{51}$$

$$A' = \left\{\boldsymbol{\theta} \in \Theta^{\otimes n} : \frac{1}{n}\sum_{i=1} c(\theta_i) \le 1, T(\pi_{\boldsymbol{\theta}}) \ge \mu' - \frac{\gamma\delta}{3}\right\}. \tag{52}$$

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 $\nu, \nu'$ 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 $\nu$.

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 $\nu_1$-a.e. $\boldsymbol{\theta}$ and $T(\pi_{\boldsymbol{\theta}}) \ge \mu + \frac{\gamma\delta}{3}$ for $\nu'_1$-a.e. $\boldsymbol{\theta}$. Applying the TV version of Le Cam's method for quadratic risk (see [44, Lemma 1]) yields the following minimax lower bound:

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

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

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

Optimizing over the choice of $\pi_0, \pi'_0$ thus yields (48). □

EXAMPLE 2. As a quick application, consider the problem of estimating the $\ell_1$-norm of a vector in the Gaussian location model [27, 4], 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 [4] that $R^*_{\mathrm{det}}(n) = \Theta((\frac{\log\log n}{\log n})^2)$. (In fact, the sharp constant as $n \to \infty$ was also found). To see how this result follows from Theorem 7, we first verify the assumptions A1-A6. Let $c \equiv 0$ and $\Pi = \mathcal{P}([-1,1])$ which is weakly compact. Let $\mathcal{F}$ be the set of all continuous bounded real-valued functions on $[-1,1]$. Denote by $P$ the Gaussian convolution kernel. It remains to verify A4b. Indeed, $\pi \mapsto T(\pi) = \pi h$ is weakly continuous because $h$ is continuous and bounded on $[-1,1]$. Next, for any continuous bound $g : [-1,1] \to \mathbb{R}$, $\pi Pg = \pi(g * \varphi)$, where $*$ denotes convolution and $\varphi$ is the standard normal density. Since $g * \varphi$ is continuous bounded, $\pi Pg$ is weakly continuous in $\pi$. Note that $K_V = 1$, applying Theorem 7 yields $c\delta^2_{\chi^2}(\frac{1}{\sqrt{n}}) - \frac{1}{4n} \le R^*_{\mathrm{det}}(n) \le \delta^2_{\chi^2}(\frac{1}{\sqrt{n}})$ for constant $c$, where

$$\delta_{\chi^2}(t) = \sup\left\{\int |\theta|\pi'(d\theta) - \int |\theta|\pi(d\theta) : \chi^2(\pi' * N(0,1) \| \pi * N(0,1)) \le t^2\right\}, \tag{53}$$

where 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 the supplementary material [31, Section 6])

$$\delta_{\chi^2}(t) \asymp \frac{\log\log\frac{1}{t}}{\log\frac{1}{t}}, \tag{54}$$

recovering the result of [4].

However, if the parameter space is unbounded with $\Theta = \mathbb{R}$ we have $K_V = \infty$ and lower bound in Theorem 7 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 [4] that $R^*_{\mathrm{det}}(n) \asymp \frac{1}{\log n}$.

**4. Exponential families.** Let us revisit the setting of Theorem 1 in the special case when $\Theta$ and $\mathcal{X}$ are both finite. Given a closed convex set $\Pi \subset \mathcal{P}(\Theta)$, 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 $\mu$. Hence, the statistical problem at hand becomes: Given $X_i \overset{iid}{\sim} \mu \in M_0$ 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 $\mathcal{X} = [d] \triangleq \{1, \dots, d\}$. Then for $j \in [d]$ let $\phi_j(x) = \mathbf{1}\{x = j\}$, and denote $\phi(x) = (\phi_1(x) \dots, \phi_d(x))$. For every $\gamma = (\gamma_1, \dots, \gamma_d) \in \mathbb{R}^d$ 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 $P_\gamma$ forms an exponential family with *natural parameters* $\gamma$ and *mean parameters* $\mu_{\mathrm{f}}(\gamma) = \mathbb{E}_{X\sim P_\gamma}[\phi(X)]$. Theorem 1 then shows that there exists $g$ such that the empirical-mean estimator

$$\hat{T}(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \tag{55}$$

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

It turns out that this result can be extended to other exponential families (i.e. different choices of $\phi : \mathcal{X} \to \mathbb{R}^d$) for which the (near) optimality of empirical-mean estimators continues to hold. This extends the result of [23] 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 $(\nu, \phi, \Gamma)$, where $\nu$ 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 $\phi_i$, i.e., $\mathcal{F} = \{\langle h, \phi\rangle : h \in \mathbb{R}^d\}$. We make two standing assumptions on the exponential family:[12]

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

$$\mathrm{Var}_{X\sim P_{\gamma_0}}(\langle \phi(X), h\rangle) > 0 \qquad \forall h \in \mathbb{R}^d \setminus \{0\}. \tag{56}$$

In addition to the natural parameter $\gamma$, we define the *mean parameter* $\mu$ via the forward map

$$\mu_{\mathrm{f}}(\gamma) \triangleq \mathbb{E}_{X\sim P_\gamma}[\phi(X)].$$

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

It is well known (see e.g. [2]) that inside $\Gamma$ the function $\gamma \mapsto C(\gamma)$ is infinitely differentiable, whose first two derivatives give the mean and covariance of $\phi(X)$:

$$\mu_{\mathrm{f}}(\gamma) = \nabla C(\gamma), \qquad \frac{\partial \mu_{\mathrm{f}}}{\partial \gamma} = \mathrm{Hess}\, C(\gamma) = \mathrm{Cov}_{P_\gamma}[\phi(X)] \triangleq \Sigma(\gamma)\,. \tag{57}$$

The non-degeneracy assumption (56) implies

$$\Sigma(\gamma) \succ 0 \qquad \forall \gamma \in \Gamma\,. \tag{58}$$

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

$$\mu_{\mathrm{f}}(\gamma_{\mathrm{r}}(\mu)) = \mu \qquad \forall \mu \in M\,.$$

It is also known that Jacobian of $\gamma_{\mathrm{r}}$ can be computed as

$$\frac{\partial \gamma_{\mathrm{r}}(\mu)}{\partial \mu} = \Sigma^{-1}(\gamma_{\mathrm{r}}(\mu))\,. \tag{59}$$

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

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

$$R_n^*(\Gamma_0) = \inf_{\hat{T}} \sup_{\gamma \in \Gamma_0} \mathbb{E}_{X_i \overset{iid}{\sim} P_\gamma}[|\hat{T}(X_1, \ldots, X_n) - T(\gamma)|^2]\,. \tag{60}$$

The main finding in this section is that for estimating linear functionals of the mean parameter $\mu$, under certain convexity assumptions (that are strictly weaker than those in [23]), the minimax quadratic risk is characterized by certain moduli of continuity within universal constant factors. To this end, let $\omega_H$ denote the modulus of continuity of $T$ on $M_0$ with respect to the Hellinger distance, i.e.

$$\omega_H(t) \triangleq \sup_{\gamma, \gamma' \in \Gamma_0} \{T(\gamma) - T(\gamma') : H(P_\gamma, P_{\gamma'}) \le t\}, \tag{61}$$

THEOREM 8. *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 $(\nu, \phi, \Gamma)$ 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*

$$\mu \mapsto \sqrt{\mathrm{Var}_{\tilde{P}_\mu}[\phi]} \text{ is concave in } \mu \in M_0 \text{ for all } \phi \in \mathcal{F}. \tag{62}$$

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

$$T(\gamma) = \langle h, \mu_{\mathrm{f}}(\gamma) \rangle \tag{63}$$

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

$$c_0 \omega_H(1/\sqrt{n}) \le \sqrt{R_n^*(\Gamma_0)} \le c_1 \omega_H(1/\sqrt{n}),, \tag{64}$$

*and this rate is achieved by the estimator of the type (55) with $g \in \mathcal{F}$.*

The proof is given in the supplementary material [31, Section 5]. We stress that constants $c_0, c_1$ in (64) do not depend on dimension of the exponential family, and thus as in [23] we can think of the above result as essentially non-parametric.

REMARK 2. Note that in the setting of the preceding Theorem 8, 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 (55). Indeed, it gives an unbiased estimator by design. However, the problem is that $\mathrm{Var}_{P_\gamma}[\phi_0(X)]$ might be prohibitively large (such as in population recovery in [32]). The main discovery here is that the concavity condition (62) guarantees existence of some other $g \neq \phi_0$ such that the empirical average of $g$ is minimax rate-optimal.

REMARK 3 (The double role of Assumption (62)). To shed some light on how assumption (62) relates to the tightness of empirical-mean estimators, we observe that the Fisher information matrix for parameter $\gamma$ is given by $I_F(\gamma) = \Sigma(\gamma)$, while for parameter $\mu$ we get $I_F(\mu) = \tilde{\Sigma}^{-1}(\mu)$. Let us consider one-dimensional case $d = 1$, in which case estimating $T(\gamma)$ is equivalent to estimating the mean $\mu$ itself. An intermediate result in the proof of the upper bound (see (95) in the supplemental material [31]) shows that $R_n^*(M_0) \le \frac{1}{n \min_\mu I_F(\mu)}$. At the same time from the Bayesian Cramér-Rao lower bound (van Trees inequality) [16], we expect that one could derive a lower bound which is matching up to a constant multiplier. However, the multiplier will necessarily depend on how rapidly $I_F(\mu)$ grows around its minimum.[13] How, then, are we able to get the minimax rate to within absolute global constants? It turns out that Assumption (62) implies a universal estimate on how rapidly $I_F(\mu)$ is allowed to grow (this is the key new ingredient of Theorem 8 compared to Theorem 1 encoded in the inequality (99) in [31]). Thus, assumption (62) enters our proof in two crucial ways: for the applicability of the minimax theorem and for taming the behavior of Fisher information. While for the former we could replace (62) by quasi-concavity, for the latter the stronger requirement of concavity is necessary for our method. We suspect, thus, that Assumption (62) may be the ultimate limit of how far the present method can be pushed.

4.1. *Comparison to Juditsky-Nemirovski [23].* As opposed to the squared loss (60), Juditsky-Nemirovski [23] considered the $\epsilon$-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 8 proved for the quadratic risk can be translated to the $\epsilon$-quantile loss similarly as done in Corollary 2.

COROLLARY 9. *In the setting of Theorem 8, whenever $e^{-2n} \le \epsilon \le 2^{-8}$ we have up to absolute constants of proportionality*

$$R_{n,\epsilon}^*(\Gamma_0) \asymp \omega_H\left(\sqrt{\frac{1}{n}\ln\frac{1}{\epsilon}}\right). \tag{65}$$

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 8. □

[13] Indeed, let $\mu^*$ denote the minimizer of $I_F(\mu)$, which we assume is in the interior of $M_0$. The van Trees inequality [37, Theorem 2.13] states that for large enough $n$ $R_n^*(M_0) \ge (n\mathbb{E}[I_F(\mu^* + n^{-1/4}U)] + \pi^2\sqrt{n})^{-1}$, where $U$ denotes a random variable on $[-1,1]$ with density $\cos^2\frac{\pi u}{2}$. Thus, if $I_F(\mu^*) > 0$ and continuous we have the matching lower bound $R_n^* \ge \frac{1+o(1)}{nI_F(\mu^*)}$. However, the magnitude of $o(1)$ term will be the larger the more rapidly $I_F(\mu)$ grows in the neighborhood of $\mu^*$.

We now discuss results of [23]. The following assumptions are made in [23] (later called *a simple observation schemes* in [22, Section 2.4.2])

1. The exponential family $(\nu, \phi, \Gamma)$ has $\Gamma = \mathbb{R}^d$, i.e. the natural parameters $\gamma$ can range over the entire space $\mathbb{R}^d$.
2. The functional $T(\gamma) = T(A(\xi))$ is linear in $\xi$, where $\gamma = A(\xi)$ is a reparametrization such that the map

$$\xi \mapsto C(A(\xi) + a) - C(A(\xi)) \text{ is concave for every } a \in \mathbb{R}^d. \tag{66}$$

Under these assumptions, it is shown that (cf. [23, Theorem 3.1 and Proposition 3.1])

$$\frac{1}{2}\omega_H\left(\sqrt{2\left(1 - e^{-\frac{1}{2n}\log\frac{1}{4\epsilon}}\right)}\right) \le R^*_{n,\epsilon} \le \frac{1}{2}\omega_H\left(\sqrt{2\left(1 - e^{-\frac{1}{n}\log\frac{2}{\epsilon}}\right)}\right); \tag{67}$$

in particular, whenever $\exp(-2n) \le \epsilon < \frac{1}{5}$, we have (65) (also within absolute constants). Thus, to compare our results with [23] we need to compare the assumptions. It turns out (see the supplementary material [31, Section 5] for a proof) that (66) is equivalent to the following requirement

$$\xi \mapsto \mu_{\mathrm{f}}(A(\xi)) \text{ is affine and } \mu \mapsto \mathrm{Var}_{P_\mu}[\phi(X)] \text{ is concave in } \mu \in M_0 \text{ for all } \phi \in \mathcal{F}. \tag{68}$$

This equivalence shows that our condition (62) is strictly weaker than (68). Let us consider a simple example showing difference between (62) and (68).

EXAMPLE 3 (Exponential distribution). Let $\mathcal{X} = \mathbb{R}^d_+$, $\gamma \in \mathbb{R}^d_+$ and take $P_\gamma(dx) = \prod_{i=1}^d \gamma_i e^{-\gamma_i x_i}\mathbf{1}_{\{x_i>0\}}dx_i$, 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 $\ell_p$-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

$$\omega_H(t) \asymp t d^{\max(\frac{1}{2}-\frac{1}{p},0)} \tag{69}$$

up to absolute (i.e. $p$-independent) constants. (For $p \le 2$ the worst pair $(\mu, \mu')$ are 1-sparse (with a single nonzero), whereas for $p > 2$ they are scaled constant vectors.) From Theorem 8 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 (62) holds, the condition (68) imposed by [23] 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 [23]).

The above example thus shows that the extension from (68) to (62) 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 (62) holds but not (68). For larger dimension $d > 1$, the family $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma$ a $d \times d$ positive definite matrix, does not satisfy either (68) or (62). (However, in this case a partial remedy is possible – see [22, Section 3.4.1].)

Finally, we point out a limitation of our method. For that we consider an example where behavior of $\epsilon$-quantile and quadratic risks are very different.

EXAMPLE 4. Consider the goal of estimating the bias $p$ of $X_i \overset{iid}{\sim} \mathrm{Bern}(p)$ where $p \in \{1/4, 1/3\}$. In this case $M_0$ is not convex and the modulus of continuity $\omega_H(t) = 0$ for $t < H(\mathrm{Bern}(1/4), \mathrm{Bern}(1/3))$, but this does not contradict (65) 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 $c_1, c_2 > 0$. Thus, characterization of $R^*_n$ in (64) cannot hold.

Consequently, we see that generally $R_n^* \asymp \omega_H(1/\sqrt{n})$ is not possible without convexity of $M_0$. It turns out that methods of [22, Section 3.1.4] are possible to extend to some non-convex cases, such as when $M_0$ is a union of finitely many convex sets (in which case $R_{n,\epsilon}^*$ is again characterized in terms of $\omega_H(\cdot)$). Our method, however, obtains the result on $R_{n,\epsilon}^*$ from that of $R_n^*$ and thus manifestly fails to attack such cases, including that of the previous example.

**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 MIT-IBM Watson AI Lab.

We thank Prof. A. Rakhlin and Prof. A. Tsybakov for pointing out [9] and [23], 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.

## REFERENCES

[1] BORWEIN, J. M. and ZHUANG, D. (1986). On Fan's minimax theorem. *Mathematical Programming* **34** 232–234.

[2] BROWN, L. D. (1986). Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. In *Lecture Notes-Monograph Series*, (S. S. Gupta, ed.) **9** Institute of Mathematical Statistics, Hayward, CA.

[3] BROWN, L. D. and LOW, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics* **24** 2524–2535.

[4] CAI, T. T. and LOW, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics* **39** 1012–1041.

[5] COLLIER, O., COMMINGES, L. and TSYBAKOV, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics* **45** 923–958.

[6] COLLIER, O., COMMINGES, L., TSYBAKOV, A. B. and VERZELEN, N. (2016). Optimal adaptive estimation of linear functionals under sparsity. *arXiv preprint arXiv:1611.09744*.

[7] DE, A., O'DONNELL, R. and SERVEDIO, R. A. (2020). Sharp Bounds for Population Recovery. *Theory Of Computing* **16** 1–20.

[8] DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics* **22** 238–270.

[9] DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence, II. *The Annals of Statistics* **19** 668–701.

[10] DUNFORD, N. and SCHWARTZ, J. T. (1958). *Linear Operators: General theory* **1**. Interscience Publishers, New York.

[11] DVIR, Z., RAO, A., WIGDERSON, A. and YEHUDAYOFF, A. (2012). Restriction access. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* 19–33. ACM.

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

[13] FAN, K. (1953). Minimax theorems. *Proceedings of the National Academy of Sciences* **39** 42–47.

[14] FISHER, R. A., CORBET, A. S. and WILLIAMS, C. B. (1943). 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* 42–58.

[15] GIBBS, A. L. and SU, F. E. (2002). On choosing and bounding probability metrics. *International statistical review* **70** 419–435.

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

[17] GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric estimation under shape constraints* **38**. Cambridge University Press.

[18] GROENEBOOM, P. and WELLNER, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation* **19**. Springer Science & Business Media.

[19] HUANG, J. and WELLNER, J. A. (1997). Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics* 123–169. Springer.

[20] IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1984). On the nonparametric estimation of a value of a linear functional in the Gaussian white noise. *Theory of Probability & Its Applications* **29** 19–32.

[21] JANA, S., POLYANSKIY, Y. and WU, Y. (2020). Extrapolating the profile of a finite population. In *Proceedings of Conference on Learning Theory (COLT)*. arXiv:2005.10561.
[22] JUDITSKY, A. and NEMIROVSKI, A. (2020). *Statistical Inference via Convex Optimization* **69**. Princeton University Press.
[23] JUDITSKY, A. B. and NEMIROVSKI, A. S. (2009). Nonparametric estimation by convex programming. *The Annals of Statistics* **37** 2278–2300.
[24] KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53** 457–481.
[25] LE CAM, L. (1973). Convergence of Estimates Under Dimensionality Restrictions. *The Annals of Statistics* **1** 38 – 53.
[26] LE CAM, L. (1986). *Asymptotic methods in statistical decision theory*. Springer-Verlag, New York, NY.
[27] LEPSKI, O., NEMIROVSKI, A. and SPOKOINY, V. (1999). On estimation of the $L_r$ norm of a regression function. *Probability Theory and Related Fields* **113** 221–253.
[28] LUGOSI, G. and MENDELSON, S. (2019). Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society* **22** 925–965.
[29] ORLITSKY, A., SURESH, A. T. and WU, Y. (2016). Optimal Prediction of the Number of Unseen Species. *Proceedings of the National Academy of Sciences (PNAS)* **113** 13283-13288.
[30] POLYANSKIY, Y., SURESH, A. T. and WU, Y. (2017). Sample complexity of population recovery. In *Proceedings of Conference on Learning Theory (COLT)*. arXiv:1702.05574v3.
[31] POLYANSKIY, Y. and WU, Y. (2019). Supplement to "Dualizing Le Cam's method for functional estimation I: Theory and examples".
[32] POLYANSKIY, Y. and WU, Y. (2023). Dualizing Le Cam's method for functional estimation II: Applications to estimating the unseens.
[33] POLYANSKIY, Y. and WU, Y. (forthcoming). *Information Theory: From Coding to Learning*. Cambridge University Press Draft: https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf.
[34] RASKHODNIKOVA, S., RON, D., SHPILKA, A. and SMITH, A. (2009). Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing* **39** 813–842.
[35] ROBBINS, H. (1951). 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.
[36] STRASSER, H. (1985). *Mathematical theory of statistics: Statistical experiments and asymptotic decision theory*. Walter de Gruyter, Berlin, Germany.
[37] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY.
[38] VALIANT, P. (2011). Testing symmetric properties of distributions. *SIAM Journal on Computing* **40** 1927–1968.
[39] WIGDERSON, A. and YEHUDAYOFF, A. (2012). Population recovery and partial identification. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on* 390–399. IEEE.
[40] WOLFOWITZ, J. (1957). The minimum distance method. *The Annals of Mathematical Statistics* 75–88.
[41] WU, Y. and YANG, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory* **62** 3702-3720.
[42] WU, Y. and YANG, P. (2018). Sample complexity of the distinct element problem. *Mathematical Statistics and Learning* **1** 37-72.
[43] WU, Y. and YANG, P. (2020). Optimal estimation of Gaussian mixtures with denoised method of moments. *The Annals of Statistics* **48** 1981–2007. arxiv:1807.07237.
[44] YU, B. (1997). Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* 423–435.
[45] ZHANG, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statistica Sinica* **7** 181–193.
[46] ZHANG, C.-H. (2003). Compound decision theory and empirical Bayes methods. *The Annals of Statistics* **31** 379–390.

## SUPPLEMENTARY MATERIAL

**Supplementary material for "Dualizing Le Cam's method for functional estimation I: Theory and examples"**

(; .pdf). Due to space constraints, some technical results are given in the supplementary documents [31].

*Submitted to the Annals of Statistics*

# SUPPLEMENTARY MATERIAL FOR "DUALIZING LE CAM'S METHOD FOR FUNCTIONAL ESTIMATION, WITH APPLICATIONS TO ESTIMATING THE UNSEENS"

BY YURY POLYANSKIY* AND YIHONG WU†

*Massachusetts Institute of Technology* and Yale University†*

*Department of Statistics and Data Science*
*Yale University*
*New Haven, CT 06511*
*yihong.wu@yale.edu*

*Department of Electrical Engineering and Computer Science*
*Massachusetts Institute of Technology*
*Cambridge, MA 02139*
*yp@mit.edu*

## 5. Proof of Theorem 8.

LEMMA 10 (Auxiliary convex analysis). *Let $X$ and $Y$ be a dual pair of finite-dimensional vector spaces and $\Pi$ 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*[1]

$$f(x, cy) = |c| f(x,y)\,.$$

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

$$d(x'\|x) \triangleq \sup\{\langle x - x', y\rangle : f(x,y) \le 1\}, \tag{70}$$

$$d_S(x', x) \triangleq d(x', (x+x')/2), \tag{71}$$

$$\delta_0(t) \triangleq \inf_y \sup_{x\in\Pi} t f(x,y) + |\langle x, g - y\rangle|, \tag{72}$$

$$\delta_1(t) \triangleq \sup_{x,x'\in\Pi} \{\langle x - x', g\rangle : d(x'\|x) \le t\}, \tag{73}$$

$$\delta_2(t) \triangleq \sup_{x,x'\in\Pi} \{\langle x - x', g\rangle : d_S(x', x) \le t\}. \tag{74}$$

*We claim the following:*

1. $d_S(x', x) = d_S(x, x')$
2. $d_S(x', x) \le d(x\|x')$
3. $\frac{1}{2}\delta_2(t) \le \delta_1(t) \le \delta_2(t)$

---

*MSC2020 subject classifications:* Primary 62G05; secondary 62C20,62G86,62G07.

*Keywords and phrases:* Functional estimation, minimax optimality, convex optimization and duality, modulus of continuity, linear programming, estimating the unseens, distinct elements problem, species problem..

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

4. *And the key result:*

$$\frac{1}{2}\delta_1(t) \le \delta_0(t) \le \delta_1(t). \tag{75}$$

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

$$\langle x-x', y\rangle \le t f(x,y). \tag{76}$$

We further lower bound

$$\delta_0(t) \ge \frac{1}{2}\inf_y t(f(x,y)+f(x',y)) + |\langle x-x', g-y\rangle| \tag{77}$$

$$\ge \frac{1}{2}\inf_y tf(x,y) + |\langle x-x', g-y\rangle| \tag{78}$$

$$\ge \frac{1}{2}\langle x-x', g\rangle + \frac{1}{2}\inf_y tf(x,y) - \langle x-x', y\rangle, \tag{79}$$

where in the first step we used convexity of $|\cdot|$, in the second positivity of $f$ and in the last step $|a| \ge a$. From (76) we conclude that $\delta_0 \ge \frac{1}{2}\delta_1$.

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

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

We now apply minimax theorem to get

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

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

$$\delta_0(t) \le \sup_{x\in\Pi, x'\in\Pi} \inf_y tf(x,y) + \langle x-x', g-y\rangle \tag{80}$$

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

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

Consequently, the right-hand side of (80) evaluates to exactly $\delta_1(t)$.

□

PROOF OF THEOREM 8. 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.

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

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

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 $P_\gamma$ in terms of the expectations of unit-variance functions in $\mathcal{F}$;[2] an explicit expression for $d$ is given in (93) below. The modulus of continuity of $T$ with respect to $d_J$ and $d$ will also play a role:

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

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

We start by establishing the following comparison

$$\omega_J(t) \le \omega_H(t)\,. \tag{81}$$

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 (81).

A routine two-point argument yields the lower bound

$$\omega_H(c_3/\sqrt{n}) \le \sqrt{R_n^*(\Gamma_0)} \tag{82}$$

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:

$$\sqrt{R_n^*(\Gamma_0)} \le \omega_d(1/\sqrt{n})\,. \tag{83}$$

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

$$\omega_J(t) \ge \omega_d(c_2 t)\,. \tag{84}$$

- Subadditivity property of $\omega_d$:

$$\omega_d(ct) \ge c\omega_d(t), \quad \forall 0 \le c \le 1\,. \tag{85}$$

- Together (81), (82) and the three steps above imply

$$\omega_H(\frac{c_0}{\sqrt{n}}) \le \sqrt{R_n^*} \le \omega_d(\frac{c_2}{c_2\sqrt{n}}) \le \frac{1}{c_2}\omega_H(\frac{1}{\sqrt{n}})\,. \tag{86}$$

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

$$\omega_H(\frac{c_3}{\sqrt{n}}) \ge c_0\omega_H(\frac{1}{\sqrt{n}}) \tag{87}$$

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

We proceed to proving the above claims. Let us extend the family $\mathcal{F}$ to $\mathcal{F}^* = \operatorname{span}\{\mathcal{F}, 1\}$ by adding constants. Similarly, we extend $\phi$ to $\phi^*(x) = (1, \phi(x)) \in \mathbb{R}^{d+1}$ by adding a constant coordinate. (Note that as exponential family $\phi^*$ no longer satisfies non-degeneracy condition (56)). 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 $\gamma^*$ are coupled by this relation.) Recall that $\mathbb{E}_\gamma[\phi(X)] = \mu_{\mathrm{f}}(\gamma) = \mu$ 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:

$$\inf_{g\in\mathcal{F}^*}\sup_{\gamma\in\Gamma_0}\sqrt{\mathbb{E}_\gamma[(T-\hat{T})^2]} \le \inf_{g\in\mathcal{F}^*}\sup_{\gamma\in\Gamma_0}\frac{1}{\sqrt{n}}\sqrt{\operatorname{Var}_{P_\gamma}[g(X)]} + |\mathbb{E}_{P_\gamma}[g(X)] - T(\gamma)| \tag{88}$$

$$= \inf_{\phi\in\mathcal{F}^*}\sup_{\mu\in M_0}\frac{1}{\sqrt{n}}\sqrt{\operatorname{Var}_{\tilde{P}_\mu}[g(X)]} + |\,\langle \gamma^* - h^*, \mu^*\rangle\,| \tag{89}$$

$$= \delta_0(1/\sqrt{n})\,, \tag{90}$$

where $\delta_0$ is defined as (similar to $\delta_{\mathrm{bv}}$ previously defined in (24))

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

This definition coincides with $\delta_0$ defined in Lemma 10 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 $\langle x, y\rangle = \sum_{i=0}^n x_i y_i$. 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 $f(x, y) = \sqrt{\operatorname{Var}_{\tilde{P}_\mu}[g(X)]}$
- $e_0 = (1, 0, \ldots, 0)$ 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 10 shows

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

This completes the proof of (83).

We proceed to proving (84). We start with some preparatory remarks. A simple calculation reveals that

$$d_J(P_{\gamma_1}, P_{\gamma_2}) = \langle \gamma_1 - \gamma_2, \mu_{\mathrm{f}}(\gamma_1) - \mu_{\mathrm{f}}(\gamma_2)\rangle \,. \tag{91}$$

Similarly, we have the following expression for $d$:

$$d(\tilde{P}_{\mu'}\|\tilde{P}_\mu) = \sup_{a\in\mathbb{R}^n}\{\langle \mu - \mu', a\rangle : \left\langle \tilde{\Sigma}(\mu)a, a\right\rangle \le 1\} \tag{92}$$

$$= \sqrt{\left\langle \tilde{\Sigma}^{-1}(\mu)\Delta, \Delta\right\rangle}\,, \quad \Delta = \mu - \mu'\,, \tag{93}$$

where we used the identity

$$\sup_{y\in\mathbb{R}^n}\{\langle y,b\rangle : \langle Ay,y\rangle \le 1\} = \sqrt{\langle A^{-1}b,b\rangle}\,, \tag{94}$$

which follows from the Cauchy-Schwarz inequality: $\langle y,b\rangle^2 = \langle A^{\frac12}y, A^{-\frac12}b\rangle^2 \le \langle A^{-1}b,b\rangle\langle Ay,y\rangle$.

Thus, we get a more explicit formula for $\omega_d$:

$$\omega_d(t) = \sup_{\mu_1,\mu_2\in M_0}\left\{\langle \Delta, h\rangle : \left\langle \tilde\Sigma^{-1}(\mu_2)\Delta,\Delta\right\rangle \le t^2\,,\ \Delta=\mu_1-\mu_2\right\}. \tag{95}$$

This expression clearly shows (85).

We next establish a key inequality connecting the behavior of $\tilde\Sigma^{-1}(\lambda\mu_1+\bar\lambda\mu_0)$ with the assumption (62). Consider the following chain of inequalities: for any $a\in\mathbb{R}^n$,

$$\left\langle \tilde\Sigma^{-1}(\lambda\mu_1+\bar\lambda\mu_0)a,a\right\rangle^{\frac12} = \sup_y\left\{\langle y,a\rangle : \left\langle\tilde\Sigma(\lambda\mu_1+\bar\lambda\mu_0)y,y\right\rangle^{\frac12}\le 1\right\} \tag{96}$$

$$\le \sup_y\left\{\langle y,a\rangle : \lambda\left\langle\tilde\Sigma(\mu_1)y,y\right\rangle^{\frac12} + \bar\lambda\left\langle\tilde\Sigma(\mu_2)y,y\right\rangle^{\frac12}\le 1\right\} \tag{97}$$

$$\le \sup_y\left\{\langle y,a\rangle : \lambda\left\langle\tilde\Sigma(\mu_1)y,y\right\rangle^{\frac12}\le 1\right\} \tag{98}$$

$$= \frac{1}{\lambda}\left\langle\tilde\Sigma^{-1}(\mu_1)a,a\right\rangle^{\frac12}, \tag{99}$$

where in (96) we used (94), in (97) we applied (62), in (98) we omitted the second term, which is non-negative by (58), and in (99) we used (94) again.

Next, we obtain an upper bound on $d_J(P_{\gamma_1},P_{\gamma_2})$ by continuing from (91). We denote $\mu_i=\mu_{\rm f}(\gamma_i), i=1,2$ 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 $\gamma_\lambda \triangleq \gamma_{\rm r}(\lambda\mu_1+\bar\lambda\mu_2)$ and

$$\dot\gamma_\lambda = \frac{d}{d\lambda}\gamma_\lambda = \sum_{j=1}^n \frac{\partial\gamma_{\rm r}}{\partial\mu_j}(\mu_{1,j}-\mu_{2,j}) \overset{(59)}{=} \Sigma(\gamma_\lambda)^{-1}\Delta\,. \tag{100}$$

Then we have

$$d_J(P_{\gamma_1},P_{\gamma_2}) = \int_0^1 d\lambda\left\langle\tilde\Sigma^{-1}(\lambda\mu_1+\bar\lambda\mu_2)\Delta,\Delta\right\rangle \tag{101}$$

$$\le \int_0^{1/2} d\lambda\frac{1}{\bar\lambda}\left\langle\tilde\Sigma^{-1}(\mu_2)\Delta,\Delta\right\rangle + \int_{1/2}^1 d\lambda\frac{1}{\lambda}\left\langle\tilde\Sigma^{-1}(\mu_1)\Delta,\Delta\right\rangle \tag{102}$$

$$= \ln 2\cdot\left\langle(\tilde\Sigma^{-1}(\mu_2)+\tilde\Sigma^{-1}(\mu_1))\Delta,\Delta\right\rangle, \tag{103}$$

where (101) is from (100), (102) is from (99) and (103) is by computing the integrals.

Finally, consider a pair $\mu_1,\mu_2\in M_0$ in the optimization (95), i.e. such that

$$\left\langle\tilde\Sigma(\mu_2)\Delta,\Delta\right\rangle \le t^2\,, \tag{104}$$

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

$$\mu_1' = \frac23\mu_1+\frac13\mu_2\,,\quad \mu_2' = \frac13\mu_1+\frac23\mu_2\,, \tag{105}$$

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

$$\langle \mu_1' - \mu_2', g\rangle = \frac{1}{3}\langle \Delta, g\rangle \,. \tag{106}$$

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

$$d_J(\tilde{P}_{\mu_1'}, \tilde{P}_{\mu_2'}) \le c' t^2 \,, \tag{107}$$

which, together with (107) would clearly establish (84). Notice that from (105) and (99) we have

$$\left\langle \tilde{\Sigma}^{-1}(\mu_1')\Delta, \Delta \right\rangle \le 3 \left\langle \tilde{\Sigma}^{-1}(\mu_2)\Delta, \Delta \right\rangle \tag{108}$$

$$\left\langle \tilde{\Sigma}^{-1}(\mu_2')\Delta, \Delta \right\rangle \le \frac{3}{2} \left\langle \tilde{\Sigma}^{-1}(\mu_2)\Delta, \Delta \right\rangle \,. \tag{109}$$

Hence, the left-hand side in (103) is upper-bounded by a constant multiple of $\langle \tilde{\Sigma}(\mu_2)\Delta, \Delta\rangle$, which, in view of (104), shows (107) and, hence, (84).

We complete the proof by showing (87). Notice that $\omega_H(1/\sqrt{n}) \le \omega_H(c_0/\sqrt{\lfloor c_0^2 n\rfloor})$ and then from (64) we have for all $n \ge 2/c_0^2$ and $c_4 = \frac{\sqrt{2}}{c_0}$:

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

where (a) follows from (85) with $c = \frac{c_2 c_3}{c_4}$; (b) follows from (84) and (81). This completes the proof of (87). □

PROOF OF (66) $\iff$ (68). To show this equivalence, first notice the representation

$$C(\gamma + a) - C(\gamma) = \langle \mu_{\mathrm{f}}(\gamma), a\rangle + \int_0^1 (1-s) a^T \Sigma(\gamma + sa) a \, ds \tag{110}$$

since $\nabla C(\gamma) = \mu_{\mathrm{f}}(\gamma)$ and $\operatorname{Hess} C(\gamma) = \Sigma(\gamma)$ as in (57). Thus, from here (68) clearly imply (66) by virtue of

$$a^T \Sigma(\gamma) a = \operatorname{Var}_{P_\gamma}[\langle \phi(X), a\rangle] \,. \tag{111}$$

Conversely, (66) 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. (110), we conclude that $\xi \mapsto \langle \mu_{\mathrm{f}}(A(\xi)), a\rangle$ is concave for any $a$ (in particular, for $-a$ as well), and hence $\xi \mapsto \mu_{\mathrm{f}}(A(\xi))$ must be affine. Continuing, again from (66) we must have that

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

is concave for any $\epsilon \ne 0$. Taking the limit as $\epsilon \to 0$, cf. (110), we conclude that $\xi \mapsto a^T \Sigma(A(\xi)) a$ must be concave, which implies the second claim in (68) in view of (111). □

## 6. Additional proofs.

PROOF OF PROPOSITION 6. Recall that $\Theta = [0,1]$ and $\mathcal{X} = [0,1] \times \{0,1\}$ for Case 1 and $\mathcal{X} = [0,1]^2 \times \{0,1\}^2$ for Case 2. By assumption, $\Pi$ is a convex and weakly closed set of distributions on $[0,1]$. The goal is to estimate either $T_c(\pi) \triangleq F_\pi(s_0) = \pi([0,s_0])$ or $T_m(\pi) \triangleq \int_{[0,1]} \theta\pi(d\theta)$.

We aim to apply Theorem 1 by verifying its assumptions. Assumptions A1 holds because $T_c(\pi)$ and $T_m(\pi)$ are both linear in $\pi$. For A2, $\Pi$ is convex by assumption. 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 $\Pi$ with the weak topology. Since $\Pi$ 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 by definition, while for $\pi \mapsto T_c(\pi)$ weak continuity is implied by the assumption (indeed, if $\pi_n \xrightarrow{w} \pi$ then $F_{\pi_n}(s_0) \to F_\pi(s_0)$ due to assumption on $s_0$ being a point of continuity of $F_\pi$). To complete the verification of A4b, it suffices to verify that for any continuous $\phi$ on $\mathcal{X}$ the functional $\pi P\phi$ is weakly continuous in $\pi$. 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

$$\pi P\phi = \int_{[0,1]} \pi(d\theta) f(\theta), \quad f(\theta) \triangleq \int_{[0,1]} da g_1(a)\phi(a,1)1\{\theta \le a\} + \phi(a,0)1\{\theta > a\} \tag{112}$$

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 (112) is weakly continuous as integral of a continuous function $f$. □

PROOF OF (54). 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 $\chi^2$-divergence (32), for any feasible solution $\pi, \pi'$ of (53), 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 $k$th moment of $\pi$. By existing results in approximation theory (see [4]), 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 $|\int |\theta| \pi'(d\theta) - \int |\theta| \pi'(d\theta)| \le \frac{C}{k} + \sqrt{tk! C^k}$. Choosing $k = c \frac{\log \frac{1}{t}}{\log\log \frac{1}{t}}$ for some small constant $c$ proves the upper bound of (54).

To show the lower bound part, by the duality between best polynomial approximation and moment matching (see e.g. [41, 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 [4]. Furthermore, matching first $k$ moments implies that the corresponding Gaussian mixture are close in $\chi^2$-divergence [4]: $\chi^2(\pi' * N(0,1) \| \pi * N(0,1)) \le \frac{C^k}{k!}$. Choosing $k = c \frac{\log \frac{1}{t}}{\log\log \frac{1}{t}}$ for some large constant $c$ proves the desired lower bound. □

## 7. Classical applications.

7.1. *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 $\pi$ is a distribution on $[-1,1]$ with density $\rho$ 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 $\rho$, 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 $\pi$ on $[-1,1]$ with its density $\rho$, we can set $T(\pi) = \rho(0)$ and view $\Pi$ 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 $\Pi$ and $C[-1,1]$ with the topology of uniform convergence, then $\Pi$ becomes a closed convex subset of $C[-1,1]$ and the Arzela-Ascoli theorem [10, IV.6.7] implies that $\Pi$ 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 $\Pi$ 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 of continuity here satisfies the following:

LEMMA 11. *There exist constants $c_0, c_1$ depending on $\beta$ 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 $\epsilon$, 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 $(\beta, L)$. 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 $(\beta, L)$ and $\epsilon = t^{\frac{2\beta}{2\beta+1}}$. □

Applying Theorem 1, we recover the classical result:

$$\inf_{\hat{T}} \sup_{\rho \in \mathcal{P}(\beta,L)} \mathbb{E}_{X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} f} |\hat{T}(X_1,\ldots,X_n) - \rho(0)|^2 \asymp n^{-\frac{2\beta}{2\beta+1}}. \tag{113}$$

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. [37, Section 1.2]).

7.2. *White Gaussian noise model.* In this section we revisit the Gaussian white noise model and re-derive the classical result of [20, 8] on the rate optimality (within constant factors) of linear estimators from convex duality. This derivation is self-contained without recourse to the general result on exponential family in Section 4. Let

$$dX_t = f(t)dt + \sigma dB_t, \quad t \in [0,1], \tag{114}$$

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 linear 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 8 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

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

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 $\sup_{f\in\mathcal{F}} \langle f, g\rangle - T(f) \ge 0 \ge \inf_{f\in\mathcal{F}} \langle f, g\rangle - T(f)$.[3] Therefore, we have

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

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

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

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

$$D(P_f \| P_{f'}) = \frac{1}{2\sigma^2}\|f - f'\|_2^2, \tag{115}$$

where $P_f$ denotes the law of $\{X_t : t \in [0,1]\}$ as in (114), and $C$ is an absolute constant. Thus we have shown that

$$\frac{\omega(\sigma)}{C} \le \sqrt{R^*(\sigma)} \le \omega(\sigma) \tag{116}$$

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

[3] 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.