A constrained risk inequality for general losses

John Duchi **Feng Ruan**
Stanford University

Abstract

We provide a general constrained risk inequality that applies to arbitrary non-decreasing losses, extending a result of Brown and Low [*Ann. Stat. 1996*]. Given two distributions P_0 and P_1, we find a lower bound for the risk of estimating a parameter $\theta(P_1)$ under P_1 given an upper bound on the risk of estimating the parameter $\theta(P_0)$ under P_0. The inequality is a useful tool, as its proof relies only on the Cauchy-Schwartz inequality, it applies to general losses, including optimality gaps in stochastic convex optimization, and it transparently gives risk lower bounds on super-efficient and adaptive estimators.

1 Introduction

In the theory of optimality for statistical estimators, we wish to develop the tightest lower bounds on estimation error possible. With this in mind, Cai and Low [4] highlight three desiderata make a completely satisfying efficiency benchmark or lower bound: it is distribution specific, in the sense that the lower bound is a function of the specific distribution P generating the data; the lower bound is uniformly achievable, in that there exist estimators achieving the lower bound uniformly over P in a class $\mathcal{P}$ of distributions; and there is a super-efficiency result, so that if an estimator $\widehat{\theta}$ achieves better risk than that indicated by the lower bound at a particular distribution P_0, there exist other distributions P_1 where the estimator has worse risk than the bound. While the Stein phenomenon [14] shows that satisfying all three of these desiderata precisely is impossible when estimating three- or higher-dimensional quantities, in the case of estimation of a real-valued functional $\theta(P)$ of a distribution P, one can often develop such results. Our purpose is to show a transparent proof of such lower bounds via a "hardest one-dimensional subproblem" argument [15]. Our hope is that this perspective is useful for explanation of the failures of super-efficient estimators, such as the Hodges' estimator, which must achieve inflated error away from points at which they are superefficient, or for researchers who wish to develop lower bounds for estimation.

In classical one-parameter families of distributions, such as location families or exponential families, the Fisher Information bounds satisfy our three desiderata of locality, achievability, and impossibility of super-efficiency, and in classical parametric problems, no estimator can be super-efficient on more than a set of measure zero points [9, 17, 18]. Similarly satisfying results hold in other problems. In the case of estimation of the value of a convex function f in white noise, for example, Cai and Low [4] provide precisely such a result, characterizing a local modulus of continuity with properties analogous to the Fisher information. For certain optimization problems, Chatterjee, Duchi, Lafferty, and Zhu [5] give a computational analogue of the Fisher Information that governs the difficulty of finding the minimizer of a function (though their techniques cannot provide optimality bounds for function values themselves).

Key to many of these results, and to understanding nonparametric functional estimation more broadly, is the *constrained risk inequality* of Brown and Low [3]. They develop a two-point inequality especially well-suited to providing lower bounds for adaptive nonparametric function estimation problems, and they also show that it gives quantitative bounds on the mean-squared error of super-efficient estimators for one-parameter problems, such as Gaussian mean estimation. Their work, however, relies strongly on using the squared error loss—that is, the quality of an estimator $\widehat{\theta}$ for a parameter θ is measured by $\mathbb{E}[(\widehat{\theta}-\theta)^2]$. In many applications, it is interesting to evaluate the error in other metrics, such as absolute error or the probability of deviation of the estimator $\widehat{\theta}$ away from the parameter θ by more than a specified amount. We

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

extend Brown and Low's work [3] by providing constrained risk inequalities that apply to general losses, allowing quantitative super-efficiency bounds: an estimator that overperforms on one distribution must necessarily suffer (much) worse loss on others. Our proofs rely only on the Cauchy-Schwarz inequality, so we can decouple the argument from the particular loss. There are more general results on lower bounds that demonstrate tradeoffs must exist, such as Lepskii's results on adaptivity in Gaussian white noise models [10] or [16, Theorem 6, App. A1]. While (similar to [3]) our approach does not always provide sharp constants, the risk inequality allows us to provide finite sample lower bounds for estimation under general losses, which brings us closer to the celebrated local asymptotic minimax theorem of Le Cam and Hájek [e.g. 9, 18, Ch. 8.7]. To illustrate our results, we provide applications to estimation of a normal mean, optimization of Lipschitz convex functions, and certain nonparametric estimation problems, deferring proofs to Section 5.

2 The constrained risk inequalities

We begin with the simplest version of our setting. Let P be a distribution on a sample space $\mathcal{Z}$, and let $\theta(P) \in \mathbb{R}^k$ be a parameter of interest. For predicting a point $v \in \mathbb{R}^k$ when the distribution is P, the estimator suffers loss

$$L(v, P) := \ell(\|v - \theta(P)\|_2), \tag{1}$$

where $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-decreasing scalar loss function. For $Z \sim P$ and an estimator $\widehat{\theta}$ of $\theta(P)$ based on Z, the risk of $\widehat{\theta}$ is then

$$R(\widehat{\theta}, P) := \mathbb{E}_P\left[L(\widehat{\theta}, P)\right] = \mathbb{E}_P\left[\ell(\|\widehat{\theta}(Z) - \theta(P)\|_2)\right].$$

The result to come relies on the similarity of two distributions to one another, and accordingly, we define the χ^2-affinity by

$$\rho\left(P_1 \| P_0\right) := \int \frac{dP_1^2}{dP_0} = \mathbb{E}_0\left[\frac{dP_1^2}{dP_0^2}\right] = \mathbb{E}_1\left[\frac{dP_1}{dP_0}\right].$$
$$D_{\chi^2}\left(P_1 \| P_0\right) := \rho\left(P_1 \| P_0\right) - 1.$$

where $\mathbb{E}_0$ and $\mathbb{E}_1$ denote expectation under P_0 and P_1, respectively. Introduce the notation shorthand $(x)_+ = \max(x, 0)$ for any scalar $x \in \mathbb{R}$. With these definitions, we have the following theorem, which gives a lower bound for the risk of the estimator $\widehat{\theta}$ on a distribution P_1 given an upper bound for its risk under P_0.

Theorem 1. *Assume $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ in the loss (1) is convex. Let $\theta_0 = \theta(P_0)$ and $\theta_1 = \theta(P_1)$, and define the separation $\Delta = 2\ell(\frac{1}{2}\|\theta_0 - \theta_1\|_2)$. If the estimator $\widehat{\theta}$ satisfies $R(\widehat{\theta}, P_0) \le \delta$, then*

$$R(\widehat{\theta}, P_1) \ge \left(\Delta^{1/2} - (\rho(P_1\|P_0) \cdot \delta)^{1/2}\right)_+^2. \tag{2}$$

A few corollaries are possible. The first applies to more general (non-convex) loss functions.

Corollary 1. *Let the conditions of Theorem 1 hold, except that $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ is an arbitrary non-decreasing function. Define $\Delta = \ell(\frac{1}{2}\|\theta_0 - \theta_1\|_2)$. If the estimator $\widehat{\theta}$ satisfies $R(\widehat{\theta}, P_0) \le \delta$, then*

$$R(\widehat{\theta}, P_1) \ge \left(\Delta^{1/2} - \left(\rho\left(P_1\|P_0\right)\delta\right)^{1/2}\right)_+^2.$$

We can also give a corollary with slightly sharper constants, which applies to the case that we measure error using a power loss.

Corollary 2. *In addition to the conditions of Theorem 1, assume $\ell(t) = t^k$ for some $k \in (0, \infty)$, and define $\Delta = \|\theta_0 - \theta_1\|_2$. If the estimator $\widehat{\theta}$ satisfies $R(\widehat{\theta}, P_0) \le \delta^k$, then*

$$R(\widehat{\theta}, P_1) \ge \begin{cases} \left(\Delta^{k/2} - (\rho(P_1\|P_0) \cdot \delta^k)^{1/2}\right)_+^2 & \text{if } k \le 2 \\ \left(\Delta - (\rho(P_1\|P_0) \cdot \delta^2)^{1/2}\right)_+^k & \text{if } k \ge 2. \end{cases} \tag{3}$$

An extension of Theorem 1 applies in somewhat more general scenarios, which, for example, will be useful in Section 3.2 when we consider stochastic convex optimization. An extension of the loss (1) is to consider any general loss $L(v, P)$ satisfying $\inf_v L(v, P) = 0$ for all distributions P, and then define the *loss distance* between distributions P_0, P_1 by

$$d_L(P_0, P_1) := \inf_v \left\{L(v, P_0) + L(v, P_1)\right\}. \tag{4}$$

For the risk $R(\widehat{\theta}, P) = \mathbb{E}_P[L(\widehat{\theta}, P)]$, we then have the following theorem.

Theorem 2. *Let the loss $L : \mathbb{R}^k \times \mathcal{P} \to \mathbb{R}_+$ be a general loss with loss distance d_L (4). If the estimator $\widehat{\theta}$ satisfies $R(\widehat{\theta}, P_0) \le \delta$, then*

$$R(\widehat{\theta}, P_1) \ge \left(d_L(P_0, P_1)^{1/2} - \left(\rho\left(P_1\|P_0\right)\delta\right)^{1/2}\right)_+^2.$$

3 Examples

We provide several examples that apply to estimation of one-dimensional quantities to illustrate our results. For the first two, we consider Gaussian mean estimation, where the results are simplest and cleanest to state, and which immediately demonstrate the failure of the Hodges' estimator. We also consider stochastic convex optimization, where we give an example application of the constrained risk inequality in Theorem 2. For the last set of examples, we consider super-efficient estimation in family of nonparametric models.

3.1 Gaussian mean estimation

We begin with two examples on one-dimensional Gaussian mean estimation to show the benefits of moving beyond squared error. For the first, we consider a zero-one loss function indicating whether the estimated mean is near the true mean. Fix $\sigma^2 > 0$ and let $X_1, \ldots, X_n$ be i.i.d. $P_\theta = \mathsf{N}(\theta, \sigma^2)$, and let $\ell(t) = \mathbf{1}\{|t| \ge \sigma/\sqrt{n}\}$, so that

$$R(\widehat{\theta}, P_\theta^n) = P_\theta^n \left(|\widehat{\theta}(X_1, \ldots, X_n) - \theta| \ge \frac{\sigma}{\sqrt{n}} \right),$$

where P_θ^n denotes the n-fold product of $X_i \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$. Now, let $\delta_n \in [0,1], \delta_n \to 0$ be an otherwise arbitrary sequence, and let $0 < c < 1$ be a fixed constant. Define the sequence of local alternative parameter spaces

$$\Theta_n := \left\{ \theta \in \mathbb{R} \mid 2\frac{\sigma}{\sqrt{n}} \le |\theta| \le \frac{\sigma}{\sqrt{n}} \sqrt{c \log \frac{1}{\delta_n}} \right\}.$$

We then have the following proposition.

Proposition 1. *Let $\widehat{\theta}_n : \mathbb{R}^n \to \mathbb{R}$ be a sequence of estimators satisfying $R(\widehat{\theta}_n, P_0^n) \le \delta_n$ for all n. Then*

$$\begin{aligned}
&\liminf_n \inf_{\theta \in \Theta_n} R(\widehat{\theta}_n, P_\theta^n) \\
&= \liminf_n \inf_{\theta \in \Theta_n} P_\theta^n \left(\sqrt{n} |\widehat{\theta}_n(X_1, \ldots, X_n) - \theta| \ge \sigma \right) \\
&= 1.
\end{aligned}$$

Remark The Le Cam–Hájek asymptotic minimax theorem (cf. [17, 9]) implies that for any symmetric, quasiconvex loss $\ell : \mathbb{R}^k \to \mathbb{R}_+$, if $\{P_\theta\}_{\theta \in \Theta}$ is a suitably regular family of distributions with Fisher information matrices I_θ, then for any $\theta_0 \in \operatorname{int} \Theta$ there exist sequences of prior densities $\pi_{n,c}$ supported on $\{\theta \in \mathbb{R}^k \mid \|\theta - \theta_0\|_2 \le c/\sqrt{n}\}$ such that

$$\begin{aligned}
&\liminf_{c \to \infty} \liminf_n \inf_{\widehat{\theta}_n} \int \mathbb{E}_\theta[\ell(\sqrt{n}(\widehat{\theta}_n - \theta))] d\pi_{n,c}(\theta) \\
&\qquad \ge \mathbb{E}[\ell(Z)] \quad \text{where} \quad Z \sim \mathsf{N}(0, I_{\theta_0}^{-1})
\end{aligned} \tag{5}$$

(see [9, Lemma 6.6.5] and also [17, Eq. (9)]). This in turn implies that for Lebesgue-almost-all θ, we have $\limsup_n \mathbb{E}_\theta[\ell(\sqrt{n}(\widehat{\theta}_n - \theta))] \ge \mathbb{E}[\ell(Z)]$ for $Z \sim \mathsf{N}(0, I_\theta^{-1})$. For the indicator loss $\ell(t) = \mathbf{1}\{|t| \ge \sigma\}$, these results imply that $\limsup_n P_\theta(|\sqrt{n}(\widehat{\theta}_n - \theta)| \ge \sigma) \ge 2\Phi(-1)$ for almost all θ in our normal mean setting, where Φ is the standard normal CDF. Proposition 1 strengthens this: if there exists a point of superefficiency with asymptotic probability of error 0, then there exists a large set of points with asymptotic probability of error 1.

Proof. Assume that n is large enough that $c \log \frac{1}{\delta_n} \ge 2$, and let $\theta \in \Theta_n$. A calculation then yields that

$$\rho(P_\theta^n \| P_0^n) = \exp\left(\frac{n\theta^2}{\sigma^2} \right) \le \exp\left(\frac{c\sigma^2 n \log \frac{1}{\delta_n}}{\sigma^2 n} \right) = \delta_n^{-c}.$$

We also have that $\ell(\frac{1}{2}|\theta|) = \mathbf{1}\{|\theta| \ge 2\sigma/\sqrt{n}\} = 1$, and substituting this into Corollary 1, we obtain $R(\widehat{\theta}, P_\theta^n) \ge \left(1 - \delta_n^{(1-c)/2}\right)_+^2$. As $c < 1$, this quantity tends to 1 as $n \to \infty$. □

Let us consider Corollary 2 for our second application. In this case, we consider estimating a Gaussian mean given $X_i \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$, but we use the absolute error $L(\theta, P) = |\theta - \theta(P)|$ as our loss as opposed to the typical mean squared error.

Proposition 2. *Let $\widehat{\theta} : \mathbb{R}^n \to \mathbb{R}$ be an estimator such that $R(\widehat{\theta}, P_0^n) \le \frac{\epsilon}{\sqrt{n}}$. Then for all $\alpha \in [0,1]$, there exists θ such that*

$$R(\widehat{\theta}, P_\theta^n) \ge \sqrt{\frac{\alpha}{n}} \left(\sqrt[4]{\log \frac{1}{\epsilon}} - \sqrt[4]{\frac{\epsilon^{2-2\alpha}}{\alpha}} \right)_+^2.$$

In particular, if $\epsilon \le 10^{-2}$, then there exists θ with $R(\widehat{\theta}, P_\theta^n) \ge \frac{1}{4}\sqrt{\frac{\log \frac{1}{\epsilon}}{n}}$.

Proof. Let $\alpha \in [0,1]$, to be chosen presently. Let $\theta \ge 0$ with $\theta^2 = \frac{\alpha}{n} \log \frac{1}{\epsilon}$. Then we have $\rho(P_\theta^n \| P_0^n) = \exp(n\theta^2) = \frac{1}{\epsilon^\alpha}$ and that $\Delta = |\theta|$ in the notation of Corollary 2. The corollary then implies

$$\begin{aligned}
R(\widehat{\theta}, P_\theta^n) &\ge \left(\sqrt{\theta} - \sqrt{\epsilon^{-\alpha}\epsilon/\sqrt{n}} \right)_+^2 \\
&= \frac{\sqrt{\alpha}}{\sqrt{n}} \left(\sqrt[4]{\log \frac{1}{\epsilon}} - \sqrt[4]{\frac{\epsilon^{2-2\alpha}}{\alpha}} \right)_+^2.
\end{aligned}$$

The second result of the proposition follows by taking $\alpha = 1/8$ and using the numerical fact that that $\sqrt[4]{\log \frac{1}{\epsilon}} - \sqrt[4]{8\epsilon^{7/4}} \ge \sqrt[4]{\log \frac{1}{\epsilon}/2}$ for $\epsilon \le 10^{-2}$. □

As an example consequence of Proposition 2, consider the Hodges' estimator

$$\widehat{\theta}_n^{\text{Hodges}} := \begin{cases} \overline{X}_n & \text{if } |\overline{X}_n| \ge n^{-1/4} \\ 0 & \text{otherwise,} \end{cases}$$

where $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$. At $\theta = 0$, this estimator satisfies

$$\begin{aligned}
\mathbb{E}[|\widehat{\theta}_n^{\text{Hodges}}|] &= \mathbb{E}[|\overline{X}_n| \mathbf{1}\{|\overline{X}_n| \ge n^{-1/4}\}] \\
&\le \sqrt{\frac{1}{n}} \cdot \sqrt{P_0(|\overline{X}_n| \ge n^{-1/4})} \\
&\le \sqrt{\frac{2}{n} \exp\left(-\frac{\sqrt{n}}{2} \right)}
\end{aligned}$$

by the standard tail bound that $\mathbb{P}(|Z| \geq t) \leq 2\exp(-t^2/2\sigma^2)$ for $Z \sim \mathsf{N}(0, \sigma^2)$. In particular, for all large enough n, there is a $\theta \in [0, n^{-1/2}]$ such that

$$\mathbb{E}_\theta[|\widehat{\theta}_n^{\text{Hodges}} - \theta|] \geq \frac{1}{8n^{1/4}} \gg \frac{1}{\sqrt{n}}.$$

3.2 Stochastic convex optimization

The generality of Theorem 2 may at first glance be opaque, but specializing it allows us to demonstrate certain instance-specific bounds for stochastic convex optimization, a building block of much of machine learning [19, 2]. Here, we receive i.i.d. data $X_i \overset{\text{iid}}{\sim} P$, and wish to minimize $f_P(\theta) := \mathbb{E}[F(\theta, X)]$ over $\theta \in \Theta$, where $F(\cdot, x)$ is convex for each x and Θ is a closed convex set, and the loss is the optimality gap

$$L(\theta, P) := f_P(\theta) - \inf_{\theta^\star \in \Theta} f_P(\theta^\star).$$

The loss distance (4), $d_L(P_0, P_1) = \inf_{\theta\in\Theta}\{f_0(\theta) + f_1(\theta) - f_0^\star - f_1^\star\}$ is then identical to a criterion Agarwal et al. [1] develop for stochastic optimization.

We consider 1-Lipschitz-continuous objectives, which are the "standard" for stochastic convex optimization [12, 1, 7], where the minimax rate is $1/\sqrt{n}$. Let the data $X \in \{-1, 1\}$ and let $F(\theta, x) = (1 - x\theta)_+$, the hinge loss for support vector machines, setting $P_\gamma(X = x) = \frac{1+x\gamma}{2}$, for $\gamma \in (0, 2/5)$. Then for $\epsilon \in [0, 1]$, $\rho(P_{\gamma-\epsilon} \| P_\gamma) = 1 + \frac{4\epsilon^2}{1-\gamma^2} \leq \exp(5\epsilon^2)$, and

$$d_L(P_\gamma, P_{\gamma-\epsilon}) = \begin{cases} 0 & \text{if } \epsilon \leq \gamma \\ 2\gamma & \text{if } \epsilon > \gamma. \end{cases}$$

Now, suppose an estimator $\widehat{\theta}_n$ given n i.i.d. observations X_i satisfies $R(\widehat{\theta}_n, P_\gamma) \leq \frac{\delta}{\sqrt{n}}$, where $\delta < 1$. Then Theorem 2 implies that if $\gamma < \epsilon$,

$$\begin{aligned} R(\widehat{\theta}_n, P_{\gamma-\epsilon}) &= \mathbb{E}_{\gamma-\epsilon}[f_{\gamma-\epsilon}(\widehat{\theta}_n) - f_{\gamma-\epsilon}^\star] \\ &\geq \left(\sqrt{2\gamma} - \sqrt{\exp(5n\epsilon^2)\delta/\sqrt{n}}\right)_+^2. \end{aligned}$$

Choosing $\epsilon_n^2 = \frac{\log(1/\delta)}{5n}$ then implies the next corollary.

Corollary 3. *Let F be the hinge loss and P_γ be as above, and let $\delta > 0$. If $\gamma^2 \in [\frac{1}{n}, \frac{\log(1/\delta)}{5n}]$ and $R(\widehat{\theta}_n, P_\gamma) \leq \frac{\delta}{\sqrt{n}}$, then*

$$R(\widehat{\theta}_n, P_{\gamma-\epsilon_n}) \geq (\sqrt{2} - 1)^2\gamma.$$

In particular, there are settings of P_0 where achieving better than the minimax rate necessarily leads to degraded convergence by a factor of $\sqrt{\log(1/\delta)}$. Nonetheless, other scenarios allow stronger convergence guarantees: the empirical risk minimizer $\widehat{\theta}_n = \mathrm{argmin}_\theta \sum_{i=1}^n (1 - X_i\theta)_+$ always achieves $R(\widehat{\theta}_n, P) \leq O(1)/\sqrt{n}$ [13, Ch. 5]. Letting $\theta_\gamma = \mathrm{argmin}_\theta \mathbb{E}_{P_\gamma}[(1 - \theta X)_+] = \mathrm{sign}(\gamma)$, Hoeffding's inequality implies $P_\gamma(\widehat{\theta}_n \neq \mathrm{sign}(\gamma)) \leq \exp(-n\gamma^2/2)$. Moreover, $R(\theta, P_\gamma) \leq 2\gamma$ for all $\theta \in [-1, 1]$, and so we obtain

$$R(\widehat{\theta}_n, P_\gamma) \leq O(1) \min\left\{\gamma, \frac{1}{\sqrt{n}}, \exp(-n\gamma^2/2)\right\}.$$

In brief, while Corollary 3 shows that beating the benchark minimax rate $1/\sqrt{n}$ implies efficiency losses for *some* distributions, it also highlights the importance of more nuanced notions of complexity: more local notions are essential.

3.3 Super-efficient estimation in nonparametric models

It is often interesting to derive efficiency lower bounds outside of standard parametric models; it is our experience that students are frequently curious about such quantities, especially when they have seen only Fisher-information-based lower bounds. Conveniently, we can also apply our results to estimation of functionals in general non-parametric models. In this case, we focus on quantities where the classical asymptotic normality results apply, so that there do indeed exist classically efficient estimators and an analogue of the Le Cam–Hájek local asymptotic minimax theorems. We first present a general result that applies to appropriately smooth parameters of the underlying distribution, which we subsequently specialize to estimation of the mean of an arbitrary distribution with finite variance. We adapt the classical idea of Stein [15], which constructs hardest one-dimensional subproblems, following the treatment of van der Vaart [18, Chapter 25].

To set the stage, consider estimation of a parameter $\theta(P_0) \in \mathbb{R}$ of a distribution P_0 on the space $\mathcal{Z}$. Letting $\mathcal{P}$ denote the collection of all distributions on $\mathcal{Z}$, we consider sub-models $\mathcal{P}_0 \subset \mathcal{P}$ around P_0 defined in terms of local perturbations of P_0. In particular, let $\mathcal{G} \subset L^2(P_0)$ consist of those functions $g : \mathcal{Z} \to \mathbb{R}$ satisfy $\mathbb{E}_0[g(Z)] = 0$ and $\mathbb{E}_0[g(Z)^2] < \infty$. For bounded functions $g \in \mathcal{G}$, we may consider tilts of the distribution P_0 of the form

$$dP(z) = (1 + tg(z))dP_0(z)$$

for small t; however, as g may be unbounded, we require a bit more care. Following [18, Example 25.16], we let $\phi : \mathbb{R} \to [0, 2]$ be any $\mathcal{C}^3$ function satisfying $\phi(1) = 1$, $\phi'(1) = 1$, and for which both $\|\phi'\|_\infty \leq K$ and $\|\phi''\|_\infty \leq K$ for a constant K; for example, $\phi(t) = 2/(1 + e^{-2t})$ suffices. For any $g \in \mathcal{G}$, define the tilted distribution

$$dP_{t,g}(z) := \frac{1}{C_t}\phi(tg(z))dP_0(z) \quad \text{where} \quad C_t = \int \phi(tg(z))dP_0(z). \tag{6}$$

The following lemma describes the divergence of $P_{t,g}$ from P_0 (see Section 5.5 for proof).

Lemma 1. *Let $g \in \mathcal{G}$ and P_0 and $P_{t,g}$ be as defined in Eq. (6). Then*

$$D_{\chi^2}(P_{t,g} \| P_0) = 1 + t^2\mathbb{E}_0[g(Z)^2] + o(t^2) \quad \text{and} \quad |C_t - 1| \le \frac{K}{2}t^2\mathbb{E}_0[g(Z)^2].$$

With this setting, let us assume that our parameter θ of interest is smooth in the underlying perturbation (6), meaning that there exists an *influence function* $\dot{\theta}_0 : \mathcal{Z} \to \mathbb{R}$, $\dot{\theta}_0 \in L^2(P_0)$, with $\mathbb{E}_0[\dot{\theta}_0(Z)] = 0$ such that

$$\theta(P_{t,g}) = \theta(P_0) + t\mathbb{E}_0[\dot{\theta}_0(Z)g(Z)] + o(t) \tag{7}$$

as $t \to 0$, that is, $\theta(P_{t,g})$ has a linear first-order expansion in L^2 based on $\dot{\theta}_0$. For example, the mean $\theta(P) = \mathbb{E}_P[Z]$ has the identity mapping $\dot{\theta}_0(Z) = Z - \mathbb{E}_P[Z]$. For more on such linear expansions and their importance and existence, see [18, Chapter 25]. In short, however, the influence function allows extension of the Fisher Information from classical problems, and by defining $I_0^{-1} := \mathbb{E}_{P_0}[\dot{\theta}_0(Z)^2]$, one has the analogue of the local minimax lower bound (5) that there exist sequences of prior densities π_n supported on $\{t \in \mathbb{R} \mid |t| \le 1/\sqrt{n}\}$ such that

$$\sup_{g \in \mathcal{G}} \liminf_n \inf_{\widehat{\theta}_n} \int \mathbb{E}_{P_{t,g}^n}[\ell(\sqrt{n}(\widehat{\theta}_n - \theta(P_{t,g})))]d\pi_n(t) \ge \mathbb{E}[\ell(Z)] \quad \text{where} \quad Z \sim \mathsf{N}(0, I_0^{-1}). \tag{8}$$

The supremum above may be taken to be over only scalar multiples of the function $\dot{\theta}_0$.

3.3.1 Non-convergence in probability: the general case

We now come to our super-efficiency result, which we will specialize to the nonparametric mean presently. Essentially the weakest typical form of convergence of estimators is convergence in probability, which is of course implied by convergence in mean-square or absolute error. As our general constrained risk inequality (Corollary 1) handles this case without challenge, and because lower bounds on the probability of error are strong, we focus on the zero-one error. Let $K < \infty$ be an arbitrary constant, and for each n, define the loss function $\ell(t) = \mathbf{1}\{\sqrt{n}|t| \ge K\}$, so that

$$R(\widehat{\theta}, P^n) = P^n\left(\sqrt{n}|\widehat{\theta}(Z_1, \ldots, Z_n) - \theta(P)| \ge K\right).$$

Under the assumption that $\widehat{\theta}_n$ is a super-efficient sequence of estimators under P_0, we will show that for essentially *all* non-trivial local alternatives, defined by the tilting (6), the estimators $\widehat{\theta}_n$ have probability of error tending to 1.

Making this more precise, consider the subset

$$\mathcal{G}_0 := \{g \in \mathcal{G} \mid \mathbb{E}_0[\dot{\theta}_0(Z)g(Z)] \neq 0, \mathbb{E}_0[g(Z)^2] \le 1\}, \tag{9}$$

that is, those functions $g \in \mathcal{G}$ for which the perturbation of $\theta(P_0)$ to $\theta(P_{t,g})$ is non-trivial as $t \to 0$, by the first-order expansion (7). Let us suppose that $R(\widehat{\theta}_n, P_0^n) \le \delta_n$ for all n, where $\delta_n \to 0$ and $\frac{1}{n}\log\frac{1}{\delta_n} \to 0$ (this last assumption is simply to make our argument simpler). Now, let $B > 2$ and $c \in (0, 1)$ be otherwise arbitrary constants, and for each $g \in \mathcal{G}_0$, define the set of local alternative distributions

$$\mathcal{P}_{n,g} := \left\{P_{t,g} \in \mathcal{P} \mid \frac{K^2}{n} \cdot \frac{B^2}{\mathbb{E}_0[\dot{\theta}_0(Z)g(Z)]^2} \le t^2, \quad \text{and} \quad t^2 \le \frac{c}{n}\log\frac{1}{\delta_n}\right\}. \tag{10}$$

We have the following proposition.

Proposition 3. *Let $\widehat{\theta}_n : \mathcal{Z}^n \to \mathbb{R}$ be a sequence of estimators satisfying $R(\widehat{\theta}_n, P_0^n) \le \varepsilon_n$, where $\epsilon_n \to 0$. Let $\delta_n \ge \varepsilon_n$ be any sequence satisfying $\delta_n \to 0$ and $n^{-1}\log\delta_n \to 0$. Then*

$$\begin{aligned}&\inf_{g \in \mathcal{G}_0} \liminf_n \inf_{P \in \mathcal{P}_{n,g}} R(\widehat{\theta}_n, P^n) \\ &= \inf_{g \in \mathcal{G}_0} \liminf_n \inf_{P \in \mathcal{P}_{n,g}} P^n\left(\sqrt{n}|\widehat{\theta}_n(Z_1^n) - \theta(P)| \ge K\right) \\ &= 1.\end{aligned}$$

Remark This result parallels Proposition 1, applying to nonparametric estimators. In comparison with the local asymptotic minimax result (8), we see the stronger result that super-efficiency at a single distribution for the zero-one error implies that asymptotically, the loss is as large as possible for a wide range of alternative distributions.

Proof. Fix $g \in \mathcal{G}_0$, and let $\theta_t = \theta(P_{t,g})$ and $\theta_0 = \theta(P_0)$ be parameters of interest. For shorthand, define $\Delta = \mathbb{E}_0[\dot{\theta}_0(Z)g(Z)] \neq 0$, so that $\theta_t = \theta_0 + (1 + o(1))t\Delta$ as $t \to 0$. By Lemma 1, we have that

$$\rho(P_{t,g}^n \| P_0^n) = \left(1 + (1 + o(1))t^2\mathbb{E}_0[g(Z)^2]\right)^n$$

as $t \to 0$, so that if $\mathbb{E}_0[g(Z)^2] \le 1$,

$$\begin{aligned}&\sup_t \left\{\rho(P_{t,g}^n \| P_0^n) \mid t^2 \le \frac{c}{n}\log\frac{1}{\delta_n}\right\} \\ &\le \left(1 + (1 + o(1))\frac{c}{n}\log\frac{1}{\delta_n}\right)^n \\ &\le \exp\left((1 + o(1))c\log\frac{1}{\delta_n}\right) = \delta_n^{-c+o(1)}\end{aligned} \tag{11}$$

as $n \to \infty$. Note that as $B > 2$, by the definition (7) of an influence function, we have for all t satisfying $\frac{BK}{|\Delta|} \le \sqrt{n}|t| \le \sqrt{c \log \frac{1}{\delta_n}}$ that

$$\begin{aligned}
&\ell(|\theta_t - \theta_0|/2) \\
&= \mathbf{1}\{\sqrt{n}|\theta_t - \theta_0| \ge 2K\} \\
&= \mathbf{1}\left\{\sqrt{n}\left|\frac{BK}{\sqrt{n}}(1 + o(1))\Delta\right| \ge 2K\right\} \\
&= \mathbf{1}\{|BK \pm o(1)| \ge 2K\} = 1 \ \text{ for large enough } n,
\end{aligned}$$

where the final equality holds because $B > 2$. Applying Corollary 1 and inequality (11), we thus obtain for large enough n, all $P \in \mathcal{P}_{n,g}$ satisfy

$$R(\widehat{\theta}_n, P^n) \ge \left(1 - \sqrt{\delta_n^{-c+o(1)}\delta_n}\right)_+^2,$$

which tends to 1 as $n \to \infty$ whenever $\delta_n \to 0$ and $c < 1$. □

3.3.2 Non-convergence in probability for the mean

Proposition 3 is abstract, so we make it more concrete by considering mean estimation for distributions with variance 1. Let P_0 be a distribution on $\mathbb{R}$ with $\mathbb{E}_0[Z] = 0$ and $\mathrm{Var}_0(Z) = 1$. In this case, the influence function is the identity mapping $\dot{\theta}_0(z) = z$. Let $0 < K < \infty$ be any constant. In this case, the family $\mathcal{G}_0$ of nontrivial perturbations (9) is precisely those with nonzero covariance with the random variable Z,

$$\mathcal{G}_0 = \Big\{g : \mathbb{R} \to \mathbb{R} \mid \mathbb{E}_0[g(Z)] = 0, \mathbb{E}_0[g(Z)^2] \le 1, \\ \text{and } \mathbb{E}_0[Zg(Z)] \ne 0\Big\}.$$

We thus have the following corollary, which applies to the tilted families $\mathcal{P}_{n,g}$ as above (10).

Corollary 4. *Let $\widehat{\theta}_n : \mathcal{Z}^n \to \mathbb{R}$ be any sequence of estimators such that $P_0^n(\sqrt{n}|\widehat{\theta}_n| \ge K) \le \varepsilon_n$, where $\epsilon_n \to 0$. Let $\delta_n \ge \varepsilon_n$ be any sequence satisfying $\delta_n \to 0$ and $n^{-1}\log\delta_n \to 0$. Then*

$$\inf_{g \in \mathcal{G}_0} \liminf_n \inf_{P \in \mathcal{P}_{n,g}} P^n\left(\sqrt{n}|\widehat{\theta}_n - \mathbb{E}_P[Z]| \ge K\right) = 1.$$

In short, we see the expected result: if any estimator achieves even the in-probability convergence $\widehat{\theta}_n = o_P(1/\sqrt{n})$ at $\theta = 0$, then there must be a large collection of distributions where the *best* performance of the estimator across the entire collection must be worse than the typical $\sqrt{n}$-rate of convergence.

4 Discussion

We have provided an extension of Brown and Low's constrained risk inequality [3], showing how to provide risk inequalities for general losses. Our results on efficient non-parametric estimators in Section 3.3 immediately extend beyond 0-1 losses. For example, consider estimating a parameter $\theta(P_0)$ of a distribution P_0 where θ has influence function $\dot{\theta}_0 : \mathbb{R} \to \mathbb{R}$, and assume the estimator sequence $\widehat{\theta}_n : \mathbb{R}^n \to \mathbb{R}$ satisfies

$$\mathbb{E}_{P_0^n}\left[|\widehat{\theta}_n - \theta(P_0)|\right] \le \sqrt{\frac{\delta_n}{n}}$$

where $\delta_n \to 0$. Then for the family $\mathcal{G}_0$ consisting of $g : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}_0[g(Z)] = 0$, $\mathbb{E}_0[g(Z)^2] \le 1$, and $\mathbb{E}_0[\dot{\theta}_0(Z)g(Z)] \ne 0$, we can consider an analogue of the tilted family (10) where for $0 < c_0 < c_1 < 1$ we define

$$\mathcal{P}_{n,g} = \left\{P_{t,g} \mid c_0 \frac{\log\frac{1}{\delta_n}}{n} \le t^2 \le c_1 \frac{\log\frac{1}{\delta_n}}{n}\right\}.$$

Then by Corollary 2 and an argument analogous to that for Proposition 2, there exists a numerical constant $K > 0$ such that for all $g \in \mathcal{G}_0$,

$$\begin{aligned}
&\liminf_n \inf_{P \in \mathcal{P}_{n,g}} \sqrt{\frac{n}{\log\frac{1}{\delta_n}}}\mathbb{E}_{P^n}\left[|\widehat{\theta}_n - \theta(P)|\right] \\
&\ge K|\mathbb{E}_0[\dot{\theta}_0(Z)g(Z)]| > 0.
\end{aligned}$$

The one-dimensional lower bounds we have provided are, we hope, transparent—relying only on the Cauchy-Schwarz inequality—and easy to apply to a range of estimation settings, making them well-suited to pedagogical situations. It is possible to follow Brown and Low's work [3] to give non-adaptivity results in nonparametric function estimation [cf. 6, 10, 11, 3, 16], with relatively straightforward derivations (though of course, these results are known). We hope that our constrained risk inequalities for general losses may lead to easier understanding of such issues in other areas as well.

5 Proofs

5.1 Proof of Theorem 1

It is no loss of generality to assume that $\widehat{\theta}(z) \in [\theta_0, \theta_1] = \{t\theta_0 + (1-t)\theta_1 \mid t \in [0,1]\}$ for all z: letting $\mathrm{proj}(\theta) = \mathrm{argmin}_{\theta'}\{\|\theta - \theta'\|_2 \mid \theta' \in [\theta_0, \theta_1]\}$ be the projection of θ onto the segment $[\theta_0, \theta_1]$, then $\|\mathrm{proj}(\theta) - \theta_i\|_2 \le \|\theta - \theta_i\|_2$ for $i \in \{0,1\}$ by standard properties of convex projections [8].

For any $\theta \in [\theta_0, \theta_1]$, which must satisfy $\theta = t\theta_0 + (1-t)\theta_1$, we have

$$\begin{aligned}
&\sqrt{\ell(\|\theta - \theta_0\|_2)} + \sqrt{\ell(\|\theta - \theta_1\|_2)} \\
&= \sqrt{\ell((1-t)\|\theta_0 - \theta_1\|_2)} + \sqrt{\ell(t\|\theta_0 - \theta_1\|_2)} \\
&\geq \sqrt{\ell((1-t)\|\theta_0 - \theta_1\|_2) + \ell(t\|\theta_0 - \theta_1\|_2)} \\
&\geq \sqrt{2\ell\left(\frac{1}{2}\|\theta_0 - \theta_1\|_2\right)} \qquad (12)
\end{aligned}$$

as $\ell(ta) + \ell((1-t)a)$ is minimized by $t = \frac{1}{2}$ for any $a \geq 0$ (by convexity of ℓ). Using the majorization inequality (12) and our without loss of generality assumption that $\widehat{\theta}(z) \in [\theta_0, \theta_1]$ for all $z \in \mathcal{Z}$, we thus have

$$\begin{aligned}
&\mathbb{E}_1\left[\ell(\|\widehat{\theta} - \theta_0\|_2)^{1/2}\right] + \mathbb{E}_1\left[\ell(\|\widehat{\theta} - \theta_1\|_2)^{1/2}\right] \\
&\geq \sqrt{2\ell\left(\frac{1}{2}\|\theta_0 - \theta_1\|_2\right)} = \Delta^{1/2}. \qquad (13)
\end{aligned}$$

Now, using the Cauchy–Schwarz inequality and rearranging inequality (13), we have

$$\begin{aligned}
R(\widehat{\theta}, P_1) &= \mathbb{E}_1\left[\ell(\|\widehat{\theta} - \theta_1\|_2)\right] \\
&\geq \mathbb{E}_1\left[\ell(\|\widehat{\theta} - \theta_1\|_2)^{1/2}\right]^2 \\
&\geq \left(\Delta^{1/2} - \mathbb{E}_1\left[\ell(\|\widehat{\theta} - \theta_0\|_2)^{1/2}\right]\right)_+^2.
\end{aligned}$$

Finally, a likelihood ratio change of measure yields that

$$\begin{aligned}
&\mathbb{E}_1\left[\ell(\|\widehat{\theta} - \theta_0\|_2)^{1/2}\right] \\
&= \mathbb{E}_0\left[\frac{dP_1}{dP_0}\ell(\|\widehat{\theta} - \theta_0\|_2)^{1/2}\right] \\
&\leq \mathbb{E}_0\left[\frac{dP_1^2}{dP_0^2}\right]^{1/2} \mathbb{E}_0\left[\ell(\|\widehat{\theta} - \theta_0\|_2)\right]^{1/2} \\
&= \left(\rho\left(P_1 \| P_0\right) R(\widehat{\theta}, P_0)\right)^{1/2}.
\end{aligned}$$

This gives the lower bound (2) once we use that $R(\widehat{\theta}, P_0) \leq \delta$.

5.2 Proof of Corollary 1

The proof is nearly identical to that of Theorem 1, with one minor change. Instead of the majorization inequality (12), we have for all $t \in [0, 1]$ that

$$\begin{aligned}
&\ell(t\|\theta_0 - \theta_1\|_2) + \ell((1-t)\|\theta_0 - \theta_1\|_2) \\
&\geq \ell\left(\frac{1}{2}\|\theta_0 - \theta_1\|_2\right).
\end{aligned}$$

Substituting this and the definition $\Delta = \ell(\frac{1}{2}\|\theta_0 - \theta_1\|_2)$, then following the proof of Theorem 1, *mutatis mutandis*, gives the corollary.

5.3 Proof of Corollary 2

The proof is again identical to Theorem 1, except that we consider separately the cases $k \in (0, 2]$ and $k > 2$. In the first case that $0 < k \leq 2$, we replace the majorization inequality (12) for $\theta = t\theta_0 + (1-t)\theta_1$, where $t \in [0, 1]$, with the inequality

$$\begin{aligned}
&L(\theta, P_0)^{1/2} + L(\theta, P_1)^{1/2} \\
&= \left[(1-t)^{k/2} + t^{k/2}\right] \|\theta_0 - \theta_1\|_2^{k/2} \geq \|\theta_0 - \theta_1\|_2^{k/2}.
\end{aligned}$$

Using $\Delta = \|\theta_0 - \theta_1\|_2$ and tracing the proof of Theorem 1 then gives the first inequality (3). For the second inequality, the case $k \in (2, \infty)$, we may apply the first case that $k \leq 2$ and Hölder's inequality. Indeed, by the assumption that $R(\widehat{\theta}, P_0) \leq \delta^k$, we have

$$\mathbb{E}_0\left[\|\widehat{\theta} - \theta_0\|_2^2\right] \leq \mathbb{E}_0\left[\|\widehat{\theta} - \theta_0\|_2^k\right]^{2/k} \leq \delta^2.$$

Applying the result for $k = 2$ in the first case of inequality (3) yields

$$\begin{aligned}
R(\widehat{\theta}, P_1) &\geq \mathbb{E}_1\left[\|\widehat{\theta} - \theta_1\|_2^2\right]^{k/2} \\
&\geq \left(\Delta - (\rho(P_1 \| P_0)\delta^2)^{1/2}\right)_+^k.
\end{aligned}$$

5.4 Proof of Theorem 2

By analogy with inequality (12) in the proof of Theorem 1, for any v we have

$$\begin{aligned}
\sqrt{L(v, P_0)} + \sqrt{L(v, P_1)} &\geq \sqrt{L(v, P_0) + L(v, P_1)} \\
&\geq \sqrt{d_L(P_0, P_1)},
\end{aligned}$$

and so (analogizing inequality (13)) we have

$$\mathbb{E}_1\left[L(\widehat{\theta}, P_0)^{1/2}\right] + \mathbb{E}_1\left[L(\widehat{\theta}, P_1)^{1/2}\right] \geq \sqrt{d_L(P_0, P_1)}.$$

Then using Cauchy-Schwarz and a likelihood ratio change of measure exactly as in the proof of Theorem 1 yields

$$\begin{aligned}
R(\widehat{\theta}, P_1) \geq \mathbb{E}_1\left[L(\widehat{\theta}, P_1)\right] &\geq \mathbb{E}_1\left[L(\widehat{\theta}, P_1)^{1/2}\right]^2 \\
&\geq \left(d_L(P_0, P_1)^{1/2} - \mathbb{E}_1\left[L(\widehat{\theta}, P_0)^{1/2}\right]\right)_+^2 \\
&\geq \left(d_L(P_0, P_1)^{1/2} - \left(\rho\left(P_1 \| P_0\right) R(\widehat{\theta}, P_0)\right)^{1/2}\right)_+^2.
\end{aligned}$$

5.5 Proof of Lemma 1

By the boundedness assumptions on ϕ' and ϕ'', Taylor's theorem implies that

$$\begin{aligned}
&|\phi(t) - 1| \leq \|\phi'\|_\infty |t| \leq K|t| \\
&\text{and } \ |\phi(t) - 1 - t| \leq \frac{1}{2}\|\phi''\|_\infty t^2 \leq \frac{1}{2}Kt^2
\end{aligned}$$

for all $t \in \mathbb{R}$. Thus we have

$$\begin{aligned} C_t &= \int \phi(tg(z)) dP_0(z) \\ &\leq \int (1 + tg(z)) dP_0(z) + \frac{K}{2} \int t^2 g(z)^2 dP_0(z) \\ &\leq 1 + \frac{Kt^2}{2} \mathbb{E}_0[g(Z)^2], \end{aligned}$$

and similarly

$$\begin{aligned} C_t &= \int \phi(tg(z)) dP_0(z) \\ &\geq \int (1 + tg(z)) dP_0(z) - \frac{K}{2} \int t^2 g(z)^2 dP_0(z) \\ &\geq 1 - \frac{Kt^2}{2} \mathbb{E}_0[g(Z)^2]. \end{aligned}$$

Let $\sigma^2 = \mathbb{E}_0[g(Z)^2]$ for shorthand. Considering the χ^2-divergence, we have

$$D_{\chi^2}\left(P_{t,g} \| P_0\right) = \int (\phi(tg(z))/C_t - 1)^2 dP_0(z),$$

and the integrand has the bound

$$\begin{aligned} \left(\frac{\phi(tg(z))}{C_t} - 1\right)^2 &\leq \left(\frac{1 + K|tg(z)|}{1 - Kt^2\sigma^2} - 1\right)^2 \\ &\leq \frac{2K^2t^2}{(1 - Kt^2\sigma^2)^2} (g(z)^2 + t^2\sigma^4), \end{aligned}$$

and

$$\begin{aligned} &\lim_{t \to 0} \frac{1}{t^2} \left(\frac{\phi(tg(z))}{C_t} - 1\right)^2 \\ &= \lim_{t \to 0} \frac{1}{t^2} \left(\frac{1 + tg(z) + O(t^2)}{1 - O(t^2)} - 1\right)^2 = g(z)^2. \end{aligned}$$

Lebesgue's dominated convergence theorem implies that

$$\lim_{t \to 0} \frac{1}{t^2} D_{\chi^2}\left(P_{t,g} \| P_0\right) = \mathbb{E}_0[g(Z)^2].$$

References

[1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[2] L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale learning. *SIAM Review*, 60(2):223–311, 2018.

[3] L. D. Brown and M. G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *Annals of Statistics*, 24 (6):2524–2535, 1996.

[4] T. Cai and M. Low. A framework for estimating convex functions. *Statistica Sinica*, 25:423–456, 2015.

[5] S. Chatterjee, J. Duchi, J. Lafferty, and Y. Zhu. Local minimax complexity of stochastic convex optimization. In *Advances in Neural Information Processing Systems 29*, 2016.

[6] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90 (432):1200–1224, 1995.

[7] J. C. Duchi. Introductory lectures on stochastic convex optimization. In *The Mathematics of Data*, IAS/Park City Mathematics Series. American Mathematical Society, 2018.

[8] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II.* Springer, New York, 1993.

[9] L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[10] O. V. Lepskii. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications*, 35(3):454–466, 1990.

[11] O. V. Lepskii. Asymptotically minimax adaptive estimation I: upper bounds. Optimally adaptive estimates. *Theory of Probability and Its Applications*, 36:682–697, 1991.

[12] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[13] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory.* SIAM and Mathematical Programming Society, 2009.

[14] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206, 1956.

[15] C. Stein. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 187–195, 1956.

[16] A. Tsybakov. Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Annals of Statistics*, 26(6):2420–2469, 1998.

[17] A. W. van der Vaart. Superefficiency. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, chapter 27. Springer, 1997.

[18] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic

Mathematics. Cambridge University Press, 1998.

[19] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.