Admissibility of solution estimators for stochastic optimization

Amitabh Basu* Tu Nguyen* Ao Sun*

January 23, 2019

Abstract

We look at stochastic optimization problems through the lens of statistical decision theory. In particular, we address admissibility, in the statistical decision theory sense, of the natural sample average estimator for a stochastic optimization problem (which is also known as the *empirical risk minimization (ERM) rule* in learning literature). It is well known that for general stochastic optimization problems, the sample average estimator may not be admissible. This is known as *Stein's paradox* in the statistics literature. We show in this paper that for optimizing stochastic linear functions over compact sets, the sample average estimator *is* admissible.

1 Introduction

A large class of stochastic optimization problems can be formulated in the following way:

$$\min_{x \in X} \{f(x) := \mathbb{E}_\xi[F(x, \xi)]\}, \tag{1.1}$$

where $X \subseteq \mathbb{R}^d$ is a fixed feasible region, ξ is a random variable taking values in $\mathbb{R}^m$, and $F : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$. We wish to solve this problem with access to independent samples of ξ. Consider two classical examples:

1. Consider a learning problem with access to labeled samples $(z, y) \in \mathbb{R}^n \times \mathbb{R}$ from some distribution and the goal is to find a function $f \in \mathcal{F}$ in a finitely parametrized hypothesis class $\mathcal{F}$ (e.g., all neural network functions with a fixed architecture) that minimizes expected loss, where the loss function is given by $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. One can model this using (1.1) by setting d to be the number of parameters for $\mathcal{F}$, $m = n + 1$, $X \subseteq \mathbb{R}^d$ is the subset that describes $\mathcal{F}$ via the parameters, and $F(f, (z, y)) = \ell(f(z), y)$.

2. When $d = m$, $F(x, \xi) = \|x - \xi\|^2$, $X = \mathbb{R}^d$ and ξ is distributed with mean μ, (1.1) becomes
$$\min_{x \in \mathbb{R}^d} \mathbb{E}_\xi[\|x - \xi\|^2] = \min_{x \in \mathbb{R}^d} \|x - \mathbb{E}[\xi]\|^2 + \mathbb{V}[\xi]$$
In particular, if one knows $\mu := \mathbb{E}[\xi]$, the optimal solution is given by $x = \mu$. Thus, this stochastic optimization problem becomes equivalent to the classical statistics problem of estimating the mean of the distribution of ξ.

*Department of Applied Mathematics and Statistics, The Johns Hopkins University. Amitabh Basu and Tu Nguyen were supported by NSF grant CMMI1452820 and ONR grant N000141812096.

In this paper, we will focus on (1.1) in the setting where $X \subseteq \mathbb{R}^d$ is a given compact (not necessarily convex) set, $m = d$, ξ has a *Gaussian distribution* with *unknown* mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$ denoted by $\xi \sim N(\mu, \Sigma)$, and $F(x, \xi) = \xi^T x$. So, we assume that the distribution of ξ is in the family of gaussian distributions, but we do not know which one in particular. Note that if an oracle provides μ exactly, the solution to (1.1) can be obtained by solving the *deterministic* optimization problem

$$\min\{\mu^T x : x \in X\}. \tag{1.2}$$

As mentioned before, although we do not know μ, we assume that we have access to data points $\xi^1, \xi^2, \ldots$ drawn independently from $N(\mu, \Sigma)$. **For sake of exposition, we first treat the case where Σ is the identity matrix. In Section 4, we will remove this assumption.** Note also that we do not make any convexity assumption on the feasible region. Thus, along with linear or convex optimization, we also capture non-convex feasible regions like mixed-integer non-linear optimization or linear complementarity constraints. We view the problem of solving (1.1) within a statistical decision theory framework. We briefly review relevant terminology from statistical decision theory below.

1.1 Statistical decision theory and admissibility

Statistical decision theory is a mathematical framework for modeling decision making in the face of uncertain or incomplete information. The starting point is to model the uncertainty or partial information by a set of *states of nature* denoted by Θ. One wishes to choose from a given set of possible *actions* $\mathcal{A}$; this is the decision making process. The interpretation is that given a state $\theta \in \Theta$, one wishes to choose an action that performs best in this state of nature. To take our stochastic optimization setting, the states of nature are given by $\mu \in \mathbb{R}^d$ and the set of actions $\mathcal{A}$ is the feasible region X. In other words, given that the uncertain objective is $F(x, \xi)$ with $\xi \sim N(\mu, I)$, one needs to select $x \in X$ that minimizes $f(x) := \mathbb{E}_\xi[F(x, \xi)]$. Within the general framework of statistical decision theory, one defines a *loss function* $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}_+$ to evaluate the performance of a particular action $a \in \mathcal{A}$ against a particular state of nature $\theta \in \Theta$[1]. The interpretation is that the smaller $\mathcal{L}(\theta, a)$ is, the better $a \in \mathcal{A}$ does with respect to the state $\theta \in \Theta$. In our setting of stochastic optimization, we take an action $\hat{x} \in X$. The natural way to evaluate its performance is to see how well it does with respect to problem (1.1), i.e., how close is $f(\hat{x})$ to the optimal value of (1.1).

Therefore, the following becomes a natural loss function for stochastic optimization:

$$\begin{aligned}\mathcal{L}(\mu, x) &:= f(x) - f(x(\mu)) \\ &= \mathbb{E}_{\xi\sim N(\mu,I)}[F(x, \xi)] - \mathbb{E}_{\xi\sim N(\mu,I)}[F(x(\mu), \xi)],\end{aligned} \tag{1.3}$$

[1]We caution the reader that the use of the words "loss" and "risk" in statistical decision theory are somewhat different from their use in machine learning literature. In machine learning, the function $F(x, \xi)$ is usually referred to as "loss" and the function $f(x)$ is referred to as "risk" in (1.1). Thus Example 1. above becomes a "risk minimization problem" with an associated "empirical risk minimization (ERM)" problem when one replaces the expectation by a sample average.

where $x(\mu)$ is an optimal solution to (1.1) when $\xi \sim N(\mu, I)$. In the case of an uncertain linear objective $F(x, \xi) = \xi^T x$, this would reduce to

$$\mathcal{L}(\mu, x) = \mu^T x - \mu^T x(\mu), \tag{1.4}$$

where $x(\mu)$ is an optimal solution to (1.2).

The *statistical* aspect of statistical decision theory comes from the fact that the state θ is not revealed directly, but only through data/observations based on θ that can be noisy or incomplete. Thus, one has to take the optimal decision $a \in \mathcal{A}$ (one that minimizes the loss function) without knowing θ but with only partial or uncertain information about θ. This is formalized by postulating a random variable whose distribution depends on θ that takes values in a *sample space* χ. This gives a parameterized family of probability distributions $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ on χ. After observing a realization $y \in \chi$ of this random variable, one forms an opinion about what the possible state is and one chooses an action $a \in \mathcal{A}$. Formally, we have a set of *decision rules*, i.e., maps $\delta : \chi \to \mathcal{A}$ giving an action $\delta(y) \in \mathcal{A}$ when data $y \in \chi$ is observed. To take our particular setting of stochastic optimization, one observes data points $\xi^1, \xi^2, \ldots, \xi^n$ that are i.i.d. realizations of $\xi \in N(\mu, I)$; thus, $\chi = \underbrace{\mathbb{R}^m \times \mathbb{R}^m \times \ldots \times \mathbb{R}^m}_{n \text{ times}}$ with distributions $\mathcal{P} := \{\underbrace{N(\mu, I) \times N(\mu, I) \times \ldots \times N(\mu, I)}_{n \text{ times}} : \mu \in \mathbb{R}^d\}$ on χ parameterized by the states $\mu \in \mathbb{R}^d$.

Finally, one evaluates decision rules by averaging over the data; formally, one defines the *risk function*

$$\mathcal{R}(\theta, \delta) := \mathbb{E}_{y \sim P_\theta}[\mathcal{L}(\theta, \delta(y))].$$

One can think of the risk function as mapping a decision rule to a nonnegative function on the class of distributions $\mathcal{P}$, or alternatively, a nonnegative function on the parameter space Θ; this function is sometimes called *the risk of the decision rule*. A decision rule is "good" if its risk has "low" values. The "best" possible decision rule would be a δ^* such that for any other decision rule δ', $\mathcal{R}(\theta, \delta^*) \leq \mathcal{R}(\theta, \delta')$ for all $\theta \in \Theta$, i.e., δ^* has risk that pointwise dominates the risk of any other decision rule. Usually such universally dominating decision rules do not exist.

We say that δ' *weakly dominates* δ if $\mathcal{R}(\theta, \delta') \leq \mathcal{R}(\theta, \delta)$ for all $\theta \in \Theta$. We say that δ' *dominates* δ if, in addition, $\mathcal{R}(\hat{\theta}, \delta') < \mathcal{R}(\hat{\theta}, \delta)$ for some $\hat{\theta} \in \Theta$. A decision rule δ is said to be *inadmissible* if there exists another decision rule δ' that dominates δ. A decision rule δ is said to be *admissible* if it is not dominated by any other decision rule. In-depth discussions of general statistical decision theory can be found in [1, 2, 7, 8].

1.2 Our result

We would like to study the admissibility of natural decision rules for solving (1.1). As explained above, we put this in the decision theoretical framework by setting the sample space $\chi = \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \ldots \times \mathbb{R}^d}_{n \text{ times}}$, where n is the number of i.i.d. observations one makes for $\xi \sim N(\mu, I)$. A decision rule is now a map $\delta : \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \ldots \times \mathbb{R}^d}_{n \text{ times}} \to X$. The class of

distributions is $\mathcal{P} = \{\underbrace{N(\mu, I) \times N(\mu, I) \times \ldots \times N(\mu, I)}_{n \text{ times}} : \mu \in \mathbb{R}^d\}$. The loss function is defined as in (1.3). In this paper, we wish to study the admissibility of the *sample average* decision rule δ_{SA} defined as

$$\delta_{SA}(\xi^1, \ldots, \xi^n) \in \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} F(x, \xi^i) : x \in X \right\} \tag{1.5}$$

In other words, δ_{SA} reports an optimal solution with respect to the sample average of the objective vectors. This is a standard procedure in stochastic optimization, and often goes by the name of *sample average approximation (SAA)*; in machine learning, it goes by the name of *Empirical Risk Minimization (ERM)*. To emphasize the dependence on the number of samples n, we introduce a superscript, i.e., δ^n_{SA} will denote the estimator based on the sample average of n observations. Moreover, for any $n \in \mathbb{N}$, let Δ^n be the set of all decision rules $\delta : \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \ldots \times \mathbb{R}^d}_{n \text{ times}} \to X$ such that $\mathbb{E}_{\xi^1, \ldots, \xi^n}[\delta(\xi^1, \ldots, \xi^n)]$ exists.

In turns out that there are simple instances of problem (1.1) where the sample average estimator is *not admissible*. Consider the setting of Example 2 in the Introduction, where $F(x, \xi) = \|x - \xi\|^2$ and $X = \mathbb{R}^d$. We again assume $\xi \sim N(\mu, I)$ with unknown μ. It is a simple calculation to obtain from (1.3) that $\mathcal{L}(\mu, x) = \|x - \mu\|^2$ in this case; thus, $x = \mu$ minimizes the loss. Consequently, the problem becomes the classical problem of estimating the mean of a Gaussian from samples under "squared loss". It can also be easily checked that the sample average decision rule solving (1.5) simply returns the empirical average of the samples, i.e., $\delta^n_{SA}(\xi^1, \ldots, \xi^n) = \overline{\xi}$ where $\overline{\xi} := \frac{1}{n} \sum_{i=1}^n \xi^i$ denotes the sample average. It is well-known that this sample average decision rule is *not admissible* if $d \geq 3$; this was first observed by Stein [9] and is commonly referred to as *Stein's paradox* in statistics literature. The so-called James-Stein estimator [6] can be shown to strictly dominate the sample average estimator; see [1, 7] for an exposition.

Our main theorem shows that for optimizing an uncertain linear objective over a fixed compact set there is no "Stein's paradox" phenomenon, i.e., the sample average solution is admissible.

THEOREM 1.1. *Consider problem* (1.1) *in the setting where X is a given compact set and $F(\xi, x) = \xi^T x$, and $\xi \sim N(\mu, I)$ with unknown μ. The sample average rule now simply becomes*

$$\delta^n_{SA}(\xi^1, \ldots, \xi^n) \in \arg\min\{\overline{\xi}^T x : x \in X\} \tag{1.6}$$

where $\overline{\xi} := \frac{1}{n} \sum_{i=1}^n \xi^i$ denotes the sample average of the observed objective vectors. For any $n \in \mathbb{N}$, δ^n_{SA} is admissible within Δ^n.

1.3 Comparison with previous work

The statistical decision perspective on stochastic optimization presented here was pioneered (as far as we are aware) in [3] and [4]. Between these two papers, [3] is closer to our work because the authors deal precisely with the question of admissibility of solution estimators for stochastic optimization.

In particular, the authors of [3] consider two different stochastic optimization problems: one where $X = \mathbb{R}^d$ and $F(x, \xi) = x^T Q x + \xi^T x$ for some fixed matrix positive definite matrix Q (i.e., unconstrained convex quadratic maximization), and the second one where X is the unit ball and $F(x, \xi) = \xi^T x$. ξ is again assumed to be distributed according a normal distribution $N(\mu, I)$ with unknown mean μ. They show that the sample average approximation is *not* admissible in general for the first problem, and it is admissible for the second problem. Note that the second problem is a special case of our setting. Observe that in both these cases, there is a closed-form solution to the deterministic version of the optimization problem, which helps to bring the problem into the domain of classical statistics literature for squared loss functions. This is not true for the general optimization problem we consider here (even if we restrict X to be a polytope, we get a linear program which, in general, has no closed form solution).

Another difference between our work and [3] is the following. In [3], the question of admissibility is addressed within a smaller subset of decision rules that are "decomposable" in the sense that any decision rule is of the form $\tau \circ \kappa$, where $\kappa : \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \ldots \times \mathbb{R}^d}_{} \to \mathbb{R}^d$ maps the data $\xi^1, \ldots, \xi^n$ to a vector $\hat{\mu} \in \mathbb{R}^d$ and then $\tau : \mathbb{R}^d \to X$ is of the form $\tau(\hat{\mu}) \in \arg\min\{\hat{\mu}^T x : x \in X\}$. In other words, one first estimates the mean of the uncertain objective (using any appropriate decision rule) and then uses this estimate to solve a deterministic optimization problem. In the follow-up work [4], the authors call such decision rules *Separate estimation-optimization (Separate EO) schemes* and more general decision rules as *Joint estimation-optimization (Joint EO) schemes*. Note that proving inadmissibility within separate EO schemes implies inadmissibility within joint EO schemes. On the other hand, establishing admissibility within joint EO schemes means defending against a larger class of decision rules.

In this paper, we establish admissibility of the sample average estimator within general decision rules (joint EO schemes in the terminology of [4]). The only condition we put on the decision rules is that of integrability, which is a minimum requirement needed to even define the risk of a decision rule.

2 Technical Tools

We first recall a basic fact from calculus.

LEMMA 2.1. *Let $F : \mathbb{R}^m \to \mathbb{R}$ be a twice continuously differentiable map such that $F(0) = 0$. Suppose $\nabla^2 F(0)$ is not negative semidefinite; in other words, there is a direction $d \in \mathbb{R}^d$ of positive curvature, i.e., $d^T \nabla^2 F(0) d > 0$. Then there exists $z \in \mathbb{R}^m$ such that $F(z) > 0$.*

Proof. If $\nabla F(0) \neq 0$, then there exist $\lambda > 0$ such that $F(z) > 0$ for $z = \lambda \nabla F(0)$ since $F(0) = 0$. Else, if $\nabla F(0) = 0$ then there exists $\lambda > 0$ such that $F(z) > 0$ for $z = \lambda d$, where d is the direction of positive curvature at 0. □

We will need the following central definition and result from statistics. See e.g., Section 6, Chapter 1 in [7].

DEFINITION 2.2. *A* statistic *is a function $T : \chi \to \mathbb{R}^m$, i.e., it is any function that maps the data to a vector (or a scalar if $m = 1$). Let $\mathcal{P}$ be a family of distributions on the sample space*

χ. *A* sufficient statistic for $\mathcal{P}$ *is a statistic on* χ *such that the conditional distribution on* χ *given* $T = t$ *does not depend on the distribution from* $\mathcal{P}$*, for all* $t \in \mathbb{R}^m$.

PROPOSITION 2.3. *Let* $\chi = \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \ldots \times \mathbb{R}^d}_{n \text{ times}}$ *and let* $\mathcal{P} = \{\underbrace{N(\mu, I) \times N(\mu, I) \times \ldots \times N(\mu, I)}_{n \text{ times}} : \mu \in \mathbb{R}^d\}$*, i.e.,* $(\xi^1, \ldots, \xi^n) \in \chi$ *are i.i.d samples from the normal distribution* $N(\mu, I)$*. Then* $T(\xi^1 \ldots, \xi^n) = \overline{\xi} := \frac{1}{n}\sum_{i=1}^n \xi^i$ *is a sufficient statistic for* $\mathcal{P}$.

We will also need the following useful property for the family of normal distributions $\{N(\mu, I) : \mu \in \mathbb{R}^m\}$. Indeed the following result is true for any *exponential family* of distributions; see Theorem 5.8, Chapter 1 in [7] for details.

THEOREM 2.4. *Let* $f : \mathbb{R}^m \to \mathbb{R}^d$ *be any integrable function. The function*

$$h(\mu) := \int_{\mathbb{R}^m} f(y) e^{\frac{-n\|y-\mu\|^2}{2}} dy$$

is continuous and has derivatives of all orders with respect to μ*, which can be obtained by differentiating under the integral sign.*

Below, for any vector v, v_j will denote the j-th coordinate, and for any matrix $A \in \mathbb{R}^{p \times q}$, A_{ij} will denote entry in the i-th row and j-th column.

3 Proof of Theorem 1.1

Proof of Theorem 1.1. As introduced in the previous sections, $\overline{\xi}$ will denote the sample average of $\xi^1, \ldots, \xi^n$. Consider an arbitrary decision rule $\delta \in \Delta^n$. Consider the conditional expectation

$$\eta(y) = \mathbb{E}_{\xi^1,\ldots,\xi^n}[\delta(\xi^1, \ldots, \xi^n) | \overline{\xi} = y].$$

Observe that $\eta(y) \in \text{conv}(X)$ (i.e., the convex hull of X, which is compact since X is compact) since δ maps into X. Moreover, since $\overline{\xi}$ is a sufficient statistic for the family of normal distributions by Proposition 2.3, $\eta(y)$ does not depend on μ. This is going to be important below. To maintain intuitive notation, we will also say that δ^n_{SA} is given by $\delta^n_{SA}(\xi^1, \ldots, \xi^n) = \eta^*(\overline{\xi})$, where $\eta^*(y)$ returns a point in $\arg\max\{\ y^T x \ : \ x \in X\}$. Using the law of total expectation,

$$\begin{aligned} R(\mu, \delta) &= \mathbb{E}_{\xi^1,\ldots,\xi^n}[\mathcal{L}(\mu, \delta(\xi^1, \ldots, \xi^n))] \\ &= \mathbb{E}_{\xi^1,\ldots,\xi^n}[\mu^T \delta(\xi^1, \ldots, \xi^n)] - \mu^T x(\mu) \\ &= \mathbb{E}_y[\mathbb{E}_{\xi^1,\ldots,\xi^n}[\mu^T \delta(\xi^1, \ldots, \xi^n) | \overline{\xi} = y]] - \mu^T x(\mu) \\ &= \mathbb{E}_y[\mu^T \eta(y)] - \mu^T x(\mu) \end{aligned}$$

If $\eta = \eta^*$ almost everywhere, then $R(\mu, \delta) = R(\mu, \delta^n_{SA})$, and we would be done. So in the following, we assume that $\eta \neq \eta^*$ on a set of strictly positive measure. This implies the following

CLAIM 3.1. *For all* $y \in \mathbb{R}^d$, $y^T \eta(y) \geq y^T \eta^*(y)$ *and the set* $\{y \in \mathbb{R}^d \ : \ y^T \eta(y) > y^T \eta^*(y)\}$ *is of strictly positive measure.*

Proof. Since X is compact, $\mathrm{conv}(X)$ is a compact, convex set and $\min\{y^T x \ : \ x \in \mathrm{conv}(X)\} = \min\{y^T x \ : \ x \in X\}$ for every $y \in \mathbb{R}^d$. Therefore, since $\eta(y) \in \mathrm{conv}(X)$ and $\eta^*(y) \in \arg\max\{\ y^T x \ : \ x \in X\}$, we have $y^T\eta(y) \geq y^T\eta^*(y)$ for all $y \in \mathbb{R}^d$.

Since $\mathrm{conv}(X)$ is a compact, convex set, the set of $y \in \mathbb{R}^d$ such that $|\arg\min\{y^T x \ : \ x \in \mathrm{conv}(X)\}| > 1$ is of zero Lebesgue measure. Let $S \subseteq \mathbb{R}^d$ be the set of $y \in \mathbb{R}^d$ such that $\arg\min\{y^T x \ : \ x \in \mathrm{conv}(X)\}$ is a singleton, i.e., there is a unique optimal solution; so $\mathbb{R}^d \setminus S$ has zero Lebesgue measure. Let $D := \{y \in \mathbb{R}^d : \eta(y) \neq \eta^*(y)\}$. Since we assume that D has strictly positive measure, $D \cap S$ must have strictly positive measure. Consider any $y \in D \cap S$. Since $\min\{y^T x \ : \ x \in X\} = \min\{y^T x \ : \ x \in \mathrm{conv}(X)\}$, we must have $\arg\min\{y^T x \ : \ x \in X\} \subseteq \arg\min\{y^T x \ : \ x \in \mathrm{conv}(X)\}$. Since $y \in S$, $\arg\min\{y^T x \ : \ x \in \mathrm{conv}(X)\}$ is a singleton and thus $\eta^*(y)$ is the unique optimum for $\min\{y^T x \ : \ x \in \mathrm{conv}(X)\}$. Since $y \in D$, $\eta(y) \neq \eta(y^*)$, and therefore $y^T\eta(y) > y^T\eta^*(y)$. Thus, we have the second part of the claim. □

Now consider the function $F : \mathbb{R}^d \to \mathbb{R}$ defined by

$$F(\mu) := R(\mu, \delta) - R(\mu, \delta^n_{SA}). \tag{3.1}$$

To show that δ^n_{SA} is admissible, it suffices to show that there exists $\bar{\mu} \in \mathbb{R}^d$ such that $F(\bar{\mu}) > 0$. For any $\mu \in \mathbb{R}^d$, we have from above

$$\begin{aligned} F(\mu) &= R(\mu, \delta) - R(\mu, \delta^n_{SA}) \\ &= \mathbb{E}_y[\mu^T\eta(y)] - \mathbb{E}_y[\mu^T\eta^*(y)] \\ &= \mu^T \textstyle\int_{\mathbb{R}^d} (\frac{n}{2\pi})^{n/2} \eta(y) e^{\frac{-n\|y-\mu\|^2}{2}} dy - \mu^T \int_{\mathbb{R}^d} (\frac{n}{2\pi})^{n/2} \eta^*(y) e^{\frac{-n\|y-\mu\|^2}{2}} dy \\ &= \mu^T \textstyle\int_{\mathbb{R}^d} (\frac{n}{2\pi})^{n/2} (\eta(y) - \eta^*(y)) e^{\frac{-n\|y-\mu\|^2}{2}} dy \end{aligned}$$

where in the second to last equality, we have used the fact that $\overline{\xi}$ has distribution $N(\mu, \frac{1}{n}I)$. Note that the formula above immediately gives $F(0) = 0$. We will employ Lemma 2.1 on $F(\mu)$ to show the existence of $\bar{\mu} \in \mathbb{R}^d$ such that $F(\bar{\mu}) > 0$. For this purpose, we need to compute the gradient $\nabla F(\mu)$ and Hessian $\nabla^2 F(\mu)$. We alert the reader that in these calculations, it is crucial that $\eta(y)$ does not depend on μ (due to sufficiency of the sample average) and hence it is to be considered as a constant when computing the derivatives below. For ease of calculation, we introduce the following functions $E, G^1, \ldots, G^d : \mathbb{R}^d \to \mathbb{R}^d$:

$$\begin{aligned} E(\mu) &:= (\tfrac{n}{2\pi})^{n/2} \textstyle\int_{\mathbb{R}^d} (\eta(y) - \eta^*(y)) e^{\frac{-n\|y-\mu\|^2}{2}} dy, \\ G^i(\mu) &:= (\tfrac{n}{2\pi})^{n/2} \textstyle\int_{\mathbb{R}^d} y_i (\eta(y) - \eta^*(y)) e^{\frac{-n\|y-\mu\|^2}{2}} dy, \\ i &= 1, \ldots, d. \end{aligned}$$

So $F(\mu) = \mu^T E(\mu)$. We also define the map $G : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ as

$$G(\mu)_{ij} = (G^i(\mu))_j.$$

Claim 3.2. *For any $\mu \in \mathbb{R}^d$, $\nabla F(\mu) = E(\mu) + nG(\mu)\mu - n(\mu^T E(\mu))\mu$. (Note that $G(\mu)\mu$ is a matrix-vector product.)*

Proof of Claim. This is a straightforward calculation. Consider the i-th coordinate of $\nabla F(\mu)$, i.e., the i-th partial derivative

$$\begin{aligned}
\frac{\partial F}{\partial \mu_i} &= \frac{\partial(\sum_j \mu_j E(\mu)_j)}{\partial \mu_i} \\
&= E(\mu)_i + \textstyle\sum_{j=1}^d \mu_j \frac{\partial E(\mu)_j}{\partial \mu_i} \\
&= E(\mu)_i + \textstyle\sum_{j=1}^d \mu_j \big(\int_{\mathbb{R}^d} (\frac{n}{2\pi})^{n/2} (\eta(y) - \eta^*(y))_j \frac{\partial(e^{\frac{-n\|y-\mu\|^2}{2}})}{\partial \mu_i} dy \big) \\
&= E(\mu)_i + \mu^T \textstyle\int_{\mathbb{R}^d} (\frac{n}{2\pi})^{n/2} (\eta(y) - \eta^*(y)) \frac{\partial(e^{\frac{-n\|y-\mu\|^2}{2}})}{\partial \mu_i} dy \\
&= E(\mu)_i + \mu^T \textstyle\int_{\mathbb{R}^d} (\frac{n}{2\pi})^{n/2} (\eta(y) - \eta^*(y)) e^{\frac{-n\|y-\mu\|^2}{2}} (n(y_i - \mu_i)) dy \\
&= E(\mu)_i + n\mu^T G^i(\mu) - n(\mu^T E(\mu))\mu_i
\end{aligned}$$

where in the third equality, we have used Theorem 2.4 and the fact that $\eta(y), \eta^*(y)$ do not depend on μ by sufficiency of the sample average (Proposition 2.3). The last expression above corresponds to the i-th coordinate of $E(\mu) + nG(\mu)\mu - n(\mu^T E(\mu))\mu$. Thus, we are done. $\diamond$

Claim 3.3. $\nabla^2 F(0) = n(G(0)^T + G(0))$.

Proof of Claim. Let us compute $\frac{\partial^2 F}{\partial \mu_i \mu_j}$ using the expression for $\frac{\partial F}{\partial \mu_i}$ from Claim 3.2.

$$\begin{aligned}
\frac{\partial^2 F}{\partial \mu_i \mu_j} &= \frac{\partial(E(\mu)_i)}{\partial \mu_j} + n\frac{\partial(\mu^T G^i(\mu))}{\partial \mu_j} - n\frac{\partial((\mu^T E(\mu))\mu_i)}{\partial \mu_j} \\
&= \frac{\partial(E(\mu)_i)}{\partial \mu_j} + n(G^i(\mu))_j + n\mu^T \frac{\partial(G^i)}{\partial \mu_j} - n\frac{\partial(\mu^T E(\mu))}{\partial \mu_j}\mu_i - n(\mu^T E(\mu))\gamma_{ij}
\end{aligned}$$

where γ_{ij} denotes the Kronecker delta function, i.e., $\gamma_{ij} = 1$ if $i = j$ and 0 otherwise. At $\mu = 0$, the above simplifies to

$$\left.\frac{\partial^2 F}{\partial \mu_i \mu_j}\right|_{\mu=0} = \left.\frac{\partial(E(\mu)_i)}{\partial \mu_j}\right|_{\mu=0} + n(G^i(0))_j. \tag{3.2}$$

Let us now investigate $\frac{\partial(E(\mu)_i)}{\partial \mu_j}$. By applying Theorem 2.4 and the sufficiency of $\overline{\xi}$ again, we obtain

$$\begin{aligned}
\frac{\partial(E(\mu)_i)}{\partial \mu_j} &= \textstyle\int_{\mathbb{R}^d} (\frac{n}{2\pi})^{n/2} (\eta(y) - \eta^*(y))_i \frac{\partial(e^{\frac{-n\|y-\mu\|^2}{2}})}{\partial \mu_i} dy \\
&= \textstyle\int_{\mathbb{R}^d} (\frac{n}{2\pi})^{n/2} (\eta(y) - \eta^*(y))_i e^{\frac{-n\|y-\mu\|^2}{2}} (n(y_j - \mu_j)) dy \\
&= n(\frac{n}{2\pi})^{n/2} \textstyle\int_{\mathbb{R}^d} y_j (\eta(y) - \eta^*(y))_i e^{\frac{-n\|y-\mu\|^2}{2}} dy \\
&\quad -n(\frac{n}{2\pi})^{n/2} \mu_j \textstyle\int_{\mathbb{R}^d} (\eta(y) - \eta^*(y))_i e^{\frac{-n\|y-\mu\|^2}{2}} dy \\
&= n(G^j(\mu))_i - n\mu_j E(\mu)_i
\end{aligned}$$

Therefore, at $\mu = 0$, we obtain that $\left.\frac{\partial(E(\mu)_i)}{\partial \mu_j}\right|_{\mu=0} = nG^j(0)_i$. Putting this back into (3.2), and using the definition of the matrix $G(\mu)$, we obtain

$$\left.\frac{\partial^2 F}{\partial \mu_i \mu_j}\right|_{\mu=0} = nG^j(0)_i + n(G^i(0))_j = n(G(0)_{ji} + G(0)_{ij}).$$

Thus, we obtain that $\nabla^2 F(0) = n(G(0)^T + G(0))$. $\diamond$

CLAIM 3.4. *There exists a direction of positive curvature for $\nabla^2 F(0)$, i.e., there exists $d \in \mathbb{R}^d$ such that $d^T \nabla^2 F(0) d > 0$.*

Proof of Claim. Consider the trace $\mathrm{Tr}(\nabla^2 F(0))$ of the Hessian at $\mu = 0$. By Claim 3.3,

$$\begin{aligned}
\mathrm{Tr}(\nabla^2 F(0)) &= 2n \,\mathrm{Tr}(G(0)) \\
&= 2n \textstyle\sum_{i=1}^d (\frac{n}{2\pi})^{n/2} \int_{\mathbb{R}^d} y_i (\eta(y) - \eta^*(y))_i e^{\frac{-n\|y-\mu\|^2}{2}} dy \\
&= 2n (\frac{n}{2\pi})^{n/2} \int_{\mathbb{R}^d} y^T (\eta(y) - \eta^*(y)) e^{\frac{-n\|y-\mu\|^2}{2}} dy
\end{aligned}$$

By Claim 3.1, $y^T(\eta(y) - \eta^*(y)) \geq 0$ for any $y \in \mathbb{R}^d$ and $y^T(\eta(y) - \eta^*(y)) > 0$ on a set of strictly positive measure. Therefore, $\int_{\mathbb{R}^d} y^T (\eta(y) - \eta^*(y)) e^{\frac{-n\|y\|^2}{2}} dy > 0$.

Therefore, the trace of $\nabla^2 F(0)$ is strictly positive. Since the trace equals the sum of the eigenvalues of $\nabla^2 F(0)$ (see Section 1.2.5 in [5]), we must have at least one strictly positive eigenvalue. The corresponding eigenvector is a direction of positive curvature. $\diamond$

As noted earlier, $F(0) = 0$. Combining Claim 3.4 and Lemma 2.1, there exists $\bar{\mu} \in \mathbb{R}^d$ such that $F(\bar{\mu}) > 0$. $\square$

4 General covariance for the Gaussian

The discussion in the previous sections focused on the family of normal distributions with the identity as the covariance matrix. However, Theorem 1.1 extends to the case where one allows any positive definite covariance matrix Σ for the normal distributions. In this case, we again consider the function $F(\mu)$ defined in (3.1) and prove that there exists $\bar{\mu}$ such that $F(\bar{\mu}) > 0$. The only difference is that in the formulas one must substitute the distribution $\overline{\xi} \sim N(\mu, \frac{1}{n}\Sigma)$, i.e., the density function everywhere must be

$$g_{\mu,\Sigma}(y) := \frac{1}{\sqrt{\sigma}} \left(\frac{n}{2\pi}\right)^{n/2} \exp\left(-\frac{n}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right),$$

where σ is the determinant of Σ. Redefining

$$\begin{aligned}
E(\mu) &:= \tfrac{1}{\sqrt{\sigma}} \left(\tfrac{n}{2\pi}\right)^{n/2} \int_{\mathbb{R}^d} (\eta(y) - \eta^*(y)) \exp\left(-\tfrac{n}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right) dy, \\
G^i(\mu) &:= \tfrac{1}{\sqrt{\sigma}} \left(\tfrac{n}{2\pi}\right)^{n/2} \int_{\mathbb{R}^d} y_i (\eta(y) - \eta^*(y)) \exp\left(-\tfrac{n}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right) dy,\ i = 1, \ldots, d,
\end{aligned}$$

letting $G(\mu)$ be the matrix with $G^i(\mu)$ as rows, and adapting the calculations from the previous section reveals that

$$\nabla^2 F(0) = n(\Sigma^{-1} G(0) + G(0)^T \Sigma^{-1}) \tag{4.1}$$

Claim 3.1 again shows that the trace $\mathrm{Tr}(G(0)) = \int_{\mathbb{R}^d} y^T (\eta(y) - \eta^*(y)) g_{0,\Sigma}(y) dy > 0$. This shows that $G(0)$ has an eigenvalue λ with positive *real* part (since $G(0)$ is not guaranteed to be symmetric, its eigenvalues and eigenvectors may be complex). Let the corresponding

(possibly complex) eigenvector be v, i.e., $G(0)v = \lambda v$ and $Re(\lambda) > 0$ (denoting the real part of λ). Following standard linear algebra notation, for any matrix/vector M, M^* will denote its Hermitian conjugate [5] (which equals the transpose if the matrix has real entries). We now consider

$$\begin{aligned} v^*\nabla^2 F(0)v &= n(v^*(\Sigma^{-1}G(0) + G(0)^T\Sigma^{-1})v) \\ &= n(v^*\Sigma^{-1}G(0)v + v^*G(0)^T\Sigma^{-1}v) \\ &= n(v^*\Sigma^{-1}G(0)v + v^*G(0)^*\Sigma^{-1}v) \\ &= n(\lambda(v^*\Sigma^{-1}v) + \lambda^*(v^*\Sigma^{-1}v)) \\ &= 2n(v^*\Sigma^{-1}v)Re(\lambda) \end{aligned}$$

Since Σ is positive definite, so is Σ^{-1}. Therefore $v^*\Sigma^{-1}v > 0$ and we obtain that $v^*\nabla^2 F(0)v > 0$. Since $\nabla^2 F(0)$ is a symmetric matrix, all its eigenvalues are real and in particular its largest eigenvalue γ_d is positive because

$$\gamma_d = \max_{x \in \mathbb{C}^d \setminus \{0\}} \frac{x^*\nabla^2 F(0)x}{x^*x} \geq \frac{v^*\nabla^2 F(0)v}{v^*v} > 0.$$

Thus, $\nabla^2 F(0)$ has a direction of positive curvature and Lemma 2.1 implies that there exists $\bar{\mu} \in \mathbb{R}^d$ such that $F(\bar{\mu}) > 0$.

5 Future Work

To the best of our knowledge, a thorough investigation of the admissibility of solution estimators for stochastic optimization problems has not been undertaken in the statistics or optimization literature. There are several questions that one may immediately pose for future research:

1. Does the admissibility result continue to hold for nonlinear objectives? For example, what if $F(x, \xi) = x^TQx + \xi^Tx$ for some fixed PSD matrix Q, and X is a compact, convex set? We believe new ideas beyond the techniques introduced in this paper are needed to analyze the admissibility of the sample average estimator for this convex quadratic program. This problem is interesting from a financial engineering perspective, where the stochastic optimization problem seeks to minimize a coherent risk measure over a convex set. The simplest such measure is a weighted sum of the expectation and the variance of the returns, which can be modeled using the above $F(x, \xi)$.

2. What about piecewise linear objectives $F(x, \xi)$? Such objectives show up in the stochastic optimization literature under the name of *news-vendor type problems*. The current techniques of this paper do not easily apply directly to this setting either.

3. For learning problems, such as neural network training with squared or logistic loss, what can be said about the admissibility of the sample average rule, which usually goes under the name of "empirical risk minimization"? Is the empirical risk minimization rule an admissible rule in the sense of statistical decision theory? It would be very interesting if the answer actually depends on the hypothesis class that is being learnt. It is also possible that decision rules that take the empirical risk objective and report a *local* optimum can be shown to dominate decision rules that report the global optimum, under

certain conditions. This would be an interesting perspective on the debate whether local solutions are "better" in a theoretical sense than global optima.

Acknowledgments. We are extremely grateful to Prof. Daniel Naiman for helping us understand basic admissibility results from statistics and numerous follow-up discussions related to this work. We also owe a great philosophical debt to Prof. Carey Priebe who always nudged us to think about optimization in statistical/probabilistic terms. His aphorism "The solution to an optimization problem is a random variable!" made us think about optimization in a new light. Finally, the main inspiration for this work came from listening to a beautiful talk by Prof. Gérard Cornuéjols explaining his work from [3] and [4].

References

[1] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[2] Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volume I*, volume 117. CRC Press, 2015.

[3] Danial Davarnia and Gérard Cornuéjols. From estimation to optimization via shrinkage. *Operations Research Letters*, 45(6):642–646, 2017.

[4] Danial Davarnia, Burak Kocuk, and Gérard Cornuéjols. Bayesian solution estimators in stochastic optimization. *http://www.optimization-online.org/DB_HTML/2017/11/6318.html*, 2018.

[5] R. Horn and C. Johnson. *Matrix Analysis*. John Wiley & Sons, second edition, 1985.

[6] William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

[7] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[8] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[9] Charles 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*, volume 1, pages 197–206, 1956.