OPTIMAL ESTIMATORS FOR THRESHOLD-BASED QUALITY MEASURES

AARON ABRAMS, SANDY GANZELL, HENRY LANDAU, ZEPH LANDAU, JAMES POMMERSHEIM, AND ERIC ZASLOW

Abstract. We consider a problem in parametric estimation: given n samples from an unknown distribution, we want to estimate which distribution, from a given one-parameter family, produced the data. Following Schulman and Vazirani [12], we evaluate an estimator in terms of the chance of being within a specified tolerance of the correct answer, in the worst case. We provide optimal estimators for several families of distributions on $\mathbb{R}$. We prove that for distributions on a compact space, there is always an optimal estimator that is translation-invariant, and we conjecture that this conclusion also holds for any distribution on $\mathbb{R}$. By contrast, we give an example showing it does not hold for a certain distribution on an infinite tree.

1. Introduction

Estimating probability distribution functions is a central problem in statistics. Specifically, beginning with an unknown probability distribution on an underlying space X, one wants to be able to do two things: first, given some empirical data sampled from the unknown probability distribution, estimate which one of a presumed set of possible distributions produced the data; and second, obtain bounds on how good this estimate is. For example, the *maximum likelihood estimator* selects the distribution that maximizes the probability (among those under consideration) of producing the observed data. Depending on what properties of the estimator one is trying to evaluate, this may or may not be optimal. An extensive literature, dating back to the early 20th century, addresses problems of this sort; see for example [2, 3, 6, 8, 10, 13].

In this paper we consider one such problem. We presume samples are coming from an unknown "translate" of a fixed known distribution. The challenge is to guess the translation parameter. More precisely, we are given a distribution μ on a space X, along with an action of a group G on X, which defines a set of translated distributions μ_θ as follows:

$$\mu_\theta(A) = \mu(\{x : \theta x \in A\}) \tag{1}$$

for $A \subset X$. Thus in this context an *estimator* is a (measurable) function $e : X^n \to G$; the input $\mathbf{x} = (x_1, \ldots, x_n)$ is the list of samples, and the output $e(\mathbf{x})$ is the estimate of θ, the translation parameter. For the majority of the paper we will study the case of $G = \mathbb{R}$ acting by translations (changes in

location) on $X = \mathbb{R}$, and the group action will be written additively, as seen beginning in Section 2.

We are interested in finding good estimators; thus we need a way of measuring an estimator's quality. A common way to do this is to measure the *mean squared error*, in which case an optimal estimator minimizes this error. Various results are known in this case; for instance the maximum likelihood estimator (which agrees with the *sample mean estimator*),

$$e(\mathbf{x}) = \left(\frac{1}{n}\sum x_i\right) - \mathbb{E}(\mu),$$

minimizes the mean squared error if μ is a Gaussian distribution on $\mathbb{R}$.

In this paper we investigate a different and natural measure of quality whereby we consider an estimator to succeed or fail according to whether or not its estimate is within a certain threshold $\delta > 0$ of the correct answer. We then define the quality of the estimator to be the chance of success in the worst case. This notion was introduced in [12] to analyze certain approximation algorithms in computer science. Precisely, the *δ-quality* of e is defined as

$$\begin{aligned} Q_\delta(e) &= \inf_\theta Q_\delta^\theta(e) \\ &= \inf_\theta \Pr\{d(e(\mathbf{x}),\theta) < \delta : x_i \text{ are chosen from } \mu_\theta\} \\ &= \inf_\theta \mu_\theta^n(\{\mathbf{x} : d(e(\mathbf{x}),\theta) < \delta\}), \end{aligned} \tag{2}$$

where d is a metric on X and μ_θ^n is the product measure $\mu_\theta \times \cdots \times \mu_\theta$ on X^n.[1] Note that, depending on context, it is sometimes advantageous to define the quality using a closed interval rather than an open one; for example in the discrete case we could then interpret $Q_0^\theta(e)$ as the probability that e is exactly equal to θ. We write $Q(e)$ when the value of δ is unambiguous. For fixed δ, an (n-sample) estimator e is *optimal* if $Q_\delta(e) \geq Q_\delta(e')$ for all (n-sample) estimators e'. Many authors use the term *minimax* to describe optimal estimators. Note that much of the literature on this subject uses the notion of loss functions and the associated risk $R = 1 - Q$; our point of view is equivalent but more convenient for our purposes.

Motivated initially by analyzing an approximate algorithm for determining the average matching size in a graph, Schulman and Vazirani [12] introduce the stronger notion of a *majorizing estimator*, which is optimal (by the above definition) simultaneously for all $\delta > 0$. This was previously studied by Pitman [10], who considered several different optimality criteria and, for each one, constructed optimal "shift-invariant" estimators (defined below). Schulman and Vazirani focus on the Gaussian distribution and prove that the mean estimator is the unique majorizing estimator in this case.

[1] In the case of perverse measures, μ, we must consider the probability as the sup of the intersection of the set $\{d(e(\mathbf{x}),\theta) < \delta\}$ with all measurable sets. We will ignore this caveat throughout. Indeed, we primarily focus on absolutely continuous measures (as [4] and [7] have done, for example) and purely atomic measures.

In the first part of this paper we investigate the optimal estimators for several different classes of distributions on $\mathbb{R}$. We conjecture that there is always an optimal estimator e that is *shift-invariant*, i.e. e satisfies

$$e(x_1 + c, \ldots, x_n + c) = e(x_1, \ldots, x_n) + c$$

for all $c, x_i \in \mathbb{R}$. These estimators are typically easier to analyze than general estimators, because the quality is the same everywhere, i.e. $Q(e) = Q^\theta(e)$ for every θ. Conditions under which invariant minimax estimators can be obtained have been studied, for example, in [1], [9] and [11]. Indeed, some of our existence results follow from the quite general Hunt-Stein theorem [11, Theorem 9.5.5], but we give constructions that are very natural and explicit. We obtain general bounds on the quality of shift-invariant estimators (Section 2) and general estimators (Section 3), and then we apply these bounds to several families of distributions (Section 4). In each case, we are able to construct an optimal estimator that is shift-invariant. These examples include the Gaussian and exponential distributions, among others.

These results motivate our study of shift-invariant estimators on other spaces; these are estimators that are equivariant with respect to the induced diagonal action of G on either the left or the right on X^n. That is, a *left-invariant* estimator satisfies

$$e(g\mathbf{x}) = ge(\mathbf{x}) \tag{3}$$

where

$$g(x_1, \ldots, x_n) = (gx_1, \ldots, gx_n).$$

Right-invariance is defined similarly.

In Section 5 we show that on a compact space X, if e is an estimator for μ, then there is always a shift-invariant estimator with quality at least as high as that of e. The idea is to construct a shift-invariant estimator s as an average of the translates of e; this is essentially a simple proof of a special case of the Hunt-Stein theorem. As there is no invariant probability measure on $\mathbb{R}$, the proof does not extend to the real case.

Finally, in the last section, we give an example due to L. Schulman which shows that (on non-compact spaces) there may be no shift-invariant estimator that is optimal. It continues to be an interesting problem to determine conditions under which one can guarantee the existence of a shift-invariant estimator that is optimal.

The authors thank V. Vazirani and L. Schulman for suggesting the problem that motivated this paper along with subsequent helpful discussions. We are grateful to the reviewers for many helpful suggestions, especially for correcting our proof of Lemma 7. As always, we thank Julie Landau, without whose support and down-the-line backhand this work would not have been possible.

2. The real case: shift-invariant estimators

Let $G = X = \mathbb{R}$, and consider the action of G on X by translations. Because much of this paper is concerned with this context, we spell out once more the parameters of the problem. We assume $\delta > 0$ is fixed throughout. We are given a probability distribution μ on $\mathbb{R}$, and we are to guess which distribution μ_θ produced a given collection $\mathbf{x} = (x_1, \ldots, x_n)$ of data, where $\mu_\theta(A) = \mu(\{x : x + \theta \in A\})$. An estimator is a function $e : \mathbb{R}^n \to \mathbb{R}$, and we want to maximize its quality, which is given by

$$\begin{aligned} Q(e) = \inf_\theta Q^\theta(e) &= \inf_\theta \Pr\{|e(\mathbf{x}) - \theta| < \delta : x_i \text{ are chosen from } \mu_\theta\} \\ &= \inf_\theta \mu_\theta^n(\{\mathbf{x} : |e(\mathbf{x}) - \theta| < \delta\}). \end{aligned}$$

First some notation. We will write the group action additively and likewise the induced diagonal action of G on $\mathbb{R}^n$; in other words if $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $a \in \mathbb{R}$, then $\mathbf{x} + a$ denotes the point $(x_1 + a, \ldots, x_n + a) \in \mathbb{R}^n$. Similarly if $Y \subset \mathbb{R}^n$ and $A \subset \mathbb{R}$ then $Y + A = \{\mathbf{y} + a : \mathbf{y} \in Y, a \in A\}$. We also use the "interval notation" $(\mathbf{x}+a, \mathbf{x}+b)$ for the set $\{\mathbf{x}+t : a < t < b\}$; this is a segment of length $(b-a)\sqrt{n}$ in $\mathbb{R}^n$ if a and b are finite. If $f : \mathbb{R}^n \to \mathbb{R}$ is any function, and $\theta \in \mathbb{R}$, define $f_\theta(\mathbf{x}) = f(\mathbf{x} - \theta)$. If $f : \mathbb{R} \to \mathbb{R}$, then define $f^{[n]} : \mathbb{R}^n \to \mathbb{R}$ by $f^{[n]}(\mathbf{x}) = f(x_1)f(x_2)\cdots f(x_n)$.

We now establish our upper bound on the quality of shift-invariant estimators. Note that a shift-invariant estimator has the property that $e(\mathbf{x} - e(\mathbf{x})) = 0$. Also note that a shift-invariant estimator is determined uniquely by its values on the coordinate hyperplane

$$X_0 = \{\mathbf{x} \in \mathbb{R}^n : x_1 = 0\},$$

and that a shift-invariant estimator exists for any choice of such values on X_0. In addition, for e shift-invariant,

$$\mu_\theta^n(\{\mathbf{x} : |e(\mathbf{x}) - \theta| < \delta\}) = \mu_\theta^n(\{\mathbf{x} : |e(\mathbf{x} - \theta)| < \delta\}) = \mu^n(\{\mathbf{x} : |e(\mathbf{x})| < \delta\}),$$

so the quality can be ascertained by setting $\theta = 0$.

Definition 1. *For fixed n, let $\mathcal{A}$ denote the collection of all Borel subsets A of the form*

$$A = \bigcup_{\mathbf{x} \in X_0} \mathbf{x} + (f(\mathbf{x}) - \delta, f(\mathbf{x}) + \delta),$$

where $f : X_0 \to \mathbb{R}$ is a Borel function. For fixed μ and n, define

$$S_{\mu,n} = S_{\mu,n}(\delta) = \sup_{A \in \mathcal{A}} \{\mu^n(A)\}.$$

Theorem 2. *Let μ and n be given. Then any shift-invariant n-sample estimator e satisfies $Q(e) \le S_{\mu,n}$.*

Proof: Due to the observation above, it suffices to bound the quality of e at $\theta = 0$. But this quality is just $\mu^n(A)$ where $A = e^{-1}((-\delta, \delta))$. Note that

$$A = \bigcup_{\mathbf{x} \in X_0} (\mathbf{x} - e(\mathbf{x}) - \delta, \mathbf{x} - e(\mathbf{x}) + \delta),$$

and in particular $A \in \mathcal{A}$. Thus the quality of e is at most $S_{\mu,n}$. □

Theorem 3. *Let μ and n be given. If the* sup *in Definition 1 is achieved, then there is a shift-invariant n-sample estimator with quality $S_{\mu,n}$. For any $\epsilon > 0$, there is a shift-invariant n-sample estimator e with quality greater than $S_{\mu,n} - \epsilon$.*

Proof: For a given $A \in \mathcal{A}$, we will define a shift-invariant estimator e with the property that $A \subset e^{-1}((-\delta, \delta))$. Then $Q(e) \geq \mu^n(A)$. The theorem then follows from the definition of $S_{\mu,n}$.

So, fix $A \in \mathcal{A}$. For each $\mathbf{x} \in \mathbb{R}^n$ choose the supremum of those m's and the infimum of $M \geq m$ such that $A \cap (\mathbf{x} + \mathbb{R}) \subset [\mathbf{x} + m, \mathbf{x} + M]$. Define $e(\mathbf{x})$ on X_0 by $e(\mathbf{x}) = (M + m)/2$, and then extend e to all of $\mathbb{R}^n$ to make it shift-invariant. Now clearly $A \subseteq e^{-1}([-\delta, \delta])$, since if $\mathbf{x} \in A$ then $|M - m| \leq 2\delta$. This completes the proof. □

3. The real case: general estimators

In this section we obtain a general upper bound on the quality of randomized estimators, still in the case $G = X = \mathbb{R}$. The arguments are similar to those of the previous section.

Again δ is fixed throughout. A *randomized estimator* is a function $X^n \times \Omega \to \mathbb{R}$ where Ω is a probability space of estimators; thus for fixed $\omega \in \Omega$, $e = \widetilde{e}(\cdot, \omega)$ is an estimator. The δ-quality of a randomized estimator $\widetilde{e}$ is

$$Q(\widetilde{e}) = \inf_{\theta} Q^{\theta}(\widetilde{e})$$

where

$$Q^{\theta}(\widetilde{e}) = \mathbb{E}\mu_{\theta}^{n} \left\{ \mathbf{x} \mid |\widetilde{e}(\mathbf{x}) - \theta| < \delta \right\}.$$

Definition 4. *For fixed n, let*

$$\mathcal{B} = \{B \subset \mathbb{R}^n : B \cap (B + 2k\delta) = \emptyset \text{ for all nonzero integers } k\}.$$

For fixed μ and n, define

$$T_{\mu,n} = T_{\mu,n}(\delta) = \sup_{B \in \mathcal{B}} \{\mu^n(B)\}.$$

Comparing with Definition 1, we observe that $\mathcal{A} \subset \mathcal{B}$ and hence $S_{\mu,n} \leq T_{\mu,n}$.

Theorem 5. *Let μ and n be given. Then any n-sample randomized estimator $\widetilde{e}$ satisfies $Q(\widetilde{e}) \leq T_{\mu,n}$.*

Proof: We will give a complete proof in the case that μ is defined by a density function f, and then indicate the modifications required for the general case. The difference is purely technical; the ideas are the same.

Consider first a non-randomized estimator e. The performance of e at θ is $\mu_{\theta}^{n}\left(e^{-1}((\theta - \delta, \theta + \delta))\right)$. To simplify notation we will let $I_{e,\theta}$ denote the set

$e^{-1}((\theta-\delta,\theta+\delta))$ and we will suppress the subscript e when no ambiguity exists. Since $Q(e)$ is an infimum, the average performance of e at the k points $\theta_i = 2\delta i$, $(i = 1, 2, \ldots, k)$ is at least $Q(e)$:

$$Q(e) \leq \frac{1}{k}\sum_{i=1}^{k} \mu_{\theta_i}^n(I_{\theta_i}) \tag{4}$$

Now we use the density function f. Recall that $f_\theta(x) = f(x-\theta)$. Define $\widetilde{f}$ on $\mathbb{R}^n$ by

$$\widetilde{f}(\mathbf{x}) = \max_i\{f_{\theta_i}^{[n]}(\mathbf{x})\} = \max_i\{f_{\theta_i}(x_1)f_{\theta_i}(x_2)\cdots f_{\theta_i}(x_n)\}.$$

Since the I_{θ_i} are disjoint, we now have

$$\begin{aligned}
Q(e) &\leq \frac{1}{k}\sum_{i=1}^{k}\int_{I_{\theta_i}} f_{\theta_i}^{[n]}(\mathbf{x})d\mathbf{x} \\
&\leq \frac{1}{k}\sum_{i=1}^{k}\int_{I_{\theta_i}} \widetilde{f}(\mathbf{x})d\mathbf{x} \\
&\leq \frac{1}{k}\int_{\bigcup I_{\theta_i}} \widetilde{f}(\mathbf{x})d\mathbf{x} \\
&\leq \frac{1}{k}\int_{\mathbb{R}^n} \widetilde{f}(\mathbf{x})d\mathbf{x} \\
&= \frac{1}{k}\int\limits_{\{x_1\leq 0\}} \widetilde{f}(\mathbf{x})d\mathbf{x} + \frac{1}{k}\int\limits_{\{0<x_1<2\delta k\}} \widetilde{f}(\mathbf{x})d\mathbf{x} + \frac{1}{k}\int\limits_{\{x_1\geq 2\delta k\}} \widetilde{f}(\mathbf{x})d\mathbf{x}. \qquad (5)
\end{aligned}$$

We will bound the middle term by $T_{\mu,n}$ and show that the first and last terms go to zero (independently of e) as k gets large. The bound on the middle term is a consequence of the following claim.

Claim. For any $a \in \mathbb{R}$,

$$\int_{\{a\leq x_1\leq a+2\delta\}} \widetilde{f}(\mathbf{x})d\mathbf{x} \leq T_{\mu,n}.$$

To prove the claim, set $Z = \{\mathbf{x} \in \mathbb{R}^n : a \leq x_1 < a+2\delta\}$, and set $Z_i = \{\mathbf{x} \in Z : i \text{ is the smallest index such that } \widetilde{f}(\mathbf{x}) = f_{\theta_i}^{[n]}(\mathbf{x})\}$. Thus the

Z_i are disjoint and cover Z. Now

$$\begin{aligned}\int_{\{a\le x_1\le a+2\delta\}} \widetilde{f}(\mathbf{x})d\mathbf{x} &= \sum_i \int_{Z_i} \widetilde{f}(\mathbf{x})d\mathbf{x} = \sum_i \int_{Z_i} f_{\theta_i}^{[n]}(\mathbf{x})d\mathbf{x} \\ &= \sum_i \int_{Z_i-\theta_i} f^{[n]}(\mathbf{x})d\mathbf{x} \\ &= \int_{\bigcup(Z_i-\theta_i)} f^{[n]}(\mathbf{x})d\mathbf{x} \\ &\le T_{\mu,n}.\end{aligned}$$

The last equality follows from the fact that the $Z_i - \theta_i$ are disjoint (recall that $\theta_i = 2\delta i$), and the final step follows because the set $B = \cup(Z_i - \theta_i)$ is in $\mathcal{B}$. This proves the claim. □

Next we show that $\frac{1}{k}\int_{\{x_1\le 0\}} \widetilde{f}(\mathbf{x})d\mathbf{x}$ approaches zero as k grows. Recall that $\theta_i = 2\delta i$, and set $z_i = \int_{\{x_1\le 0\}} f_{2\delta i}^{[n]}(\mathbf{x})d\mathbf{x} = \int_{\{x_1\le -2\delta i\}} f^{[n]}(\mathbf{x})d\mathbf{x}$. The function f is a probability density function, so f is nonnegative and has total integral 1. The Dominated Convergence Theorem then implies that the sequence $\{z_i\}$ is decreasing to 0. Bounding $\widetilde{f}(\mathbf{x})$ by $\sum_{i=1}^k f_{\theta_i}^{[n]}(\mathbf{x})$ we have

$$\frac{1}{k}\int_{\{x_1\le 0\}} \widetilde{f}(\mathbf{x})d\mathbf{x} \le \frac{1}{k}\int_{\{x_1\le 0\}} \sum_{i=1}^k f_{\theta_i}^{[n]}(\mathbf{x})d\mathbf{x} = \frac{1}{k}\sum_{i=1}^k z_i \to 0.$$

A similar argument shows that the term $\frac{1}{k}\int_{\{x_1\ge 2\delta k\}} \widetilde{f}(\mathbf{x})d\mathbf{x}$ goes to 0 as k grows. Since (5) holds for all k, we have $Q(e) \le T_{\mu,n}$ for any estimator e.

We have shown that for any $\epsilon > 0$, we can find k depending on ϵ and f such that the average performance of an arbitrary estimator e on the k points $\theta_i = 2\delta i$ is bounded above by $T_{\mu,n} + 2\epsilon$. Now, for a randomized estimator $\widetilde{e}$, the quality is bounded above by its average performance on the same k points, and that performance can be no better than the best estimator's performance. We conclude that $Q(\widetilde{e}) \le T_{\mu,n} + 2\epsilon$, and the theorem follows.

The proof is now complete in the case that μ has a density f. In general, the argument requires minor technical adjustments. The first step that requires modification is the definition of the function $\widetilde{f}$. Let

$$\nu = \sum_{i=1}^k \mu_{\theta_i}^n, \quad g_i = \frac{d\mu_{\theta_i}^n}{d\widetilde{\mu}} \quad \text{and} \quad \widetilde{f} = \max(g_1, \ldots g_k).$$

Then $\widetilde{f}\cdot\nu = \widetilde{\mu}$ and we work with $\widetilde{\mu}$ rather than $\widetilde{f}$, and the remainder of the argument goes through with corresponding changes. □

4. The real case: examples

We have obtained bounds on quality for general estimators and for shift-invariant ones. In this section we give several situations where the bounds coincide, and therefore the optimal shift-invariant estimators constructed in Section 2 are in fact optimal estimators, as promised by the Hunt-Stein theorem. These examples include many familiar distributions, and they provide evidence for the following conjecture.

Conjecture 1. *Let μ be a distribution on $\mathbb{R}$. Then there is an optimal estimator for μ that is shift-invariant.*

4.1. **Warm-up: unimodal densities, one sample.** Our first class of examples generalizes Gaussian distributions and many others. The argument works only with one sample, but we will refine it in 4.2. Note that the optimal estimator in this case is the maximum likelihood estimator.

We say that a density function is *unimodal* if for all y, $\{x : f(x) \geq y\}$ is convex.

Example 4.1. *Let μ be defined by a unimodal density function f. Then there is a shift-invariant one-sample estimator that is optimal.*

Proof: We first show that $T_{\mu,1} = S_{\mu,1}$. It follows from the definition of $\mathcal{B}$ that any set $B \in \mathcal{B}$ must have Lebesgue measure less than or equal to 2δ. Since f is unimodal, $\int_B f(x)dx$ is maximized by concentrating B around the peak of f; thus the best B will be an interval that includes the peak of f. But any interval in $\mathcal{B}$ is contained in $\mathcal{A}$ and thus $T_{\mu,1} \leq S_{\mu,1}$. Since $S_{\mu,1} \leq T_{\mu,1}$ always, we have $T_{\mu,1} = S_{\mu,1}$.

Now, recalling that $S_{\mu,1}$ and $T_{\mu,1}$ are defined as suprema, we observe that the above argument shows that if one is achieved then so is the other. Therefore the result follows from Theorems 3 and 5. □

4.2. **A sufficient condition.** The next class is more restrictive than the preceding, but with the stronger hypothesis we get a result for arbitrary n. Any Gaussian distribution continues to satisfy the hypothesis.

Example 4.2. *Let μ be a distribution defined by a density function of the form $f = e^{\lambda(x)}$ with $\lambda'(x)$ continuous and decreasing. Then for any n, there is a shift-invariant n-sample estimator that is optimal.*

Proof: For any fixed $\mathbf{x} \in X_0$, we define a function $h_{\mathbf{x}} : \mathbb{R} \longrightarrow \mathbb{R}$ by

$$h_{\mathbf{x}}(t) = f^{[n]}(\mathbf{x} + t) = e^{\lambda(x_1+t)+\cdots+\lambda(x_n+t)}.$$

Since

$$h'_{\mathbf{x}}(t) = e^{\lambda(x_1+t)+\cdots+\lambda(x_n+t)}(\lambda'(x_1 + t) + \cdots + \lambda'(x_n + t))$$

and λ' is decreasing, it is clear that for each $\mathbf{x}$, $h'_{\mathbf{x}}(t) = 0$ for at most one value of t. Since $h_{\mathbf{x}}(t) \to 0$ as $t \to \pm\infty$, it follows that for any $\mathbf{x}$, $h_{\mathbf{x}}$ is a unimodal function of t.

Now the argument is similar to Example 4.1. We will show that $T_{\mu,n} = S_{\mu,n}$. Since $f^{[n]}$ restricted to each orbit $\mathbf{x} + \mathbb{R}$ is unimodal as we have just shown, a set $B \in \mathcal{B}$ on which the integral of $f^{[n]}$ is maximized is obtained by choosing an interval from each orbit. To make this more precise, for each $\mathbf{x} \in X_0$, let $t_{\mathbf{x}}$ be the center of the length 2δ interval $I = (t_{\mathbf{x}} - \delta, t_{\mathbf{x}} + \delta)$ that maximizes $\int_I h_{\mathbf{x}}\, dt$. Then let

$$A = \bigcup_{\mathbf{x} \in X_0} (\mathbf{x} + t_x - \delta, \mathbf{x} + t_x + \delta).$$

Now $A \in \mathcal{A}$, and moreover $\mu^n(A) \geq \mu^n(B)$ for any $B \in \mathcal{B}$, because $\int_{A \cap (\mathbf{x}+\mathbb{R})} f^{[n]} \geq \int_{B \cap (\mathbf{x}+\mathbb{R})} f^{[n]}$ for each $\mathbf{x} \in X_0$.

Thus $\sup_{B \in \mathcal{B}}\{\mu^n(B)\}$ is achieved by $B = A \in \mathcal{A}$, and it follows that $S_{\mu,n} = T_{\mu,n}$ and that the best shift-invariant estimator is optimal. □

4.3. **Monotonic distributions on $\mathbb{R}^+$.** The third class of examples generalizes the exponential distribution, defined by the density $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$. The optimal estimator in this case[2] is *not* the maximum likelihood estimator.

Example 4.3. *Let μ be defined by a density function f that is decreasing for $x \geq 0$ and identically zero for $x < 0$. Then for any n, there is a shift-invariant n-sample estimator that is optimal.*

Proof: We construct the estimator as follows: for $\mathbf{x} \in \mathbb{R}^n$, define $e(\mathbf{x}) = \min\{x_1, ..., x_n\} - \delta$. Note that this is shift-invariant; therefore $Q(e)$ can be computed at $\theta = 0$. That is, it suffices to show that $Q^0(e) = T_{\mu,n}$.

Let $B = \{\mathbf{x} \in \mathbb{R}^n : 0 \leq \min\{x_1, ..., x_n\} < 2\delta\}$. Note that $B = e^{-1}([-\delta, \delta))$, and so $\mu^n(B)$ is the quality of e. Note also that $B \in \mathcal{B}$ (in fact $B \in \mathcal{A}$), so certainly $\mu^n(B) \leq T_{\mu,n}$. We will show that any $C \in \mathcal{B}$ can be modified to a set $C' \in \mathcal{B}$ such that $C' \subset B$ and $\mu^n(C) \leq \mu^n(C')$. It then follows that $T_{\mu,n} \leq \mu^n(B)$, and this will complete the proof.

So, let $C \in \mathcal{B}$, and define $C' = \{\mathbf{x} \in B : \mathbf{x} + 2k\delta \in C \text{ for some } k \in \mathbb{Z}\}$. Note that k is determined uniquely by $\mathbf{x}$. Now $C' \subset B$ is in $\mathcal{B}$, and by our hypotheses on f, if $\mathbf{x} \in B$ then $f^{[n]}(\mathbf{x}) \geq f^{[n]}(\mathbf{x} + 2k\delta)$ for every integer k. Therefore $\mu^n(C') - \mu^n(C) = \int_{C'}[f^{[n]}(\mathbf{x}) - f^{[n]}(\mathbf{x} + 2k\delta)]d\mu^n \geq 0$. □

4.4. **Discrete distributions.** Here we discuss purely atomic distributions on finite sets of points. Because we are only trying to guess within δ of the correct value of θ, there are many possible choices of estimators with the same quality. Among the optimal ones is the maximum likelihood estimator.

[2]Note that in a typical estimation problem involving a family of exponential distributions, one is trying to estimate the "scale" parameter λ rather than the "location" θ.

Example 4.4. *Let μ be a distribution on a finite set of points Z. There is a shift-invariant one-sample estimator that is optimal. Furthermore, if all of the pairwise distances between points of Z are distinct, then for every n there is a shift-invariant n-sample estimator that is optimal.*

Proof: We first treat the case $n = 1$. Since μ is discrete, the supremum defining $S_{\mu,1}$ is attained; therefore by Theorems 3 and 5 it suffices to show that every estimator has quality at most $S_{\mu,1}$.

Let $Z = \{z_1, \ldots, z_r\}$, and for any $z \in Z$, let p_z denote the mass at z. For a finite set, we use $|\cdot|$ to denote the cardinality. Suppose that e is any estimator.

Lemma 6. *Let Y be any finite subset of $\mathbb{R}$. Then*

$$Q(e) \leq S_{\mu,1} \frac{|Y + Z|}{|Y|},$$

where $Y + Z$ denotes the set $\{y + z \mid y \in Y, z \in Z\}$.

Proof of Lemma: To prove the lemma we estimate the average quality $Q^\theta(e)$ over $\theta \in Y$. We have

$$\sum_{\theta \in Y} Q^\theta(e) = \sum_{\theta \in Y} \mu(e^{-1}(\theta - \delta, \theta + \delta) - \theta) = \sum_{\theta \in Y} \sum_z p_z$$

with the inner part of the last sum taken over those $z \in Z$ that lie in $e^{-1}(\theta - \delta, \theta + \delta) - \theta$. Using x to denote $\theta + z$, this condition becomes $e(x) \in (\theta - \delta, \theta + \delta)$, and the right hand side above may be rewritten as

$$\sum_{\theta \in Y} \sum_z p_z = \sum_{\theta \in Y} \sum_x p_{x-\theta} = \sum_{x \in Y+Z} \sum_\theta p_{x-\theta},$$

with the inner sum now taken over all θ with $e(x) \in (\theta - \delta, \theta + \delta)$. This latter condition implies that z is within δ of $x - e(x)$. But by definition, $S_{\mu,1}$ is the maximum measure of any interval of length 2δ. Hence, for any fixed $x \in Y + Z$, the inner sum is at most $S_{\mu,1}$, and the entire sum is thus bounded above by $S_{\mu,1} \cdot |Y+Z|$. Dividing by $|Y|$ gives a bound for the average quality over Y, and since $Q(e)$ is defined as an infimum the lemma follows. □

We now apply the lemma to complete the Example. Let k be a natural number, and let

$$Y_k = \{h_1 z_1 + \cdots + h_r z_r : h_i \in \mathbb{Z} \text{ and } 0 \leq h_i < k\}.$$

Note that $Y_k \subseteq Y_{k+1}$ and $|Y_k| \leq k^r$. It follows that for any $\epsilon > 0$ there exists k such that $|Y_{k+1}|/|Y_k| < 1 + \epsilon$, for otherwise $|Y_k|$ would grow at least exponentially in k. Using the fact that $Y_k + Z \subseteq Y_{k+1}$, the lemma applied to Y_k implies that $Q(e) \leq S_{\mu,1}(1 + \epsilon)$. Therefore $Q(e) \leq S_{\mu,1}$, and this finishes the case $n = 1$.

Lastly, we consider an arbitrary n. If we are given samples $x_1, ..., x_n$ and if any $x_i \neq x_j$ for some i and j, then by our hypothesis the shift θ is uniquely determined. Thus we may assume that any optimal estimator picks the right

θ in these cases, and the only question is what value the estimator returns if all the samples are identical. The above analysis of the one sample case can be used to show that any optimal shift-invariant estimator is optimal. □

5. The compact case

So far we have dealt only with distributions on $\mathbb{R}$, where the shift parameter is a translation. In every specific case that we have analyzed, we have found a shift-invariant estimator among the optimal estimators. In this section we prove that if $G = X$ is a compact group acting on itself by (left) multiplication, then at least for measures defined by density functions, there is always a shift-invariant estimator as good as any given estimator. In Section 6 we show that the compactness hypothesis cannot be eliminated entirely; we do not know how much it can be weakened, if at all.

We will continue to use both G and X as notations, in order to emphasize the distinction between the two roles played by this object. Eaton [3] discusses estimators in a context in which the group G acts on both the parameter space Θ and the sample space X. In his work, the sample space X is an arbitrary homogeneous space (i.e., a space with a transitive G-action). In this generality, shift-invariant estimators may not exist, since there may not even exist a function from X^n to Θ that preserves the G action. For this reason, we choose to identify the sample space with the group G.

As usual G acts diagonally on X^n; we denote the orbit space by Y. An element $\mathbf{y}$ of Y is an equivalence class $\mathbf{y} = [\mathbf{x}] = \{(gx_1, \ldots, gx_n) : g \in G\}$, which we identify with G via $(gx_1, \ldots, gx_n) \mapsto gx_1$. For $\mathbf{x} = (x_1, \ldots, x_n) \in X$ we denote by $\mathbf{x}_0$ the point $x_1^{-1}\mathbf{x}$; thus $\mathbf{x}_0$ is in the orbit of $\mathbf{x}$ and has first coordinate 1. The set $X_0 = \{\mathbf{x}_0 : \mathbf{x} \in X\} \subset X^n$ is naturally identified with Y.

Equip $G = X$ with a left- and right-invariant metric d, meaning that $d(gx, gy) = d(x, y) = d(xg, yg)$ for all $x, y, g \in G$. Let $B(g) = B_\delta(g)$ denote the ball of radius δ around $g \in G$. If S is a subset of a measure space (T, α) then we denote the measure of S variously by $\alpha(S)$, $\int_S d\alpha$, or $\int_T \chi_S \, d\alpha$. (The notation χ_S refers to the characteristic function of the set S.)

Fix δ and n, and let μ be an arbitrary measure on X. The following technical lemma says that to evaluate an integral over X^n, we can integrate over each G-orbit and then integrate the result over the orbit space.

Lemma 7. *There exist measures ν on Y and $\alpha_{\mathbf{y}}$ on each orbit $\mathbf{y}$ such that for any function F on X^n,*

$$\int_{X^n} F(\mathbf{x}) \, d\mu^n = \int_Y \int_G F(g\mathbf{x}_0) \, d\alpha_{\mathbf{y}} \, d\nu.$$

Proof: Let $\varphi : X^n \to X^n$ be defined by $\varphi(\mathbf{x}) = (x_1, x_1^{-1}x_2, \ldots, x_1^{-1}x_n)$ and π be the image of μ^n with respect to φ, i.e, $\pi(A) = \mu^n\{\varphi \in A\}$ for all

Borel $A \subset X^n$. Then

$$\int \widetilde{F}\left(\varphi(\mathbf{x})\right) d\mu^n(\mathbf{x}) = \int \widetilde{F}(\mathbf{y}) d\pi(\mathbf{y})$$

for nonnegative Borel functions $\widetilde{F}$ on X^n. Taking $\widetilde{F}(\mathbf{y}) = F(y_1, y_1y_2, \ldots, y_1y_n)$ yields

$$\begin{aligned}\int F(\mathbf{x}) d\mu^n(\mathbf{x}) &= \int F(y_1, y_1y_2, \ldots, y_1y_n) d\pi(\mathbf{y}) \\ &= \int d\widetilde{\nu}(y_2, \ldots, y_n) \int F(y_1, y_1y_2, \ldots, y_1y_n) d\widetilde{\alpha}_{(y_2,\ldots,y_n)}(y_1),\end{aligned}$$

for some measures $\widetilde{\nu}$ and $\widetilde{\alpha}$. The right-hand side can then be written as

$$\int_{X_0} d\nu(\mathbf{x}_0) \int_G F(g\mathbf{x}_0) d\alpha_{\mathbf{x}_0}(g),$$

where ν is an image of $\widetilde{\nu}$ with respect to the function $(y_2, \ldots, y_n) \mapsto (1, y_2, \ldots, y_n)$ and $\alpha_{\mathbf{x}_0} = \widetilde{\alpha}_{(y_2,\ldots,y_n)}$ for $\mathbf{x}_0 = (1, y_2, \ldots, y_n)$, completing the proof. □

Lemma 8. *If s is a shift-invariant (n-sample) estimator then*

$$Q(s) = \int_Y \int_G \chi_{B(s(\mathbf{x}_0)^{-1})} \, d\alpha_{\mathbf{x}_0} \, d\nu.$$

Proof: Since s is shift-invariant, its quality can be computed at the identity. Thus $Q(s) = Q^1(s) = \mu^n(s^{-1}(B(1))) = \int_{X^n} \chi_{s^{-1}(B(1))} \, d\mu^n$. By Lemma 7, this integral can be decomposed as

$$\int_Y \int_G \chi_{s^{-1}(B(1))}(g\mathbf{x}_0) \, d\alpha_{\mathbf{x}_0} \, d\nu.$$

Now, note that $g\mathbf{x}_0 \in s^{-1}(B(1))$ if and only if $gs(\mathbf{x}_0) \in B(1)$ if and only if $g \in B(s(\mathbf{x}_0)^{-1})$. Thus the integral above is the same as the one in the statement of the lemma, and we are done. □

We are now ready to prove the result. Note that we do not prove that optimal estimators exist—only that if they exist, then one of them is shift-invariant.

Theorem 9. *Let $G = X$ be a compact group, let δ and n be given, and let μ be defined by a density function. If $e : X^n \to G$ is any estimator then there exists a shift-invariant estimator s with $Q(s) \geq Q(e)$.*

Proof: Let $e : X^n \to G$ be any estimator. For each group element $\gamma \in G$, we define a shift-invariant estimator s_γ that agrees with e on the coset γX_0:

$$s_\gamma(g, gx_2, \ldots, gx_n) = g\gamma^{-1} e(\gamma, \gamma x_2, \ldots, \gamma x_n).$$

We will show that there exists γ such that $Q(s_\gamma) \geq Q(e)$.

Denote by ρ the invariant (Haar) measure on G. Since $Q(e)$ is defined as an infimum, we have

$$\begin{aligned} Q(e) \le \int_{\theta\in G} Q^\theta(e)\,d\rho &= \int_{\theta\in G}\int_{X^n} \chi_{\theta^{-1}e^{-1}(B(\theta))}(\mathbf{x})\,d\mu^n\,d\rho \\ &= \int_{X^n}\int_{\theta\in G} \chi_{\theta^{-1}e^{-1}(B(\theta))}(\mathbf{x})\,d\rho\,d\mu^n \\ &= \int_Y\int_G\int_{\theta\in G} \chi_{\theta^{-1}e^{-1}(B(\theta))}(g\mathbf{x}_0)\,d\rho\,d\alpha_{\mathbf{x}_0}\,d\nu, \end{aligned} \tag{6}$$

where the last equality comes from Lemma 7. The condition that $g\mathbf{x}_0 \in \theta^{-1}e^{-1}(B(\theta))$ is equivalent to $d(e(\theta g\mathbf{x}_0),\theta) < \delta$. Now we make the substitution $\gamma = \theta g$. Thus $\theta = \gamma g^{-1}$, and the condition becomes $d(e(\gamma\mathbf{x}_0),\gamma g^{-1}) < \delta$, or, by invariance of the metric, $d(\gamma^{-1}e(\gamma\mathbf{x}_0), g^{-1}) < \delta$. This says that $g^{-1} \in B(\gamma^{-1}e(\gamma\mathbf{x}_0))$, or equivalently, $g \in B(e(\gamma\mathbf{x}_0)^{-1}\gamma)$.

This allows us to rewrite the triple integral (6), using the measure-preserving transformation $\theta \mapsto \gamma = \theta g$, as

$$\begin{aligned} \int_Y\int_G\int_{\theta\in G} \chi_{\theta^{-1}e^{-1}(B(\theta))}\,d\rho\,d\alpha_{\mathbf{x}_0}\,d\nu &= \int_Y\int_G\int_{\gamma\in G} \chi_{B(e(\gamma\mathbf{x}_0)^{-1}\gamma)}\,d\rho\,d\alpha_{\mathbf{x}_0}\,d\nu \\ &= \int_{\gamma\in G}\left(\int_Y\int_G \chi_{B(e(\gamma\mathbf{x}_0)^{-1}\gamma)}\,d\alpha_{\mathbf{x}_0}\,d\nu\right)d\rho \end{aligned}$$

Now, comparing with Lemma 8, we see that the inner integral above is exactly the quality of the shift-invariant estimator s_γ.

We therefore have

$$Q(e) \le \int_\gamma Q(s_\gamma)\,d\rho,$$

or in other words, the average quality of the shift-invariant estimators $\{s_\gamma\}$ is at least $Q(e)$. Therefore at least one of the s_γ satisfies $Q(s_\gamma) \ge Q(e)$. □

6. A non-shift-invariant example

In the following example, suggested by L. Schulman, the optimal shift-invariant estimator is not optimal. This provides an interesting complement to Conjecture 1. Lehman and Casella [9, Section 5.3] also give examples of this phenomenon.

Let X be the infinite trivalent tree, which we view as the Cayley graph of the group $G = (\mathbb{Z}/2\mathbb{Z}) * (\mathbb{Z}/2\mathbb{Z}) * (\mathbb{Z}/2\mathbb{Z}) = \langle a, b, c \mid a^2 = b^2 = c^2 = 1\rangle$. In words, G is a discrete group generated by three elements a, b, and c, each of order two and with no other relations. Each non-identity element of G can be written uniquely as a finite word in the letters a, b, c with no letter appearing twice in a row; we refer to such a word as the *reduced form* of the group element. (We write 1 for the identity element of G.) Multiplication in the group is performed by concatenating words and then canceling any repeated letters in pairs. Evidently G is infinite. The Cayley graph X is a

graph with one vertex labeled by each group element of G, and with an edge joining vertices w and x if and only if $w = xa$, $w = xb$, or $w = xc$. Note that this relation is symmetric, since a, b, and c each have order 2. Each vertex of X has valence 3, and X is connected and contains no circuits, i.e. it is a tree. Finally, X becomes a metric space by declaring each edge to have length one.

Because of how we defined the edges of X, G acts naturally on the left of X: given $g \in G$, the map $g : X \to X$ defined by $g(x) = gx$ is an isometry of X. So if $\delta > 0$ is given, μ is a probability distribution, $\theta \in G$, and e is an estimator, then the shift μ_θ and the quality $Q_\delta(e)$ are defined as usual by (1) and (2).

We are ready to present the example. Suppose $0 < d < 1$ is fixed, and let μ be the probability distribution with atoms of weight $1/3$ at the three vertices a, b, c. Thus for $\theta \in G$, the distribution μ_θ has atoms of weight $1/3$ at the three neighbors $\theta a, \theta b, \theta c$ of the vertex θ in X.

Example 6.1. *There is an optimal one-sample estimator with quality* $2/3$*, but the quality of any shift-invariant one-sample estimator is at most* $1/3$*.*

Proof: Consider the one-sample estimator e that truncates the last letter of the sample (unless the sample is the identity, in which case we arbitrarily assign the value a). That is, for a vertex x of X,

$$e(x) = \begin{cases} w \text{ if } x = w\ell \text{ is reduced, and } \ell = a, b, \text{ or } c \\ a \text{ if } x = 1. \end{cases}$$

Geometrically, this estimator takes a sample x and, unless $x = 1$, guesses that the shift is the (unique) neighbor of x that is closer to 1.

We compute the quality of e. Note $Q^1(e) = 1$, because if $\theta = 1$ then the sample will be a, b, or c, and the estimator is guaranteed to guess correctly. In fact $Q^a(e) = 1$ also, as is easily verified. For any other shift θ, the sample is $\theta\ell$ for $\ell = a, b$, or c, and the estimator guesses correctly exactly when ℓ differs from the last letter of θ. So $Q^\theta(e) = 2/3$, and $Q(e) = \inf_\theta Q^\theta(e) = 2/3$.

It is easy to see that this estimator is optimal. Suppose e' is another estimator and $Q(e') > 2/3$. Since each local quality $Q^\theta(e')$ is either 0, $1/3$, $2/3$, or 1, we must have $Q^\theta(e') = 1$ for all θ. This means e' always guesses correctly. But since there are different values of θ that can produce the same sample, this is impossible.

Observe that the estimator e above is neither left- nor right-invariant. For instance right-invariance fails, as $e(ba \cdot a) = e(b) = 1 \neq ba = e(ba) \cdot a$, and the same example shows the failure of left-invariance: $e(b \cdot aa) = id \neq ba = be(aa)$.

Indeed, we conclude by showing that the quality of any shift-invariant one-sample estimator e' is at most $1/3$. Suppose $e'(1) = w$. If e' is left-invariant, it follows that $e'(x) = xw$ for all x; if e' is right-invariant it follows that $e'(x) = wx$ for all x.

Since $\delta < 1$, the quality of e' at $\theta = 1$ is equal to the probability that $e'(x) = 1$, given that x was sampled from μ. With equal probability x is a, b, or c; since at most one of wa, wb, wc and one of aw, bw, cw can equal 1, we conclude that $Q(e') \leq Q^1(e') \leq 1/3$. □

We remark that this example readily generalizes to other finitely generated groups with infinitely many ends: the key is that e is a two-to-one map but with only one sample, a shift-invariant estimator is necessarily one-to-one.

References

[1] Berger, J. O., Statistical Decision Theory and Bayesian Analysis, second edition, Springer-Verlag New York, Inc., 1985.

[2] Bondar, J. V. and P. Milnes, *Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups*, Probability Theory and Related Fields, vol 57 (1981), pp. 103–128.

[3] Eaton, Morris L., Group invariance applications in statistics. Regional Conference Series in Probability and Statistics, Volume 1, Institute of Mathematical Statistics (Hayward, CA) and American Statistical Association (Alexandria, VA), 1989.

[4] Farrell, R. H., *Estimation of a location parameter in the absolutely continuous case*, Annals of Mathematical Statistics, vol 35 (1964), pp. 949–998

[5] Federer, H. Geometric Measure Theory. Die Grundlehren der mathematischen Wissenschaften, Band 153, Springer-Verlag New York, Inc., 1969.

[6] Fisher, R. A., *Two new properties of mathematical likelihood,* Proceedings of the Royal Society of London, Series A, vol 144 no 852 (Mar. 29, 1934), pp. 285–307.

[7] Kamberova, G. and M. Mintz, *Minimax rules under zero-one loss for a restricted location parameter*, Journal of Statistical Planning and Inference, vol 79 (1999), no. 2, pp. 205–221.

[8] Kiefer, J. *Invariance, minimax sequential estimation, and continuous time processes*, Annals of Mathematical Statistics, vol 28 (1957), pp. 573–601.

[9] Lehman, E. L. and G. Casella, Theory of Point Estimation, second edition, Springer-Verlag Inc. (New York), 1998.

[10] Pitman, E. J. G., *The estimation of the location and scale parameters of a continuous population of any given form,* Biometrika, vol 30 no 3/4 (Jan., 1939), pp. 391–421.

[11] Robert, C. P. , The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, second edition, Springer Science & Business Media, LLC. (New York), 2007.

[12] Schulman, L. and V. Vazirani, *A computationally motivated definition of parametric estimation and its applications to the Gaussian distribution,* Combinatorica, vol 25 no 4 (2005), pp. 465–486.

[13] Wijsman, Robert A., Invariant measures on groups and their use in statistics. Lecture Notes—Monograph Series, Volume 14, Institute of Mathematical Statistics (Hayward, CA), 1990.