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 [2], 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.

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 \( \mu \) on a space \( X \), along with an action of a group \( G \) on \( X \), which defines a set of translated distributions \( \mu_\theta \) as follows:

\[
(1) \quad \mu_\theta(A) = \mu(\{x : \theta x \in A\})
\]

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

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

minimizes the mean squared error if \( \mu \) 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 [2] to analyze certain approximation algorithms in computer science. Precisely, the \( \delta \)-quality of \( e \) is defined as

\[ Q_\delta(e) = \inf_{\theta} Q_{\delta}(e) \]

\[ = \inf_{\theta} \Pr \{ d(\hat{\mu}(x), \theta) < \delta : x_i \text{ are chosen from } \mu_\theta \} \]

\[ = \inf_{\theta} \mu\left( \{ x : d(\hat{\mu}(x), \theta) < \delta \} \right) , \]

where \( d \) is a metric on \( X \) and \( \mu^n_\theta \) is the product measure \( \mu_\theta \times \cdots \times \mu_\theta \) on \( X^n \). We write \( Q(e) \) when the value of \( \delta \) is unambiguous. For fixed \( \delta \), an \( (n\text{-sample}) \) estimator \( e \) is optimal if \( Q_\delta(e) \geq Q_\delta(e') \) for all \( (n\text{-sample}) \) estimators \( e' \).

Motivated initially by analyzing an approximate algorithm for determining the average matching size in a graph, Schulman and Vazirani [2] introduce the notion of a majorizing estimator, which is optimal (by the above definition) simultaneously for all \( \delta > 0 \). They focus on the Gaussian distribution and prove that the mean estimator is the unique majorizing estimator in this case.

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 \( \theta \). 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 have analyzed, we are able to construct an optimal estimator that is

---

1In the case of perverse measures, \( \mu \), we must consider the probability as the sup of the intersection of the set \( \{ d(\mu(x), \theta) < \delta \} \) with all measurable sets. We will ignore this caveat throughout.
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

$$e(gx) = ge(x)$$

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 $\mu$ is given by a density function and $e$ is an estimator for $\mu$, 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 $e$. 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 seems to be an interesting problem to determine conditions under which one can guarantee the existence of a shift-invariant estimator which is optimal.

The authors thank V. Vazirani and L. Schulman for suggesting the problem that motivated this paper along with subsequent helpful discussions. 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 $\mu$ on $\mathbb{R}$, and we are to guess which distribution $\mu_\theta$ produced a given collection $x = (x_1, \ldots, x_n)$ of data, where $\mu_\theta(A) = \mu\{x : x + \theta \in A\}$. An estimator is a function $e : \mathbb{R}^n \rightarrow \mathbb{R}$, and we want to maximize its quality, which is given by

$$Q(e) = \inf_{\theta} Q^\theta(e) = \inf_{\theta} \Pr\{|e(x) - \theta| < \delta : x_i \text{ are chosen from } \mu_\theta\}$$

$$= \inf_{\theta} \mu_\theta^\delta(\{x : |e(x) - \theta| < \delta\}).$$

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 $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $a \in \mathbb{R}$, then $x + a$ 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 $Y + A = \{y + a : y \in Y, a \in A\}$. We also use the “interval notation” $(x+a, x+b)$ for the set $\{x+t : a < t < b\}$; this is a segment of length $(b-a)\sqrt{n}$ in $\mathbb{R}^n$ if $a$ and $b$ are finite. If $f : \mathbb{R}^n \rightarrow \mathbb{R}$
is any function, and \( \theta \in \mathbb{R} \), define \( f_\theta(x) = f(x - \theta) \). If \( f : \mathbb{R} \to \mathbb{R} \), then define \( f^{[n]} : \mathbb{R}^n \to \mathbb{R} \) by \( f^{[n]}(x) = f(x_1)f(x_2) \cdots f(x_n) \).

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

\[
X_0 = \{ x \in \mathbb{R}^n : x_1 = 0 \},
\]

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

\[
\mu^0_\theta (\{ x : |e(x) - \theta| < \delta \}) = \mu^0_\theta (\{ x : |e(x - \theta)| < \delta \}) = \mu^0 (\{ x : |e(x)| < \delta \}),
\]

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

**Definition 1.** For fixed \( n \), let \( A \) denote the collection of all measurable subsets of \( \mathbb{R}^n \) whose intersection with each of the lines \( x + \mathbb{R} \) is contained in a segment of length \( 2\delta \sqrt{n} \). Precisely,

\[
A = \{ A \subset \mathbb{R}^n : \text{ for all } x \in \mathbb{R}^n, \quad A \cap (x + \mathbb{R}) \subseteq [y - \delta, y + \delta] \text{ for some } y \in x + \mathbb{R} \}.
\]

For fixed \( \mu \) and \( n \), define

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

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

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 \( A = e^{-1}([\delta, 2\delta]) \). Note that

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

and in particular \( A \in A \). Thus the quality of \( e \) is at most \( S_{\mu,n} \). \( \square \)

**Theorem 3.** Let \( \mu \) 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 \( A \in A \), we will define a shift-invariant estimator \( e \) with the property that \( A \subset e^{-1}([\delta, 2\delta]) \). Then \( Q(e) \geq \mu^n(A) \). The theorem then follows from the definition of \( S_{\mu,n} \).

So, fix \( A \in A \). For each \( x \in \mathbb{R}^n \) choose the supremum of those \( m \)'s and the infimum of \( M \geq m \) such that \( A \cap (x + \mathbb{R}) \subset (x + m, x + M) \). Define \( e(x) \) on \( X_0 \) by \( e(x) = (M + m)/2 \), and then extend \( e \) to all of \( \mathbb{R}^n \) to make it shift-invariant. Now clearly \( A \subset e^{-1}([\delta, 2\delta]) \), since if \( x \in A \) then \( |M + m| \leq 2\delta \). This completes the proof. \( \square \)
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 $\delta$ is fixed throughout. A randomized estimator is a distribution over estimators, i.e. given $x$, an estimator is chosen in some specified but random way, and then the selected estimator is evaluated at $x$. The $\delta$-quality of a randomized estimator $\bar{e}$ is

$$Q(\bar{e}) = Q_\delta(\bar{e}) = \inf_{\theta} \mathbb{E}_{e \in \bar{e}}(Q^\theta_\delta(e)).$$

**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 $\mu$ 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 $A \subset B$ and hence $S_{\mu,n} \leq T_{\mu,n}$.

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

Proof: We will give a complete proof in the case that $\mu$ 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 $\theta$ is $\mu^\theta_\delta(\theta^{-1}((\theta - \delta, \theta + \delta)))$. To simplify notation we will let $I_{e,\theta}$ denote the set $\mu^{-1}((\theta - \delta, \theta + \delta))$ 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 $Q(e)$:

$$Q(e) \leq \frac{1}{k} \sum_{i=1}^{k} \mu^\theta_\delta(I_{\theta_i}).$$

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

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

Observe that $\tilde{f}(x)$ approaches 0 as any $x_j$ approaches $\pm \infty$. 


Since the $I_{\theta_i}$ are disjoint, we now have

$$Q(e) \leq \frac{1}{k} \sum_{i=1}^{k} \int_{I_{\theta_i}} f^{[n]}(x)dx$$

$$= \frac{1}{k} \int_{\bigcup I_{\theta_i}} f^{[n]}(x)dx$$

$$\leq \frac{1}{k} \int_{\bigcup I_{\theta_i}} \tilde{f}(x)dx$$

$$\leq \frac{1}{k} \int_{\mathbb{R}^n} \tilde{f}(x)dx$$

(5) $$= \frac{1}{k} \int_{\{x_1 \leq 0\}} \tilde{f}(x)dx + \frac{1}{k} \int_{\{0 < x_1 < 2\delta_k\}} \tilde{f}(x)dx + \frac{1}{k} \int_{\{x_1 \geq 2\delta_k\}} \tilde{f}(x)dx.$$

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 $a \in \mathbb{R}$,

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

To prove the claim, set $Z = \{x \in \mathbb{R}^n : a \leq x_1 \leq a + 2\delta\}$, and set $Z_i = \{x \in Z : i$ is the smallest index such that $f(x) = f^{[n]}(x)\}$. Thus the $Z_i$ are disjoint and cover the interior of $Z$. Now

$$\int_{\{a \leq x_1 \leq a + 2\delta\}} \tilde{f}(x)dx = \sum_i \int_{Z_i} \tilde{f}(x)dx = \sum_i \int_{Z_i} f^{[n]}(x)dx$$

$$= \sum_i \int_{Z_i - \theta_i} f^{[n]}(x)dx$$

$$= \int_{\bigcup (Z_i - \theta_i)} f^{[n]}(x)dx$$

$$\leq T_{\mu,n}.$$  

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 $B = \bigcup (Z_i - \theta_i)$ is in $B$. This proves the claim. $\square$

Next we show that $\frac{1}{k} \int_{\{x_1 \leq 0\}} \tilde{f}(x)dx$ approaches zero as $k$ grows. Recall that $\theta_i = 2\delta i$, and set $z_i = \int_{\{x_1 \leq 0\}} f^{[n]}(x)dx = \int_{\{x_1 \leq -2\delta i\}} f^{[n]}(x)dx$. The function $f$ is a probability density function, so $f$ is nonnegative, has total integral 1, and approaches 0 as $|x| \to \infty$. Thus the sequence $\{z_i\}$ is decreasing.
to 0. Bounding $\tilde{f}(x)$ by $\sum_{i=1}^{k} f^{[n]}_{\theta_i}(x)$ we have

$$\frac{1}{k} \int_{\{x_1 \leq 0\}} \tilde{f}(x) dx \leq \frac{1}{k} \int_{\{x_1 \leq 0\}} \sum_{i=1}^{k} f^{[n]}_{\theta_i}(x) dx = \frac{1}{k} \sum_{i=1}^{k} z_i \to 0.$$ \[1\]

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

We have shown that for any $\epsilon > 0$, we can find $k$ depending on $\epsilon$ 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 $\tilde{e}$, 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 $Q(\tilde{e}) \leq T_{\mu,n} + 2\epsilon$, and the theorem follows.

The proof is now complete in the case that $\mu$ has a density $f$. In general, the argument requires minor technical adjustments. The first step that requires modification is the definition of the function $\tilde{f}$; without a density one can still define a “supremum” $\tilde{\mu}$ of the measures $\mu^n_{\theta_i}$ as follows:

$$\tilde{\mu}(A) = \sup \sum_{j} \max_{i} \mu^n_{\theta_i}(A_j),$$

where the supremum is taken over all countable partitions $A = \cup A_j$ of the set $A$. One then works with $\tilde{\mu}$ rather than $\tilde{f}$, and the remainder of the argument goes through with corresponding changes. \[2\]

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. These examples include many familiar distributions, and they provide evidence for the following conjecture.

**Conjecture 1.** Let $\mu$ be a distribution on $\mathbb{R}$. Then there is an optimal estimator for $\mu$ 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.

**Example 4.1.** Let $\mu$ be defined by a density function $f$ with a unique local maximum. 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 $B$ that any set $B \in B$ must have Lebesgue measure less than or equal to $2\delta$. Since $f$ is unimodal, $\int_B f(x)$ 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 \( B \) is contained in \( A \) 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.

\[ \Box \]

### 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 \( \mu \) 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 \( x \in X_0 \), we define a function \( h_x : \mathbb{R} \rightarrow \mathbb{R} \) by

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

Since

\[
h_x'(t) = e^{\lambda(x_1 + t) + \cdots + \lambda(x_n + t)} (\lambda'(x_1 + t) + \cdots + \lambda'(x_n + t))
\]

and \( \lambda' \) is decreasing, it is clear that for each \( x \), \( h_x'(t) = 0 \) for at most one value of \( t \). Since \( h_x(t) \to 0 \) as \( t \to \pm \infty \), it follows that for any \( x \), \( h_x \) 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 \( x + \mathbb{R} \) 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 \( x \in X_0 \), let \( t_x \) be the center of the length \( 2\delta \) interval \( I = (t_x + \delta, t_x - \delta) \) that maximizes \( \int_I h_x \, dt \). Then let

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

Now \( A \in \mathcal{A} \), and moreover \( \mu^n(A) \geq \mu^n(B) \) for any \( B \in \mathcal{B} \), because

\[
\int_{A \cap (x + \mathbb{R})} f^{[n]} \geq \int_{B \cap (x + \mathbb{R})} f^{[n]} \text{ for each } x \in X_0.
\]

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. \( \Box \)

### 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 is not the maximum likelihood estimator. (Note that in a typical estimation problem involving a family of exponential distributions, one is trying to estimate \( \lambda \) rather than a “horizontal” shift \( \theta \).)
Example 4.3. Let \( \mu \) 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 \( x \in \mathbb{R}^n \), define \( e(x) = \min\{x_1, \ldots, x_n\} - \delta \). 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 \( B = \{ x \in \mathbb{R}^n : 0 \leq \min\{x_1, \ldots, x_n\} < 2\delta \} \). 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 \( C' = \{ x \in B : x + 2k\delta \in C \) for some \( k \in \mathbb{Z} \} \). Note that \( k \) is determined uniquely by \( x \). Now \( C' \subset B \) is in \( \mathcal{B} \), and by our hypotheses on \( f \), if \( x \in B \) then \( f(x) \geq f(x + 2k\delta) \) for every integer \( k \). Therefore \( \mu^n(C') - \mu^n(C) = \int_{C'} [f(x) - f(x + 2k\delta)] d\mu^n \geq 0 \). \( \square \)

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

Example 4.4. Let \( \mu \) 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 \( \mu \) is discrete, the supremum defining \( S_{\mu,1} \) is attained; therefore by Theorems 3 and 5 it suffices to show that every estimator shows 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) ∈ (θ − δ, θ + δ), and the right hand side above may be rewritten as
\[ \sum_{\theta \in Y} \sum_z p_z = \sum_{\theta \in Y} \sum_{x \in Y + Z} p_{x - \theta} = \sum_{x \in Y + Z} p_{x - \theta}, \]
with the inner sum now taken over all θ with e(x) ∈ (θ − δ, θ + δ). This latter condition implies that z is within δ of x − e(x). But by definition, S_{µ,1} is the maximum measure of any interval of length 2δ. Hence, for any fixed x ∈ Y + Z, the inner sum is at most S_{µ,1}, and the entire sum is thus bounded above by S_{µ,1} |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 ⊆ Y_{k+1} and |Y_k| ≤ k^r. It follows that for any ε > 0 there exists k such that |Y_{k+1}|/|Y_k| < 1 + ε, for otherwise |Y_k| would grow at least exponentially in k. Using the fact that Y_k + Z ⊆ Y_{k+1}, the lemma applied to Y_k implies that Q(e) ≤ S_{µ,1}(1 + ε). Therefore Q(e) ≤ S_{µ,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 ≠ 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. As usual G acts diagonally on X^n; we denote the orbit space by Y. An element y of Y is an equivalence class y = [x] = \{ (gx_1, ..., gx_n) : g \in G \}, which we identify with G via (gx_1, ..., gx_n) ↦ gx_1. For x = (x_1, ..., x_n) ∈ X we denote by x_0 the point x_1^{-1}x; thus x_0 is in the orbit of x and has first coordinate 1. The set X_0 = \{ x_0 : x \in X \} ⊂ 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 \( \delta \) around \( g \in G \). If \( S \) is a subset of a measure space \((T, \alpha)\) then we denote the measure of \( S \) variously by \( \alpha(S) \), \( \int_S \, d\alpha \), or \( \int_T \chi_S \, d\alpha \). (The notation \( \chi_S \) refers to the characteristic function of the set \( S \).)

Fix \( \delta \) and \( n \), and let \( \mu \) be an arbitrary measure on \( X \) defined by a density function \( f \). Our explicit use of the density function begins and ends with the following technical fact about \( \mu \), which 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 \( \nu \) on \( Y \) and \( \alpha_y \) on each orbit \( y \) such that
\[
\int_{X^n} F(x) \, d\mu^n = \int_Y \int_G F(gx_0) \, d\alpha_y \, d\nu.
\]

Proof: The quotient map \( X^n \to Y \) is a Lipschitz fibration with nonzero Jacobian \( J \), so locally we can apply the co-area formula of Federer [Theorem 3.2.12 of [1]] to the function \( F \cdot f / J \). Using a partition of unity we get the global result 
\[
\int_{X^n} F(x) \, d\mu^n = \int_Y \int_G F(x) \, d\alpha_y \, d\nu.
\]
Setting \( g = x_1 \) finishes the proof, since \( x = x_1x_0 \).

**Lemma 8.** If \( s \) is a shift-invariant \((n\text{-sample})\) estimator then
\[
Q(s) = \int_Y \int_G \chi_{B(s(x_0)^{-1})} \, d\alpha_y \, d\nu.
\]

Proof: Since \( s \) is shift-invariant, its quality can be computed at the identity. Thus
\[
Q(s) = Q(s(s^{-1}(1))) = \int_{X^n} \chi_{s^{-1}(B(1))} \, d\mu^n.
\]
By Lemma 7, this integral can be decomposed as
\[
\int_Y \int_G \chi_{s^{-1}(B(1))}(gx_0) \, d\alpha_y \, d\nu.
\]
Now, note that \( gx_0 \in s^{-1}(B(1)) \) if and only if \( gs(x_0) \in B(1) \) if and only if \( g \in B(s(x_0)^{-1}) \). 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 \( \delta \) and \( n \) be given, and let \( \mu \) 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 \( s_\gamma \) that agrees with \( e \) on the coset \( \gamma X_0 \):
\[
s_\gamma(g, gx_2, \ldots, gx_n) = g\gamma^{-1}e(\gamma, \gamma x_2, \ldots, \gamma x_n).
\]
We will show that there exists $\gamma$ such that $Q(s_\gamma) \geq Q(e)$. 

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

\[
Q(e) \leq \int_{\theta \in G} Q^\theta(e) \, d\rho = \int_{\theta \in G} \int_{X^n} \chi_{\theta^{-1}e^{-1}(B(\theta))}(x) \, d\mu^n \, d\rho = \int_{X^n} \int_{\theta \in G} \chi_{\theta^{-1}e^{-1}(B(\theta))}(x) \, d\rho \, d\mu^n
\]

\[
\overset{(6)}{=} \int_{Y} \int_{G} \int_{\theta \in G} \chi_{\theta^{-1}e^{-1}(B(\theta))}(g\xi_0) \, d\rho \, d\alpha_y \, d\nu,
\]

where the last equality comes from Lemma 7. The condition that $g\xi_0 \in \theta^{-1}e^{-1}(B(\theta))$ is equivalent to $d(e(\theta g\xi_0), \theta) < \delta$. Now we make the substitution $\gamma = \theta g$. Thus $\theta = \gamma g^{-1}$, and the condition becomes $d(e(\gamma \xi_0), \gamma g^{-1}) < \delta$, or, by invariance of the metric, $d(\gamma^{-1} e(\gamma \xi_0), g^{-1}) < \delta$. This says that $g^{-1} \in B(\gamma^{-1} e(\gamma \xi_0))$, or equivalently, $g \in B(e(\gamma \xi_0) - \gamma g)$. 

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

\[
\int_{Y} \int_{G} \int_{\theta \in G} \chi_{\theta^{-1}e^{-1}(B(\theta))} \, d\rho \, d\alpha_y \, d\nu = \int_{Y} \int_{G} \int_{\gamma \in G} \chi_{B(e(\gamma \xi_0) - \gamma)} \, d\rho \, d\alpha_y \, d\nu
\]

\[
= \int_{\gamma \in G} \left( \int_{Y} \int_{G} \chi_{B(e(\gamma \xi_0) - \gamma)} \, d\alpha_y \, d\nu \right) \, d\rho
\]

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

We therefore have

\[
Q(e) \leq \int_{\gamma} Q(s_\gamma) \, d\rho,
\]

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 $s_\gamma$ satisfies $Q(s_\gamma) \geq Q(e)$. □

6. A non-shift-invariant example

The following example was suggested by L. Schulman, and provides an interesting complement to Conjecture 1.

Let $X$ be the infinite trivalent tree, which we view as the Cayley graph of the group $G = (\mathbb{Z}/2\mathbb{Z}) \ast (\mathbb{Z}/2\mathbb{Z}) \ast (\mathbb{Z}/2\mathbb{Z}) = \langle a, b, c \mid a^2 = b^2 = c^2 = 1 \rangle$. In words, $G$ is a discrete group generated by three elements $a, b, c$, 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 $g : X \rightarrow X$ defined by $g(x) = gx$ is an isometry of $X$. So if $\delta > 0$ is given, $\mu$ is a probability distribution, $\theta \in G$, and $e$ is an estimator, then the shift $\mu_\theta$ and the quality $Q_\delta(e)$ are defined as usual by (1) and (2).

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

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

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 $a, b, \text{ or } c$, and the estimator is guaranteed to guess correctly. In fact $Q^\theta(e) = 1$ also, as is easily verified. For any other shift $\theta$, the sample is $\theta \ell$ for $\ell = a, b, \text{ or } c$, and the estimator guesses correctly exactly when $\ell$ differs from the last letter of $\theta$. 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 $\theta$. This means $e'$ always guesses correctly. But since there are different values of $\theta$ 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 $\mu$. 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
