Proceedings of Machine Learning Research vol 247:1–15, 2024 37th Annual Conference on Learning Theory

An information-theoretic lower bound in time-uniform estimation

John Duchi JDUCHI@STANFORD.EDU
Departments of Statistics and Electrical Engineering, Stanford University

Saminul Haque SAMINULH@STANFORD.EDU
Department of Computer Science, Stanford University

Editors: Shipra Agrawal and Aaron Roth

Abstract

We present an information-theoretic lower bound for the problem of parameter estimation with time-uniform coverage guarantees. Via a new a reduction to sequential testing, we obtain stronger lower bounds that capture the hardness of the time-uniform setting. In the case of location model estimation, logistic regression, and exponential family models, our $\Omega(\sqrt{n^{-1}\log\log n})$ lower bound is sharp to within constant factors in typical settings.

1. Introduction

Let $X_1, X_2, \dots$ be independent samples from a common distribution $P \in \mathcal{P}$ over a domain $\mathcal{X}$. Given a parameter $\theta : \mathcal{P} \to \Theta \subset \mathbb{R}$ and an error-bound function $t : \mathbb{N} \to \mathbb{R}_+$, we say a sequence of estimators $\{\hat{\theta}_n\}_{n\in\mathbb{N}}$ is an $(\alpha, t(\cdot))$-*estimator* if for all $P \in \mathcal{P}$,

$$\mathbb{P}_{X_i \overset{\text{iid}}{\sim} P}\left(|\hat{\theta}_n(X_{1:n}) - \theta(P)| \le t(n), \text{ for all } n \in \mathbb{N}\right) \ge 1 - \alpha. \tag{1}$$

We provide fundamental limits on $t(\cdot)$ that imply there exist no $(\alpha, t(\cdot))$-estimators.

Our motivation comes from the problem of estimating time-uniform confidence sequences (Darling and Robbins, 1967; Lai, 1976), where the analyst must return a confidence set C_n at every n such that $\mathbb{P}(\theta(P) \in C_n \text{ for all } n) \ge 1 - \alpha$. Darling and Robbins (1967) introduced confidence sequences as a means of allowing statistical inference without committing *a priori* to a fixed sample size (Robbins, 1970). Their time-uniformity has made confidence sequences are a popular tool in sequential data analysis, such as in clinical testing (Jennison and Turnbull, 1989; Lai, 1984), A/B testing (Johari et al., 2022; Howard et al., 2021), and bandit arm identification (Jamieson et al., 2014; Kaufmann et al., 2016).

Methods for producing confidence sequences often see the size of the confidence intervals decay at a rate of $\Theta(\sqrt{n^{-1}\log\log n})$ (Darling and Robbins, 1967; Howard et al., 2021; Jamieson et al., 2014), ignoring dependence on parameters other than n. This stands in contrast to the fixed sample size setting, where the confidence intervals instead are of size $\Theta(n^{-1/2})$; thus, these methods incur a $\Theta(\sqrt{\log\log n})$ factor to achieve time-uniform coverage. Farrell (1964) (and Jamieson et al. (2014), via a reduction to Farrell's results) shows this cost is necessary in exponential families. Their lower bound technique proceeds in two steps: first, they show an optimal estimator takes the form of a sufficient statistic, and second, they use the law-of-the-iterated-logarithm to show that this statistic has fluctuations on the order of $\Theta(\sqrt{n^{-1}\log\log n})$ infinitely often. This argument crucially relies on the particular structure of the one-parameter exponential family model, which allows one to argue that an optimal estimator necessarily thresholds a sufficient statistic, making it unclear how, or even if, the bounds generalize to other scenarios. In this paper, we elicit the same $\Theta(\sqrt{\log\log n})$ cost through information-theoretic techniques, which have the benefit of extending to broader families of problems via now familiar reductions (e.g. Wainwright, 2019, Chapter 14).

Our techniques rely on standard information-theoretic results, such as Le Cam's two point method, the Bretagnolle-Huber inequality, and the reduction from estimation to testing. Our novel technique in proving this lower bound is to reduce to a sequence of testing problems that grows iteratively harder. By accumulating the errors from these tests, we obtain the extra $\Theta(\sqrt{\log\log n})$ factor in the lower bound.

Notation We use $\mathcal{V}$ to denote the set of infinite binary sequences $\{0,1\}^\infty$. For $v \in \mathcal{V}$ and $l, k \in \mathbb{N}$, we use $v_{k:l}$ to denote the $(l-k+1)$-length binary string $(v_k, \ldots, v_l)$. If $l < k$, we understand $v_{k:l}$ to be the zero-length null sequence. Given finite-length bit strings $b \in \{0,1\}^k, b' \in \{0,1\}^l$, we use $b \oplus b'$ to denote the concatenation $(b_1, \ldots, b_k, b'_1, \ldots, b'_l) \in \{0,1\}^{k+l}$. We use $\mathsf{Uni}^{\mathcal{V}}$ to denote the uniform distribution over $\mathcal{V}$. Formally, if $V \sim \mathsf{Uni}^{\mathcal{V}}$, then for any $k \in \mathbb{N}$ and $b \in \{0,1\}^k$, $\mathbb{P}(v_{1:k} = b) = 2^{-k}$. Also for any $b \in \{0,1\}^\star$, we use $\mathsf{Uni}^{\mathcal{V}}_b$ to denote the distribution $\mathsf{Uni}^{\mathcal{V}}$ conditioned on the event that b is a prefix. We will use the property that for any $b \in \{0,1\}^\star$, $\mathsf{Uni}^{\mathcal{V}}_b = \frac{1}{2}\mathsf{Uni}^{\mathcal{V}}_{b\oplus(0)} + \frac{1}{2}\mathsf{Uni}^{\mathcal{V}}_{b\oplus(1)}$.

2. Reduction from estimation to testing

We adapt the classical approach to information-theoretic lower bounds of reducing estimation to testing—showing that if one can solve the estimation problem (1), then one can distinguish between many distributions. Accordingly, let $\{P_v\}_{v\in\mathcal{V}}$ be a family of probability distributions and for each $v \in \mathcal{V}$, let $\mathbb{P}_v$ be the probability space where $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P_v$. Under this setting, we define the following time-uniform test:

Definition 1 **($(\alpha, \{n_k\}_{k\in\mathbb{N}})$-test)** *A sequence of $\{0,1\}$-valued randomized tests $\{\hat{V}_k\}_{k\in\mathbb{N}}$ is an $(\alpha, \{n_k\})$-test if for all $v \in \mathcal{V}$, $\mathbb{P}_v(\hat{V}_k(X_{1:n_k}) = v_k$ for all $k \in \mathbb{N}) \geq 1-\alpha$.*

The testing problem will be to find an $(\alpha, \{n_k\})$-test given $\{P_v\}$. Conversely, we will show that if $\{n_k\}$ does not grow fast enough, then no $(\alpha, \{n_k\})$-test can exist. By reducing the original estimation problem (1) to this testing problem, the lower bound we find on $\{n_k\}$ will confer lower bounds on $t(\cdot)$ for the original testing problem.

2.1. Reduction from estimation to testing

We return to the estimation setting (1), where we have distributions $\mathcal{P}$ and a parameter $\theta : \mathcal{P} \to \Theta$. For a sub-family $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$, say that $\{P_v\}$ has *parameter separation* $\{\delta_k\}_{k\in\mathbb{N}}$ if for all $k \in \mathbb{N}$ and all $v, v' \in \mathcal{V}$ such that the prefixes $v_{1:k} \neq v'_{1:k}$, we have $\delta_k \leq |\theta(P_v) - \theta(P_{v'})|$. This yields the following proposition which reduces estimation to testing.

Proposition 2 *Suppose $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$ has parameter separation $\{\delta_k\}_{k\in\mathbb{N}}$ and $\hat{\theta}$ is an $(\alpha, t(\cdot))$-estimator. Then for $n_k = \inf\{n : t(n) < \delta_k/2\}$ there exists an $(\alpha, \{n_k\})$-test.*

Proof For ease of notation, we use $\theta(v)$ to denote $\theta(P_v)$ and use $\hat{\theta}_n$ to refer to $\hat{\theta}_n(X_{1:n})$. We now construct a test sequence $\hat{V} = \{\hat{V}_k\}_{k\in\mathbb{N}}$. First define the projection function

$$\hat{v}(\vartheta) = \operatorname*{argmin}_{v\in\mathcal{V}} |\vartheta - \theta(v)|,$$

where we break potential ties arbitrarily, and then take $\hat{V}_k(X_{1:n_k}) = \hat{v}(\hat{\theta}_{n_k})_k$.

To see that $\hat{V}$ is an $(\alpha, \{n_k\})$-test sequence, let $v \in \mathcal{V}$ and consider the event $E = \cap_{n \in \mathbb{N}} \{|\hat{\theta}_n - \theta(v)| \leq t(n)\}$. By the time-uniform utility of $\hat{\theta}$, $\mathbb{P}_v(E) \geq 1 - \alpha$, and so it suffices to show that on the event E, $\hat{V}_k = v_k$ for all $k \in \mathbb{N}$. It follows by definition of $\hat{v}$ and n_k that for all $k \in \mathbb{N}$,

$$|\hat{\theta}_{n_k} - \theta(\hat{v}(\hat{\theta}_{n_k}))| \leq |\hat{\theta}_{n_k} - \theta(v)| \leq t(n_k) < \tfrac{1}{2}\delta_k.$$

Therefore, $|\theta(\hat{v}(\hat{\theta}_{n_k})) - \theta(v)| < \delta_k$ by triangle inequality, which implies by the definition of δ_k that $v_{1:k} = \hat{v}(\hat{\theta}_{n_k})_{1:k}$ and, in particular, $v_k = \hat{v}(\hat{\theta}_{n_k})_k = \hat{V}_k$. ■

2.2. Hardness of the testing problem

We say that $\{P_v\}$ has *distribution closeness* $\{\Delta_k\}_{k \in \mathbb{N}}$ if for all $k \in \mathbb{N}$ and for all $v, v' \in \mathcal{V}$ such that $v_{1:k} = v'_{1:k}$, we have that $D_{\text{kl}}(P_v \| P_{v'}) \leq \Delta_k$. Close distributions imply a lower bound on sequences $\{n_k\}$ for which $(\alpha, \{n_k\})$-tests exist.

Theorem 3 *Suppose $\{P_v\}_{v \in \mathcal{V}}$ is a testing problem with distribution closeness $\{\Delta_k\}$. If there exists an $(\alpha, \{n_k\})$-test, then for infinitely many $k \in \mathbb{N}$, $n_k > (1 - \alpha)\Delta_k^{-1} \log k$.*

Combining Theorem 3 with Proposition 2 immediately gives a lower bound for the time-uniform estimation problem.

Corollary 4 *Let $\mathcal{P}$ be a family of distributions and $\theta : \mathcal{P} \to \mathbb{R}$ be a parameter function. Suppose there exists a sub-family $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ with parameter separation $\{\delta_k\}$ and distribution closeness $\{\Delta_k\}$. If there exists an $(\alpha, t(\cdot))$-estimator sequence, then for infinitely many $k \in \mathbb{N}$,*

$$t\big((1 - \alpha)\Delta_k^{-1} \log k\big) \geq \delta_k / 2.$$

Proof of Theorem 3 We first argue that we may take the tests to be deterministic without loss of generality. Consider the augmented problem setting of $\tilde{\mathcal{X}} = \mathcal{X} \times [0, 1]$ and $\tilde{P}_v = P_v \otimes \mathsf{Uni}([0, 1])$, so sample is supplemented with independent uniform randomness. Also define $\tilde{\mathbb{P}}_v$ as the probability space corresponding to i.i.d. draws $\tilde{X}_1, \tilde{X}_2, \ldots$ from $\tilde{P}_v$.

We may express any randomized test $\hat{V}_k(X_{1:n_k})$ as a deterministic function f_k of $X_{1:n_k}$ and some independent noise U_k distributed according to $\mathsf{Uni}([0, 1])$. Then, we construct the deterministic test on $\tilde{\mathcal{X}}^{n_k}$ by $\tilde{V}_k((X_i, U_i)_{1:n_k}) = f_k(X_{1:n_k}, U_{n_k})$. Applying this for all $k \in \mathbb{N}$ yields a deterministic test-sequence $\tilde{V}$ on $\mathcal{X}$ that has the same distribution as $\hat{V}$, and so if $\hat{V}$ is an $(\alpha, t(\cdot))$-test, then so is $\tilde{V}$. Finally, observe that $D_{\text{kl}}(\tilde{P}_v \| \tilde{P}_{v'}) = D_{\text{kl}}(P_v \| P_{v'})$ by chain rule of the KL-divergence, so the family $\{\tilde{P}_v\}_{v \in \mathcal{V}}$ also has distribution closeness $\{\Delta_k\}$. Therefore, proving the lower bound for deterministic tests on $\{\tilde{P}_v\}_{v \in \mathcal{V}}$ will imply the same lower bound for randomized tests on $\{P_v\}_{v \in \mathcal{V}}$.

We now define

$$\mathsf{CondErr}(\tilde{V}, v, k) := \tilde{\mathbb{P}}_v(\tilde{V}_k \neq v_k \mid \tilde{V}_{1:k-1} = v_{1:k-1}),$$

which is the probability an $(\alpha, \{n_k\})$-test $\tilde{V}$ makes an error at test k conditioned on the event that all previous tests were correct. The following lemma gives a lower bound on $\mathsf{CondErr}(\tilde{V}, v, k)$ over uniform distributions:

Lemma 5 *Let $\tilde{V}$ be a deterministic $(\alpha, \{n_k\})$-test sequence for the testing problem $\{\tilde{P}_v\}$. For all $k \in \mathbb{N}$ and all $b \in \{0,1\}^{k-1}$,*

$$\int \mathsf{CondErr}(\tilde{V}, v, k) d\mathsf{Uni}_b^{\mathcal{V}}(v) \geq \frac{1}{4} \exp\left(-\frac{\Delta_k n_k}{1-\alpha}\right).$$

Proof Because for each $j \in \mathbb{N}$, $\tilde{V}_j$ is a deterministic function of $\tilde{X}_{1:n_j}$, for each $b' \in \{0,1\}^j$ we may define a set $A_{b'} \subset \tilde{\mathcal{X}}^{n_j}$ such that $\{\tilde{V}_{1:j} = b'\} = \{\tilde{X}_{1:n_j} \in A_{b'}\}$. Let $\mathcal{V}_b = \{v \in \mathcal{V} \mid v_{1:k-1} = b\}$ and note that $\mathcal{V}_b$ is the support of $\mathsf{Uni}_b^{\mathcal{V}}$. For all $v \in \mathcal{V}_b$, we thus have the following equality of events in $\tilde{\mathbb{P}}_v$:

$$\{\tilde{V}_{1:k-1} = v_{1:k-1}\} = \{\tilde{V}_{1:k-1} = b\} = \{\tilde{X}_{1:n_{k-1}} \in A_b\} = \{\tilde{X}_{1:n_k} \in A_b \times \tilde{\mathcal{X}}^{n_k - n_{k-1}}\}. \quad (2)$$

Define $S := A_b \times \tilde{\mathcal{X}}^{n_k - n_{k-1}} \subset \tilde{\mathcal{X}}^{n_1}$ and notice that $A_{b\oplus(0)}, A_{b\oplus(1)}$ are relative complements in S. Then because $\mathsf{Uni}_b^{\mathcal{V}} = \frac{1}{2}\mathsf{Uni}_{b\oplus(0)}^{\mathcal{V}} + \frac{1}{2}\mathsf{Uni}_{b\oplus(1)}^{\mathcal{V}}$, we get

$$\begin{aligned}
\int \mathsf{CondErr}(\tilde{V}, v, k) d\mathsf{Uni}_b^{\mathcal{V}}(v) &= \int \tilde{\mathbb{P}}_v(\tilde{V}_k \neq v_k \mid \tilde{V}_{k-1} = v_{k-1}) d\mathsf{Uni}_b^{\mathcal{V}}(v) \\
&= \frac{1}{2} \int \tilde{\mathbb{P}}_v(\tilde{V}_k = 1 \mid \tilde{V}_{k-1} = v_{k-1}) d\mathsf{Uni}_{b\oplus(0)}^{\mathcal{V}}(v) \\
&\quad + \frac{1}{2} \int \tilde{\mathbb{P}}_v(\tilde{V}_k = 0 \mid \tilde{V}_{k-1} = v_{k-1}) d\mathsf{Uni}_{b\oplus(1)}^{\mathcal{V}}(v) \\
&\stackrel{(\star)}{=} \frac{1}{2} \int \tilde{\mathbb{P}}_v(\tilde{X}_{1:n_k} \in A_{b\oplus(1)} \mid \tilde{X}_{1:n_k} \in S) d\mathsf{Uni}_{b\oplus(0)}^{\mathcal{V}}(v) \\
&\quad + \frac{1}{2} \int \tilde{\mathbb{P}}_v(\tilde{X}_{1:n_k} \in A_{b\oplus(0)} \mid \tilde{X}_{1:n_k} \in S) d\mathsf{Uni}_{b\oplus(1)}^{\mathcal{V}}(v) \\
&= \frac{1}{2} \int \tilde{P}_v^{n_k}(A_{b\oplus(1)} \mid S) d\mathsf{Uni}_{b\oplus(0)}^{\mathcal{V}}(v) \\
&\quad + \frac{1}{2} \int \tilde{P}_v^{n_k}(A_{b\oplus(0)} \mid S) d\mathsf{Uni}_{b\oplus(1)}^{\mathcal{V}}(v),
\end{aligned} \quad (3)$$

where equality $(\star)$ follows from the identity (2).

To proceed, we use Le Cam's method and the Bretagnolle-Huber inequality below (see, e.g. Yu (1997, Lemma 1) and Tsybakov (2009, Lemma 2.6) for proofs):

Lemma 6 (Le Cam's method) *Let P, Q be distributions over a domain $\mathcal{X}$. For any $S \subset \mathcal{X}$,*

$$P(S) + Q(S^c) \geq 1 - \sup_S (Q(S) - P(S)) = 1 - \|P - Q\|_{\mathrm{TV}}.$$

Lemma 7 (Bretagnolle-Huber) *Let P, Q be distributions over a domain $\mathcal{X}$. Then*

$$\|P - Q\|_{\mathrm{TV}} \leq 1 - \frac{1}{2} \exp(-D_{\mathrm{kl}}(P \| Q)).$$

Applying Lemma 6 to Eq. (3) implies the lower bound

$$\begin{aligned}
&\int \mathsf{CondErr}(\tilde{V}, v, k) d\mathsf{Uni}_b^{\mathcal{V}}(v) \\
&\quad \geq \frac{1}{2}\left(1 - \left\| \int \tilde{P}_v^{n_k}(\cdot \mid S) d\mathsf{Uni}_{b\oplus(0)}^{\mathcal{V}}(v) - \int \tilde{P}_v^{n_k}(\cdot \mid S) d\mathsf{Uni}_{b\oplus(1)}^{\mathcal{V}}(v) \right\|_{\mathrm{TV}}\right).
\end{aligned}$$

Then applying Lemma 7 and using the convexity of the KL-divergence, we obtain the lower bound

$$\int \mathsf{CondErr}(\tilde{V}, v, k) d\mathsf{Uni}_b^{\mathcal{V}}(v) \geq \frac{1}{4} \exp\left(- \sup_{v,v' \in \mathcal{V}_b} D_{\mathrm{kl}}\left(\tilde{P}_v^{n_k}(\cdot \mid S) \| \tilde{P}_{v'}^{n_k}(\cdot \mid S) \right) \right). \tag{4}$$

To conclude the proof, we use the following lemma to bound the KL-divergence between conditional distributions:

Lemma 8 (KL-divergence of conditional distributions) *Let P, Q be two distributions over a domain $\mathcal{X}$. Let S be a measurable subset of $\mathcal{X}$, and let P_S, Q_S denote the conditional distributions of P given S and Q given S, respectively. Then $D_{\mathrm{kl}}(P_S \| Q_S) \leq \frac{1}{P(S)} D_{\mathrm{kl}}(P \| Q)$.*

Proof By the chain rule for KL-divergence,

$$\begin{aligned} D_{\mathrm{kl}}(P \| Q) &= D_{\mathrm{kl}}(\mathsf{Bern}(P(S)) \| \mathsf{Bern}(Q(S))) + P(S) D_{\mathrm{kl}}(P_S \| Q_S) + P(S^c) D_{\mathrm{kl}}(P_{S^c} \| Q_{S^c}) \\ &\geq P(S) D_{\mathrm{kl}}(P_S \| Q_S). \end{aligned}$$

Dividing by $P(S)$ proves the claim. ∎

By applying Lemma 8 and the assumption $D_{\mathrm{kl}}\left(\tilde{P}_v \| \tilde{P}_{v'}\right) \leq \Delta_k$, we have

$$\sup_{v,v' \in \mathcal{V}_b} D_{\mathrm{kl}}\left(\tilde{P}_v^{n_k}(\cdot \mid S) \| \tilde{P}_{v'}^{n_k}(\cdot \mid S) \right) \leq \sup_{v,v' \in \mathcal{V}_b} \frac{D_{\mathrm{kl}}\left(\tilde{P}_v^{n_k} \| \tilde{P}_{v'}^{n_k} \right)}{\tilde{P}_v^{n_k}(S)} \leq \frac{\Delta_k n_k}{1-\alpha}.$$

Substituting this into Eq. (4) proves Lemma 5. ∎

Iteratively applying Lemma 5, we obtain a lower bound on the cumulative conditional errors:

Lemma 9 *Let $\tilde{V}$ be a deterministic $(\alpha, \{n_k\})$-test sequence for the testing problem $\{\tilde{P}_v\}$. For all $\ell \in \mathbb{N}$, there exists a distribution $b \in \{0,1\}^{\ell-1}$ such that*

$$\int \sum_{k=1}^{\ell} \mathsf{CondErr}(\tilde{V}, v, k) d\mathsf{Uni}_b^{\mathcal{V}}(v) \geq \frac{1}{4} \sum_{k=1}^{\ell} \exp\left(-\frac{\Delta_k n_k}{1-\alpha} \right).$$

Proof We proceed inductively.

Base Case ($\ell = 1$). This case follows immediately by applying Lemma 5 with $k = 1$.

Inductive step. Let $\ell \geq 2$. Assume the claim holds for $\ell - 1$ and let $b' \in \{0,1\}^{\ell-2}$ such that

$$\int \sum_{k=1}^{\ell-1} \mathsf{CondErr}(\tilde{V}, v, k) d\mathsf{Uni}_{b'}^{\mathcal{V}}(v) \geq \frac{1}{4} \sum_{k=1}^{\ell-1} \exp\left(-\frac{\Delta_k n_k}{1-\alpha} \right). \tag{5}$$

It will suffice to find $b \in \{0,1\}^{\ell-1}$ that satisfies both

$$\int \sum_{k=1}^{\ell-1} \mathsf{CondErr}(\tilde{V}, v, k) d\mathsf{Uni}_b^{\mathcal{V}}(v) \geq \frac{1}{4} \sum_{k=1}^{\ell-1} \exp\left(-\frac{\Delta_k n_k}{1-\alpha} \right) \tag{6}$$

and

$$\int \mathsf{CondErr}(\tilde{V}, v, k) d\mathsf{Uni}_b^{\mathcal{V}}(v) \geq \frac{1}{4} \exp\left(-\frac{\Delta_\ell n_\ell}{1-\alpha}\right). \tag{7}$$

Because $\mathsf{Uni}_{b'}^{\mathcal{V}} = \frac{1}{2}\mathsf{Uni}_{b'\oplus(0)}^{\mathcal{V}} + \frac{1}{2}\mathsf{Uni}_{b'\oplus(1)}^{\mathcal{V}}$, the inductive hypothesis (Eq. (5)) implies

$$\max_{i\in\{0,1\}} \int \sum_{k=1}^{\ell-1} \mathsf{CondErr}(\tilde{V}, v, k) d\mathsf{Uni}_{b'\oplus(i)}^{\mathcal{V}}(v) \geq \frac{1}{4} \sum_{k=1}^{\ell-1} \exp\left(-\frac{\Delta_k n_k}{1-\alpha}\right). \tag{8}$$

Let $i \in \{0, 1\}$ be any bit obtaining the maximum in the left-hand side of Eq. (8). By taking $b = b' \oplus (i)$, we ensure that b satisfies Eq. (6). Applying Lemma 5 with $k = \ell$, we also have that b satisfies Eq. (7). This completes the induction and proves the claim. ■

Conversely, the time-uniform conditions on $\tilde{V}$ yield an upper-bound on the sum of $\mathsf{CondErr}$.

Lemma 10 *Let $\tilde{V}$ be an $(\alpha, \{n_k\})$-test sequence. For all $v \in \mathcal{V}$,*

$$\sum_{k=1}^{\infty} \mathsf{CondErr}(\tilde{V}, v, k) \leq \frac{\alpha}{1-\alpha}. \tag{9}$$

Proof By definition of $\hat{V}$, $\mathbb{P}_v(\hat{V}_{1:k} = v_{1:k}) \geq 1 - \alpha$ for all $v \in \mathcal{V}$ and all $k \in \mathbb{N}$. Furthermore, for $v \in \mathcal{V}$,

$$\begin{aligned} \mathbb{P}_v(\hat{V}_{1:k} = v_{1:k}) &= \mathbb{P}_v(\hat{V}_{1:k-1} = v_{1:k-1})\mathbb{P}_v(\hat{V}_k = v_k \mid \hat{V}_{1:k-1} = v_{1:k-1}) \\ &= \mathbb{P}_v(\hat{V}_{1:k-1} = v_{1:k-1}) - \mathbb{P}_v(\hat{V}_{1:k-1} = v_{1:k-1})\mathbb{P}_v(\hat{V}_k \neq v_k \mid \hat{V}_{1:k-1} = v_{1:k-1}) \\ &\leq \mathbb{P}_v(\hat{V}_{1:k-1} = v_{1:k-1}) - (1-\alpha)\mathsf{CondErr}(\hat{V}, v, k). \end{aligned}$$

Iterating the above inequality thus yields $\mathbb{P}_v(\hat{V}_{1:k} = v_{1:k}) \leq 1-(1-\alpha)\sum_{j=1}^{k} \mathsf{CondErr}(\hat{V}, v, j)$. Because $\mathbb{P}_v(\hat{V}_{1:k} = v_{1:k}) \geq 1 - \alpha$, we have that $\sum_{j=1}^{k} \mathsf{CondErr}(\hat{V}, v, j) \leq \frac{\alpha}{1-\alpha}$. Taking $k \to \infty$ proves the result. ■

Combining Lemmas 9 and 10, we obtain

$$\frac{1}{4} \sum_{k=1}^{\infty} \exp\left(-\frac{\Delta_k n_k}{1-\alpha}\right) \leq \frac{\alpha}{1-\alpha}.$$

Theorem 3 therefore follows because the summability of the sequence implies that $\exp(-\frac{\Delta_k n_k}{1-\alpha}) < \frac{1}{k}$, or $\log k < \frac{\Delta_k n_k}{1-\alpha}$, infinitely often. ■

3. Applying the lower bound

The abstract theorem and Corollary 4 combine to yield a fairly straightforward recipe for proving lower bounds for time-uniform estimation problems. For a parametric model $\{P_\theta\}_{\theta\in\Theta}$, where $\Theta \subset \mathbb{R}$, the basic idea proceeds in three steps:

1. Demonstrate the (typical) scaling that $D_{\text{kl}}\left(P_\theta \| P_{\theta'}\right) \le M(\theta - \theta')^2$ for some $M < \infty$
2. Develop a particular well-separated collection of parameters $\{\theta(v)\} \subset \Theta$ by, for $v \in \mathcal{V}$, writing a ternary expansion of $\theta(v) = 2\sum_{i=1}^\infty v_i 3^{-i}$, which guarantees that if v, v' first disagree at position k, then $D_{\text{kl}}\left(P_\theta \| P_{\theta'}\right) \lesssim (3^{-k})^2$, while $|\theta(v) - \theta(v')| \gtrsim 3^{-k}$.
3. Apply Corollary 4.

This approach is familiar from classical "two-point" lower bounds that rely on reducing estimation to testing (e.g. Yu, 1997; Duchi, 2023). The quadratic scaling of the KL-divergence is indeed common: in (essentially) any setting in which the Fisher Information I_θ of the model $\{P_\theta\}$ exists, classical information geometry considerations show that $D_{\text{kl}}\left(P_\theta \| P_{\theta'}\right) = \frac{1}{2} I_\theta(\theta' - \theta)^2 + o(|\theta' - \theta|^2)$, (see, e.g., Cover and Thomas, 2006, Exercise 11.7). Our lower bounds require the quadratic bound only locally, so we expect them to apply in most estimation problems.

Let us formalize these steps. Let Θ be a convex subset of $\mathbb{R}$ and $\theta_0 \in \Theta$, and let $r > 0$ be (a radius) such that $[\theta_0, \theta_0 + r] \subset \Theta$. To apply the lower bound, we wish to choose a parameter mapping $\theta : \mathcal{V} \to \Theta$ to maximize the parameter separation $\{\delta_k\}$ while minimizing the distribution closeness $\{\Delta_k\}$. In our parametric setting, given a mapping $\theta : \mathcal{V} \to \Theta$, we can take

$$\delta_k = \inf_{v_{1:k} \neq v'_{1:k}} |\theta(v) - \theta(v')| \qquad \text{and} \qquad \Delta_k = \sup_{v_{1:k} = v'_{1:k}} M(\theta(v) - \theta(v'))^2.$$

To that end, we take a shifted Cantor set mapping

$$\theta(v) = \theta_0 + r \cdot 2\sum_{i=1}^\infty v_i 3^{-i},$$

where by inspection $\theta(v) \in [\theta_0, \theta_0 + r] \subset \Theta$. Upper and lower bounds on $|\theta(v) - \theta(v')|$ for $v, v' \in \mathcal{V}$ are then immediate. Indeed, letting $v, v' \in \mathcal{V}$ differ for the first time at index k,

$$\begin{aligned} |\theta(v) - \theta(v')| = 2r\left|\sum_{i=k}^\infty (v_i - v'_i)3^{-i}\right| &= 2r \cdot 3^{-k}\left|\sum_{i=0}^\infty (v_{i+k} - v'_{i+k})3^{-i}\right| \\ &\le 2r \cdot 3^{-k}\sum_{i=0}^\infty 3^{-i} = 3^{1-k} r. \end{aligned}$$

So long as the KL-bound $D_{\text{kl}}\left(P_{\theta_0} \| P_{\theta_0 + \delta}\right) \le M\delta^2$ holds for $\delta \le \frac{2}{3}r$, whenever $v_{1:k} = v'_{1:k}$ we have $|\theta(v) - \theta(v')| \le 3^{-k} r$, and we may thus choose distributional closeness

$$\Delta_k = M 9^{-k} r^2.$$

On the other hand, when v, v' differ for the first time in index k, then again by the triangle inequality,

$$|\theta(v) - \theta(v')| \ge 2r \cdot 3^{-k}\left(|v_k - v'_k| - \left|\sum_{i=1}^\infty (v_{i+k} - v'_{i+k})3^{-i}\right|\right) \ge 3^{-k} r$$

because $v_k \neq v'_k$. Therefore, if $v_{1:k} \neq v'_{1:k}$, then $|\theta(v) - \theta(v')| \geq 3^{-k} r$ and so we may take the parameter separation $\delta_k = 3^{-k} r$. Applying Corollary 4 thus yields the following proposition.

Proposition 11 *Let $\Theta \subset \mathbb{R}$ be convex and the family $\{P_\theta\}_{\theta \in \Theta}$ be such that for some $r > 0$ and $\theta_0 \in \Theta$ with $[\theta_0, \theta_0 + r] \in \Theta$, $D_{\rm kl}\left(P_{\theta_0} \| P_{\theta_0 + \delta}\right) \leq M\delta^2$ for $\delta \leq r$. Then any $(\alpha, t(\cdot))$-estimator sequence satisfies*

$$t(n) \geq \sqrt{\frac{1-\alpha}{8M}} \cdot \sqrt{\frac{\log \log n}{n}} \text{ infinitely often.}$$

Proof By Corollary 4, if there exists an $(\alpha, t(\cdot))$-estimator, then for $N_k = (1-\alpha)\Delta_k^{-1} \log k$, we use the identification of the distributional closeness $\Delta_k = M\delta_k^2$ with the squared parameter separation to obtain

$$t(N_k) \geq \frac{1}{2}\delta_k = \frac{1}{2}\sqrt{\frac{\Delta_k}{M}} = \frac{1}{2}\sqrt{\frac{(1-\alpha)\log k}{MN_k}}$$

for infinitely many $k \in \mathbb{N}$. Because $N_k = \frac{1-\alpha}{Mr^2} 9^k \log k$, for any sufficiently large $k \in \mathbb{N}$, we have both $N_k \leq 10^k$ and $\log \log N_k \leq \log k + \log \log 10 \leq 2 \log k$, or $\log k \geq \frac{1}{2} \log \log N_k$. Taking n from among the indices $\{N_k\}_{k \in \mathbb{N}}$,

$$t(n) \geq \frac{1}{2}\sqrt{\frac{(1-\alpha)\log \log n}{2Mn}} \text{ infinitely often}$$

as desired. ■

While Proposition 11 demonstrates the $\log \log n$ penalty of time-uniform estimation, a more standard technique demonstrates the correct dependence on the confidence α.

Proposition 12 *Let the conditions of Proposition 11 hold. Then for $\alpha < \frac{1}{4}$, any $(\alpha, t(\cdot))$-estimator sequence satisfies*

$$t(n) \geq \frac{1}{2}\sqrt{\frac{\log(1/4\alpha)}{Mn}} \text{ for all } n \in \mathbb{N}.$$

As the proof is a more-or-less standard two point lower bound, we defer it to Section 5.

3.1. Location models

We first demonstrate how the recipe above applies to obtain lower bounds on (time-uniformly) estimating the parameter of a location model without particular reliance on the structure of the family beyond mild integrability conditions. Let f be a density on $\mathbb{R}$, and consider the family $\{P_\theta\}_{\theta \in \mathbb{R}}$ with densities of the form $x \mapsto f(x - \theta)$. Then under appropriate integrability conditions on the derivatives of the density, the Fisher Information $I_\theta = \int \frac{f'(x-\theta)^2}{f(x-\theta)} dx$ is constant (Lehmann and Casella, 1998), and we also have the following quadratic expansion (e.g. Kullback, 1968, Section 2.6).

Lemma 13 *For sufficiently regular location families, $D_{\rm kl}\left(P_\theta \| P_{\theta'}\right) = \frac{1}{2} I_\theta (\theta - \theta')^2 + o((\theta - \theta')^2)$.*

Example 1 (Location families from regular densities) *Let f be a suitably regular density with Fisher Information $I_f = \int \frac{f'(x)^2}{f(x)} dx$. We may take $M = \frac{1}{2} I_f$ in Proposition 11, and so for any $(\alpha, t(\cdot))$-estimator, $t(n) \geq \frac{1}{2}\sqrt{\frac{(1-\alpha)\log\log n}{I_f \cdot n}}$ infinitely often.*

We can be more explicit when it is easy to compute the KL-divergence.

Example 2 (Gaussian location model) *Consider $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$. Then for $P_\theta = \mathsf{N}(\mu, \sigma^2)$, we have $D_{\mathrm{kl}}(P_\theta \| P_{\theta'}) = \frac{1}{2\sigma^2}(\theta - \theta')^2$, and so taking $M = \frac{1}{2\sigma^2}$ Proposition 11 yields that for any $(\alpha, t(\cdot))$-estimator, $t(n) \geq \frac{\sigma}{2}\sqrt{\frac{(1-\alpha)\log\log n}{n}}$ infinitely often.*

Example 3 (Cauchy location model) *Consider the density $p_\theta(x) = 1/(\pi\sigma[1 + (\frac{x-\theta}{\sigma})^2])$. Then Chyzak and Nielsen (2019, Theorem 1) show that*

$$D_{\mathrm{kl}}(P_\theta \| P_{\theta'}) = \log\left(1 + \frac{(\mu - \mu')^2}{4\sigma^2}\right) \leq \frac{(\mu - \mu')^2}{4\sigma^2},$$

and so in Proposition 11 we take $M = \frac{1}{4\sigma^2}$ to obtain that for any $(\alpha, t(\cdot))$-estimator, $t(n) \geq \sigma\sqrt{\frac{(1-\alpha)\log\log n}{2n}}$ infinitely often.

3.2. Generalized linear models

Exponential family models provide another clear demonstration of the lower bound technique. Consider two domains $\mathcal{X}$ and $\mathcal{Y}$. Let Q be a distribution over $\mathcal{X}$, μ be a reference measure over $\mathcal{Y}$, and $T : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ be a statistic of y given x. We define the conditional exponential family model, or one-parameter generalized linear model,

$$p_\theta(y \mid x) = \exp\left(\theta T(y, x) - A(\theta \mid x)\right) \;\; \text{where} \;\; A(\theta \mid x) = \log \int_{\mathcal{Y}} \exp(\theta T(y \mid x)) d\mu(y) \tag{10}$$

and p_θ is a density with respect to μ. Then we define P_θ over $\mathcal{X} \times \mathcal{Y}$ such $P_\theta(X \in \cdot) = Q$ and $P_\theta(Y \in \cdot \mid X = x)$ has density $p_\theta(y|x)$ with respect to μ. The log-partition function $A(\theta \mid x)$ is analytic on its open convex domain (Brown, 1986), and

$$\begin{aligned} D_{\mathrm{kl}}(P_\theta \| P_{\theta'}) &\overset{(i)}{=} \mathbb{E}_{X \sim Q}\left[D_{\mathrm{kl}}(P_\theta(Y \in \cdot \mid X) \| P_{\theta'}(Y \in \cdot \mid X))\right] \\ &\overset{(ii)}{=} \mathbb{E}_{X \sim Q}\left[A(\theta' \mid X) - A(\theta \mid X) - A'(\theta \mid X)(\theta' - \theta)\right], \end{aligned}$$

where equality (i) follows from the chain rule and equality (ii) is an immediate calculation for exponential family models. Recognizing the Bregman divergence between θ and θ' according to the (convex) log-partition function $A(\cdot \mid x)$ in this last display, for all $\theta, \theta + \delta \in \operatorname{dom} A(\cdot \mid x)$ we obtain

$$D_{\mathrm{kl}}(P_\theta(Y \in \cdot \mid x) \| P_{\theta+\delta}(Y \in \cdot \mid x)) = \frac{1}{2} A''(\theta + t \mid x)\delta^2$$

for some $t \in [0, \delta]$. Because $A''(\theta \mid x) = \mathrm{Var}_\theta(Y \mid X = x)$ in the conditional model (10), Proposition 11 implies the following corollary.

Corollary 14 *Assume the expected variance $\mathbb{E}[\mathrm{Var}_\theta(Y \mid X)]$ is continuous and finite in a neighborhood of some θ_0 in the one-parameter GLM (10). Let $M = \mathbb{E}[\mathrm{Var}_{\theta_0}(Y \mid X)]$. Then any $(\alpha, t(\cdot))$-estimator satisfies*

$$t(n) \ge \frac{1}{4}\sqrt{\frac{(1-\alpha)\log\log n}{Mn}} \text{ infinitely often.}$$

Proof Let $M_\theta = \mathbb{E}[\mathrm{Var}_\theta(Y \mid X)]$. Then by continuity and Taylor's theorem, $D_{\mathrm{kl}}\left(P_{\theta_0} \| P_{\theta_0+\delta}\right) = \frac{1}{2}\mathbb{E}[A''(\theta_0 + t\delta \mid X)]\delta^2$ for some $t \in [0, 1]$. Using continuity, there exists a neighborhood of θ_0 such that $M_\theta \le 2M$ for all θ in the neighborhood. ■

We can make these calculations explicit for logistic regression:

Example 4 (Logistic regression) *Take $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{-1, 1\}$, μ to be the counting measure on $\mathcal{Y}$ and $T(y, x) = yx$. Then the log-partition $A(\theta \mid x) = \log(e^{-x\theta} + e^{x\theta})$ satisfies*

$$A'(\theta \mid x) = \frac{x(e^{x\theta} - e^{-x\theta})}{e^{-x\theta} + e^{x\theta}} \text{ and } A''(\theta \mid x) = x^2 - \frac{x^2(e^{x\theta} - e^{-x\theta})^2}{(e^{-x\theta} + e^{x\theta})^2} \le x^2.$$

Taking $M = \mathbb{E}[X^2]$, any $(\alpha, t(\cdot))$-estimator for θ satisfies $t(n) \ge \frac{1}{4}\sqrt{\frac{(1-\alpha)\log\log n}{\mathbb{E}[X^2]n}}$ infinitely often.

3.3. Models with nuisance parameters

As we note above, the techniques we develop apply to any family with an appropriate (Fisher) information measure, meaning that they transparently apply to problems with nuisance parameters and in which we estimate some functional of the distribution. To treat this in a fairly general case, we consider semiparametric models, following van der Vaart (1998, Chapter 25). For a distribution $P \in \mathcal{P}$, we define the tangent space $\dot{\mathcal{P}}_P$ to be the set of functions $g \in L^2(P)$ such that for some neighborhood $O \subset \mathbb{R}$ of 0, there exists a one-dimensional parametric submodel $\{P_h\}_{h \in O} \subset \mathcal{P}$ satisfying $P_0 = P$ and

$$\int (\sqrt{dP_h} - \sqrt{dP} - \tfrac{1}{2}hg\sqrt{dP})^2 = o(h^2).$$

We say one-dimensional parametric submodels satisfying this are *quadratic mean differentiable (QMD) at P with score g* (van der Vaart, 1998, Chapter 7.2); any such model admits the quadratic approximation $D_{\mathrm{kl}}\left(P \| P_h\right) = \frac{h^2}{2}\int g^2 dP + o(h^2)$, and g *a fortiori* has mean 0 under P. We make the standard assumption that the parameter θ is differentiable at P with respect to $\dot{\mathcal{P}}_P$, meaning that it has an *influence function* $\dot{\theta}$ whose coordinate functions belong to the closure of the linear span of $\dot{\mathcal{P}}_P$ such that for all $\{P_h\} \subset \mathcal{P}$ that are QMD at P with score g, we have

$$\theta(P_h) - \theta(P) = h\int \dot{\theta}(x)g(x)dP(x) + o(h) \text{ as } h \to 0. \tag{11}$$

While these conditions are abstract, they arise naturally in parametric models, and they also appear fairly generically in empirical risk minimization problems (M-estimation), highlighting the power of the lower bounds Theorem 3 and Proposition 11 imply. We present two examples before providing the general bound.

Example 5 (Parametric models) *Let $\{P_\beta\}_{\beta \in B}$ be a parametric model family for some open $B \subset \mathbb{R}^d$, where we wish to estimate the parameter $\theta(P_\beta) = \varphi(\beta)$ for a differentiable $\varphi : \mathbb{R}^d \to \mathbb{R}$. The simplest example consists of the coordinate functions $\varphi(\beta) = \beta_j$, the jth coordinate of β. Let $\dot{\ell}_\beta = \nabla_\beta \log p_\beta$ be the gradient of the log likelihood and $I_\beta = P_\beta \dot{\ell}_\beta \dot{\ell}_\beta^T$ be the Fisher information matrix, where we assume that I_β is positive definite. Van der Vaart (1998, Example 25.15) shows that $\dot{\theta}(x) = \nabla\varphi(\beta)^T I_\beta^{-1} \dot{\ell}_\beta(x)$ is the influence function.*

Example 6 (Risk minimizers and M-estimators) *Let $\mathcal{P}$ be the collection of distributions on a domain $\mathcal{X}$ and consider a loss function $\ell : \Theta \times \mathcal{X} \to \mathbb{R}_+$ where $\Theta \subset \mathbb{R}^d$ and ℓ is convex in θ for all $x \in \mathcal{X}$. Define the risk minimizer $\theta : \mathcal{P} \to \Theta$ by*

$$\theta(P) = \operatorname*{argmin}_{\theta \in \Theta} \left\{ L(\theta) := \mathbb{E}_P[\ell(\theta, X)] = \int \ell(\theta, x) dP(x) \right\}.$$

Assuming sufficient smoothness of L and ℓ, the associated influence function is $\dot{\theta}(x) = -\nabla^2 L(\theta(P))^{-1} \nabla_\theta \ell(\theta(P), x)$ (e.g. Duchi and Ruan, 2021, Prop. 1). Following the same approach as Example 5, an individual coordinate $\theta_j(P)$ has influence

$$\dot{\theta}_j(x) = -e_j^T \nabla^2 L(\theta(P))^{-1} \nabla_\theta \ell(\theta(P), x).$$

Regardless, we have the following lower bound whenever an influence function exists.

Proposition 15 *Let $\theta : \mathcal{P} \to \mathbb{R}$ have influence $\dot{\theta}$ at $\dot{\mathcal{P}}_P$. Then any $(\alpha, t(\cdot))$-estimator of θ satisfies*

$$t(n) \geq \frac{1}{4} \sqrt{(1-\alpha) \int \dot{\theta}(x)^2 dP(x)} \cdot \sqrt{\frac{\log \log n}{n}} \quad \textit{infinitely often.}$$

Proof Because minimax estimation over all of $\mathcal{P}$ is at least as hard as minimax estimation over any one-dimensional subfamily, we will demonstrate minimax lower bounds over a QMD subfamily with score $g : \mathcal{X} \to \mathbb{R}$ and then choose an appropriate g.

Let $\{P_h\}_{h \in H}$ be a QMD subfamily with score g, where H is a neighborhood of 0, and we assume that $|\int \dot{\theta}(x) g(x) dP(x)| > 0$ in the first-order expansion (11) (as we will choose g later, this is no loss of generality). We reduce estimation of θ over $\{P_h\}$ to estimation of $h \in \mathbb{R}$ itself. For any $\delta > 0$, by restricting H to a small enough neighborhood of 0, the derivative (11) implies that for any $h, h' \in H$ we have

$$\theta(P_h) - \theta(P_{h'}) = \theta(P_h) - \theta(P) + \theta(P) - \theta(P_{h'}) = (h - h') \int \dot{\theta}(x) g(x) dP(x) \pm \delta.$$

Now, given an estimator $\hat{\theta}_n$, define $\hat{h}_n = \operatorname{argmin}_{h' \in H} |\hat{\theta}_n - \theta(P_{h'})|$ (or a value arbitrarily close to minimizing $|\hat{\theta}_n - \theta(P_{h'})|$), so that

$$|\theta(P_{\hat{h}_n}) - \theta(P_h)| \leq |\theta(P_{\hat{h}_n}) - \hat{\theta}_n| + |\hat{\theta}_n - \theta(P_h)| \leq 2|\hat{\theta}_n - \theta(P_h)|.$$

Rearranging, $|\hat{h}_n - h| \leq \frac{2}{|\int \dot{\theta}(x) g(x) dP(x)| - \delta} |\hat{\theta}_n - \theta(P_h)|$, and we have shown that any $(\alpha, t(\cdot))$-estimator for θ on $\{P_h\}$ yields an $(\alpha, \frac{2}{|\int \dot{\theta}(x) g(x) dP(x)| - \delta} t(\cdot))$-estimator $\hat{h}_n$ for h on $\{P_h\}$.

Because $D_{\rm kl}\left(P \| P_h\right) = \frac{h^2}{2}\int g^2 dP + o(h^2)$, Proposition 11 implies that for any $M > \frac{1}{2}\int g^2 dP$, any $(\alpha, \tilde{t}(\cdot))$-estimator for h satisfies

$$\tilde{t}(n) \ge \sqrt{\frac{1-\alpha}{8M}} \cdot \sqrt{\frac{\log\log n}{n}} \text{ infinitely often.}$$

Putting these two observations together and taking $\delta \searrow 0$ and $M \searrow \frac{1}{2}\int g^2 dP$, we conclude that if $\hat{\theta}_n$ is an $(\alpha, t(\cdot))$-estimator for θ on $\{P_h\}$, then

$$t(n) \ge \frac{|\int \dot{\theta}(x) g(x) dP(x)|}{2} \cdot \sqrt{\frac{1-\alpha}{4\int g^2 dP}} \cdot \sqrt{\frac{\log\log n}{n}} \text{ infinitely often.}$$

This lower bound over $\{P_h\}$ trivially provides a minimax lower bound for $\mathcal{P}$, so choosing $g = \dot{\theta}$ yields the proposition. ■

Returning to Examples 5 and 6, Proposition 15 provides immediate lower bounds. In Example 5, we have influence $\dot{\theta} = \nabla\varphi(\beta)^T I_\beta^{-1}\dot{\ell}_\beta$, yielding

$$\mathbb{E}_{P_\beta}[\dot{\theta}(X)^2] = \nabla\varphi(\beta)^T \mathbb{E}_{P_\beta}\left[I_\beta^{-1}\dot{\ell}_\beta\dot{\ell}_\beta^T I_\beta^{-1}\right]\nabla\varphi(\beta) = \nabla\varphi(\beta)^T I_\beta^{-1}\nabla\varphi(\beta),$$

where I_β is the Fisher information for the model P_β. Applying Proposition 15 and taking the supremum over $\beta \in B$, we obtain that for all $(\alpha, t(\cdot))$-estimators,

$$t(n) \ge \frac{1}{4}\sqrt{(1-\alpha)\sup_{\beta\in B}\nabla\varphi(\beta)^T I_\beta^{-1}\nabla\varphi(\beta)} \cdot \sqrt{\frac{\log\log n}{n}} \text{ infinitely often.}$$

Similarly, in Example 6, if we define the asymptotic efficient covariance

$$\Sigma_P = \nabla^2 L(\theta(P))^{-1}\text{Cov}(\nabla\ell(\theta(P), X))\nabla^2 L(\theta(P))^{-1},$$

any $(\alpha, t(\cdot))$-estimator of $\varphi(\theta(P))$ for a differentiable $\varphi : \mathbb{R}^d \to \mathbb{R}$ must satisfy

$$t(n) \ge \frac{1}{4}\sqrt{(1-\alpha)\nabla\varphi(\theta(P))^T\Sigma_P\nabla\varphi(\theta(P))} \cdot \sqrt{\frac{\log\log n}{n}} \text{ infinitely often.}$$

4. Attaining the lower bound

Via a suitable doubling strategy, estimators achieving high-probability convergence guarantees for fixed sample sizes essentially immediately provide time-uniform guarantees with an additional $\log\log n$ penalty, matching our lower bounds. Readers should not take this section as advocating particular estimators—substantial work (e.g. Howard et al., 2020) on time-uniform sequences develops stronger estimators—but as a relatively simple proof of concept (see also Kirichenko and Grunwald (2021, Thm. 3.1)). Given a sequence of estimators $\hat{\theta}_n$, consider the *time uniform* estimator-sequence

$$\hat{\theta}_n^{\rm tu} := \hat{\theta}_{2^{\lfloor \log_2 n\rfloor}} = \begin{cases} \hat{\theta}_n(X_{1:n}) & \text{if } n = 2^k, \text{ for some } k \in \mathbb{N} \\ \hat{\theta}_{n-1}^{\rm tu} & \text{otherwise,} \end{cases} \tag{12}$$

so that $\hat{\theta}^{\text{tu}}$ is $\hat{\theta}_n$ evaluated at the most recent power of two sample size. So long as $\hat{\theta}$ is sufficiently accurate, $\hat{\theta}_n^{\text{tu}}$ is a time-uniform estimator. In stating the proposition, we say that $\hat{\theta}$ has *high-probability deviation bound* F for $F : \mathbb{R}_+ \times \mathbb{N} \to \mathbb{R}_+$ if

$$\mathbb{P}\left(|\hat{\theta}_n(X_{1:n}) - \theta(P)| \le F(\log(1/\alpha), n)\right) \ge 1 - \alpha \text{ for all } \alpha \in [0,1],\ n \in \mathbb{N},$$

where F is increasing in its first argument and decreasing in its second. In typical cases, such as sub-Gaussian mean estimation, we expect $F(t,n) = C\sqrt{t/n}$ for some C.

Proposition 16 *Let $\hat{\theta}_n$ have high probability deviation bound F. Then $\{\hat{\theta}_n^{\text{tu}}\}$ is an $(\alpha, t(\cdot))$-estimator for*

$$t(n) = F\big(2\log\log_2 n + \log(1/\alpha) + 1/2, n/2\big).$$

Proof Let E denote the event that for all $k \in \mathbb{N}$, $|\hat{\theta}_{2^k} - \theta(P)| \le F(\log(\pi^2 k^2/6\alpha), 2^k)$. By a union bound, $\mathbb{P}(E^c) \le \frac{6}{\pi^2}\sum_{k=1}^{\infty}\frac{\alpha}{k^2} = \alpha$ by the Basel problem, so E occurs with probability at least $1-\alpha$. Now for $n \in \mathbb{N}$, let $k = k_n = \lfloor \log_2 n \rfloor$ be the unique natural number such that $2^k \le n < 2^{k+1}$. By definition, $\hat{\theta}_n = \hat{\theta}_{2^k}$ and so on E,

$$|\hat{\theta}_n^{\text{tu}} - \theta(P)| \le F\left(\log\frac{\pi^2 k^2}{6\alpha}, 2^k\right) \quad \text{for all } n.$$

Now use that $k \le \log_2 n$ implies $\log k^2 = 2\log k \le 2\log\log_2 n$, $\log\frac{\pi^2}{6} < \frac{1}{2}$, and that F is decreasing in its second argument and $2^k > n/2$. ■

As a simple example, consider estimating a sub-Gaussian mean, where we recall that X is σ^2-sub-Gaussian if $\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le \exp(\frac{\lambda^2\sigma^2}{2})$ for all $\lambda \in \mathbb{R}$.

Example 7 (Sub-Gaussian mean estimation) *Let $X_i \stackrel{\text{iid}}{\sim} P$ be σ^2-sub-Gaussian with mean $\theta = \theta(P)$. The sample mean $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies $\mathbb{P}(|\hat{\theta}_n - \theta| \ge t) \le 2\exp(-\frac{nt^2}{2\sigma^2})$, that is, $\hat{\theta}_n$ has high probability deviation bound $F(\log\frac{1}{\alpha}, n) = \sqrt{\frac{2\sigma^2\log(2/\alpha)}{n}}$. The time-uniform estimator (12) is thus an $(\alpha, t(\cdot))$-estimator for*

$$t(n) = O(1)\cdot\sigma\sqrt{\frac{\log\log n + \log(1/\alpha)}{n}}$$

by Proposition 16, matching the lower bound in Example 2 to within constant factors; Proposition 12 shows its sharpness generally.

5. Proof of Proposition 12

For $\theta \in \Theta$, let $\mathbb{P}_\theta$ represent the probability space corresponding to $X_1, X_2, \ldots \stackrel{\text{iid}}{\sim} P_\theta$. Considering testing between two points θ_0 and $\theta_\delta = \theta_0 + \delta \in \Theta$, where $\delta > 0$ is a quantity we choose later. For any test T_n based on a sample of size n, Le Cam's lemma (Lemma 6) and the Bretagnolle-Huber inequality (Lemma 7) give that

$$\mathbb{P}_{\theta_0}(T_n \ne \theta_0) + \mathbb{P}_{\theta_\delta}(T_n \ne \theta_\delta) \ge \frac{1}{2}\exp(-D_{\text{kl}}\left(P_{\theta_0}^n \| P_{\theta_\delta}^n\right)) \ge \frac{1}{2}\exp(-nM\delta^2),$$

where we use the assumption that $D_{\text{kl}}\left(P_{\theta_0} \| P_{\theta_0+\delta}\right) \le M\delta^2$. One such test is

$$T_n := \begin{cases} \theta_0 & \text{if } |\theta_0 - \hat{\theta}_n| \le |\hat{\theta}_n - \theta_\delta| \\ \theta_\delta & \text{otherwise.} \end{cases}$$

By the triangle inequality $|\hat{\theta}_n - \theta_0| \le \frac{\delta}{2}$ implies $|\hat{\theta}_0 - \theta_n| \le |\hat{\theta}_n - \theta_\delta|$ and $T_n = \theta_0$, and similarly, $|\hat{\theta}_n - \theta_\delta| < \delta/2$ implies $T_n = \theta_\delta$. Thus

$$\mathbb{P}_{\theta_0}(|\hat{\theta}_n - \theta_0| > \delta/2) + \mathbb{P}_{\theta_\delta}(|\hat{\theta}_n - \theta_\delta| \ge \delta/2) \ge \mathbb{P}_{\theta_0}(T_n \ne \theta_0) + \mathbb{P}_{\theta_\delta}(T_n \ne \theta_\delta) \ge \frac{1}{2}\exp(-nM\delta^2),$$

and so

$$\sup_{\theta \in \Theta} \mathbb{P}_\theta(|\hat{\theta}_n - \theta| \ge \delta/2) \ge \max_{\theta \in \{\theta_0, \theta_\delta\}} \mathbb{P}_\theta(|\hat{\theta}_n - \theta| \ge \delta/2) \ge \frac{1}{4}\exp(-nM\delta^2).$$

By choosing $\delta = \sqrt{\log(1/4\alpha')/Mn}$ for any $\alpha' > \alpha$, we have

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta\left(|\hat{\theta}_n - \theta| \le \frac{1}{2}\sqrt{\frac{\log(1/4\alpha')}{Mn}}\right) \le 1 - \alpha' < 1 - \alpha.$$

On the other hand, the time-uniform condition gives that for any n,

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta(|\hat{\theta}_n - \theta| \le t(n)) \ge \inf_{\theta \in \Theta} \mathbb{P}_\theta(|\hat{\theta}_m - \theta| \le t(m), \text{ for all } m \in \mathbb{N}) \ge 1 - \alpha.$$

This implies that $t(n) \ge \frac{1}{2}\sqrt{\log(1/4\alpha')/Mn}$. Taking $\alpha' \searrow \alpha$ thus proves the claim.

Acknowledgments

Thank you to the anonymous reviewers, who provided the simple proof of Lemma 5 and encouraged us to include the material in Section 3.3. This work was partially supported by the Office of Naval Research grant N00014-22-1-2669 and the Stanford DAWN Consortium.

References

Lawrence D. Brown. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, California, 1986.

Frédéric Chyzak and Frank Nielsen. A closed-form formula for the Kullback-Leibler divergence between Cauchy distributions. *arXiv:1905.10965 [cs.IT]*, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

Donald A Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66, 1967.

John C. Duchi. Stats311/EE377: Information theory and statistics. Course at Stanford University, Fall 2023. URL http://web.stanford.edu/class/stats311.

John C. Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *Annals of Statistics*, 49(1):21–48, 2021.

Roger H. Farrell. Asymptotic behavior of expected sample size in certain one sided tests. *Annals of Mathematical Statistics*, 35(1):36–72, 1964.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2):1055–1080, 2021.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sebastien Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the Twenty Seventh Annual Conference on Computational Learning Theory*, pages 423–439. PMLR, 2014.

Christopher Jennison and Bruce W Turnbull. Interim analyses: the repeated confidence interval approach. *Journal of the Royal Statistical Society, Series B*, 51(3):305–334, 1989.

Ramesh Johari, Pete Koomen, Leo Pekelis, and David J. Walsh. Always valid inference: Continuous monitoring of A/B tests. *Operations Research*, pages 1806–1821, 2022.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Alisa Kirichenko and Peter Grunwald. Minimax rates without the fixed sample size assumption. *arXiv:2006.11170v2 [math.ST]*, 2021.

Solomon Kullback. *Information Theory and Statistics*. Dover Publications, 1968.

Tze-Leung Lai. On confidence sequences. *Annals of Statistics*, 4(2):265–280, 1976.

Tze-Leung Lai. Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: a sequential approach. *Communications in Statistics – Theory and Methods*, 13(19):2355–2368, 1984.

Erich L. Lehmann and George Casella. *Theory of Point Estimation, Second Edition*. Springer, 1998.

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.