Estimation with Quantized Parameter Side-Information

Mathis Wetterwald,
LTCI, Telecom Paris IP Paris
91120 Palaiseau, France
mathis.wetterwald@telecom-paris.fr

Ofer Shayevitz,
Tel-Aviv University
Tel Aviv, Israel
ofersha@eng.tau.ac.il

Michèle Wigger,
LTCI, Telecom Paris IP Paris
91120 Palaiseau, France
michele.wigger@telecom-paris.fr

***Abstract*—This paper presents upper and lower bounds on the minimax risk of a parameter estimation problem where the estimator not only observes independent and identically distributed observations but also a quantized version of the parameter, where the quantization has a limitation on the number of levels, but can be optimized otherwise. Our upper and lower bounds have similar decay rates for large number of observation samples and quantization levels.**

I. Introduction

In the classical setting of parameter estimation, one is given n independent and identically distributed (i.i.d.) samples X^n from some unknown distribution P_θ, that belongs to some known parametric family $\{P_\theta\}_{\theta\in\mathcal{J}}$, where in this paper we will consider $\mathcal{J}$ a bounded interval of $\mathbb{R}$. The goal is then to provide an estimate $\hat\theta = \hat\theta(X^n)$ for the true underlying parameter θ, while guaranteeing that the guess is optimal in some sense. Typically, the quality of the estimation is measured with respect to a nonnegative loss function $\ell : \mathcal{J}\times\mathcal{J} \to \mathbb{R}_+$ (often taken to be quadratic), and in particular by the expected loss, or statistical *risk*, $R_\theta(\hat\theta) = \mathbb{E}_\theta[\ell(\theta,\hat\theta)]$. Then, one is usually interested in either 1) uniform bounds on the minimal possible risk of *unbiased estimators* (i.e., satisfying $\mathbb{E}_\theta[\hat\theta] = \theta$), 2) the *Bayesian estimators* that minimize the expected risk $\mathbb{E}_{\theta\sim\pi}[R_\theta(\hat\theta)]$ over some prior θ on the parameter space, or 3) *minimax estimators* that minimize the worst case risk over the parameter space, i.e., uniformly guaranteeing the minimax risk $R^* = \min_{\hat\theta}\max_{\theta\in\mathcal{J}} R_\theta(\hat\theta)$. The problem of parameter estimation under these paradigms has been extensively studied in the literature, and is by now very well understood [1]–[3]. Below, we limit our discussion to the minimax setting.

In this paper, we consider a case where the estimating agent is given access to a quantized version of the parameter θ, prior to seeing the samples. In particular, suppose that the agent can evaluate a k-level quantizer $g : \mathcal{J} \to \{1,\dots,k\}$ of its choosing, and then, after seeing the samples X^n from P_θ, should provide an estimate $\hat\theta(X^n, g(\theta))$. What is the best quantizer g the agent can choose, and what is the effect on the minimax risk? When there are no samples, then this question pertains to simple minimax quantization of the parameter space. For instance, if $\mathcal{J} = [0,1]$ and under quadratic loss, the optimal quantization is simply uniform and the minimax risk is $1/k^2$. Below, we characterize the minimax risk in the case where n samples from P_θ are given. It turns out that the minimax risk still improves roughly by a factor of $1/k^2$, but the (essentially) optimal quantizer is somewhat different; it is essentially cyclically uniform with a cycle that corresponds to the standard deviation of the no-quantizer estimation error.

Related work. Settings of a related flavor have been considered by various authors, especially in the context of distributed parameter estimation with rate constraints, where samples governed by the same parameter are observed by remote parties. Such works include Alswhede and Burnashive [4], Han and Amari [5], Berger and Zhang [6], Zhang et al [7], Braverman et al [8], Hadar and Shayevitz [9], and Hadar et al [10]. Loosely speaking, the theoretical difficulties tackled by these works stem either from the fact that each terminal has a small number of samples, or by the fact that the parameter determines the relation between remote samples and cannot be estimated locally. Our work focuses on a somewhat different regime, where one terminal has direct access to the parameter but needs to efficiently encode (quantize) this information to best help the local estimation of the other.

II. Problem Setup and Main Result

Consider a bounded interval $\mathcal{J} \subseteq \mathbb{R}$ and for each $\theta \in \mathcal{J}$ a pmf P_θ over a discrete finite set $\mathcal{X}$. For ease of notation, let $\mathcal{X} = \{1,\dots,m\}$, for m a given positive integer.

We consider the two-terminal setup in Figure 1. A helper

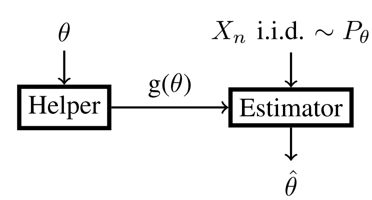

Fig. 1. Parameter estimation with coded side-information from a perfectly informed helper.

perfectly observes the parameter $\theta \in \mathcal{J}$ that is to be estimated at the estimator. To facilitate the estimator's task, the helper can send a k-valued message M, for $k \geq 2$, to the estimator

$$M = g^{(n,k)}(\theta) \in \{1,\dots,k\}. \tag{1}$$

The estimator observes n independent and identically distributed drawings $X^n = (X_1, \ldots, X_n)$ of the pmf P_θ. Based on X^n and M it guesses the parameter θ:

$$\hat{\theta} = \phi^{(n,k)}(M, X^n). \tag{2}$$

We are interested in characterizing the best possible Bayes risk over all helper strategies $g^{(n,k)}$ and estimators $\phi^{(n,k)}$:

$$R^\star_{n,k} := \min_{g^{(n,k)},\phi^{(n,k)}} \max_\theta \mathbb{E}_{P_\theta}\left[\left(\theta - \hat{\theta}\right)^2\right]. \tag{3}$$

We shall need the following assumptions.

- *Assumption 1:* There exists a real number $D > 0$ such that for sufficiently close $\theta, \theta' \in \mathcal{J}$:

$$\|P_\theta - P_{\theta'}\|_1 \le D|\theta - \theta'|. \tag{4}$$

  and there exists an open interval $\mathcal{I}$ such that

$$p_{0,\mathcal{I}} := \inf_{\theta\in\mathcal{I}} \min_{x\in\mathcal{X}} P_\theta(x) > 0. \tag{5}$$

- *Assumption 2:* There exists a real number $F > 0$ such that for $\theta, \theta' \in \mathcal{J}$, if P_θ and $P_{\theta'}$ are sufficiently close:

$$\|P_\theta - P_{\theta'}\|_1 \ge F|\theta - \theta'|. \tag{6}$$

Our main result are upper and lower bounds on $R^\star_{n,k}$ under above assumptions. Notice that these assumptions are in particular satisfied for the canonical problem of P_θ the Bernoulli-θ distribution and $\mathcal{J} = [0, 1]$.

Theorem 1 (Converse)**.** *Under above Assumption 1, for sufficiently large values of n:*

$$C_{\text{lower}} \cdot \frac{1}{nk^2} \le R^\star_{n,k}, \tag{7}$$

where

$$C_{\text{lower}} = \frac{1}{4}(m-1)^{1-m} e^{-\frac{(1+D)^2}{p_{0,\mathcal{I}}}-1} m^{\frac{m}{2}} \frac{1}{(2\pi)^{\frac{m-1}{2}}}. \tag{8}$$

Theorem 2 (Achievability)**.** *Under Assumptions 1 and 2, if $k > \frac{D+F}{2F}$ and for n large enough:*

$$R^\star_{n,k} \le \frac{8m|\mathcal{J}|^2}{nk^2} + \frac{m^2 \ln(nk^2)}{n\left((k-1/2)F - D/2\right)^2}. \tag{9}$$

Moreover, if $\mathcal{J} = [0, 1]$ and P_θ is Bernoulli-θ, then for $k > 2$:

$$R^\star_{n,k} \le \frac{321 \ln(k^2)}{n(k/2-1)^2}. \tag{10}$$

Remark 1. *Ignoring the helper information allows the estimator to achieve a minimax risk in the order of $1/n$. Ignoring the data, a simple helper strategy (partitioning the parameter space $\mathcal{J}$ into equally-long subintervals) allows for a minimax risk of $1/k^2$. In that sense, above bounds allow to combine the benefits of the helper information and the observations X^n.*

Notice further that for moderate or large values of k and n, the upper bounds in Theorem 2 strictly improve over above simple strategies.

III. Achievability Proof (Proof of Theorem 2)

A. Scheme

Fix a length

$$L := \sqrt{\frac{\ln(nk^2)}{n}} \frac{m}{(k-\frac{1}{2})\frac{F}{2} - \frac{D}{4}}. \tag{11}$$

As $k > \frac{D+F}{2F}$, we have $L > 0$. Partition the parameter interval $\mathcal{J}$ into disjoint subintervals of length L, where the last subinterval might be smaller. Denote these intervals by $I_1, \ldots, I_N$, for

$$N := \left\lceil \frac{|\mathcal{J}|}{L} \right\rceil, \tag{12}$$

and the mid-points of the subintervals by $\theta_1^{\text{mid}}, \ldots, \theta_N^{\text{mid}}$.

We propose the following strategy for the helper and the estimator.

Helper: If $\theta \in I_j$, the helper declares

$$M = j \mod k. \tag{13}$$

Estimator: Upon observing a sequence X^n of type π_{X^n} and helper message M, the estimator declares the mid-point θ_j^{mid} that among all the mid-points indicated by the helper message is closest to π_{X^n}. I.e., it sets

$$j^* = \underset{j\colon (j \mod k = M)}{\operatorname{argmin}} \|\pi_{X^n} - P_{\theta_j^{\text{mid}}}\|_1 \tag{14}$$

and declares

$$\hat{\theta} = \theta_{j^*}^{\text{mid}}. \tag{15}$$

If above the argmin is not unique, pick one of the minimizers at random.

In the following subsection, we analyze this scheme and show that it achieves the performances in Theorem 2. In fact, for the Bernoulli-θ result, we will need to modify the choice of L in (11).

Figure 2 plots the minimax risk $R_{n,k}$ achieved by the proposed helper and estimator strategy for $L = 2\sqrt{\ln(nk^2)/(nk^2)}$, $\mathcal{J} = [0, 1]$, $P_\theta = (\theta, \theta - \theta^2, (1-\theta)^2)$ This numerical evaluation confirms that $R_{n,k}$, when seen as a function of k, is nearly proportional to $\frac{1}{k^2}$ (because the slope of the line in the logspace is close to -2). Numerical evaluation thus suggests that $R_{n,k}$ has a decay rate as suggested by our lower bound in Theorem 1 (a little worse than $\frac{1}{k^2}$, the better the higher n gets). The same numerical evaluation also suggests that $R_{n,k}$ decays proportionally to $\frac{1}{n}$, again as suggested by Theorem 1.

B. Analysis

To upper bound the expected mean squared error of our proposed strategy $(g^{(n,k)}, \phi^{(n,k)})$, we start by noting that for an arbitrary parameter $\theta \in \mathcal{J}$, if

$$\|\boldsymbol{\pi}_{X^n} - P_\theta\|_1 < L\frac{k-\frac{1}{2}}{2}F - \frac{LD}{4}, \tag{16}$$

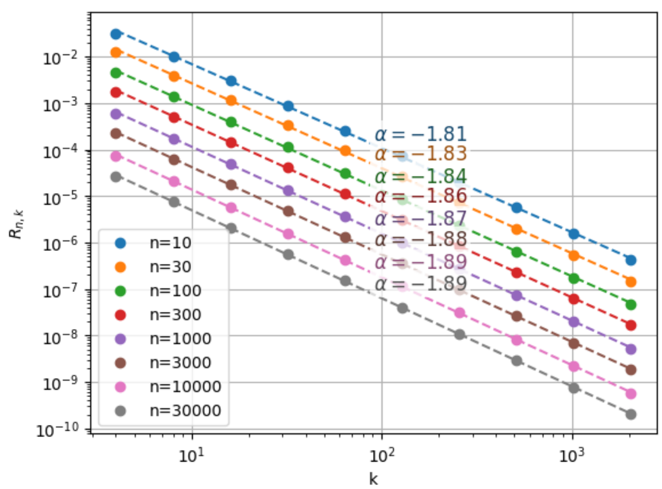

Fig. 2. Minimax risk $R_{n,k}$ of the estimator described in (13)–(15) for varying n and k.

then j^* (as defined in Eq. (14)) is the correct interval containing θ and thus $|\theta^{\text{mid}}_{j^*} - \theta| \le L/2$. Indeed, for any $j \neq j^*$ such that $j \equiv j^* \mod k$:

$$\begin{aligned}
&\|\boldsymbol{\pi}_{X^n} - P_{\theta^{\text{mid}}_{j^*}}\|_1 \\
&\le \|\boldsymbol{\pi}_{X^n} - P_\theta\|_1 + \|P_{\theta^{\text{mid}}_{j^*}} - P_\theta\|_1 && (17)\\
&< L\frac{k-\frac{1}{2}}{2}F - \frac{LD}{4} + \frac{LD}{2} && (18)\\
&= LF\left(k - \frac{1}{2}\right) - \left(L\frac{k-\frac{1}{2}}{2}F - \frac{LD}{4}\right) && (19)\\
&< \|P_\theta - P_{\theta^{\text{mid}}_j}\|_1 - \|\boldsymbol{\pi}_{X^n} - P_\theta\|_1 && (20)\\
&\le \|\boldsymbol{\pi}_{X^n} - P_{\theta^{\text{mid}}_j}\|_1. && (21)
\end{aligned}$$

We can then write:

$$\begin{aligned}
&\mathbb{E}_{P_\theta}\left[|\hat{\theta} - \theta|^2\right] \\
&\le P_\theta\left(\|\pi_{X^n} - \theta\|_1 > L\frac{k-\frac{1}{2}}{2}F - \frac{LD}{4}\right)|\mathcal{J}|^2 \\
&\quad + P_\theta\left(\|\pi_{X^n} - \theta\|_1 \le L\frac{k-\frac{1}{2}}{2}F - \frac{LD}{4}\right)\frac{L^2}{4} && (22)\\
&\le 8me^{-\frac{nL^2}{m^2}((k-1/2)F/2-D/4)^2}|\mathcal{J}|^2 + \frac{L^2}{4} && (23)\\
&= \frac{8m|\mathcal{J}|^2}{nk^2} + \frac{m^2\ln(nk^2)}{n((k-1/2)F-D/2)^2}, && (24)
\end{aligned}$$

where the second inequality holds by bounding probabilities either by 1 or as in [11, Remark to Lemma 2.12] and the last equality by the choice of L in (11).

Refined Analysis for the Bernoulli-θ example leading to (10)*:*

Modify the length L in the scheme to

$$L = \sqrt{\frac{2\ln(k)}{n}}(k/2-1)^{-1}, \tag{25}$$

which is positive for $k > 2$. Notice that for this binary example:

$$|\boldsymbol{\pi}_{X^n}(0) - (1-\theta)| = |\boldsymbol{\pi}_{X^n}(1) - \theta|, \tag{26}$$

and thus δ-typicality as defined in [11] is equivalent to the condition $|\boldsymbol{\pi}_{X^n}(1) - \theta| \le \delta$.

To upper bound the expected mean squared error of our proposed strategy $(g^{(n,k)}, \phi^{(n,k)})$, we start by noting that for an arbitrary parameter $\theta \in \mathcal{J}$, if

$$|\boldsymbol{\pi}_{X^n}(1) - \theta| \le L\frac{k-2}{2}, \tag{27}$$

then j^* is the correct interval containing θ and thus $|\theta^{\text{mid}}_{j^*} - \theta| \le L/2$. In a similar way, for $i = 1, 2, \ldots$, if

$$\begin{aligned}
|\boldsymbol{\pi}_{X^n}(1) - \theta| \in \Bigg(&L\frac{k-2}{2} + (i-1)Lk, \\
&L\frac{k-2}{2} + iLk\Bigg], && (28)
\end{aligned}$$

then j^* represents the interval with $g(j^*) = g(\theta)$ that is at most i-th closest to the interval $I_{g(\theta)}$, and with our strategy the squared error is bounded by

$$|\hat{\theta} - \theta|^2 \le \left(\frac{L}{2} + iLk\right)^2. \tag{29}$$

We can then write, with $|\pi_{X^n}(1) - \theta| = d_\theta$:

$$\begin{aligned}
&\mathbb{E}_{P_\theta}\left[|\hat{\theta} - \theta|^2\right] \\
&\le P_\theta\left(d_\theta \le L\frac{k-2}{2}\right)\frac{L^2}{4} \\
&\quad + 2\sum_{i=1}^{N-1} P_\theta\Big(d_\theta \in \Big(L\frac{k-2}{2} + (i-1)Lk, \\
&\qquad L\frac{k-2}{2} + iLk\Big]\Big)\left(\frac{L}{2} + iLk\right)^2 && (30)\\
&\overset{(a)}{\le} \frac{L^2}{4} + 2P_\theta\Big(d_\theta > L\frac{k-2}{2}\Big)\left(\frac{L}{2} + Lk\right)^2 \\
&\quad + 2\sum_{i=2}^{N-1} P_\theta\Big(d_\theta > L\frac{k-2}{2} + (i-1)Lk\Big) \\
&\qquad \times \underbrace{\left(\left(\frac{L}{2} + iLk\right)^2 - \left(\frac{L}{2} + (i-1)Lk\right)^2\right)}_{=L^2k(1+(2i-1)k)} && (31)\\
&\overset{(b)}{\le} \frac{L^2}{4} + 32me^{-nL^2(k/2-1)^2}(Lk)^2 \\
&\quad + 16mL^2k\sum_{i=2}^{N-2}(1-k+2ik)e^{-nL^2(\frac{k}{2}-1+(i-1)k)^2} && (32)\\
&\overset{(c)}{\le} 65L^2 + 16mL^2k\sum_{i=2}^{N-2}(2ik)e^{-nL^2(i-1)^2k^2} \\
&\quad + 16mL^2k \times 2ke^{-nL^2(k/2-1)^2} && (33)\\
&\overset{(d)}{\le} 65L^2 + 32mL^2\left(1 + k^2\sum_{i=2}^{\infty} ie^{-nL^2(i-1)^2k^2}\right) && (34)
\end{aligned}$$

where we notice that the infinite sum in (34) converges because $e^{-nL^2k^2} < 1$. In above sequence of inequalities:

- (a) holds by writing $P_\theta(|\pi_{X^n}(1) - \theta| \in (a,b]) = P_\theta(|\pi_{X^n}(1)-\theta| > a) - P_\theta(|\pi_{X^n}(1)-\theta| > b)$, rearranging the terms in the sum and dropping negative terms;
- (b) holds by bounding the probabilities using [11, Remark to Lemma 2.12];
- (c) holds by using the value of L to simplify the first exponential term, by getting $i=1$ out of the sum, and by dropping for the other summands the negative term $(1-k)$ as well as the positive term $k/2-1$ in the exponent;
- (d) holds by adding nonnegative terms to the sum, and computing the case $i=1$.

Plugging the choice of L in (25), into (34) yields:

$$\mathbb{E}_{P_\theta}\left[|\hat{\theta}-\theta|^2\right] \leq \frac{\ln(k^2)}{n(k/2-1)^2}\left(65 + 32m(k^2A+1)\right), \tag{35}$$

where here $m=2$ and for $k \geq 2$,

$$A = \sum_{i=2}^{\infty} i e^{-nL^2(i-1)^2k^2} < \sum_{\ell=0}^{\infty}(\ell+2)\left(\frac{1}{k^2}\right)^{(\ell+1)^2} \tag{36}$$
$$\leq \frac{1}{k^2}\sum_{\ell=0}^{\infty}(\ell+2)\left(\frac{1}{4}\right)^{\ell^2} < \frac{2.77}{k^2},$$

and can be bounded by $3/k^2$.

IV. Converse Proof (Proof of Theorem 1)

Before presenting the proof, we make some definitions and state an auxiliary lemma. Let $\mathbb{B}_{\boldsymbol{p}}(r)$ be the $\mathbb{L}_1$-ball of type-vectors $\boldsymbol{\pi} \in \mathcal{P}_n$ of radius r around a probability vector $\boldsymbol{p} \in \mathcal{P}$:

$$\mathbb{B}_{\boldsymbol{p}}(r) \triangleq \{\boldsymbol{\pi} \in \mathcal{P}_n : \ \|\boldsymbol{p}-\boldsymbol{\pi}\|_1 \ \leq r\} \tag{37}$$

Let $\mathcal{I}$ be any subinterval of the parameter space $\mathcal{J}$ of length $|\mathcal{I}| = n^{-1/2}$ with

$$p_0 \triangleq \min_{\theta\in\mathcal{I}}\min_{x\in\mathcal{X}} P_\theta(x) > 0. \tag{38}$$

Define the region of types

$$\mathcal{R}_{\mathcal{I}} = \bigcup_{\theta'\in\mathcal{I}} \mathbb{B}_{P_{\theta'}}\left(\frac{1}{\sqrt{n}}\right). \tag{39}$$

We then have the following lemma:

Lemma 1. *1) For sufficiently large n, the number of n-types in region $\mathcal{R}_{\mathcal{I}}$ is*

$$|\mathcal{R}_{\mathcal{I}}| \geq (m-1)^{-(m-1)} \cdot \sqrt{n}^{m-1}. \tag{40}$$

2) For all $\theta' \in \mathcal{I}$ and $\boldsymbol{\pi} \in \mathcal{R}_{\mathcal{I}}$:

$$P_{\theta'}(\boldsymbol{\pi}) \geq \frac{C}{\sqrt{n}^{m-1}} =: P_{\min}, \tag{41}$$

where $C = e^{-(1+D)^2/p_0 - 1} m^{\frac{m}{2}} \frac{1}{(2\pi)^{\frac{m-1}{2}}}$

Proof. See Appendix A. □

We can now prove the desired lower bound in the theorem. Pick $\theta_1, \theta_2 \in \mathcal{I}$ with same helper message, $g(\theta_1) = g(\theta_2) = j$ and distance:

$$|\theta_1 - \theta_2| \geq \frac{1}{k\sqrt{n}}. \tag{42}$$

Notice that such parameter values must exist because the helper function takes only k possible values and because $\mathcal{I}$ is of length $1/\sqrt{n}$.

The desired lower bound is then obtained through a modified 2-point Le-Cam method where the analysis is restricted to the set $\mathcal{R}_{\mathcal{I}}$:

$$2R^\star_{k,n} \geq \mathbb{E}\left[|\theta_1 - \hat{\theta}_1|^2\right] + \mathbb{E}\left[|\theta_2 - \hat{\theta}_2|^2\right] \tag{43}$$
$$\geq \sum_{\boldsymbol{\pi}\in\mathcal{R}_{\mathcal{I}}} P_{\min} \cdot \Big(|\theta_1 - \hat{\theta}(j,\boldsymbol{\pi})|^2$$
$$+ |\theta_2 - \hat{\theta}(j,\boldsymbol{\pi})|^2\Big) \tag{44}$$
$$\geq \frac{1}{2}|\mathcal{R}_{\mathcal{I}}| P_{min} \cdot |\theta_1 - \theta_2|^2 \tag{45}$$
$$\geq \frac{1}{2}(m-1)^{-m+1} C |\theta_1 - \theta_2|^2 \tag{46}$$
$$\geq \frac{1}{2}(m-1)^{-m+1} C \cdot \frac{1}{k^2 n}. \tag{47}$$

Acknowledgements

The work of M. Wetterwald and M. Wigger has been supported by the ERC under grant agreement 101125691.

Appendix A
Proof of Lemma 1

Proof of 1): By definition of $\mathcal{R}_{\mathcal{I}}$:

$$\mathbb{B}_{P_{\theta'}}\left(\frac{1}{\sqrt{n}}\right) \subseteq \mathcal{R}_{\mathcal{I}} \tag{48}$$

Moreover, by Assumption (38), for sufficiently large n, the number of n-types in ball $\mathbb{B}_{P_{\theta'}}(r)$ is lower bounded by $(r/(m-1))^{m-1}$. Setting $r = n^{-1/2}$, this concludes the proof of 1).

Proof of 2): For any $\theta \in \mathcal{I}$ and type $\boldsymbol{\pi} \in \mathcal{R}_I$:

$$P_\theta(\boldsymbol{\pi}) = \binom{n}{n\boldsymbol{\pi}(1), ..., n\boldsymbol{\pi}(m)} \prod_{i\in\mathcal{X}} P_\theta(i)^{n\boldsymbol{\pi}(i)}. \tag{49}$$

We use Stirling's and Herbert Robbins' bounds to write

$$e^{\frac{1}{12n+1}} \frac{n^n}{e^n}\sqrt{2\pi n} < n! < e^{\frac{1}{12n}} \frac{n^n}{e^n}\sqrt{2\pi n} \tag{50}$$

and thus to conclude:

$$P_\theta(\boldsymbol{\pi}) \geq \frac{e^{\frac{1}{12n+1}}}{\prod_{i=1}^m e^{\frac{1}{12n\boldsymbol{\pi}(i)}}} \cdot \frac{n^n}{\prod_{i=1}^m (n\boldsymbol{\pi}(i))^{n\boldsymbol{\pi}(i)}}$$
$$\cdot \frac{\sqrt{2\pi n}}{\prod_{i=1}^m \sqrt{2\pi n \boldsymbol{\pi}(i)}} \prod_{i=1}^m P_\theta(i)^{n\boldsymbol{\pi}(i)} \tag{51}$$
$$\geq \frac{e^{\frac{1}{12n+1}}}{e^{\frac{m}{12(np_0 - \sqrt{n})}}} \cdot \frac{1}{\prod_{i=1}^m \left(\frac{\boldsymbol{\pi}(i)}{P_\theta(i)}\right)^{n\boldsymbol{\pi}(i)}}$$

$$\cdot \frac{1}{\sqrt{2\pi}^{m-1}} \cdot \sqrt{\frac{n}{\prod_{i=1}^m n\boldsymbol{\pi}(i)}} \tag{52}$$

where we used that $\sum_i \boldsymbol{\pi}(i) = 1$ and

$$\frac{e^{\frac{1}{12n+1}}}{\prod_{i=1}^m e^{\frac{1}{12n\boldsymbol{\pi}_i}}} \geq \frac{e^{\frac{1}{12n+1}}}{e^{\frac{m}{12(np_0-\sqrt{n})}}}. \tag{53}$$

Moreover, since the geometric mean is upper bounded by the arithmetic mean,

$$\prod_{i=1}^m \boldsymbol{\pi}(i) \leq \left(\frac{\sum_{i=1}^m \boldsymbol{\pi}(i)}{m}\right)^m = m^{-m} \tag{54}$$

and thus :

$$\sqrt{\frac{n}{\prod_{i=1}^m n\boldsymbol{\pi}(i)}} \geq n^{-\frac{m-1}{2}} m^{\frac{m}{2}}. \tag{55}$$

By the triangle inequality and Assumption 1, we have for any $\theta \in \mathcal{I}$ and $\boldsymbol{\pi} \in \mathcal{R}_{\mathcal{I}}$:

$$\|\boldsymbol{\pi} - P_\theta\|_1 \leq \|\boldsymbol{\pi} - P_{\theta'}\|_1 + \|P_{\theta'} - P_\theta\|_1 \tag{56}$$

$$\leq \frac{1}{\sqrt{n}} + D|\theta' - \theta| \tag{57}$$

$$\leq \frac{1+D}{\sqrt{n}}, \tag{58}$$

where θ' is the parameter so that $\boldsymbol{\pi} \in \mathbb{B}_{P_{\theta'}}(1/\sqrt{n})$.

We therefore have:

$$\log\left(\prod_{i=1}^m \left(\frac{\boldsymbol{\pi}(i)}{P_\theta(i)}\right)^{n\boldsymbol{\pi}(i)}\right)$$

$$= \sum_{i=1}^m n\boldsymbol{\pi}(i) \log\left(\frac{\boldsymbol{\pi}(i)}{P_\theta(i)}\right) \tag{59}$$

$$= \sum_{i=1}^m n\boldsymbol{\pi}(i) \log\left(1 + \frac{\boldsymbol{\pi}(i) - P_\theta(i)}{P_\theta(i)}\right) \tag{60}$$

$$\leq \sum_{i=1}^m n((\boldsymbol{\pi}(i) - P_\theta(i)) + P_\theta(i)) \frac{\boldsymbol{\pi}(i) - P_\theta(i)}{P_\theta(i)} \tag{61}$$

$$= n\sum_{i=1}^m (\boldsymbol{\pi}(i) - P_\theta(i)) + n\sum_{i=1}^m \frac{(\boldsymbol{\pi}(i) - P_\theta(i))^2}{P_\theta(i)} \tag{62}$$

$$\leq 0 + (1+D)^2/p_0. \tag{63}$$

Combining above observations, and noting that for n such that $n \geq \frac{4}{p_0^2}$ and $n \geq \frac{m^2}{144}$,

$$\frac{e^{\frac{1}{12n+1}}}{e^{\frac{m}{12(np_0-\sqrt{n})}}} \geq \frac{1}{e^{\frac{m}{12\sqrt{n}}}} \geq e^{-1}, \tag{64}$$

we obtain that for sufficiently large values of n:

$$P_{\theta'}(\boldsymbol{\pi}) \geq e^{E-1} m^{\frac{m}{2}} \frac{1}{(2\pi)^{\frac{m-1}{2}}} n^{-\frac{m-1}{2}} \tag{65}$$

for $E = -(1+D)^2/p_0$. This proves part 2).

References

[1] H. L. Van Trees and K. L. Bell, *Detection estimation and modulation theory, part I: detection, estimation, and filtering theory*. John Wiley & Sons, 2013.

[2] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics, New York, NY: Springer-Verlag, 1986.

[3] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. New York, NY: Cambridge University Press, 2025.

[4] R. Ahlswede and M. V. Burnashev, "On Minimax Estimation in the Presence of Side Information About Remote Data," *The Annals of Statistics*, vol. 18, no. 1, pp. 141 – 171, 1990.

[5] T. S. Han and S.-i. Amari, "Parameter estimation with multiterminal data compression," *IEEE transactions on Information Theory*, vol. 41, no. 6, pp. 1802–1833, 1995.

[6] Z. Zhang and T. Berger, "Estimation via compressed information," *IEEE transactions on Information theory*, vol. 34, no. 2, pp. 198–211, 2002.

[7] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[8] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1011–1020, 2016.

[9] U. Hadar and O. Shayevitz, "Distributed estimation of gaussian correlations," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5323–5338, 2019.

[10] U. Hadar, J. Liu, Y. Polyanskiy, and O. Shayevitz, "Communication complexity of estimating correlations," in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 792–803, 2019.

[11] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.