Fisher Information and Mutual Information Constraints

Leighton Pate Barnes and Ayfer Özgür
Stanford University, Stanford, CA 94305
Email: {lpb, aozgur}@stanford.edu

Abstract—We consider the processing of statistical samples $X \sim P_\theta$ by a channel $p(y|x)$, and characterize how the statistical information from the samples for estimating the parameter $\theta \in \mathbb{R}^d$ can scale with the mutual information or capacity of the channel. We show that if the statistical model has a sub-Gaussian score function, then the trace of the Fisher information matrix for estimating θ from Y can scale at most linearly with the mutual information between X and Y. We apply this result to obtain minimax lower bounds in distributed statistical estimation problems, and obtain a tight preconstant for Gaussian mean estimation. We then show how our Fisher information bound can also imply mutual information or Jensen-Shannon divergence based distributed strong data processing inequalities.

I. INTRODUCTION

In this work, we consider the processing of statistical samples $X \sim P_\theta$ by a channel $p(y|x)$, and try to understand how the statistical information from the samples for estimating the parameter θ can scale with the mutual information or capacity of the channel. In particular, we begin by looking at the Fisher information for estimating θ from the processed data Y. Fisher information describes the curvature of the statistical model as one moves around the parameter space, and immediately implies lower bounds for the error in estimating θ from the statistical samples via the well-known Cramér-Rao lower bound [1], [2] for unbiased estimators. It can also imply lower bounds for arbitrarily biased estimators in an asymptotic sense [3], or in a Bayesian setting via the van Trees inequality [4].

Fisher information satisfies a data processing inequality in the sense that it must decrease during processing [5]. In our main result, we develop a strong data processing inequality that more precisely quantifies the maximum possible Fisher information from Y – in terms of the mutual information between X and Y – after processing the data. This is a generalization of recent results on Fisher information due to the authors in both the communication constrained [6] and privacy constrained [7] settings. Other works such as [8] have also considered statistical inference under general channels and obtain some bounds for communication and privacy constrained channels in particular. In concurrent work by the same authors [9], they obtain bounds for mutual information constrained channels that are similar to those in the present paper. Their work differs in that they only treat discrete output

spaces $\mathcal{Y}$, they do not obtain lower bounds for fully interactive communication protocols, and they use a slightly different sub-Gaussianity assumption. One application where general mutual information constrained channels are needed is when communication of statistical samples is done over an analog channel such as an additive white Gaussian noise channel or Gaussian multiple access channel, such as in recent works [10] and [11] which seek to jointly study the communication and estimation problem.

In our main result, we show that if the score of the statistical model is sub-Gaussian – an assumption that is also used in previous works [6], [7] – then the trace of the Fisher information matrix for estimating θ from Y can scale at most linearly with the mutual information between X and Y. We also show by the example of the Gaussian location model that the preconstant we obtain is optimal and cannot be improved. We then apply this result in a distributed setting where there are n nodes each with a sample X_i taken independently from P_θ, and where the nodes can communicate in multiple rounds of communication via a public blackboard. We show how Fisher information from the total blackboard transcript can be similarly bounded, and develop minimax lower bounds in this distributed estimation setting. Finally, in the last section, we show how our Fisher information upper bound can also imply mutual information or Jensen-Shannon divergence based distributed strong data processing inequalities similar to those from [12].

II. PRELIMINARIES

Suppose that $\{P_\theta\}_{\theta \in \Theta}$ is a family of probability distributions parametrized by $\theta \in \Theta \subseteq \mathbb{R}^d$ that is dominated by some sigma-finite measure μ. Let p_θ be the density of P_θ with respect to μ. Let X be a statistical sample drawn from P_θ, and let Y be the output of a channel with transition probability density $p(y|x)$ (with respect to some dominating measure ν on the sample space $\mathcal{Y}$) when x is the input. In this paper we analyze the Fisher information for estimating θ from the processed sample Y and show how it can scale with the mutual information $I_\theta(X;Y)$.[1]

Recall some Fisher information basics. The score for X is defined as
$$S_\theta(X) = \nabla_\theta \log p_\theta(X) .$$

[1]We use $I_\theta(X;Y)$ to denote the mutual information between X and Y when X is drawn from P_θ. In the case that there is a prior distribution on θ this is the same as $I(X;Y|\theta)$.

The Fisher information matrix for the processed samples Y is

$$I_Y(\theta) = \mathbb{E}\left[(\nabla_\theta \log p_\theta(Y))(\nabla_\theta \log p_\theta(Y))^T\right]$$

and we can characterize the trace of this matrix by

$$\mathsf{Tr}(I_Y(\theta)) = \mathbb{E}_Y \|\mathbb{E}_X[S_\theta(X)|Y]\|_2^2 . \tag{1}$$

The decomposition in (1) is one of the main tools used by the authors in [6], [7], and we defer its proof to those references. The proof is a straightforward computation, but requires the interchange of limiting operations, specifically the interchange between integration over the sample space $\mathcal{X}$ and differentiation with respect to the parameter components θ_j. This interchange can be justified under the following regularity conditions (see [13] §26 Lemma 1):

(i) The square-root density $\sqrt{p_\theta(x)}$ is continuously differentiable with respect to each component θ_j at μ-almost all x.

(ii) The Fisher information for each component θ_j, $\mathbb{E}\left[\left(\frac{\partial}{\partial \theta_j}\log p_\theta(X)\right)^2\right]$, exists and is a continuous function of θ_j.

One consequence of these two conditions is that the score random vector has mean zero, i.e.,

$$\mathbb{E}[S_\theta(X)] = 0 .$$

One additional regularity condition is needed in the present paper to interchange limits because we are not assuming that the channel $p(y|x)$ has a finite output alphabet $\mathcal{Y}$ (such as in [6]), or a point-wise local differential privacy condition (such as in [7]):

(iii) The channel $p(y|x)$ is square integrable in the sense that $\int p(y|x)^2 p_\theta(x) d\mu(x) < \infty$ at ν-almost all y for each θ.

Finally, recall some basic facts about sub-Gaussian random variables. We say that a mean-zero random variable X is sub-Gaussian with parameter N if

$$\mathbb{E}\left[e^{\lambda X}\right] \le e^{\frac{\lambda^2 N^2}{2}}$$

for all $\lambda \in \mathbb{R}$. Furthermore, such random variables X satisfy

$$\mathbb{E}\left[e^{\frac{\lambda X^2}{2N^2}}\right] \le \frac{1}{\sqrt{1-\lambda}}$$

for each $\lambda \in [0,1)$ (see, for example, [14] §2.1 and §2.4).

III. FISHER INFORMATION BOUND

In this section we state and prove the following main result concerning how Fisher information can scale with the mutual information between X and Y.

Theorem 1. *Suppose that $\langle u, S_\theta(X)\rangle$ is sub-Gaussian with parameter N for any unit vector $u \in \mathbb{R}^d$. Under regularity conditions (i)-(iii) above,*

$$\mathsf{Tr}(I_Y(\theta)) \le 2N^2 I_\theta(X;Y) .$$

Note that if $Y \in [1:2^k]$, i.e. Y is constrained to being a k-bit message, then $I_\theta(X;Y) \le H(Y) \le k$ and we recover

the sub-Gaussian case from Theorem 2 of [6] as a special case. Similarly, if $p(y|x)$ is a locally differentially private mechanism in the sense that $\frac{p(y|x)}{p(y|x')} \le e^\varepsilon$ for any x, x', y, then $I_\theta(X;Y) = O(\min\{\varepsilon, \varepsilon^2\})$ which recovers Propositions 2 and 4 from [7], again as a special case.

Note also that the mutual information $I_\theta(X;Y)$ is upper bounded by the capacity of the channel $p(y|x)$ since the capacity is the maximum mutual information over all possible input distributions. In this way we can also think of the above theorem as an upper bound on Fisher information in terms of the capacity of the channel that processes the samples.

Proof. We begin by "lifting" the problem to higher dimensions by considering a new dB-dimensional statistical model

$$\mathbf{X} \sim \mathbf{P}_\theta = \prod_{i=1}^B P_{\theta_i}$$

where $\theta = (\theta_1, \ldots, \theta_B)$, $\mathbf{X} = (X_1, \ldots, X_B)$, and $\mathbf{Y} = (Y_1, \ldots, Y_B)$. Each X_i is drawn independently according to P_{θ_i} from the original d-dimensional model, and each Y_i is the corresponding output of the channel. Note that

$$\mathsf{Tr}(I_{\mathbf{Y}}(\theta)) = \sum_{i=1}^B \mathsf{Tr}(I_{Y_i}(\theta_i)) \tag{2}$$

and that when $\theta_0 = \theta_1 = \ldots = \theta_B$ we have

$$\mathsf{Tr}(I_{\mathbf{Y}}(\theta)) = B\mathsf{Tr}(I_Y(\theta_0)) . \tag{3}$$

We will therefore proceed by analyzing $\mathsf{Tr}(I_{\mathbf{Y}}(\theta))$ evaluated at the specific θ values with $\theta_0 = \theta_1 = \ldots = \theta_B$. Note that by taking scaled sums of independent sub-Gaussian random variables, the new dB-dimensional model has a score function that is sub-Gaussian with the same constant N as that of the original model. We use the decomposition shown in (1) for the new dB-dimensional model:

$$\mathsf{Tr}(I_{\mathbf{Y}}(\theta)) = \mathbb{E}_{\mathbf{Y}}\|\mathbb{E}_{\mathbf{X}}[S_\theta(\mathbf{X})|\mathbf{Y}]\|_2^2$$
$$= \mathbb{E}_{\mathbf{Y}}\left[\left(\mathbb{E}_{\mathbf{X}}\left[\frac{p(\mathbf{Y}|\mathbf{X})}{p_\theta(\mathbf{Y})}\langle u_{\mathbf{Y}}, S_\theta(\mathbf{X})\rangle\right]\right)^2\right] \tag{4}$$

where

$$u_{\mathbf{Y}} = \frac{\mathbb{E}_{\mathbf{X}}[S_\theta(\mathbf{X})|\mathbf{Y}]}{\|\mathbb{E}_{\mathbf{X}}[S_\theta(\mathbf{X})|\mathbf{Y}]\|_2} .$$

The key point, and the reason for doing the lifting step, is that the ratio inside the expectation in (4) concentrates around $e^{BI_{\theta_0}(X;Y)}$ as B gets large. More concretely, by the strong law of large numbers,

$$\frac{1}{B}\log\frac{p(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y})} = \frac{1}{B}\log\prod_{i=1}^B \frac{p(y_i|x_i)}{p_{\theta_0}(y_i)}$$
$$= \frac{1}{B}\sum_{i=1}^B \log\frac{p(y_i|x_i)}{p_{\theta_0}(y_i)}$$

converges almost surely to

$$\mathbb{E}_{X,Y}\left[\log\frac{p(Y|X)}{p_{\theta_0}(Y)}\right] = I_{\theta_0}(X;Y)$$

as $B \to \infty$. Therefore

$$\frac{p(\mathbf{Y}|\mathbf{X})}{p_\theta(\mathbf{Y})} \le e^{B(I_{\theta_0}(X;Y)+\varepsilon_B)}$$

with probability at least $1 - \varepsilon_B$ for some ε_B that converges to zero as $B \to \infty$. Let A_B be the event $\left\{ (\mathbf{x},\mathbf{y}) \in \mathcal{X}^B \times \mathcal{Y}^B : \frac{p(\mathbf{y}|\mathbf{x})}{p_{\theta_0}(\mathbf{y})} \le e^{B(I_{\theta_0}(X;Y)+\varepsilon_B)} \right\}$ and let A_B^C be its complement. Following from (4),

$$\mathbb{E}_{\mathbf{Y}}\left[\left(\mathbb{E}_{\mathbf{X}}\left[\frac{p(\mathbf{Y}|\mathbf{X})}{p_\theta(\mathbf{Y})}\langle u_{\mathbf{Y}}, S_\theta(\mathbf{X})\rangle \right] \right)^2 \right]$$

$$\le \mathbb{E}_{\mathbf{Y}}\left[\mathbb{E}_{\mathbf{X}}\left[\frac{p(\mathbf{Y}|\mathbf{X})}{p_\theta(\mathbf{Y})}\langle u_{\mathbf{Y}}, S_\theta(\mathbf{X})\rangle^2 \right] \right]$$

$$= \iint_{A_B} \frac{p(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y})}\langle u_{\mathbf{y}}, S_\theta(\mathbf{x})\rangle^2 p_\theta(\mathbf{x})p_\theta(\mathbf{y})\, d\mu(\mathbf{x})d\nu(\mathbf{y}) \quad (5)$$

$$+ \iint_{A_B^C} \frac{p(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y})}\langle u_{\mathbf{y}}, S_\theta(\mathbf{x})\rangle^2 p_\theta(\mathbf{x})p_\theta(\mathbf{y})\, d\mu(\mathbf{x})d\nu(\mathbf{y})$$

$$\qquad\qquad (6)$$

We bound the term (5) as follows:

$$\exp \iint 1_{A_B}(\mathbf{x},\mathbf{y}) \frac{p(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y})} \frac{\lambda}{2} \left(\frac{\langle u_{\mathbf{y}}, S_\theta(\mathbf{x})\rangle}{N} \right)^2$$
$$\qquad\qquad\qquad\qquad \cdot p_\theta(\mathbf{x})p_\theta(\mathbf{y})\, d\mu(\mathbf{x})d\nu(\mathbf{y})$$

$$\le \iint \frac{p(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y})} \exp \left(1_{A_B}(\mathbf{x},\mathbf{y}) \frac{\lambda}{2} \left(\frac{\langle u_{\mathbf{y}}, S_\theta(\mathbf{x})\rangle}{N} \right)^2 \right)$$
$$\qquad\qquad\qquad\qquad \cdot p_\theta(\mathbf{x})p_\theta(\mathbf{y})\, d\mu(\mathbf{x})d\nu(\mathbf{y})$$

$$\le \varepsilon_B + \iint_{A_B} \frac{p(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y})} \exp \left(\frac{\lambda}{2} \left(\frac{\langle u_{\mathbf{y}}, S_\theta(\mathbf{x})\rangle}{N} \right)^2 \right)$$
$$\qquad\qquad\qquad\qquad \cdot p_\theta(\mathbf{x})p_\theta(\mathbf{y})\, d\mu(\mathbf{x})d\nu(\mathbf{y})$$

$$\le \varepsilon_B + e^{B(I_{\theta_0}(X;Y)+\varepsilon_B)} \iint_{A_B} \exp \left(\frac{\lambda}{2} \left(\frac{\langle u_{\mathbf{y}}, S_\theta(\mathbf{x})\rangle}{N} \right)^2 \right)$$
$$\qquad\qquad\qquad\qquad \cdot p_\theta(\mathbf{x})p_\theta(\mathbf{y})\, d\mu(\mathbf{x})d\nu(\mathbf{y})$$

$$\le \varepsilon_B + \frac{e^{B(I_{\theta_0}(X;Y)+\varepsilon_B)}}{\sqrt{1-\lambda}} .$$

for $0 \le \lambda < 1$. Taking logs,

$$\iint_{A_B} \frac{p(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y})}\langle u_{\mathbf{y}}, S_\theta(\mathbf{x})\rangle^2 p_\theta(\mathbf{x})p_\theta(\mathbf{y})\, d\mu(\mathbf{x})d\nu(\mathbf{y})$$
$$\le \frac{2N^2}{\lambda} \log \left(\varepsilon_B + \frac{e^{B(I_{\theta_0}(X;Y)+\varepsilon_B)}}{\sqrt{1-\lambda}} \right) . \quad (7)$$

For the other term (6),

$$\iint_{A_B^C} \frac{p(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y})}\langle u_{\mathbf{y}}, S_\theta(\mathbf{x})\rangle^2 p_\theta(\mathbf{x})p_\theta(\mathbf{y})\, d\mu(\mathbf{x})d\nu(\mathbf{y})$$

$$\le \iint_{A_B^C} p_\theta(\mathbf{x},\mathbf{y}) \|S_\theta(\mathbf{x})\|_2^2 \, d\mu(\mathbf{x})d\nu(\mathbf{y})$$

$$\le \left(\varepsilon_B \iint p_\theta(\mathbf{x},\mathbf{y}) \|S_\theta(\mathbf{x})\|_2^4 d\mu(\mathbf{x})d\nu(\mathbf{y}) \right)^{\frac{1}{2}} \quad (8)$$

$$= \left(\varepsilon_B \int p_\theta(\mathbf{x}) \|S_\theta(\mathbf{x})\|_2^4 d\mu(\mathbf{x}) \right)^{\frac{1}{2}} \le \left(c_1 \varepsilon_B (dB)^2 N^4 \right)^{\frac{1}{2}}$$
$$\qquad\qquad (9)$$

for some absolute constant c_1. To get (8) we have used the Cauchy-Schwarz inequality, and (9) follows by using the sub-Gaussianity of each of the $(dB)^2$ different terms in $\|S_\theta(X)\|_2^4$ to bound their fourth moments. Combining (7) and (9),

$$\mathsf{Tr}(I_Y(\theta_0)) = \frac{1}{B}\mathsf{Tr}(I_{\mathbf{Y}}(\theta))$$
$$\le \frac{2N^2}{\lambda B} \log \left(\varepsilon_B + \frac{e^{B(I_{\theta_0}(X;Y)+\varepsilon_B)}}{\sqrt{1-\lambda}} \right)$$
$$+ (c_1 \varepsilon_B d^2 N^4)^{\frac{1}{2}}$$

and taking $B \to \infty$ and then $\lambda \to 1$ gives the final result. $\qquad\square$

IV. DISTRIBUTED STATISTICAL ESTIMATION

The upper bound on Fisher information from Section III can be of particular interest in a distributed setting, where there are n distinct nodes each with a sample X_i taken i.i.d. from P_θ. In this setting, the nodes communicate information about their samples via a "public blackboard" in multiple rounds of communication, and then a centralized estimator uses the blackboard transcript after communication to construct an estimate $\hat\theta$ of the parameter θ.

More formally, on round $t = 1,\ldots,T$ of communication, each node $i = 1,\ldots,n$ communicates a random variable $Y_{i,t}$ according to its local data X_i and the previous information written on the transcript $Y_{i-1,t}, Y_{i-2,t},\ldots,Y_{1,t},\Pi_{t-1},\ldots,\Pi_1$ via a channel $p(y_{i,t}|x_i, y_{i-1,t},\ldots,y_{1,t},\pi_{t-1},\ldots,\pi_1)$. Here we define $\pi_t = (y_{1,t},\ldots,y_{n,t})$ to be the information written to the public blackboard after communication on round t. The estimator $\hat\theta$ is then a function of the total blackboard transcript

$$\Pi = (\Pi_1,\ldots,\Pi_T) .$$

By using Theorem 1, we have the following upper bound on Fisher information from the transcript Π.

Corollary 1. *Suppose that $\langle u, S_\theta(X)\rangle$ is sub-Gaussian with parameter N for any unit vector $u \in \mathbb{R}^d$. Under regularity conditions (i)-(iii) above,*

$$\mathsf{Tr}(I_\Pi(\theta)) \le 2N^2 I_\theta(X_1,\ldots,X_n;\Pi) .$$

Proof. Using the equivalent of (1) in this interactive communication setting,

$$\mathsf{Tr}(I_\Pi(\theta)) = \sum_{i=1}^{n} \mathbb{E}_\Pi \left\| \frac{\mathbb{E}_{X_i}[S_\theta(X_i)p_{i,\Pi}(X_i)]}{\mathbb{E}_{X_i}[p_{i,\Pi}(X_i)]} \right\|^2 \quad (10)$$

where

$$p_{i,\pi}(x_i) = \prod_t p(y_{i,t}|x_i, y_{i-1,t}, \ldots, y_{1,t}, \pi_{t-1}, \ldots, \pi_1) \ .$$

The decomposition (10) is detailed in Appendix E of [7]. Using Theorem 1 above, each term in the sum in (10) can be upper bounded by $2N^2 I_\theta(X_i; \Pi)$. Thus by the independence of the X_i,

$$\mathsf{Tr}(I_\Pi(\theta)) \leq 2N^2 I_\theta(X_1, \ldots, X_n; \Pi) \ .$$

$\square$

We now consider the special case of the Gaussian location model where $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ and $\Theta = [-1, 1]^d$. It can be readily checked that in this case $S_\theta(X) \sim \mathcal{N}\left(0, \frac{1}{\sigma^2}I_d\right)$. This leads to the following upper bound on Fisher information, where we will see shortly that the constant on the right-hand side is optimal.

Corollary 2. *Suppose that $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$. Then*

$$\mathsf{Tr}(I_\Pi(\theta)) \leq \frac{2}{\sigma^2} I_\theta(X_1, \ldots, X_n; \Pi) \ .$$

By using Corollary 2 along with the multivariate van Trees inequality from [4], we immediately get the following lower bound on the minimax risk in estimating θ from the transcript Π.

Corollary 3. *Suppose that $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ and $\Theta = [-1, 1]^d$. Then*

$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta}(\Pi) - \theta\|_2^2 \geq \frac{d^2}{\frac{2}{\sigma^2}\sup_{\theta \in \Theta} I_\theta(X_1, \ldots, X_n; \Pi) + \pi^2 d} \ .$$

In order to see that the constant 2 in Corollaries 2 and 3 is optimal, consider the following example. Let $d = 1$ and $T = 1$ and suppose that communication is done independently over additive white Guassian noise channels so that $Y_i = X_i + W_i$ where $W_i \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$ and $\Pi = (Y_1, \ldots, Y_n)$. In this case

$$I_\theta(X_1, \ldots, X_n; \Pi) = \frac{n}{2}\log\left(1 + \frac{\sigma^2}{\sigma^2_{\text{noise}}}\right)$$

and Corollary 3 gives a lower bound of

$$\sup_{\theta \in \Theta} \mathbb{E}[(\hat{\theta}(\Pi) - \theta)^2] \geq \frac{\sigma^2}{n\log\left(1 + \frac{\sigma^2}{\sigma^2_{\text{noise}}}\right)}$$

assuming that n is large enough so that the second term in denominator is negligible. The simple averaging estimator $\hat{\theta}(\Pi) = \frac{1}{n}\sum_{i=1}^n Y_i$ is unbiased and has variance $\frac{\sigma^2}{n}\left(1 + \frac{\sigma^2_{\text{noise}}}{\sigma^2}\right)$. In the regime where $\sigma^2_{\text{noise}} >> \sigma^2$, both the lower bound and expected squared error become $\frac{\sigma^2_{\text{noise}}}{n}$, and thus any constant less than 2 in Corollary 2 would lead to a contradiction.

The distributed estimation setting considered here is the same setting that is considered in past works [6], [7], except that here we do not assume the channels $p(y_{i,t}|x_i, y_{i-1,t}, \ldots, y_{1,t}, \pi_{t-1}, \ldots, \pi_1)$ have a particular communication or privacy-constrained structure, and instead leave

them general and have the constraint be the total mutual information $I_\theta(X_1, \ldots, X_n; \Pi)$. The distributed data-processing inequality from [12] can also apply to general mutual information constrained settings, but requires a bounded likelihood ratio assumption that we do not make. One of the benefits of our Fisher information bound is that it can be applied to the Gaussian location model directly, without the need to truncate the Gaussians so that they satisfy this bounded likelihood ratio assumption. In the subsequent section we show how our Fisher information bounds can also imply Jensen-Shannon divergence based strong data processing inequalities that are very similar to those from [12].

V. RELATION TO DIVERGENCE BOUNDS

Because of Fisher information's interpretation as a second-order approximation of KL divergence (or Jensen-Shannon divergence) as one moves around the parameter space Θ, the above upper bounds on Fisher information can also imply similar upper bounds on the divergence between two distributions from $\{Q_\theta\}_{\theta \in \Theta}$ where $Q_\theta = P_\theta^n \circ P_{\Pi|X_1,\ldots,X_n}$ is the induced distribution for Π.

Instead of working with KL divergence directly, we will instead analyze the related Jensen-Shannon divergence because of its nicer properties. In particular, its square root satisfies the triangle inequality which we describe and use below. Let

$$\mathsf{JS}(P\|Q) = \mathsf{KL}\left(P\left\|\frac{P+Q}{2}\right.\right) + \mathsf{KL}\left(Q\left\|\frac{P+Q}{2}\right.\right)$$

be the Jensen-Shannon divergence between distributions P and Q where $\mathsf{KL}(P\|Q)$ is the usual KL divergence. The square-root of the Jensen-Shannon divergence satisfies the triangle inequality [15] in that

$$\sqrt{\mathsf{JS}(P\|Q)} \leq \sqrt{\mathsf{JS}(P\|R)} + \sqrt{\mathsf{JS}(Q\|R)}$$

for any distributions $P, Q,$ and R.

In this section we will need the following additional regularity condition.

(iv) Suppose that $\mathsf{JS}(Q_\theta\|Q_{\theta+\Delta\theta})$ can be represented by its Taylor expansion

$$\mathsf{JS}(Q_\theta\|Q_{\theta+\Delta\theta}) = \frac{1}{4}\Delta\theta^T I_\Pi(\theta)\Delta\theta + O(\|\Delta\theta\|^3)$$

in such a way that the constants in the Big-O term can be made independent of the choice of $\theta \in \Theta$.

For the Jensen-Shannon divergence we will prove the following data processing bound.

Theorem 2. *Suppose $\theta_0, \theta_1 \in \Theta$ are such that $\theta_\lambda = \lambda\theta_1 + (1-\lambda)\theta_0$ for $\lambda \in [0, 1]$ are contained in Θ. Under the assumptions in Theorem 1 above and regularity condition (iv),*

$$\mathsf{JS}(Q_{\theta_0}\|Q_{\theta_1}) \leq \frac{\|\theta_1 - \theta_0\|_2^2 N^2}{2}\int_0^1 I_{\theta_\lambda}(X_1, \ldots, X_n; \Pi)d\lambda \ .$$

Note that if we consider the random variable V to represent a prior that chooses the parameter θ_0 or θ_1 each with probability $\frac{1}{2}$, then

$$I(V; \Pi) = \frac{1}{2}\mathsf{JS}(Q_{\theta_0}\|Q_{\theta_1})$$

and we can write the result from Theorem 2 as

$$I(V;\Pi) \leq \frac{\|\theta_1 - \theta_0\|_2^2 N^2}{4} \int_0^1 I_{\theta_\lambda}(X_1,\ldots,X_n;\Pi)d\lambda . \quad (11)$$

We compare this to the following theorem. Let $\beta(P_{\theta_0}, P_{\theta_1})$ be the strong data processing inequality (SDPI) constant defined to be the minimum value β such that

$$I(V;\Pi) \leq \beta I(X;\Pi)$$

where $V \to X \to \Pi$ is a Markov chain and $V \sim \text{Bern}(1/2)$ picks whether X is drawn from P_{θ_0} or P_{θ_1} as above.

Theorem 3 ([12] Theorem 1.1). *Suppose $P_{\theta_0}, P_{\theta_1}$ are two probability distributions with $\frac{1}{c}P_{\theta_0} \leq P_{\theta_1} \leq cP_{\theta_0}$ for some constant $c \geq 1$. Let $\beta(P_{\theta_0}, P_{\theta_1})$ be the SDPI constant defined above. Then*

$$I(V;\Pi) \leq Kc\beta(P_{\theta_0}, P_{\theta_1}) \min_{v \in \{\theta_0, \theta_1\}} I_v(X_1,\ldots,X_n;\Pi) \quad (12)$$

for a universal constant K.

Note that in Theorem 3 we have only presented the special case $V \sim \text{Bern}(1/2)$ so that the left-hand side can be written as the mutual information, whereas the general theorem does not assume even probabilities and instead the left-hand side is written as a Hellinger distance term.

To compare (11) and (12) consider what happens in the Gaussian case where $P_\theta \sim \mathcal{N}(\theta, \sigma I_d)$. In this case the SDPI constant $\beta(P_{\theta_0}, P_{\theta_1}) = O\left(\frac{\|\theta_0 - \theta_1\|_2^2}{\sigma^2}\right)$ as used in [12] and shown in [16]. In this case the right-hand sides of the two bounds are very similar with only some minor differences. In particular, (12) has the minimum mutual information over the two parameter values, but has an extra factor of c to compensate. In contrast, (11) is written in terms on the average mutual information along the linear path from θ_0 to θ_1 in the parameter space. One benefit of (11) over (12) is that it does not need this bounded likelihood ratio assumption, instead assuming the sub-Gaussian score function property from Theorem 1, and therefore it can be applied to the Gaussian case directly without any need for truncation.

A. Proof of Theorem 2

Using a multivariate Taylor expansion of $\mathsf{JS}(Q_\theta\|Q_{\theta+\Delta\theta})$ at $\Delta\theta = 0$,

$$\mathsf{JS}(Q_\theta\|Q_{\theta+\Delta\theta}) = \frac{1}{4}\Delta\theta^T I_\Pi(\theta)\Delta\theta + O(\|\Delta\theta\|^3) . \quad (13)$$

Given two parameter values $\theta_0, \theta_1 \in \Theta \subseteq \mathbb{R}^d$ we can upper bound the Jensen-Shannon divergence between Q_{θ_0} and Q_{θ_1} as follows. For any real number $0 \leq \lambda \leq 1$ we define $\theta_\lambda = \lambda\theta_1 + (1-\lambda)\theta_0$. By the triangle inequality,

$$\sqrt{\mathsf{JS}(Q_{\theta_0}\|Q_{\theta_1})} \leq \sum_{i=1}^M \sqrt{\mathsf{JS}\left(Q_{\theta_{\frac{i-1}{M}}}\Big\|Q_{\theta_{\frac{i}{M}}}\right)}$$

and then by squaring and using Jensen's inequality,

$$
\begin{aligned}
\mathsf{JS}&(Q_{\theta_0}\|Q_{\theta_1}) \\
&\leq \left(\sum_{i=1}^M \sqrt{\mathsf{JS}\left(Q_{\theta_{\frac{i-1}{M}}}\Big\|Q_{\theta_{\frac{i}{M}}}\right)}\right)^2 \\
&\leq M \sum_{i=1}^M \mathsf{JS}\left(Q_{\theta_{\frac{i-1}{M}}}\Big\|Q_{\theta_{\frac{i}{M}}}\right) \\
&= \frac{M}{4} \sum_{i=1}^M \left(\frac{\theta_1 - \theta_0}{M}\right)^T I_\Pi\left(\theta_{\frac{i-1}{M}}\right)\left(\frac{\theta_1 - \theta_0}{M}\right) \\
&\quad + MO\left(\left\|\frac{\theta_1 - \theta_0}{M}\right\|^3\right)
\end{aligned} \quad (14)
$$

where the last line (14) follows from (13). Continuing from (14),

$$
\begin{aligned}
\mathsf{JS}(Q_{\theta_0}\|Q_{\theta_1}) \leq &\frac{1}{4}\|\theta_1 - \theta_0\|_2^2 \frac{1}{M}\sum_{i=1}^M \mathsf{Tr}\left(I_\Pi\left(\theta_{\frac{i-1}{M}}\right)\right) \\
&+ MO\left(\left\|\frac{\theta_1 - \theta_0}{M}\right\|^3\right) .
\end{aligned}
$$

Taking $M \to \infty$ we get that

$$\mathsf{JS}(Q_{\theta_0}\|Q_{\theta_1}) \leq \frac{1}{4}\|\theta_1 - \theta_0\|_2^2 \int_0^1 \mathsf{Tr}(I_\Pi(\theta_\lambda)) d\lambda .$$

where we have used the Riemann integrability of the entries of $I_\Pi(\theta_\lambda)$ which can be shown using regularity conditions (i) and (ii). The result then follows from Corollary 1.

B. One-Parameter Families of Distributions

In this section we apply Theorem 2 to the case where we are given two distributions μ_0 and μ_1 that do satisfy the bounded likelihood ratio assumption, even if they are not part of a parametric family. Suppose the two distributions have densities $f_0(x)$ and $f_1(x)$, respectively. In this case we can define a one-parameter family of distributions between the two using an exponential twist such as in [17]. For $\theta \in [0,1]$ we define a new density f_θ by

$$f_\theta(x) = \frac{1}{C_\theta} f_1^\theta(x) f_0^{1-\theta}(x)$$

where $C_\theta = \int f_1^\theta(x) f_0^{1-\theta}(x)dx$ in order to normalize the density. The score function for this one-parameter family is

$$S_\theta(x) = \log \frac{f_1(x)}{f_0(x)} - \frac{C_\theta'}{C_\theta} .$$

If, like in [12], we make the bounded likelihood ratio assumption

$$\frac{1}{c}f_0(x) \leq f_1(x) \leq cf_0(x) ,$$

then $|S_\theta(x)| \leq 2\log c$ for all x. This implies the score function is sub-Gaussian with a parameter that is $O(\log c)$, and yields

$$I(V;\Pi) \leq K(\log c)^2 \int_0^1 I_\lambda(X_1,\ldots,X_n;\Pi)d\lambda$$

for an absolute constant K.

REFERENCES

[1] H. Cramér, *Mathematical Methods of Statistics*. Princeton Univ. Press, 1946.

[2] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, vol. 37, 1945.

[3] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[4] R. D. Gill and B. Y. Levit, "Applications of the van trees inequality: a Bayesian cramér-rao bound," *Bernoulli*, vol. 1, no. 1/2, pp. 059–079, 1995.

[5] R. Zamir, "A proof of the fisher information inequality via a data processing argument," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1246–1250, 1998.

[6] L. P. Barnes, Y. Han, and A. Özgür, "Lower bounds for learning distributions under communication constraints via fisher information," *Journal of Machine Learning Resarch*, vol. 21, no. 236, pp. 1–30, 2020.

[7] L. P. Barnes, W.-N. Chen, and A. Özgür, "Fisher information under local differential privacy," *IEEE Journal of Selected Areas in Information Theory*, vol. 1, no. 3, pp. 645 – 659, 2020.

[8] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints: Lower bounds from chi-square contraction," in *Proceedings of the 32nd Conference on Learning Theory*, vol. 99. Phoenix, USA: PMLR, 25–28 Jun 2019, pp. 3–17.

[9] J. Acharya, C. Canonne, and H. Tyagi, "General lower bounds for interactive high-dimensionalestimation under information constraints," arXiv preprint, arXiv:*2010.06562*, 2020.

[10] C.-Z. Lee, L. P. Barnes, and A. Özgür, "Over-the-air statistical estimation," *Proceedings of the 2020 IEEE Global Communications Conference*.

[11] ——, "Lower bounds for over-the-air statistical estimation," *Submitted to the 2021 IEEE International Symposium on Information Theory*.

[12] 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*. ACM, 2016, p. 1011–1020.

[13] A. A. Borovkov, *Mathematical Statistics*. Gordon and Breach Science Publishers, 1998.

[14] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Serieis in Statistical and Probabilistic Mathematics, 2019.

[15] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[16] M. Raginsky, "Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, 2016.

[17] A. G. Dabak and D. H. Johnson, "Relations between kullback-leibler distance and fisher information," *Available at http://ece.rice.edu/~dhj/distance.pdf*.