Lecture 14: Communication/privacy constrained estimation

Lecturer: Yanjun Han

May 12, 2021

Today's plan

constrained estimation:

- communication and privacy constraints
- distributed estimation and types of communication protocols
- tool I: strong data processing inequality
- tool II: van Trees inequality + quantized Fisher information
- tool III: direct modeling + Assouad's lemma

Constrained estimation

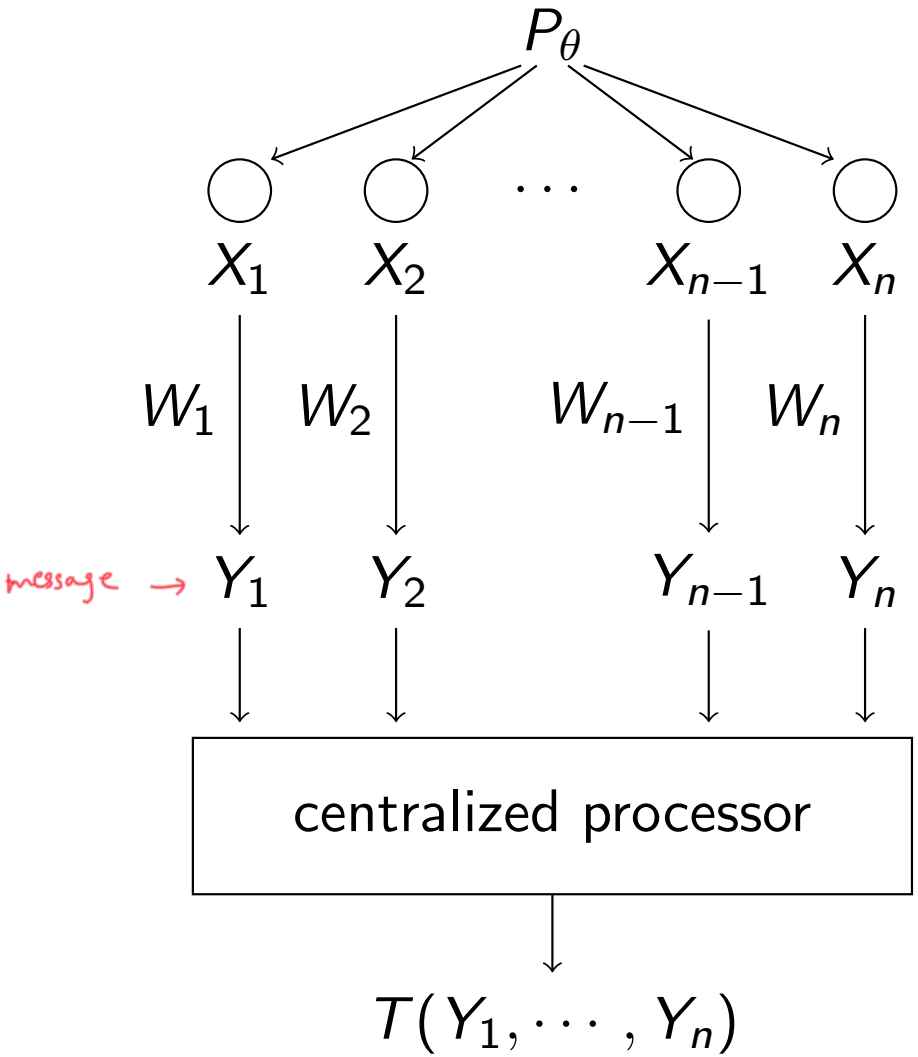
- usual statistical model: $X \sim P_\theta$
- however, the learner only has access to Y, which is the output of sending X via a channel $W \in \mathcal{W}$ up to learner's choice
- target: given loss $L(\theta, T)$, jointly design W and $T(Y)$
- communication-constrained channel $\mathcal{W}_k$:

$$W \in \mathcal{W}_k : \mathcal{X} \to \{0,1\}^k$$

- privacy-constrained channel $\mathcal{W}_\varepsilon$:

$$W \in \mathcal{W}_\varepsilon : \mathcal{X} \to \mathcal{Y} \qquad \sup_{x,x' \in \mathcal{X}} \sup_{A \subseteq \mathcal{Y}} \frac{W(A \mid x)}{W(A \mid x')} \leq e^\varepsilon.$$

Distributed estimation



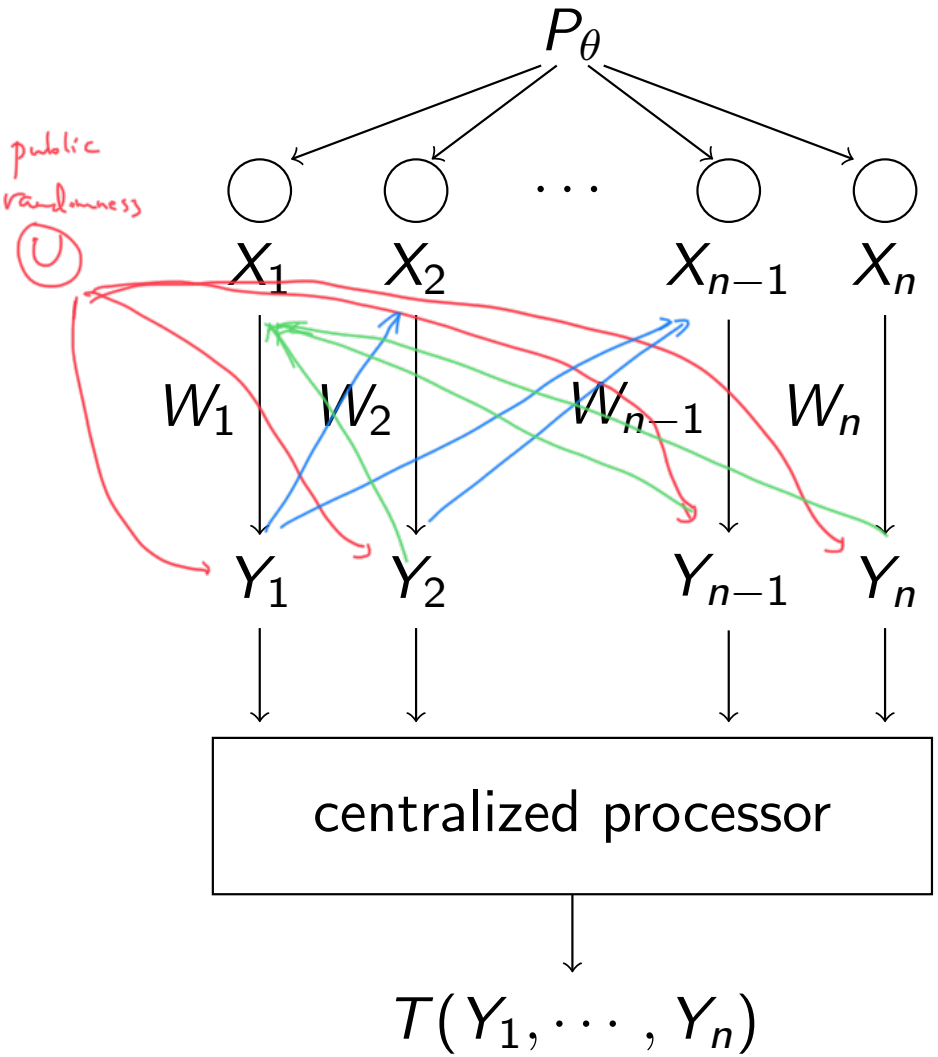
minimax risk:

$$\inf_T \inf_{W_1,\cdots,W_n} \sup_\theta \mathbb{E}_\theta[L(\theta, T(Y^n))]$$

examples:

- distributed Gaussian mean estimation $P_\theta = \mathcal{N}(\theta, I_d)$
- distributed discrete distribution estimation $P_\theta = (\theta_1, \cdots, \theta_d)$
- distributed uniformity testing

Communication protocols

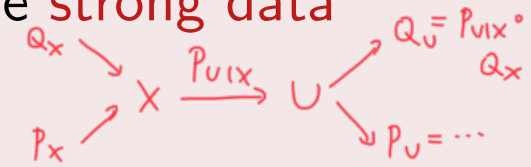


- simultaneous message passing (SMP) with private randomness
- SMP with public randomness
- sequential communication protocol
- blackboard (interactive) communication protocol

Tool I: strong data processing inequality

Definition

Given a f-divergence and $P_{U|X}$ (or in addition P_X), the strong data processing coefficient is defined as



$$\eta_f(P_{U|X}) \triangleq \sup_{P_X, Q_X} \frac{D_f(P_{U|X} \circ Q_X, P_{U|X} \circ P_X)}{D_f(Q_X, P_X)}$$

$$\eta_f(P_{U|X}, P_X) \triangleq \sup_{Q_X} \frac{D_f(P_{U|X} \circ Q_X, P_{U|X} \circ P_X)}{D_f(Q_X, P_X)}$$

- data-processing inequality: $\eta_f \leq 1$
- Polyanskiy and Wu (2017):

$$\eta_{\chi^2} = \eta_{\mathsf{KL}} \leq \eta_f \leq \eta_{\mathsf{TV}} = \sup_{x,x'} \|P_{U|X=x} - P_{U|X=x'}\|_{\mathsf{TV}}$$

- Ordentlich and Polyanskiy (2021): suffice to consider binary (P_X, Q_X) for any $\eta_f(P_{U|X})$

The KL case

Lemma

$$\eta_{\mathsf{KL}}(P_{U|X}) = \sup_{P_{XY}:U-X-Y} \frac{I(U;Y)}{I(X;Y)}$$

$$\eta_{\mathsf{KL}}(P_{U|X}, P_X) = \sup_{P_{Y|X}:U-X-Y} \frac{I(U;Y)}{I(X;Y)}$$

examples:

- $\mathcal{X} = \mathcal{U} = \{0,1\}$ and $\mathbb{P}(U \neq x \mid X = x) = \varepsilon$:

$$\eta_{\mathsf{KL}}(P_{U|X}) = (1-2\varepsilon)^2$$

- (X,U) joint Gaussian with correlation $\rho \in [-1,1]$:

$$\eta_{\mathsf{KL}}(P_{U|X}, P_X) = \rho^2$$

Why mutual information?

- upper bounds of $I(X;Y)$ available for both communication and privacy constrained estimation
- communication-constrained channel $\mathcal{W}_k$:

$$I(X;Y) \leq k$$

$I(X;Y) \leq \sup_x D_{KL}(P_{Y|X=x} \| \text{Unif}(y)) \leq k$

- privacy-constrained channel $\mathcal{W}_\varepsilon$:

$$I(X;Y) \leq \min\{\varepsilon, 2\varepsilon^2\}$$

$\leq \max_{x,x'} D_{KL}(\underbrace{P_{Y|x}}_{p} \| \underbrace{P_{Y|x'}}_{q})$

$e^{-\varepsilon} \leq \frac{p(y)}{q(y)} \leq e^{\varepsilon}$

1) $D_{KL}(p\|q) = \sum_x p(x) \underbrace{\log \frac{p(x)}{q(x)}}_{\leq \varepsilon} \leq \varepsilon$

2) $D_{KL}(p\|q) = \sum_x \left[p(x) \log \frac{p(x)}{q(x)} - p(x) + q(x) \right]$

$= \sum_x q(x) \left[\underbrace{\frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} - \frac{p(x)}{q(x)} + 1}_{\leq O(\varepsilon^2)} \right] \leq 2\varepsilon^2$

A simple example

- setting: $X_1, \cdots, X_n \sim \mathsf{Bern}(p)$ with $p \in [0,1]$, and $W_i(Y_i \mid X_i) \in \mathcal{W}_\varepsilon$
- claim:

$$\inf_T \inf_{W_1,\cdots,W_n} \sup_p \mathbb{E}_p[(T(Y^n) - p)^2] \gtrsim \frac{1}{n} + \frac{1}{n\varepsilon^2}$$

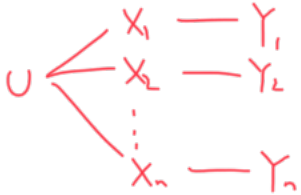
- analysis: let $U \sim \mathsf{Unif}(\{0,1\})$, and $P_{X_i|U=0} = \mathsf{Bern}(1/2 - \delta)$, $P_{X_i|U=1} = \mathsf{Bern}(1/2 + \delta)$
- strong data-processing inequality: $U - X^n - Y^n$

$$I(U; Y^n) \le \eta_{\mathsf{KL}}(P_{U|X^n}, P_{X^n}) \cdot I(X^n; Y^n)$$

- however, computing $\eta_{\mathsf{KL}}(P_{U|X^n}, P_{X^n})$ is very complicated...

Tensorization under SMP

- key observation: under SMP, $(Y_1, \cdots, Y_n) \mid U$ are independent
- upper bound of mutual information:



$$
\begin{aligned}
I(U; Y^n) &\le \sum_{i=1}^{n} I(U; Y_i) \\
&\le \sum_{i=1}^{n} \eta_{\mathrm{KL}}(P_{U|X_i}, P_{X_i}) \cdot I(X_i; Y_i) \\
&\le n \cdot (2\delta)^2 \cdot 2\varepsilon^2 \\
&= 8n\delta^2\varepsilon^2
\end{aligned}
$$

- two-point method: choosing $\delta^2 \asymp (n\varepsilon^2)^{-1}$

Solution for the interactive protocols

Theorem (Braverman, Garg, Ma, Nguyen, Woodruff, 2016)

Suppose that $U \sim \mathrm{Unif}(\{0,1\})$ and $P_{X|U=0}(x)/P_{X|U=1}(x) \in [1/c, c]$ for some $c > 1$ and any $x \in \mathcal{X}$. Then if Y^n is obtained from X^n via any interactive communication protocol, we have

$$H^2(P_{Y^n|U=0}, P_{Y^n|U=1}) \lesssim_c \eta_{\mathrm{KL}}(P_{U|X}, P_X) \cdot I(X^n; Y^n)$$

upper bound on $I(X^n; Y^n)$ under interactive protocols:

- communication constraints: $I(X^n; Y^n) \le \log |\mathcal{Y}^n| = nk$
- privacy constraints:

$$I(X^n; Y^n) = \sum_t I(X^n; Y_t \mid Y^{t-1}) \le n \min\{\varepsilon, 2\varepsilon^2\}$$

Product Bernoulli example

- setting: $X_1, \cdots, X_n \sim \prod_{i=1}^{d} \mathrm{Bern}(p_i)$, with $(p_1, \cdots, p_d) \in [0,1]^d$
- constraints: communication or privacy
- claim:

$$\inf_{T} \inf_{W_1, \cdots, W_n \in \mathcal{W}} \sup_{p} \mathbb{E}_p[\| T(Y^n) - p \|_2^2] \gtrsim \begin{cases} \frac{d}{n} \cdot \frac{d}{\min\{k, d\}} & \text{if } \mathcal{W} = \mathcal{W}_k \\ \frac{d}{n} \cdot \frac{d}{\min\{\varepsilon^2, \varepsilon, d\}} & \text{if } \mathcal{W} = \mathcal{W}_\varepsilon \end{cases}$$

- analysis under SMP: apply Fano's method to $U \sim \mathrm{Unif}(\{1/2 \pm \delta\}^d)$, but need to compute $\eta_{\mathrm{KL}}(P_{U|X}, P_X)$
- analysis under interactive protocol: direct sum argument + previous data processing inequality

Tensorization property of SDPI

Theorem

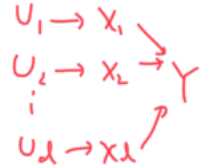
$$\eta_{\mathrm{KL}}\left(\prod_i P_{U_i|X_i}, \prod_i P_{X_i}\right) = \max_i \eta_{\mathrm{KL}}(P_{U_i|X_i}, P_{X_i})$$

proof:

$$\begin{aligned}
\frac{I(Y;U^d)}{I(Y;X^d)} &= \frac{\mathbb{E}_Y[D_{\mathrm{KL}}(P_{U^d|Y}\|P_{U^d})]}{\mathbb{E}_Y[D_{\mathrm{KL}}(P_{X^d|Y}\|P_{X^d})]} \\
&= \frac{\sum_i \mathbb{E}_{(U^{i-1},Y)}[D_{\mathrm{KL}}(P_{U_i|U^{i-1},Y}\|P_{U_i})]}{\sum_i \mathbb{E}_{(X^{i-1},Y)}[D_{\mathrm{KL}}(P_{X_i|X^{i-1},Y}\|P_{X_i})]} \\
&\le \frac{\sum_i \mathbb{E}_{(U^{i-1},Y)}\mathbb{E}_{X^{i-1}|U^{i-1},Y}[D_{\mathrm{KL}}(P_{U_i|X^{i-1},Y}\|P_{U_i})]}{\sum_i \mathbb{E}_{(X^{i-1},Y)}[D_{\mathrm{KL}}(P_{X_i|X^{i-1},Y}\|P_{X_i})]} \\
&\le \max_{i,y,x^{i-1}} \frac{D_{\mathrm{KL}}(P_{U_i|X^{i-1}=x^{i-1},Y=y}\|P_{U_i})}{D_{\mathrm{KL}}(P_{X_i|X^{i-1}=x^{i-1},Y=y}\|P_{X_i})} \\
&\le \max_i \eta_{\mathrm{KL}}(P_{U_i|X_i}, P_{X_i})
\end{aligned}$$

$P_{U_i|U^{i-1}} = P_{U_i}$

$P_{U_i|U^{i-1},Y} = \sum_{x^{i-1}} P_{U_i|X^{i-1},Y} \times P_{X^{i-1}|U^{i-1},Y}$



Analysis under interactive protocol

- Assouad's lemma under interactive protocol:

$$\sum_{j=1}^{d} \mathbb{P}(\widehat{v}_j(Y^n) \neq v_j) \geq \frac{1}{2} \sum_{j=1}^{d} (1 - \|P_{+j} - P_{-j}\|_{\mathrm{TV}})$$

  where P_{+j}, P_{-j} are the distributions of the final message Y^n under $v_j = 1$ and $v_j = -1$, respectively

- idea of upper bounding $\|P_{+j} - P_{-j}\|_{\mathrm{TV}}$: if P_θ has a product structure, then every node could discard all but the j-th coordinate of X

$= P_{\theta_1} \times P_{\theta_2} \times \cdots \times P_{\theta_d}$

- consequence: reduce to 1-dimensional problem
- so-called "direct-sum" result in literature

Gaussian location model

- setting: $X_1, \cdots, X_n \sim \mathcal{N}(\theta, I_d)$, with $\theta \in \mathbb{R}^d$
- constraints: communication or privacy
- result:

$$\inf_{T} \inf_{W_1, \cdots, W_n \in \mathcal{W}} \sup_{\theta} \mathbb{E}_\theta[\| T(Y^n) - \theta \|_2^2] \gtrsim \begin{cases} \frac{d}{n} \cdot \frac{d}{\min\{k, d\}} & \text{if } \mathcal{W} = \mathcal{W}_k \\ \frac{d}{n} \cdot \frac{d}{\min\{\varepsilon^2, \varepsilon, d\}} & \text{if } \mathcal{W} = \mathcal{W}_\varepsilon \end{cases}$$

- analysis exactly the same, with SDPI specialized to Gaussian model

Tool II: quantized Fisher information

- recall the definition of Fisher information matrix:

$$I_X(\theta) = \mathbb{E}_{X \sim P_\theta}\left[\left(\frac{\partial \log p(X \mid \theta)}{\partial \theta}\right)\left(\frac{\partial \log p(X \mid \theta)}{\partial \theta}\right)^\top\right]$$

- quantized Fisher information for Y:

$$I_Y(\theta) = \mathbb{E}_{Y \sim W \circ P_\theta}\left[\left(\frac{\partial \log p(Y \mid \theta)}{\partial \theta}\right)\left(\frac{\partial \log p(Y \mid \theta)}{\partial \theta}\right)^\top\right]$$

- relationship between $p(Y \mid \theta)$ and $p(X \mid \theta)$:

$$p(y \mid \theta) = \sum_{x \in \mathcal{X}} W(y \mid x) p(x \mid \theta)$$

Lemma

$$I_Y(\theta) = \mathbb{E}_Y\left[\left(\mathbb{E}_\theta[\dot{\ell}_\theta(X) \mid Y]\right)\left(\mathbb{E}_\theta[\dot{\ell}_\theta(X) \mid Y]\right)^\top\right]$$

van Trees inequality

Theorem (van Trees inequality, a.k.a. Bayesian Cramér–Rao bound)

For a differentiable prior density π on $[a, b]$ with $\pi(a) = \pi(b) = 0$, it holds that

$$\mathbb{E}_{\theta\sim\pi}\mathbb{E}_{X\sim P_\theta}[(T(X) - \theta)^2] \geq \frac{1}{\mathbb{E}_{\theta\sim\pi}[I_X(\theta)] + I(\pi)},$$

where $I(\pi) = \int_a^b [\pi'(t)]^2/\pi(t)dt$ is the Fisher information of π.

Proof: consider two expressions of the quantity

$$S = \int_a^b \int_{\mathcal{X}} (T(x) - \theta)\frac{\partial[p(x \mid \theta)\pi(\theta)]}{\partial\theta} dx d\theta \quad = \; |$$

$$S^2 \leq \left(\int_a^b \int_{\mathcal{X}} (T(x)-\theta)^2 \pi(\theta) p(x|\theta)\, dx d\theta\right)\left(\int_a^b \int_{\mathcal{X}} \left(\frac{\partial[p(x|\theta)\pi(\theta)]}{\partial\theta}\right)^2 \Big/ \pi(\theta)p(x|\theta)^{dx d\theta}\right)$$

$$= \mathbb{E}_{\theta\sim\pi}[I(\theta)] + I(\pi)$$

Corollaries and generalizations

- Corollary 1: for 1-D model,

$$\sup_{\theta\in[a,b]} \mathbb{E}_{X\sim P_\theta}[(T(X)-\theta)^2] \geq \frac{1}{\sup_{\theta\in[a,b]} I_X(\theta) + 4\pi^2/(b-a)^2}$$

- Corollary 2: for high-dimensional model,

$$\sup_{\theta\in[a,b]^d} \mathbb{E}_{X\sim P_\theta}[\|T(X)-\theta\|_2^2] \geq \frac{d^2}{\sup_{\theta\in[a,b]^d} \mathrm{Tr}(I_X(\theta)) + 4d\pi^2/(b-a)^2}$$

- application in constrained estimation: upper bound

$$\sup_{\theta\in[a,b]^d} \sup_{W\in\mathcal{W}} \mathrm{Tr}(I_{Y^n}(\theta))$$

Tensorization property

Even under an interactive protocol, we have

$$\mathrm{Tr}(I_{Y^n}(\theta)) \leq n \cdot \sup_{W\in\mathcal{W}} \mathrm{Tr}(I_Y(\theta))$$

Upper bounds of individual Fisher information

Upper bound I (bounded variance)

If $\dot{\ell}_\theta(x)$ has a bounded second moment σ^2 along any direction under $x \sim P_\theta$, then

$$\sup_{W \in \mathcal{W}} \mathrm{Tr}(I_Y(\theta)) \leq \begin{cases} 2^k \sigma^2 & \text{if } \mathcal{W} = \mathcal{W}_k, \\ \min\{e^\varepsilon, \varepsilon^2\}\sigma^2 & \text{if } \mathcal{W} = \mathcal{W}_\varepsilon. \end{cases}$$

Upper bound II (sub-Gaussian)

If $\dot{\ell}_\theta(x)$ is subGaussian with parameter σ^2 under $x \sim P_\theta$, then

$$\sup_{W \in \mathcal{W}} \mathrm{Tr}(I_Y(\theta)) \lesssim \begin{cases} k\sigma^2 & \text{if } \mathcal{W} = \mathcal{W}_k, \\ \min\{\varepsilon, \varepsilon^2\}\sigma^2 & \text{if } \mathcal{W} = \mathcal{W}_\varepsilon. \end{cases}$$

Example I: Gaussian location model

- van Trees inequality:

$$R^\star_{n,d} \geq \frac{d^2}{n \cdot \sup_{\theta \in [a,b]^d} \sup_{W \in \mathcal{W}} \mathrm{Tr}(I_Y(\theta)) + 4d\pi^2/(b-a)^2}$$

- score function $\dot{\ell}_\theta(x) = \theta - x$ is 1-sub-Gaussian
- communication constraint: $\mathrm{Tr}(I_Y(\theta)) \lesssim k$, giving

$$R^\star_{n,d,k} \gtrsim \frac{d}{n} \cdot \frac{d}{\min\{d, k\}}$$

- privacy constraint: $\mathrm{Tr}(I_Y(\theta)) \lesssim \min\{\varepsilon, \varepsilon^2\}$, giving

$$R^\star_{n,d,\varepsilon} \gtrsim \frac{d}{n} \cdot \frac{d}{\min\{d, \varepsilon, \varepsilon^2\}}$$

Example II: discrete distribution model

- setting: $X_1, \cdots, X_n \sim \theta = (\theta_1, \cdots, \theta_d)$, which is a probability vector
- van Trees inequality:

$$R^\star_{n,d} \geq \frac{d^2}{n \cdot \sup_{\theta \in [a,b]^d} \sup_{W \in \mathcal{W}} \mathrm{Tr}(I_Y(\theta)) + 4d\pi^2/(b-a)^2}$$

- score function only finite second moment d
- communication constraint: $\mathrm{Tr}(I_Y(\theta)) \lesssim d2^k$, giving

$$R^\star_{n,d,k} \gtrsim \frac{1}{n} \cdot \frac{d}{\min\{d, 2^k\}}$$

- privacy constraint: $\mathrm{Tr}(I_Y(\theta)) \lesssim \min\{e^\varepsilon, \varepsilon^2\} d$, giving

$$R^\star_{n,d,k} \gtrsim \frac{1}{n} \cdot \frac{d}{\min\{d, \varepsilon^2, e^\varepsilon\}}$$

Tool III: direct modeling + Assouad's lemma

- setting: $X_1, \cdots, X_n \sim p = (p_1, \cdots, p_d)$
- target: uniformity testing $p = \mathrm{unif}_d$ vs. $\|p - \mathrm{unif}_d\|_{\mathrm{TV}} \geq \delta$
- result: table of sample complexities

	$\mathcal{W}_k$	$\mathcal{W}_\varepsilon$
SMP w/ public coin (this lecture)	$\frac{\sqrt{d}}{\delta^2} \cdot \sqrt{\frac{d}{\min\{d, 2^k\}}}$	$\frac{\sqrt{d}}{\delta^2} \cdot \frac{\sqrt{d}}{\min\{\sqrt{d}, \varepsilon^2\}}$
SMP w/ private coin (future lecture)	$\frac{\sqrt{d}}{\delta^2} \cdot \frac{d}{\min\{d, 2^k\}}$	$\frac{\sqrt{d}}{\delta^2} \cdot \frac{d}{\min\{d, \varepsilon^2\}}$

Direct modeling the channel

- assume $\mathcal{W} = \mathcal{W}_k$
- Paninski's construction: for $v \in \{\pm 1\}^{d/2}$, let

$$X \sim P_v = \left(\frac{1 + v_1\delta}{d}, \frac{1 - v_1\delta}{d}, \cdots, \frac{1 + v_{d/2}\delta}{d}, \frac{1 - v_{d/2}\delta}{d} \right)$$

- distribution of $Y \sim Q_v$: for $y \in \{0,1\}^k$,

$$Q_v(y) = \sum_{i \in [d/2]} \left(\frac{1 + v_i\delta}{d} W(y \mid 2i-1) + \frac{1 - v_i\delta}{d} W(y \mid 2i) \right)$$

χ^2-method

- the central quantity in χ^2-method:

$$\sum_{y\in\{0,1\}^k} \frac{Q_v(y)Q_{v'}(y)}{Q_0(y)} = 1 + \frac{2\delta^2}{d} v^\top H(W) v',$$

where $H \in \mathbb{R}^{d/2\times d/2}$ with

$$H(W)_{i,j} = \sum_y \frac{(W(y \mid 2i-1) - W(y \mid 2i))(W(y \mid 2j-1) - W(y \mid 2j))}{\sum_x W(y \mid x)}$$

- χ^2-method:

$$\chi^2(\mathbb{E}_v[Q_v^{\otimes n}], Q_0) \lesssim \frac{n^2\delta^4}{d^2} \max_{W\in\mathcal{W}} \|H(W)\|_F^2$$

- upper bounding the Frobenius norm: $H(W) \preceq H^\star(W)$, which has non-negative entries and all rows summing to 2. Perron–Frobenius gives $\rho(H^\star) \le 2$, and $\|H^\star\|_F^2 \le \mathrm{Tr}(H^\star) \le 2^{k+1}$.

References: strong data processing inequality

- Maxim Raginsky. "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels." *IEEE Transactions on Information Theory* 62.6 (2016): 3355-3389.
- Yury Polyanskiy and Yihong Wu. "Strong data-processing inequalities for channels and Bayesian networks." *Convexity and Concentration.* Springer, New York, NY, 2017. 211-249.
- John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. "Local privacy and statistical minimax rates." 2013 IEEE 54th Annual Symposium on Foundations of Computer Science. IEEE, 2013.
- Mark Braverman, Ankit Garg, Tengyu Ma, Huy L. Nguyen, and David 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.

References: more recent approaches

- Yanjun Han, Ayfer Özgür, and Tsachy Weissman. "Geometric lower bounds for distributed parameter estimation under communication constraints." *Conference On Learning Theory*. PMLR, 2018.
- Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. "Lower bounds for learning distributions under communication constraints via Fisher information." *Journal of Machine Learning Research* 21.236 (2020): 1-30.
- Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. "Inference under information constraints I: Lower bounds from chi-square contraction." *IEEE Transactions on Information Theory* 66.12 (2020): 7835-7855.
- FOCS'20 tutorial: `http://www.cs.columbia.edu/~ccanonne/tutorial-focs2020/`

Next lecture: scandiction (Tsachy)