Distributed Statistical Estimation of High-Dimensional and Nonparametric Distributions

Yanjun Han (Stanford EE)

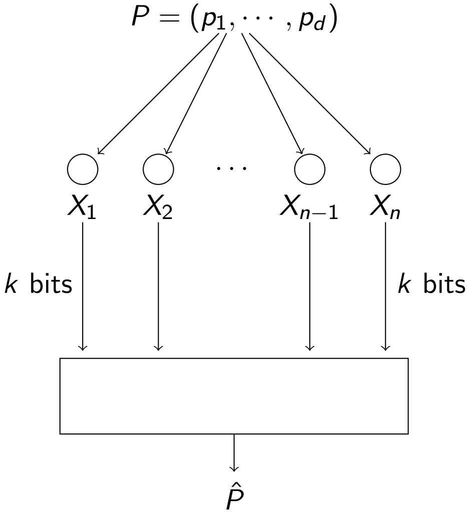
Joint work with:

Pritam Mukherjee	Stanford EE
Ayfer Özgür	Stanford EE
Tsachy Weissman	Stanford EE

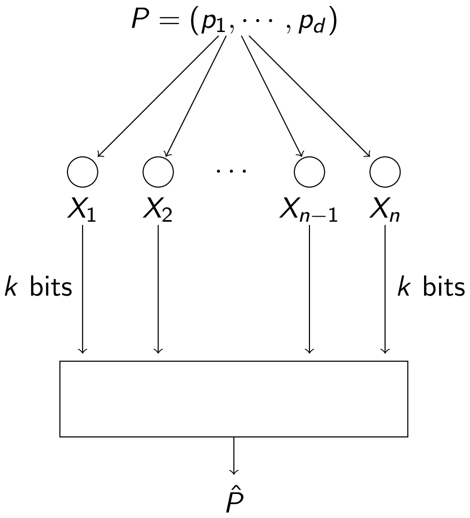
July 16, 2018

Outline

Distributed Distribution Estimation



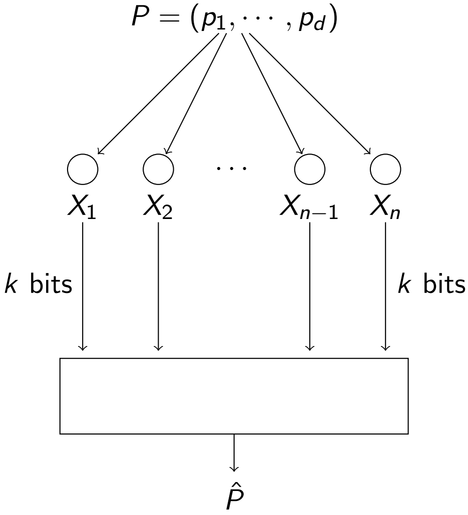
Distributed Distribution Estimation



Parameters:

- n: number of sensors
- k: number of bits
- d: dimensionality

Distributed Distribution Estimation



Parameters:

- n: number of sensors
- k: number of bits
- d: dimensionality

Goal: characterize

$$\inf_{\text{schemes}} \sup_P \mathbb{E}_P \|\hat{P} - P\|_1$$

Main Results

Theorem

The minimax ℓ_1 risk for distributed distribution estimation is

$$\inf_{schemes} \sup_{P} \mathbb{E}\|\hat{P} - P\|_1 \asymp \sqrt{\frac{d}{n}} \cdot \left(\sqrt{\frac{d}{2^k}} \vee 1 \right).$$

Main Results

Theorem

The minimax ℓ_1 risk for distributed distribution estimation is

$$\inf_{schemes} \sup_{P} \mathbb{E}\|\hat{P} - P\|_1 \asymp \sqrt{\frac{d}{n}} \cdot \left(\sqrt{\frac{d}{2^k}} \vee 1 \right).$$

Implications:

- require $k \geq \log_2 d - O(1)$ to achieve centralized performance
- $\frac{d}{2^k}$ distributed sensors $\Leftrightarrow$ 1 centralized sensor

Related Works

Gaussian location model (and its variants):

- lots of works: Duchi et al.'13, Zhang et al.'13, Shamir'14, Garg et al.'14, Braverman et al.'16
- $\frac{d}{k}$ distributed sensors $\Leftrightarrow$ 1 centralized sensor
- tool: strong data processing inequality

Related Works

Gaussian location model (and its variants):

- lots of works: Duchi et al.'13, Zhang et al.'13, Shamir'14, Garg et al.'14, Braverman et al.'16
- $\frac{d}{k}$ distributed sensors $\Leftrightarrow$ 1 centralized sensor
- tool: strong data processing inequality

Discrete distribution estimation:

- require $\Omega(n \log d)$ bits in total to achieve centralized performance (Diakonikolas et al.'17)
- minimax risk for $k \ll \log d$ is missing

Achievability: Grouping Idea

Split $\{1, 2, \cdots, d\}$ into groups:

$$\underbrace{1, 2, \cdots, 2^k - 1}_{G_1}, \underbrace{2^k, 2^k + 1, \cdots, 2(2^k - 1)}_{G_2}, \cdots, \underbrace{d - 2^k + 2, \cdots, d}_{G_m}$$

Achievability: Grouping Idea

Split $\{1, 2, \cdots, d\}$ into groups:

$$\underbrace{1, 2, \cdots, 2^k - 1}_{G_1}, \underbrace{2^k, 2^k + 1, \cdots, 2(2^k - 1)}_{G_2}, \cdots, \underbrace{d - 2^k + 2, \cdots, d}_{G_m}$$

- protocol: each sensor is responsible for one group

Achievability: Grouping Idea

Split $\{1, 2, \cdots, d\}$ into groups:

$$\underbrace{1, 2, \cdots, 2^k - 1}_{G_1}, \underbrace{2^k, 2^k + 1, \cdots, 2(2^k - 1)}_{G_2}, \cdots, \underbrace{d - 2^k + 2, \cdots, d}_{G_m}$$

- protocol: each sensor is responsible for one group
- estimator $\hat{P}$: empirical distribution within each group

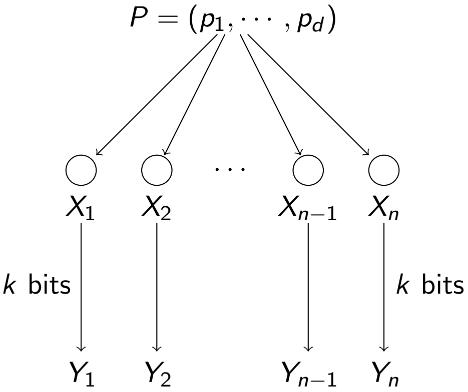
Achievability: Grouping Idea

Split $\{1, 2, \cdots, d\}$ into groups:

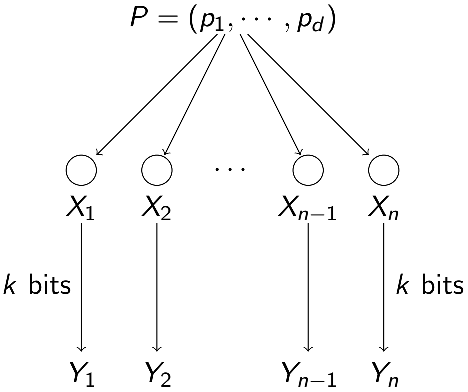
$$\underbrace{1, 2, \cdots, 2^k - 1}_{G_1}, \underbrace{2^k, 2^k + 1, \cdots, 2(2^k - 1)}_{G_2}, \cdots, \underbrace{d - 2^k + 2, \cdots, d}_{G_m}$$

- protocol: each sensor is responsible for one group
- estimator $\hat{P}$: empirical distribution within each group
- n distributed sensors $\Rightarrow \frac{n}{m} \asymp \frac{n2^k}{d}$ centralized sensors

Characterizing all Schemes



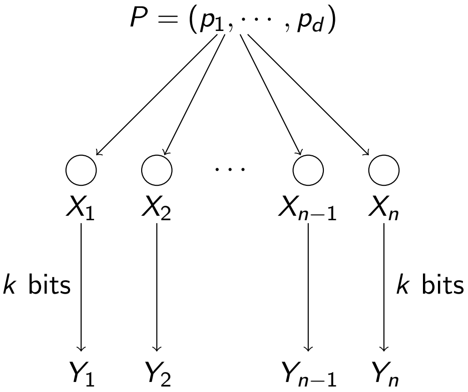
Characterizing all Schemes



For any $i \in [n], s \in [2^k]$:

- $\mathbb{P}(Y_i = s | X_i) \triangleq a_{i,s}(X_i)$

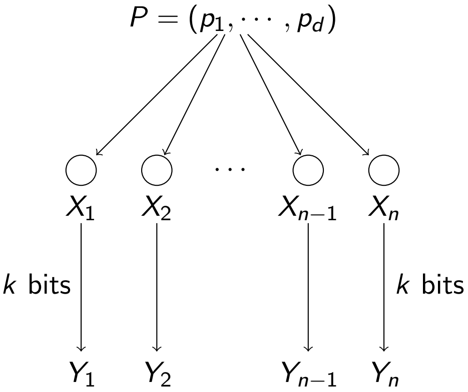
Characterizing all Schemes



For any $i \in [n], s \in [2^k]$:

- $\mathbb{P}(Y_i = s|X_i) \triangleq a_{i,s}(X_i)$
- $\mathbb{P}(Y_i = s) = \mathbb{E}_P a_{i,s}(X_i)$

Characterizing all Schemes



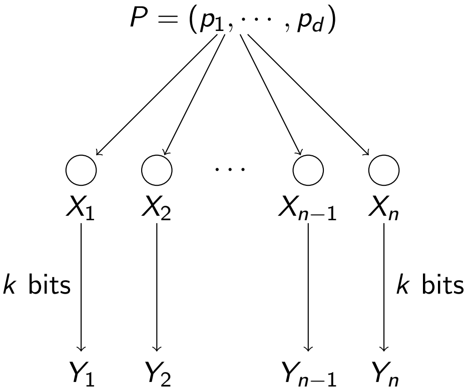
For any $i \in [n], s \in [2^k]$:

- $\mathbb{P}(Y_i = s|X_i) \triangleq a_{i,s}(X_i)$
- $\mathbb{P}(Y_i = s) = \mathbb{E}_P a_{i,s}(X_i)$

Feasible schemes:

- $a_{i,s} \in [0, 1]$

Characterizing all Schemes



For any $i \in [n], s \in [2^k]$:

- $\mathbb{P}(Y_i = s | X_i) \triangleq a_{i,s}(X_i)$
- $\mathbb{P}(Y_i = s) = \mathbb{E}_P a_{i,s}(X_i)$

Feasible schemes:

- $a_{i,s} \in [0, 1]$
- $\sum_s a_{i,s} \equiv 1$

Proof of Lower Bound

Paninski's construction:

- $U \sim \mathrm{Unif}(\{\pm 1\}^{\frac{d}{2}})$

Proof of Lower Bound

Paninski's construction:

- $U \sim \text{Unif}(\{\pm 1\}^{\frac{d}{2}})$
- $X \sim P_U = (\frac{1}{d} + \delta U_1, \frac{1}{d} - \delta U_1, \cdots, \frac{1}{d} + \delta U_{d/2}, \frac{1}{d} - \delta U_{d/2})$

Proof of Lower Bound

Paninski's construction:

- $U \sim \text{Unif}(\{\pm 1\}^{\frac{d}{2}})$
- $X \sim P_U = (\frac{1}{d} + \delta U_1, \frac{1}{d} - \delta U_1, \cdots, \frac{1}{d} + \delta U_{d/2}, \frac{1}{d} - \delta U_{d/2})$
- Y generated by X based on previous scheme

Proof of Lower Bound

Paninski's construction:

- $U \sim \text{Unif}(\{\pm 1\}^{\frac{d}{2}})$
- $X \sim P_U = (\frac{1}{d} + \delta U_1, \frac{1}{d} - \delta U_1, \cdots, \frac{1}{d} + \delta U_{d/2}, \frac{1}{d} - \delta U_{d/2})$
- Y generated by X based on previous scheme

Fano's inequality for $U - X - Y$:

$$\sup_P \mathbb{E}_P \|\hat{P} - P\|_1 \geq \frac{d\delta}{8}\left(1 - \frac{I(U;Y) + \ln 2}{d/8}\right)$$

Upper Bound of $I(U;Y)$

$$I(U;Y) \leq \sum_{i=1}^{n} I(U;Y_i)$$

Upper Bound of $I(U;Y)$

$$I(U;Y) \le \sum_{i=1}^{n} I(U;Y_i)$$
$$\le \sum_{i=1}^{n} \mathbb{E}_U D(P_{Y_i|U} \| P_{Y_i|U=\mathbf{0}})$$

Upper Bound of $I(U;Y)$

$$\begin{aligned} I(U;Y) &\le \sum_{i=1}^{n} I(U;Y_i) \\ &\le \sum_{i=1}^{n} \mathbb{E}_U D(P_{Y_i|U} \| P_{Y_i|U=\mathbf{0}}) \\ &\le \sum_{i=1}^{n} \mathbb{E}_U \chi^2(P_{Y_i|U} \| P_{Y_i|U=\mathbf{0}}) \end{aligned}$$

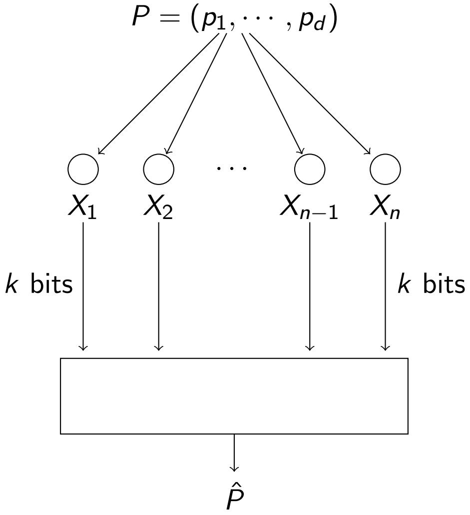
Upper Bound of $I(U;Y)$

$$
\begin{aligned}
I(U;Y) &\le \sum_{i=1}^{n} I(U;Y_i) \\
&\le \sum_{i=1}^{n} \mathbb{E}_U D(P_{Y_i|U} \| P_{Y_i|U=\mathbf{0}}) \\
&\le \sum_{i=1}^{n} \mathbb{E}_U \chi^2(P_{Y_i|U} \| P_{Y_i|U=\mathbf{0}}) \\
&= \sum_{i=1}^{n} \sum_{s=1}^{2^k} \mathbb{E}_U \frac{(\mathbb{E}_{P_U} a_{i,s}(X) - \mathbb{E}_{P_\mathbf{0}} a_{i,s}(X))^2}{\mathbb{E}_{P_\mathbf{0}} a_{i,s}(X)}
\end{aligned}
$$

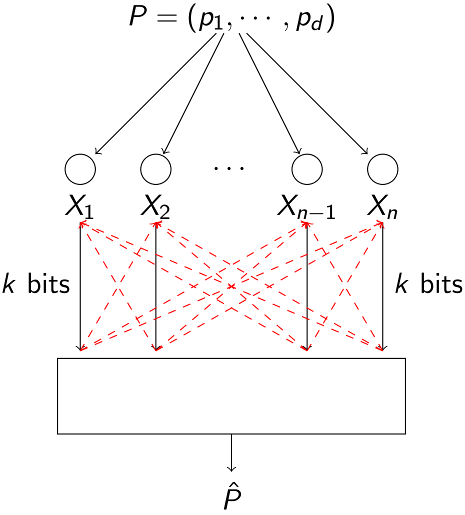
Upper Bound of $I(U; Y)$

$$
\begin{aligned}
I(U;Y) &\le \sum_{i=1}^{n} I(U;Y_i) \\
&\le \sum_{i=1}^{n} \mathbb{E}_U D(P_{Y_i|U} \| P_{Y_i|U=\mathbf{0}}) \\
&\le \sum_{i=1}^{n} \mathbb{E}_U \chi^2(P_{Y_i|U} \| P_{Y_i|U=\mathbf{0}}) \\
&= \sum_{i=1}^{n} \sum_{s=1}^{2^k} \mathbb{E}_U \frac{(\mathbb{E}_{P_U} a_{i,s}(X) - \mathbb{E}_{P_\mathbf{0}} a_{i,s}(X))^2}{\mathbb{E}_{P_\mathbf{0}} a_{i,s}(X)} \\
&\le n2^k \cdot 2\delta^2
\end{aligned}
$$

Blackboard Communication Protocol

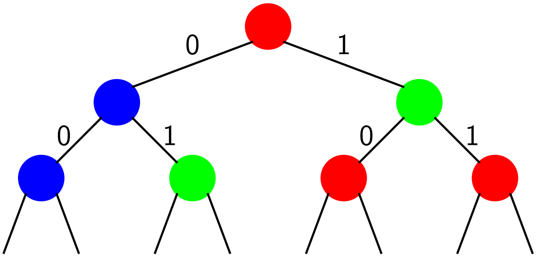


Blackboard Communication Protocol



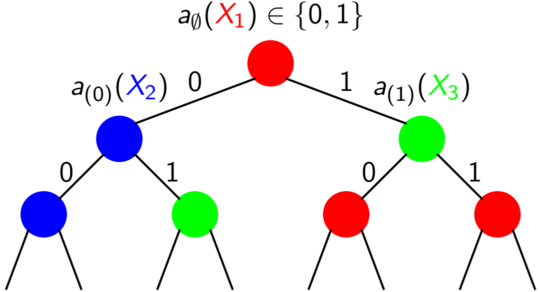
Blackboard Communication Protocol (Cont'd)

Red - Sensor 1, Blue - Sensor 2, Green - Sensor 3



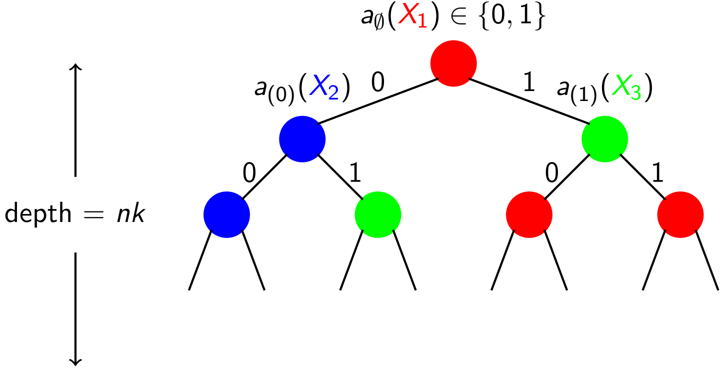
Blackboard Communication Protocol (Cont'd)

Red - Sensor 1, Blue - Sensor 2, Green - Sensor 3



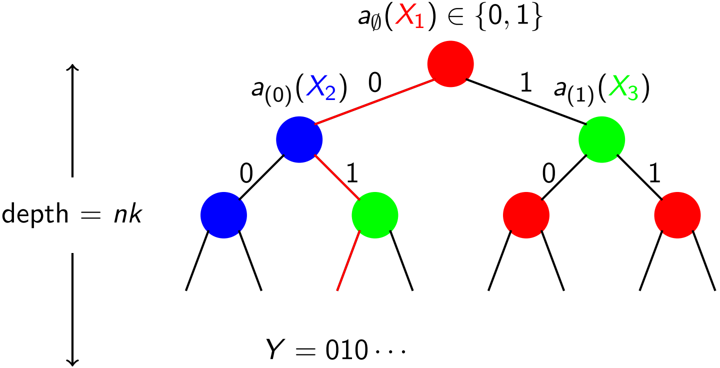
Blackboard Communication Protocol (Cont'd)

Red - Sensor 1, Blue - Sensor 2, Green - Sensor 3



Blackboard Communication Protocol (Cont'd)

Red - Sensor 1, Blue - Sensor 2, Green - Sensor 3



Nonparametric Density Estimation

Let $H^s[0,1]$ be the class of all s-Lipschitz probability densities supported on $[0,1]$, where $0 < s \le 1$.

Theorem

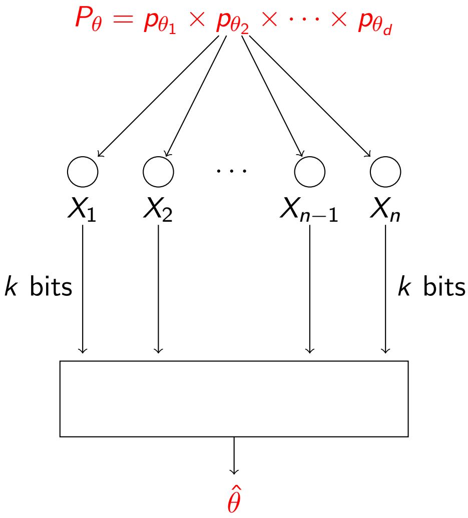
Under k-bit communication constraints,

$$\inf_{\text{schemes}} \sup_{f \in H^s[0,1]} \mathbb{E}_f \|\hat{f} - f\|_1 \asymp (n \cdot 2^k)^{-\frac{s}{2(s+1)}} \vee n^{-\frac{s}{2s+1}}.$$

Corollary

Centralized performance is achieved iff $k \ge \frac{1}{2s+1} \log_2 n - O(1)$.

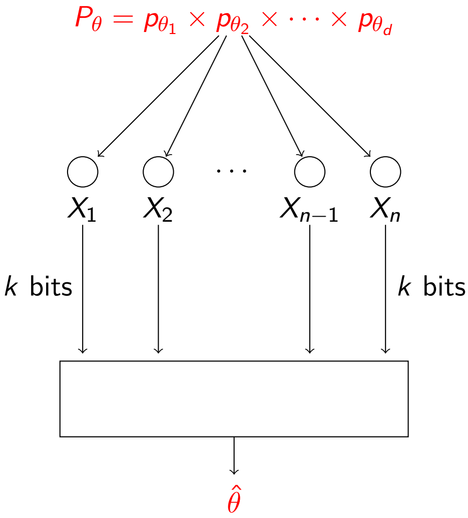
General Distributed Estimation



Parameters:

- n: number of sensors
- k: number of bits
- d: dimensionality

General Distributed Estimation



Parameters:

- n: number of sensors
- k: number of bits
- d: dimensionality

Goal: characterize

$$\inf_{\text{schemes}} \sup_{\theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2$$

General Lower Bounds

Theorem (Han, Özgür, Weissman'18)

Fix any θ_0*, let* $S(X)$ *be the score function of* (p_θ) *around* $\theta = \theta_0$*:*

$$S(X) = \frac{\partial}{\partial\theta} \log p_\theta(X)\Big|_{\theta=\theta_0}.$$

Assuming mild regularity conditions,

$$\inf_{\text{schemes}} \sup_{\theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \gtrsim \frac{d}{n\text{Var}(S(X))} \vee \frac{d^2}{n2^k\text{Var}(S(X))} \vee \frac{d^2}{nk\|S(X)\|_{\psi_2}^2}.$$

Comparison with SDPI

Strong data processing inequality (SDPI):

$$I(U;Y) \le \gamma^*(U,X) I(X;Y)$$

Comparison with SDPI

Strong data processing inequality (SDPI):

$$I(U;Y) \leq \gamma^*(U,X) I(X;Y)$$

- $U - X$ determined by statistical model $X \sim P_U$, $X - Y$ subject to communication constraints

Comparison with SDPI

Strong data processing inequality (SDPI):

$$I(U;Y) \le \gamma^*(U,X) I(X;Y)$$

- $U - X$ determined by statistical model $X \sim P_U$, $X - Y$ subject to communication constraints
- leads to tight results in Gaussian location model

Comparison with SDPI

Strong data processing inequality (SDPI):

$$I(U;Y) \le \gamma^*(U,X)I(X;Y)$$

- $U - X$ determined by statistical model $X \sim P_U$, $X - Y$ subject to communication constraints
- leads to tight results in Gaussian location model
- can only result in linear dependence on k, while our dependence is exponential

Comparison with SDPI

Strong data processing inequality (SDPI):

$$I(U;Y) \le \gamma^*(U,X) I(X;Y)$$

- $U - X$ determined by statistical model $X \sim P_U$, $X - Y$ subject to communication constraints
- leads to tight results in Gaussian location model
- can only result in linear dependence on k, while our dependence is exponential
- unclear operational meaning

Geometric Inequalities

Let $X = (X_1, \cdots, X_d)$ be a random vector with independent and zero-mean entries.

Geometric Inequalities

Let $X = (X_1, \cdots, X_d)$ be a random vector with independent and zero-mean entries.

Geometric Inequalities (Han, Özgür, Weissman'18)

- If $\mathrm{Var}(X_i) \le \sigma^2$ for any i:

$$\|\mathbb{E}[X|A]\|_2^2 \le \sigma^2 \cdot \frac{1 - \mathbb{P}(A)}{\mathbb{P}(A)}, \qquad \forall A \subset \mathbb{R}^d$$

- If each X_i is σ^2-sub-Gaussian:

$$\|\mathbb{E}[X|A]\|_2^2 \le C\sigma^2 \cdot \log \frac{1}{\mathbb{P}(A)}, \qquad \forall A \subset \mathbb{R}^d$$

Geometric Inequalities

Let $X = (X_1, \cdots, X_d)$ be a random vector with independent and zero-mean entries.

Geometric Inequalities (Han, Özgür, Weissman'18)

- If $\mathrm{Var}(X_i) \le \sigma^2$ for any i:

$$\|\mathbb{E}[X|A]\|_2^2 \le \sigma^2 \cdot \frac{1-\mathbb{P}(A)}{\mathbb{P}(A)}, \qquad \forall A \subset \mathbb{R}^d$$

- If each X_i is σ^2-sub-Gaussian:

$$\|\mathbb{E}[X|A]\|_2^2 \le C\sigma^2 \cdot \log \frac{1}{\mathbb{P}(A)}, \qquad \forall A \subset \mathbb{R}^d$$

Thank you!