Cookbook: Lower Bounds for Statistical Inference in Distributed and Constrained Settings

Jayadev Acharya, Clement Canonne, Himanshu Tyagi

FOCS 2020

Part II: Lower bounds for learning

Two recalls

$$\mathbf{p}(Y = y) = \sum_x \mathbf{p}(x) \cdot W(y|x) = \mathbb{E}_{X \sim \mathbf{p}}[W(y|X)]$$

For distributions $\mathbf{p}, \mathbf{q}$

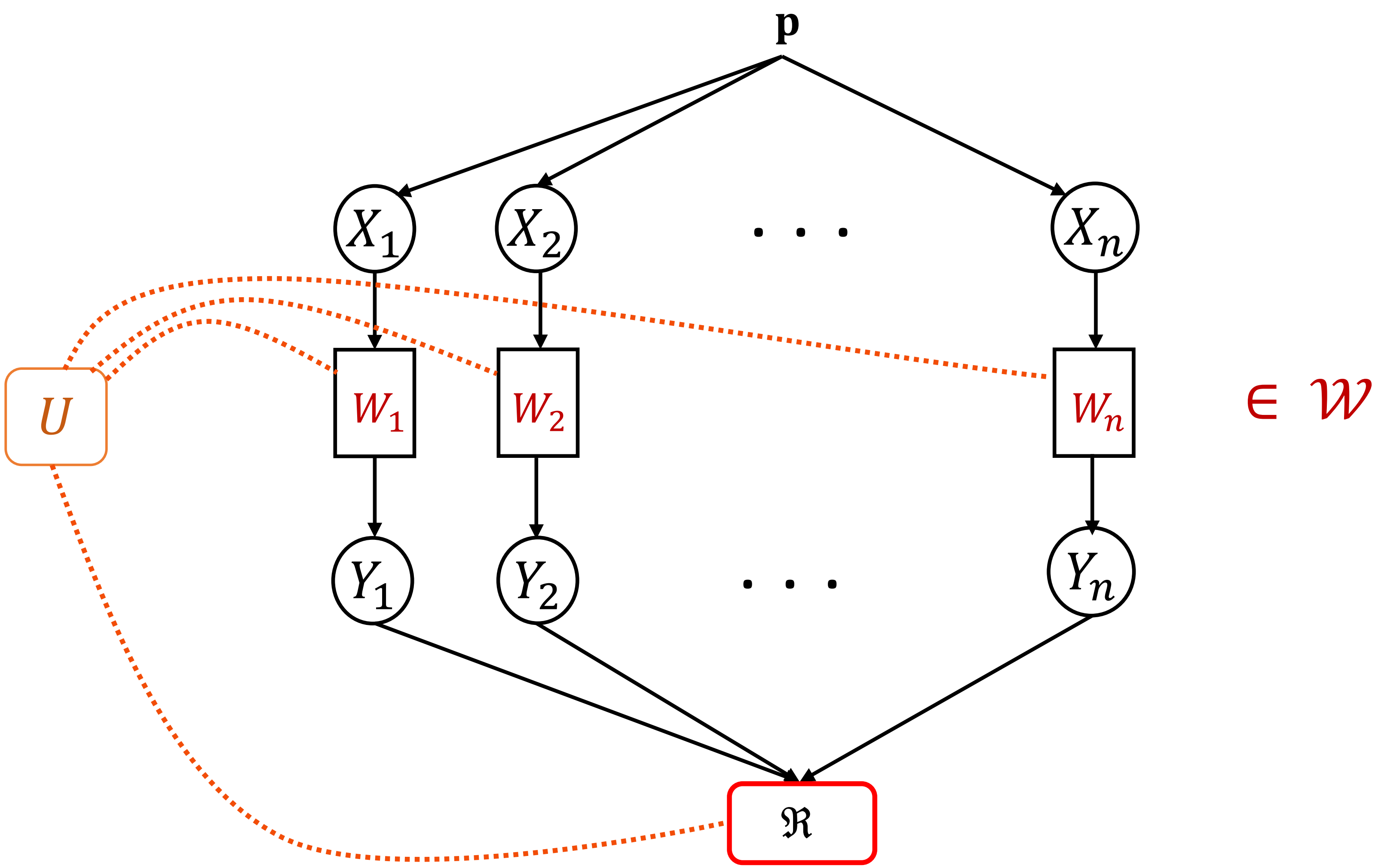
$\mathrm{D}(\mathbf{p} \parallel \mathbf{q}) \coloneqq \sum_x \mathbf{p}(x) \log[\mathbf{p}(x)/\mathbf{q}(x)]$

$\mathrm{d}_{\chi^2}(\mathbf{p}, \mathbf{q}) \coloneqq \sum_x \big(\mathbf{p}(x) - \mathbf{q}(x)\big)^2 / \mathbf{q}(x)$

$$\mathrm{D}(\mathbf{p} \parallel \mathbf{q}) \leq \mathrm{d}_{\chi^2}(\mathbf{p}, \mathbf{q})$$

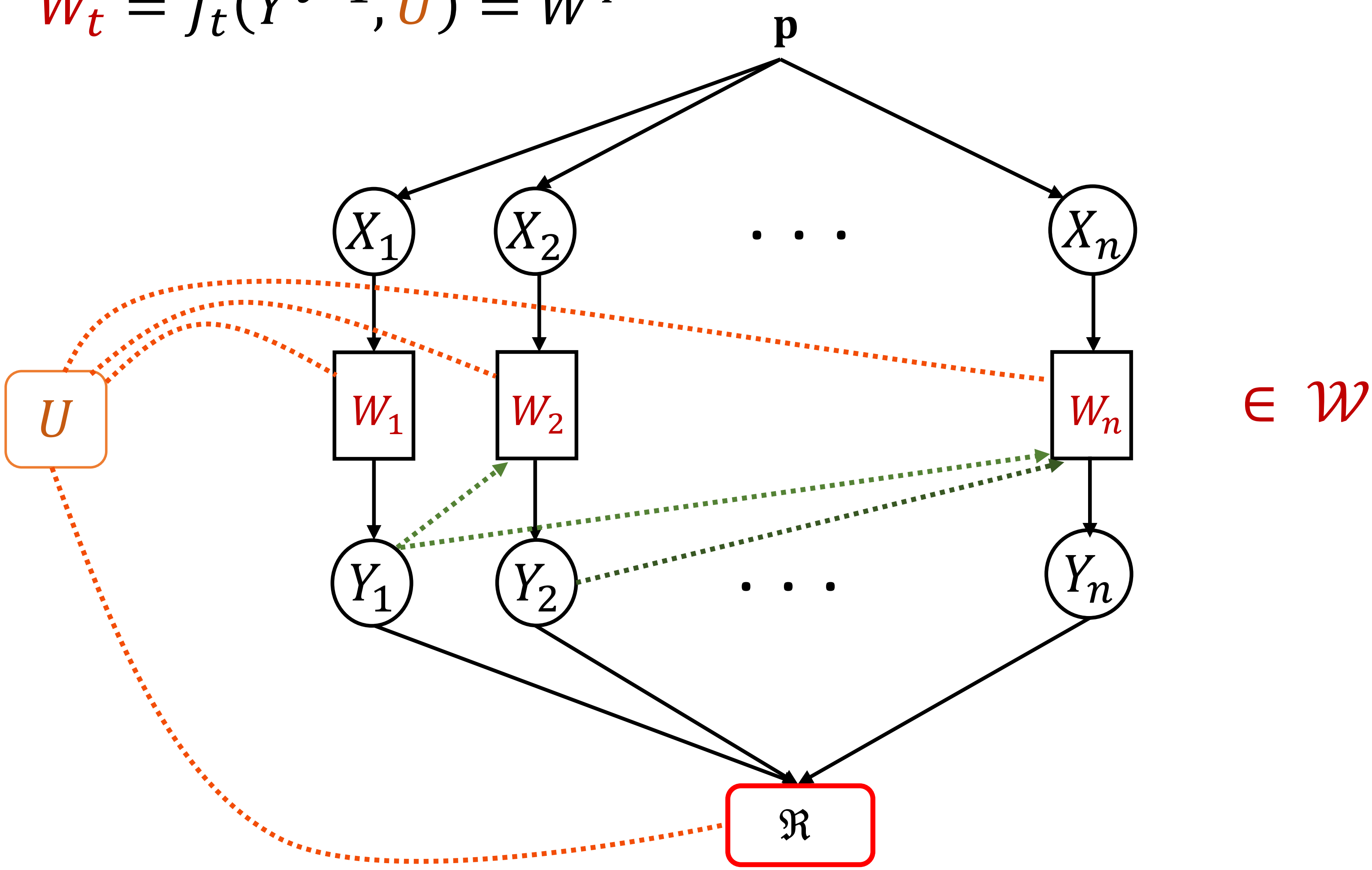
SMP protocols

$W_t = f_t(U)$



(Sequentially) interactive protocols

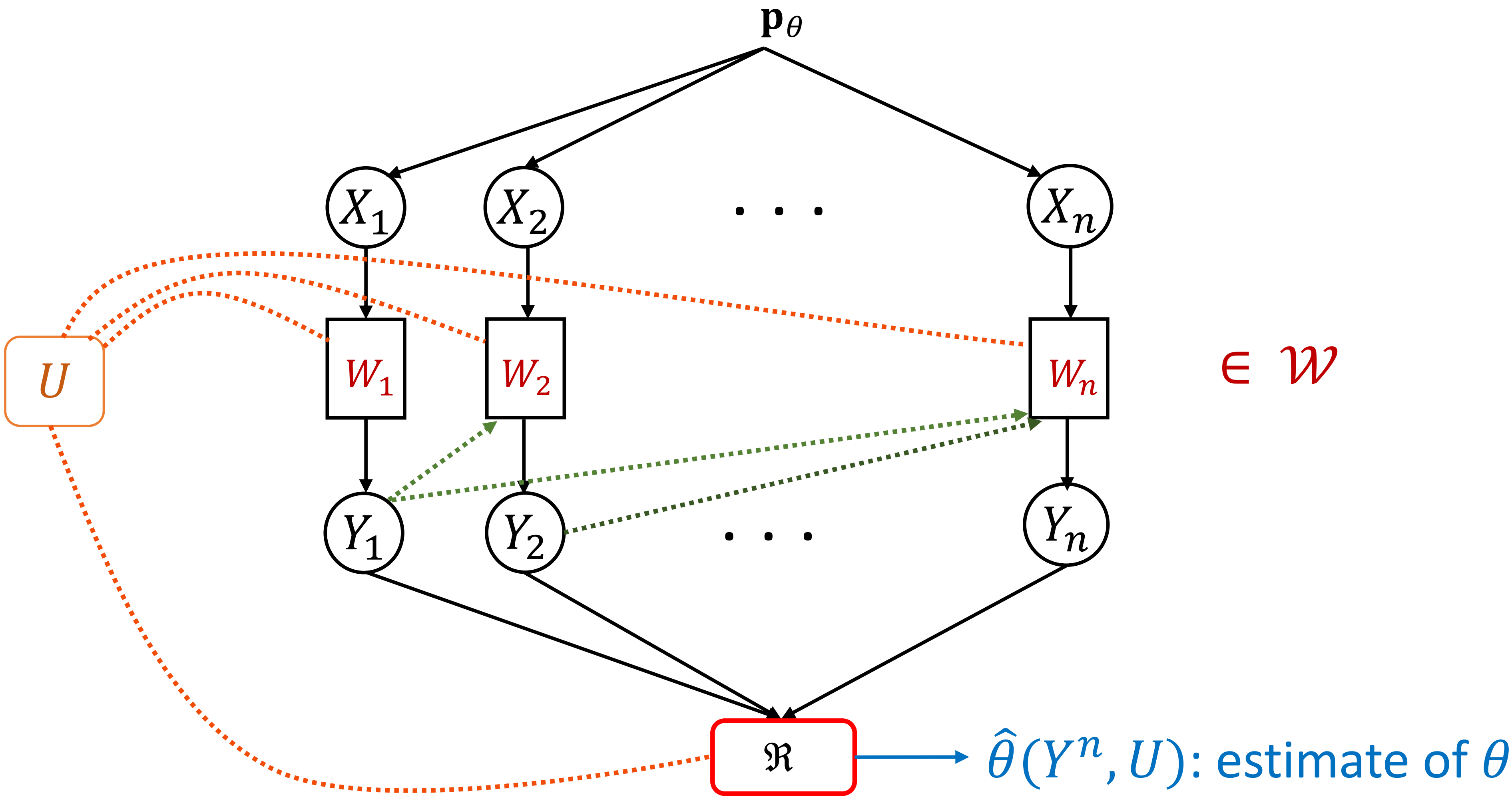
$$W_t = f_t(Y^{t-1}, U) = W^{Y^{t-1}}$$



Estimation

$\Theta \subseteq \mathbb{R}^d$: space of parameters

$\mathcal{P}_\Theta = \{\mathbf{p}_\theta : \theta \in \Theta\}$, distributions over $\mathcal{X}$ indexed by Θ



Objective: ℓ_p estimation of θ

Given $\varepsilon > 0, p \geq 1$

$$\sup_{\theta} \mathbb{E}_{\mathbf{p}_\theta}\left[\ell_p\left(\hat{\theta}(Y^n, U), \theta\right)^p\right]^{1/p} \leq \varepsilon$$

where

$$\ell_p(u, v)^p = \sum |u_i - v_i|^p$$

Sample complexity: Smallest n for which such a $\hat{\theta}$ exists

Example: Discrete distributions (Δ_k)

- Parameter space

$$\Theta = \{\theta \in [0,1]^k \colon \sum \theta_i = 1\}$$

- Underlying domain

$$\mathcal{X} = \{1, \dots, k\}$$

- For $x \in \mathcal{X}$

$$\mathbf{p}_\theta(x) = \theta_x$$

θ denotes the probability mass function of $\mathbf{p}_\theta$

Example: Product Bernoulli ($\mathcal{B}_d$)

- Parameter space

$$\Theta = \{\theta \in [-1,1]^d\}$$

- Underlying domain

$$\mathcal{X} = \{-1, 1\}^d$$

- For $x = (x_1, \ldots, x_d) \in \mathcal{X}$

$$\mathbf{p}_\theta(x) = \prod_i \mathbf{p}_{\theta_i}(x_i)$$

$$\mathbf{p}_{\theta_i}(1) = \frac{1+\theta_i}{2}, \mathbf{p}_{\theta_i}(-1) = \frac{1-\theta_i}{2}$$

Therefore, $\mathbb{E}_\theta[x_i] = \theta_i$

θ is the distribution mean

Example: Gaussians ($\mathcal{G}_d$)

- Parameter space

$$\Theta = \{\theta \in [-1,1]^d\}$$

- Underlying domain

$$\mathcal{X} = \mathbb{R}^d$$

$$\mathbf{p}_\theta = N(\theta, \mathbb{I})$$

θ is the distribution mean

Estimation Tasks in this tutorial

Discrete distribution estimation (Δ_k)

Product Bernoulli mean estimation $(\mathcal{B}_d)$

Gaussian mean estimation $(\mathcal{G}_d)$

- Results qualitatively same as $\mathcal{B}_d$ (details can be messy)

Aim of the tutorial

Provide general methods

Discrete distribution estimation (Δ_k)

$p = 1$: ℓ_1 distance

Product Bernoulli estimation $(\mathcal{B}_d)$

$p = 2$: Euclidean distance

Information constraints

1. Communication constraints

ℓ-bit communication constraints

$$\mathcal{W}_\ell = \{W : \mathcal{X} \to \{0,1\}^\ell\}$$

2. Local differential privacy constraints

ϱ-LDP channels

$$\mathcal{W}_\varrho = \left\{W : \max_{\{x,x' \in \mathcal{X}, y \in \mathcal{Y}\}} \frac{W(y|x)}{W(y|x')} \leq e^\varrho\right\}$$

Sample complexity for the applications

Problem		
Δ_k, ℓ_1	$\frac{k}{\varepsilon^2} \cdot \frac{k}{\min\{2^\ell, k\}}$	$\frac{k}{\varepsilon^2} \cdot \frac{k}{\varrho^2}$
$\mathcal{B}_d, \ell_2$	$\frac{d}{\varepsilon^2} \cdot \frac{d}{\min\{\ell, d\}}$	$\frac{d}{\varepsilon^2} \cdot \frac{d}{\varrho^2}$
$\mathcal{G}_d, \ell_2$	$\frac{d}{\varepsilon^2} \cdot \frac{d}{\min\{\ell, d\}}$	$\frac{d}{\varepsilon^2} \cdot \frac{d}{\varrho^2}$

Centralized sample complexity × blow-up due to constraints

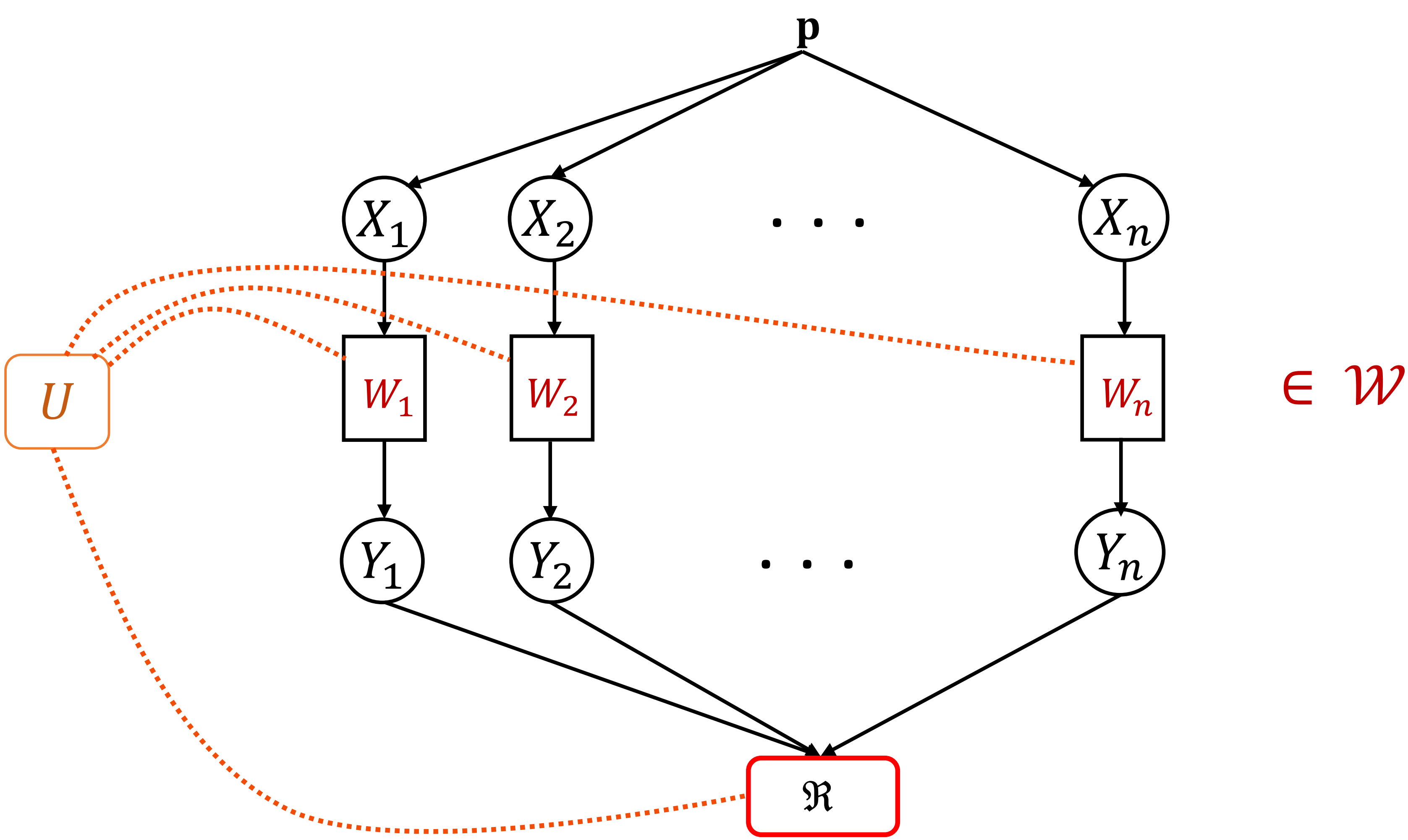
Lower bounds for SMP protocols

Reference	Distribution	Constraints
[GMN14, ZDJW14]	$\mathcal{G}_d$	
[DJW17]	$\Delta_k, \mathcal{G}_d$	
[HOW18, HMOW18]	$\Delta_k, \mathcal{B}_d, \mathcal{G}_d$	
[ACT19]	Δ_k	 

SMP vs interactive

SMP protocols:

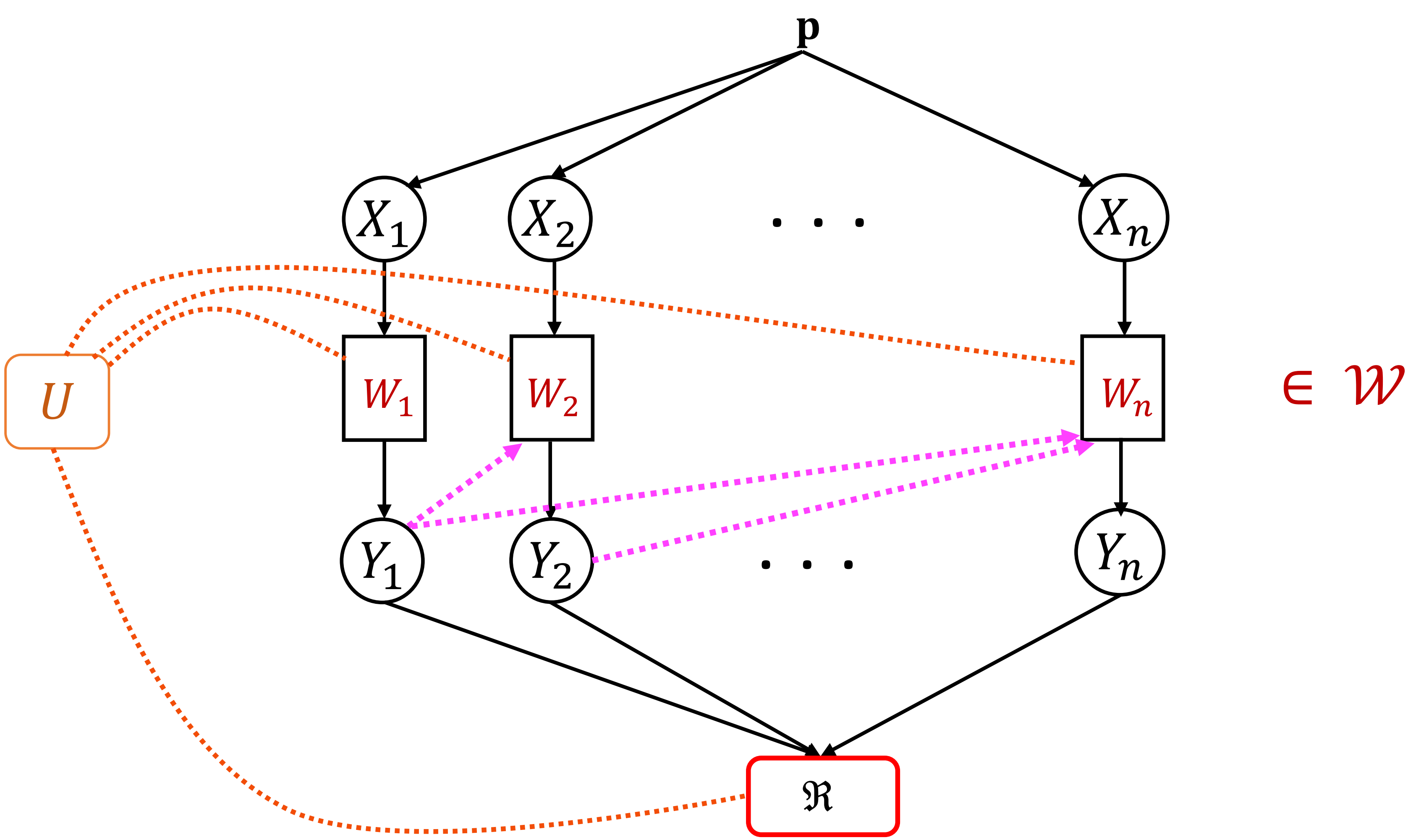
Fixed U lower bounds for **independent** channels



SMP vs interactive

Interactive protocols:

Several dependencies possible, harder to handle



The plan for this hour

Three methods to establish interactive lower bounds

1. Classic Cramer-Rao/van Trees inequality [BHO19, BCO20]
    - Unified results for $\Delta_k, \mathcal{B}_d, \mathcal{G}_d$
    - Results hold for ℓ_2 loss
2. Strong Data Processing + Assouad's method [BGMNW16, DR19]
    - Lower bounds for $\mathcal{B}_d, \mathcal{G}_d$ under ℓ_2 loss
    - Naturally extends to other ℓ_p loss functions
3. Chi-squared contractions + Assouad's method [ACLST20, ACT20]
    - Unified bounds for $\Delta_k, \mathcal{B}_d, \mathcal{G}_d$
    - Works under ℓ_p for $p \geq 1$

Plan for each part

1. General methodology
2. Application

Proving lower bounds in statistical inference

Recall the goal

$$\sup_\theta \mathbb{E}_{\mathbf{p}_\theta}\left[\ell_p\left(\hat{\theta}(Y^n,U),\theta\right)^p\right]^{1/p} \leq \varepsilon$$

1. Prior π is a distribution over Θ $\quad \pi \to \Theta \to X \to Y$
2. Show that for any $\hat{\theta}$

$$\mathbb{E}_\pi\left[\mathbb{E}_{\mathbf{p}_\theta}\left[\ell_p\left(\hat{\theta}(Y^n,U),\theta\right)^p\right]^{1/p}\right] > \varepsilon$$

$\mathbb{E} \leq \max \Rightarrow$ a lower bound on n

All lower bounds will involve choosing a π at some point

Observation: Given π, suffices to prove lower bound for a fixed $U = u$ and denote (Y^n, u) by Y^n

Proving lower bounds in statistical inference

1. Design a prior π over Θ

2. Show that for any $\hat{\theta}$

$$\mathbb{E}_{\pi}\left[\mathbb{E}_{\mathbf{p}_{\theta}}\left[\ell_p\left(\hat{\theta}(Y^n),\theta\right)^p\right]^{1/p}\right] > \varepsilon$$

1. CR/van Trees inequality

Outline for method 1

- Univariate Cramer Rao (CR) bound
  - Bounds error in terms of Fisher information
- High-dimensional CR
- Bayesian CR bound/van Trees inequality
- Error bounds under information constraints
- Application

Cramer Rao bound (for $d = 1$)

$$\Theta \subseteq \mathbb{R}$$

Fisher information:

$$\mathrm{I}_X(\theta) \coloneqq \mathbb{E}_{X\sim \mathbf{p}_\theta}\left[\left(\frac{\partial}{\partial\theta}\ln \mathbf{p}_\theta(X)\right)^2\right]$$

$\hat{\theta}$: any unbiased estimator of θ, i.e., $\mathbb{E}\left[\hat{\theta}\right] = \theta$

Theorem.

$$\mathrm{Var}\left(\hat{\theta}\right) \geq \frac{1}{\mathrm{I}_X(\theta)}$$

Example: Bernoulli mean estimation

$\mathcal{X} = \{0,1\}, \mathbf{p}_\theta(1) = \theta, X_1, \dots, X_n \sim_{iid} \mathbf{p}_\theta$

$$\mathrm{I}_{X_1}(\theta) = \frac{1}{\theta(1-\theta)}$$

By additivity of Fisher information

$$\mathrm{I}_{X^n}(\theta) = n \cdot \mathrm{I}_{X_1}(\theta) = n \cdot \frac{1}{\theta(1-\theta)}$$

For any unbiased $\hat{\theta}(X^n)$

$$\mathrm{Var}\big(\hat{\theta}\big) \geq \frac{\theta(1-\theta)}{n}$$

Achieved by

$$\hat{\theta} = (X_1 + \cdots + X_n)/n$$

Multivariate Cramer Rao bound

$$\Theta \subseteq \mathbb{R}^d$$

$d \times d$ Fisher information matrix:

$$\left(\mathrm{I}_X(\theta)\right)_{i_1,i_2} := \mathbb{E}_X\left[\frac{\partial^2}{\partial\theta_{i_1}\partial\theta_{i_2}} \ln \mathbf{p}_\theta(X)\right]$$

Theorem (CR). For any unbiased $\hat{\theta}(X)$

$$\mathrm{Cov}\left(\hat{\theta}(X)\right) \geq \left(\mathrm{I}_X(\theta)\right)^{-1}$$

Corollary.

$$\ell_2\left(\hat{\theta}(X), \theta\right)^2 = \sum\left(\hat{\theta}_i - \theta_i\right)^2 \geq \mathrm{Tr}\left(\left(\mathrm{I}_X(\theta)\right)^{-1}\right) \geq \frac{d^2}{\mathrm{Tr}\left(\mathrm{I}_X(\theta)\right)}$$

Last step uses $\mathrm{Tr}(A) \cdot \mathrm{Tr}(A^{-1}) \geq d^2$ for p.s.d. A

van Trees inequality [vT68, GL95]

Unbiasedness a strong assumption

van Trees inequality: a Bayesian CR bound

- π: a prior distribution over Θ
- Lower bound for error under π

van Trees inequality [vT68, GL95]

Let $\pi := \pi_1 \times \cdots \times \pi_d$ be a product prior over $\Theta \subseteq \mathbb{R}^d$, i.e.,

$$\pi(\theta) = \pi_1(\theta_1) \ldots \pi_d(\theta_d)$$

Theorem. Under some mild assumptions

$$\mathbb{E}_\pi \mathbb{E}_{X \sim \mathbf{p}_\theta} \left[\ell_2\left(\hat{\theta}(X), \theta\right)^2 \right] \geq \frac{d^2}{\mathbb{E}_\pi\left[\mathrm{Tr}\left(\mathrm{I}_X(\theta)\right)\right] + \mathrm{I}(\pi)},$$

where $\mathrm{I}(\pi) = \mathrm{I}(\pi_1) + \cdots + \mathrm{I}(\pi_d)$

$$\mathrm{I}(\pi_i) := \mathbb{E}_{\pi_i}\left[\left(\frac{\partial}{\partial \theta_i} \ln \pi_i(\theta_i)\right)^2\right]$$

[GL95] R. D. Gill, B. Y. Levit, "Applications of the van Trees inequality: a Bayesian Cramer-Rao bound" *Bernoulli, 1995*

van Trees inequality [vT68, GL95]

Theorem. Under some mild assumptions

$$\mathbb{E}_{\pi}\mathbb{E}_{X\sim\mathbf{p}_{\theta}}\left[\ell_2\left(\hat{\theta}(Y^n),\theta\right)^2\right] \geq \frac{d^2}{\mathbb{E}_{\pi}\left[\mathrm{Tr}\left(\mathrm{I}_{Y^n}(\theta)\right)\right] + \mathrm{I}(\pi)}.$$

Design a π to upper bound

$$\mathbb{E}_{\pi}\left[\mathrm{Tr}\left(\mathrm{I}_{Y^n}(\theta)\right)\right] + \mathrm{I}(\pi)$$

[GL95] R. D. Gill, B. Y. Levit, "Applications of the van Trees inequality: a Bayesian Cramer-Rao bound" *Bernoulli, 1995*

$\mathbb{E}_\pi\left[\mathrm{Tr}\left(\mathrm{I}_{Y^n}(\theta)\right)\right]$ under interactive protocols [BHO19]

Fix θ. By the chain rule of Fisher information,

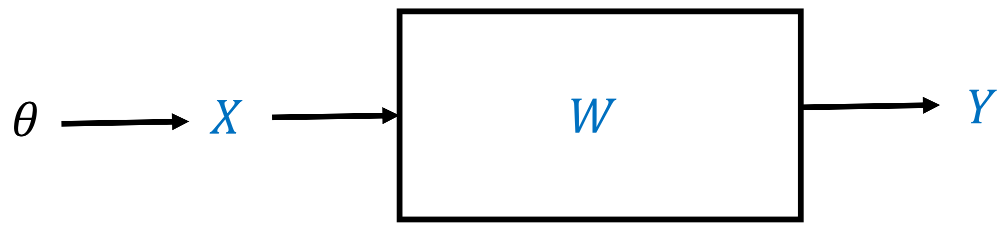
$$\mathrm{Tr}\left(\mathrm{I}_{Y^n}(\theta)\right) = \sum_t \mathbb{E}_{Y^{t-1}}\left[\mathrm{Tr}\left(\mathrm{I}_{Y_t|Y^{t-1}}(\theta)\right)\right]$$

Given θ, X_t indep Y^{t-1}. Using this

$$\mathrm{Tr}\left(\mathrm{I}_{Y^n}(\theta)\right) \le n \cdot \sup_{W \in \mathcal{W}} \mathrm{Tr}\left(\mathrm{I}_Y(\theta)\right)$$

Consider worst θ in the support of π

$$\mathbb{E}_\pi\left[\mathrm{Tr}\left(\mathrm{I}_{Y^n}(\theta)\right)\right] \le n \cdot \sup_{\theta \in \mathrm{supp}\,(\pi)} \sup_{W \in \mathcal{W}} \mathrm{Tr}\left(\mathrm{I}_Y(\theta)\right)$$



$\mathrm{I}(\pi)$

Fact [Borovkov95]**.** Given $A = [a - \Delta, a + \Delta] \subset \Theta$ there exists μ s.t.

$$\mathrm{I}(\mu) = \frac{3.14159265358\ldots^2}{\Delta^2}.$$

This is the smallest possible value.

Choosing $\pi = \mu \times \cdots \times \mu$ (each $\pi_i = \mu$), $\mathrm{I}(\pi) = d \cdot 3.14\ldots^2 / \Delta^2$.

Information-constrained lower bounds

$$\mathbb{E}_\pi\left[\mathrm{Tr}\left(\mathrm{I}_{(Y^n,U)}(\theta)\right)\right] + \mathrm{I}(\pi) \leq n \cdot \sup_{\theta\in A^d} \sup_{W\in\mathcal{W}} \mathrm{Tr}\big(\mathrm{I}_Y(\theta)\big) + \frac{d\cdot 3.15^2}{\Delta^2}$$

Therefore,

$$\varepsilon^2 \geq \frac{d^2}{n \cdot \sup_{\theta\in A^d} \sup_{W\in\mathcal{W}} \mathrm{Tr}\big(\mathrm{I}_Y(\theta)\big) + \frac{d\cdot 3.15^2}{\Delta^2}}$$

Application 1: $\mathcal{B}_d$

$A = [-0.5, 0.5]$

[BHO19] $$\sup_{\theta \in A^d} \sup_{W \in \mathcal{W}_\ell} \mathrm{Tr}\big(\mathrm{I}_Y(\theta)\big) = O(\ell)$$

$$n \geq \frac{d^2}{\varepsilon^2 \cdot \ell}$$

[BCO19] $$\sup_{\theta \in A^d} \sup_{W \in \mathcal{W}_\varrho} \mathrm{Tr}\big(\mathrm{I}_Y(\theta)\big) = O(\varrho^2)$$

$$n \geq \frac{d^2}{\varepsilon^2 \cdot \varrho^2}$$

Application 2: Δ_k under ℓ_2

$A = [\frac{1}{4k}, \frac{1}{3k}]$

 [BHO19] $\sup_{\theta \in A^d} \sup_{W \in \mathcal{W}_\ell} \mathrm{Tr}\big(\mathrm{I}_Y(\theta)\big) = O\big(k \cdot 2^\ell\big)$

$$n \geq \frac{k}{\varepsilon^2 \cdot 2^\ell}$$

 [BCO19] $\sup_{\theta \in A^d} \sup_{W \in \mathcal{W}_\varrho} \mathrm{Tr}\big(\mathrm{I}_Y(\theta)\big) = O(k \cdot \varrho^2)$

$$n \geq \frac{k}{\varepsilon^2 \cdot \varrho^2}$$

Conclusion

- Tight bounds for ℓ_2 estimation
- Works for Δ_k, $\mathcal{B}_d$, and under 
- *Does not yield* ℓ_1 *bounds*

Detour: Assouad's method

Method 2 and 3 use classic Assouad's method

The method

$\mathcal{Z} := \{-1, 1\}^m$ for some m

$\Theta_{\mathcal{Z}} = \{\theta_z : z \in \mathcal{Z}\} \subseteq \Theta$, such that

$$\ell_p(\theta_z, \theta_{z'})^p \geq \frac{d_{\mathrm{ham}}(z, z')}{m} \cdot \varepsilon^p$$

$\theta_z, \theta_{z'}$ are far if z, z' are far

Prior π is the uniform distribution over $\Theta_{\mathcal{Z}}$

- $Z \sim_{uar} \mathcal{Z}$
- $\theta = \theta_Z$

Estimate θ_Z under $\pi \Rightarrow$ Estimate Z in Hamming distance

$\Rightarrow Y^n$ gives information about Z

Assouad's method

Theorem. If

$$\mathbb{E}_{\pi}\left[\mathbb{E}_{\mathbf{p}_{\theta}}\left[\ell_p\left(\hat{\theta}(Y^n), \theta\right)^p\right]^{1/p}\right] \leq \frac{\varepsilon}{10},$$

then

$$\sum_i I(Z_i \wedge Y^n) = \Omega(m).$$

Example: $\mathcal{B}_d$ under ℓ_2

$m = d, \mathcal{Z} = \{-1, 1\}^d$

For $z \in \mathcal{Z}$,

$$\theta_z \coloneqq \frac{2\varepsilon}{\sqrt{d}} \cdot z = \mathbb{E}_{X \sim \mathbf{p}_z}[X]$$

$$\Pr(X_i = 1) = 0.5 + \frac{\varepsilon z_i}{\sqrt{d}}$$

Therefore,

$$\ell_2(\theta_z, \theta_{z'})^2 = d_{\text{ham}}(z, z') \cdot \frac{16\varepsilon^2}{d}$$

$\mathbf{p}_{\theta_z}$ denoted by $\mathbf{p}_z$

Example: Δ_k under ℓ_1 [Paninski08]

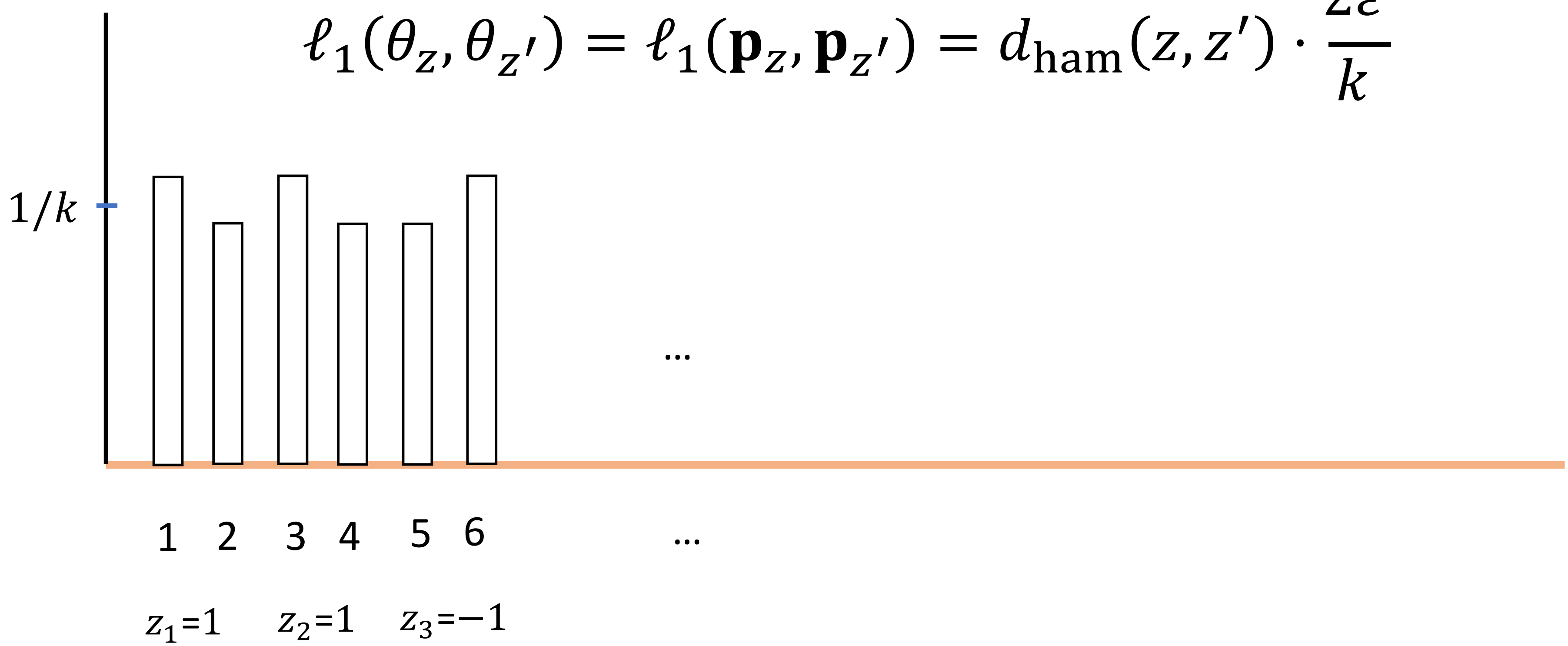
$m = k/2, \mathcal{Z} = \{-1,1\}^{k/2}$

For $j = 1, \ldots, k/2$, let

$$\mathbf{p}_z(2j-1) = \frac{1 + z_j \varepsilon}{k}, \mathbf{p}_z(2j) = \frac{1 - z_j \varepsilon}{k}$$

$\theta_z \in [0,1]^k, \theta_z(j) = \mathbf{p}_z(j)$

$$\ell_1(\theta_z, \theta_{z'}) = \ell_1(\mathbf{p}_z, \mathbf{p}_{z'}) = d_{\text{ham}}(z, z') \cdot \frac{2\varepsilon}{k}$$

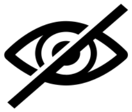


2. SDPI + Assouad

Background

[DJWZ14, GMN14] use SDPI for SMP protocols

[BGMNW16] generalize to interactive protocols for 

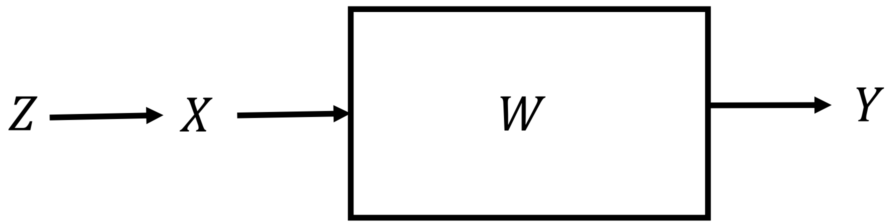
[DR19] 

Outline for method 2

- For $d = 1$:
  - Strong data processing constant
  - Distributed SDPI for interactive protocols
- Extend to $d > 1$ by a direct sum result
- Application

Strong data processing constant $(d = 1)$

$\mathbf{p}_1, \mathbf{p}_{-1}$ two distributions
Let $Z \sim_{uar} \{-1,1\}$, and $X \sim \mathbf{p}_Z$



Guess Z from Y

$\boldsymbol{\beta}(\mathbf{p}_1, \mathbf{p}_{-1})$ be smallest $\boldsymbol{\beta}$ such that for any $Z - X - Y$

$$I(Z \wedge Y) \leq \boldsymbol{\beta} \cdot I(X \wedge Y)$$

Y tells about Z at most $\boldsymbol{\beta}$ fraction of what it tells about X

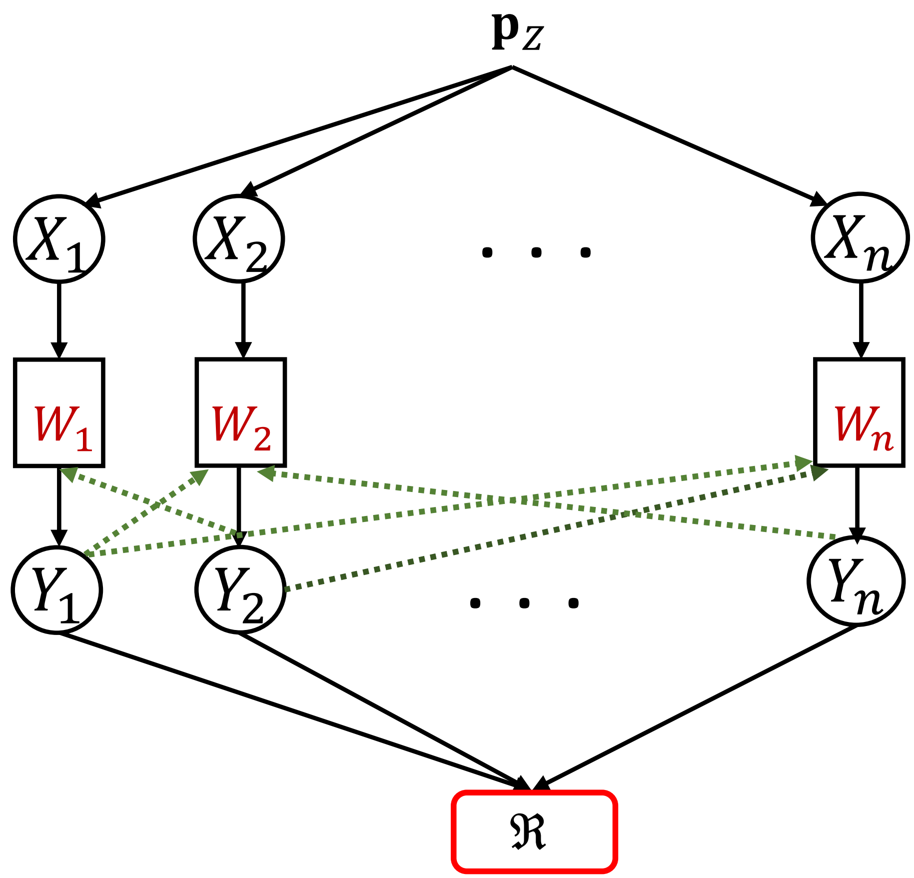
Can be shown:

- $I(X \wedge Y) \leq \ell$ for 
- $I(X \wedge Y) \leq O(\varrho^2)$ for 

A distributed SDPI [BGMNW15]

$Z \sim_{uar} \{-1,1\}$ and $X^n \sim \mathbf{p}_Z$

Guess Z from Y^n



A distributed SDPI [BGMNW15]

Theorem. Suppose $\mathbf{p}_{-1}(x) = \Theta(\mathbf{p}_{+1}(x))$. Then for any *blackboard* protocol

$$I(Z \wedge Y^n) = O(\boldsymbol{\beta} \cdot I(X^n \wedge Y^n)).$$

Y^n tells about Z at most $\boldsymbol{\beta}$ fraction of what it tells about X^n

SMP protocols: follows by tensorization of SDPI [Raginsky14]

Interactive protocols: cut-paste property of Hellinger distance from communication complexity [BYJKS04, Jayaram09]

Example: Bernoulli

$$\mathbf{p}_{\pm 1} = \mathrm{Bern}(0.5 \pm \delta)$$

$$\beta(\mathbf{p}_{+1}, \mathbf{p}_{-1}) = O(\delta^2)$$

$$I(Z \wedge Y^n) = O(\delta^2 \cdot I(X^n \wedge Y^n)\ )$$

Under information constraints

$$I(X^n \wedge Y^n) \leq n\ell$$

[BGMNW15] $I(Z \wedge Y^n) = \Omega(1) \Rightarrow \quad n = \Omega\left(\frac{1}{\ell\delta^2}\right).$

$$I(X^n \wedge Y^n) = O(n\varrho^2)$$

[DR19] $I(Z \wedge Y^n) = \Omega(1) \Rightarrow \quad n = \Omega\left(\frac{1}{\varrho^2\delta^2}\right).$

Generalization to $d > 1$

$\mathcal{Z} = \{-1,1\}^d, \delta = \frac{\varepsilon}{\sqrt{d}}$ $\Theta_{\mathcal{Z}} = \mathrm{Bern}\left(0.5 \pm \frac{\varepsilon}{\sqrt{d}}\right)^{\oplus d}$

[BGMNW15] Prove a direct sum result

$\sum_i I(Z_i \wedge Y^n) = \Omega(d) \Rightarrow$ $n = \Omega\left(d \cdot \frac{d}{\ell \varepsilon^2}\right)$

[DR19] Use the direct sum result

$\sum_i I(Z_i \wedge Y^n) = \Omega(d) \Rightarrow$ $n = \Omega\left(d \cdot \frac{d}{\varrho^2 \varepsilon^2}\right)$

Conclusion

- Tight lower bounds for estimation $\mathcal{B}_d, \mathcal{G}_d$ under ℓ_2

- Can naturally extend to other ℓ_p loss

- *Does not yield desired bounds Δ_k*

3. χ^2 contraction+ Assouad

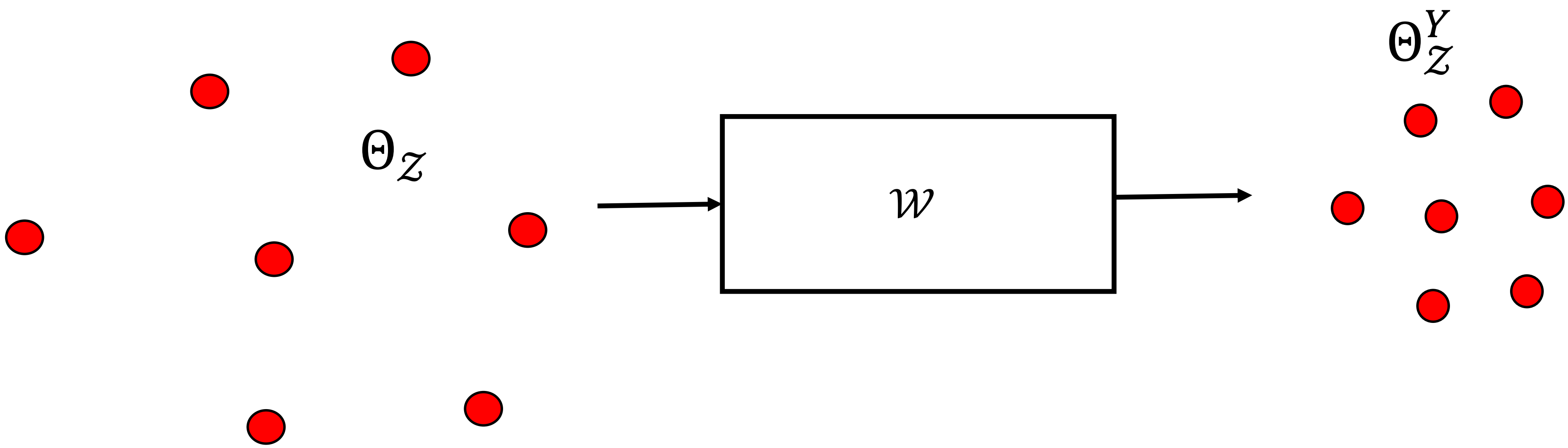
Outline for method 3

- Bounding mutual information by chi-squared contractions
- Bounding the chi squared contraction
- General plug and play bounds
- Application to Δ_k
- Extensions to $\mathcal{B}_d$

Bounding mutual information

By Assouad's method,

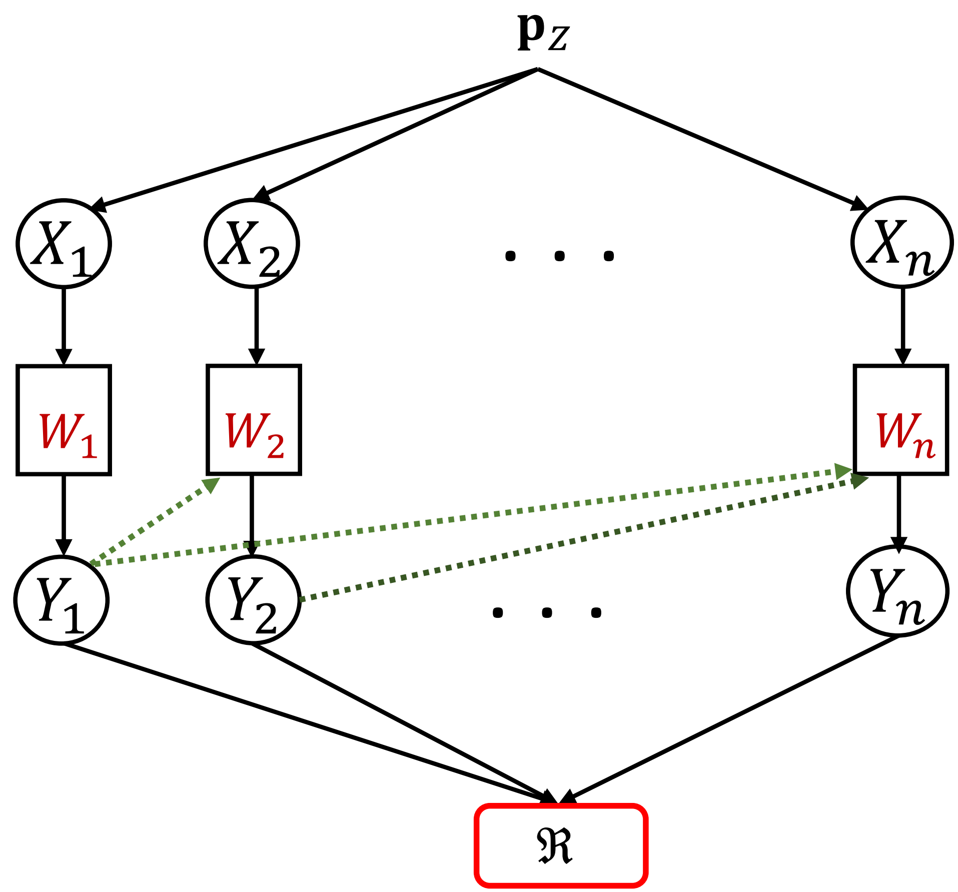
Bound $\sum_i I(Z_i \wedge Y^n)$ as a function of $\mathcal{W}$



Notation

$\mathbf{p}_z$: shorthand for $\mathbf{p}_{\theta_z}$

$\mathbf{p}_z^{Y^n}$: distribution of Y^n when input distribution $\mathbf{p}_z$



Information bound on one coordinate

Fix $i \in [m]$

Bound $I(Z_i \wedge Y^n)$

average output distribution fixing $Z_i = \pm 1$:

$$\mathbf{p}_{+i}^{Y^n} := \frac{1}{2^{k-1}} \sum_{z: z_i = +1} \mathbf{p}_z^{Y^n}$$

$$\mathbf{p}_{-i}^{Y^n} := \frac{1}{2^{k-1}} \sum_{z: z_i = -1} \mathbf{p}_z^{Y^n}$$

$I(Z_i \wedge Y^n)$ is large $\Leftrightarrow$ $\mathbf{p}_{+i}^{Y^n}$ and $\mathbf{p}_{-i}^{Y^n}$ must be far

$\Rightarrow$ bound distance between $\mathbf{p}_{+i}^{Y^n}$ and $\mathbf{p}_{-i}^{Y^n}$

How do channels shrink the distance?

Difficulty in handling distributions [ACLST20]

$$D\left(\mathbf{p}_{+i}^{Y^n} || \mathbf{p}_{-i}^{Y^n}\right) = \sum_t \mathbb{E}_{\mathbf{p}_{+i}^{Y^{t-1}}} \left[D\left(\mathbf{p}_{+i}^{Y_t | Y^{t-1}} || \mathbf{p}_{-i}^{Y_t | Y^{t-1}} \right) \right]$$

1. **Interactivity** in the protocols to choose channels
2. $\mathbf{p}_{+i}$ and $\mathbf{p}_{-i}$ **mixture** distributions, complicated expressions

Delicate to handle (see discussion in [ACLST20])

Convexity to handle mixtures [ACLST20]

$z \in \{-1,1\}^m$, $z^{\oplus i}$ obtained by flipping the ith coordinate of z

Theorem.

$$I(Z_i \wedge Y^n) \leq \frac{1}{2^{m+1}} \sum_{z \in \{-1,1\}^m} D\left(\mathbf{p}_z^{Y^n} || \mathbf{p}_{z^{\oplus i}}^{Y^n}\right) = \frac{1}{2} \mathbb{E}_Z\left[D\left(\mathbf{p}_Z^{Y^n} || \mathbf{p}_{Z^{\oplus i}}^{Y^n}\right)\right]$$

Proof. Convexity of divergence to the definitions of $\mathbf{p}_{+i}^{Y^n}$ and $\mathbf{p}_{-i}^{Y^n}$ ■

Information about Z_i bounded by average divergence in message distribution upon changing only Z_i

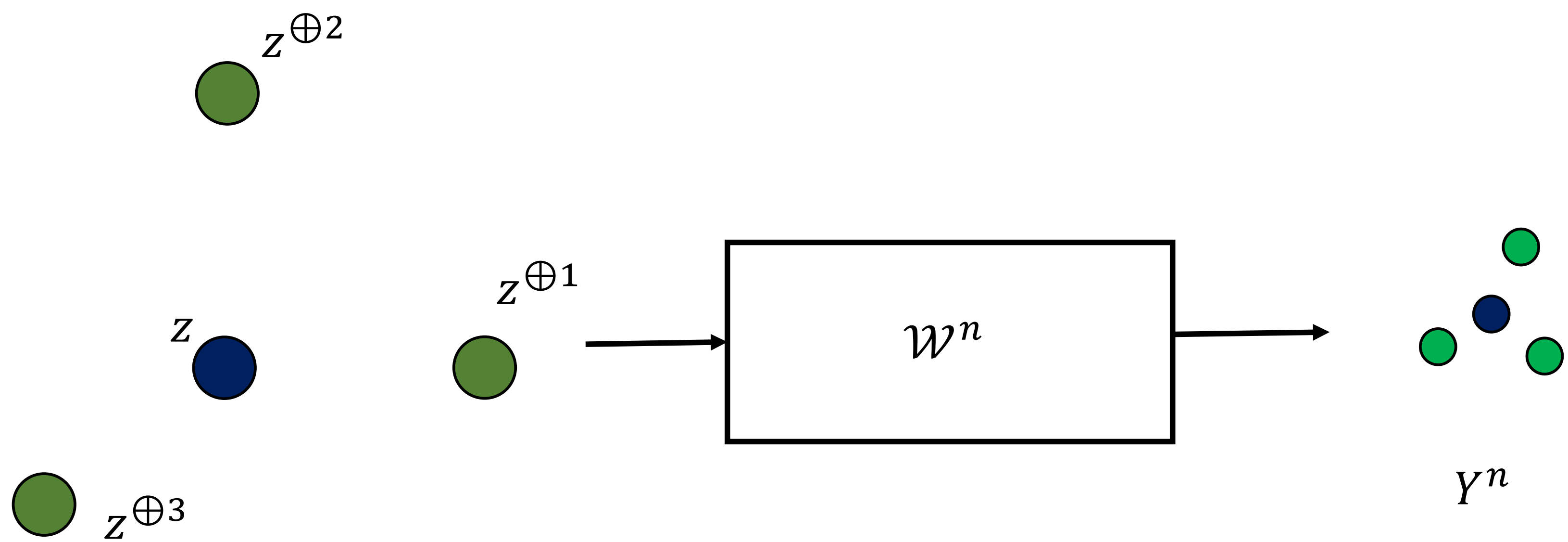
Focus on one z

By linearity of expectations

$$\sum_i I(Z_i \wedge Y^n) \leq \frac{1}{2}\,\mathbb{E}_Z\left[\sum_i D\left(\mathbf{p}_Z^{Y^n} \| \mathbf{p}_{Z\oplus i}^{Y^n}\right)\right]$$

Therefore,

$$\sum_i I(Z_i \wedge Y^n) \leq \frac{1}{2}\max_z\left[\sum_i D\left(\mathrm{p}_z^{Y^n} \| \mathrm{p}_{z\oplus i}^{Y^n}\right)\right]$$

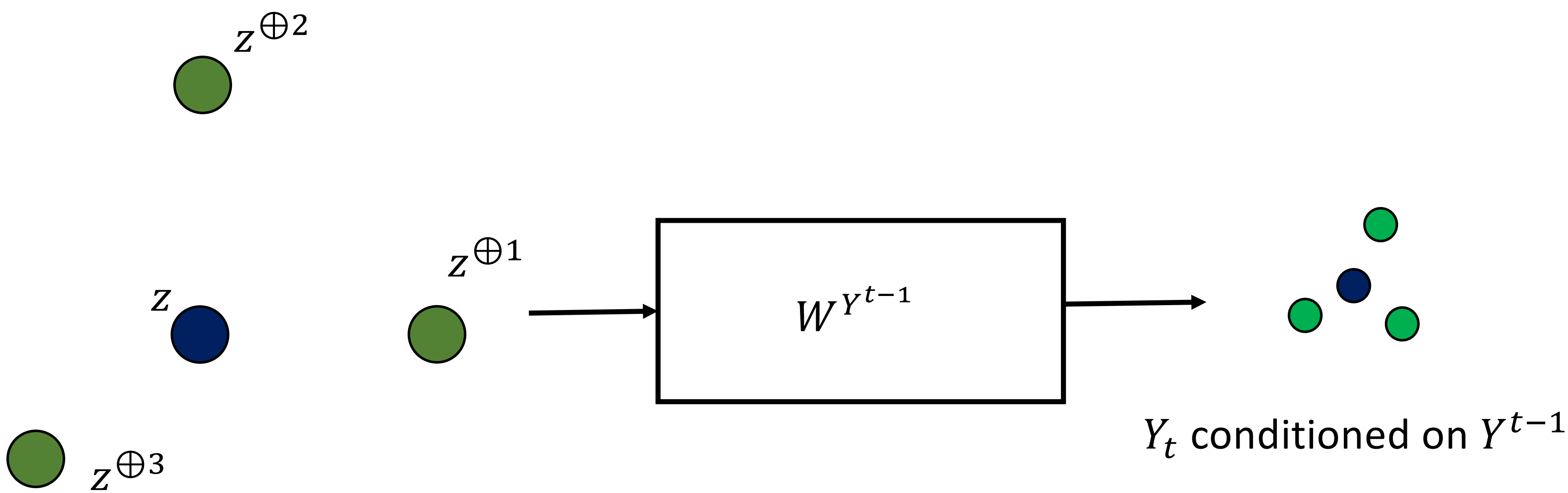


Bounding $\sum_i D\left(\mathrm{p}_z^{Y^n} || \mathrm{p}_{z\oplus i}^{Y^n}\right)$

By the chain rule of divergence

$$\sum_i D\left(\mathbf{p}_z^{Y^n} || \mathbf{p}_{z\oplus i}^{Y^n}\right) = \sum_t \mathbb{E}_{\mathbf{p}_z^{Y^{t-1}}}\left[\sum_i D\left(\mathbf{p}_z^{Y_t|Y^{t-1}} || \mathbf{p}_{z\oplus i}^{Y_t|Y^{t-1}}\right)\right].$$

- $\mathbf{p}_z^{Y_t|Y^{t-1}}$: Distribution of Y_t with input $\mathbf{p}_z$ conditioned on Y^{t-1}
- Channel at player t a function only of Y^{t-1}, denoted $W^{Y^{t-1}}$



Bounding $\sum_i D\left(\mathbf{p}_z^{Y_t|Y^{t-1}} || \mathbf{p}_{z\oplus i}^{Y_t|Y^{t-1}}\right)$

Y^{t-1} **fixed** (conditioning on Y^{t-1}), denote $W^{Y^{t-1}}$ by W_t

$$\mathbf{p}_z^{Y_t}(y) \coloneqq \mathbf{p}_z^{Y_t|Y^{t-1}}(y) = \sum_x \mathbf{p}_z(x) W_t(y|x) = \mathbb{E}_{X_t \sim \mathbf{p}_z}[W_t(y|X_t)]$$

Since $\mathrm{KL} \le \chi^2$, plugging the expression above

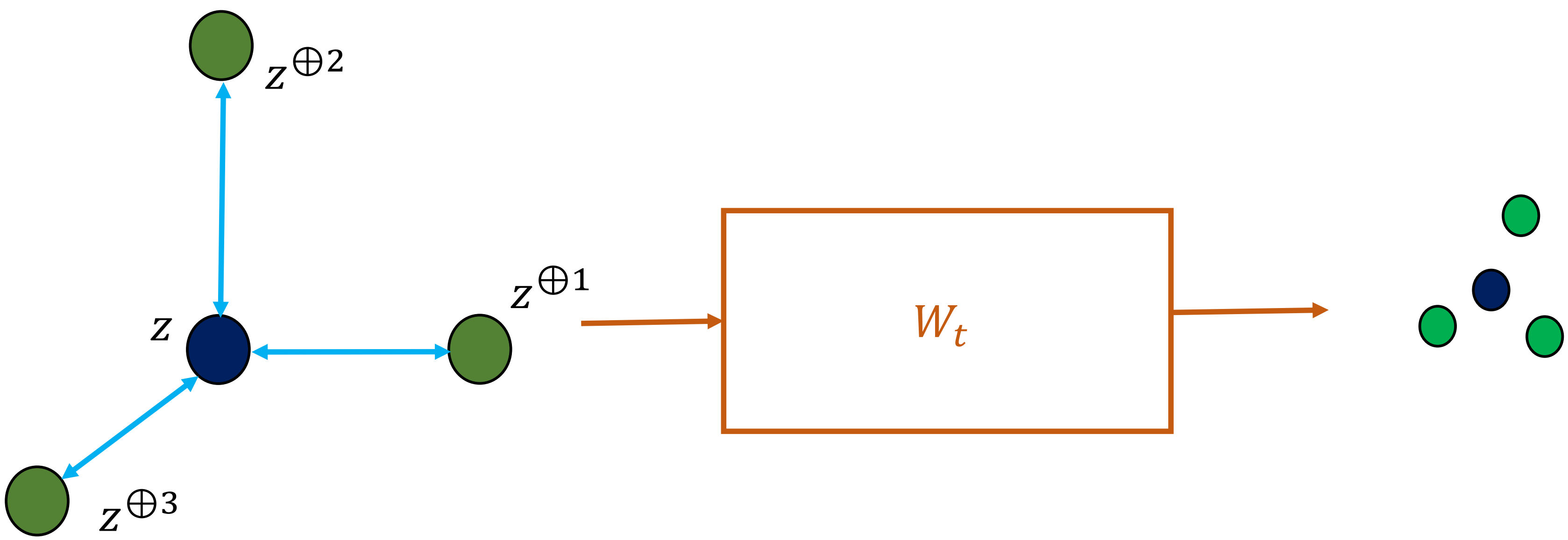
$$\sum_i D\left(\mathbf{p}_z^{Y_t} || \mathbf{p}_{z\oplus i}^{Y_t}\right) \le \sum_i \sum_y \frac{\left(\mathbf{p}_z^{Y_t}(y) - \mathbf{p}_{z\oplus i}^{Y_t}(y)\right)^2}{\mathbf{p}_{z\oplus i}^{Y_t}(y)}$$

$$= \sum_i \sum_y \frac{\left(\sum_x (\mathbf{p}_z(x) - \mathbf{p}_{z\oplus i}(x)) W_t(y|x)\right)^2}{\mathbb{E}_{\mathbf{p}_{z\oplus i}}[W_t(y|X)]}$$

An explicit bound at one user

$$\sum_i D\left(\mathbf{p}_z^{Y_t|Y^{t-1}} || \mathbf{p}_{z\oplus i}^{Y_t|Y^{t-1}}\right) = \sum_i \sum_y \frac{\left(\sum_x \left(\mathbf{p}_z(x) - \mathbf{p}_{z\oplus i}(x)\right) W_t(y|x)\right)^2}{\mathbb{E}_{\mathbf{p}_{z\oplus i}}[W_t(y|X)]}$$

Explicit bound on mutual information in terms of channels



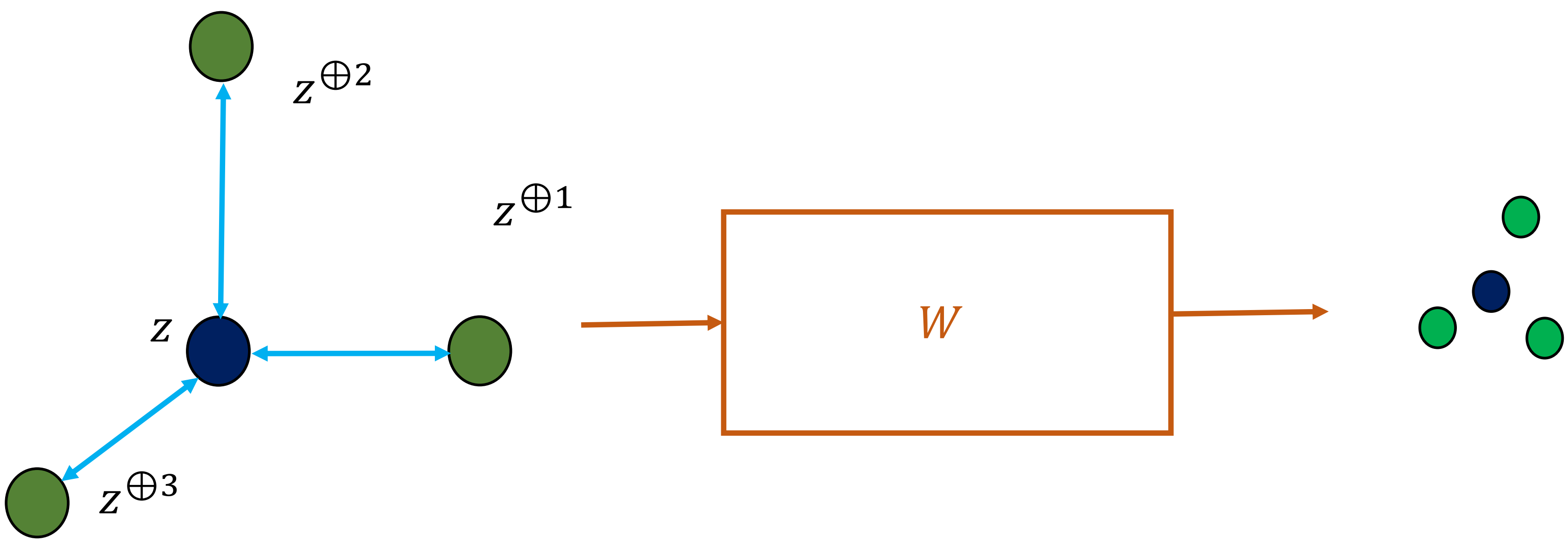
Average information bound [ACLST20, ACT20]

Theorem.

$$\sum_i I(Z_i \wedge Y^n) \leq n \cdot \sup_z \sup_{W \in \mathcal{W}} \sum_i H(W, z, i)$$

where

$$H(W, z, i) := \sum_y \frac{\left(\sum_x \left(\mathbf{p}_z(x) - \mathbf{p}_{z^{\oplus i}}(x)\right) W(y|x)\right)^2}{\mathbb{E}_{\mathbf{p}_{z^{\oplus i}}}[W(y|X)]}$$



Applications

[ACLST20]

- Bounds for estimating Δ_k under ℓ_1
- Applications to testing distributions (in Part 3 of the tutorial)

[ACT20]

- Plug and play bounds for establishing lower bounds
- Bounds for estimating $\Delta_k, \mathcal{B}_d, \mathcal{G}_d$ under ℓ_p for $p \geq 1$

Estimating Δ_k under ℓ_1

Example: Δ_k under ℓ_1 [ACLST20]

Plugging into the definition of Paninski construction:

$$\sum_i H(W,z,i) = \frac{\varepsilon^2}{k} \cdot \sum_i \sum_y \frac{\left(W(y|2i-1) - W(y|2i)\right)^2}{\sum_x W(y|x)}$$

Used for testing in next part:

$$\| H(W) \|_* := \sum_i \sum_y \frac{\left(W(y|2i-1) - W(y|2i)\right)^2}{\sum_x W(y|x)}$$

A few lines of computation

$$\sup_{W \in \mathcal{W}_\ell} \| H(W) \|_* = \min\{2^\ell, k\}$$

$$\sup_{W \in \mathcal{W}_\varrho} \| H(W) \|_* = O(\varrho^2)$$

Example: Δ_k under ℓ_1 [ACLST20]

Plugging in the theorem and requiring

$\sum_i I(Z_i \wedge Y^n) = \Omega(k) \Rightarrow \qquad n = \Omega\left(\frac{k^2}{\varepsilon^2 \min\{2^\ell, k\}}\right)$

$\sum_i I(Z_i \wedge Y^n) = \Omega(k) \Rightarrow \qquad n = \Omega\left(\frac{k^2}{\varepsilon^2 \varrho^2}\right)$

Plug and play bounds

Towards plug and play bounds [ACT20]

$$\sum_i \sum_y \frac{\left(\sum_x \left(\mathbf{p}_z(x) - \mathbf{p}_{z\oplus i}(x)\right) W(y|x)\right)^2}{\mathbb{E}_{\mathbf{p}_{z\oplus i}}[W(y|X)]}$$

Suppose some nice properties hold (and they do for $\Delta_k, \mathcal{B}_d$)

A1 **(nice densities):** For some α, $\frac{\mathbf{p}_z(x)-\mathbf{p}_{z\oplus i}(x)}{\mathbf{p}_z(x)} = \alpha \cdot \phi_{z,i}(x)$

A2 **(Boundedness):** For some κ, $\sup_{y,W} \frac{\mathbb{E}_{X\sim\mathbf{p}_z}[W(y|X)]}{\mathbb{E}_{X\sim\mathbf{p}_{z\oplus i}}[W(y|X)]} \leq \kappa$

A3 **(orthonormality):** $\mathbb{E}_{X\sim\mathbf{p}_z}\left[\phi_{z,i}(X)\phi_{z,j}(X)\right] = \delta_{ij}$

A variance plug and play bound [ACT20]

Theorem. Under A1, A2, and A3,

$$\sum_i I(Z_i \wedge Y^n) \leq O\left(n\alpha^2 \cdot \sup_z \sup_{W \in \mathcal{W}} \sum_y \frac{\mathrm{Var}_{X \sim \mathbf{p}_z}\big(W(y|X)\big)}{\mathbb{E}_{X \sim \mathbf{p}_z}[W(y|X)]}\right)$$

Variance of message distribution

Applications:

$$\sum_y \frac{\mathrm{Var}_{X \sim \mathbf{p}_z}\big(W(y|X)\big)}{\mathbb{E}_{X \sim \mathbf{p}_z}[W(y|X)]} \leq |\mathcal{Y}| = 2^{\ell}$$

$$\sum_y \frac{\mathrm{Var}_{X \sim \mathbf{p}_z}\big(W(y|X)\big)}{\mathbb{E}_{X \sim \mathbf{p}_z}[W(y|X)]} \leq O(\varrho^2)$$

Applications [ACT20]

For Δ_k, Paninski construction, $\alpha = \varepsilon/\sqrt{k}$

$\sum_i I(Z_i \wedge Y^n) = \Omega(k) \Rightarrow \quad n = \Omega\left(\frac{k^2}{\varepsilon^2 \min\{2^\ell, k\}}\right)$

$\sum_i I(Z_i \wedge Y^n) = \Omega(k) \Rightarrow \quad n = \Omega\left(\frac{k^2}{\varepsilon^2 \varrho^2}\right)$

For $\mathcal{B}_d$, $\alpha = \varepsilon/\sqrt{d}$

$\sum_i I(Z_i \wedge Y^n) = \Omega(d) \Rightarrow \quad n = \Omega\left(\frac{d^2}{\varrho^2 \varepsilon^2}\right)$

An information plug and play bound [ACT20]

A4 **(subgaussianity):** For $X \sim \mathbf{p}_z$ $[\phi_{z,1}(X), \ldots, \phi_{z,m}]$ is σ^2-subgaussian

Theorem. Under A1, A2, and A3, A4

$$\sum_i I(Z_i \wedge Y^n) \leq O\left(n\alpha^2\sigma^2 \cdot \sup_z \sup_{W \in \mathcal{W}} H(p_Z^Y)\right),$$

where Y is message distribution with input $\mathbf{p}_z$**,** and H is the entropy

Application [ACT20]

For $\mathcal{B}_d$, $\alpha = \varepsilon/\sqrt{d}$, $\sigma = 1$

$\sum_i I(Z_i \wedge Y^n) = \Omega(d) \Rightarrow$ $n = \Omega\left(\frac{d^2}{\varepsilon^2 \ell}\right)$

Conclusion

Three methods to prove lower bounds on distributed estimation

Cramer Rao bounds + Fisher information

Distributed strong data processing + Assouad’s method

Chi-squared contraction + Assouad’s method

Coming up

Hypothesis testing under information constraints

Thanks!