Statistical Inference in Distributed or Constrained Settings: **Techniques and Recipes**


An Easy Dinner Classic

Classic Split Pea Soup
Serves: 8 | *Serving Size:* 1 and 1/2 cups

Ingredients:

2 cups split peas
8 cups water
1 bay leaf
1/4 teaspoon salt
2 cups chopped carrots
1 cup chopped celery
1 cup chopped onion
1 large potato, diced
1 teaspoon thyme leaves
1/2 teaspoon pepper

Directions:

Combine split peas, water, bay leaf, and salt in a large kettle.

Bring to a boil, reduce heat and simmer for 2 hours, stirring occasionally.

Add the remaining ingredients and continue to simmer until the vegetables are tender, about 30 minutes longer.

Nutrition Information:

Serves 8. Each 1 and 1/2 cup serving has 89 calories, 0 g fat, 0 g saturated fat, 0 g trans fat, 0 mg cholesterol, 233 mg sodium, 18 g carbohydrate, 6 g fiber, 4 g sugar, and 5 g protein.

Each serving also contains 108% DV vitamin A, 10% DV vitamin C, 4% DV calcium, and 5% DV iron.

Chef's Tips:

Frozen peas make a great garnish. You can add them to the soup during the last 10 minutes of cooking.

As the split peas are cooking, check to make sure that there is enough water and that the split peas do not stick. Add more water if the soup becomes too thick.

Peas are low in fat, but naturally high in protein and fiber.

BROUGHT TO YOU BY:

©www.foodandhealth.com

Conference on Learning Theory 2021

What’s on the menu?

I. Appetizers	Jayadev
II. MC 1	Jayadev
III. MC 2	Himanshu
IV. DIY Desserts	Clément

Chefs: Jayadev Acharya, Clément Canonne, Himanshu Tyagi

COLT 2021

Appetizers

- Statistical Inference
- Distributed / constrained settings
- Problems and examples
- Related work and pointers

Main Course – I: Discrete distributions

- A puzzle to solve **all** problems under communication constraints
- Lower bounds for interactive estimation for arbitrary channels
  - Tight bounds under communication, privacy as application

Main Course – II: General distributions Himanshu

Unified method to prove “interactive” lower bounds

- Discrete, high-dimensional, nonparametric, etc
- Communication, privacy, etc
- General plug-n-play methods

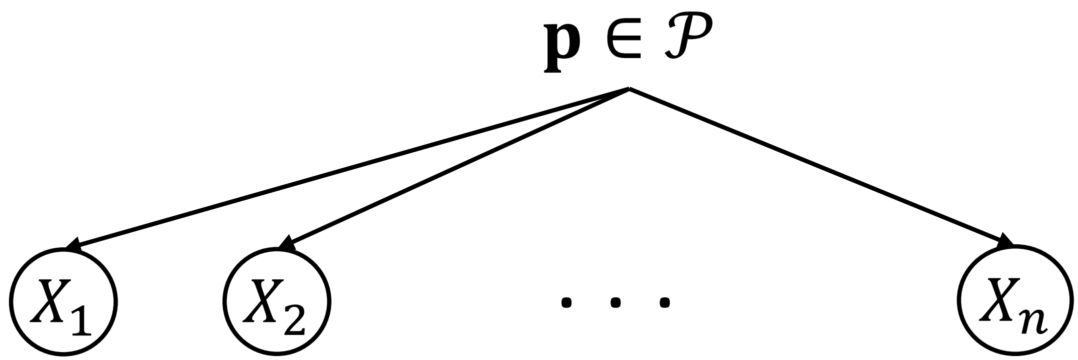
DIY desserts: Recitation

Clément

- How to apply the lower bounds
- Several exercises

Statistical Inference

$\mathcal{P}$: family of distributions over $\mathcal{X}$



Given $X^n := (X_1, \ldots, X_n)$: i.i.d. samples from an unknown $\mathbf{p}$

Solve some inference task about $\mathbf{p}$

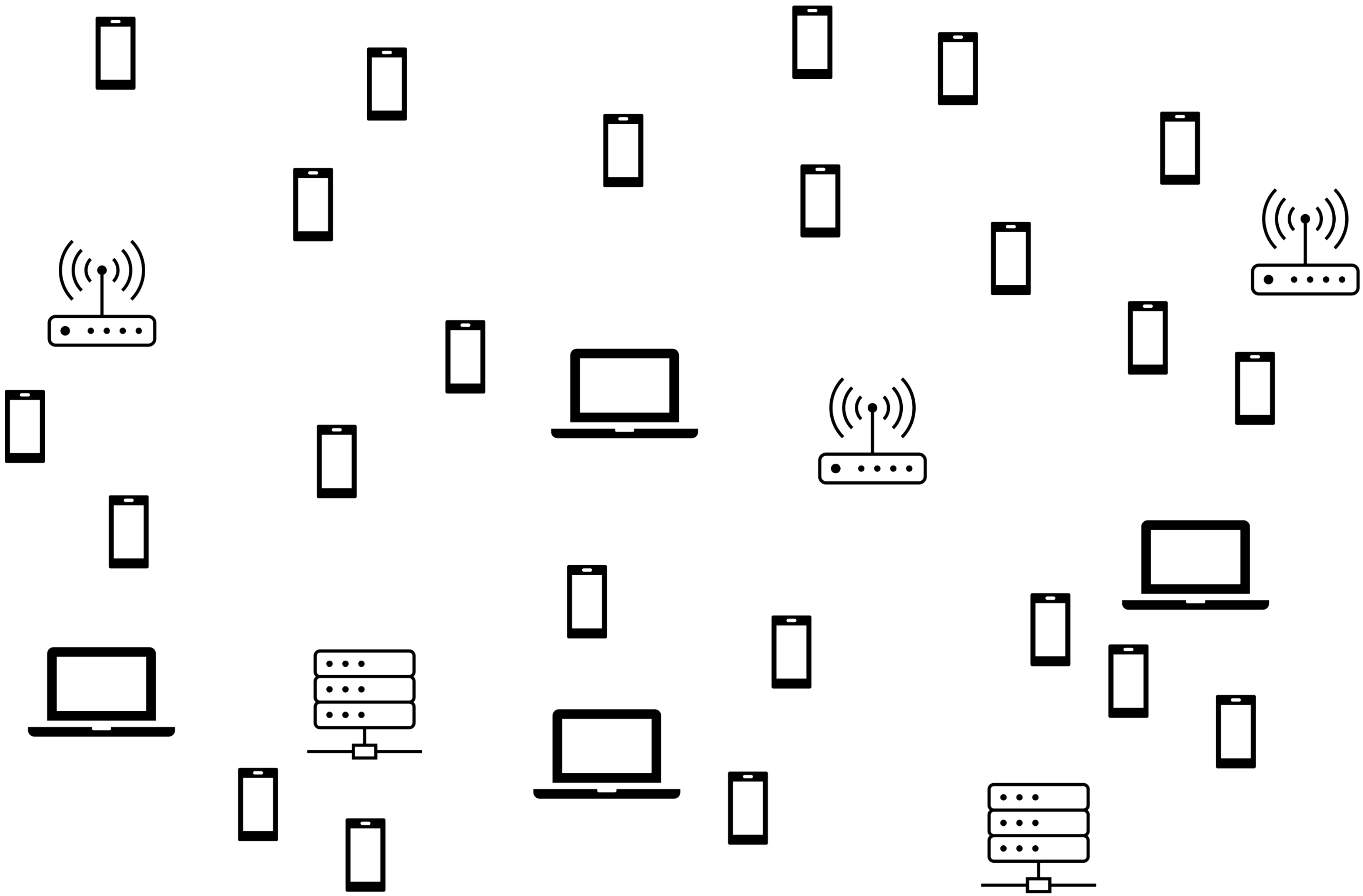
Sample complexity: smallest n to solve the task

This is inference in central setting

Information Constraints

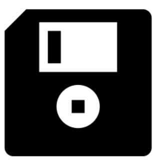
Distributed or Constrained Settings

No direct access to X_is



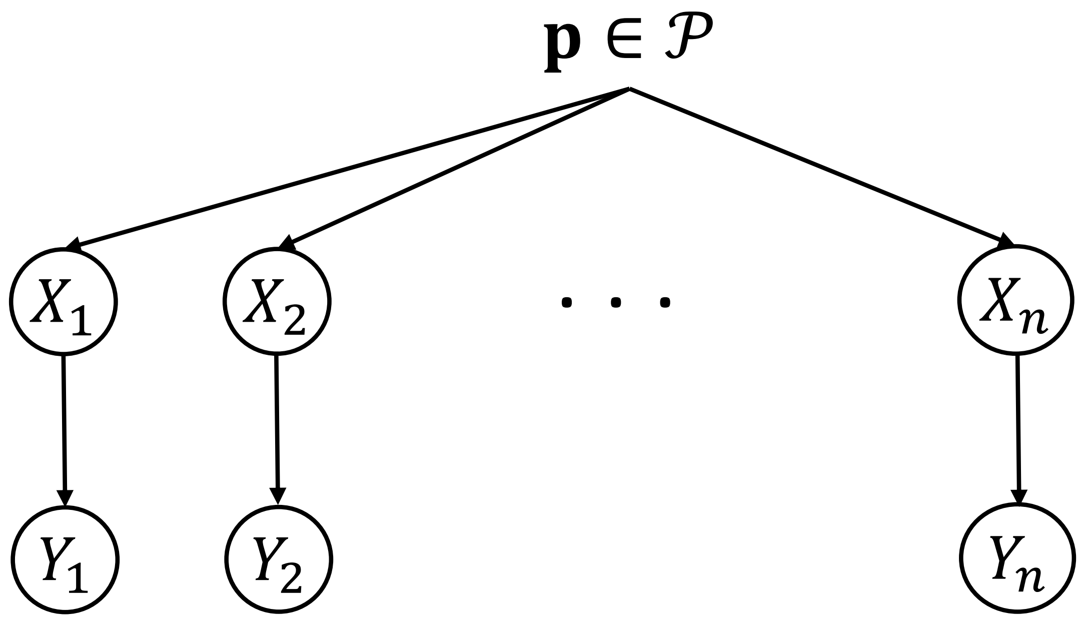
("Motivation" slide)

Statistical Inference under **constraints**



Local constraints

Statistical Inference

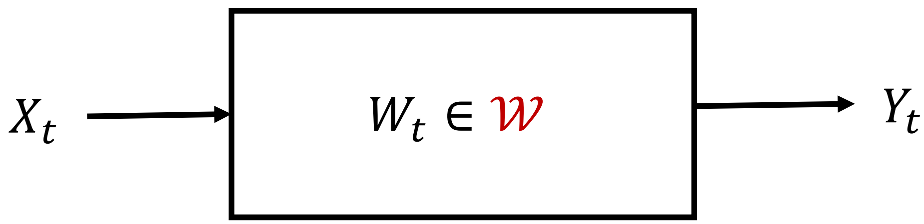


The messages are what we observe with constraints

Modeling the constraints

[ACT20c]

n users, user t observes X_t and sends message Y_t

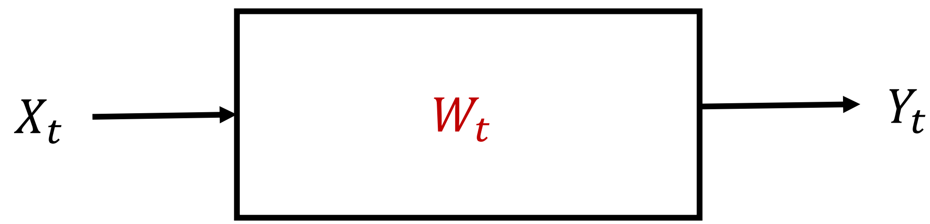


$$W_t(y|x) \coloneqq \Pr(Y_t = y | X_t = x)$$

$W_t \in \mathcal{W}$: a set of **allowed** (randomized) channels $\Leftrightarrow$ the constraints

The algorithm/protocol dictates how user t chooses W_t from $\mathcal{W}$

Modeling the local information constraints



When $X_t \sim \mathbf{p}$

$$\mathbf{p}^{W_t}(Y_t = y) := \sum_x \mathbf{p}(x) W_t(y|x) = \mathbb{E}[W_t(y|X)]$$

Example 1: Communication constraints

[Shamir14,HMÖW18,ACT20d…]

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

Each X_t is mapped to ℓ bits.

Bandwidth constraints

Example 2: Local Differential Privacy (LDP)

[Warner65, EPR03, KLNRS11]

$W: \mathcal{X} \to \{0,1\}^*$ is ϱ-**LDP** if $\forall x, x' \in \mathcal{X}, \forall y,$

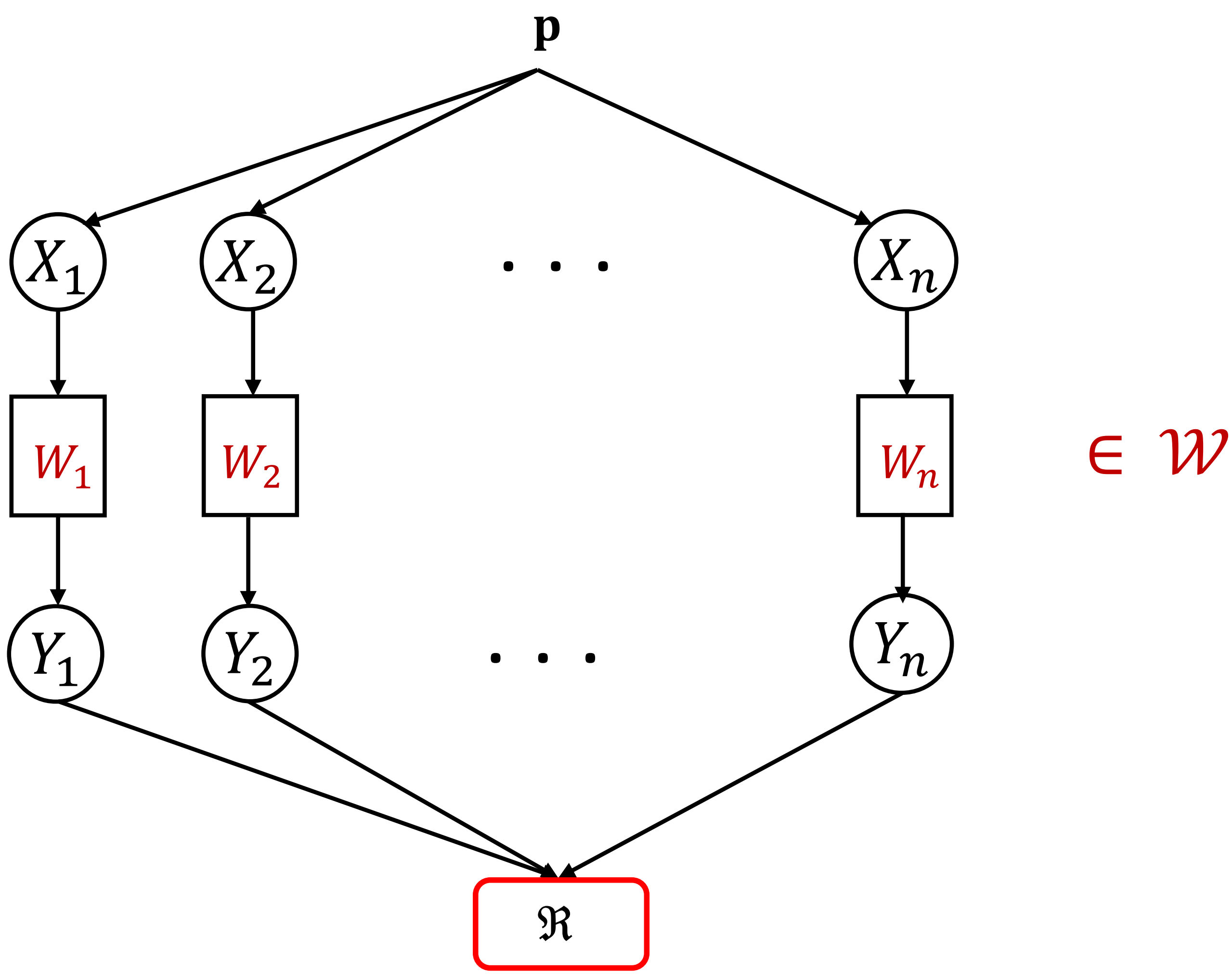
$$\frac{W(y|x)}{W(y|x')} \leq e^{\varrho} \approx 1 + \varrho$$

$$\mathcal{W}_{\varrho} = \{\text{all } \varrho - \text{LDP channels}\}$$

Privacy guarantees even "against" the server

The Protocols

Distributed Statistical Inference



Given $Y^n := Y_1, \ldots, Y_n$, solve the inference task

Distributed statistical inference

Once we decide $W^n := W_1, \ldots, W_n$,

$$\mathbf{p}^{W^n}(Y^n) = \prod_t \mathbf{p}^{W_t}(Y_t)$$

How to choose $W_1, W_2, \ldots, W_n \in \mathcal{W}$ to minimize n?

The protocols

Simultaneous Message Passing (SMP)/Non-interactive schemes

W_ts are chosen simultaneously

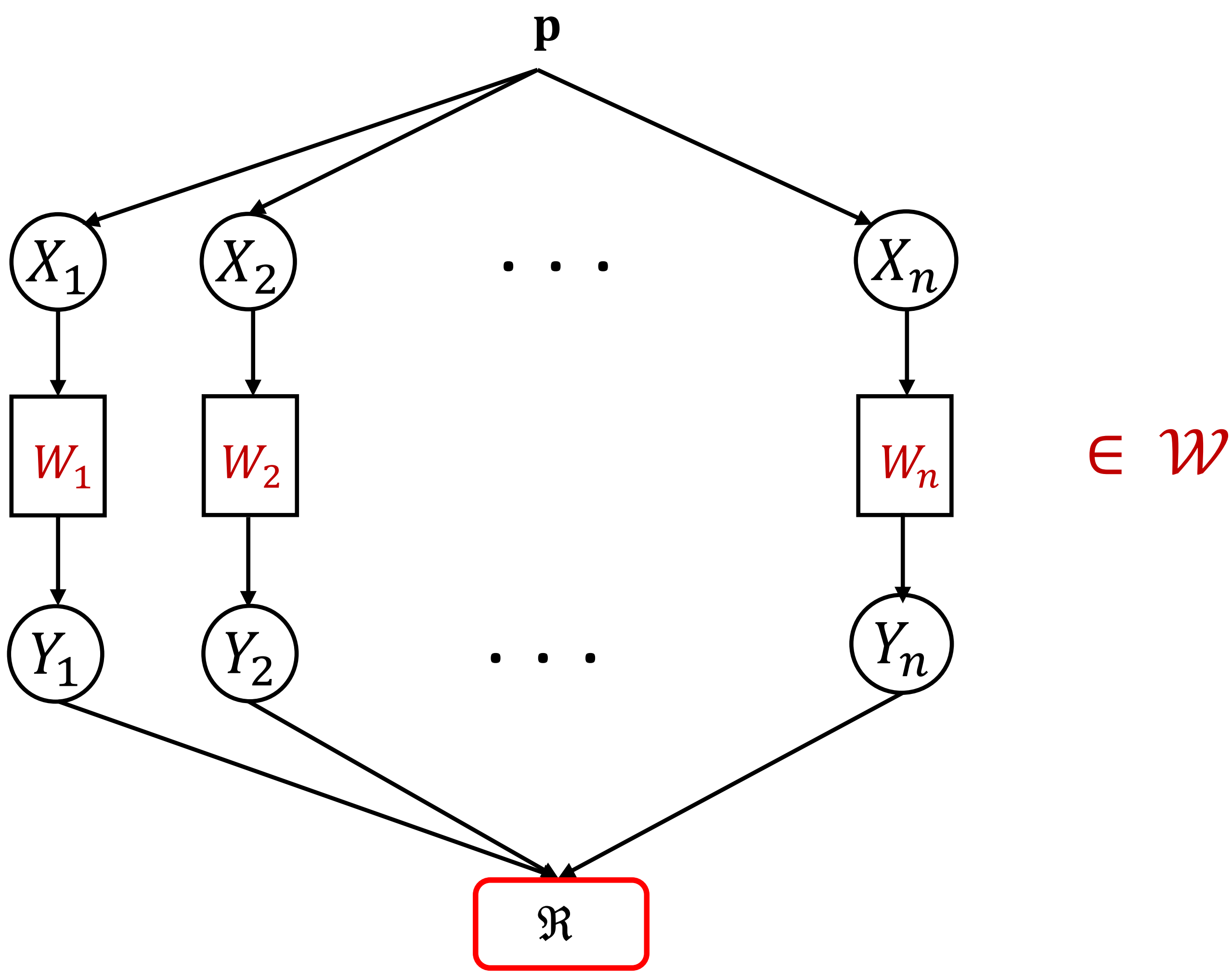
private-coin SMP (no shared randomness)

W_ts are chosen independently

$Y_1, Y_2, \ldots, Y_n$ are independent

e.g., $W_1, \ldots, W_n$ are fixed

Private-coin SMP protocols



Noninteractive ("simultaneous message-passing"),
no common randomness

The protocols

Simultaneous Message Passing (SMP)/Non-interactive schemes

W_ts are chosen simultaneously

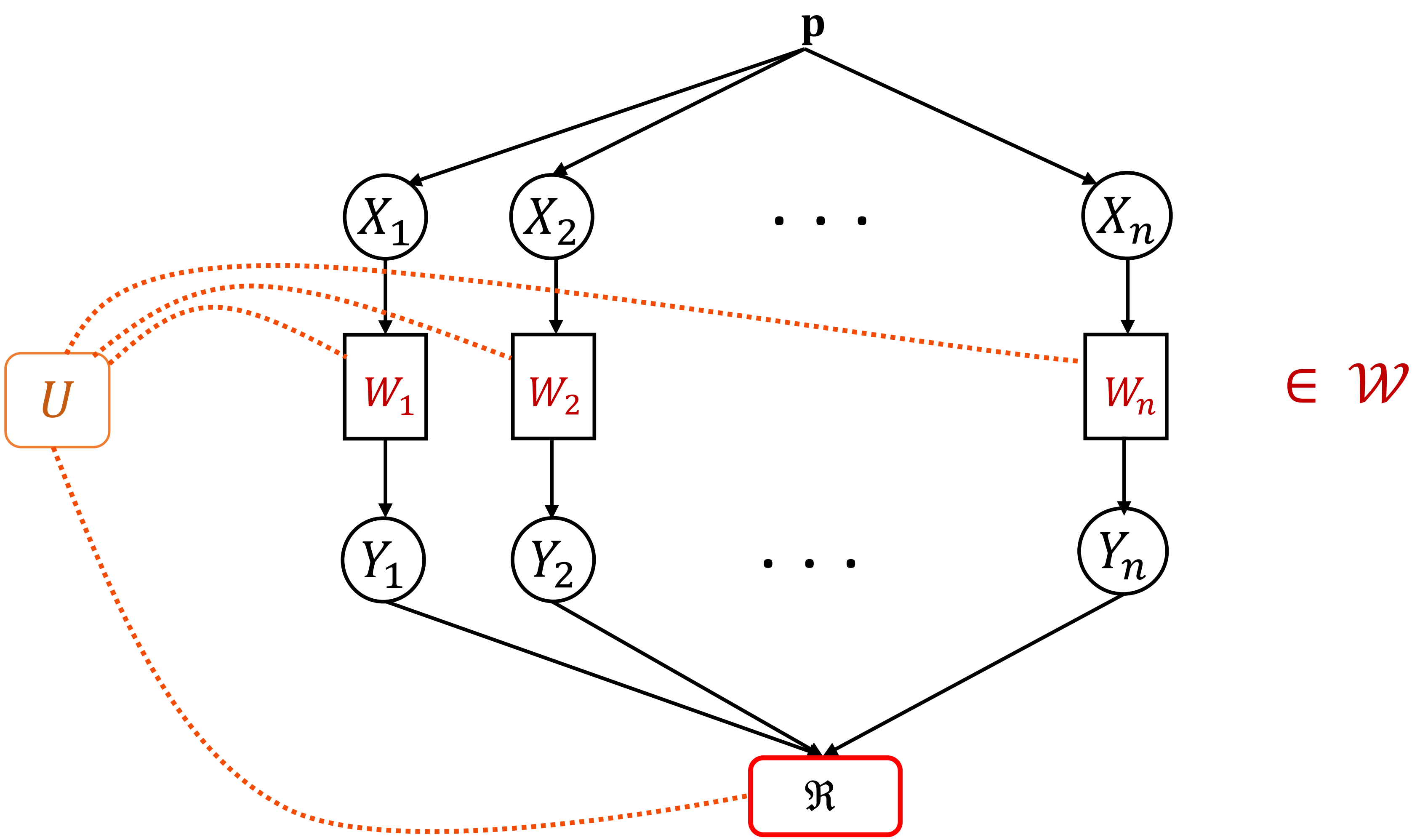
public-coin SMP (shared randomness)

U: common random string available to all users and referee

W_t is a function of U

$Y_1, Y_2, \dots, Y_n$ are independent **given** U

Public-coin SMP protocols



Noninteractive ("simultaneous message-passing"),
but common random seed

The protocols

Interactive schemes

W_ts can depend on previous messages

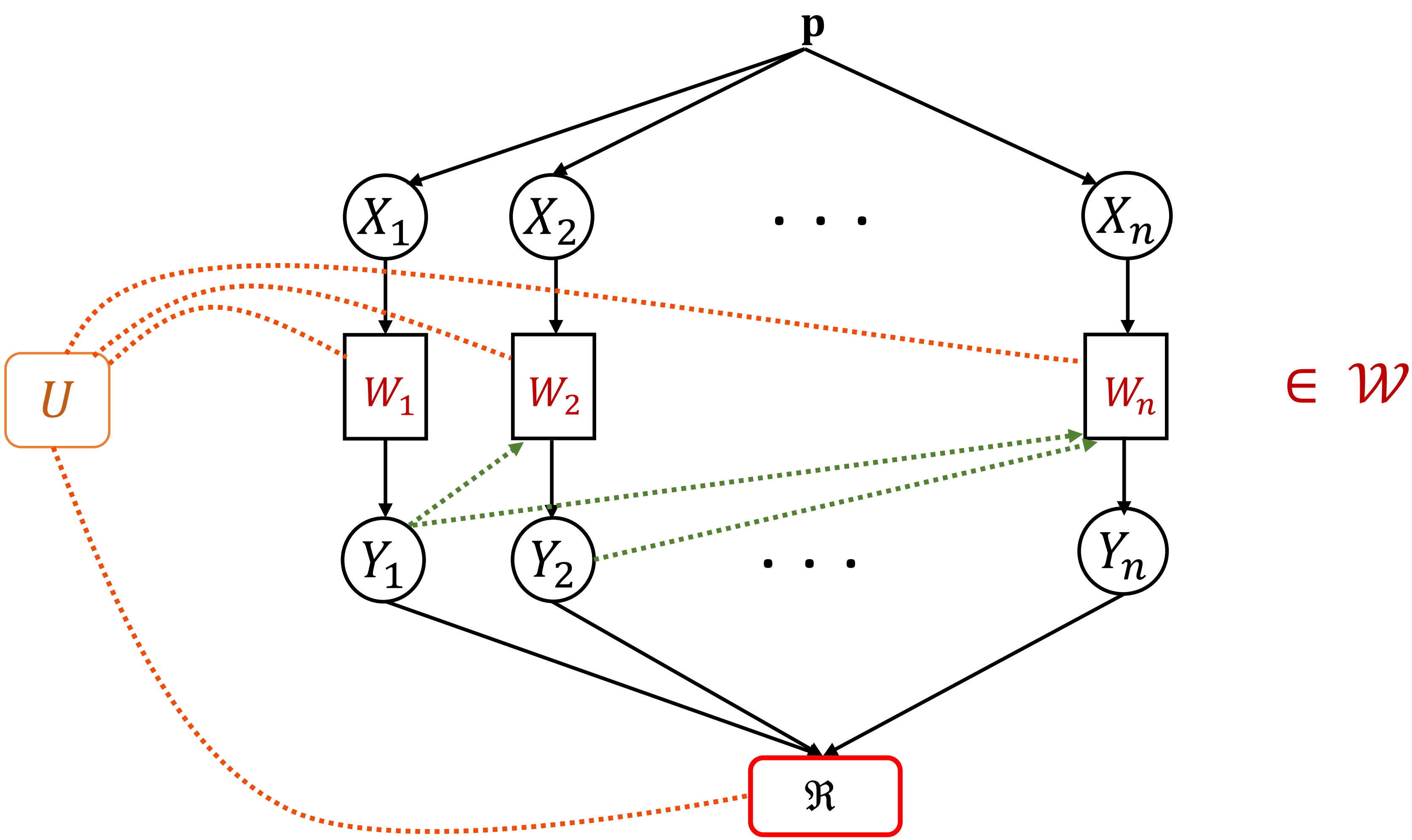
sequentially interactive protocols

U: common random string available to all users and referee

for $t = 1, \dots, n$

W_t is a function of (U, Y^{t-1})

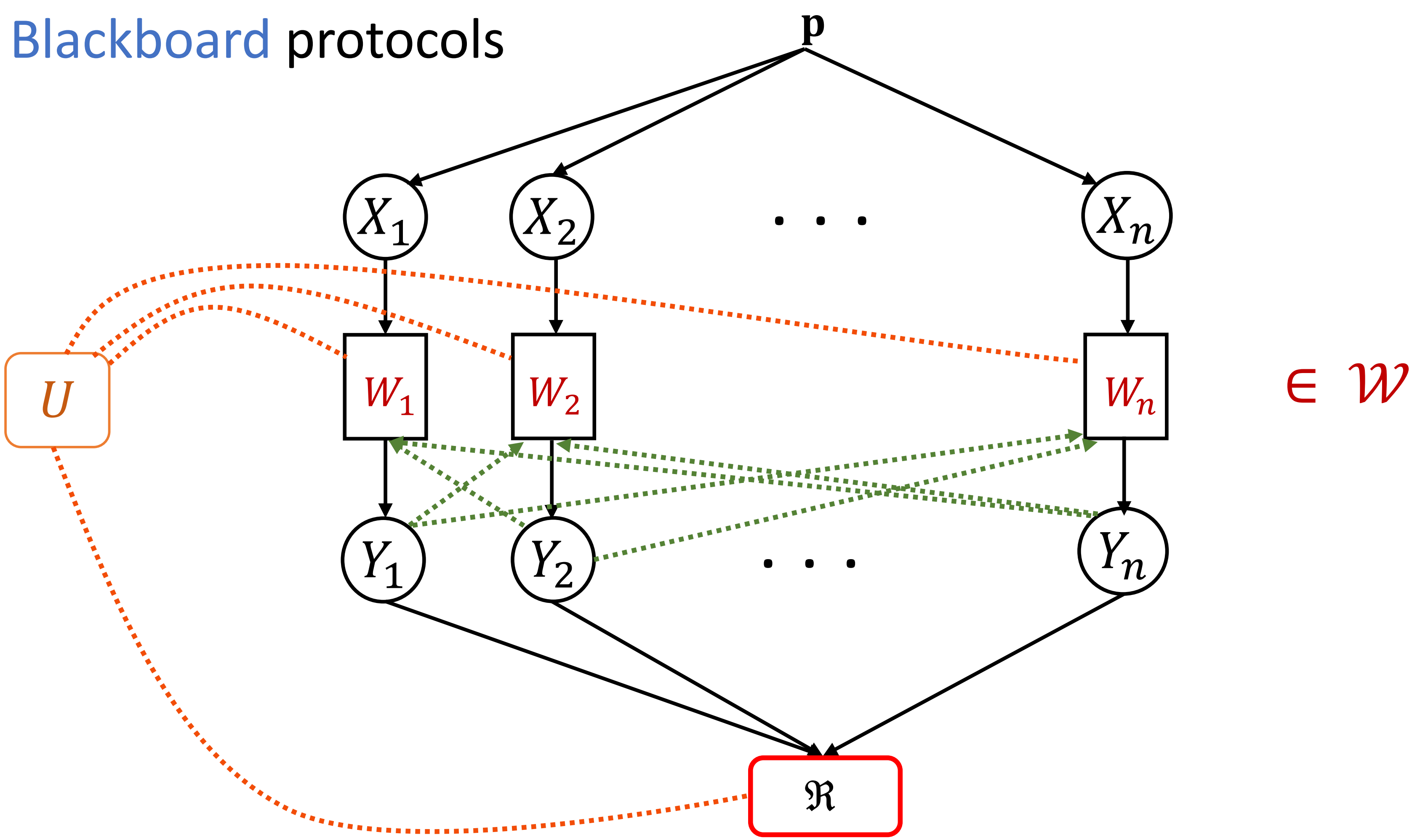
Sequentially Interactive protocols



Interactive ("one-pass, sequential"),
and common random seed

Types of protocols

Blackboard protocols



Fully interactive ("many passes"),
and common random seed

Types of protocols

Each of these models is **at least as powerful** as the previous

private-coin $\preccurlyeq$ public-coin $\preccurlyeq$ sequentially interactive $\preccurlyeq$ blackboard

Each has its pros and cons (both in theory *and* practice) and may require different techniques to analyze.

Questions about setting?

The Problems

Parameter/density estimation

Goodness-of-fit / Hypothesis testing

Sample complexity: smallest n to solve the task

Example 1: Discrete distributions

$\mathcal{P} = \Delta_d$: distbs on $[d] \coloneqq \{1 \dots d\}$

Goal: output $\widehat{\mathbf{p}}$ such that

$$\mathbb{E}[\mathrm{TV}(\widehat{\mathbf{p}}, \mathbf{p})] \le \varepsilon$$

Sample complexity = $\Theta\left(\frac{d}{\varepsilon^2}\right)$

(without constraints)

$\mathbf{q}$: a reference distribution

Goal: Test

$$\mathbf{p} = \mathbf{q} \text{ vs } \mathrm{TV}(\mathbf{p}, \mathbf{q}) > \varepsilon$$

Sample complexity = $\Theta\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$

(without constraints) [Paninski08]

$$TV(\mathbf{p}, \mathbf{q}) \coloneqq \sup_{S \subseteq [k]} \big(\mathbf{p}(S) - \mathbf{q}(S)\big) = \frac{1}{2} \ell_1(\mathbf{p}, \mathbf{q})$$

Example 2: High dimensional distributions

$$\mathcal{P} = \{\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) : \boldsymbol{\mu} \in \mathrm{R}^d\}$$

Goal: output $\widehat{\boldsymbol{\mu}}$ such that

$$\mathbb{E}[|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2] \leq \varepsilon^2$$

Sample complexity = $\Theta\left(\frac{d}{\varepsilon^2}\right)$

(without constraints)

Goal: Test

$$\boldsymbol{\mu} = \mathbf{0} \text{ vs } |\boldsymbol{\mu}|_2 > \varepsilon$$

Sample complexity = $\Theta\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$

(without constraints)

*detecting signal vs noise

Other families: product Bernoulli

Research goals

Establish sample complexity bounds for …

- Different $\mathcal{W}$s
- Estimation/Testing/other properties
- Private-coin SMP/public-coin SMP/interactive
- Discrete/high-dimensional/non-parametric

Mix-n-match?

Already a bit too much … each interesting in its own right … !

For example … discrete distribution testing

$\mathcal{W}_\varrho$, [AminJosephMao '20, BerrettButucea'20, AcharyaCanonneLiuSunTyagi'20]:

Private-coin SMP $\ll$ public-coin SMP $\approx$ SMP/interactive

$\mathcal{W}_\ell$, [AcharyaCanonneLiuSunTyagi'20]:

Private-coin SMP $\ll$ public-coin SMP $\approx$ SMP/interactive

General $\mathcal{W}$, [AcharyaCanonneLiuSunTyagi'20]:

Private-coin SMP $\ll$ public-coin SMP $\ll$ SMP/interactive

Similarly for Gaussian mean testing … [AcharyaCanonneTyagi'20, SzaboVuursteenVanZanten'20]

Parameter/density estimation

~~Goodness-of-fit / Hypothesis testing~~

Part 3 of tutorial (link)

Learn about Ingster's method from HT!

Establishing tight results for SMP protocols generally easier …

$Y_1, \ldots, Y_n$ independent (given U)

See general discussion in

[ACLST20] J. Acharya, C. Canonne, Y. Liu, Z. Sun, H. Tyagi, "Interactive inference under information constraints" *arXiv: 2007.10976 (in submission)*

Methods to establish interactive lower bounds

1. Cramer-Rao/van Trees inequality [BarnesHanOzgur19, BarnesChenOzgur20, SarbuZaidi21]
   - Unified results for $\Delta_d, \mathcal{B}_d, \mathcal{G}_d$
   - Results hold for ℓ_2 loss

2. Strong Data Processing + Assouad's method [BravermanGardMaNguyenWoodruff16, DuchiRogers19]
   - 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 [AcharysCanonneLiuSunTyagi20, AcharyaCanonneSunTyagi20]
   - Unified bounds for $\Delta_d, \mathcal{B}_d, \mathcal{G}_d$
   - Works under ℓ_p for $p \geq 1$
   - For arbitrary channels

Pointers

Part 2 of tutorial (link)

Cramer-Rao/van Trees inequality

Strong Data Processing + Assouad's method

Next two parts …

MC1:

- Discrete distributions
  - Simulate and infer for upper bounds
  - Lower bounds

MC2:

- Unified approach for general distributions and channel families

MC 1: Discrete Distributions

Discrete distribution estimation

$\mathcal{P} = \Delta_d$: distbs on $[d] \coloneqq \{1 \dots d\}$

Goal: output $\widehat{\mathbf{p}}$ such that

$$\mathbb{E}[\mathrm{TV}(\widehat{\mathbf{p}}, \mathbf{p})] \leq \varepsilon$$

Sample complexity = $\Theta\left(\frac{d}{\varepsilon^2}\right)$ (without constraints)

Empirical distribution works - DIY

$X_1, \ldots, X_n \sim \mathbf{p}$, $N_x :=$ # times x appears

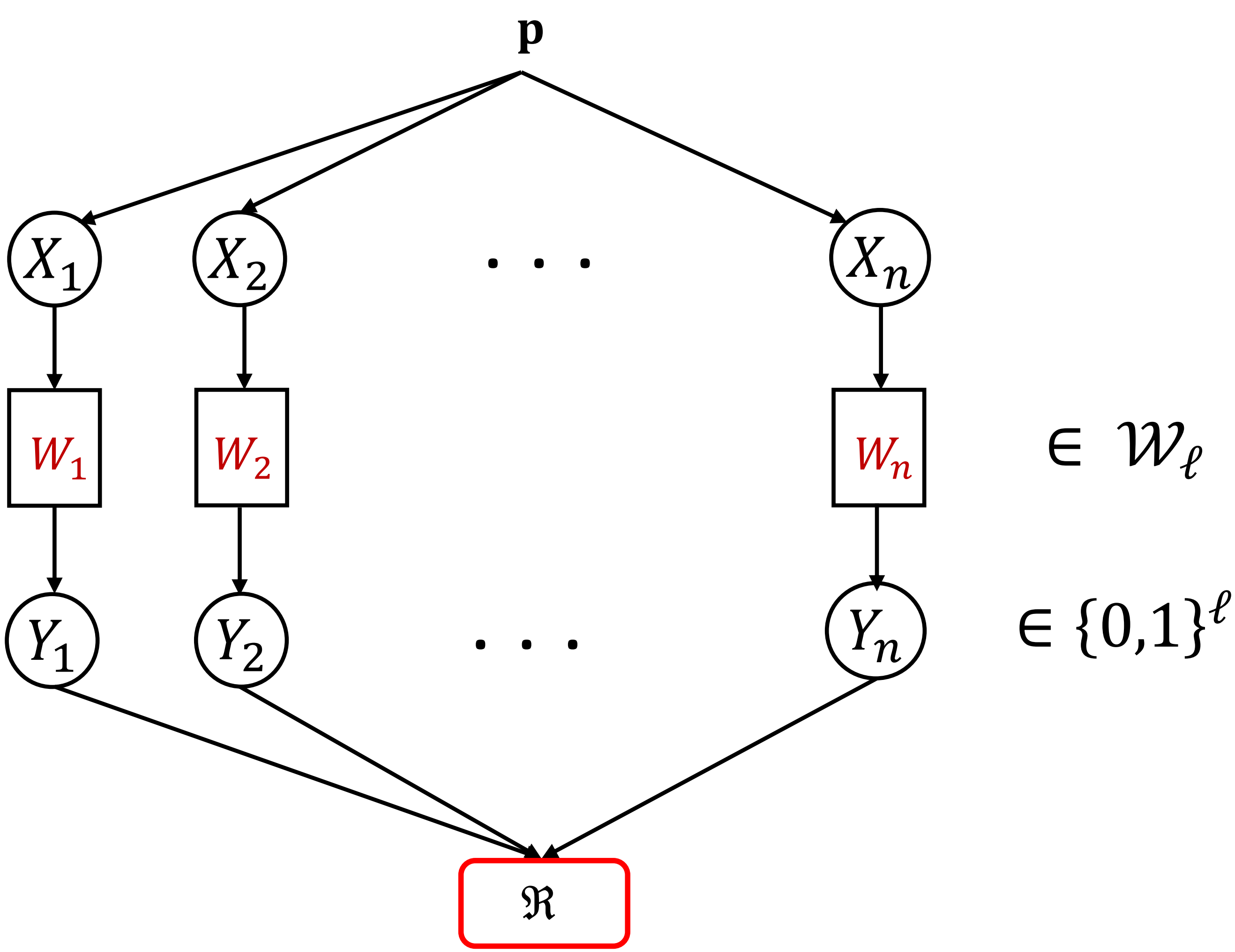
Empirical distribution: $\widehat{\mathbf{p}}(x) = N_x / n$

$N_x \sim \mathrm{Bin}\big(n, \mathbf{p}(x)\big)$

$$\mathbb{E}\left[\big(\widehat{\mathbf{p}}(x) - \mathbf{p}(x)\big)^2\right] = \frac{\mathbf{p}(x)\big(1 - \mathbf{p}(x)\big)}{n} \Rightarrow \mathbb{E}[\ell_2^2(\widehat{\mathbf{p}}, \mathbf{p})] \leq \frac{1}{n}$$

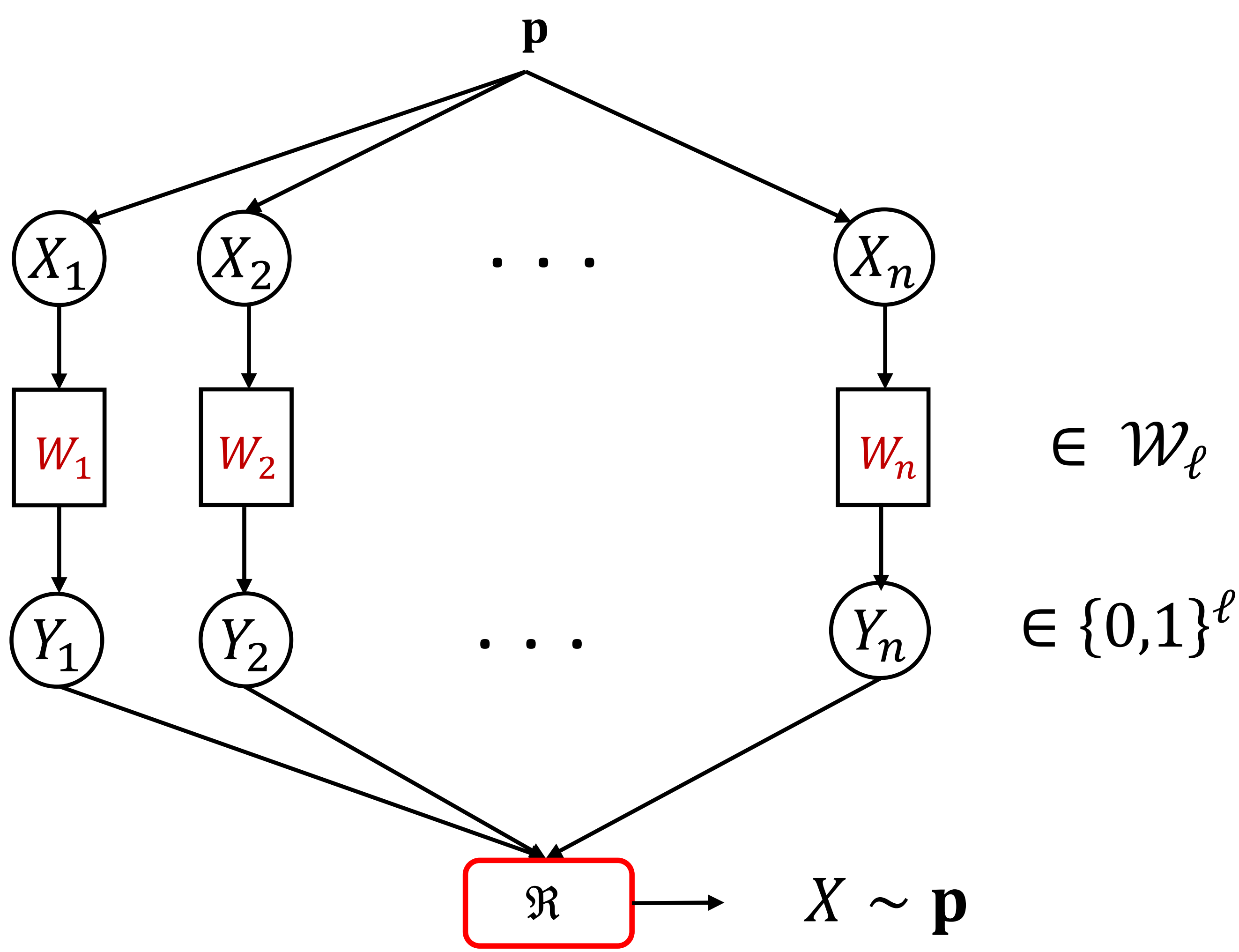
$$\begin{aligned} \mathbb{E}[\ell_1(\widehat{\mathbf{p}}, \mathbf{p})]^2 &\leq \mathbb{E}[\ell_1(\widehat{\mathbf{p}}, \mathbf{p})^2] && \textit{(Jensen)} \\ &\leq d \cdot \mathbb{E}[\ell_2^2(\widehat{\mathbf{p}}, \mathbf{p})] && \textit{(Cauchy Schwarz)} \\ &\leq \frac{d}{n} \end{aligned}$$

Under communication constraints



A simulation puzzle …

Goal: To simulate a sample from messages



One simulation to solve them all …

Theorem. Suppose simulation is possible with $f(d, \ell)$ samples.

Let T be some task with sample complexity $T(d, \varepsilon)$.

Then T can be solved with $f(d, \ell) \cdot T(d, \varepsilon)$ samples under $\mathcal{W}_\ell$.

What is $f(d, \log_2 d)$?

One simulation to solve them all …

Theorem. There is a private-coin SMP protocol with

$$f(d, \ell) \approx \max\left\{\frac{d}{2^{\ell}}, 1\right\}.$$

No protocol (even interactive) can do better!

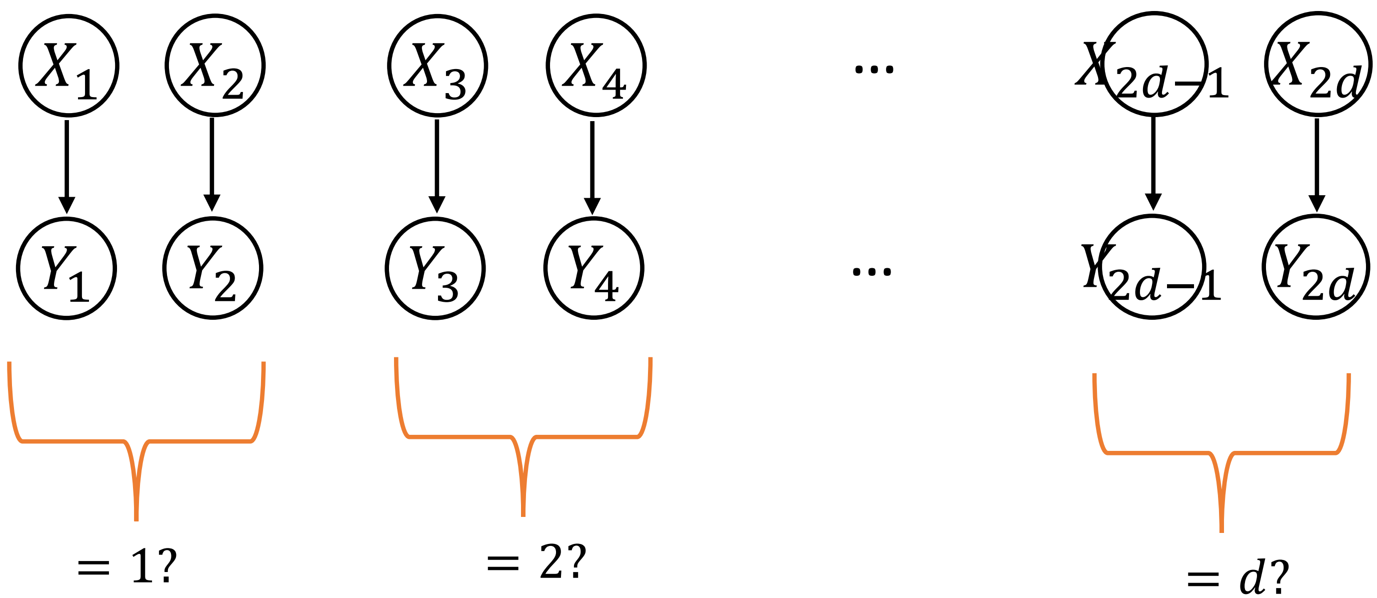
Estimation with $\Theta\left(\frac{d}{\varepsilon^2} \cdot \frac{d}{2^{\ell}}\right)$ and testing with $\Theta\left(\frac{\sqrt{d}}{\varepsilon^2} \cdot \frac{d}{2^{\ell}}\right)$

Algorithm for one-bit

Take $2d$ players and pair them into d groups:

- First pair tell if their input is symbol 1
- Second tell if their input is symbol 2
- And so on …

Algorithm for one-bit



$$Y_{2i-1} = I\{X_{2i-1} = i\}$$
$$Y_{2i} = I\{X_{2i} = i\}$$

Algorithm for one-bit

- Output $i \in [d]$ if:
  - Player $2i - 1$ is the **only** odd player sending 1
  - Player $2i$ sends 0
- If no such i, output $\perp$

Conditioned on not outputting $\perp$, output $\sim p$

Algorithm for one-bit

Player $2i - 1$ is the **only** odd player sending 1

$$\Pr(Y_{2i-1} = 1, Y_{2i'-1} = 0 \; for \; i' \neq i) = \mathbf{p}(i) \prod_{i' \neq i} (1 - \mathbf{p}(i'))$$

Player $2i$ sends 0

$$\Pr(Y_{2i} = 0) = (1 - \mathbf{p}(i))$$

$$\Pr(\text{output } i | \text{ not } \bot) = \mathbf{p}(i) \cdot \prod_{i' \in [d]} \left(1 - \mathbf{p}(i')\right) \propto \mathbf{p}(i)$$

Corollary

Inference Task	Centralized	One-bit private-SMP
Estimation	$\Theta\left(\frac{d}{\varepsilon^2}\right)$	$\Theta\left(\frac{d^2}{\varepsilon^2}\right)$
Testing	$\Theta\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$	$\Theta\left(\frac{d^{3/2}}{\varepsilon^2}\right)$

Corollary

Inference Task	Centralized	One-bit private-SMP	One-bit public-SMP
Estimation	$\Theta\left(\frac{d}{\varepsilon^2}\right)$	$\Theta\left(\frac{d^2}{\varepsilon^2}\right)$	$\Theta\left(\frac{d^2}{\varepsilon^2}\right)$
Testing:	$\Theta\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$	$\Theta\left(\frac{d^{3/2}}{\varepsilon^2}\right)$	$\Theta\left(\frac{d}{\varepsilon^2}\right)$

Bounds are tight … simulate and infer is optimal for private-coin SMP

Related work

Under SMP protocols these bounds are tight for communication constraints [HanMukherjeeOzgur19, AcharyaCanonneTyagi'19] and LDP [DuchiJordanWainwright14]

Sample complexity with interactivity and general channels?

[ACLST20] J. Acharya, C. Canonne, Y. Liu, Z. Sun, H. Tyagi, "Interactive inference under information constraints" *arXiv: 2007.10976 (in submission)*

Reminder of my time: prove lower bounds

Recipe:

- Design **hard instances** that has some structure
- Show that problem is hard within these
- Assouad's method and reduction to testing
- Bound "information contraction" due to constraints

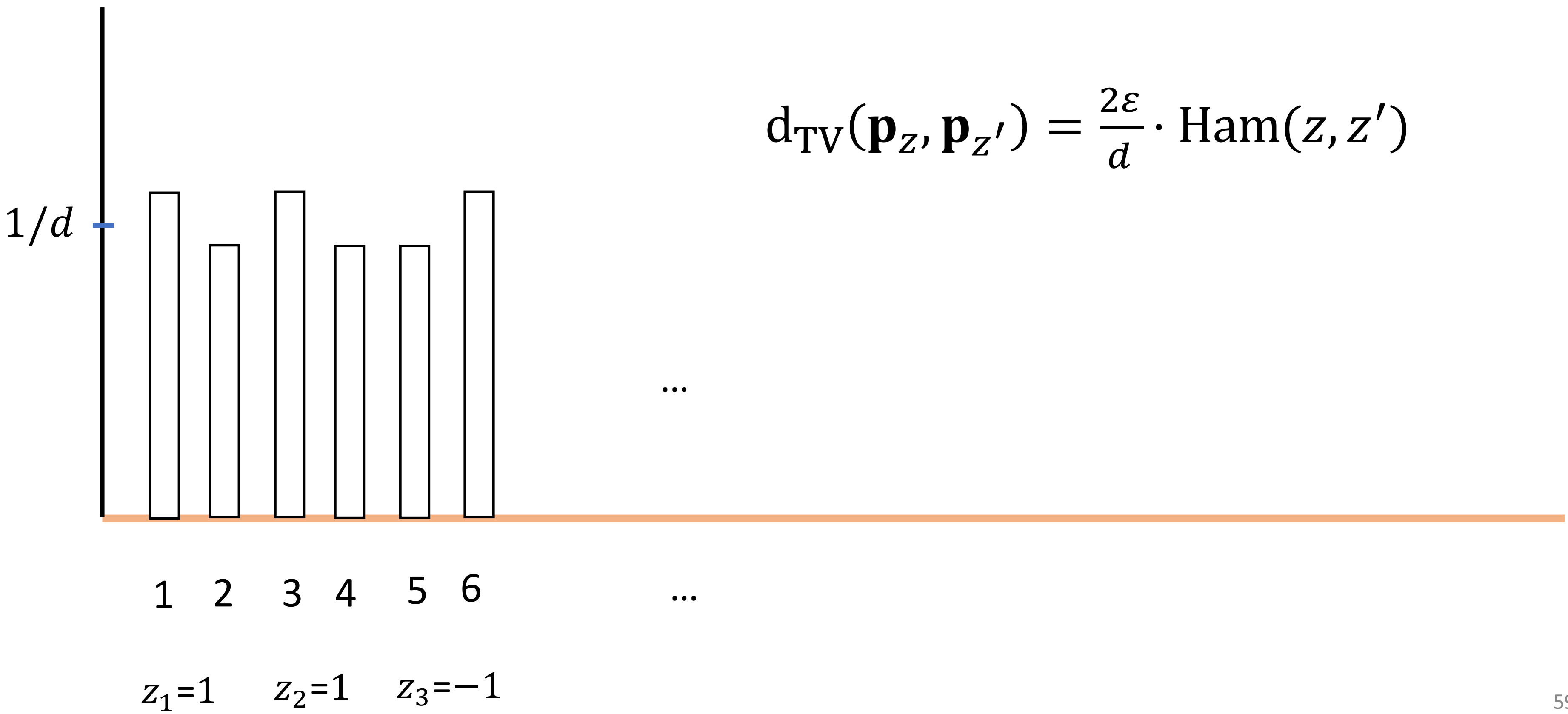
A hard instance

A hard instance

[Paninski'08] Let $\mathcal{Z} = \{-1,1\}^{d/2}$, and $\mathcal{P}_{\mathcal{Z}} = \{\mathbf{p}_z : z \in \mathcal{Z}\}$, where

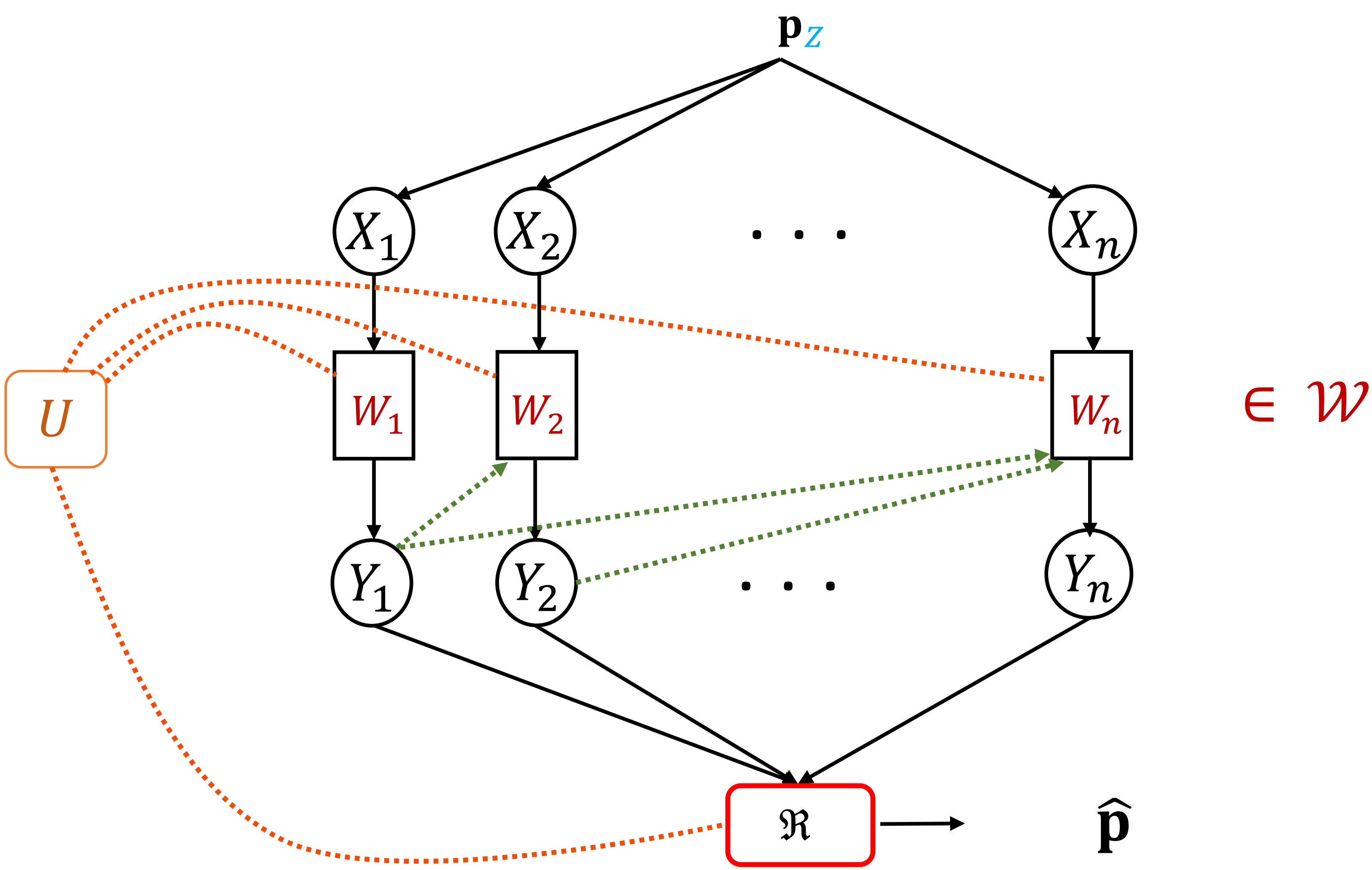
$$\mathbf{p}_z(2i-1) = \frac{1 + z_i \cdot 2\varepsilon}{d}, \qquad \mathbf{p}_z(2i) = \frac{1 - z_i \cdot 2\varepsilon}{d}, \qquad i = 1, \dots, d/2.$$

$$\mathrm{d_{TV}}(\mathbf{p}_z, \mathbf{p}_{z'}) = \frac{2\varepsilon}{d} \cdot \mathrm{Ham}(z, z')$$



Learning lower bounds

$Z = (Z_1, \dots, Z_{d/2}) \sim_{uar} \mathcal{Z}$, ie, each $Z_i \sim^{iid} \mathrm{Bern}(0.5)$



Learning lower bounds

Exercise: Let $z \in \mathcal{Z}$ and $\widehat{\mathbf{p}}$ satisfies $\mathrm{d}_{\mathrm{TV}}(\widehat{\mathbf{p}}, \mathbf{p}_z) < \frac{\varepsilon}{10}$.

Then,

$$z^* = \arg\min_{z'} \mathrm{d}_{\mathrm{TV}}(\widehat{\mathbf{p}}, \mathbf{p}_{z'})$$

satisfies

$$\mathrm{Ham}(z, z^*) < \frac{d}{10}.$$

From learning to testing

Assoud's method

If we can estimate $\mathbf{p}_Z \in_{\text{uar}} \mathcal{P}_{\mathcal{Z}}$, then we can estimate Z!

Theorem. Pick $Z \sim_{uar} \mathcal{Z}$.

If

$$\mathbb{E}_Z\left[\mathbb{E}_{\mathbf{p}_Z}[\mathrm{d}_{\mathrm{TV}}(\widehat{\mathbf{p}}(Y^n, U), \mathbf{p}_Z)]\right] < \frac{\varepsilon}{10}$$

then there exists an estimator $\hat{Z}(Y^n, U)$ such that

$$\sum_{1 \le i \le d/2} \Pr(\hat{Z}_i = Z_i) > 0.8 \times \frac{d}{2}.$$

- Note: We could write this bound as $\sum_i I(Z_i \wedge Y^n | U) = \Omega(d)$

Assouad's method

Exercise. If

$$\sum_{1 \le i \le d/2} \Pr(\hat{Z}_i = Z_i) > 0.8 \times \frac{d}{2},$$

then there exists a subset $S \subseteq \{1, \dots, d/2\}$ with $|S| > d/6$ s.t. if $i \in S$,

$$\Pr(\hat{Z}_i = Z_i) > 0.7.$$

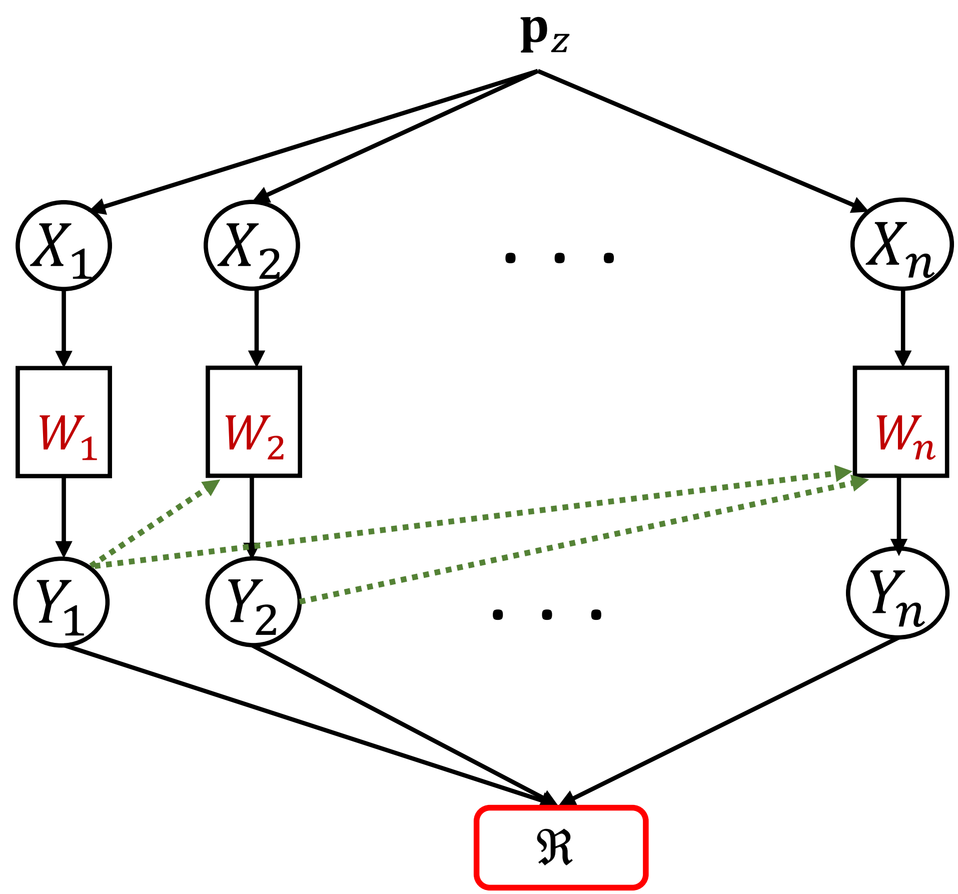
Now we need a lower bound on n for this to happen

Information bound on one coordinate

Notation

Fix $i \in [d/2]$, when can we figure Z_i?

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



Information bound on one coordinate

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

When $Z_i = 1$: $\mathbf{p}_{+i}^{Y^n} \coloneqq \frac{1}{2^{d/2-1}} \sum_{z: z_i = +1} \mathbf{p}_z^{Y^n}$

When $Z_i = -1$: $\mathbf{p}_{-i}^{Y^n} \coloneqq \frac{1}{2^{d/2-1}} \sum_{z: z_i = -1} \mathbf{p}_z^{Y^n}$

If we can guess Z_i from Y^n

$\Leftrightarrow$ $\mathrm{d_{TV}}\left(\mathbf{p}_{+i}^{Y^n}, \mathbf{p}_{-i}^{Y^n}\right)$ must be large

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

Total variation and hypothesis testing

$\mathbf{p}_1, \mathbf{p}_2$ be any two distributions over $\mathcal{Y}$

$j \in \{1,2\}$ be picked at random

Given $Y \sim \mathbf{p}_j$, design a $\hat{j}(Y)$ that is a guess for j

For any $\hat{j}(Y)$:

$$\Pr(\hat{j}(Y) = j) \leq \frac{1}{2}\left(1 + d_{\mathrm{TV}}(\mathbf{p}_1, \mathbf{p}_2)\right)$$

Information bound on one coordinate

In our case, $\mathbf{p}_1 = \mathbf{p}_{+i}^{Y^n}, \mathbf{p}_2 = \mathbf{p}_{-i}^{Y^n}$, and

$$\Pr(\hat{Z}_i = Z_i) > 0.7 \Rightarrow d_{\mathrm{TV}}\left(\mathbf{p}_{+i}^{Y^n}, \mathbf{p}_{-i}^{Y^n}\right) \geq 0.4$$

Since this holds for at least $d/6$ coordinates,

$$\sum_i d_{\mathrm{TV}}\left(\mathbf{p}_{+i}^{Y^n}, \mathbf{p}_{-i}^{Y^n}\right)^2 \geq \frac{d}{6} \times 0.16.$$

Some ingredients

$$D(\mathbf{p}_1||\mathbf{p}_2) := \sum_y \mathbf{p}_1(y) \log \frac{\mathbf{p}_1(y)}{\mathbf{p}_2(y)}, \chi^2(\mathbf{p}_1, \mathbf{p}_2) := \sum_y \frac{\left(\mathbf{p}_1(y) - \mathbf{p}_2(y)\right)^{\mathbf{2}}}{\mathbf{p}_2(y)}$$

Pinsker's inequality, convexity of logarithms:

$$2 \cdot \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_1, \mathbf{p}_2)^2 \leq D(\mathbf{p}_1||\mathbf{p}_2) \leq \chi^2(\mathbf{p}_1, \mathbf{p}_2)$$

Chain rule of KL divergence: If $\mathbf{p}_1$ and $\mathbf{p}_2$ are over $\mathcal{Y}_1 \times \mathcal{Y}_2$:

$$\begin{aligned} & D(\mathbf{p}_1(Y_1, Y_2)||\mathbf{p}_2(Y_1, Y_2)) \\ &= D(\mathbf{p}_1(Y_1)||\mathbf{p}_2(Y_1)) + \mathbb{E}_{Y_1}[D(\mathbf{p}_1(Y_2|Y_1)||\mathbf{p}_2(Y_2|Y_1))] \end{aligned}$$

KL $\leq$ chi-squared (DIY)

Since $\log(1+x) \leq x$ (why?)

$$D(\mathbf{p}||\mathbf{q}) := \sum_x \mathbf{p}(x) \log\left(1 + \frac{\mathbf{p}(x) - \mathbf{q}(x)}{\mathbf{q}(x)}\right)$$

$$\leq \sum_x \mathbf{p}(x) \frac{\big(\mathbf{p}(\boldsymbol{x}) - \mathbf{q}(\boldsymbol{x})\big)}{\mathbf{q}(\boldsymbol{x})} = \chi^2(\mathbf{p}, \mathbf{q})$$

Exercise: Prove the chain rule of KL.

Why go to KL?

By Pinsker's inequality,

$$4 \cdot d_{\mathrm{TV}}\left(\mathbf{p}_{+i}^{Y^n}, \mathbf{p}_{-i}^{Y^n}\right)^2 \leq \left(D\left(\mathbf{p}_{+i}^{Y^n} || \mathbf{p}_{-i}^{Y^n}\right) + D\left(\mathbf{p}_{-i}^{Y^n} || \mathbf{p}_{+i}^{Y^n}\right)\right)$$

Summing over i,

$$\sum_i \left(D\left(\mathbf{p}_{+i}^{Y^n} || \mathbf{p}_{-i}^{Y^n}\right) + D\left(\mathbf{p}_{-i}^{Y^n} || \mathbf{p}_{+i}^{Y^n}\right)\right)$$
$$\geq \sum_i 4 \cdot d_{\mathrm{TV}}\left(\mathbf{p}_{+i}^{Y^n}, \mathbf{p}_{-i}^{Y^n}\right)^2 \geq 4 \cdot \frac{d}{6} \times 0.16 \geq \frac{d}{10}$$

$\mathbf{p}_{+i}^{Y^n}$ are mixture distributions!

Handling mixtures is painful, leads to issues to extend SMP lower bounds to interactive setting

Convexity to the rescue

Exercise: KL divergence is convex.

For any distributions $\mathbf{p}_1, \mathbf{p}_2$ and $\mathbf{q}_1, \mathbf{q}_2$ and $\lambda \in [0,1]$,

$$D(\lambda \mathbf{p}_1 + (1-\lambda)\mathbf{q}_1 || \lambda \mathbf{p}_2 + (1-\lambda)\mathbf{q}_2)$$
$$\leq \lambda \cdot D(\mathbf{p}_1 || \mathbf{p}_2) + (1-\lambda) \cdot D(\mathbf{p}_1 || \mathbf{p}_2)$$

Prove using concavity of logarithms

Convexity to handle mixtures

$z \in \{-1,1\}^{k/2}$, $z^{\oplus i}$ obtained by flipping the ith coordinate of z

Theorem.

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

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 when all others are fixed!

Convexity to handle mixtures

Summing over i

$$\frac{d}{20} \leq \mathbb{E}_Z\left[\sum_i D\left(\mathbf{p}_Z^{Y^n} || \mathbf{p}_{Z\oplus i}^{Y^n}\right)\right]$$

- For given Z the sum is divergences when changing one coordinate

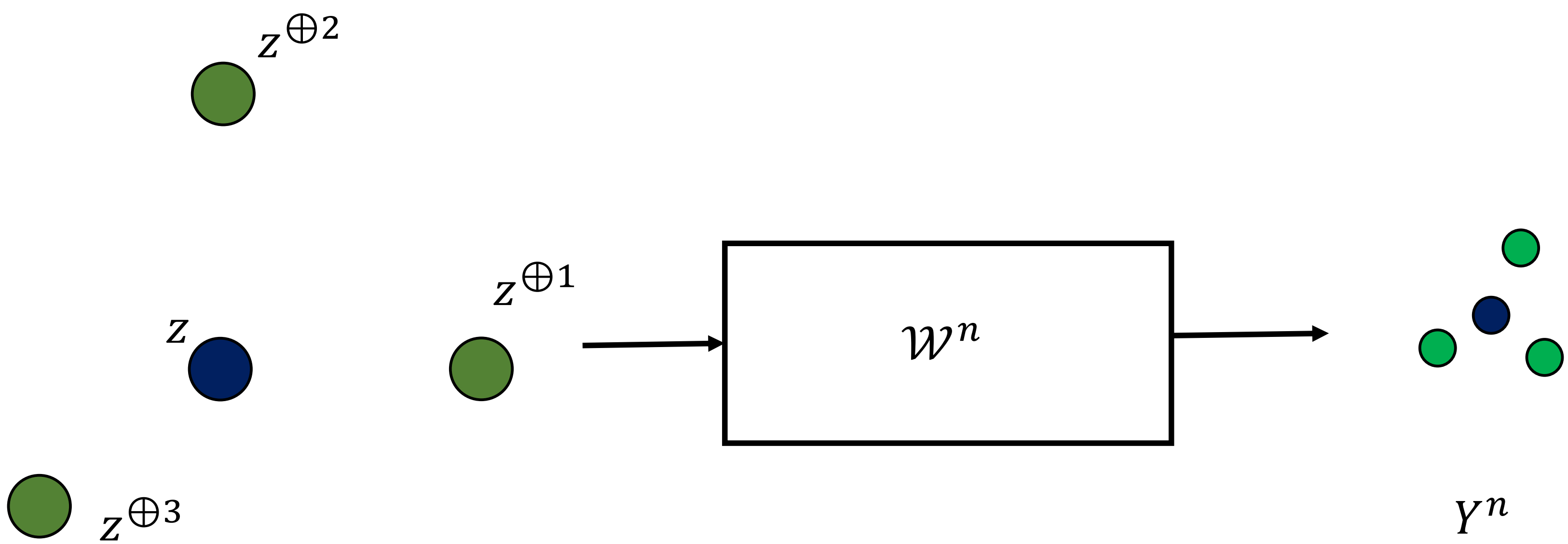
Focus on one z

By expectation<max, and linearity of expectations,

$$\frac{d}{20} \le \max_{z} \left[\sum_{i} D\left(\mathrm{p}_{z}^{Y^n} || \mathrm{p}_{z^{\oplus i}}^{Y^n} \right) \right]$$

** the following is the original bound in terms of MI:

$$\sum_{i} I(Z_i \wedge Y^n) \le \frac{1}{2} \cdot \max_{z} \left[\sum_{i} D\left(\mathrm{p}_{z}^{Y^n} || \mathrm{p}_{z^{\oplus i}}^{Y^n} \right) \right]$$

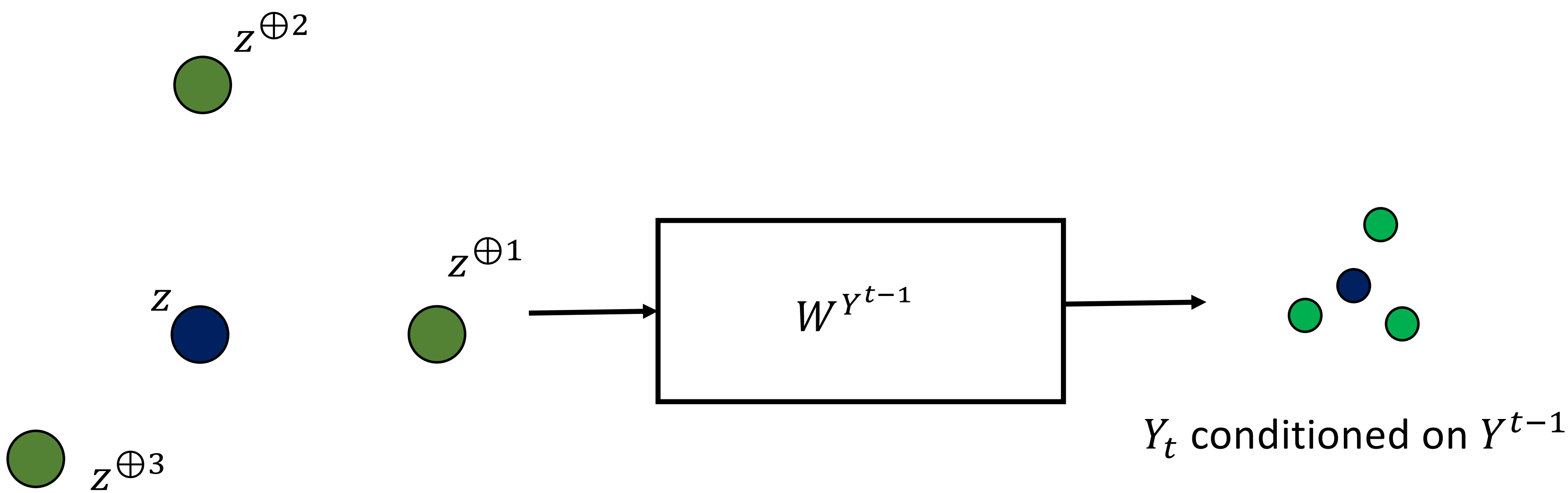


Bounding $\sum_i D\left(\mathrm{p}_z^{Y^n} || \mathrm{p}_{z\oplus i}^{Y^n}\right)$

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 $W^{Y^{t-1}}$

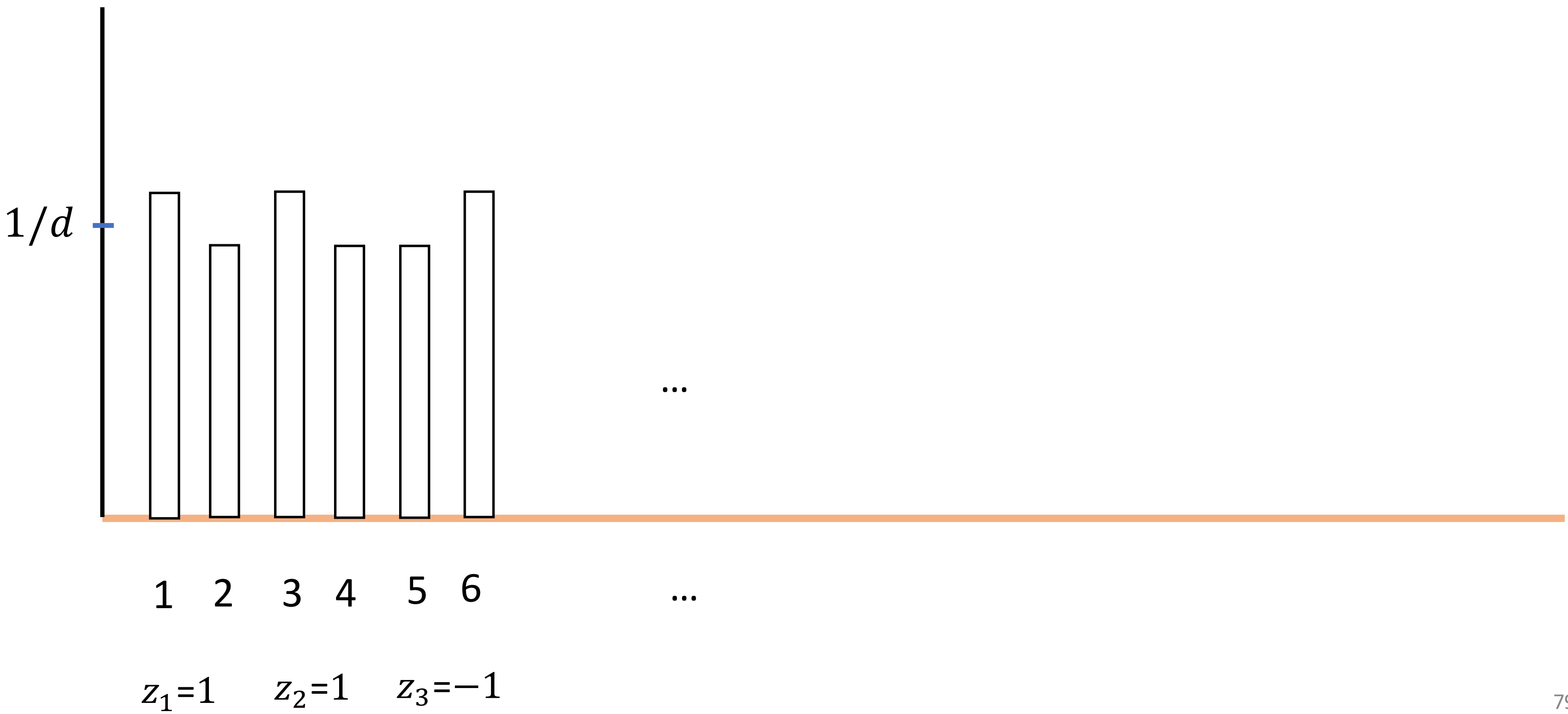


Recall

For $z \in \{-1,1\}^{d/2}$,

$$\mathbf{p}_z(2i-1) = \frac{1 + z_i 2\varepsilon}{d}, \qquad \mathbf{p}_z(2i) = \frac{1 - z_i 2\varepsilon}{d}, \qquad i = 1, \dots, d/2.$$

$\mathbf{p}_z$ and $\mathbf{p}_{z \oplus i}$ differ only on $2i-1$ and $2i$

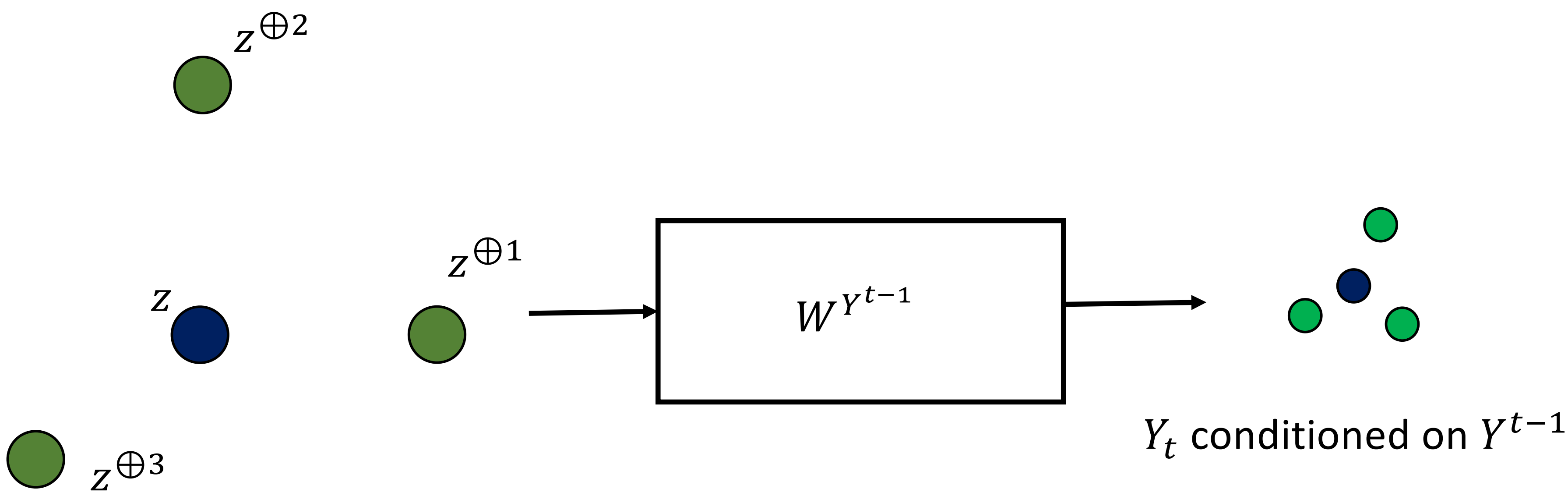


Bounding $\sum_i D\left(\mathbf{p}_z^{Y_t|Y^{t-1}} || \mathbf{p}_{z\oplus i}^{Y_t|Y^{t-1}}\right)$

$\mathbf{p}_z$ and $\mathbf{p}_{z\oplus i}$ differ only on $2i-1$ and $2i$ by $4\varepsilon z_i/d$

- Fix Y^{t-1}

$$\mathbf{p}_z^{Y_t|Y^{t-1}}(y) = \mathbf{p}_{z\oplus i}^{Y_t|Y^{t-1}}(y) + \frac{4\varepsilon z_i}{d}\left(W^{Y^{t-1}}(y|2i-1) - W^{Y^{t-1}}(y|2i)\right)$$



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

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

$$\sum_i D\left(\mathbf{p}_Z^{Y_t|Y_{t-1}} || \mathbf{p}_{Z\oplus i}^{Y_t|Y_{t-1}}\right) \leq \sum_i \sum_y \frac{\left(\mathbf{p}_Z^{Y_t}(y) - \mathbf{p}_{Z\oplus i}^{Y_t}(y)\right)^2}{\mathbf{p}_{Z\oplus i}^{Y_t}(y)}$$

$$\leq \frac{8\varepsilon^2}{d} \cdot \sum_i \sum_y \frac{\left(W(y|2i-1) - W(y|2i)\right)^2}{\sum_x W(y|x)}$$

Recall

$$\mathbf{p}_Z(2i-1) = \frac{1 + Z_i \varepsilon}{d}, \qquad \mathbf{p}_Z(2i) = \frac{1 - Z_i \varepsilon}{d}$$

$|W(y|2i-1) - W(y|2i)|$ large $\Leftrightarrow$ seeing y tells about Z_i

An average information contraction bound

Theorem. [ACLST20] Under any **interactive protocol**,

$$\sum_i I(Z_i \wedge Y^n) \leq n \cdot \frac{8\varepsilon^2}{d} \cdot \sup_{W \in \mathcal{W}} \sum_i \sum_y \frac{\left(W(y|2i-1) - W(y|2i)\right)^2}{\sum_x W(y|x)}$$

Theorem. If there exists an estimator then

$$\frac{d}{20} \leq n \cdot \frac{8\varepsilon^2}{d} \cdot \sup_{W \in \mathcal{W}} \sum_i \sum_y \frac{\left(W(y|2i-1) - W(y|2i)\right)^2}{\sum_x W(y|x)}$$

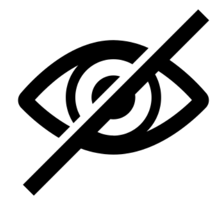
Applications

For any $W \in \mathcal{W}_\ell$

$$\sum_i \sum_y \frac{\left(W(y|2i-1) - W(y|2i)\right)^2}{\sum_x W(y|x)} \leq 2^\ell$$

For any $W \in \mathcal{W}_\varrho, \varrho \leq 1$

$$\sum_i \sum_y \frac{\left(W(y|2i-1) - W(y|2i)\right)^2}{\sum_x W(y|x)} = O(\varrho^2)$$



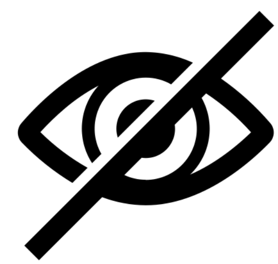
Interactive lower bound for estimation

$$\frac{d}{20} \leq n \cdot \frac{8\varepsilon^2}{d} \cdot 2^{\ell}$$

$$n = \Omega\left(\frac{d^2}{2^{\ell}\varepsilon^2}\right)$$

$$\frac{d}{20} \leq n \cdot \frac{8\varepsilon^2}{d} \cdot \varrho^2$$

$$n = \Omega\left(\frac{d^2}{\varepsilon^2\varrho^2}\right)$$



Plug-n-play bounds

$H(W)$ is a $\frac{d}{2} \times \frac{d}{2}$ PSD matrix:

$$\left(H(W)\right)_{ij} :=$$

$$\sum_{y \in Y} \frac{\left(W(y|2i-1) - W(y|2i)\right)\left(W(y|2j-1) - W(y|2j)\right)}{\sum_j W(y|j)}$$

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

Plug-n-play bounds

$$\| \mathcal{W} \| \stackrel{\text{def}}{=} \max_{W \in \mathcal{W}} \| H(W) \|$$

Testing:

Classic	Private-coin SMP	Public-coin SMP	Sequentially Interactive
$\Omega\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$	$\Omega\left(\frac{d^{3/2}}{\varepsilon^2 \| \mathcal{W} \|_*}\right)$	$\Omega\left(\frac{d}{\varepsilon^2 \| \mathcal{W} \|_{\mathrm{F}}}\right)$	$\Omega\left(\frac{d}{\varepsilon^2 \sqrt{\| \mathcal{W} \|_{OP} \| \mathcal{W} \|_*}}\right)$

Estimation

Classic	Sequentially Interactive
$\Omega\left(\frac{d}{\varepsilon^2}\right)$	$\Omega\left(\frac{d^2}{\varepsilon^2 \| \mathcal{W} \|_*}\right)$

Next 45 minutes:

Reinforcement Learning by Himanshu Tyagi …

References (click to go)

2021 (6)

On Learning Parametric Distributions from Quantized Samples. Septimia Sarbu; and Abdellatif Zaidi. In *Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT'21)*, June 2021.
bibtex ▾

Inference Under Information Constraints III: Local Privacy Constraints. Jayadev Acharya; Clément L. Canonne; Cody Freitag; Ziteng Sun; and Himanshu Tyagi. *IEEE J. Sel. Areas Inf. Theory*, 2(1): 253–267. 2021.
bibtex ▾

Unified lower bounds for interactive high-dimensional estimation under information constraints. Jayadev Acharya; Clément L. Canonne; Zuteng Sun; and Himanshu Tyagi. *CoRR*, abs/2010.06562. 2021.
bibtex ▾

Information-constrained optimization: can adaptive processing of gradients help?. Jayadev Acharya; Clément L. Canonne; Prathamesh Mayekar; and Himanshu Tyagi. *CoRR*, abs/2104.00979. 2021.
bibtex ▾

Optimal Rates for Nonparametric Density Estimation under Communication Constraints. Jayadev Acharya; Clément L. Canonne; Aditya Vikram Singh; and Himanshu Tyagi. *CoRR*, abs/2107.10078. 2021.
bibtex ▾

Local Differential Privacy Is Equivalent to Contraction of E_γ-Divergence. Shahab Asoodeh; Maryam Aliakbarpour; and Flávio P. Calmon. *CoRR*, abs/2102.01258. 2021.
bibtex ▾

2020 (12)

Interactive Inference under Information Constraints. Jayadev Acharya; Clément L. Canonne; Yuhan Liu; Ziteng Sun; and Himanshu Tyagi. *CoRR*, abs/2007.10976. 2020.
Paper bibtex ▾

Fisher Information Under Local Differential Privacy. Leighton Pate Barnes; Wei-Ning Chen; and Ayfer Özgür. *IEEE J. Sel. Areas Inf. Theory*, 1(3): 645–659. 2020.
bibtex ▾

Geometric Lower Bounds for Distributed Parameter Estimation under Communication Constraints. Yanjun Han; Ayfer Özgür; and Tsachy Weissman. *ArXiv e-prints*, abs/1802.08417v3. September 2020.
bibtex ▾

Inference under information constraints I: Lower bounds from chi-square contraction. Jayadev Acharya; Clément L. Canonne; and Himanshu Tyagi. *IEEE Trans. Inform. Theory*, 66(12): 7835–7855. 2020. Preprint available at arXiv:abs/1812.11476.
Paper doi bibtex ▾

Inference Under Information Constraints II: Communication Constraints and Shared Randomness. Jayadev Acharya; Clément L. Canonne; and Himanshu Tyagi. *IEEE Trans. Inf. Theory*, 66(12): 7856–7877. 2020.
bibtex ▾

Domain Compression and its Application to Randomness-Optimal Distributed Goodness-of-Fit. Jayadev Acharya; Clément L. Canonne; Yanjun Han; Ziteng Sun; and Himanshu Tyagi. In *COLT*, volume 125, of *Proceeding of Machine Learning Research*, pages 3–40, 2020. PMLR
bibtex ▾

Distributed Signal Detection under Communication Constraints. Jayadev Acharya; Clément L. Canonne; and Himanshu Tyagi. In *COLT*, volume 125, of *Proceedings of Machine Learning Research*, pages 41–63, 2020. PMLR
bibtex ▾

Lecture notes on: Information-theoretic methods for high-dimensional statistics. Yihong Wu. 2020.
Paper bibtex ▾

Lower bounds for learning distributions under communication constraints via fisher information. Leighton Pate Barnes; Yanjun Han; and Ayfer Özgür. *J. Mach. Learn. Res.*, 21: Paper No. 236, 30. 2020.
bibtex ▾

Private Identity Testing for High-Dimensional Distributions. Clément L. Canonne; Gautam Kamath; Audra McMillan; Jonathan Ullman; and Lydia Zakynthinou. In *Advances in Neural Information Processing Systems 33*, 2020. Preprint available at arXiv:abs/1905.11947
bibtex ▾

Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. Thomas Berrett; and Cristina Butucea. In *NeurIPS*, 2020.
bibtex ▾

Local differential privacy: elbow effect in optimal density estimation and adaptation over Besov ellipsoids. Cristina Butucea; Amandine Dubois; Martin Kroll; and Adrien Saumard. *Bernoulli*, 26(3): 1727–1764. 2020.
Paper doi bibtex ▾

References

2019 (5)

Locally Private Gaussian Estimation. Matthew Joseph; Janardhan Kulkarni; Jieming Mao; and Steven Z. Wu. In H. Wallach; H. Larochelle; A. Beygelzimer; F.; E. Fox; and R. Garnett., editor(s), *Advances in Neural Information Processing Systems 32*, pages 2984–2993. Curran Associates, Inc., 2019.
bibtex

Fisher Information for Distributed Estimation under a Blackboard Communication Protocol. Leighton P. Barnes; Yanjun Han; and Ayfer Özgür. In *ISIT*, pages 2704–2708, 2019. IEEE
bibtex

Lower Bounds for Locally Private Estimation via Communication Complexity. John Duchi; and Ryan Rogers. In Alina Beygelzimer; and Daniel Hsu., editor(s), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, of *Proceedings of Machine Learning Research*, pages 1161–1191, Phoenix, USA, June 2019. PMLR
bibtex

Hadamard Response: Estimating Distributions Privately, Efficiently, and with Little Communication. Jayadev Acharya; Ziteng Sun; and Huanyu Zhang. In Kamalika Chaudhuri; and Masashi Sugiyama., editor(s), *Proceedings of Machine Learning Research*, volume 89, pages 1120–1129, 16–18 Apr 2019. PMLR
Paper bibtex

Communication and Memory Efficient Testing of Discrete Distributions. Ilias Diakonikolas; Themis Gouleakis; Daniel M. Kane; and Sankeerth Rao. In *COLT*, volume 99, of *Proceedings of Machine Learning Research*, pages 1070–1106, 2019. PMLR
bibtex

2018 (4)

Geometric Lower Bounds for Distributed Parameter Estimation under Communication Constraints. Yanjun Han; Ayfer Özgür; and Tsachy Weissman. In *Proceedings of the 31st Conference on Learning Theory, COLT 2018*, volume 75, of *Proceedings of Machine Learning Research*, pages 3163–3188, 2018. PMLR The arXiv (v3) version from 2020 corrects some issues and includes more results.
bibtex

Distributed Statistical Estimation of High-Dimensional and Non-parametric Distributions. Yanjun Han; Pritam Mukherjee; Ayfer Özgür; and Tsachy Weissman. In *Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT'18)*, pages 506–510, 2018.
bibtex

Minimax optimal procedures for locally private estimation. John C. Duchi; Michael I. Jordan; and Martin J. Wainwright. *J. Amer. Statist. Assoc.*, 113(521): 182–201. 2018.
bibtex

Optimal schemes for discrete distribution estimation under locally differential privacy. Min Ye; and Alexander Barg. *IEEE Trans. Inform. Theory*, 64(8): 5662–5676. 2018.
Paper doi bibtex

2017 (1)

Information-theoretic lower bounds on Bayes risk in decentralized estimation. Aolin Xu; and Maxim Raginsky. *IEEE Transactions on Information Theory*, 63(3): 1580–1600. 2017.
bibtex

2016 (1)

Communication lower bounds for statistical estimation problems via a distributed data processing inequality. Mark Braverman; Ankit Garg; Tengyu Ma; Huy L. Nguyen; and David P. Woodruff. In *Symposium on Theory of Computing Conference, STOC'16*, pages 1011–1020, 2016. ACM
bibtex

2014 (2)

On Communication Cost of Distributed Statistical Estimation and Dimensionality. Ankit Garg; Tengyu Ma; and Huy L. Nguyen. In *Advances in Neural Information Processing Systems 27*, pages 2726–2734, 2014.
bibtex

Fundamental limits of online and distributed algorithms for statistical learning and estimation. Ohad Shamir. In *Advances in Neural Information Processing Systems 27*, pages 163–171, 2014.
bibtex

2013 (1)

Information-theoretic lower bounds for distributed statistical estimation with communication constraints. Yuchen Zhang; John Duchi; Michael I. Jordan; and Martin J. Wainwright. In *Advances in Neural Information Processing Systems 26*, pages 2328–2336, 2013.
bibtex

2009 (1)

Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. Martin J. Wainwright. *IEEE Trans. Inform. Theory*, 55(12): 5728–5741. 2009.
Paper doi bibtex

Some references and previous work

Some references and previous work

Too many for a single slide, or two. Starts, more or less, with Tsitsiklis’89, picks up again in the mid-2000’s with a slightly different focus: local privacy, various types of communication constraints, ML-related motivations…

For a detailed bibliography:

www.cs.columbia.edu/~ccanonne/tutorial-focs2020/bibliography.html

Now you all say … Phew!