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

Jayadev Acharya, **Clément Canonne**, Himanshu Tyagi

FOCS 2020

Part I: What is this all about?

Techniques and recipes for distributed learning and testing under constraints

Techniques and recipes for distributed **learning** and **testing** under constraints

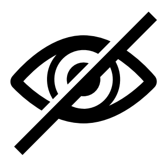
Parameter/density estimation

Goodness-of-fit / identity testing

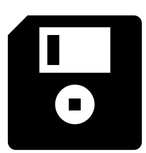
Techniques and recipes for distributed **learning** and **testing** under **constraints**

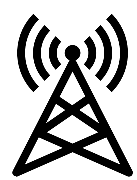
Parameter/density estimation

Goodness-of-fit / hypothesis testing



Local constraints

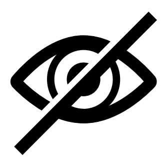




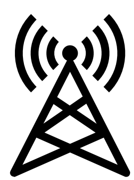
Techniques and recipes for distributed **learning** and **testing** under **constraints**

Parameter/density estimation

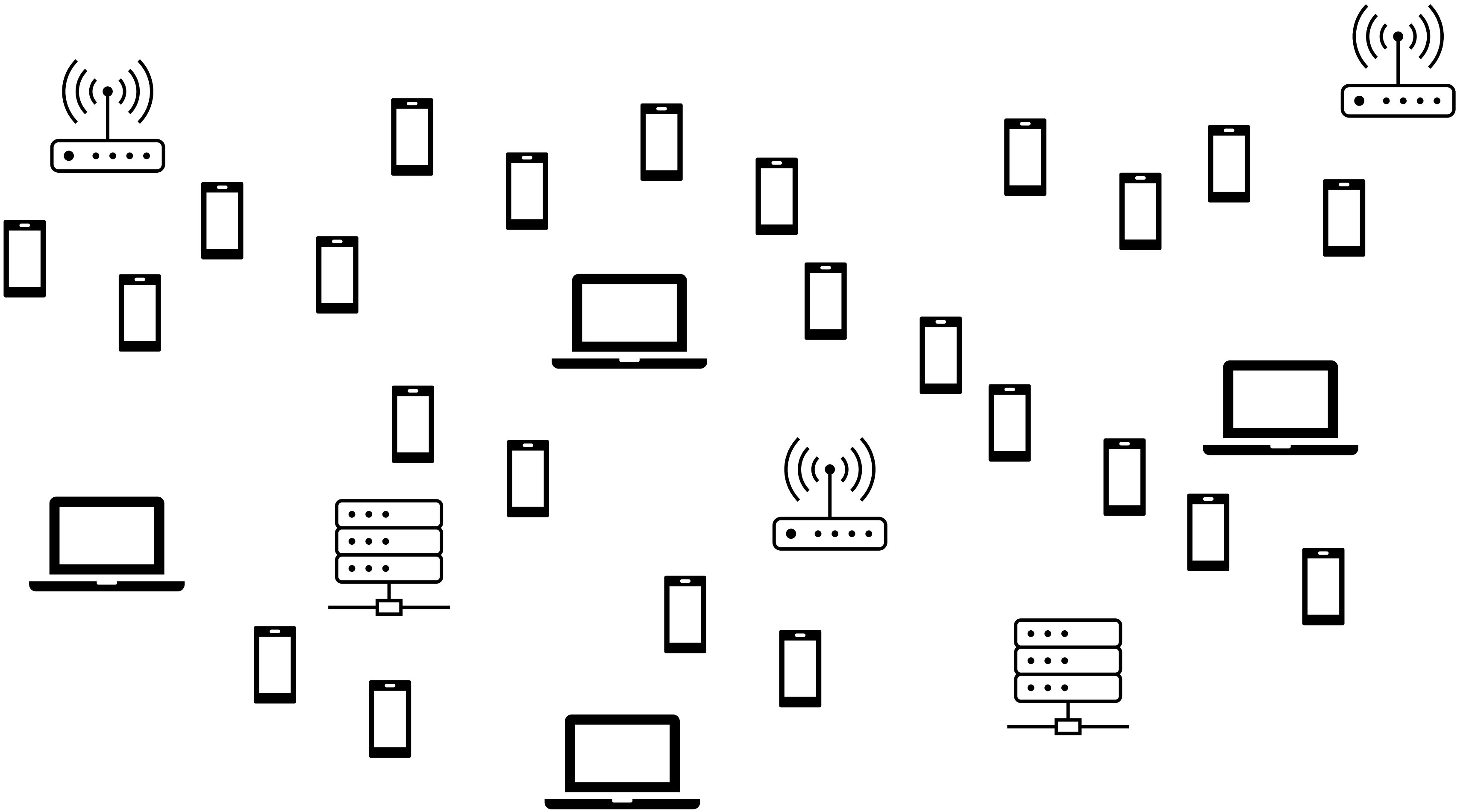
Goodness-of-fit / hypothesis testing



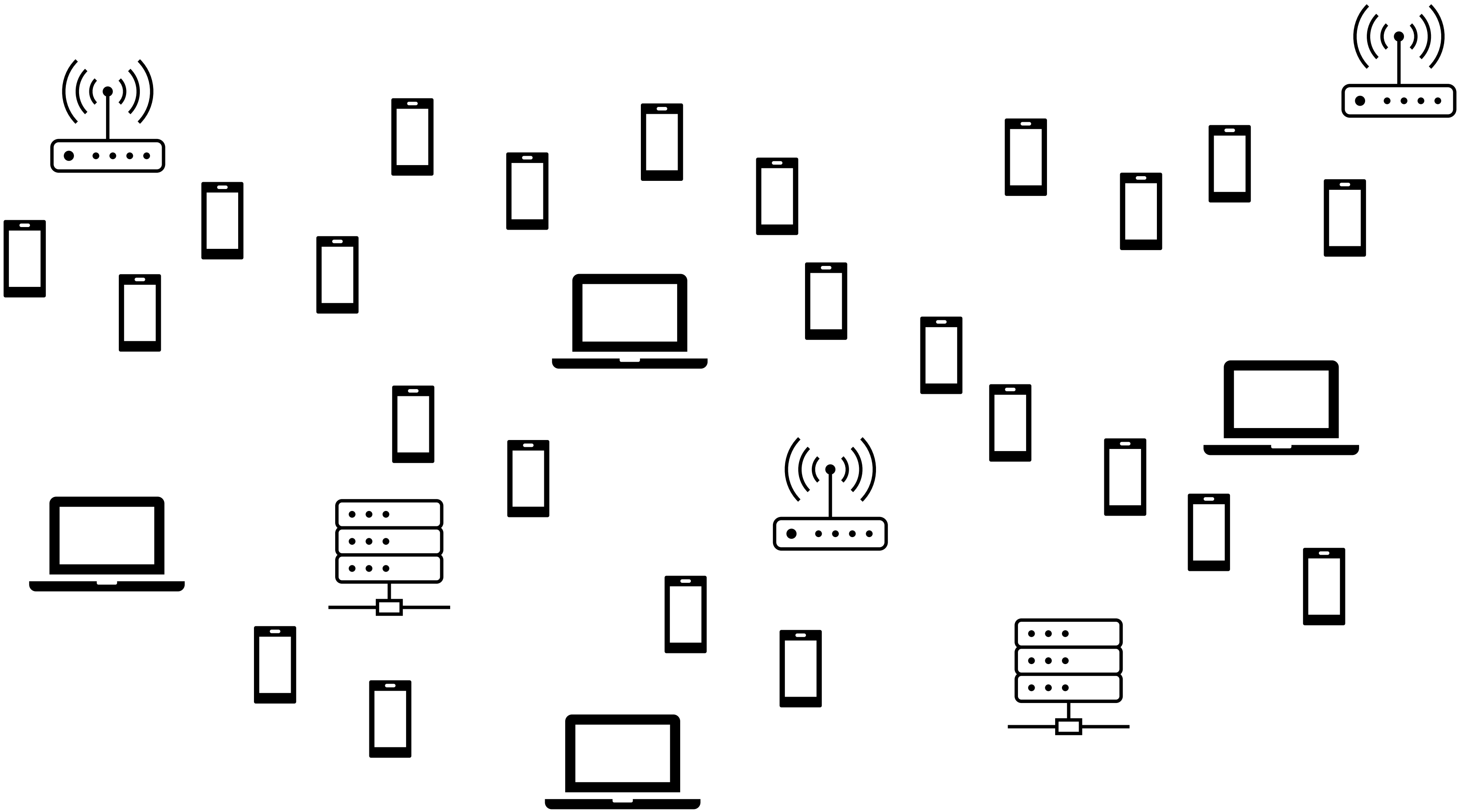
Local constraints



Techniques and recipes for **distributed** learning and testing under constraints

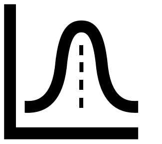


Techniques and recipes for **distributed** learning and testing under constraints



("Motivation" slide)

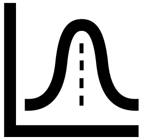
An example: high-dimensional Gaussian



$X^n := (X_1, \ldots, X_n)$: samples from an unknown $\mathcal{N}(\boldsymbol{\mu}, \mathbb{I}_{\mathbf{d}})$.

Goal: learn $\boldsymbol{\mu}$ to ℓ_2 error ε.

An example: high-dimensional Gaussian



$X^n := (X_1, \dots, X_n)$: samples from an unknown $\mathcal{N}(\boldsymbol{\mu}, \mathbb{I}_{\mathbf{d}})$.

Goal: learn $\boldsymbol{\mu}$ to ℓ_2 error ε.

Theorem. Without constraints, in the centralized setting, $n = \Theta\left(\frac{d}{\varepsilon^2}\right)$ samples are necessary and sufficient.

An example: high-dimensional Gaussian

$X^n \coloneqq (X_1, \ldots, X_n)$: samples from an unknown $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I_d})$.

Goal: learn $\boldsymbol{\mu}$ to ℓ_2 error ε.

Theorem. Without constraints, in the centralized setting, $n = \Theta\left(\frac{d}{\varepsilon^2}\right)$ samples are necessary and sufficient.

"Folklore/easy"

An example: high-dimensional Gaussian

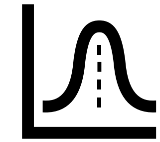
$X^n := (X_1, \dots, X_n)$: samples from an unknown $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I_d})$.

Goal: learn $\boldsymbol{\mu}$ to ℓ_2 error ε.

Theorem. Without constraints, in the centralized setting, $n = \Theta\left(\frac{d}{\varepsilon^2}\right)$ samples are necessary and sufficient.

But how do we prove an analogue under local privacy (LDP)? Under communication constraints? With/without interaction?

An example: high-dimensional Gaussian



$X^n := (X_1, \dots, X_n)$: samples from an unknown $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I_d})$.

Goal: learn $\boldsymbol{\mu}$ to ℓ_2 error ε.

Theorem. Without constraints, in the centralized setting, $n = \Theta\left(\frac{d}{\varepsilon^2}\right)$ samples are necessary and sufficient.

But how do we prove an analogue under local privacy (LDP)? Under communication constraints? With/without interaction?

Theorem. Under ρ–LDP, $n = \Theta\left(\frac{d^2}{\varepsilon^2 \rho^2}\right)$ samples are necessary and sufficient.

Goal of this tutorial, refined

General, **re-usable** techniques to establish **lower bounds** on the **sample complexity** of such **distributed**/**constrained** statistical problems (in various settings)

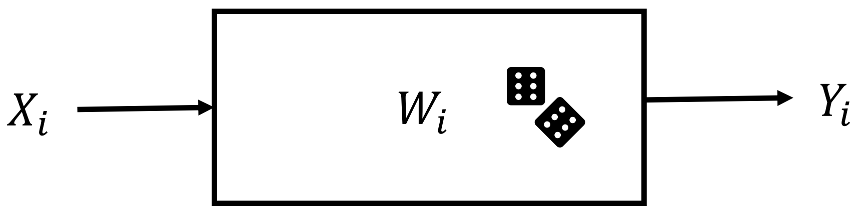
Some definitions before we start

Modeling the constraints [ACT20c]

n users, user i observes X_i and sends message Y_i

$\mathcal{X}$: domain of the unknown $\mathbf{p}$

$\mathcal{Y}$: message space



$$W_i(y|x) \coloneqq \Pr(Y_i = y | X_i = x)$$

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

The algorithm/protocol dictates how user i chooses W_i from $\mathcal{W}$

Example 1: Communication constraints

[Sha14,HMÖW18,ACT20d...]

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

Each X_i is mapped to ℓ bits.

Tight bandwidth constraints

Example 2: Local Differential Privacy (LDP)

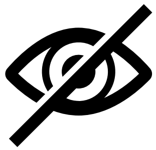
[Warner65, EPR03, KLNRS11]

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

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

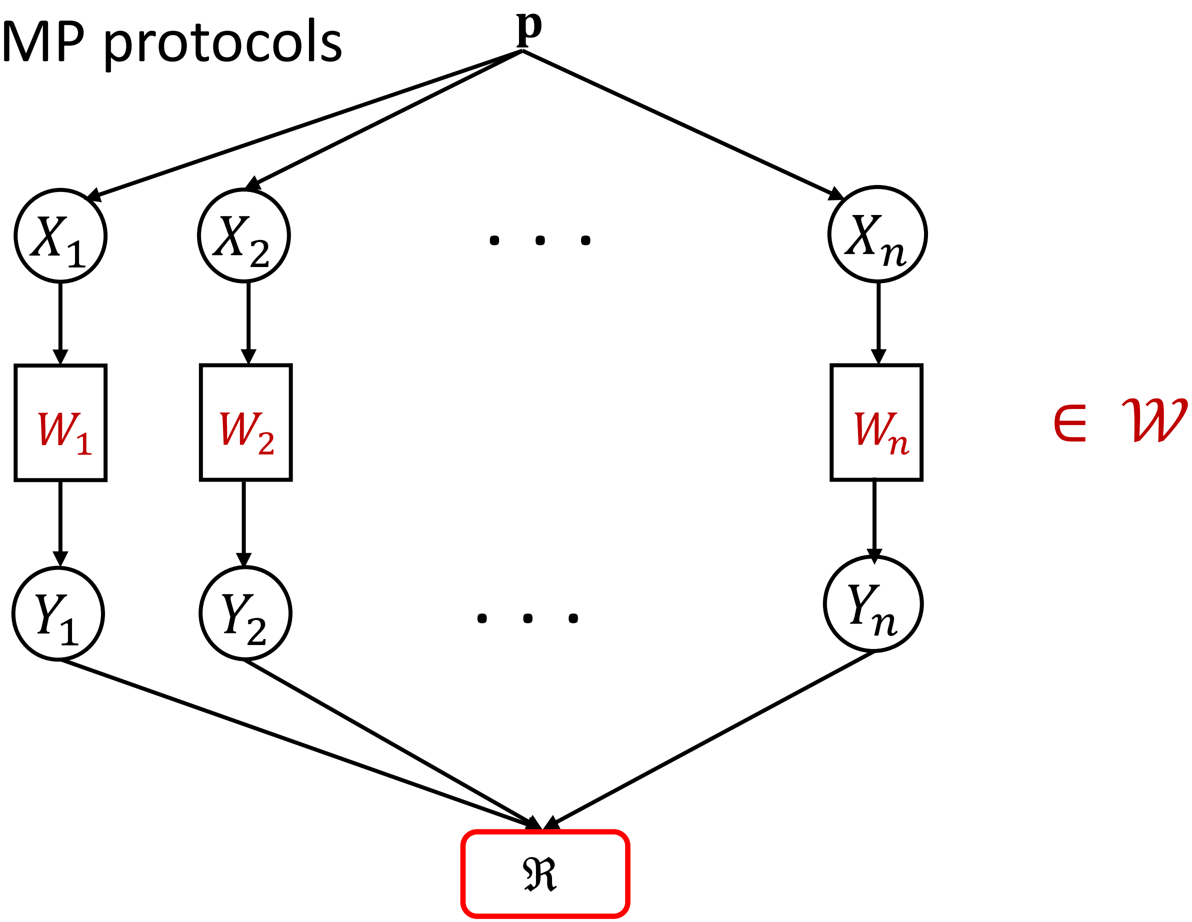
$$\mathcal{W}_{\varrho} = \{\text{all } \varrho - \text{LDP channels}\}$$

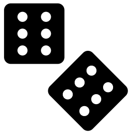
Privacy guarantees even "against" the server



Types of protocols

Private-coin SMP protocols

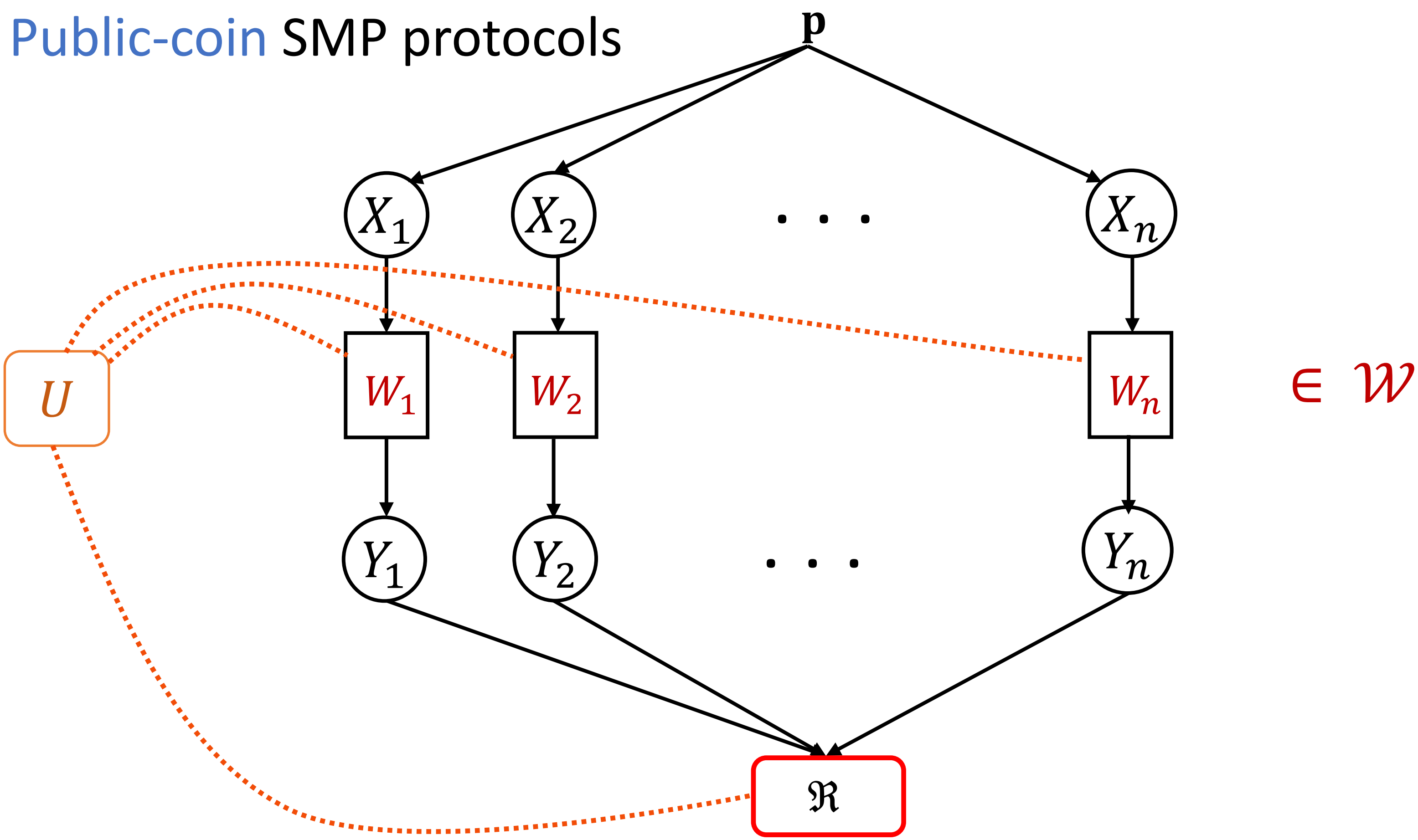


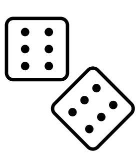


Noninteractive ("simultaneous message-passing"), no common random seed

Types of protocols

Public-coin SMP protocols

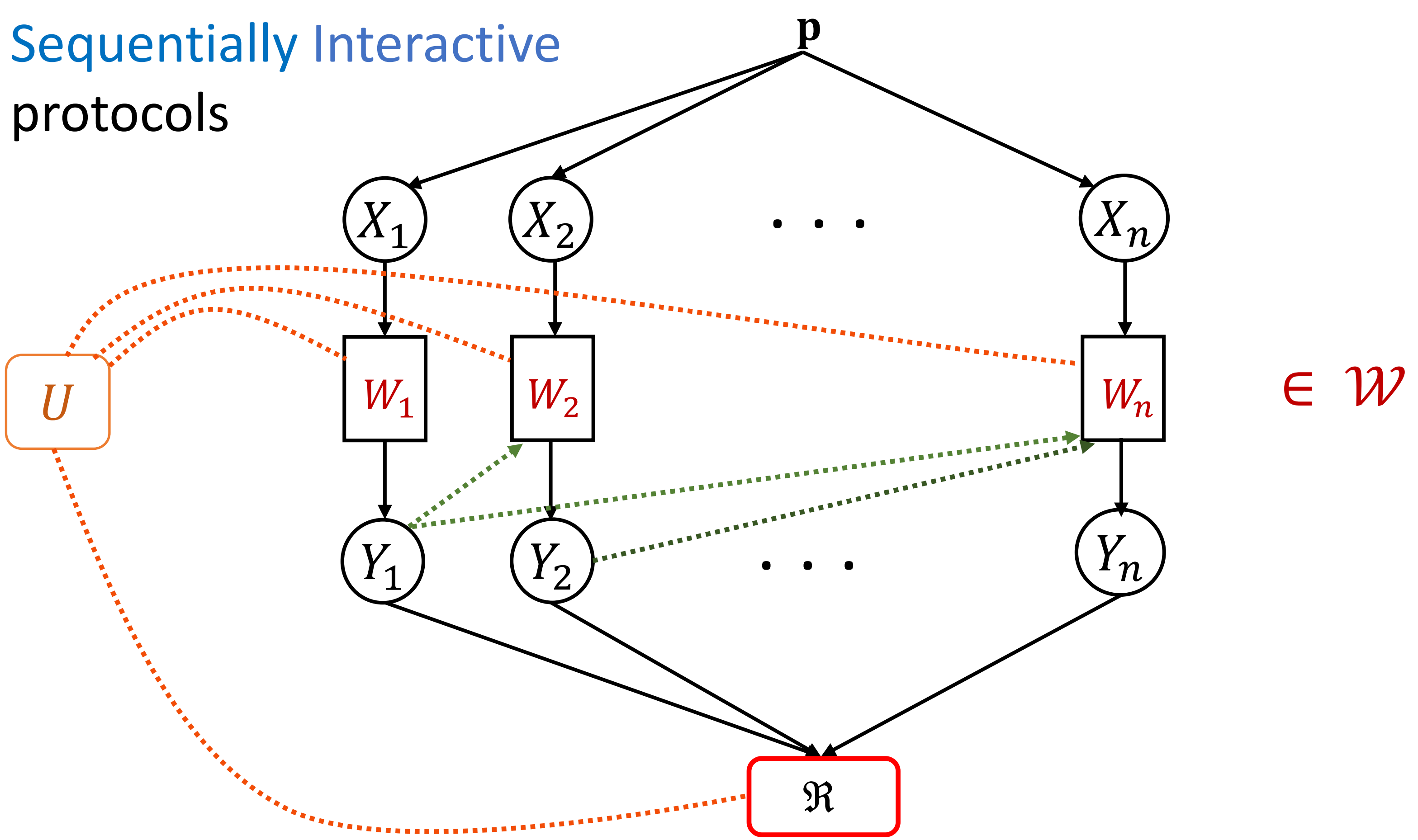


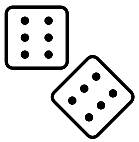


Noninteractive ("simultaneous message-passing"), *but* common random seed

Types of protocols

Sequentially Interactive protocols

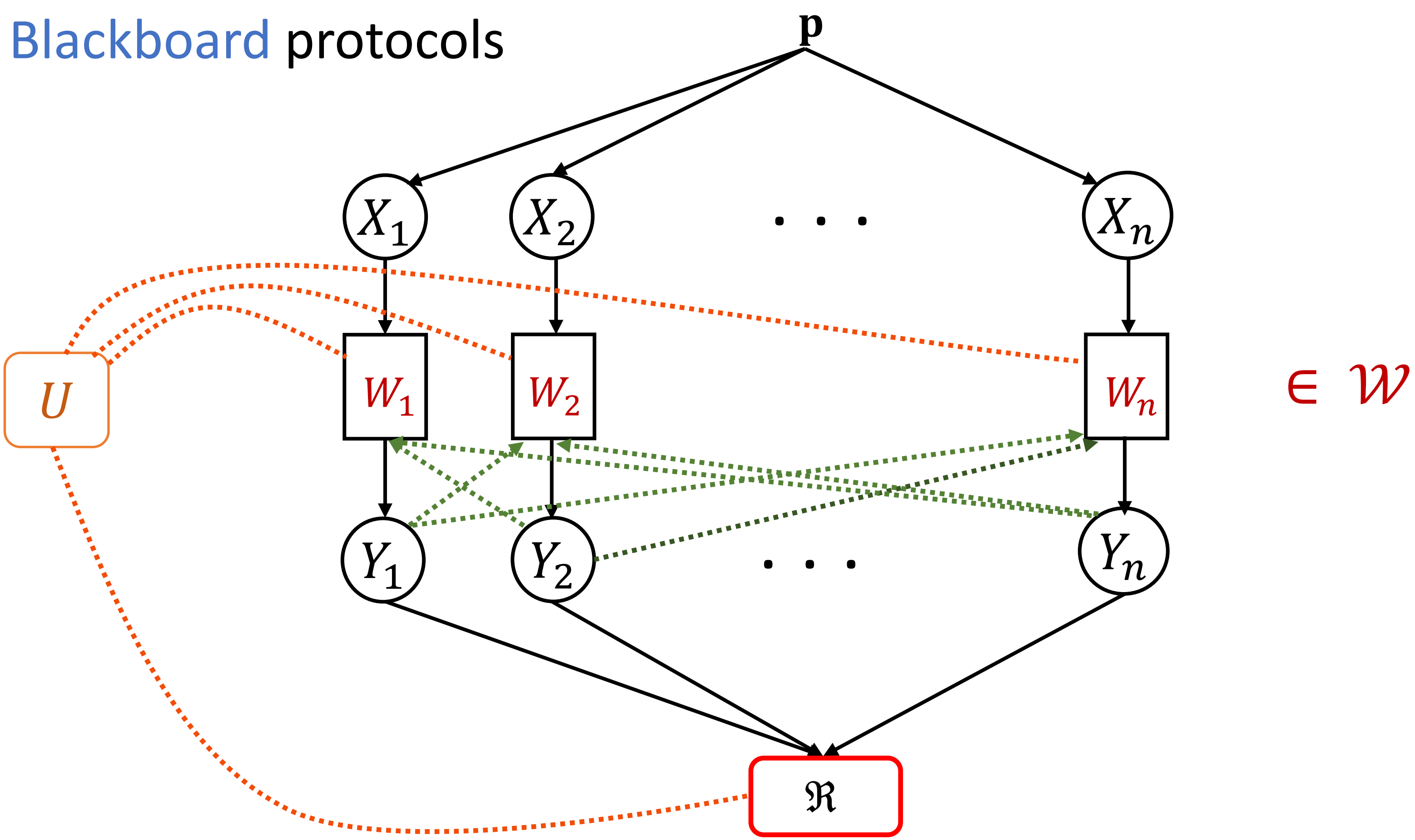


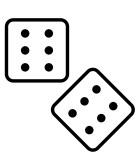


Interactive ("one-pass, sequential"), and common random seed

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.

Types of problems

ε: **accuracy** parameter

Estimation (learning): Design $\widehat{\mathbf{p}}(Y^n)$ such that

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

$\mathrm{d}(\cdot,\cdot)$ is a **distance/loss** ⇝ e.g., total variation or parameter distance

Types of problems

ε: **accuracy** parameter

Estimation (learning): Design $\hat{\mathbf{p}}(Y^n)$ such that

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

$\mathrm{d}(\cdot,\cdot)$ is a **distance/loss** ⇝ e.g., total variation or parameter distance

Hypothesis testing: given two sets of "yes" and "no" distributions $\mathcal{H}_0, \mathcal{H}_\varepsilon$ "with separation ε,"* design $T(Y^n)$ such that

$$\Pr(T(Y^n) = 0) > 0.9, \text{ if } \mathbf{p} \in \mathcal{H}_0$$
$$\Pr(T(Y^n) = 1) > 0.9, \text{ if } \mathbf{p} \in \mathcal{H}_\varepsilon$$

* I.e., $\mathrm{d}(\mathbf{p}, \mathbf{q}) > \varepsilon$ for every $\mathbf{p} \in \mathcal{H}_0$, $\mathbf{q} \in \mathcal{H}_\varepsilon$

Types of problems: Estimation

1. **Distribution learning**

Dimension = $k-1,$ Accuracy = ε

$\boldsymbol{p}$: unknown distribution on $\mathcal{X}=[k]$, distance/loss: total variation*

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

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

* $TV(\boldsymbol{p}, \boldsymbol{q}) = \sup_{S\subseteq[k]} (\boldsymbol{p}(S) - \boldsymbol{q}(S))$

Types of problems: Estimation

2. High-dimensional mean learning

Dimension = d, Accuracy = ε

$\boldsymbol{p}$ assumed to product distribution over $\mathcal{X} = \mathbb{R}^d$ w mean $\boldsymbol{\mu} = \mathbb{E}_{\boldsymbol{p}}[X]$, distance/loss: ℓ_2

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

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

*Families of interest: Gaussians, Product Bernoulli

Types of problems: Hypothesis testing

1. Identity testing

Dimension = $k-1$, Accuracy = ε

$\mathbf{q}$: reference distribution on $\mathcal{X}=[k]$, distance/loss: total variation

$$\mathcal{H}_0 = \{\boldsymbol{q}\}, \qquad \mathcal{H}_\varepsilon = \{\boldsymbol{p}' : TV(\boldsymbol{p}', \boldsymbol{q}) > \varepsilon\}$$

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

Types of problems: Hypothesis testing

2. High-dimensional mean testing

Dimension = d, Accuracy = ε

$\boldsymbol{p}$ assumed to product distribution over $\mathcal{X} = \mathbb{R}^d$, distance/loss: ℓ_2

$$\mathcal{H}_0 = \{\boldsymbol{p}' : \mathbb{E}_{\boldsymbol{p}'}[X] = \boldsymbol{0}\}, \qquad \mathcal{H}_\varepsilon = \{\boldsymbol{p}' : \left\|\mathbb{E}_{\boldsymbol{p}'}[X]\right\|_2 > \varepsilon\}$$

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

*Families of interest: Gaussian, Product Bernoulli

Some references and previous work

Some references and previous work

(not in that order)

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

Plan for the tutorial

~~I. Introduction~~	Clément
II. Lower Bounds for Estimation	Jayadev
III. Lower Bounds for Testing	Himanshu
IV. Some upper bounds, and discussion	Clément