On the Computational and Statistical Interface and “Big Data”

Michael I. Jordan
University of California, Berkeley

February 17, 2015

What Is the Big Data Phenomenon?

- Science in confirmatory mode (e.g., particle physics)
- Science in exploratory mode (e.g., astronomy, genomics)
- Measurement of human activity, particularly online activity, is generating massive datasets that can be used (e.g.) for personalization and for creating markets
- Sensor networks are becoming pervasive

What Are the Conceptual/Mathematical Issues?

- The need to control statistical risk under constraints on algorithmic runtime

What Are the Conceptual/Mathematical Issues?

- The need to control statistical risk under constraints on algorithmic runtime
  - how do risk and runtime trade off as a function of the amount of data?

What Are the Conceptual/Mathematical Issues?

- The need to control statistical risk under constraints on algorithmic runtime
  - how do risk and runtime trade off as a function of the amount of data?
- Statistical with distributed and streaming data
  - how is inferential quality impacted by communication constraints?

What Are the Conceptual/Mathematical Issues?

- The need to control statistical risk under constraints on algorithmic runtime
  - how do risk and runtime trade off as a function of the amount of data?
- Statistical with distributed and streaming data
  - how is inferential quality impacted by communication constraints?
- The tradeoff between statistical risk and privacy (and other externalities)

What Are the Conceptual/Mathematical Issues?

- The need to control statistical risk under constraints on algorithmic runtime
  - how do risk and runtime trade off as a function of the amount of data?
- Statistical with distributed and streaming data
  - how is inferential quality impacted by communication constraints?
- The tradeoff between statistical risk and privacy (and other externalities)
- Many other issues that require a blend of statistical thinking (e.g., a focus on sampling, confidence intervals, evaluation, diagnostics, causal inference) and computational thinking (e.g., scalability, abstraction)

Our Approach

- Take (classical) statistical decision theory as a mathematical point of departure
- Treat computation, communication, privacy, etc as constraints on statistical risk
- This induces tradeoffs among these quantities and the number of data points

Our Approach

- Take (classical) statistical decision theory as a mathematical point of departure
- Treat computation, communication, privacy, etc as constraints on statistical risk
- This induces tradeoffs among these quantities and the number of data points

- Under the hood: geometry, information theory and optimization

Background

- In the 1930's, Wald laid the foundations of statistical decision theory
- Given a family of probability distributions $\mathcal{P}$, a parameter $\theta(P)$ for each $P \in \mathcal{P}$, an estimator $\hat{\theta}$, and a loss $l(\hat{\theta}, \theta(P))$, define the risk:

$$R_P(\hat{\theta}) := \mathbb{E}_P \left[l(\hat{\theta}, \theta(P)) \right]$$

- Minimax principle [Wald, '39, '43]: choose estimator minimizing worst-case risk:

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[l(\hat{\theta}, \theta(P)) \right]$$

Part I: Privacy and Minimax Risk

with John Duchi and Martin Wainwright
University of California, Berkeley

Privacy and Risk

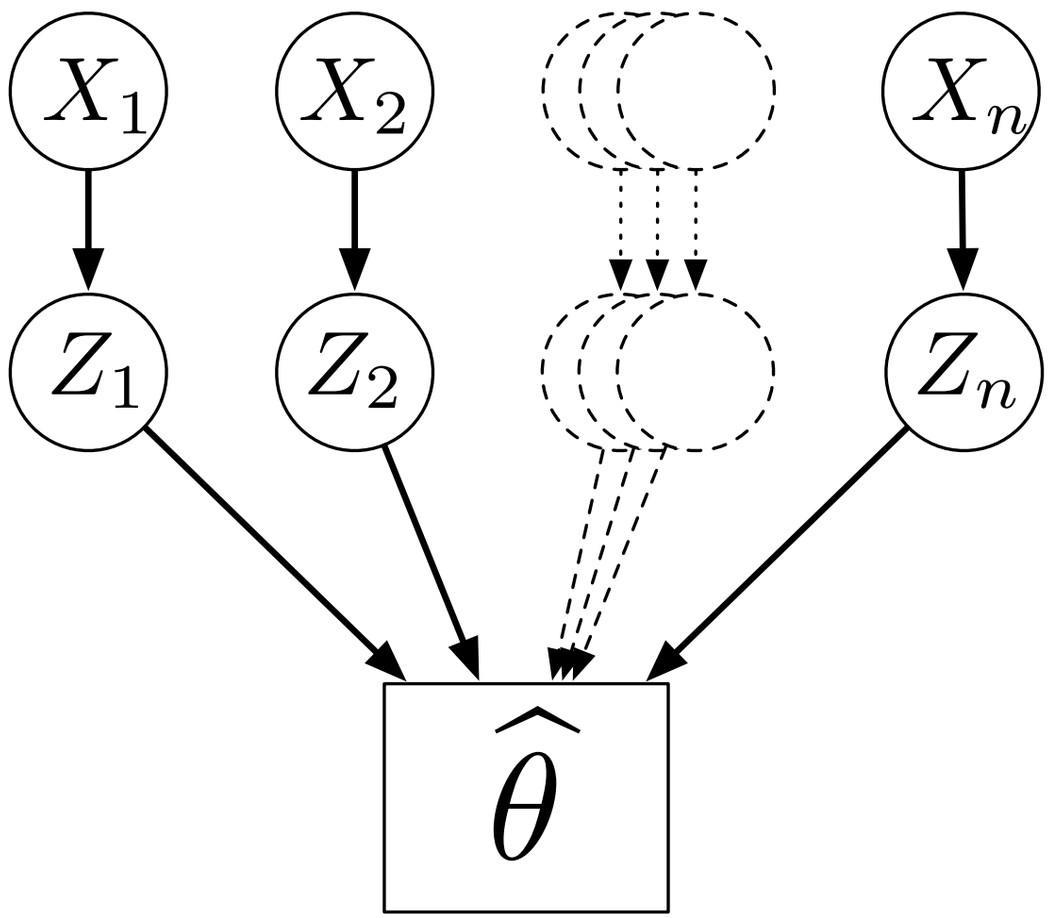
- Individuals are not generally willing to allow their personal data to be used without control on how it will be used and now much privacy loss they will incur
- We will quantify “privacy loss” via differential privacy
- We then treat differential privacy as a constraint on inference via statistical decision theory
- This yields (personal) tradeoffs between privacy loss and inferential gain

A model of privacy

Local privacy: providers do not trust collector [Warner 65, Evfimievski et al. 03]

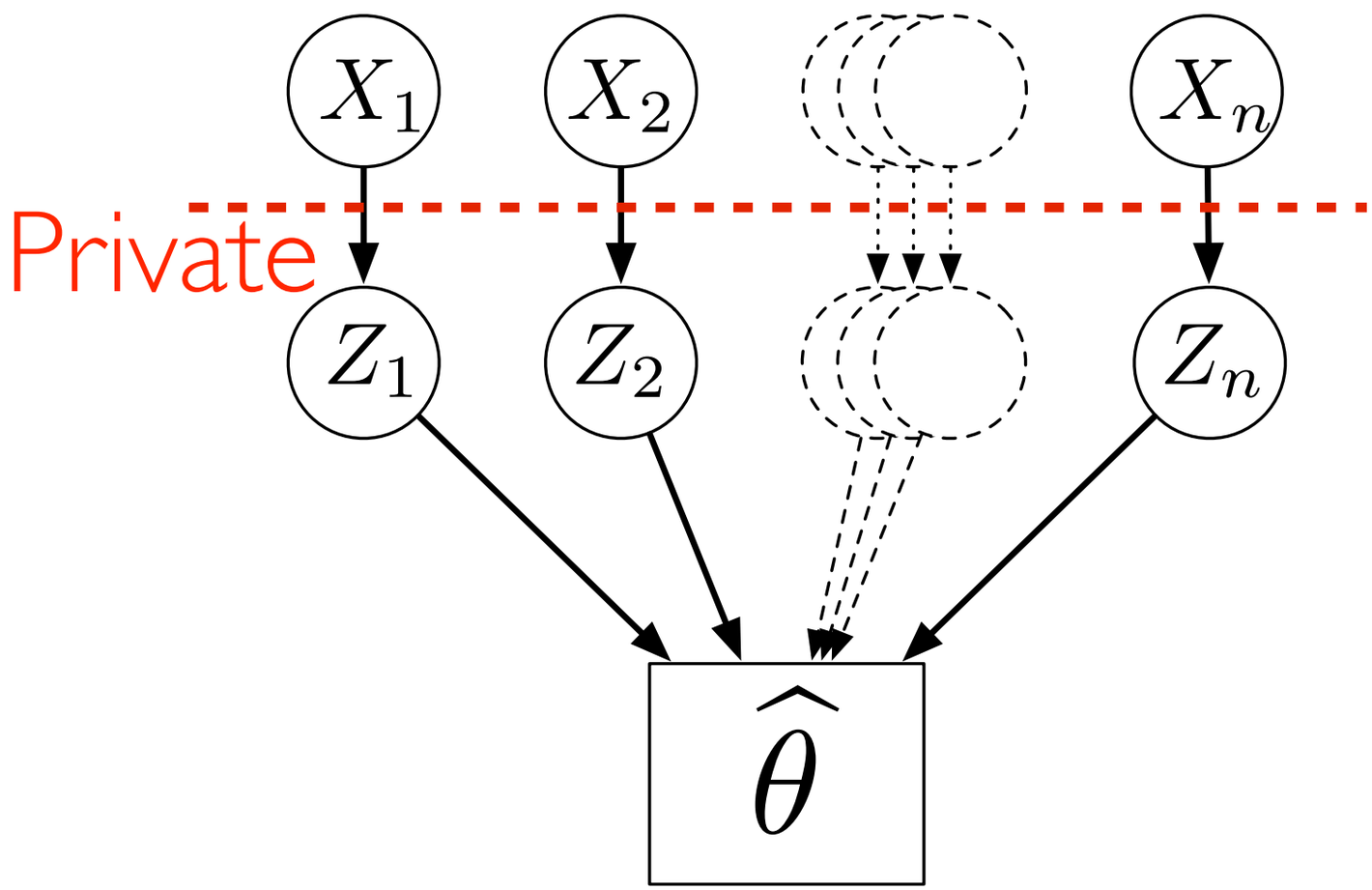
A model of privacy

Local privacy: providers do not trust collector [Warner 65, Evfimievski et al. 03]



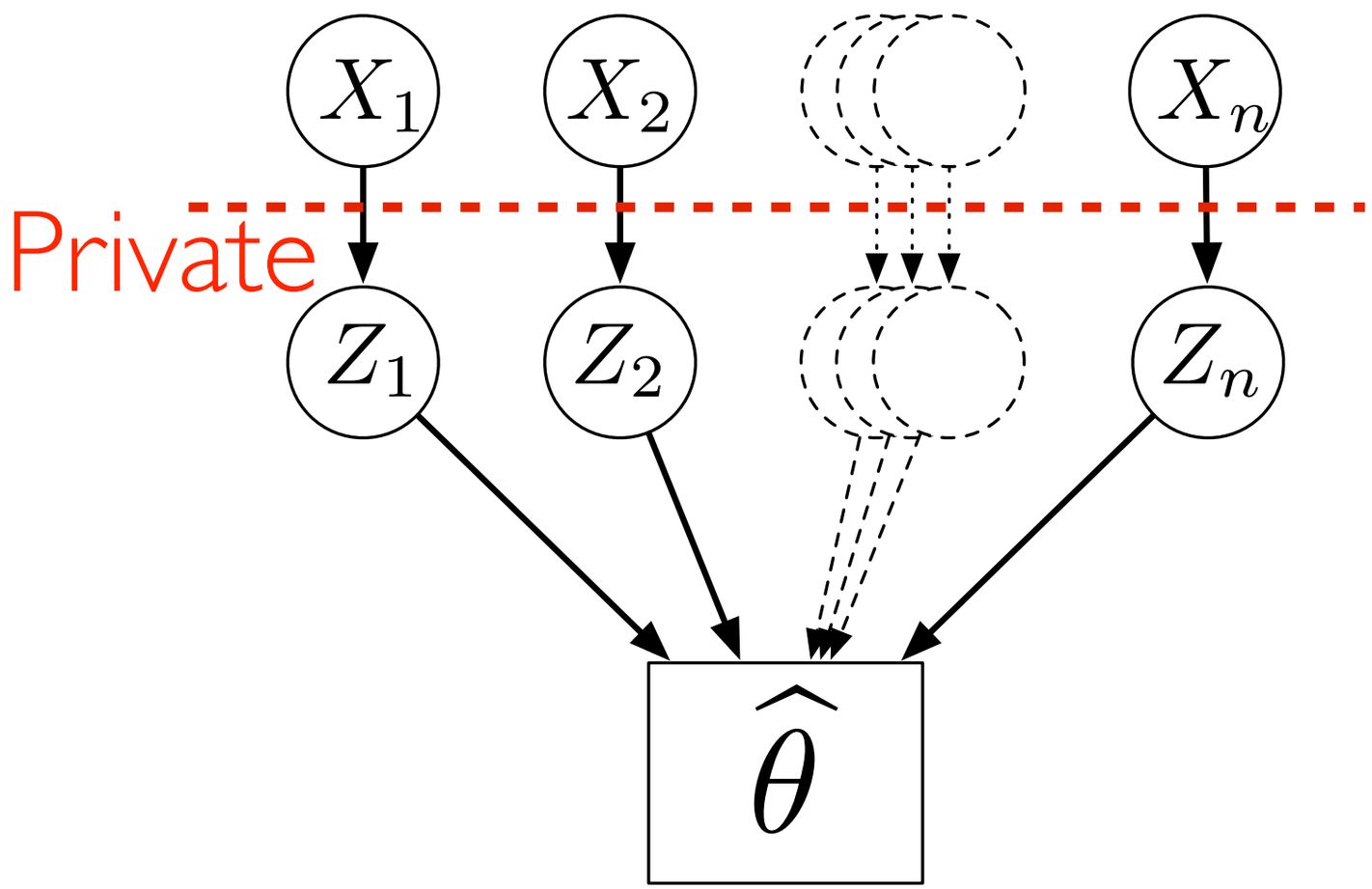
A model of privacy

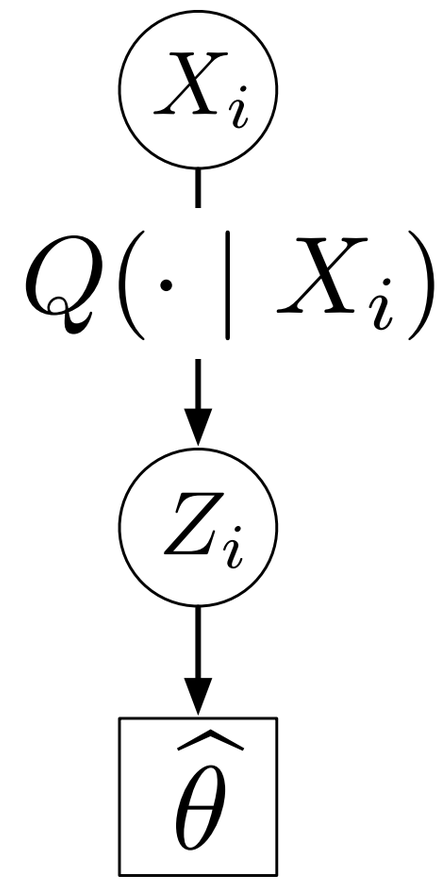
Local privacy: providers do not trust collector [Warner 65, Evfimievski et al. 03]



A model of privacy

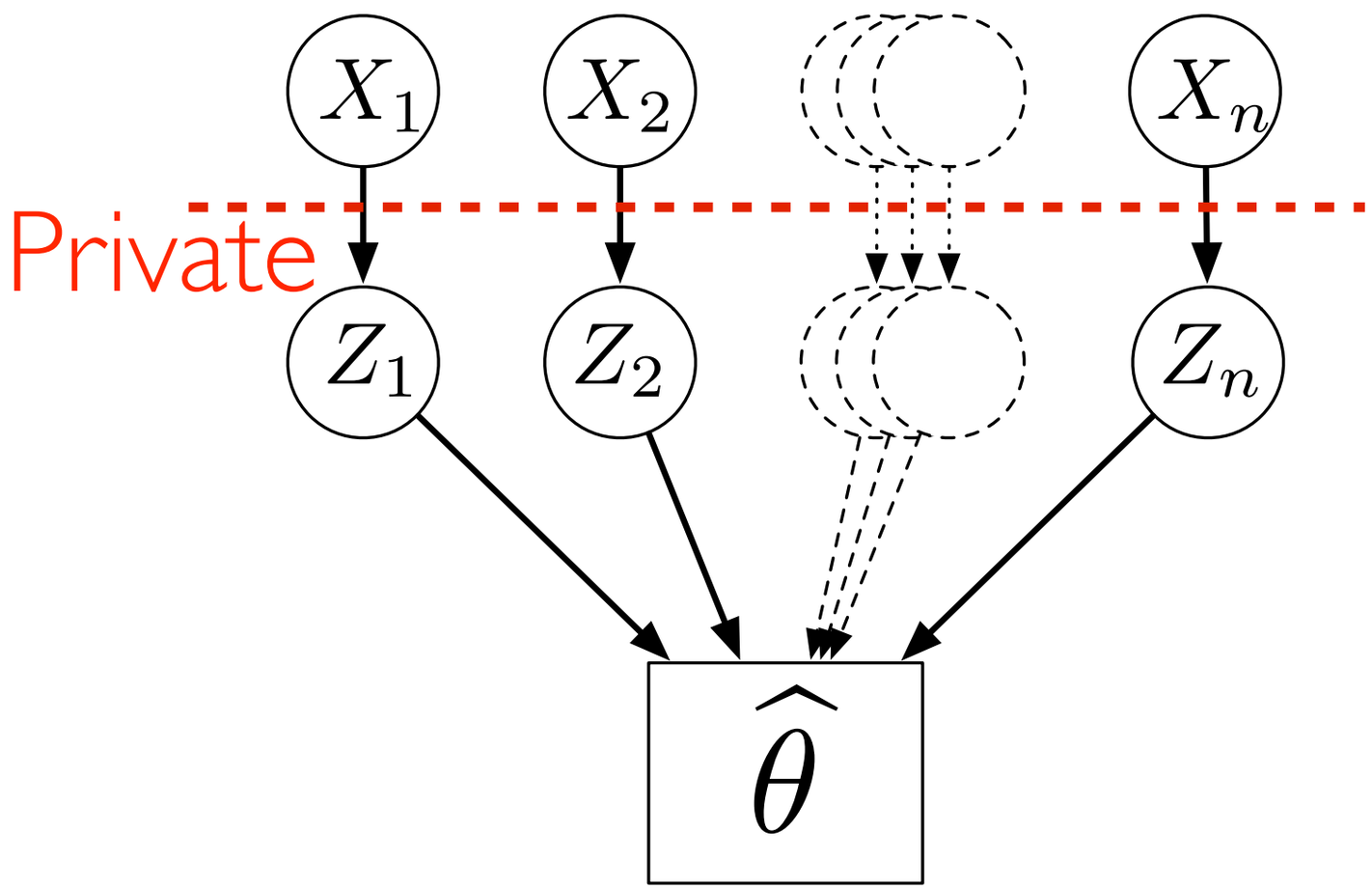
Local privacy: providers do not trust collector [Warner 65, Evfimievski et al. 03]

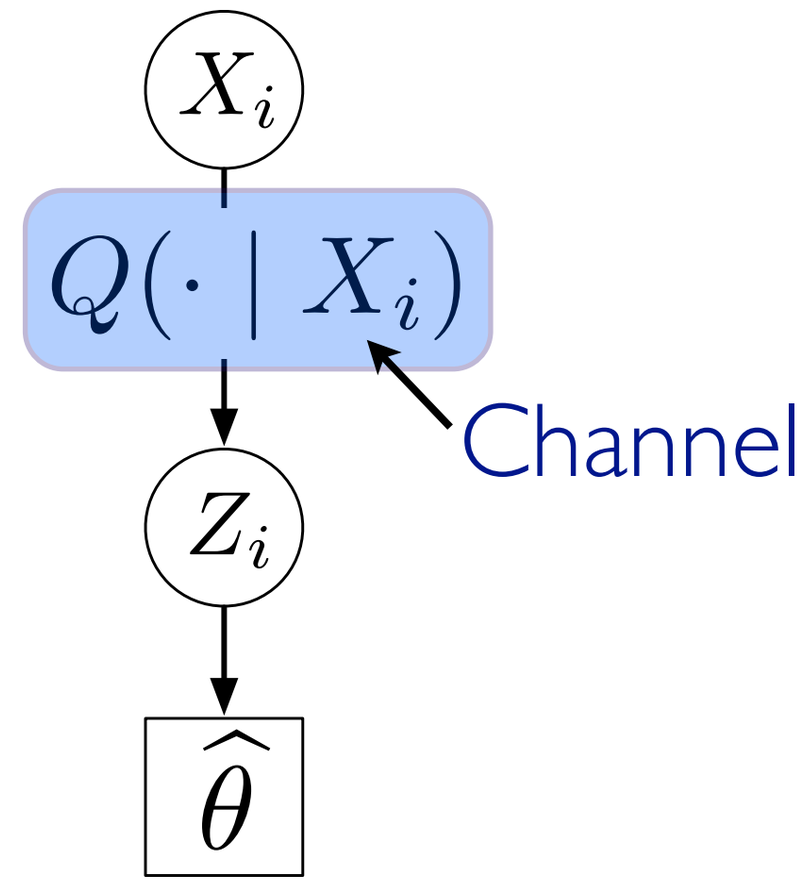




A model of privacy

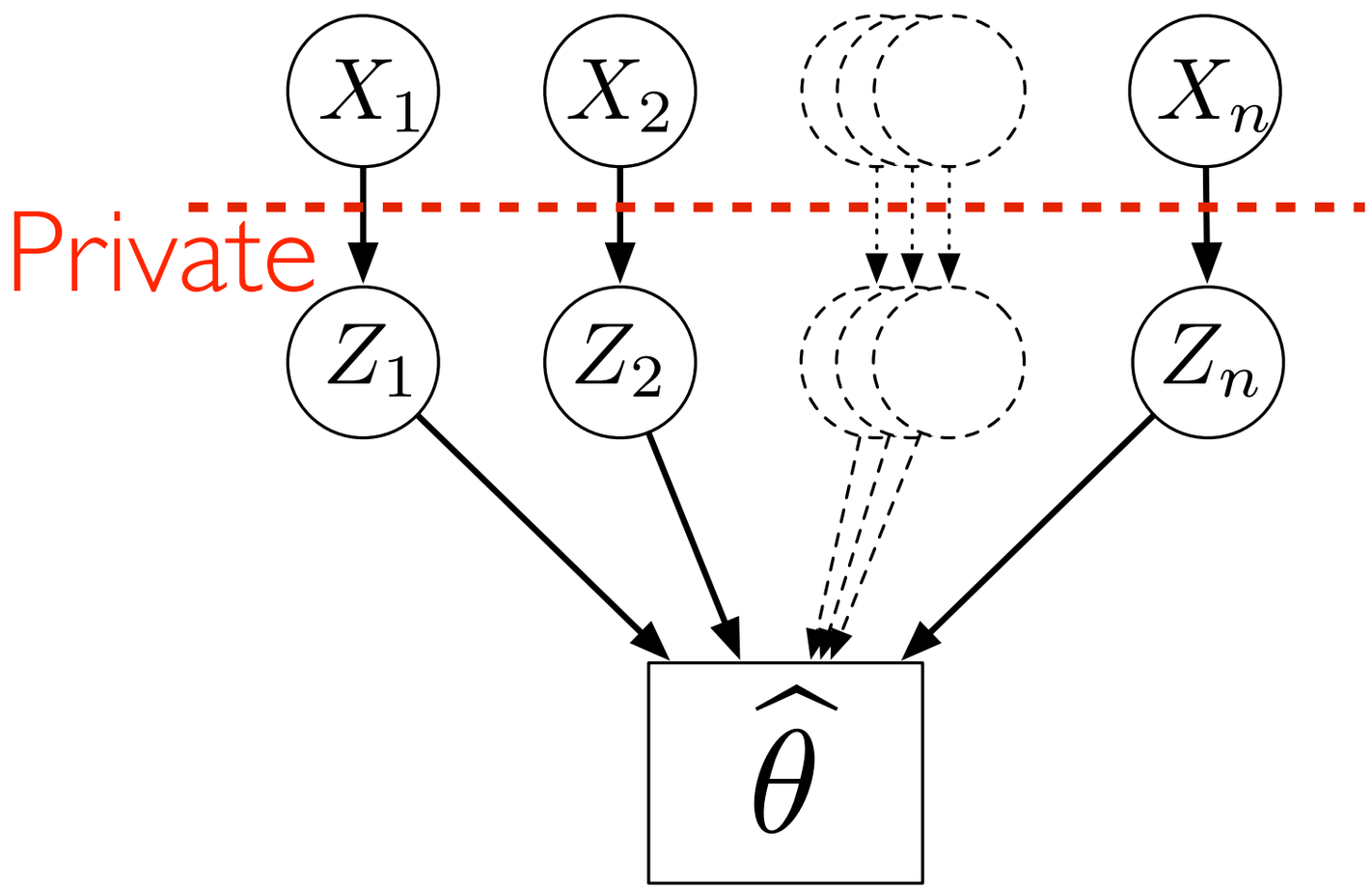
Local privacy: providers do not trust collector [Warner 65, Evfimievski et al. 03]

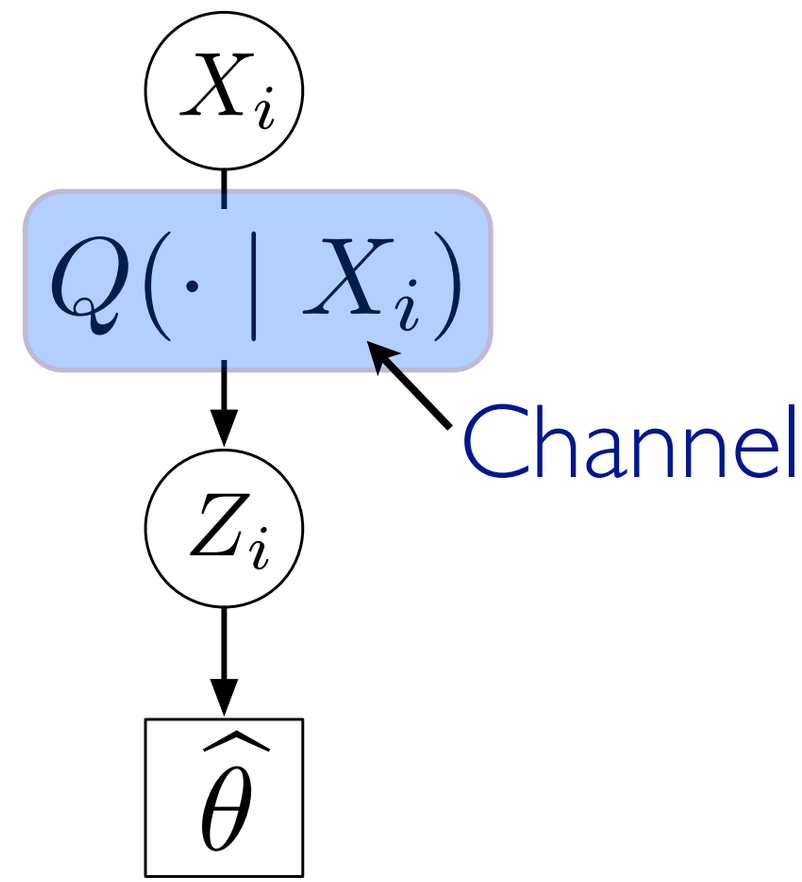




A model of privacy

Local privacy: providers do not trust collector [Warner 65, Evfimievski et al. 03]



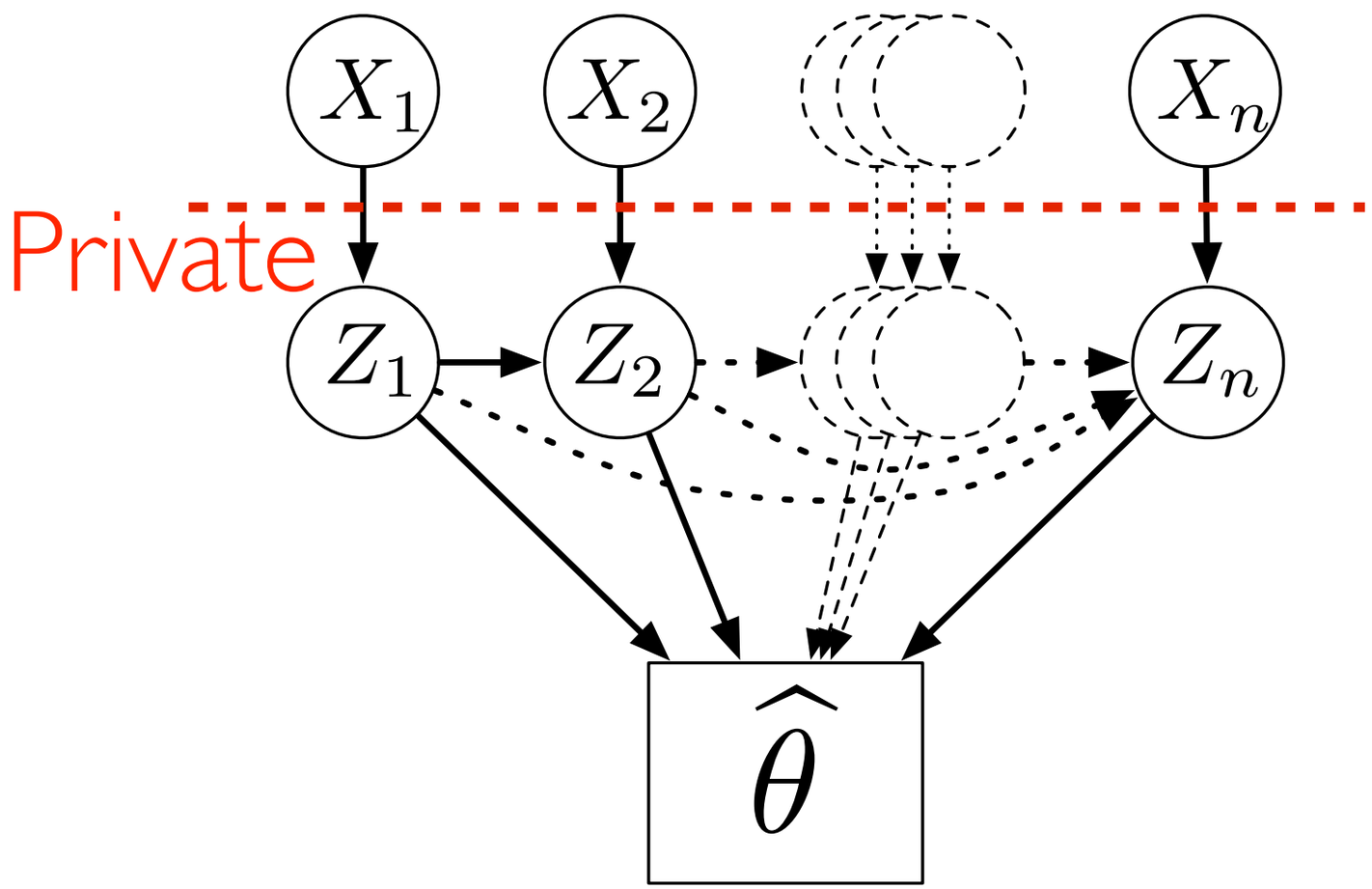


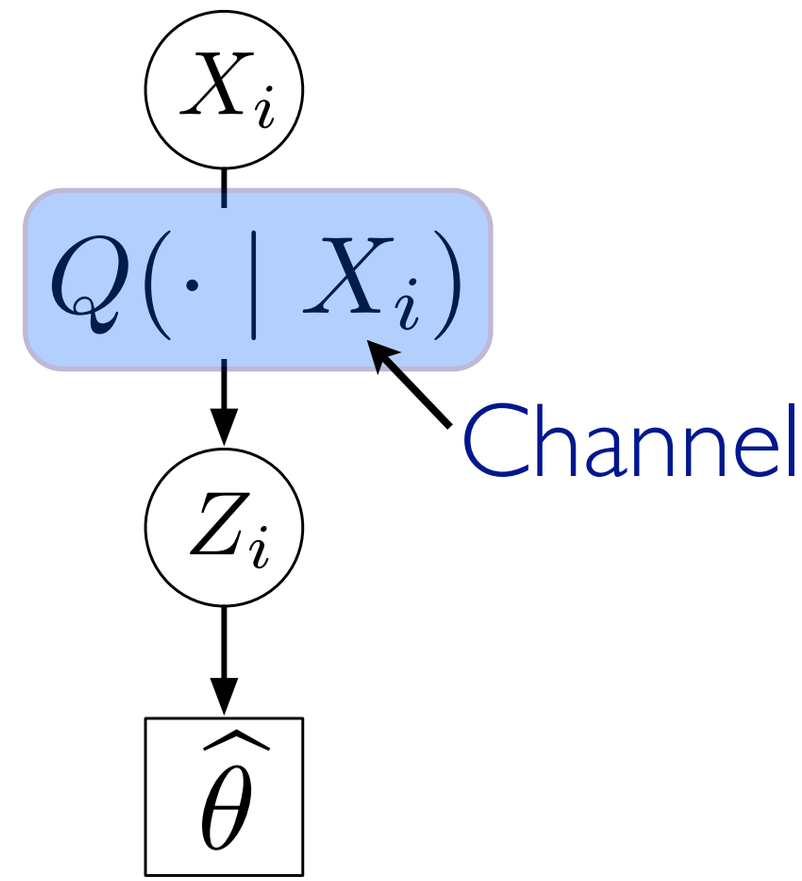
Individuals $i \in \{1, \ldots, n\}$ with private data $X_i \overset{\text{iid}}{\sim} P$

Estimator $Z_1^n \mapsto \widehat{\theta}(Z_1^n)$

A model of privacy

Local privacy: providers do not trust collector [Warner 65, Evfimievski et al. 03]





Individuals $i \in \{1, \ldots, n\}$ with private data $X_i \overset{\text{iid}}{\sim} P$

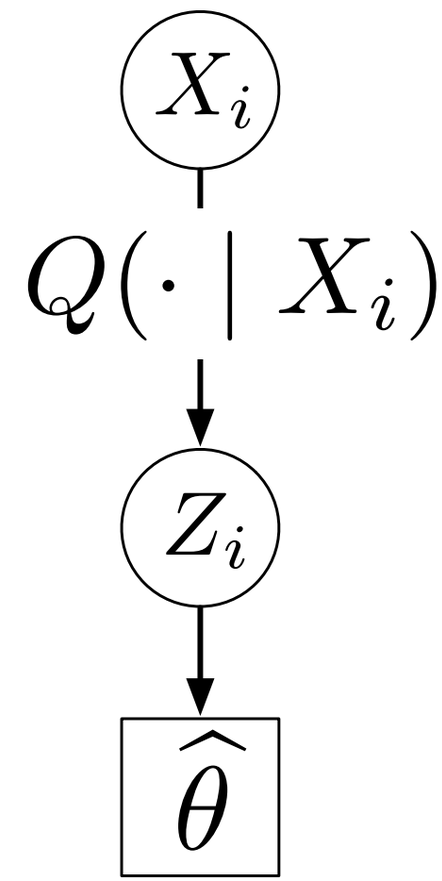
Estimator $Z_1^n \mapsto \widehat{\theta}(Z_1^n)$

Definitions of privacy

Definition: channel Q is α-differentially private if

$$\sup_{S, x \in \mathcal{X}, x' \in \mathcal{X}} \frac{Q(Z \in S \mid x)}{Q(Z \in S \mid x')} \leq \exp(\alpha)$$

[Dwork, McSherry, Nissim, Smith 06]

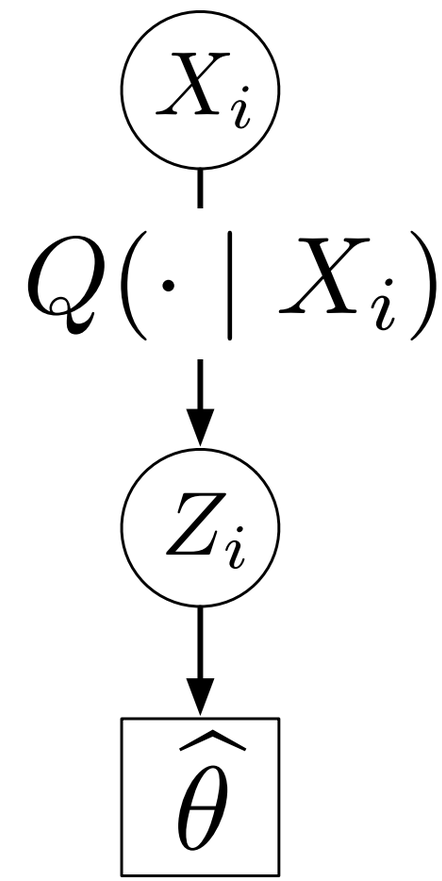


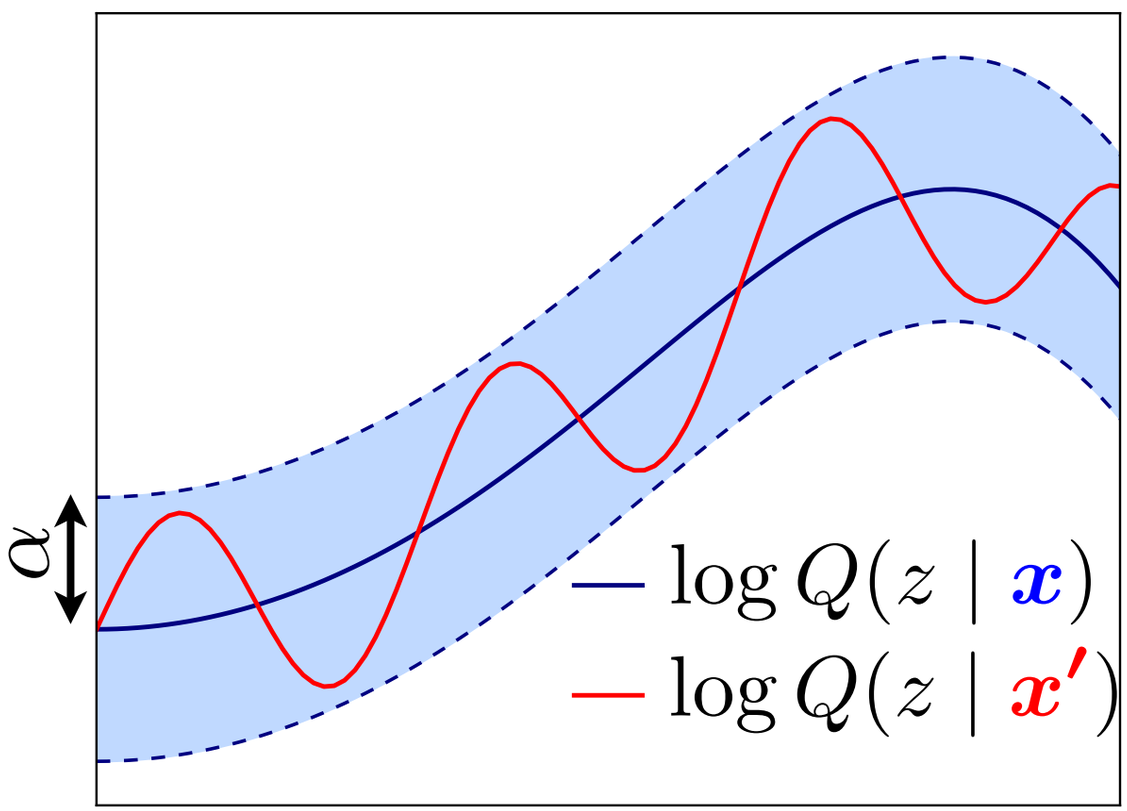
Definitions of privacy

Definition: channel Q is α-differentially private if

$$\sup_{S, x \in \mathcal{X}, x' \in \mathcal{X}} \frac{Q(Z \in S \mid x)}{Q(Z \in S \mid x')} \leq \exp(\alpha)$$

[Dwork, McSherry, Nissim, Smith 06]





Private Minimax Risk

Central object of study: minimax risk

- Parameter $\theta(P)$ of distribution
- Family of distributions $\mathcal{P}$
- Loss ℓ measuring error

Minimax risk

$$\mathfrak{M}_n(\theta(\mathcal{P}), \ell) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\ell(\widehat{\theta}(X_1^n), \theta(P)) \right]$$

Private Minimax Risk

Central object of study: minimax risk

- Parameter $\theta(P)$ of distribution
- Family of distributions $\mathcal{P}$
- Loss ℓ measuring error
- Family $\mathcal{Q}_\alpha$ of private channels

α-private Minimax risk

$$\mathfrak{M}_n(\theta(\mathcal{P}), \ell, \alpha) := \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q} \left[\ell(\widehat{\theta}(Z_1^n), \theta(P)) \right]$$

Private Minimax Risk

Central object of study: minimax risk

- Parameter $\theta(P)$ of distribution
- Family of distributions $\mathcal{P}$
- Loss ℓ measuring error
- Family $\mathcal{Q}_\alpha$ of private channels

α-private Minimax risk

$$\mathfrak{M}_n(\theta(\mathcal{P}), \ell, \alpha) := \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q} \left[\ell(\widehat{\theta}(Z_1^n), \theta(P)) \right]$$

Best α-private channel

Private Minimax Risk

Central object of study: minimax risk

- Parameter $\theta(P)$ of distribution
- Family of distributions $\mathcal{P}$
- Loss ℓ measuring error
- Family $\mathcal{Q}_\alpha$ of private channels

α-private Minimax risk

$$\mathfrak{M}_n(\theta(\mathcal{P}), \ell, \alpha) := \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}\left[\ell(\widehat{\theta}(Z_1^n), \theta(P))\right]$$

Best α-private channel

Minimax risk under privacy constraint

Vignette: private mean (location) estimation

Example: estimate reasons for hospital visits

Patients admitted to hospital for substance abuse

Estimate prevalence of different substances

Vignette: private mean (location) estimation

Example: estimate reasons for hospital visits

Patients admitted to hospital for substance abuse

Estimate prevalence of different substances

1	Alcohol
1	Cocaine
0	Heroin
0	Cannabis
0	LSD
0	Amphetamines

Proportions $\boldsymbol{\theta} = \begin{array}{l} \theta_1 = .45 \\ \theta_2 = .32 \\ \theta_3 = .16 \\ \theta_4 = .20 \\ \theta_5 = .00 \\ \theta_6 = .02 \end{array}$

Vignette: mean estimation

Consider estimation of mean $\theta(P) := \mathbb{E}_P[X] \in \mathbb{R}^d$, errors measured in ℓ_∞-norm, i.e. $\mathbb{E}[\|\widehat{\theta} - \theta\|_\infty]$ for

$$\mathcal{P}_d := \left\{ \text{distributions } P \text{ supported on } [-1, 1]^d \right\}$$

Vignette: mean estimation

Consider estimation of mean $\theta(P) := \mathbb{E}_P[X] \in \mathbb{R}^d$, errors measured in ℓ_∞-norm, i.e. $\mathbb{E}[\|\widehat{\theta} - \theta\|_\infty]$ for

$$\mathcal{P}_d := \left\{ \text{distributions } P \text{ supported on } [-1, 1]^d \right\}$$

Proposition:

Minimax rate

$$\mathfrak{M}_n(\mathcal{P}_d, \|\cdot\|_\infty) \asymp \min\left\{1, \frac{\sqrt{\log d}}{\sqrt{n}}\right\}$$

(achieved by sample mean)

Vignette: mean estimation

Consider estimation of mean $\theta(P) := \mathbb{E}_P[X] \in \mathbb{R}^d$, errors measured in ℓ_∞-norm, i.e. $\mathbb{E}[\|\widehat{\theta} - \theta\|_\infty]$ for

$$\mathcal{P}_d := \left\{\text{distributions } P \text{ supported on } [-1, 1]^d\right\}$$

Proposition:

Private minimax rate for $\alpha = O(1)$

$$\mathfrak{M}_n(\mathcal{P}_d, \|\cdot\|_\infty, \alpha) \asymp \min\left\{1, \frac{\sqrt{d \log d}}{\sqrt{n\alpha^2}}\right\}$$

Vignette: mean estimation

Consider estimation of mean $\theta(P) := \mathbb{E}_P[X] \in \mathbb{R}^d$, errors measured in ℓ_∞-norm, i.e. $\mathbb{E}[\|\widehat{\theta} - \theta\|_\infty]$ for

$$\mathcal{P}_d := \left\{ \text{distributions } P \text{ supported on } [-1, 1]^d \right\}$$

Proposition:

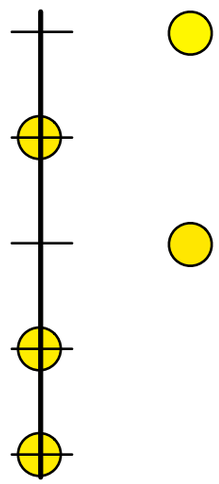
Private minimax rate for $\alpha = O(1)$

$$\mathfrak{M}_n(\mathcal{P}_d, \|\cdot\|_\infty, \alpha) \asymp \min\left\{1, \frac{\sqrt{d \log d}}{\sqrt{n\alpha^2}}\right\}$$

Note: Effective sample size $n \mapsto n\alpha^2/d$

Optimal mechanism?

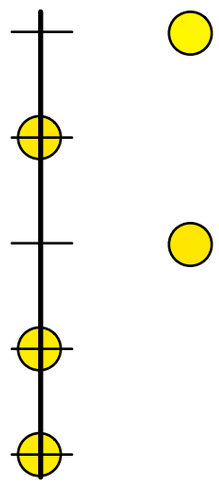
$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



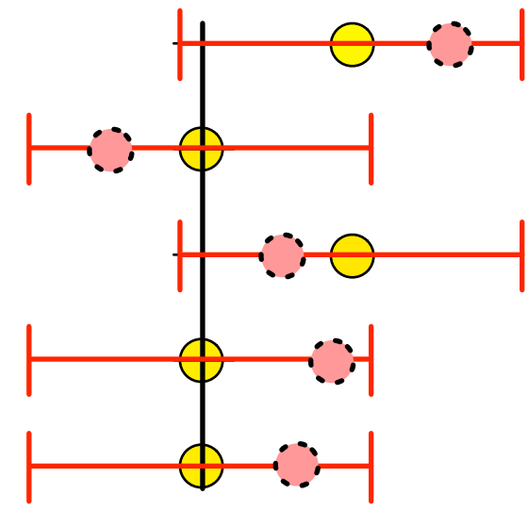
Non-private observation

Optimal mechanism?

$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



$$Z = X + W = \begin{bmatrix} 1 + W_1 \\ 0 + W_2 \\ 1 + W_3 \\ 0 + W_4 \\ 0 + W_5 \end{bmatrix}$$

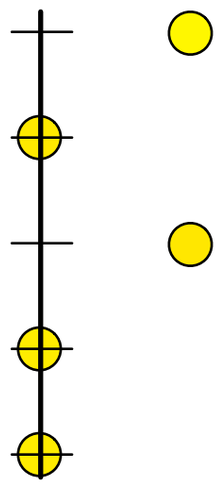


Non-private observation

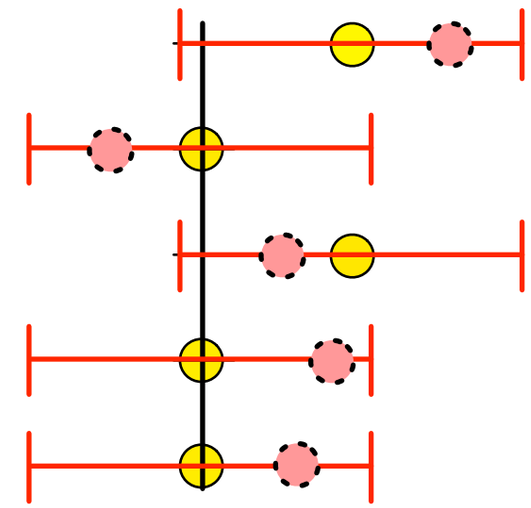
Idea 1: add independent noise (e.g. standard Laplace mechanism)

Optimal mechanism?

$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



$$Z = X + W = \begin{bmatrix} 1 + W_1 \\ 0 + W_2 \\ 1 + W_3 \\ 0 + W_4 \\ 0 + W_5 \end{bmatrix}$$



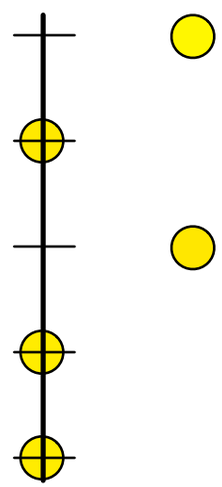
Non-private observation

Idea 1: add independent noise (e.g. standard Laplace mechanism)

Problem: magnitude much too large
(this is unavoidable: *provably sub-optimal*)

Optimal mechanism

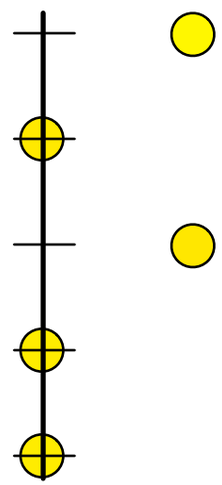
$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



Non-private observation

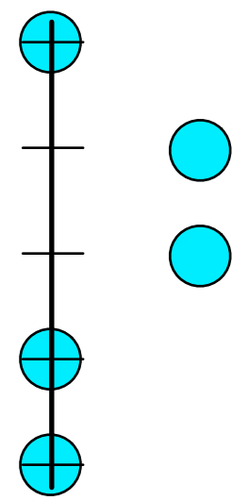
Optimal mechanism

$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



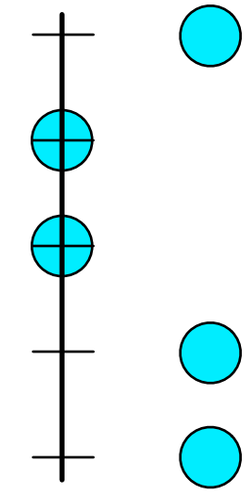
Non-private observation

$$v = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



View 1

$$1 - v = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

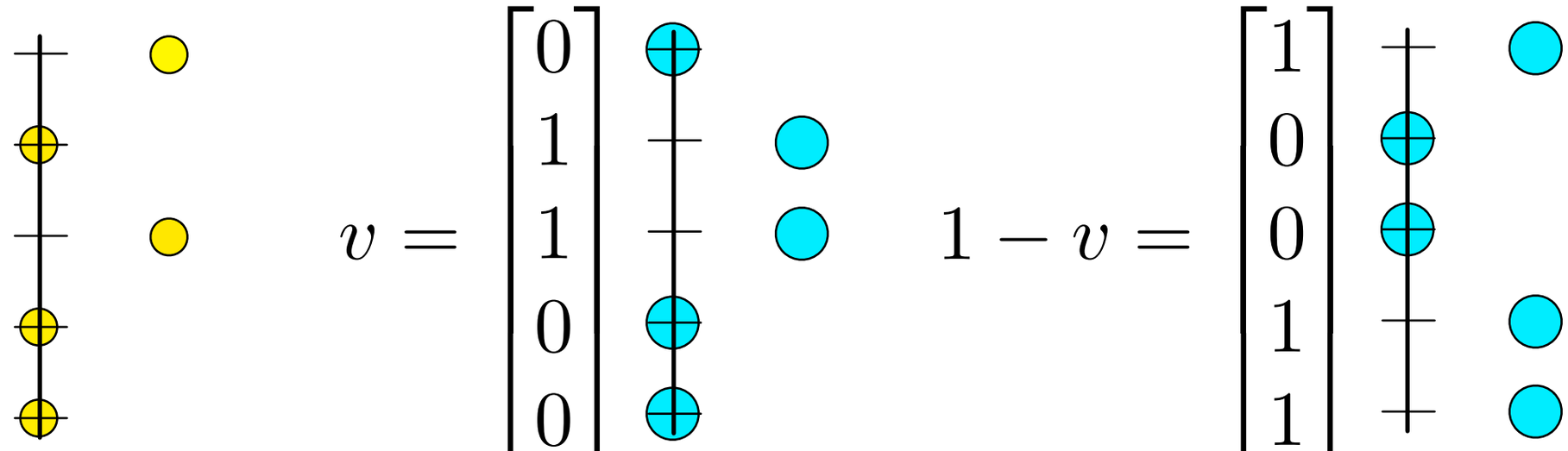


View 2

- Draw v uniformly in $\{0,1\}^d$

Optimal mechanism

$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad v = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad 1 - v = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$



Non-private observation

View 1 (closer: 3 overlap)

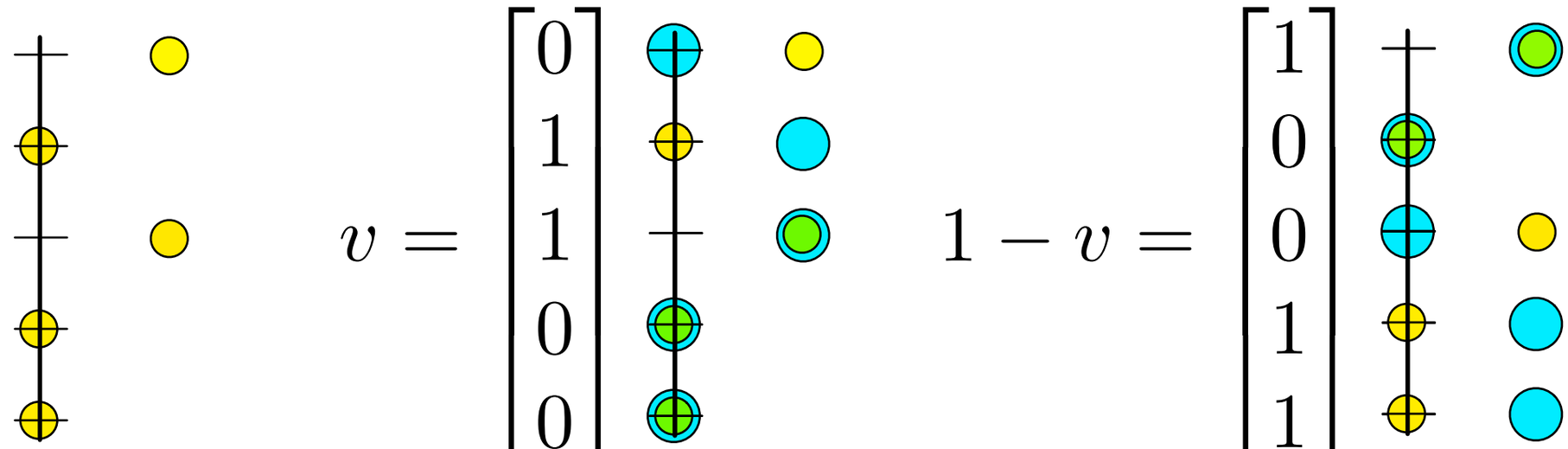
View 2 (farther: 2 overlap)

- Draw v uniformly in $\{0,1\}^d$
- With probability $\dfrac{e^{\alpha}}{1+e^{\alpha}}$

  choose closer of v and $1-v$ to X
- otherwise, choose farther

Optimal mechanism

$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad v = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad 1 - v = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$



Non-private observation

View 1 (closer: 3 overlap)

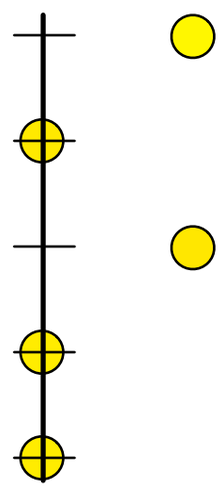
View 2 (farther: 2 overlap)

- Draw v uniformly in $\{0,1\}^d$
- With probability $\dfrac{e^{\alpha}}{1+e^{\alpha}}$

  choose closer of v and $1-v$ to X
- otherwise, choose farther

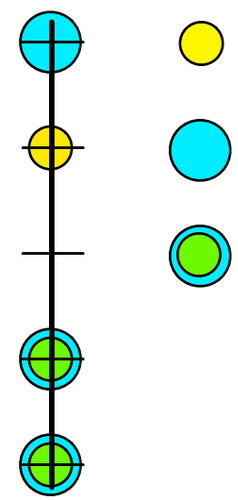
Optimal mechanism

$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



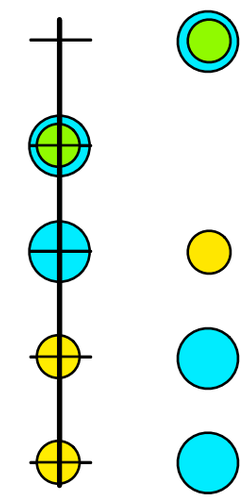
Non-private observation

$$v = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



View 1
(closer: 3 overlap)

$$1 - v = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$



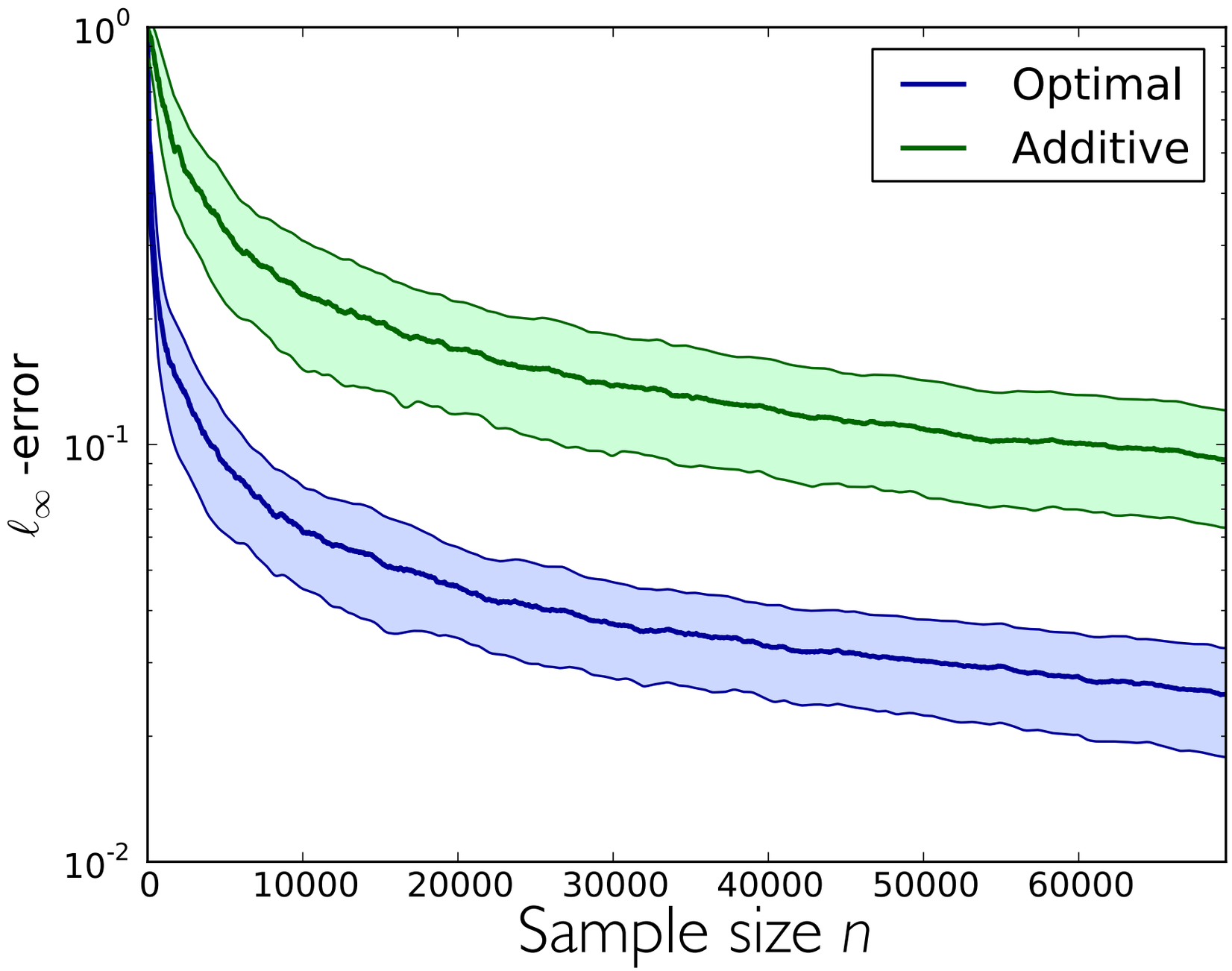
View 2
(farther: 2 overlap)

- Draw v uniformly in $\{0,1\}^d$
- With probability $\dfrac{e^\alpha}{1+e^\alpha}$

  choose closer of v and $1-v$ to X
- otherwise, choose farther

At end:
Compute sample average and de-bias

Empirical evidence



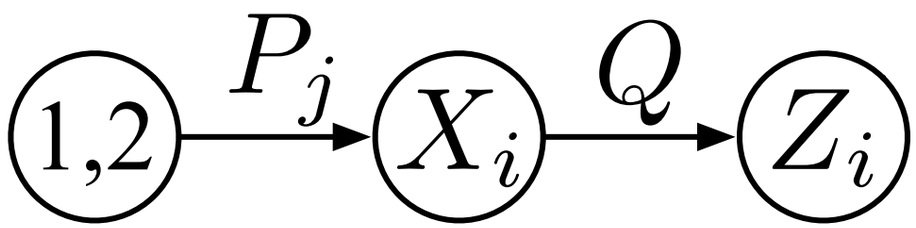
Data source: Drug Abuse Warning Network

Estimate proportion of emergency room visits involving different substances

Sample size reductions

Given α-private channel Q, pair $\{P_1, P_2\}$ induces marginal

$$M_j(S) := \int Q(S \mid x_1, \ldots, x_n) dP_j^n(x_1, \ldots, x_n)$$



Sample size reductions

Given α-private channel Q, pair $\{P_1, P_2\}$ induces marginal

$$M_j(S) := \int Q(S \mid x_1, \ldots, x_n) dP_j^n(x_1, \ldots, x_n)$$

$$1,2 \xrightarrow{P_j} X_i \xrightarrow{Q} Z_i$$

Question: How much "contraction" does privacy induce?

Sample size reductions

Given α-private channel Q, pair $\{P_1, P_2\}$ induces marginal

$$M_j(S) := \int Q(S \mid x_1, \ldots, x_n) dP_j^n(x_1, \ldots, x_n)$$

$$1,2 \xrightarrow{P_j} X_i \xrightarrow{Q} Z_i$$

Question: How much "contraction" does privacy induce?

Theorem (data processing): for any α-private channel and i.i.d. sample of size n

$$D_{\mathrm{kl}}(M_1 \| M_2) + D_{\mathrm{kl}}(M_2 \| M_1) \leq 4n(e^{\alpha} - 1)^2 \|P_1 - P_2\|_{\mathrm{TV}}^2$$

Sample size reductions

Given α-private channel Q, pair $\{P_1, P_2\}$ induces marginal

$$M_j(S) := \int Q(S \mid x_1, \ldots, x_n) dP_j^n(x_1, \ldots, x_n)$$

$$1,2 \xrightarrow{P_j} X_i \xrightarrow{Q} Z_i$$

Question: How much "contraction" does privacy induce?

Theorem (data processing): for any α-private channel and i.i.d. sample of size n

$$D_{\mathrm{kl}}\left(M_1 \| M_2\right) + D_{\mathrm{kl}}\left(M_2 \| M_1\right) \leq 4n\boldsymbol{(e^{\alpha} - 1)^2} \left\| P_1 - P_2 \right\|_{\mathrm{TV}}^2$$

Note: $n \mapsto \boldsymbol{n\alpha^2}$ for $\alpha \lesssim 1$

Final remarks: privacy

Rough technique: Reduction of estimation to testing, then apply information-theoretic testing lower bounds

[Le Cam, Hasminskii, Ibragimov, Assouad, Birge, Barron, Yu, ...]

Additional examples

- Fixed-design regression
- Convex risk minimization
- Multinomial estimation
- Nonparametric density estimation

Almost always: effective sample size reduction $n \mapsto n\alpha^2$

In d-dimensional problems: $n \mapsto \dfrac{n\alpha^2}{d}$

Key: Allows identification of new optimal mechanisms

Part II: Communication and Minimax Risk

with John Duchi, Martin Wainwright and Yuchen Zhang
University of California, Berkeley

Communication-constraints

[Yao 79; Abelson 80; Tsitsiklis and Luo 87; Han & Amari 98; Tatikonda & Mitter 04; ...]

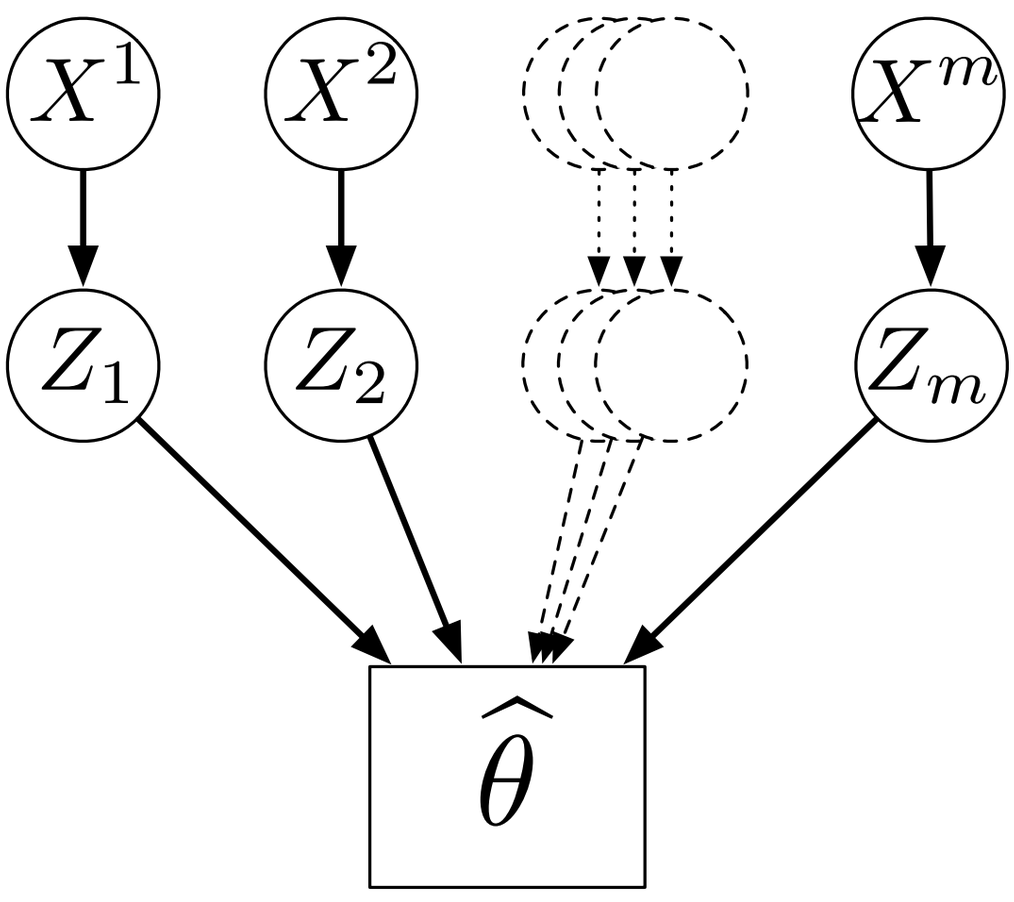
Communication-constraints

- Large data necessitates distributed storage
- Independent data collection (hospitals)
- Privacy?

[Yao 79; Abelson 80; Tsitsiklis and Luo 87; Han & Amari 98; Tatikonda & Mitter 04; ...]

Communication-constraints

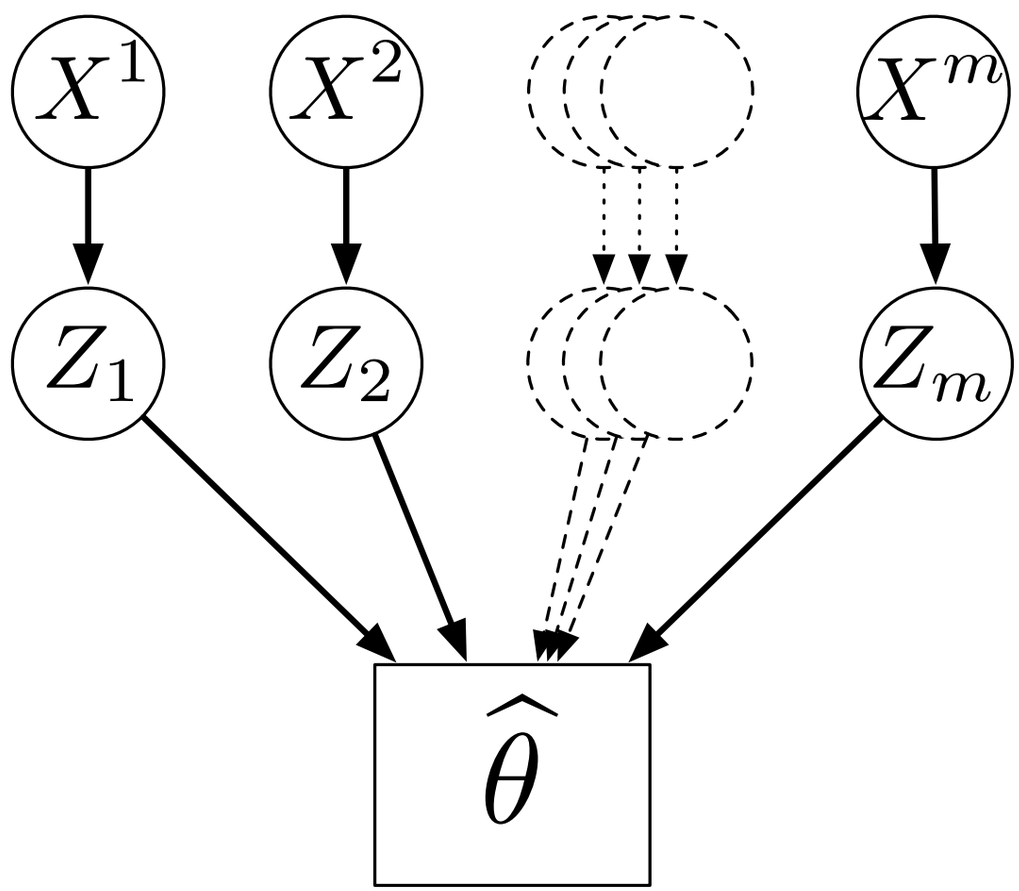
- Large data necessitates distributed storage
- Independent data collection (hospitals)
- Privacy?



[Yao 79; Abelson 80; Tsitsiklis and Luo 87; Han & Amari 98; Tatikonda & Mitter 04; ...]

Communication-constraints

- Large data necessitates distributed storage
- Independent data collection (hospitals)
- Privacy?



Setting: each of m agents has sample of size n

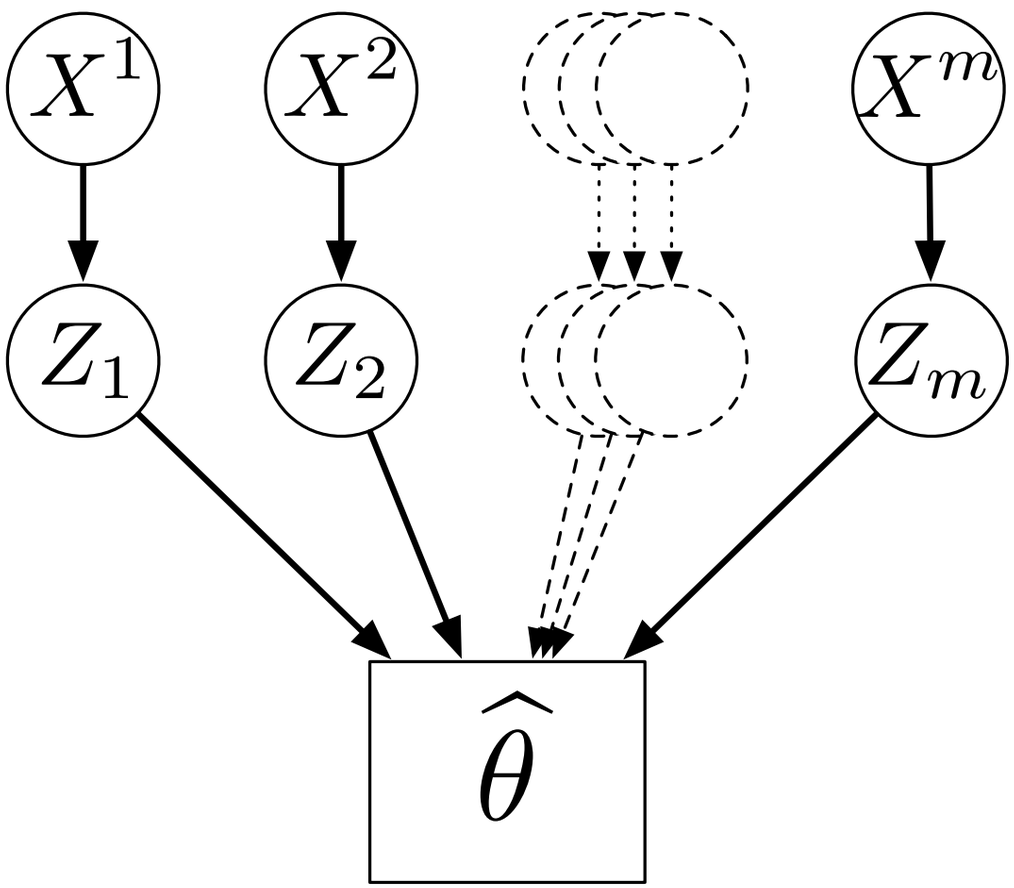
$$X^i = (X_1^i, X_2^i, \ldots, X_n^i)$$

Messages Z_i to fusion center

[Yao 79; Abelson 80; Tsitsiklis and Luo 87; Han & Amari 98; Tatikonda & Mitter 04; ...]

Communication-constraints

- Large data necessitates distributed storage
- Independent data collection (hospitals)
- Privacy?



Setting: each of m agents has sample of size n

$$X^i = (X_1^i, X_2^i, \ldots, X_n^i)$$

Messages Z_i to fusion center

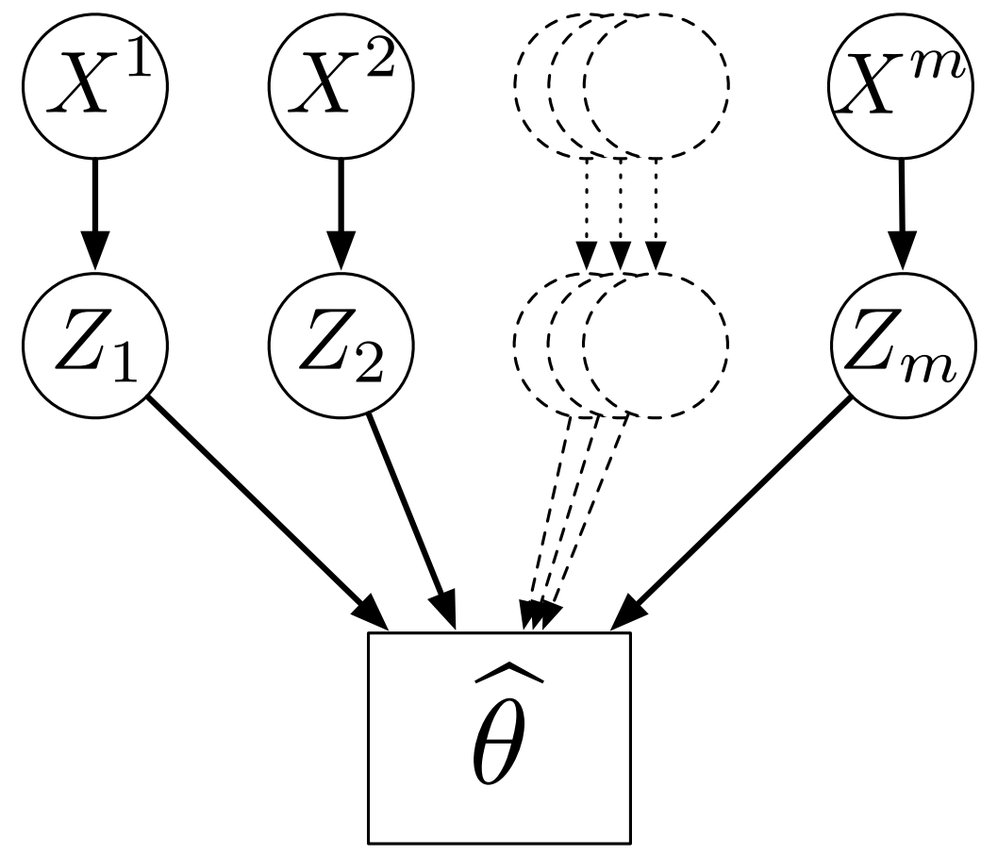
Question: tradeoffs between communication and statistical utility?

[Yao 79; Abelson 80; Tsitsiklis and Luo 87; Han & Amari 98; Tatikonda & Mitter 04; ...]

Minimax communication

Central object of study:

- Parameter $\theta(P)$ of distribution
- Family of distributions $\mathcal{P}$
- Loss $\|\cdot\|_2^2$



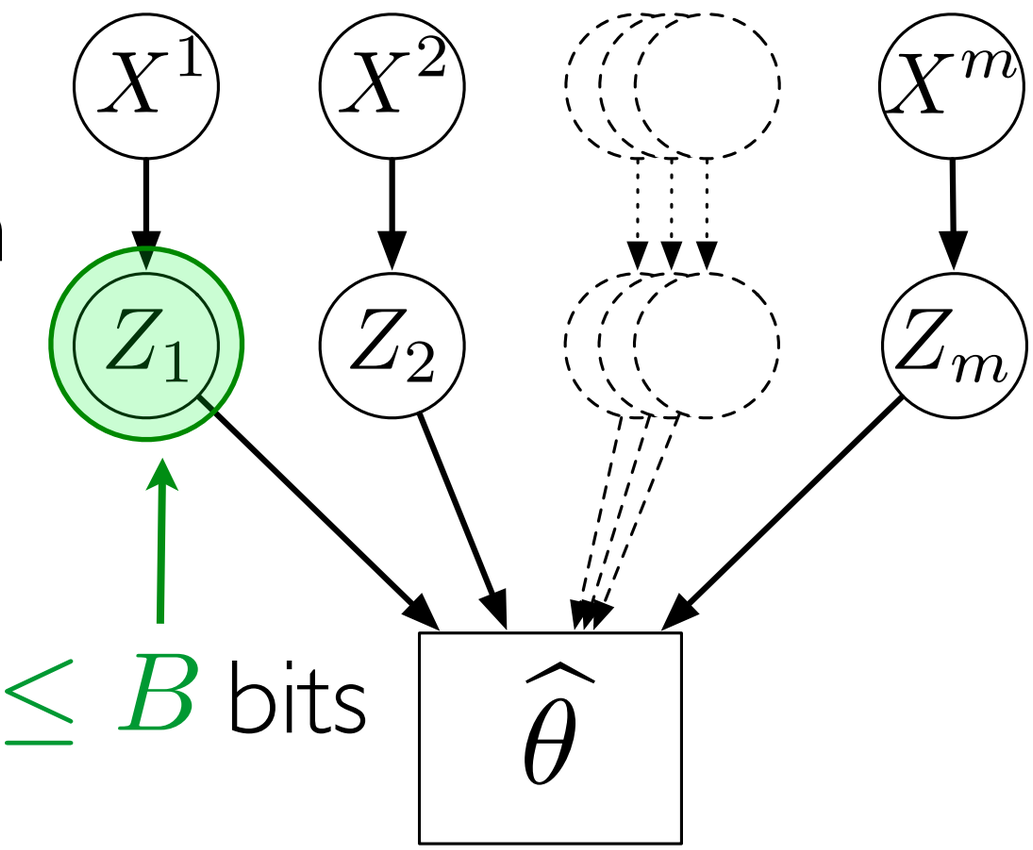
Minimax risk with B-bounded communication

$$\mathfrak{M}_n(\theta(\mathcal{P}), B) := \inf_{\pi \in \Pi_B} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\|\widehat{\theta}(Z_1^m) - \theta(P)\|_2^2 \right]$$

Minimax communication

Central object of study:

- Parameter $\theta(P)$ of distribution
- Family of distributions $\mathcal{P}$
- Loss $\|\cdot\|_2^2$



Constrained to be $\leq B$ bits

Minimax risk with B-bounded communication

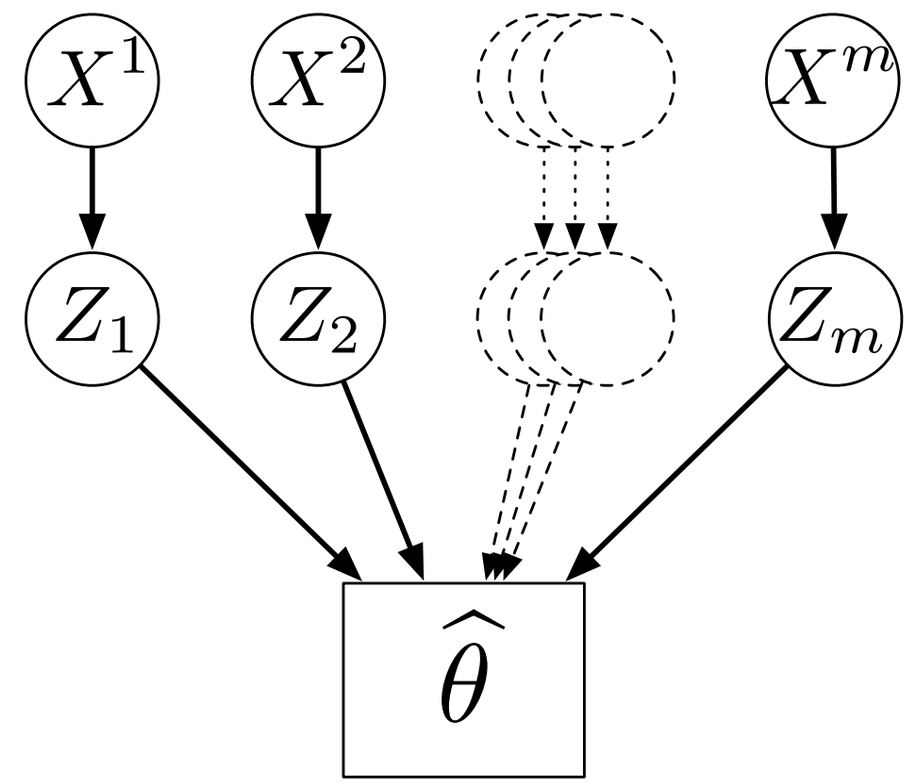
$$\mathfrak{M}_n(\theta(\mathcal{P}), B) := \inf_{\pi \in \Pi_B} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\|\widehat{\theta}(Z_1^m) - \theta(P)\|_2^2 \right]$$

Best protocol $Z_i = \pi(X^i)$ with Z_i smaller than B bits

Vignette: mean estimation

Consider estimation in normal location family, $\theta \in [-1, 1]^d$

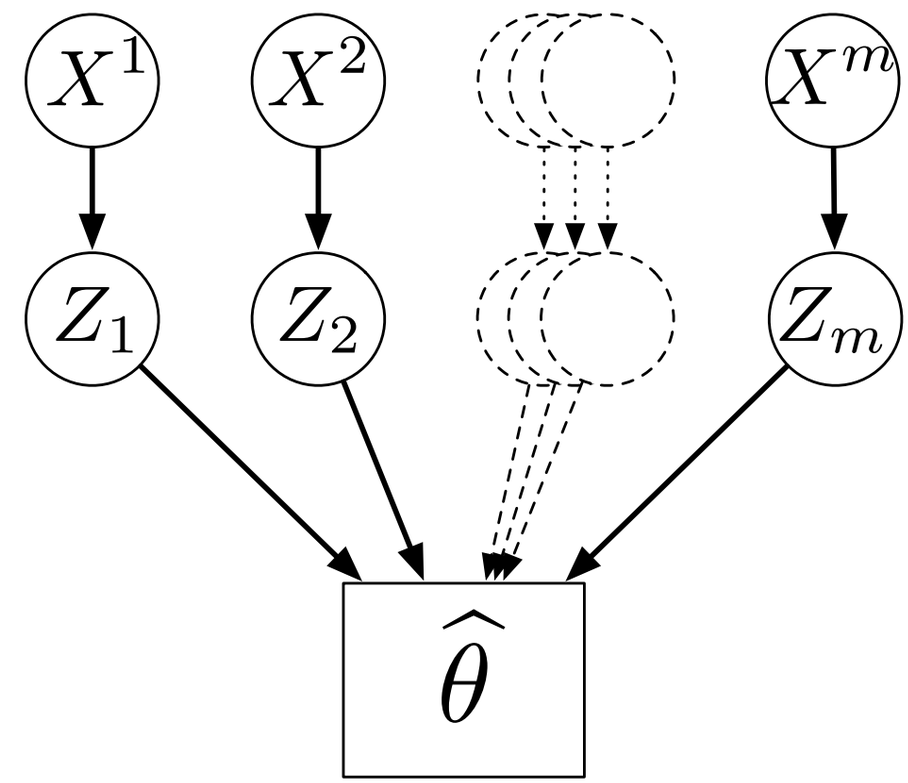
$$X_j^i \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2 I_{d \times d})$$



Vignette: mean estimation

Consider estimation in normal location family, $\theta \in [-1, 1]^d$

$$X_j^i \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2 I_{d \times d})$$



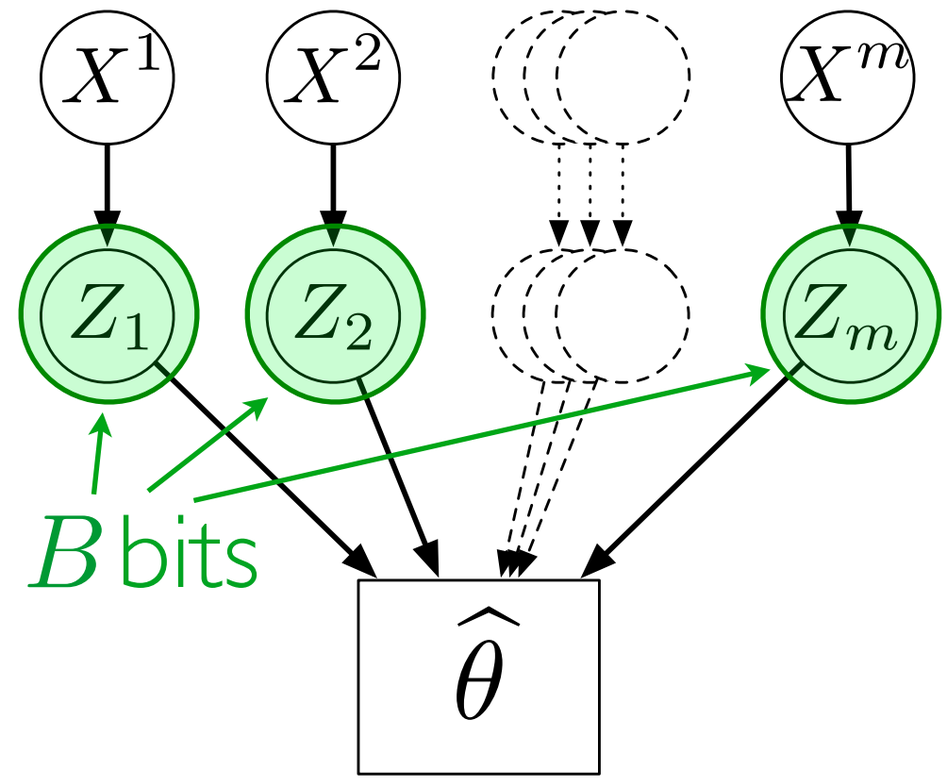
Theorem: when each agent has sample of size n
Minimax rate

$$\mathbb{E}[\|\widehat{\theta}(X^1, \ldots, X^m) - \theta\|_2^2] \asymp \frac{\sigma^2 d}{nm}$$

Vignette: mean estimation

Consider estimation in normal location family, $\theta \in [-1, 1]^d$

$$X_j^i \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2 I_{d \times d})$$



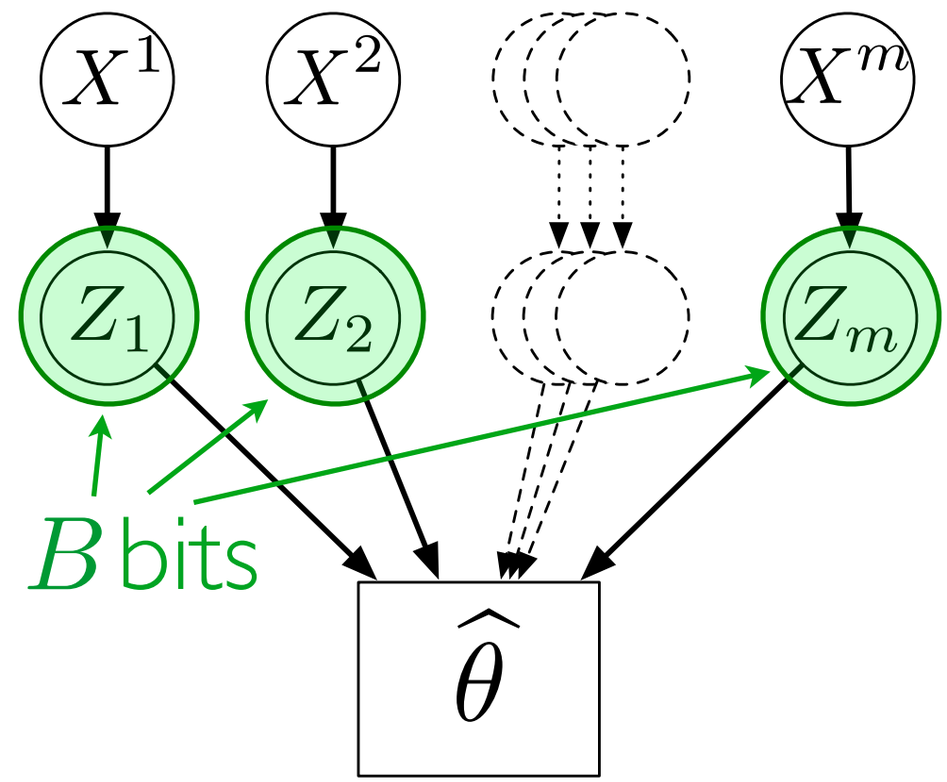
Theorem: when each agent has sample of size n
Minimax rate with B-bounded communication

$$\frac{d}{B \wedge d} \frac{1}{\log m} \frac{\sigma^2 d}{nm} \lesssim \mathfrak{M}_n(\mathcal{N}_d, B) \lesssim \frac{d \log m}{B \wedge d} \frac{\sigma^2 d}{nm}$$

Vignette: mean estimation

Consider estimation in normal location family, $\theta \in [-1, 1]^d$

$$X_j^i \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2 I_{d \times d})$$



Theorem: when each agent has sample of size n
Minimax rate with B-bounded communication

$$\frac{d}{B \wedge d} \frac{1}{\log m} \frac{\sigma^2 d}{nm} \lesssim \mathfrak{M}_n(\mathcal{N}_d, B) \lesssim \frac{d \log m}{B \wedge d} \frac{\sigma^2 d}{nm}$$

Consequence: each sends $\approx d$ bits for optimal estimation