Privacy Aware Learning

John C. Duchi, Michael I. Jordan, Martin J. Wainwright

University of California, Berkeley

NIPS 2012

An example

An example

Setting: We want to construct a good image classifier, say, of military personnel

Petraeus in a market

Source: Department of Defense

Petraeus thinking

Source: CBS News

Pictures Petraeus might not want shared

Pictures Petraeus might not want shared

Maybe he shouldn't share that?

Source: Rolling Stone

But I want a good image classifier

What should we do?

Start to develop theory of learning from private data

Start to develop theory of learning from private data

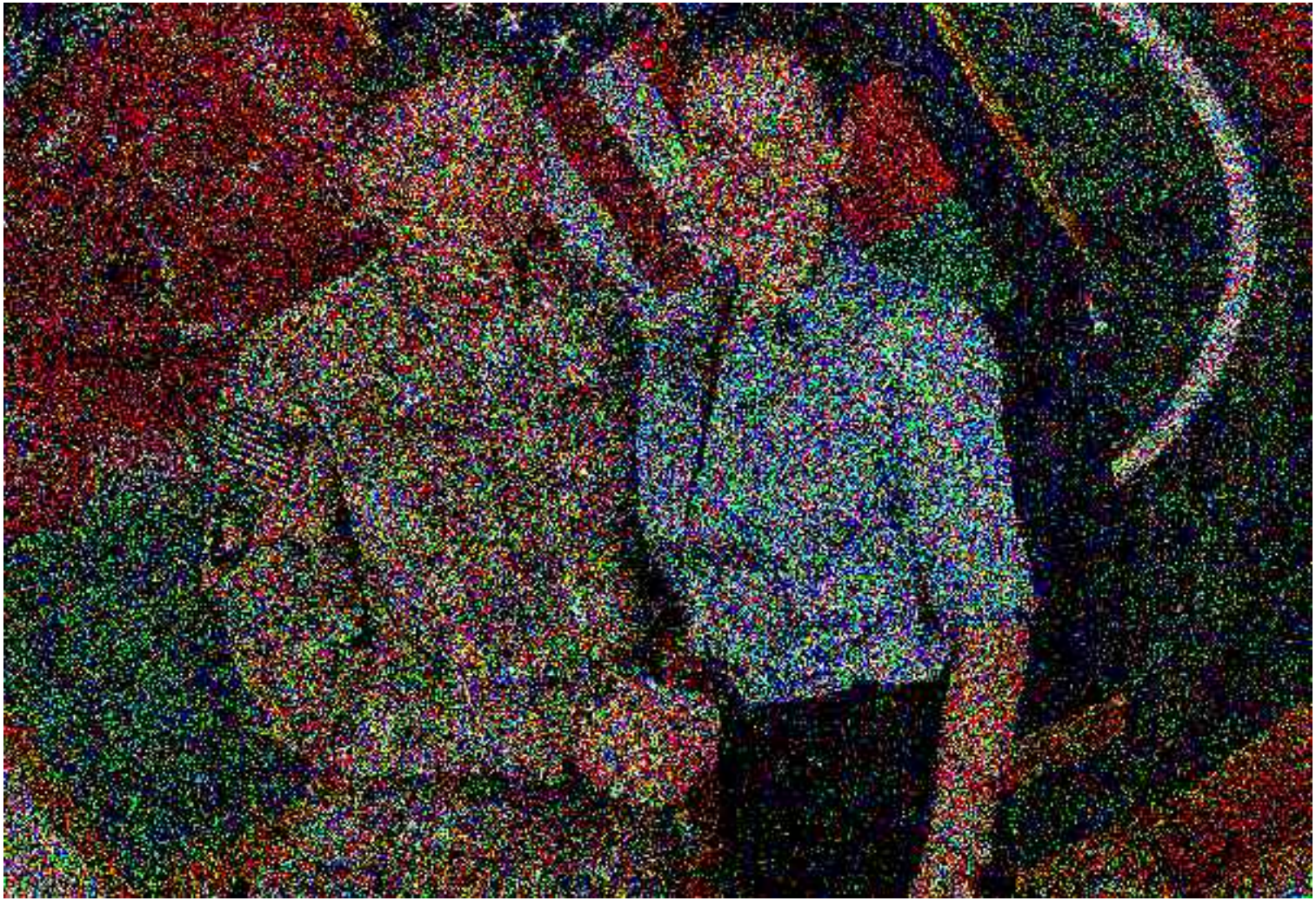
Instead of this

Start to develop theory of learning from private data

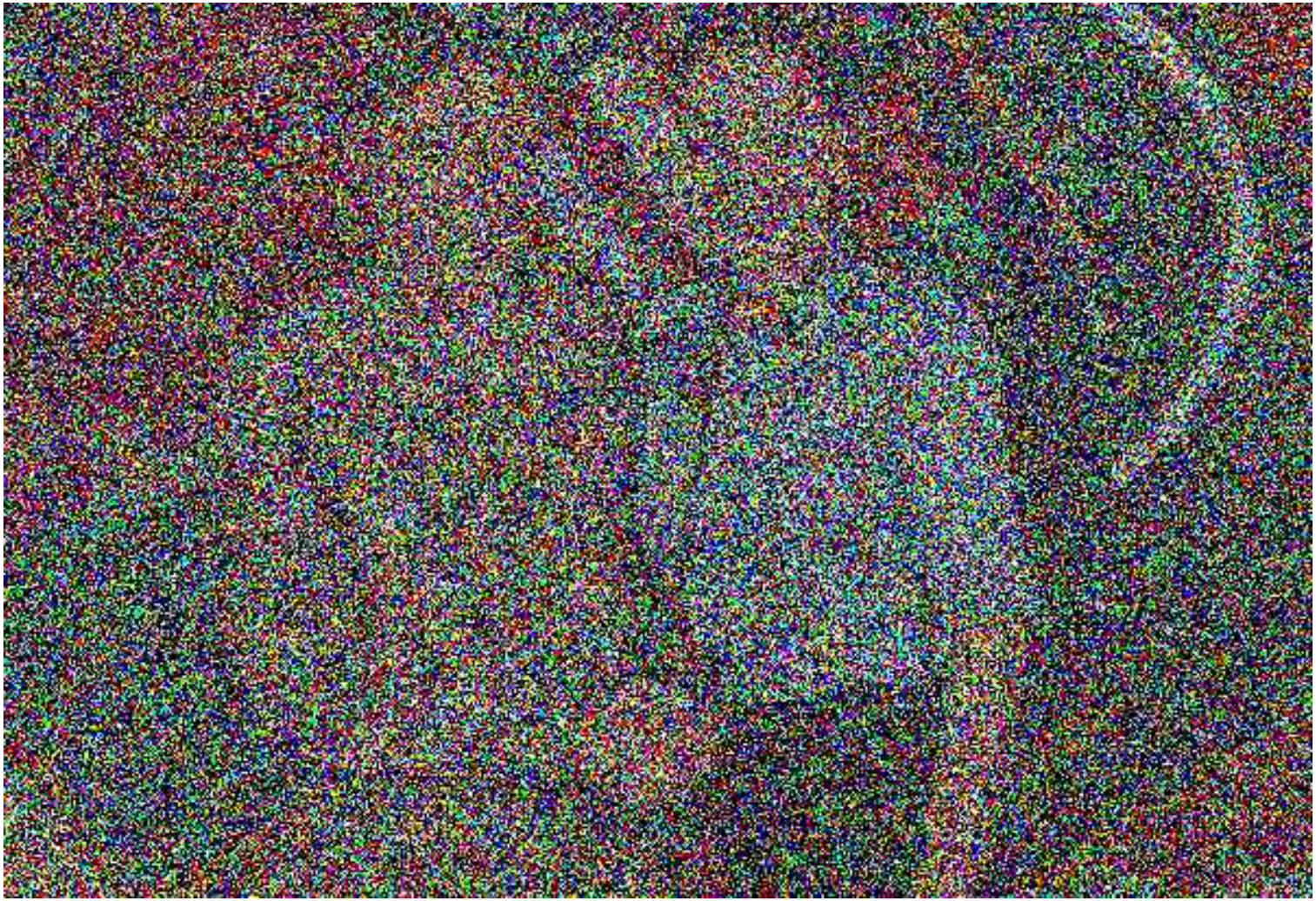
Start to develop theory of learning from private data

Learn from

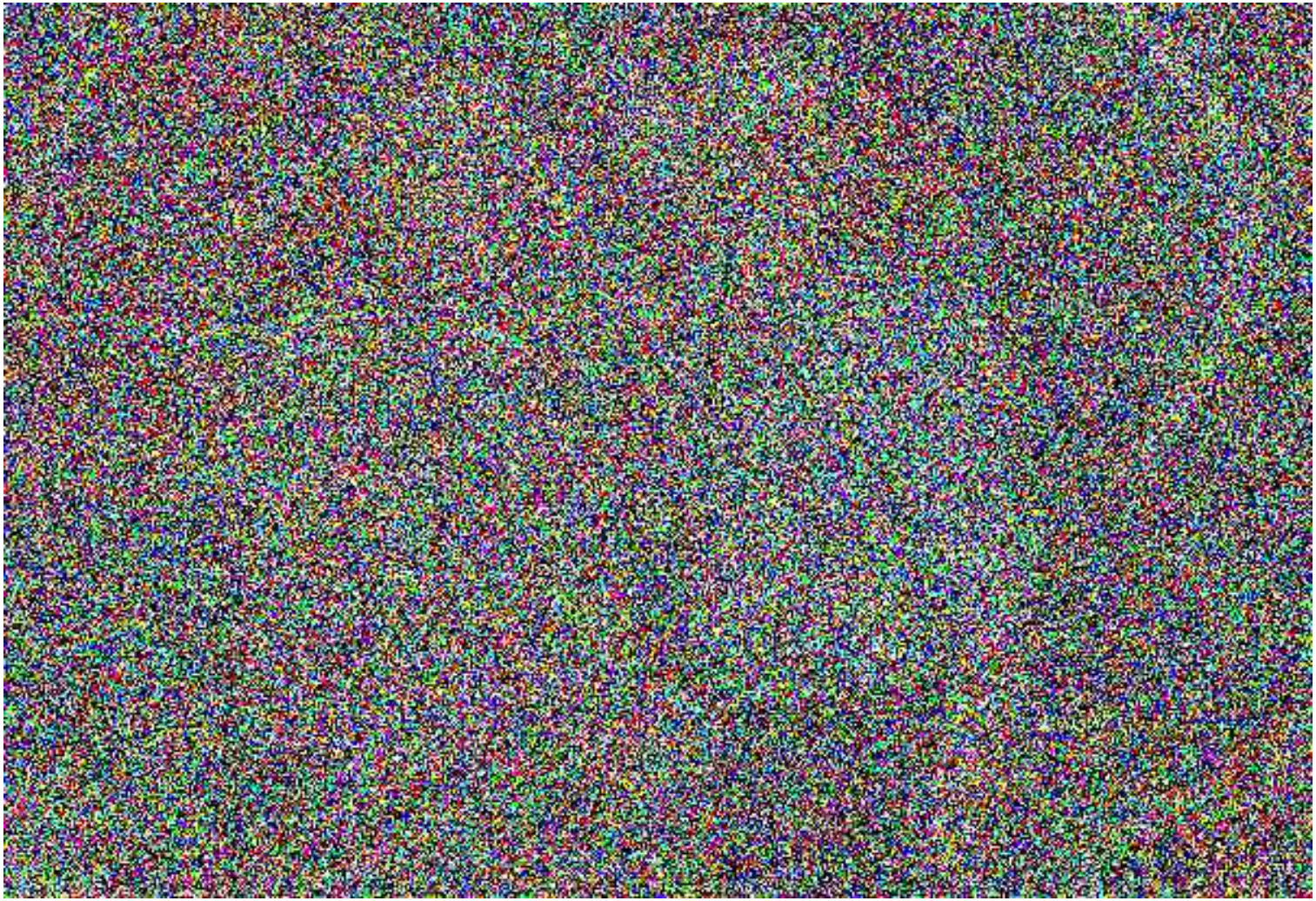
Start to develop theory of learning from private data



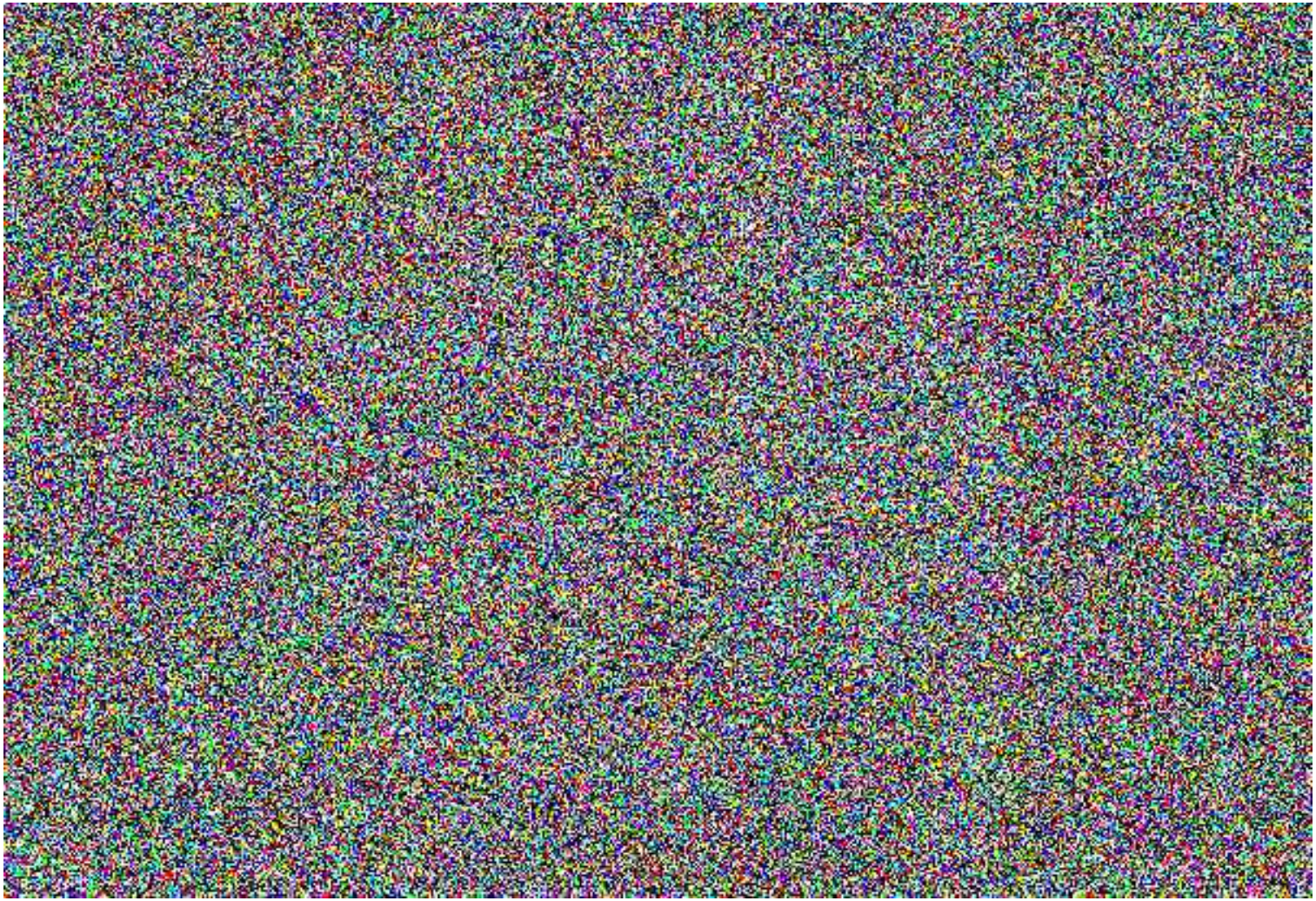
Start to develop theory of learning from private data



Start to develop theory of learning from private data

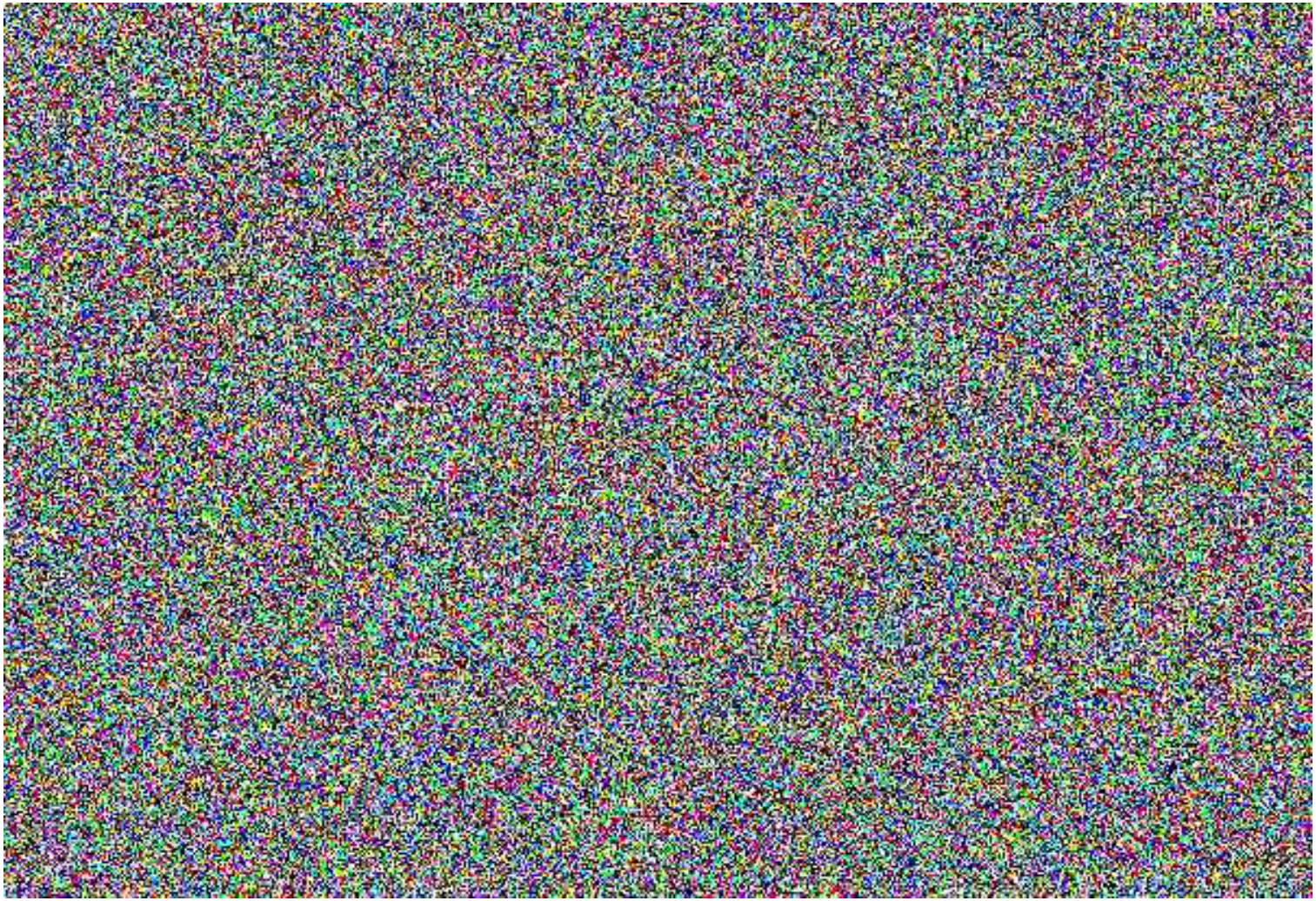


Start to develop theory of learning from private data



Start to develop theory of learning from private data

Start to develop theory of learning from private data



Outline

I Problem statement and motivating examples

II The privacy game

III Statistical estimation tradeoffs

IV Conclusions and future work

Tradeoffs

What are the tradeoffs between
maintaining privacy and statistical estimation?

Tradeoffs

What are the tradeoffs between
maintaining privacy and statistical estimation?

Fine-grained tradeoffs between privacy and utility

Setting

- Get samples $X_1, \ldots, X_n$
- Have a parameter θ we want to infer
- Measure performance of parameter θ with loss $\ell(\theta; X)$

Example: breast cancer prediction

- Data in (x, y) pairs (regressor $x \in \{-1, 1\}^d$, label $y \in \{\pm 1\}$)

$$x = \begin{bmatrix} 1 & -1 & 1 & -1 & \cdots & 1 \\ \text{Clump} & \text{Uniform} & \text{Adhesian} & \text{Chromatin} & \cdots & \text{Mitoses} \end{bmatrix}$$

$$y = \begin{cases} +1 & \text{if cancerous} \\ -1 & \text{if not} \end{cases}$$

Example: breast cancer prediction

- Data in (x, y) pairs (regressor $x \in \{-1, 1\}^d$, label $y \in \{\pm 1\}$)

$$x = \begin{bmatrix} 1 & -1 & 1 & -1 & \cdots & 1 \\ \text{Clump} & \text{Uniform} & \text{Adhesian} & \text{Chromatin} & \cdots & \text{Mitoses} \end{bmatrix}$$

$$y = \begin{cases} +1 & \text{if cancerous} \\ -1 & \text{if not} \end{cases}$$

- **Goal:** Find θ so that $\mathrm{sign}(\theta^\top x) = y$

Example: breast cancer prediction

- Data in (x, y) pairs (regressor $x \in \{-1, 1\}^d$, label $y \in \{\pm 1\}$)

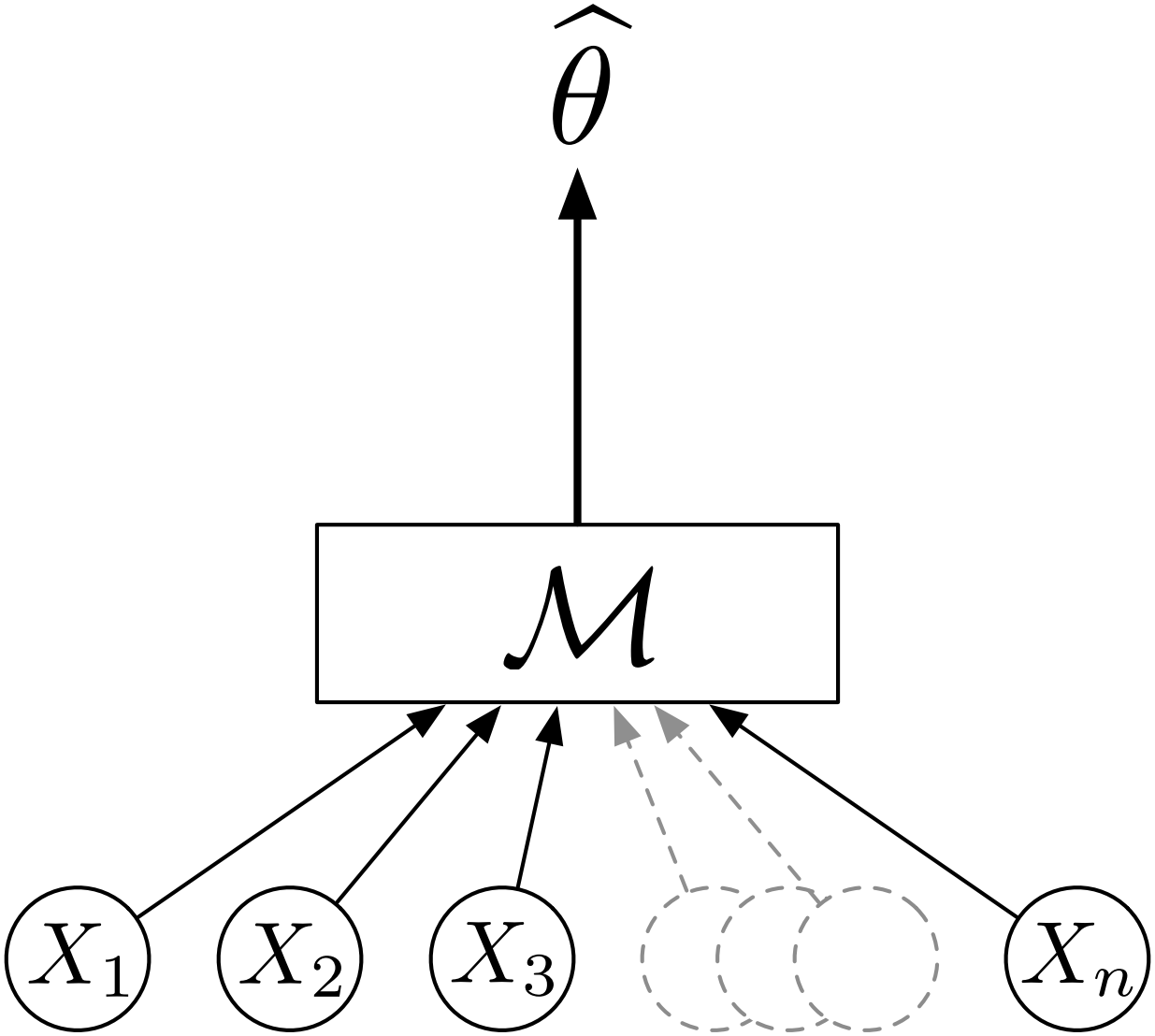
$$x = \begin{bmatrix} 1 & -1 & 1 & -1 & \cdots & 1 \\ \text{Clump} & \text{Uniform} & \text{Adhesian} & \text{Chromatin} & \cdots & \text{Mitoses} \end{bmatrix}$$

$$y = \begin{cases} +1 & \text{if cancerous} \\ -1 & \text{if not} \end{cases}$$

- **Goal:** Find θ so that $\mathrm{sign}(\theta^\top x) = y$
- **Loss:**

$$\ell(\theta; \{x, y\}) = \left[1 - y\theta^\top x\right]_+$$

Setting



Formal setting

Goal: minimize a risk R measuring performance of a parameter θ:

Formal setting

Goal: minimize a risk R measuring performance of a parameter θ:

$$\begin{aligned} \text{minimize} \quad & R(\theta) = \mathbb{E}[\ell(\theta; X)] \\ \text{subject to} \quad & \theta \in \Theta \end{aligned}$$

using samples $X_1, \ldots, X_n$.

Formal setting

Goal: minimize a risk R measuring performance of a parameter θ:

$$\begin{aligned} \text{minimize} \quad & R(\theta) = \mathbb{E}[\ell(\theta; X)] \\ \text{subject to} \quad & \theta \in \Theta \end{aligned}$$

using samples $X_1, \ldots, X_n$.

Question: Can we find $\widehat{\theta}$ so that $R(\widehat{\theta}) - R(\theta^\star)$ small without learning about $X_1, \ldots, X_n$?

Formal setting

Goal: minimize a risk R measuring performance of a parameter θ:

$$\begin{aligned} \text{minimize} \quad & R(\theta) = \mathbb{E}[\ell(\theta; X)] \\ \text{subject to} \quad & \theta \in \Theta \end{aligned}$$

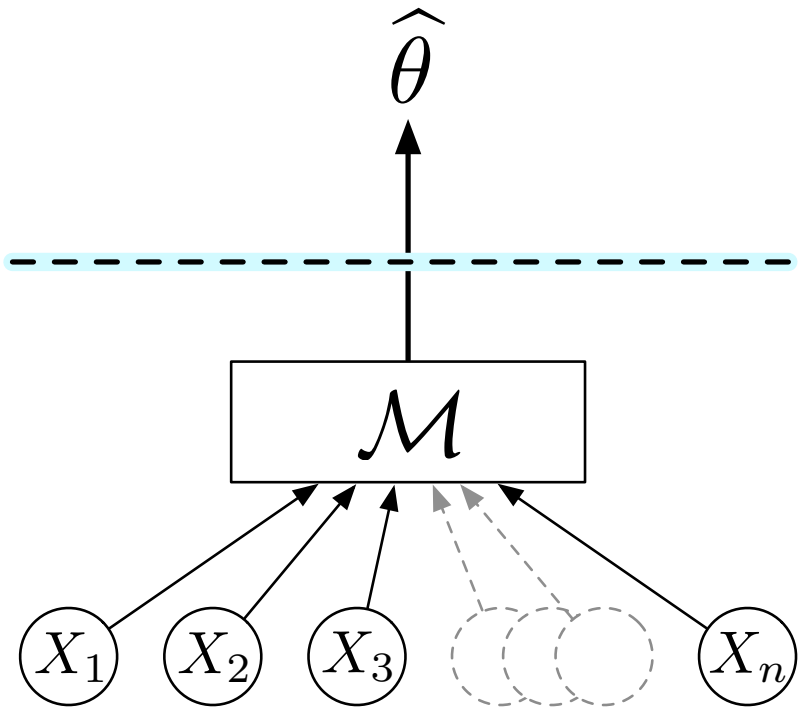
using samples $X_1, \ldots, X_n$.

Question: Can we find $\widehat{\theta}$ so that $R(\widehat{\theta}) - R(\theta^\star)$ small without learning about $X_1, \ldots, X_n$?

From but not *about*

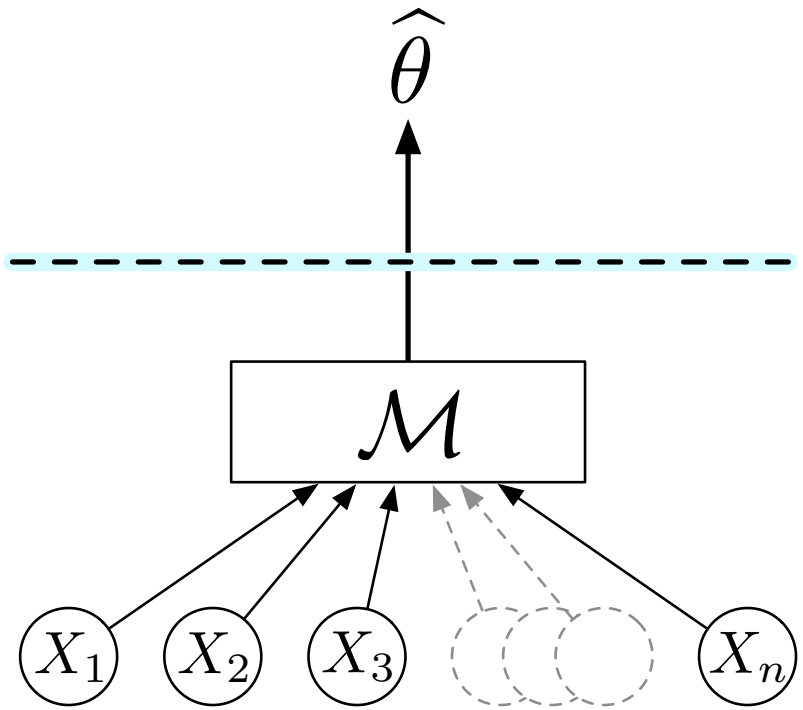
Formalizing privacy

Prior work: Lots (Chaudhuri and collaborators, Dwork et al., Wasserman and Zhou)



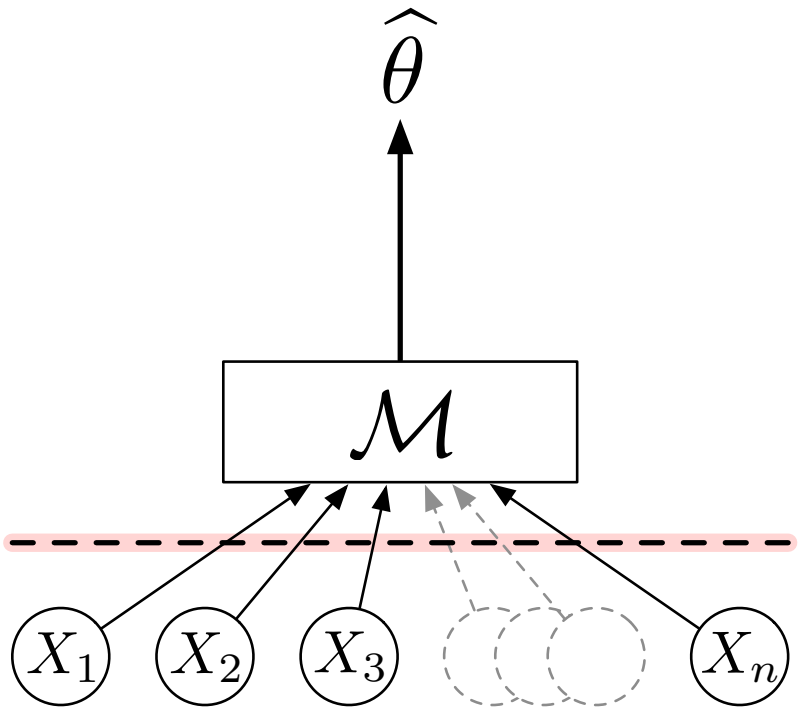
Formalizing privacy

Prior work: Lots (Chaudhuri and collaborators, Dwork et al., Wasserman and Zhou)
Local Privacy: Changing privacy barrier (Evfimievski et al. 2003, Warner 1965)



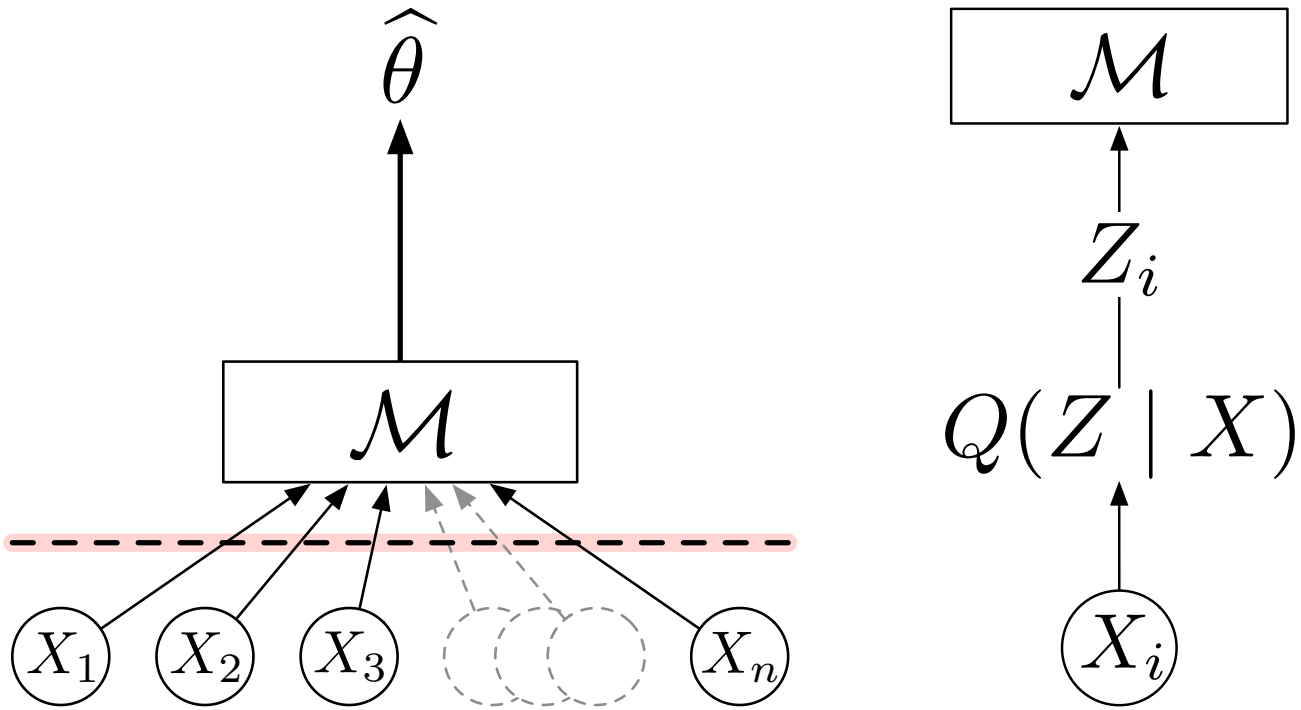
Formalizing privacy

Prior work: Lots (Chaudhuri and collaborators, Dwork et al., Wasserman and Zhou)
Local Privacy: Changing privacy barrier (Evfimievski et al. 2003, Warner 1965)



Formalizing privacy

Prior work: Lots (Chaudhuri and collaborators, Dwork et al., Wasserman and Zhou)
Local Privacy: Changing privacy barrier (Evfimievski et al. 2003, Warner 1965)



How do we get privacy?

Example: Classification

- Data pairs (x, y) with $x \in \{-1, 1\}^d$, label $y \in \{\pm 1\}$

$$x = \begin{bmatrix} 1 & -1 & 1 & 1 & \cdots & -1 \\ \text{Clump} & \text{Uniform} & \text{Adheres} & \text{Chromatin} & \cdots & \text{Mitoses} \end{bmatrix}$$

$$y = \begin{cases} +1 & \text{if cancerous} \\ -1 & \text{if not} \end{cases}$$

How do we get privacy?

Example: Classification

- Data pairs (x, y) with $x \in \{-1, 1\}^d$, label $y \in \{\pm 1\}$

$$x = \begin{bmatrix} 1 & -1 & 1 & 1 & \cdots & -1 \\ \text{Clump} & \text{Uniform} & \text{Adheres} & \text{Chromatin} & \cdots & \text{Mitoses} \end{bmatrix}$$

$$y = \begin{cases} +1 & \text{if cancerous} \\ -1 & \text{if not} \end{cases}$$

Idea: Add independent random noise W to coordinates of x:

$$Z_i = X_i + W$$

How do we get privacy?

Example: Classification

- Data pairs (x, y) with $x \in \{-1, 1\}^d$, label $y \in \{\pm 1\}$

$$x = \begin{bmatrix} 1 & -1 & 1 & 1 & \cdots & -1 \\ \text{Clump} & \text{Uniform} & \text{Adheres} & \text{Chromatin} & \cdots & \text{Mitoses} \end{bmatrix}$$

$$y = \begin{cases} +1 & \text{if cancerous} \\ -1 & \text{if not} \end{cases}$$

Idea: Add independent random noise W to coordinates of x:

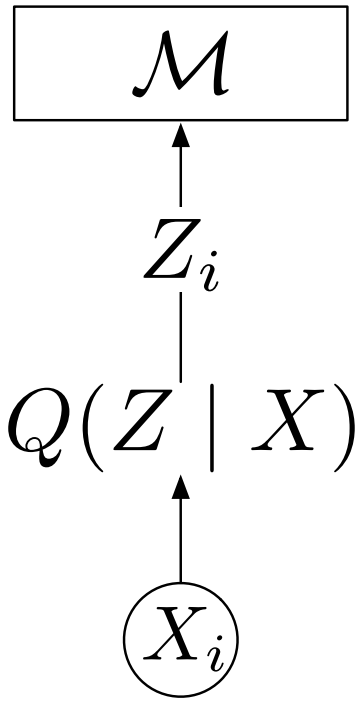
$$Z_i = X_i + W$$

Problem: This is *highly suboptimal*, dimension dependence blows up

Communication model

Local Privacy: Communication model to study minimization of

$$R(\theta) = \mathbb{E}[\ell(\theta; X)]$$

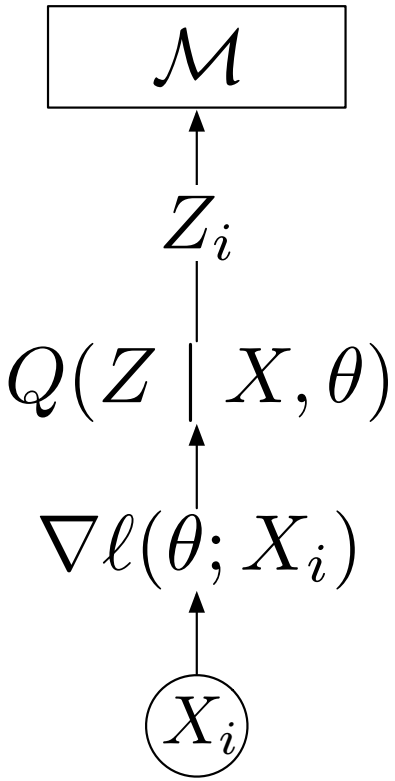


Communication model

Local Privacy: Communication model to study minimization of

$$R(\theta) = \mathbb{E}[\ell(\theta; X)]$$

- Communicate $\nabla \ell(\theta; X_i)$

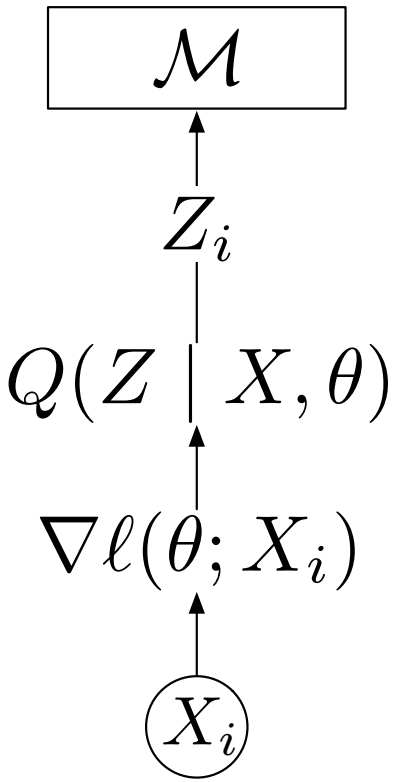


Communication model

Local Privacy: Communication model to study minimization of

$$R(\theta) = \mathbb{E}[\ell(\theta; X)]$$

- Communicate $\nabla \ell(\theta; X_i)$
  - Want to minimize, $\nabla \ell$ is sufficient
  - Use stochastic optimization techniques with $\nabla \ell(\theta; x_i)$



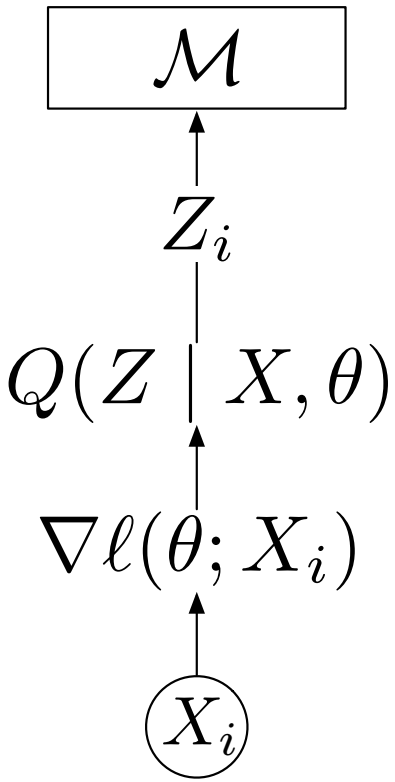
Communication model

Local Privacy: Communication model to study minimization of

$$R(\theta) = \mathbb{E}[\ell(\theta; X)]$$

- Communicate $\nabla\ell(\theta; X_i)$
  - Want to minimize, $\nabla\ell$ is sufficient
  - Use stochastic optimization techniques with $\nabla\ell(\theta; x_i)$
- Really communicate Z_i with property

$$\mathbb{E}_Q[Z_i \mid \theta, X_i] = \nabla\ell(\theta; X_i)$$



Main Contributions

Contribution 1: Optimal types of noise to guarantee privacy

Main Contributions

Contribution 1: Optimal types of noise to guarantee privacy

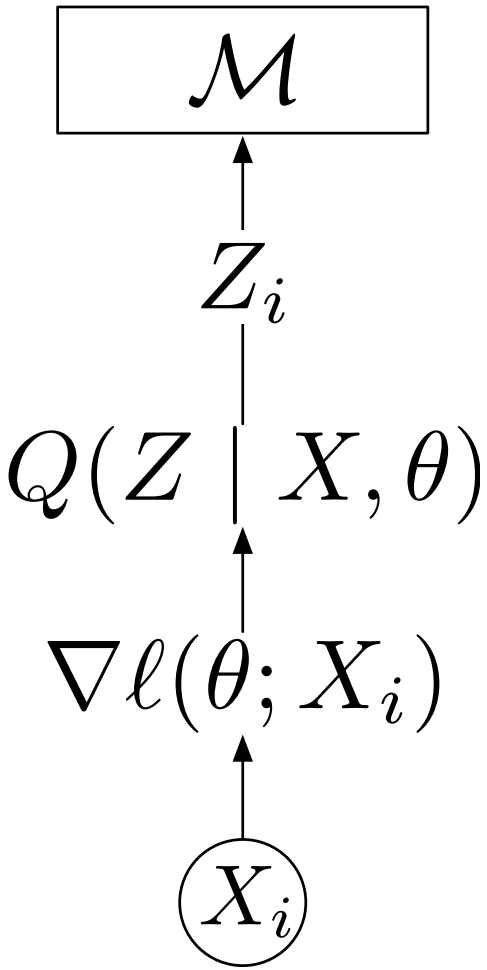
Contribution 2: Sharp upper and lower bounds on convergence rates as a function of privacy

Privacy Saddle Points

Optimal Local Privacy:

Maximize privacy of Q subject to

$$\mathbb{E}_Q[Z \mid \theta, X] = \nabla \ell(\theta; X)$$



Privacy saddle points

Goal: Maximize privacy of Z for X subject to $\widehat{\theta}$ being learnable (some constraints on Z)

Privacy saddle points

Goal: Maximize privacy of Z for X subject to $\widehat{\theta}$ being learnable (some constraints on Z)

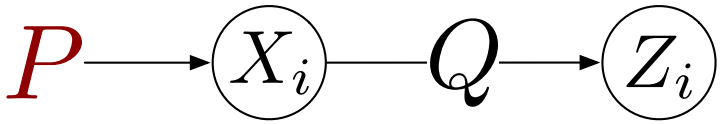
Privacy metric: mutual information

Privacy saddle points

Goal: Maximize privacy of Z for X subject to $\widehat{\theta}$ being learnable (some constraints on Z)

Privacy metric: mutual information

$$\sup_{P \in \mathcal{P}} I(P; Q)$$

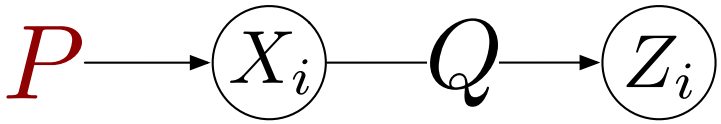


Privacy saddle points

Goal: Maximize privacy of Z for X subject to $\widehat{\theta}$ being learnable (some constraints on Z)

Privacy metric: mutual information

$$\sup_{P \in \mathcal{P}} I(P; Q)$$



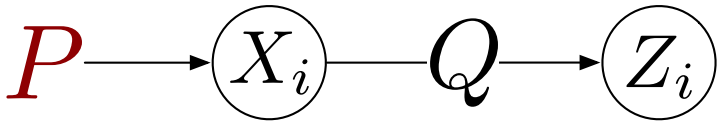
Worst case information measure

Privacy saddle points

Goal: Maximize privacy of Z for X subject to $\widehat{\theta}$ being learnable (some constraints on Z)

Privacy metric: mutual information

$$\sup_{P \in \mathcal{P}} I(P; Q)$$



Worst case information measure

Strategy: We provide general solution to

$$\underset{Q}{\text{minimize}} \sup_{P \in \mathcal{P}} I(P; Q)$$

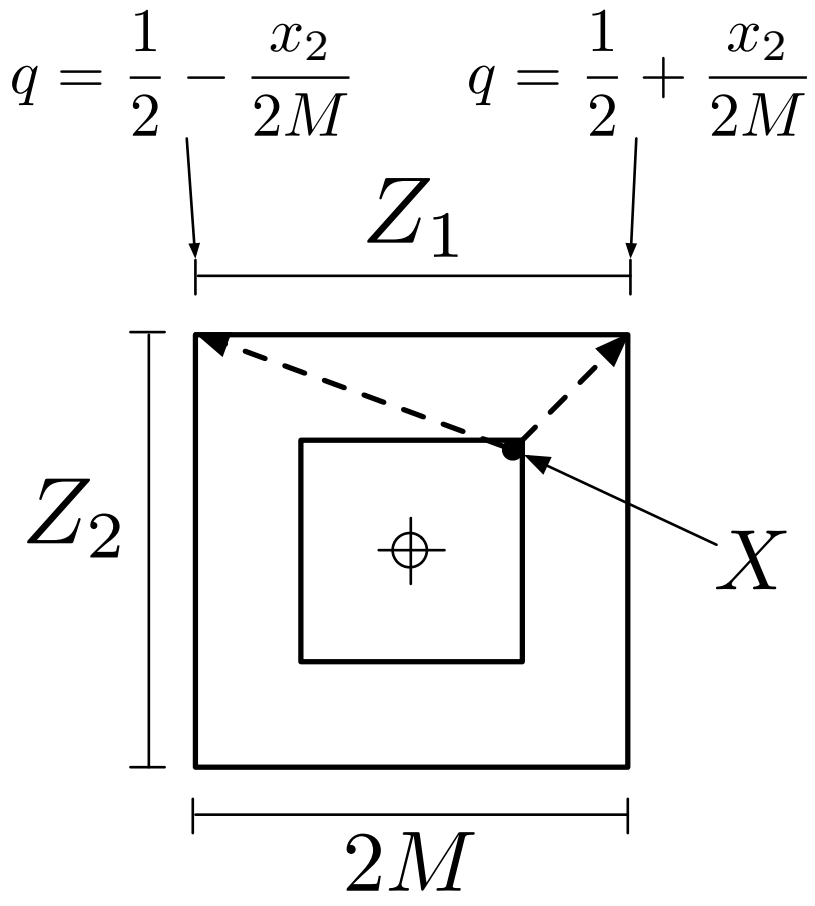
over distributions Q with larger support than P

Mutual information saddle point example

Setting: Data $x \in \{-1, 1\}^d$, allow z to be in $\|z\|_\infty \leq M$.

Mutual information saddle point example

Setting: Data $x \in \{-1, 1\}^d$, allow z to be in $\|z\|_\infty \leq M$.



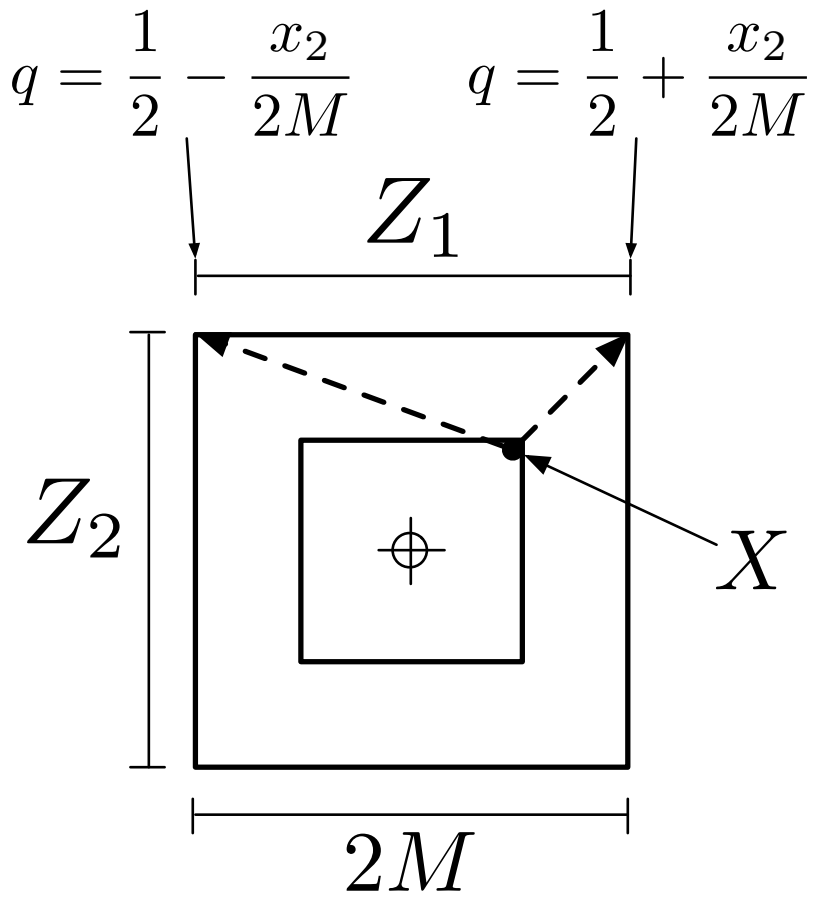
Mutual information saddle point example

Setting: Data $x \in \{-1, 1\}^d$, allow z to be in $\|z\|_\infty \le M$.

Optimal distribution Q given X:

- Independent coordinates $z_i \in \{-M, M\}$
- Distribution

$$Q^*(Z_i = M \mid X) = \frac{1}{2} + \frac{X_i}{2M}$$



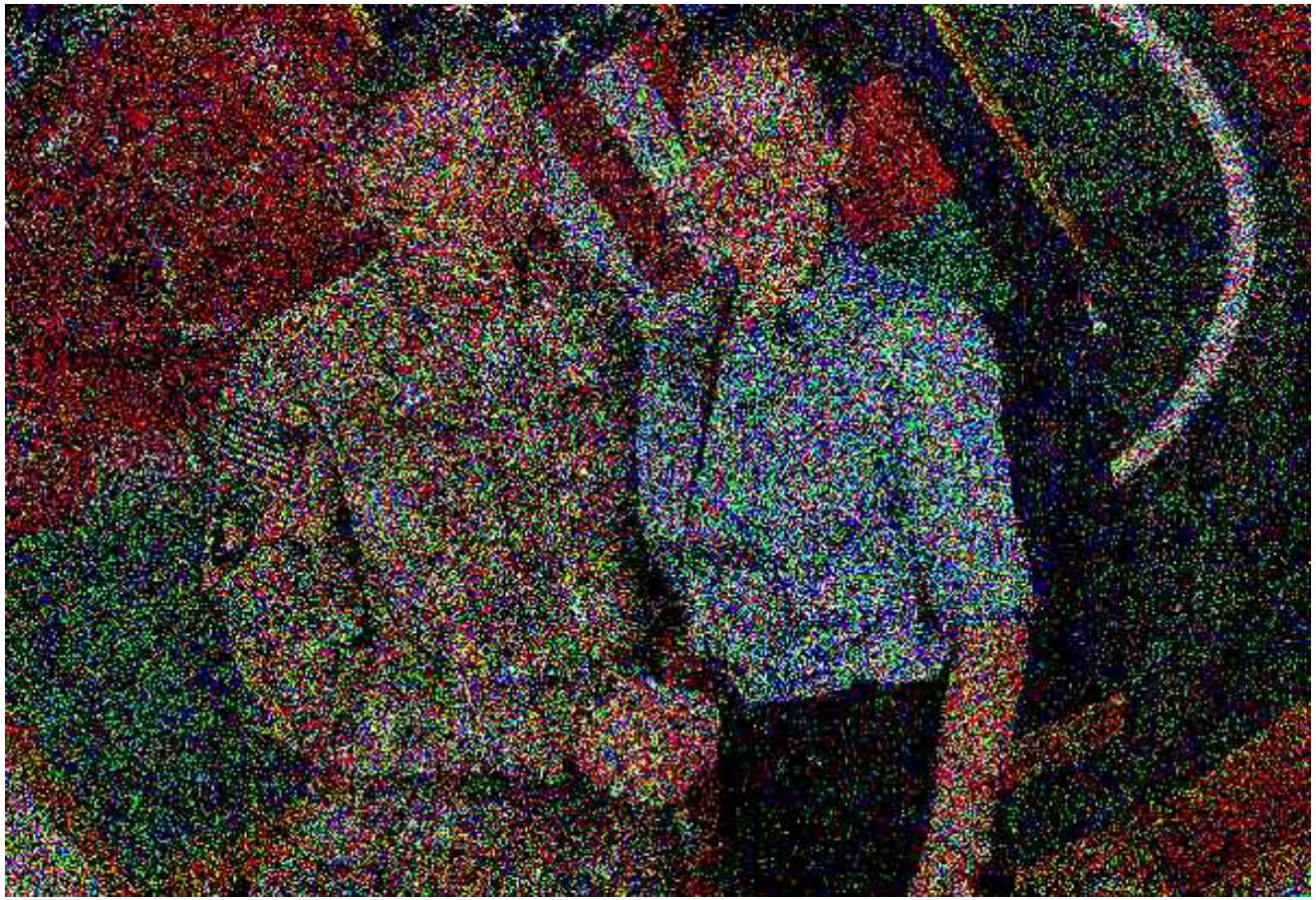
Example of optimal perturbation

Example of optimal perturbation

1 bit per bit

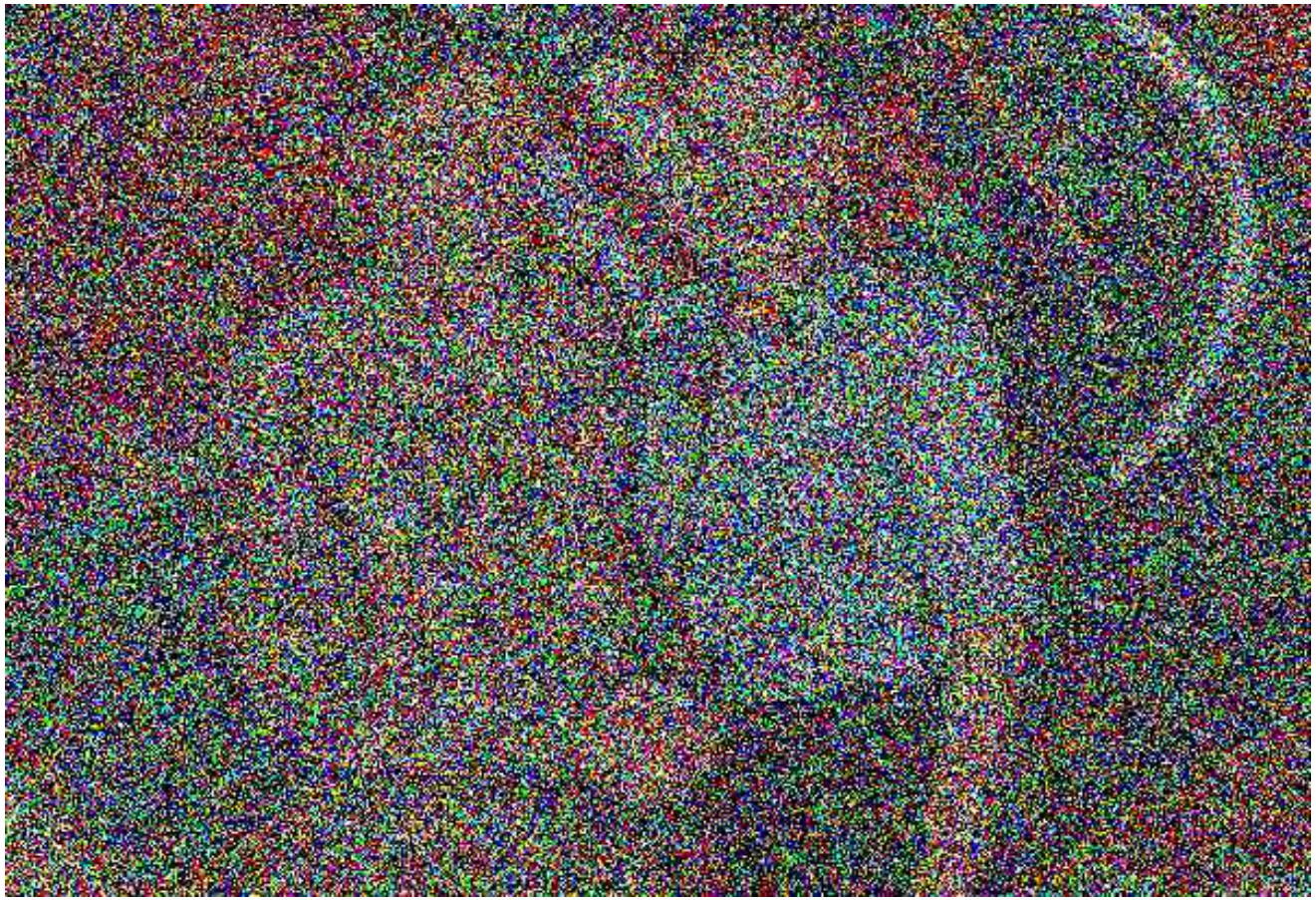
Example of optimal perturbation

.13 bits per bit
(2× slower
convergence)



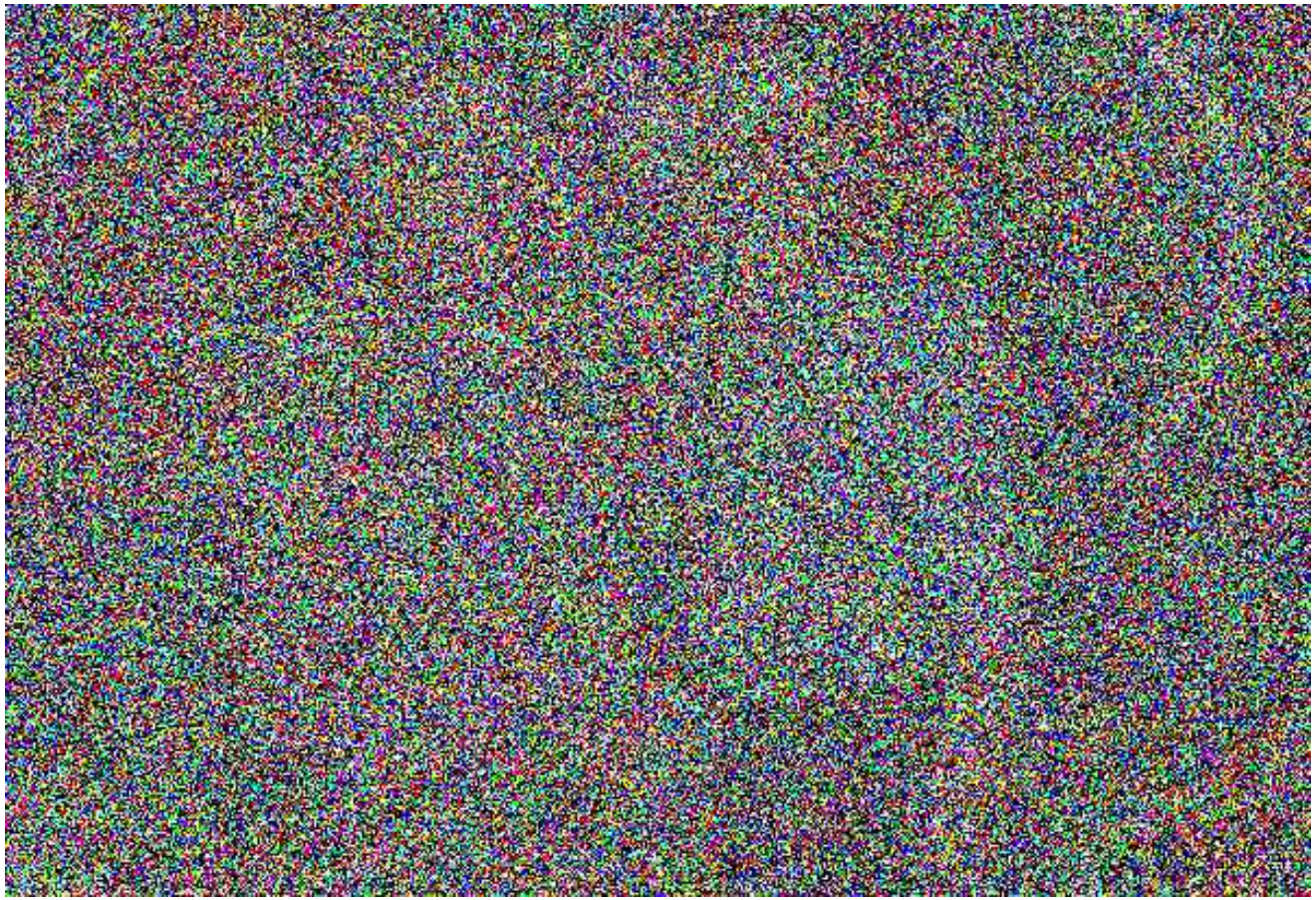
Example of optimal perturbation

.033 bits per bit
(4$\times$ slower
convergence)



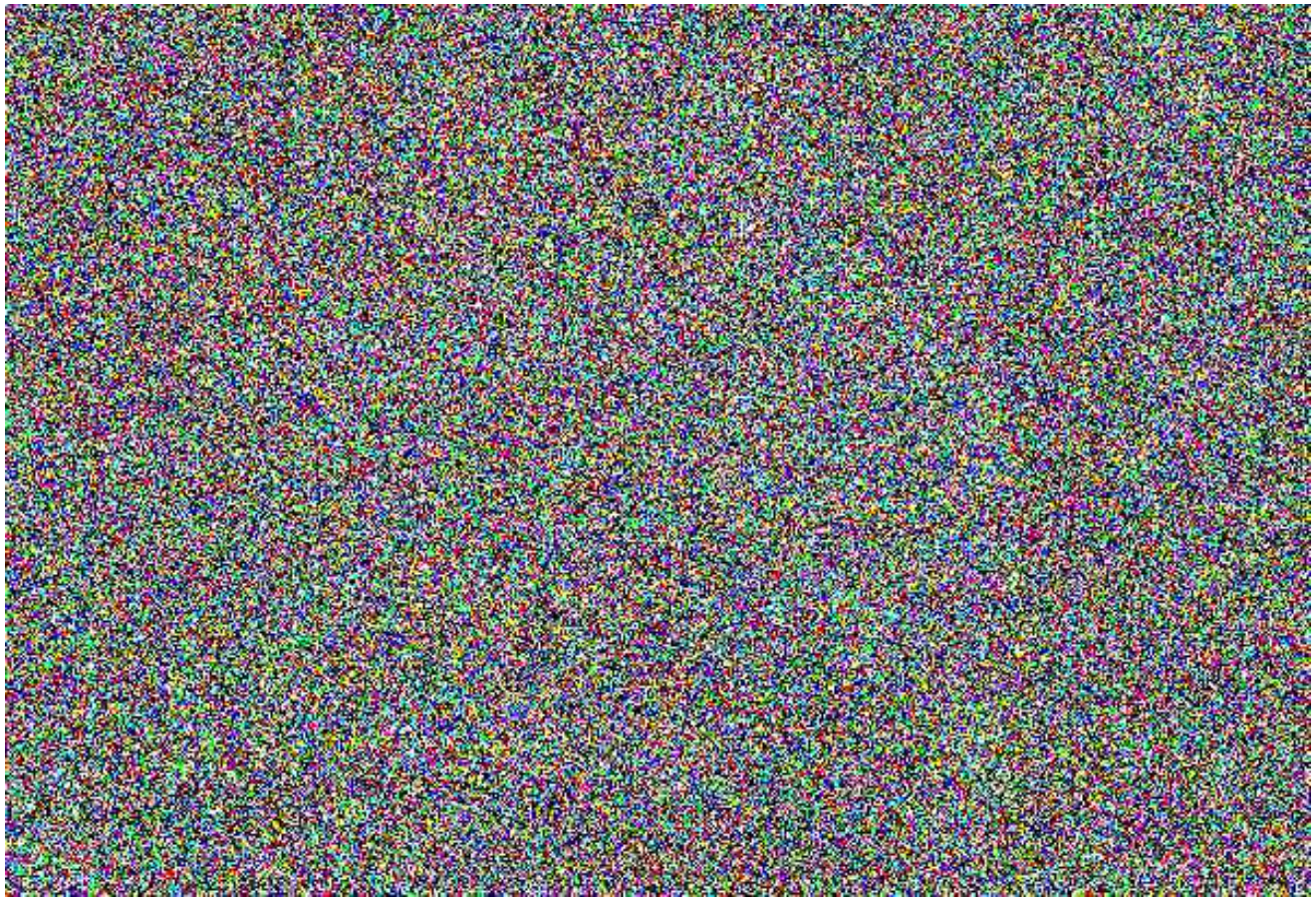
Example of optimal perturbation

.0081 bits per bit
(8× slower
convergence)



Example of optimal perturbation

.002 bits per bit
(16× slower
convergence)

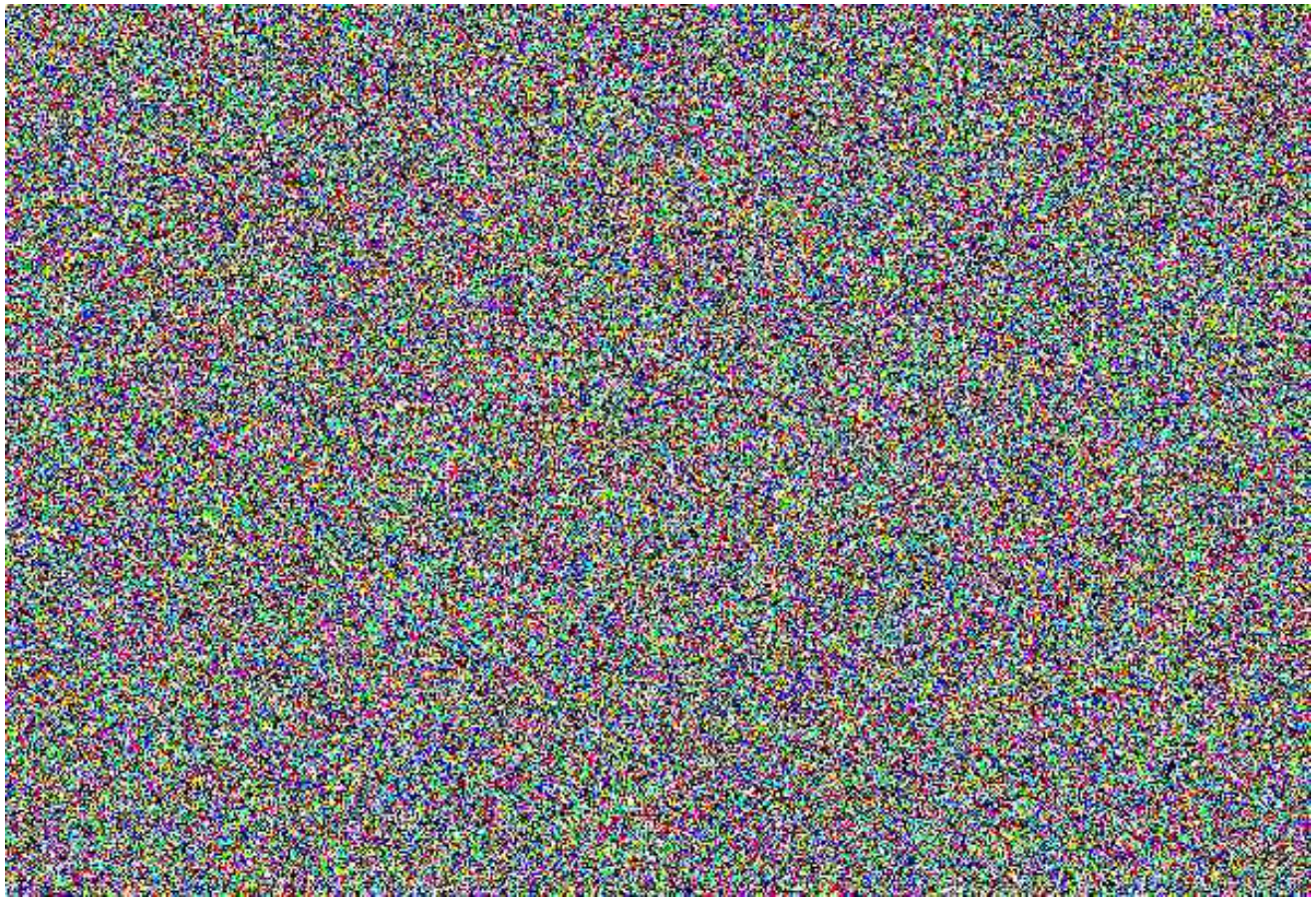


Example of optimal perturbation

.0005 bits per bit
(32× slower
convergence)

Example of optimal perturbation

.00013 bits per bit
(64$\times$ slower
convergence)



Statistical estimation and convergence rates

Exhibiting tradeoffs

Goal: Understand tradeoff between mutual information bound

$$I^* := \min_Q \sup_P I(X; Z)$$

and number of samples n

Reminder: $\widehat{\theta}$ is our estimate, based on $X_1, \ldots, X_n$,

$$R(\theta) := \mathbb{E}[\ell(\theta; X)]$$

Exhibiting tradeoffs

Goal: Understand tradeoff between mutual information bound

$$I^* := \min_Q \sup_P I(X; Z)$$

and number of samples n for risk minimization problems

Exhibiting tradeoffs

Goal: Understand tradeoff between mutual information bound

$$I^* := \min_Q \sup_P I(X; Z)$$

and number of samples n for risk minimization problems

Theorem: Effective sample size for d dimensional problem is made worse by

$$n \mapsto \frac{nI^*}{d}$$

Exhibiting tradeoffs

Goal: Understand tradeoff between mutual information bound

$$I^* := \min_Q \sup_P I(X; Z)$$

and number of samples n for risk minimization problems

Theorem: Effective sample size for d dimensional problem is made worse by

$$n \mapsto \frac{nI^*}{d}$$

- Lower bound holds for all methods
- Upper bound achieved by stochastic approximation

Exhibiting tradeoffs

Have mutual information

$$I^* := \min_Q \sup_P I(X; Z)$$

Theorem: Optimality gap for d dimensional problem

$$\Omega(1)\frac{1}{\sqrt{n}} \leq \mathbb{E}[R(\widehat{\theta})] - R(\theta^\star) \leq \mathcal{O}(1)\frac{1}{\sqrt{n}}$$

Exhibiting tradeoffs

Have mutual information

$$I^* := \min_Q \sup_P I(X; Z)$$

Theorem: Optimality gap for d dimensional problem

$$\Omega(1)\frac{\sqrt{d}}{\sqrt{nI^*}} \leq \mathbb{E}[R(\widehat{\theta})] - R(\theta^\star) \leq \mathcal{O}(1)\frac{\sqrt{d}}{\sqrt{nI^*}}$$

Exhibiting tradeoffs

Have mutual information

$$I^* := \min_Q \sup_P I(X; Z)$$

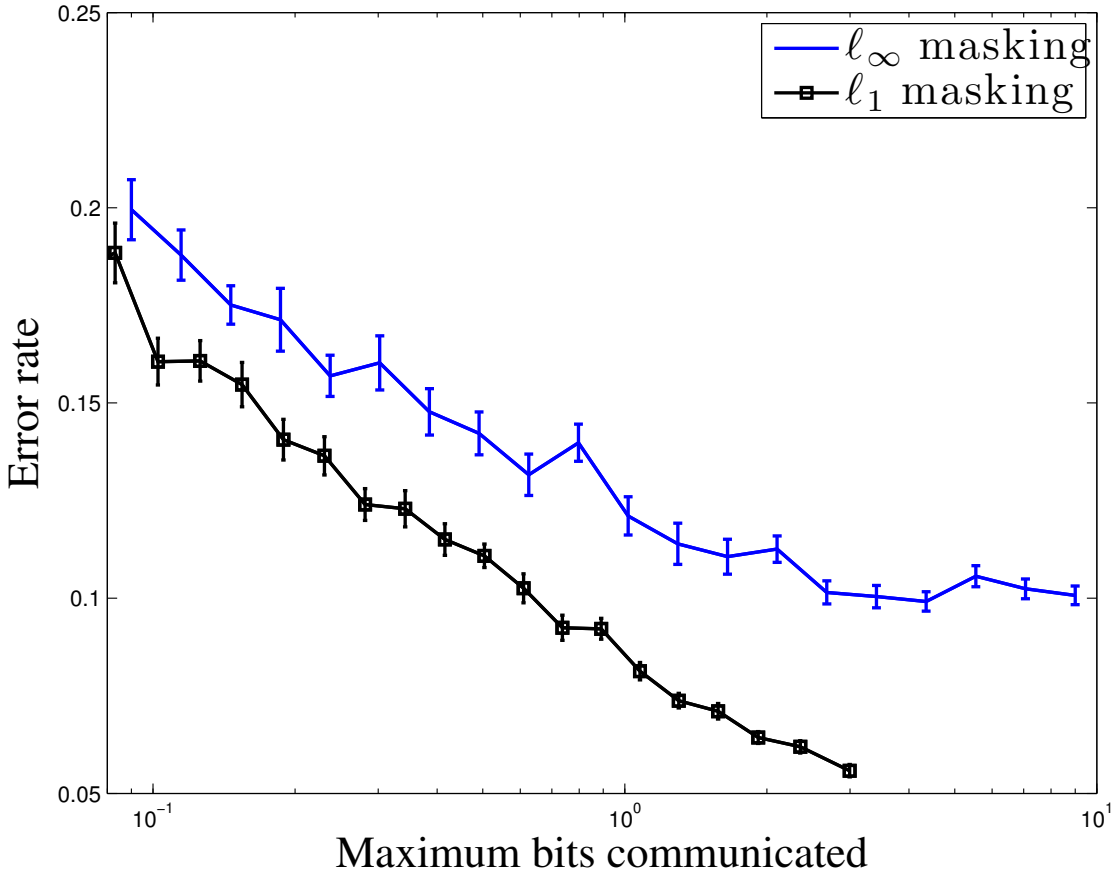
Theorem: Optimality gap for d dimensional problem

$$\Omega(1)\frac{\sqrt{d}}{\sqrt{nI^*}} \leq \mathbb{E}[R(\widehat{\theta})] - R(\theta^\star) \leq \mathcal{O}(1)\frac{\sqrt{d}}{\sqrt{nI^*}}$$

- Lower bound holds for all methods
- Upper bound achieved by stochastic approximation

Experimental example: breast cancer prediction

- Regressors x are markers for breast cancer, labels y are presence/absence of tumor
- Measure predictive performance: count $\mathrm{sign}(\theta^\top x_i) = y_i$



Conclusions and future work

1. Have given sharp rates of convergence when providers of data play "privacy game." Extensions to differential privacy as well.

Conclusions and future work

1. Have given sharp rates of convergence when providers of data play "privacy game." Extensions to differential privacy as well.
2. In a soon to be on `arXiv` paper, we generalize this: no more privacy game, essentially all statistical estimators

Conclusions and future work

1. Have given sharp rates of convergence when providers of data play "privacy game." Extensions to differential privacy as well.
2. In a soon to be on `arXiv` paper, we generalize this: no more privacy game, essentially all statistical estimators
3. Is it possible to release a perturbed version of the data $X_1, \ldots, X_n$?

Conclusions and future work

1. Have given sharp rates of convergence when providers of data play "privacy game." Extensions to differential privacy as well.
2. In a soon to be on `arXiv` paper, we generalize this: no more privacy game, essentially all statistical estimators
3. Is it possible to release a perturbed version of the data $X_1, \ldots, X_n$?
4. What if all we care about is protecting some function $\varphi(X_i)$?

Conclusions and future work

1. Have given sharp rates of convergence when providers of data play "privacy game." Extensions to differential privacy as well.
2. In a soon to be on `arXiv` paper, we generalize this: no more privacy game, essentially all statistical estimators
3. Is it possible to release a perturbed version of the data $X_1, \ldots, X_n$?
4. What if all we care about is protecting some function $\varphi(X_i)$?

Thanks!

Conclusions and future work

1. Have given sharp rates of convergence when providers of data play "privacy game." Extensions to differential privacy as well.
2. In a soon to be on `arXiv` paper, we generalize this: no more privacy game, essentially all statistical estimators
3. Is it possible to release a perturbed version of the data $X_1, \ldots, X_n$?
4. What if all we care about is protecting some function $\varphi(X_i)$?

Exhibiting tradeoffs

Have mutual information

$$I^* := \min_Q \sup_P I(X; Z)$$

Theorem: There are constants a, b with $b/a = \mathcal{O}(1)$ dependent only on learning problem such that

$$\frac{\sqrt{d}}{\sqrt{nI^*}} a \leq \mathbb{E}[R(\widehat{\theta})] - R(\theta^\star) \leq \frac{\sqrt{d}}{\sqrt{nI^*}} b$$

Exhibiting tradeoffs

Have mutual information

$$I^* := \min_Q \sup_P I(X; Z)$$

Theorem: There are constants a, b with $b/a = \mathcal{O}(1)$ dependent only on learning problem such that

$$\frac{\sqrt{d}}{\sqrt{nI^*}} a \leq \mathbb{E}[R(\widehat{\theta})] - R(\theta^\star) \leq \frac{\sqrt{d}}{\sqrt{nI^*}} b$$

- Lower bound holds for all methods
- Upper bound achieved by stochastic approximation

Mutual information saddle points

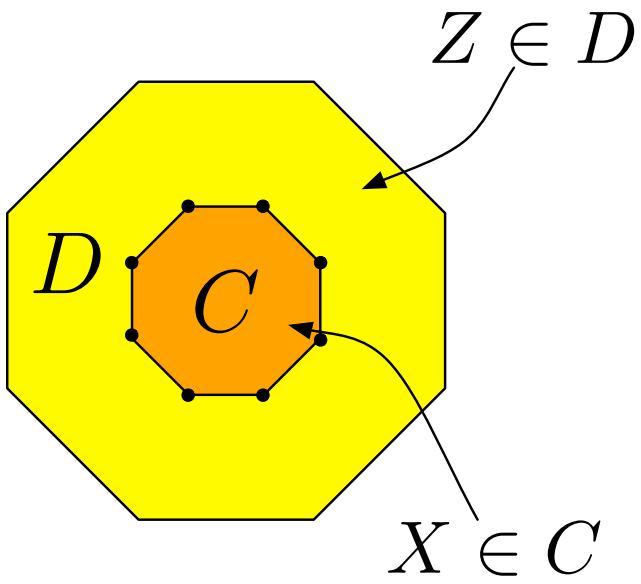
Goal: Channel Q^* (where $Z \sim Q^*(\cdot \mid X)$) so that

$$\min_{Q} \max_{P,\ell} I(P, Q) \geq \max_{P,\ell} I(P, Q^*)$$

Mutual information saddle points

Goal: Channel Q^* (where $Z \sim Q^*(\cdot \mid X)$) so that

$$\min_{Q} \max_{P,\ell} I(P, Q) \geq \max_{P,\ell} I(P, Q^*)$$

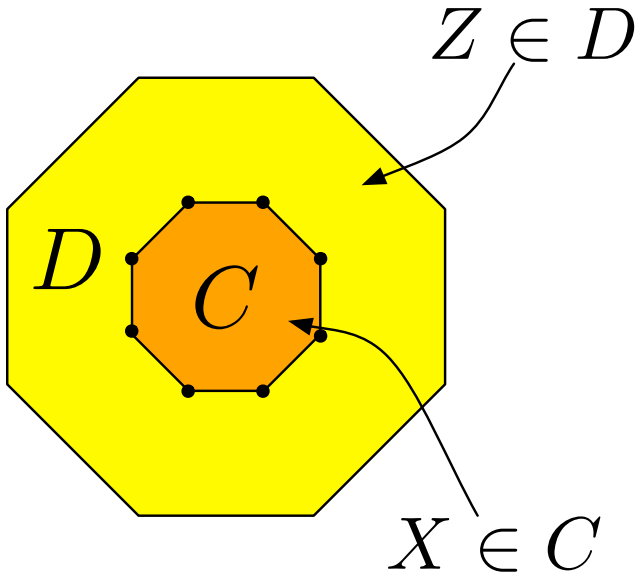


Mutual information saddle points

Goal: Channel Q^* (where $Z \sim Q^*(\cdot \mid X)$) so that

$$\min_{Q} \max_{P,\ell} I(P, Q) \geq \max_{P,\ell} I(P, Q^*)$$

Theorem: Let $\nabla \ell(\theta; X) \in C$, $Z \in D$.



Mutual information saddle points

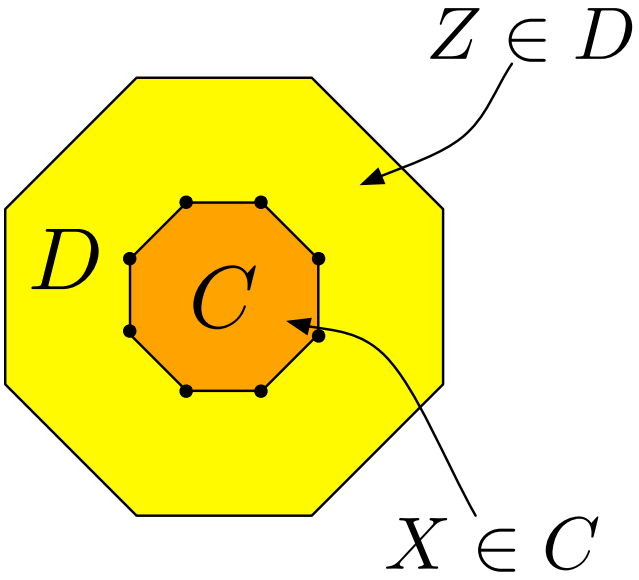
Goal: Channel Q^* (where $Z \sim Q^*(\cdot \mid X)$) so that

$$\min_{Q} \max_{P,\ell} I(P, Q) \geq \max_{P,\ell} I(P, Q^*)$$

Theorem: Let $\nabla \ell(\theta; X) \in C$, $Z \in D$. If

- P^* is uniform on extreme points of C
- Q^* supported on extreme points of D, maximizes entropy of Z given X

Also Q^* is unique



Mutual information saddle points

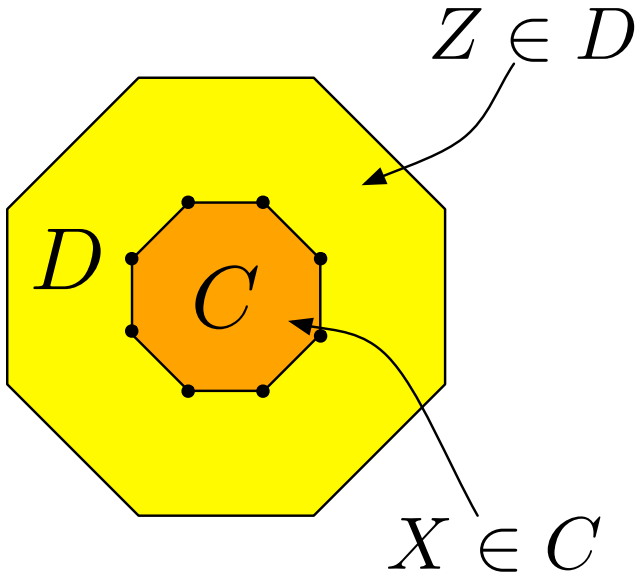
Goal: Channel Q^* (where $Z \sim Q^*(\cdot \mid X)$) so that

$$\min_Q \max_{P,\ell} I(P, Q) \geq \max_{P,\ell} I(P, Q^*)$$

Theorem: Let $\nabla \ell(\theta; X) \in C$, $Z \in D$. If

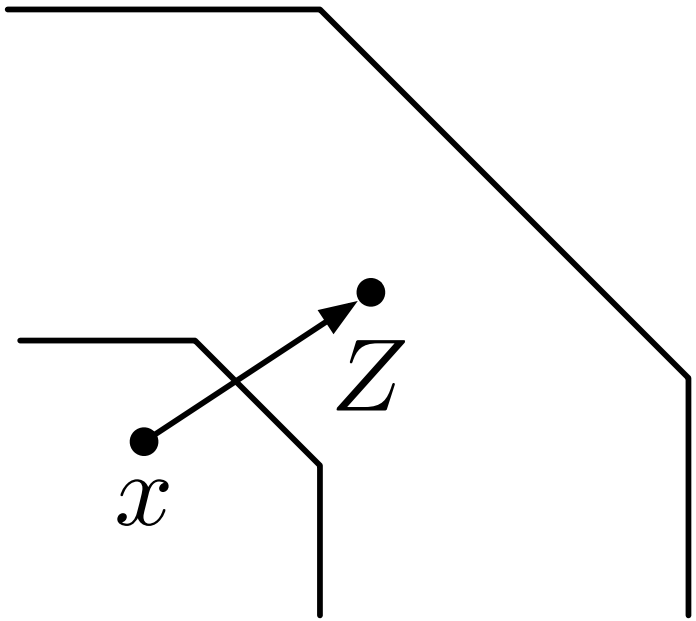
- P^* is uniform on extreme points of C
- Q^* supported on extreme points of D, maximizes entropy of Z given X

Also Q^* is unique



$$\min_Q \max_{P,\ell} I(X; Z) = \max_{P,\ell} \min_Q I(X; Z) = I(X^*; Z^*).$$

Privacy intuition



Privacy intuition