Information-theoretic methods in statistical machine learning

Martin Wainwright

UC Berkeley
Department of EECS, and Department of Statistics

Based on joint work with:

John Duchi, Stanford University
Michael Jordan, UC Berkeley
Mert Pilanci, UC Berkeley

Introduction

Era of massive data sets

Leads to new issues that are both statistical and computational in nature.

Introduction

Era of massive data sets

Leads to new issues that are both statistical and computational in nature.

1. Statistical issues
   - concentration of measure
   - curse of dimensionality
   - importance of "low-dimensional" structure

Introduction

Era of massive data sets

Leads to new issues that are both statistical and computational in nature.

1. Statistical issues
   - concentration of measure
   - curse of dimensionality
   - importance of “low-dimensional” structure
2. Algorithmic issues
   - Increasing importance of privacy
   - Memory and storage constraints
   - Computational constraints

Introduction

Era of massive data sets

Leads to new issues that are both statistical and computational in nature.

1. Statistical issues
   - concentration of measure
   - curse of dimensionality
   - importance of "low-dimensional" structure
2. Algorithmic issues
   - Increasing importance of privacy
   - Memory and storage constraints
   - Computational constraints

This lecture

Some vignettes in which information theory has an important role to play.

Issue A: Privacy versus statistical utility

Many sources of data have both statistical utility and privacy concerns.

(a) Personal genome project

Issue A: Privacy versus statistical utility

Many sources of data have both statistical utility and privacy concerns.

(a) Personal genome project

(b) Privacy breach
Scientific American, August 2013

Issue A: Privacy versus statistical utility

Many sources of data have both statistical utility and privacy concerns.

(a) Personal genome project

(b) Privacy breach
Scientific American, August 2013

Question

How to obtain principled tradeoffs between these competing criteria?

Issue B: Need for distributed estimators

Many modern datasets are too large to be stored on a single machine.

Google server farms

Netflix data base

Issue B: Need for distributed estimators

Many modern datasets are too large to be stored on a single machine.

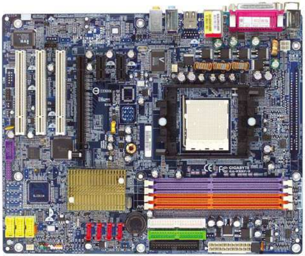
Google server farms

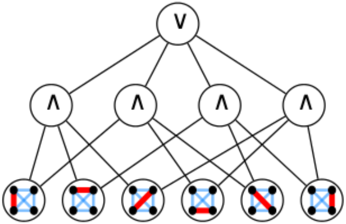
Netflix data base

Question

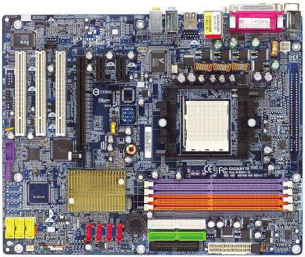
How to design statistical estimators that operate only on small pieces data?
Fundamental tradeoffs between centralized and distributed estimators?

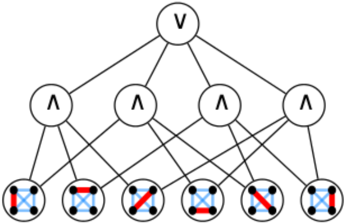
Issue C: Computational vs. statistical efficiency





Issue C: Computational vs. statistical efficiency

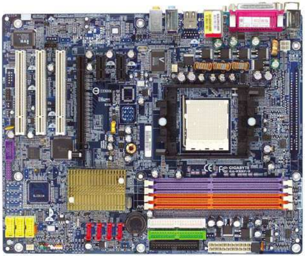


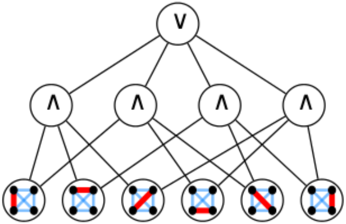


Question

What are the trade-offs between computational and statistical efficiency?

Issue C: Computational vs. statistical efficiency





More specific questions

- How to reduce computional complexity while retaining statistical optimality?
- When is there a gap between polynomial-time and exponential-time algorithms?
- Differences in hierarchies of polynomial computation?

§1. Statistics and privacy

Privacy concerns with many types of data:

- your personal genome
- your sexual behaviour
- a company's designs and algorithms
- your financial data

§1. Statistics and privacy

Privacy concerns with many types of data:

- your personal genome
- your sexual behaviour
- a company's designs and algorithms
- your financial data

Off-set by potential benefits from statistical aggregation:

- biological basis of disease
- epidemiological control
- reduced use of energy/materials
- improved economic forecasting

§1. Statistics and privacy

Privacy concerns with many types of data:

- your personal genome
- your sexual behaviour
- a company's designs and algorithms
- your financial data

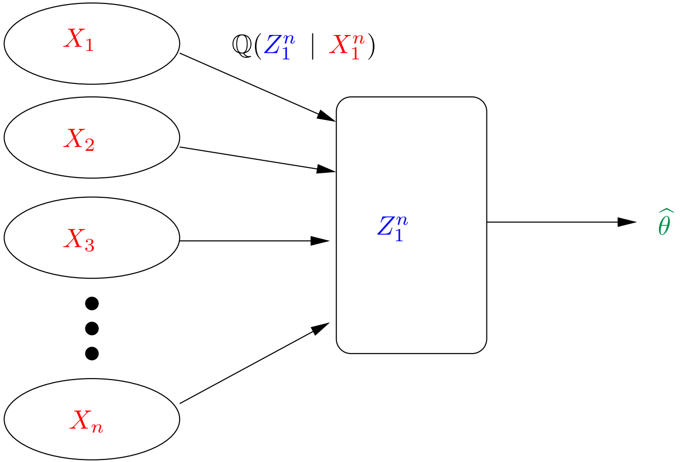
Off-set by potential benefits from statistical aggregation:

- biological basis of disease
- epidemiological control
- reduced use of energy/materials
- improved economic forecasting

Question

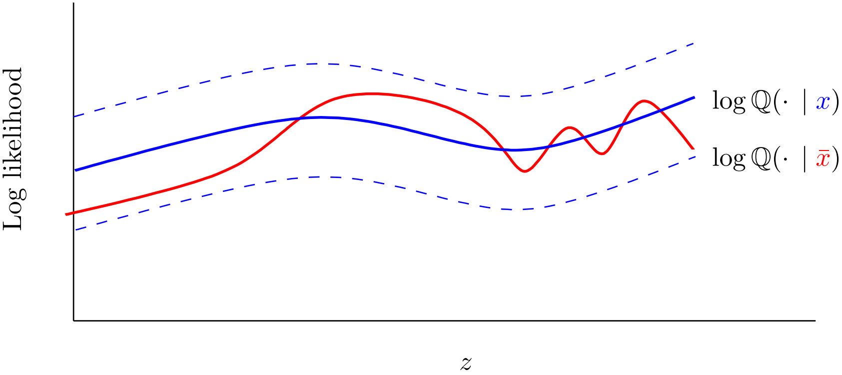
What are the fundamental trade-offs between privacy and statistical utility?

Basic model of local privacy



- each individual $i \in \{1, 2, \ldots, n\}$ has personal data $X_i \sim \mathbb{P}_{\theta^*}$
- conditional distribution $\mathbb{Q}$ between private data X_1^n and public data Z_1^n
- estimator $Z_1^n \mapsto \widehat{\theta}$ of unknown parameter θ^*.

Local privacy at level α



Definition

Conditional distribution $\mathbb{Q}$ is locally α-differentially private if

$$e^{-\alpha} \le \sup_z \frac{\mathbb{Q}(z \mid x_1^n)}{\mathbb{Q}(z \mid \bar{x}_1^n)} \le e^{\alpha} \qquad \text{for all } x_1^n \text{ and } \bar{x}_1^n \text{ such that } d_{\text{HAM}}(x_1^n, \bar{x}_1^n) = 1.$$

(Dwork et al., 2006)

Hypothesis testing interpretation

- consider two data sets x_1^n and $\bar{x}_1^n$ that differ in at least one co-ordinate
- given privatized observations Z_1^n, an adversary wants to test between:
  H_0: Underlying data set is $x_1^n = \{x_1, \ldots, x_n\}$
  H_1: Underlying data set is $\bar{x}_1^n = \{\bar{x}_1, \ldots, \bar{x}_n\}$

Hypothesis testing interpretation

- consider two data sets x_1^n and $\bar{x}_1^n$ that differ in at least one co-ordinate
- given privatized observations Z_1^n, an adversary wants to test between:

  H_0: Underlying data set is $x_1^n = \{x_1, \ldots, x_n\}$

  H_1: Underlying data set is $\bar{x}_1^n = \{\bar{x}_1, \ldots, \bar{x}_n\}$

α-privacy limits testing accuracy

For any test function $\psi : \mathcal{Z}_1^n \to \{0, 1\}$:

$$\frac{1}{2} \sum_{j=0}^{1} \mathbb{P}\big[\psi(Z_1^n) \neq j,\ H = H_j\big] \geq \frac{1}{1 + e^{\alpha}}.$$

Consequently, α-privacy provides a bound on the disclosure risk.

(Wasserman & Zhou, 2011)

Testing error versus privacy level

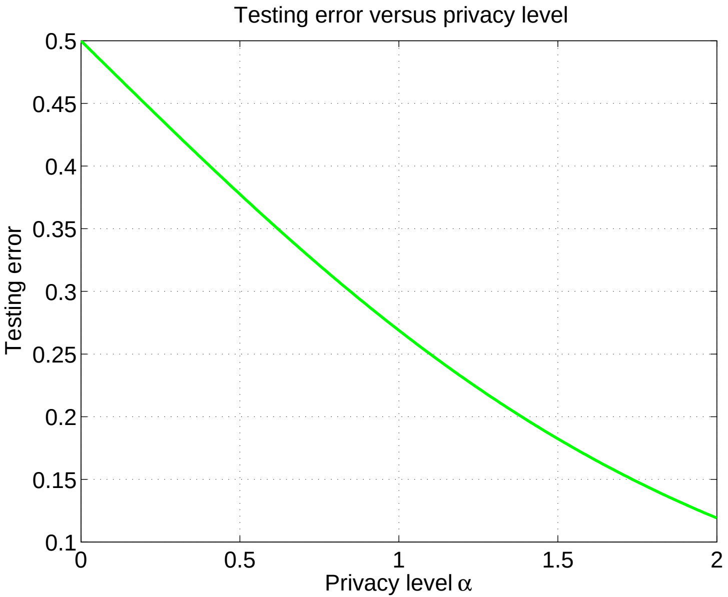
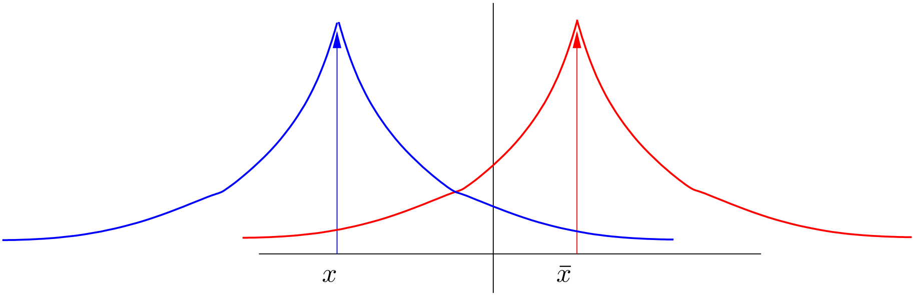


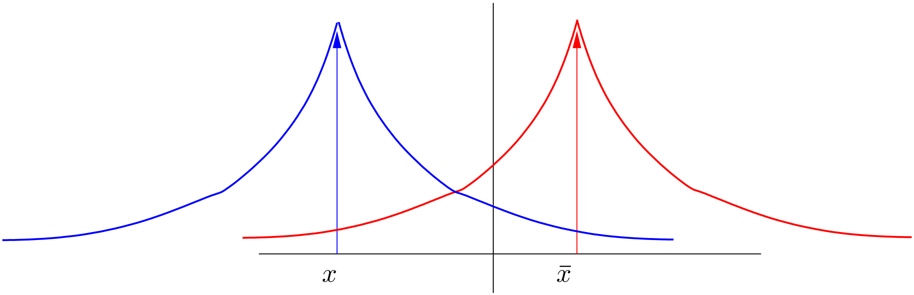
Illustration of Laplacian mechanism



Add α-Laplacian noise (Dwork et al., 2006)

$$Z = x + W, \qquad \text{where } W \text{ has density} \propto e^{-\alpha\,|w|}$$

Illustration of Laplacian mechanism



Add α-Laplacian noise (Dwork et al., 2006)

$$Z = x + W, \qquad \text{where } W \text{ has density } \propto e^{-\alpha\,|w|}$$

For all $x, x' \in [-1/2, 1/2]$:

$$\sup_{z \in \mathbb{R}} \left| \log \frac{\mathbb{Q}(z \mid x)}{\mathbb{Q}(z \mid \bar{x})} \right| \;=\; \alpha \left| \sup_{z \in \mathbb{R}} |z - x| - |z - \bar{x}| \right| \;\leq\; \alpha.$$

Various mechanisms for α-privacy

Choices from past work:

- randomized response in survey questions (Warner, 1965)
- Laplacian noise (Dwork et al., 2006)
- exponential mechanism (McSherry & Talwar, 2007)

Various mechanisms for α-privacy

Choices from past work:

- randomized response in survey questions (Warner, 1965)
- Laplacian noise (Dwork et al., 2006)
- exponential mechanism (McSherry & Talwar, 2007)

Some past work on privacy and estimation:

- local differential privacy and PAC learning (Kasiviswanathan et al., 2008)
- linear queries over discrete-valued data sets (Hardt & Rothblum, 2010)
- global differential privacy and histogram estimators (Hall et al., 2011)
- lower bounds for certain 1-D statistics (Chaudhuri & Hsu, 2012)

Various mechanisms for α-privacy

Choices from past work:

- randomized response in survey questions (Warner, 1965)
- Laplacian noise (Dwork et al., 2006)
- exponential mechanism (McSherry & Talwar, 2007)

Some past work on privacy and estimation:

- local differential privacy and PAC learning (Kasiviswanathan et al., 2008)
- linear queries over discrete-valued data sets (Hardt & Rothblum, 2010)
- global differential privacy and histogram estimators (Hall et al., 2011)
- lower bounds for certain 1-D statistics (Chaudhuri & Hsu, 2012)

Question

Can we provide a general characterization of the trade-offs between α-privacy and statistical utility?

Will measure statistical utility in terms of minimax risk

Minimax optimality with α-privacy

- family of distributions $\{\mathbb{P} \in \mathcal{F}\}$, and functional $\mathbb{P} \mapsto \theta(\mathbb{P})$
- samples $X_1^n \equiv \{X_1, \ldots, X_n\} \sim \mathbb{P}$ and estimator $X_1^n \mapsto \widehat{\theta}(X_1^n)$
- loss function (e.g., squared error, 0-1 error, ℓ_1-error)

$$(\widehat{\theta}, \theta) \qquad \mapsto \qquad \underbrace{\mathcal{L}(\widehat{\theta}, \theta)}_{\text{quality of } \widehat{\theta} \text{ as estimate of } \theta}$$

Minimax optimality with α-privacy

- family of distributions $\{\mathbb{P} \in \mathcal{F}\}$, and functional $\mathbb{P} \mapsto \theta(\mathbb{P})$
- samples $X_1^n \equiv \{X_1, \ldots, X_n\} \sim \mathbb{P}$ and estimator $X_1^n \mapsto \widehat{\theta}(X_1^n)$
- loss function (e.g., squared error, 0-1 error, ℓ_1-error)

$$(\widehat{\theta}, \theta) \qquad \mapsto \qquad \underbrace{\mathcal{L}(\widehat{\theta}, \theta)}_{\text{quality of } \widehat{\theta} \text{ as estimate of } \theta}$$

Ordinary minimax risk:

$$\mathfrak{M}_n(\mathcal{F}) := \underbrace{\inf_{\widehat{\theta}}}_{\text{Best estimator}} \underbrace{\sup_{\mathbb{P} \in \mathcal{F}}}_{\text{Worst-case distribution}} \mathbb{E}\Big[\mathcal{L}\big(\widehat{\theta}(X_1^n), \theta(\mathbb{P})\big)\Big]$$

Minimax optimality with α-privacy

- family of distributions $\{\mathbb{P} \in \mathcal{F}\}$, and functional $\mathbb{P} \mapsto \theta(\mathbb{P})$
- samples $X_1^n \equiv \{X_1, \ldots, X_n\} \sim \mathbb{P}$ and estimator $X_1^n \mapsto \widehat{\theta}(X_1^n)$
- loss function (e.g., squared error, 0-1 error, ℓ_1-error)

$$(\widehat{\theta}, \theta) \quad \mapsto \quad \underbrace{\mathcal{L}(\widehat{\theta}, \theta)}_{\text{quality of } \widehat{\theta} \text{ as estimate of } \theta}$$

Ordinary minimax risk:

$$\mathfrak{M}_n(\mathcal{F}) := \underbrace{\inf_{\widehat{\theta}}}_{\text{Best estimator}} \underbrace{\sup_{\mathbb{P} \in \mathcal{F}}}_{\text{Worst-case distribution}} \mathbb{E}\Big[\mathcal{L}\big(\widehat{\theta}(X_1^n), \theta(\mathbb{P})\big)\Big]$$

Minimax risk with α-privacy

Estimators now depend on privatized samples Z_1^n

$$\mathfrak{M}_n(\alpha; \mathcal{F}) := \underbrace{\inf_{\mathbb{Q} \in \mathcal{Q}_\alpha}}_{\text{Best } \alpha\text{-private channel}} \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}\Big[\mathcal{L}\big(\widehat{\theta}(Z_1^n), \ \theta(\mathbb{P})\big)\Big]$$

Illustration: Non-parametric density estimation

Suppose that we want to estimate the quantity $\mathbb{P} \mapsto \theta(\mathbb{P}) \equiv$ density f

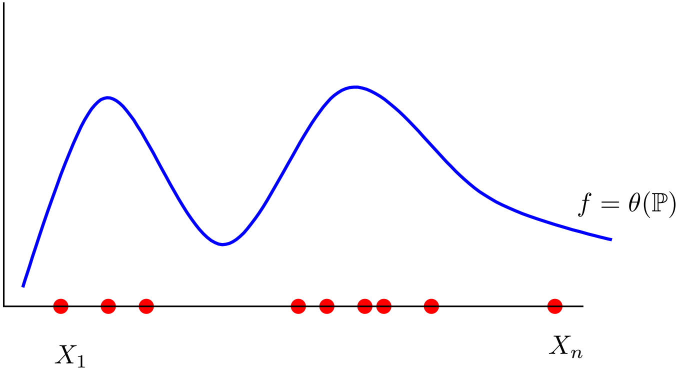
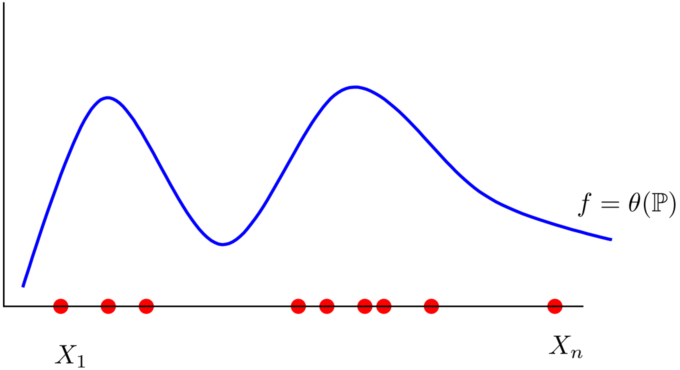


Illustration: Non-parametric density estimation

Suppose that we want to estimate the quantity $\mathbb{P} \mapsto \theta(\mathbb{P}) \equiv$ density f



Classical fact

Ordinary minimax rates depend on number of derivatives $\beta > 1/2$ of density f:

$$\mathfrak{M}_n\big(\mathcal{F}(\beta)\big) \asymp \left(\frac{1}{n}\right)^{\frac{2\beta}{2\beta+1}}$$

(Ibragimov & Hasminskii, 1978; Stone, 1980)

Optimal rates for α-private density estimation

Consider density estimation based on α-private views $(Z_1, \dots, Z_n)$ of original samples $(X_1, \dots, X_n)$.

Optimal rates for α-private density estimation

Consider density estimation based on α-private views $(Z_1, \ldots, Z_n)$ of original samples $(X_1, \ldots, X_n)$.

Theorem (Duchi, W. & Jordan, 2013)

For all privacy levels $\alpha \in (0, 1/4]$ and smoothness levels $\beta > 1/2$:

$$\mathfrak{M}_n\big(\alpha; \mathcal{F}(\beta)\big) \asymp \left(\frac{1}{\alpha^2 n}\right)^{\frac{2\beta}{2\beta+2}}$$

Optimal rates for α-private density estimation

Consider density estimation based on α-private views $(Z_1, \ldots, Z_n)$ of original samples $(X_1, \ldots, X_n)$.

Theorem (Duchi, W. & Jordan, 2013)

For all privacy levels $\alpha \in (0, 1/4]$ and smoothness levels $\beta > 1/2$:

$$\mathfrak{M}_n\big(\alpha; \mathcal{F}(\beta)\big) \asymp \Big(\frac{1}{\alpha^2 n}\Big)^{\frac{2\beta}{2\beta+2}}$$

- can give a simple/explicit scheme that achieves this optimal rate.

Optimal rates for α-private density estimation

Consider density estimation based on α-private views $(Z_1, \ldots, Z_n)$ of original samples $(X_1, \ldots, X_n)$.

Theorem (Duchi, W. & Jordan, 2013)

For all privacy levels $\alpha \in (0, 1/4]$ and smoothness levels $\beta > 1/2$:

$$\mathfrak{M}_n\big(\alpha; \mathcal{F}(\beta)\big) \asymp \Big(\frac{1}{\alpha^2 n}\Big)^{\frac{2\beta}{2\beta+2}}$$

- can give a simple/explicit scheme that achieves this optimal rate.
- contrast with classical rate $(1/n)^{\frac{2\beta}{2\beta+1}}$: Penalty for privacy can be significant!

Optimal rates for α-private density estimation

Consider density estimation based on α-private views $(Z_1, \ldots, Z_n)$ of original samples $(X_1, \ldots, X_n)$.

Theorem (Duchi, W. & Jordan, 2013)

For all privacy levels $\alpha \in (0, 1/4]$ and smoothness levels $\beta > 1/2$:

$$\mathfrak{M}_n\big(\alpha; \mathcal{F}(\beta)\big) \asymp \left(\frac{1}{\alpha^2 n}\right)^{\frac{2\beta}{2\beta+2}}$$

- can give a simple/explicit scheme that achieves this optimal rate.
- contrast with classical rate $(1/n)^{\frac{2\beta}{2\beta+1}}$: Penalty for privacy can be significant!

Example: How many samples $N(\epsilon)$ to achieve error $\epsilon = 0.01$ for Lipschitz densities $(\beta = 1)$?

Classical case $N \approx 1,000$ versus Private case $N \approx 10,000$.

How to measure “size” of function classes?

How to measure "size" of function classes?

- A 2δ-packing of $\mathcal{F}$ is a collection $\{f^1, \ldots, f^M\} \subset \mathcal{F}$ such that

$$\|f^j - f^k\|_2 > 2\delta \qquad \text{for all } j \neq k.$$

- The packing number $M(2\delta)$ is the cardinality of the largest such set.

How to measure "size" of function classes?

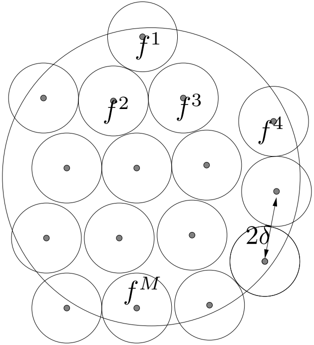
- A 2δ-packing of $\mathcal{F}$ is a collection $\{f^1, \ldots, f^M\} \subset \mathcal{F}$ such that

$$\|f^j - f^k\|_2 > 2\delta \qquad \text{for all } j \neq k.$$

- The packing number $M(2\delta)$ is the cardinality of the largest such set.

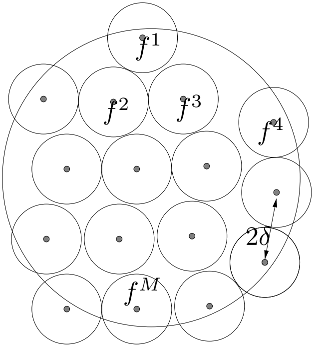
- Packing/covering entropy: introduced in seminal work by Kolmogorov
- Deterministic concept but connected to Shannon entropy via volume ratios
- Central object in proving minimax lower bounds (e.g., Hasminskii & Ibragimov, 1978; Birge, 1983; Yu, 1997; Yang & Barron, 1999)

From metric entropy to Fano's inequality



- construct a 2δ-packing of densities $\{f^1, \ldots, f^M\}$ with $\log M(\delta) \asymp (1/\delta)^{1/\beta}$ elements
- draw packing index $V \in \{1, \ldots M\}$ uniformly at random
- conditioned on $V = j$, draw n i.i.d. samples $\{X_1, \ldots, X_n\} \sim f^j$

From metric entropy to Fano's inequality



- construct a 2δ-packing of densities $\{f^1, \ldots, f^M\}$ with $\log M(\delta) \asymp (1/\delta)^{1/\beta}$ elements
- draw packing index $V \in \{1, \ldots M\}$ uniformly at random
- conditioned on $V = j$, draw n i.i.d. samples $\{X_1, \ldots, X_n\} \sim f^j$

Mutual information

Relative discriminability controlled by mutual information

$$I(X_1, \ldots, X_n; V) = \frac{1}{M} \sum_{j=1}^{M} D(\bar{\mathbb{P}} \parallel \mathbb{P}^j) \qquad \text{where} \qquad \underbrace{\bar{\mathbb{P}} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}^j}_{\text{Mixture distribution}}$$

A quantitative data processing inequality

$$\boxed{V} \longrightarrow \boxed{X} \xrightarrow{\mathbb{Q}} \boxed{Z}$$

- packing index $V \in \{1, 2, \ldots, M\}$
- non-private variables $(X \mid V = j) \sim \mathbb{P}_j$
- mixture distribution $\bar{\mathbb{P}} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_j$.

A quantitative data processing inequality

$$V \longrightarrow X \xrightarrow{\mathbb{Q}} Z$$

- packing index $V \in \{1, 2, \ldots, M\}$
- non-private variables $(X \mid V = j) \sim \mathbb{P}_j$
- mixture distribution $\bar{\mathbb{P}} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_j$.

Theorem (Duchi, W. & Jordan, 2013)

For any non-interactive α-private channel $\mathbb{Q}$, we have

$$\frac{I(Z_1, \ldots, Z_n; V)}{n} \leq (e^{\alpha} - 1)^2 \underbrace{\sup_{\|\gamma\|_\infty \leq 1} \left\{ \frac{1}{M} \sum_{j=1}^{M} \Big(\int_{\mathcal{X}} \gamma(x) \big(d\mathbb{P}_j(x) - d\bar{\mathbb{P}}(x) \big) \Big)^2 \right\}}_{\textit{dimension-dependent contraction}}$$

§2: Fast algorithms via randomized approximations

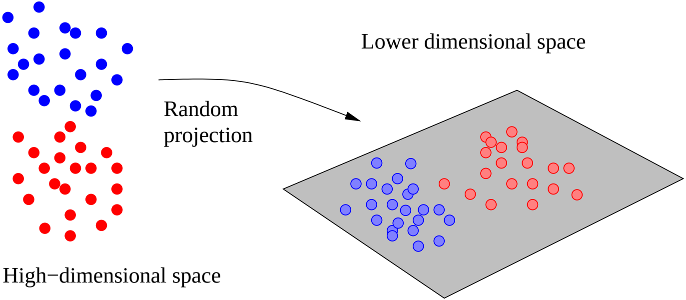
Massive data sets require very fast algorithms but with rigorous guarantees.

§2: Fast algorithms via randomized approximations

Massive data sets require very fast algorithms but with rigorous guarantees.

A general purpose tool:

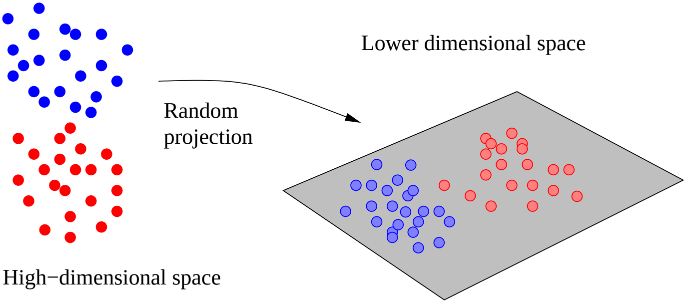
- Choose a random subspace of "low" dimension m.
- Project data into subspace, and solve reduced dimension problem.



§2: Fast algorithms via randomized approximations

A general purpose tool:

- Choose a random subspace of "low" dimension m.
- Project data into subspace, and solve reduced dimension problem.



Widely used in practice:

- Johnson & Lindenstrauss (1984): for Hilbert spaces
- various algorithms in theoretical computer science: (e.g., Vempala, 2004)

Randomized sketching of constrained least-squares

Original program based on data vector $y \in \mathbb{R}^n$ and data matrix $A \in \mathbb{R}^{n \times d}$:

$$x^{\text{LS}} = \arg\min_{x \in \mathcal{C}} \underbrace{\|Ax - y\|_2^2}_{f(x)}$$

where $\mathcal{C}$ is a compact, convex set in $\mathbb{R}^d$.

Randomized sketching of constrained least-squares

Original program based on data vector $y \in \mathbb{R}^n$ and data matrix $A \in \mathbb{R}^{n \times d}$:

$$x^{\text{LS}} = \arg\min_{x \in \mathcal{C}} \underbrace{\|Ax - y\|_2^2}_{f(x)}$$

where $\mathcal{C}$ is a compact, convex set in $\mathbb{R}^d$.

Given a sketching matrix $S \in \mathbb{R}^{m \times n}$, consider sketched version

$$\widehat{x} = \arg\min_{x \in \mathcal{C}} \|SAx - Sy\|_2^2$$

Randomized sketching of constrained least-squares

Original program based on data vector $y \in \mathbb{R}^n$ and data matrix $A \in \mathbb{R}^{n \times d}$:

$$x^{\text{LS}} = \arg\min_{x \in \mathcal{C}} \underbrace{\|Ax - y\|_2^2}_{f(x)}$$

where $\mathcal{C}$ is a compact, convex set in $\mathbb{R}^d$.

Given a sketching matrix $S \in \mathbb{R}^{m \times n}$, consider sketched version

$$\widehat{x} = \arg\min_{x \in \mathcal{C}} \|SAx - Sy\|_2^2$$

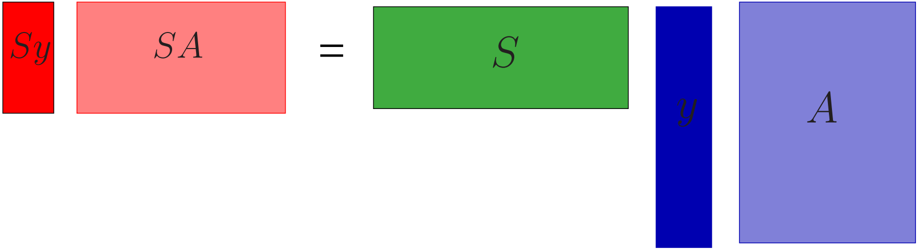
Question:

How many projections m for a required for a δ-approximation?

Cost approx.: $\underbrace{f(x^{\text{LS}})}_{\text{Optimal value}} \leq \underbrace{f(\widehat{x})}_{\text{Sketch value}} \leq \underbrace{(1+\delta)^2}_{\text{Approx. factor}} f(x^{\text{LS}})$.

Solution approx.: $\underbrace{\|\widehat{x} - x^{\text{LS}}\|_A^2}_{\text{Sketch error}} \leq \underbrace{\delta^2}_{\text{Approx. factor}} \|x^{\text{LS}}\|_A^2$.

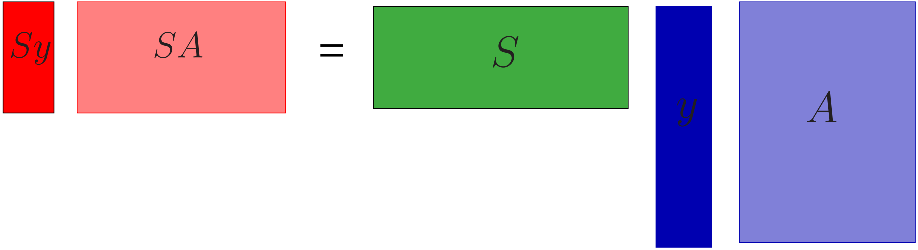
The classical sketch



Original problem based on data $(y, A) \in \mathbb{R}^n \times \mathbb{R}^{n \times d}$:

$$x^* = \arg\min_{x \in \mathbb{R}^d} \|Ax - y\|_2^2$$

The classical sketch



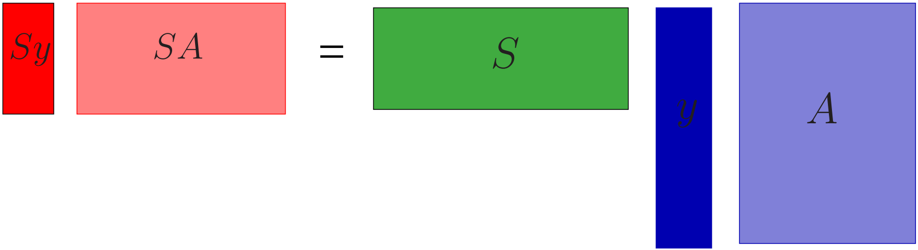
Original problem based on data $(y, A) \in \mathbb{R}^n \times \mathbb{R}^{n \times d}$:

$$x^* = \arg\min_{x \in \mathbb{R}^d} \|Ax - y\|_2^2$$

Given sketched data $(Sy, SA) \in \mathbb{R}^m \times \mathbb{R}^{m \times d}$, compute sketched solution:

$$\widehat{x} = \arg\min_{x \in \mathcal{C}} \|SAx - Sy\|_2^2.$$

The classical sketch



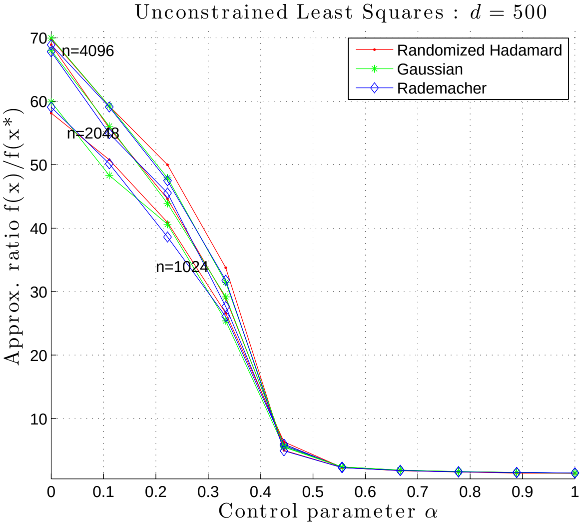
Given sketched data $(Sy, SA) \in \mathbb{R}^m \times \mathbb{R}^{m \times d}$, compute sketched solution:

$$\widehat{x} = \arg\min_{x \in \mathcal{C}} \|SAx - Sy\|_2^2.$$

Some past work:

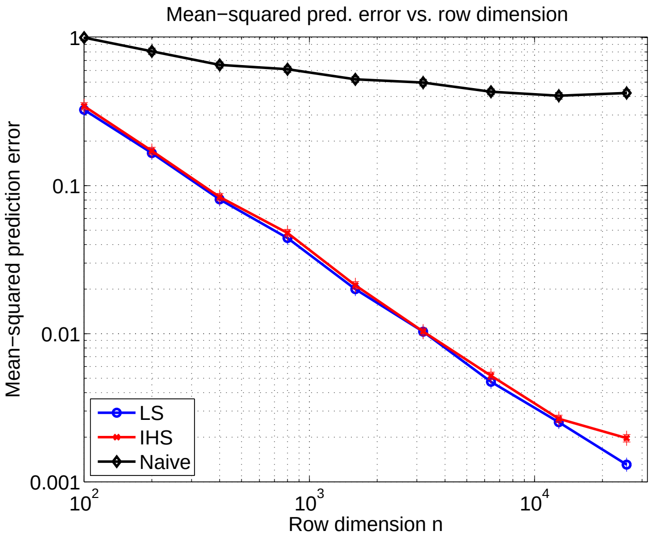
- δ-cost approximation using $m \gtrsim \frac{1}{\delta^2}\,\mathrm{rank}(A) \log(d)$ samples (Sarlos, 2006; Mahoney, 2011)
- various extensions and variants (Boutsidis & Drineas, 2009; Drineas et al., 2011)
- sharp results for general convex constraint sets (Pilanci & W., 2014)

Cost approximation: Unconstrained LS



Sketch size $m \gtrsim \alpha\, d$

Solution approximation: Unconstrained LS



Sketch size $m \gtrsim d \log(n)$

An information-theoretic lower bound

Consider random ensemble of least-squares problems

$$x^{\text{LS}} = \arg\min_{x \in \mathcal{C}} \Big\{ \|y - Ax\|_2^2 \Big\} \qquad \text{where } y = Ax^* + w$$

with $w \sim N(0, \sigma^2 I_n)$.

An information-theoretic lower bound

Consider random ensemble of least-squares problems

$$x^{\text{LS}} = \arg\min_{x \in \mathcal{C}} \left\{ \|y - Ax\|_2^2 \right\} \qquad \text{where } y = Ax^* + w$$

with $w \sim N(0, \sigma^2 I_n)$.

Theorem (Pilanci & W, 2014)

For a broad class of random sketching matrices $S \in \mathbb{R}^{m \times n}$, any estimator $\widetilde{x}$ based on the pair (SA, Sy) has MSE lower bounded as

$$\sup_{x^* \in \mathcal{C} \cap \mathbb{B}_2(1)} \mathbb{E}_{S,w} \left[\|\widetilde{x} - x^*\|_A^2 \right] \gtrsim \frac{\sigma^2 \, \log M}{\min\{m, n\}},$$

where M is the $1/2$-packing number of $\mathcal{C} \cap \mathbb{B}_2(1)$.

An information-theoretic lower bound

Consider random ensemble of least-squares problems

$$x^{\text{LS}} = \arg\min_{x \in \mathcal{C}} \left\{ \|y - Ax\|_2^2 \right\} \qquad \text{where } y = Ax^* + w$$

with $w \sim N(0, \sigma^2 I_n)$.

Theorem (Pilanci & W, 2014)

For a broad class of random sketching matrices $S \in \mathbb{R}^{m \times n}$, any estimator $\widetilde{x}$ based on the pair (SA, Sy) has MSE lower bounded as

$$\sup_{x^* \in \mathcal{C} \cap \mathbb{B}_2(1)} \mathbb{E}_{S,w}\big[\|\widetilde{x} - x^*\|_A^2\big] \gtrsim \frac{\sigma^2 \, \log M}{\min\{m, n\}},$$

where M is the $1/2$-packing number of $\mathcal{C} \cap \mathbb{B}_2(1)$.

Special case: For unconstrained least-squares, this theorem implies that

$$\sup_{x^* \in \mathbb{B}_2(1)} \mathbb{E}_{S,w}\big[\|\widetilde{x} - x^*\|_A^2\big] \gtrsim \frac{\sigma^2 \, d}{\min\{m, n\}}.$$

A different approach: Hessian sketch

Initial idea: Severe loss of information is caused by sketching data vector y. So let's sketch only data matrix A, and solve program

$$\widetilde{x} = \arg\min_{x \in \mathcal{C}} \Big\{ \frac{1}{2m} \|SAx\|_2^2 - \langle A^T y, \, x \rangle \Big\}.$$

A different approach: Hessian sketch

Initial idea: Severe loss of information is caused by sketching data vector y. So let's sketch only data matrix A, and solve program

$$\widetilde{x} = \arg\min_{x \in \mathcal{C}} \Big\{ \frac{1}{2m} \|SAx\|_2^2 - \langle A^T y, \, x \rangle \Big\}.$$

Resulting bound suffers from same problem as classical sketch:

$$\|\widetilde{x} - \widehat{x}\|_A^2 \precsim \delta^2 \; \|x^{\text{LS}}\|_A^2.$$

A different approach: Hessian sketch

Initial idea: Severe loss of information is caused by sketching data vector y. So let's sketch only data matrix A, and solve program

$$\widetilde{x} = \arg\min_{x \in \mathcal{C}} \Big\{ \frac{1}{2m} \|SAx\|_2^2 - \langle A^T y, \, x \rangle \Big\}.$$

Resulting bound suffers from same problem as classical sketch:

$$\|\widetilde{x} - \widehat{x}\|_A^2 \precsim \delta^2 \, \|x^{\text{LS}}\|_A^2.$$

....but this procedure can be iterated!

A different approach: Hessian sketch

Initial idea: Severe loss of information is caused by sketching data vector y. So let's sketch only data matrix A, and solve program

$$\widetilde{x} = \arg\min_{x \in \mathcal{C}} \Big\{ \frac{1}{2m} \|SAx\|_2^2 - \langle A^T y,\, x \rangle \Big\}.$$

Resulting bound suffers from same problem as classical sketch:

$$\|\widetilde{x} - \widehat{x}\|_A^2 \precsim \delta^2 \; \|x^{\text{LS}}\|_A^2.$$

....but this procedure can be iterated!

Iterative Hessian sketch

Given an iteration number $N \geq 1$:

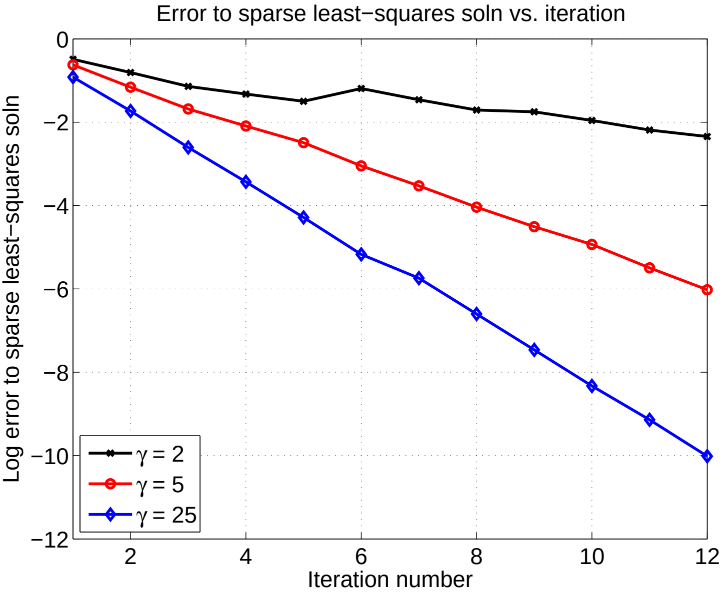
(1) Initialize at $x^0 = 0$.

(2) For iterations $t = 0, 1, 2, \ldots, N-1$, generate an independent sketch matrix $S^{t+1} \in \mathbb{R}^{m \times n}$, and perform the update

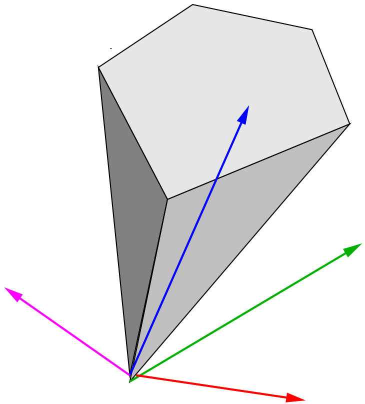
$$x^{t+1} = \arg\min_{x \in \mathcal{C}} \Big\{ \frac{1}{2m} \|S^{t+1} A(x - x^t)\|_2^2 - \langle A^T(y - Ax^t),\, x \rangle \Big\}.$$

(3) Return the estimate $\widehat{x} = x^N$.

Empirically: geometric convergence is observed



Gaussian width of transformed tangent cone



Gaussian width of cone $A\mathcal{K} = \{A\Delta \in \mathbb{R}^n \mid \Delta \in \mathcal{K}\}$

$$\mathcal{W}(A\mathcal{K}) := \mathbb{E}\Big[\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \langle w,\, z\rangle\Big] \quad \text{where} \quad \begin{array}{c} w \sim N(0, I_{n\times n}) \\ \mathcal{S}^{n-1} = \{z \in \mathbb{R}^n \mid \|z\|_2 = 1\} \end{array}$$

Main result for sub-Gaussian sketches

Tangent cone at x^*:

$$\mathcal{K} = \{\Delta \in \mathbb{R}^d \mid \Delta = t(x - x^*) \in \mathcal{C} \quad \text{for some } t \geq 0.$$

Width of transformed cone $A\mathcal{K} \cap \mathcal{S}^{n-1}$:

$$\mathcal{W}(A\mathcal{K}) = \mathbb{E}\Big[\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \langle w,\, z\rangle\Big].$$

Theorem (Pilanci & W., 2014)

For any $\delta \in (0,1)$, performing $\log(2/\delta)$ steps of iterative Hessian sketch using a sub-Gaussian sketch dimension lower bounded as

$$m \gtrsim \mathcal{W}^2(A\mathcal{K})$$

suffices to ensure that the sketched solution is δ-optimal with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.

Sketching using randomized orthonormal systems

Step 1: Choose some fixed orthonormal matrix $H \in \mathbb{R}^{n \times n}$.
Example: Hadamard matrices

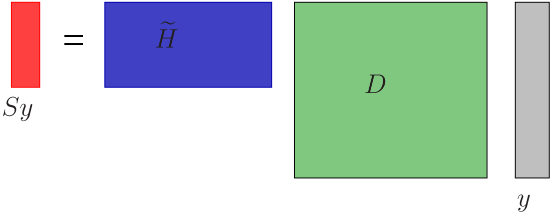
$$H_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad H_{2^t} = \underbrace{H_2 \otimes H_2 \otimes \cdots \otimes H_2}_{\text{Kronecker product } t \text{ times}}$$

(E.g., Ailon & Liberty, 2010)

Sketching using randomized orthonormal systems

Step 1: Choose some fixed orthonormal matrix $H \in \mathbb{R}^{n \times n}$.
Example: Hadamard matrices

$$H_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad H_{2^t} = \underbrace{H_2 \otimes H_2 \otimes \cdots \otimes H_2}_{\text{Kronecker product } t \text{ times}}$$



Step 2:

(A) Multiply data vector y with a diagonal matrix of random signs $\{-1, +1\}$

(B) Choose m rows of H to form sub-sampled matrix $\widetilde{H} \in \mathbb{R}^{m \times n}$

(C) Requires $\mathcal{O}(n \log m)$ time to compute sketched vector $Sy = \widetilde{S}\, Dy$.

(E.g., Ailon & Liberty, 2010)

Summary

Many challenges with massive data setes

- data collection and privacy concerns
- randomized algorithms for fast optimization
- distributed algorithms for statistical inference

Summary

Many challenges with massive data setes

- data collection and privacy concerns
- randomized algorithms for fast optimization
- distributed algorithms for statistical inference

Some papers/pre-prints:

- Duchi, W. & Jordan (2013). Local privacy and statistical minimax rates. http://arxiv.org/abs/1302.3203, Presented in part at FOCS 2013.
- W. (2014). Constrained forms of statistical minimax: Communication, computation and privacy. *Proceedings of the International Congress of Mathematicians.*
- Pilanci & W (2014). Randomized sketches of convex programs with sharp guarantees. *Arxiv preprint, April*, Presented in part at ISIT 2014.
- Pilanci & W (2014). Iterative Hessian sketch: Fast and accurate solution approximation for constrained least squares. *Arxiv preprint, November*.