On Computational Thinking, Inferential Thinking and Data Science

Michael I. Jordan
University of California, Berkeley

August 7, 2018

Some Perspective

- What's new?
  - not just large data sets, but data sets containing abundant data on each individual
  - where "individual" can be a person, a gene, a region of the sky, a habitat, etc
  - and where there are long tails

Some Perspective

- What’s new?
  - – not just large data sets, but data sets containing abundant data on each individual
  - – where “individual” can be a person, a gene, a region of the sky, a habitat, etc
  - – and where there are long tails
- What’s challenging?
  - – timely, trustable, and transparent inference and decision-making at the level of individuals

Some Perspective

- What's new?
  - not just large data sets, but data sets containing abundant data on each individual
  - where "individual" can be a person, a gene, a region of the sky, a habitat, etc
  - and where there are long tails
- What's challenging?
  - timely, trustable, and transparent inference and decision-making at the level of individuals
- What's required?
  - not just a library of algorithms, but a blend of computational thinking and inferential thinking

A Job Description, circa 2015

- Your Boss: "*I need a Big Data system that will replace our classic service with a personalized service*"

A Job Description, circa 2015

- Your Boss: "*I need a Big Data system that will replace our classic service with a personalized service*"
- "*It should work reasonably well for anyone and everyone; I can tolerate a few errors but not too many dumb ones that will embarrass us*"

A Job Description, circa 2015

- Your Boss: "*I need a Big Data system that will replace our classic service with a personalized service*"
- "*It should work reasonably well for anyone and everyone; I can tolerate a few errors but not too many dumb ones that will embarrass us*"
- "*It should run just as fast as our classic service*"

A Job Description, circa 2015

- Your Boss: "*I need a Big Data system that will replace our classic service with a personalized service*"
- "*It should work reasonably well for anyone and everyone; I can tolerate a few errors but not too many dumb ones that will embarrass us*"
- "*It should run just as fast as our classic service*"
- "*It should only improve as we collect more data; in particular it shouldn't slow down*"

A Job Description, circa 2015

- Your Boss: "*I need a Big Data system that will replace our classic service with a personalized service*"
- "*It should work reasonably well for anyone and everyone; I can tolerate a few errors but not too many dumb ones that will embarrass us*"
- "*It should run just as fast as our classic service*"
- "*It should only improve as we collect more data; in particular it shouldn't slow down*"
- "*There are serious privacy concerns of course, and they vary across the clients*"

What's Required

- Data Science requires a thorough blending of computational thinking and inferential thinking

What's Required

- Data Science requires a thorough blending of computational thinking and inferential thinking
- Computational thinking means (inter alia)
  - abstraction, modularity, scalability, robustness, etc.

What’s Required

- Data Science requires a thorough blending of computational thinking and inferential thinking
- Computational thinking means (inter alia)
  - abstraction, modularity, scalability, robustness, etc.
- Inferential thinking means (inter alia)
  - considering the real-world phenomenon behind the data
  - considering the sampling pattern that gave rise to the data
  - developing procedures that will go “backwards” from the data to the underlying phenomenon

The Challenges are Daunting

- The core theories in computer science and statistics developed separately and there is an oil and water problem
- Core statistical theory doesn't have a place for runtime and other computational resources
- Core computational theory doesn't have a place for statistical risk

Outline

- Inference under privacy constraints
- Inference under communication constraints
- Lower bounds, the variational perspective and symplectic integration

Part I: Inference and Privacy

with John Duchi and Martin Wainwright

Privacy and Data Analysis

- 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
- “Privacy loss” can be quantified (say) via differential privacy
- We want to trade privacy loss against the value we obtain from “data analysis”
- The question becomes that of quantifying such value and juxtaposing it with privacy loss

Privacy

query

$\downarrow$

database

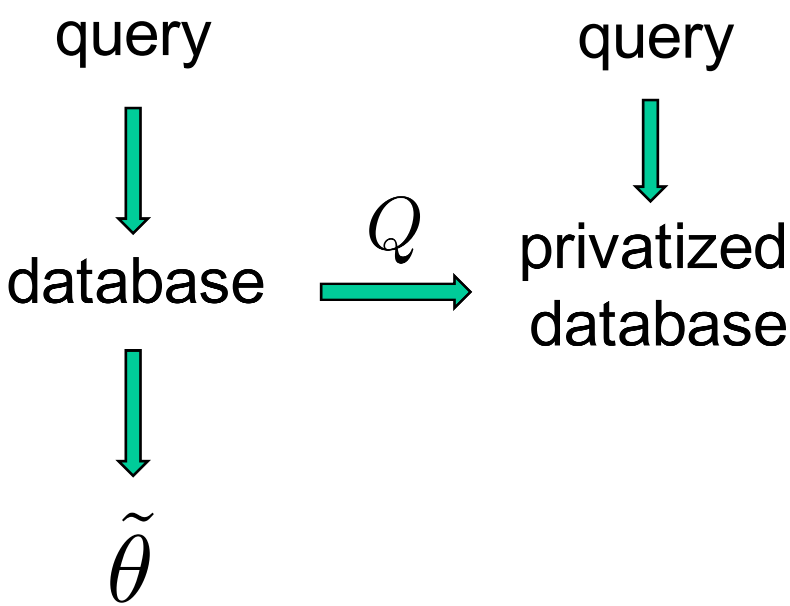
Privacy

$$
\begin{array}{c}
\text{query} \\
\downarrow \\
\text{database} \\
\downarrow \\
\tilde{\theta}
\end{array}
$$

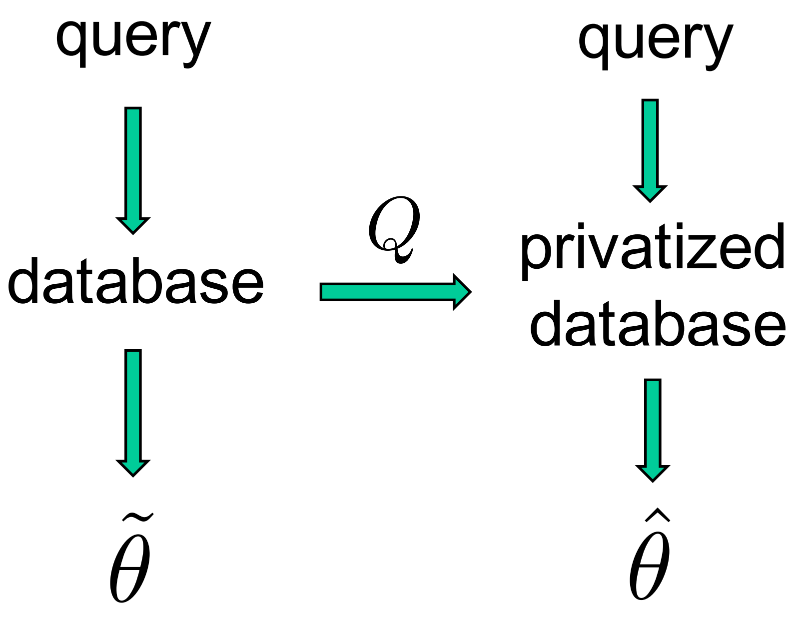
Privacy

$$
\begin{tikzcd}
\text{query} \arrow[d] & \\
\text{database} \arrow[r, "Q"] \arrow[d] & \text{privatized database} \\
\tilde{\theta} &
\end{tikzcd}
$$

Privacy

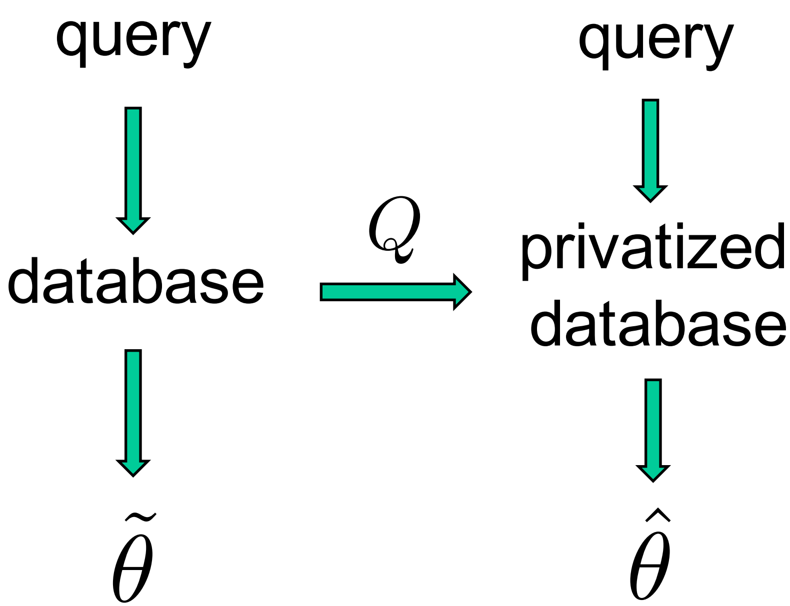


Privacy



query
query
Q
database
privatized database
$\tilde{\theta}$
$\hat{\theta}$

Privacy



Classical problem in differential privacy: show that $\hat{\theta}$ and $\tilde{\theta}$ are close under constraints on Q

Inference

query

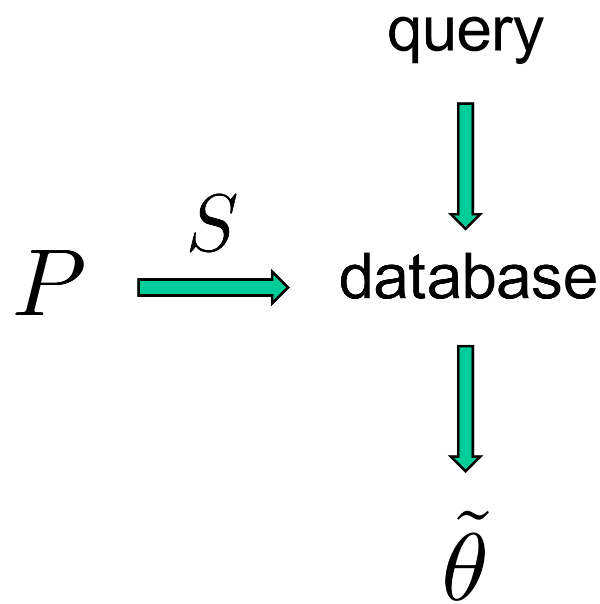
$\downarrow$

database

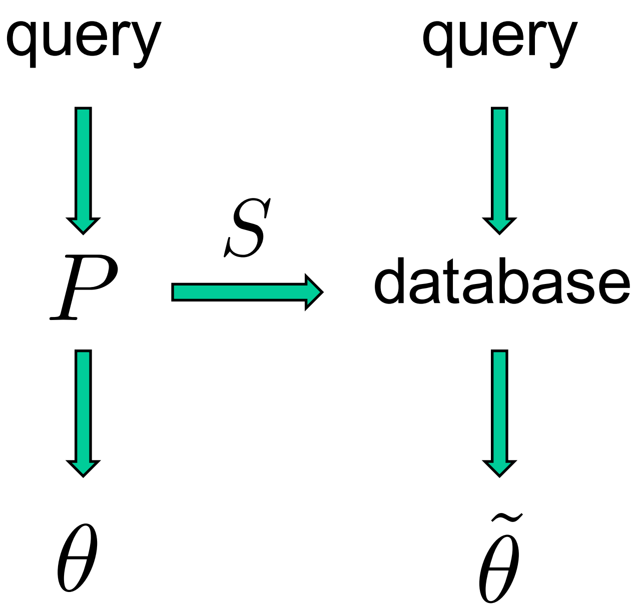
$\downarrow$

$$\tilde{\theta}$$

Inference



Inference



Inference

$$\begin{tikzcd} \text{query} \arrow[d] & \text{query} \arrow[d] \\ P \arrow[r, "S"] \arrow[d] & \text{database} \arrow[d] \\ \theta & \tilde{\theta} \end{tikzcd}$$

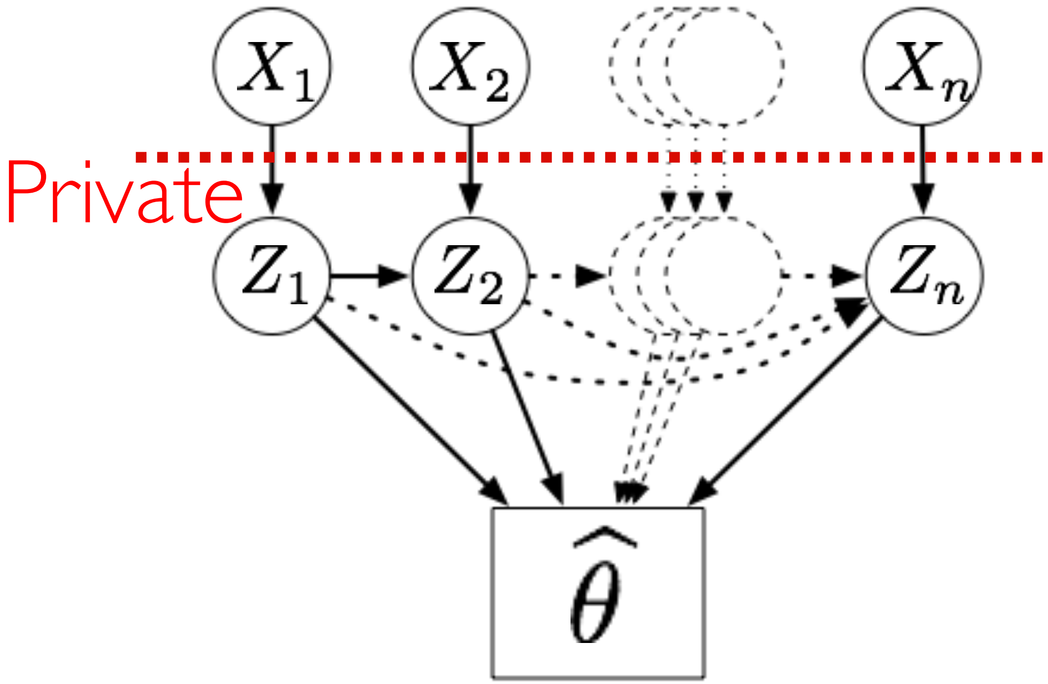
Classical problem in statistical theory: show that $\tilde{\theta}$ and θ are close under constraints on S

Privacy and Inference

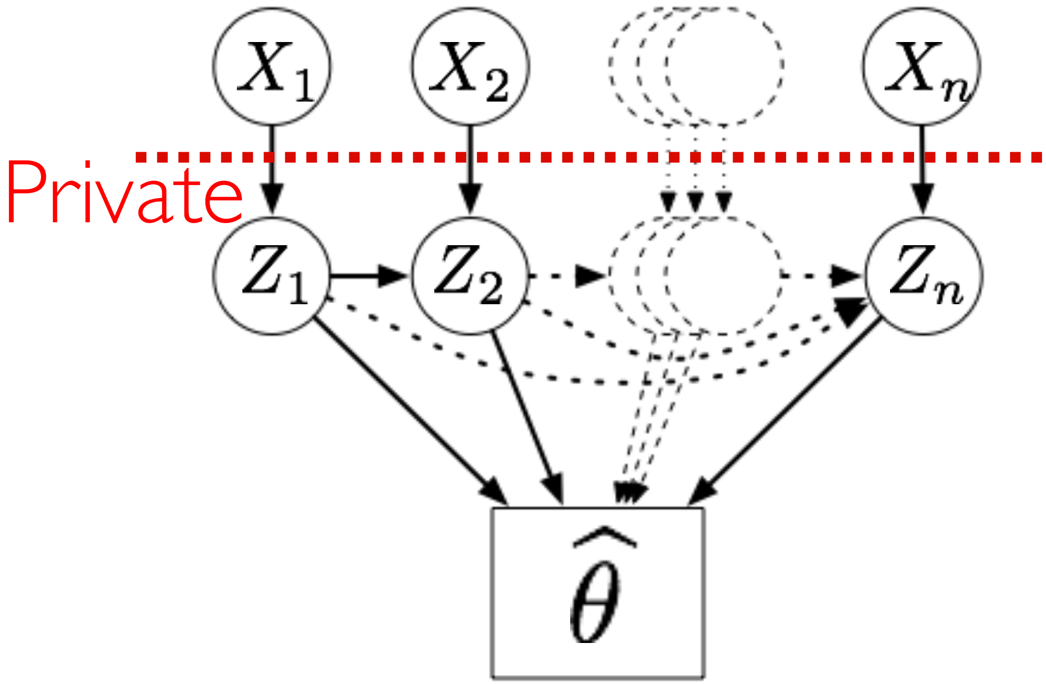
$$
\begin{tikzcd}
\text{query} \arrow[d] & \text{query} \arrow[d] & \text{query} \arrow[d] \\
P \arrow[r, "S"] \arrow[d] & \text{database} \arrow[r, "Q"] \arrow[d] & \text{privatized database} \arrow[d] \\
\theta & \tilde{\theta} & \hat{\theta}
\end{tikzcd}
$$

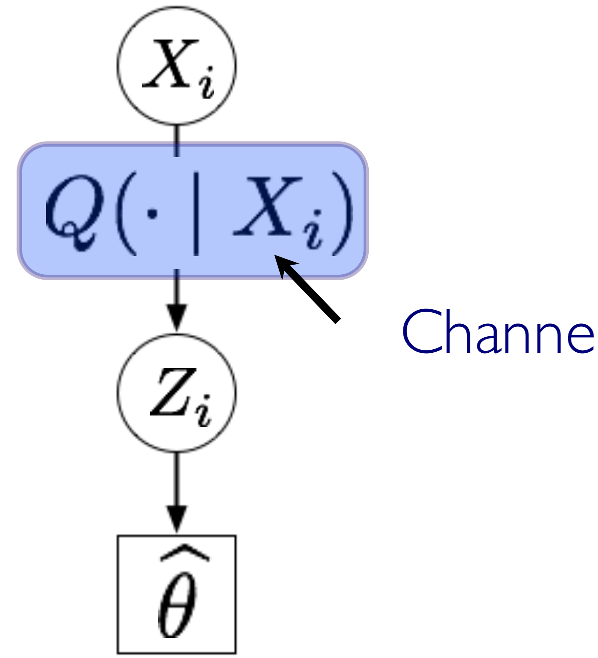
The privacy-meets-inference problem: show that θ and $\hat{\theta}$ are close under constraints on Q and on S

Local Privacy



Local Privacy





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

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

Private 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 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 $\theta = \begin{array}{ll} \theta_1 & = .45 \\ \theta_2 & = .32 \\ \theta_3 & = .16 \\ \theta_4 & = .20 \\ \theta_5 & = .00 \\ \theta_6 & = .02 \end{array}$

Vignette: Mean Estimation

Consider estimation of mean $\theta(P) := \mathbb{E}_P[X] \in \mathbb{R}^d$, with errors measured in ℓ_∞-norm, 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 $\theta(P) := \mathbb{E}_P[X] \in \mathbb{R}^d$, with errors measured in ℓ_∞-norm, 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 $\quad n \mapsto n\alpha^2/d$

Additional Examples

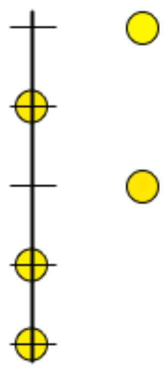
- Fixed-design regression
- Convex risk minimization
- Multinomial estimation
- Nonparametric density estimation

- Almost always, the effective sample size reduction is:

$$n \mapsto \frac{n\alpha^2}{d}$$

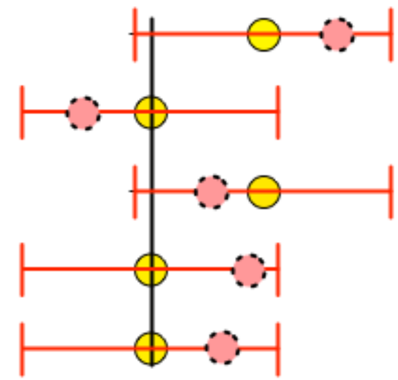
Optimal mechanism?

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



Non-private observation

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



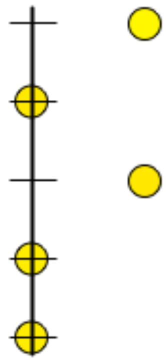
Idea 1: add independent noise
(e.g. Laplace mechanism)

[Dwork et al. 06]

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

Optimal mechanism

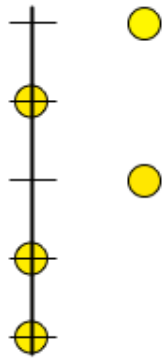
$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



Non-private
observation

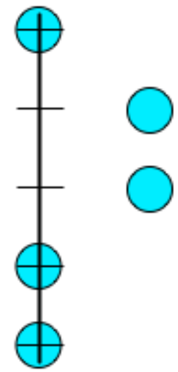
Optimal mechanism

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



Non-private observation

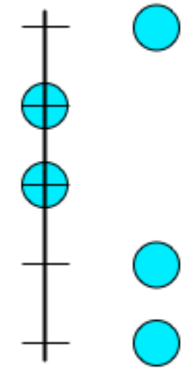
$$v = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



View 1

(closer: 3 overlap)

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



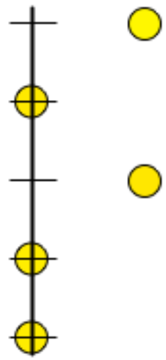
View 2

(farther: 2 overlap)

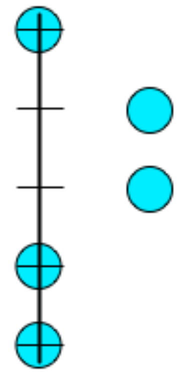
- Draw v uniformly in $\{0, 1\}^d$

Optimal mechanism

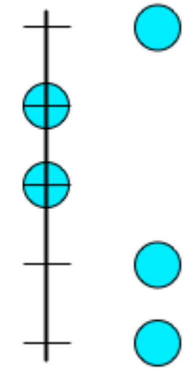
$$X = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$



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



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



Non-private observation

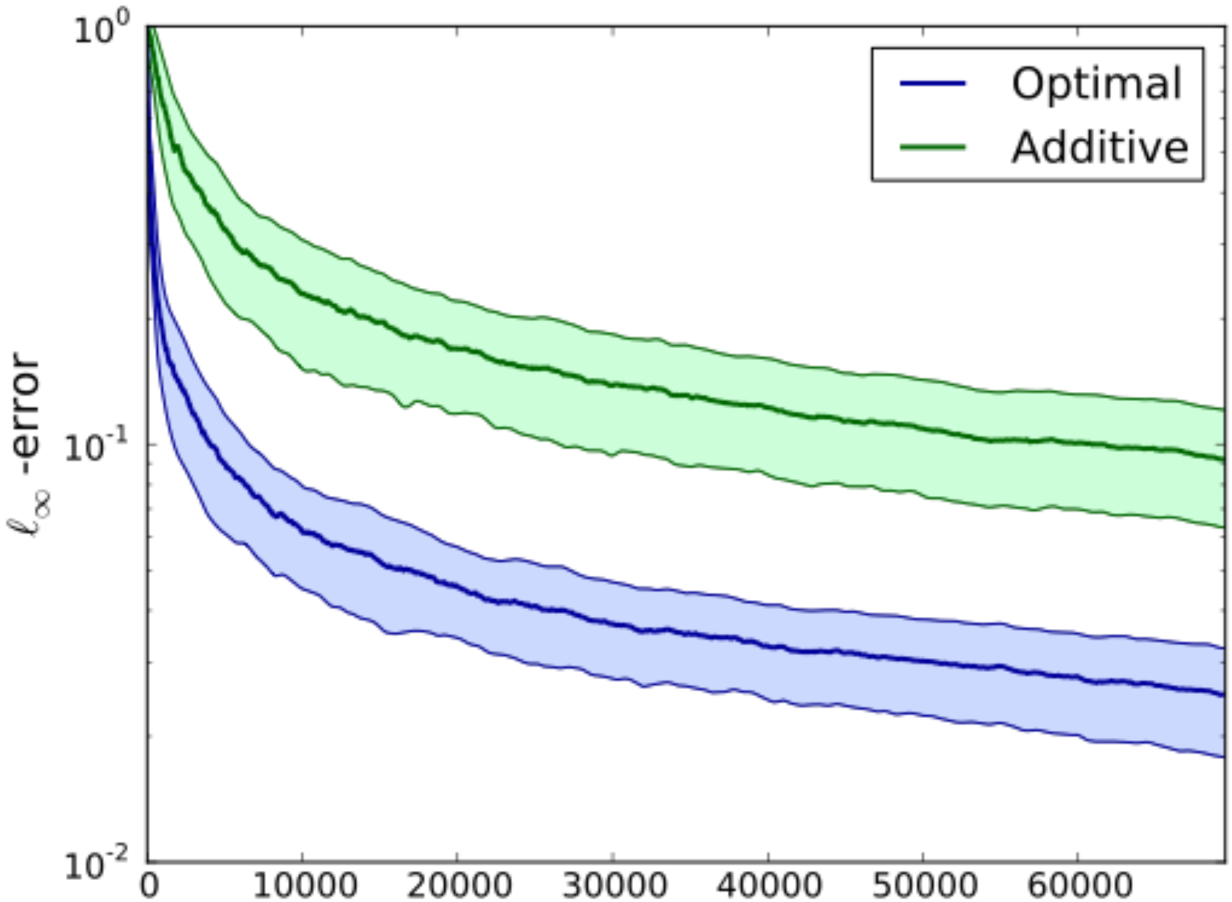
View 1 (closer: 3 overlap)

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

Empirical evidence



Sample size *n*

Data source:
Drug Abuse Warning Network

Estimate proportion of emergency room visits involving different substances

Computation and Inference

- How does inferential quality trade off against classical computational resources such as time and space?

Computation and Inference

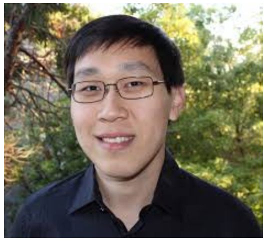
- How does inferential quality trade off against classical computational resources such as time and space?
- Hard!

Computation and Inference: Mechanisms and Bounds

- Tradeoffs via convex relaxations
  - linking runtime to convex geometry and risk to convex geometry
- Tradeoffs via concurrency control
  - optimistic concurrency control
- Bounds via optimization oracles
  - number of accesses to a gradient as a surrogate for computation
- Bounds via communication complexity
- Tradeoffs via subsampling
  - bag of little bootstraps, variational consensus Monte Carlo

Part II: Variational, Hamiltonian and Symplectic Perspectives on Acceleration

with Andre Wibisono, Ashia Wilson and Michael Betancourt

The Important Role of Optimization

- Optimization has been playing an increasingly important role in Data Science

The Important Role of Optimization

- Optimization has been playing an increasingly important role in Data Science
- It not only supplies algorithms

The Important Role of Optimization

- Optimization has been playing an increasingly important role in Data Science
- It not only supplies algorithms
- But it also supplies lower bounds, and thereby fundamental understanding

The Important Role of Optimization

- Optimization has been playing an increasingly important role in Data Science
- It not only supplies algorithms
- But it also supplies lower bounds, and thereby fundamental understanding

- But, perhaps surprisingly, optimization is still an immature field, and open problems abound

Interplay between Differentiation and Integration

- The 300-yr-old fields: Physics, Statistics
  - cf. Lagrange/Hamilton, Laplace expansions, saddlepoint expansions
- The numerical disciplines
  - e.g.,. finite elements, Monte Carlo

Interplay between Differentiation and Integration

- The 300-yr-old fields: Physics, Statistics
  - cf. Lagrange/Hamilton, Laplace expansions, saddlepoint expansions
- The numerical disciplines
  - e.g.,. finite elements, Monte Carlo
- Optimization?

Interplay between Differentiation and Integration

- The 300-yr-old fields: Physics, Statistics
  - cf. Lagrange/Hamilton, Laplace expansions, saddlepoint expansions
- The numerical disciplines
  - e.g.,. finite elements, Monte Carlo
- Optimization?
  - to date, almost entirely focused on differentiation

Accelerated gradient descent

Setting: Unconstrained convex optimization

$$\min_{x \in \mathbb{R}^d} f(x)$$

- Classical gradient descent:

$$x_{k+1} = x_k - \beta \nabla f(x_k)$$

  obtains a convergence rate of $O(1/k)$

- Accelerated gradient descent:

$$\begin{aligned} y_{k+1} &= x_k - \beta \nabla f(x_k) \\ x_{k+1} &= (1-\lambda_k) y_{k+1} + \lambda_k y_k \end{aligned}$$

  obtains the (optimal) convergence rate of $O(1/k^2)$

Accelerated methods: Continuous time perspective

- Gradient descent is discretization of gradient flow

$$\dot{X}_t = -\nabla f(X_t)$$

  (and mirror descent is discretization of natural gradient flow)

- Su, Boyd, Candes '14: Continuous time limit of accelerated gradient descent is a second-order ODE

$$\ddot{X}_t + \frac{3}{t}\dot{X}_t + \nabla f(X_t) = 0$$

- These ODEs are obtained by taking continuous time limits. Is there a deeper generative mechanism?

Our work: A general variational approach to acceleration

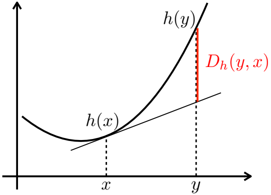
A systematic discretization methodology

Bregman Lagrangian

Define the **Bregman Lagrangian**:

$$\mathcal{L}(x, \dot{x}, t) = e^{\gamma_t + \alpha_t} \left(D_h(x + e^{-\alpha_t}\dot{x}, x) - e^{\beta_t} f(x) \right)$$

- Function of position x, velocity $\dot{x}$, and time t
- $D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle$ is the Bregman divergence
- h is the convex distance-generating function
- f is the convex objective function

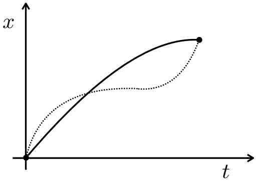


Bregman Lagrangian

$$\mathcal{L}(x, \dot{x}, t) = e^{\gamma_t + \alpha_t} \left(D_h(x + e^{-\alpha_t} \dot{x}, x) - e^{\beta_t} f(x) \right)$$

Variational problem over curves:

$$\min_X \int \mathcal{L}(X_t, \dot{X}_t, t) \, dt$$



Optimal curve is characterized by **Euler-Lagrange** equation:

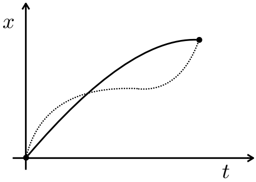
$$\frac{d}{dt} \left\{ \frac{\partial \mathcal{L}}{\partial \dot{x}}(X_t, \dot{X}_t, t) \right\} = \frac{\partial \mathcal{L}}{\partial x}(X_t, \dot{X}_t, t)$$

Bregman Lagrangian

$$\mathcal{L}(x, \dot{x}, t) = e^{\gamma_t + \alpha_t} \left(D_h(x + e^{-\alpha_t} \dot{x}, x) - e^{\beta_t} f(x) \right)$$

Variational problem over curves:

$$\min_X \int \mathcal{L}(X_t, \dot{X}_t, t)\, dt$$



Optimal curve is characterized by **Euler-Lagrange** equation:

$$\frac{d}{dt} \left\{ \frac{\partial \mathcal{L}}{\partial \dot{x}}(X_t, \dot{X}_t, t) \right\} = \frac{\partial \mathcal{L}}{\partial x}(X_t, \dot{X}_t, t)$$

E-L equation for Bregman Lagrangian under ideal scaling:

$$\ddot{X}_t + (e^{\alpha_t} - \dot{\alpha}_t)\dot{X}_t + e^{2\alpha_t + \beta_t} \left[\nabla^2 h(X_t + e^{-\alpha_t} \dot{X}_t) \right]^{-1} \nabla f(X_t) = 0$$

General convergence rate

Theorem

Theorem Under ideal scaling, the E-L equation has convergence rate

$$f(X_t) - f(x^*) \leq O(e^{-\beta_t})$$

Proof. Exhibit a Lyapunov function for the dynamics:

$$\mathcal{E}_t = D_h\left(x^*,\, X_t + e^{-\alpha_t}\dot{X}_t\right) + e^{\beta_t}(f(X_t) - f(x^*))$$

$$\dot{\mathcal{E}}_t = -e^{\alpha_t+\beta_t} D_f(x^*, X_t) + (\dot{\beta}_t - e^{\alpha_t})e^{\beta_t}(f(X_t) - f(x^*)) \;\leq\; 0$$

□

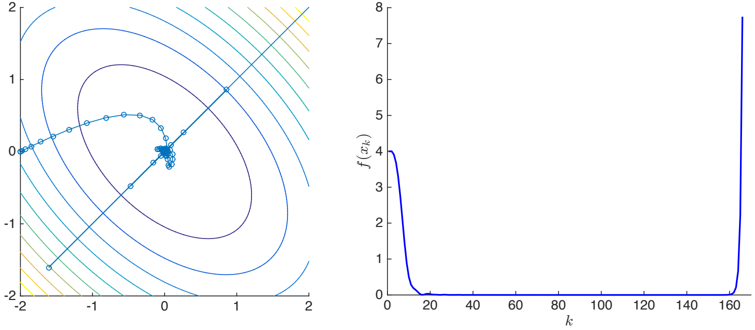
Note: Only requires convexity and differentiability of f, h

Naive discretization doesn't work

$$x_{k+1} = \frac{p}{k+p} z_k + \frac{k}{k+p} x_k$$

$$z_k = \arg\min_z \left\{ Cpk^{(p-1)} \langle \nabla f(x_k), z \rangle + \frac{1}{\epsilon} D_h(z, z_{k-1}) \right\}$$

Cannot obtain a convergence guarantee, and empirically unstable



Mysteries

- Why can't we discretize the dynamics when we are using exponentially fast clocks?
- What happens when we arrive at a clock speed that we can discretize?
- How do we discretize once it's possible?

Mysteries

- Why can’t we discretize the dynamics when we are using exponentially fast clocks?
- What happens when we arrive at a clock speed that we can discretize?
- How do we discretize once it’s possible?

- The answers are to be found in symplectic integration

Symplectic Integration

- Consider discretizing a system of differential equations obtained from physical principles
- Solutions of the differential equations generally conserve various quantities (energy, momentum, volumes in phase space)
- Is it possible to find discretizations whose solutions exactly conserve these same quantities?
- Yes!
  - from a long line of research initiated by Jacobi, Hamilton, Poincare' and others

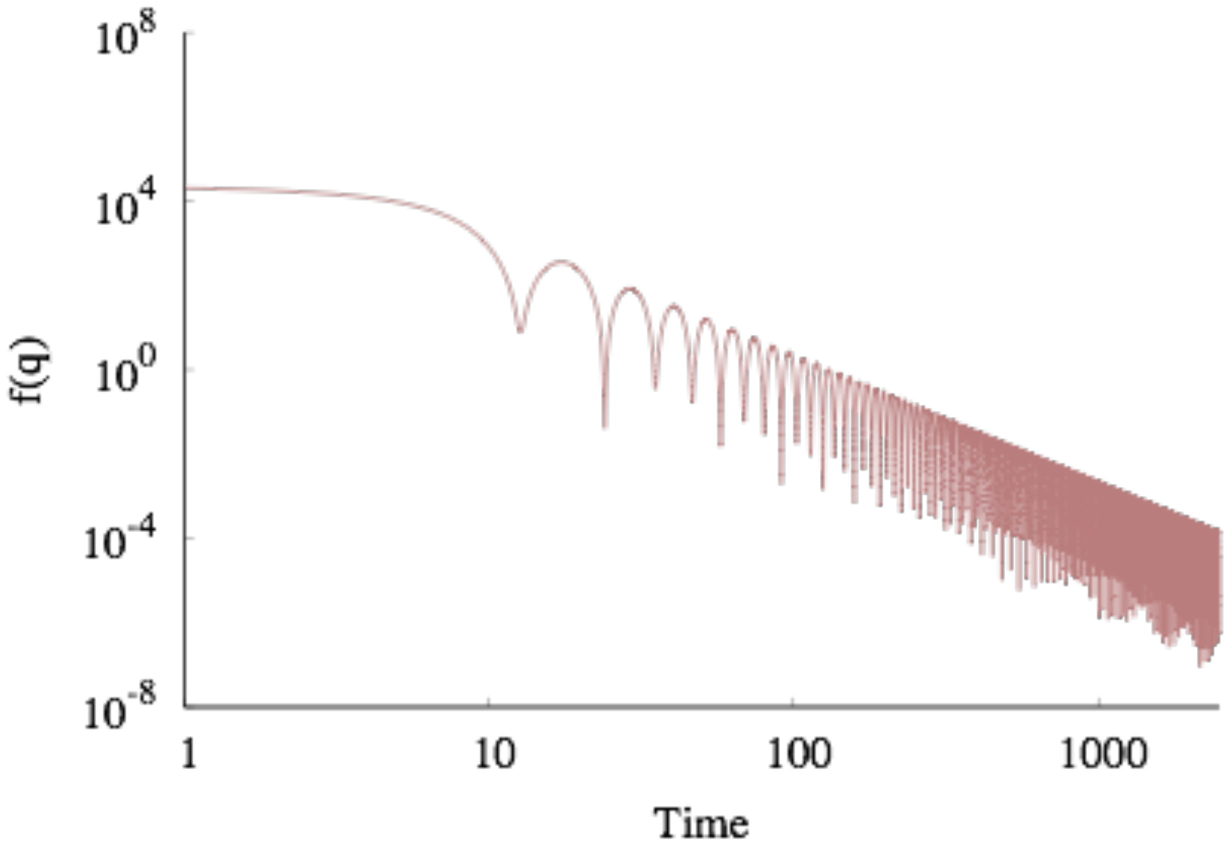
Towards A Symplectic Perspective

- We've discussed discretization of Lagrangian-based dynamics
- Discretization of Lagrangian dynamics is often fragile and requires small step sizes
- We can build more robust solutions by taking a Legendre transform and considering a *Hamiltonian* formalism:

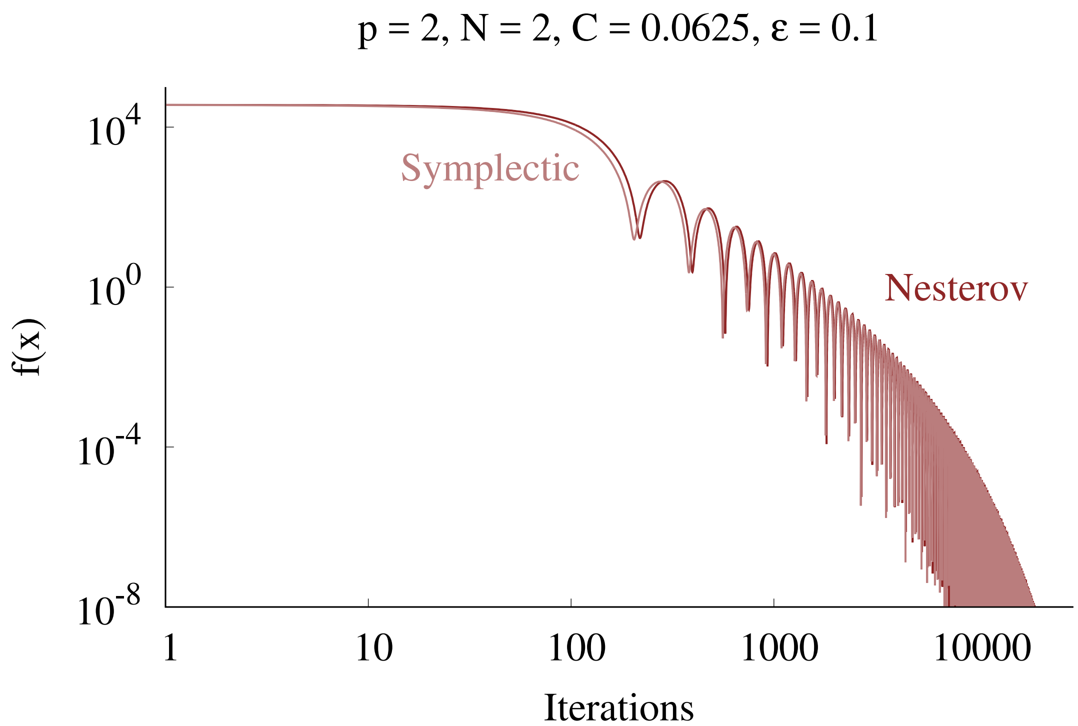
$$L(q, v, t) \rightarrow H(q, p, t, \mathcal{E})$$

$$\left(\frac{\mathrm{d}q}{\mathrm{d}t}, \frac{\mathrm{d}v}{\mathrm{d}t}\right) \rightarrow \left(\frac{\mathrm{d}q}{\mathrm{d}\tau}, \frac{\mathrm{d}p}{\mathrm{d}\tau}, \frac{\mathrm{d}t}{\mathrm{d}\tau}, \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}\tau}\right)$$

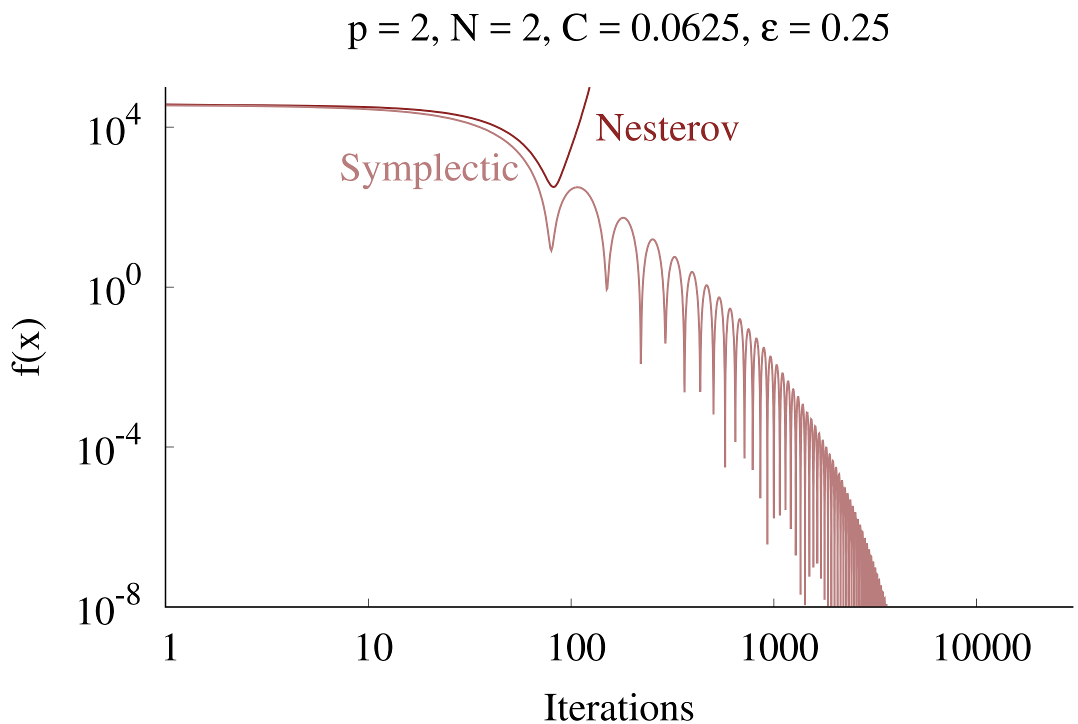
Symplectic Integration of Bregman Hamiltonian



Symplectic vs Nesterov



Symplectic vs Nesterov



Discussion

- Data and inferential problems will be everywhere in computer science, and will fundamentally change the field
- Many conceptual and mathematical challenges arising in taking this effort seriously, in addition to systems challenges and "outreach" challenges
- Facing these challenges will require a rapprochement between computational thinking and inferential thinking
- This effort is just beginning!

Reference

- Wibisono, A., Wilson, A. and Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences, 133*, E7351-E7358.