Noise Robustness of Optimization Algorithms via Statistical Queries

Cristóbal Guzmán

guzman@cwi.nl

CWI Scientific Meeting
27-11-2015

Motivation

Motivation

- In algorithms: Faster is better, right?

Motivation

- In algorithms: Faster is better, right?
- Sometimes, other objectives are as important

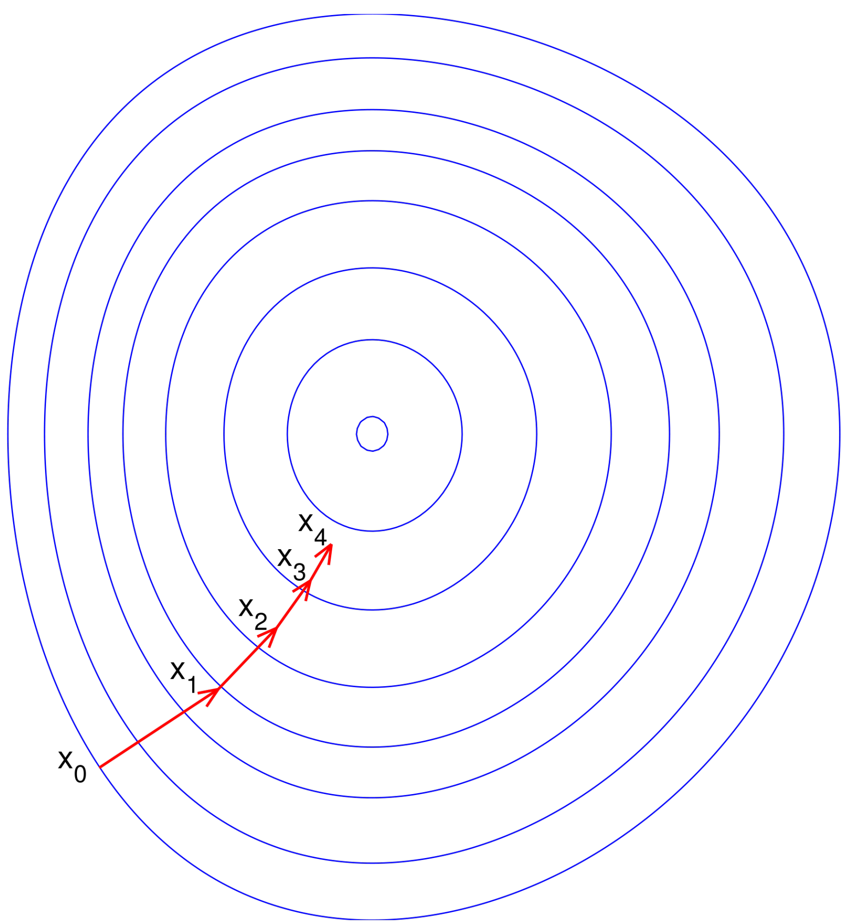
Motivation

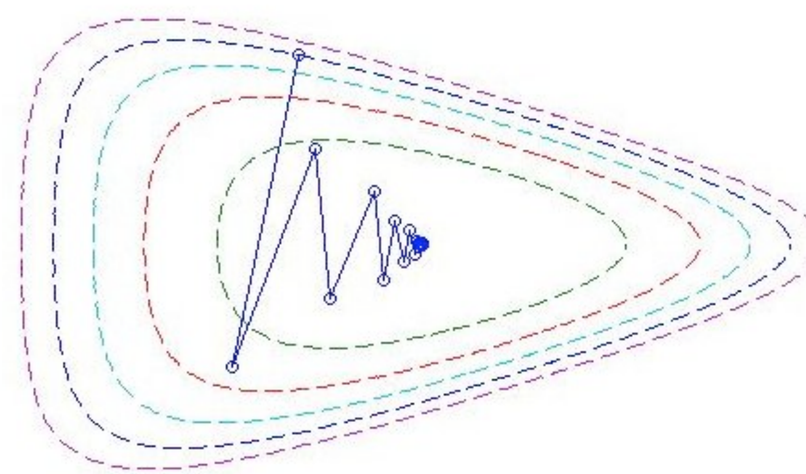
- In algorithms: Faster is better, right?
- Sometimes, other objectives are as important
- This time, we care about robustness

Example: Gradient Methods

Goal: Find $f^* = \min_{x \in X} f(x)$

where $f : X \to \mathbb{R}$ smooth convex function, $X \subseteq \mathbb{R}^d$ compact convex set





Example: Gradient Methods

Goal: Find $f^* = \min_{x \in X} f(x)$

- Gradient Descent *[Euler:1770]*

$$x^{t+1} = x^t - \gamma_t \nabla f(x^t)$$

Example: Gradient Methods

Goal: Find $f^* = \min_{x \in X} f(x)$

- Gradient Descent *[Euler:1770]*

$$x^{t+1} = x^t - \gamma_t \nabla f(x^t) \qquad \Rightarrow \qquad f(x^T) - f^* = O(1/T)$$

Example: Gradient Methods

Goal: Find $f^* = \min_{x \in X} f(x)$

- Gradient Descent *[Euler:1770]*

$$x^{t+1} = x^t - \gamma_t \nabla f(x^t) \qquad \Rightarrow \qquad f(x^T) - f^* = O(1/T)$$

- Accelerated Gradient Method *[Nesterov:1983]*

$$\begin{aligned} x^t &= y^t - \gamma_t \nabla f(y^t) \\ y^{t+1} &= x^t + \frac{\alpha_t - 1}{\alpha_{t+1}}(x^t - x^{t-1}) \end{aligned}$$

Example: Gradient Methods

Goal: Find $f^* = \min_{x \in X} f(x)$

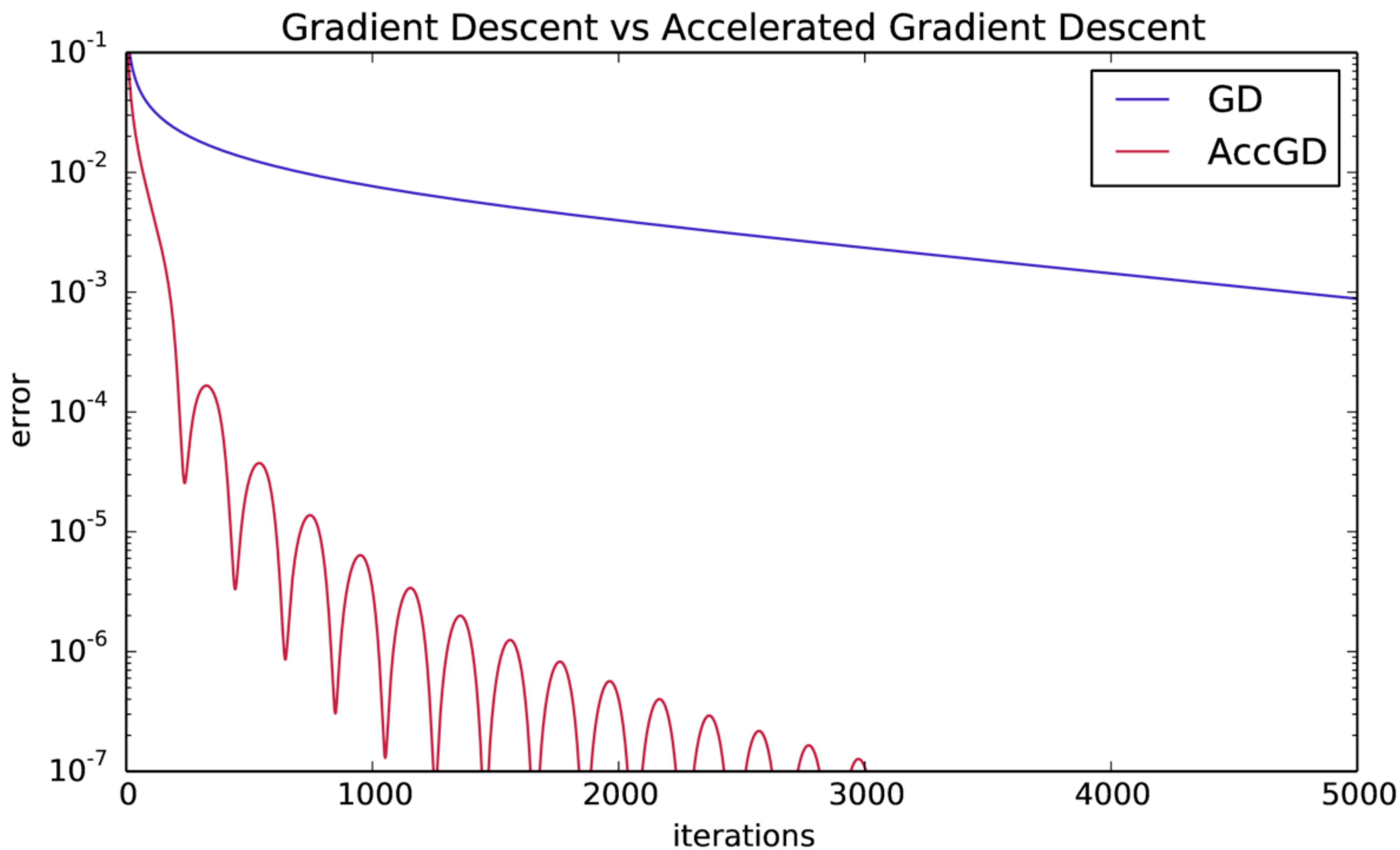
- Gradient Descent *[Euler:1770]*

$$x^{t+1} = x^t - \gamma_t \nabla f(x^t) \qquad \Rightarrow \qquad f(x^T) - f^* = O(1/T)$$

- Accelerated Gradient Method *[Nesterov:1983]*

$$\begin{aligned} x^t &= y^t - \gamma_t \nabla f(y^t) \\ y^{t+1} &= x^t + \frac{\alpha_t - 1}{\alpha_{t+1}}(x^t - x^{t-1}) \end{aligned} \Rightarrow \qquad f(x^T) - f^* = O(1/T^2)$$

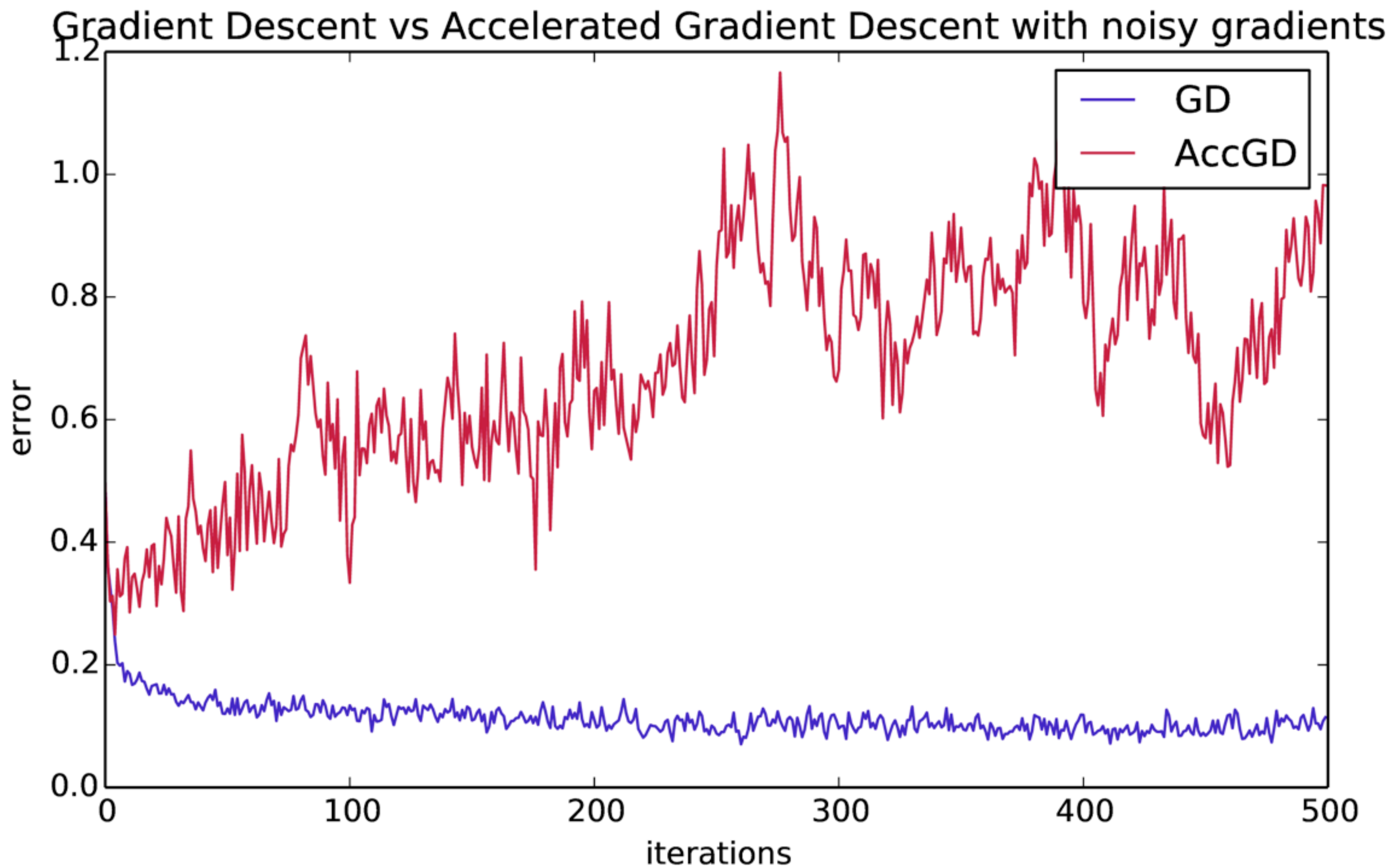
Example: Gradient Methods



Source: Moritz Hardt's blog

$$g(x) = \nabla f(x)$$

Example: Gradient Methods



Source: Moritz Hardt's blog

$$g(x) = \nabla f(x) + \xi$$

$$\xi \sim \mathcal{N}(0, \sigma^2), \quad \sigma = 0.1$$

Why Noise Robustness?

- There are other important reasons why acceleration could be harmful

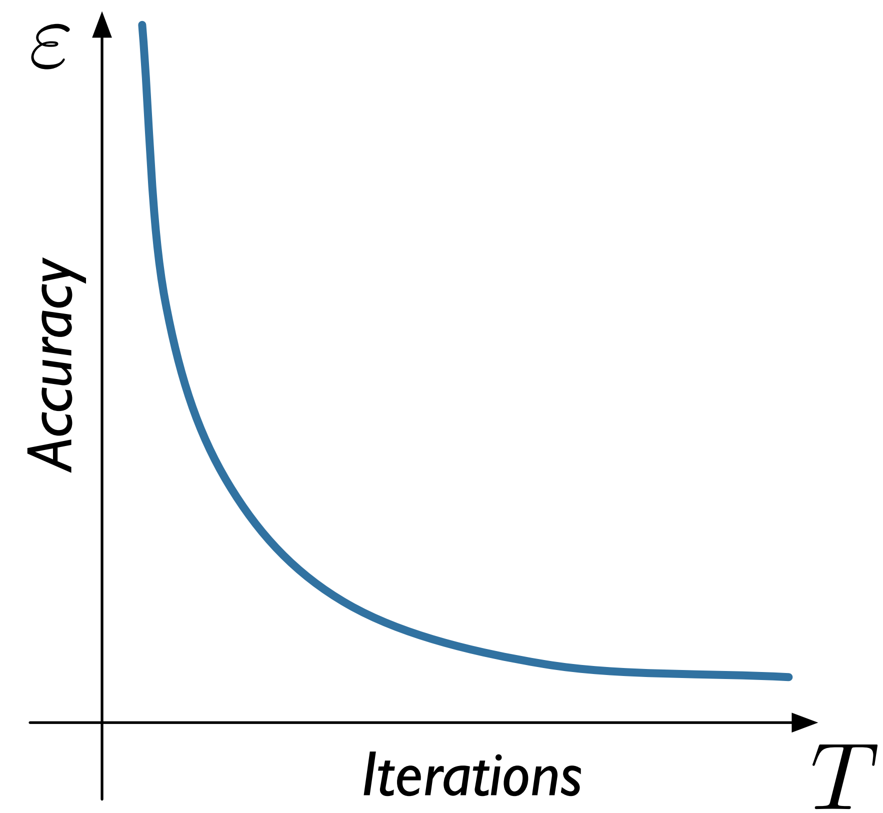
Why Noise Robustness?

- There are other important reasons why acceleration could be harmful
  - Overfitting, privacy leakage, etc.

Why Noise Robustness?

- There are other important reasons why acceleration could be harmful
  - Overfitting, privacy leakage, etc.

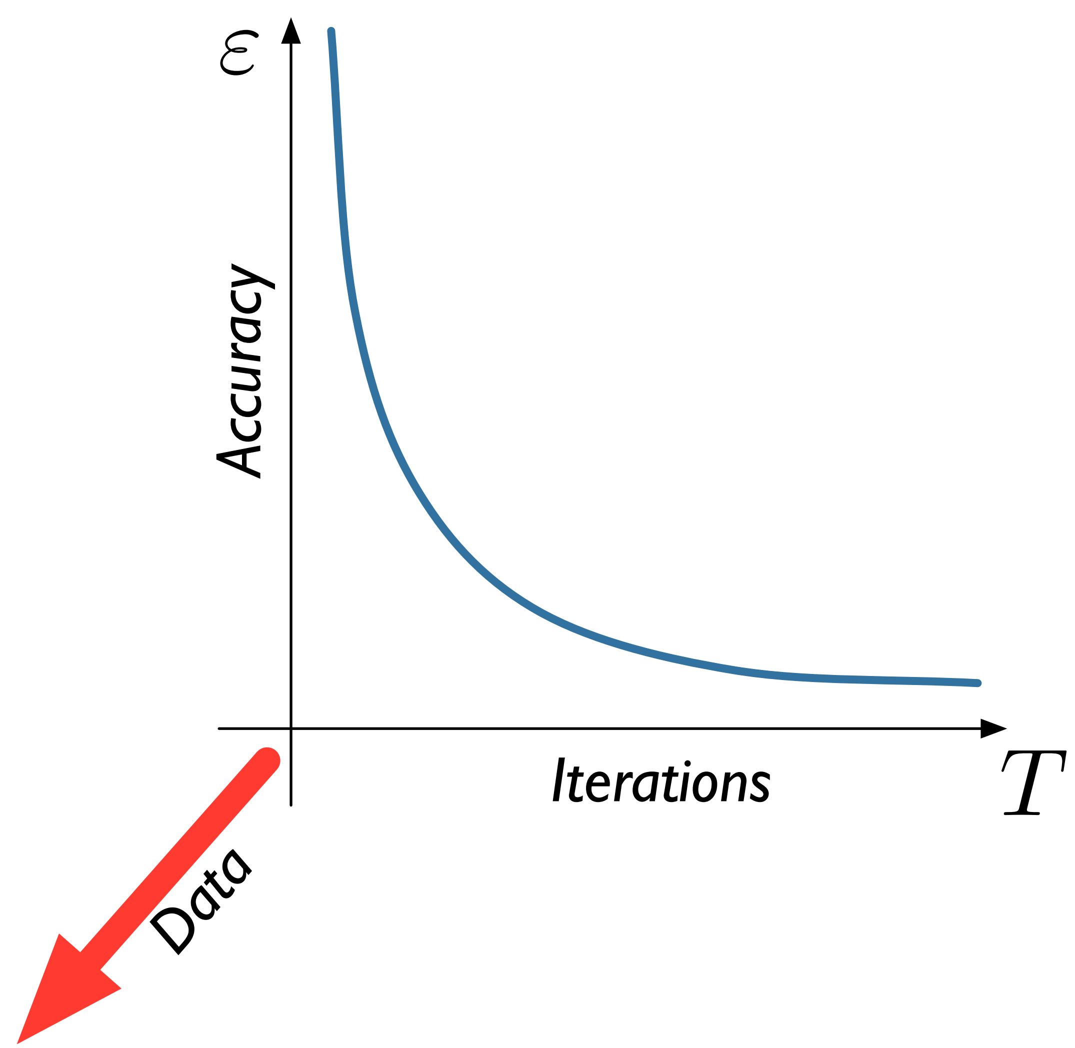
- Goal: A theory of optimization algorithms that incorporates noise sensitivity



Why Noise Robustness?

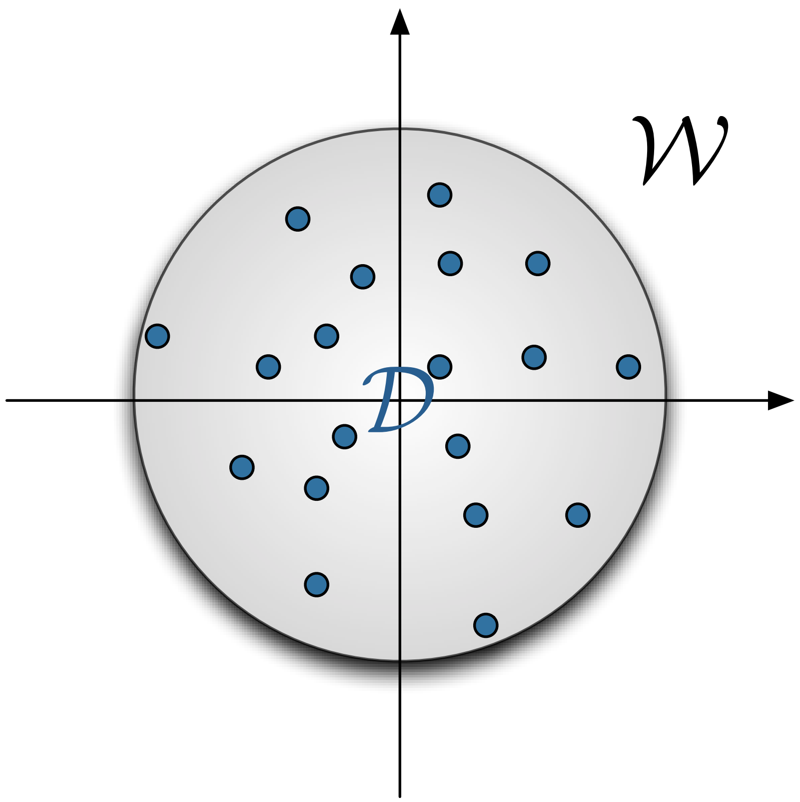
- There are other important reasons why acceleration could be harmful
    - Overfitting, privacy leakage, etc.

- Goal: A theory of optimization algorithms that incorporates noise sensitivity



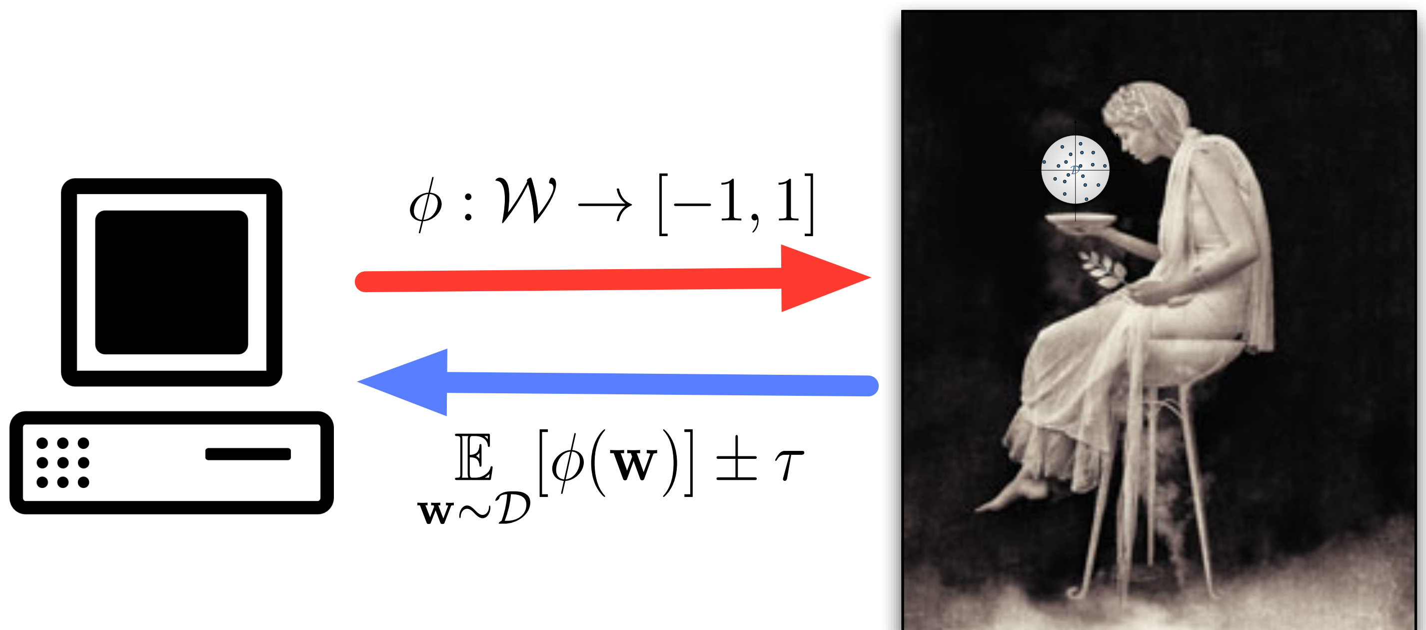
Statistical Queries

- Unknown distribution $\mathcal{D}$ supported on $\mathcal{W}$



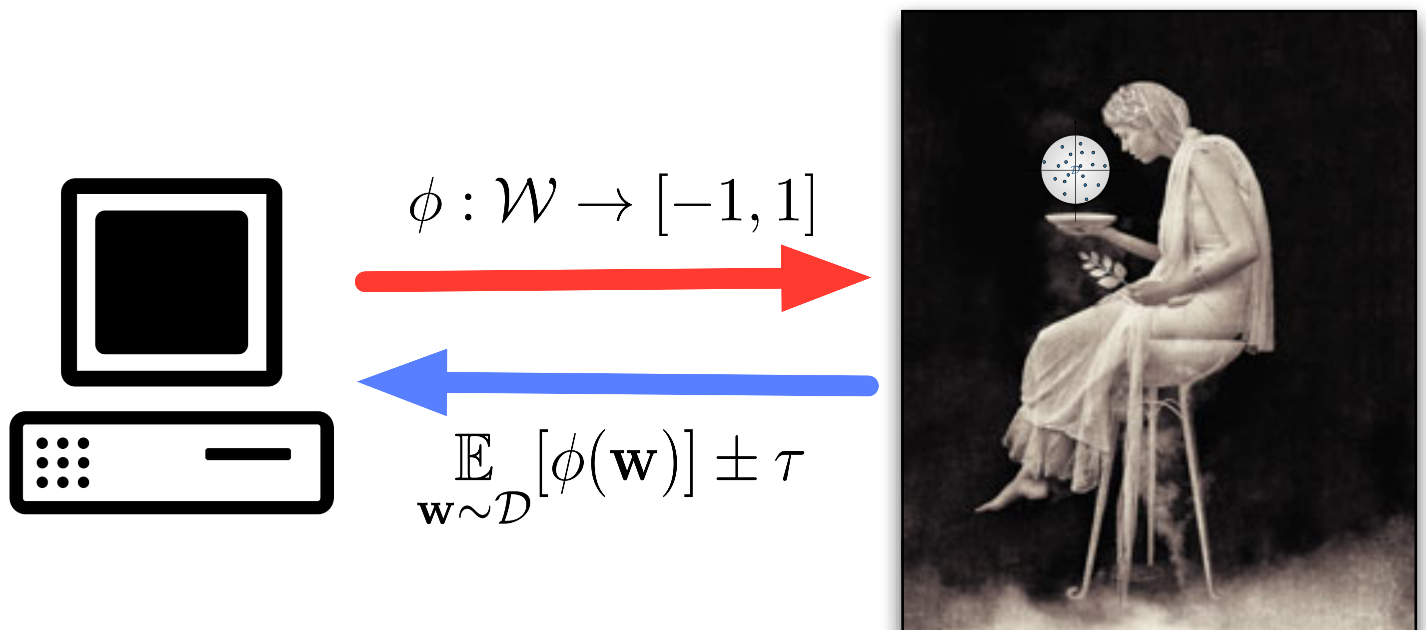
Statistical Queries

- Unknown distribution $\mathcal{D}$ supported on $\mathcal{W}$
- Algorithm has oracle access to $\mathcal{D}$



Statistical Queries

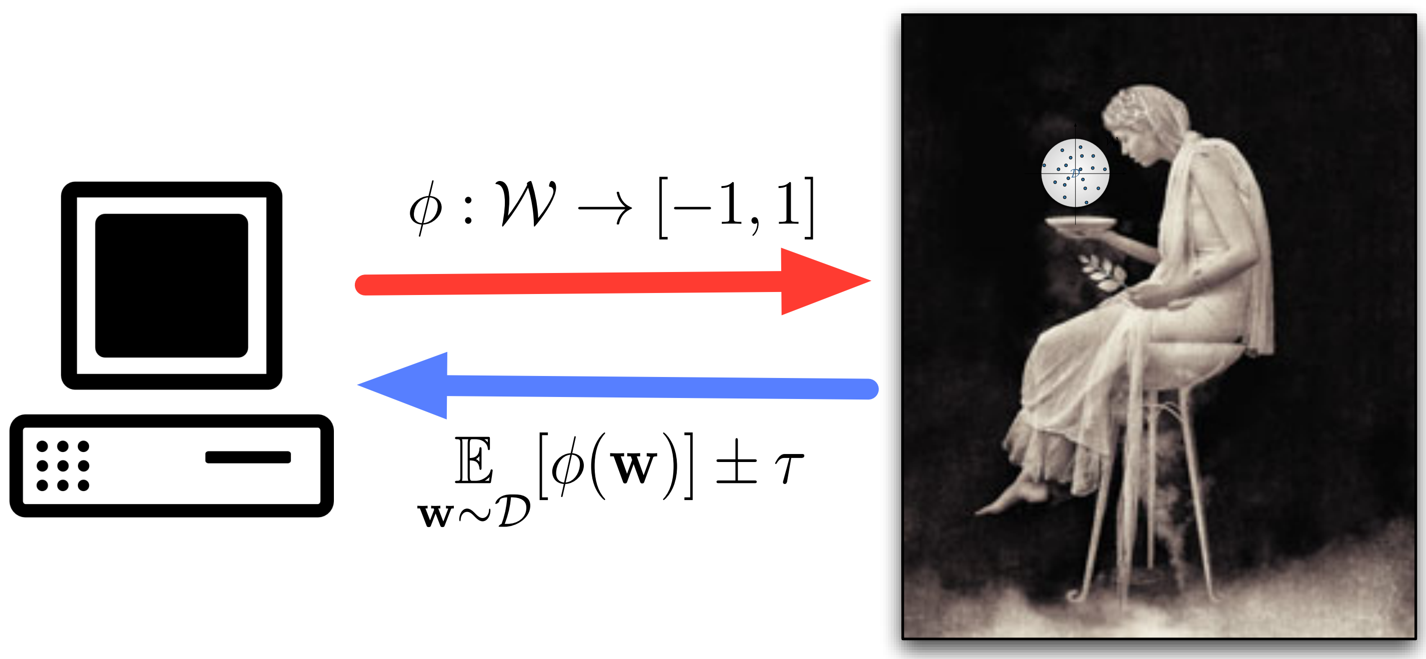
- Unknown distribution $\mathcal{D}$ supported on $\mathcal{W}$
- Algorithm has oracle access to $\mathcal{D}$



An oracle query can be emulated by $1/\tau^2$ samples

Statistical Queries

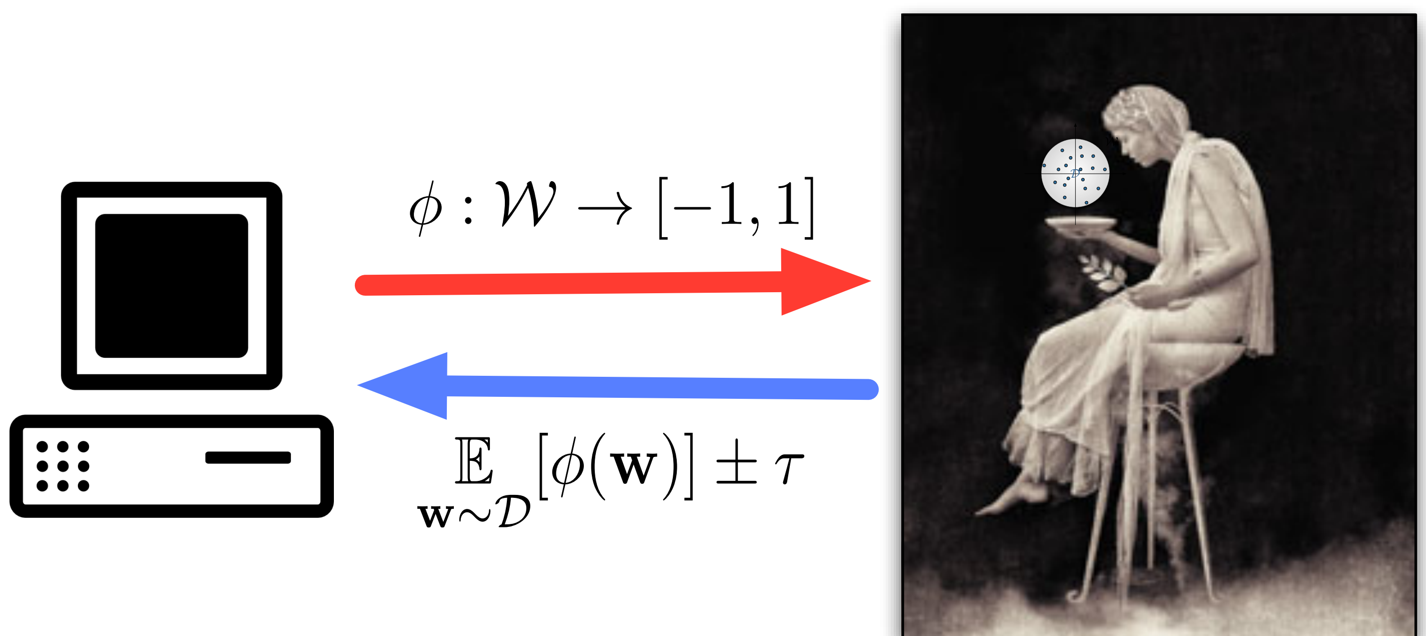
- Unknown distribution $\mathcal{D}$ supported on $\mathcal{W}$
- Algorithm has oracle access to $\mathcal{D}$



- Goals
    - Efficiency ~ small number of queries

Statistical Queries

- Unknown distribution $\mathcal{D}$ supported on $\mathcal{W}$
- Algorithm has oracle access to $\mathcal{D}$



- Goals
  - Efficiency ~ small number of queries
  - Noise tolerance ~ τ as large as possible

Why Statistical Queries?

SQ algorithms have additional properties, or provide the basis for designing algorithms with additional features:

Why Statistical Queries?

SQ algorithms have additional properties, or provide the basis for designing algorithms with additional features:

- Noise tolerance *[Kearns:1994]*
- Differential privacy *[Blum, Dwork, McSherry, Nissim:2005], [Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith, 2011]*
- Distributed computation *[Balcan, Blum, Fine, Mansour:2012]*
- Generalization in adaptive data analysis *[Dwork, Feldman, Hardt, Pitassi, Reingold, Roth:2014]*

Statistical Query Convex Opt.

- Stochastic convex optimization

$$\min_{x \in X} \mathop{\mathbb{E}}_{\mathbf{w} \sim \mathcal{D}} [f(x, \mathbf{w})]$$

Statistical Query Convex Opt.

- Stochastic convex optimization

$$\min_{x \in X} \mathop{\mathbb{E}}_{\mathbf{w} \sim \mathcal{D}} [f(x, \mathbf{w})]$$

- Statistical queries
    - Function value: $\phi(\cdot) = f(x, \cdot)$

Statistical Query Convex Opt.

- Stochastic convex optimization

$$\min_{x \in X} \mathop{\mathbb{E}}_{\mathbf{w} \sim \mathcal{D}} [f(x, \mathbf{w})]$$

- Statistical queries
  - Function value: $\phi(\cdot) = f(x, \cdot)$
  - Gradient: $\phi(\cdot) = \dfrac{\partial f(x, \cdot)}{\partial x_i} \quad i = 1, \ldots, d$

Statistical Query Convex Opt.

- Stochastic convex optimization

$$\min_{x \in X} \mathop{\mathbb{E}}_{\mathbf{w} \sim \mathcal{D}} [f(x, \mathbf{w})]$$

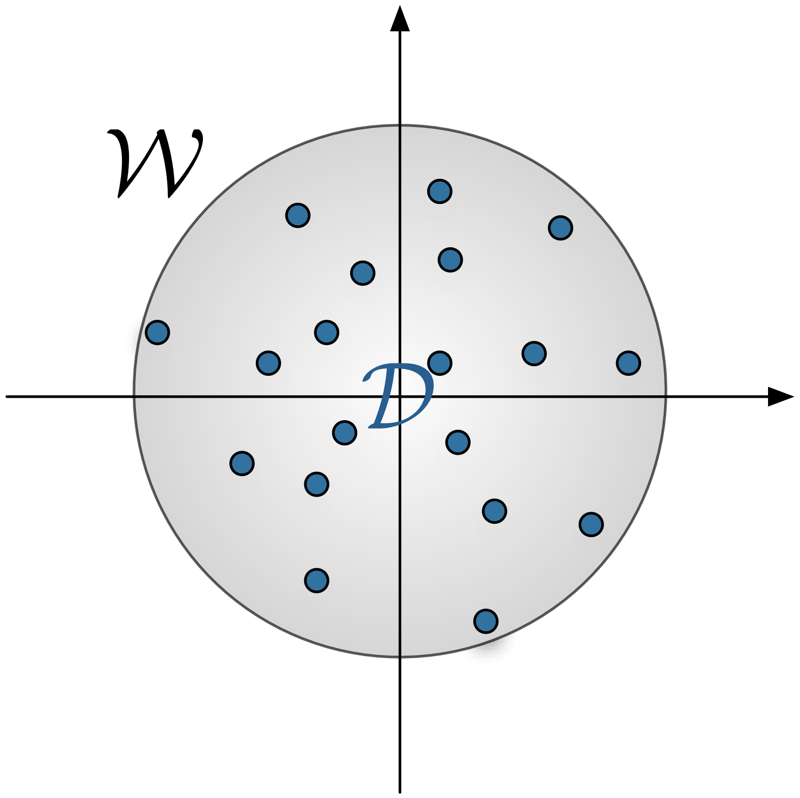
- Statistical queries
  - Function value: $\phi(\cdot) = f(x, \cdot)$
  - Gradient: $\phi(\cdot) = \dfrac{\partial f(x, \cdot)}{\partial x_i} \quad i = 1, \ldots, d$
- Difficulty: estimation in 2-norm accumulates errors

$$g_i = \mathop{\mathbb{E}}_{\mathbf{w} \sim \mathcal{D}} \left[\frac{\partial f(x, \mathbf{w})}{\partial x_i} \right] \pm \tau \quad \Rightarrow \quad \| \nabla \mathbb{E}_{\mathbf{w}} [f(x, \mathbf{w})] - g \|_2 \approx \sqrt{d} \tau$$

Optimal Estimation Algorithm

[Feldman, G., Vempala:2015]

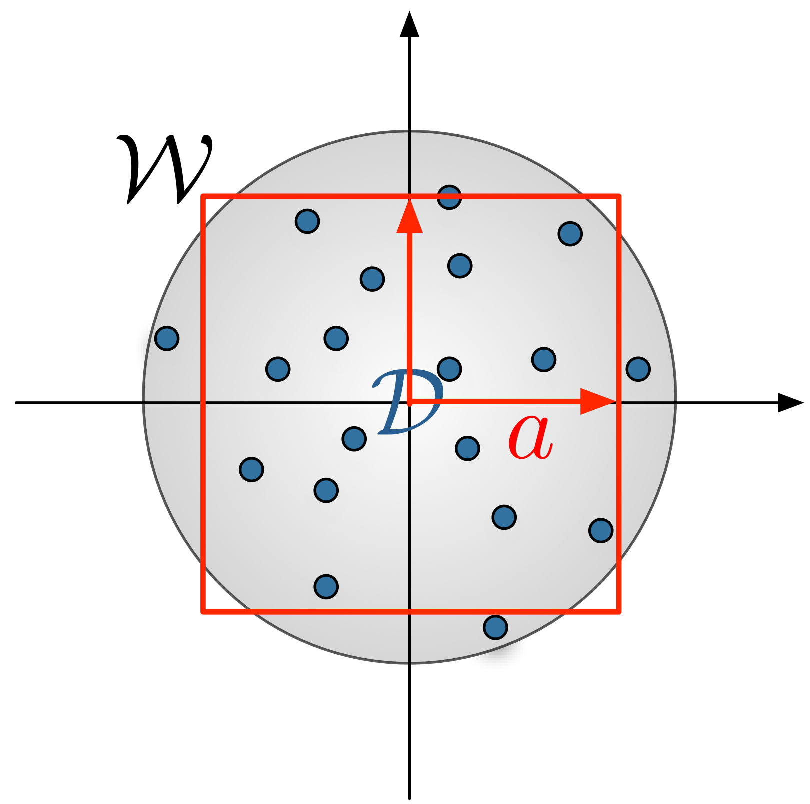
Q: Is it possible to avoid noise accumulation in gradient estimation?



Optimal Estimation Algorithm

[Feldman, G., Vempala:2015]

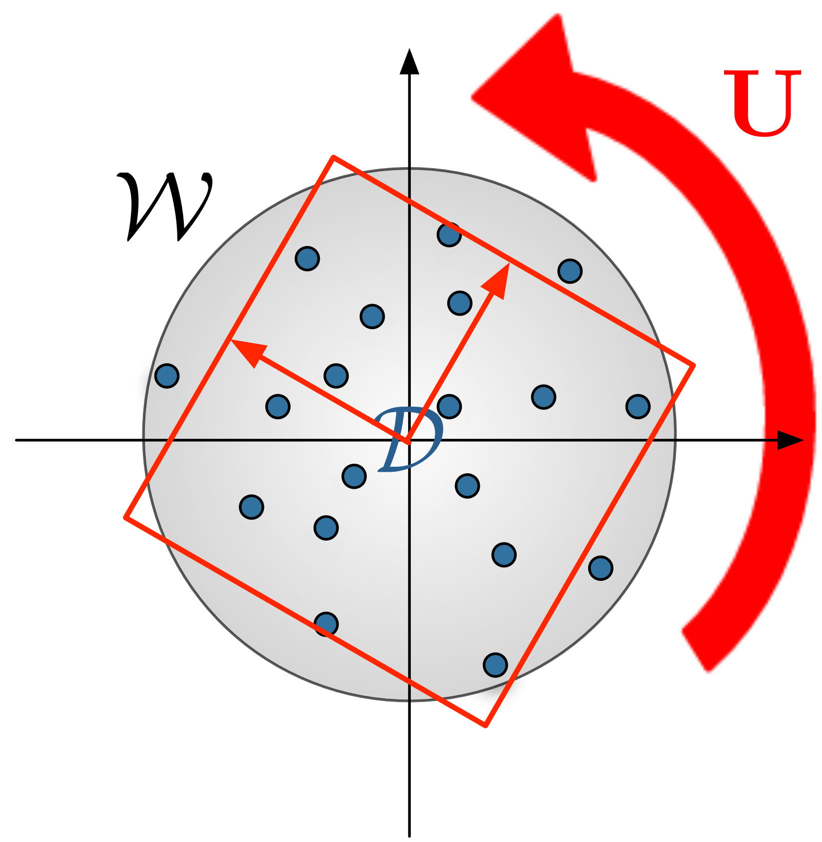
Q: Is it possible to avoid noise accumulation in gradient estimation?



Optimal Estimation Algorithm

[Feldman, G., Vempala:2015]

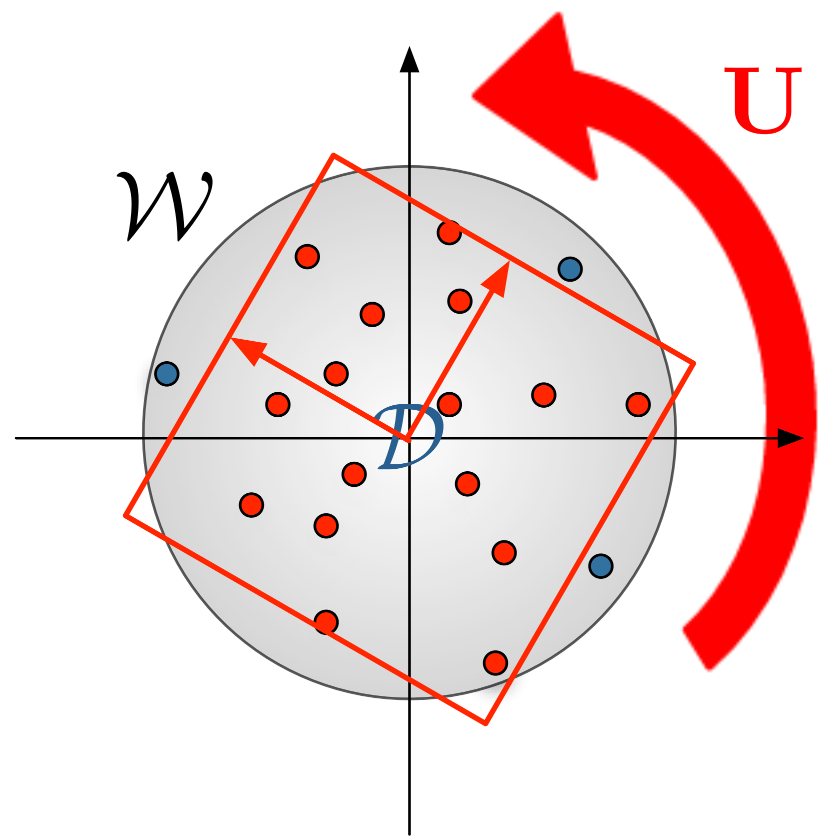
Q: Is it possible to avoid noise accumulation in gradient estimation?



Optimal Estimation Algorithm

[Feldman, G., Vempala:2015]

Q: Is it possible to avoid noise accumulation in gradient estimation?



With high probability $\|\nabla \mathbb{E}_{\mathbf{w}}[f(x, \mathbf{w})] - g\|_2 \approx O(\sqrt{\log d})\tau$

Further Consequences

[Feldman, G., Vempala:2015]

New statistical query algorithms for optimization and machine learning

Further Consequences

[Feldman, G., Vempala:2015]

New statistical query algorithms for optimization and machine learning

- Gradient methods: mirror-descent, accelerated method, strongly convex minimization
- Polynomial-time algorithms: center of gravity, interior-point method
- Improved algorithms for high-dimensional classification (Perceptron) and regression
- Improved differentially-private convex optimization algorithms

Further Consequences

[Feldman, G., Vempala:2015]

New statistical query algorithms for optimization and machine learning

- Gradient methods: mirror-descent, accelerated method, strongly convex minimization
- Polynomial-time algorithms: center of gravity, interior-point method
- Improved algorithms for high-dimensional classification (Perceptron) and regression
- Improved differentially-private convex optimization algorithms

We provide new structural lower bounds for convex optimization

Thank you