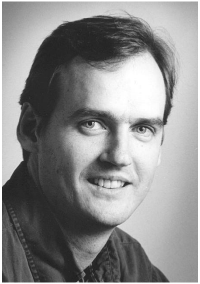
Information-Theoretic Lower Bounds on the Oracle Complexity of Convex Optimization

Alekh Agarwal Peter Bartlett Pradeep Ravikumar
Martin Wainwright
UC Berkeley UT Austin



Convex optimization

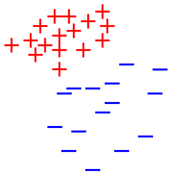
- Convex optimization arises in control, signal processing, machine learning, finance etc.
- Several known algorithms such as gradient descent, Newton method, interior point methods etc.
- Upper bounds on computational complexities for specific methods well-studied.
- Relatively little research on fundamental hardness of convex optimization.
- Minimum computation needed by *any* algorithm to solve a convex optimization problem.

A Motivating Example

- Classical statistics studies *sample complexity* to obtain a certain estimation error.
- Example: binary classification using Support Vector Machines (SVM).

A Motivating Example

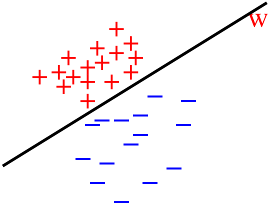
- Classical statistics studies *sample complexity* to obtain a certain estimation error.
- Example: binary classification using Support Vector Machines (SVM).
  - Samples $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathbb{R}^d \times \{-1, 1\})^n$ drawn *i.i.d.*.
  - Learn a mapping $f : \mathbb{R}^d \mapsto \{-1, 1\}$ to predict y given x.



A Motivating Example

- Classical statistics studies *sample complexity* to obtain a certain estimation error.
- Example: binary classification using Support Vector Machines (SVM).
  - Samples $\{(x_1, y_1), \dots, (x_n, y_n)\} \in (\mathbb{R}^d \times \{-1, 1\})^n$ drawn *i.i.d.*.
  - Learn a mapping $f : \mathbb{R}^d \mapsto \{-1, 1\}$ to predict y given x.
  - Predict using $\text{sign}({w_{\text{opt}}}^T x)$.
  - Optimal w_{opt} minimizes the criterion:

$$w_{\text{opt}} = \arg\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i w^T x_i\} + \frac{\lambda}{2} \|w\|^2.$$



Estimation error vs. computational budget

- Learning theory studies error bounds:

$$\mathbb{P}(y \neq \text{sign}({w_{\text{opt}}}^T x)) \leq \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i w^T x_i\} + \mathcal{O}\left(\sqrt{\frac{\ln 1/\delta}{n}}\right)$$

with probability $\geq 1 - \delta$.

Estimation error vs. computational budget

- Learning theory studies error bounds:

$$\mathbb{P}(y \neq \text{sign}({w_{\text{opt}}}^T x)) \leq \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i w^T x_i\} + \mathcal{O}\left(\sqrt{\frac{\ln 1/\delta}{n}}\right)$$

  with probability $\geq 1 - \delta$.
- Sample complexity natural when samples are few.
- Often assumed that computation is abundant.
  - Given enough samples, w_{opt} can be computed efficiently.

Challenges with large datasets

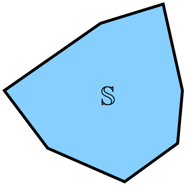
- Large and high-dimensional datasets shift bottleneck from samples to computation.
- w_{opt} result of non-linear non-smooth optimization problem.
- Interested in decay of estimation error with increasing computational budget.
- Algorithm independent understanding of computational complexity.

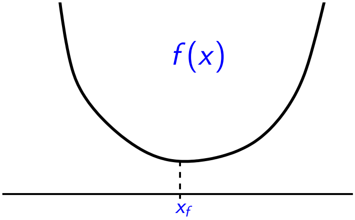
Optimization for Estimation

- Many estimators expressed as results of optimization problems.
- Most learning algorithms based on minimizing a convex objective function.
- Examples:
  - binary classification (e.g. SVM, logistic regression, boosting etc.)
  - least squares regression (e.g. ridge, lasso etc.)
  - non-parametric estimation (kernel ridge regression, basis pursuit etc.)
- Complexity of optimization: essential for understanding statistical complexity.

Convex Optimization setup

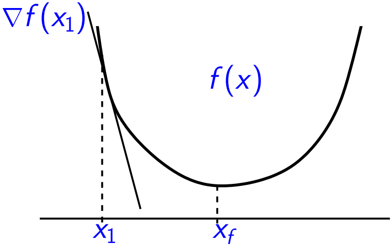
- **Optimization Problem:** $\min_{x \in \mathbb{S}} f(x) = f(x_f)$.
- $\mathbb{S}$ is a convex, compact set in $\mathbb{R}^d$.
- f is an (unknown) function picked from a class $\mathcal{F}$.
- We assume $\mathcal{F}$ is some subset of all convex functions.
- Algorithm told $\mathbb{S}$ and $\mathcal{F}$.
- **Goal**: Find x such that $f(x) - f(x_f) \leq \epsilon$.





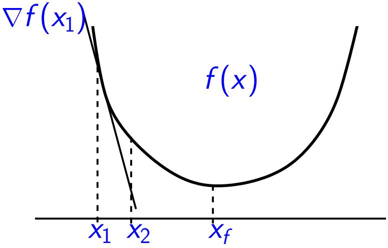
First-order oracle model of complexity

- Work within oracle complexity model [NY'83].
- Optimization proceeds in rounds $t = 1, \dots, T$.
- At time t, an algorithm $\mathcal{M}$ proposes x_t as its guess for x_f.
- Oracle returns $(f(x_t), \nabla f(x_t))$.



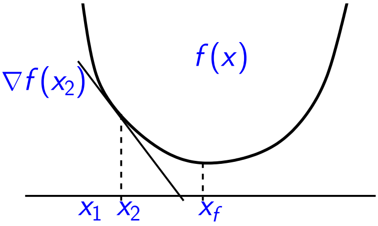
First-order oracle model of complexity

- Work within oracle complexity model [NY'83].
- Optimization proceeds in rounds $t = 1, \ldots, T$.
- At time t, an algorithm $\mathcal{M}$ proposes x_t as its guess for x_f.
- Oracle returns $(f(x_t), \nabla f(x_t))$.



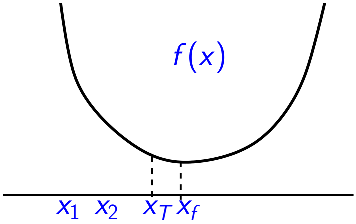
First-order oracle model of complexity

- Work within oracle complexity model [NY'83].
- Optimization proceeds in rounds $t = 1, \ldots, T$.
- At time t, an algorithm $\mathcal{M}$ proposes x_t as its guess for x_f.
- Oracle returns $(f(x_t), \nabla f(x_t))$.



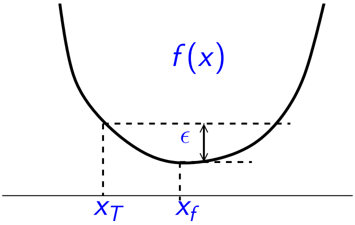
First-order oracle model of complexity

- Work within oracle complexity model [NY'83].
- Optimization proceeds in rounds $t = 1, \ldots, T$.
- At time t, an algorithm $\mathcal{M}$ proposes x_t as its guess for x_f.
- Oracle returns $(f(x_t), \nabla f(x_t))$.
- Algorithms such as gradient descent, ellipsoid method, quasi-Newton methods etc.



Oracle model contd.

- **Optimization error:** $\epsilon_T(\mathcal{M}, f) = f(x_T) - f(x_f)$.
- **Oracle Complexity:**
  Smallest $T(\epsilon, \mathcal{M}, f)$ such that $f(x_T) - f(x_f) \leq \epsilon$.
- **Minimax Complexity:**
  $$\underbrace{\inf_{\mathcal{M}}}_{\textit{Best algorithm}} \quad \underbrace{\sup_{f \in \mathcal{F}}}_{\textit{worst function}} \quad T(\epsilon, \mathcal{M}, f).$$
- Equivalently, for a fixed T study $\inf_{\mathcal{M}} \ \sup_{f \in \mathcal{F}} \ \epsilon_T(\mathcal{M}, f)$.



Stochastic first-order oracle model of complexity

- At time t, an algorithm $\mathcal{M}$ proposes x_t as its guess for x_f.
- Oracle returns $(\widehat{f}(x_t), \widehat{z}(x_t))$.
- Unbiased function values: $\mathbb{E}\widehat{f}(x_t) = f(x_t)$.
- Unbiased gradients: $\mathbb{E}\widehat{z}(x_t) = \nabla f(x_t)$.
- Bounded variance: $\mathbb{E}\|\widehat{z}(x_t)\|_1^2 \leq \sigma^2$.
- Algorithms such as stochastic gradient descent, mirror descent, stocastic approximation procedures etc.

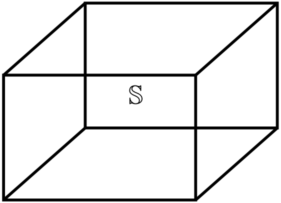
Stochastic Oracle model contd.

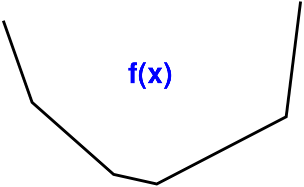
- **Optimization error:** $\epsilon_T(\mathcal{M}, f) = \mathbb{E}f(x_T) - f(x_f)$.
- **Oracle Complexity:**
  Smallest $T(\epsilon, \mathcal{M}, f)$ such that $\mathbb{E}f(x_T) - f(x_f) \leq \epsilon$.
- **Minimax Complexity:**
  $$\underbrace{\inf_{\mathcal{M}}}_{\textit{Best algorithm}} \quad \underbrace{\sup_{f \in \mathcal{F}}}_{\textit{worst function}} \quad T(\epsilon, \mathcal{M}, f).$$
- Equivalently, for a fixed T study $\inf_{\mathcal{M}} \ \sup_{f \in \mathcal{F}} \ \mathbb{E}\ \epsilon_T(\mathcal{M}, f)$.

Complexity lower bounds for convex, Lipschitz functions

- Let $\mathcal{F}_{\mathsf{CV}}(\mathbb{S}, L)$ be the class of all convex functions $f : \mathbb{S} \mapsto \mathbb{R}$ such that

  $$|f(x) - f(y)| \leq L\|x - y\|_\infty, \text{ equivalently } \|\nabla f(x)\|_1 \leq L \quad \forall x, y \in \mathbb{S}.$$





Complexity lower bounds for convex, Lipschitz functions

- Let $\mathcal{F}_{\mathsf{CV}}(\mathbb{S}, L)$ be the class of all convex functions $f \ : \ \mathbb{S} \mapsto \mathbb{R}$ such that

$$|f(x)-f(y)| \leq L\|x-y\|_\infty, \text{ equivalently } \|\nabla f(x)\|_1 \leq L \quad \forall x, y \in \mathbb{S}.$$

Theorem

No method can produce an ϵ-approximate optimizer for every convex, Lipschitz function in fewer than $\mathcal{O}\left(\frac{rL^2 d}{\epsilon^2}\right)$ queries.

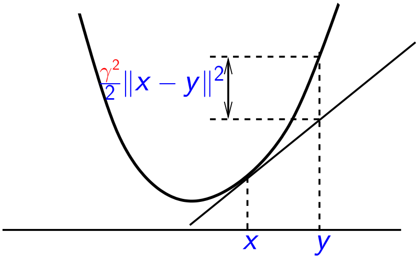
- r is the radius of the largest ℓ_∞ ball contained in $\mathbb{S}$.
- Lower bound achieved by stochastic gradient descent.

Complexity lower bounds for strongly convex functions

- Let $\mathcal{F}_{\mathsf{SCV}}(\mathbb{S}, L, \gamma)$ be the class of all functions $f \in \mathcal{F}_{\mathsf{CV}}(\mathbb{S}, L)$ such that

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\gamma^2}{2}\|x - y\|_2^2.$$

- Functions with lower bounded curvature, widely studied in optimization.



Complexity lower bounds for strongly convex functions

- Let $\mathcal{F}_{\text{SCV}}(\mathbb{S}, L, \gamma)$ be the class of all functions $f \in \mathcal{F}_{\text{CV}}(\mathbb{S}, L)$ such that

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\gamma^2}{2} \|x - y\|_2^2.$$

Theorem

No method can produce an ϵ-approximate optimizer for every strongly convex, Lipschitz function in fewer than $\mathcal{O}\left(\frac{L^2}{\gamma^2 \epsilon}\right)$ queries.

- Lower bound attained by stochastic gradient descent.

Lower bounds for convex functions with sparse optima

- Let $\mathcal{F}_{\mathsf{sp}}(\mathbb{S}, L, k)$ be the class of all convex functions f such that x_f has at most k non-zero entries and

  $|f(x) - f(y)| \leq L\|x - y\|_1,$ equivalently $\|\nabla f(x)\|_\infty \leq L \quad \forall x, y \in \mathbb{S}.$

Lower bounds for convex functions with sparse optima

- Let $\mathcal{F}_{\mathsf{sp}}(\mathbb{S}, L, k)$ be the class of all convex functions f such that x_f has at most k non-zero entries and

$$|f(x)-f(y)| \leq L\|x-y\|_1, \text{ equivalently } \|\nabla f(x)\|_\infty \leq L \quad \forall x, y \in \mathbb{S}.$$

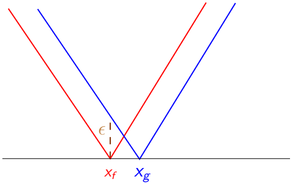
Theorem

No method can produce an ϵ-approximate optimizer for every function in $\mathcal{F}_{sp}(\mathbb{S}, L, k)$ in fewer than $\mathcal{O}\left(\frac{L^2 k^2 \log \frac{d}{k}}{\epsilon^2}\right)$ queries.

- Much milder logarithmic dependence on dimension d.
- Lower bound attained by the method of mirror descent ([NY'83], [BT'03]).

Proof intuition

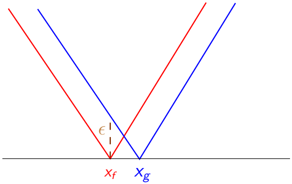
- Proofs based on identifying a hard subset of functions.
- Lower bound based on optimizing every function in hard subset well.
- Want a hard subset of functions with
  - Any two functions *far enough* so no algorithm can get lucky.



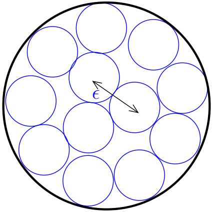
$$g(x_f) - g(x_g) \leq \epsilon$$

Proof intuition

- Proofs based on identifying a hard subset of functions.
- Lower bound based on optimizing every function in hard subset well.
- Want a hard subset of functions with
  - Any two functions *far enough* so no algorithm can get lucky.
  - *Large enough* number of functions to force a lot of queries.



$g(x_f) - g(x_g) \leq \epsilon$



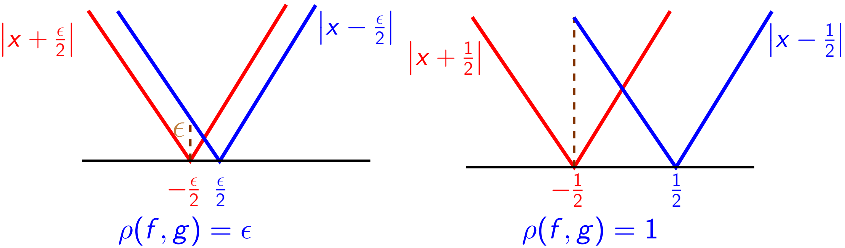
Large packing set of functions.

The ρ semimetric

Definition

$$\rho(f,g) = \inf_{x \in \mathbb{S}} \Big[f(x) + g(x) \Big] - f(x_f) - g(x_g).$$

- $\rho(f,g) \geq 0$, doesn't obey triangle inequality.
- $\rho(f,g) = 0$ if and only if $x_f = x_g$.
- Measures how different f and g are for optimization.



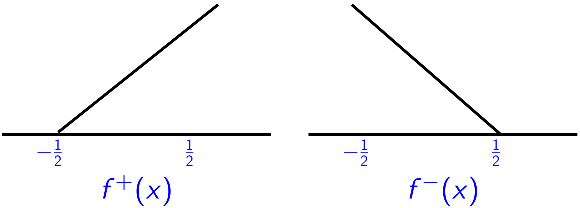
Proof Outline

- Design a ρ-separated subclass of $\mathcal{F}$.
- Algorithm needs to identify oracle's f.
- Stochastic first-order oracle corrupts $(f(x_t), \nabla f(x_t))$ with noise.
- Identifying f equivalent to estimating f from noisy samples.
- Use sample complexity results for the estimation problem to lower bound number of queries.

A ρ-separated subclass of $\mathcal{F}_{\mathsf{CV}}$

- Let $\mathbb{S} = [-1/2, 1/2]^d$
- Define $f_i^+(x) = |1/2 + x(i)|$, $f_i^-(x) = |1/2 - x(i)|$.
- For $\alpha \in \{-1, 1\}^d$ define

$$g_\alpha(x) = \frac{1}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f_i^+(x) + \left(\frac{1}{2} - \alpha_i \delta\right) f_i^-(x)$$



Conclusions

- Obtain tight minimax lower bounds on oracle complexity for stochastic convex optimization.
- Clean information theoretic proofs through reduction to a parameter estimation problem.
- Identify the ρ semimetric natural for optimization.
- Bounds show optimality of stochastic gradient descent and stochastic mirror descent for certain problems.

Thank You