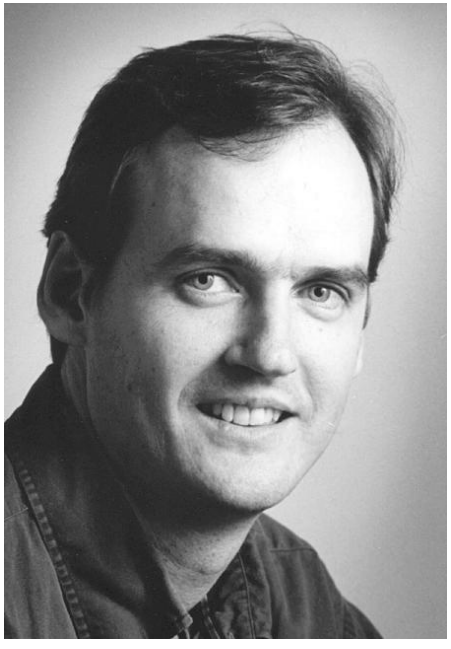
Information-Theoretic Lower Bounds on the Oracle Complexity of Sparse Convex Optimization

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



Sparse Convex Optimization

- High-dimensional convex optimization arises in computational biology, collaborative filtering, computational astronomy etc.
- Sparsity preferred/enforced for statistical and computational reasons.
- High-dimensional linear regression: $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$.
  - Lasso estimator:

$$\hat{\theta}_n \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n}(y_i - x_i^T \theta)^2 + \lambda \|\theta\|_1 \right\}.$$

  - ℓ_1 norm penalty enforces sparsity.
  - Sparsity essential to show statistical consistency when $d \gg n$.

Complexity of sparse convex optimization

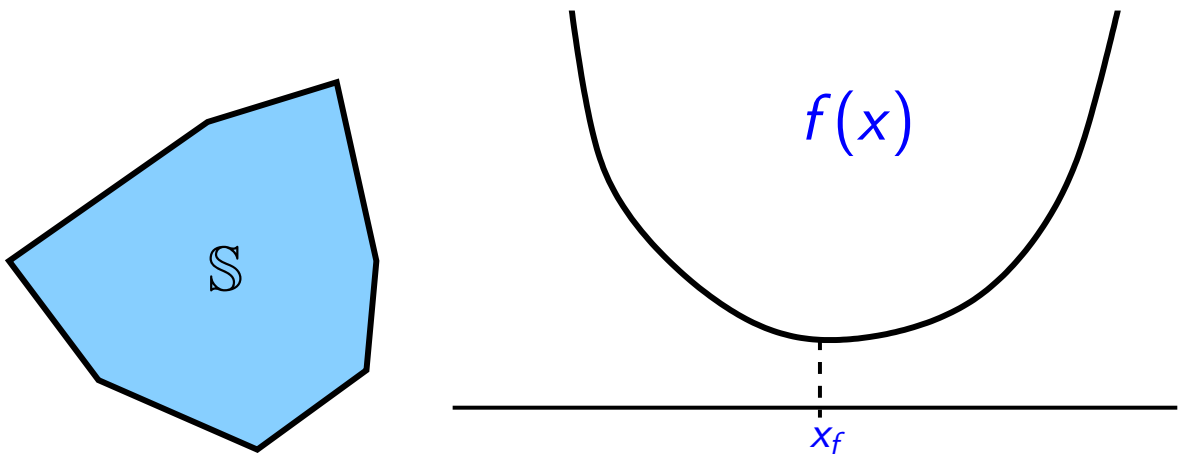
- Specialized algorithms:
  - Mirror descent (Nemirovski & Yudin, 1983; Beck & Teboulle, 2003).
  - Forward-backward splitting (Duchi & Singer, 2009).
  - Regularized dual averaging (Xiao, 2009).
  - Accelerated variants, extensions to mirror descent etc.
- Upper bounds on computational complexities for specific methods well-studied.

Complexity of sparse convex optimization

- Specialized algorithms:
  - Mirror descent (Nemirovski & Yudin, 1983; Beck & Teboulle, 2003).
  - Forward-backward splitting (Duchi & Singer, 2009).
  - Regularized dual averaging (Xiao, 2009).
  - Accelerated variants, extensions to mirror descent etc.
- Upper bounds on computational complexities for specific methods well-studied.
- No known results on fundamental hardness of sparse convex optimization.
- Minimum computation needed by *any* algorithm to solve a convex optimization problem with sparse optimum.

Convex Optimization setup

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



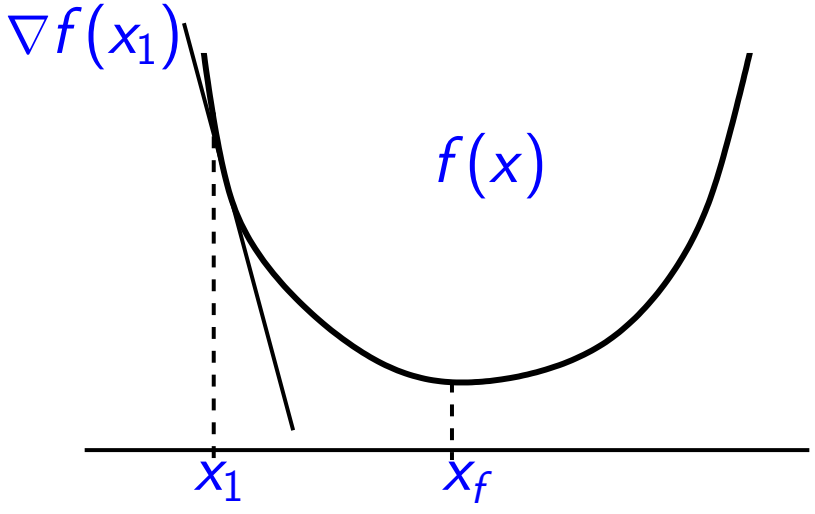
Stochastic first-order oracle model of complexity

- 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 $(\widehat{f}(x_t), \widehat{z}(x_t))$.

Stochastic 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 $(\widehat{f}(x_t), \widehat{z}(x_t))$.

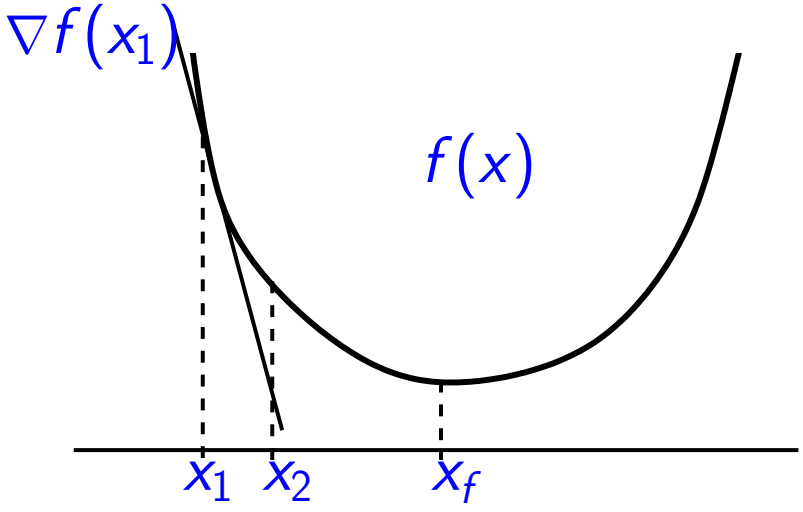
$$\underbrace{\mathbb{E}[\widehat{f}(x_t)] = f(x_t)}_{\text{unbiased function values}} \quad , \quad \underbrace{\mathbb{E}[\widehat{z}(x_t)] \in \partial f(x_t)}_{\text{unbiased subgradients}} \quad , \text{ and } \quad \underbrace{\mathbb{E}\left[\|\widehat{z}(x_t)\|_p^2\right] \leq \sigma^2}_{\text{bounded noise}} .$$



Stochastic 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 $(\widehat{f}(x_t), \widehat{z}(x_t))$.

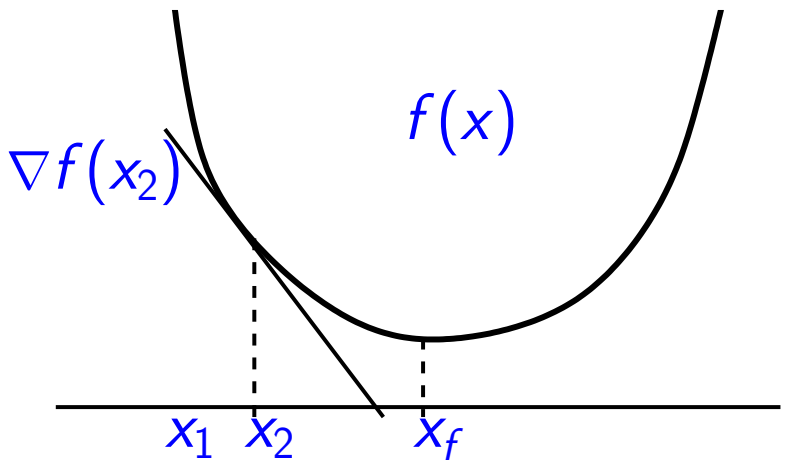
$$\underbrace{\mathbb{E}[\widehat{f}(x_t)] = f(x_t)}_{\text{unbiased function values}} \quad , \quad \underbrace{\mathbb{E}[\widehat{z}(x_t)] \in \partial f(x_t)}_{\text{unbiased subgradients}} \quad , \text{ and } \quad \underbrace{\mathbb{E}\left[\|\widehat{z}(x_t)\|_p^2\right] \leq \sigma^2}_{\text{bounded noise}} .$$



Stochastic first-order oracle model of complexity

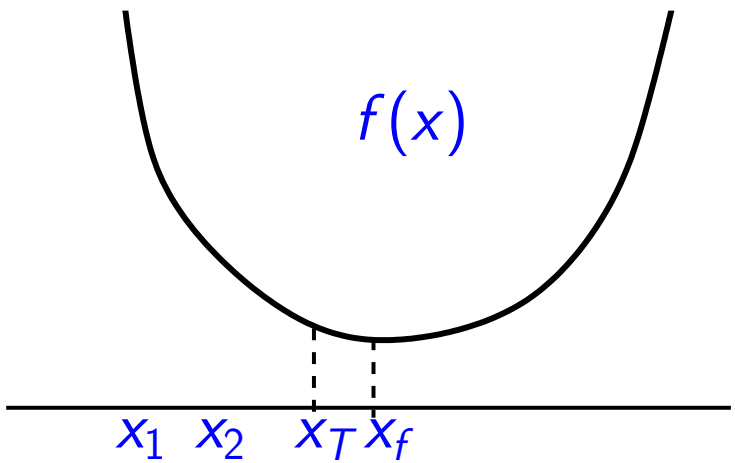
- Work within oracle complexity model [NY'83].
- Optimization proceeds in rounds $t = 1, \ldots, T$.
- At time t, an algorithm $\mathcal{M}$ proposes x_t as its guess for x_f.
- Oracle returns $(\widehat{f}(x_t), \widehat{z}(x_t))$.

$$\underbrace{\mathbb{E}[\widehat{f}(x_t)] = f(x_t)}_{\text{unbiased function values}} \quad , \quad \underbrace{\mathbb{E}[\widehat{z}(x_t)] \in \partial f(x_t)}_{\text{unbiased subgradients}} \quad , \text{ and } \quad \underbrace{\mathbb{E}\left[\|\widehat{z}(x_t)\|_p^2\right] \leq \sigma^2}_{\text{bounded noise}} .$$



Stochastic first-order oracle model of complexity

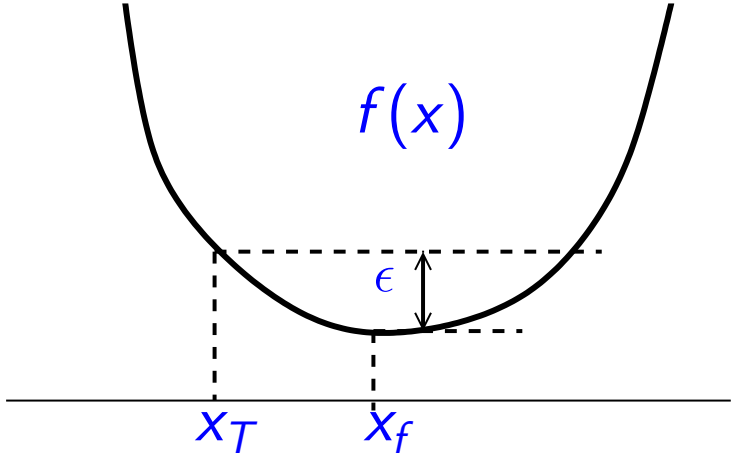
- Work within oracle complexity model [NY'83].
- Optimization proceeds in rounds $t = 1, \ldots, T$.
- At time t, an algorithm $\mathcal{M}$ proposes x_t as its guess for x_f.
- Oracle returns $(\widehat{f}(x_t), \widehat{z}(x_t))$.
- Algorithms such as stochastic gradient descent, mirror descent, FOBOS,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



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:**

$$\underbrace{\inf_{\mathcal{M}}}_{\textit{Best algorithm}} \quad \underbrace{\sup_{f \in \mathcal{F}}}_{\textit{worst function}} \quad T(\epsilon, \mathcal{M}, f).$$



Previous work on oracle complexity

- Minimax Complexity: $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ for convex, Lipschitz functions (Nemirovski & Yudin, 1983).
- Minimax Complexity: $\mathcal{O}\left(\frac{d^2}{\epsilon^2}\right)$ for convex, Lipschitz functions (ABRW, 2009).
- Prohibitive complexity in high-dimensions.
- Can we do better for sparse problems?

Lower bound 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 bound for convex functions with sparse optima

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

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

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.

- Mild logarithmic dependence on dimension d.
- Lower bound attained by the method of mirror descent.

Proof Outline

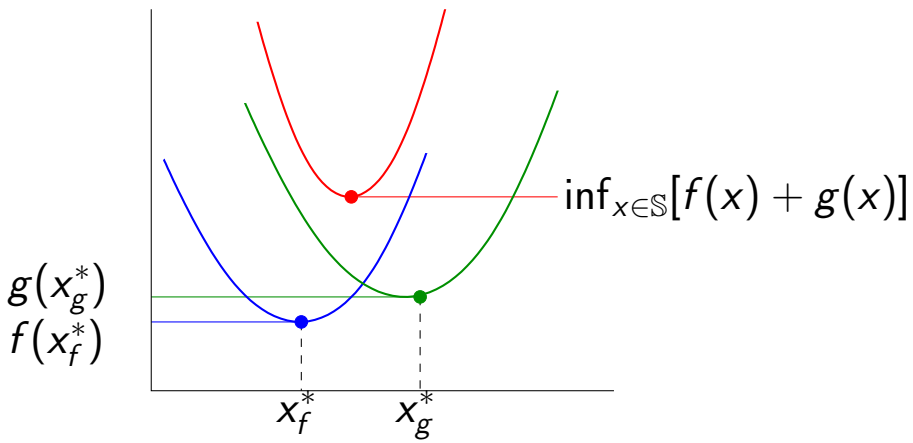
- Proofs based on identifying a hard subset of functions in $\mathcal{F}$.
- A large subset of *well-separated* functions.
- Reduction from optimization to function identification.
- Function estimated from noisy samples provided by oracle.
- Lower bound number of samples needed to solve the estimation problem.

Function separation in the ρ semimetric

Definition

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

- $\rho(f,g) > 0$ unless $x_f = x_g$.
- Measures how different f and g are for optimization.

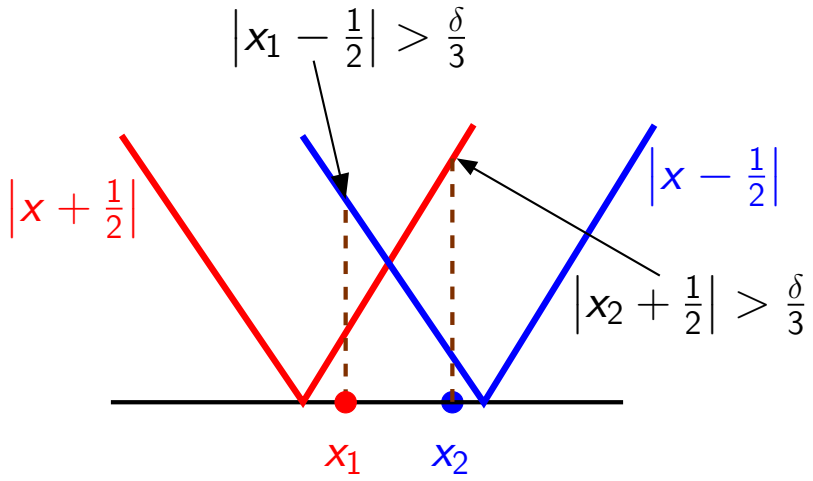


From optimization to function identification

Lemma

Let $\inf_{f \neq g \in \mathcal{F}} \rho(f, g) = \delta$. *Then for any* $x \in \mathbb{S}$, *there is at most one function* $f \in \mathcal{F}$ *such that*

$$f(x) - f(x_f) \leq \frac{\delta}{3}.$$



Any point is sub-optimal on either f or g.

- Can identify oracle's function based on algorithm's output.

Designing the function class

- Consider functions indexed by $\alpha \in \{-1, 0, +1\}^d$.
- For any α with at most k non-zero entries, define:

$$g_\alpha(x) = \sum_{i=1}^{d} \underbrace{\left(\frac{1}{2} + \alpha_i \delta\right)}_{\text{coin bias}} \underbrace{f_i^+(x)}_{\text{base function}} + \left(\frac{1}{2} - \alpha_i \delta\right) \underbrace{f_i^-(x)}_{\text{base function}} .$$

- Base functions f_i^+, f_i^- depend only on i_{th} coordinate.
- Sparsity of α ensures sparsity of optimum.
- ρ-separation ensured by picking a packing set of α's.

Designing a stochastic first-order oracle

- Associate a coin with each coordinate i.
- i_{th} coin has bias $\frac{1}{2} + \alpha_i \delta$.
- Oracle tosses each coin, observes outcomes $b_{i,t}$.
- Returns value and gradient of the function:

$$\hat{g}_\alpha(x) = \sum_{i=1}^{d} b_{i,t} f_i^+(x) + (1 - b_{i,t}) f_i^-(x).$$

- Satisfies unbiasedness, bounded noise.
- Relates optimization to estimating bias of coins.

Conclusions

- Obtain tight minimax lower bounds on oracle complexity for sparse 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 some popularly used methods in oracle complexity model.

Thank You