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

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, Martin J. Wainwright

Presenter: Wei-Cheng Lee
Advisor: I-Hsiang Wang

Group Meeting Presentation

Outline

Contribution of the Paper

- Provides a method to relate problems of stochastic convex optimization to statistical parameter estimation.

Contribution of the Paper

- Provides a method to relate problems of stochastic convex optimization to statistical parameter estimation.
- Use new method to extend classical analysis results on minimax lower bounds and relax the constraints.

Contribution of the Paper

- Provides a method to relate problems of stochastic convex optimization to statistical parameter estimation.
- Use new method to extend classical analysis results on minimax lower bounds and relax the constraints.
- Provides the rst tight lower bound on the oracle complexity of sparse optimization.

Structure of the Paper

- Introduce oracle optimization model and corresponding terminology.

Structure of the Paper

- Introduce oracle optimization model and corresponding terminology.
- State lower bounds of minimax complexity under certain class of functions and oracles and list what algorithms can guarantee to attain the lower bound.

Structure of the Paper

- Introduce oracle optimization model and corresponding terminology.
- State lower bounds of minimax complexity under certain class of functions and oracles and list what algorithms can guarantee to attain the lower bound.
- Use a general proof technique to relate the optimization problems to parameter estimation.

Structure of the Paper

- Introduce oracle optimization model and corresponding terminology.
- State lower bounds of minimax complexity under certain class of functions and oracles and list what algorithms can guarantee to attain the lower bound.
- Use a general proof technique to relate the optimization problems to parameter estimation.
- Fit each theorem to the proof technique.

Outline

Convex Optimization in the Oracle Model

- Problem : Consider a convex f over a convex set $\mathbb{S} \subseteq \mathbb{R}^d$ which is feasible. We want to find

$$x_f^* \in argmin_{x \in \mathbb{S}} f(x)$$

- function class: $\mathbb{S} \subseteq \mathbb{R}^d$, $p \in [1, \infty]$ and convex $f : \mathbb{S} \to \mathbb{R}$.

Convex Optimization in the Oracle Model

- Problem : Consider a convex f over a convex set $\mathbb{S} \subseteq \mathbb{R}^d$ which is feasible. We want to find

$$x_f^* \in argmin_{x \in \mathbb{S}} f(x)$$

- function class: $\mathbb{S} \subseteq \mathbb{R}^d$, $p \in [1, \infty]$ and convex $f : \mathbb{S} \to \mathbb{R}$.
  1. $\mathcal{F}_{cv}(\mathbb{S}, L, p)$:

$$|f(x) - f(y)| \leq L||x - y||_q$$

     for all $x, y \in \mathbb{S}$ where $\frac{1}{q} = 1 - \frac{1}{p}$.

Convex Optimization in the Oracle Model

- Problem : Consider a convex f over a convex set $\mathbb{S} \subseteq \mathbb{R}^d$ which is feasible. We want to find

$$x_f^* \in argmin_{x \in \mathbb{S}} f(x)$$

- function class: $\mathbb{S} \subseteq \mathbb{R}^d$, $p \in [1, \infty]$ and convex $f : \mathbb{S} \to \mathbb{R}$.
  1. $\mathcal{F}_{cv}(\mathbb{S}, L, p)$:

     $$|f(x) - f(y)| \leq L||x - y||_q$$

     for all $x, y \in \mathbb{S}$ where $\frac{1}{q} = 1 - \frac{1}{p}$.
  2. $\mathcal{F}_{scv}(\mathbb{S}, p; L, \gamma) \subseteq \mathcal{F}_{cv}(\mathbb{S}, L, p)$, *where* $\frac{L}{r^2} \geq \frac{r}{4} d^{\frac{1}{p}}$:

     $$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \alpha(1-\alpha)\frac{\gamma^2}{2}||x - y||_2^2$$

     for all $x, y \in \mathbb{S}$.

Convex Optimization in the Oracle Model

- Problem : Consider a convex f over a convex set $\mathbb{S} \subseteq \mathbb{R}^d$ which is feasible. We want to find

$$x_f^* \in argmin_{x \in \mathbb{S}} f(x)$$

- function class: $\mathbb{S} \subseteq \mathbb{R}^d$, $p \in [1, \infty]$ and convex $f : \mathbb{S} \to \mathbb{R}$.
  1. $\mathcal{F}_{cv}(\mathbb{S}, L, p)$:
  $$|f(x) - f(y)| \leq L||x - y||_q$$
  for all $x, y \in \mathbb{S}$ where $\frac{1}{q} = 1 - \frac{1}{p}$.
  2. $\mathcal{F}_{scv}(\mathbb{S}, p; L, \gamma) \subseteq \mathcal{F}_{cv}(\mathbb{S}, L, p), where\ \frac{L}{r^2} \geq \frac{r}{4} d^{\frac{1}{p}}$:
  $$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) - \alpha(1 - \alpha)\frac{\gamma^2}{2}||x - y||_2^2$$
  for all $x, y \in \mathbb{S}$.
  3. $\mathcal{F}_{sp}(k; \mathbb{S}, L)$: $k \leq \left\lfloor \frac{d}{2} \right\rfloor$ satisfy L-Lipschitz in l_∞ norm and there exists some
  $$x_f^* \in argmin_{x \in \mathbb{S}} f(x) \quad \text{satisfying} ||x^*||_0 \leq k.$$

Oracle and Optimization Method

- Oracle is a function defined as $\phi : \mathbb{S} \times \mathcal{F} \rightarrow \mathcal{I}$ that answers query $x \in \mathbb{S}$ by returning $\phi(x) \in \mathcal{I}$. We constrain our oracles as first-order stochastic oracles satisfying $\phi(x, f) = (\hat{f}(x), \hat{z}(x))$ such that

$$\underbrace{\mathbb{E}[\hat{f}(x)] = f(x)}_{\textcolor{red}{unbiased\ \ function\ \ values}} \quad , \quad \underbrace{\mathbb{E}[\hat{z}(x)] \in \partial f(x)}_{\textcolor{red}{unbiased\ \ sub\ \ gradients}} \quad and \quad \underbrace{\mathbb{E}[||\hat{z}(x)||_p^2] \leq \sigma^2}_{\textcolor{red}{bounded\ \ variance}}.$$

$\mathbb{O}_{p,\sigma}$ to denote the class of all previous oracles with (p, σ)

Oracle and Optimization Method

- Oracle is a function defined as $\phi : \mathbb{S} \times \mathcal{F} \to \mathcal{I}$ that answers query $x \in \mathbb{S}$ by returning $\phi(x) \in \mathcal{I}$. We constrain our oracles as first-order stochastic oracles satisfying $\phi(x, f) = (\hat{f}(x), \hat{z}(x))$ such that

$$\underbrace{\mathbb{E}[\hat{f}(x)] = f(x)}_{\textit{unbiased function values}} \quad , \quad \underbrace{\mathbb{E}[\hat{z}(x)] \in \partial f(x)}_{\textit{unbiased sub gradients}} \quad \textit{and} \quad \underbrace{\mathbb{E}[||\hat{z}(x)||_p^2] \leq \sigma^2}_{\textit{bounded variance}}.$$

  $\mathbb{O}_{p,\sigma}$ to denote the class of all previous oracles with (p, σ)
- Optimization method: For a given oracle ϕ, $\mathbb{M}_T$ denote the class of all optimization methods that make T queries according to procedure:
  1. the Method $\mathcal{M} \in \mathbb{M}_T$ queries oracle to reveal $\phi(x_t, f)$ for any iteration $t = 1, ..., T$.
  2. the Method $\mathcal{M} \in \mathbb{M}_T$ decides next step x_{t+1} using information $\{\phi(x_1, f), ..., \phi(x_t, f)\}$.

Errors

- For any $\mathcal{M} \in \mathbb{M}_T$ define optimization error on f after T steps as

$$\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi) := f(x_T) - \min_{x \in \mathbb{S}} = f(x_T) - f(x_f^*)$$

If the oracle is stochastic, x_T is random depending on the oracle, we thus consider

$$\mathbb{E}[\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi)]$$

Errors

- For any $\mathcal{M} \in \mathbb{M}_T$ define optimization error on f after T steps as

$$\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi) := f(x_T) - \min_{x \in \mathbb{S}} = f(x_T) - f(x_f^*)$$

  If the oracle is stochastic, x_T is random depending on the oracle, we thus consider

$$\mathbb{E}[\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi)]$$

- We are interesting in this minimax error which is defined as

$$\epsilon_T^*(\mathcal{F}, \mathbb{S}; \phi) := \underbrace{\inf_{\mathcal{M} \in \mathbb{M}_T}}_{best\ \ method} \underbrace{\sup_{f \in \mathcal{F}}}_{worst\ \ function} \mathbb{E}[\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi)]$$

  The lower bound of $\epsilon_T^*(\mathcal{F}, \mathbb{S}; \phi)$ can be interpreted as fundamental hardness of the problem under this oracle and function class.

Outline

Theorem 1

Theorem 1: Let $\mathbb{S} \subset \mathbb{R}^d$ be a convex set such that $\mathbb{S} \supseteq \mathbb{B}_\infty(r)$ for some $r > 0$. Then, there exists a universal constant $c_0 > 0$ such that the minimax oracle complexity over the class $\mathcal{F}_{\text{cv}}(\mathbb{S}, L, p)$ satisfies the following lower bounds.

(a) For $1 \leq p \leq 2$

$$\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon_T^*(\mathcal{F}_{\text{cv}}, \mathbb{S}; \phi) \geq \min\left\{ c_0 L\, r \sqrt{\frac{d}{T}}, \frac{Lr}{144} \right\}. \tag{9}$$

(b) For $p > 2$

$$\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon_T^*(\mathcal{F}_{\text{cv}}, \mathbb{S}; \phi) \geq \min\left\{ c_0 L\, r \frac{d^{1-\frac{1}{p}}}{\sqrt{T}}, \frac{L d^{1-1/p} r}{72} \right\}. \tag{10}$$

Theorem 2

Theorem 2: Let $\mathbb{S} = \mathbb{B}_\infty(r)$. Then, there exist universal constants $c_1, c_2 > 0$ such that the minimax oracle complexity over the class $\mathcal{F}_{\text{scv}}(\mathbb{S}, p; L, \gamma)$ satisfies the following lower bounds.

(a) For $p = 1$, the oracle complexity $\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon^*(\mathcal{F}_{\text{scv}}, \phi)$ is lower bounded by

$$\min\left\{c_1 \frac{L^2}{\gamma^2 T},\ c_2 L r \sqrt{\frac{d}{T}},\ \frac{L^2}{1152\gamma^2 d},\ \frac{Lr}{144}\right\}. \tag{12}$$

(b) For $p > 2$, the oracle complexity $\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon^*(\mathcal{F}_{\text{scv}}, \phi)$ is lower bounded by

$$\min\left\{c_1 \frac{L^2 d^{1-2/p}}{\gamma^2 T},\ c_2 \frac{L r d^{1-1/p}}{\sqrt{T}},\ \frac{L^2 d^{1-2/p}}{1152\gamma^2},\ \frac{L r d^{1-1/p}}{144}\right\}. \tag{13}$$

Theorem 3

Theorem 3: Let $\mathcal{F}_{\text{sp}}$ be the class of all convex functions that are L-Lipschitz with respect to the $||\cdot||_\infty$ norm and that have a k-sparse optimizer. Let $\mathbb{S} \subset \mathbb{R}^d$ be a convex set with $\mathbb{B}_\infty(r) \subseteq \mathbb{S}$. Then there exists a universal constant $c > 0$ such that for all $k \leq \lfloor \frac{d}{2} \rfloor$, we have

$$\sup_{\phi \in \mathbb{O}_{\infty,L}} \epsilon^*(\mathcal{F}_{\text{sp}}, \phi) \geq \min\left\{ cLr\sqrt{\frac{k^2 \log \frac{d}{k}}{T}}, \frac{Lkr}{432} \right\}. \quad (14)$$

Proof Method:Bottom up + Top down

- To better understand why the problem of convex optimization can relate to statistical parameter estimation, we not only state the general proof technique of theorems but also take a small function class of theorem 1 as a demonstrating example.
- function class of interests: $\mathbb{S} = \mathbb{B}_\infty(\frac{1}{2})$ for $p \in [1, 2]$

Proof Method:Bottom up + Top down

- To better understand why the problem of convex optimization can relate to statistical parameter estimation, we not only state the general proof technique of theorems but also take a small function class of theorem 1 as a demonstrating example.
- function class of interests: $\mathbb{S} = \mathbb{B}_\infty(\frac{1}{2})$ for $p \in [1, 2]$
- Proof steps:
  1. Construct a difficult enough subclass of functions.
  2. Use a set of chosen function class to divide $\mathbb{S}$ to show that optimizing well is equivalent to function identification.
  3. Give certain oracle and relate it to coin tossing.
  4. Using Fano's inequality to bridge coin-tossing and stochastic convex optimization.

Constructing a Difficult Enough Subclass of Functions-1

- Construct a subclass of $\mathcal{G} \subseteq \mathcal{F}$ that we use to derive the lower bounds. Any subclass is parametrized by $\mathcal{V} = \{\alpha^1, ..., \alpha^M\} \subseteq \{-1, +1\}^d$ be a subset of vertices of the hypercube such that

$$\triangle H(\alpha^j, \alpha^k) \geq \frac{d}{4}, \quad \textit{for all } j \neq k$$

- This $\mathcal{V}$ is called a $\frac{d}{4}$-packing in the Hamming norm. We are guaranteed to construct such a set with $|\mathcal{V}| \geq (2/\sqrt{e})^{\frac{d}{2}}$.

Constructing a Difficult Enough Subclass of Functions-1

- Construct a subclass of $\mathcal{G} \subseteq \mathcal{F}$ that we use to derive the lower bounds. Any subclass is parametrized by $\mathcal{V} = \{\alpha^1, ..., \alpha^M\} \subseteq \{-1, +1\}^d$ be a subset of vertices of the hypercube such that

$$\triangle H(\alpha^j, \alpha^k) \geq \frac{d}{4}, \quad \textit{for all } j \neq k$$

- This $\mathcal{V}$ is called a $\frac{d}{4}$-packing in the Hamming norm. We are guaranteed to construct such a set with $|\mathcal{V}| \geq (2/\sqrt{e})^{\frac{d}{2}}$.
- Let $\mathcal{G}_{base} = \{f_i^+, f_i^-, i = 1, ..., d\}$ denote some $2d$ functions carefully chosen on the $\mathbb{S}$. For a giving tolerance $\delta \in (0, \frac{1}{4}]$, we define for each $\alpha \in \mathcal{V}$ the function $x \mapsto g_\alpha(x)$ as

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

where c is used to force each $g_\alpha(x)$ lies in $\mathcal{F}$.

Constructing a Difficult Enough Subclass of Functions-2

- We now focus ourselves on function class $\mathcal{G}(\delta) := \{g_\alpha, \alpha \in \mathcal{V}\}$. In the proof, we need to carefully chose base functions to ensure that $\mathcal{G} \subseteq \mathcal{F}$.
- On $\mathbb{S} = \mathbb{B}_\infty(\frac{1}{2})$ for $p \in [1, 2]$:
- We specify base function $(f_i^+(x), f_i^-(x))$ as

$$f_i^+(x) := |x(i) + \frac{1}{2}|, \quad and \quad f_i^-(x) := |x(i) - \frac{1}{2}|$$

- One must make sure each $g_\alpha(x)$ lies in $\mathcal{F}$ and attain minimum in $\mathbb{S} = \mathbb{B}_\infty(\frac{1}{2})$.

Optimizing Well is Equivalent to Function Identification-1

- Our idea is that if a method can optimize over $\mathcal{G}(\delta)$ up to certain tolerance, then it is capable of identifying which function $g_\alpha \in \mathcal{G}(\delta)$.

Optimizing Well is Equivalent to Function Identification-1

- Our idea is that if a method can optimize over $\mathcal{G}(\delta)$ up to certain tolerance, then it is capable of identifying which function $g_\alpha \in \mathcal{G}(\delta)$.
- We define a discrepancy function $\rho(f, g)$ as

$$\rho(f, g) := inf_{x \in \mathcal{S}}[f(x) + g(x) - f(x_f^*) - g(x_g^*)].$$

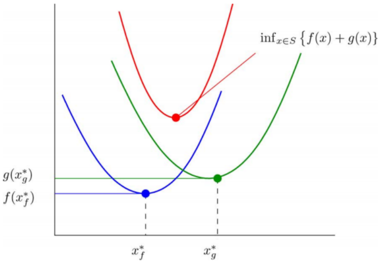


Fig. 1. Illustration of the discrepancy function $\rho(f, g)$. The functions f and g achieve their minimum values $f(x_f^*)$ and $g(x_g^*)$ at the points x_f^* and x_g^*, respectively.

Optimizing Well is Equivalent to Function Identification-2

- For the subclass $\mathcal{G}(\delta)$, we quantity how densely it is packed:

$$\psi(\mathcal{G}(\delta)) := \min_{\alpha \neq \beta \in \mathcal{V}} \rho(g_\alpha, g_\beta)$$

Optimizing Well is Equivalent to Function Identification-2

- For the subclass $\mathcal{G}(\delta)$, we quantity how densely it is packed:

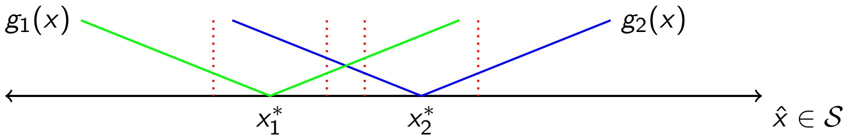
$$\psi(\mathcal{G}(\delta)) := \min_{\alpha \neq \beta \in \mathcal{V}} \rho(g_\alpha, g_\beta)$$

- We will have following two lemmas:

Lemma (1)

For any $\tilde{x} \in \mathbb{S}$, there can be at most one function $g_\alpha \in \mathcal{G}(\delta)$ such that

$$g_\alpha(\tilde{x}) - \inf_{x \in \mathbb{S}} g_\alpha(x) \leq \frac{\psi(\delta)}{3}$$



Optimizing Well is Equivalent to Function Identification-3

Lemma (2)

For a unknown function $g_\alpha^* \in \mathcal{G}(\delta)$*,if based on* $\phi(X_1^T; g_{\alpha^*}) := \{\phi(x_t; g_{\alpha^*}),\ \ t = 1, 2, ..., T\}$*, there exists a method* $\mathcal{M}_T$ *that achieves minimax error satisfying*

$$\mathbb{E}[\epsilon_T(\mathcal{M}, \mathcal{G}(\delta), \mathbb{S}, \phi)] \leq \frac{\psi(\delta)}{9}$$

Based on such method $\mathcal{T}$*, one can construct a hypothesis test* $\tilde{\alpha} : \phi(x_1^T; g_\alpha^*) \rightarrow \mathcal{V}$ *such that* $\max_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi[\tilde{\alpha}(\mathcal{M}_T) \neq \alpha^*] \leq \frac{1}{3}$

Optimizing Well is Equivalent to Function Identification-3

Lemma (2)

For a unknown function $g_\alpha^* \in \mathcal{G}(\delta)$*,if based on* $\phi(X_1^T; g_{\alpha^*}) := \{\phi(x_t; g_{\alpha^*}),\ \ t = 1, 2, ..., T\}$*, there exists a method* $\mathcal{M}_T$ *that achieves minimax error satisfying*

$$\mathbb{E}[\epsilon_T(\mathcal{M}, \mathcal{G}(\delta), \mathbb{S}, \phi)] \leq \frac{\psi(\delta)}{9}$$

Based on such method $\mathcal{T}$*, one can construct a hypothesis test* $\tilde{\alpha} : \phi(x_1^T; g_\alpha^*) \to \mathcal{V}$ *such that* $\max_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi[\tilde{\alpha}(\mathcal{M}_T) \neq \alpha^*] \leq \frac{1}{3}$

- The above two lemmas can be interpreted as relating how well a convex minimization algorithm do to function identification. When base function is fixed, we can evaluate $\frac{\psi(\delta)}{9}$ and set it to ϵ to get a hypothesis test probability upper-bound of $\mathbb{P}_\phi[\tilde{\alpha}(\mathcal{M}_T) \neq \alpha^*] \leq \frac{1}{3}$.

3) Oracle Answers and Coin Tosses: We now describe stochastic first order oracles ϕ for which the samples $\phi(x_1^T; g_\alpha)$ can be related to coin tosses. In particular, we associate a coin with each dimension $i \in \{1, 2, \ldots, d\}$, and consider the set of coin bias vectors lying in the set

$$\Theta(\delta) = \{(1/2 + \alpha_1 \delta, \ldots, 1/2 + \alpha_d \delta) \mid \alpha \in \mathcal{V}\}. \tag{22}$$

Given a particular function $g_\alpha \in \mathcal{G}(\delta)$—or equivalently, vertex $\alpha \in \mathcal{V}$—we consider two different types of stochastic first-order oracles ϕ, defined as follows.

Oracle A

Given a particular function $g_\alpha \in \mathcal{G}(\delta)$—or equivalently, vertex $\alpha \in \mathcal{V}$—we consider two different types of stochastic first-order oracles ϕ, defined as follows.

Oracle A: 1-dimensional unbiased gradients

(a) Pick an index $i \in \{1, \ldots, d\}$ uniformly at random.
(b) Draw $b_i \in \{0, 1\}$ according to a Bernoulli distribution with parameter $1/2 + \alpha_i \delta$.
(c) (c) For the given input $x \in \mathbb{S}$, return the value $\widehat{g}_{\alpha,A}(x)$ and a sub-gradient $\widehat{z}_{\alpha,A}(x) \in \partial\widehat{g}_{\alpha,A}(x)$ of the function

$$\widehat{g}_{\alpha,A} := c\big[b_i f_i^+ + (1 - b_i) f_i^-\big].$$

By construction, the function value and gradients returned by Oracle A are unbiased estimates of those of g_α. In particular, since each coordinate i is chosen with probability $1/d$, the expectation $\mathbb{E}\big[\widehat{g}_{\alpha,A}(x)\big]$ is given by

$$\frac{c}{d} \sum_{i=1}^{d} \big[\mathbb{E}[b_i] f_i^+(x) + \mathbb{E}[1 - b_i] f_i^-(x)\big] \;=\; g_\alpha(x)$$

with a similar relation for the gradient. Furthermore, as long as the base functions f_i^+ and f_i^- have gradients bounded by 1, we have $\mathbb{E}[\|\widehat{z}_{\alpha,A}(x)\|_p] \leq c$ for all $p \in [1, \infty]$.

Oracle B

Oracle B: d-dimensional unbiased gradients

(a) For $i = 1, \ldots, d$, draw $b_i \in \{0, 1\}$ according to a Bernoulli distribution with parameter $1/2 + \alpha_i \delta$.

(b) For the given input $x \in \mathbb{S}$, return the value $\widehat{g}_{\alpha,B}(x)$ and a sub-gradient $\widehat{z}_{\alpha,B}(x) \in \partial \widehat{g}_{\alpha,B}(x)$ of the function

$$\widehat{g}_{\alpha,B} := \frac{c}{d} \sum_{i=1}^{d} \left[b_i f_i^{+} + (1 - b_i) f_i^{-} \right].$$

As with Oracle A, this oracle returns unbiased estimates of the function values and gradients. We frequently work with functions f_i^{+}, f_i^{-} that depend only on the ith coordinate $x(i)$. In such cases, under the assumptions $|\frac{\partial f_i^{+}}{\partial x(i)}| \leq 1$ and $|\frac{\partial f_i^{-}}{\partial x(i)}| \leq 1$, we have

$$\begin{aligned} \|\widehat{z}_{\alpha,B}(x)\|_p^2 &= \frac{c^2}{d^2} \left(\sum_{i=1}^{d} \left| b_i \frac{\partial f_i^{+}(x)}{\partial x(i)} + (1 - b_i) \frac{\partial f_i^{-}(x)}{\partial x(i)} \right|^p \right)^{2/p} \\ &\leq c^2 d^{2/p - 2}. \end{aligned} \tag{23}$$

Lower Bounds on Coin-Tossing-1

4) Lower Bounds on Coin-Tossing: Finally, we use information-theoretic methods to lower bound the probability of correctly estimating the true parameter $\alpha^* \in \mathcal{V}$ in our model. At each round of either Oracle A or Oracle B, we can consider a set of d coin tosses, with an associated vector $\theta^* = (\frac{1}{2} + \alpha_1^* \delta, \ldots, \frac{1}{2} + \alpha_d^* \delta)$ of parameters. At any round, the output of Oracle A can (at most) reveal the instantiation $b_i \in \{0, 1\}$ of a randomly chosen index, whereas Oracle B can at most reveal the entire vector $(b_1, b_2, \ldots, b_d)$. Our goal is to lower bound the probability of estimating the true parameter α^*, based on a sequence of length T. As noted previously in remarks following Theorem 1, this part of our proof exploits classical techniques from statistical minimax theory, including the use of Fano's inequality (see, e.g., [9]–[12]) and Le Cam's bound (see, e.g., [12] and [20]).

Lower Bounds on Coin-Tossing-2

Lemma 3: Suppose that the Bernoulli parameter vector α^* is chosen uniformly at random from the packing set $\mathcal{V}$, and suppose that the outcome of $\ell \leq d$ coins chosen uniformly at random is revealed at each round $t = 1, \ldots, T$. Then for any $\delta \in (0, 1/4]$, any hypothesis test $\widehat{\alpha}$ satisfies

$$\mathbb{P}[\widehat{\alpha} \neq \alpha^*] \geq 1 - \frac{16\ell T\delta^2 + \log 2}{\frac{d}{2}\log(2/\sqrt{e})} \tag{24}$$

where the probability is taken over both randomness in the oracle and the choice of α^*.

Note that we will apply the lower bound (24) with $\ell = 1$ in the case of Oracle A, and $\ell = d$ in the case of Oracle B.

Proof: For each time $t = 1, 2, \ldots, T$, let U_t denote the randomly chosen subset of size ℓ, $X_{t,i}$ be the outcome of oracle's coin toss at time t for coordinate i and let $Y_t \in \{-1, 0, 1\}^d$ be a random vector with entries

$$Y_{t,i} = \begin{cases} X_{t,i}, & \text{if } i \in U_t, \text{ and} \\ -1 & \text{if } i \notin U_t. \end{cases}$$

By Fano's inequality [19], we have the lower bound

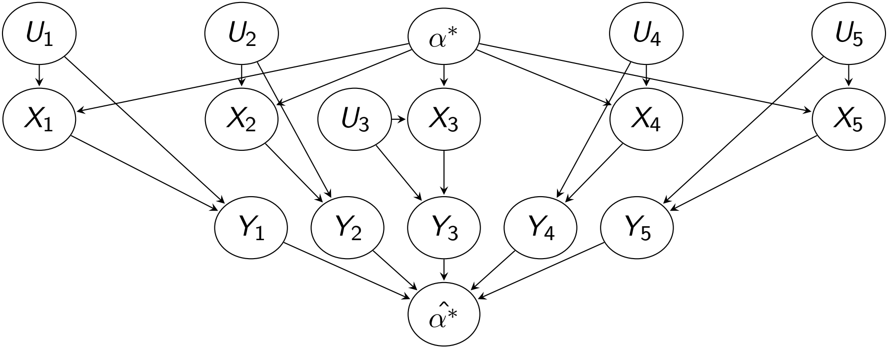
$$\mathbb{P}[\widehat{\alpha} \neq \alpha^*] \geq 1 - \frac{I(\{(U_t, Y_t\}_{t=1}^T; \alpha^*) + \log 2}{\log |\mathcal{V}|}$$

Fano's Inequality and Graphical model

Theorem (Fano's Inequality)

If X is uniformly chosen at random on space $\mathcal{X}$ then for any Markov Chain $X \to Y \to \hat{X}$ (Given Y, X and $\hat{X}$ are conditionally independent) we have

$$P(X \neq \hat{X}) \geq 1 - \frac{I(X : Y) + \log 2}{\log |X|}$$



Conclusion

- We show how to relate problems of stochastic convex optimization to statistical parameter estimation.

Conclusion

- We show how to relate problems of stochastic convex optimization to statistical parameter estimation.
- To prove lower bound of different function class, we need to carefully chose different base functions.

Conclusion

- We show how to relate problems of stochastic convex optimization to statistical parameter estimation.
- To prove lower bound of different function class, we need to carefully chose different base functions.
- Some of the lower bounds can be achieved by certain algorithms (e.g. SGD, Mirror descent) which characterizes how well those algorithms are.