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

Paper by Agarwal et al. Presented by Raphael A. Meyer

NYU Tandon

Overview

NYU

Introduction & The Results

Stochastic Convex Optimization

- Really common tool in the real world

Stochastic Convex Optimization

- Really common tool in the real world
- Stochastic methods run faster than deterministic

Stochastic Convex Optimization

- Really common tool in the real world
- Stochastic methods run faster than deterministic
- Theoretical results depends on the shape of the objective

NYU

Stochastic Convex Optimization

- Really common tool in the real world
- Stochastic methods run faster than deterministic
- Theoretical results depends on the shape of the objective
  - Lipschitz

Stochastic Convex Optimization

- Really common tool in the real world
- Stochastic methods run faster than deterministic
- Theoretical results depends on the shape of the objective
  - Lipschitz
  - Strong convexity

Stochastic Convex Optimization

- Really common tool in the real world
- Stochastic methods run faster than deterministic
- Theoretical results depends on the shape of the objective
  - Lipschitz
  - Strong convexity
  - Strongly smooth

NYU

Stochastic Convex Optimization

- Really common tool in the real world
- Stochastic methods run faster than deterministic
- Theoretical results depends on the shape of the objective
  - Lipschitz
  - Strong convexity
  - Strongly smooth
- Are our algorithms optimal for these settings?

NYU

Stochastic Convex Optimization

- Really common tool in the real world
- Stochastic methods run faster than deterministic
- Theoretical results depends on the shape of the objective
  - Lipschitz
  - Strong convexity
  - Strongly smooth
- Are our algorithms optimal for these settings?
  - Yes. They are.

NYU

Stochastic Convex Optimization

- Really common tool in the real world
- Stochastic methods run faster than deterministic
- Theoretical results depends on the shape of the objective
  - Lipschitz
  - Strong convexity
  - Strongly smooth
- Are our algorithms optimal for these settings?
  - Yes. They are.
  - There exists a Lipschitz/Strong Convex objective function and stochastic gradient such that $O(\frac{1}{\sqrt{T}})/O(\frac{1}{T})$ iterations are required.

NYU

Stochastic First Order Optimization

Given:

- Ability to compute $g(\mathbf{x})$

Stochastic First Order Optimization

Given:

- Ability to compute $g(\mathbf{x})$
- Ability to randomly estimate $\mathbf{z}(\mathbf{x}) \approx \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})$

NYU

Stochastic First Order Optimization

Given:

- Ability to compute $g(\mathbf{x})$
- Ability to randomly estimate $\mathbf{z}(\mathbf{x}) \approx \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})$
- Convex domain $\mathbb{S}$

Stochastic First Order Optimization

Given:

- Ability to compute $g(\mathbf{x})$
- Ability to randomly estimate $\mathbf{z}(\mathbf{x}) \approx \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})$
- Convex domain $\mathbb{S}$
- Maximum iterations T

Stochastic First Order Optimization

Given:

- Ability to compute $g(\mathbf{x})$
- Ability to randomly estimate $\mathbf{z}(\mathbf{x}) \approx \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})$
- Convex domain $\mathbb{S}$
- Maximum iterations T

Stochastic First Order Optimization

Given:

- Ability to compute $g(\mathbf{x})$
- Ability to randomly estimate $\mathbf{z}(\mathbf{x}) \approx \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})$
- Convex domain $\mathbb{S}$
- Maximum iterations T

Solve the problem:

$$\min_{\mathbf{x} \in \mathbb{S}} g(\mathbf{x})$$

By computing $g(\mathbf{x})$ and $\mathbf{z}(\mathbf{x})$ at most T times

NYU

Stochastic First Order Optimization

Given:

- Ability to compute $g(\mathbf{x})$
- Ability to randomly estimate $\mathbf{z}(\mathbf{x}) \approx \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})$
- Convex domain $\mathbb{S}$
- Maximum iterations T

Solve the problem:

$$\min_{\mathbf{x} \in \mathbb{S}} g(\mathbf{x})$$

By computing $g(\mathbf{x})$ and $\mathbf{z}(\mathbf{x})$ at most T times

- This "access to $g(\mathbf{x}_i)$ and $\mathbf{z}(\mathbf{x}_i)$" is called *Oracle Access*

NYU

Stochastic First Order Optimization

Given:

- Ability to compute $g(\mathbf{x})$
- Ability to randomly estimate $\mathbf{z}(\mathbf{x}) \approx \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})$
- Convex domain $\mathbb{S}$
- Maximum iterations T

Solve the problem:

$$\min_{\mathbf{x} \in \mathbb{S}} g(\mathbf{x})$$

By computing $g(\mathbf{x})$ and $\mathbf{z}(\mathbf{x})$ at most T times

- This "access to $g(\mathbf{x}_i)$ and $\mathbf{z}(\mathbf{x}_i)$" is called *Oracle Access*
- Minimizing Oracle queries is minimizing the rounds of gradient descent

NYU

Result 1: Lipschitz

- Let $\mathbb{S}$ be a convex domain

Result 1: Lipschitz

- Let $\mathbb{S}$ be a convex domain
- Let $\mathcal{F}_{cv}(\mathbb{S}, L)$ be the set of convex functions g defined on $\mathbb{S}$ such that

$$\|\frac{\partial g}{\partial \mathbf{x}}\|_2 \leq L$$

Result 1: Lipschitz

- Let $\mathbb{S}$ be a convex domain
- Let $\mathcal{F}_{cv}(\mathbb{S}, L)$ be the set of convex functions g defined on $\mathbb{S}$ such that
$$\|\frac{\partial g}{\partial \mathbf{x}}\|_2 \leq L$$
- The oracle $\phi(\mathbf{x}, g)$ maps to a random pair $(\hat{g}(\mathbf{x}), \hat{\mathbf{z}}(\mathbf{x}))$ such that
$$\mathbb{E}[\hat{g}(\mathbf{x})] = g(\mathbf{x}) \qquad \mathbb{E}[\hat{\mathbf{z}}(\mathbf{x})] = \frac{\partial g}{\partial \mathbf{x}} \qquad \mathbb{E}[\|\hat{\mathbf{z}}(\mathbf{x})\|_2 \leq L]$$

NYU

Result 1: Lipschitz

- Let $\mathbb{S}$ be a convex domain
- Let $\mathcal{F}_{cv}(\mathbb{S}, L)$ be the set of convex functions g defined on $\mathbb{S}$ such that
$$\|\frac{\partial g}{\partial \mathbf{x}}\|_2 \leq L$$
- The oracle $\phi(\mathbf{x}, g)$ maps to a random pair $(\hat{g}(\mathbf{x}), \hat{\mathbf{z}}(\mathbf{x}))$ such that
$$\mathbb{E}[\hat{g}(\mathbf{x})] = g(\mathbf{x}) \qquad \mathbb{E}[\hat{\mathbf{z}}(\mathbf{x})] = \frac{\partial g}{\partial \mathbf{x}} \qquad \mathbb{E}[\|\hat{\mathbf{z}}(\mathbf{x})\|_2 \leq L]$$
- Let $\mathcal{M}_T \in \mathbb{M}_T$ denote any algorithm that makes T queries to $\phi(\mathbf{x}_i, g)$ and returns some $\mathbf{x}_T$

NYU

Result 1: Lipschitz

- Let $\mathbb{S}$ be a convex domain
- Let $\mathcal{F}_{cv}(\mathbb{S}, L)$ be the set of convex functions g defined on $\mathbb{S}$ such that
$$\|\frac{\partial g}{\partial \mathbf{x}}\|_2 \leq L$$
- The oracle $\phi(\mathbf{x}, g)$ maps to a random pair $(\hat{g}(\mathbf{x}), \hat{\mathbf{z}}(\mathbf{x}))$ such that
$$\mathbb{E}[\hat{g}(\mathbf{x})] = g(\mathbf{x}) \qquad \mathbb{E}[\hat{\mathbf{z}}(\mathbf{x})] = \frac{\partial g}{\partial \mathbf{x}} \qquad \mathbb{E}[\|\hat{\mathbf{z}}(\mathbf{x})\|_2 \leq L]$$
- Let $\mathcal{M}_T \in \mathbb{M}_T$ denote any algorithm that makes T queries to $\phi(\mathbf{x}_i, g)$ and returns some $\mathbf{x}_T$
- Let $\varepsilon_T(\mathcal{M}_T, g, \phi) = g(\mathbf{x}_T) - \inf_{\mathbf{x}\in\mathbb{S}} g(\mathbf{x})$ be the error after T iterations using ϕ

NYU

Result 1: Lipschitz

Theorem 1

- When $\mathbb{S}$ contains the ℓ_∞ ball $B_\infty(r)$,

Result 1: Lipschitz

Theorem 1

- When $\mathbb{S}$ contains the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min\left\{ c_0 L r \sqrt{\frac{d}{T}}, \frac{Lr}{144} \right\}$$

NYU

Result 1: Lipschitz

Theorem 1

- When $\mathbb{S}$ contains the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min \left\{ c_0 L r \sqrt{\frac{d}{T}}, \frac{Lr}{144} \right\}$$

NYU

Result 1: Lipschitz

Theorem 1

- When $\mathbb{S}$ contains the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min\left\{ c_0 Lr\sqrt{\frac{d}{T}}, \frac{Lr}{144} \right\}$$

- ε should scale linearly with r

NYU

Result 1: Lipschitz

Theorem 1

- When $\mathbb{S}$ contains the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min\left\{ c_0 Lr\sqrt{\frac{d}{T}}, \frac{Lr}{144} \right\}$$

- ε should scale linearly with r
- Characterizes the first few and last many iterations

NYU

Result 1: Lipschitz

Theorem 1

- When $\mathbb{S}$ contains the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min\left\{ c_0 L r \sqrt{\frac{d}{T}}, \frac{Lr}{144} \right\}$$

- ε should scale linearly with r
- Characterizes the first few and last many iterations
- Mirror Descent achieves this rate of $O(Lr\sqrt{\frac{d}{T}})$

NYU

Result 1: Lipschitz

Theorem 1

- When $\mathbb{S}$ contains the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min \left\{ c_0 L r \sqrt{\frac{d}{T}}, \frac{Lr}{144} \right\}$$

- ε should scale linearly with r
- Characterizes the first few and last many iterations
- Mirror Descent achieves this rate of $O(Lr\sqrt{\frac{d}{T}})$
- Can we get better rates if we add any assumptions about g?

NYU

Result 2: Strong Convex & Lipschitz

- $\mathcal{F}_{scv}(\mathbb{S}, L, \gamma)$ is the set of L-Lipschitz functions that are γ-strongly convex:

NYU

Result 2: Strong Convex & Lipschitz

- $\mathcal{F}_{scv}(\mathbb{S}, L, \gamma)$ is the set of L-Lipschitz functions that are γ-strongly convex:
- For all $\alpha \in [0, 1]$, and all $\mathbf{x}, \mathbf{y} \in \mathbb{S}$,

$$g(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \geq \alpha g(\mathbf{x}) + (1-\alpha) g(\mathbf{y}) + \alpha(1-\alpha)\frac{\gamma^2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$$

NYU

Result 2: Strong Convex & Lipschitz

- $\mathcal{F}_{scv}(\mathbb{S}, L, \gamma)$ is the set of L-Lipschitz functions that are γ-strongly convex:
- For all $\alpha \in [0,1]$, and all $\mathbf{x}, \mathbf{y} \in \mathbb{S}$,

$$g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \geq \alpha g(\mathbf{x}) + (1-\alpha)g(\mathbf{y}) + \alpha(1-\alpha)\frac{\gamma^2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$$

- We are always at least γ-bowl shaped

NYU

Result 2: Strong Convex & Lipschitz

- $\mathcal{F}_{scv}(\mathbb{S}, L, \gamma)$ is the set of L-Lipschitz functions that are γ-strongly convex:
- For all $\alpha \in [0, 1]$, and all $\mathbf{x}, \mathbf{y} \in \mathbb{S}$,

$$g(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \geq \alpha g(\mathbf{x}) + (1-\alpha) g(\mathbf{y}) + \alpha(1-\alpha)\frac{\gamma^2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$$

- We are always at least γ-bowl shaped
- Our Hessian has minimum Eigenvalue $\geq \gamma$

NYU

Result 2: Strong Convex & Lipschitz

- $\mathcal{F}_{scv}(\mathbb{S}, L, \gamma)$ is the set of L-Lipschitz functions that are γ-strongly convex:
- For all $\alpha \in [0,1]$, and all $\mathbf{x}, \mathbf{y} \in \mathbb{S}$,

$$g(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}) \geq \alpha g(\mathbf{x})+(1-\alpha)g(\mathbf{y})+\alpha(1-\alpha)\frac{\gamma^2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$$

- We are always at least γ-bowl shaped
- Our Hessian has minimum Eigenvalue $\geq \gamma$
  - In 1 dimension, the 2^{nd} derivative is $\geq \gamma$

NYU

Result 2: Strong Convex & Lipschitz

- $\mathcal{F}_{scv}(\mathbb{S}, L, \gamma)$ is the set of L-Lipschitz functions that are γ-strongly convex:
- For all $\alpha \in [0,1]$, and all $\mathbf{x}, \mathbf{y} \in \mathbb{S}$,

$$g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \geq \alpha g(\mathbf{x}) + (1-\alpha)g(\mathbf{y}) + \alpha(1-\alpha)\frac{\gamma^2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$$

- We are always at least γ-bowl shaped
- Our Hessian has minimum Eigenvalue $\geq \gamma$
    - In 1 dimension, the 2^{nd} derivative is $\geq \gamma$
- Why is $\mathcal{F}_{scv}(\mathbb{S}, L, \gamma)$ really weird?

NYU

Result 2: Strong Convex & Lipschitz

- $\mathcal{F}_{scv}(\mathbb{S}, L, \gamma)$ is the set of L-Lipschitz functions that are γ-strongly convex:
- For all $\alpha \in [0, 1]$, and all $\mathbf{x}, \mathbf{y} \in \mathbb{S}$,

$$g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \geq \alpha g(\mathbf{x}) + (1-\alpha)g(\mathbf{y}) + \alpha(1-\alpha)\frac{\gamma^2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$$

- We are always at least γ-bowl shaped
- Our Hessian has minimum Eigenvalue $\geq \gamma$
  - In 1 dimension, the 2^{nd} derivative is $\geq \gamma$
- Why is $\mathcal{F}_{scv}(\mathbb{S}, L, \gamma)$ really weird?
  - $r \leq \frac{4L}{\gamma^2\sqrt{d}}$

NYU

Result 2: Strong Convex & Lipschitz

Theorem 2

- When $\mathbb{S}$ equals the ℓ_∞ ball $B_\infty(r)$,

Result 2: Strong Convex & Lipschitz

Theorem 2

- When $\mathbb{S}$ equals the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{scv}} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min\left\{ c_1 \frac{L^2}{\gamma^2 T}, c_2 L r \sqrt{\frac{d}{T}}, \frac{L^2}{1152\gamma^2 d}, \frac{Lr}{144} \right\}$$

Result 2: Strong Convex & Lipschitz

Theorem 2

- When $\mathbb{S}$ equals the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{scv}} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min\left\{c_1 \frac{L^2}{\gamma^2 T}, c_2 Lr\sqrt{\frac{d}{T}}, \frac{L^2}{1152\gamma^2 d}, \frac{Lr}{144}\right\}$$

Result 2: Strong Convex & Lipschitz

Theorem 2

- When $\mathbb{S}$ equals the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{scv}} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min \left\{ c_1 \frac{L^2}{\gamma^2 T}, c_2 L r \sqrt{\frac{d}{T}}, \frac{L^2}{1152 \gamma^2 d}, \frac{Lr}{144} \right\}$$

- ε should scale linearly with r

NYU

Result 2: Strong Convex & Lipschitz

Theorem 2

- When $\mathbb{S}$ equals the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{scv}} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min\left\{ c_1 \frac{L^2}{\gamma^2 T}, c_2 Lr\sqrt{\frac{d}{T}}, \frac{L^2}{1152\gamma^2 d}, \frac{Lr}{144} \right\}$$

- ε should scale linearly with r
- Characterizes the first few and last many iterations

Result 2: Strong Convex & Lipschitz

Theorem 2

- When $\mathbb{S}$ equals the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{scv}} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min\left\{ c_1 \frac{L^2}{\gamma^2 T}, c_2 Lr\sqrt{\frac{d}{T}}, \frac{L^2}{1152\gamma^2 d}, \frac{Lr}{144} \right\}$$

- ε should scale linearly with r
- Characterizes the first few and last many iterations
- Algorithms almost achieve this rate of $O(\frac{L^2}{\gamma^2 T})$

NYU

Result 2: Strong Convex & Lipschitz

Theorem 2

- When $\mathbb{S}$ equals the ℓ_∞ ball $B_\infty(r)$,
- There exists an oracle $\phi(\mathbf{x}, g)$ such that

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{scv}} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \min\left\{c_1 \frac{L^2}{\gamma^2 T}, c_2 L r \sqrt{\frac{d}{T}}, \frac{L^2}{1152\gamma^2 d}, \frac{Lr}{144}\right\}$$

- ε should scale linearly with r
- Characterizes the first few and last many iterations
- Algorithms almost achieve this rate of $O(\frac{L^2}{\gamma^2 T})$
- If $\gamma \approx 0$, we retrieve the Lipschitz lower bound

NYU

Information Theory

- Beautiful subfield of mathematics

Information Theory 101

- Beautiful subfield of mathematics
- Describes when there is enough random information to communicate some structure

Information Theory 101

- Beautiful subfield of mathematics
- Describes when there is enough random information to communicate some structure
  - How many times do I need to flip a coin to figure out if heads or tails is more likely?

Information Theory 101

- Beautiful subfield of mathematics
- Describes when there is enough random information to communicate some structure
  - How many times do I need to flip a coin to figure out if heads or tails is more likely?
- Powerful tool for making sharp lower bounds on the number of samples needed from a random distribution

Flipping Coin

- Suppose coin α is heads with probability $\frac{1}{2} + \delta$

Flipping Coin

- Suppose coin α is heads with probability $\frac{1}{2} + \delta$
- Suppose coin β is heads with probability $\frac{1}{2} - \delta$

Flipping Coin

- Suppose coin α is heads with probability $\frac{1}{2} + \delta$
- Suppose coin β is heads with probability $\frac{1}{2} - \delta$
- I randomly pick up one coin.

Flipping Coin

- Suppose coin α is heads with probability $\frac{1}{2} + \delta$
- Suppose coin β is heads with probability $\frac{1}{2} - \delta$
- I randomly pick up one coin.
- How many times T do I need to flip the coin for you to know if I am using α or β?

NYU

Flipping Coin

- Suppose coin α is heads with probability $\frac{1}{2} + \delta$
- Suppose coin β is heads with probability $\frac{1}{2} - \delta$
- I randomly pick up one coin.
- How many times T do I need to flip the coin for you to know if I am using α or β?
  - Chernoff Bound: $T = O(\frac{1}{\delta^2})$ rounds are suffice to have 90% certainty.

NYU

Flipping Coin

- Suppose coin α is heads with probability $\frac{1}{2} + \delta$
- Suppose coin β is heads with probability $\frac{1}{2} - \delta$
- I randomly pick up one coin.
- How many times T do I need to flip the coin for you to know if I am using α or β?
  - Chernoff Bound: $T = O(\frac{1}{\delta^2})$ rounds are suffice to have 90% certainty.
  - Can we prove this is optimal?

NYU

- Suppose coin α is heads with probability $\frac{1}{2}+\delta$
- Suppose coin β is heads with probability $\frac{1}{2}-\delta$
- I randomly pick up one coin.
- How many times T do I need to flip the coin for you to know if I am using α or β?
  - Chernoff Bound: $T = O(\frac{1}{\delta^2})$ rounds are suffice to have 90% certainty.
  - Can we prove this is optimal?
- The smaller δ is, the less information we get from every coin flip.

Le Cam's Lemma

Intuition: The hardness of finding the true parameter that generates random data is lower bounded by the hardness of distinguishing between two parameters.

Le Cam's Lemma

Intuition: The hardness of finding the true parameter that generates random data is lower bounded by the hardness of distinguishing between two parameters.

Le Cam's Lemma

Let $\mathbb{P} = \{\mathcal{P}_\theta\}$ be a set of probability distributions parameterized by a vector $\theta \in \Theta$. Let S be a sample from some $\mathcal{P}_\theta$. Let $\hat{\theta}(S)$ map S to any element of Θ. Let $d : \Theta \times \Theta \to \mathbb{R}$ be an error metric.

NYU

Le Cam's Lemma

Intuition: The hardness of finding the true parameter that generates random data is lower bounded by the hardness of distinguishing between two parameters.

Le Cam's Lemma

Let $\mathbb{P} = \{\mathcal{P}_\theta\}$ be a set of probability distributions parameterized by a vector $\theta \in \Theta$. Let S be a sample from some $\mathcal{P}_\theta$. Let $\hat{\theta}(S)$ map S to any element of Θ. Let $d : \Theta \times \Theta \to \mathbb{R}$ be an error metric.Then, for any $\mathcal{P}_{\theta_1}, \mathcal{P}_{\theta_2} \in \mathbb{P}$,

Le Cam's Lemma

Intuition: The hardness of finding the true parameter that generates random data is lower bounded by the hardness of distinguishing between two parameters.

Le Cam's Lemma

Let $\mathbb{P} = \{\mathcal{P}_\theta\}$ be a set of probability distributions parameterized by a vector $\theta \in \Theta$. Let S be a sample from some $\mathcal{P}_\theta$. Let $\hat{\theta}(S)$ map S to any element of Θ. Let $d : \Theta \times \Theta \to \mathbb{R}$ be an error metric.Then, for any $\mathcal{P}_{\theta_1}, \mathcal{P}_{\theta_2} \in \mathbb{P}$,

$$\inf_{\hat{\theta}} \sup_{\mathcal{P}_\theta \in \mathbb{P}} \mathbb{E}_{S \sim \mathcal{P}_\theta} [d(\hat{\theta}(S), \theta)] \geq \frac{d(\theta_1, \theta_2)}{4} \int_S \min\{\Pr[\mathcal{P}_{\theta_1} = S], \Pr[\mathcal{P}_{\theta_2} = S]\} dS$$

Flipping Coin

- Suppose coin α is heads with probability $\frac{1}{2} + \delta$
- Suppose coin β is heads with probability $\frac{1}{2} - \delta$
- I randomly pick up one coin.
- How many times T do I need to flip the coin for you to know if I am using α or β?
  - Chernoff Bound: $T = O(\frac{1}{\delta^2})$ rounds are suffice to have 90% certainty.

NYU

Flipping Coin

- Suppose coin α is heads with probability $\frac{1}{2} + \delta$
- Suppose coin β is heads with probability $\frac{1}{2} - \delta$
- I randomly pick up one coin.
- How many times T do I need to flip the coin for you to know if I am using α or β?
  - Chernoff Bound: $T = O(\frac{1}{\delta^2})$ rounds are suffice to have 90% certainty.
- For $\delta \in (0, 1/4)$, for any estimator $\hat{\alpha}$, we have

$$\sup_{\alpha^* \in \{\frac{1}{2}+\delta, \frac{1}{2}-\delta\}} \Pr[\hat{\alpha} \neq \alpha^*] \geq 1 - \sqrt{8T\delta^2}$$

NYU

Flipping Coin

- Suppose coin α is heads with probability $\frac{1}{2} + \delta$
- Suppose coin β is heads with probability $\frac{1}{2} - \delta$
- I randomly pick up one coin.
- How many times T do I need to flip the coin for you to know if I am using α or β?
  - Chernoff Bound: $T = O(\frac{1}{\delta^2})$ rounds are suffice to have 90% certainty.
- For $\delta \in (0, 1/4)$, for any estimator $\hat{\alpha}$, we have

$$\sup_{\alpha^* \in \{\frac{1}{2}+\delta, \frac{1}{2}-\delta\}} \Pr[\hat{\alpha} \neq \alpha^*] \geq 1 - \sqrt{8T\delta^2}$$

- We get constant probability for $T = \Omega(\frac{1}{\delta^2})$, so the Chernoff bound is optimal!

NYU

Flipping Coins

- Suppose α is a stack of d δ-biased coins

Flipping Coins

- Suppose α is a stack of d δ-biased coins
- Suppose β is a stack of d δ-biased coins

Flipping Coins

- Suppose α is a stack of d δ-biased coins
- Suppose β is a stack of d δ-biased coins
- I randomly pick up one stack of coins

Flipping Coins

- Suppose α is a stack of d δ-biased coins
- Suppose β is a stack of d δ-biased coins
- I randomly pick up one stack of coins
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack

Flipping Coins

- Suppose α is a stack of d δ-biased coins
- Suppose β is a stack of d δ-biased coins
- I randomly pick up one stack of coins
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack
- How many times T do I need to pick and flip a coin for you to know if I am using α or β?

NYU

Flipping Coins

- Suppose α is a stack of d δ-biased coins
- Suppose β is a stack of d δ-biased coins
- I randomly pick up one stack of coins
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack
- How many times T do I need to pick and flip a coin for you to know if I am using α or β?
- Observation: This depends on how similar α and β are.

NYU

Flipping Coins

- Suppose α is a stack of d δ-biased coins
- Suppose β is a stack of d δ-biased coins
- I randomly pick up one stack of coins
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack
- How many times T do I need to pick and flip a coin for you to know if I am using α or β?
- Observation: This depends on how similar α and β are.
- In the best case, this should require $T \approx \frac{1}{\delta^2}$

NYU

Flipping Many Coins

- Suppose $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ are stacks of d δ-biased coins

Flipping Many Coins

- Suppose $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ are stacks of d δ-biased coins
- I randomly pick up one stack of coins α^*

NYU

Flipping Many Coins

- Suppose $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ are stacks of d δ-biased coins
- I randomly pick up one stack of coins α^*
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack

Flipping Many Coins

- Suppose $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ are stacks of d δ-biased coins
- I randomly pick up one stack of coins α^*
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack
- How many times T do I need to pick and flip a coin for you to know which α_j I am using?

NYU

Flipping Many Coins

- Suppose $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ are stacks of d δ-biased coins
- I randomly pick up one stack of coins α^*
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack
- How many times T do I need to pick and flip a coin for you to know which α_j I am using?
- This is hard to reason about, but there should be a minimum T needed

NYU

Flipping Many Coins

- Suppose $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ are stacks of d δ-biased coins
- I randomly pick up one stack of coins α^*
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack
- How many times T do I need to pick and flip a coin for you to know which α_j I am using?
- This is hard to reason about, but there should be a minimum T needed
- Observation: This depends on the similarities between α_j and α_k for all j, k

NYU

Fano's Inequality

Let $\theta_1, \ldots, \theta_k$ describe probability distributions. Nature picks some θ^* uniformly at random. Dataset S is generated from θ^*.

Fano's Inequality

Fano's Inequality

Let $\theta_1, \ldots, \theta_k$ describe probability distributions. Nature picks some θ^* uniformly at random. Dataset S is generated from θ^*.Then, any estimate $\hat{\theta}$ of θ^* given only S has

$$\Pr[\hat{\theta} \neq \theta^*] \geq 1 - \frac{\mathbb{I}(\theta^*, S) + \log 2}{\log k}$$

Where $\mathbb{I}(\theta^*, S)$ is the mutual information between θ^* and S.

NYU

Fano's Inequality

Let $\theta_1, \dots, \theta_k$ describe probability distributions. Nature picks some θ^* uniformly at random. Dataset S is generated from θ^*.Then, any estimate $\hat{\theta}$ of θ^* given only S has

$$\Pr[\hat{\theta} \neq \theta^*] \geq 1 - \frac{\mathbb{I}(\theta^*, S) + \log 2}{\log k}$$

Where $\mathbb{I}(\theta^*, S)$ is the mutual information between θ^* and S.

Intuition: if S does not reveal much about θ^*, and if there are many candidates $\hat{\theta}$, then we cannot find θ^* reliably.

Flipping Many Coins

- Suppose $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ are stacks of d δ-biased coins
- I randomly pick up one stack of coins α^*
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack
- How many times T do I need to pick and flip a coin for you to know which α_j I am using?

NYU

Flipping Many Coins

- Suppose $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ are stacks of d δ-biased coins
- I randomly pick up one stack of coins α^*
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack
- How many times T do I need to pick and flip a coin for you to know which α_j I am using?
- The coin flips do not reveal much about α^* because δ is small.

NYU

Flipping Many Coins

- Suppose $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ are stacks of d δ-biased coins
- I randomly pick up one stack of coins α^*
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack
- How many times T do I need to pick and flip a coin for you to know which α_j I am using?
- The coin flips do not reveal much about α^* because δ is small.
- For $\delta \in (0, 1/4)$, any test $\hat{\alpha}$ has

$$\Pr[\hat{\alpha} \neq \alpha^*] \geq 1 - \frac{16T\delta^2 + \log 2}{\log k}$$

NYU

Flipping Many Coins

- Suppose $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ are stacks of d δ-biased coins
- I randomly pick up one stack of coins α^*
- I sample $i \sim [d]$, and flip the i^{th} coin in my stack
- How many times T do I need to pick and flip a coin for you to know which α_j I am using?
- The coin flips do not reveal much about α^* because δ is small.
- For $\delta \in (0, 1/4)$, any test $\hat{\alpha}$ has

$$\Pr[\hat{\alpha} \neq \alpha^*] \geq 1 - \frac{16T\delta^2 + \log 2}{\log k}$$

- If we pick $k = O(c^d)$, then we need $T = \Omega(\frac{d}{\delta^2})$ for 90% confidence.

NYU

Stochastic Optimization

Connecting Optimization & Coins

$$\sup_{\phi} \inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \Omega(\frac{1}{\sqrt{T}})$$

NYU

Connecting Optimization & Coins

$$\sup_{\phi} \inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \Omega(\frac{1}{\sqrt{T}})$$

- How can we make stochastic optimization look like recovering α from coin flips?

NYU

Connecting Optimization & Coins

$$\sup_{\phi} \inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \Omega(\frac{1}{\sqrt{T}})$$

- How can we make stochastic optimization look like recovering α from coin flips?
- We can design the oracle $\phi(\mathbf{x}_t, g)$

NYU

Connecting Optimization & Coins

$$\sup_{\phi} \inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \Omega(\frac{1}{\sqrt{T}})$$

- How can we make stochastic optimization look like recovering α from coin flips?
- We can design the oracle $\phi(\mathbf{x}_t, g)$
- We can restrict to a finite subset of $\mathcal{F}_{cv}(\mathbb{S}, L)$

NYU

Connecting Optimization & Coins

$$\sup_{\phi} \inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{g \in \mathcal{F}_{cv}(\mathbb{S}, L)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \Omega(\frac{1}{\sqrt{T}})$$

- How can we make stochastic optimization look like recovering α from coin flips?
- We can design the oracle $\phi(\mathbf{x}_t, g)$
- We can restrict to a finite subset of $\mathcal{F}_{cv}(\mathbb{S}, L)$
- Show that for some oracle $\phi(\mathbf{x}_t, g)$ and $\mathcal{G}(\delta) \subseteq \mathcal{F}_{cv}(\mathbb{S}, L)$, we have

$$\inf_{\mathcal{M}_T \in \mathbb{M}_T} \max_{g \in \mathcal{G}(\delta)} \mathbb{E}[\varepsilon_T(\mathcal{M}_T, g, \phi)] \geq \Omega(\frac{1}{\sqrt{T}})$$

NYU

Proof Intuition: Lipschitz

- We want to achieve error $\varepsilon > 0$ on $\mathcal{F}_{cv}$

Proof Intuition: Lipschitz

- We want to achieve error $\varepsilon > 0$ on $\mathcal{F}_{cv}$
- We pick $\mathcal{G}(\delta) \subseteq \mathcal{F}_{cv}$ such that

NYU

Proof Intuition: Lipschitz

- We want to achieve error $\varepsilon > 0$ on $\mathcal{F}_{cv}$
- We pick $\mathcal{G}(\delta) \subseteq \mathcal{F}_{cv}$ such that
  - If we can get error ε on all $g \in \mathcal{G}(\delta)$, then we can recover a stack of δ-biased coins α^*

NYU

Proof Intuition: Lipschitz

- We want to achieve error $\varepsilon > 0$ on $\mathcal{F}_{cv}$
- We pick $\mathcal{G}(\delta) \subseteq \mathcal{F}_{cv}$ such that
  - If we can get error ε on all $g \in \mathcal{G}(\delta)$, then we can recover a stack of δ-biased coins α^*
  - $|\mathcal{G}(\delta)| = \Theta(c^d)$

NYU

Proof Intuition: Lipschitz

- We want to achieve error $\varepsilon > 0$ on $\mathcal{F}_{cv}$
- We pick $\mathcal{G}(\delta) \subseteq \mathcal{F}_{cv}$ such that
  - If we can get error ε on all $g \in \mathcal{G}(\delta)$, then we can recover a stack of δ-biased coins α^*
  - $|\mathcal{G}(\delta)| = \Theta(c^d)$
  - $\delta = \Theta(\varepsilon)$

NYU

Proof Intuition: Lipschitz

- We want to achieve error $\varepsilon > 0$ on $\mathcal{F}_{cv}$
- We pick $\mathcal{G}(\delta) \subseteq \mathcal{F}_{cv}$ such that
  - If we can get error ε on all $g \in \mathcal{G}(\delta)$, then we can recover a stack of δ-biased coins α^*
  - $|\mathcal{G}(\delta)| = \Theta(c^d)$
  - $\delta = \Theta(\varepsilon)$
- Recall that recovering a stack of coins takes $T = \Omega(\frac{d}{\delta^2})$ samples

NYU

Proof Intuition: Lipschitz

- We want to achieve error $\varepsilon > 0$ on $\mathcal{F}_{cv}$
- We pick $\mathcal{G}(\delta) \subseteq \mathcal{F}_{cv}$ such that
  - If we can get error ε on all $g \in \mathcal{G}(\delta)$, then we can recover a stack of δ-biased coins α^*
  - $|\mathcal{G}(\delta)| = \Theta(c^d)$
  - $\delta = \Theta(\varepsilon)$
- Recall that recovering a stack of coins takes $T = \Omega(\frac{d}{\delta^2})$ samples
- Since $\mathcal{G}(\delta)$ yields $\delta = \Theta(\varepsilon)$, we know $T = \Omega(\frac{d}{\varepsilon^2})$ and $\varepsilon = \Omega(\sqrt{\frac{d}{T}})$

NYU

Proof Intuition: Strong Convexity

- We want to achieve error $\varepsilon > 0$ on $\mathcal{F}_{scv}$
- We pick $\mathcal{G}(\delta) \subseteq \mathcal{F}_{cv}$ such that
  - If we can get error ε on all $g \in \mathcal{G}(\delta)$, then we can recover a stack of δ-biased coins α^*
  - $|\mathcal{G}(\delta)| = \Theta(c^d)$
  - $\delta = \Theta(\sqrt{\varepsilon})$
- Recall that recovering a stack of coins takes $T = \Omega(\frac{d}{\delta^2})$ samples
- Since $\mathcal{G}(\delta)$ yields $\delta = \Theta(\sqrt{\varepsilon})$, we know $T = \Omega(\frac{d}{\varepsilon})$ and $\varepsilon = \Omega(\frac{\sqrt{d}}{T})$

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$.

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \left|x + \frac{1}{2}\right|$ and $f^-(x) := \left|x - \frac{1}{2}\right|$

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \left|x + \frac{1}{2}\right|$ and $f^-(x) := \left|x - \frac{1}{2}\right|$
- Desmos Graph Link

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \left|x + \frac{1}{2}\right|$ and $f^-(x) := \left|x - \frac{1}{2}\right|$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \left|x + \frac{1}{2}\right|$ and $f^-(x) := \left|x - \frac{1}{2}\right|$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:
  - Sample $i \sim [d]$ uniformly

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \left|x + \frac{1}{2}\right|$ and $f^-(x) := \left|x - \frac{1}{2}\right|$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:
  - Sample $i \sim [d]$ uniformly
  - Return $f^+(x_i)$ and its gradient w.p. $\frac{1}{2} + \alpha_i \delta$

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + \left(\frac{1}{2} - \alpha_i \delta\right) f^-(x_i)$
- Where $f^+(x) := \left|x + \frac{1}{2}\right|$ and $f^-(x) := \left|x - \frac{1}{2}\right|$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:
  - Sample $i \sim [d]$ uniformly
  - Return $f^+(x_i)$ and its gradient w.p. $\frac{1}{2} + \alpha_i \delta$
  - Return $f^-(x_i)$ and its gradient w.p. $\frac{1}{2} - \alpha_i \delta$

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + \left(\frac{1}{2} - \alpha_i \delta\right) f^-(x_i)$
- Where $f^+(x) := \left|x + \frac{1}{2}\right|$ and $f^-(x) := \left|x - \frac{1}{2}\right|$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:
  - Sample $i \sim [d]$ uniformly
  - Return $f^+(x_i)$ and its gradient w.p. $\frac{1}{2} + \alpha_i \delta$
  - Return $f^-(x_i)$ and its gradient w.p. $\frac{1}{2} - \alpha_i \delta$
- This is the same as flipping a randomly chosen coin from α

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + \left(\frac{1}{2} - \alpha_i \delta\right) f^-(x_i)$
- Where $f^+(x) := \left|x + \frac{1}{2}\right|$ and $f^-(x) := \left|x - \frac{1}{2}\right|$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:
  - Sample $i \sim [d]$ uniformly
  - Return $f^+(x_i)$ and its gradient w.p. $\frac{1}{2} + \alpha_i \delta$
  - Return $f^-(x_i)$ and its gradient w.p. $\frac{1}{2} - \alpha_i \delta$
- This is the same as flipping a randomly chosen coin from α
- If we optimize dimension i of $\mathbf{x}$, then we know the i^{th} coin of α

Construction of the αs

- Let us pick $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ such that

$$\alpha_i \text{ and } \alpha_j \text{ are equal at at most } \frac{d}{4} \text{ indices}$$

NYU

Construction of the αs

- Let us pick $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ such that

$$\alpha_i \text{ and } \alpha_j \text{ are equal at at most } \frac{d}{4} \text{ indices}$$

- Then we can take $k = |\mathcal{V}| = (2/\sqrt{e})^{d/2}$

Construction of the αs

- Let us pick $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ such that

$$\alpha_i \text{ and } \alpha_j \text{ are equal at at most } \frac{d}{4} \text{ indices}$$

- Then we can take $k = |\mathcal{V}| = (2/\sqrt{e})^{d/2}$
- Then if $\mathbb{E}_\phi[\varepsilon_T(\mathcal{M}_T, g_\alpha, \phi)] \leq \frac{\delta}{8}$ for all $\alpha \in \mathcal{V}$, then we uniquely decode α with probability $\frac{2}{3}$

Construction of the αs

- Let us pick $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ such that

$$\alpha_i \text{ and } \alpha_j \text{ are equal at at most } \frac{d}{4} \text{ indices}$$

- Then we can take $k = |\mathcal{V}| = (2/\sqrt{e})^{d/2}$
- Then if $\mathbb{E}_\phi[\varepsilon_T(\mathcal{M}_T, g_\alpha, \phi)] \leq \frac{\delta}{8}$ for all $\alpha \in \mathcal{V}$, then we uniquely decode α with probability $\frac{2}{3}$
  - Algebra and Markov's Inequality

NYU

Construction of the αs

- Let us pick $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ such that

$$\alpha_i \text{ and } \alpha_j \text{ are equal at at most } \frac{d}{4} \text{ indices}$$

- Then we can take $k = |\mathcal{V}| = (2/\sqrt{e})^{d/2}$
- Then if $\mathbb{E}_\phi[\varepsilon_T(\mathcal{M}_T, g_\alpha, \phi)] \leq \frac{\delta}{8}$ for all $\alpha \in \mathcal{V}$, then we uniquely decode α with probability $\frac{2}{3}$
  - Algebra and Markov's Inequality
- So, given $\varepsilon > 0$, we can use $\delta = 8\varepsilon$.

NYU

Theorem 1

- Recall that for $\delta \in (0, 1/4)$, any test $\hat{\alpha}$ has

$$\Pr[\hat{\alpha} \neq \alpha^*] \geq 1 - \frac{16T\delta^2 + \log 2}{\log |\mathcal{V}|}$$

Theorem 1

- Recall that for $\delta \in (0, 1/4)$, any test $\hat{\alpha}$ has

$$\Pr[\hat{\alpha} \neq \alpha^*] \geq 1 - \frac{16T\delta^2 + \log 2}{\log |\mathcal{V}|}$$

- Further, for $\delta = 8\varepsilon$, we have

$$\frac{1}{3} \geq \Pr[\hat{\alpha} \neq \alpha^*]$$

NYU

Theorem 1

- Recall that for $\delta \in (0, 1/4)$, any test $\hat{\alpha}$ has

$$\Pr[\hat{\alpha} \neq \alpha^*] \geq 1 - \frac{16T\delta^2 + \log 2}{\log |\mathcal{V}|}$$

- Further, for $\delta = 8\varepsilon$, we have

$$\frac{1}{3} \geq \Pr[\hat{\alpha} \neq \alpha^*]$$

- So, we have

$$\frac{1}{3} \geq 1 - \frac{16T \cdot (8\varepsilon)^2 + \log 2}{\log |\mathcal{V}|}$$

$$\frac{1}{3} \geq 1 - \frac{c_0 T\varepsilon^2 + \log 2}{\frac{d}{2}\log(2/\sqrt{e})}$$

$$\varepsilon \geq \Omega\left(\sqrt{\frac{d}{T}}\right)$$

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0, 1]$.

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0, 1]$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + \left(\frac{1}{2} - \alpha_i \delta\right) f^-(x_i)$

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0, 1]$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \theta \left| x + \frac{1}{2} \right| + \frac{1-\theta}{4} \left(x + \frac{1}{2}\right)^2$

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0, 1]$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \theta \left|x + \frac{1}{2}\right| + \frac{1-\theta}{4} \left(x + \frac{1}{2}\right)^2$
- $f^-(x) := \theta \left|x - \frac{1}{2}\right| + \frac{1-\theta}{4} \left(x - \frac{1}{2}\right)^2$

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0, 1]$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \theta \left|x + \frac{1}{2}\right| + \frac{1-\theta}{4} \left(x + \frac{1}{2}\right)^2$
- $f^-(x) := \theta \left|x - \frac{1}{2}\right| + \frac{1-\theta}{4} \left(x - \frac{1}{2}\right)^2$
- Desmos Graph Link

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1,1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0,1]$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d}\sum_{i=1}^{d}\left(\frac{1}{2}+\alpha_i\delta\right) f^{+}(x_i) + (\frac{1}{2}-\alpha_i\delta)f^{-}(x_i)$
- Where $f^{+}(x) := \theta\left|x+\frac{1}{2}\right| + \frac{1-\theta}{4}\left(x+\frac{1}{2}\right)^2$
- $f^{-}(x) := \theta\left|x-\frac{1}{2}\right| + \frac{1-\theta}{4}\left(x-\frac{1}{2}\right)^2$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0, 1]$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + \left(\frac{1}{2} - \alpha_i \delta\right) f^-(x_i)$
- Where $f^+(x) := \theta \left|x + \frac{1}{2}\right| + \frac{1-\theta}{4} \left(x + \frac{1}{2}\right)^2$
- $f^-(x) := \theta \left|x - \frac{1}{2}\right| + \frac{1-\theta}{4} \left(x - \frac{1}{2}\right)^2$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:
    - Sample $i \sim [d]$ uniformly

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0, 1]$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \theta \left|x + \frac{1}{2}\right| + \frac{1-\theta}{4}\left(x + \frac{1}{2}\right)^2$
- $\qquad f^-(x) := \theta \left|x - \frac{1}{2}\right| + \frac{1-\theta}{4}\left(x - \frac{1}{2}\right)^2$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:
  - Sample $i \sim [d]$ uniformly
  - Return $f^+(x_i)$ and its gradient w.p. $\frac{1}{2} + \alpha_i \delta$

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0, 1]$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \theta \left|x + \frac{1}{2}\right| + \frac{1-\theta}{4} \left(x + \frac{1}{2}\right)^2$
- $f^-(x) := \theta \left|x - \frac{1}{2}\right| + \frac{1-\theta}{4} \left(x - \frac{1}{2}\right)^2$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:
  - Sample $i \sim [d]$ uniformly
  - Return $f^+(x_i)$ and its gradient w.p. $\frac{1}{2} + \alpha_i \delta$
  - Return $f^-(x_i)$ and its gradient w.p. $\frac{1}{2} - \alpha_i \delta$

NYU

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1, 1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0, 1]$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f^+(x_i) + (\frac{1}{2} - \alpha_i \delta) f^-(x_i)$
- Where $f^+(x) := \theta \left|x + \frac{1}{2}\right| + \frac{1-\theta}{4} \left(x + \frac{1}{2}\right)^2$
- $\qquad f^-(x) := \theta \left|x - \frac{1}{2}\right| + \frac{1-\theta}{4} \left(x - \frac{1}{2}\right)^2$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:
  - Sample $i \sim [d]$ uniformly
  - Return $f^+(x_i)$ and its gradient w.p. $\frac{1}{2} + \alpha_i \delta$
  - Return $f^-(x_i)$ and its gradient w.p. $\frac{1}{2} - \alpha_i \delta$
- This is the same as flipping a randomly chosen coin α_i

Construction for Lipschitz $\mathcal{G}(\delta)$

- Let $\alpha \in \{-1,1\}^n$. Fix $\delta > 0$. Fix $\theta \in [0,1]$.
- Define $g_\alpha(\mathbf{x}) := \frac{c}{d}\sum_{i=1}^{d}\left(\frac{1}{2}+\alpha_i\delta\right) f^+(x_i) + (\frac{1}{2}-\alpha_i\delta) f^-(x_i)$
- Where $f^+(x) := \theta\left|x+\frac{1}{2}\right| + \frac{1-\theta}{4}\left(x+\frac{1}{2}\right)^2$
- $\qquad f^-(x) := \theta\left|x-\frac{1}{2}\right| + \frac{1-\theta}{4}\left(x-\frac{1}{2}\right)^2$
- Desmos Graph Link
- Let our oracle $\phi(\mathbf{x}, g_\alpha)$ be:
  - Sample $i \sim [d]$ uniformly
  - Return $f^+(x_i)$ and its gradient w.p. $\frac{1}{2}+\alpha_i\delta$
  - Return $f^-(x_i)$ and its gradient w.p. $\frac{1}{2}-\alpha_i\delta$
- This is the same as flipping a randomly chosen coin α_i
- If we optimize dimension i of $\mathbf{x}$, then we know α_i

Construction of the αs

- Let us pick $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ such that

$$\alpha_i \text{ and } \alpha_j \text{ are equal at at most } \frac{d}{4} \text{ indices}$$

Construction of the αs

- Let us pick $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ such that

$$\alpha_i \text{ and } \alpha_j \text{ are equal at at most } \frac{d}{4} \text{ indices}$$

- Then we can take $k = |\mathcal{V}| = (2/\sqrt{e})^{d/2}$

NYU

Construction of the αs

- Let us pick $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ such that

$$\alpha_i \text{ and } \alpha_j \text{ are equal at at most } \frac{d}{4} \text{ indices}$$

- Then we can take $k = |\mathcal{V}| = (2/\sqrt{e})^{d/2}$
- Then if $\mathbb{E}_\phi[\varepsilon_T(\mathcal{M}_T, g_\alpha, \phi)] \leq \frac{C_0\delta^2}{1-\theta}$, then we uniquely decode α with probability $\frac{2}{3}$

Construction of the αs

- Let us pick $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ such that

$$\alpha_i \text{ and } \alpha_j \text{ are equal at at most } \frac{d}{4} \text{ indices}$$

- Then we can take $k = |\mathcal{V}| = (2/\sqrt{e})^{d/2}$
- Then if $\mathbb{E}_\phi[\varepsilon_T(\mathcal{M}_T, g_\alpha, \phi)] \leq \frac{C_0\delta^2}{1-\theta}$, then we uniquely decode α with probability $\frac{2}{3}$
  - Algebra and Markov's Inequality

NYU

Construction of the αs

- Let us pick $\mathcal{V} = \{\alpha_1, \ldots, \alpha_k\}$ such that

$$\alpha_i \text{ and } \alpha_j \text{ are equal at at most } \frac{d}{4} \text{ indices}$$

- Then we can take $k = |\mathcal{V}| = (2/\sqrt{e})^{d/2}$
- Then if $\mathbb{E}_\phi[\varepsilon_T(\mathcal{M}_T, g_\alpha, \phi)] \leq \frac{C_0\delta^2}{1-\theta}$, then we uniquely decode α with probability $\frac{2}{3}$
  - Algebra and Markov's Inequality
- So, given $\varepsilon > 0$, we can use $\delta = \sqrt{C_1\varepsilon(1-\theta)}$.

NYU

Theorem 2

- Recall that for $\delta \in (0, 1/4)$, any test $\hat{\alpha}$ has

$$\Pr[\hat{\alpha} \neq \alpha^*] \geq 1 - \frac{16T\delta^2 + \log 2}{\log |\mathcal{V}|}$$

Theorem 2

- Recall that for $\delta \in (0, 1/4)$, any test $\hat{\alpha}$ has

$$\Pr[\hat{\alpha} \neq \alpha^*] \geq 1 - \frac{16T\delta^2 + \log 2}{\log |\mathcal{V}|}$$

- Further, for $\delta = \sqrt{C_1 \varepsilon (1 - \theta)}$, we have

$$\frac{1}{3} \geq \Pr[\hat{\alpha} \neq \alpha^*]$$

NYU

Theorem 2

- Recall that for $\delta \in (0, 1/4)$, any test $\hat{\alpha}$ has

$$\Pr[\hat{\alpha} \neq \alpha^*] \geq 1 - \frac{16T\delta^2 + \log 2}{\log |\mathcal{V}|}$$

- Further, for $\delta = \sqrt{C_1 \varepsilon (1 - \theta)}$, we have

$$\frac{1}{3} \geq \Pr[\hat{\alpha} \neq \alpha^*]$$

- So, for θ not too large, we have

$$\frac{1}{3} \geq 1 - \frac{16T \cdot (C_1 \varepsilon (1 - \theta))^2 + \log 2}{\log |\mathcal{V}|}$$

$$\varepsilon \geq \Omega\left(\frac{d}{(1 - \theta)T}\right)$$

NYU