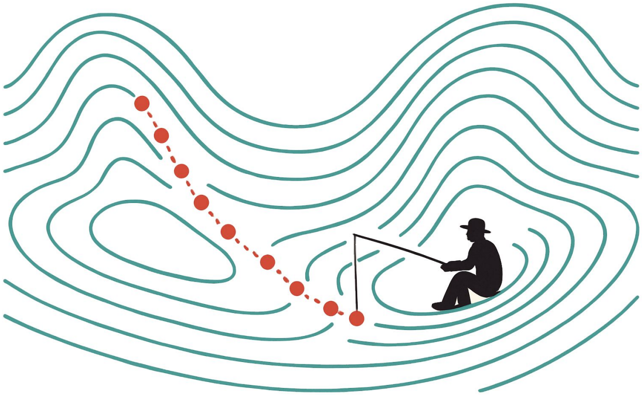


Can SGD Select Good Fishermen?

Alkis Kalavasis, Anay Mehrotra, Felix Zhou

Yale University

Image generated by Sora.

Collaborators

Alkis Kalavasis
(Yale University)

Anay Mehrotra
(Yale University)

Table of Contents

Table of Contents

Hunters vs Fishermen

In a small village, two mutually exclusive occupations are available: hunting and fishing.

Hunters vs Fishermen

In a small village, two mutually exclusive occupations are available: hunting and fishing.

Simple question:

> *What makes a good fisherman and what makes a good hunter?*

Hunters vs Fishermen

In a small village, two mutually exclusive occupations are available: hunting and fishing.

Simple question:

What makes a good fisherman and what makes a good hunter?

1. Collect random sample of hunters and fishermen from the village.

Hunters vs Fishermen

In a small village, two mutually exclusive occupations are available: hunting and fishing.

Simple question:

What makes a good fisherman and what makes a good hunter?

1. Collect random sample of hunters and fishermen from the village.
2. Record relevant features and income.

Hunters vs Fishermen

In a small village, two mutually exclusive occupations are available: hunting and fishing.

Simple question:

What makes a good fisherman and what makes a good hunter?

1. Collect random sample of hunters and fishermen from the village.
2. Record relevant features and income.
3. Estimate parameters of 2 linear models, one per occupation.

Hunters vs Fishermen

The resulting linear fits can be biased due to selection bias.

Hunters vs Fishermen

The resulting linear fits can be biased due to selection bias.

- Better hunters will opt to hunt, and vice-versa.

Yale

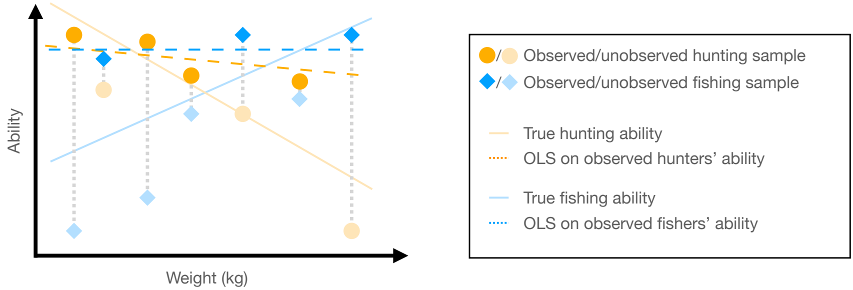
Hunters vs Fishermen

The resulting linear fits can be biased due to selection bias.

- Better hunters will opt to hunt, and vice-versa.
- Never observe fishing earnings of an individual better at hunting!

The resulting linear fits can be biased due to selection bias.

- Better hunters will opt to hunt, and vice-versa.
- Never observe fishing earnings of an individual better at hunting!



Courtesy of Cherapanamjeri, Daskalakis, Ilyas, Zampetakis [CDIZ23].

History: Inference under Selection Bias

Rich history in Statistics and Econometrics, starting with foundational works of Roy [Roy51], Heckman [Hec79], Willis and Rosen [WR79], Fair and Jaffe [FJ72], and has since found many applications:

- Causal inference and imitation learning [Hec90]

History: Inference under Selection Bias

Rich history in Statistics and Econometrics, starting with foundational works of Roy [Roy51], Heckman [Hec79], Willis and Rosen [WR79], Fair and Jaffe [FJ72], and has since found many applications:

- Causal inference and imitation learning [Hec90]
- Learning from strategically reported data [HMPW16; DRSW+18; KR20],

History: Inference under Selection Bias

Rich history in Statistics and Econometrics, starting with foundational works of Roy [Roy51], Heckman [Hec79], Willis and Rosen [WR79], Fair and Jaffe [FJ72], and has since found many applications:

- Causal inference and imitation learning [Hec90]
- Learning from strategically reported data [HMPW16; DRSW+18; KR20],
- Learning from auction data [AH02; AH07; CDIZ22].

History: Inference under Selection Bias

Rich history in Statistics and Econometrics, starting with foundational works of Roy [Roy51], Heckman [Hec79], Willis and Rosen [WR79], Fair and Jaffe [FJ72], and has since found many applications:

- Causal inference and imitation learning [Hec90]
- Learning from strategically reported data [HMPW16; DRSW+18; KR20],
- Learning from auction data [AH02; AH07; CDIZ22].
- Studies of participation in the labor force [Hec74; Han76; Nel77; Hec79; Cog80; Han80]

History: Inference under Selection Bias

Rich history in Statistics and Econometrics, starting with foundational works of Roy [Roy51], Heckman [Hec79], Willis and Rosen [WR79], Fair and Jaffe [FJ72], and has since found many applications:

- Causal inference and imitation learning [Hec90]
- Learning from strategically reported data [HMPW16; DRSW+18; KR20],
- Learning from auction data [AH02; AH07; CDIZ22].
- Studies of participation in the labor force [Hec74; Han76; Nel77; Hec79; Cog80; Han80]
- Studies of migration and income [NZ80; Bor87]

History: Inference under Selection Bias

Rich history in Statistics and Econometrics, starting with foundational works of Roy [Roy51], Heckman [Hec79], Willis and Rosen [WR79], Fair and Jaffe [FJ72], and has since found many applications:

- Causal inference and imitation learning [Hec90]
- Learning from strategically reported data [HMPW16; DRSW+18; KR20],
- Learning from auction data [AH02; AH07; CDIZ22].
- Studies of participation in the labor force [Hec74; Han76; Nel77; Hec79; Cog80; Han80]
- Studies of migration and income [NZ80; Bor87]
- Studies of the effect of unions on wages [Lee78; AF82]

History: Inference under Selection Bias

Rich history in Statistics and Econometrics, starting with foundational works of Roy [Roy51], Heckman [Hec79], Willis and Rosen [WR79], Fair and Jaffe [FJ72], and has since found many applications:

- Causal inference and imitation learning [Hec90]
- Learning from strategically reported data [HMPW16; DRSW+18; KR20],
- Learning from auction data [AH02; AH07; CDIZ22].
- Studies of participation in the labor force [Hec74; Han76; Nel77; Hec79; Cog80; Han80]
- Studies of migration and income [NZ80; Bor87]
- Studies of the effect of unions on wages [Lee78; AF82]
- Studies of returns on education [GHH78; KLMT79; WR79]

Goal: recover unknown regressors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k \in \mathbb{R}^d$ to error $\varepsilon > 0$ given observations $(\boldsymbol{x}_1, y_1^{\max}), \ldots, (\boldsymbol{x}_n, y_n^{\max})$.

Model: Regression under Self-Selection Biases

Goal: recover unknown regressors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k \in \mathbb{R}^d$ to error $\varepsilon > 0$ given observations $(\boldsymbol{x}_1, y_1^{\max}), \ldots, (\boldsymbol{x}_n, y_n^{\max})$.

DEFINITION 1 (MAXIMUM SELF-SELECTION MODEL [CDIZ23])

An observation $(\boldsymbol{x}, y^{\max})$ is generated as follows:

Yale

Model: Regression under Self-Selection Biases

Goal: recover unknown regressors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k \in \mathbb{R}^d$ to error $\varepsilon > 0$ given observations $(\boldsymbol{x}_1, y_1^{\max}), \ldots, (\boldsymbol{x}_n, y_n^{\max})$.

DEFINITION 1 (MAXIMUM SELF-SELECTION MODEL [CDIZ23])

An observation $(\boldsymbol{x}, y^{\max})$ is generated as follows:

1. $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$.

Yale

Goal: recover unknown regressors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k \in \mathbb{R}^d$ to error $\varepsilon > 0$ given observations $(\boldsymbol{x}_1, y_1^{\max}), \ldots, (\boldsymbol{x}_n, y_n^{\max})$.

DEFINITION 1 (MAXIMUM SELF-SELECTION MODEL [CDIZ23])

An observation $(\boldsymbol{x}, y^{\max})$ is generated as follows:

1. $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$.
2. $y_i = \boldsymbol{w}_i^\top \boldsymbol{x} + \xi_i$ where $\xi_i \sim_{i.i.d.} \mathcal{N}(0, 1)$.

Model: Regression under Self-Selection Biases

Goal: recover unknown regressors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k \in \mathbb{R}^d$ to error $\varepsilon > 0$ given observations $(\boldsymbol{x}_1, y_1^{\max}), \ldots, (\boldsymbol{x}_n, y_n^{\max})$.

DEFINITION 1 (MAXIMUM SELF-SELECTION MODEL [CDIZ23])

An observation $(\boldsymbol{x}, y^{\max})$ is generated as follows:

1. $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$.
2. $y_i = \boldsymbol{w}_i^\top \boldsymbol{x} + \xi_i$ where $\xi_i \sim_{i.i.d.} \mathcal{N}(0, 1)$.
3. Observe $(\boldsymbol{x}, y^{\max})$ where $y^{\max} = \max\{y_1, \ldots, y_k\}$.

Table of Contents

Prior Works: Regression under Self-Selection Biases

Efficient algorithms for *finite samples* were not known until recently.

Prior Works: Regression under Self-Selection Biases

Efficient algorithms for *finite samples* were not known until recently.

[CDIZ23] Cherapanamjeri, Daskalakis, Ilyas, Zampetakis [STOC'23] designed a moment-based algorithm with $\mathrm{poly}(d) \cdot \exp(k/\varepsilon)$ sample complexity and running time.

Prior Works: Regression under Self-Selection Biases

Efficient algorithms for *finite samples* were not known until recently.

[CDIZ23] Cherapanamjeri, Daskalakis, Ilyas, Zampetakis [STOC'23] designed a moment-based algorithm with $\mathrm{poly}(d) \cdot \exp(k/\varepsilon)$ sample complexity and running time.

[GM24] Gaitonde and Mossel also used moments to design an algorithm with $\mathrm{poly}(d, k, 1/\varepsilon)$ sample complexity but $\mathrm{poly}(d) + (1/\varepsilon)^{\tilde{O}(k)}$ running time.

Prior Works: Regression under Self-Selection Biases

Efficient algorithms for *finite samples* were not known until recently.

[CDIZ23] Cherapanamjeri, Daskalakis, Ilyas, Zampetakis [STOC'23] designed a moment-based algorithm with $\mathrm{poly}(d) \cdot \exp(k/\varepsilon)$ sample complexity and running time.

[GM24] Gaitonde and Mossel also used moments to design an algorithm with $\mathrm{poly}(d, k, 1/\varepsilon)$ sample complexity but $\mathrm{poly}(d) + (1/\varepsilon)^{\tilde{O}(k)}$ running time.

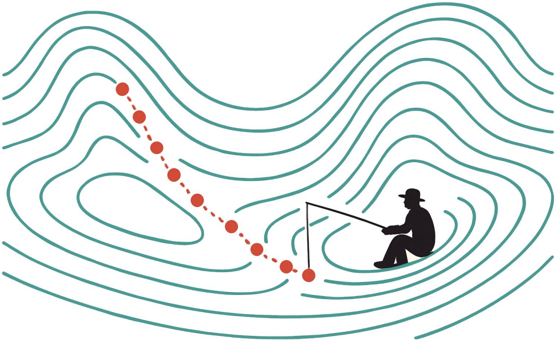
Question: Polynomial number of samples is sufficient. *Can we design an algorithm with polynomial running time?*

Table of Contents

Our Results

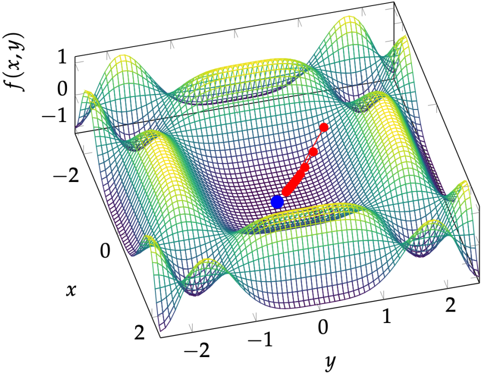
THEOREM 1 (KALAVASIS, MEHROTRA, Z. '25)

There is an algorithm for regression under self-selection bias with $\text{poly}(d, 1/\varepsilon, k)$ sample complexity and $\text{poly}(d, 1/\varepsilon) + 2^{\tilde{O}(k)}$ running time.



THEOREM 2 (KALAVASIS, MEHROTRA, Z. '25)

There is an SGD-based local convergence algorithm for regression under self-selection bias with $\text{poly}(d, 1/\varepsilon, k)$ sample complexity and running time, given a $\text{poly}(1/k)$-warm start.

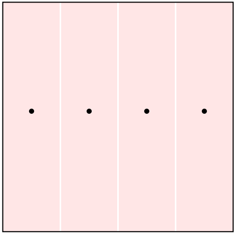


Key Idea

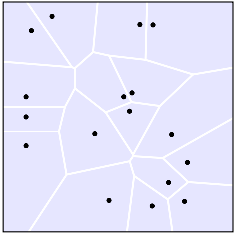
1. Unexpected connection to learning with "*coarse* observations" [Fotakis, Kalavasis, Kontonis, Tzamos, 2021].

Key Idea

1. Unexpected connection to learning with "*coarse* observations" [Fotakis, Kalavasis, Kontonis, Tzamos, 2021].
   - Simplifying example: instead of observing $z \sim \mathcal{N}(\boldsymbol{\mu}^\star, I)$, we observe a set from some given partition containing z. Can we recover $\boldsymbol{\mu}^\star$?

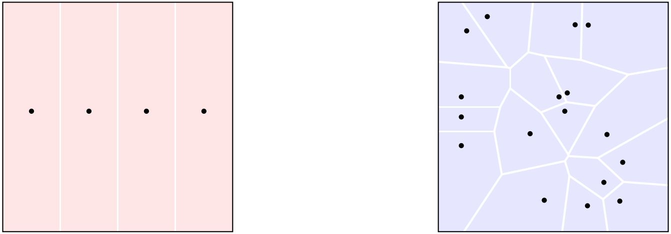


(a) Non-Identifiable Case



(b) Convex Partition Case

1. Unexpected connection to learning with "*coarse* observations" [Fotakis, Kalavasis, Kontonis, Tzamos, 2021].
    - Simplifying example: instead of observing $z \sim \mathcal{N}(\boldsymbol{\mu}^\star, I)$, we observe a set from some given partition containing z. Can we recover $\boldsymbol{\mu}^\star$?



(a) Non-Identifiable Case

(b) Convex Partition Case

1. Unexpected connection to learning with "*coarse* observations" [Fotakis, Kalavasis, Kontonis, Tzamos, 2021].
   - Simplifying example: instead of observing $z \sim \mathcal{N}(\boldsymbol{\mu}^\star, I)$, we observe a set from some given partition containing z. Can we recover $\boldsymbol{\mu}^\star$?

2. Run stochastic gradient descent (SGD) on the "coarse negative log-likelihood function."

Table of Contents

Table of Contents

Learning with Coarse Observations

Goal: recover unknown parameter $\boldsymbol{\theta}^\star$ given coarsened observations $P_1, \ldots, P_n$ from a given partition $\mathcal{P}$ of $\mathbb{R}^d$.

Learning with Coarse Observations

Goal: recover unknown parameter $\boldsymbol{\theta}^\star$ given coarsened observations $P_1, \ldots, P_n$ from a given partition $\mathcal{P}$ of $\mathbb{R}^d$.

DEFINITION 2 (COARSE LEARNING MODEL [FKKT21])

An observation $P \in \mathcal{P}$ is generated as follows:

Learning with Coarse Observations

Goal: recover unknown parameter $\boldsymbol{\theta}^\star$ given coarsened observations $P_1, \dots, P_n$ from a given partition $\mathcal{P}$ of $\mathbb{R}^d$.

DEFINITION 2 (COARSE LEARNING MODEL [FKKT21])

An observation $P \in \mathcal{P}$ is generated as follows:

1. $\boldsymbol{z} \sim q_{\boldsymbol{\theta}^\star}$.

Learning with Coarse Observations

Goal: recover unknown parameter $\boldsymbol{\theta}^\star$ given coarsened observations $P_1, \ldots, P_n$ from a given partition $\mathcal{P}$ of $\mathbb{R}^d$.

DEFINITION 2 (COARSE LEARNING MODEL [FKKT21])

An observation $P \in \mathcal{P}$ is generated as follows:

1. $\boldsymbol{z} \sim q_{\boldsymbol{\theta}^\star}$.
2. Observe unique $P \in \mathcal{P}$ s.t. $P \ni \boldsymbol{z}$.

Yale

Self-Selection Partition

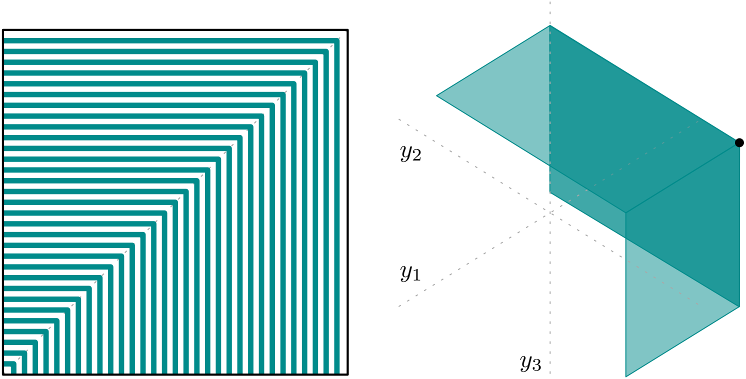
- $\boldsymbol{\theta}^\star = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)$ and $q_{\boldsymbol{\theta}^\star}$ is distribution of $\boldsymbol{z} = (\boldsymbol{x}, y^{\max})$.

Self-Selection Partition

- $\boldsymbol{\theta}^\star = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)$ and $q_{\boldsymbol{\theta}^\star}$ is distribution of $\boldsymbol{z} = (\boldsymbol{x}, y^{\max})$.
- Observing $(\boldsymbol{x}, y^{\max}) \equiv \{\boldsymbol{x}\} \times P_{y^{\max}}$ where $P_{y^{\max}} \in \mathcal{P}_{\max}$ below.

Self-Selection Partition

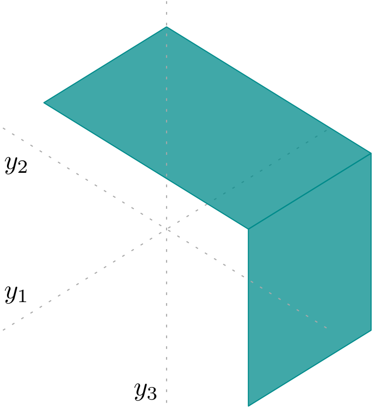
- $\boldsymbol{\theta}^\star = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)$ and $q_{\boldsymbol{\theta}^\star}$ is distribution of $\boldsymbol{z} = (\boldsymbol{x}, y^{\max})$.
- Observing $(\boldsymbol{x}, y^{\max}) \equiv \{\boldsymbol{x}\} \times P_{y^{\max}}$ where $P_{y^{\max}} \in \mathcal{P}_{\max}$ below.



Self-selection partition for $k = 2$ and a single observation for $k = 3$.

Remark

Reduction to coarse learning is general and captures other problems such as regression with "second-price auction data".



A single observation of second-price auction data for $k = 3$.

Table of Contents

Coarse Negative Log-Likelihood

Facts about the general coarse NLL

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta})$$

Coarse Negative Log-Likelihood

Facts about the general coarse NLL

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = -\mathbb{E}_{P \sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \left[\log q^{\mathcal{P}}_{\boldsymbol{\theta}}(P)\right]$$

Yale

Coarse Negative Log-Likelihood

Facts about the general coarse NLL

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = -\mathbb{E}_{P \sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \left[\log q^{\mathcal{P}}_{\boldsymbol{\theta}}(P)\right] = -\sum_{P \in \mathcal{P}} q_{\boldsymbol{\theta}^\star}(P) \cdot \log \int_{\boldsymbol{z} \in P} q_{\boldsymbol{\theta}}(\boldsymbol{z})\,.$$

Yale

Coarse Negative Log-Likelihood

Facts about the general coarse NLL

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = -\mathbb{E}_{P\sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}}\left[\log q^{\mathcal{P}}_{\boldsymbol{\theta}}(P)\right] = -\sum_{P\in\mathcal{P}} q_{\boldsymbol{\theta}^\star}(P)\cdot\log\int_{\boldsymbol{z}\in P} q_{\boldsymbol{\theta}}(\boldsymbol{z})\,.$$

- $\boldsymbol{\theta}^\star$ is a stationary point of $\mathcal{L}_{\mathcal{P}}$

Coarse Negative Log-Likelihood

Facts about the general coarse NLL

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = -\mathbb{E}_{P \sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}}\left[\log q^{\mathcal{P}}_{\boldsymbol{\theta}}(P)\right] = -\sum_{P \in \mathcal{P}} q_{\boldsymbol{\theta}^\star}(P) \cdot \log \int_{\boldsymbol{z} \in P} q_{\boldsymbol{\theta}}(\boldsymbol{z})\,.$$

- $\boldsymbol{\theta}^\star$ is a stationary point of $\mathcal{L}_{\mathcal{P}}$

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}}}[\boldsymbol{z}] - \mathbb{E}_{P \sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}} | P}\left[\boldsymbol{z}\right]\,.$$

Coarse Negative Log-Likelihood

Facts about the general coarse NLL

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = -\mathbb{E}_{P \sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \left[\log q^{\mathcal{P}}_{\boldsymbol{\theta}}(P)\right] = -\sum_{P \in \mathcal{P}} q_{\boldsymbol{\theta}^\star}(P) \cdot \log \int_{\boldsymbol{z} \in P} q_{\boldsymbol{\theta}}(\boldsymbol{z}) \,.$$

- $\boldsymbol{\theta}^\star$ is a stationary point of $\mathcal{L}_{\mathcal{P}}$

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}}}[\boldsymbol{z}] - \mathbb{E}_{P \sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}} | P}[\boldsymbol{z}] \,.$$

- $\mathcal{L}_{\mathcal{P}}$ is convex* if each $P \in \mathcal{P}$ is convex (Brascamp–Lieb Inequality)

Coarse Negative Log-Likelihood

Facts about the general coarse NLL

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = -\mathbb{E}_{P \sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}}\left[\log q^{\mathcal{P}}_{\boldsymbol{\theta}}(P)\right] = -\sum_{P \in \mathcal{P}} q_{\boldsymbol{\theta}^\star}(P) \cdot \log \int_{\boldsymbol{z} \in P} q_{\boldsymbol{\theta}}(\boldsymbol{z}) \,.$$

- $\boldsymbol{\theta}^\star$ is a stationary point of $\mathcal{L}_{\mathcal{P}}$

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}}}[\boldsymbol{z}] - \mathbb{E}_{P \sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}} | P}[\boldsymbol{z}] \,.$$

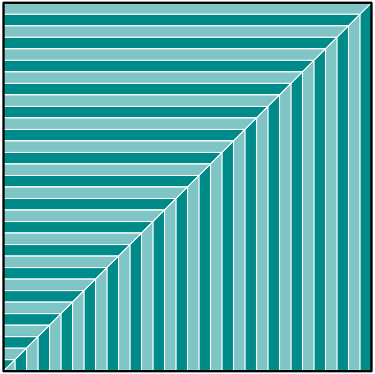
- $\mathcal{L}_{\mathcal{P}}$ is convex* if each $P \in \mathcal{P}$ is convex (Brascamp–Lieb Inequality)

$$\boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}}}[\boldsymbol{z}] - \mathbb{E}_{P \sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \mathrm{Cov}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}} | P}[\boldsymbol{z}] \,.$$

Yale

Remark

If we also observe *index* $i_{\max}$ of the regressor attaining $y^{\max}$, the partition becomes convex, and we can straightforwardly recover the efficient algorithm of [CDIZ23] for the *known-index* variant of self-selection.



Known-index self-selection partition for $k = 2$.

Table of Contents

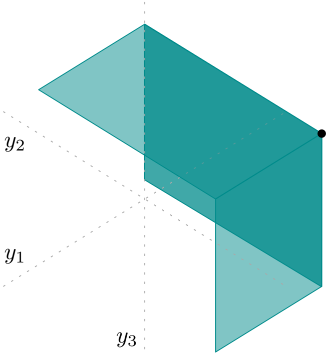
Challenges

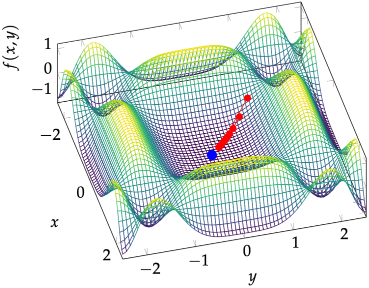
1. How can we compute $\boldsymbol{\nabla}\mathcal{L}_{\mathcal{P}_{\max}} = \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}}[\boldsymbol{z}] - \mathbb{E}_{P\sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}|P}[\boldsymbol{z}]$?

Challenges

1. How can we compute $\nabla \mathcal{L}_{\mathcal{P}_{\max}} = \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}}}[\boldsymbol{z}] - \mathbb{E}_{P \sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}} | P}[\boldsymbol{z}]$?

2. Self-selection partition $\mathcal{P}_{\max}$ is <u>not</u> convex, hence $\mathcal{L}_{\mathcal{P}_{\max}}$ potentially has many local minima.



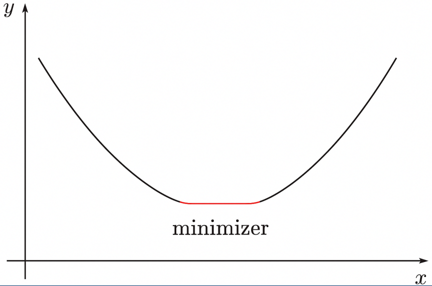


Challenges

1. How can we compute $\nabla \mathcal{L}_{\mathcal{P}_{\max}} = \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}}[\boldsymbol{z}] - \mathbb{E}_{P\sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}|P}[\boldsymbol{z}]$?

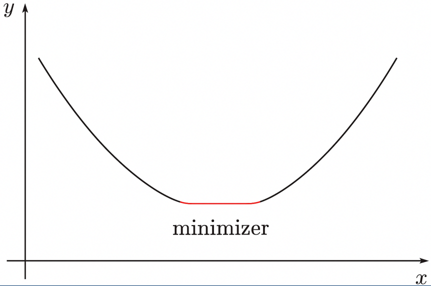
2. Self-selection partition $\mathcal{P}_{\max}$ is not convex, hence $\mathcal{L}_{\mathcal{P}_{\max}}$ potentially has many local minima.

3. $\mathcal{L}_{\mathcal{P}_{\max}}$ can be "flat" near $\boldsymbol{\theta}^\star$, and SGD may be unable to recover $\boldsymbol{\theta}^\star$.



Challenges

1. How can we compute $\nabla \mathcal{L}_{\mathcal{P}_{\max}} = \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}}[\boldsymbol{z}] - \mathbb{E}_{P\sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}|P}[\boldsymbol{z}]$?
   - Unbiased estimates given samples $x, P_{y^{\max}}$. ✓
2. Self-selection partition $\mathcal{P}_{\max}$ is <u>not</u> convex, hence $\mathcal{L}_{\mathcal{P}_{\max}}$ potentially has many local minima.
3. $\mathcal{L}_{\mathcal{P}_{\max}}$ can be "flat" near $\boldsymbol{\theta}^\star$, and SGD may be unable to recover $\boldsymbol{\theta}^\star$.



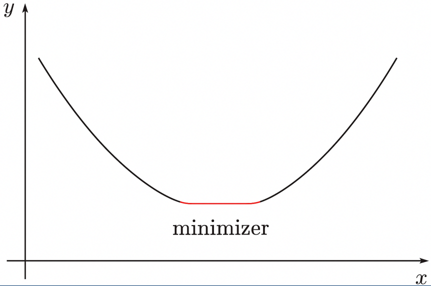
Yale

Challenges

1. How can we compute $\nabla \mathcal{L}_{\mathcal{P}_{\max}} = \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}}[\boldsymbol{z}] - \mathbb{E}_{P\sim q^{\mathcal{P}}_{\boldsymbol{\theta}^\star}} \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}|P}[\boldsymbol{z}]$?
   - Unbiased estimates given samples $x, P_{y^{\max}}$. ✓

2. Self-selection partition $\mathcal{P}_{\max}$ is <u>not</u> convex, hence $\mathcal{L}_{\mathcal{P}_{\max}}$ potentially has many local minima.
   - We show it is *locally* convex about $\boldsymbol{\theta}^\star$. ✓(with warm start)

3. $\mathcal{L}_{\mathcal{P}_{\max}}$ can be "flat" near $\boldsymbol{\theta}^\star$, and SGD may be unable to recover $\boldsymbol{\theta}^\star$.



Quadratic Growth from "Information Preservation"

Claim: It suffices to show $\mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}}) \geq \Omega(\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|)$.

Quadratic Growth from "Information Preservation"

Claim: It suffices to show $\mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}}) \geq \Omega(\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|)$.

$$\alpha \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^2 \leq \mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}})^2$$

Quadratic Growth from "Information Preservation"

Claim: It suffices to show $\mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}}) \geq \Omega(\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|)$.

$$\begin{aligned}\alpha \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^2 &\leq \mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}})^2 \\ &\leq \mathrm{KL}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}} \| q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}}) && \text{Pinsker's}\end{aligned}$$

Quadratic Growth from "Information Preservation"

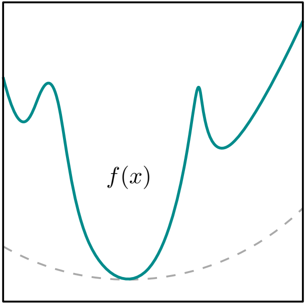
Claim: It suffices to show $\mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}}) \geq \Omega(\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|)$.

$$\begin{aligned} \alpha \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^2 &\leq \mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}})^2 \\ &\leq \mathrm{KL}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}} \| q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}}) && \text{Pinsker's} \\ &= \mathcal{L}_{\mathcal{P}_{\max}}(\boldsymbol{\theta})\,. && \text{by definition} \end{aligned}$$

Quadratic Growth from "Information Preservation"

Claim: It suffices to show $\mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}}) \geq \Omega(\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|)$.

$$
\begin{aligned}
\alpha \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^2 &\leq \mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}})^2 \\
&\leq \mathrm{KL}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}} \| q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}}) && \text{Pinsker's} \\
&= \mathcal{L}_{\mathcal{P}_{\max}}(\boldsymbol{\theta})\,. && \text{by definition}
\end{aligned}
$$

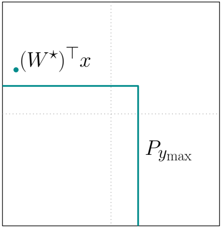


Proof Intuition of "Information Preservation"

- Identity an event $\mathcal{E}$ such that $\boldsymbol{w}_{i_{\max}}^{\top}\boldsymbol{x} \gg \boldsymbol{w}_j^{\top}\boldsymbol{x}$ for all $j \neq i_{\max}$.

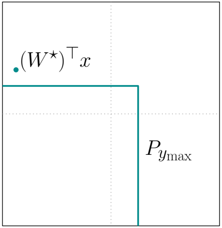
Proof Intuition of "Information Preservation"

- Identity an event $\mathcal{E}$ such that $\boldsymbol{w}_{i_{\max}}^\top \boldsymbol{x} \gg \boldsymbol{w}_j^\top \boldsymbol{x}$ for all $j \neq i_{\max}$.
- Conditional on $\mathcal{E}$, $P_{y^{\max}}$ "looks" like a convex set under $q_{\boldsymbol{\theta}^\star}$.



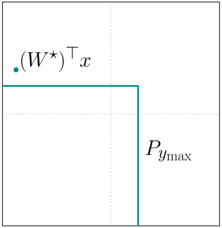
Proof Intuition of "Information Preservation"

- Identity an event $\mathcal{E}$ such that $\boldsymbol{w}_{i_{\max}}^\top \boldsymbol{x} \gg \boldsymbol{w}_j^\top \boldsymbol{x}$ for all $j \neq i_{\max}$.
- Conditional on $\mathcal{E}$, $P_{y^{\max}}$ "looks" like a convex set under $q_{\boldsymbol{\theta}^\star}$.
- Conditional on $\mathcal{E}$ and under regularity conditions, prove that $\mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}} \mid \mathcal{E}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}} \mid \mathcal{E}) \geq \Omega(\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|)$.



Proof Intuition of "Information Preservation"

- Identity an event $\mathcal{E}$ such that $\boldsymbol{w}_{i_{\max}}^\top \boldsymbol{x} \gg \boldsymbol{w}_j^\top \boldsymbol{x}$ for all $j \neq i_{\max}$.
- Conditional on $\mathcal{E}$, $P_{y^{\max}}$ "looks" like a convex set under $q_{\boldsymbol{\theta}^\star}$.
- Conditional on $\mathcal{E}$ and under regularity conditions, prove that $\mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}} \mid \mathcal{E}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}} \mid \mathcal{E}) \geq \Omega(\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|)$.
- Conclude using the fact $\mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}}) \geq \Pr[\mathcal{E}] \cdot \mathrm{TV}(q_{\boldsymbol{\theta}}^{\mathcal{P}_{\max}} \mid \mathcal{E}, q_{\boldsymbol{\theta}^\star}^{\mathcal{P}_{\max}} \mid \mathcal{E})$.



Conclusion

- Study the geometry of regression with self-selection bias through the lenses of the coarse learning framework.

Conclusion

- Study the geometry of regression with self-selection bias through the lenses of the coarse learning framework.
- Leads to an SGD-based local convergence algorithm, which improves on the running time of [GM24].

Conclusion

- Study the geometry of regression with self-selection bias through the lenses of the coarse learning framework.
- Leads to an SGD-based local convergence algorithm, which improves on the running time of [GM24].

Is there a fully-polynomial (SGD-based) algorithm for regression with self-selection bias?

That's All!

felix-zhou.com

felix.zhou@yale.edu