what makes a good fisherman?

Constantinos Daskalakis
EECS & CSAIL, MIT
what makes a good fisherman?

“I bet you have to understand the Poisson distribution”

Gautam Kamath
what makes a good fisherman?

“A very deep net?”

Clément Canonne
what makes a good fisherman?

“Or a wide one :)

Rohan Sukumaran
what makes a good fisherman?

“attention is all you need?”

Kwang-Sung Jun
what makes a good fisherman?

our work

Yeshwanth Cherapanamjeri
UC Berkeley

Andrew Ilyas
MIT

Manolis Zampetakis
UC Berkeley

Homonymous paper: https://arxiv.org/abs/2205.03246

Brief Mention: https://arxiv.org/abs/2205.02060
Estimation of Standard Auction Models (EC’22)
What makes a good fisherman?

a statistical approach towards an answer

• Collect data: \( \{(x^{(j)}, y^{(j)})\}_{j=1}^{n} \) where
  - \( x^{(j)} \): features of individual \( j \) (e.g. height, weight, speed, training)
  - \( y^{(j)} \): daily catch of individual \( j \)

• Fit model: \( y \sim f_\theta(x) \), where \( \{f_\theta\}_\theta \) is some distribution family
  - e.g. linear regression: \( y = x^T w + \eta \), where \( \eta \sim \mathcal{N}(0, \sigma^2) \)

• Issue with this approach?
  - missing data from all *unrealized* fishermen
  - “self-selection bias”
Self-selection bias in a village

[Roy’51]

be a fisherman or be a hunter?

Feature vector

\[ x = \begin{bmatrix} 
\text{height} \\
\text{weight} \\
\vdots \\
\text{speed} 
\end{bmatrix} \]

Tries hunting

Income from hunting

\[ y_{\text{hunt}} = x^T w_{\text{hunt}} + \mathcal{N}(0, \sigma_{\text{hunt}}^2) \]

Tries fishing

Income from fishing

\[ y_{\text{fish}} = x^T w_{\text{fish}} + \mathcal{N}(0, \sigma_{\text{fish}}^2) \]
Self-selection bias in a village

[Roy’51]

be a fisherman or be a hunter?

Income from hunting

\[ y_{\text{hunt}} = x^T w_{\text{hunt}} + \mathcal{N}(0, \sigma_{\text{hunt}}^2) \]

Income from fishing

\[ y_{\text{fish}} = x^T w_{\text{fish}} + \mathcal{N}(0, \sigma_{\text{fish}}^2) \]

Feature vector

\[ x = \begin{bmatrix} \text{height} \\ \text{weight} \\ \vdots \\ \text{speed} \end{bmatrix} \]
Self-selection bias in a village

[Roy’51]

be a fisherman or be a hunter?

Income from hunting

\[ y_{\text{hunt}} = x^\top w_{\text{hunt}} + \mathcal{N}(0, \sigma_{\text{hunt}}^2) \]

Income from fishing

\[ y_{\text{fish}} = x^\top w_{\text{fish}} + \mathcal{N}(0, \sigma_{\text{fish}}^2) \]

Feature vector

\[ x = \begin{bmatrix} \text{height} \\ \text{weight} \\ \vdots \\ \text{speed} \end{bmatrix} \]
Self-selection bias in a village

[Roy’51]

be a fisherman or be a hunter?

Tries hunting

Income from hunting

\[ y_{\text{hunt}} = x^T w_{\text{hunt}} + \mathcal{N}(0, \sigma_{\text{hunt}}^2) \]

Tries fishing

Income from fishing

\[ y_{\text{fish}} = x^T w_{\text{fish}} + \mathcal{N}(0, \sigma_{\text{fish}}^2) \]
Self-selection bias in a village  

[Roy’51]  
be a fisherman or be a hunter?

- **Fishing income** of individual with features $x \in \mathbb{R}^d$:  
  $$I_f(x) = x^T w_f + \eta_f$$  
  where $\eta_f \sim \mathcal{N}(0, \sigma_f^2)$ is idiosyncratic shock independent of everything

- **Hunting income** of individual with features $x \in \mathbb{R}^d$:  
  $$I_h(x) = x^T w_h + \eta_h$$  
  where $\eta_h \sim \mathcal{N}(0, \sigma_h^2)$ is idiosyncratic shock independent of everything

- **Strategic optimization**: choose profession $p(x)$ w/ largest income $I(x)$  
  $$I(x) = \max\{I_f(x), I_h(x)\}$$  
  $$p(x) = \arg\max\{I_f(x), I_h(x)\}$$

- **Observations**: $\{(x^{(j)}, I(x^{(j)}), p(x^{(j)}))\}_{j=1}^n$

- **Goal**: estimate $w_f, w_h$

**In other settings:**
- think different professions (e.g. programmers vs nurses)
- think different majors at a university (e.g. math vs music)
- think choosing best out of more than two professions, majors, …
Example 2: Selective SAT Score Reporting College Admissions

- **SAT subject test scores** of individual with features \( x \in \mathbb{R}^d \) for subjects A and B:
  \[
  S_A(x) = f_{\theta_A}(x) + \eta_A \\
  S_B(x) = f_{\theta_B}(x) + \eta_B
  \]
  where \( \eta_A, \eta_B \) are zero-mean idiosyncratic shocks

- **Historical data** for each test:
  \( p_A(S_A) \): probability of being accepted to college when submitting subject A score
  \( p_B(S_B) \): probability of being accepted to college when submitting subject B score

- **Selective reporting:** choose which score to report to maximize probability of acceptance
  \[
  S(x) = \arg\max \{ p_A(S_A(x)), p_B(S_B(x)) \} \\
  p(x) = \max \{ p_A(S_A(x)), p_B(S_B(x)) \}
  \]

- **Observations:** \( \{(x^{(j)}, S(x^{(j)}), p(x^{(j)}))\}_{j=1}^n \)

- **Goal:** estimate \( \theta_A, \theta_B \)
Example 3: Market Disequilibrium

[Fair & Jaffee’72]

- **Supply** on houses with features $x \in \mathbb{R}^d$:
  
  $S(x) = x^T w_S + \eta_S,$

  where $\eta_S \sim \mathcal{N}(0, \sigma_S^2)$ is idiosyncratic shock independent of everything

- **Demand** on houses with features $x \in \mathbb{R}^d$:
  
  $D(x) = x^T w_D + \eta_D,$

  where $\eta_D \sim \mathcal{N}(0, \sigma_D^2)$ is idiosyncratic shock independent of everything

- **Market Disequilibrium**: $S(x) \neq D(x)$
  
  houses w/ features $x$ closing: $C(x) = \min\{S(x), D(x)\}$

- **Observations**: $\{(x^{(j)}, C(x^{(j)}))\}_{j=1}^n$
  
  do not observe whether sell or buy side is tight at $x^{(j)}$

- **Goal**: estimate $w_S, w_D$
Example 4: Non-parametric setting

Estimating Auction Models [Guerre-Perrigne-Vuong’00; Athey-Haile’02;...]

• **Repeated** single-item auction involving $k$ populations of bidders
• **Asymmetric independent private values (IPV):**
  • values $v_1(t) \sim F_1, ..., v_k(t) \sim F_k$
  • independent across time and bidders
• Assume **Bayesian Nash equilibrium:**
  • bids $b_1(t) \sim G_1, ..., b_k(t) \sim G_k$
  • Independent across time and bidders
• **Observations:** $\{(i(t), p(t))\}_{t=1}^n$
  • i.e. only observe winner and price paid
• First-price auction: $i(t) = \arg\max_i\{b_i(t)\}, p(t) = \max_i\{b_i(t)\}$
• Second-price auction: $i(t) = \arg\max_i\{b_i(t)\}, p(t) = \max_{i \neq i(t)}\{b_i(t)\}$
• **Goal:** non-parametric bid distribution estimation
  (and maybe also value distribution estimation)
Self-Selection Bias

\[ Z \rightarrow \mathcal{O}(Z) \]

Observed data \( \mathcal{O}(Z) \) is not all underlying data \( Z \) of interest, but the output of some strategic process which operated on \( Z \) and selected some of the data.

\[ Z_1 \rightarrow \mathcal{O}(Z_1) \]

\[ Z_2 \rightarrow \mathcal{O}(Z_2) \]

\[ Z_k \rightarrow \mathcal{O}(Z_1, \ldots, Z_k) \]
Self-Selection Bias

Theory and Applications

- participation in the labor force [Roy’51; Heckman’74,’79; Nelson’77; Cogan’14; Hanoch’14; Hanoch-Smith’14]
- retirement decisions [Gordon-Blinder’80]
- returns to education [Griliches-Hall-Hausman’78; Kenny-Lee-Maddala-Trost’79; Willis-Rosen’79]
- effects of unions on wages [Lee’78; Abowd-Farber’82]
- migration and income [Nakosteen-Zimmer’80; Borjas’87]
- physician and lawyer behavior [Poirier’81; Weisbrod’83]
- tenure choice and the demand for housing [Lee-Trost’78; Rosen’79; King’80]
- market disequilibrium models [Fair & Jaffee’72, Goldfeld-Quandt’75]
- identification of auction models under partial observability [Guerre-Perrigne-Vuong’00; Athey-Haile’02,’07]
- [Maddala’86; Cameron-Trivedi’05; Brooks’19] textbook introductions
- **intimate relationship to:** max affine regression, mixture of regressions, logit/probit regression, truncated regression
- **theoretical understanding:** mostly asymptotic sample regime, identification results in some settings
This Talk

- Linear Regression
  - Known index (efficient algorithm)
  - Unknown index (identifiability + efficient algorithm)
- Non-parametric Density Estimation
  - efficient algorithm for auctions
- Future directions
This Talk

- **Linear Regression**
  - Known index (efficient algorithm)
  - Unknown index (identifiability + efficient algorithm)
- Non-parametric Density Estimation
  - efficient algorithm for auctions
- Future directions
Data Generation Process:

1. Sample covariates $x \sim \mathcal{D}_x$ (e.g. individual’s features)

2. Compute potential outcomes $y_i = w_i^T x + \eta_i$, (e.g. skills) where $\eta_i$ mean-zero noise independent of everything

3. Select $i_* = S(y_1, \ldots, y_k)$ (e.g. argmax)

4. Add $(x, i_*, y_{i_*})$ to training set OR add $(x, y_{i_*})$ to training set
   - known-index setting
   - unknown-index setting

Linear Regression w/ Self-Selection

$w_1, w_2, \ldots, w_k \in \mathbb{R}^d$

Ground-truth weights (unknown)

$S(\cdot): \mathbb{R}^k \rightarrow [k]$

Selection function (known)
Linear Regression w/ Self-Selection

comparison to mixture of linear regressions

• Linear regression w/ self-selection:
  • Sample $x \sim \mathcal{D}_x$
  • Compute $y_i = w_i^T x + \eta_i$, for all $i = 1, \ldots, k$
  • Select $i_\ast = S(y_1, \ldots, y_k)$
  • Output $(x, y_{i_\ast})$ and perhaps also $i_\ast$

  Non-trivial even when $i_\ast$ shown
  Choice of which $i_\ast$ is shown is endogenous

• Mixtures of linear regressions [Day'69, Kiefer'78, Aitkin-Wilson'80, De Veaux'89, Jordan-Jacobs'94,…, Li-Liang'18, Kwon-Caramanis’20, Chen-Li-Song’20, Diakonikolas-Kane’20]:
  • Sample $x \sim \mathcal{D}_x$
  • Compute $y_i = w_i^T x + \eta_i$, for all $i = 1, \ldots, k$
  • Select $i_\ast = \{i, \text{with probability } \pi_i\}$
  • Output $(x, y_{i_\ast})$

  Trivial if $i_\ast$ were revealed ($k$ independent OLSs)
  Choice of which $i_\ast$ is shown is exogenous
Linear Regression w/ Self-Selection

comparison to max-affine regression, probit regression

• Linear regression w/ self-selection:
  • Sample $x \sim \mathcal{D}_x$
  • Compute $y_i = w_i^T x + \eta_i$, for all $i = 1, \ldots, k$
  • Select $i_* = S(y_1, \ldots, y_k)$
  • Output $(x, y_{i_*})$ and perhaps also $i_*$

• Probit Regression
  • Sample $x \sim \mathcal{D}_x$
  • Compute $y_i = w_i^T x + \eta_i$, for all $i = 1, \ldots, k$
  • Select $i_* = \text{argmax}(y_1, \ldots, y_k)$
  • Output $(x, i_*)$

When all of $(x, i_*, y_{i_*})$ observed, observe more information compared to probit regression

Want to accommodate more general selection rules $S(\cdot)$

Want finite rates

Maximum likelihood is convex wrt $w_i - w_j$ differences

No finite rates are known
Linear Regression w/ Self-Selection

\[ w_1, w_2, \ldots, w_k \in \mathbb{R}^d \]

Ground-truth weights \textit{(unknown)}

\[ S(\cdot): \mathbb{R}^k \rightarrow [k] \]
Selection function \textit{(known)}

Data Generation Process:

1. Sample covariates \( x \sim D_x \) (e.g. individual's features)

2. Compute potential outcomes \( y_i = w_i^T x + \eta_i \) (e.g. skills)
   where \( \eta_i \) mean-zero noise independent of everything

3. Select \( i_\ast = S(y_1, \ldots, y_k) \) (e.g. argmax)

4. Add \( (x, i_\ast, y_{i_\ast}) \) to training set \textit{known-index setting}
   OR add \( (x, y_{i_\ast}) \) to training set \textit{unknown-index setting}
Result: Known-Index Setting

**Theorem:** Under some mild assumptions (discussed soon), with \( n \) observations \( \{ (x^{(j)}, y_{i_*}^{(j)}, i_*^{(j)}) \}_j \) we can construct estimates \( \{ \hat{w}_1, \ldots, \hat{w}_k \} \) that approximate the true parameters at a rate of

\[
\max_{i \in [k]} \| \hat{w}_i - w_i \|_2^2 \leq O \left( \frac{\log(n)}{n} \right),
\]

where \( O(\cdot) \) hides polynomial dependencies in other problem parameters (discussed soon).

The running time is also polynomial in \( n \) and problem parameters (discussed soon).
Estimation: Known-Index Setting

Assumption 1: Regression Assumptions

Standard assumptions on feature and weight vectors:

(i) *Bounded feature norm:* constant $C$ for which $\| x^{(j)} \|_2 \leq C$

(ii) *Bounded weight norm:* constant $B$ for which $\| w_i \|_2 \leq B$

(iii) *Covariate thickness:* $\frac{1}{n} \sum_{j=1}^{n} x^{(j)} x^{(j)} \succeq I_d$
Assumption 2: Survival Probability

If $\mathbb{P}(i^* = \ell) \approx 0$, impossible to efficiently estimate $W_\ell$
(e.g., a village with no hunters!)

Assumption: There exists a constant $\alpha > 0$ such that every potential outcome is observed with probability at least $\alpha/k$
(parametrize sample/time complexity w.r.t. $a$ - will be poly$(1/\alpha)$)
Estimation: Known-Index Setting

Assumption 3: Convex self-selecting sets

In our village example, we observe the $\arg\max\{y_1, y_2\}$:

$$\arg\max\{y_1, y_2\}$$

But we can handle more complex selection functions:

**Our assumption:** Convex-inducing slices

i.e. for all $i$, all values of $y_i$, the set of $y_{-i}$ such that

$$i = S(y_i, y_{-i})$$

is convex

**Satisfied by:**

- maximum function (e.g. choosing profession)
- for any monotonic $f_1, \ldots, f_k$: $\arg\max_{i \in [k]} f_i(y_i)$ (e.g. SAT score reporting)
**Result: Known-Index Setting**

**Theorem:** Under the discussed assumptions, with \( n \) observations \( \{(x^{(j)}, y_{i^*}^{(j)}, i_{i^*}^{(j)})\}_j \) we can construct estimates \( \{\hat{w}_1, \ldots, \hat{w}_k\} \) that approximate the true parameters at a rate of

\[
\max_{i \in [k]} \| \hat{w}_i - w_i \|_2^2 \leq \text{poly}(k, 1/\alpha, \sigma, 1/\sigma, B, C) \cdot \frac{\log(n)}{n}.
\]

The running time is \( \text{poly}(n, d, k, 1/\alpha, \sigma, 1/\sigma, B, C) \).
Approach: Known-Index Setting

Write objective function based on population log-likelihood

\[ \bar{\ell}(W) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{(y_{i_*}, i_*) \sim P_{W^*}(\cdot | x = x(j))} \left[ \log P_{W}(y_{i_*}, i_* | x = x(j)) \right] \]

GOALs:
- show \( \bar{\ell}(W) \) strongly concave
- show unique maximum of \( \bar{\ell}(W) \) occurs at \( W = W^* \)
- show near-optimum of \( \bar{\ell}(W) \) can be computed, even though we do not have infinitely many samples
- show that this can be done efficiently
Approach: Known-Index Setting

Write objective function based on population log-likelihood

\[
\overline{\ell}(W) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{(y_{i_*}, i_*) \sim Pr_{W^*}(\cdot|x=x(j))}[\log Pr_W(y_{i_*}, i_* | x = x(j))] 
\]

Gradient of objective function:

\[
\nabla_W \overline{\ell}(W) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{(y_{i_*}, i_*) \sim Pr_{W^*}(\cdot|x=x(j))}[\cdots] - \\
\mathbb{E}_{(y_{i_*}, i_*) \sim Pr_{W^*}(\cdot|x=x(j))}[\mathbb{E}_{y_{-i_*} \sim Pr_{W}(\cdot|S_i(y_{i_*}))}[\cdots]] 
\]

Hessian:

\[
\nabla_W^2 \overline{\ell}(W) \leq -\Omega(\alpha/\sigma k \cdot I) 
\]

(uses slice convexity)

(\text{so } \overline{\ell}(W) \text{ strongly convex})

Optimality:

\[
\nabla_W \overline{\ell}(W^*) = 0 
\]

(easy)

(\text{so } W^* \text{ global optimum})

I have a sample for the outer distribution and can simulate samples from the inner one to get an unbiased estimate of this term

slice of \(S\): all \(y_{-i_*}\) such that \(i^* = S(y_{i_*}, y_{-i_*})\)
Approach: Known-Index Setting

Gradient of objective function:

\[ \nabla_W \bar{\ell}(W) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}(y_{i^*}, i^*_j) \sim Pr_{W^*}(\cdot | x = x(j)) \left[ \cdots \right] - \]

\[ \mathbb{E}(y_{i^*}, i^*_j) \sim Pr_{W^*}(\cdot | x = x(j)) \left[ \mathbb{E} y_{i^*_j} \sim Pr_{W}(\cdot | S_{i^*_j}(y_{i^*_j})) \left[ \cdots \right] \right] \]

I have a sample for the outer distribution and can simulate samples from the inner one to get an unbiased estimate of this term.

I have a sample for the outer distribution and can simulate samples from the inner one to get an unbiased estimate of this term.

Perform Projected Stochastic Gradient Descent:

\[ W_0, W_t, W_\infty = W^* \]

Challenge: sampling from conditional distribution \( Pr_w(\cdot | S_{i^*_j}(y_{i^*_j})) \) using rejection sampling might take exponential time.

It might be that under \( W \) it is very unlikely that \( i^*_j \) is selected when it has value \( y_{i^*_j} \).

Solution: Langevin dynamics
This Talk

- Linear Regression
  - Known index (efficient algorithm)
  - Unknown index (identifiability + efficient algorithm)
- Non-parametric Density Estimation
  - efficient algorithm for auctions
- Future directions
This Talk

- Linear Regression
  - Known index (efficient algorithm)
  - Unknown index (identifiability + efficient algorithm)
- Non-parametric Density Estimation
  - efficient algorithm for auctions
- Future directions
Unknown-Index Setting

Data Generation Process: For $j = 1, \ldots, n$

1. Sample covariates $x^{(j)} \sim \mathcal{D}$
2. Compute potential outcomes $y_i = w_i^T x^{(j)} + \eta_i$, for $i = 1, \ldots, k$
3. Select $i_* = S(y_1, \ldots, y_k)$; set $y^{(j)} = y_{i_*}$
4. Add $(x^{(j)}, y^{(j)})$ to training set (\(i_*\) is hidden)

Challenge: no more concave log-likelihood!
(In fact, even identifiability is not obvious)

Our Results: identification results for max selector, general $k$
polynomial time/sample complexity for $k = 2$
This Talk

- Linear Regression
  - Known index (efficient algorithm)
  - Unknown index (identifiability + efficient algorithm)
- Non-parametric Density Estimation
  - efficient algorithm for auctions
- Future directions
This Talk

- Linear Regression
  - Known index (efficient algorithm)
  - Unknown index (identifiability + efficient algorithm)
- Non-parametric Density Estimation
  - efficient algorithm for auctions
- Future directions
Setting: Non-Parametric Case

Instead of $k$ linear models, we have $k$ distributions $\{F_1, \ldots, F_k\}$

$$y_i \sim F_i \quad \implies \quad i^* = \arg\max y_i \quad \implies \quad y = y_{i^*}$$

Example: (repeated) first-price auctions under independent private values
Results: Non-Parametric Case

**Theorem**: Can compute estimate $F_i$ such that $\mathcal{W}(F_i, \hat{F}_i) \leq \varepsilon$

w.p. $1 - \delta$ using $O\left(\left(\frac{2}{\lambda \varepsilon}\right)^{4k} \frac{\log(1/\delta)}{\varepsilon^2}\right)$ samples

**Also in our paper**: Results for second-price auctions, value distribution estimation

(Exponential dependence inevitable, don’t observe bids near zero)

If we only care about the “effective support” (i.e., range $[p, 1]$ with constant probability $\gamma$ of observing winning bid):

**Theorem**: The same algorithm yields $\sup_{y \in [p, 1]} |F_i(y) - \hat{F}_i(y)| \leq \varepsilon$

w.p. $1 - \delta$ using $O\left(\frac{\log(k/\delta)}{\gamma^4 \varepsilon^2}\right)$ samples

Also in our paper:

Results for second-price auctions, value distribution estimation

Extends line of work in Econometrics where only identification results had been obtained [Athey-Haile’02] unless the setting is symmetric [Morganti’11, Menzel-Morganti’13, Guerre-Perrigne-Vuong’00] or parametric [Donald-Paarsch’96, Athey-Haile’06, Athey-Levin-Seira’11]
Summary: Contributions

- Estimation results for linear-regression w/ self-selection bias under both known and unknown index

- Estimation results for non-parametric density estimation w/ self-selection bias, and applications to estimating auction models

- **Future directions:** estimation problem is wide open
  
  - Beyond linear regression
  
  - Unknown/less structured noise model
  
  - Unknown selection rules
Estimation Question (today): Observed data $\mathcal{O}(\mathbf{Z})$ is not all underlying data $\mathbf{Z}$ of interest, but the output of some strategic process which operated on $\mathbf{Z}$.

Normative Question: What learning/optimization procedure should agents run in such an environment to map observations to decisions?

- when utilities are concave: many answers in GT and Online Learning
- when utilities are not concave, e.g. because strategies are compactly represented by DNNs?

[Daskalakis-Skoulakis-Zampetakis’21]: Even local Nash equilibria are intractable (even in two-player zero-sum case, even if game is perfectly known).

need to rethink Deep Learning in the multi-agent setting!