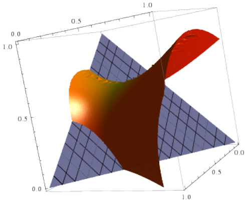
Multi-Observation Losses



Bo Waggoner
Microsoft Research, NYC

Columbia
April 4, 2019

Based on joint work with Rafael Frongillo (U. Colorado, Boulder), Tom Morgan (Harvard), Sebastian Casalaina-Martin (U. Colorado, Boulder), Nishant Mehta (U. Victoria).

$$\operatorname*{argmin}_{r \in \mathcal{R}} \mathop{\mathbb{E}}_{y \sim p} \ell(r, y)$$

$$\underset{r \in \mathcal{R}}{\operatorname{argmin}} \underset{\substack{y_1, y_2 \sim p \\ \text{i.i.d.}}}{\mathbb{E}} \ell(r, y_1, y_2)$$

$$\operatorname*{argmin}_{r \in \mathcal{R}} \mathop{\mathbb{E}}_{\substack{y_1, \ldots, y_m \sim p \\ \text{i.i.d.}}} \ell(r, y_1, \ldots, y_m)$$

Outline

1. Background: information elicitation
   what do you get when you minimize a loss?
2. Paper 1: Multi-Observation Elicitation (COLT 2017)
   what changes with multi-observation losses?
3. Paper 2: Multi-Observation Regression (AISTATS 2019)
   what ML problems can they solve?

Background: information elicitation

What do you get when you minimize a loss?

$$\Gamma(p) := \operatorname*{argmin}_{r \in \mathcal{R}} \mathop{\mathbb{E}}_{y \sim p} \ell(r, y) \tag{1}$$

Examples:

- $\ell(r, y) = (r - y)^2$

Background: information elicitation

What do you get when you minimize a loss?

$$\Gamma(p) := \operatorname*{argmin}_{r \in \mathcal{R}} \mathop{\mathbb{E}}_{y \sim p} \ell(r, y) \tag{1}$$

Examples:

- $\ell(r, y) = (r - y)^2$ $\quad\Gamma(p) = \mathbb{E}_{y \sim p}\, y$ *(mean)*

Background: information elicitation

What do you get when you minimize a loss?

$$\Gamma(p) := \operatorname*{argmin}_{r \in \mathcal{R}} \mathop{\mathbb{E}}_{y \sim p} \ell(r, y) \tag{1}$$

Examples:

- $\ell(r, y) = (r - y)^2$ $\qquad \Gamma(p) = \mathbb{E}_{y \sim p}\, y$ *(mean)*
- $\ell(r, y) = |r - y|$

Background: information elicitation

What do you get when you minimize a loss?

$$\Gamma(p) := \operatorname*{argmin}_{r \in \mathcal{R}} \mathop{\mathbb{E}}_{y \sim p} \ell(r, y) \tag{1}$$

Examples:

- $\ell(r, y) = (r - y)^2$ — $\Gamma(p) = \mathbb{E}_{y \sim p}\, y$ *(mean)*
- $\ell(r, y) = |r - y|$ — *median*

Background: information elicitation

What do you get when you minimize a loss?

$$\Gamma(p) := \underset{r \in \mathcal{R}}{\operatorname{argmin}} \underset{y \sim p}{\mathbb{E}} \ell(r, y) \tag{1}$$

Examples:

- $\ell(r, y) = (r - y)^2$ — $\Gamma(p) = \mathbb{E}_{y \sim p}\, y$ *(mean)*
- $\ell(r, y) = |r - y|$ — *median*
- $\ell(r, y) = \begin{cases} 0 & r = y \\ 1 & \text{otherwise} \end{cases}$

Background: information elicitation

What do you get when you minimize a loss?

$$\Gamma(p) := \operatorname*{argmin}_{r\in\mathcal{R}} \mathop{\mathbb{E}}_{y\sim p} \ell(r,y) \tag{1}$$

Examples:

- $\ell(r,y) = (r-y)^2$ — $\Gamma(p) = \mathbb{E}_{y\sim p}\, y$ *(mean)*
- $\ell(r,y) = |r-y|$ — *median*
- $\ell(r,y) = \begin{cases} 0 & r = y \\ 1 & \text{otherwise} \end{cases}$ — *mode*

Background: information elicitation

What do you get when you minimize a loss?

$$\Gamma(p) := \operatorname*{argmin}_{r \in \mathcal{R}} \mathop{\mathbb{E}}_{y \sim p} \ell(r, y) \tag{1}$$

Examples:

- $\ell(r, y) = (r - y)^2$ $\Gamma(p) = \mathbb{E}_{y \sim p}\, y$ *(mean)*
- $\ell(r, y) = |r - y|$ *median*
- $\ell(r, y) = \begin{cases} 0 & r = y \\ 1 & \text{otherwise} \end{cases}$ *mode*
- $\ell(r, y) = ??$ *variance*

Background: information elicitation

What do you get when you minimize a loss?

$$\Gamma(p) := \operatorname*{argmin}_{r \in \mathcal{R}} \mathop{\mathbb{E}}_{y \sim p} \ell(r, y) \tag{1}$$

- $\Gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}}$ is a **property** of the distribution p
- Γ is **elicitable** if there exists ℓ such that (1) holds

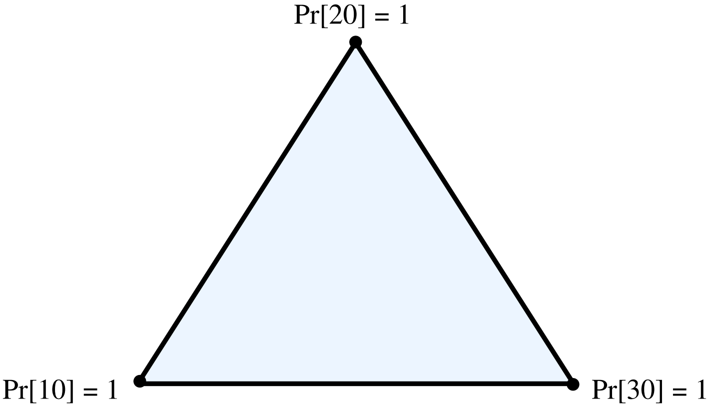
The variance is not elicitable

Proposition (Folklore)

There is no loss function that elicits the variance of p.

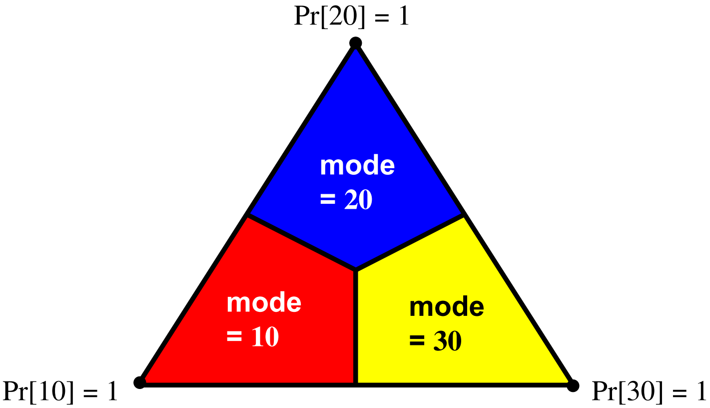
Information elicitation - the picture

The simplex $\Delta_{\mathcal{Y}}$ for $\mathcal{Y} = \{10, 20, 30\}$:



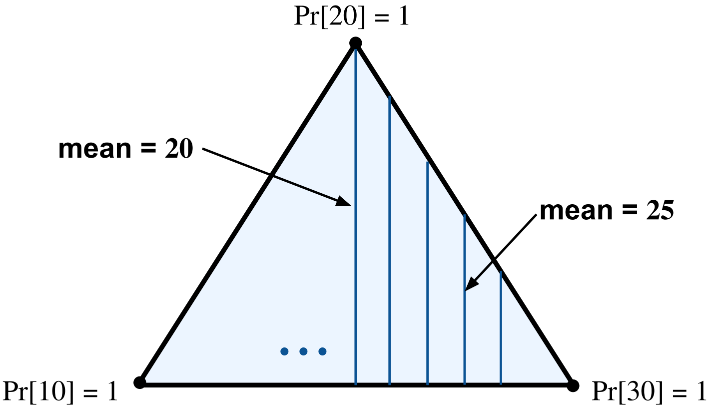
Information elicitation - the picture

- A property is a partition of the simplex.
- The **level set** of r is $\{p : \Gamma(p) = r\}$.



Information elicitation - the picture

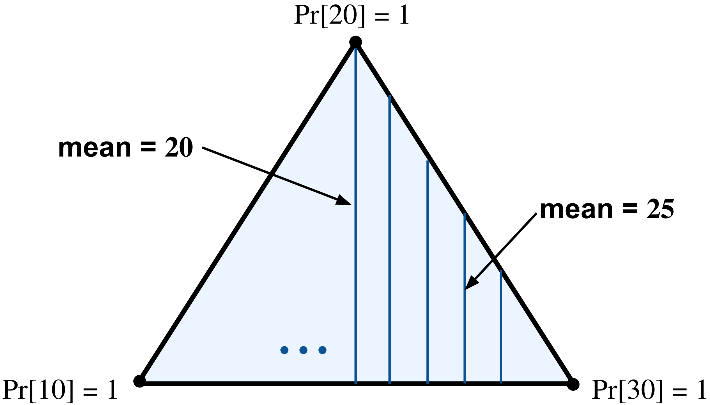
- A property is a partition of the simplex.
- The **level set** of r is $\{p : \Gamma(p) = r\}$.



Key basic fact

Theorem

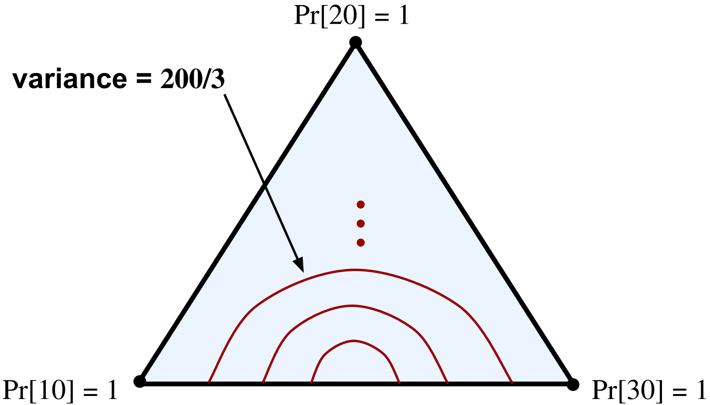
If a property is elicitable, then all of its level sets are convex sets.



Key basic fact

Theorem

If a property is elicitable, then all of its level sets are convex sets.



Non-elicitable properties

Known: there is no loss function eliciting the variance.
Suggestions?

[1] e.g. Frongillo and Kash, 2015

Non-elicitable properties

Known: there is no loss function eliciting the variance.
Suggestions?

Indirect elicitation: elicit **mean** and **second moment**, then calculate.
$\implies$ the **elicitation complexity**[1] of the variance is 2.

[1] e.g. Frongillo and Kash, 2015

Non-elicitable properties

Known: there is no loss function eliciting the variance.
Suggestions?

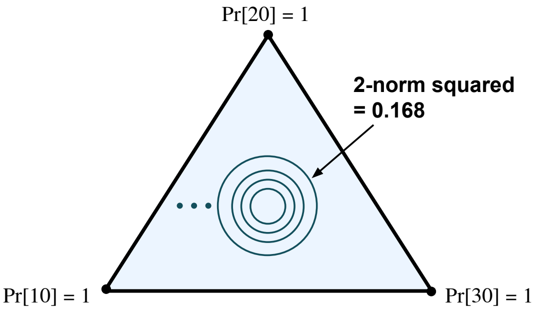
Indirect elicitation: elicit **mean** and **second moment**, then calculate.
$\implies$ the **elicitation complexity**[1] of the variance is 2.

Note: always possible to elicit entire distribution and calculate.
$\implies$ elicitation complexity $\leq |\mathcal{Y}| - 1$ for all properties.

[1] e.g. Frongillo and Kash, 2015

Final case study: 2-norm

Consider $\Gamma(p) = \|p\|_2^2 = \sum_y p_y^2$. *Measures non-uniformity of p*



Fact: [Frongillo and Kash, 2015] The elicitation complexity of the 2-norm is $|\mathcal{Y}| - 1$.

Paper 1: (im)possibilities

Multi-Observation Elicitation. COLT 2017.

Casalaina-Martin, Frongillo, Morgan, Waggoner.

Paper 1: (im)possibilities

Multi-Observation Elicitation. COLT 2017.

Casalaina-Martin, Frongillo, Morgan, Waggoner.

Goals:

- Propose multi-observation losses.
- Give **upper bounds** avoiding prior impossibilities.
- Develop theory of losses from algebraic geometry.
- Use it to prove **lower bounds**.

Example 1: Variance

Claim 1: Let
$$f(y_1, y_2) = \frac{1}{2}\left(y_1 - y_2\right)^2 .$$
Then $\mathbb{E}_{y_1, y_2 \sim p}\, f(y_1, y_2) = \mathrm{Var}(p)$.

Example 1: Variance

Claim 1: Let

$$f(y_1, y_2) = \frac{1}{2}\left(y_1 - y_2\right)^2.$$

Then $\mathbb{E}_{y_1, y_2 \sim p}\, f(y_1, y_2) = \mathrm{Var}(p)$.

Claim 2: The multi-observation loss function

$$\ell(r, y_1, y_2) = \big(r - f(y_1, y_2)\big)^2$$

elicits the variance of p.

Example 2: 2-norm

Consider $\Gamma(p) = \|p\|_2^2 = \sum_y p_y^2$.

Example 2: 2-norm

Consider $\Gamma(p) = \|p\|_2^2 = \sum_y p_y^2$.

Claim 3: Let

$$f(y_1, y_2) = \mathbf{1}\left[y_1 = y_2\right].$$

Then $\mathbb{E}_{y_1, y_2 \sim p}\, f(y_1, y_2) = \|p\|_2^2$.

Example 2: 2-norm

Consider $\Gamma(p) = \|p\|_2^2 = \sum_y p_y^2$.

Claim 3: Let
$$f(y_1, y_2) = \mathbf{1}\left[y_1 = y_2\right].$$
Then $\mathbb{E}_{y_1, y_2 \sim p}\, f(y_1, y_2) = \|p\|_2^2$.

Claim 4: The multi-observation loss function
$$\ell(r, y_1, y_2) = \big(r - f(y_1, y_2)\big)^2$$
elicits the 2-norm squared of p.

Wait a minute!

What about the following transformation?

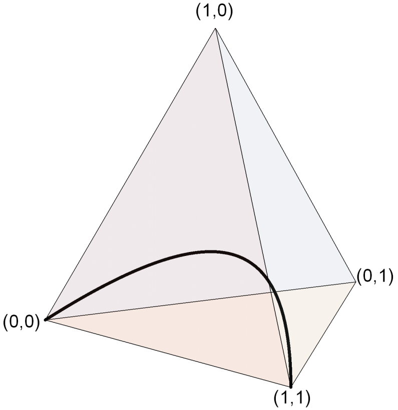
Let $p' = p \times p$ (distributions over i.i.d. pairs).
Then

$$\mathop{\mathbb{E}}_{y_1, y_2 \sim p} \ell(r, y_1, y_2) = \mathop{\mathbb{E}}_{\bar{y} \sim p'} \ell(r, \bar{y}).$$

So can't we reduce multi-observation elicitation to standard elicitation?

No. . .

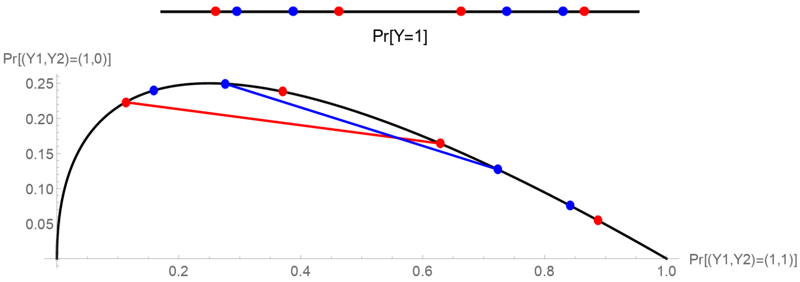
tetrahedron = distributions on $\{0,1\} \times \{0,1\}$
arc = i.i.d. distributions



. . . but this can provide lower bounds

Proposition

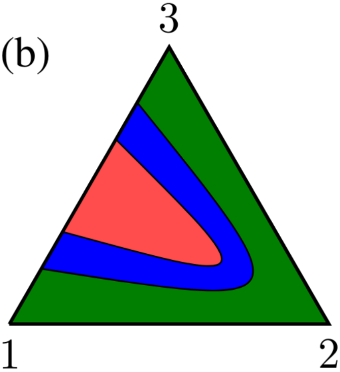
The fourth-central moment is not elicitable with any ≤ 2 observation loss function.



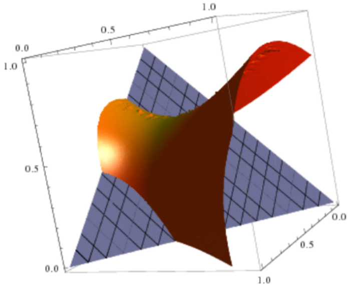
Key geometric idea: variance example

Level sets of m-observation elicitable properties can be non-convex. . .
. . . but they must be **projections** from convex level sets in $\Delta_{\mathcal{Y}}^m$.

(b)



(c)



Lower bound on number of observations

Theorem

If Γ is a m-observation-elicitable and "nice", then its level sets are all sets of zeros of some degree-at-most-m polynomial in p.

Lower bound on number of observations

Theorem

If Γ is a m-observation-elicitable and "nice", then its level sets are all sets of zeros of some degree-at-most-m polynomial in p.

Example (variance): $\left\{p : \sum_y p_y y^2 - \left(\sum_y p_y y\right)^2 = \frac{200}{3}\right\}$

Example (k-norm): $\left\{p : \sum_y p_y^k = 0.168\right\}$.

Lower bound on number of observations

Theorem

If Γ is a m-observation-elicitable and "nice", then its level sets are all sets of zeros of some degree-at-most-m polynomial in p.

Theorem (Real Nullstellensatz, extremely roughly)

A linear function can't vanish on a circle.

Lower bound on number of observations

Theorem

If Γ is a m-observation-elicitable and "nice", then its level sets are all sets of zeros of some degree-at-most-m polynomial in p.

Theorem (Real Nullstellensatz, very roughly)

If a level set consists of zeros of a degree-m polynomial, and the polynomial g vanishes on that level set, and some other conditions hold, then g has degree at least m.

Lower bound on number of observations

Theorem

If Γ is a m-observation-elicitable and "nice", then its level sets are all sets of zeros of some degree-at-most-m polynomial in p.

Theorem (Real Nullstellensatz, very roughly)

If a level set consists of zeros of a degree-m polynomial, and the polynomial g vanishes on that level set, and some other conditions hold, then g has degree at least m.

Corollary

To elicit the k norm requires a k-observation loss.

Summary and elicitation complexity

Two measures of complexity:

- **dimensionality**: how many parameters need to be elicited?
- **observations**: how many observations used in the loss function?

Summary and elicitation complexity

Two measures of complexity:

- **dimensionality**: how many parameters need to be elicited?
- **observations**: how many observations used in the loss function?

Nontrivial example: nth central moment is elicitable with $\sqrt{n}$ parameters and $\sqrt{n}$ observations.

Best we can do separately: n and n.

Summary and elicitation complexity

Two measures of complexity:

- **dimensionality**: how many parameters need to be elicited?
- **observations**: how many observations used in the loss function?

Theorem (Key example)

The 2-norm requires $|\mathcal{Y}| - 1$ parameters if using traditional loss functions, but just one parameter using the multi-observation loss

$$\ell(r, y_1, y_2) = \Big(r - \mathbf{1}[y_1 = y_2]\Big)^2.$$

Paper 2: generalized regression

Multi-Observation Regression. AISTATS 2019.

Frongillo, Mehta, Morgan, Waggoner.

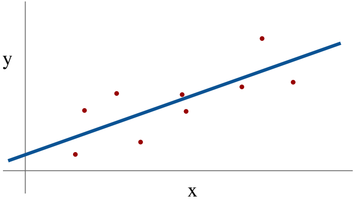
Paper 2: generalized regression

Multi-Observation Regression. AISTATS 2019.

Frongillo, Mehta, Morgan, Waggoner.

Setup:

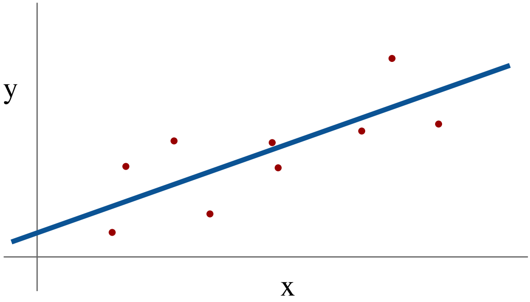
- Unknown distribution on (x, y) pairs
- Draw set of i.i.d. samples
- Goal: learn hypothesis $f : \mathcal{X} \to \mathcal{R}$ *map x to "summary" of y*
- Example: map x to **expected** y



Dominant paradigm: ERM

Example: least squares,

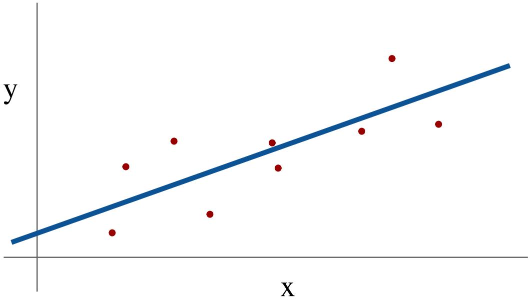
$$\underset{f}{\operatorname{argmin}} \sum_{(x,y)} (f(x) - y)^2$$



Dominant paradigm: ERM

More generally,

$$\operatorname*{argmin}_f \sum_{x,y} \ell(f(x), y).$$



Problem: non-elicitable properties!

Given x, we might want to predict...

- variance of y — *economics, biology, social science*
- upper confidence bound on y — *robust design (engineering)*
- risk measures — *finance*
- 2-norm of y — *economics, biology*
- ...

Problem: non-elicitable properties!

Given x, we might want to predict…

- variance of y *economics, biology, social science*
- upper confidence bound on y *robust design (engineering)*
- risk measures *finance*
- 2-norm of y *economics, biology*
- …

Prior paradigm does not directly apply!

Default solution: Fit a separate model for each parameter.

Problem: non-elicitable properties!

Given x, we might want to predict...

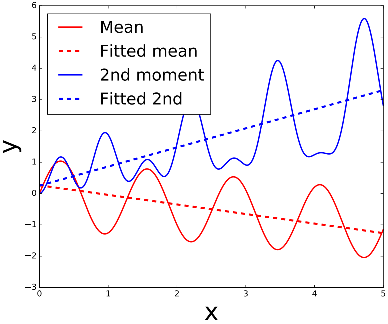
- variance of y *economics, biology, social science*
- upper confidence bound on y *robust design (engineering)*
- risk measures *finance*
- 2-norm of y *economics, biology*
- ...

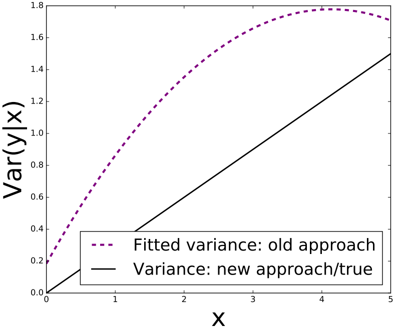
Prior paradigm does not directly apply!

Default solution: Fit a separate model for each parameter.

Problems: may need many parameters; VC-dimension issues...

Potential VC issues





Solution (?): Multi-observation losses

Proposal: Just fit a multi-observation loss!

$$\min_f \sum_{x,y} \ell(f(x), y_1, y_2)$$

Problem?

Solution (?): Multi-observation losses

Proposal: Just fit a multi-observation loss!

$$\min_f \sum_{x,y} \ell(f(x), y_1, y_2)$$

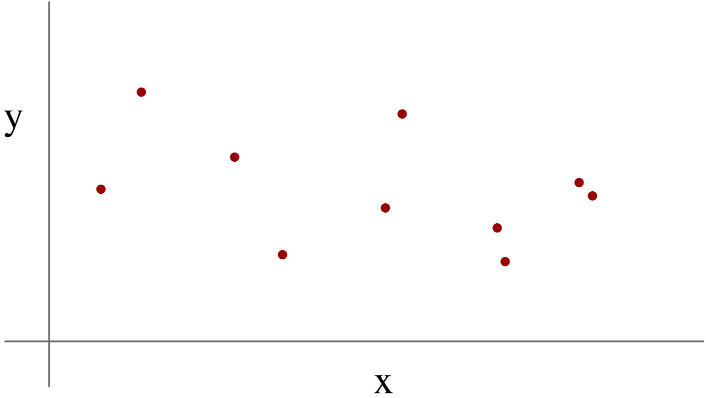
Problem?

We only have (x, y) samples, not (x, y_1, y_2)!

Fitting multi-observation losses

Clump data into **metasamples** $(x, y_1, \ldots, y_m)$,
then do empirical risk minimization:

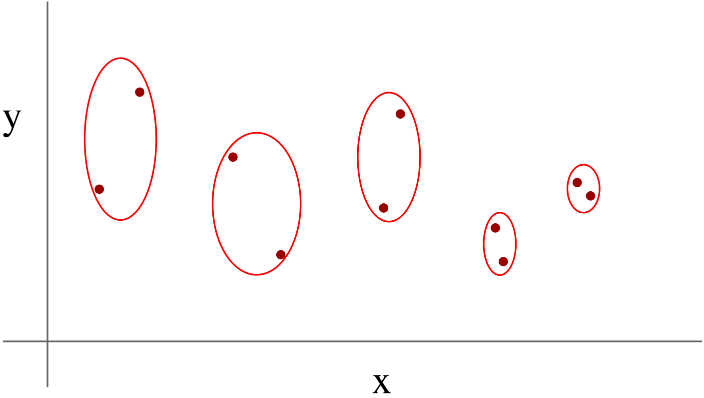
$$\min_f \sum_{\text{metasamples}} \ell(f(x), y_1, \ldots, y_m).$$



Fitting multi-observation losses

Clump data into **metasamples** $(x, y_1, \ldots, y_m)$,
then do empirical risk minimization:

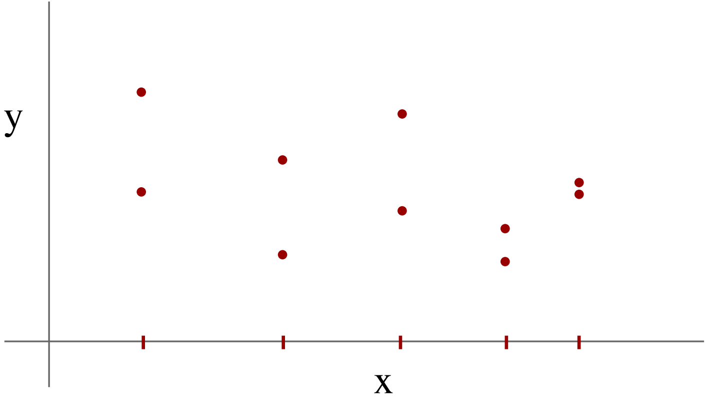
$$\min_f \sum_{\text{metasamples}} \ell(f(x), y_1, \ldots, y_m).$$



Fitting multi-observation losses

Clump data into **metasamples** $(x, y_1, \ldots, y_m)$,
then do empirical risk minimization:

$$\min_f \sum_{\text{metasamples}} \ell(f(x), y_1, \ldots, y_m).$$



Theory

Lipschitz assumption: $\Pr[y \mid x]$ changes slowly in x.

Unbiased algorithm:

1. Sample $x_1, \ldots, x_n$ i.i.d. *ignore their y's*
2. Draw "enough" fresh (x, y) pairs
3. Use maximum matching to assign ys to nearby original x_i.

Theory

Lipschitz assumption: $\Pr[y \mid x]$ changes slowly in x.

Unbiased algorithm:

1. Sample $x_1, \ldots, x_n$ i.i.d. *ignore their y's*
2. Draw "enough" fresh (x, y) pairs
3. Use maximum matching to assign ys to nearby original x_i.

Theorem (Informal)

With probability $1-\delta$, for $x \in [0,1]$, we draw $\tilde{O}(n)$ samples and

$$\textit{Risk(alg)} \leq \textit{Risk(opt)} + O\left(\textit{Rademacher complexity}\right) + O\left(\frac{1}{\sqrt{n}}\right).$$

Some proof ideas

1. With high probability, for all but $O(\sqrt{n})$ metasamples $(x, y_1, \ldots, y_m)$, all y_j were sampled "nearby".
   holds for arbitrary distributions

Some proof ideas

1. With high probability, for all but $O(\sqrt{n})$ metasamples $(x, y_1, \dots, y_m)$, all y_j were sampled "nearby".
   holds for arbitrary distributions

2. "Corrupted samples".

Some proof ideas

1. With high probability, for all but $O(\sqrt{n})$ metasamples $(x, y_1, \ldots, y_m)$, all y_j were sampled "nearby".
   holds for arbitrary distributions

2. "Corrupted samples".
   - y_j was sampled from a distribution close to $\Pr[y \mid x]$.

Some proof ideas

1. With high probability, for all but $O(\sqrt{n})$ metasamples $(x, y_1, \ldots, y_m)$, all y_j were sampled "nearby".
   holds for arbitrary distributions

2. "Corrupted samples".
   - y_j was sampled from a distribution close to $\Pr[y \mid x]$.
   - View that distribution as a mixture of $\Pr[y \mid x]$ and arbitrary.

Some proof ideas

1. With high probability, for all but $O(\sqrt{n})$ metasamples $(x, y_1, \ldots, y_m)$, all y_j were sampled "nearby".
   holds for arbitrary distributions

2. "Corrupted samples".
   - y_j was sampled from a distribution close to $\Pr[y \mid x]$.
   - View that distribution as a mixture of $\Pr[y \mid x]$ and arbitrary.
   - With good probability, all y_j in metasample came from $\Pr[y \mid x]$.

Some proof ideas

1. With high probability, for all but $O(\sqrt{n})$ metasamples $(x, y_1, \ldots, y_m)$, all y_j were sampled "nearby".
   holds for arbitrary distributions

2. "Corrupted samples".
   - y_j was sampled from a distribution close to $\Pr[y \mid x]$.
   - View that distribution as a mixture of $\Pr[y \mid x]$ and arbitrary.
   - With good probability, all y_j in metasample came from $\Pr[y \mid x]$.
   - Only lose $O(\sqrt{n})$ metasamples to bad mixtures.

Simulations

Setup:

- Draw $x \sim U[0, 1]$
- Draw $y = g(x) + N(0, 1)$
- Goal: fit $\mathrm{Var}(y \mid x)$ *answer* $= 1$

Simulations

Setup:

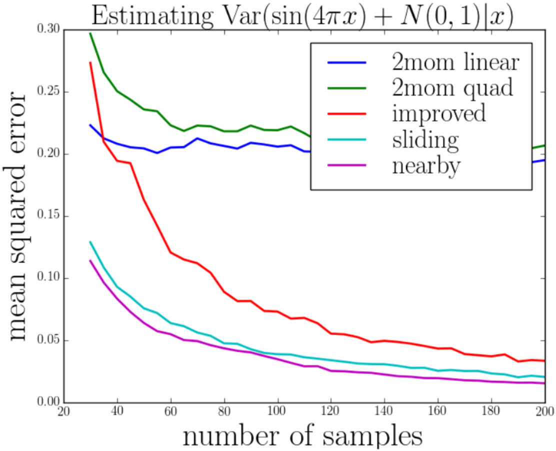
- Draw $x \sim U[0,1]$
- Draw $y = g(x) + N(0,1)$
- Goal: fit $\mathrm{Var}(y \mid x)$ *answer* $= 1$

Algorithms:

- "2mom linear" - fit linear functions to moments
- "2mom quad" - fit quadratics to moments
- "improved" - our theoretically-rigorous algorithm
- our other clustering algorithms

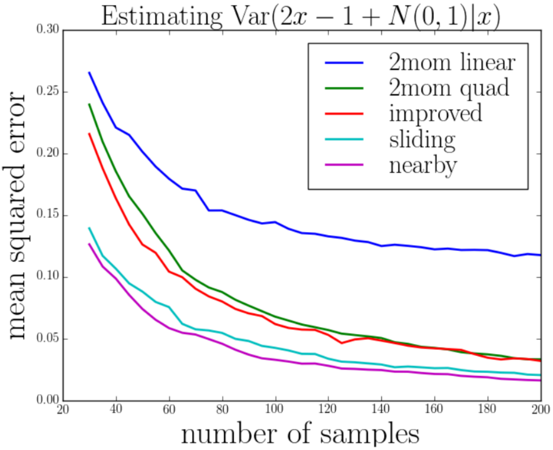
Observations from simulations

Difficult task:
As expected, default approaches perform very poorly.



Observations from simulations

Easy task:
Multi-observation approach can still be a better choice.



Summary

- Studied **multi-observation losses** $\ell(x, y_1, \ldots, y_m)$

Summary

- Studied **multi-observation losses** $\ell(x, y_1, \ldots, y_m)$
- **Elicitation complexity**: number of parameters and/or observations needed

Summary

- Studied **multi-observation losses** $\ell(x, y_1, \ldots, y_m)$
- **Elicitation complexity**: number of parameters and/or observations needed
- Multiple observations can lower number of parameters needed

Summary

- Studied **multi-observation losses** $\ell(x, y_1, \ldots, y_m)$
- **Elicitation complexity**: number of parameters and/or observations needed
- Multiple observations can lower number of parameters needed
- Techniques for lower-bounding number of observations needed

Summary

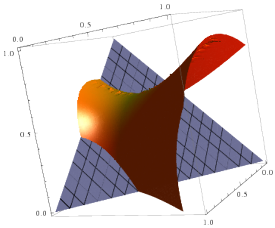
- Studied **multi-observation losses** $\ell(x, y_1, \ldots, y_m)$
- **Elicitation complexity**: number of parameters and/or observations needed
- Multiple observations can lower number of parameters needed
- Techniques for lower-bounding number of observations needed
- Algorithms for **metasamples** and multi-obs. ERM

Summary

- Studied **multi-observation losses** $\ell(x, y_1, \ldots, y_m)$
- **Elicitation complexity**: number of parameters and/or observations needed
- Multiple observations can lower number of parameters needed
- Techniques for lower-bounding number of observations needed
- Algorithms for **metasamples** and multi-obs. ERM
- Examples with huge improvement in sample complexity

Future directions (1/2)

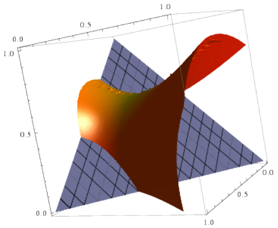
- Elicitation complexity: more upper and lower bounds
  central moments, multiple parameters & observations



Future directions (1/2)

- Elicitation complexity: more upper and lower bounds
  central moments, multiple parameters & observations

- Algorithms (or assumptions) in **high dimensions**
  information-theoretic barriers in general

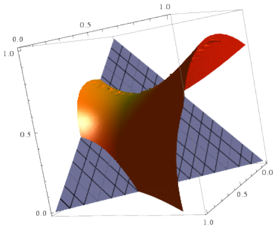


Future directions (1/2)

- Elicitation complexity: more upper and lower bounds
  central moments, multiple parameters & observations

- Algorithms (or assumptions) in **high dimensions**
  information-theoretic barriers in general

- Partner with practitioners $\to$ useful applications



Future directions (2/2)

- Multi-Observation Elicitation, COLT 2017.
- Multi-Observation Regression, AISTATS 2019.
- . . . next in the franchise?

Future directions (2/2)

- Multi-Observation Elicitation, COLT 2017.
- Multi-Observation Regression, AISTATS 2019.
- . . . next in the franchise?

Proceedings of Machine Learning Research vol XX:1–1, 2019

Multi-Observation: Apocalypse

Abstract

No algorithm could have predicted *this*. . .