Multi-Observation Regression

Rafael Frongillo, CU Boulder
Nishant Mehta, University of Victoria
Tom Morgan, Harvard
Bo Waggoner, CU Boulder

Multi-Observation Elicitation

Standard Property Elicitation (single-observation losses)

$$\arg\min_{r \in \mathbb{R}^d} \mathrm{E}_{Y \sim P}[\ell(r, Y)]$$

Squared loss elicits the mean of P

Variance is not elicitable by a single observation loss!

Multi-Observation Elicitation (multi-observation losses)

$$\arg\min_{r \in \mathbb{R}^d} \mathrm{E}_{(Y_1,\dots,Y_m)\sim P^m}[\ell(r, Y_1, \dots, Y_m)]$$

Variance is now elicitable, using $m = 2$

$$\ell(r, Y_1, Y_2) = \left(r - \frac{1}{2}(Y_1 - Y_2)^2\right)^2$$

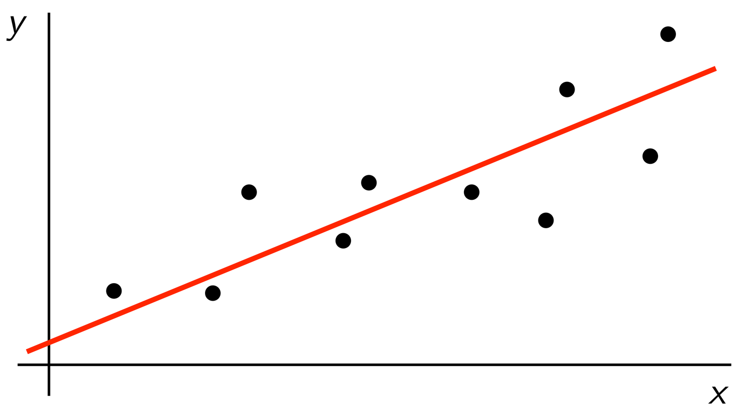
The 2-norm is also elicitable. Let $\mathcal{Y} = \{1, 2, \dots, K\}$

Squared 2-norm: $\sum_{j=1}^{K} P_j^2$

$$\ell(r, Y_1, Y_2) = \left(r - 1\{Y_1 = Y_2\}\right)^2$$

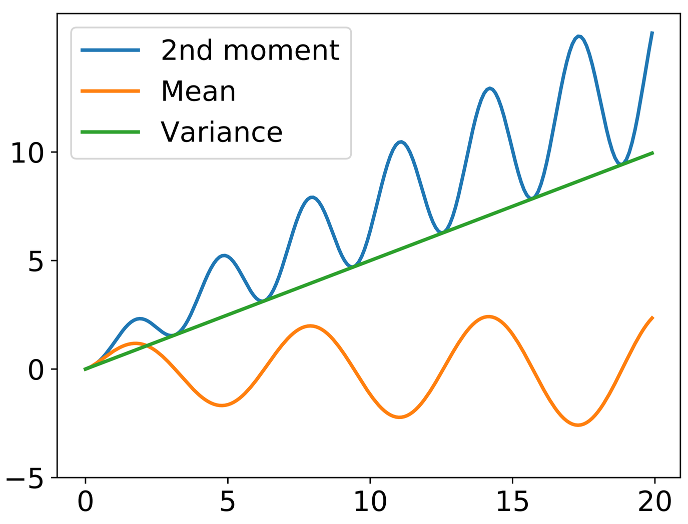
Multi-Observation Regression

Classical regression



Regress on y using ERM with squared loss

How about e.g. variance or 2-norm?



Traditional approach predicts variance by regressing on y and y^2 separately.

This might have high sample complexity and is suboptimal when the property varies in a simple way.

Solution: Regress directly on property via multi-observation loss!

So we try to minimize $R(f) := \mathrm{E}_{X\sim\mathcal{D}}\left[\mathrm{E}_{\mathbf{Y}\sim\mathcal{D}_X^m}\left[\ell_f(X, \mathbf{Y})\right]\right]$

How? Use ERM: $\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell_f\left(\underbrace{X_i}_{\text{i.i.d.} \sim \mathcal{D}}, \underbrace{(Y_{i,1}, \dots, Y_{i,m})}_{\text{drawn from } \mathcal{D}_{X_i}^m}\right)$

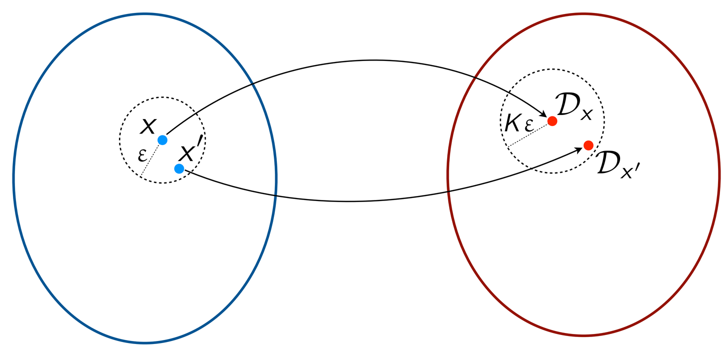
Problem: We don't have a multi-observation sampling oracle.
So, we don't have meta-samples!

We need to make do with a classical sampling oracle.
(classical, single-observation samples)

"Meta" Algorithm for ERM with Meta-Samples

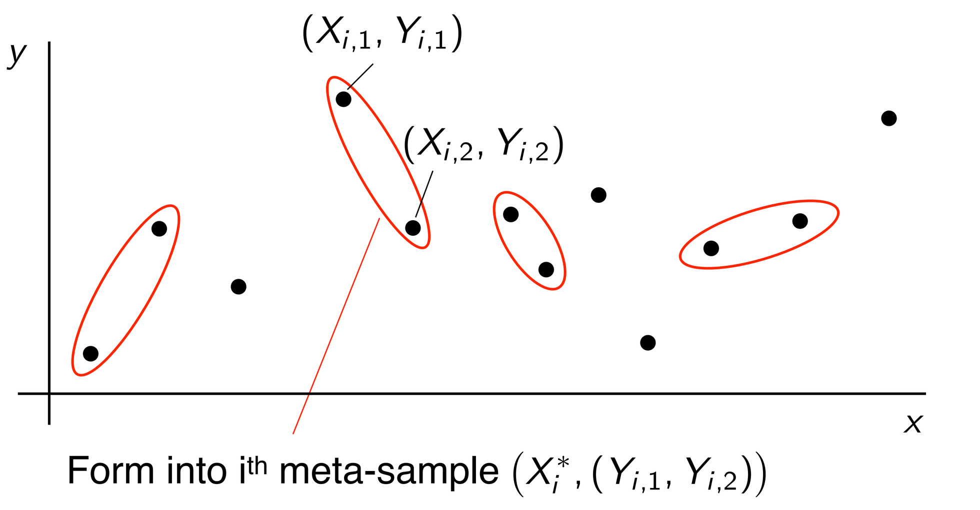
Solution: Form approximate meta-samples using extra samples and assume that the conditional distribution is slowly changing:

$$\exists K \text{ such that, for all } x, x' \in \mathcal{X}: \|\mathcal{D}_x - \mathcal{D}_{x'}\|_{TV} \le K\|x - x'\|_2$$



(1) Collect large number of classical samples

(2) Clump together samples with nearby x-values



(3) Run ERM on meta-samples $\left(X_i^*, \underbrace{(\tilde{Y}_{i,1}, \dots, \tilde{Y}_{i,m})}_{\text{Approximately drawn from } \mathcal{D}_{X_i^*}^m}\right)_{i\in[n]}$

Algorithm for Constructing Meta-Samples

Algorithm 1

Given: n, m, N, ε

Sample n points $X_1^*, \dots, X_n^*$ i.i.d. from $\mathcal{D}$

For $j = 1, \dots, m$

  Sample $k = \frac{N}{m}$ points $X_1^{(j)}, \dots, X_k^{(j)}$ i.i.d. from $\mathcal{D}$

  Find a maximum matching $M^{(j)}$ between $X_1^*, \dots, X_n^*$ and $X_1^{(j)}, \dots, X_k^{(j)}$, where X_i^* and $X_{i'}^{(j)}$ are adjacent iff $\|X_i^* - X_{i'}^{(j)}\| \le \varepsilon$

  If $|M^{(j)}| < n$

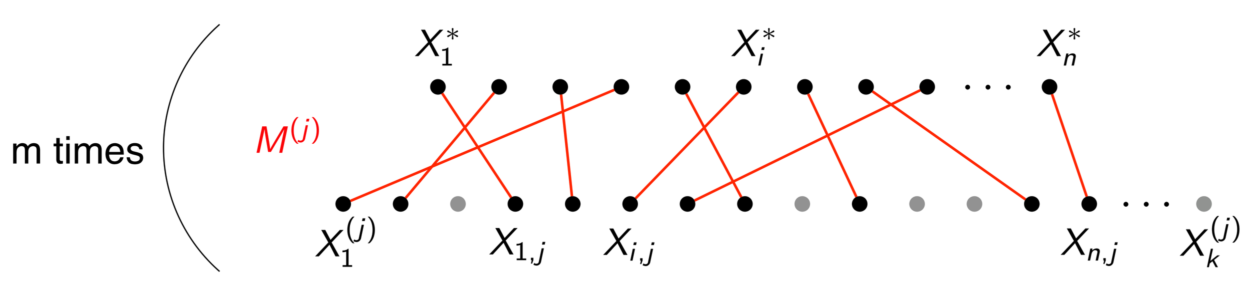
    Arbitrarily match remaining X_i^*'s (ignoring distance constraints)

  For $i = 1, \dots, n$

    Let $X_{i,j}$ denote the match of X_i^* in $M^{(j)}$

    Sample a label $\tilde{Y}_{i,j}$ from $\mathcal{D}_{X_{i,j}}$

Return $(X_i^*, (\tilde{Y}_{i,1}, \dots, \tilde{Y}_{i,m}))_{i\in[n]}$



Guarantee:
We say X_i^* is ε-well-matched by the set of matchings $M_1, \dots, M_m$ if X_i^* is matched to an ε-close point in each matching.

Matching Lemma
If $N = \tilde{\Omega}\left(md^{(d+2)/2}n^{(d+1)/2}\right)$, then with probability at least $1 - \delta$: all but $\tilde{O}(\sqrt{n})$ points X_i^* are $(1/\sqrt{n})$-well-matched by $M_1, \dots, M_m$.

General case excess risk bound

Theorem
Assume that the conditional distribution is slowly changing.
Let the loss be L-Lipschitz, and take $N = \tilde{\Omega}\left(md^{(d+2)/2}n^{(d+1)/2}\right)$.
If Algorithm 1 is run with input $(n, m, N, 1/\sqrt{n})$, and if ERM is run on the resulting meta-sample, then with probability at least $1 - \delta$,

$$R(\hat{f}_{\mathrm{ERM}}) - R(f^*) \le 2L\mathcal{R}_n(\mathcal{F}) + 2B\left(2\sqrt{\log\frac{4}{\delta}} + mK\right)\frac{1}{\sqrt{n}}$$

Rademacher complexity
Upper bound on loss

Proof Sketch

Idea: Think of ERM as being run on corrupted samples
Let $\varepsilon = 1/\sqrt{n}$

(1) From Matching Lemma, only $O(\sqrt{n})$ points X_i^* fail to be ε-well-matched. The labels for these points are not even approximately drawn from $\mathcal{D}_{X_i^*}$, so we consider the corresponding $O(\sqrt{n})$ meta-samples as corrupted.

(2) For each well-matched point X_i^*:

  We have $\|\mathcal{D}_{X_{i,j}} - \mathcal{D}_{X_i^*}\|_{TV} \le K\varepsilon$ for all $j \in [m]$

  Think of $\tilde{Y}_{ij}$ as sampled as follows:

    First, draw Z_{ij} from Bernoulli$(K\varepsilon)$.

    Then, $\tilde{Y}_{ij} = \text{sample from} \begin{cases} \mathcal{D}_{X_i^*} & \text{if } Z_{ij} = 0 \\ Q_{ij} & \text{if } Z_{ij} = 1 \end{cases}$

    Arbitrary "bad" mixture component

(3) With high probability, at most $O(\sqrt{n})$ of meta-samples from well-matched points have any label coming from a bad mixture component. In all, only $O(\sqrt{n})$ of meta-samples are corrupted.

Simulations

Setup: Let $X \sim U([0,1])$ and $Y = g(X) + \xi$ for $\xi \sim \mathrm{N}(0,1)$
Goal: Predict $\mathrm{Var}[Y \mid X]$
Algorithms:

"2mom linear" - fit linear functions to moments
"2mom quad" - fit quadratic functions to moments
"unbiased" - our algorithm (has theoretical guarantees)
"sliding" / "nearby" - other, non-theoretically rigorous algorithms