Label Complexity in Machine Learning

Malik Magdon-Ismail
Professor, CS, RPI

- Machine Learning

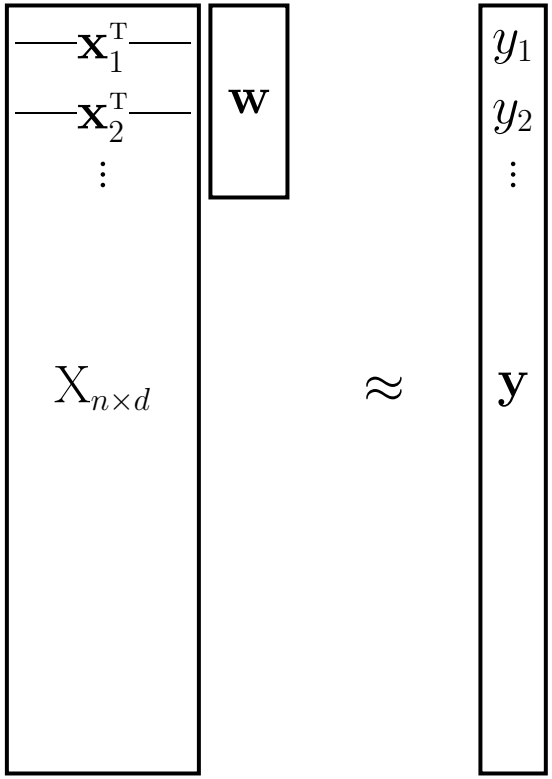
  Data (in particular labels) as a scarce resource.

  Compute can be prohibitive, n^2, nd^2.
- Quantum Algorithms

  Tight success probabilities for Quantum-order finding.

Linear Regression

$\mathrm{X} \in \mathbb{R}^{n \times d}$, $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^n$ $\qquad d \ll n \ll e^d$



$\mathbf{w}_*$ minimizes:

$$\|\mathrm{X}\mathbf{w} - \mathbf{y}\|_2^2 = \sum_{i=1}^{n} (\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i)^2.$$

$$\mathbf{w}_* = \mathrm{X}^{\dagger}\mathbf{y} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathbf{y}.$$

Under standard statistical sampling settings, e.g. $(\mathbf{x}, y) \overset{\text{i.i.d.}}{\sim} P(\mathbf{x}, y)$,

$$\mathbb{E}_P\left[\|\mathbf{w}_*^{\mathrm{T}}\mathbf{x} - y\|_2^2\right] = \min_{\mathbf{w}} \mathbb{E}_P\left\{\left[\|\mathbf{w}^{\mathrm{T}}\mathbf{x} - y\|_2^2\right]\right\} + O\left(\frac{d}{n}\right).$$

ϵ-Coreset

Subset of the data:

$$(\mathbf{x}_{i_1}, y_{i_1}), (\mathbf{x}_{i_2}, y_{i_2}), (\mathbf{x}_{i_3}, y_{i_3}), \ldots, (\mathbf{x}_{i_m}, y_{i_m}).$$

Weights:

$$\nu_{i_1}, \nu_{i_2}, \nu_{i_3}, \ldots, \nu_{i_m}.$$

Solve the weighted regression:

$$\hat{\mathbf{w}} \leftarrow \operatorname*{argmin}_{\mathbf{w}} \sum_{\ell=1}^{m} \nu_{i_\ell} (\mathbf{w}^{\mathrm{T}} \mathbf{x}_{i_\ell} - y_{i_\ell})^2$$

Require:

$$\|\mathrm{X}\hat{\mathbf{w}} - y\|_2^2 \leq (1+\epsilon)\|\mathrm{X}\mathbf{w}_* - y\|_2^2.$$

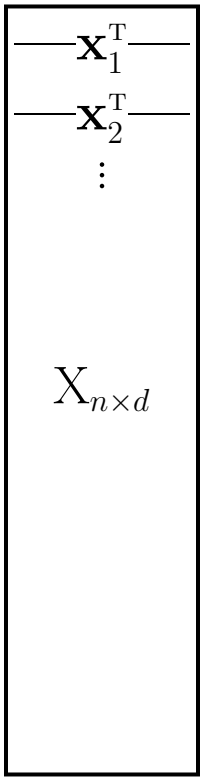
$\epsilon \leq d/n.$

Motivation: compact; efficient; more robust.

Algorithms exist: given $\mathrm{X}, \mathbf{y}, \epsilon \to m \in O(d/\epsilon)$ [Drineas, et al. 2006, Boutsidis, M-I, 2013]

Label Complexity Problem

Construct coreset given only X. (**y**-oblivious; active regression)



Is it even possible?
Deterministic coreset: **not possible**.
Random coresets; large norm coresets: **not good**.

Sampling and Rescaling

Pick $\mathbf{x} \in \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \sim \{p_1, p_2, \ldots, p_n\}$ and obtain the label $y(\mathbf{x})$.

$$\hat{\mathbf{w}} \leftarrow \operatorname*{argmin}_{\mathbf{w}} \underbrace{\left\{ \frac{1}{p(\mathbf{x})} \big(\mathbf{w}^{\mathrm{T}} \mathbf{x} - y(\mathbf{x}) \big)^2 \right\}}_{\mathcal{E}(\mathbf{x})}.$$

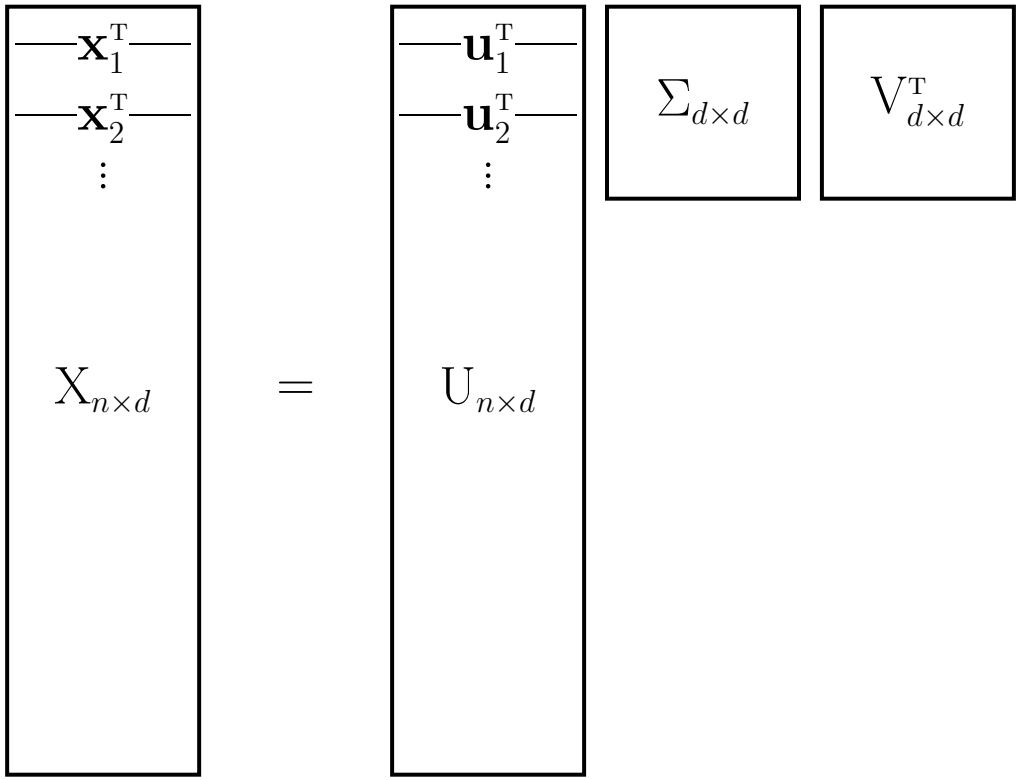
$$\mathbb{E}[\mathcal{E}(\mathbf{x})] = \sum_{i=1}^{n} \cancel{p_i} \times \frac{1}{\cancel{p_i}} (\mathbf{w}^{\mathrm{T}} \mathbf{x}_i - y_i)^2 = \sum_{i=1}^{n} (\mathbf{w}^{\mathrm{T}} \mathbf{x}_i - y_i)^2.$$

Minimize a quantity whose expectation is the quantity to be minimized.

Does this work?
No.

Repeat, sample $m > 1$ times and use inverse probability weights?
Yes, with good sampling probabilities $p_1, \ldots, p_n$.

Leverage Scores



leverage score $\ell_i = \|\mathbf{u}_i\|_2^2$.
i.i.d sampling probability $p_i = \ell_i/d$.

Theorem (Chen, Price, 2018; Dong, M-I, 2026)

For any X *and* any *unknown* $\mathbf{y}$, $m \approx 2d/\epsilon + 21d\ln d$ *implies*

$$\mathbb{E}[\|\mathrm{X}\hat{\mathbf{w}} - y\|_2^2] \leq (1+\epsilon)\|\mathrm{X}\mathbf{w}_* - y\|_2^2.$$

Rejection

$$\mathrm{X} = \begin{bmatrix} 7 & 5 & 3 \\ 4 & 8 & 1 \\ 0 & 1 & 0 \\ 8 & 7 & 6 \\ 5 & 8 & 8 \\ 7 & 3 & 2 \\ 2 & 5 & 7 \\ 3 & 2 & 9 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 2 \\ 0 \\ -2 \\ 2 \\ -2 \end{bmatrix}$$

$$\mathrm{X}^{-} = \begin{bmatrix} 7 & 5 & 3 \\ 4 & 8 & 1 \\ 8 & 7 & 6 \\ 5 & 8 & 8 \\ 7 & 3 & 2 \\ 2 & 5 & 7 \\ 3 & 2 & 9 \end{bmatrix}, \qquad \mathbf{y}^{-} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 0 \\ -2 \\ 2 \\ -2 \end{bmatrix}$$

$$\|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|_2^2 = \mathbf{9.6113}$$

$$\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 = \mathbf{9.6163}$$

$$\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 = (1 + 0.00052)\|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|_2^2$$

Rejecting One Point Is Always Possible

Theorem (M-I, Gittens, Dong, preprint)

Rejecting One Point Is Always Possible

Theorem (M-I, Gittens, Dong, preprint)

For any $\mathrm{X}, \mathbf{y}$*:*

Rejecting One Point Is Always Possible

Theorem (M-I, Gittens, Dong, preprint)

For any $\mathrm{X}, \mathbf{y}$*:*

- *There is an* $(\mathbf{x}_i, y_i)$ *to throw out and get* $\mathrm{X}^-, \mathbf{y}^-$ *giving* $\hat{\mathbf{w}}$ *such that*

$$\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 \leq \left(1 + \frac{d}{(n-d)^2}\right) \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|_2^2.$$

- *The result is tight.*

Rejecting One Point Is Always Possible

Theorem (M-I, Gittens, Dong, preprint)

For any $\mathrm{X}, \mathbf{y}$*:*

- *There is an* $(\mathbf{x}_i, y_i)$ *to throw out and get* $\mathrm{X}^-, \mathbf{y}^-$ *giving* $\hat{\mathbf{w}}$ *such that*

$$\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 \leq \left(1 + \frac{d}{(n-d)^2}\right) \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|_2^2.$$

- *The result is tight.*
- *There is an algorithm* $\mathcal{A}$ *to find* $(\mathbf{x}_i, y_i)$ **without knowing y** *(randomized),*

$$\mathbb{E}_{\mathcal{A}}\left[\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2\right] \leq \left(1 + \frac{d}{(n-d)^2}\right) \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|_2^2.$$

Rejecting One Point Is Always Possible

Theorem (M-I, Gittens, Dong, preprint)

For any $\mathrm{X}, \mathbf{y}$*:*

- *There is an* $(\mathbf{x}_i, y_i)$ *to throw out and get* $\mathrm{X}^-, \mathbf{y}^-$ *giving* $\hat{\mathbf{w}}$ *such that*

$$\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 \leq \left(1 + \frac{d}{(n-d)^2}\right) \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|_2^2.$$

- *The result is tight.*
- *There is an algorithm* $\mathcal{A}$ *to find* $(\mathbf{x}_i, y_i)$ **without knowing y** *(randomized),*

$$\mathbb{E}_{\mathcal{A}}\left[\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2\right] \leq \left(1 + \frac{d}{(n-d)^2}\right) \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|_2^2.$$

- *The algorithm* $\mathcal{A}$ *is simple and efficient: reject* $(\mathbf{x}_i, y_i)$ *with probability*

$$p_i = \frac{1}{\mathcal{Z}} \frac{(1-\ell_i)^2}{\ell_i}.$$

We can reduce label complexity by **one** from n to $n-1$!

Proof Sketch: Rejecting One Point

Proof Sketch: Rejecting One Point

1. Perturbation from $\mathrm{X}, \mathbf{y}$ to $\mathrm{X}^{-}, \mathbf{y}^{-}$ after throwing out $(\mathbf{x}_i, y_i)$.

$$\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|^2 = \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|^2 + \frac{\ell_i}{(1-\ell_i)^2}\|\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_* - \mathbf{y}_i\|^2.$$

Proof Sketch: Rejecting One Point

1. Perturbation from $\mathrm{X}, \mathbf{y}$ to $\mathrm{X}^-, \mathbf{y}^-$ after throwing out $(\mathbf{x}_i, y_i)$.

$$\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|^2 = \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|^2 + \frac{\ell_i}{(1-\ell_i)^2}\|\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_* - \mathbf{y}_i\|^2.$$

2. Reject $\mathbf{x}_i$ with $p_i = \dfrac{1}{\mathcal{Z}}\dfrac{(1-\ell_i)^2}{\ell_i}$, where the "partition function" $\mathcal{Z} = \displaystyle\sum_{i=1}^{n} \frac{(1-\ell_i)^2}{\ell_i}$.

Proof Sketch: Rejecting One Point

1. Perturbation from $\mathrm{X}, \mathbf{y}$ to $\mathrm{X}^-, \mathbf{y}^-$ after throwing out $(\mathbf{x}_i, y_i)$.

$$\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|^2 = \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|^2 + \frac{\ell_i}{(1-\ell_i)^2}\|\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_* - \mathbf{y}_i\|^2.$$

2. Reject $\mathbf{x}_i$ with $p_i = \dfrac{1}{\mathcal{Z}}\dfrac{(1-\ell_i)^2}{\ell_i}$, where the "partition function" $\mathcal{Z} = \displaystyle\sum_{i=1}^{n} \frac{(1-\ell_i)^2}{\ell_i}$.

3. Compute the expected loss,

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}\left[\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|^2\right] &= \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|^2 + \frac{1}{\mathcal{Z}}\sum_{i=1}^{n} \cancel{\frac{(1-\ell_i)^2}{\ell_i}} \cancel{\frac{\ell_i}{(1-\ell_i)^2}}\|\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_* - \mathbf{y}_i\|^2 \\ &= \left(1 + \frac{1}{\mathcal{Z}}\right)\|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|^2.\end{aligned}$$

Proof Sketch: Rejecting One Point

1. Perturbation from $\mathrm{X}, \mathbf{y}$ to $\mathrm{X}^-, \mathbf{y}^-$ after throwing out $(\mathbf{x}_i, y_i)$.

$$\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|^2 = \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|^2 + \frac{\ell_i}{(1-\ell_i)^2}\|\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_* - \mathbf{y}_i\|^2.$$

2. Reject $\mathbf{x}_i$ with $p_i = \dfrac{1}{\mathcal{Z}}\dfrac{(1-\ell_i)^2}{\ell_i}$, where the "partition function" $\mathcal{Z} = \displaystyle\sum_{i=1}^{n} \frac{(1-\ell_i)^2}{\ell_i}$.

3. Compute the expected loss,

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}\left[\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|^2\right] &= \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|^2 + \frac{1}{\mathcal{Z}}\sum_{i=1}^{n} \cancel{\frac{(1-\ell_i)^2}{\ell_i}} \cancel{\frac{\ell_i}{(1-\ell_i)^2}} \|\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_* - \mathbf{y}_i\|^2 \\ &= \left(1 + \frac{1}{\mathcal{Z}}\right)\|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|^2.\end{aligned}$$

4. Lower bound $\mathcal{Z}$. For any left singular matrix U, $\mathcal{Z} \geq (n-d)^2/d$.

Note: $\mathcal{Z}$ can be computed given U.
What about removing more than one point?

Example: Rejecting One Point

$$\mathrm{X} = \begin{bmatrix} 7 & 5 & 3 \\ 4 & 8 & 1 \\ 0 & 1 & 0 \\ 8 & 7 & 6 \\ 5 & 8 & 8 \\ 7 & 3 & 2 \\ 2 & 5 & 7 \\ 3 & 2 & 9 \end{bmatrix} \qquad \boldsymbol{\ell} = \begin{bmatrix} 0.2927 \\ 0.6296 \\ 0.0215 \\ 0.2993 \\ 0.3683 \\ 0.4654 \\ 0.3344 \\ 0.5889 \end{bmatrix}$$

for any X :

$$\frac{1}{\mathcal{Z}} \leq \frac{d}{(n-d)^2} = 0.12$$

for this X :

$$\frac{1}{\mathcal{Z}} = \frac{1}{\sum_{i=1}^{n} (1-\ell_i)^2/\ell_i} = 0.019$$

For any $\mathbf{y}$:

$$\mathbb{E}_{\mathcal{A}}\left[\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2\right] \leq (1 + \mathbf{0.019}) \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|_2^2.$$

Pop Quiz: Why not just throw out the red point with lowest leverage?

Algorithm: Reject k Points

Let A denote a subset of k rows. Let $\mathrm{U_A}$ be the corresponding rows of U.

Reject A with probability

$$p_{\mathrm{A}} = \frac{1}{\mathcal{Z}} \frac{(1 - \|\mathrm{U_A}\|_2^2)^2}{\|\mathrm{U_A}\|_2^2}$$

One can efficiently sample A according to p_{A}.

Reducing Label Complexity by $\Omega(\sqrt{n})$

$$p_{\mathrm{A}} = \frac{1}{\mathcal{Z}} \frac{(1 - \|\mathrm{U}_{\mathrm{A}}\|_2^2)^2}{\|\mathrm{U}_{\mathrm{A}}\|_2^2}$$

Theorem (M-I, Gittens, Dong, preprint)

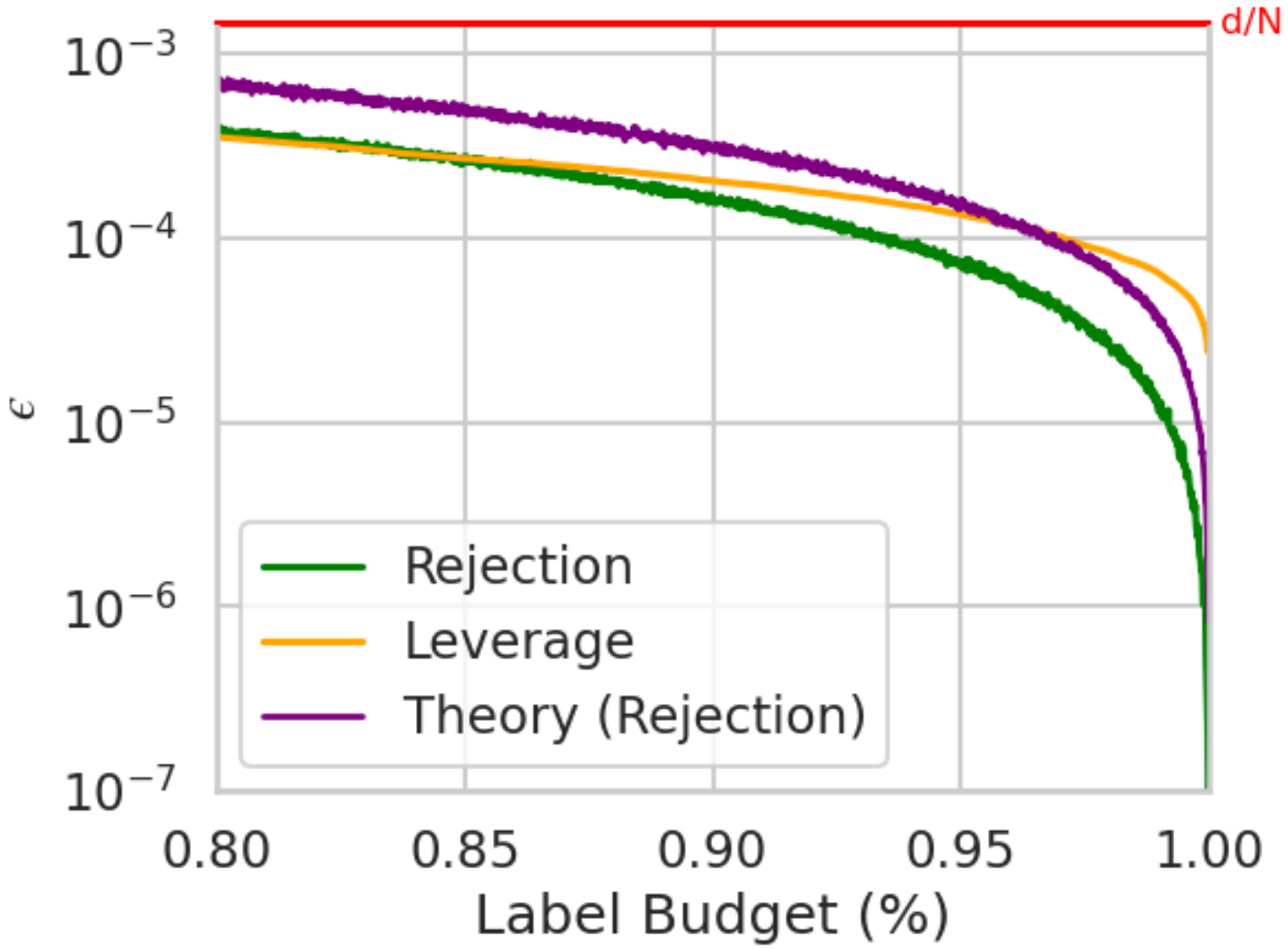
For <u>*any*</u> $\mathrm{X}, \mathbf{y}$*, reject the rows* A *with probability* p_{A}*. For* $k < n/d$*,*

$$\mathbb{E}_{\mathcal{A}}\left[\|\mathrm{X}\hat{\mathbf{w}} - \mathbf{y}\|^2\right] \leq \left(1 + \frac{dk^2}{(n - dk)^2}\right) \|\mathrm{X}\mathbf{w}_* - \mathbf{y}\|^2.$$

Notes:

- Set $k \in \Omega(\sqrt{n})$ to get $(1 + d/n)$-approximation.
- With additional assumptions one can reduce label complexity by $\sim 0.3n$.
  e.g. if the residual is isotropic, such as i.i.d. Gaussians.
- Not trivial to sample from p_{A}, $O(n^2/d)$ in practice.

Illustration on Real Data



Conclusion & Open Questions

- Lots of nice little linear algebra exercises.
- Rejection is better than sampling for tight approximation to linear regression.
  Reduces label complexity by $\Omega(n)$ in practice.
- Given X, ϵ, what is $\mathsf{labelcomplexity}(\mathrm{X}, \epsilon)$?
  Can one tighten the slack in our bound?
- Remove conditional assumptions on rejection algorithm to get label reduction $\Omega(n)$?
- What about nonlinear families, non-convex objectives?
  Next step: linear families, convex objective.