Active Linear Regression for ℓ_p Norms and Beyond

Taisuke Yasuda
CMU

based on work with

Cameron Musco
UMass Amherst

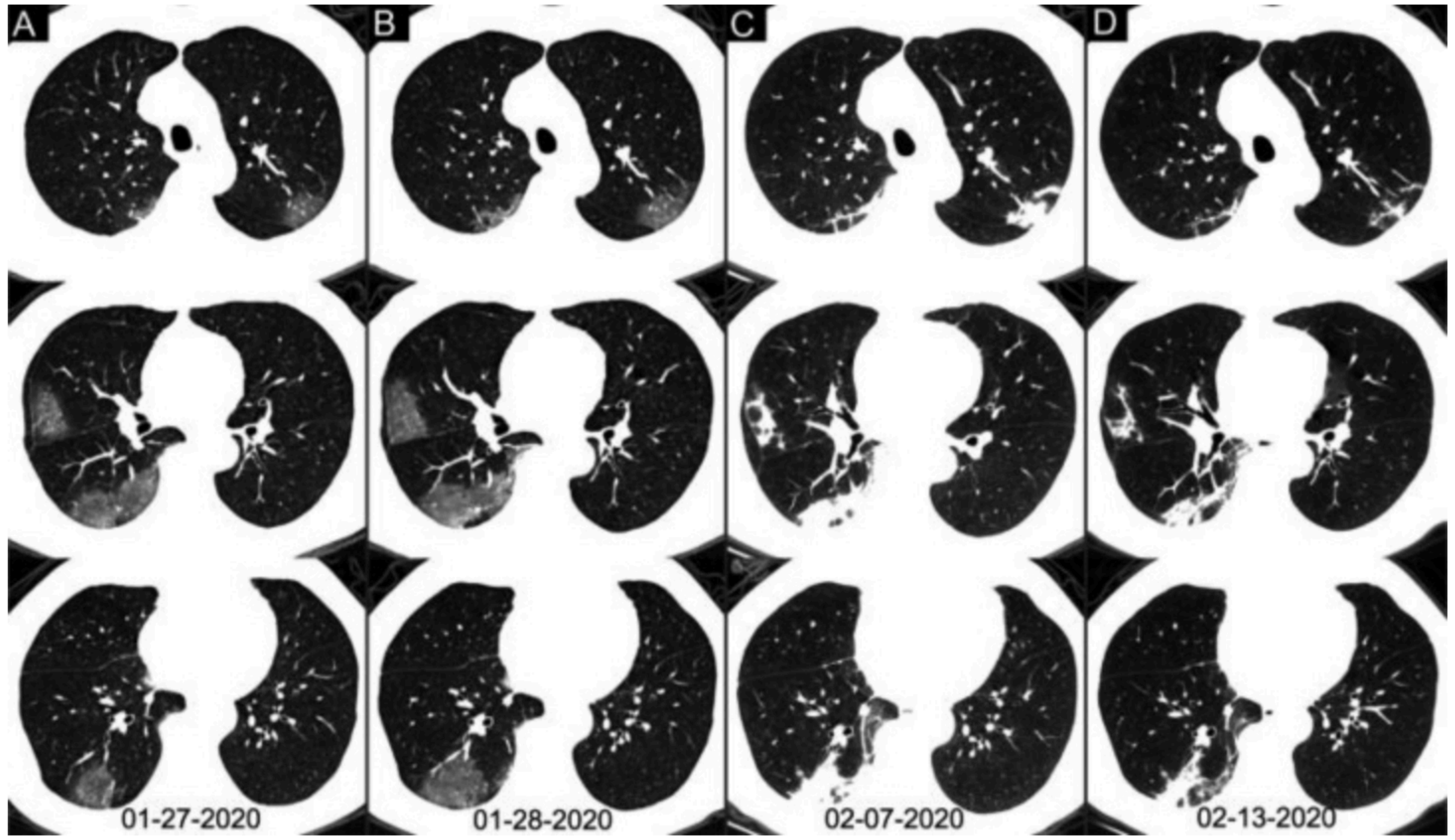
Christopher Musco
NYU

David P. Woodruff
CMU

FOCS, November 3, 2022

Active Learning

- Oftentimes, *largest bottleneck* in machine learning applications is the *collection of labels*
- Active learning aims to *minimize the number of label entries read*
- Most basic question in active learning: *active linear regression*



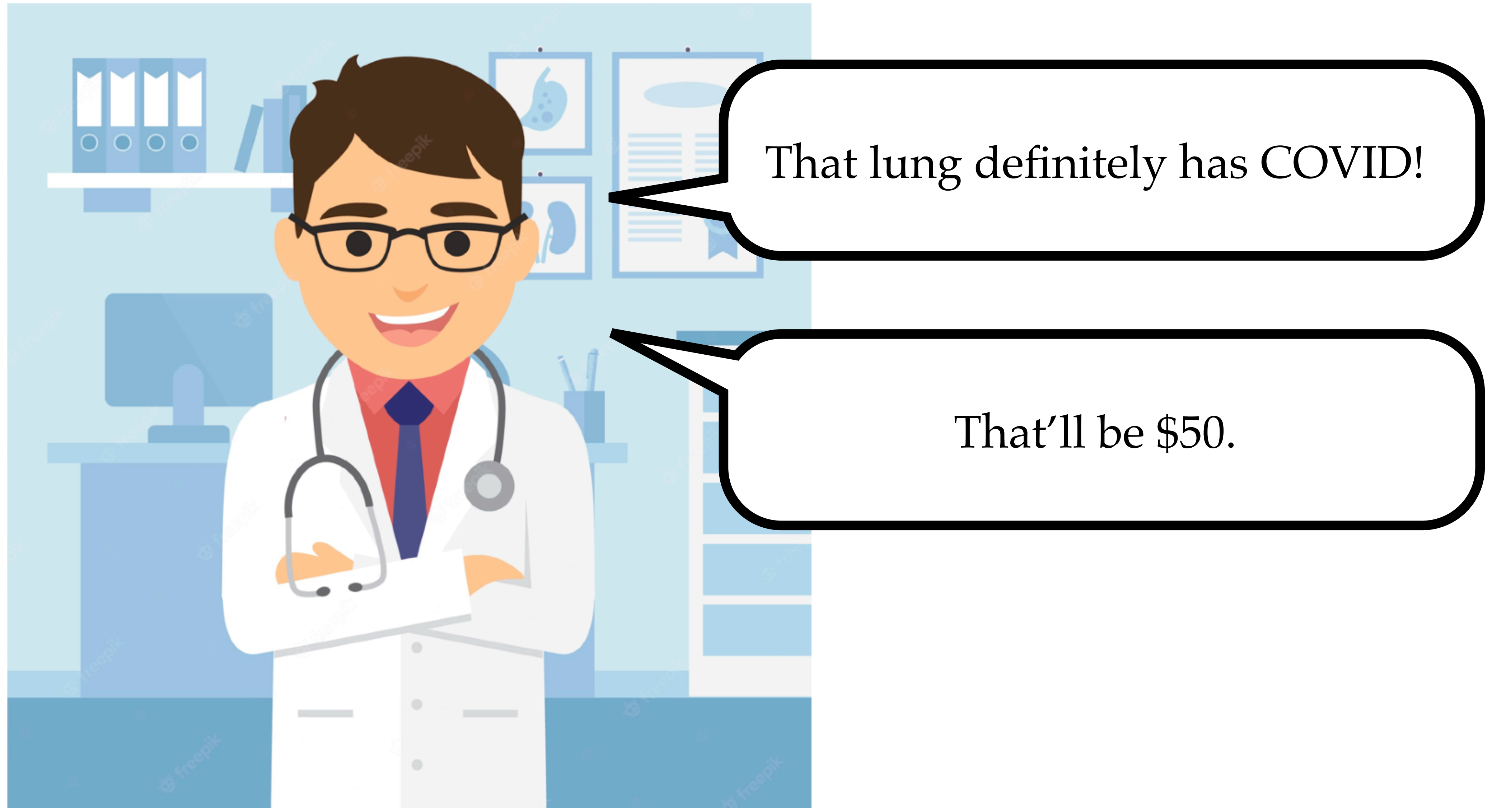


Image from Ai, Tao, et al. "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases." *Radiology* (2020).

Active ℓ_p Linear Regression

Problem Setting

- ℓ_p norm: $\|y\|_p = \left(\sum_{i=1}^{n} |y_i|^p \right)^{1/p}$
- Given:
  - Design matrix $A \in \mathbb{R}^{n \times d}$ with n examples and d features
  - Query access to label vector $b \in \mathbb{R}^n$
- Output:
  - Coefficient vector $\hat{x} \in \mathbb{R}^d$ such that $\|A\hat{x} - b\|_p^p \leq (1 + \varepsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\|_p^p$

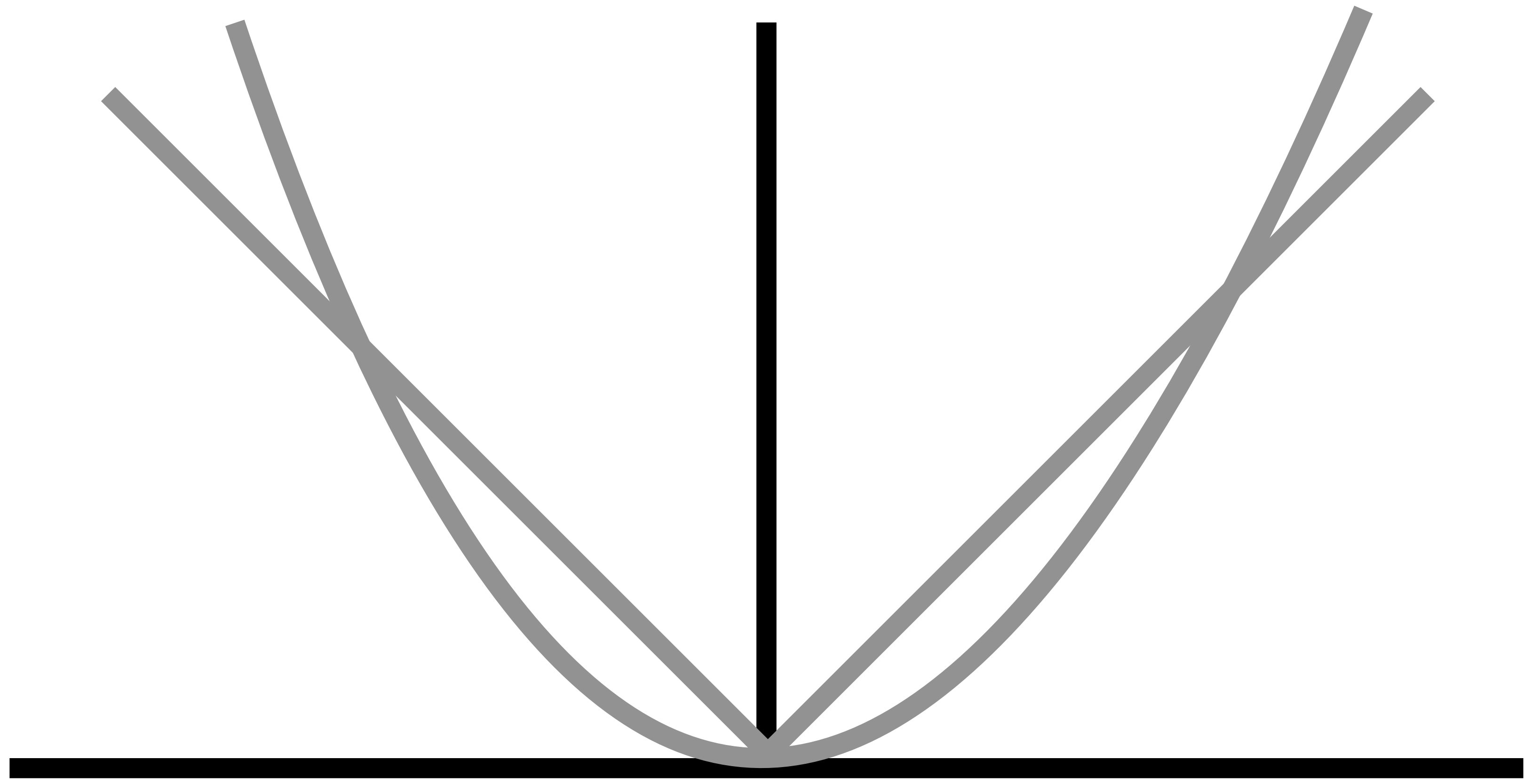
Active ℓ_p Linear Regression

Prior Work

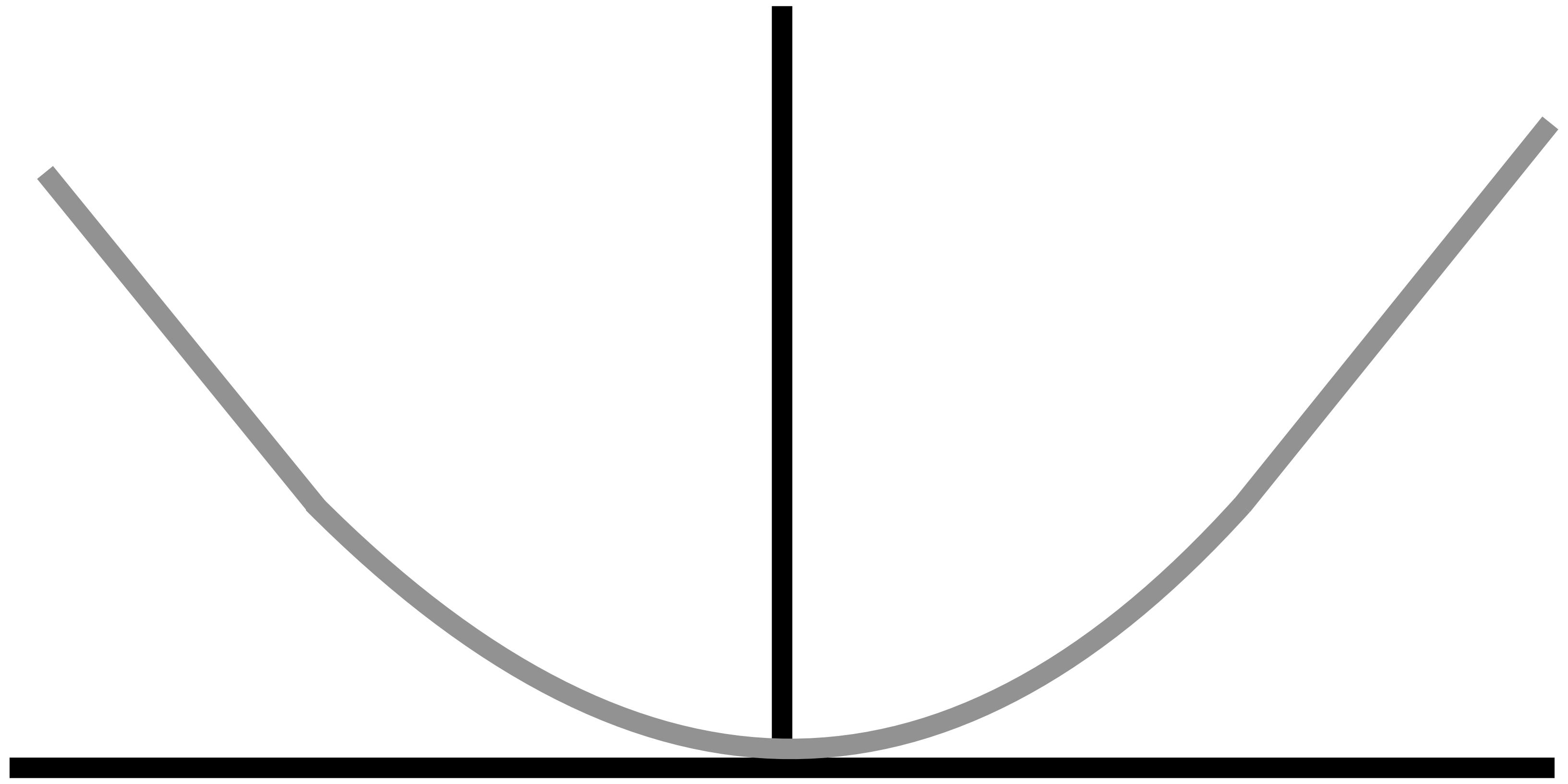
	Query Complexity	Work
$p = 2$	$\Theta\left(\frac{d}{\varepsilon}\right)$	[Chen, Price 2019]
$p = 1$	$\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$	[Chen, Derezinski 2021] [Parulekar, Parulekar, Price 2021]
$1 < p < 2$	$\tilde{O}\left(\frac{d^2}{\varepsilon^2}\right)$	[Chen, Derezinski 2021]
$1 < p < 2$	$\tilde{\Theta}\left(\frac{d}{\varepsilon}\right)$	[Musco, Musco, Woodruff, Y 2022]
$p > 2$	$\tilde{O}\left(\frac{d^{p/2}}{\varepsilon^p}\right), \Omega\left(d^{p/2} + \frac{1}{\varepsilon^{p-1}}\right)$	[Musco, Musco, Woodruff, Y 2022]
$0 < p < 1$	$\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$	[Musco, Musco, Woodruff, Y 2022]

Beyond ℓ_p Norms

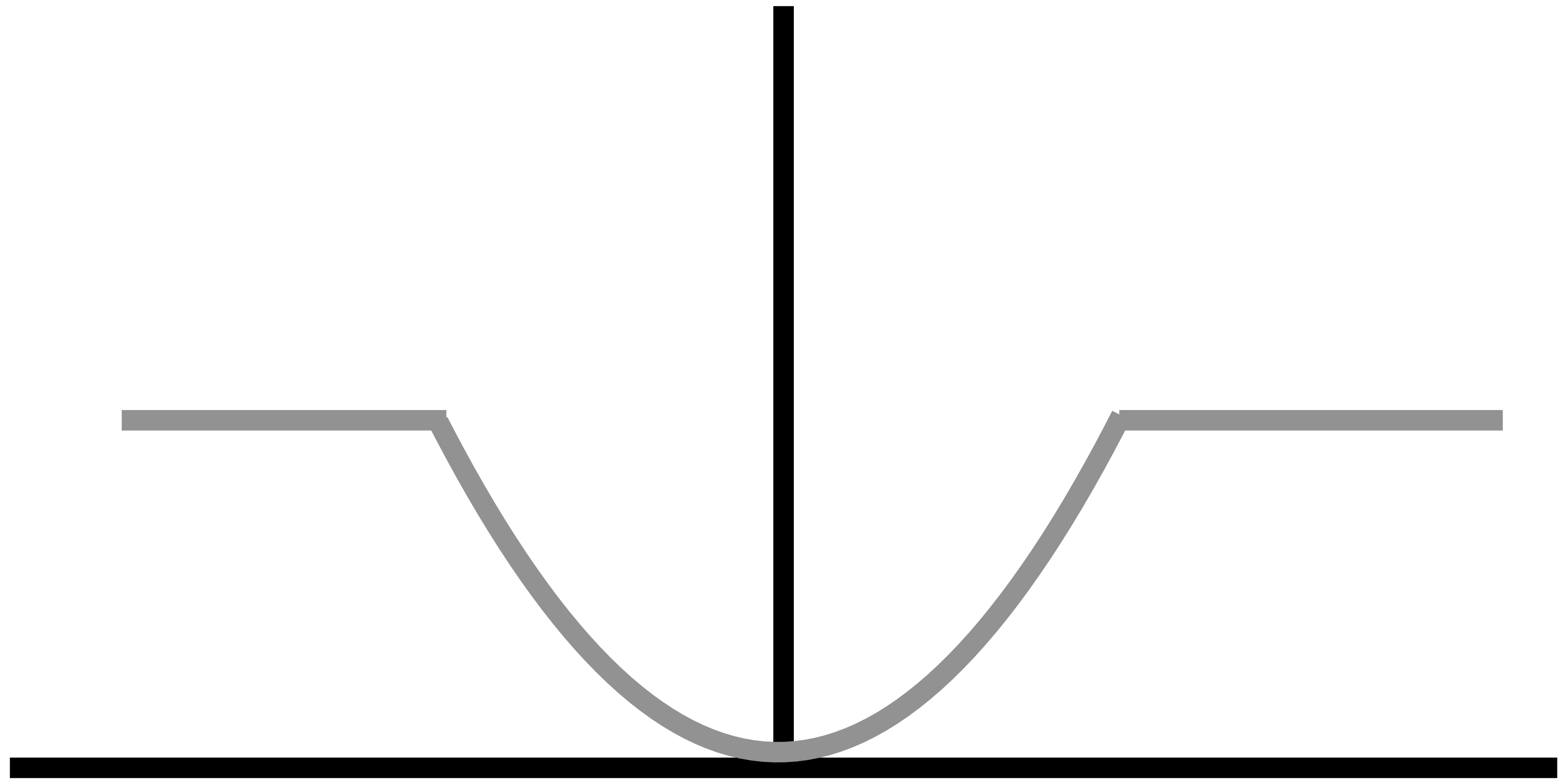
Our Results



ℓ_p loss: $\dfrac{d}{\varepsilon}$



Huber loss: $\dfrac{d^{4-2\sqrt{2}}}{\mathrm{poly}(\varepsilon)} < \dfrac{d^{1.172}}{\mathrm{poly}(\varepsilon)}$



Tukey loss: $\dfrac{d^2}{\mathrm{poly}(\varepsilon)}$

Applications of Our Techniques

- Sparsification for M-estimators, Orlicz norms
- Sparsification for the Huber loss, γ functions
- Kronecker product regression
- Robust subspace approximation
- …

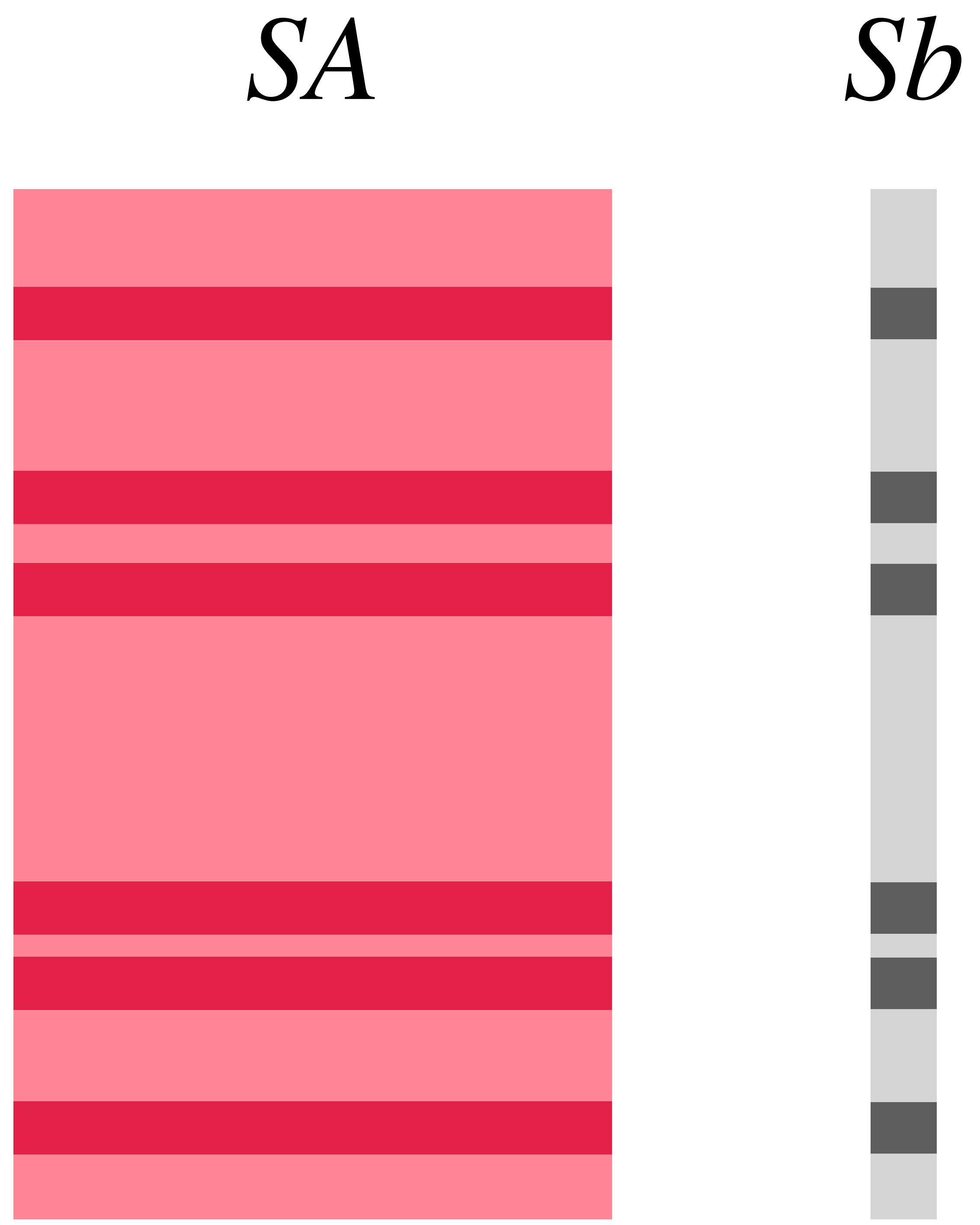
Proof Sketch, $1 < p < 2$

Plan for the Proof

- Part 1: $\tilde{O}\left(\frac{d}{\varepsilon^2}\right)$ bound for $1 < p < 2$
  - Sensitivity sampling + partitions by sensitivity
- Part 2: $\tilde{O}\left(\frac{d}{\varepsilon}\right)$ bound for $1 < p < 2$
  - Strong convexity + iteration

High-Level Algorithmic Approach

- Step 1: Select important training examples
- Step 2: Read the labels corresponding to important training examples
- Step 3: Solve the smaller problem



Sensitivity Sampling

How to select important training examples

- SA should approximate A: $\|SAx\|_p^p = (1 \pm \varepsilon)\|Ax\|_p^p$ for all $x \in \mathbb{R}^d$
- Sampling each row proportionally to its *sensitivity* can get this!

$$\sigma_i(A) := \sup_x \frac{|\langle a_i, x\rangle|^p}{\|Ax\|_p^p}$$

Sensitivity Score: Largest fraction of ℓ_p norm captured by the ith coordinate

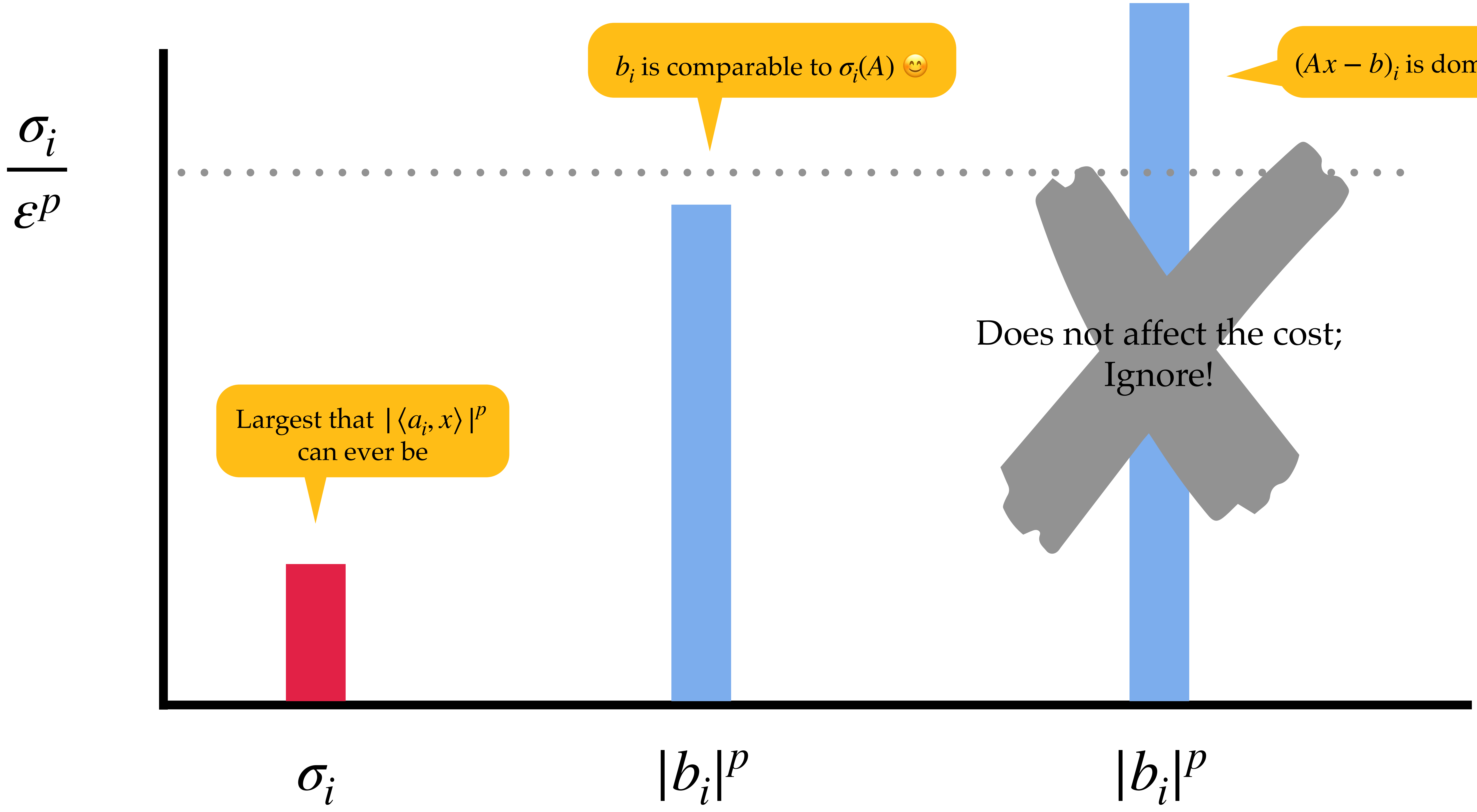
- Key fact: $\sum_{i=1}^{n} \sigma_i(A) \leq d$

Not too many rows can be too important

Problem: b is not a column of A $\rightarrow$
sensitivities of A do not capture large entries of b!

Partitions by Sensitivity

Labels that are too large are not important!



Note: WLOG $\|Ax\|_p^p = O(1)$, $\|b\|_p^p = O(1)$, OPT $= 1$

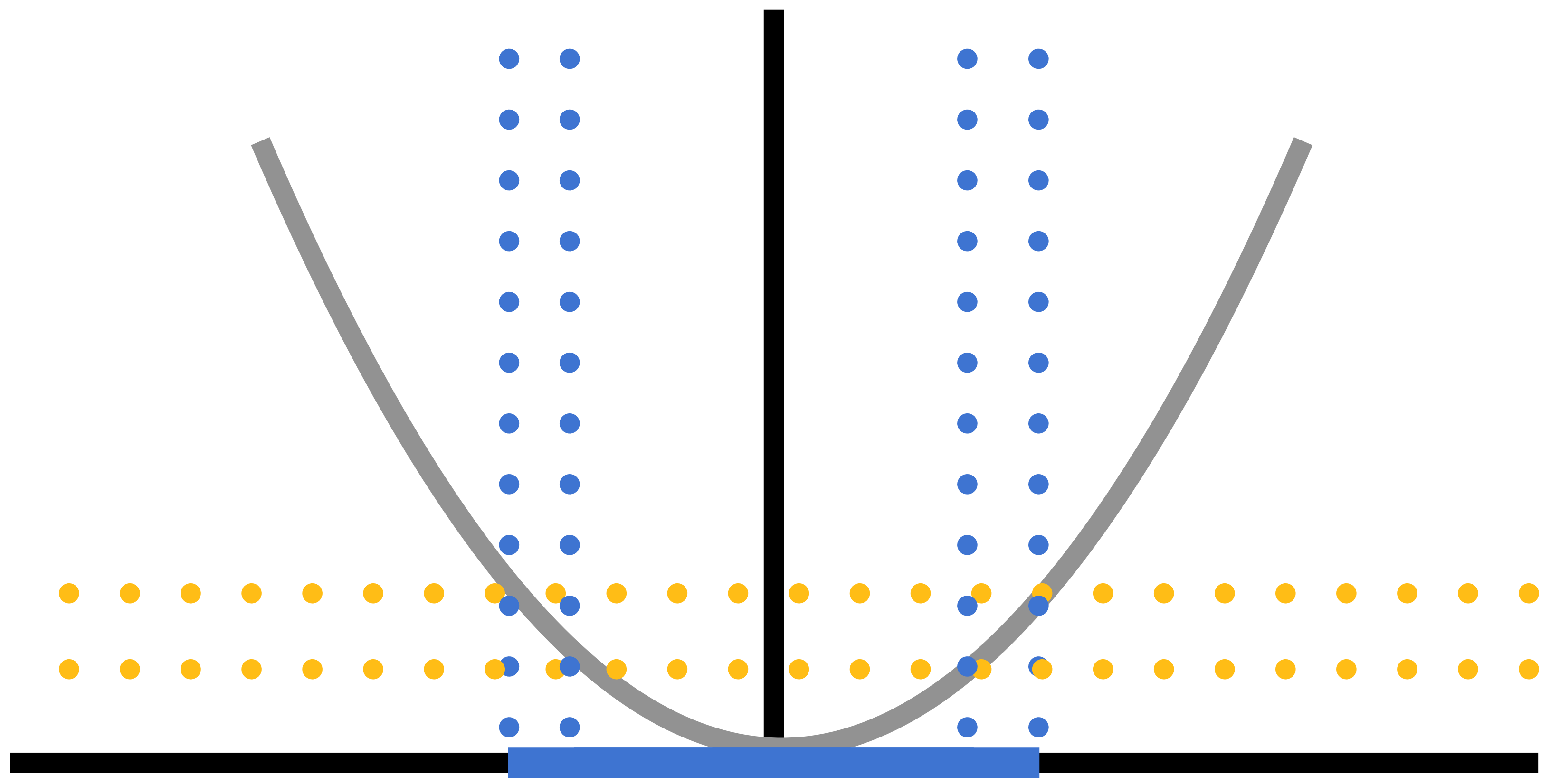
Part 1 Summary

- Sensitivity sampling: selecting important rows of A
  - Problem: this preserves ℓ_p norm objective for Ax, but not for $Ax - b$!
- Partitions by sensitivity: ignoring large entries
  - If b_i is much larger than σ_i, then it cannot affect near-optimal solutions
  - For the purpose of analysis, we can zero out such b_i
  - Sensitivity sampling argument goes through!
- Optimized chaining argument $\rightarrow$ $\tilde{O}(d/\varepsilon^2)$ bound

Can we do better?

Ideas for Optimizing the Argument

- First obtain a $\sqrt{\varepsilon}$-approx. soln. in $\tilde{O}(d/\varepsilon)$ queries
- The ℓ_p loss is *strongly convex* for $p \in (1,2)$
  - If x is nearly optimal, then x is close to x^*
- If x and x^* are close, then we can prove an improved bound on the sensitivity sampling approximation
- Iterate!



Ideas for Optimizing the Argument

Accuracy Boosting

Lemma. If a $(1+\varepsilon)$ approximation can be obtained by making d/ε^{β} queries, then a $(1+\varepsilon)$ approximation can be obtained by making $d/\varepsilon^{\frac{2\beta}{1+\beta}}$ queries.

Recurrence: $\beta_1 = 2$ (by Part 1)

$$\beta_{i+1} = \frac{2\beta_i}{1+\beta_i}$$

$$\beta_i = 1 + \frac{1}{2^i - 1}$$

$\beta_i \to 1$, so we get an algorithm with $\tilde{O}(d/\varepsilon)$ queries! 🥳

Conclusion

- We study *active ℓ_p linear regression*, in which the number of queries to the target vector is minimized
- We give an algorithm based on *sensitivity sampling and partitions by sensitivity* which achieves optimal dependence on d
- For $1 < p < 2$, we achieve optimal dependence on ε by a *strong convexity and iteration argument*

arXiv link