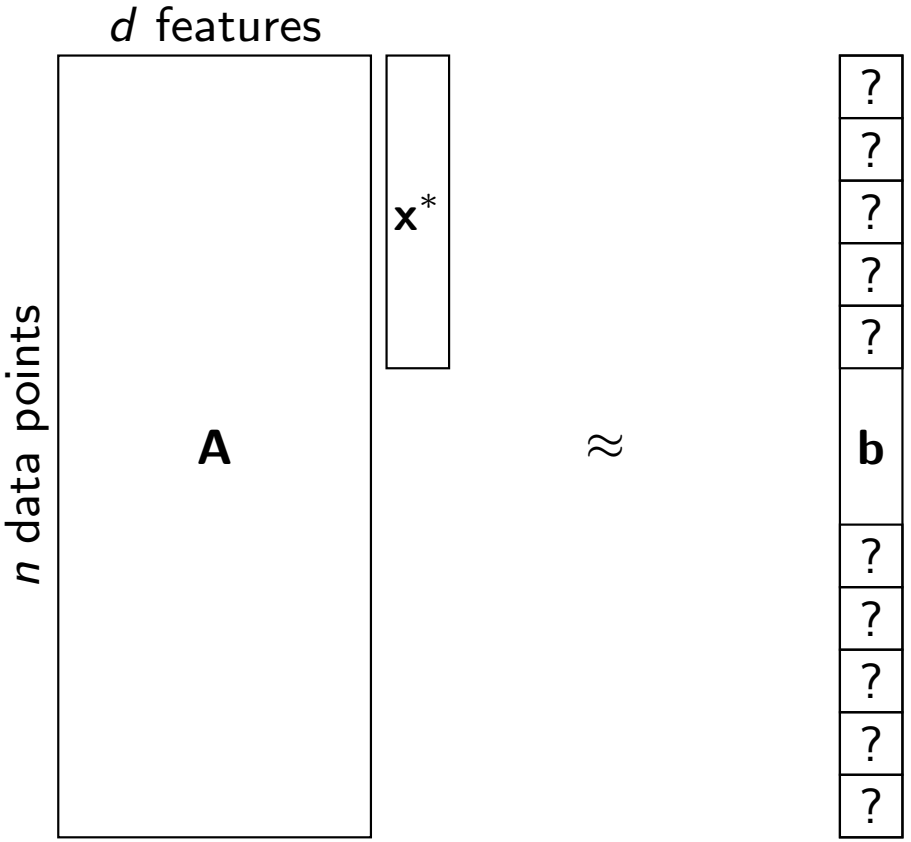
Improved Active Learning via Dependent Leverage Score Sampling

Atsushi Shimizu, Xiaoou Cheng, Christopher Musco, Jonathan Weare

New York University

May 10, 2024

Setup

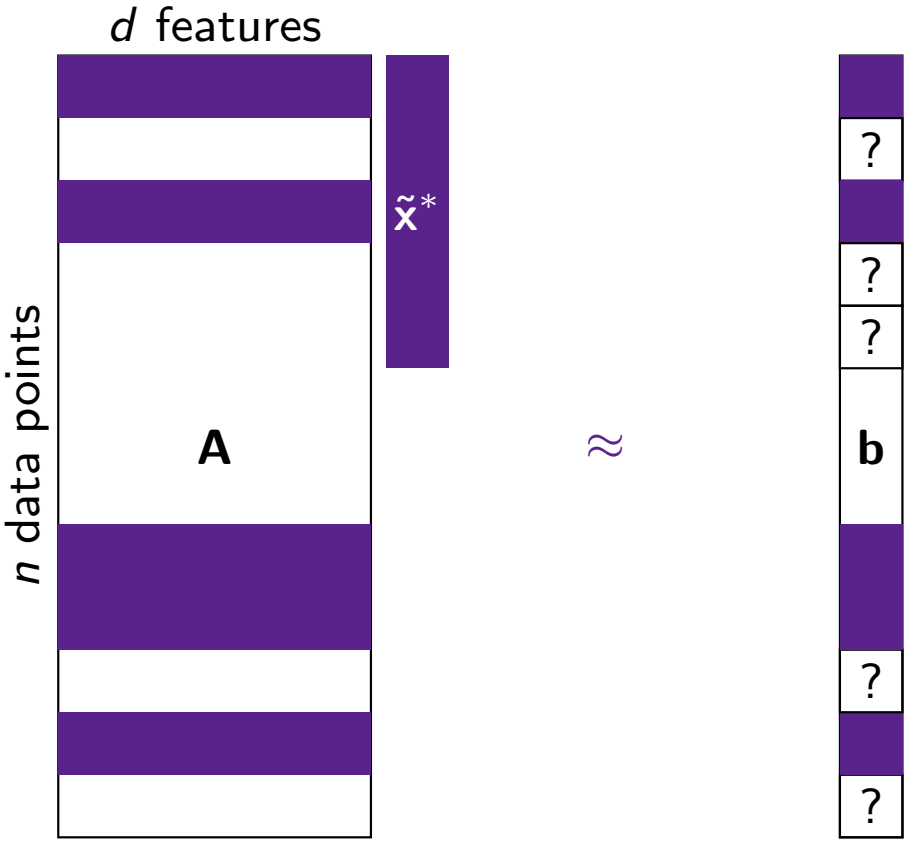


$n \gg d$; $\mathbf{A}$ is known. $\mathbf{b}$ is unknown

We want to solve a least-square problem: find $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$.

Query access to each entry in $\mathbf{b}$ is expensive.

Setup



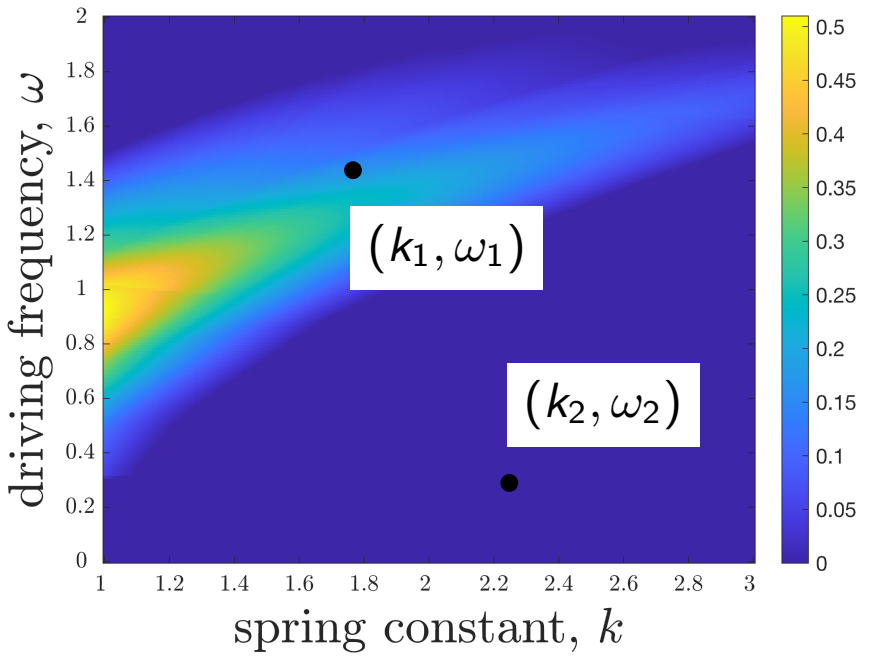
$n \gg d$; $\mathbf{A}$ is known. $\mathbf{b}$ is unknown

We want to solve a least-square problem: find $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$.

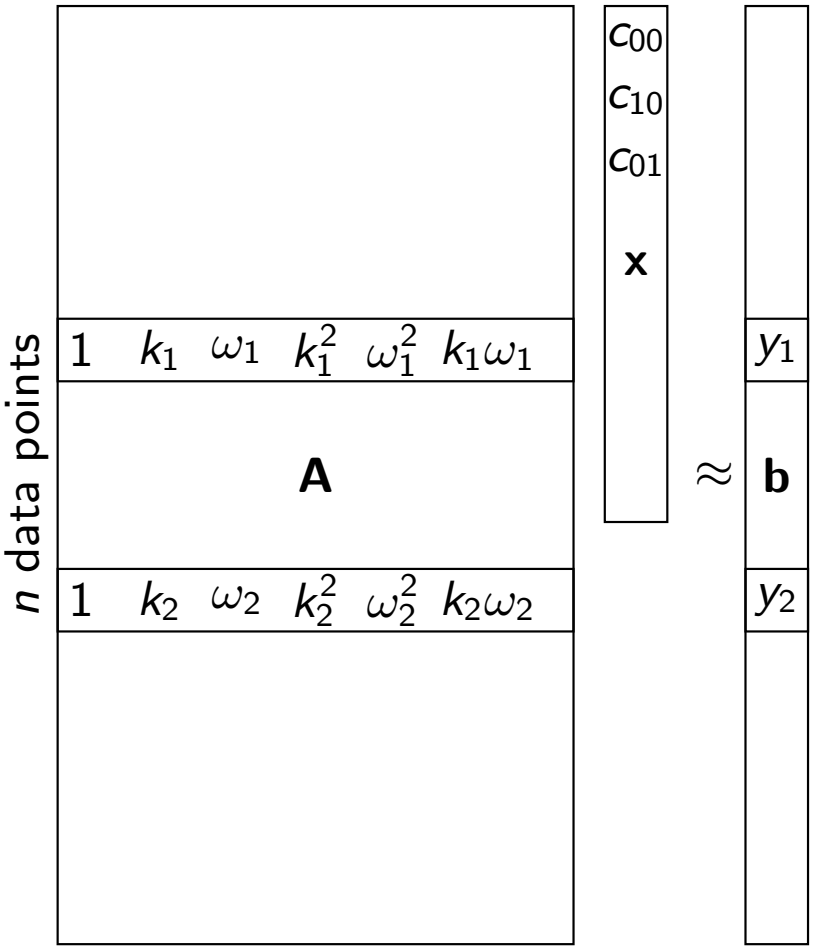
Query access to each entry in $\mathbf{b}$ is expensive.

Query a subset of entries and deal with a smaller linear system. The solution is $\tilde{\mathbf{x}}^*$.

Example



Target Function y: Quantity of Interest (QoI)

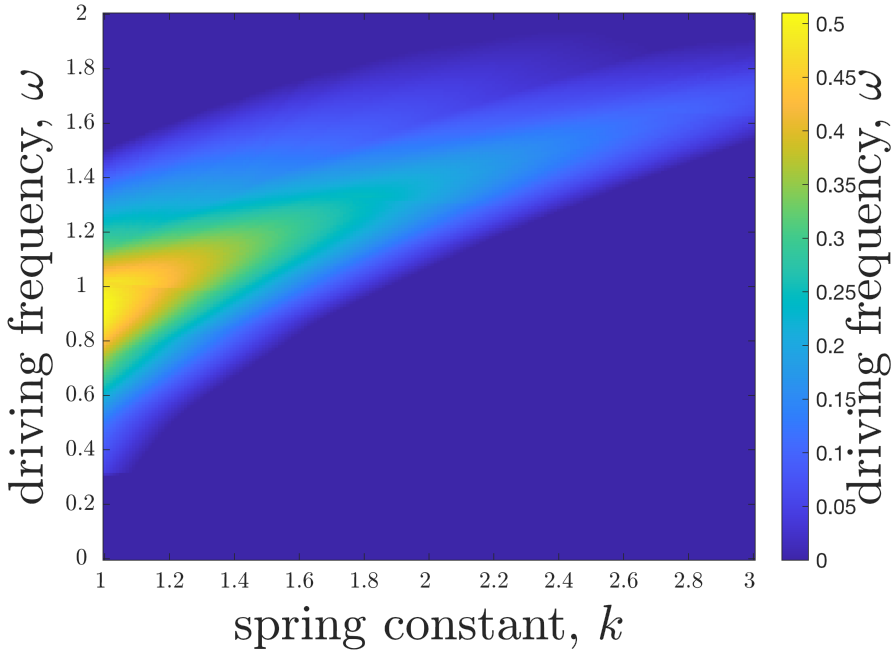


Want to fit a polynomial

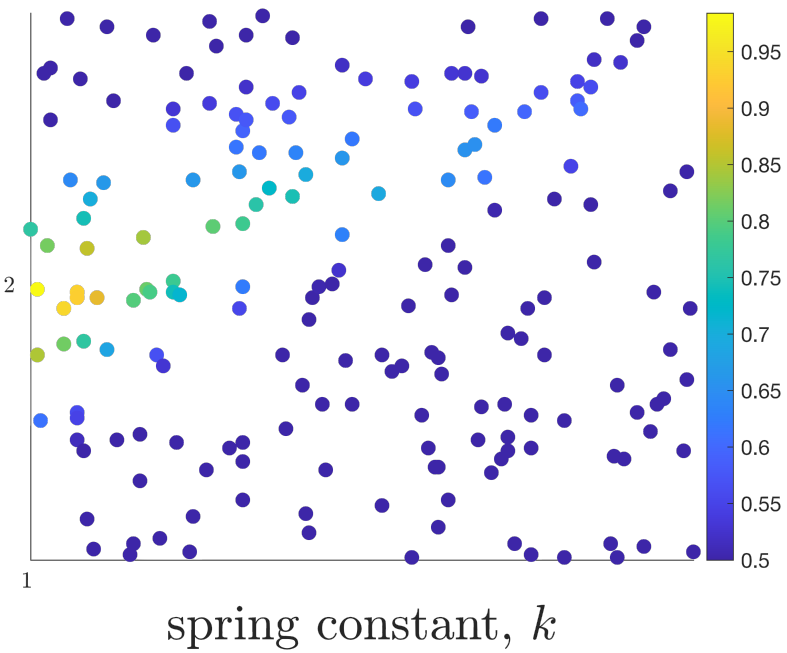
$$y \approx c_{00} + c_{10}k + c_{01}\omega + c_{20}k^2 + c_{11}k\omega + c_{02}\omega^2$$

QoI with <u>each</u> pair of (k, ω) requires solving a PDE $\rightsquigarrow$ expensive

Example



Target Function y: Quantity of Interest (QoI)



Chosen Points

Want to fit a polynomial

$$y \approx c_{00} + c_{10}k + c_{01}\omega + c_{20}k^2 + c_{11}k\omega + c_{02}\omega^2$$

QoI with <u>each</u> pair of (k, ω) requires solving a PDE $\rightsquigarrow$ expensive

Active Linear Regression: Goal

Concrete Goal:

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$, use a small number of queries to a target $\mathbf{b} \in \mathbb{R}^n$ to find $\tilde{\mathbf{x}}^*$ such that

$$\|\mathbf{A}\tilde{\mathbf{x}}^* - \mathbf{b}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2,$$

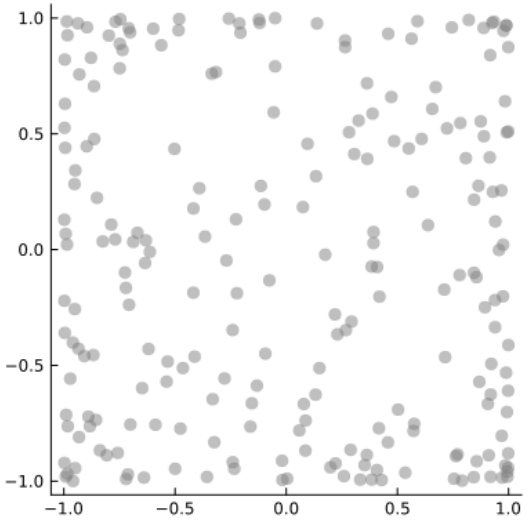
where $\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$.

There are no assumptions on $\mathbf{b}$: "agnostic learning".

e.g. we do not assume $\mathbf{b} = \mathbf{A}\mathbf{x} + \mathcal{N}(0, I)$

Existing Methods

- Bernoulli Sampling + Leverage Score



Complexity: $O(d \log d + d/\epsilon)$

- (Chen and Price 2019) proposes an algorithm with $O(d)$ complexity: hard to implement, worse empirical performance

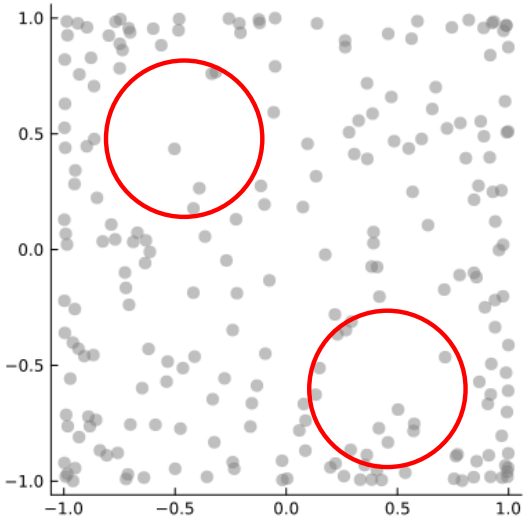
We aim for:

Better empirical performance

Matched or improved theoretical bound

Existing Methods and Our Method

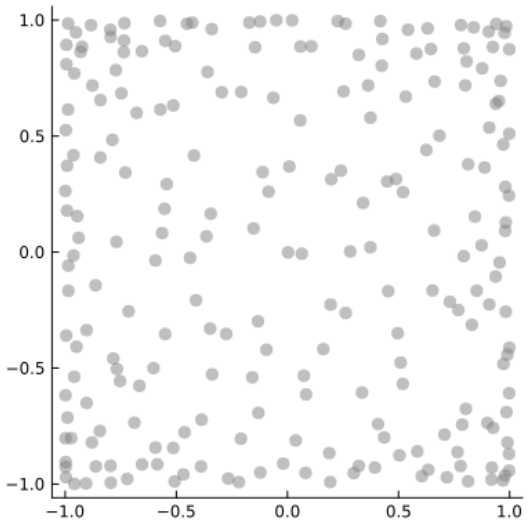
- Bernoulli Sampling + Leverage Score



Complexity: $O(d \log d + d/\epsilon)$

- (Chen and Price 2019) proposes an algorithm with $O(d)$ complexity: hard to implement, worse empirical performance

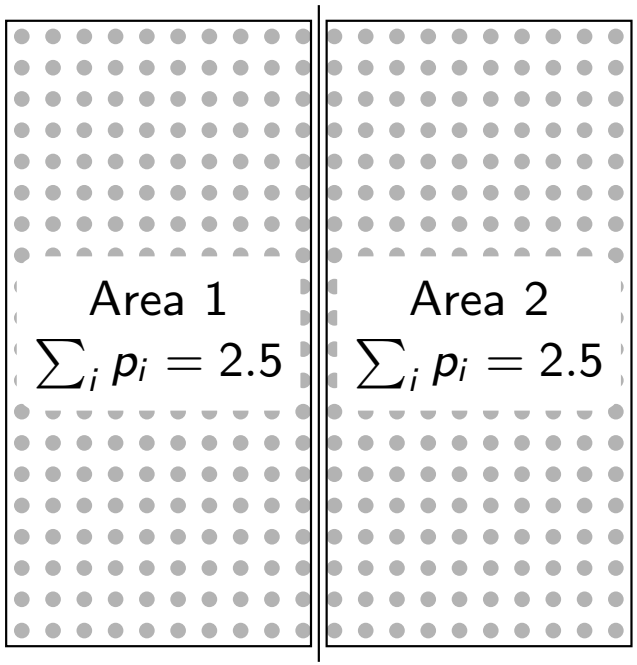
Pivotal Sampling + Leverage Score



Complexity: $O(d \log d + d/\epsilon)$
or
$O(d + d/\epsilon)$

– when fitting a degree d polynomial on an interval (a special case of linear regression with $d + 1$ features)

Spatially-Aware Pivotal Sampling: Key Idea – Stratification



Suppose we decide to choose 5 points.
Point i is included with probability p_i.

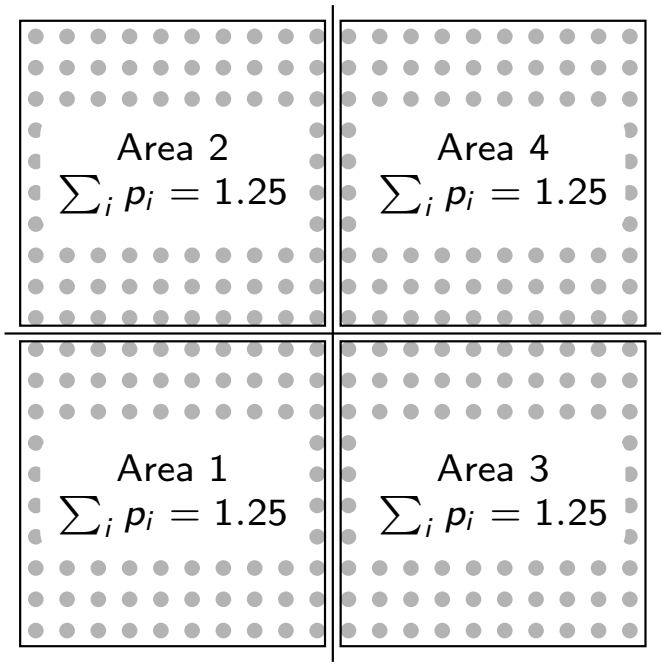
- Bernoulli Sampling: on average, choose 2.5 points from each side.

  independence $\rightsquigarrow$ not unlikely to have all 5 points from the same side

- Stratification: negative correlation

  The number of points coming from each side is of minimal variance $\rightsquigarrow$ either 2 or 3 points

Spatially-Aware Pivotal Sampling: Key Idea – Stratification



- Stratification: negative correlation

  The number of points coming from each side is of minimal variance.

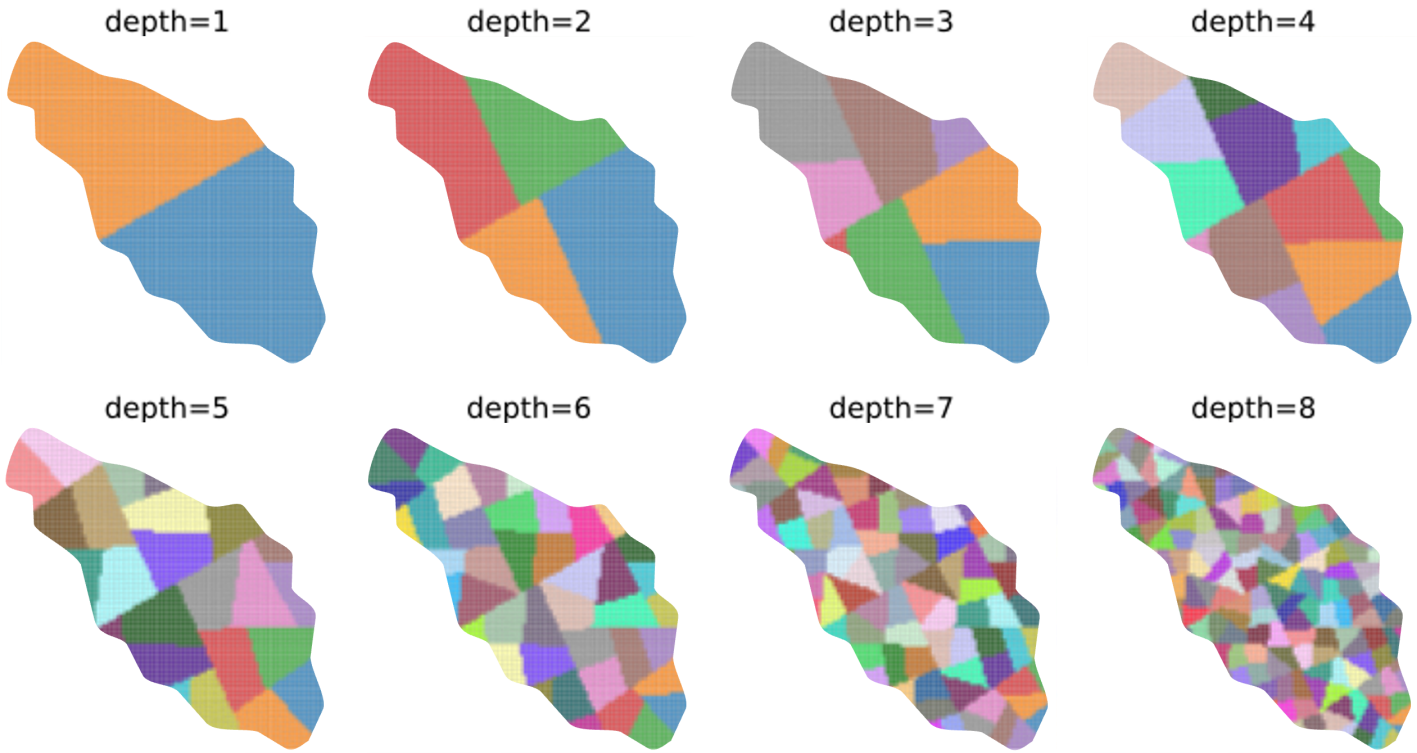
  We would like to partition the space recursively and guarantee minimal variance on each subregion.

  $\rightsquigarrow$ The sampled points are well-spread across the whole space, and still included with probability p_i.

  Binary tree based pivotal sampling (Deville and Tille 1998), combined with the right choice of tree, allows you to carry out this stratification hierarchically.

PCA Variant

Instead of looking at each axis in order, we can also partition the data points with PCA.



Main Result 1: Matched Bound for Any One-sided ℓ_∞ Independence Sampling Distribution

We show that, as long as

- rows from $\mathbf{A}$ are sampled with marginal probabilities proportional to the leverage scores
- pivotal sampling method is used

Then the complexity of the sampling method matches that of independent leverage score sampling.

Theorem

With high probability, pivotal sampling + leverage score method returns $\tilde{\mathbf{x}}^*$ *satisfying* $\|\mathbf{A}\tilde{\mathbf{x}}^* - \mathbf{b}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2$ *while only observing* $O(d \log d + d/\epsilon)$ *entries in* $\mathbf{b}$.

Main Result 1: Matched Bound for Any One-sided ℓ_∞ Independence Sampling Distribution

We show that, as long as

- rows from $\mathbf{A}$ are sampled with marginal probabilities proportional to the leverage scores
- ~~pivotal sampling method is used~~
- the sampling strategy obeys a weak "one-sided ℓ_∞ independence" condition (which includes pivotal sampling)

Then the complexity of the sampling method matches that of independent leverage score sampling.

Theorem

With high probability, any one-sided ℓ_∞ independence sampling method + leverage score returns $\tilde{\mathbf{x}}^*$ satisfying $\|\mathbf{A}\tilde{\mathbf{x}}^* - \mathbf{b}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2$ while only observing $O(d \log d + d/\epsilon)$ entries in $\mathbf{b}$.

Main Result 1: Matched Bound for Any One-sided ℓ_∞ Independence Sampling Distribution

We show that, as long as

- rows from $\mathbf{A}$ are sampled with marginal probabilities proportional to the leverage scores
- ~~pivotal sampling method is used~~
- the sampling strategy obeys a weak "one-sided ℓ_∞ independence" condition (which includes pivotal sampling)

Then the complexity of the sampling method matches that of independent leverage score sampling.

Theorem

With high probability, any one-sided ℓ_∞ independence sampling method + leverage score returns $\tilde{\mathbf{x}}^*$ satisfying $\|\mathbf{A}\tilde{\mathbf{x}}^* - \mathbf{b}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2$ while only observing $O(d \log d + d/\epsilon)$ entries in $\mathbf{b}$.

Main Result 1: Matched Bound for Any One-sided ℓ_∞ Independence Sampling Distribution

We show that, as long as

- rows from $\mathbf{A}$ are sampled with marginal probabilities proportional to the leverage scores
- ~~pivotal sampling me~~ very weak: even allows positive correlation between samples
- the sampling strategy obeys a weak "one-sided ℓ_∞ independence" condition (which includes pivotal sampling)

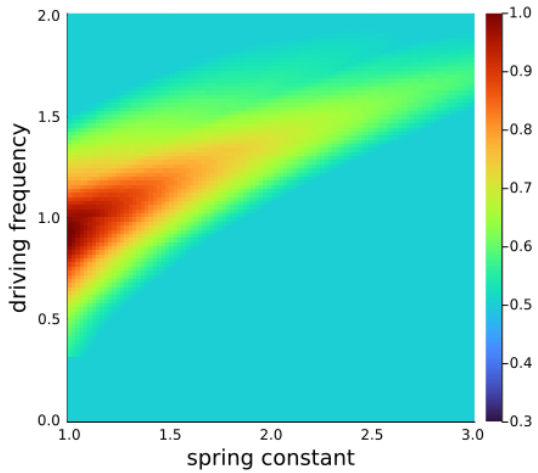
Then the complexity of the sampling method matches that of independent leverage score sampling.

Theorem

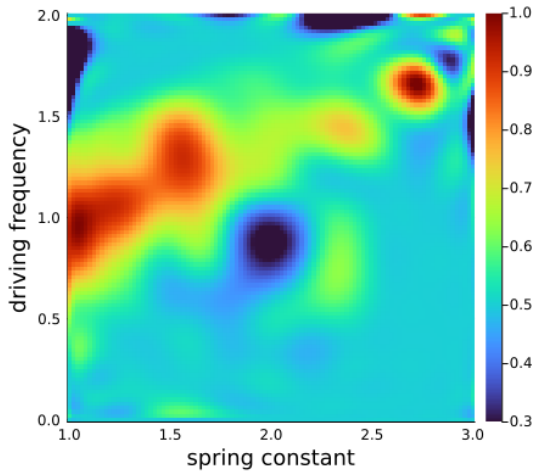
With high probability, any one-sided ℓ_∞ independence sampling method + leverage score returns $\tilde{\mathbf{x}}^*$ satisfying $\|\mathbf{A}\tilde{\mathbf{x}}^* - \mathbf{b}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2$ while only observing $O(d \log d + d/\epsilon)$ entries in $\mathbf{b}$.

Simulation Results: 2D Oscillator

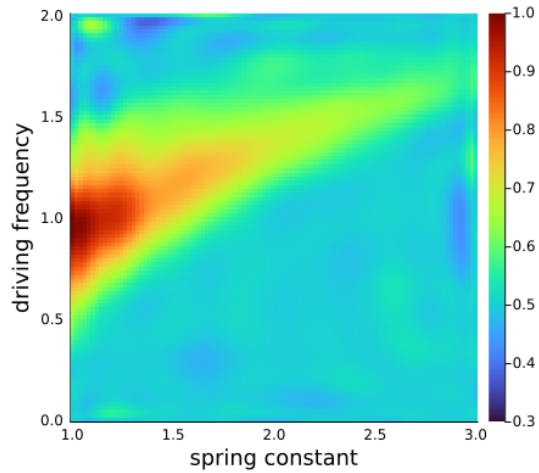
The maximum displacement of a 2D damped harmonic oscillator, as a function of driving frequency and spring constant:



(a) Target Function.



(b) Bernoulli Sampling.

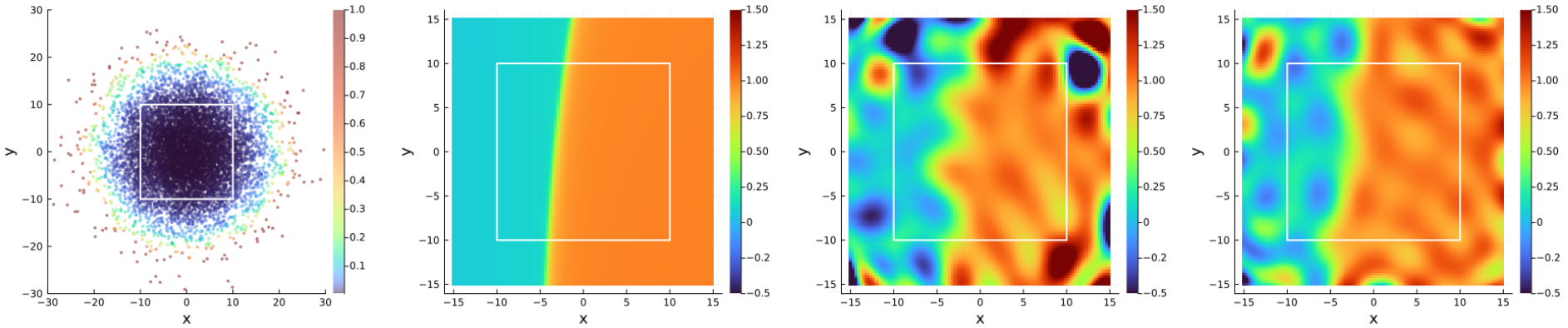


(c) Pivotal Sampling (our method).

Figure 1: (a) is the target value. Both (b) and (c) draw 250 samples using leverage score sampling and perform polynomial regression of degree 20. (b) uses Bernoulli sampling while (c) uses our pivotal sampling method.

Simulation Results: Surface Reaction

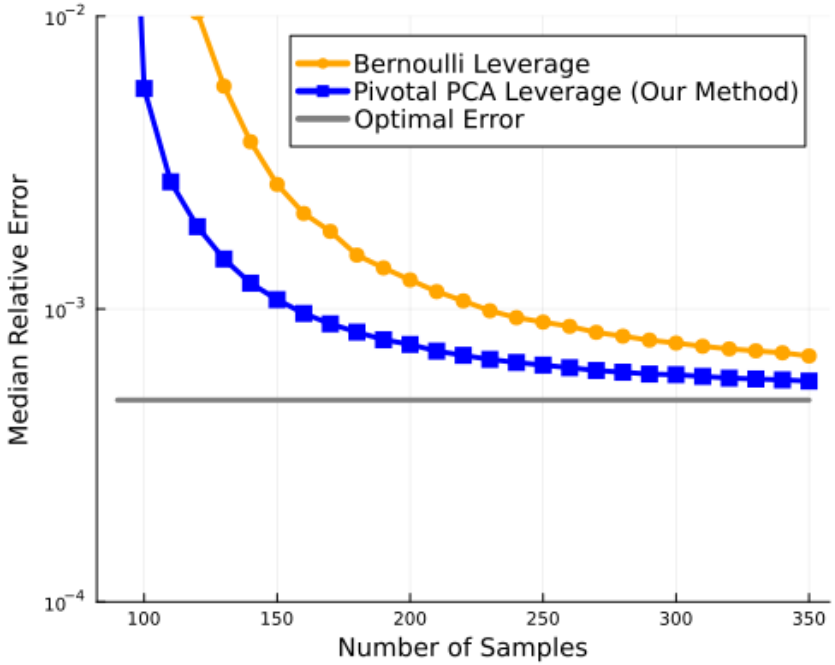
A chemical surface coverage problem:



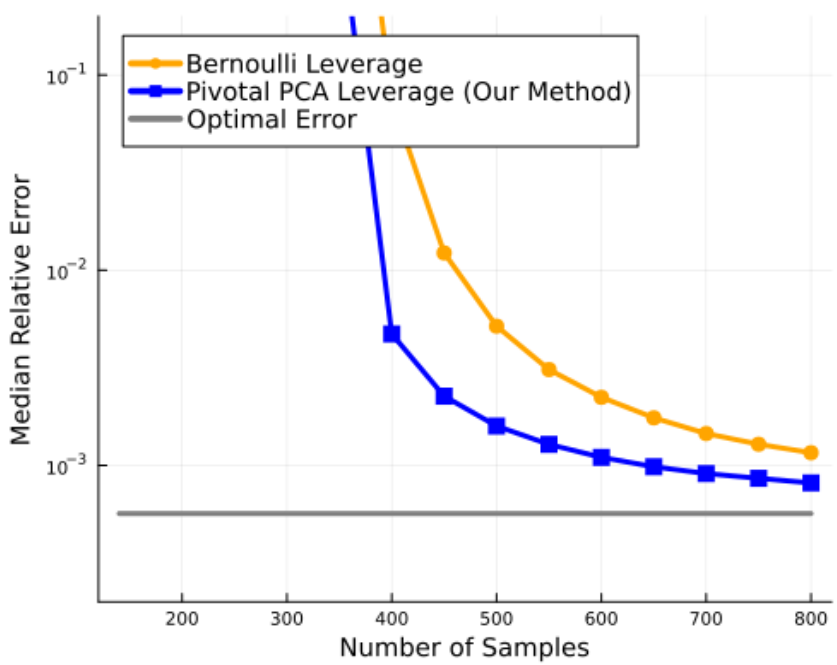
(a) Data Points Colored by Leverage Score. (b) Target Function. (c) Bernoulli Sampling. (d) Pivotal Sampling.

Figure 2: Approximation of the surface reaction model when 425 samples are used to fit a degree 25 polynomial. To help readers focus on the area near the origin, we draw $[-10, 10]^2$ box.

Summary across Examples with an Increasing Number of Samples



(a) Damped Harmonic Oscillator, degree 12.



(b) Surface Reaction, degree 25.

Figure 3: Results for active polynomial regression for the damped harmonic oscillator QoI and the surface reaction model with polynomials.

Our leverage-score based pivotal method outperforms standard Bernoulli leverage score sampling, suggesting the benefits of spatially-aware sampling.

Main Result 2: Improved Bound for Polynomial Regression

In the important special case of polynomial regression, we prove a second result specific to pivotal sampling, showing an improvement on complexity by a log factor.
For this, we consider:

- any function $b : [\ell, u] \to \mathbb{R}$ defined on an interval $[\ell, u] \subset \mathbb{R}$
- fitting b with a degree d polynomial based on evaluations of the function at some points in $[\ell, u]$.

Theorem

With high probability, pivotal sampling + leverage score method constructs a degree d polynomial $\tilde{p}$ satisfying

$$\|\tilde{p} - b\|_2^2 \leq (1+\epsilon) \min_{\text{degree } d \text{ polynomial } p} \|p - b\|_2^2$$

while only collecting $O(d)$ points in $[\ell, u]$.
Here $\|f\|_2^2$ denotes the average squared magnitude $\int_\ell^u f(x)^2 dx$ of a function f.

Compare to: $\Theta(d \log d)$ for Bernoulli sampling + leverage score method

Take-Home Message

Method	Complexity Guarantee
Bernoulli Sampling + Leverage Score	$\Theta(d \log d + d/\epsilon)$
Pivotal Sampling + Leverage Score	$O(d \log d + d/\epsilon)$
Pivotal Sampling + Leverage Score for 1d poly. reg.	$O(d + d/\epsilon)$

Take-Home Message

Method	Complexity Guarantee
Bernoulli Sampling + Leverage Score	$\Theta(d \log d + d/\epsilon)$
One-sided ℓ_∞ Independence Sampling + Leverage Score	$O(d \log d + d/\epsilon)$
Pivotal Sampling + Leverage Score for 1d poly. reg.	$O(d + d/\epsilon)$

Atsushi Shimizu

Christopher Musco

Jonathan weare

arXiv: 2310.04966

Check out our poster #153!