Minimax and Bayesian experimental design: Bridging the gap between statistical and worst-case approaches to least squares regression

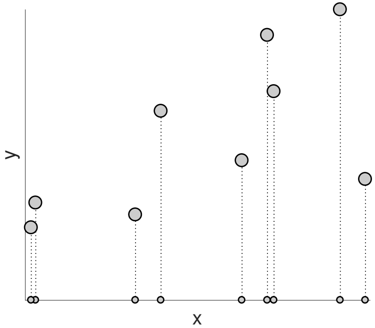
Michael W. Mahoney
ICSI and Department of Statistics, UC Berkeley

Joint work with Michał Dereziński, Feynman Liang, Manfred Warmuth, and Ken Clarkson

September 2019

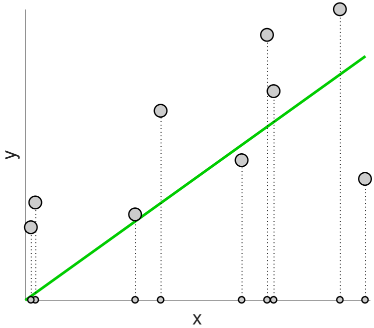
Outline

Bias of the least-squares estimator



$$S = (x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$$

Bias of the least-squares estimator

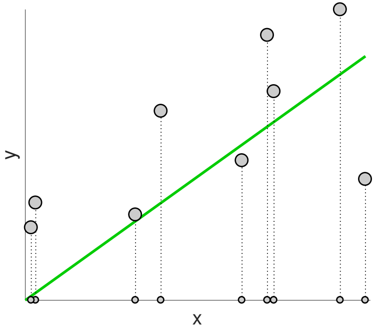


$$S = (x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$$

Statistical regression

$$y = x \cdot w^* + \xi, \quad \mathbb{E}[\,\xi\,] = 0$$

Bias of the least-squares estimator



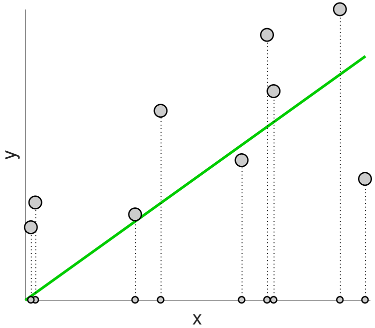
$$S = (x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$$

Statistical regression

$$y = x \cdot w^* + \xi, \quad \mathbb{E}[\,\xi\,] = 0$$

$$w^*(S) = \operatorname*{argmin}_w \sum_i (x_i \cdot w - y_i)^2$$

Bias of the least-squares estimator



$$S = (x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$$

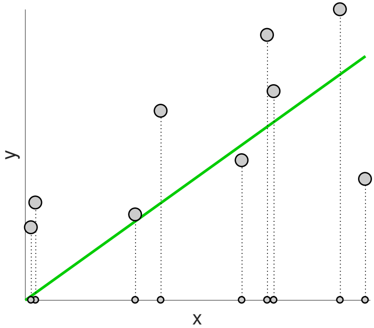
Statistical regression

$$y = x \cdot w^* + \xi, \quad \mathbb{E}[\,\xi\,] = 0$$

$$w^*(S) = \operatorname*{argmin}_w \sum_i (x_i \cdot w - y_i)^2$$

Unbiased! $\mathbb{E}\left[w^*(S)\right] = w^*$

Bias of the least-squares estimator



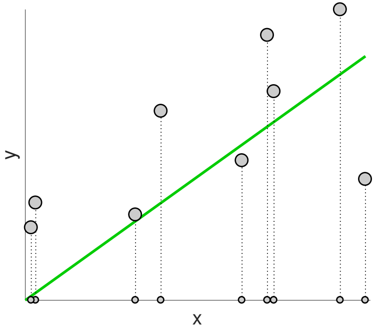
$$S = (x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$$

Worst-case regression

$$w^* = \underset{w}{\operatorname{argmin}}\ \mathbb{E}_{\mathrm{D}}\big[(x \cdot w - y)^2\big]$$

$$w^*(S) = \underset{w}{\operatorname{argmin}} \sum_i (x_i \cdot w - y_i)^2$$

Bias of the least-squares estimator



$$S = (x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$$

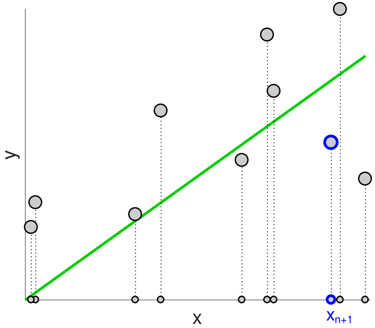
Worst-case regression

$$w^* = \underset{w}{\operatorname{argmin}}\ \mathbb{E}_{\mathrm{D}}\left[(x \cdot w - y)^2\right]$$

$$w^*(S) = \underset{w}{\operatorname{argmin}} \sum_i (x_i \cdot w - y_i)^2$$

Biased! $\quad \mathbb{E}\left[w^*(S)\right] \neq w^*$

Correcting the worst-case bias

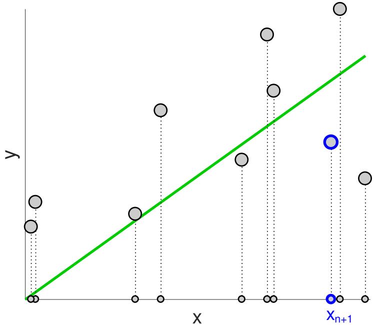


$$S = (x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$$

Worst-case regression

Sample $\quad x_{n+1} \sim x^2 \cdot \mathrm{D}_{\mathcal{X}}$

Correcting the worst-case bias

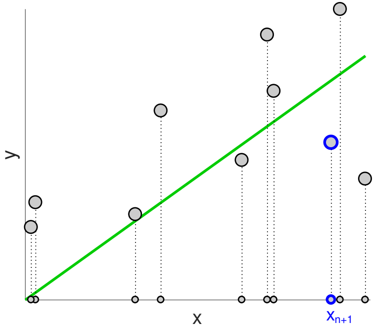


$$S = (x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$$

Worst-case regression

Sample	x_{n+1}	$\sim$	$x^2 \cdot \mathrm{D}_{\mathcal{X}}$
Query	y_{n+1}	$\sim$	$\mathrm{D}_{\mathcal{Y}\mid x = x_{n+1}}$

Correcting the worst-case bias



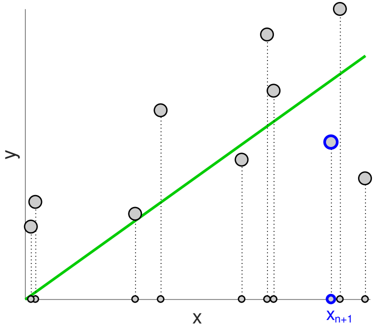
$$S = (x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$$

Worst-case regression

Sample	x_{n+1}	$\sim$	$x^2 \cdot \mathrm{D}_{\mathcal{X}}$
Query	y_{n+1}	$\sim$	$\mathrm{D}_{\mathcal{Y}\mid x=x_{n+1}}$

$$S' \leftarrow S \cup (x_{n+1}, y_{n+1})$$

Correcting the worst-case bias



$$S = (x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$$

Worst-case regression

Sample	x_{n+1}	$\sim$	$x^2 \cdot \mathrm{D}_{\mathcal{X}}$
Query	y_{n+1}	$\sim$	$\mathrm{D}_{\mathcal{Y}\mid x=x_{n+1}}$

$$S' \leftarrow S \cup (x_{n+1}, y_{n+1})$$

Unbiased! $\mathbb{E}\left[w^*(S')\right] = w^*$

In general: *add dimension many points*

Derezinski and Warmuth

Worst-case regression in d dimensions

$$S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}, \qquad (\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$$

Estimate the optimum

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \ \mathbb{E}_{\mathrm{D}}\big[(\mathbf{x}^\top \mathbf{w} - y)^2\big]$$

Volume rescaled sampling

Sample $\overbrace{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+d}}^{d \text{ points}} \sim \det\begin{pmatrix} -\mathbf{x}_{n+1}^\top- \\ \ldots \\ -\mathbf{x}_{n+d}^\top- \end{pmatrix}^2 \cdot (\mathrm{D}_{\mathcal{X}})^d$

Query $y_{n+i} \sim \mathrm{D}_{\mathcal{Y}|\mathbf{x}=\mathbf{x}_{n+i}} \quad \forall_{i=1..d}$

Add $S_\circ = (\mathbf{x}_{n+1}, y_{n+1}), \ldots, (\mathbf{x}_{n+d}, y_{n+d})$ to S

Theorem $\mathbb{E}\big[\mathbf{w}^*(S \cup S_\circ)\big] = \mathbf{w}^*$

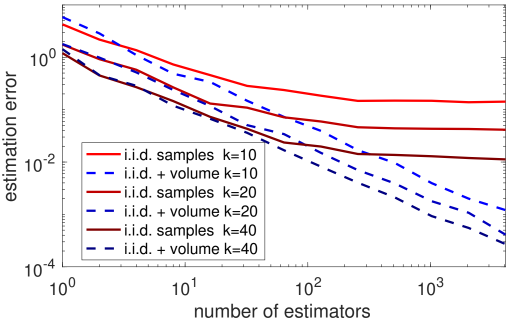
even though $\mathbb{E}\big[\mathbf{w}^*(S)\big] \neq \mathbf{w}^*$

Effect of correcting the bias

Let $\widehat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^*(S_t)$, for independent samples $S_1, ..., S_T$

Question: Is the estimation error $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|$ converging to 0?

$$\text{Example:} \quad \mathbf{x}^\top = (x_1, \ldots, x_5) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad y = \underbrace{\sum_{i=1}^{5} x_i + \frac{x_i^3}{3}}_{\text{nonlinearity}} + \epsilon,$$



Discussion

- First-of-a-kind unbiased estimator for random designs, different than RandNLA sampling theory
- Augmentation uses a determinantal point process (DPP) we call volume-rescaled sampling
- There are many efficient DPP algorithms
- A new mathematical framework for computing expectations

Key application: Experimental design

- Bridge the gap between statistical and worst-case perspectives

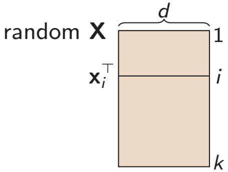
Outline

Volume-rescaled sampling

Derezinski and Warmuth

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ – i.i.d. random vectors sampled from $\mathbf{x} \sim \mathrm{D}_{\mathcal{X}}$

$\mathrm{D}_{\mathcal{X}}^k$ – distribution of $\mathbf{X}$



Volume-rescaled sampling of size k from $\mathrm{D}_{\mathcal{X}}$:

$$\mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^k(\mathbf{X}) \propto \det(\mathbf{X}^\top \mathbf{X})\, \mathrm{D}_{\mathcal{X}}^k(\mathbf{X})$$

Note: For $k = d$, we have $\det(\mathbf{X}^\top \mathbf{X}) = \det(\mathbf{X})^2$

Question: What is the normalization factor of $\mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^k$?

$$\mathbb{E}_{\mathrm{D}_{\mathcal{X}}^k}[\det(\mathbf{X}^\top \mathbf{X})] = ??$$

Can find it through a new proof of the Cauchy-Binet formula!

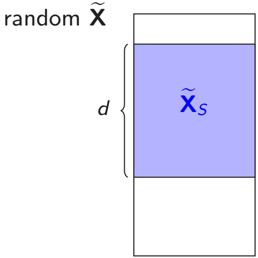
Let $\widetilde{\mathbf{X}} \sim \mathrm{VS}^k_{\mathrm{D}_\mathcal{X}}$ and $S \subseteq [k]$ be a random size d set such that

$$\Pr(S \,|\, \widetilde{\mathbf{X}}) \propto \det(\widetilde{\mathbf{X}}_S)^2.$$

Then:

- $\widetilde{\mathbf{X}}_S \sim \mathrm{VS}^d_{\mathrm{D}_\mathcal{X}}$,
- $\widetilde{\mathbf{X}}_{[k]\backslash S} \sim \mathrm{D}^{k-d}_\mathcal{X}$,
- S is uniformly random,

and the three are independent.



Consequences for least squares

Derezinski and Warmuth

Theorem ([DWH19])

Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)\} \overset{\text{i.i.d.}}{\sim} D^k$, *for any* $k \geq 0$.

$$\text{Sample} \qquad \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_d \ \sim \ \mathrm{VS}^d_{\mathrm{D}_\mathcal{X}},$$

$$\text{Query} \qquad \widetilde{y}_i \ \sim \ D_{\mathcal{Y}|\mathbf{x}=\widetilde{\mathbf{x}}_i} \quad \forall_{i=1..d}.$$

Then for $S_\circ = \{(\widetilde{\mathbf{x}}_1, \widetilde{y}_1), \ldots, (\widetilde{\mathbf{x}}_d, \widetilde{y}_d)\}$,

$$\begin{aligned} \mathbb{E}\big[\mathbf{w}^*(S \cup S_\circ)\big] &= \mathbb{E}_{S \sim D^k}\big[\mathbb{E}_{S_\circ \sim \mathrm{VS}^d_D}[\mathbf{w}^*(S \cup S_\circ)]\big] \\ \text{(decomposition)} \quad &= \mathbb{E}_{\tilde{S} \sim \mathrm{VS}^{k+d}_D}\big[\mathbf{w}^*(\tilde{S})\big] \\ (d\text{-modularity}) \quad &= \mathbf{w}^*. \end{aligned}$$

Outline

Classical statistical regression

We consider n parameterized experiments: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$.
Each experiment has a real random outcome Y_i for $i = 1..n$.

Classical setup:

$$Y_i = \mathbf{x}_i^\top \mathbf{w}^* + \xi_i, \quad \mathbb{E}[\xi_i] = 0, \quad \mathrm{Var}[\xi_i] = \sigma^2, \quad \mathrm{cov}[\xi_i, \xi_j] = 0, \ \ i \neq j$$

The *ordinary least squares* estimator $\mathbf{w}_{\mathrm{LS}} = \mathbf{X}^+ Y$ satisfies:

(unbiasedness) $$\mathbb{E}[\mathbf{w}_{\mathrm{LS}}] = \mathbf{w}^*,$$

(mean squared error) $$\overbrace{\mathbb{E}\,\|\mathbf{w}_{\mathrm{LS}} - \mathbf{w}^*\|^2}^{\mathsf{MSE}(\mathbf{w}_{\mathrm{LS}})} = \sigma^2 \mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^{-1}\big)$$

letting $b = \mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^{-1}\big)$ $$= \frac{b}{n} \cdot \mathbb{E}\,\|\boldsymbol{\xi}\|^2$$

(mean squared prediction error) $$\overbrace{\mathbb{E}\,\|\mathbf{X}(\mathbf{w}_{\mathrm{LS}} - \mathbf{w}^*)\|^2}^{\mathsf{MSPE}(\mathbf{w}_{\mathrm{LS}})} = \sigma^2 d$$

$$= \frac{d}{n} \cdot \mathbb{E}\,\|\boldsymbol{\xi}\|^2$$

Experimental design in classical setting (summary)

Suppose we have a budget of k experiments out of the n choices.
Goal: Select a subset of k experiments $S \subseteq [n]$
Question: How large does k need to be so that:

$$\overbrace{\text{Excess estimation error}}^{\text{MSE or MSPE}} \;\leq\; \epsilon \cdot \overbrace{\text{Total noise}}^{\mathbb{E}\,\|\boldsymbol{\xi}\|^2} \quad ?$$

Denote $L^* = \mathbb{E}\,\|\boldsymbol{\xi}\|^2 = n\sigma^2$.

Prior result:
There is a design $(S, \widehat{\mathbf{w}})$ of size k s.t. $\mathbb{E}[\widehat{\mathbf{w}}_S] = \mathbf{w}^*$ and:

$$\begin{aligned} \mathrm{MSE}(\widehat{\mathbf{w}}_S) - \mathrm{MSE}(\mathbf{w}_{\mathrm{LS}}) \;&\leq\; \epsilon \cdot L^*, \quad \text{for } k \geq d + b/\epsilon, \\ \mathrm{MSPE}(\widehat{\mathbf{w}}_S) - \mathrm{MSPE}(\mathbf{w}_{\mathrm{LS}}) \;&\leq\; \epsilon \cdot L^*, \quad \text{for } k \geq d + d/\epsilon, \end{aligned}$$

where $b = \mathrm{tr}((\mathbf{X}^\top\mathbf{X})^{-1})$.

Experimental design in general setting (summary)

No assumptions on Y_i.
We define $\mathbf{w}^* \stackrel{def}{=} \mathbb{E}[\mathbf{w}_{\mathrm{LS}}] = \mathbf{X}^+\mathbb{E}[Y]$.
Define "total noise" as $L^* \stackrel{def}{=} \mathbb{E}\,\|\boldsymbol{\xi}\|^2$, where $\boldsymbol{\xi} \stackrel{def}{=} \mathbf{X}^\top\mathbf{w}^* - Y$.

Theorem 1 (MSE).
There is a random design $(S, \widehat{\mathbf{w}})$ such that $\mathbb{E}[\widehat{\mathbf{w}}_S] = \mathbf{w}^*$ and

$$\mathrm{MSE}(\widehat{\mathbf{w}}_S) - \mathrm{MSE}(\mathbf{w}_{\mathrm{LS}}) \;\leq\; \epsilon \cdot L^*, \quad \text{for } k = O(d\log n + b/\epsilon),$$

where $b = \mathrm{tr}((\mathbf{X}^\top\mathbf{X})^{-1})$.

Theorem 2 (MSPE).
There is a random design $(S, \widehat{\mathbf{w}})$ such that $\mathbb{E}[\widehat{\mathbf{w}}_S] = \mathbf{w}^*$ and

$$\mathrm{MSPE}(\widehat{\mathbf{w}}_S) - \mathrm{MSPE}(\mathbf{w}_{\mathrm{LS}}) \;\leq\; \epsilon \cdot L^*, \quad \text{for } k = O(d\log n + d/\epsilon).$$

Classical experimental design

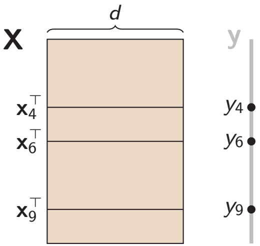
Consider n parameterized experiments: $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$.
Each experiment has a real random response y_i such that:

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, \sigma^2)$$

Goal: Select $k \ll n$ experiments to best estimate $\mathbf{w}^*$

Select $S = \{4, 6, 9\}$

Receive y_4, y_6, y_9



A-optimal design

Find an unbiased estimator $\widehat{\mathbf{w}}$ with smallest *mean squared error*:

$$\min_{\widehat{\mathbf{w}}} \max_{\mathbf{w}^*} \underbrace{\mathbb{E}_{\widehat{\mathbf{w}}}\Big[\|\widehat{\mathbf{w}} - \mathbf{w}^*\|^2\Big]}_{\mathrm{MSE}[\widehat{\mathbf{w}}]} \quad \text{subject to} \quad \mathbb{E}\big[\widehat{\mathbf{w}}\big] = \mathbf{w}^* \quad \forall_{\mathbf{w}^*}$$

Given every $y_1, \ldots, y_n$, the optimum is *least squares*: $\widehat{\mathbf{w}} = \mathbf{X}^\dagger \mathbf{y}$

$$\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] = \mathrm{tr}\big(\mathrm{Var}[\mathbf{X}^\dagger \mathbf{y}]\big) = \sigma^2 \mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^{-1}\big)$$

A-optimal design: $\displaystyle\min_{S:\,|S|\le k} \mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\big)$

Typical required assumption: $y_i = \mathbf{x}_i^\top \mathbf{w}^* + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$

A-optimal design

Find an unbiased estimator $\widehat{\mathbf{w}}$ with smallest *mean squared error*:

$$\min_{\widehat{\mathbf{w}}} \max_{\mathbf{w}^*} \underbrace{\mathbb{E}_{\widehat{\mathbf{w}}}\left[\|\widehat{\mathbf{w}} - \mathbf{w}^*\|^2\right]}_{\mathrm{MSE}[\widehat{\mathbf{w}}]} \quad \text{subject to} \quad \mathbb{E}\left[\widehat{\mathbf{w}}\right] = \mathbf{w}^* \quad \forall_{\mathbf{w}^*}$$

Given set $\{y_i : i \in S\}$, the optimum is *least squares*: $\widehat{\mathbf{w}} = \mathbf{X}_S^\dagger \mathbf{y}_S$

$$\mathrm{MSE}\left[\mathbf{X}_S^\dagger \mathbf{y}_S\right] = \mathrm{tr}\left(\mathrm{Var}[\mathbf{X}_S^\dagger \mathbf{y}_S]\right) = \sigma^2 \mathrm{tr}\left((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\right)$$

A-optimal design: $\displaystyle \min_{S:\,|S| \le k} \mathrm{tr}\left((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\right)$

Typical required assumption: $y_i = \mathbf{x}_i^\top \mathbf{w}^* + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$

A-optimal design: a simple guarantee

Theorem (Avron and Boutsidis, 2013)
For any $\mathbf{X}$ and $k \geq d$ there is S of size k such that:

$$\mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\big) \leq \frac{n-d+1}{k-d+1} \underbrace{\mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^{-1}\big)}_{\text{(denoted } \phi\text{)}}$$

Corollary If $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}$ where $\mathrm{Var}[\boldsymbol{\xi}] = \sigma^2 \mathbf{I}$ and $\mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$ then

$$\underbrace{\mathrm{tr}\big(\mathrm{Var}[\mathbf{X}_S^\dagger \mathbf{y}_S]\big)}_{\sigma^2 \mathrm{tr}((\mathbf{X}_S^\top \mathbf{X}_S)^{-1})} \leq \sigma^2 \frac{n-d+1}{k-d+1} \phi \leq \underbrace{\frac{\phi}{k-d+1}}_{\epsilon} \cdot \underbrace{\mathrm{tr}\big(\mathrm{Var}[\boldsymbol{\xi}]\big)}_{n\sigma^2}$$

$$k = d + \phi/\epsilon \quad \text{and} \quad \mathrm{MSE}[\mathbf{X}_S^\dagger \mathbf{y}_S] \leq \epsilon \cdot \mathrm{tr}(\mathrm{Var}[\boldsymbol{\xi}])$$

A-optimal design: a simple guarantee

Theorem (Avron and Boutsidis, 2013)
For any $\mathbf{X}$ and $k \geq d$ there is S of size k such that:

$$\mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\big) \leq \frac{n-d+1}{k-d+1} \underbrace{\mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^{-1}\big)}_{\text{(denoted } \phi)}$$

Corollary If $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}$ where $\overbrace{\mathrm{Var}[\boldsymbol{\xi}] = \sigma^2 \mathbf{I} \text{ and } \mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}}^{\text{Is this necessary?}}$ then

$$\underbrace{\mathrm{tr}\big(\mathrm{Var}[\mathbf{X}_S^\dagger \mathbf{y}_S]\big)}_{\sigma^2 \mathrm{tr}((\mathbf{X}_S^\top \mathbf{X}_S)^{-1})} \leq \sigma^2 \frac{n-d+1}{k-d+1} \phi \leq \underbrace{\frac{\phi}{k-d+1}}_{\epsilon} \cdot \underbrace{\mathrm{tr}\big(\mathrm{Var}[\boldsymbol{\xi}]\big)}_{n\sigma^2}$$

$$k = d + \phi/\epsilon \quad \text{and} \quad \mathrm{MSE}[\mathbf{X}_S^\dagger \mathbf{y}_S] \leq \epsilon \cdot \mathrm{tr}(\mathrm{Var}[\boldsymbol{\xi}])$$

General response model (What if ξ_i is not $\mathcal{N}(0, \sigma^2)$?)

$\mathcal{F}_n$ - all random vectors in $\mathbb{R}^n$ with finite second moment

$$\mathbf{y} \in \mathcal{F}_n$$

$$\mathbf{w}^* \overset{def}{=} \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{y}}\left[\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2\right] = \mathbf{X}^\dagger \mathbb{E}[\mathbf{y}],$$

$$\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}} \overset{def}{=} \mathbf{y} - \mathbf{X}\mathbf{w}^* = \mathbf{y} - \mathbf{X}\mathbf{X}^\dagger \mathbb{E}[\mathbf{y}]$$ - deviation from best linear predictor

Two special cases:

1. Statistical regression: $\mathbb{E}\left[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\right] = \mathbf{0}$ (mean-zero noise)
2. Worst-case regression: $\mathrm{Var}\left[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\right] = \mathbf{0}$ (deterministic $\mathbf{y}$)

Random experimental designs

Statistical: Fixed S is ok
Worst-case: Fixed S can be exploited by the adversary

Definition

A *random experimental design* $(S, \widehat{\mathbf{w}})$ of size k is:

1. a random set variable $S \subseteq \{1..n\}$ such that $|S| \leq k$
2. a (jointly with S) random function $\widehat{\mathbf{w}} : \mathbb{R}^{|S|} \to \mathbb{R}^d$

Mean squared error of a random experimental design $(S, \widehat{\mathbf{w}})$:

$$\mathrm{MSE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] = \mathbb{E}_{S,\widehat{\mathbf{w}},\mathbf{y}}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}^*\|^2\big]$$

$\mathcal{W}_k(\mathbf{X})$ - family of *unbiased* random experimental designs $(S, \widehat{\mathbf{w}})$:

$$\mathbb{E}_{S,\widehat{\mathbf{w}},\mathbf{y}}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] = \underbrace{\mathbf{X}^\dagger\mathbb{E}[\mathbf{y}]}_{\mathbf{w}^*} \qquad \text{for all } \mathbf{y} \in \mathcal{F}_n$$

Main result

Theorem

For any $\epsilon > 0$, *there is a random experimental design* $(S, \widehat{\mathbf{w}})$ *of size*

$$k = O(d \log n + \phi/\epsilon), \quad \textit{where} \quad \phi = \mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^{-1}\big),$$

such that $(S, \widehat{\mathbf{w}}) \in \mathcal{W}_k(\mathbf{X})$ *(unbiasedness) and for any* $\mathbf{y} \in \mathcal{F}_n$

$$\mathrm{MSE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] - \mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] \leq \epsilon \cdot \mathbb{E}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big]$$

Toy example: $\quad \mathrm{Var}[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}] = \sigma^2 \mathbf{I}, \quad \mathbb{E}[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}] = \mathbf{0}$

1. $\mathbb{E}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big] = \mathrm{tr}\big(\mathrm{Var}[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}]\big)$
2. $\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] = \frac{\phi}{n} \cdot \mathrm{tr}\big(\mathrm{Var}[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}]\big)$

Main result

Theorem

For any $\epsilon > 0$, *there is a random experimental design* $(S, \widehat{\mathbf{w}})$ *of size*

$$k = O(d \log n + \phi/\epsilon), \quad \textit{where} \quad \phi = \mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^{-1}\big),$$

such that $(S, \widehat{\mathbf{w}}) \in \mathcal{W}_k(\mathbf{X})$ *(unbiasedness) and for any* $\mathbf{y} \in \mathcal{F}_n$

$$\mathrm{MSE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] - \mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] \leq \epsilon \cdot \mathbb{E}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big]$$

Toy example: $\quad \mathrm{Var}[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}] = \sigma^2 \mathbf{I}, \quad \mathbb{E}[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}] = \mathbf{0}$

1. $\mathbb{E}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big] = \mathrm{tr}\big(\mathrm{Var}[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}]\big)$
2. $\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] = \frac{\phi}{n} \cdot \mathrm{tr}\big(\mathrm{Var}[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}]\big)$

Important special instances

1. *Statistical regression:* $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}, \quad \mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$

$$\mathrm{MSE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] - \mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] \le \epsilon \cdot \mathrm{tr}\big(\mathrm{Var}[\boldsymbol{\xi}]\big)$$

   - Weighted regression: $\mathrm{Var}[\boldsymbol{\xi}] = \mathrm{diag}\big([\sigma_1^2, \ldots, \sigma_n^2]\big)$
   - Generalized regression: $\mathrm{Var}[\boldsymbol{\xi}]$ is arbitrary
   - Bayesian regression: $\mathbf{w}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

2. *Worst-case regression:* $\mathbf{y}$ is any fixed vector in $\mathbb{R}^n$

$$\mathbb{E}_{S,\widehat{\mathbf{w}}}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}^*\|^2\big] \le \epsilon \cdot \|\mathbf{y} - \mathbf{X}\mathbf{w}^*\|^2$$

where $\mathbf{w}^* = \mathbf{X}^\dagger \mathbf{y}$

Main result: proof outline

1. Volume sampling:
   - to get unbiasedness and expected bounds
   - control MSE in tail of distribution

   1.1 well-conditioned matrices

   1.2 unbiased estimators

2. Error bounds via i.i.d. sampling:
   - to bound sample size k
   - control MSE in bulk of the distribution

   2.1 Leverage score sampling: $\mathrm{Pr}(i) \overset{def}{=} \frac{1}{d}\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i$

   2.2 Inverse score sampling: $\mathrm{Pr}(i) \overset{def}{=} \frac{1}{\phi}\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-2}\mathbf{x}_i$ (new)

3. Proving expected error bounds for least squares

Volume sampling

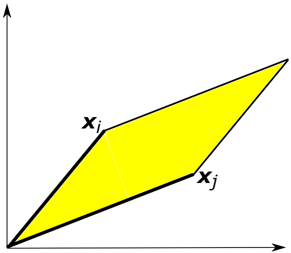
Definition

Given a full rank matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ we define volume sampling $\mathrm{VS}(\mathbf{X})$ as a distribution over sets $S \subseteq [n]$ of size d:

$$\Pr(S) = \frac{\det(\mathbf{X}_S)^2}{\det(\mathbf{X}^\top \mathbf{X})}.$$

$\Pr(S) \sim$ squared volume of the parallelepiped spanned by $\{\mathbf{x}_i : i \in S\}$

Computational cost:
$O(\mathrm{nnz}(\mathbf{X}) \log n + d^4 \log d)$



Unbiased estimators via volume sampling

Under arbitrary response model, any i.i.d. sampling is biased

Theorem ([DWH19])

Volume sampling corrects the least squares bias of i.i.d. sampling.

Let $q = (q_1, \ldots, q_n)$ be some i.i.d. importance sampling.

$$\text{volume + i.i.d.} \quad \overbrace{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_d}}^{\sim \mathrm{VS}(\mathbf{X})}, \overbrace{\mathbf{x}_{i_{d+1}}, \mathbf{x}_{i_{d+2}}, \ldots, \mathbf{x}_{i_k}}^{\sim q^{k-d}}$$

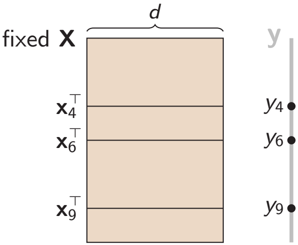
$$\mathbb{E}\left[\operatorname*{argmin}_{\mathbf{w}} \sum_{t=1}^{k} \frac{1}{q_{i_t}} (\mathbf{x}_{i_t}^\top \mathbf{w} - y_{i_t})^2\right] = \mathbf{w}^*_{\mathbf{y}|\mathbf{X}}$$

Key idea: volume-rescaled importance sampling

Simple volume-rescaled sampling:

- Let $\mathrm{D}_{\mathcal{X}}$ be a uniformly random $\mathbf{x}_i$
- $(\mathbf{X}_S, \mathbf{y}_S) \sim \mathrm{VS}_D^k$ and $\widehat{\mathbf{w}} = \mathbf{X}_S^{\dagger}\mathbf{y}_S$.

Then, $\mathbb{E}[\widehat{\mathbf{w}}] = \mathbf{w}^*_{\mathbf{y}|\mathbf{X}}$.



Problem: Not robust to worst-case noise
Solution: Volume-rescaled importance sampling

- Let $p = (p_1, \dots, p_n)$ be an importance sampling distribution,
- Define $\widetilde{\mathbf{x}} \sim \mathrm{D}_{\mathcal{X}}$ as $\widetilde{\mathbf{x}} = \frac{1}{\sqrt{p_i}}\mathbf{x}_i$ for $i \sim p$.

Then, for $(\widetilde{\mathbf{X}}_S, \widetilde{\mathbf{y}}_S) \sim \mathrm{VS}_D^k$ and $\widehat{\mathbf{w}} = \widetilde{\mathbf{X}}_S^{\dagger}\widetilde{\mathbf{y}}_S$, we have $\mathbb{E}[\widehat{\mathbf{w}}] = \mathbf{w}^*_{\mathbf{y}|\mathbf{X}}$.

Importance sampling for experimental design

1. *Leverage score sampling*: $\Pr(i) = p_i^{\mathrm{lev}} \overset{def}{=} \frac{1}{d}\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i$

   A standard sampling method for worst-case linear regression.

2. *Inverse score sampling*: $\Pr(i) = p_i^{\mathrm{inv}} \overset{def}{=} \frac{1}{\phi}\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-2}\mathbf{x}_i$.

   A novel sampling technique essential for achieving $O(\phi/\epsilon)$ sample size.

Minimax A-optimality and Minimax experimental design

Definition

Minimax A-optimal value for experimental design:

$$R_k^*(\mathbf{X}) \overset{def}{=} \min_{(S,\widehat{\mathbf{w}})\in\mathcal{W}_k(\mathbf{X})} \ \max_{\mathbf{y}\in\mathcal{F}_n\setminus\mathrm{Sp}(\mathbf{X})} \frac{\mathrm{MSE}\left[\widehat{\mathbf{w}}(\mathbf{y}_S)\right] - \mathrm{MSE}\left[\mathbf{X}^\dagger\mathbf{y}\right]}{\mathbb{E}\left[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\right]}$$

Fact. $\mathbf{X}^\dagger\mathbf{y}$ is the *minimum variance unbiased estimator* for $\mathcal{F}_n$:

$$\begin{aligned} \text{if} \quad & \mathbb{E}_{\mathbf{y},\widehat{\mathbf{w}}}\left[\widehat{\mathbf{w}}(\mathbf{y})\right] = \mathbf{X}^\dagger\mathbb{E}[\mathbf{y}] \qquad && \forall_{\mathbf{y}\in\mathcal{F}_n} \\ \text{then} \quad & \mathrm{Var}\left[\widehat{\mathbf{w}}(\mathbf{y})\right] \succeq \mathrm{Var}\left[\mathbf{X}^\dagger\mathbf{y}\right] && \forall_{\mathbf{y}\in\mathcal{F}_n} \end{aligned}$$

- If $d \le k \le n$, then $R_k^*(\mathbf{X}) \in [0,\infty)$
- If $k \ge C\cdot d\log n$, then $R_k^*(\mathbf{X}) \le C\cdot\phi/k$ for some C
- If $k^2 < \epsilon nd/3$, then $R_k^*(\mathbf{X}) \ge (1-\epsilon)\cdot\phi/k$ for some $\mathbf{X}$

Alternative: mean squared *prediction* error

Definition. $\mathrm{MSPE}\big[\widehat{\mathbf{w}}\big] = \mathbb{E}\big[\|\mathbf{X}(\widehat{\mathbf{w}} - \mathbf{w}^*)\|^2\big]$ (V-optimality)

Theorem

There is $(S, \widehat{\mathbf{w}})$ *of size* $k = O(d\log n + d/\epsilon)$ *s.t. for any* $\mathbf{y} \in \mathcal{F}_n$,

$$\mathrm{MSPE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] - \mathrm{MSPE}\big[\mathbf{X}^\dagger\mathbf{y}\big] \le \epsilon \cdot \mathbb{E}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big]$$

Follows from the MSE bound by reduction to $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$.

Then $\mathrm{MSPE}\big[\widehat{\mathbf{w}}\big] = \mathrm{MSE}\big[\widehat{\mathbf{w}}\big]$ and $\phi = d$.

Minimax V-optimal value:

$$\min_{(S,\widehat{\mathbf{w}})\in\mathcal{W}_k(\mathbf{X})} \; \max_{\mathbf{y}\in\mathcal{F}_n\backslash\mathrm{Sp}(\mathbf{X})} \frac{\mathrm{MSPE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] - \mathrm{MSPE}\big[\mathbf{X}^\dagger\mathbf{y}\big]}{\mathbb{E}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big]}$$

Questions about minimax experimental design

1. Can $R_k^*(\mathbf{X})$ be found, exactly or approximately?
2. What happens in the regime of $k \leq C \cdot d \log n$?
3. Can we restrict $\mathcal{W}_k(\mathbf{X})$ to only tractable experimental designs?
4. Does the minimax-value change when you restrict $\mathcal{F}_n$?
    4.1 Weighted regression
    4.2 Generalized regression
    4.3 Bayesian regression
    4.4 Worst-case regression

Reduction to worst-case regression

Theorem

W.l.o.g. we can replace random $\mathbf{y} \in \mathcal{F}_n$ *with fixed* $\mathbf{y} \in \mathbb{R}^n$*:*

$$R_k^*(\mathbf{X}) = \min_{(S,\widehat{\mathbf{w}}) \in \mathcal{W}_k(\mathbf{X})} \max_{\mathbf{y} \in \mathbb{R}^n \setminus \mathrm{Sp}(\mathbf{X})} \frac{\mathbb{E}_{S,\widehat{\mathbf{w}}}\left[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{X}^\dagger \mathbf{y}\|^2\right]}{\|\mathbf{y} - \mathbf{X}\mathbf{X}^\dagger \mathbf{y}\|^2}$$

Suppose $(S, \widehat{\mathbf{w}})$ for all fixed response vectors $\mathbf{y} \in \mathbb{R}^n$ satisfies

$$\mathbb{E}\left[\widehat{\mathbf{w}}(\mathbf{y}_S)\right] = \mathbf{X}^\dagger \mathbf{y} \quad \text{and} \quad \mathbb{E}\left[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{X}^\dagger \mathbf{y}\|^2\right] \le \epsilon \cdot \|\mathbf{y} - \mathbf{X}\mathbf{X}^\dagger \mathbf{y}\|^2.$$

Then, for all random response vectors $\mathbf{y} \in \mathcal{F}_n$ and $\mathbf{w}^* \in \mathbb{R}^d$,

$$\underbrace{\mathbb{E}\left[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}^*\|^2\right]}_{\mathrm{MSE}[\widehat{\mathbf{w}}(\mathbf{y}_S)]} \le \underbrace{\mathbb{E}\left[\|\mathbf{X}^\dagger \mathbf{y} - \mathbf{w}^*\|^2\right]}_{\mathrm{MSE}[\mathbf{X}^\dagger \mathbf{y}]} + \epsilon \cdot \mathbb{E}\left[\|\mathbf{y} - \mathbf{X}\mathbf{w}^*\|^2\right].$$

Outline

Bayesian experimental design

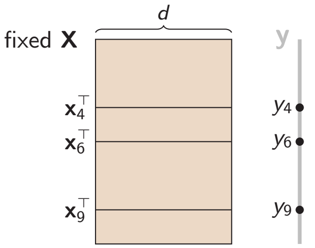
Consider n parameterized experiments: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$.
Each experiment has a real random response y_i such that:

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, \sigma^2), \quad \mathbf{w}^* \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{A}^{-1})$$

Goal: Select $k \ll n$ experiments to best estimate $\mathbf{w}^*$

Select $S = \{4, 6, 9\}$

Receive y_4, y_6, y_9



Bayesian A-optimal design

Given the Bayesian assumptions, we have

$$\mathbf{w} \mid \mathbf{y}_S \;\sim\; \mathcal{N}\Big(\, (\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A})^{-1} \mathbf{X}_S^\top \mathbf{y}_S,\;\; \sigma^2 (\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A})^{-1} \,\Big),$$

Bayesian A-optimality criterion:

$$f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) = \mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A})^{-1}\big).$$

Goal: Efficiently find subset S of size k such that:

$$f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \le (1+\epsilon) \cdot \underbrace{\min_{S' : |S'| = k} f_{\mathbf{A}}(\mathbf{X}_{S'}^\top \mathbf{X}_{S'})}_{\mathrm{OPT}_k}$$

Relaxation to a semi-definite program

SDP relaxation

The following can be found via an SDP solver in polynomial time:

$$p^* \;=\; \operatorname*{argmin}_{p_1,\ldots,p_n} \; f_{\mathbf{A}}\Big(\sum_{i=1}^{n} p_i \mathbf{x}_i \mathbf{x}_i^\top\Big),$$

$$\text{subject to} \quad \forall_i \;\; 0 \le p_i \le 1, \quad \sum_i p_i = k.$$

The solution p^* satisfies $f_{\mathbf{A}}\big(\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top\big) \le \mathrm{OPT}_k$.

Question: For what k can we efficiently round this to S of size k?

Efficient rounding for effective dimension many points

Definition

Define **A**-effective dimension as $d_{\mathbf{A}} = \mathrm{tr}\big(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \mathbf{A})^{-1}\big) \le d$.

Theorem ([DLM19])

If $k = \Omega\big(\frac{d_{\mathbf{A}}}{\epsilon} + \frac{\log 1/\epsilon}{\epsilon^2}\big)$, then there is a polynomial time algorithm that finds subset S of size k such that

$$f_{\mathbf{A}}\big(\mathbf{X}_S^\top\mathbf{X}_S\big) \le (1+\epsilon)\cdot \mathrm{OPT}_k.$$

Remark: Extends to other Bayesian criteria: C/D/V-optimality.

Key idea: Rounding with **A**-regularized volume-rescaled sampling, a new kind of determinantal point process.

Comparison with prior work

	Criteria	Bayesian	$k = \Omega(\cdot)$
[WYS17]	A,V	✗	$\frac{d^2}{\epsilon}$
[AZLSW17]	A,C,D,E,G,V	✓	$\frac{d}{\epsilon^2}$
[NSTT19]	A,D	✗	$\frac{d}{\epsilon} + \frac{\log 1/\epsilon}{\epsilon^2}$
our result [DLM19]	A,C,D,V	✓	$\frac{d_{\mathbf{A}}}{\epsilon} + \frac{\log 1/\epsilon}{\epsilon^2}$

Outline

Conclusions

Unbiased estimators for least squares, uses volume sampling

Recent developments:

- Experimental design without any noise assumptions, i.e., arbitrary response
- Minimax experimental design: bridging the gap bw statistical and worst-case perspectives
- Applications in Bayesian experimental design: bridging the gap bw experimental design and determinantal point processes

Going beyond least squares:

- extensions to non-square losses,
- applications in distributed optimization.

References

Haim Avron and Christos Boutsidis.
Faster subset selection for matrices and applications.
SIAM Journal on Matrix Analysis and Applications, 34(4):1464–1499, 2013.

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang.
Near-optimal design of experiments via regret minimization.
In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 126–135, Sydney, Australia, August 2017.

Michał Dereziński, Kenneth L. Clarkson, Michael W. Mahoney, and Manfred K. Warmuth.
Minimax experimental design: Bridging the gap between statistical and worst-case approaches to least squares regression.
In Proceedings of the 32nd Conference on Learning Theory, 2019.

Michał Dereziński, Feynman Liang, and Michael W. Mahoney.
Distributed estimation of the inverse Hessian by determinantal averaging.
arXiv e-prints (to appear), June 2019.

Michał Dereziński and Manfred K. Warmuth.
Reverse iterative volume sampling for linear regression.
Journal of Machine Learning Research, 19(23):1–39, 2018.

Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu.
Correcting the bias in least squares regression with volume-rescaled sampling.
In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, 2019.

Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat.
Proportional volume sampling and approximation algorithms for a -optimal design.
In Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1369–1386