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
Correcting the bias in least squares regression

Volume-rescaled sampling

Minimax experimental design

Bayesian experimental design

Conclusions
Bias of the least-squares estimator

\[ S = (x_1, y_1), \ldots, (x_n, y_n) \overset{i.i.d.}{\sim} D \]
Bias of the least-squares estimator

Statistical regression

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

\[ S = (x_1, y_1), \ldots, (x_n, y_n) \text{ i.i.d.} \sim D \]
Bias of the least-squares estimator

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

Statistical regression

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

\[ w^*(S) = \arg\min_w \sum_i (x_i \cdot w - y_i)^2 \]
Bias of the least-squares estimator

Statistical regression

\[ S = (x_1, y_1), \ldots, (x_n, y_n) \text{ i.i.d. } D \]

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

\[ w^*(S) = \arg\min_w \sum_i (x_i \cdot w - y_i)^2 \]

Unbiased! \[ \mathbb{E}[w^*(S)] = w^* \]
Bias of the least-squares estimator

\[ S = (x_1, y_1), \ldots, (x_n, y_n) \sim \text{i.i.d.} D \]

Worst-case regression

\[ w^* = \arg\min_w \mathbb{E}_D[(x \cdot w - y)^2] \]

\[ w^*(S) = \arg\min_w \sum_i (x_i \cdot w - y_i)^2 \]
Bias of the least-squares estimator

\[ S = (x_1, y_1), \ldots, (x_n, y_n) \sim D \]

**Worst-case regression**

\[ w^* = \arg\min_w \mathbb{E}_D[(x \cdot w - y)^2] \]

\[ w^*(S) = \arg\min_w \sum_i (x_i \cdot w - y_i)^2 \]

Biased! \[ \mathbb{E}[w^*(S)] \neq w^* \]
Correcting the worst-case bias

\[ S = (x_1, y_1), \ldots, (x_n, y_n) \text{ i.i.d. } \sim D \]

Worst-case regression

Sample \[ x_{n+1} \sim x^2 \cdot D \chi \]
Correcting the worst-case bias

\[ S = (x_1, y_1), \ldots, (x_n, y_n) \text{ i.i.d. } \sim D \]

Worst-case regression

- Sample: \( x_{n+1} \sim x^2 \cdot D_x \)
- Query: \( y_{n+1} \sim D_{y|x=x_{n+1}} \)
Correcting the worst-case bias

\[ S = (x_1, y_1), \ldots, (x_n, y_n) \ \text{i.i.d.} \sim D \]

Worst-case regression

Sample \[ x_{n+1} \sim x^2 \cdot D_x \]
Query \[ y_{n+1} \sim D_{y|x=x_{n+1}} \]

\[ S' \leftarrow S \cup (x_{n+1}, y_{n+1}) \]
Correcting the worst-case bias

\[ S = (x_1, y_1), \ldots, (x_n, y_n) \, \text{i.i.d.} \, \sim D \]

Worst-case regression

Sample \( x_{n+1} \sim x^2 \cdot D_x \)

Query \( y_{n+1} \sim D_{Y|x=x_{n+1}} \)

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

Unbiased! \[ \mathbb{E}[w^*(S')] = w^* \]
In general: *add dimension many points*

Derezinski and Warmuth

**Worst-case regression** in $d$ dimensions

$$S = (x_1, y_1), \ldots, (x_n, y_n)^{i.i.d.} \sim D, \quad (x, y) \in \mathbb{R}^d \times \mathbb{R}$$

**Estimate the optimum**

$$w^* = \arg\min_{w \in \mathbb{R}^d} \mathbb{E}_D[(x^T w - y)^2]$$

**Volume rescaled sampling**

Sample $d$ points $x_{n+1}, \ldots, x_{n+d} \sim \det\left(\begin{array}{c} -x_{n+1}^T \\
\vdots \\
-x_{n+d}^T 
\end{array}\right) \cdot (D_X)^d$

Query $y_{n+i} \sim D_{Y|x=x_{n+i}} \quad \forall i = 1..d$

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

**Theorem**

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

even though

$$\mathbb{E}[w^*(S)] \neq w^*$$
Effect of correcting the bias

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

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

Example: $x^\top = (x_1, \ldots, x_5) \sim \mathcal{N}(0, 1)$, $y = \sum_{i=1}^{5} x_i + \frac{x_i^3}{3} + \epsilon$, nonlinearity

![Graph showing the effect of correcting the bias with different numbers of estimators and samples.](image-url)
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

Correcting the bias in least squares regression

Volume-rescaled sampling

Minimax experimental design

Bayesian experimental design

Conclusions
Volume-rescaled sampling

Derezinski and Warmuth

\( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \) — i.i.d. random vectors sampled from \( \mathbf{x} \sim D_X \)

\( D_X^k \) — distribution of \( \mathbf{X} \)

Volume-rescaled sampling of size \( k \) from \( D_X \):

\[
\text{VS}^k_{D_X}(\mathbf{X}) \propto \det(\mathbf{X}^\top \mathbf{X}) \, D_X^k(\mathbf{X})
\]

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

**Question:** What is the normalization factor of \( \text{VS}^k_{D_X} \) ?

\[
\mathbb{E}_{D_X^k} [\det(\mathbf{X}^\top \mathbf{X})] = ??
\]

Can find it through a new proof of the Cauchy-Binet formula!
Let $\tilde{\mathbf{X}} \sim VS_{D\tilde{\mathbf{X}}}^k$ and $S \subseteq [k]$ be a random size $d$ set such that

$$\Pr(S \mid \tilde{\mathbf{X}}) \propto \det(\tilde{\mathbf{X}}_S)^2.$$  

Then:

- $\tilde{\mathbf{X}}_S \sim VS_{D\tilde{\mathbf{X}}}^d$,
- $\tilde{\mathbf{X}}_{[k]\setminus S} \sim D_{\tilde{\mathbf{X}}}^{k-d}$,
- $S$ is uniformly random,

and the three are independent.
Theorem ([DWH19])

Let \( S = \{ (x_1, y_1), \ldots, (x_k, y_k) \} \) i.i.d. \( D^k \), for any \( k \geq 0 \).

Sample \( \tilde{x}_1, \ldots, \tilde{x}_d \sim VS_{D^X}^d \),

Query \( \tilde{y}_i \sim D_{Y|X=\tilde{x}_i} \forall i=1..d \).

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

\[
E \left[ w^*(S \cup S_\circ) \right] = E_{S \sim D^k} \left[ E_{S_\circ \sim VS_{D^X}^d} \left[ w^*(S \cup S_\circ) \right] \right] \\
(\text{decomposition}) = E_{\tilde{S} \sim VS_{D^X}^{k+d}} \left[ w^*(\tilde{S}) \right] \\
(\text{d-modularity}) = w^*.
\]
Outline

Correcting the bias in least squares regression

Volume-rescaled sampling

Minimax experimental design

Bayesian experimental design

Conclusions
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 \( i = 1 \ldots n \).

**Classical setup:**

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

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

- **(unbiasedness)** \( \mathbb{E}[\mathbf{w}_{LS}] = \mathbf{w}^* \),
- **(mean squared error)** \( \text{MSE}(\mathbf{w}_{LS}) = \mathbb{E}\|\mathbf{w}_{LS} - \mathbf{w}^*\|^2 = \sigma^2 \text{tr}\left( (\mathbf{X}^\top \mathbf{X})^{-1} \right) \)
- letting \( b = \text{tr}\left( (\mathbf{X}^\top \mathbf{X})^{-1} \right) \)
  \[
  = \frac{b}{n} \cdot \mathbb{E}\|\xi\|^2
  \]
- **(mean squared prediction error)** \( \text{MSPE}(\mathbf{w}_{LS}) = \mathbb{E}\|\mathbf{X}(\mathbf{w}_{LS} - \mathbf{w}^*)\|^2 = \sigma^2 d \)
  \[
  = \frac{d}{n} \cdot \mathbb{E}\|\xi\|^2
  \]
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:

\[
\text{MSE or MSPE} \leq \epsilon \cdot \mathbb{E} \|\xi\|^2
\]

Denote \( L^* = \mathbb{E} \|\xi\|^2 = n\sigma^2 \).

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

\[
\begin{align*}
\text{MSE}(\hat{w}_S) - \text{MSE}(w_{LS}) &\leq \epsilon \cdot L^*, \quad \text{for } k \geq d + b/\epsilon, \\
\text{MSPE}(\hat{w}_S) - \text{MSPE}(w_{LS}) &\leq \epsilon \cdot L^*, \quad \text{for } k \geq d + d/\epsilon,
\end{align*}
\]

where \( b = \text{tr}((X^TX)^{-1}) \).
Experimental design in general setting (summary)

No assumptions on $Y_i$.
We define $w^* \overset{\text{def}}{=} \mathbb{E}[w_{LS}] = X^+\mathbb{E}[Y]$.
Define “total noise” as $L^* \overset{\text{def}}{=} \mathbb{E} \|\xi\|^2$, where $\xi \overset{\text{def}}{=} X^T w^* - Y$.

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

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

where $b = \text{tr}((X^T X)^{-1})$.

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

$$\text{MSPE}(\hat{w}_S) - \text{MSPE}(w_{LS}) \leq \epsilon \cdot L^*, \quad \text{for } k = O(d \log n + d/\epsilon).$$
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:

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

**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 $\hat{w}$ with smallest mean squared error:

$$\min_{\hat{w}} \max_{w^*} \mathbb{E}_{\hat{w}} [\| \hat{w} - w^* \|^2] \quad \text{subject to} \quad \mathbb{E}[\hat{w}] = w^* \quad \forall w^*$$

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

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

A-optimal design: $\min_{S: |S| \leq k} \text{tr} \left( (X_S^\top X_S)^{-1} \right)$

Typical required assumption: $y_i = x_i^\top w^* + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$
A-optimal design

Find an unbiased estimator $\hat{w}$ with smallest mean squared error:

$$\min_{\hat{w}} \max_{w^*} \mathbb{E}_\hat{w}[||\hat{w} - w^*||^2] \quad \text{subject to} \quad \mathbb{E}[(\hat{w})] = w^* \quad \forall w^*$$

MSE[$\hat{w}$]

Given set $\{y_i : i \in S\}$, the optimum is least squares: $\hat{w} = X_{S y}^\dagger$

$$\text{MSE}[X_{S y}^\dagger] = \text{tr}(\text{Var}[X_{S y}^\dagger]) = \sigma^2 \text{tr}((X_{S}^\top X_{S})^{-1})$$

A-optimal design: $\min_{S: |S| \leq k} \text{tr}((X_S^\top X_S)^{-1})$

Typical required assumption: $y_i = x_i^\top w^* + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$
A-optimal design: a simple guarantee

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

$$\operatorname{tr}((X_S^T X_S)^{-1}) \leq \frac{n - d + 1}{k - d + 1} \frac{\operatorname{tr}((X^T X)^{-1})}{k - d + 1}$$

(denoted $\phi$)

**Corollary** If $y = Xw^* + \xi$ where $\operatorname{Var}[\xi] = \sigma^2 I$ and $\mathbb{E}[\xi] = 0$ then

$$\frac{\operatorname{tr}((X_S^T y_S)^\dagger)}{\sigma^2 \operatorname{tr}((X_S^T X_S)^{-1})} \leq \sigma^2 \frac{n - d + 1}{k - d + 1} \phi \leq \frac{\phi}{k - d + 1} \cdot \frac{\epsilon}{n \sigma^2} \cdot \operatorname{tr}((\operatorname{Var}[\xi])$$

$$k = d + \phi/\epsilon \quad \text{and} \quad \operatorname{MSE}[X_S^\dagger y_S] \leq \epsilon \cdot \operatorname{tr}((\operatorname{Var}[\xi])$$
A-optimal design: a simple guarantee

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

$$\text{tr}((X_S^\top X_S)^{-1}) \leq \frac{n-d+1}{k-d+1} \text{tr}((X^\top X)^{-1})$$

(denoted $\phi$)

**Corollary** If $y = Xw^* + \xi$ where $\text{Var}[\xi] = \sigma^2 I$ and $E[\xi] = 0$ then

$$\text{tr}(\text{Var}[X_S^\dagger y_S]) \leq \sigma^2 \frac{n-d+1}{k-d+1} \phi \leq \frac{\phi}{k-d+1} \cdot \text{tr}(\text{Var}[\xi])$$

$$k = d + \phi/\epsilon \quad \text{and} \quad \text{MSE}[X_S^\dagger y_S] \leq \epsilon \cdot \text{tr}(\text{Var}[\xi])$$
General response model (What if $\xi_i$ is not $\mathcal{N}(0, \sigma^2)$?)

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

$y \in \mathcal{F}_n$

$w^* \overset{\text{def}}{=} \arg\min_w \mathbb{E}_y[\|Xw - y\|^2] = X^\dagger \mathbb{E}[y]$,

$\xi_{y|X} \overset{\text{def}}{=} y - Xw^* = y - XX^\dagger \mathbb{E}[y] \quad \text{- deviation from best linear predictor}$

Two special cases:

1. Statistical regression: $\mathbb{E}[\xi_{y|X}] = 0$ (mean-zero noise)
2. Worst-case regression: $\text{Var}[\xi_{y|X}] = 0$ (deterministic $y$)
Random experimental designs

**Statistical:** Fixed $S$ is ok

**Worst-case:** Fixed $S$ can be exploited by the adversary

**Definition**

A *random experimental design* $(S, \hat{w})$ 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 $\hat{w} : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^d$

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

$$\text{MSE}[\hat{w}(y_S)] = \mathbb{E}_{S,\hat{w},y}[||\hat{w}(y_S) - w^*||^2]$$

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

$$\mathbb{E}_{S,\hat{w},y}[\hat{w}(y_S)] = \underbrace{X^\dagger \mathbb{E}[y]}_{w^*} \quad \text{for all } y \in \mathcal{F}_n$$
Main result

**Theorem**

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

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

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

$$ \text{MSE}[\hat{w}(y_S)] - \text{MSE}[X^\dagger y] \leq \epsilon \cdot \mathbb{E}[\|\xi_y|X\|^2] $$

**Toy example:**

$$ \text{Var}[\xi_y|X] = \sigma^2 I, \quad \mathbb{E}[\xi_y|X] = 0 $$

1. $\mathbb{E}[\|\xi_y|X\|^2] = \text{tr}((\text{Var}[\xi_y|X]))$
2. $\text{MSE}[X^\dagger y] = \frac{\phi}{n} \cdot \text{tr}(\text{Var}[\xi_y|X])$
Main result

**Theorem**

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

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

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

$$\text{MSE}[^{\hat{w}}(y_S)] - \text{MSE}[X^\dagger y] \leq \epsilon \cdot \mathbb{E}[\|\xi_y|_X\|^2]$$

**Toy example:** \hspace{1cm} $\text{Var}[^{\xi_y}|_X] = \sigma^2 I, \quad \mathbb{E}[^{\xi_y}|_X] = 0$

1. \(\mathbb{E}[\|^{\xi_y}|_X\|^2] = \text{tr}(\text{Var}[^{\xi_y}|_X])\)
2. \(\text{MSE}[X^\dagger y] = \frac{\phi}{n} \cdot \text{tr}(\text{Var}[^{\xi_y}|_X])\)
Important special instances

1. **Statistical regression:** \( y = Xw^* + \xi, \quad \mathbb{E}[\xi] = 0 \)

\[
\text{MSE}[\hat{w}(y_S)] - \text{MSE}[X^\dagger y] \leq \epsilon \cdot \text{tr}(\text{Var}[\xi])
\]

- **Weighted regression:** \( \text{Var}[\xi] = \text{diag}([\sigma^2_1, \ldots, \sigma^2_n]) \)

- **Generalized regression:** \( \text{Var}[\xi] \) is arbitrary

- **Bayesian regression:** \( w^* \sim \mathcal{N}(0, I) \)

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

\[
\mathbb{E}_{S,\hat{w}}[\|\hat{w}(y_S) - w^*\|^2] \leq \epsilon \cdot \|y - Xw^*\|^2
\]

where \( w^* = X^\dagger y \)
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: $\Pr(i) \overset{\text{def}}{=} \frac{1}{d} x_i^\top (X^\top X)^{-1} x_i$
   2.2 Inverse score sampling: $\Pr(i) \overset{\text{def}}{=} \frac{1}{\phi} x_i^\top (X^\top X)^{-2} x_i$  (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 \( \text{VS}(\mathbf{X}) \) 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:**

\( O(\text{nnz}(\mathbf{X}) \log n + d^4 \log d) \)
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.

$$
\mathbb{E} \left[ \argmin_w \sum_{t=1}^k \frac{1}{q_{i_t}} (x_{i_t}^\top w - y_{i_t})^2 \right] = w_y^* \mid x
$$
Simple volume-rescaled sampling:

- Let $D_X$ be a uniformly random $x_i$.
- $(X_S, y_S) \sim V_{S_D}^k$ and $\hat{w} = X_S^\dagger y_S$.

Then, $E[\hat{w}] = w_y^* x$.

**Problem:** Not robust to worst-case noise

**Solution:** Volume-rescaled importance sampling

- Let $p = (p_1, \ldots, p_n)$ be an importance sampling distribution,
- Define $\tilde{x} \sim D_X$ as $\tilde{x} = \frac{1}{\sqrt{p_i}} x_i$ for $i \sim p$.

Then, for $(\tilde{X}_S, \tilde{y}_S) \sim V_{S_D}^k$ and $\tilde{w} = \tilde{X}_S^\dagger \tilde{y}_S$, we have $E[\tilde{w}] = w_y^* x$. 
Importance sampling for experimental design

1. **Leverage score sampling**: \( \Pr(i) = p_i^{\text{lev}} \overset{\text{def}}{=} \frac{1}{d} x_i ^\top (X ^\top X)^{-1} x_i \)

   A standard sampling method for worst-case linear regression.

2. **Inverse score sampling**: \( \Pr(i) = p_i^{\text{inv}} \overset{\text{def}}{=} \frac{1}{\phi} x_i ^\top (X ^\top X)^{-2} x_i \).

   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:

\[
R_k^*(X) \overset{\text{def}}{=} \min_{(S, \hat{w}) \in \mathcal{W}_k(X)} \max_{y \in \mathcal{F}_n \setminus \text{Sp}(X)} \frac{\text{MSE}[\hat{w}(y_S)] - \text{MSE}[X^\dagger y]}{\mathbb{E}[\|\xi_y x\|^2]}
\]

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

- if \(\mathbb{E}_{y, \hat{w}}[\hat{w}(y)] = X^\dagger \mathbb{E}[y] \quad \forall y \in \mathcal{F}_n\)
- then \(\text{Var}[\hat{w}(y)] \succeq \text{Var}[X^\dagger y] \quad \forall y \in \mathcal{F}_n\)

- If \(d \leq k \leq n\), then \(R_k^*(X) \in [0, \infty)\)
- If \(k \geq C \cdot d \log n\), then \(R_k^*(X) \leq C \cdot \phi/k\) for some \(C\)
- If \(k^2 < \epsilon nd/3\), then \(R_k^*(X) \geq (1-\epsilon) \cdot \phi/k\) for some \(X\)
Alternative: mean squared prediction error

**Definition.** \( \text{MSPE}[\hat{w}] = \mathbb{E}[\|X(\hat{w} - w^*)\|^2] \) (V-optimality)

**Theorem**

*There is \((S, \hat{w})\) of size \(k = O(d \log n + d/\epsilon)\) s.t. for any \(y \in F_n\),*

\[
\text{MSPE}[\hat{w}(y_S)] - \text{MSPE}[X^\dagger y] \leq \epsilon \cdot \mathbb{E}[\|\xi_y|X\|^2]
\]

Follows from the MSE bound by reduction to \(X^\top X = I\).

Then \( \text{MSPE}[\hat{w}] = \text{MSE}[\hat{w}] \) and \( \phi = d \).

**Minimax V-optimal value:**

\[
\min_{(S, \hat{w}) \in W_k(X)} \max_{y \in F_n \setminus \text{Sp}(X)} \frac{\text{MSPE}[\hat{w}(y_S)] - \text{MSPE}[X^\dagger y]}{\mathbb{E}[\|\xi_y|X\|^2]}
\]
Questions about minimax experimental design

1. Can $R^*_k(X)$ be found, exactly or approximately?

2. What happens in the regime of $k \leq C \cdot d \log n$?

3. Can we restrict $W_k(X)$ 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 $y \in \mathcal{F}_n$ with fixed $y \in \mathbb{R}^n$:

$$R^*_k(X) = \min_{(S, \hat{w}) \in \mathcal{W}_k(X)} \max_{y \in \mathbb{R}^n \setminus \text{Sp}(X)} \frac{\mathbb{E}_{S, \hat{w}} \left[ \| \hat{w}(y_S) - X^\dagger y \|_2^2 \right]}{\| y - XX^\dagger y \|_2^2}$$

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

$$\mathbb{E}[\hat{w}(y_S)] = X^\dagger y \quad \text{and} \quad \mathbb{E}[\| \hat{w}(y_S) - X^\dagger y \|_2^2] \leq \epsilon \cdot \| y - XX^\dagger y \|_2^2.$$  

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

$$\underbrace{\mathbb{E}[\| \hat{w}(y_S) - w^* \|_2^2]}_{\text{MSE}[\hat{w}(y_S)]} \leq \underbrace{\mathbb{E}[\| X^\dagger y - w^* \|_2^2]}_{\text{MSE}[X^\dagger y]} + \epsilon \cdot \mathbb{E}[\| y - Xw^* \|_2^2].$$
Correcting the bias in least squares regression

Volume-rescaled sampling

Minimax experimental design

Bayesian experimental design

Conclusions
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:

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

**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

$$w | y_S \sim \mathcal{N}\left( (X_S^TX_S + A)^{-1}X_S^Ty_S, \sigma^2(X_S^TX_S + A)^{-1} \right),$$

Bayesian A-optimality criterion:

$$f_A(X_S^TX_S) = \text{tr}((X_S^TX_S + A)^{-1}).$$

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

$$f_A(X_S^TX_S) \leq (1 + \epsilon) \cdot \min_{S': |S'|=k} \underbrace{f_A(X_{S'}^TX_{S'})}_{\text{OPT}_k}$$
SDP relaxation

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

\[ p^* = \arg\min_{p_1, \ldots, p_n} \ f_A \left( \sum_{i=1}^n p_i x_i x_i^\top \right), \]

subject to \( \forall i \ 0 \leq p_i \leq 1, \sum_i p_i = k. \)

The solution \( p^* \) satisfies \( f_A \left( \sum_i p_i x_i x_i^\top \right) \leq \text{OPT}_k. \)

**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 $d_A = \text{tr}(X^\top X(X^\top X + A)^{-1}) \leq d$.

**Theorem ([DLM19])**

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

$$f_A(X^\top_S X_S) \leq (1 + \epsilon) \cdot \text{OPT}_k.$$

**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

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Bayesian</th>
<th>$k = \Omega(\cdot)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[WYS17]</td>
<td>A,V</td>
<td>X</td>
</tr>
<tr>
<td>[AZLSW17]</td>
<td>A,C,D,E,G,V</td>
<td>✓</td>
</tr>
<tr>
<td>[NSTT19]</td>
<td>A,D</td>
<td>X</td>
</tr>
<tr>
<td><strong>our result</strong> [DLM19]</td>
<td>A,C,D,V</td>
<td>✓</td>
</tr>
</tbody>
</table>
Correcting the bias in least squares regression

Volume-rescaled sampling

Minimax experimental design

Bayesian experimental design

Conclusions
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 between statistical and worst-case perspectives
- Applications in Bayesian experimental design: bridging the gap between 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.

Near-optimal design of experiments via regret minimization.

Michał Dereźniński, 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.

Michał Dereźniński, Feynman Liang, and Michael W. Mahoney.
Distributed estimation of the inverse Hessian by determinantal averaging.

Michał Dereźniński and Manfred K. Warmuth.
Reverse iterative volume sampling for linear regression.

Michał Dereźniński, Manfred K. Warmuth, and Daniel Hsu.
Correcting the bias in least squares regression with volume-rescaled sampling.

Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat.
Proportional volume sampling and approximation algorithms for a -optimal design.