# Minimax experimental design

Bridging the gap between statistical and worst-case approaches to least squares regression


Michał Dereziński — Berkeley UNIVERSITY OF CALIFORNIA

Kenneth L. Clarkson — IBM

Michael W. Mahoney — Berkeley UNIVERSITY OF CALIFORNIA

Manfred K. Warmuth — UNIVERSITY OF CALIFORNIA SANTA CRUZ, Google

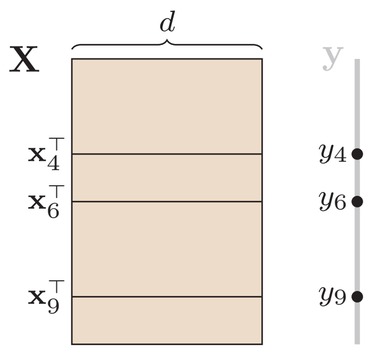

## Classical experimental design

Consider $n$ vectors: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$.
Each $\mathbf{x}_i$ has a hidden random response $y_i$:

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

**Task:** Select $k \ll n$ responses to query

Select $S = \{4, 6, 9\}$
Receive $y_4, y_6, y_9$




**Goal:** Find the *best unbiased estimator* of $\mathbf{w}^*$

## A-optimal design

Minimize the *mean squared error*:

$$\min_{\widehat{\mathbf{w}}} \quad \underbrace{\mathbb{E}_{\widehat{\mathbf{w}}}\left[\|\widehat{\mathbf{w}} - \mathbf{w}^*\|^2\right]}_{\mathrm{MSE}[\widehat{\mathbf{w}}]} \quad \text{s.t.} \quad \mathbb{E}\left[\widehat{\mathbf{w}}\right] = \mathbf{w}^*$$

**Classical result:** given $\{y_i : i \in S\}$,
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: $\min_{S:\,|S|\le k} \mathrm{tr}\left((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\right)$

---

**Alternative:** *Mean squared prediction error*

$$\mathrm{MSPE}\left[\widehat{\mathbf{w}}\right] = \mathbb{E}_{\widehat{\mathbf{w}}}\left[\|\mathbf{X}(\widehat{\mathbf{w}} - \mathbf{w}^*)\|^2\right]$$

V-optimal design: $\min_{S:\,|S|\le k} \underbrace{\mathrm{tr}\left(\mathbf{X}(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{X}^\top\right)}_{\propto \mathrm{MSPE}[\mathbf{X}_S^\dagger \mathbf{y}_S]}$

All of our results extend to V-optimal design

## Prior: A simple guarantee

**Theorem** (from [**?**]) For any $\mathbf{X}$ and $k \ge d$ there is $S$ of size $k$ s.t.:

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

Let $\boldsymbol{\xi}^\top = [\xi_1, \ldots, \xi_n]$ be the noise vector

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

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

For $k = d + \phi/\epsilon$ there is $S$ of size $k$ s.t.
$\mathrm{MSE}[\mathbf{X}_S^\dagger \mathbf{y}_S] \le \epsilon \cdot \mathrm{tr}(\mathrm{Var}[\boldsymbol{\xi}])$

**Question:** What if we drop the assumptions on the noise vector $\boldsymbol{\xi}$?

## Arbitrary response model

$\mathbf{y} \in \mathcal{F}_n$ - random vector in $\mathbb{R}^n$ s.t. $\mathbb{E}[\|\mathbf{y}\|^2] < \infty$

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

Two special cases:

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

## Random experimental designs

**Definition 1** *We define a random experimental design $(S, \widehat{\mathbf{w}})$ of size $k$ as:*

1. *a random $S \subseteq \{1..n\}$ s.t. $|S| \le k$*
2. *a jointly random function $\widehat{\mathbf{w}} : \mathbb{R}^{|S|} \to \mathbb{R}^d$*

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

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

## Main result: experimental design with arbitrary responses

**Theorem 1** *There is a random experimental design of size*

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

*such that for any random response vector $\mathbf{y} \in \mathcal{F}_n$ we have*

$$\textit{(unbiasedness)} \quad \mathbb{E}\left[\widehat{\mathbf{w}}(\mathbf{y}_S)\right] = \mathbf{w}^*_{\mathbf{y}|\mathbf{X}},$$

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

**Important examples**

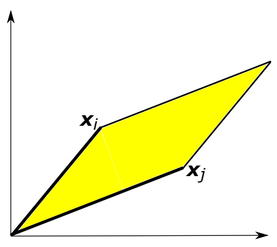
1. *Statistical regression:*
   $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}, \quad \mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$
   - Weighted regression: $\mathrm{Var}[\boldsymbol{\xi}] = \mathrm{diag}\left([\sigma_1^2, \ldots, \sigma_n^2]\right)$
   - Generalized regression: $\mathrm{Var}[\boldsymbol{\xi}]$ is arbitrary
2. *Worst-case regression:*
   $\mathbf{y}$ is any fixed vector in $\mathbb{R}^n$

## Volume sampling

**Definition 2** *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: $O(nd^2)$

**Theorem** (from [**?**]) Volume sampling corrects the bias of any i.i.d. sampling $q = (q_1, \ldots, q_n)$:

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

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

## Leverage and inverse scores

$\mathbf{z}_i$ – $i$th column of pseudoinverse $\underbrace{(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top}_{\mathbf{X}^\dagger}$

1. *Leverage score sampling*:
   $\Pr(i) = p_i^{\mathrm{lev}} \overset{def}{=} \frac{1}{d}\mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x}_i = \frac{1}{d}\mathbf{x}_i^\top \mathbf{z}_i$
   used to obtain a subspace embedding
2. *Inverse score sampling* (new):
   $\Pr(i) = p_i^{\mathrm{inv}} \overset{def}{=} \frac{1}{\phi}\mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-2}\mathbf{x}_i = \frac{1}{\phi}\|\mathbf{z}_i\|^2$
   used to get the $\phi/\epsilon$ rate for the MSE

Controlling the least squares estimator:

$$\widehat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \sum_{t=1}^{k} \frac{1}{q_{i_t}} (\mathbf{x}_{i_t}^\top \mathbf{w} - y_{i_t})^2$$
$$= \underbrace{\left(\sum_{t=1}^{k} \frac{1}{q_{i_t}} \mathbf{z}_{i_t} \mathbf{x}_{i_t}^\top\right)^{-1}}_{\text{needs leverage scores}} \underbrace{\sum_{t=1}^{k} \frac{1}{q_{i_t}} \mathbf{z}_{i_t} y_{i_t}}_{\text{needs inverse scores}}$$

**Strategy**: Use a mixture of both $p^{\mathrm{lev}}$ and $p^{\mathrm{inv}}$

## Minimax A-optimal value for experimental design

**Definition 3** *Define the minimax A-optimal value for experimental design with unbiased estimators as:*

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

**Proposition 2** *Least squares is the minimum variance unbiased estimator for $\mathcal{F}_n$:*

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

**Theorem 3** *The following statements are true about the minimax A-optimal value:*

1. *If $d \le k \le n$, then $R_k^*(\mathbf{X}) \in [0, \infty)$ for a full rank $\mathbf{X}$.*
2. *If $k \ge C \cdot d \log n$, then $R_k^*(\mathbf{X}) \le C \cdot \phi/k$ for some $C$.* *(bound from Theorem **??**)*
3. *If $k^2 < \epsilon n d/3$, then $R_k^*(\mathbf{X}) \ge (1-\epsilon) \cdot \phi/k$ for some $\mathbf{X}$.* *(matching lower bound)*

**References**


[1] H. Avron, C. Boutsidis. *Faster subset selection for matrices and applications.* JMAA 2013.
[2] M. Dereziński, M. Warmuth, D. Hsu. *Correcting the bias in least squares regression with volume-rescaled sampling.* AISTATS'19.