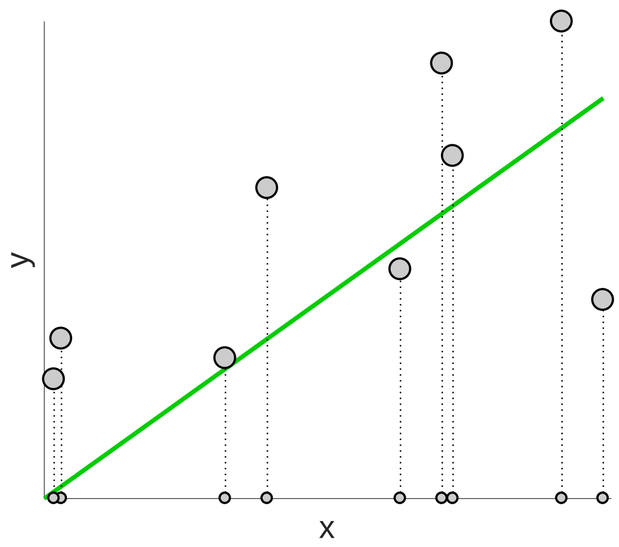
Correcting the bias in least squares regression with volume-rescaled sampling

Michał Dereziński — Berkeley UNIVERSITY OF CALIFORNIA

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

Daniel Hsu — COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

Random design regression



Dataset: $\mathbb{S} = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$

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

Estimator: $\mathbf{w}^*(\mathbb{S}) = \operatorname{argmin}_{\mathbf{w}} \sum_{(\mathbf{x}_i, y_i) \in S} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$

Bias of the least squares

$\mathbf{x}_i \sim \mathrm{D}_{\mathcal{X}}$ - random input vector in $\mathbb{R}^d$
$y_i \sim \mathrm{D}_{y|\mathbf{x}=\mathbf{x}_i}$ - random response variable in $\mathbb{R}$

$$\mathbf{w}^*(\mathbb{S}) = \underbrace{\Big(\sum_i \mathbf{x}_i \mathbf{x}_i^\top\Big)^{-1}}_{\text{inverse covariance}} \sum_i \mathbf{x}_i y_i$$

Random inverse covariance introduces bias:

$$\mathbb{E}\big[\mathbf{w}^*(\mathbb{S})\big] \neq \mathbf{w}^*_{\mathrm{D}}$$

Correcting the bias

Our approach: Augment dataset with d points drawn with *volume-rescaled sampling*

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

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

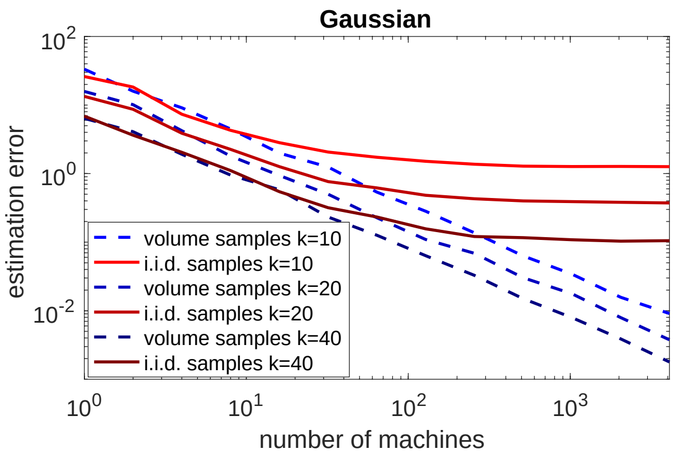
Add $\mathbb{S}_\circ = (\mathbf{x}_{n+1}, y_{n+1}), \ldots, (\mathbf{x}_{n+d}, y_{n+d})$

Theorem $\mathbb{E}\big[\mathbf{w}^*\big(\langle \mathbb{S}, \mathbb{S}_\circ \rangle\big)\big] = \mathbf{w}^*$

Example: Averaging estimators

$\mathbf{x}^\top = (x_1, \ldots, x_d) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\quad y = \xi(\mathbf{x}) + \epsilon$, $\xi(\cdot)$ is a non-linear function and $\epsilon \sim \mathcal{N}(0,1)$.

estimation error: $\left\| \frac{1}{T} \sum_{t=1}^T \mathbf{w}^*(\mathbb{S}_t) \;-\; \mathbf{w}^*_{\mathrm{D}} \right\|^2$



Plot uses $d = 5$ and $\xi(\mathbf{x}) = \sum_{i=1}^d x_i + \frac{x_i^3}{3}$

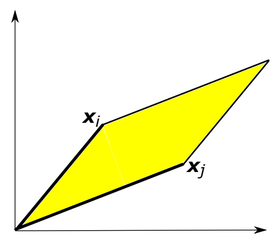
Volume-rescaled sampling

Definition 1 *Given* $\mathrm{D}_{\mathcal{X}}$ *and any* $k \geq d$,

$$\mathrm{VS}^k_{\mathrm{D}_{\mathcal{X}}}(\mathbf{x}_1, \ldots, \mathbf{x}_k) \;\propto\; \det\Big(\sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i^\top\Big) \cdot \prod_{i=1}^k \mathrm{D}_{\mathcal{X}}(\mathbf{x}_i)$$

When $k = d$, then

$$\det\Big(\sum_{i=1}^d \mathbf{x}_i \mathbf{x}_i^\top\Big) = \det\begin{pmatrix} -\mathbf{x}_1^\top- \\ \cdots \\ -\mathbf{x}_d^\top- \end{pmatrix}^2 = \text{volume}^2 \ldots$$



… of the parallelepiped spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_d$

Augmentation property

Theorem 1 *Let* $\mathbb{S}_\circ \sim \mathrm{VS}^d_{\mathrm{D}_{\mathcal{X}}}$, $\mathbb{S} \sim \mathrm{D}^{k-d}_{\mathcal{X}}$ *and* σ *be a uniformly random permutation of* $\{1..n\}$. *Then*

$$\sigma\big(\langle \mathbb{S}_\circ, \mathbb{S} \rangle\big) \sim \mathrm{VS}^k_{\mathrm{D}_{\mathcal{X}}}$$

volume + i.i.d. $\Rightarrow$ $\mathrm{VS}^k_{\mathrm{D}_{\mathcal{X}}}$

$\overbrace{\mathbf{x}_1, \ldots, \mathbf{x}_d}^{\mathrm{VS}^d_{\mathrm{D}_{\mathcal{X}}}}, \overbrace{\mathbf{x}_{d+1}, \ldots, \mathbf{x}_k}^{\mathrm{D}^{k-d}_{\mathcal{X}}}$

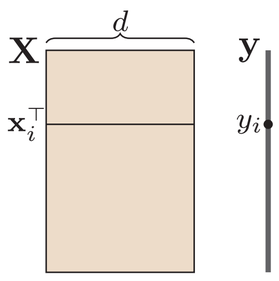
$\mathbf{x}_{\sigma_1}, \ldots, \mathbf{x}_{\sigma_k}$

Unbiasedness via augmentation

$(\mathbf{X}, \mathbf{y}) \sim \mathrm{VS}^k_{\mathrm{D}}$:

$\mathbf{x}_1, \ldots, \mathbf{x}_k \sim \mathrm{VS}^k_{\mathrm{D}_{\mathcal{X}}}$,

$y_i \sim \mathrm{D}_{y|\mathbf{x}=\mathbf{x}_i} \quad \forall_i$.



$\mathbb{S} = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k) \overset{\text{i.i.d.}}{\sim} \mathrm{D}$
$\mathbb{S}_\circ = (\widetilde{\mathbf{x}}_1, \widetilde{y}_1), \ldots, (\widetilde{\mathbf{x}}_d, \widetilde{y}_d) \sim \mathrm{VS}^d_{\mathrm{D}}$

Theorem 2 *For any distribution* D,

$$\mathbb{E}_{\mathrm{VS}^k_{\mathrm{D}}}\big[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}\big] = \Big(\mathbb{E}_{\mathrm{D}_{\mathcal{X}}}\big[\mathbf{x}\mathbf{x}^\top\big]\Big)^{-1} \mathbb{E}_{\mathrm{D}}\big[\mathbf{x}\, y\big]$$

$$\begin{aligned} \mathbb{E}\big[\mathbf{w}^*\big(\langle \mathbb{S}_\circ, \mathbb{S} \rangle\big)\big] &= \mathbb{E}_{\mathbb{S} \sim D^k}\Big[\mathbb{E}_{\mathbb{S}_\circ \sim \mathrm{VS}^d_{\mathrm{D}}}\big[\mathbf{w}^*\big(\langle \mathbb{S}_\circ, \mathbb{S} \rangle\big)\big]\Big] \\ \text{(Theorem 1)} \quad &= \mathbb{E}_{\widetilde{\mathbb{S}} \sim \mathrm{VS}^{k+d}_{\mathrm{D}}}\big[\,\mathbf{w}^*(\,\widetilde{\mathbb{S}}\,)\,\big] \\ \text{(Theorem 2)} \quad &= \mathbf{w}^*_{\mathrm{D}} \end{aligned}$$

Sampling complexity

Conditioning number of $\mathrm{D}_{\mathcal{X}}$:

$$K_{\mathrm{D}_{\mathcal{X}}} = \sup_{\widetilde{\mathbf{x}} \in \mathrm{supp}(\mathrm{D}_{\mathcal{X}})} \widetilde{\mathbf{x}}^\top \Big(\mathbb{E}_{\mathrm{D}_{\mathcal{X}}}\big[\mathbf{x}\mathbf{x}^\top\big]\Big)^{-1} \widetilde{\mathbf{x}}.$$

Theorem 3 *The complexity of* $\mathrm{VS}^d_{\mathrm{D}_{\mathcal{X}}}$ *is:*

1. $O(K_{\mathrm{D}_{\mathcal{X}}} d \log d)$ *i.i.d. samples from* $\mathrm{D}_{\mathcal{X}}$,
2. $O(K_{\mathrm{D}_{\mathcal{X}}} d^3 \log d)$ *arithmetic operations*

Gaussian algorithm

Theorem 4 *Suppose that* $\mathrm{D}_{\mathcal{X}} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_{2d+2} \overset{\text{i.i.d.}}{\sim} \mathrm{D}_{\mathcal{X}}$ *and*

$$\widetilde{\mathbf{x}}_i \overset{\text{def}}{=} \Big(\sum_{j=d+1}^{2d+2} \mathbf{x}_j \mathbf{x}_j^\top\Big)^{\frac{1}{2}} \Big(\sum_{j=1}^{d} \mathbf{x}_j \mathbf{x}_j^\top\Big)^{-\frac{1}{2}} \mathbf{x}_i.$$

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

General volume-rescaled sampling algorithm

Construct an approximate covariance matrix $\widehat{\mathbf{\Sigma}}$:

$$(1-\epsilon) \cdot \mathbb{E}_{\mathrm{D}_{\mathcal{X}}}\big[\mathbf{x}\mathbf{x}^\top\big] \preceq \widehat{\mathbf{\Sigma}} \preceq (1+\epsilon) \cdot \mathbb{E}_{\mathrm{D}_{\mathcal{X}}}\big[\mathbf{x}\mathbf{x}^\top\big], \quad \text{where} \quad \epsilon = \frac{1}{\sqrt{2d}}$$

Leverage score sampling w.r.t. matrix $\widehat{\mathbf{\Sigma}}$:

$$\mathrm{Lev}_{\widehat{\mathbf{\Sigma}}, \mathcal{X}}(\mathbf{x}) \;\propto\; l_{\widehat{\mathbf{\Sigma}}}(\mathbf{x}) \cdot \mathrm{D}_{\mathcal{X}}(\mathbf{x}), \quad \text{where} \quad l_{\widehat{\mathbf{\Sigma}}}(\mathbf{x}) = \mathbf{x}^\top \widehat{\mathbf{\Sigma}}^{-1} \mathbf{x}$$

Key idea: Distortion-free intermediate sampling [DWH18, Der19]

- Sample a larger intermediate sample — **here:** d^2 i.i.d. leverage score samples
- Downsample to the target size — **here:** d samples

Does not distort the target distribution even though $\widehat{\mathbf{\Sigma}} \neq \mathbb{E}_{\mathrm{D}_{\mathcal{X}}}\big[\mathbf{x}\mathbf{x}^\top\big]$

Input: $\widehat{\mathbf{\Sigma}}$, $\mathrm{Lev}_{\widehat{\mathbf{\Sigma}}, \mathcal{X}}$
repeat
  $\mathbf{x}_1, \ldots, \mathbf{x}_{d^2} \overset{\text{i.i.d.}}{\sim} \mathrm{Lev}_{\widehat{\mathbf{\Sigma}}, \mathcal{X}}$
  $\widetilde{\mathbf{X}} \leftarrow \Big[\frac{1}{\sqrt{d\, l_{\widehat{\mathbf{\Sigma}}}(\mathbf{x}_i)}} \cdot \mathbf{x}_i^\top\Big]_{d^2 \times d}$
  Sample $Acc \sim \mathrm{Bernoulli}\Big(\frac{\det(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})}{\det(\widehat{\mathbf{\Sigma}})}\Big)$
until $Acc = \text{true}$
$S \leftarrow \mathrm{VolSamp}(\widetilde{\mathbf{X}})$
return $\mathbf{X}_S$

$\mathrm{VolSamp}\big(\mathbf{X} \in \mathbb{R}^{n \times d}\big)$: [DW18]
  $S \leftarrow \{1..n\}$
  while $|S| > d$
    $\forall_{i \in S} \quad q_i \leftarrow \frac{\det(\mathbf{X}_{S \setminus i}^\top \mathbf{X}_{S \setminus i})}{(|S| - d)\det(\mathbf{X}_S^\top \mathbf{X}_S)}$
    Sample $i \sim (q_i)_{i \in S}$
    $S \leftarrow S \setminus \{i\}$
  end
return S

DW18 Dereziński, Warmuth. *Reverse iterative volume sampling for linear regression*. JMLR, 2018

DWH18 Dereziński, Warmuth, Hsu. *Leveraged volume sampling for linear regression*. NeurIPS 2018

Der18 Dereziński. *Fast determinantal point processes via distortion-free intermediate sampling*. 2018