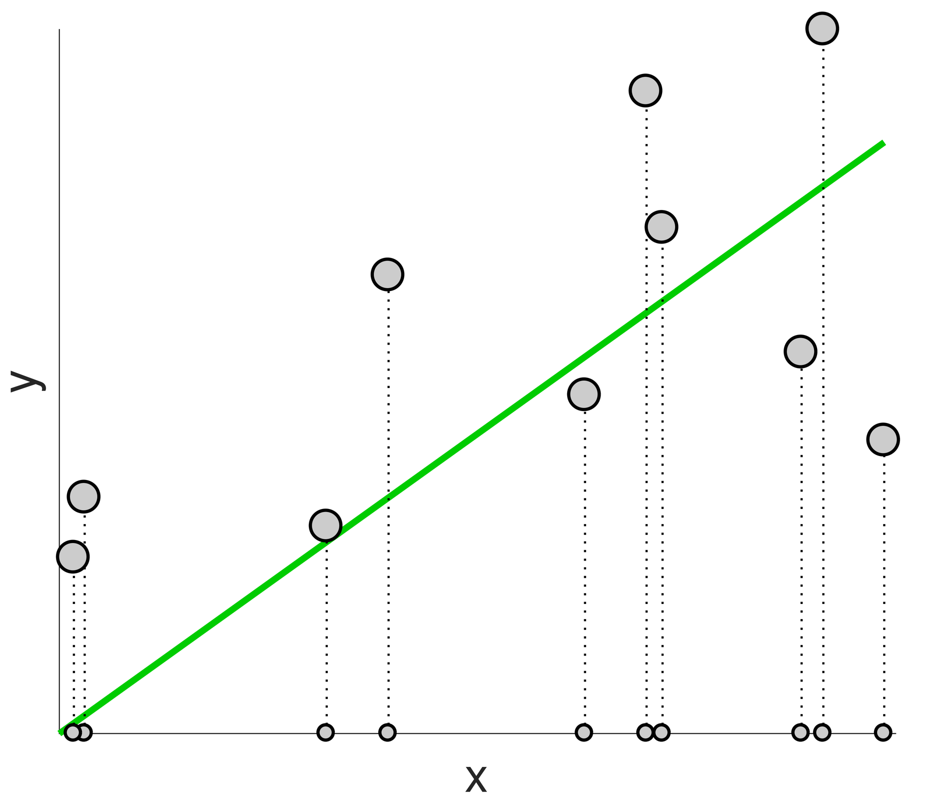
Unbiased estimates for linear regression via volume sampling

Michał Dereziński   Manfred K. Warmuth

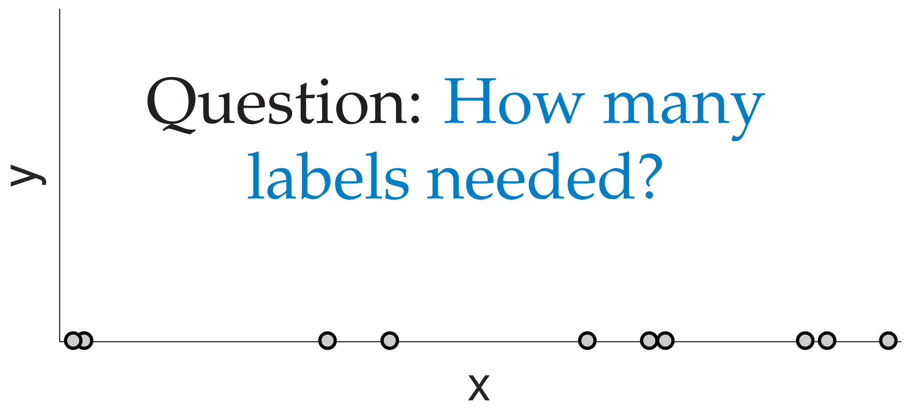
Linear regression



$$L(w) = \sum_i (x_i w - y_i)^2$$

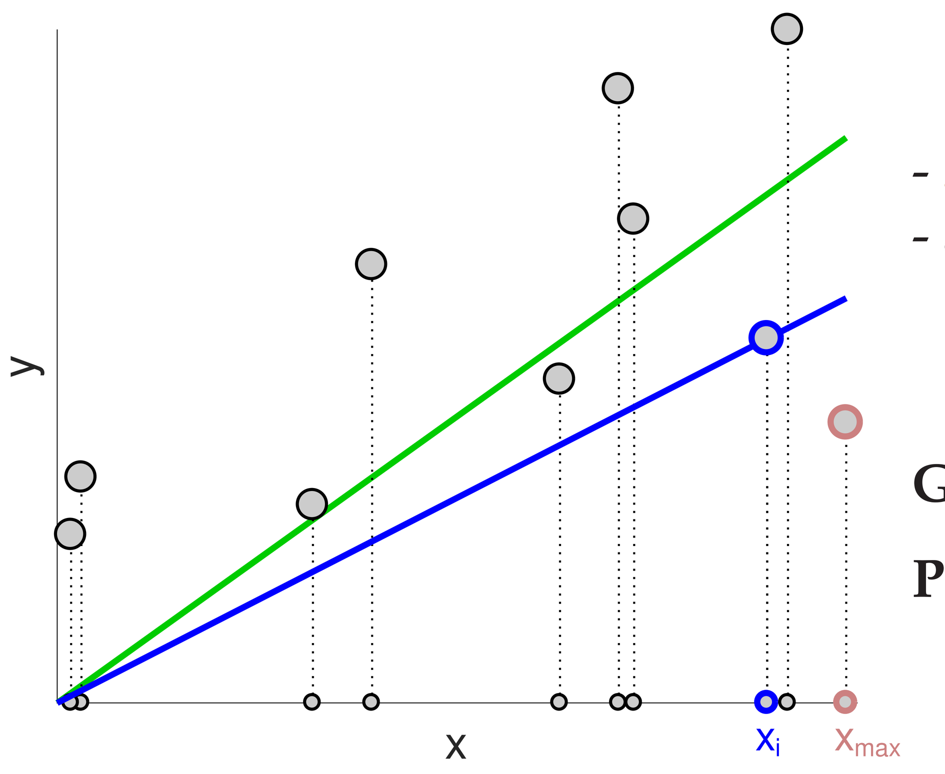
$$w^* = \underset{w}{\operatorname{argmin}}\, L(w)$$

Expensive labels



- All x_i given
- Labels y_i unknown
- Learner can ask for a subset of labels

Answer: one label



- $x_{\max}$ (furthest from 0) is bad
- any deterministic choice bad

Good: 1 label y_i drawn $\sim x_i^2$

Predict: $w_i^* = \frac{y_i}{x_i}$

Loss formula: $\mathbb{E}_i[L(w_i^*)] = 2\,L(w^*)$

Unbiasedness: $\mathbb{E}_i[w_i^*] = \sum_i \underbrace{\frac{x_i^2}{\|\mathbf{x}\|^2}}_{P(i)} \underbrace{\frac{y_i}{x_i}}_{w_i^*} = w^*$

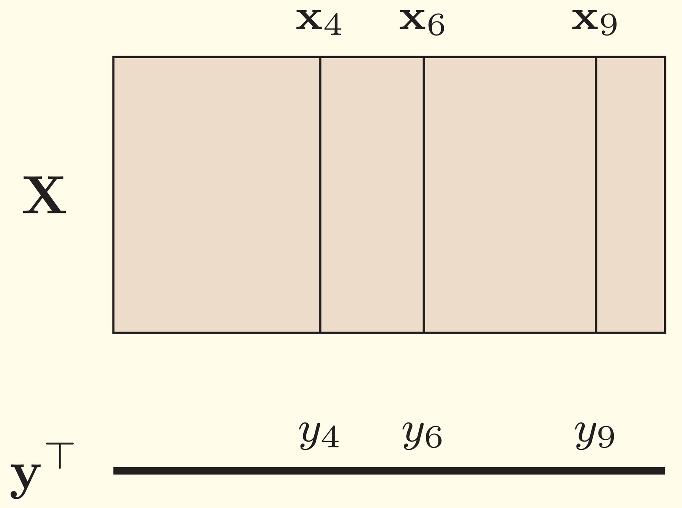
Subsampling for linear regression

Given: n points $\mathbf{x}_i \in \mathbb{R}^d$ with hidden labels $y_i \in \mathbb{R}$

Goal: Minimize loss $L(\mathbf{w}) = \sum_i (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$ over all n points

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

Receive y_4, y_6, y_9

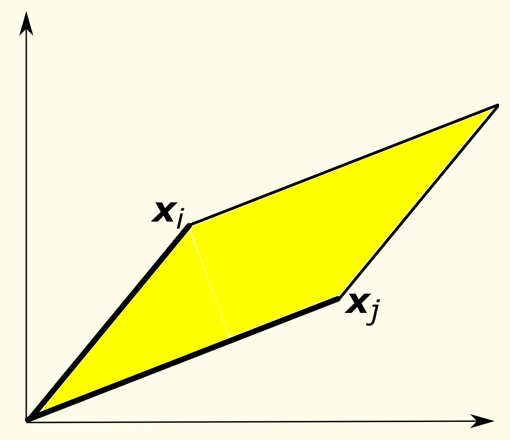


Simple strategy: Solve the subproblem, $\mathbf{w}^*(S) = \mathbf{X}_S^{+\top} \mathbf{y}_S$

Volume sampling

$S \subseteq \{1..n\}$ chosen w.p.

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



Distribution over all d-element subsets S:

$$P(S) = \det(\mathbf{X}_S \mathbf{X}_S^\top) \,/\, Z$$

Normalization factor obtained via Cauchy-Binet formula:

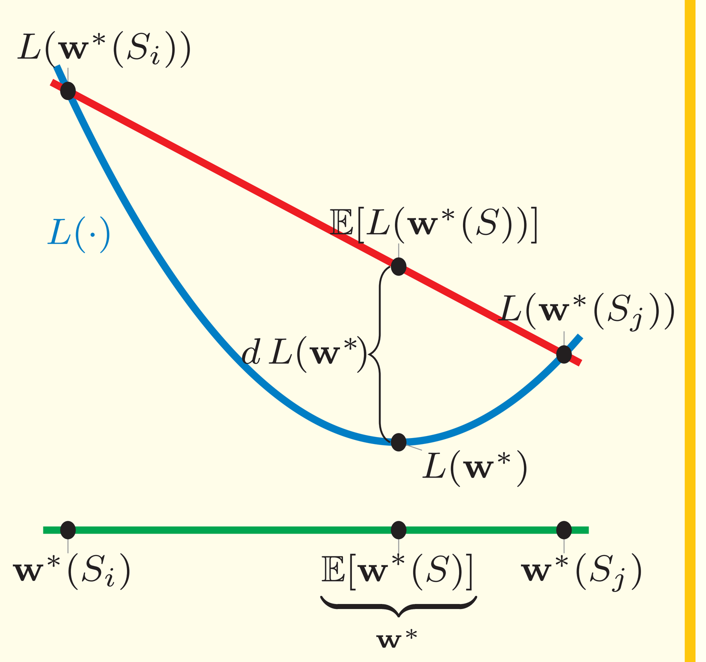
$$Z = \sum_{S:|S|=d} \det(\mathbf{X}_S \mathbf{X}_S^\top) = \det(\mathbf{X}\mathbf{X}^\top)$$

Loss expectation formula

Theorem
For a volume-sampled set S of size d,

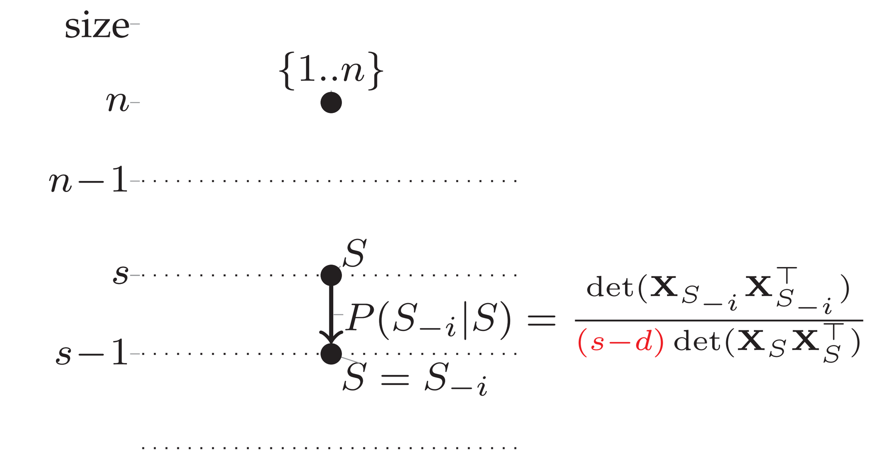
$$\mathbb{E}\,[L(\mathbf{w}^*(S))] = (d+1)\, L(\underbrace{\mathbf{w}^*}_{\mathbb{E}[\mathbf{w}^*(S)]}),$$

if $\mathbf{X}$ is in general position



- distribution does not depend on labels
- no range restrictions!

Reverse iterative volume sampling



Start with $S = \{1..n\}$

Sample index $i \in S$

Go to set $S_{-i} = S - \{i\}$

Repeat until desired size

Runtime: $O(n^2 d)$

$$P\left(\begin{matrix}\{1..n\}\\ \downarrow\\ \downarrow\\ S_\circ\end{matrix}\right) = \overbrace{\prod_{\text{edges of path}} P(S_{-i}|S)}^{\text{telescoping product}} = \frac{1}{(n-d)\dots(s-d+1)} \frac{\det(\mathbf{X}_{S_\circ}\mathbf{X}_{S_\circ}^\top)}{\det(\mathbf{X}\mathbf{X}^\top)}$$

$$P(S_\circ) = \frac{\overbrace{(n-s)!}^{\#\text{ of paths}}}{(n-d)\dots(s-d+1)} \frac{\det(\mathbf{X}_{S_\circ}\mathbf{X}_{S_\circ}^\top)}{\det(\mathbf{X}\mathbf{X}^\top)} = \frac{1}{\binom{n-d}{s-d}} \frac{\det(\mathbf{X}_{S_\circ}\mathbf{X}_{S_\circ}^\top)}{\det(\mathbf{X}\mathbf{X}^\top)}$$

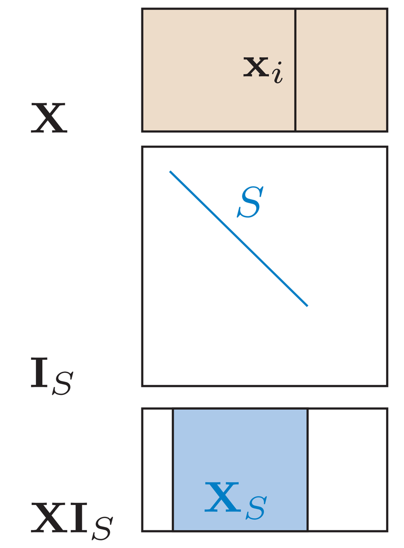
Unbiased estimator for pseudo-inverse $\mathbf{X}^+$

Key trick: To each subset S assign a formula $\mathbf{F}(S)$ st
$\mathbf{F}(S) = \sum_{i\in S} P(S_{-i}|S)\mathbf{F}(S_{-i})$.

Then: $\mathbb{E}_S[\mathbf{F}(S)] = \mathbf{F}(\{1..n\})$

Expectation formulas for $(\mathbf{X}\mathbf{I}_S)^+$

1. $\mathbb{E}[(\mathbf{X}\mathbf{I}_S)^+] = \mathbf{X}^+$ (unbiasedness)
2. $\mathbb{E}[\underbrace{(\mathbf{X}_S\mathbf{X}_S^\top)^{-1}}_{((\mathbf{X}\mathbf{I}_S)^+)^2}] = \frac{n-d+1}{s-d+1} \underbrace{(\mathbf{X}\mathbf{X}^\top)^{-1}}_{(\mathbf{X}^+)^2}$ (variance)



Corollary: $\mathbb{E}[\mathbf{w}^*(S)] = \mathbb{E}[(\mathbf{X}\mathbf{I}_S)^{+\top}\mathbf{y}] = \mathbf{X}^{+\top}\mathbf{y} = \mathbf{w}^*$

Averaging unbiased estimators

Let $\hat{\mathbf{y}}(S) = \mathbf{X}^\top \mathbf{w}^*(S)$. If $\mathbf{w}^*(S)$ is unbiased ($\mathbb{E}[\mathbf{w}^*(S)] = \mathbf{w}^*$), then:

$$\overbrace{\mathbb{E}[L(\mathbf{w}^*(S))] \le (1+c)\,L(\mathbf{w}^*)}^{\text{loss bound}} \iff \overbrace{\mathbb{E}[\|\hat{\mathbf{y}}(S) - \mathbb{E}[\hat{\mathbf{y}}(S)]\|^2] \le c\,L(\mathbf{w}^*)}^{\text{variance bound}}$$

Take average of k i.i.d. samples of size s: $\overline{\mathbf{w}}^* = \frac{1}{k}\sum_{j=1}^k \mathbf{w}^*(S_j)$,

$\mathbb{E}\left[L\left(\overline{\mathbf{w}}^*\right)\right] \le \left(1 + \frac{c}{k}\right) L(\mathbf{w}^*) \implies \frac{s\,c}{\epsilon}$ labels achieve $(1+\epsilon)L(\mathbf{w}^*)$

With size d volume sampling, we need d^2/ϵ labels. Is d/ϵ possible?

Open: Is there unbiased estimator with $s = O(d)$ and $c = O(1)$?