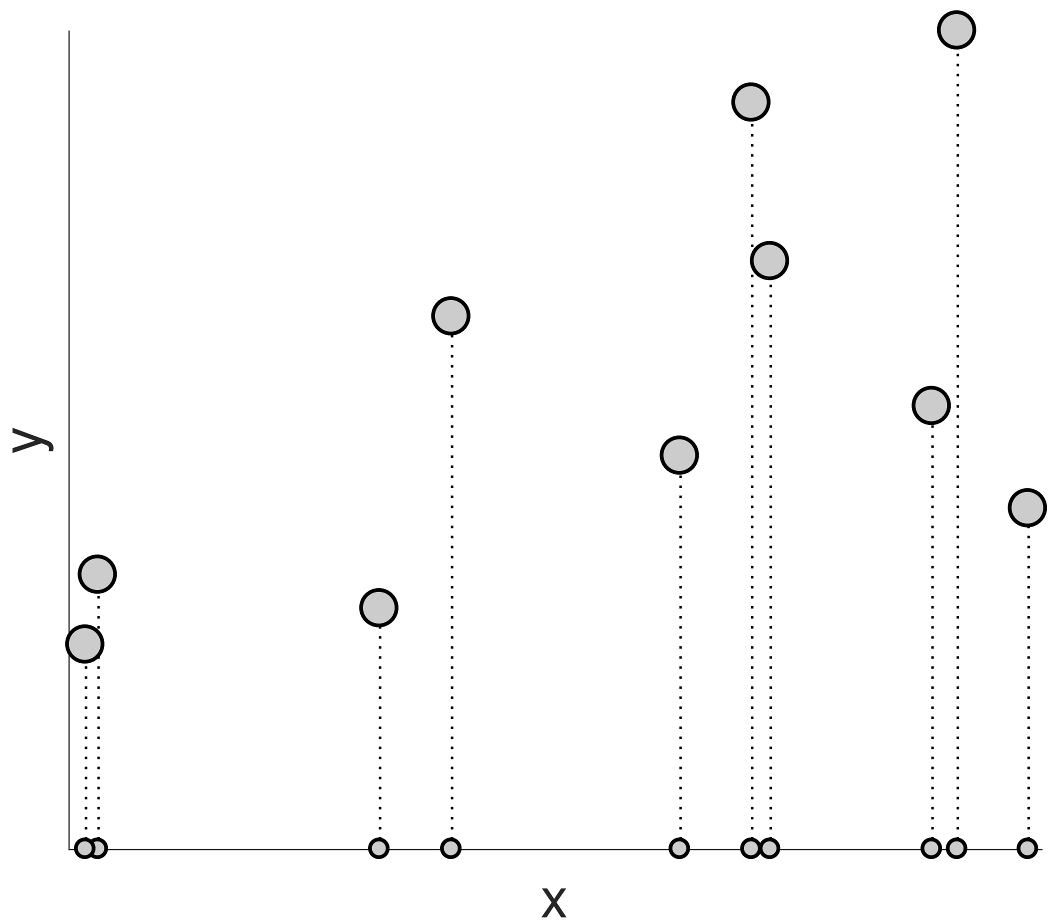
Unbiased estimates for linear regression via volume sampling

Michał Dereziński and Manfred Warmuth

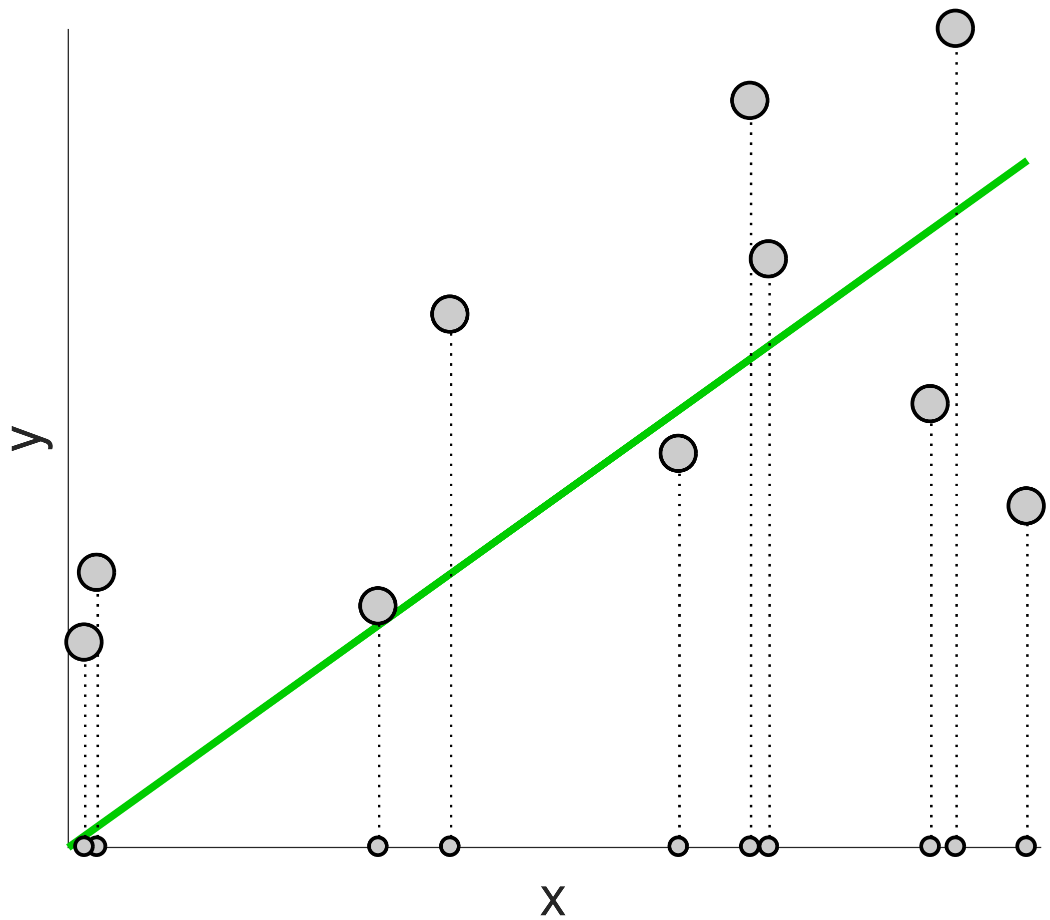
UCSC, Applied Math Seminar, 10-9-17

Outline

Linear regression

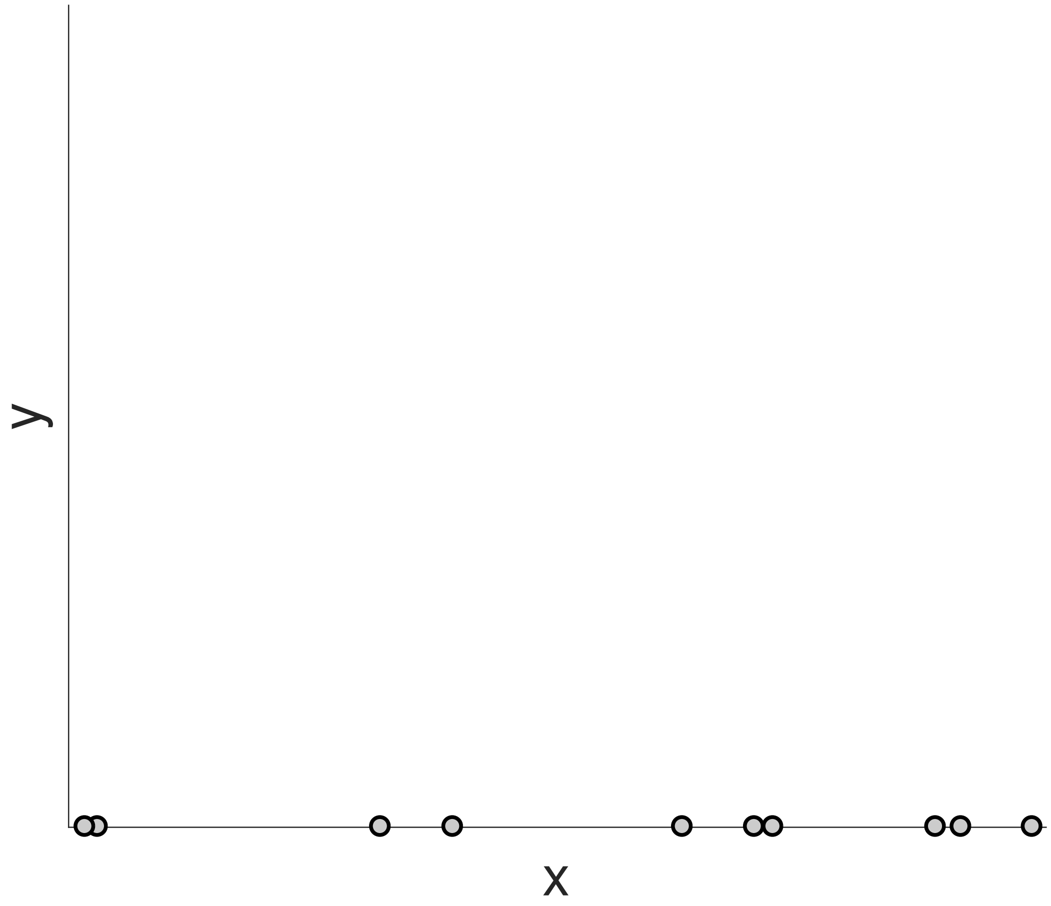


Optimal solution



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

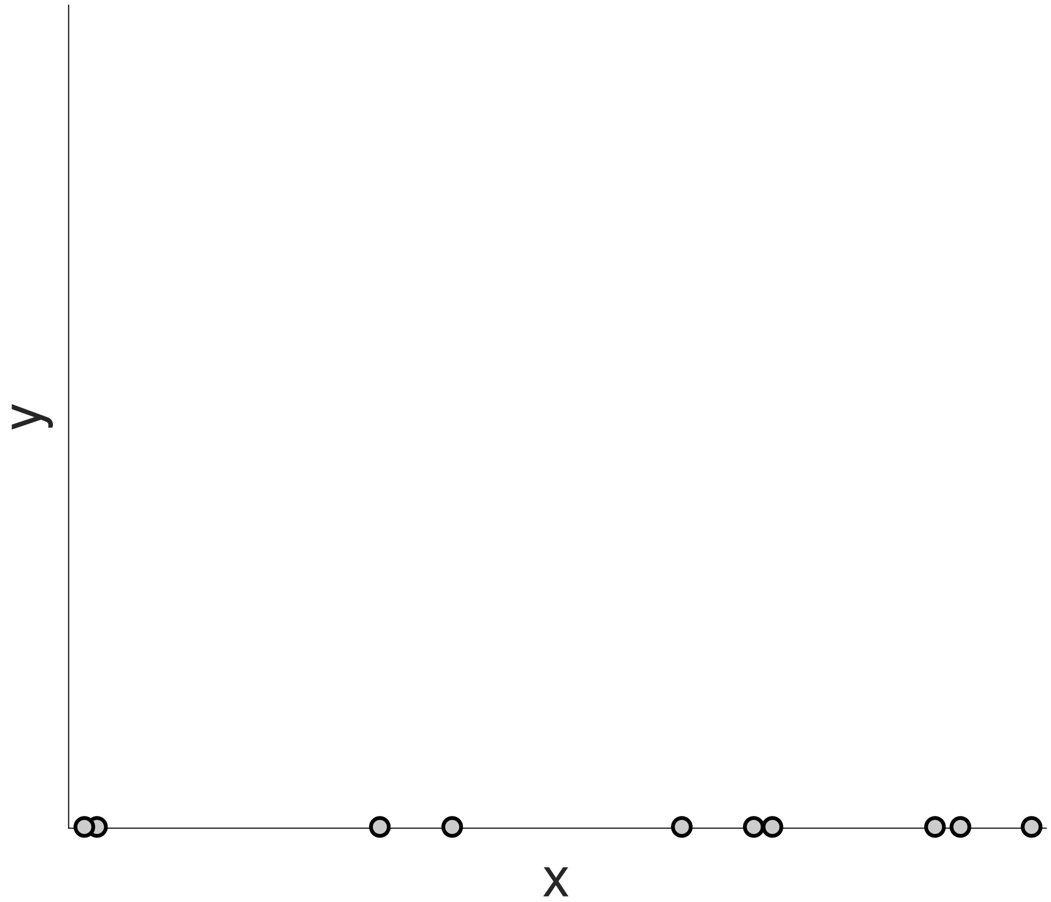
How many labels needed to get close to optimum?



- All x_i given
- But labels y_i unknown

Guess how many needed?

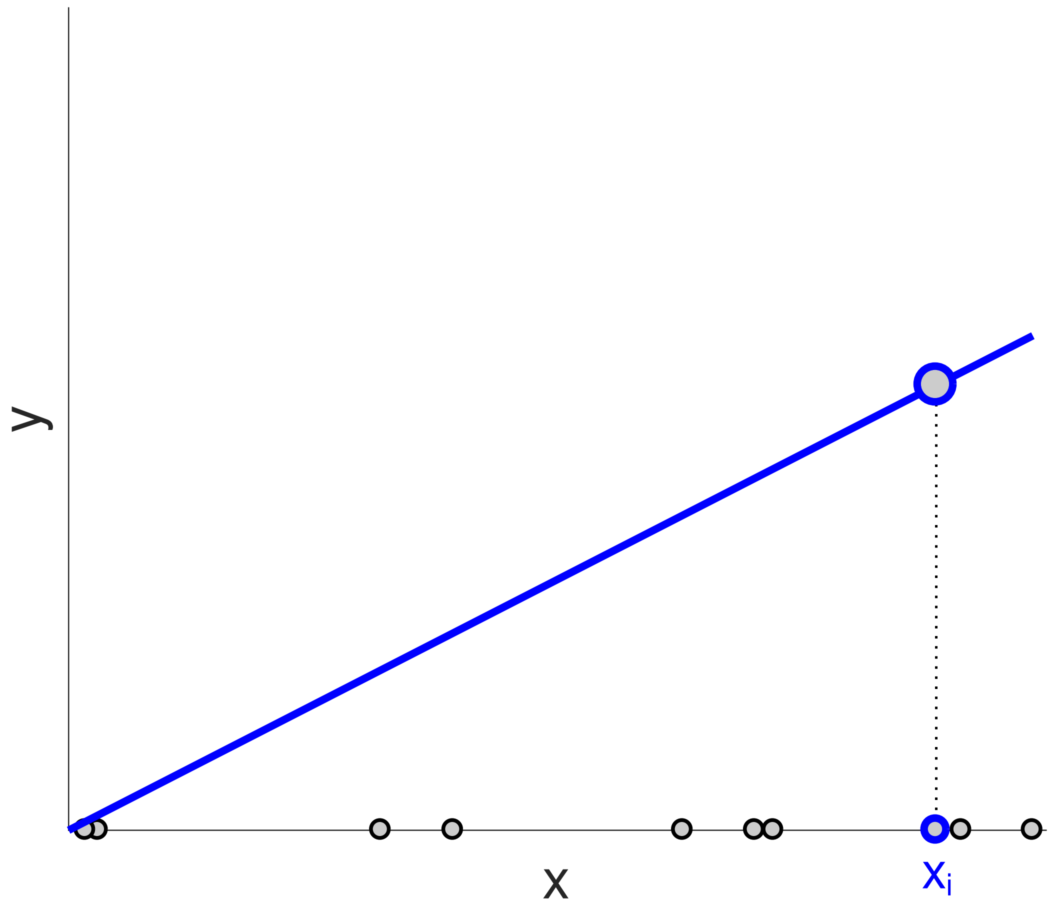
How many labels needed to get close to optimum?



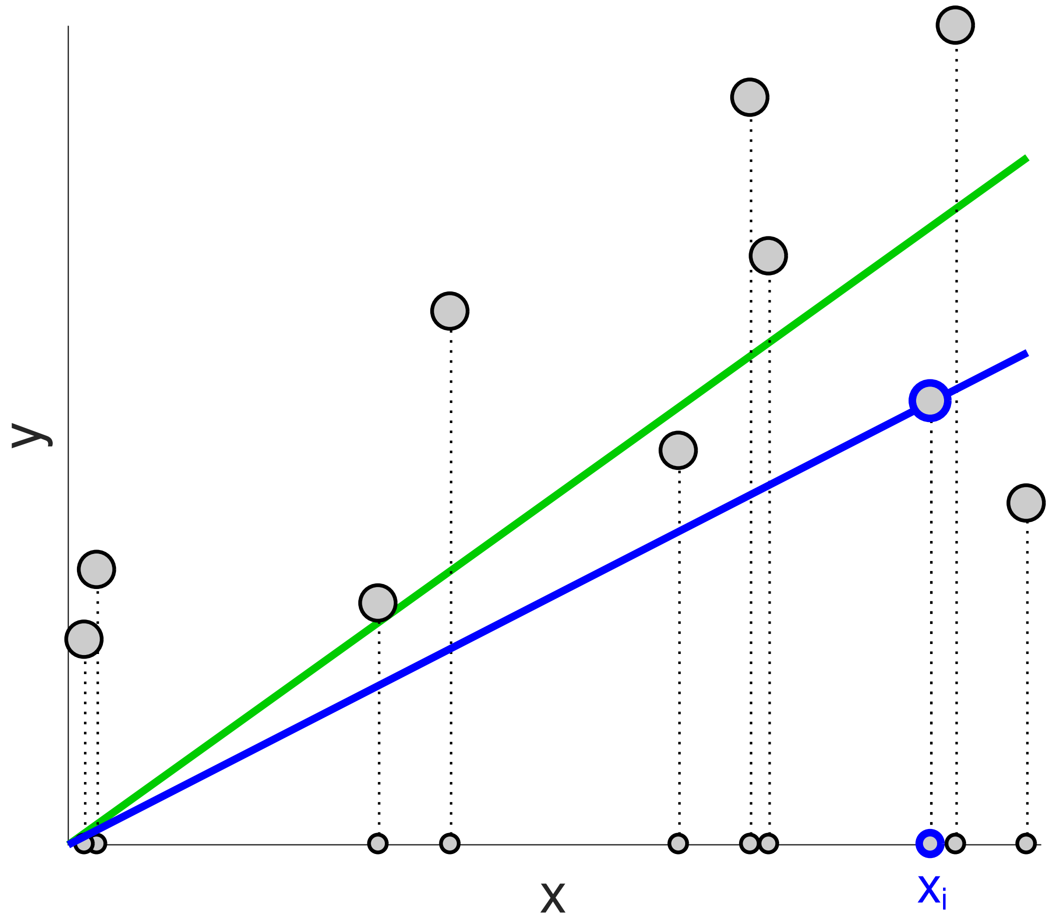
- All x_i given
- But labels y_i unknown

Guess how many needed?

Answer: 1 label

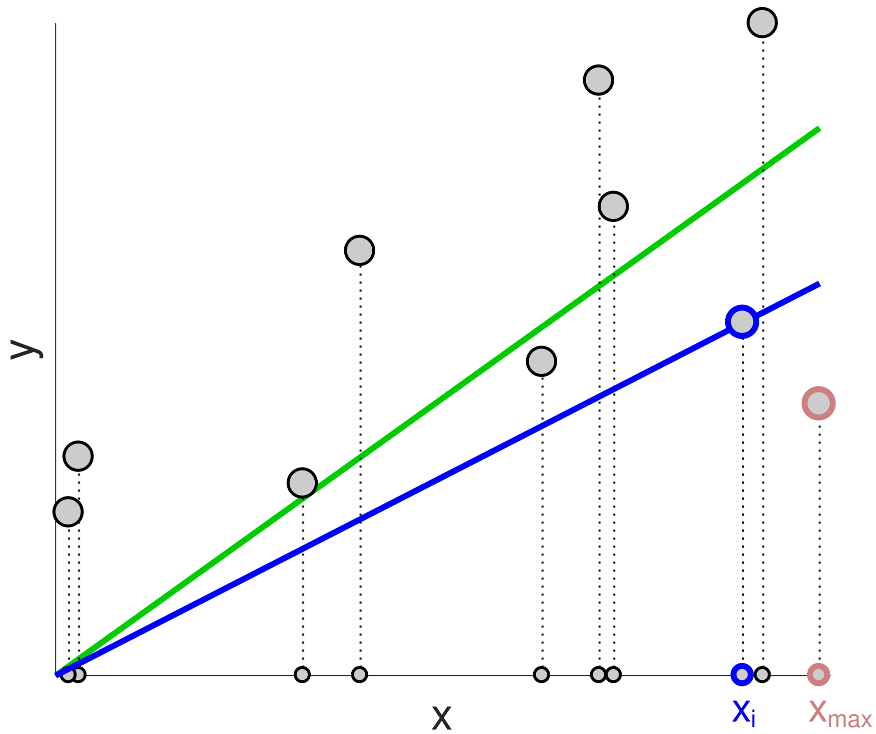


How good is 1 label?



Loss of estimate $=$ 2 $\times$ Loss of optimum

Which one?



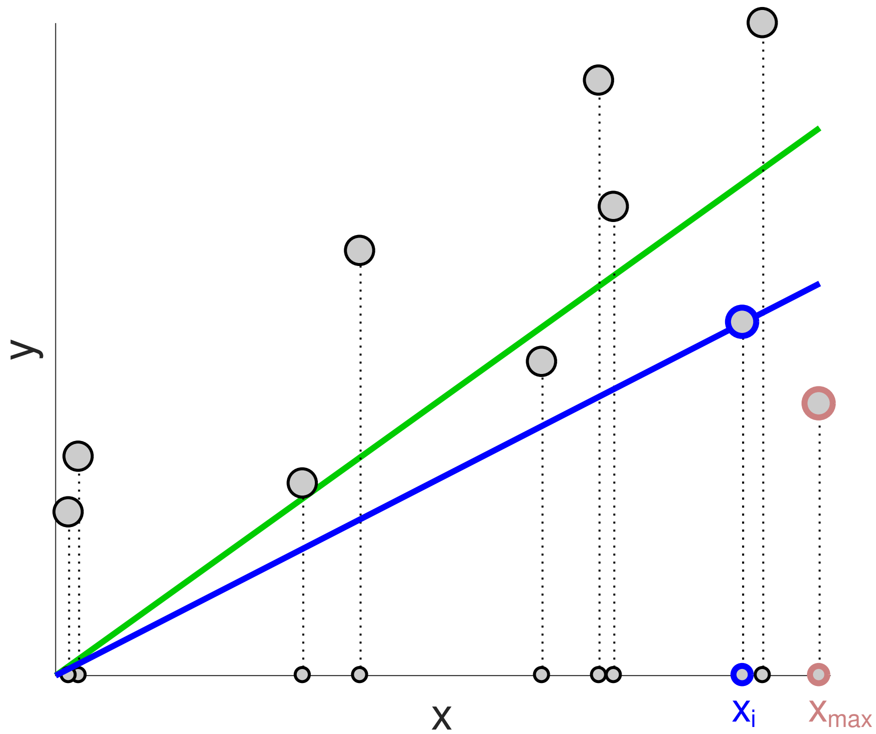
- x_{max} (furthest from 0) is bad
- any deterministic choice is bad

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

$$\mathbb{E}_i \sum_j \Big(\underbrace{\frac{y_i}{x_i}}_{w_i^*} x_j - y_j\Big)^2 = 2 \sum_j (w^* x_j - y_j)^2$$

$$\mathbb{E}_i\, w_i^* = \sum_i \overbrace{\frac{x_i^2}{\|\mathbf{x}\|^2}}^{P(i)} \overbrace{\frac{y_i}{x_i}}^{w_i^*} = w^*$$

Which one?



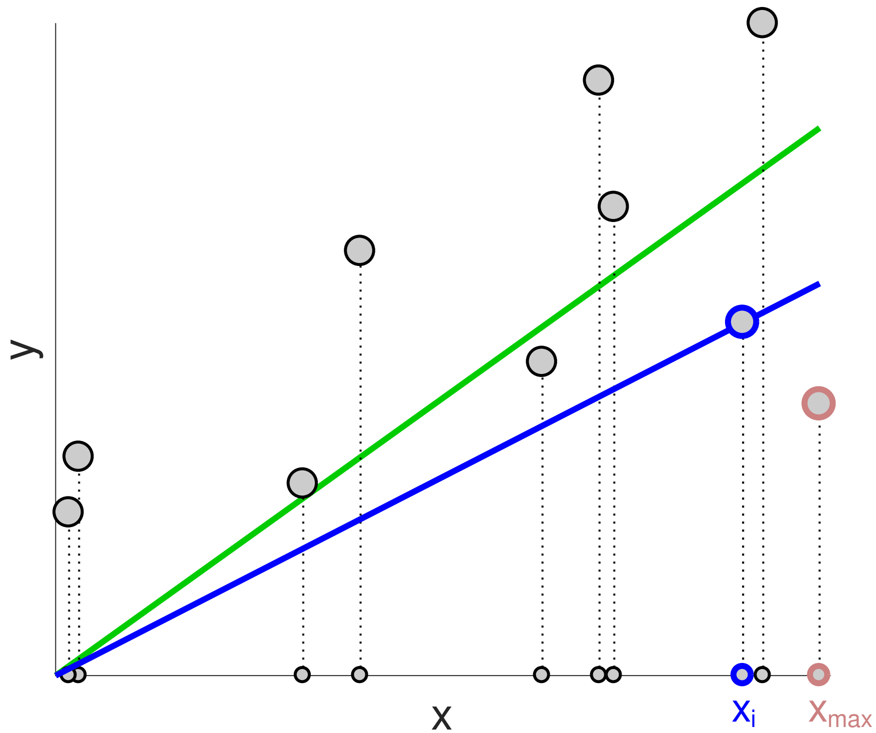
- x_{max} (furthest from 0) is bad
- any deterministic choice is bad

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

$$\mathbb{E}_i \sum_j (\underbrace{\frac{y_i}{x_i}}_{w_i^*} x_j - y_j)^2 = 2 \sum_j (w^* x_j - y_j)^2$$

$$\mathbb{E}_i\, w_i^* = \sum_i \overbrace{\frac{x_i^2}{\|\mathbf{x}\|^2}}^{P(i)} \overbrace{\frac{y_i}{x_i}}^{w_i^*} = w^*$$

Which one?



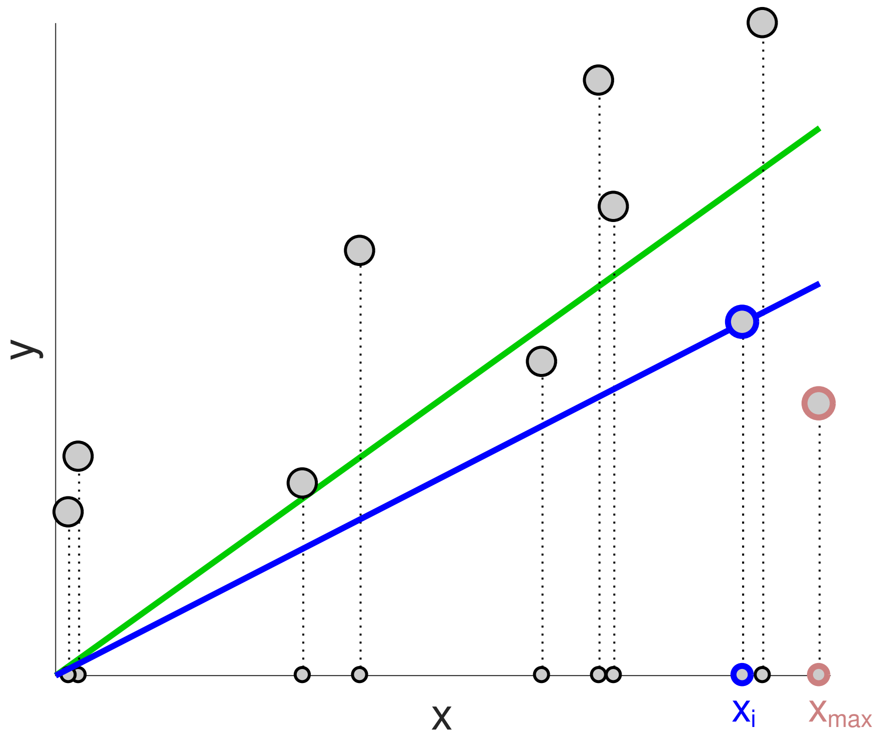
- x_{max} (furthest from 0) is bad
- any deterministic choice is bad

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

$$\mathbb{E}_i \sum_j (\underbrace{\frac{y_i}{x_i}}_{w_i^*} x_j - y_j)^2 = 2 \sum_j (w^* x_j - y_j)^2$$

$$\mathbb{E}_i\, w_i^* = \sum_i \overbrace{\frac{x_i^2}{\|\mathbf{x}\|^2}}^{P(i)} \overbrace{\frac{y_i}{x_i}}^{w_i^*} = w^*$$

Which one?



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

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

$$\mathbb{E}_i \sum_j (\underbrace{\frac{y_i}{x_i}}_{w_i^*} x_j - y_j)^2 = 2 \sum_j (w^* x_j - y_j)^2$$

$$\mathbb{E}_i \, w_i^* = \sum_i \overbrace{\frac{x_i^2}{\|\mathbf{x}\|^2}}^{P(i)} \overbrace{\frac{y_i}{x_i}}^{w_i^*} = w^*$$

General: subsampling for linear regression

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

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

Receive y_4, y_6, y_9

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

Strategy: Solve subproblem $(\mathbf{X}_S, \mathbf{y}_S)$, obtaining:

$$\mathbf{w}^*(S) = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i \in S} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = \mathbf{X}_S^{+\top} \mathbf{y}_S$$

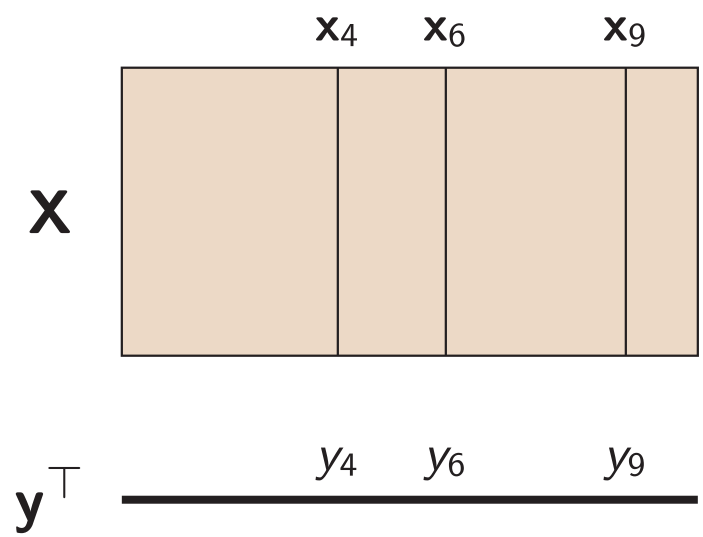
$$\mathbf{X}_S^{+} = \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top)^{-1} \quad \text{- pseudo-inverse of } \mathbf{X}_S$$

General: subsampling for linear regression

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

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

Receive y_4, y_6, y_9



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

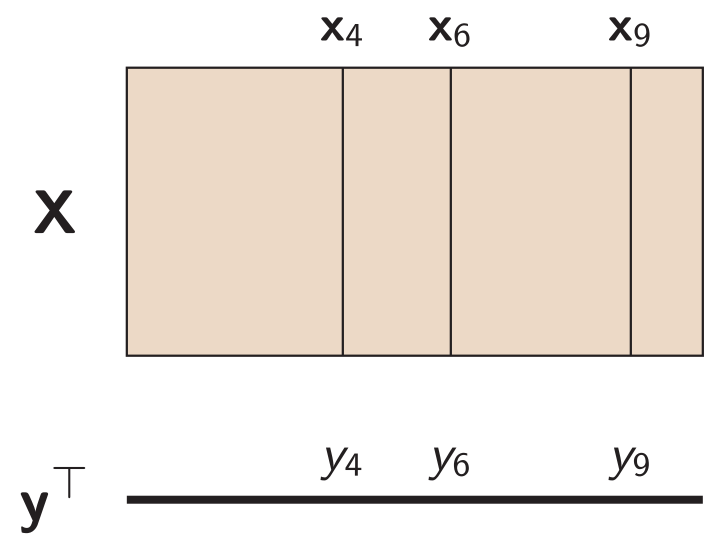
Strategy: Solve subproblem $(\mathbf{X}_S, \mathbf{y}_S)$, obtaining:

$$\mathbf{w}^*(S) = \operatorname*{argmin}_{\mathbf{w}} \sum_{i \in S} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = \mathbf{X}_S^{+\top} \mathbf{y}_S$$

$$\mathbf{X}_S^{+} = \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top)^{-1} \qquad \text{- pseudo-inverse of } \mathbf{X}_S$$

General: subsampling for linear regression

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



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

Receive y_4, y_6, y_9

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

Strategy: Solve subproblem $(\mathbf{X}_S, \mathbf{y}_S)$, obtaining:

$$\mathbf{w}^*(S) = \operatorname*{argmin}_{\mathbf{w}} \sum_{i \in S} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = \mathbf{X}_S^{+\top} \mathbf{y}_S$$

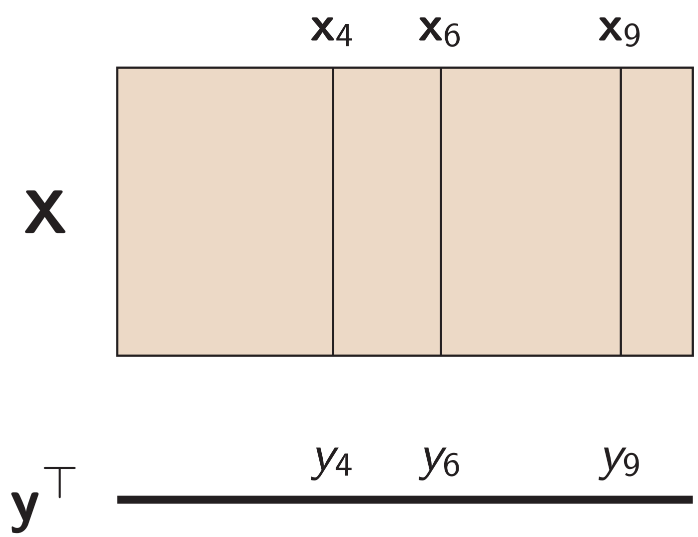
$$\mathbf{X}_S^+ = \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top)^{-1} \quad \text{- pseudo-inverse of } \mathbf{X}_S$$

General: subsampling for linear regression

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

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

Receive y_4, y_6, y_9



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

Strategy: Solve subproblem $(\mathbf{X}_S, \mathbf{y}_S)$, obtaining:

$$\mathbf{w}^*(S) = \operatorname*{argmin}_{\mathbf{w}} \sum_{i \in S} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = \mathbf{X}_S^{+\top} \mathbf{y}_S$$

$$\mathbf{X}_S^{+} = \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top)^{-1} \quad \text{- pseudo-inverse of } \mathbf{X}_S$$

Are dimension many labels sufficient?

Claim

There is no good deterministic algorithm for selecting d labels.

1-dimensional example:

$$\begin{array}{cc} & \begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & & \mathbf{x}_n \end{array} \\ \mathbf{X} = & \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \\ \mathbf{y}^\top = & \begin{pmatrix} 0 & 1 & \cdots & 1 \end{pmatrix} \end{array}$$

Deterministic pick $S = \{1\}$, receive $y_1 = 0$
Deterministic predictor $\mathbf{w}^*(\{1\}) = 0$
Optimal predictor $\mathbf{w}^* = \frac{n-1}{n} = 1 - \frac{1}{n}$

$$\underbrace{L(\overbrace{\mathbf{w}^*(\{1\})}^{0})}_{n-1} = n \underbrace{L(\overbrace{\mathbf{w}^*}^{\frac{n-1}{n}})}_{\frac{n-1}{n}}$$

With uniform choice of $S : |S| = 1$, $\mathbb{E}[L(\mathbf{w}^*(S))] = 2L(\mathbf{w}^*)$

Our Result

A randomized algorithm can achieve $\mathbb{E}[L(\mathbf{w}^*(S))] = (d+1)\, L(\mathbf{w}^*)$

Are dimension many labels sufficient?

Claim

There is no good deterministic algorithm for selecting d labels.

1-dimensional example:

$$\begin{array}{cc} & \begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & & \mathbf{x}_n \end{array} \\ \mathbf{X} = & \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \\ \mathbf{y}^\top = & \begin{pmatrix} 0 & 1 & \cdots & 1 \end{pmatrix} \end{array}$$

Deterministic pick $S = \{1\}$, receive $y_1 = 0$
Deterministic predictor $\mathbf{w}^*(\{1\}) = 0$
Optimal predictor $\mathbf{w}^* = \frac{n-1}{n} = 1 - \frac{1}{n}$

$$\underbrace{L(\overbrace{\mathbf{w}^*(\{1\})}^{0})}_{n-1} = n \underbrace{L(\overbrace{\mathbf{w}^*}^{\frac{n-1}{n}})}_{\frac{n-1}{n}}$$

With uniform choice of $S : |S| = 1,\ \mathbb{E}[L(\mathbf{w}^*(S))] = 2L(\mathbf{w}^*)$

Our Result

A randomized algorithm can achieve $\mathbb{E}[L(\mathbf{w}^*(S))] = (d+1)\, L(\mathbf{w}^*)$

Are dimension many labels sufficient?

Claim

There is no good deterministic algorithm for selecting d labels.

1-dimensional example:

$$\begin{array}{rcl} & & \begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & & \mathbf{x}_n \end{array} \\ \mathbf{X} & = & \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \\ \mathbf{y}^\top & = & \begin{pmatrix} 0 & 1 & \cdots & 1 \end{pmatrix} \end{array}$$

Deterministic pick $S = \{1\}$, receive $y_1 = 0$
Deterministic predictor $\mathbf{w}^*(\{1\}) = 0$
Optimal predictor $\mathbf{w}^* = \frac{n-1}{n} = 1 - \frac{1}{n}$

$$\underbrace{L(\overbrace{\mathbf{w}^*(\{1\})}^{0})}_{n-1} = n\ \underbrace{L(\overbrace{\ \mathbf{w}^*\ }^{\frac{n-1}{n}})}_{\frac{n-1}{n}}$$

With uniform choice of $S : |S| = 1,\ \mathbb{E}[L(\mathbf{w}^*(S))] = 2L(\mathbf{w}^*)$

Our Result

A randomized algorithm can achieve $\mathbb{E}[L(\mathbf{w}^*(S))] = (d+1)\, L(\mathbf{w}^*)$

Are dimension many labels sufficient?

Claim

There is no good deterministic algorithm for selecting d labels.

1-dimensional example:

$$\begin{array}{rcl} & & \begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & & \mathbf{x}_n \end{array} \\ \mathbf{X} & = & \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \\ \mathbf{y}^\top & = & \begin{pmatrix} 0 & 1 & \cdots & 1 \end{pmatrix} \end{array}$$

Deterministic pick $S = \{1\}$, receive $y_1 = 0$

Deterministic predictor $\mathbf{w}^*(\{1\}) = 0$

Optimal predictor $\mathbf{w}^* = \frac{n-1}{n} = 1 - \frac{1}{n}$

$$\underbrace{L(\overbrace{\mathbf{w}^*(\{1\})}^{0})}_{n-1} = n\ \underbrace{L(\overbrace{\mathbf{w}^*}^{\frac{n-1}{n}})}_{\frac{n-1}{n}}$$

With uniform choice of $S : |S| = 1,\ \mathbb{E}[L(\mathbf{w}^*(S))] = 2L(\mathbf{w}^*)$

Our Result

A randomized algorithm can achieve $\mathbb{E}[L(\mathbf{w}^*(S))] = (d+1)\, L(\mathbf{w}^*)$

Towards Volume Sampling

$L(\mathbf{w}^*) = 1.81$

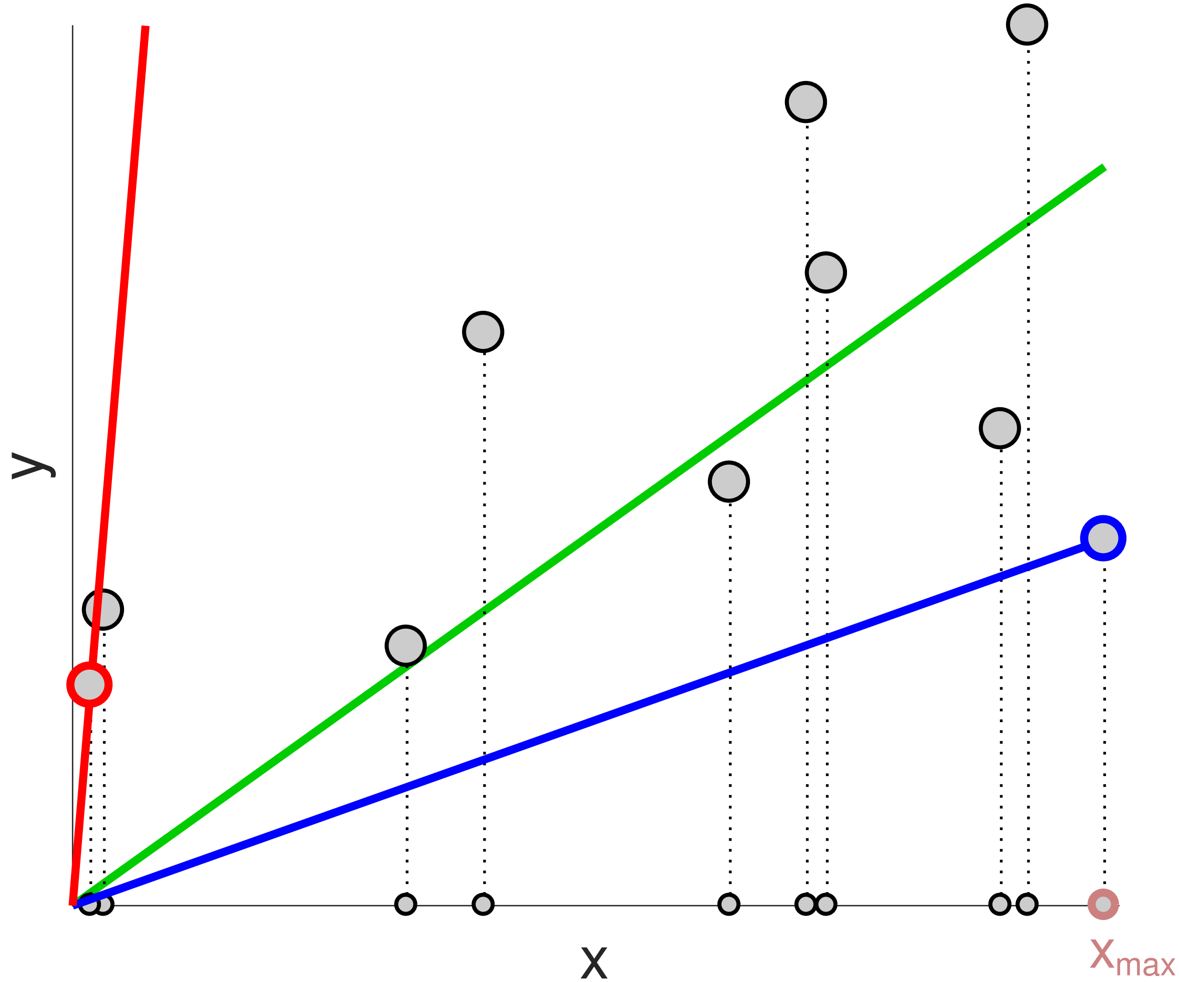
$L(\mathbf{w}^*(S_1)) = 4.03$

$L(\mathbf{w}^*(S_2)) = 2199$

For $P(S) \propto \|\mathbf{x}_S\|^2$

$\mathbb{E}[L(\mathbf{w}^*(S))] = 3.61$

$= 2L(\mathbf{w}^*)$



Instances with larger norm $\|\mathbf{x}\|^2$ are more informative

What generalizes $\|\mathbf{x}\|^2$?

Volume Sampling[1]

Generalize norms to sets of examples

$$\mathbf{X}_S = \begin{pmatrix} | & | \\ \mathbf{x}_i & \mathbf{x}_j \\ | & | \end{pmatrix}$$

Distribution over all d-element subsets S:

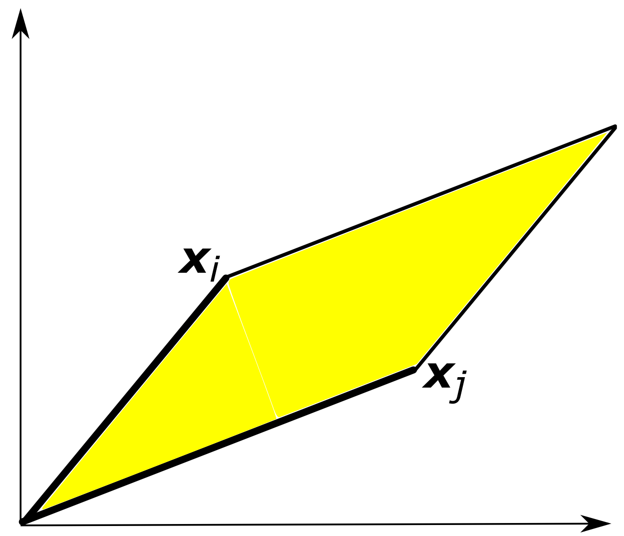
$$P(S) = \det(\mathbf{X}_S\mathbf{X}_S^\top) \,/\, Z$$

Also well defined for any $|S| \geq d$.

$\det(\mathbf{X}_S\mathbf{X}_S^\top) =$ **squared** volume of parallelepiped $\mathcal{P}(\mathbf{x}_i, \mathbf{x}_j)$

Note: Normalization factor Z can be derived using Cauchy-Binet formula:

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



[1] Deshpande, Rademacher, Vempala, Wang. 2006

Volume Sampling[1]

Generalize norms to sets of examples

Distribution over all d-element subsets S:

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

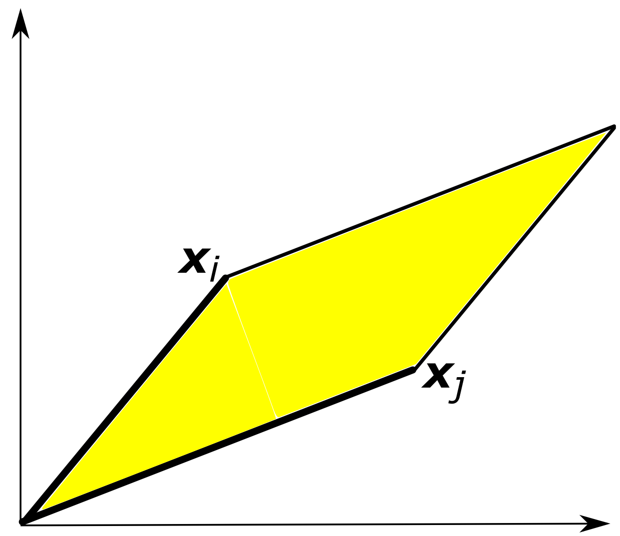
Also well defined for any $|S| \geq d$.

Note: Normalization factor Z can be derived using Cauchy-Binet formula:

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

$$\mathbf{X}_S = \begin{pmatrix} | & | \\ \mathbf{x}_i & \mathbf{x}_j \\ | & | \end{pmatrix}$$

$\det(\mathbf{X}_S \mathbf{X}_S^\top) =$ **squared** volume of parallelepiped $\mathcal{P}(\mathbf{x}_i, \mathbf{x}_j)$



[1] Deshpande, Rademacher, Vempala, Wang. 2006

How many examples needed?

We will show that using volume sampling,
d **labels suffice** to achieve a multiplicative approximation

Thm: For any full rank matrix $\mathbf{X}$, $d - 1$ **labels do not suffice**

Proof idea: Adversary has freedom to set the label of one additional point while $L(\mathbf{w}^*) = 0$ and algorithm has positive loss

Outline

Main results

For a volume-sampled d-element set S,

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

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

- Sampling distribution does not depend on the labels
- No range restrictions! **No dependence on** n

Recall model:
- Adversary picks $\mathbf{X}$
- Learner picks subset of label indices
- Adversary picks all labels

Main Results in a Picture