Unbiased estimates for linear regression via volume sampling

Michał Dereżiński and Manfred Warmuth

UCSC, Applied Math Seminar, 10-9-17
Outline

Introduction

Overview of Results

Main Proof Method

Further Research Directions

Proof of Loss Expectation Formula

Appendix
Linear regression
Optimal solution

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

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

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

\[
\mathbb{E}_i \sum_j \left( \frac{y_i}{x_i} x_j - y_j \right)^2 = 2 \sum_j (w^* x_j - y_j)^2
\]

\[
\mathbb{E}_i w_i^* = \sum_i \frac{x_i^2}{\|x\|^2} \frac{y_i}{x_i} = w^*
\]
Which one?

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

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

\[
\mathbb{E}_i \sum_j \left( \frac{y_i}{x_i} x_j - y_j \right)^2 = 2 \sum_j \left( w^* x_j - y_j \right)^2
\]

\[
\mathbb{E}_i w_i^* = \sum_i \frac{x_i^2}{\|x\|^2} \frac{y_i}{x_i} = w^*
\]
Which one?

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

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

$$
\mathbb{E}_i \sum_j \left( \frac{y_i}{x_j} x_j - y_j \right)^2 = 2 \sum_j (w^* x_j - y_j)^2
$$

$$
\mathbb{E}_i w_i^* = \sum_i \frac{x_i^2}{\|x\|^2} \frac{y_i}{x_i} = w^*
$$
Which one?

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

**Good:** 1 label \( y_i \) drawn \( \sim x_i^2 \)

\[
\mathbb{E}_i \sum_j \left( \frac{y_i}{x_j} \right) x_j - y_j)^2 = 2 \sum_j (w^* x_j - y_j)^2
\]

\[
\mathbb{E}_i w_i^* = \sum_i \frac{x_i^2}{\|x\|^2} \frac{y_i}{x_i} = w^*
\]
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:

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

$$\mathbf{X}_S^+ = \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top)^{-1}$$ - pseudo-inverse of $\mathbf{X}_S$
General: subsampling for linear regression

**Given:** \( n \) points \( x_i \in \mathbb{R}^d \) with hidden labels \( y_i \in \mathbb{R} \)

Select \( S = \{4, 6, 9\} \)

Receive \( y_4, y_6, y_9 \)

**Goal:** Minimize loss \( L(w) = \sum_i (x_i^\top w - y_i)^2 \) over all \( n \) points

**Strategy:** Solve subproblem \((X_S, y_S)\), obtaining:

\[
\begin{align*}
w^*(S) &= \arg\min_w \sum_{i \in S} (x_i^\top w - y_i)^2 = X_S^+ y_S \\
X_S^+ &= X_S^\top (X_S X_S^\top)^{-1} - \text{pseudo-inverse of } X_S
\end{align*}
\]
General: subsampling for linear regression

**Given:** \( n \) points \( x_i \in \mathbb{R}^d \) with hidden labels \( y_i \in \mathbb{R} \)

Select \( S = \{4, 6, 9\} \)

Receive \( y_4, y_6, y_9 \)

**Goal:** Minimize loss \( L(w) = \sum_i (x_i^T w - y_i)^2 \) over all \( n \) points

**Strategy:** Solve subproblem \( (X_S, y_S) \), obtaining:

\[
 w^*(S) = \arg\min_w \sum_{i \in S} (x_i^T w - y_i)^2 = X_S^{+T} y_S
\]

\[
 X_S^+ = X_S^T (X_S X_S^T)^{-1} \quad \text{- pseudo-inverse of } X_S
\]
General: subsampling for linear regression

**Given**: $n$ points $x_i \in \mathbb{R}^d$ with hidden labels $y_i \in \mathbb{R}$

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

Receive $y_4, y_6, y_9$

**Goal**: Minimize loss $L(w) = \sum_i (x_i^T w - y_i)^2$ over all $n$ points

**Strategy**: Solve subproblem $(X_S, y_s)$, obtaining:

$$w^*(S) = \arg\min_w \sum_{i \in S} (x_i^T w - y_i)^2 = X_S^{+\top} y_s$$

$$X_S^+ = X_S^\top (X_S X_S^\top)^{-1}$$ - pseudo-inverse of $X_S$
Are dimension many labels sufficient?

Claim

There is no good deterministic algorithm for selecting \(d\) labels.

1-dimensional example:

\[
\begin{align*}
\mathbf{x} &= (x_1 \ 1 \ \cdots \ 1) \\
y^T &= (0 \ 1 \ \cdots \ 1)
\end{align*}
\]

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

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

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

\[
L\left(\mathbf{w}^*(\{1\})\right) = \frac{0}{n-1} \quad L\left(\mathbf{w}^*\right) = \frac{n-1}{n}
\]

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:

$$
\begin{align*}
X & = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \\
y^T & = \begin{pmatrix} 0 & 1 & \cdots & 1 \end{pmatrix}
\end{align*}
$$

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

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

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

$$
\begin{align*}
L(w^*({\{1\}})) & = n L(w^*) \\
\frac{n-1}{n} & = \frac{n-1}{n}
\end{align*}
$$

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

Our Result

A randomized algorithm can achieve $\mathbb{E}[L(w^*(S))] = (d + 1) L(w^*)$
Are dimension many labels sufficient?

Claim

There is no good deterministic algorithm for selecting $d$ labels.

1-dimensional example:

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$$

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

$$Y^T = \begin{pmatrix} 0 & 1 & \cdots & 1 \end{pmatrix}$$

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

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

$$L\left( w^*(\{1\}) \right) = n L\left( w^* \right)$$

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

Our Result

A randomized algorithm can achieve $\mathbb{E}[L(w^*(S))] = (d + 1) L(w^*)$
Are dimension many labels sufficient?

**Claim**

There is no good deterministic algorithm for selecting $d$ labels.

**1-dimensional example:**

\[
X = \begin{pmatrix}
  x_1 & x_2 & \cdots & x_n
\end{pmatrix}
\]

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

\[
y^\top = \begin{pmatrix}
  0 & 1 & \cdots & 1
\end{pmatrix}
\]

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

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

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

**Our Result**

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

$L(w^*) = 1.81$

$L(w^*(S_1)) = 4.03$

$L(w^*(S_2)) = 2199$

For $P(S) \propto \|x_s\|^2$

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

$= 2L(w^*)$

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

What generalizes $\|x\|^2$?
Volume Sampling

Generalize norms to sets of examples

Distribution over all $d$-element subsets $S$:

$$P(S) = \frac{\det(X_S X_S^\top)}{Z}$$

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

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

$$Z = \sum_{S:|S|=d} \det(X_S X_S^\top) = \det(XX^\top)$$

$X_S = \begin{pmatrix} x_i & x_j \end{pmatrix}$

$\det(X_S X_S^\top)$ = squared volume of parallelepiped $P(x_i, x_j)$

$^{1}$Deshpande, Rademacher, Vempala, Wang. 2006
Volume Sampling

Generalize norms to sets of examples

Distribution over all $d$-element subsets $S$:

$$P(S) = \frac{\det(X_S X_S^\top)}{Z}$$

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

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

$$Z = \sum_{S:|S|=d} \det(X_S X_S^\top) = \det(XX^\top)$$

1Deshpande, Rademacher, Vempala, Wang. 2006
How many examples needed?

We will show that using volume sampling, \( d \) labels \textbf{suffice} to achieve a multiplicative approximation.

**Thm:** For any full rank matrix \( X \), \( d - 1 \) labels do not suffice.

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

Introduction

Overview of Results

Main Proof Method

Further Research Directions

Proof of Loss Expectation Formula

Appendix
Main results

For a volume-sampled $d$-element set $S$, 

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

if $X$ is in general position

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

Recall model:
- Adversary picks $X$
- Learner picks subset of label indices
- Adversary picks all labels