Subsampling for Ridge Regression via Regularized Volume Sampling

Michał Dereziński and Manfred Warmuth

University of California at Santa Cruz

AISTATS’18, 4-9-18
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: one label
Which one?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$E_i \sum_j \left( x_j \frac{y_i}{x_i} - x_j w^* \right)^2 = \sum_j (x_j w^* - y_j)^2$$

$$E_i w_i^* = \sum_i \frac{x_i^2}{\|x\|^2} \frac{y_i}{x_i} = w^*$$
Generalization: Volume Sampling

- Given \( n \) points \( x_i \in \mathbb{R}^d \)
- Choose subset \( S \) of \( d \) indices
- Get labels \( y_i \in \mathbb{R} \) for \( i \in S \)
- Find \( w^*(S) \): solution for the \( d \) examples \( \{(x_i, y_i) : i \in S\} \)

Key prior result [DW17]: for any \((x_1, y_1), \ldots, (x_n, y_n)\),

\[
\mathbb{E}_S \sum_j (x_j^T w^*(S) - x_j^T w^*)^2 = d \sum_j (x_j^T w^* - y_j)^2
\]

when \( S \) chosen \( \sim \)

squared volume of parallelopiped
spanned by the \( \{x_i : i \in S\} \)

\(^1\)DW. NIPS’17
Generalization: Volume Sampling

- Given \( n \) points \( x_i \in \mathbb{R}^d \)
- Choose subset \( S \) of \( d \) indices
- Get labels \( y_i \in \mathbb{R} \) for \( i \in S \)
- Find \( w^*(S) \): solution for the \( d \) examples \( \{(x_i, y_i) : i \in S\} \)

**Key prior result** [DW17]: for any \( (x_1, y_1), \ldots, (x_n, y_n) \),

\[
\mathbb{E}_S \sum_j (x_j^\top w^*(S) - x_j^\top w^*)^2 = d \sum_j (x_j^\top w^* - y_j)^2
\]

when \( S \) chosen \( \sim \) squared volume of parallelopiped spanned by the \( \{x_i : i \in S\} \)

**Unbiasedness:**

\[
\mathbb{E}_S w^*(S) = w^*
\]

---

1DW. NIPS'17
Generalization: Volume Sampling

- Given $n$ points $x_i \in \mathbb{R}^d$
- Choose subset $S$ of $d$ indices
- Get labels $y_i \in \mathbb{R}$ for $i \in S$
- Find $w^*(S)$: solution for the $d$ examples $\{(x_i, y_i) : i \in S\}$

Key prior result [DW17]: for any $(x_1, y_1), \ldots, (x_n, y_n)$,

$$
\mathbb{E}_S \sum_j (x_j^\top w^*(S) - x_j^\top w^*)^2 = d \sum_j (x_j^\top w^* - y_j)^2
$$

when $S$ chosen $\sim$ squared volume of parallelopiped spanned by the $\{x_i : i \in S\}$

Open:
Replace factor $d$ with $\epsilon$, using small $|S| \geq d$

---

$^1$DW. NIPS’17
Two types of volume sampling

Let \( \mathbf{X} \in \mathbb{R}^{d \times n} \) be full rank, \( n \gg d \)

Distribution over all \( s \)-element subsets \( S \):

\[
P(S) \sim \begin{cases} 
\det(\mathbf{X}_S^\top \mathbf{X}_S) & \text{if } s \leq d \ [\text{DRVW06}]^1 \\
\det(\mathbf{X}_S \mathbf{X}_S^\top) & \text{if } s \geq d \ [\text{AB13}]^2 
\end{cases}
\]

\[
\mathbf{X}_S = \begin{pmatrix} x_i & x_j \end{pmatrix}
\]

\[
\det(\mathbf{X}_S \mathbf{X}_S^\top) = \text{squared volume of parallelepiped } \mathcal{P}(x_i, x_j)
\]

---

^1[DRVW06] Deshpande et al. SODA’06
^2[AB13] Avron and Boutsidis. JMAA’13
Prior work: efficient volume sampling algorithms

Sampling sets of size $s \leq d$:

1. Deshpande and Rademacher (FOCS’10) $O(snd^{\omega} \log d)$
2. Guruswami and Sinop (SODA’12) $O(snd^2)$

Sampling sets of size $s \geq d$:

1. Li et al. (NIPS’17) (approximate MCMC) $\tilde{O}(s^3 nd^2)$
2. DW (NIPS’17) $O(n^2 d)$

We give an $O(nd^2)$ algorithm for $s \geq d$
Prior work: efficient volume sampling algorithms

Sampling sets of size $s \leq d$:

1. Deshpande and Rademacher (FOCS’10) $O(s^{nd^\omega} \log d)$
2. Guruswami and Sinop (SODA’12) $O(s^{nd^2})$

Sampling sets of size $s \geq d$:

1. Li et al. (NIPS’17) (approximate MCMC) $\tilde{O}(s^3 nd^2)$
2. DW (NIPS’17) $O(n^2 d)$

We give an $O(nd^2)$ algorithm for $s \geq d$
Our contributions: subsampling for linear regression

1. Statistical

1.1 *Dimension-free error bounds* (random noise assumptions)

\[
\text{Error} = O\left(\frac{\text{degrees of freedom}}{\text{subset size } s}\right)
\]

1.2 *Lower bounds*

- volume sampling achieves near-optimal error bounds
- no i.i.d. sampling works as well for small sample size \( s \)

2. Computational

2.1 *Regularized volume sampling*

2.2 *Faster sampling algorithm* \( O(nd^2) \) for any \( s \geq d \)
Our contributions: subsampling for linear regression

1. **Statistical**
   
   1.1 *Dimension-free error bounds* (random noise assumptions)
   
   \[
   \text{Error} = O\left(\frac{\text{degrees of freedom}}{\text{subset size } s}\right)
   \]

   1.2 *Lower bounds*
   
   ▶ volume sampling achieves near-optimal error bounds
   
   ▶ no i.i.d. sampling works as well for small sample size \(s\)

2. **Computational**

   2.1 *Regularized volume sampling*

   2.2 *Faster sampling algorithm* \(O(nd^2)\) for any \(s \geq d\)
Our contributions: subsampling for linear regression

1. Statistical
   1.1 *Dimension-free error bounds* (random noise assumptions)
      \[
      \text{Error} = O\left(\frac{\text{degrees of freedom}}{\text{subset size } s}\right)
      \]
   1.2 *Lower bounds*
      - volume sampling achieves near-optimal error bounds
      - no i.i.d. sampling works as well for small sample size \( s \)

2. Computational
   2.1 *Regularized volume sampling*
   2.2 *Faster sampling algorithm* \( O(nd^2) \) for any \( s \geq d \)
Statistical model: linear labels plus bounded noise

Fixed design matrix: $\mathbf{X} \in \mathbb{R}^{d \times n}$, linear model: $\mathbf{w}^* \in \mathbb{R}^d$

$$\mathbf{y} = \mathbf{X}^\top \mathbf{w}^* + \xi,$$

where $\mathbb{E}[\xi] = 0$, and $\text{Var}[\xi] \preceq \sigma^2 \mathbf{I}$

Goal: Minimize Mean Squared Prediction Error using few labels

$$\text{MSPE}(\hat{\mathbf{w}}) = \mathbb{E} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \hat{\mathbf{w}} - \mathbf{x}_i^\top \mathbf{w}^*)^2$$
Statistical model: linear labels plus bounded noise

Fixed design matrix: \( \mathbf{X} \in \mathbb{R}^{d \times n} \), linear model: \( \mathbf{w}^* \in \mathbb{R}^d \)

\[
y = \mathbf{X}^\top \mathbf{w}^* + \xi, \quad \text{where} \quad \mathbb{E}[\xi] = 0, \quad \text{and} \quad \text{Var}[\xi] \preceq \sigma^2 \mathbf{I}
\]

**Goal:** Minimize *Mean Squared Prediction Error* using few labels

\[
\text{MSPE}(\hat{\mathbf{w}}) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \hat{\mathbf{w}} - \mathbf{x}_i^\top \mathbf{w}^*)^2 \right]
\]
Dimension-free error bound

Ridge estimator: \( \hat{w}^*_\lambda(S) = \operatorname{argmin}_w \| X_S^T w - y_S \|^2 + \lambda \| w \|^2 \) \[ \left( X_S X_S^T + \lambda I \right)^{-1} X_S y_S \]

Main theorem

If \( \lambda \leq \frac{\sigma^2}{\| w^* \|^2} \), then there is a distribution over subsets \( S \) of size \( s \) s.t. \( \text{MSPE}(\hat{w}^*_\lambda(S)) \leq \frac{\sigma^2 d_\lambda}{s - d_\lambda + 1} = O \left( \frac{\sigma^2 d_\lambda}{s} \right) \), where \( d_\lambda = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \lambda} < d \), \( \{ \lambda_i \} \) - eigenvalues of \( XX^T \)

\( d_\lambda \) - degrees of freedom (statistical dimension) of \( X \), given \( \lambda \)

Note: Under decaying eigenvalues, \( d_\lambda \ll d \)
Dimension-free error bound

Ridge estimator: \[ \hat{w}^*_{\lambda}(S) = \arg\min_{w} \left\| X^T_S w - y_S \right\|^2 + \lambda \left\| w \right\|^2 \]

\[ (X_S X_S^T + \lambda I)^{-1} X_S y_S \]

Main theorem

If \( \lambda \leq \frac{\sigma^2}{\|w^*\|^2} \), then there is a distribution over subsets \( S \) of size \( s \)

\[ \text{s.t. } \text{MSPE}(\hat{w}^*_{\lambda}(S)) \leq \frac{\sigma^2 d_{\lambda}}{s - d_{\lambda} + 1} = O \left( \frac{\sigma^2 d_{\lambda}}{s} \right), \]

where \( d_{\lambda} = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \lambda} < d \), \( \{\lambda_i\} \) - eigenvalues of \( XX^T \)

\( d_{\lambda} \) - degrees of freedom (statistical dimension) of \( X \), given \( \lambda \)

Note: Under decaying eigenvalues, \( d_{\lambda} \ll d \)
Regularized volume sampling

**Naive idea:**

replace \( P(S) \sim \det(X_S X_S^T) \),

with \( P(S) \sim \det(X_S X_S^T + \lambda I) \)

**Problem:** Not clear how to sample efficiently

**Intuition:** No simple extension of the Cauchy-Binet formula

\[
\sum_{S: |S|=s} \det(X_S X_S^T) = \binom{n-d}{s-d} \det(XX^T)
\]

\[
\sum_{S: |S|=s} \det(X_S X_S^T + \lambda I) = ???
\]
Regularized volume sampling

Naive idea:

replace \( P(S) \sim \det(X_S X_S^\top) \),

with \( P(S) \sim \det(X_S X_S^\top + \lambda I) \)

Problem: Not clear how to sample efficiently

Intuition: No simple extension of the Cauchy-Binet formula

\[
\sum_{S: \ |S|=s} \det(X_S X_S^\top) = \left( \begin{array}{c} n - d \\ s - d \end{array} \right) \det(XX^\top)
\]

\[
\sum_{S: \ |S|=s} \det(X_S X_S^\top + \lambda I) = ???
\]
Reverse iterative sampling

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

Sample index $i \in S$

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

Repeat until desired size

$$
P(S_{-i} | S) \sim \frac{\det(X_{S_{-i}} X_{S_{-i}}^\top + \lambda I)}{\det(X_S X_S^\top + \lambda I)} = 1 - x_i^\top (X_S X_S^\top + \lambda I)^{-1} x_i$$

**Note:** not the same as $P(S) \sim \det(X_S X_S^\top + \lambda I)$ (unless $\lambda = 0$)
Reverse iterative sampling

Start with \( S = \{1..n\} \)

Sample index \( i \in S \)

Go to set \( S_{-i} = S - \{i\} \)

Repeat until desired size

\[
P(S_{-i}|S) \sim \frac{\det(X_{S_{-i}} X_{S_{-i}}^\top + \lambda I)}{\det(X_S X_S^\top + \lambda I)} = 1 - x_i^\top (X_S X_S^\top + \lambda I)^{-1} x_i
\]

Note: not the same as \( P(S) \sim \det(X_S X_S^\top + \lambda I) \) (unless \( \lambda = 0 \))
Reverse iterative sampling

\( \{1..n\} \) - size
- \( n \)

\( S \)
\( \downarrow \)
\( P(S_{-i}|S) \)
\( \downarrow \)
\( S = S_{-i} \)

Start with \( S = \{1..n\} \)

Sample index \( i \in S \)

Go to set \( S_{-i} = S - \{i\} \)

Repeat until desired size

\[
P(S_{-i}|S) \sim \frac{\det(X_{S_{-i}}X_{S_{-i}}^T + \lambda I)}{\det(X_SX_S^T + \lambda I)} = 1 - x_i^T(X_SX_S^T + \lambda I)^{-1}x_i
\]

**Note:** not the same as \( P(S) \sim \det(X_SX_S^T + \lambda I) \) (unless \( \lambda = 0 \))
Key Expectation Bound

Lemma

For \( \lambda \)-regularized volume sampling of set \( S \) of size \( s \)

\[
\mathbb{E}_S (X_S X_S^\top + \lambda I)^{-1} \leq \frac{n - d^\lambda + 1}{s - d^\lambda + 1} (X X^\top + \lambda I)^{-1}
\]

(for \( \lambda = 0 \), this was shown in [DW17])

Proof of main theorem

Bias-variance decomposition for fixed \( S \):

\[
\text{MSPE}(\hat{w}_\lambda^*(S) | S) = (\text{bias})^2 + \text{variance} \leq \frac{\sigma^2}{n} \text{tr}(X^\top (X_S X_S^\top + \lambda I)^{-1} X)
\]

Take expectation over \( S \) and apply the lemma.
Key Expectation Bound

Lemma

For $\lambda$-regularized volume sampling of set $S$ of size $s$

$$
\mathbb{E}_S (X_S X_S^\top + \lambda I)^{-1} \leq \frac{n - d_\lambda + 1}{s - d_\lambda + 1} (XX^\top + \lambda I)^{-1}
$$

(for $\lambda = 0$, this was shown in [DW17])

Proof of main theorem
Bias-variance decomposition for fixed $S$:

$$
\text{MSPE}(\hat{w}_\lambda^*(S) | S) = (\text{bias})^2 + \text{variance} \leq \frac{\sigma^2}{n} \text{tr}(X^\top (X_S X_S^\top + \lambda I)^{-1} X)
$$

Take expectation over $S$ and apply the lemma.
Volume sampling vs leverage score sampling

**Leverage score sampling:**
examples selected i.i.d. w.p. \( \sim x_i^T (XX^T)^{-1} x_i \).

How many labels needed for \( \text{MSPE}(\hat{w}) \leq \sigma^2 \)? (\( \sigma^2 \) - label noise)

1. **Volume sampling** \( 2d_\lambda \) labels
2. **Any i.i.d. sampling** \( \Omega(d_\lambda \ln d_\lambda) \) labels (lower bound)

Our volume sampling is as fast as computing exact leverage scores
Volume sampling vs leverage score sampling

Leverage score sampling:
examples selected i.i.d. w.p. $\sim x_i^T (XX^T)^{-1} x_i$.

How many labels needed for $\text{MSE}(\hat{w}) \leq \sigma^2$? ($\sigma^2$ - label noise)

1. Volume sampling  $2d_\lambda$ labels
2. Any i.i.d. sampling  $\Omega(d_\lambda \ln d_\lambda)$ labels (lower bound)

Our volume sampling is as fast as computing exact leverage scores
Volume sampling vs leverage score sampling

Leverage score sampling:
examples selected i.i.d. w.p. $\sim x_i^T (XX^T)^{-1} x_i$.

How many labels needed for $\text{MSPE}(\hat{w}) \leq \sigma^2$? ($\sigma^2$ - label noise)

1. Volume sampling $2d_\lambda$ labels
2. Any i.i.d. sampling $\Omega(d_\lambda \ln d_\lambda)$ labels (lower bound)

Our volume sampling is as fast as computing exact leverage scores.
**Volume sampling vs leverage score sampling**

**Leverage score sampling:**  
examples selected i.i.d. w.p. $\sim x_i^T (XX^T)^{-1} x_i$.

How many labels needed for $\text{MSPE}(\hat{w}) \leq \sigma^2$? ($\sigma^2$ - label noise)

1. **Volume sampling**  
   2$\lambda$ labels

2. **Any i.i.d. sampling**  
   $\Omega(d\lambda \ln d\lambda)$ labels (lower bound)

Our volume sampling is as fast as computing exact leverage scores
Volume sampling vs leverage scores on real data
Simple algorithm for volume sampling

\{1..n\} \quad \text{- size} \quad \text{Start with } S = \{1..n\}

\bullet \quad \text{- } n

\begin{align*}
\cdots & \quad \text{- } n-1 \\
S & \\
\downarrow \quad P(S_{-i}|S) & \quad \text{Sample index } i \in S \\
\cdots & \quad \text{- } s-1 \\
S = S_{-i} & \quad \text{Go to set } S_{-i} = S - \{i\} \\
\end{align*}

\text{Simple algorithm: } \text{Update distribution } P(S_{-i}|S) \text{ at every step}

\text{Runtime: } \underbrace{n-s}_{\text{steps}} \times \underbrace{O(n d)}_{\text{update}} = O(n^2 d)

\text{Problem: } \text{Quadratic dependence on } n
Faster algorithm via rejection sampling

Recall: \[ P(S_{-i}|S) \sim 1 - x_i^T (X_S X_S^T + \lambda I)^{-1} x_i \]

**Idea:** Rejection sampling from distribution \( P(S_{-i}|S) \)

1. Sample \( i \) uniformly from set \( S \),
2. Compute \( h_i = 1 - x_i^T (X_S X_S^T + \lambda I)^{-1} x_i \), one trial
3. With probability \( 1 - h_i \) reject and go back to 1.

**We show:** Number of trials per step is constant w.h.p.

**Runtime:** \[ \frac{n-s}{\text{steps}} \times \frac{O(1)}{\text{trials per step}} \times \frac{O(d^2)}{\text{compute } h_i} = O(nd^2) \]

**Result:** Linear dependence on \( n \)
Faster algorithm via rejection sampling

Recall: \[ P(S_{-i}|S) \sim 1 - x_i^T (X_S X_S^T + \lambda I)^{-1} x_i \]

Idea: Rejection sampling from distribution \( P(S_{-i}|S) \)

1. Sample \( i \) uniformly from set \( S \),
2. Compute \( h_i = 1 - x_i^T (X_S X_S^T + \lambda I)^{-1} x_i \),
3. With probability \( 1 - h_i \) reject and go back to 1. one trial

We show: Number of trials per step is constant \( \text{w.h.p.} \)

Runtime: \( \text{steps} \times \text{trials per step} \times \text{compute } h_i = O(n d^2) \)

Result: Linear dependence on \( n \)
Faster volume sampling on real data

*RegVol*: simple algorithm  
*FastRegVol*: rejection sampling

---

**cadata**

- **RegVol**
- **FastRegVol**

---

**MSD**

- **RegVol**
- **FastRegVol**

---

**cpusmall**

- **RegVol**
- **FastRegVol**

---

**abalone**

- **RegVol**
- **FastRegVol**
Conclusions

1. Dimension-free error bounds for subsampled regression

2. Optimal subsampling must be joint

3. Why pick volume sampling over leverage score sampling?
   3.1 Better error bounds for small sample sizes,
   3.2 Nearly the same computational efficiency.
Conclusions

1. Dimension-free error bounds for subsampled regression

2. Optimal subsampling must be joint

3. Why pick volume sampling over leverage score sampling?
   3.1 Better error bounds for small sample sizes,
   3.2 Nearly the same computational efficiency.
Conclusions

1. Dimension-free error bounds for subsampled regression

2. Optimal subsampling must be joint

3. Why pick volume sampling over leverage score sampling?
   3.1 Better error bounds for small sample sizes,
   3.2 Nearly the same computational efficiency.
Thank you