Unbiased estimates for linear regression via volume sampling

Michal Dereziński  Manfred K. Warmuth

**Linear regression**

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

\[
w^* = \arg\min_w L(w)
\]

**Expensive labels**

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

**Answer: one label**

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

**Volume sampling**

\[S \subseteq \{1..n\}\] chosen w.p. 
\[
\sim \text{squared volume of parallelepiped spanned by the } \{x_i : i \in S\}
\]

Distribution over all \(d\)-element subsets 
\[
P(S) = \frac{\det(X_S X_S^\top)}{Z}
\]

Normalization factor obtained via Cauchy-Binet formula: 
\[
Z = \sum_{S:|S|=d} \det(X_S X_S^\top) = \det(X X^\top)
\]

**Loss expectation formula**

**Theorem**

For a volume-sampled set \(S\) of size \(d\), 
\[
\mathbb{E}[L(w^*(S))] = (d + 1) \frac{1}{n} \mathbb{E}[L(w^*)]
\]

if \(X\) 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 \(S_{i+1} = S - \{i\}\)

Repeat until desired size

**Unbiased estimator for pseudo-inverse \(X^+\)**

Key trick: To each subset \(S\) assign a formula \(F(S)\) at 
\[
F(S) = \sum_{S_i \subseteq S} P(S_i|S) F(S_i).
\]

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

**Expectation formulas for \((X S)^+\)**

1. \(\mathbb{E}[(X S)^+] = X^+\) (unbiasedness)
2. \(\mathbb{E}[(X S X_S)^+ - 1] = \frac{n-d}{n-d-(d-1)} \frac{(X X^\top)^{-1}}{(X^\top)^2}\) (variance bound)

Corollary: 
\[
\mathbb{E}[w^*(S)] = \mathbb{E}[(X S)^+ y] = X^+ y = w^*
\]

**Averaging unbiased estimators**

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

loss bound
\[
\mathbb{E}[L(w^*(S))] \leq (1 + \epsilon) L(w^*) \iff \mathbb{E}[\|y(S) - E[w^*(S)]\|_2] \leq c L(w^*)
\]

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

if \(L(w^*) \leq \left(1 + \frac{c}{\epsilon}\right) L(w^*)\)

With size \(d\) volume sampling, we need \(d^2/\epsilon\) labels. Is \(d^2/\epsilon\) possible?

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