Speaker: Michal Derezinski

Joint work with Manfred Warmuth

We use volume sampling to obtain exact multiplicative regret bounds for linear regression. The learner is given a fixed set of n input vectors in $\mathbb{R}^d$ of a linear regression problem (aka fixed design). Each vector has a real hidden label. The goal is to approximately solve the linear least squares problem for all n labeled vectors while seeing only a small number of the labels.

In our most basic setup, the learner selects a random subset of d vectors from the fixed set of n vectors (without knowing any of the labels). The learner then is given the labels of the chosen subset of d vectors. We show that if the random subset is chosen based on volume sampling, then the linear least squares solution for the subset of d labeled vectors

- is an unbiased estimator of optimum solution on all n labeled vectors,
- and in expectation, the total square loss (on all n labeled vectors) of the solution found for the subset is by a factor of exactly $d+1$ larger than the minimum achievable total square loss (provided the vectors are in general position).

We next show how d/ϵ subsamples (of d points each) can be used to produce a solution with expected loss at most $1+\epsilon$ times the optimum. It is common to develop additive regret bounds for linear regression, ie. bound the expected loss of the algorithm minus the loss of the best. Instead we bound the multiplicative regret which is the additive regret divided by the optimum loss. Viewed this way, the above method based on d/ϵ subsamples of size d has multiplicative regret epsilon.

We compare our work to other approaches and give many open problems. Our results are elementary and we will give a flavor of the proof methods.

1 Intro

Given n labeled points $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, find w^* that minimizes square loss $\sum_i (x_i^T w - y_i)^2$, i.e.,

$$L(w) = \|Xw - y\|^2, \quad X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^d.$$

We can compute the exact solution using the pseudoinverse.

Our goal is to approximately minimize $L(w)$ without using all labels.

Solve subproblem (X_S, y_S) getting an approximate solution w_S^*.

Approach: choose S to get most informative labels. Instances with larger norm $\|x\|$ are more informative.

In volume sampling (Deshpande, Rademacher, Vempala, Wang), generalize the norm to sets of examples. The distribution over d-element subsets S is

$$\mathbb{P}(S) \propto \det(X_S^T X_S),$$

the squared volume of parallelopiped $P(x_i, x_j)$. d is the number of features. The normalization factor is

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

by the Cauchy-Binet formula. This is the volume spanned by the columns of the data matrix.

There is a sampling procedure based on SVD, matrix decomposition.

This is a special case of determinantal point processes.

Let $w^* = \mathbb{E}(w^*_S)$ (w^*_S is unbiased estimator of optimal). We will obtain expected loss $\mathbb{E}[L(w^*_S)]$ that is at most $dL(w^*)$ more than the optimal loss $L(w^*)$.

Note volume sampling doesn't see the labels! Linear regression is special: if there is outliers, the best solution w^* also pays.

Main result: For a volume-sampled d-element set S,

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

where $w^* = \mathbb{E}[w^*_S]$ if X is in general position.

Zeroing out everything in X except S and taking the pseudoinverse we get an unbiased estimator: If X is full rank,

$$X^+ = \mathbb{E}[(I_S X)^+],$$

Then the predictions are also unbiased

$$\mathbb{E}[w^*_S] = \mathbb{E}[(I_S X)^+ y_S] = X^+ y = w^* \tag{1}$$
$$\mathbb{E}[\widehat{y}] = Xw^* = y^*. \tag{2}$$

If X is not in general position, then we have an inequality

$$\mathbb{E}[L(w^*_S)] \leq (d+1)L(w^*).$$

Note $\mathbb{E}[L(w^*_S)]$ is sensitive to subsets S such that $\det(S_X) = 0$. Ex. consider $(X|y) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$ and $(X_\varepsilon|y) = \begin{pmatrix} 1 & 1+\varepsilon & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$.

Bias-variance decomposition gives

$$\begin{aligned}(d+1)L(w^*) &= \mathbb{E}[L(w^*_S)] = \mathbb{E}[\|\widehat{y} - y\|^2] &(3)\\ &= \mathbb{E}[\|\widehat{y} - y^*\|^2] + \|y^* - y\|^2 &(4)\\ &= \sum_{i=1}^n \mathbb{E}[(\widehat{y}_i - \mathbb{E}\widehat{y}_i)^2] + L(w^*) = \sum_{i=1}^n \mathrm{Var}[\widehat{y}_i] + L(w^*) &(5)\\ \sum_{i=1}^n \mathrm{Var}(\widehat{y}_i) &= (d+1)L(w^*) &(6)\end{aligned}$$

We can get arbitrarily close to 1 by taking many (T) samples of size d.

$$\mathbb{E}\left[L\left(\frac{1}{T}\sum_{t=1}^T w^*_{S_t}\right)\right] = \left(1 + \frac{d}{T}\right)L(w^*).$$

Note that we're calculating the w^* for each subset separately, rather than aggregating the samples.

Proof. By bias-variance decomposition,

$$= \sum_{i=1}^{n} \mathrm{Var}\left[\frac{1}{T}\sum_{t=1}^{T}\widehat{y}_i^{(t)}\right] + L(w^*). \tag{7}$$

Q: Why not just select subset S with maximal volume?

A: To protect against adversarial labeling. Example: $X = \begin{pmatrix} 1+\varepsilon \\ 1 \\ \vdots \end{pmatrix}$, $y = \begin{pmatrix} 0 \\ 1 \\ \vdots \end{pmatrix}$.

Any deterministic algorithm can be fooled.
Related work:

1. Leverage score sampling for linear regression (sample independently. We sample d examples non-independently which gives more informative samples.)
2. Hadamard transform and sketching for regression
3. Volume sampling for matrix approximation
4. Online regression regret bounds

What do additive bounds on average loss look like?

$$R(\widehat{w}_k) := \mathbb{E}[\overline{L}(\widehat{w}_k)] - \inf_w \overline{L}(w) \tag{8}$$

$$\overline{L}(w) := \frac{1}{n}\sum_{i=1}^{n} l_i(w) \tag{9}$$

$$R(\widehat{w}_k) = O\left(\frac{d \ln k}{k} Y^2\right), \qquad Y = \max_i |y_i| \tag{10}$$

This is typical bounds under uniform sampling. We show averaged volume-sampling weight vector achieves

$$R(\widehat{w}_k) = \frac{d^2}{k}\overline{L}(w^*), \qquad k \geq d. \tag{11}$$

where $k \geq d$.
Proof techniques:

- Geometry:
  - Base×height formula
    How to relate volume to linear regression? Prediction $\widehat{y}$ is projection of label vector onto column spam. BH formula gives ($\widetilde{X} = (X|y)$)

    $$\det(\widetilde{X}^T\widetilde{X}) = \det(X^T X) L(w^*).$$

  - Cauchy-Binet formula

  - New volume formula for leave-one-out loss:

- Expected pseudo-inverse:
  - determinant derivative formula
  - induction using Sherman-Morrison

Extended volume sampling: sample susbet $|S| = k \geq d$. Still $\mathbb{P}(S) \propto \det(X_S^T X_S)$. Normalization is $Z_k = \binom{n-d}{k-d} \det(X^T X)$.

Composition property: hierarchical volume sampling gives the same as one-shot volume sampling.

Pseudo-inverse formula still holds. Square inverse formula:

$$(X^T X)^{-1} = \frac{k-d+1}{n-d+1} \mathbb{E}_S[(X_S^T X_S)^{-1}].$$

Open problems: extend results to regularized regression, show $O(d/k)$ rather than $O(d^2/k)$ (likely achievable when using optimal predictor $w^*_{S_{1:T}}$), find minimax optimal sampling distribution.

(Adversary tries to labels. Adversary tries to maximize the payoff, the ratio between expected loss and best loss.)