Instructor: Adarsh Barik
Scribed by: Harshit Bisht
Lecture 7, COV878
September 3 and 19, 2025

LECTURE 7: BEYOND PAC: PRIMAL-DUAL WITNESS APPROACH

Consider a typical ML task: minimizing a parametrized function $f(w)$ subject to the constraint that the optimal parameters must be sparse. There can be multiple definitions of sparsity depending on the structure of the parameters, such as enforcing that only a small number of parameters are non-zero or if they are a matrix, enforcing the rank to be low.

Here we take the setting of optimizing a convex objective f parametrized by a vector of weights and constrained on the maximum number of non-zero/active parameters, which can also be expressed as capping the maximum L0 norm $\|w\|_0$.

$$w^* = argmin_w f(w) \text{ constrained on } \|w\|_0 \leq k$$

A similar real-world problem is noiseless compression sensing, where a we receive measurements $y = Xw^*$ generated from some unknown signal w^* through sensors represented by the matrix X. Here, $w^* \in \mathbb{R}^{n\times 1}$, $y \in \mathbb{R}^{m\times 1}$, and $X \in \mathbb{R}^{m\times n}$ with the number of measurements much smaller than the signal dimension ($m \ll n$). To deal with this data scarcity, we attempt to find the sparsest signal that could explain the current data. This problem can also be solved by the techniques discussed in the lecture.

$$w^* = argmin_w \|w\|_0 \text{ constrained on } y = Xw^*$$

1 Relaxing to L1 norm constraints

This is a difficult constraint to work with since the L0 norm is neither continuous/differentiable nor convex. Thus, we try to optimize in a relaxed setting in the hope that the optimal parameters for the relaxed problem will also be good solutions to our original problem.

The relaxation we look at is bounding the L1 norm instead $\|w\|_1 \leq R$ and solving the following problem for optimal $\hat{w}$. If we are able to show that $\hat{w}$ is very close/similar to the w^* from the original problem, the relaxation would be well justified.

$$\hat{w} = argmin_w \left[f(w) + \lambda \|w\|_1 \right]$$

Motivated from the original L0 problem, define the support of w as $supp(w) = \{i \mid w_i \neq 0\}$: the collection of indices of non-zero parameters. Similarly define "non-support" $\overline{supp}(w) = \{i \mid w_i = 0\}$. We plan to show that $\hat{w}$ and w^* are close in the following ways:

- $\|\hat{w} - w^*\|_2 \leq \epsilon$ for some positive ϵ
- $\|\hat{w}\|_0 \leq k$
- $supp(\hat{w}) = supp(w^*)$

- $sign(\hat{w}_i) = sign(w_i^*)\ \forall i \in supp(w^*)$

The final two conditions are equivalent to saying $sign(\hat{w}_i) = sign(w_i^*)\ \forall i$.

Properties of the L1 norm

Consider the L1 norm as the dual norm to the L∞ norm:

$$\|w\|_1 = \sum_i |w_i| = \max_{\|z\|_\infty \leq 1} \langle z, x \rangle$$

We can infer some coordinates for the optimal $z^* = argmax_{\|z\|_\infty \leq 1} \langle z, x \rangle$ based on coordinates of w:

$$z_i^* = \begin{cases} +1, & \text{if } w_i > 0 \\ -1, & \text{if } w_i < 0 \\ [-1, 1], & \text{if } w_i = 0 \end{cases}$$

Thus, if we conclude that $-1 < z_i^* < 1$, we can safely infer that the corresponding $w_i = 0$.

Useful matrix norm inequalities

Consider a matrix $A \in \mathbb{R}^{m \times k}$. We will define two extensions of a vector norm $\|.\|_p$ for A.

- The induced norm is the max scaling applied on a vector $|||A|||_p = \max_{\|z\|_p \leq 1} \|Az\|_p$. $|||A|||_2$ is also called the spectral norm of A.
- $\|A\|_p$ is flattening A to a vector in $\mathbb{R}^{mk \times 1}$ and applying $\|.\|_p$ to it. The L2 norm applied this way is also called the Frobenius norm $\|A\|_F = \|A\|_2$.

The following matrix norm inequalities will be useful in the rest of this lecture. They are stated without proof and can be easily verified.

1. $|||A|||_\infty = \max_{i=1}^m \|A_i\|_1$
2. For square A ($m = k$) $|||A|||_\infty \leq \sqrt{k} |||A|||_2$
3. $\|Ax\|_\infty \leq |||A|||_\infty \|x\|_\infty$
4. For some $B \in \mathbb{R}^{k \times k}$, we have $|||AB|||_\infty \leq k \|A\|_\infty \left|||B^T|||_\infty\right.$

We also define restricting a matrix/vector over some of its coordinates. Consider $a \in \mathbb{R}^{m \times 1}$, matrix $A \in \mathbb{R}^{m \times n}$, and a subset of coordinates $S \subseteq [n]$.

- We define a restricted on S $a_S \in \mathbb{R}^{|S| \times 1}$ such that coordinate a_i is included if and only if $i \in S$.
- We define restriction for matrix A_S to only include the i^{th} column of A if $i \in S$.
- Further, we define restricting both columns and rows of A as $A_{SS} = \left[A_{ij}\right]\ \forall i, j \in S$.

2 Primal-Dual witness approach

We can now express the L1 problem in a primal-dual setting as:

$$\min_{w\in\mathbb{R}^n} \max_{\|z\|_\infty\leq 1} f(w) + \lambda\langle z, w\rangle$$

We now wish to find the optimal primal-dual witness $(\hat{w}, \hat{z})$ for the above problem that also satisfy $|\hat{z}_i| < 1$ when $i \in \overline{supp}(w^*)$. This will ensure that $supp(\hat{w}) = supp(w^*)$.

Linear prediction setting

We will focus on the special case of a linear prediction problem defined below. The idea of the proof remains similar for other convex $f(w)$, but the finer details might need to be adapted to the setting.

For this setting, we make the following assumptions and solve for $\hat{w}$ as given below.

- Some sparse $w^* \in \mathbb{R}$ is fixed.
- Design matrix $X \in \mathbb{R}^{m\times n}$ is sampled with each X_{ij} an i.i.d. Rademacher random variable.
- $y \in \mathbb{R}^m$ is calculated as $y = Xw^*$.

$$\hat{w} = argmin_w \left[\frac{1}{2m}\|Xw - y\|_2^2 + \lambda\|w\|_1\right] = argmin_w \max_{\|z\|_\infty\leq 1}\left[\frac{1}{2m}\|Xw - y\|_2^2 + \lambda\langle z, w\rangle\right]$$

We note some properties for this loss function based on the properties of the setting and calculate its gradient and hessian.

$$f(w) = \frac{1}{2m}\|Xw - y\|_2^2 = \frac{1}{2m}\|X(w - w^*)\|_2^2$$

$$\nabla f(w) = \frac{1}{m}X^TX(w - w^*) \qquad \nabla^2 f(w) = \frac{1}{m}X^TX$$

Let $(\hat{w}, \hat{z})$ be the primal-dual solution obtained. By the stationarity (KKT) condition for our solution:

$$\nabla f(\hat{w}) + \lambda\hat{z} = \bar{0}$$

Construction

Let $S = \text{supp}(w^*)$ be the (unknown) support set. We construct a primal candidate $\hat{w}$ such that it is only supported on S $\hat{w} = (\hat{w}_S, \bar{0})$.

We need to show that the dual variable $\hat{z}$ is strictly feasible $\left\|\hat{z}_{\overline{S}}\right\|_\infty < 1$ on the non-support set $\overline{S} = [n] \setminus S$.

Splitting the KKT stationarity condition in support and non-support components:

$$\nabla f(\hat{w})_S + \lambda \hat{z}_S = \bar{0}$$
$$\nabla f(\hat{w})_{\overline{S}} + \lambda \hat{z}_{\overline{S}} = \bar{0}$$

The gradient components for $\hat{w}$ in support and non-support components:

$$\nabla f(\hat{w})_S = \frac{1}{m} X_S^\top X_S (\hat{w}_S - w_S^*)$$
$$\nabla f(\hat{w})_{\overline{S}} = \frac{1}{m} X_{\overline{S}}^\top X_S (\hat{w}_S - w_S^*)$$

Support coordinates

Combining the above equations in the support coordinates, we get:

$$\hat{w}_S - w_S^* = -\lambda \left(\frac{1}{m} X_S^\top X_S \right)^{-1} \hat{z}_S$$

The above equation assumes that $\frac{1}{m} X_S^\top X_S$ is invertible. Consider w_S which is the solution to the following problem:

$$w_S = argmin_{w_S \in \mathbb{R}^{|S|}} \left[f((w_S, \bar{0})) + \lambda \left\| w_S \right\|_1 \right]$$

We know that the solution for w_S will be unique ($\hat{w}_S = w_S$) if $\frac{1}{m} X_S^\top X_S$ is invertible or $\Lambda_{\min} \left(\frac{1}{m} X_S^\top X_S \right) > 0$. Notice that X is a random matrix consisting of rademacher random variables and $X_S^T X_S$ can be expressed as a sum of k random matrices corresponding to each column in X_S.

Using the Matrix Chernoff bound for random matrices, we can assert:

$$\Pr \left[\Lambda_{\min} \left(\frac{1}{m} X_S^\top X_S \right) \leq \frac{1}{2} \right] \leq k \exp \left(\frac{-m}{8k} \right)$$

As long as $m = \Omega(k \log k)$, this probability can be made vanishingly assuring us that on the support part of the primal solution is unique.

Non-support coordinates: bounding the dual variable

From the non-support part, we solve for $\hat{z}_{\overline{S}}$

$$\hat{z}_{\overline{S}} = -\frac{1}{\lambda n} X_{\overline{S}}^\top X_S (\hat{w}_S - w_S^*) = \frac{1}{m} X_{\overline{S}}^\top X_S \left(\frac{1}{m} X_S^\top X_S \right)^{-1} \hat{z}_S$$

We need to show $\left\| \hat{z}_{\overline{S}} \right\|_\infty < 1$. Using norm inequalities:

$$\left\| \hat{z}_{\overline{S}} \right\|_\infty \leq \left\| \frac{1}{m} X_{\overline{S}}^\top X_S \left(\frac{1}{m} X_S^\top X_S \right)^{-1} \right\|_\infty \left\| \hat{z}_S \right\|_\infty$$

where $\|\hat{z}_S\|_\infty \leq 1$.

Furthermore, using the norm inequality:

$$\begin{aligned}\|\hat{z}_{\overline{S}}\|_\infty &\leq |S| \left\|\frac{1}{m} X_{\overline{S}}^\top X_S\right\|_\infty \left|\left|\left| \left(\frac{1}{m} X_S^\top X_S\right)^{-1} \right|\right|\right|_\infty \\ &\leq |S|^{\frac{3}{2}} \left\|\frac{1}{m} X_{\overline{S}}^\top X_S\right\|_\infty \left|\left|\left| \left(\frac{1}{m} X_S^\top X_S\right)^{-1} \right|\right|\right|_2 \\ &\leq 2|S|^{\frac{3}{2}} \left\|\frac{1}{m} X_{\overline{S}}^\top X_S\right\|_\infty\end{aligned}$$

This condition holds if

$$\left|\left|\left|\frac{1}{m} X_{\overline{S}}^\top X_S\right|\right|\right|_\infty < \frac{1}{2|S|^{3/2}}$$

We'll show $\left(\frac{1}{m} X_{\overline{S}}^\top X_S\right)_{ij}$ is small with high probability.

$$\left(\frac{1}{m} X_{\overline{S}}^\top X_S\right)_{ij} = \frac{1}{m} \sum_{l=1}^{n} x_{li} z_{lj}$$

Note that $i \neq j$ since $i \in \overline{S}, j \in S$ and x_{li}, z_{lj} are both rademacher random variables, thus $\mathbb{E}[x_{li} z_{lj}] = 0$.

Using Hoeffding's inequality and a union bound:

$$\Pr\left[\left\|\frac{1}{m} X_{\overline{S}}^\top X_S\right\|_\infty \geq \frac{1}{2|S|^{3/2}}\right] \leq 2|\overline{S}||S| \exp\left(\frac{-m}{8|S|^3}\right)$$

The probability gets vanishingly small as long as $m = \Omega(k^3 \log n)$.

Finally, by using some norm inequalities, we can say with high probability that:

$$\|\hat{w}_S - w_S^*\|_\infty = \lambda \left\|\left(\frac{1}{m} X_S^\top X_S\right)^{-1} \hat{z}_S\right\|_\infty \leq 2\lambda|S|^3 \quad \text{with high probability}$$

Lemma If $a, b \in \mathbb{R}^n$ satisfy $\|a - b\|_\infty \leq \varepsilon$ and $\min_i |b_i| > 2\varepsilon$ then $\forall i \in [n]$, we have that $sign(a_i) = sign(b_i)$.

From the above lemma, it is clear that $\forall i \in supp(w^*)$ we have $sign(\hat{w}_i) = sign(w^*)$ as long as $\min_{i \in S} |w_i^*| \geq 4\lambda|S|^3$ (minimum weight requirement). Hence, we have justified that the solution obtained in the relaxed L1 setting for our original problem is close with high probability to the optimal solution based on L0 norm w^*, as long as certain bounds between m, n, and k hold.