10-716: Advanced Machine Learning **Spring 2019**

Lecture 14: March 7

Lecturer: Pradeep Ravikumar *Scribes: David E. Bernal, Stefani Karp, Faisal Baqai*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

In the previous lectures we have covered sparse linear regression models and analyzed them. In this lecture other kinds of structure are going to be considered. We begin this lecture by mentioning our notation.

14.1 Preliminaries

We first recall some definitions:

Given a distribution $\mathbb{P}$, we have an estimator function $\theta(\mathbb{P}) \in \mathbb{R}^d$. We also have samples $Z_i \sim \mathbb{P}, i = 1, \cdots, n$. Our goal is to satisfy $\theta^* = \theta(\mathbb{P})$, where θ^* is the true parameter, in the high dimensional case ($d > n$).

In order to analyze different methods we require two components:

- A loss function $L_n(\theta, Z_i^n)$.
- Certain structure requirements.

For the sparse linear regression case, $y_i = \boldsymbol{X}_i\theta^* + w_i$, with $w_i \sim \mathcal{N}(0, \sigma^2)$, the samples are $Z_i = (\boldsymbol{X}_i, y_i)$. In that case our loss function was $L_n(\theta, Z_i^n) = \frac{1}{2n} \|y - \boldsymbol{X}\theta\|_2^2$ and the structure requirement was sparsity of the parameter $||\theta^*||_0 = s << d$.

14.2 Non-overlapping Group Lasso

We can generalize the loss function and the structure requirements for other high dimensional cases, e.g. logistic models or generalized linear models.

14.2.1 Examples

- $$\mathbb{P}(y_i|\boldsymbol{X}_i,\theta) = \frac{exp(y_i\langle \boldsymbol{X}_i,\theta\rangle - A(\boldsymbol{X}_i,\theta))}{c_\epsilon} \tag{14.1}$$
- For any graphical model we have

$$\mathbb{P}(z|\theta) = exp(T(z)^T\theta - A(\theta)) \tag{14.2}$$

A reasonable loss function for all cases is the negative of the log-likelihood of the model parameters.

Apart from the parameter sparsity s we can use other measures of structure, for example the group sparsity.

Definition 14.1 (Group sparsity) *Given a d-dimensional parameter*

$$\theta = (\underbrace{\theta_1,\theta_2}_{g_1},\underbrace{\theta_3,\theta_4}_{g_2},\underbrace{\cdots}_{\cdots},\underbrace{\cdots}_{g_j},\underbrace{\cdots}_{\cdots},\underbrace{\theta_{d-2},\theta_{d-1},\theta_d}_{g_m}) \tag{14.3}$$

we partition it in m groups of parameters $g_j, j \in [m]$ satisfying the following properties:

- $g_i \subseteq \{1,\cdots,d\}$
- $\bigcup_i g_i = \{1,\cdots,d\}$
- $g_i \cap g_j = \emptyset, i \neq j$

Using the group sequence $G := \{g_j\}_{j=1}^m$, we define the group sparsity as

$$\|\theta\|_{G,0} = |\{j \in [m] | \exists k \in g_j : (\theta_{g_j})_k \neq 0\}| \tag{14.4}$$

Notice that the group sparsity satisfies the condition of a norm where two equal elements should result in the norm being equal to zero. This allows us to write the following equality.

$$\|\theta\|_{G,0} = |\{j \in [m]| \left\|\theta_{g_j}\right\|_0 \neq 0\}| \tag{14.5}$$
$$= (\|\theta_{g_1}\|, \|\theta_{g_2}\|, \cdots, \|\theta_{g_m}\|)_0 \tag{14.6}$$

We also notice that a natural relaxation of the group norm is the following

$$\|\theta\|_{G,1} = \sum_{j=1}^m \left|\theta_{g_j}\right| \tag{14.7}$$

In the following example, we will see how does the non-overlapping assumption on the groups and its sparsity determines the sparsity of the original parameters.

14.2.2 Example 1

Consider a parameter $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ and the following disjoint groups $g_1 = (\theta_1, \theta_2)$ and $g_2 = (\theta_3, \theta_4)$. A $\ell 2$ relaxation of the group sparsity would yield

$$\|\theta\|_{G,2} = \left\|\theta_{\{1,2\}}\right\|_2 + \left\|\theta_{\{3,4\}}\right\|_2 \tag{14.8}$$

$$= \sqrt{\theta_1^2 + \theta_2^2} + \sqrt{\theta_3^2 + \theta_4^2} \tag{14.9}$$

Notice that the overall sparsity of θ is determined by the group sparsity above, since the groups g_1 and g_2 are disjoint. From the previous equation you can also see that there is a penalty induced by the groups, i.e. if one element in a group is not zero, regardless if the other values in the group are zero, the overall group is penalized.

An example of this feature segmentation is the brain, where certain subsets of the neurons are grouped by function and/or location. Using group sparsity, we would only call a brain region inactive if all the neurons in the corresponding group were inactive.

14.3 Overlapping group Lasso

What if groups are not disjoint? Using our brain example, consider neurons that fall under multiple groups. Then the penalty on a group being inactive no longer enforces the idea of an entire group being silent together vs. active together

14.3.1 Example 2

Consider a parameter $\theta = (\theta_1, \theta_2, \theta_3)$ and the following disjoint groups $g_1 = (\theta_1, \theta_2)$ and $g_2 = (\theta_1, \theta_3)$. A L^2 relaxation of the group sparsity would yield

$$\|\theta\|_{G,2} = \left\|\theta_{\{1,2\}}\right\|_2 + \left\|\theta_{\{1,3\}}\right\|_2 \tag{14.10}$$

$$= \sqrt{\theta_1^2 + \theta_2^2} + \sqrt{\theta_1^2 + \theta_3^2} \tag{14.11}$$

Notice that now, the groups g_1 and g_2 are overlapping. Assuming that θ_2 is non-zero, and θ_3 is zero, we would have the following case. g_1 would be penalized because of θ_2, encouraging θ_1 to be non-zero while g_2 is not penalized by θ_3, encouraging θ_1 to be zero.

This small example motivates the sparsity study not directly on the groups, but on the union of their complements. This results in the following definition.

Definition 14.2 (Overlapping Group sparsity) *We define the overlapping group sparsity as*

$$\|\theta\|_G^{(0)} = \inf_{\substack{\{\beta_j\} \\ s.t.\ \theta = \sum_{j=1}^m \beta_j}} \sum_{j=1}^m \left\|\beta_{g_j}\right\|_0 \tag{14.12}$$

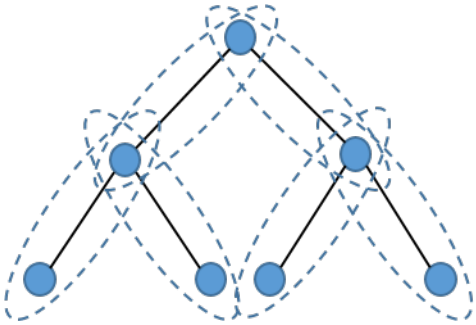

Figure 14.1: Parameter with a tree-structure, where the groups can be made along the edges of the tree (dashed lines).

Notice that the previous definition would only encourage the union of the groups to be active.

An example of parameters with overlapping group structures are those defined using a tree structure, as in Figure 14.1. In this example, if we consider the groups to be defined by the tree edges, there is an overlap in the parent nodes of each branch.

14.3.2 Matrix structured parameters and the general case

Another example of structure embedded in high dimensions would be a parameter defined as a matrix, e.g. $\Theta \in \mathbb{R}^{d_1 \times d_2}$. We could vectorize this parameter and use the techniques previously mentioned, but the fact that we already have an underlying structure can be used.

Using Singular Value Decomposition (SVD) we can express our matrix as

$$\Theta = U \Sigma V^T \tag{14.13}$$

where $U \in \mathbb{R}^{d_1 \times d'}$, $V \in \mathbb{R}^{d_2 \times d'}$ and $\Sigma \in \mathbb{R}^{d' \times d'}$. In particular, we have that both U and V being orthogonal, i.e. $UU^T = I_{d_1}, VV^T = I_{d_2}$, and that Σ is a diagonal matrix.

The elements in the diagonal of Σ, $\sigma_0 = \{\sigma_1, \sigma_2, \cdots, \sigma_{d'}\}$, are known as the singular values of Θ and are all non-negative and decreasing, i.e. $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{d'} \geq 0$.

Using the single values $\sigma_i, i \in [d']$, we can express the rank of the original matrix Θ as follows

$$rank(\Theta) = \{j \in [d'] | \sigma_j > 0\} = \|\sigma_0\|_0 \tag{14.14}$$

The rank of a parameter matrix can be seen as a parsimony notion for this particular structure. Using the spectral decomposition of the matrix, i.e. considering the vector of singular values or eigenvalues, we can analyze the matrix by analyzing the sparsity of this vector.

As above, a natural relaxation of the rank of the matrix, seen as the L^0 (pseudo)norm of the SV vector, is the L^1 if the same vector, better known as the nuclear norm of the matrix $|||\Theta|||_* = |\sigma| = \sum_{j=1}^{d'} \sigma_j(\Theta)$.

The analysis of sparsity of a matrix is not restricted to the rank or its nuclear matrix, but we can use the tools from group sparsity to derive other norms.

Using the losses and the structure information, now we can derive our estimate for the parameter in the general case as follows

$$\hat{\theta} \in \arg\min_{\theta} L_n(\theta, Z_i^n) + \lambda_n R(\theta) \tag{14.15}$$

where $R(\theta)$ is the regularization as a function of the parameters. Regularization was first added for stability of the methods, but now it can also give structure to our estimates. This avoids the curse of dimensionality, because it forces parameters to lie in a union of lower dimensional sub-spaces.

The next question that we would like to answer is, how far is our estimate, $\hat{\theta}$, from the actual parameter, θ^*? In order to answer this question first we need to answer, what is θ^*?

We can use the following definition

$$\theta^* \in \arg\min_{\theta} \mathbb{E} L_n(\theta, Z_i^n) \tag{14.16}$$

which corresponds to the M-estimation of the parameter, given that it is **minimizing** the expectation of the loss.

With the definition of θ^*, now we can define bounds on $\left\|\hat{\theta} - \theta^*\right\|$.

$$\left\|\hat{\theta} - \theta^*\right\|_p \leq \frac{1}{k} s(\theta^*)\lambda + \epsilon \tag{14.17}$$

where the norm p is arbitrary, $s(\theta^*)$ is a measure of sparsity for the parameter, e.g. element sparsity for sparse linear models, or rank for matrices, and the error ϵ are bounded with high probability. The inequality above is non-asymptotic. For details, see [negahban]

14.4 Lower Bounds

We now move to *lower* bounds - i.e., over *all* possible estimators $\hat{\theta}$, what's the best that we could possibly do?

14.4.1 Preliminaries for Main Theorem

14.4.1.1 Goal

We have some distribution $\mathbb{P}$, a general form for the state of nature, and we have some parameter θ^*. Think of it as some function of the distribution $\mathbb{P}$ (e.g., a mean). The goal is to estimate θ^* given some samples Z drawn from the distribution $\mathbb{P}$. Formally:

$$\theta^* = \theta(\mathbb{P}) \in \Theta \subseteq \mathbb{R}^d$$

$$Z_i \sim \mathbb{P} \qquad i = 1, \ldots, n \qquad Z_i \in \mathcal{Z}$$

We use the samples (our "imperfect picture" of $\mathbb{P}$) to *try* to compute $\theta(\mathbb{P})$. But we can only hope to be *close* to θ^*. There's no hope *in general* to obtain *exactly* θ^*. We can view $\hat{\theta}$ as a function that takes n samples as input:

$$\hat{\theta}(Z_i^n) \in \mathbb{R}^d$$
$$\hat{\theta} : \mathcal{Z}^n \to \mathbb{R}^d$$

14.4.1.2 Characterizing Complexity

Imagine some metric $\rho : \Theta \times \Theta \to \mathbb{R}_+$ (some notion of distance between parameters).

We want to know how far $\hat{\theta}$ is from θ^*, where we take an expectation over the samples and look at the minimax risk:

$$\underbrace{\inf_{\hat{\theta}} \sup_{\mathbb{P}} \mathbb{E}_{Z_1^n} \rho(\hat{\theta}, \theta(\mathbb{P}))}_{M_n(\theta(\mathcal{P});\rho)}$$

To make this more general, think of:

$$M_n(\theta(\mathcal{P}); \Phi \circ \rho)$$

where Φ is an increasing function - e.g., $\Phi(t) = t^2$. We use composition to make it clear that ρ *must* be a metric, and Φ is something else (any increasing function).

Note: This is a different kind of complexity than computational complexity (the latter is largely based on reductions from an instance of one class of problems to an instance of another class of problems) Instead, what's the best I can do with respect to distance from the truth, over all possible algorithms (not bounded by computation or storage)? It still involves a reduction, but that reduction is reasonably tight.

We start with an entire set of parameters (typically $\mathbb{R}^d$), and we reduce our problem to a simpler problem with a *finite* set of parameters, from which we pick one. We'll reduce our bound on minimax risk to the error of this selection (which can be seen as a hypothesis test).

Definition 14.3 (2δ-separated set) *A set $\{\theta_1, \ldots, \theta_M\}$ s.t. $\rho(\theta_j, \theta_k) \geq 2\delta \quad \forall j, k \in [M], j \neq k$.*

Imagine we have a finite subset of parameters $\{\theta_1, \ldots, \theta_M\} \subseteq \theta(\mathcal{P})$ forming a 2δ-separated set. (This is known as a packing set.)

We pick one of these M parameters at random - i.e., $J \sim \text{UNIF}\{1, \ldots, M\}$. We then draw a sample from some distribution $\mathbb{P}_{\theta_J} \in \mathcal{P}$ (i.e., a distribution s.t. $\theta(\mathbb{P}_{\theta_J}) = \theta_J$): $Z \sim \mathbb{P}_{\theta_J}$.

14.4.1.3 The Hypothesis Testing Problem

We now define a hypothesis test ψ for this problem: $\psi : \mathcal{Z} \to [M]$ (goal: use our sample Z to determine the value of J).

The probability this test ψ makes an error is defined as:

$$\mathbb{Q}[\psi(Z) \neq J] \tag{14.18}$$

where $\mathbb{Q} \equiv$ joint dist of (J, Z) from the random selection processes.

And the lowest probability of error among all hypothesis tests is:

$$\inf_{\psi} \mathbb{Q}[\psi(Z) \neq J] \tag{14.19}$$

We'll use 14.19 to lower bound the minimax risk.

What does this buy us? We typically then show that this lower bound (14.19) is at least a constant (e.g., 1/2) for sufficiently small δ, which we'll discuss much further in future lectures.

14.4.2 Main Theorem

Note: In class, the theorem was introduced with Z_1^n, though the bulk of the proof used Z instead of Z_1^n. We choose to use Z instead of Z_1^n in both the theorem and proof for consistency, noting that there is no explicit dependence on n at this point and thus Z can be thought of as an arbitrary random sample in full generality.

Theorem 14.4

$$\inf_{\hat{\theta}} \sup_{\mathbb{P}} \mathbb{E}_{Z\sim\mathbb{P}_\theta}[\Phi \circ \rho(\hat{\theta}, \theta(\mathbb{P}))] \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}[\psi(Z) \neq J]$$
$$\textit{for any } \{\theta_1, \ldots, \theta_M\} \subseteq \theta(\mathcal{P}) \textit{ s.t. } \rho(\theta_j, \theta_k) \geq 2\delta \quad j \neq k$$

Proof:

$$\sup_{\mathbb{P}} \mathbb{E}_{Z\sim\mathbb{P}_\theta}[\Phi \circ \rho(\hat{\theta}, \theta(\mathbb{P}))] \geq \mathbb{E}_{Z\sim\mathbb{P}_{\theta_j}}[\Phi \circ \rho(\hat{\theta}, \theta_j)] \qquad \forall j \in [M] \qquad \text{definition of sup}$$

$$\begin{aligned}
\implies \sup_{\mathbb{P}} \mathbb{E}_{Z\sim\mathbb{P}_\theta}[\Phi \circ \rho(\hat{\theta}, \theta(\mathbb{P}))] &\geq \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}_{Z\sim\mathbb{P}_{\theta_j}}[\Phi \circ \rho(\hat{\theta}, \theta_j)] \\
&= \sum_{j=1}^{M} \mathbb{E}_{Z\sim\mathbb{P}_{\theta_J}}[\Phi \circ \rho(\hat{\theta}, \theta_J) | J = j] P[J = j] \\
&= \mathbb{E}_J \mathbb{E}_{Z\sim\mathbb{P}_{\theta_J}}[\Phi \circ \rho(\hat{\theta}, \theta_J)] \\
&= \mathbb{E}_{(J,Z)\sim\mathbb{Q}}[\Phi \circ \rho(\hat{\theta}, \theta_J)]
\end{aligned} \tag{14.20}$$

Suppose $\rho(\hat{\theta}, \theta_{j^*}) < \delta$. $\forall k \neq j^*$, we have:

$$\begin{aligned}
\rho(\hat{\theta}, \theta_k) &\geq \rho(\theta_{j^*}, \theta_k) - \rho(\hat{\theta}, \theta_{j^*}) && \text{by triangle inequality} \\
&> 2\delta - \delta && \text{by assumption} \\
&= \delta
\end{aligned}$$

Define the hypothesis test $\psi(Z) = \inf_{j \in [M]} \rho(\hat{\theta}, \theta_j)$. Then for a fixed estimator $\hat{\theta}$, we have:

$$\rho(\hat{\theta}, \theta_J) < \delta \implies \psi(Z) = J$$

$$\psi(Z) \neq J \implies \rho(\hat{\theta}, \theta_J) \geq \delta$$

$$\begin{aligned}
\mathbb{Q}[\psi(Z) \neq J] &\leq \mathbb{Q}[\rho(\hat{\theta}, \theta_J) \geq \delta] \\
&\leq \mathbb{Q}[\Phi \circ \rho(\hat{\theta}, \theta_J) \geq \Phi(\delta)] && \text{holds for increasing } \Phi \\
&\leq \frac{\mathbb{E}_{(J,Z)\sim\mathbb{Q}}[\Phi \circ \rho(\hat{\theta}, \theta_J)]}{\Phi(\delta)} && \text{Markov's inequality} \\
\implies \Phi(\delta)\mathbb{Q}[\psi(Z) \neq J] &\leq \mathbb{E}_{(J,Z)\sim\mathbb{Q}}[\Phi \circ \rho(\hat{\theta}, \theta_J)] && (14.21)
\end{aligned}$$

Combining 14.20 and 14.21, we obtain:

$$\sup_{\mathbb{P}} \mathbb{E}_{Z\sim\mathbb{P}_\theta}[\Phi \circ \rho(\hat{\theta}, \theta(\mathbb{P}))] \geq \Phi(\delta)\mathbb{Q}[\psi(Z) \neq J] \tag{14.22}$$

Since 14.22 holds for any fixed estimator $\hat{\theta}$, we can take an inf over $\hat{\theta}$. Additionally, although we gave $\psi(Z)$ a specific structure in terms of $\hat{\theta}$ to prove 14.22, we note that taking an inf over a larger set (i.e., all hypothesis tests mapping $\mathcal{Z} \to [M]$) cannot increase the value of the RHS. Thus, we have:

$$\inf_{\hat{\theta}} \sup_{\mathbb{P}} \mathbb{E}_{Z\sim\mathbb{P}_\theta}[\Phi \circ \rho(\hat{\theta}, \theta(\mathbb{P}))] \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}[\psi(Z) \neq J] \tag{14.23}$$

which completes the proof. (which is based on [wainright]) ■

References

[wainwright] M. Wainwright, "High Dimensional Statistics," *Prerelease*, 2019

[negahban] S. Negahban, B. Yu, P. Ravikumar, M. Wainwright, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers,", *Advances in Neural Information Processing Systems*, 2009