Homework 1 Solutions

Akshay Krishnamurthy

August 28, 2017

1. **Linear Regression.** We are interested in bounding $\mathbb{E}\|\hat{\beta} - \beta\|_2^2$ and we know that $\hat{\beta}$ is the empirical risk minimizer, or the “Ordinary Least Squares” solution. Let us set up some notation: Let $X = \in \mathbb{R}^{d \times n}$ with columns being the data points, $Y \in \mathbb{R}^n$ with the corresponding regression targets, and $\epsilon \in \mathbb{R}^n$ with the corresponding noise variables. This makes $\Sigma = \frac{1}{n}XX^T \in \mathbb{R}^{d \times d}$. We know that 

$$\hat{\beta} = (n\Sigma)^{-1}XY.$$ 

On the other hand, owing to the definition of $\beta^*$, we get $(Y - \epsilon) = X^T\beta^*$, so we can write $\beta^*$ as 

$$\beta^* = (n\Sigma)^{-1}X(Y - \epsilon).$$

Combining these two, we get:

$$\mathbb{E}\|\hat{\beta} - \beta^*\|_2^2 = \mathbb{E}\|(n\Sigma)^{-1}X\epsilon\|_2^2 = \mathbb{E}\epsilon^TXX^T(n\Sigma)^{-2}X\epsilon$$

$$\leq \frac{1}{nk}\mathbb{E}\epsilon^TXX^T(n\Sigma)^{-1}X\epsilon = \frac{1}{nk}\sum_{i=1}^d \epsilon_i^2 = \frac{d\sigma^2}{nk}$$

Inequality (a) is an application of Holder’s inequality for matrices after applying trace rotation and using positive semidefiniteness.

$$\mathbb{E}\epsilon^TXX^T(n\Sigma)^{-2}X\epsilon = \mathbb{E}\text{tr} \left(XX^T(n\Sigma)^{-1}\right)$$

$$\leq \mathbb{E}\text{tr} \left(XX^T(n\Sigma)^{-1}\right) \|(n\Sigma)^{-1}\|_2$$

$$\leq \frac{1}{nk}\mathbb{E}\text{tr} \left(\epsilon^TXX^T(n\Sigma)^{-1}X\epsilon\right)$$

Equality (b) uses the fact that while $XX^T(n\Sigma)^{-1}X$ is a $n \times n$ matrix, since $(n\Sigma)^{-1}XX^T = I_d$, it is actually a projection onto a $d$ dimensional subspace (i.e. it has $d$ non-zero eigenvalues, all of which have value exactly 1). Since $\epsilon$ is a spherically symmetric gaussian vector, due to rotational invariance, it is equivalent to just project onto the first $d$ coordinates (or think of replacing the $n \times n$ matrix with a diagonal matrix where the first $d$ diagonal entries are 1 and every other entry is 0). Thus we just need to evaluate the variance of the first $d$ noise variables, which is exactly $d\sigma^2$.

2. **PAC Learning.**

(a) The idea for the algorithm is to iteratively append predicates to the end of the decision list, and at the same time filter the dataset to contain only items that do not satisfy any of the existing predicates. In more detail, at every iteration, we have a current decision list and a sample $S = \{(x, y)\}$ such that for every $x \in S$, none of the if-statements in the current decision list apply (i.e. $x$ falls off the decision list). With this subsample, we search for any single rule of the form, “If $b_i$ then $\ell_i$,” that makes no mistakes but applies to at least one example in the subsample. If no such consistent rule exists, we halt and fail. Otherwise, we append this rule to the end of the list, further filter the dataset and repeat. If all remaining examples have the same label, we append, “else $\ell$” to the end of the decision list and terminate.
3. Robust mean estimation.

(a) Let \( Z = f(X_1, \ldots, X_n) \) and define \( \Delta_i = \mathbb{E}[Z|x_i^1] - \mathbb{E}[Z|x_i^{i-1}] \). Since terms telescope, we get

\[
Z - \mathbb{E}Z = \sum_{i=1}^{n} \mathbb{E}[Z|x_i^1] - \mathbb{E}[Z|x_i^{i-1}] = \sum_{i=1}^{n} \Delta_i,
\]

And hence,

\[
\text{Var}(Z) = \mathbb{E}\left[\left(\sum_{i=1}^{n} \Delta_i\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}[\Delta_i^2] + 2 \sum_{i>j} \mathbb{E}[\Delta_i\Delta_j]
\]

For these cross terms, observe that since \( i > j \),

\[
\mathbb{E}\Delta_i\Delta_j = \mathbb{E}\mathbb{E}[\Delta_i\Delta_j|x_i^1] = \mathbb{E}\Delta_j\mathbb{E}[\Delta_i|x_i^1] = 0.
\]

Or in other words, the increments are conditionally centered. Thus we get

\[
\text{Var}(Z) = \sum_{i=1}^{n} \mathbb{E}[\Delta_i^2].
\]

Now we use independence to analyze the increments. The basic step here is to use Jensen’s inequality to pull the expectation over the variables \( X_{i+1}, \ldots, X_n \) outside of the square.

\[
\mathbb{E}[\Delta_i^2] = \mathbb{E}_{X_i^1} \left( \mathbb{E}[Z|x_i^1] - \mathbb{E}[Z|x_i^{i-1}] \right)^2 = \mathbb{E}_{X_i^1} \left( \mathbb{E}_{X_{i+1}^n} [Z - \mathbb{E}_X Z] \right)^2 \leq \mathbb{E}_{X_i^1} \left( Z - \mathbb{E}_X Z \right)^2 = \mathbb{E} \text{Var}(Z|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).
\]

In this display, instead of conditioning, I’m showing what variables we are taking expectation to in the subscript of the expectation operator. The inequality is Jensen’s inequality, \((\mathbb{E}X)^2 \leq \mathbb{E}(X^2)\).
(b) Here we apply Chebyshev’s inequality

\[ P(|X - \mu| \geq \epsilon) = P(|X - \mu|^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[|X - \mu|^2]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \]

The last step here follows from the additivity of the variance for independent random variables, or more formally

\[ \mathbb{E}[|X - \mu|^2] = \frac{1}{n^2} \sum_{i,j} \mathbb{E}(X_i - \mu)(X_j - \mu) = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}(X_i - \mu)^2 = \frac{\sigma^2}{n}. \]

Finally, setting the right hand side to be at most \( \delta \), we can set \( \epsilon = \sqrt{\frac{\sigma^2}{n\delta}} \) which proves the statement.

(c) For each of the \( k \) groups, we know by Chebyshev’s inequality that

\[ P(|\mu_j - \mu| \geq \epsilon) \leq \frac{\sigma^2}{m\epsilon^2}. \]

Now let \( b_j = 1\{|\mu_j - \mu| \leq \epsilon\} \) be a binary random variable. By Hoeffding’s inequality, with probability at least \( 1 - \delta \) we know that

\[ \frac{1}{k} \sum_{j=1}^{k} b_j - \frac{1}{k} \sum_{j=1}^{k} b_j \leq \sqrt{\frac{\log(1/\delta)}{2k}}. \]

We also know that

\[ \mathbb{E}b_j = P(|\mu_j - \mu| \leq \epsilon) \geq 1 - \frac{\sigma^2}{m\epsilon^2}. \]

Finally, observe that if \( \sum_{j=1}^{k} b_j \geq k/2 \) then at least half of the numbers in the median computation are within \( \epsilon \) of \( \mu \), which implies that \( |\hat{\mu} - \mu| \leq \epsilon \). The application of Hoeffding’s inequality, the lower bound on the mean, and this final fact, imply that for an \( \epsilon \)-accurate estimate, it is sufficient for

\[ \left( 1 - \frac{\sigma^2}{m\epsilon^2} \right) - \sqrt{\frac{\log(1/\delta)}{2k}} \geq \frac{1}{2} \]

To wrap up we just need to find values for \( k, m, \epsilon \) that let us lower bound the right hand side. Setting \( k = \lceil 8 \log(1/\delta) \rceil \), \( m = \lfloor n/k \rfloor \), and \( \epsilon = \sqrt{4\sigma^2/m} \) turns out to work. Ignoring the rounding issues, this gives \( \epsilon = O(\sqrt{\sigma^2 \log(1/\delta)}/n) \) which is precisely what we want.

To handle the rounding errors. Suppose that \( m \) is such that \( m \geq n/2k \) then we can just as well plug in \( n/2k \) into our bound, which would give

\[ |\hat{\mu} - \mu| \leq \sqrt{\frac{64\sigma^2 \log(1/\delta)}{n}}. \]

To satisfy the requirement on \( m \), we also need

\[ \frac{n}{k} - 1 \geq \frac{n}{2k} \Rightarrow n \geq 2\lceil 8 \log(1/\delta) \rceil. \]

4. **Uniform convergence.**

   (a) Setting the right hand side to be at most \( \delta \), we will get a quadratic inequality in \( t \). For convenience define \( V_n = \sum_{i=1}^{n} \mathbb{E}[X_i^2] \)

\[ t^2/2 \geq \log(1/\delta) (V_n + Mt/3). \]
At this point we can try to solve the inequality exactly but it is much simpler to just set \( t = t_1 + t_2 \) where \( t_1 \) is large enough to dominate the first part, and \( t_2 \) is large enough to dominate the second part. Specifically, if \( t_1 = \sqrt{2V_n \log(1/\delta)} \) and \( t_2 = \frac{2}{3} M \log(1/\delta) \) then,

\[
\frac{(t_1 + t_2)^2}{2} \geq \frac{t_1^2}{2} + t_1 t_2 + \frac{t_2^2}{2} = V_n \log(1/\delta) + \frac{t_2^2}{2}(2t_1 + t_2) \geq V_n \log(1/\delta) + \frac{1}{3} M t \log(1/\delta).
\]

Thus we have proved that with probability at least \( 1 - \frac{1}{3} \delta \)

\[
\sum_{i=1}^{n} X_i \leq \sqrt{2 \sum_{i=1}^{n} E X_i^2 \log(1/\delta) + \frac{2}{3} M \log(1/\delta)}.
\]

The final statement is proved by taking a union bound for the positive and negative parts, dividing through by \( n \), and observing that for iid random variables \( \sum_{i=1}^{n} E X_i^2 = nE X^2 \).

(b) We apply Bernstein’s inequality to the risk:

\[
\hat{R}_n(h) - R(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i) - R(h)
\]

This is clearly a sum of \( n \) iid centered random variables and we can set \( M = 1 \) in the application of Bernstein’s inequality, since each empirical loss term is 0,1 and the population risk is in \([0,1]\). As for the variance term, we get

\[
E(\ell(h(X), Y) - R(h))^2 \leq E \ell^2(h(X), Y) \leq E \ell(h(X), Y) = R(h).
\]

Thus via Bernstein’s inequality and a union bound over all of the hypotheses, we get that with probability at least \( 1 - \delta \)

\[
\forall h \in \mathcal{H}. |\hat{R}(h) - R(h)| \leq \sqrt{\frac{2 R(h) \log(2|\mathcal{H}|/\delta)}{n}} + \frac{2 \log(2|\mathcal{H}|/\delta)}{3n}.
\]

Now using the usual excess risk argument,

\[
R(\hat{h}_n) - R(h^*) = R(\hat{h}_n) - \hat{R}(\hat{h}_n) + \hat{R}(\hat{h}_n) - \hat{R}(h^*) + \hat{R}(h^*) - R(h^*)
\]

\[
\leq \sqrt{\frac{2 R(\hat{h}_n) \log(2|\mathcal{H}|/\delta)}{n}} + \sqrt{\frac{2 R(h^*) \log(2|\mathcal{H}|/\delta)}{n}} + \frac{4 \log(2|\mathcal{H}|/\delta)}{3n}.
\]

We just have to work with the first term from now on. For convenience \( \nu_n = 2 \log(2|\mathcal{H}|/\delta)/n \)

\[
\sqrt{R(\hat{h}_n)} \nu_n \leq \sqrt{(R(\hat{h}_n) - R(h^*) + R(h^*))} \nu_n
\]

\[
\leq \sqrt{(R(\hat{h}_n) - R(h^*))} \nu_n + \sqrt{R(h^*)} \nu_n
\]

\[
\leq \frac{1}{2} (R(\hat{h}_n) - R(h^*)) + \frac{1}{2} \nu_n + \sqrt{R(h^*)} \nu_n.
\]

Here we use two inequalities, namely \( \sqrt{a + b} \leq \sqrt{a} + \sqrt{b} \) and \( \sqrt{ab} \leq a/2 + b/2 \), which are easy to verify. Now if we plug this in above and re-arrange, we get

\[
R(\hat{h}_n) - R(h^*) \leq 4 \sqrt{\frac{2 R(h^*) \log(2|\mathcal{H}|/\delta)}{n}} + \frac{14 \log(2|\mathcal{H}|/\delta)}{3n}.
\]