0 Recap about Uniform Stability

Intuitively, an algorithm $A$ is stable if it is robust to small changes in the training sample, i.e., the variation in its output $h$ is small w.r.t the sample size.

Definition 1 (Uniform stability). Given a training set $S$ of size $m$, we build $\forall i = 1, \ldots, m$:

- $S^i = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_m\}$ by removing the $i$-th element of $S$
- $S^i = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_m\}$ by replacing the $i$-th element of $S$ by $z'_i$ drawn i.i.d. from $P$

An algorithm $L$ has uniform stability $\frac{\beta}{m}$ w.r.t a loss function $\ell$ if the following holds:

$$ \forall S, \forall i \in \{1, \ldots, m\}, \sup_z |\ell(h_S, z) - \ell(h_{S^i}, z)| \leq \frac{\beta}{m}$$

where $\beta$ is a positive constant, $h_S$ and $h_{S^i}$ are the hypothesis learned by $L$ from $S$ and $S^i$ respectively.

Theorem 2 (Generalization bound using uniform stability). Let $S$ be a training sample of size $m$ and $\delta > 0$. For any algorithm $L$ with uniform stability $\frac{\beta}{m}$ with respect to a loss function $\ell$ bounded by $M$, with probability $1 - \delta$, we have:

$$ \mathcal{R}_{h_S} \leq \hat{\mathcal{R}}_{h_S} + \frac{2\beta}{m} + (4\beta + M) \sqrt{\frac{\ln 1/\delta}{2m}}. $$

1 Uniform Stability of the Ridge Regression

We consider the following problem.

- $Z = \{z_i = (x_i, y_i)\}_{i=1}^m$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathcal{Y} = [0, B]$.
- Let $\ell(\theta, z) = (\theta^T x - y)^2 = c(\theta^T x, y)$ be the loss function.
- $R_v(\theta) = \frac{1}{m} \sum_{j=1}^m \ell(\theta, z_j) + \lambda \|\theta\|_2^2$ and $R_v^\lambda(\theta) = \frac{1}{m-1} \sum_{j=1, j \neq i}^m \ell(\theta, z_j) + \lambda \|\theta\|_2^2$
- The ridge regression problem is then defined as:

$$ \min_{\theta} R_v(\theta) $$

In class (Lecture 2, slide 55/80), we learned a bound of ridge regression using uniform stability:

Theorem 3. If $c(\theta^T x, y) \leq B^2$, then with probability $1 - \delta$, we have:

$$ \mathcal{R}_{h_\theta} \leq \hat{\mathcal{R}}_{h_\theta} + \frac{4B^2}{\lambda m} + \left(\frac{8B^2}{\lambda} + 2B\right) \sqrt{\frac{\ln 1/\delta}{2m}}. $$
The proof of Theorem 3 can be found at [1], page 18. However, the assumption \( c(\theta^T x, y) \leq B^2 \) is strong and may not be true in practice. The objective of this exercise is to prove a more general bound, where it takes into account \( \sigma \)-admissibility of the loss function \( \ell \).

**Theorem 4.** Suppose that \( \|x_i\|_2 \leq K \) and \( \ell \) is \( \sigma \)-admissible, then with probability \( 1 - \delta \), we have:

\[
R_\theta \leq \hat{R}_\theta + \frac{2\sigma^2 K^2}{\lambda m} + \left( \frac{4\sigma^2 K^2}{\lambda} + M \right) \sqrt{\frac{\ln 1/\delta}{2m}},
\]

where \( M = B^2 (K + \sqrt{\lambda})^2 \).

We do not need the definition \( \sigma \)-admissibility to prove Theorem 4, but for the sake of completion, we recall it here.

**Definition 5 (\( \sigma \)-admissibility).** A loss function \( \ell \) is \( \sigma \)-admissible if the associated cost function \( c(h(x), y) \) is convex w.r.t. its first argument and the following conditions holds

\[
\forall y_1, y_2 \in D, \forall y' \in \mathcal{Y}, |c(y_1, y') - c(y_2, y')| \leq \sigma |y_1 - y_2|,
\]

where \( D = \{ y : \exists \theta, \exists x, \theta^T x = y \} \) is the domain of the first argument of \( c \).

### 1.1 Bounding the elements

Throughout this section, we call \( \theta \) a minimizer of \( R_r \) and \( \theta^{\hat{i}} \) a minimizer of \( \hat{R}_{\hat{r}} \hat{i} \), and note that \( \ell \) is \( \sigma \)-admissible. We accept the following lemma:

**Lemma 6.** For any \( t \in [0, 1] \):

\[
\|\theta\|^2_2 - \|\theta + t \Delta \theta\|^2_2 + \|\theta^{\hat{i}}\|^2_2 - \|\theta^{\hat{i}} - t \Delta \theta\|^2_2 \leq \frac{t\sigma}{\lambda m} |(\Delta \theta)^T x_i|,
\]

where \( \Delta \theta = \theta^{\hat{i}} - \theta \).

We will also use Holder’s inequality that if with \( \frac{1}{p} + \frac{1}{q} = 1 \),

\[
\sum_i |x_iz_i| \leq \left( \sum_i |x_i|^p \right)^{1/p} \left( \sum_i |z_i|^q \right)^{1/q}.
\]

1. Show that

\[
\|\Delta \theta\|_2 \leq \frac{\sigma K}{\lambda m}.
\]

**Hint:** Use Lemma 6 to show that \( \frac{1}{2} \|\Delta \theta\|_2^2 \leq \frac{\sigma}{2\lambda m} |(\Delta \theta)^T x_i| \), then use Holder inequality to show that \( \frac{1}{2} \|\Delta \theta\|_2^2 \leq \frac{\sigma}{2\lambda m} |(\Delta \theta)^T x_i| \).

2. Find an upper bound for \( \|\theta\|_2 \). More precisely, show that

\[
\|\theta\|_2 \leq \frac{B}{\sqrt{\lambda}}.
\]

**Hint:** Use the fact that \( R_r(\theta) \leq R_r(0) \).

3. Find an upper bound for the loss function \( \ell(\theta, z) \). More precisely, show that

\[
\ell(\theta, z) \leq B^2 (K + \sqrt{\lambda})^2.
\]
1.2 Proof of Theorem 4

1. Find a stability constant for Theorem 4 by showing that
\[ |\ell(\theta, z) - \ell(\theta^i, z)| \leq \frac{\sigma^2 K^2}{\lambda m}. \]

Then complete the proof of Theorem 4.

2 Closed form of simple linear regression

We consider the linear regression problem without regulation:
\[
\min_{\theta} \hat{R}(\theta),
\]
where \(\hat{R}(\theta) = \frac{1}{m} \sum_{j=1}^{m}(\theta^T x_j - y_j)^2\). We learned in Lecture 2 that the closed-form of the linear regression problem is
\[
\theta = (X^T X)^{-1} X^T y. \tag{4}
\]
In this problem, we prove the closed-form in a simple case: \(x\) has a only one feature. Let \(Z = \{(x_i, y_i)\}_{i=1}^{m}\) where \(x_i, y_i \in \mathbb{R}\). The simple linear regression problem is then defined as:
\[
(P): \min_{a,b} \sum_{i=1}^{m}(a + bx_i - y_i)^2
\]

Let \((\hat{a}, \hat{b})\) be a solution of \((P)\), and \(x, y\) be the averages of \(x = \{x_i\}, y = \{y_i\}\) respectively.

1. Show that \(\hat{a} = \overline{y} - \hat{b}\overline{x}\).

2. Show that
\[
\hat{b} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{m}(x_i - \overline{x})^2}.
\]

3. Compare your result with the closed-form in eq (4).

3 Convex hull and linear separability

Given a set of data points \(\{x_i\}\), we can define the convex hull to be the set of all points \(x\) given by \(x = \sum_i a_i x_i\), where \(a_i \geq 0\) and \(\sum_i a_i = 1\). Consider a second set of points \(\{x'_j\}\) together with their corresponding convex hull. By definition, the two sets of points will be linearly separable if there exists a vector \(w\) and a scalar \(w_0\) such that \(w^T x_i + w_0 > 0\) for all \(x_i\), and \(w^T x'_j + w_0 < 0\) for all \(x'_j\).

1. Suppose that the convex hulls of \(\{x_i\}\) and \(\{x'_j\}\) intersect. Show that \(\{x_i\}\) and \(\{x'_j\}\) cannot be linearly separable.

2. Conversely, show that if they are linearly separable, their convex hulls do not intersect.

References