1 Relaxations

Basic idea: Replace $V_n(x_{1:t}, y_{1:t})$ by a relaxation $\text{Rel}_n(x_{1:t}, y_{1:t})$.

Let us define relaxation $\text{Rel}_n$ as any mapping $\text{Rel}_n : \bigcup_{t=0}^{n} \mathcal{X}^t \times \mathcal{Y}^t \to \mathbb{R}$. Further, we say that a relaxation is admissible if it satisfies the following two conditions.

Initial condition:

$$- \inf_{f \in F} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq \text{Rel}_n(x_{1:n}, y_{1:n})$$

Admissibility condition: For any $x_1, \ldots, x_t \in \mathcal{X}$ and any $y_1, \ldots, y_{t-1} \in \mathcal{Y}$,

$$\inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y \in \mathcal{Y}} \{ E_{\hat{y}_t \sim q_t} [\ell(\hat{y}_t, y_t)] + \text{Rel}_n(x_{1:t}, y_{1:t}) \} \leq \text{Rel}_n(x_{1:t-1}, y_{1:t-1})$$

**Proposition 1.** If $\text{Rel}_n$ is any admissible relaxation, then if we use the learning algorithm that at time $t$, given $x_t$ produces $q_t(x_t) = \arg\min_{y_t} \sup_{q \in \Delta(\mathcal{Y})} \{ E_{\hat{y}_t \sim q_t} [\ell(\hat{y}_t, y_t)] + \text{Rel}_n(x_{1:t}, y_{1:t}) \}$, then,

$$\frac{1}{n} \sum_{t=1}^{n} E_{\hat{y}_t \sim q_t} [\ell(\hat{y}_t, y_t)] - \inf_{f \in F} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq \frac{1}{n} \text{Rel}_n(\cdot)$$

**Proof.** Assume $\text{Rel}_n$ is any admissible relaxation. Also let $q_t$’s be obtained by as described above. Then, by initial condition,

$$\sum_{t=1}^{n} E_{\hat{y}_t \sim q_t(x_t)} [\ell(\hat{y}_t, y_t)] - \inf_{f \in F} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq \sum_{t=1}^{n} E_{\hat{y}_t \sim q_t} [\ell(\hat{y}_t, y_t)] + \text{Rel}_n(x_{1:n}, y_{1:n})$$

$$\leq \sum_{t=1}^{n-1} E_{\hat{y}_t \sim q_t} [\ell(\hat{y}_t, y_t)] + \sup_{y \in \mathcal{Y}} \{ E_{\hat{y}_t \sim q_n(x_n)} [\ell(\hat{y}_t, y_t)] + \text{Rel}_n(x_{1:n}, y_{1:n}) \}$$

$$= \sum_{t=1}^{n-1} E_{\hat{y}_t \sim q_t} [\ell(\hat{y}_t, y_t)] + \inf_{q_n \in \Delta(\mathcal{Y})} \sup_{y_t \in \mathcal{Y}} \{ E_{\hat{y}_t \sim q_n} [\ell(\hat{y}_t, y_t)] + \text{Rel}_n(x_{1:n}, y_{1:n}) \}$$

by admissibility condition,

$$\leq \sum_{t=1}^{n-1} E_{\hat{y}_t \sim q_t} [\ell(\hat{y}_t, y_t)] + \text{Rel}_n(x_{1:n-1}, y_{1:n-1})$$

$$\leq \ldots \leq \text{Rel}_n(\cdot)$$

\[ \square \]
2 Sequential Rademacher Relaxation

Just like we defined Rademacher complexity for statistical learning, one can define an online version of it called sequential Rademacher Complexity. Specifically, the sequential Rademacher complexity of a function class $G \subset \mathbb{R}^Z$ is defined as:

$$R^sq_n(G) := \sup_z \mathbb{E}_\epsilon \left[ \sup_{g \in G} \sum_{t=1}^n \epsilon_t g(z(\epsilon_1, \ldots, \epsilon_{t-1})) \right]$$

where $z$ is a $Z$ valued binary tree of depth $n$ where the nodes at level $t$ can be defined by mapping $z_t : \{\pm 1\}^{t-1} \mapsto Z$.

Pictorially, we can view the Rademacher complexity as:

\[
\begin{align*}
\text{Definition 1.} \quad & \text{Define the sequential Rademacher relaxation as} \\
\text{Rad}_n(x_{1:t}, y_{1:t}) := & \sup_{x,y} \mathbb{E}_{\epsilon_t, \ldots, \epsilon_n} \left[ 2 \sum_{s=t+1}^n \epsilon_s \ell(f(x_s), y_s) - \sum_{s=1}^t \ell(f(x_s), y_s) \right] \\
\text{where} \ x & \text{ above is supremum over } X \text{ valued tree of depth } n - t \ \text{and similarly } y \text{ is a } Y \text{-valued tree of depth } n - t.
\end{align*}
\]

Claim 2. $\text{Rad}_n$ is an admissible relaxation. Further using the $q_t$ corresponding to this relaxation one get that

$$\mathbb{E}[\text{Reg}_n] \leq 2R^sq_n(\ell \circ \mathcal{F})$$

Proof. As for initial condition note that,

$$\text{Rad}_n(x_{1:n}, y_{1:n}) = \sup_{f \in \mathcal{F}} \left[ -\sum_{s=1}^n \ell(f(x_s), y_s) \right] = -\inf_{f \in \mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t)$$

Now to check admissibility, note that

$$\inf_{q_t \in \Delta(Y)} \sup_{y_t \in Y} \{ \mathbb{E}_{y_t \sim q_t} [\ell(\hat{y}_t, y_t)] + \text{Rad}_n(x_{1:t}, y_{1:t}) \} = \sup_{p_t \in \Delta(Y)} \left\{ \inf_{y_t \in Y} \mathbb{E}_{y_t \sim p_t} [\ell(\hat{y}_t, y_t)] + \mathbb{E}_{y_t \sim p_t} [\text{Rad}_n(x_{1:t}, y_{1:t})] \right\}$$
\[
\begin{align*}
\sup_{y_t \in Y} \left\{ \inf_{\hat{y}_t \sim y_t} [\ell(\hat{y}_t, y_t)] + \E_{y_t \sim y_t} \sup_{x_t} \sup_{f \in \mathcal{F}} \left[ 2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(x_{s-t}(\epsilon t+1; s-1)), y_{s-t}(\epsilon t+1; s-1)) - \sum_{s=1}^{t} \ell(f(x_s), y_s) \right] \right\} \\
\sup_{y_t \in Y} \left\{ \inf_{\hat{y}_t \sim y_t} [\ell(\hat{y}_t, y_t)] + \sum_{s=t+1}^{n} \epsilon_s \ell(f(x_{s-t}(\epsilon t+1; s-1)), y_{s-t}(\epsilon t+1; s-1)) - \sum_{s=1}^{t} \ell(f(x_s), y_s) \right\} \\
\leq \sup_{y_t \in Y} \left\{ \E_{y_t \sim y_t} \sup_{x_t} \sup_{f \in \mathcal{F}} \left[ 2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(x_{s-t}(\epsilon t+1; s-1)), y_{s-t}(\epsilon t+1; s-1)) \\
+ (\ell(f(x_t), y_t) - \ell(f(x_t), y_t)) - \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right] \right\} \\
= \sup_{y_t, y_t' \in Y} \E_{y_t \sim y_t} \sup_{x_t, y_t} \sup_{f \in \mathcal{F}} \left\{ 2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(x_{s-t}(\epsilon t+1; s-1)), y_{s-t}(\epsilon t+1; s-1)) \\
+ \epsilon_t (\ell(f(x_t), y_t' - \ell(f(x_t), y_t)) - \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right\} \\
\leq \sup_{y_t, y_t' \in Y} \E_{y_t \sim y_t} \sup_{x_t, y_t} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} \epsilon_s \ell(f(x_{s-t}(\epsilon t+1; s-1)), y_{s-t}(\epsilon t+1; s-1)) \\
+ \epsilon_t (\ell(f(x_t), y_t') - \ell(f(x_t), y_t)) - \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right\} \\
+ \sup_{y_t \in Y} \E_{y_t \sim y_t} \sup_{x_t, y_t} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} \epsilon_s \ell(f(x_{s-t}(\epsilon t+1; s-1)), y_{s-t}(\epsilon t+1; s-1)) \\
- \epsilon_t (\ell(f(x_t), y_t') - \ell(f(x_t), y_t)) - \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right\}
\end{align*}
\]
= 2 \sup_{y_t \in Y} \sup_{x_t, y_{t+1}, \ldots, y_n} \sup_{f \in F} \left\{ \sum_{s=t+1}^{n} \epsilon_s \ell(f(x_s-\epsilon_t(x_{t+1}))), y_{s-t}(x_{t+1:s-1}) \right. \\
+ \epsilon_t \ell(f(x_t), y_t) - \frac{1}{2} \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right\} \\
\leq \sup_{x_t \in X} \sup_{y_t \in Y} \sup_{x_{t+1}, \ldots, y_n} \sup_{f \in F} \left\{ 2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(x_s-\epsilon_t(x_{t+1}))), y_{s-t}(x_{t+1:s-1}) \right. \\
+ \epsilon_t \ell(f(x_t), y_t) - \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right\} = \Rad_n(x_{1:t-1}, y_{1:t-1})

Put the $x_t$ that achieves the supremum as the root of a new tree of depth $n - t + 1$ and its left sub-tree is the $x^+$ tree that attains supremum when $\epsilon_t = -1$ and right sub-tree is the one that attains supremum when $\epsilon_t = -1$. Similarly for the $y$'s, hence,

$$
= \sup_{x, y} \sup_{f \in F} \left\{ 2 \sum_{s=t}^{n} \epsilon_s \ell(f(x_s-\epsilon_t(x_{1:s-1}))), y_{s-t}(x_{1:s-1}) - \sum_{s=1}^{t} \ell(f(x_s), y_s) \right\} = \Rad_n(x_{1:t-1}, y_{1:t-1})
$$

This shows admissibility. From the earlier proposition, regret is bounded by

$$
\mathbb{E}[R_n] \leq \frac{2}{n} \Rad_n(\cdot) = \frac{1}{n} \sup_{x, y} \sup_{f \in F} \left[ \sum_{s=1}^{n} \epsilon_s \ell(f(x_{s}(\epsilon_{1:s-1}))), y_{s}(\epsilon_{1:s-1}) \right] = 2 \mathcal{R}_n^q(\ell \circ \mathcal{F})
$$

\[ \square \]