Machine Learning Theory (CS 6783)

Lecture 18: Relaxations for Online Learning

1 Relaxations

Let us define relaxation $\mathbf{Rel}_n$ as any mapping $\mathbf{Rel}_n : \bigcup_{t=0}^{n} \mathcal{X}^t \times \mathcal{Y}^t \mapsto \mathbb{R}$. Further, we say that a relaxation is admissible w.r.t some function $\phi : (\mathcal{X} \times \mathcal{Y})^n \mapsto \mathbb{R}$ if it satisfies the following conditions.
1. Dominance condition :

$$\mathbf{Rel}_n\left(x_{1:n}, y_{1:n}\right) \geq -\phi((x_1, y_1), \ldots, (x_n, y_n))$$

2. Final condition :

$$\mathbf{Rel}_n\left(\cdot\right) \leq 0$$

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

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

The following proposition shows that there exists an algorithm such that

if one can find a relaxation that is admissible w.r.t. some ϕ, then there exists an algorithm for which ϕ is an upper bound on loss of algorithm if and only if one can find a relaxation admissible w.r.t. that ϕ

Proposition 1. *If $\mathbf{Rel}_n$ is any admissible relaxation w.r.t. ϕ, then if we use the learning algorithm that at time t, given x_t produces $q_t(x_t) = \operatorname{argmin}_{q \in \Delta(\mathcal{Y})} \sup_{y_t} \left\{\mathbb{E}_{\hat{y}_t \sim q_t}\left[\ell(\hat{y}_t, y_t)\right] + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)\right\}$, then,*

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t}\left[\ell(\hat{y}_t, y_t)\right] \leq \tfrac{1}{n}\phi((x_1, y_1), \ldots, (x_n, y_n))$$

Further, there can exist an algorithm with above bound on average loss only if there exists an admissible relaxation w.r.t. that ϕ.

Proof. Assume $\mathbf{Rel}_n$ is any admissible relaxation w.r.t. ϕ. Also let q_t's be obtained by as described above. Then, by dominance condition,

$$\begin{aligned}
\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t(x_t)}\left[\ell(\hat{y}_t, y_t)\right] - \phi((x_1, y_1), \ldots, (x_n, y_n)) &\leq \sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t}\left[\ell(\hat{y}_t, y_t)\right] + \mathbf{Rel}_n\left(x_{1:n}, y_{1:n}\right) \\
&\leq \sum_{t=1}^{n-1} \mathbb{E}_{\hat{y}_t \sim q_t}\left[\ell(\hat{y}_t, y_t)\right] + \sup_{y_t \in \mathcal{Y}} \left\{\mathbb{E}_{\hat{y}_n \sim q_n(x_n)}\left[\ell(\hat{y}_n, y_n)\right] + \mathbf{Rel}_n\left(x_{1:n}, y_{1:n}\right)\right\} \\
&= \sum_{t=1}^{n-1} \mathbb{E}_{\hat{y}_t \sim q_t}\left[\ell(\hat{y}_t, y_t)\right] + \inf_{q_n \in \Delta(\mathcal{Y})} \sup_{y_t \in \mathcal{Y}} \left\{\mathbb{E}_{\hat{y}_n \sim q}\left[\ell(\hat{y}_n, y_n)\right] + \mathbf{Rel}_n\left(x_{1:n}, y_{1:n}\right)\right\}
\end{aligned}$$

by admissibility condition,

$$\leq \sum_{t=1}^{n-1} \mathbb{E}_{\hat{y}_t \sim q_t} \left[\ell(\hat{y}_t, y_t)\right] + \mathbf{Rel}_n\left(x_{1:n-1}, y_{1:n-1}\right)$$
$$\leq \ldots \leq \mathbf{Rel}_n\left(\cdot\right) \leq 0$$

where last inequality is by the final condition. Hence we have that if relaxation is admissible then algorithm yields the promised bound on average expected loss. To show the only if part, simply define $\mathbf{Rel}_n\left(x_{1:n}, y_{1:n}\right) \geq \phi((x_1, y_1), \ldots, (x_n, y_n))$ (so dominance condition is satisfied automatically). Then recursively define

$$\mathbf{Rel}_n\left(x_{1:t-1}, y_{1:t-1}\right) := \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t \in \mathcal{Y}} \left\{\mathbb{E}_{\hat{y}_t \sim q_t} \left[\ell(\hat{y}_t, y_t)\right] + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)\right\}$$

(hence admissibility is satisfied). That final condition is satisfied is a simple consequence of the fact that this algorithm gotten from the relaxation is the minimax optimal algorithm since at every step its picking the minimizer strategy against the maximizing strategy opponent. Hence if there exists any algorithm with the promise bound, then the minimax algorithm also achieves this bound and so $\mathbf{Rel}_n\left(\cdot\right) \leq 0$. □

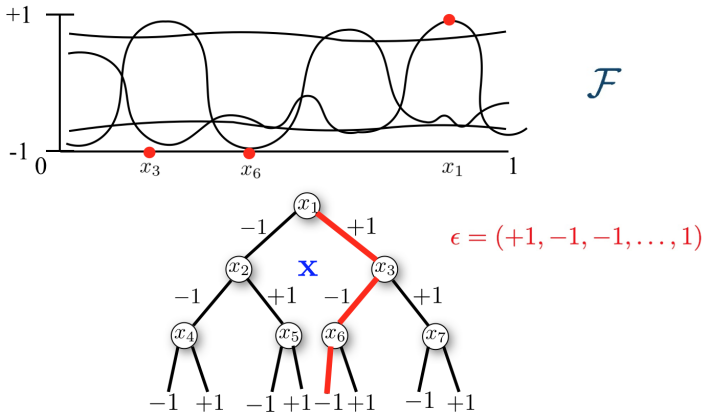
2 Sequential Rademacher Relaxation

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 $\mathcal{G} \subset \mathbb{R}^{\mathcal{Z}}$ is defined as:

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

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

Pictorially, we can view the Rademacher complexity as :



Definition 1. *Define the sequential Rademacher relaxation as*

$$\mathbf{Rad}_n\left(x_{1:t}, y_{1:t}\right) := \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f\in\mathcal{F}} \left[2\sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \ell(f(x_s), y_s) \right] - 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})$$

where $\mathbf{x}$ *above is supremum over* $\mathcal{X}$ *valued tree of depth* $n-t$ *and similarly* $\mathbf{y}$ *is a* $\mathcal{Y}$*-valued tree of depth* $n-t$*.*

We will show that the sequential Rademacher relaxation is admissible w.r.t. $\phi((x_1,y_1),\ldots,(x_n,y_n)) = \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t),y_t) + 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})$ and hence conclude that analogous to statistical learning result, the sequential Rademacher complexity of the loss class is an upper bound on regret w.r.t. arbitrary function class $\mathcal{F}\subset\mathcal{Y}^{\mathcal{X}}$.

Claim 2. $\mathbf{Rad}_n$ *is an admissible relaxation w.r.t.* ϕ *defined as:*

$$\phi((x_1,y_1),\ldots,(x_n,y_n)) = \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t),y_t) + 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})$$

Further using the q_t *corresponding to this relaxation one get that*

$$\mathbb{E}\left[\mathrm{Reg}_n(\mathcal{F})\right] \le 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})$$

Proof. As for Dominance condition note that,

$$\mathbf{Rad}_n\left(x_{1:n}, y_{1:n}\right) = \sup_{f\in\mathcal{F}} \left[-\sum_{s=1}^{n} \ell(f(x_s),y_s) \right] = -\inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t),y_t) - 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F}) = -\phi((x_1,y_1),\ldots,(x_n,y_n))$$

To check final condition, note that:

$$\mathbf{Rad}_n\left(\cdot\right) = \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{1:n}} \sup_{f\in\mathcal{F}} \left[2\sum_{s=1}^{n} \epsilon_s \ell(f(\mathbf{x}_s(\epsilon_{1:s-1})), \mathbf{y}_s(\epsilon_{1:s-1})) \right] - 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F}) = 0$$

Now to check admissibility, note that

$$\begin{aligned}
\inf_{q_t\in\Delta(\mathcal{Y})} \sup_{y_t\in\mathcal{Y}} \left\{ \mathbb{E}_{\hat{y}_t\sim q_t}\left[\ell(\hat{y}_t,y_t)\right] + \mathbf{Rad}_n\left(x_{1:t},y_{1:t}\right) \right\} &= \sup_{p_t\in\Delta(\mathcal{Y})} \inf_{\hat{y}_t\in\mathcal{Y}} \mathbb{E}_{y_t\sim p_t}\left[\ell(\hat{y}_t,y_t) + \mathbf{Rad}_n\left(x_{1:t},y_{1:t}\right)\right] \\
&= \sup_{p_t\in\Delta(\mathcal{Y})} \left\{ \inf_{\hat{y}_t\in\mathcal{Y}} \mathbb{E}_{y_t\sim p_t}\left[\ell(\hat{y}_t,y_t)\right] + \mathbb{E}_{y_t\sim p_t}\left[\mathbf{Rad}_n\left(x_{1:t},y_{1:t}\right)\right] \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{p_t \in \Delta(\mathcal{Y})} \Bigg\{ \inf_{\hat{y}_t \in \mathcal{Y}} \mathbb{E}_{y_t \sim p_t} \left[\ell(\hat{y}_t, y_t) \right] \\
&\quad + \mathbb{E}_{y_t \sim p_t} \left[\sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \left[2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \ell(f(x_s), y_s) \right] \right] \Bigg\} \\
&= \sup_{p_t \in \Delta(\mathcal{Y})} \Bigg\{ \mathbb{E}_{y_t \sim p_t} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \Bigg\{ \inf_{\hat{y}_t \in \mathcal{Y}} \mathbb{E}_{y_t \sim p_t} \left[\ell(\hat{y}_t, y_t) \right] \\
&\quad + 2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \ell(f(x_s), y_s) \Bigg\} \Bigg\} \\
&\leq \sup_{p_t \in \Delta(\mathcal{Y})} \Bigg\{ \mathbb{E}_{y_t \sim p_t} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \Bigg\{ \mathbb{E}_{y'_t \sim p_t} \left[\ell(f(x_t), y'_t) \right] \\
&\quad + 2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \ell(f(x_s), y_s) \Bigg\} \Bigg\} \\
&\leq \sup_{p_t \in \Delta(\mathcal{Y})} \mathbb{E}_{y_t, y'_t \sim p_t} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \Bigg\{ 2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) \\
&\qquad\qquad + \left(\ell(f(x_t), y'_t) - \ell(f(x_t), y_t) \right) - \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \Bigg\} \\
&= \sup_{p_t \in \Delta(\mathcal{Y})} \mathbb{E}_{y_t, y'_t \sim p_t} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \Bigg\{ 2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) \\
&\qquad\qquad + \epsilon_t \left(\ell(f(x_t), y'_t) - \ell(f(x_t), y_t) \right) - \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \Bigg\} \\
&\leq \sup_{y_t, y'_t \in \mathcal{Y}} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \Bigg\{ 2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) \\
&\qquad\qquad + \epsilon_t \left(\ell(f(x_t), y'_t) - \ell(f(x_t), y_t) \right) - \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \Bigg\} \\
&\leq \sup_{y'_t \in \mathcal{Y}} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \Bigg\{ \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) \\
&\qquad\qquad + \epsilon_t \ell(f(x_t), y'_t) - \frac{1}{2} \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \Bigg\} \\
&\quad + \sup_{y_t \in \mathcal{Y}} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \Bigg\{ \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) \\
&\qquad\qquad - \epsilon_t \ell(f(x_t), y'_t) - \frac{1}{2} \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \Bigg\}
\end{aligned}
$$

$$= 2 \sup_{y_t \in \mathcal{Y}} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) \right.$$
$$\left. + \epsilon_t \ell(f(x_t), y_t) - \frac{1}{2} \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right\}$$
$$\leq \sup_{x_t \in \mathcal{X}} \sup_{y_t \in \mathcal{Y}} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ 2 \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) \right.$$
$$\left. + \epsilon_t \ell(f(x_t), y_t) - \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right\}$$

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 $\mathbf{x}^+$ 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,

$$= \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t:n}} \sup_{f \in \mathcal{F}} \left\{ 2 \sum_{s=t}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t:s-1})), \mathbf{y}_{s-t}(\epsilon_{t:s-1})) - \sum_{s=1}^{t} \ell(f(x_s), y_s) \right\} = \mathbf{Rad}_n \left(x_{1:t-1}, y_{1:t-1} \right)$$

This shows admissibility. Finally, from the earlier proposition:

$$\mathbb{E}\left[\mathrm{Reg}_n\right] \leq 2\mathcal{R}_n^{sq}(\ell \circ \mathcal{F})$$

□