Machine Learning Theory (CS 6783)

Lecture 11: Relaxations for Online Learning

1 Relaxations

Cover's result while was very basic, actually gave us a thread that we can follow through to getting most algorithms in online learning. Let us now consider the Backward induction style idea in Cover's technique, but relax it to get a mechanism for algorithm design for general online learning. We would like to consider the general online learning framework with an arbitrary loss ℓ, a general input set $\mathcal{X}$ and outcome set $\mathcal{Y}$. Our aim is to come up with a (possibly randomized) algorithm that guarantees that against any arbitrary adversary,

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

As before, we would like to know when is this possible, what is the corresponding algorithm and how do we derive such algorithms.

Basic idea: 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 if it satisfies the following conditions.

1. Dominance condition :

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

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 $y_1, \ldots, y_{t-1} \in \mathcal{Y}$,

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

By Minimax theorem

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

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

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

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

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

by admissibility condition,

$$\begin{aligned}
&\le \sum_{t=1}^{n-1} \mathbb{E}_{\hat{y}_t\sim q_t}[\ell(\hat{y}_t, y_t)] + \mathbf{Rel}_n(x_{1:n-1}, y_{1:n-1}) \\
&\le \ldots \le \mathbf{Rel}_n(\cdot)
\end{aligned}$$

□

2 Exponential Weights Algorithm For Demonstration

We want to obtain an algorithm to guarantee that $\mathbb{E}[\mathbf{R}_n] \le C_n(\mathcal{F})$. Lets start here. Clearly, for this we can set

$$\phi(x_1, y_1, \ldots, x_n, y_n) = \min_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) + C_n(\mathcal{F})$$

The first step in relaxation is to fins a mapping that satisfies dominance condition. We can of course define our mapping as one that is equal to ϕ but then solving for the algorithm and further relaxations would become too complicated. So let us use the idea we already saw in the finite lemma proof. Let us move from max (or rather - min(- ...)) to softmax as upper bound. To this end, note that

$$\begin{aligned}
-\phi(x_1, y_1, \ldots, x_n, y_n) &= -\min_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - C_n(\mathcal{F}) \\
&= \max_{f\in\mathcal{F}} -\sum_{t=1}^{n} \ell(f(x_t), y_t) - C_n(\mathcal{F}) \\
&\le \frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}} \exp\left(-\eta\sum_{t=1}^{n} \ell(f(x_t), y_t)\right)\right) - C_n(\mathcal{F}) = \mathbf{Rel}_n(x_{1:n}, y_{1:n})
\end{aligned}$$

Ok now that we have the dominance condition let us move to admissibility. (we can fix C_n at the end to satisfy the final condition). To this end, we will start at the n'th step and the other steps will basically follow the same pattern. For the n'th step note that:

$$\inf_{q_n\in\Delta(\mathcal{Y})}\sup_{y_n\in\mathcal{Y}}\left\{\mathbb{E}_{\hat{y}_n\sim q_n}\left[\ell(\hat{y}_n,y_n)\right]+\mathbf{Rel}_n\left(x_{1:n},y_{1:n}\right)\right\}$$

$$=\inf_{q_n\in\Delta(\mathcal{Y})}\sup_{y_n\in\mathcal{Y}}\left\{\mathbb{E}_{\hat{y}_n\sim q_n}\left[\ell(\hat{y}_n,y_n)\right]+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{t=1}^{n}\ell(f(x_t),y_t)\right)\right)-C_n(\mathcal{F})\right\}$$

$$=\inf_{q_n\in\Delta(\mathcal{Y})}\sup_{y_n\in\mathcal{Y}}\left\{\mathbb{E}_{\hat{y}_n\sim q_n}\left[\ell(\hat{y}_n,y_n)\right]+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{t=1}^{n-1}\ell(f(x_t),y_t)\right)\times e^{-\eta\ell(f(x_n),y_n)}\right)-C_n(\mathcal{F})\right\}$$

$$=\inf_{q_n\in\Delta(\mathcal{Y})}\sup_{y_n\in\mathcal{Y}}\left\{\mathbb{E}_{\hat{y}_n\sim q_n}\left[\ell(\hat{y}_n,y_n)\right]+\frac{1}{\eta}\log\left(\mathbb{E}_{f\sim\hat{q}_n}\left[e^{-\eta\ell(f(x_n),y_n)}\right]\right)+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{t=1}^{n-1}\ell(f(x_t),y_t)\right)\right)\right\}$$
$$-C_n(\mathcal{F})$$

where $\hat{q}_n(f)\propto\exp\left(-\eta\sum_{t=1}^{n-1}\ell(f(x_t),y_t)\right)$

$$\leq\sup_{y_n\in\mathcal{Y}}\left\{\mathbb{E}_{f\sim\hat{q}_n}\left[\ell(f(x_n),y_n)\right]+\frac{1}{\eta}\log\left(\mathbb{E}_{f\sim\hat{q}_n}\left[e^{-\eta\ell(f(x_n),y_n)}\right]\right)+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{t=1}^{n-1}\ell(f(x_t),y_t)\right)\right)\right\}-C_n(\mathcal{F})$$

$$=\sup_{y_n\in\mathcal{Y}}\left\{\frac{1}{\eta}\log\left(\exp(\eta\mathbb{E}_{f\sim\hat{q}_n}\left[\ell(f(x_n),y_n)\right])\right)+\frac{1}{\eta}\log\left(\mathbb{E}_{f\sim\hat{q}_n}\left[e^{-\eta\ell(f(x_n),y_n)}\right]\right)+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{t=1}^{n-1}\ell(f(x_t),y_t)\right)\right)\right\}$$
$$-C_n(\mathcal{F})$$

$$=\sup_{y_n\in\mathcal{Y}}\left\{\frac{1}{\eta}\log\left(\mathbb{E}_{f\sim\hat{q}_n}\left[e^{\eta(\mathbb{E}_{f\sim\hat{q}_n}[\ell(f(x_n),y_n)])-\ell(f(x_n),y_n))}\right]\right)+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{t=1}^{n-1}\ell(f(x_t),y_t)\right)\right)\right\}$$
$$-C_n(\mathcal{F})$$

$$\leq\sup_{y_n\in\mathcal{Y}}\left\{\frac{1}{\eta}\log\left(\mathbb{E}_{f,f'\sim\hat{q}_n}\left[e^{\eta\left(\ell(f'(x_n),y_n))-\ell(f(x_n),y_n)\right)}\right]\right)+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{t=1}^{n-1}\ell(f(x_t),y_t)\right)\right)\right\}-C_n(\mathcal{F})$$

$$=\sup_{y_n\in\mathcal{Y}}\left\{\frac{1}{\eta}\log\left(\mathbb{E}_{f,f'\sim\hat{q}_n}\left[\mathbb{E}_{\epsilon_n}e^{\eta\epsilon_n\left(\ell(f'(x_n),y_n))-\ell(f(x_n),y_n)\right)}\right]\right)+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{t=1}^{n-1}\ell(f(x_t),y_t)\right)\right)\right\}-C_n(\mathcal{F})$$

$$\leq\frac{1}{\eta}\log(\exp(2\eta^2))+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{t=1}^{n-1}\ell(f(x_t),y_t)\right)\right)-C_n(\mathcal{F})$$

$$=2\eta+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{t=1}^{n-1}\ell(f(x_t),y_t)\right)\right)-C_n(\mathcal{F})=\mathbf{Rel}_n\left(x_{1:n-1},y_{1:n-1}\right)$$

Proceeding in same way we can fix

$$\mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)=2\eta(n-t)+\frac{1}{\eta}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\eta\sum_{s=1}^{t}\ell(f(x_s),y_s)\right)\right)-C_n(\mathcal{F})$$

Finally, to ensure final condition note that,

$$\mathbf{Rel}_n(\cdot) = 2\eta n + \frac{1}{\eta}\log(|\mathcal{F}|) - C_n(\mathcal{F})$$

and so

$$C_n(\mathcal{F}) = 2\eta n + \frac{1}{\eta}\log(|\mathcal{F}|)$$

Setting $\eta = \sqrt{\log|\mathcal{F}|/2n}$ yields

$$C_n(\mathcal{F}) = \sqrt{\frac{4\log|\mathcal{F}|}{n}}$$

Also note that the strategy we used was to produce $\hat{y}_t$ by first drawing $f \sim \hat{q}_t$ where

$$\hat{q}_t(f) \propto \exp\left(-\eta \sum_{s=1}^{t-1} \ell(f(x_s), y_s)\right)$$

and then setting $\hat{y}_t = f(x_t)$. This is the exponential weights algorithm.

3 Sequential Rademacher Relaxation

We already claimed this in class without proof. Perhaps, its not time to prove the result that, there exists an algorithm that guarantees that:

$$\mathbb{E}[\mathrm{Reg}_n] \le 2 \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \epsilon_t \ell(f(\mathbf{x}(\epsilon_{1:t-1}), \mathbf{y}(\epsilon_{1:t-1})))\right] =: 2\ \mathcal{R}_n^{sq}(\ell\circ\mathcal{F})$$

We will do this via the idea of relaxations. Let us define the sequential Rademacher Relaxation as follows.

Definition 1. *Define the sequential Rademacher relaxation as*

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

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

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

Proof. As for initial condition note that,

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

Now to check admissibility, note that

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

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

$$
\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}(\epsilon_{t+1:s-1})),\mathbf{y}(\epsilon_{t+1:s-1}))-\sum_{s=1}^{t}\ell(f(x_s),y_s)\right]\right]\Bigg\} - 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F}) \\
&= \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}(\epsilon_{t+1:s-1})),\mathbf{y}(\epsilon_{t+1:s-1}))-\sum_{s=1}^{t}\ell(f(x_s),y_s)\Bigg\}\Bigg\} - 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F}) \\
&\le \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}(\epsilon_{t+1:s-1})),\mathbf{y}(\epsilon_{t+1:s-1}))-\sum_{s=1}^{t}\ell(f(x_s),y_s)\Bigg\}\Bigg\} - 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F}) \\
&\le \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}(\epsilon_{t+1:s-1})),\mathbf{y}(\epsilon_{t+1:s-1})) \\
&\qquad\qquad + (\ell(f(x_t),y'_t)-\ell(f(x_t),y_t))-\sum_{s=1}^{t-1}\ell(f(x_s),y_s)\Bigg\} - 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F}) \\
&= \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}(\epsilon_{t+1:s-1})),\mathbf{y}(\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\} - 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F}) \\
&\le \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}(\epsilon_{t+1:s-1})),\mathbf{y}(\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\} - 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F}) \\
&\le \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}(\epsilon_{t+1:s-1})),\mathbf{y}(\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}(\epsilon_{t+1:s-1})),\mathbf{y}(\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\} - 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})
\end{aligned}
$$

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

Define a new mapping $\mathbf{x}^*$ and set $\mathbf{x}^*(\cdot) = x_t$, the one that achieves the supremum above and further, define $\mathbf{x}^*(+1)$ as the mapping $\mathbf{x}$ that achieves the supremum above amongst the mapping when $\epsilon_t = +1$ and $\mathbf{x}^*(-1)$ as the $\mathbf{x}$ that achieves the supremum when $\epsilon_t = -1$. Similarly define mapping $\mathbf{y}^*$. Since $\mathbf{x}^*, \mathbf{y}^*$ are one choice of such mappings, we conclude by taking supremum that

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

This shows admissibility. Also notice that $\mathbf{Rad}_n(\cdot) = 0$ and so

From the earlier proposition, regret is bounded by

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

□