Machine Learning Theory (CS 6783)

Lecture 17: Relaxations for Online Learning

1 Relaxations

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

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

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

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

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

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

by admissibility condition,

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

Dividing through by n we conclude the result. □

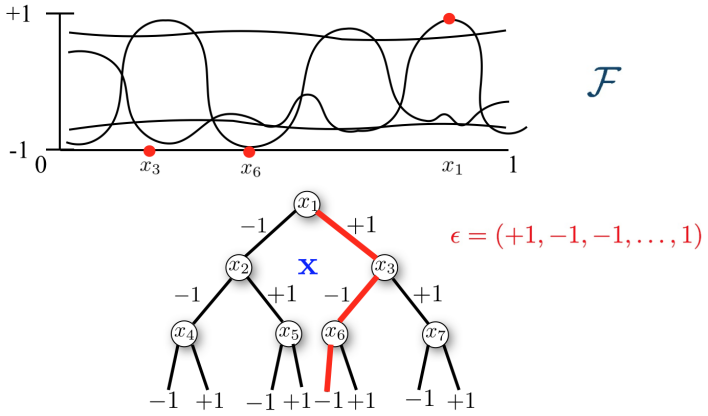
2 Sequential Rademacher Relaxation

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:

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

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 :



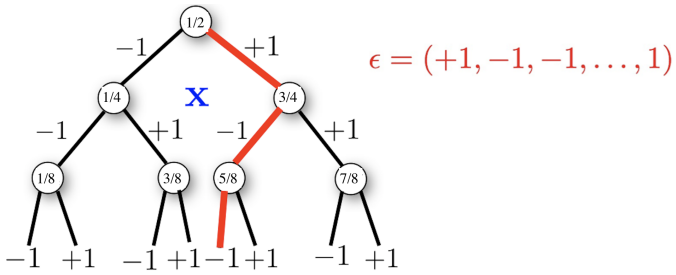
Definition 1. *Define the sequential Rademacher relaxation as*

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

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* $n-t$*.*

Remark 2.1. *I will leave this as an exercise that you can check. Pretty much all the examples of function classes* $\mathcal{F}$ *for which we obtained upper bounds on the statistical Rademacher complexity, we can obtain same upper bounds on the sequential one. This is because the term* $\sum_{t=1}^{n} \epsilon_t \ell(f(\mathbf{x}_t(\epsilon_{1:t-1})), \mathbf{y}_t(\epsilon_{1:t-1}))$ *is a martingale difference sequence (each term in the sum has*

conditional expectation of 0*) and almost all the upper bounds we proved, we only needed that the inner term was a martingale difference sequence and not sum of iid zero mean variables. The only exception to this is the example of learning thresholds. If* $\mathcal{F} \subseteq \{\pm 1\}^{[0,1]}$ *is the class of all thresholds such that every point to the right of the threshold is labeled* -1 *and to the left is labeled as* $+1$*, then we can see that for this class,* $\mathcal{R}_n^{sq}(\ell \circ \mathcal{F}) = 1$*, while statistical Rademacher complexity is* $1/\sqrt{n}$*. To see this consider the following tree:*



For this tree, we have the property that on every path, some threshold attains the label of signs on that path implies that the sequential Rademacher complexity is 1*. However VC dimension for this class is* 1 *and so statistical Rademacher complexity is* $1/\sqrt{n}$

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

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\hat{y}_t\sim q_t}\left[\ell(\hat{y}_t, y_t)\right] \leq \frac{1}{n}\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t), y_t) + 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})$$

Proof. First note that the ϕ function in this case is simply:

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

Now, as for Dominance condition, note that,

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

Next, we check the final condition. Note that:

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

Now to check admissibility, note that

$$
\begin{aligned}
&\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\}\\
&\quad=\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]\right.\\
&\qquad\left.+\mathbb{E}_{y_t\sim p_t}\left[\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]\right]-2n\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})\right\}\\
&\quad\le\sup_{p_t\in\Delta(\mathcal{Y})}\left\{\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}}\left\{\mathbb{E}_{y'_t\sim p_t}\left[\ell(f(x_t),y'_t)\right]\right.\right.\\
&\qquad\left.\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\}-2n\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})\\
&\quad\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}}\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.\\
&\qquad\left.+\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)\right\}-2n\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})\\
&\quad=\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}}\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.\\
&\qquad\left.+\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)\right\}-2n\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})\\
&\quad\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}}\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.\\
&\qquad\left.+\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)\right\}-2n\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})\\
&\quad\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}}\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.\\
&\qquad\left.+\epsilon_t\ell(f(x_t),y'_t)-\frac{1}{2}\sum_{s=1}^{t-1}\ell(f(x_s),y_s)\right\}\\
&\qquad+\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.\\
&\qquad\left.-\epsilon_t\ell(f(x_t),y'_t)-\frac{1}{2}\sum_{s=1}^{t-1}\ell(f(x_s),y_s)\right\}-2n\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})
\end{aligned}
$$

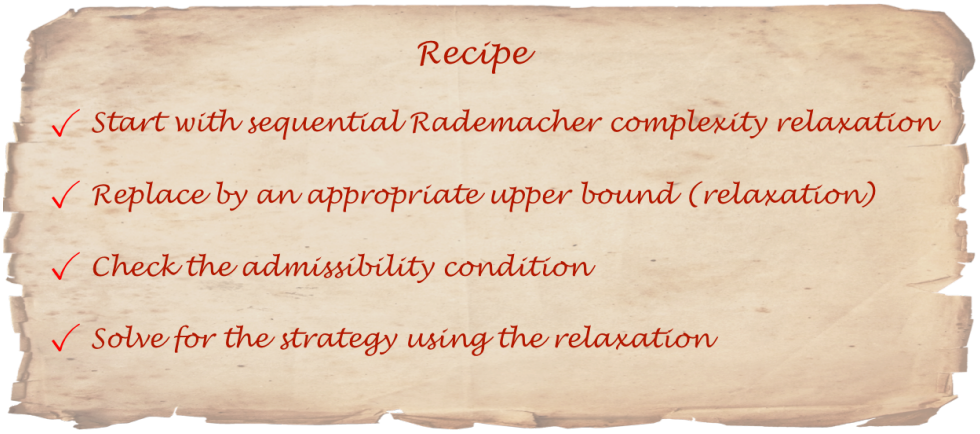
$$= 2 \sup_{y_t \in \mathcal{Y}} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{s=t+1}^{n} \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) \right.$$
$$\left. + \epsilon_t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

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,

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

This shows admissibility. From the earlier proposition and dividing throughout by n, we conclude the final statement. □

3 The Recipe



1. Write down sequential Rademacher relaxation for the given problem (in a malleable form).
2. Move to upper bound $\mathbf{Rel}_n$ such that $\forall t \in [n]$ and for all $x_{1:t}$, $y_{1:t}$,

$$\mathbf{Rad}_n\left(x_{1:t}, y_{1:t}\right) \leq \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)$$

(notice this ensures that initial condition is satisfied, this is half the work).

3. Two equivalent ways of checking admissibility condition, $\forall x_t \in \mathcal{X}$:

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

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

4. Algorithm: solve $q_t(x_t) = \operatorname*{argmin}_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t} \left\{ \mathbb{E}_{\hat{y}_t \sim q_t} \left[\ell(\hat{y}_t, y_t) \right] + \mathbf{Rel}_n \left(x_{1:t}, y_{1:t} \right) \right\}$

4 Online Linear Optimization: Euclidean space

$\mathcal{F} = \{\mathbf{f} : \|\mathbf{f}\|_2 \leq 1\}$, $\mathcal{D} = \{\nabla : \|\nabla\|_2 \leq 1\}$

Step 1

$$\begin{aligned}
\mathbf{Rad}_n \left(x_{1:t}, y_{1:t} \right) &= \sup_{\nabla} \mathbb{E}_{\epsilon_{t+1}, \ldots, \epsilon_n} \sup_{\mathbf{f} \in \mathcal{F}} \left[\left\langle \mathbf{f}, 2 \sum_{s=t+1}^{n} \epsilon_s \nabla_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \nabla_s \right\rangle \right] \\
&= \sup_{\nabla} \mathbb{E}_{\epsilon_{t+1}, \ldots, \epsilon_n} \left\| 2 \sum_{s=t+1}^{n} \epsilon_s \nabla_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \nabla_s \right\|_2 \\
&= \sup_{\nabla} \mathbb{E}_{\epsilon_{t+1}, \ldots, \epsilon_n} \sqrt{\left\| 2 \sum_{s=t+1}^{n} \epsilon_s \nabla_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \nabla_s \right\|_2^2}
\end{aligned}$$

Step 2

$$\begin{aligned}
\mathbf{Rad}_n \left(\nabla_{1:t} \right) &\leq \sup_{\nabla} \sqrt{\mathbb{E}_{\epsilon_{t+1}, \ldots, \epsilon_n} \left\| 2 \sum_{s=t+1}^{n} \epsilon_s \nabla_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \nabla_s \right\|_2^2} \\
&= \sup_{\nabla} \sqrt{\left\| \sum_{s=1}^{t} \nabla_s \right\|_2^2 + \mathbb{E}_{\epsilon_{t+1}, \ldots, \epsilon_n} \left[\text{Cross terms} \right] + 4 \mathbb{E}_{\epsilon_{t+1}, \ldots, \epsilon_n} \left[\sum_{s=t+1}^{n} \| \nabla_{s-t}(\epsilon_{t+1:s-1})) \|_2^2 \right]} \\
&= \sup_{\nabla} \sqrt{\left\| \sum_{s=1}^{t} \nabla_s \right\|_2^2 + 4 \mathbb{E}_{\epsilon_{t+1}, \ldots, \epsilon_n} \left[\sum_{s=t+1}^{n} \| \nabla_{s-t}(\epsilon_{t+1:s-1})) \|_2^2 \right]} \\
&\leq \sqrt{\left\| \sum_{s=1}^{t} \nabla_s \right\|_2^2 + 4(n-t)} =: \mathbf{Rel}_n \left(\nabla_{1:t} \right)
\end{aligned}$$

Step 3 & 4

$$\inf_{\mathbf{f}_t}\sup_{\nabla_t}\left\{\langle \mathbf{f}_t,\nabla_t\rangle+\mathbf{Rel}_n\left(\nabla_{1:t}\right)\right\}=\inf_{\mathbf{f}_t}\sup_{\nabla_t}\left\{\langle \mathbf{f}_t,\nabla_t\rangle+\sqrt{\left\|\sum_{s=1}^{t}\nabla_s\right\|_2^2+4(n-t)}\right\}$$

$$=\inf_{\mathbf{f}_t}\sup_{\nabla_t}\left\{\langle \mathbf{f}_t,\nabla_t\rangle+\sqrt{\left\|\sum_{s=1}^{t-1}\nabla_s\right\|_2^2+2\left\langle\sum_{s=1}^{t-1}\nabla_s,\nabla_t\right\rangle+\|\nabla_t\|_2^2+4(n-t)}\right\}$$

$$\leq\inf_{\mathbf{f}_t}\sup_{\nabla_t}\left\{\langle \mathbf{f}_t,\nabla_t\rangle+\sqrt{\left\|\sum_{s=1}^{t-1}\nabla_s\right\|_2^2+2\left\langle\sum_{s=1}^{t-1}\nabla_s,\nabla_t\right\rangle+4(n-t+1)}\right\}$$

Now in the above note that the second term depends on ∇_t only through $\left\langle\sum_{s=1}^{t-1}\nabla_s,\nabla_t\right\rangle$. This means that if $\mathbf{f}_t$ has any component orthogonal to $\sum_{s=1}^{t-1}\nabla_s$ then ∇_t can gain on the first term without loosing on the second term (as the component of ∇_t that increases first term is perpendicular to the second term). Hence $\mathbf{f}_t$ has to be of form $\mathbf{f}_t=-\alpha\sum_{s=1}^{t-1}\nabla_s$ for some positive α. Hence

$$\inf_{\mathbf{f}_t}\sup_{\nabla_t}\left\{\langle \mathbf{f}_t,\nabla_t\rangle+\mathbf{Rel}_n\left(\nabla_{1:t}\right)\right\}$$

$$=\inf_{\alpha>0}\sup_{\nabla_t}\left\{-\alpha\underbrace{\left\langle\sum_{s=1}^{t-1}\nabla_s,\nabla_t\right\rangle}_{\beta}+\sqrt{\left\|\sum_{s=1}^{t-1}\nabla_s\right\|_2^2+2\left\langle\sum_{s=1}^{t-1}\nabla_s,\nabla_t\right\rangle+4(n-t+1)}\right\}$$

$$\leq\inf_{\alpha>0}\sup_{\beta\in\mathbb{R}}\left\{-\alpha\beta+\sqrt{\left\|\sum_{s=1}^{t-1}\nabla_s\right\|_2^2+2\beta+4(n-t+1)}\right\}$$

Taking derivative to optimize over β for a given α we see that β is optimized when,

$$-\alpha+\frac{1}{\sqrt{\left\|\sum_{s=1}^{t-1}\nabla_s\right\|_2^2+2\beta+4(n-t+1)}}=0$$

Hence if we use

$$\alpha=\frac{1}{\sqrt{\left\|\sum_{s=1}^{t-1}\nabla_s\right\|_2^2+4(n-t+1)}}$$

then clearly the corresponding β that maximizes is at $\beta=0$. Hence,

$$\inf_{\mathbf{f}_t}\sup_{\nabla_t}\left\{\langle \mathbf{f}_t,\nabla_t\rangle+\mathbf{Rel}_n\left(\nabla_{1:t}\right)\right\}\leq\sup_{\beta\in\mathbb{R}}\left\{-\frac{\beta}{\sqrt{\left\|\sum_{s=1}^{t-1}\nabla_s\right\|_2^2+4(n-t+1)}}+\sqrt{\left\|\sum_{s=1}^{t-1}\nabla_s\right\|_2^2+2\beta+4(n-t+1)}\right\}$$

$$\leq\sqrt{\left\|\sum_{s=1}^{t-1}\nabla_s\right\|_2^2+4(n-t+1)}=\mathbf{Rel}_n\left(\nabla_{1:t-1}\right)$$

Algorithm is given by

$$\mathbf{f}_t = -\alpha \sum_{s=1}^{t-1} \nabla_s = -\frac{\sum_{s=1}^{t-1} \nabla_s}{\sqrt{\left\| \sum_{s=1}^{t-1} \nabla_s \right\|_2^2 + 4(n-t+1)}}$$

Notice that we don't need any projection, the solutions automatically have norm at most 1. The final guarantee we get is

$$\mathbb{E}\left[\mathrm{Reg}_n\right] \leq \frac{1}{n}\mathbf{Rel}_n\left(\cdot\right) = \frac{1}{n}\sqrt{4n} = \frac{2}{\sqrt{n}}$$

This gives an alternative for gradient descent and can be used for online convex optimization. For other norms, as long as the dual norm squared is a strongly-smooth function (or equivalently the norm squared is a strongly convex function) the same technique can be used where the equality due to Pythagorus theorem in the proof is replaced by inequality due to strong smoothness of norm squared. This can also be viewed as a modified, projection free form of gradient descent with automatically tuned step-sizes. The key thing to note is that the step size depends on past gradients and so if sequence is nicer, we take stronger steps.

5 Other Examples

1. Exponential weights style algorithm:

$$\mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right) = \inf_{\lambda>0}\left\{\frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\lambda\sum_{s=1}^{t}\ell(f(x_s), y_s)\right)\right) + 2\lambda(n-t)\right\}$$

2. Follow the Regularized Leader: Strongly convex R

$$\mathbf{Rel}_n\left(\nabla_{1:t}\right) = \inf_{\lambda>0}\left\{\frac{R^2}{\lambda} - \inf_{f\in\mathcal{F}}\left\{\sum_{s=1}^{t}\langle f, \nabla_s\rangle + \frac{1}{\lambda}\mathbf{R}(f)\right\} + 2\lambda(n-t)\right\}$$

3. Mirror Descent

$$\mathbf{Rel}_n\left(\nabla_{1:t}\right) = \inf_{\eta>0}\left\{\sup_{\mathbf{f}\in\mathcal{F}}\left\{\sum_{i=1}^{t}\langle \mathbf{f}, -\nabla_i\rangle + \frac{1}{\eta}\Delta_{\mathbf{R}}(\mathbf{f}|\nabla\mathbf{R}(\hat{\mathbf{y}}_t) - \eta\nabla_t)\right\} + 2\eta(n-t)\right\}$$

4. Also algorithms like follow the perturbed leader.

5. Fast rates can be obtained using this recipe by starting from an offset version of sequential Rademacher relaxation instead.