Machine Learning Theory (CS 6783)

Lecture 21: Relaxations for Online Learning

1 Recap

- We obtained a fairly general picture for online convex optimization : GD, MD, online Newton's method
- Online methods with O2B for many of these problems were optimal both in terms of learning rates and efficiency even for statistical learning
- How to derive algorithms for the general online learning setting?
- All the results in the first part of the course were non-constructive, In fact, the application of minimax theorem to the sequence meant we moved to dual game right at the start, so not easy to extract strategy.
  1. Primal game is conducive to think about designing algorithms
  2. Dual game is conducive for proving upper bounds and obtain probabilistic tools
- Closer look at Minimax rates reveal a recursive definition for value and hence automatically to a succinct form for online learning algorithms. Conditional Value :

$$\mathbf{V}_n(x_{1:t}, y_{1:t}) = \left\langle\!\!\left\langle \sup_{x_i \in \mathcal{X}} \inf_{q_i \in \Delta(\mathcal{Y})} \sup_{y_i \in \mathcal{Y}} \mathbb{E}_{\hat{y}_i \sim q_i} \right\rangle\!\!\right\rangle_{i=t+1}^{n} \left[\sum_{i=t+1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right]$$

  Minimax Strategy :

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

  Pros :
  1. We can apply minimax swap and go to the dual game from $t+1$ onwards to n, in definition of $\mathbf{V}_n(x_{1:t}, y_{1:t})$
  2. If $\mathbf{V}_n(x_{1:t}, y_{1:t})$ can be written down in a succinct form or better yet is computationally tractable, then there is hope to obtain minimax algorithm

  Cons:
  1. Except for very special problems (Eg. bit prediction example from lecture 12 for instance), $\mathbf{V}_n(x_{1:t}, y_{1:t})$ does not have a succinct form
  2. We don't know the exact minimax algorithm even for experts type problem.
- Can we use all the theory developed so far to provide a schema for designing learning algorithms for a general problem at hand?

2 Relaxations

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

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.

Initial condition :

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

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

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

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,*

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t\sim q_t}\left[\ell(\hat{y}_t, y_t)\right] - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(f(x_t), y_t) \le \tfrac{1}{n}\mathbf{Rel}_n(\cdot)$$

and if the loss is bounded then by application of Hoeffding-Azuma inequality, for any $\delta > 0$ with probability at least $1-\delta$ over our randomization,

$$\frac{1}{n}\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(f(x_t), y_t) \le \tfrac{1}{n}\mathbf{Rel}_n(\cdot) + O\left(\sqrt{\frac{\log 1/\delta}{n}}\right)$$

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

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

by admissibility condition,

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

As for the high probability version, note that is loss is bounded by some B, then by Hoeffding Azuma bound on average of martingale difference sequences,

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}_{\hat{y}_t\sim q_t(x_t)}\left[\ell(\hat{y}_t,y_t)\right]-\ell(\hat{y}_t,y_t)\right)\right|>\epsilon\right)\leq 2\exp\left(-n\epsilon^2/2B\right)$$

□

Remark 2.1. $\mathbf{V}_n$ *is an admissible relaxation and corresponding relaxation based algorithms are the minimax algorithms.*

Outline

1. From theory developed so far, arrive at relaxations
2. Each admissible relaxation comes with an algorithm
3. In fact we don't need to exactly solve for q_t as defined above, all proofs go through even if we are able to pick q_t such that,

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

3 Sequential Rademacher Relaxation

Definition 1. *Define the sequential Rademacher relaxation as*

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

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* $n-t$*.*

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

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

Proof. As for initial condition note that,

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

Now to check admissibility, note that

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

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

$$
\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,

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

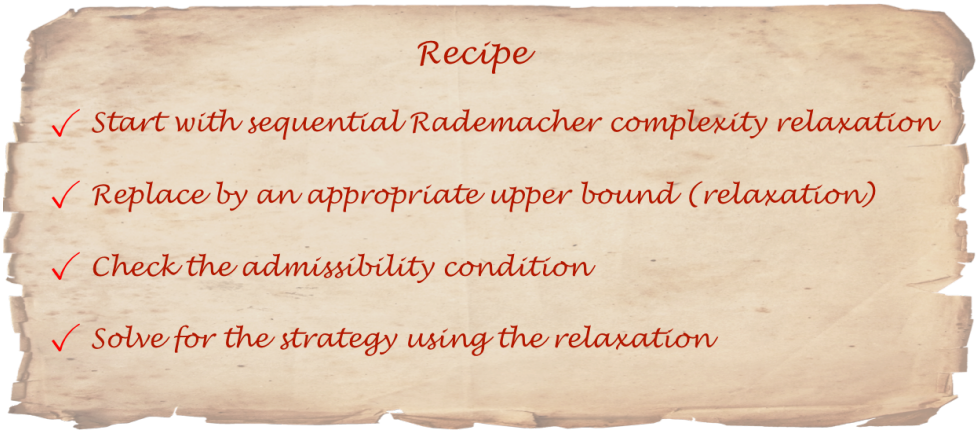
This shows admissibility. From the earlier proposition, regret is bounded by

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

□

Notice the similarity of proof above with the proof of upper bounding minimax rate by sequential Rademacher complexity. It is in fact a one step at a time version.

4 The Recipe



- Basically the proof we used to bound sequential Rademacher complexity of various problems hide in them a hierarchy of relaxations. We shall review these to derive algorithms
- We shall recover several known algorithms, some we reviewed earlier in class, but we in fact can obtain, better parameter free version of these by stopping at an earlier step in the proof of the bound so we get tighter bounds but also tractable relaxations.
- Admissibility condition easy to check based on minimax theorem once we swap roles of players for round t.