Machine Learning Theory (CS 6783)

Lecture 22: Relaxations & deriving algorithms

1 Recap

1. $\mathbf{Rel}_n : \bigcup_{t=0}^{n} \mathcal{X}^t \times \mathcal{Y}^t \mapsto \mathbb{R}$ is admissible if

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

and

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

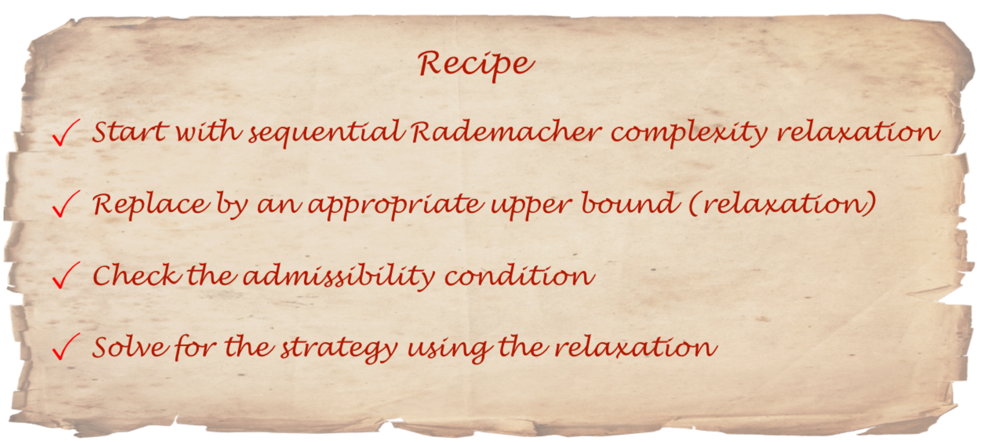
2. Algorithm : $q_t = \underset{q \in \Delta(\mathcal{Y})}{\text{argmin}} \sup_{y_t} \left\{\mathbb{E}_{\hat{y}_t \sim q}\left[\ell(\hat{y}_t, y_t)\right] + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)\right\}$

3. Bound : $\mathbb{E}\left[\mathrm{Reg}_n\right] \leq \frac{1}{n}\mathbf{Rel}_n\left(\cdot\right)$

4. Sequential Rademacher relaxation is admissible and gives the sequential Rademacher bound

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

2 The Recipe



1. Write down sequential Rademacher relaxation for the given problem (in a malleable form).

2. Mover to relaxation 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 also goes half the way in getting feasible relaxation).

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 optimization:

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

3 Example : Bit Prediction

$\mathcal{X} = \{\}$, $\mathcal{Y} = \{\pm 1\}$, $\mathcal{F} \subset \{\pm 1\}^n$, $\ell(\hat{y}, y) = \mathbb{1}_{\{\hat{y} \neq y\}} = \frac{1 - \hat{y}y}{2}$.

Steps 1 & 2 (we can use exact seq. Rademacher relaxation)

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

using the contraction lemma for removing binary tree we proved,

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

Step 3 Checking admissibility: This relaxation is basically the sequential Rademacher relaxation that just simplifies for the bit prediction problem. Hence as we already proved, it is admissible.

Step 4 Now for the algorithm :

$$\begin{aligned}
q_t &= \underset{q_t \in [0,1]}{\operatorname{argmin}} \max_{y_t \in \{\pm 1\}} \left\{\mathbb{E}_{\hat{y}_t \sim q_t}\left[\mathbb{1}_{\{\hat{y}_t \neq y_t\}}\right] + \mathbf{Rel}_n\left(y_{1:t}\right)\right\} \\
&= \inf_{q_t \in [0,1]} \max_{y_t \in \{\pm 1\}} \left\{q_t \mathbb{1}_{\{1 \neq y_t\}} + (1 - q_t) \mathbb{1}_{\{-1 \neq y_t\}} + \mathbf{Rel}_n\left(y_{1:t}\right)\right\} \\
&= \inf_{q_t \in [0,1]} \max \left\{(1 - q_t) + \mathbf{Rel}_n\left(y_{1:t-1}, +1\right), q_t + \mathbf{Rel}_n\left(y_{1:t-1}, -1\right)\right\}
\end{aligned}$$

Minimized when the two terms are equal and so,

$$q_t = \frac{1}{2}\left(1 + \mathbf{Rel}_n\left(y_{1:t-1}, +1\right) - \mathbf{Rel}_n\left(y_{1:t-1}, -1\right)\right)$$

Now note that $\mathbf{Rel}_n$ involves average over Rademacher variables but by Hoeffding bound we can replace it on each round by a single draw and pay only $O(1/\sqrt{n})$ additive factor more. Note that beyond the extra factor 2 this is the same algorithm as the minimax algorithm we derived in earlier Lecture. Finally the bound we get is

$$\mathbb{E}\left[\mathrm{Reg}_n\right] \le \tfrac{1}{n}\mathbf{Rel}_n\left(\cdot\right) = \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t f_t\right]$$

4 Finite Experts

$|\mathcal{F}| < \infty$ and ℓ bounded by say 1.

Step 1

$$\begin{aligned}\mathbf{Rad}_n\left(x_{1:t}, y_{1:t}\right) &= \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1},\dots,\epsilon_n} \sup_{f\in\mathcal{F}}\left[2\sum_{s=t+1}^n \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^t \ell(f(x_s), y_s)\right]\\ &= \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1:n}} \inf_{\lambda>0}\frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(2\lambda\sum_{s=t+1}^n \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) - \lambda\sum_{s=1}^t \ell(f(x_s), y_s)\right)\right)\end{aligned}$$

That is we move to limit of soft-max.

Step 2

$$\mathbf{Rad}_n\left(x_{1:t}, y_{1:t}\right) = \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1:n}} \inf_{\lambda>0}\frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(2\lambda\sum_{s=t+1}^n \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) - \lambda\sum_{s=1}^t \ell(f(x_s), y_s)\right)\right)$$

$$\le \inf_{\lambda>0}\sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_{\epsilon_{t+1:n}} \frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(2\lambda\sum_{s=t+1}^n \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1})) - \lambda\sum_{s=1}^t \ell(f(x_s), y_s)\right)\right)$$

$$\le \inf_{\lambda>0}\sup_{\mathbf{x},\mathbf{y}} \frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\mathbb{E}_{\epsilon_{t+1:n}}\exp\left(2\lambda\sum_{s=t+1}^n \epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1}))\right)\cdot\exp\left(-\lambda\sum_{s=1}^t \ell(f(x_s), y_s)\right)\right)$$

$$= \inf_{\lambda>0}\sup_{\mathbf{x},\mathbf{y}} \frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\mathbb{E}_{\epsilon_{t+1:n}}\left[\prod_{s=t+1}^n \mathbb{E}_{\epsilon_t}\left[\exp\left(2\lambda\epsilon_s \ell(f(\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})), \mathbf{y}_{s-t}(\epsilon_{t+1:s-1}))\right)\right]\right]\cdot\exp\left(-\lambda\sum_{s=1}^t \ell(f(x_s), y_s)\right)\right)$$

using $\mathbb{E}_\epsilon\left[e^{\epsilon x}\right] \le e^{x^2/2}$ exactly as in finite lemma proof, since loss is bounded,

$$\le \inf_{\lambda>0}\sup_{\mathbf{x},\mathbf{y}} \frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\prod_{s=t+1}^n \exp\left(2\lambda^2\right)\cdot\exp\left(-\lambda\sum_{s=1}^t \ell(f(x_s), y_s)\right)\right)$$

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

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

Step 3. (admissibility)

$$
\begin{aligned}
&\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] + \mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)\right\}\\
&\quad= \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[\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\}\right]\right\}\\
&\quad\le \inf_{\lambda>0}\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[\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]\right\}\\
&\quad= \inf_{\lambda>0}\sup_{p_t}\left\{\frac{1}{\lambda}\log(\exp(\lambda\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[\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]\right\}\\
&\quad= \inf_{\lambda>0}\sup_{p_t}\left\{\mathbb{E}_{y_t\sim p_t}\left[\frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(\lambda\inf_{\hat{y}_t\in\mathcal{Y}} \mathbb{E}_{y_t\sim p_t}\left[\ell(\hat{y}_t,y_t)\right]-\lambda\sum_{s=1}^{t}\ell(f(x_s),y_s)\right)\right) + 2\lambda(n-t)\right]\right\}\\
&\quad\le \inf_{\lambda>0}\sup_{p_t}\left\{\mathbb{E}_{y_t\sim p_t}\left[\frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(\lambda\left(\mathbb{E}_{y_t\sim p_t}\left[\ell(f(x_t),y_t)\right]-\ell(f(x_t),y_t)\right)-\lambda\sum_{s=1}^{t-1}\ell(f(x_s),y_s)\right)\right) + 2\lambda(n-t)\right]\right\}\\
&\quad\le \inf_{\lambda>0}\sup_{p_t}\left\{\frac{1}{\lambda}\log\left(\mathbb{E}_{y_t,y_t'\sim p_t}\left[\sum_{f\in\mathcal{F}}\exp\left(\lambda\left(\ell(f(x_t),y_t')-\ell(f(x_t),y_t)\right)-\lambda\sum_{s=1}^{t-1}\ell(f(x_s),y_s)\right)\right]\right) + 2\lambda(n-t)\right\}\\
&\quad= \inf_{\lambda>0}\sup_{p_t}\left\{\frac{1}{\lambda}\log\left(\mathbb{E}_{y_t,y_t'\sim p_t}\mathbb{E}_{\epsilon_t}\left[\sum_{f\in\mathcal{F}}\exp\left(\lambda\epsilon_t\left(\ell(f(x_t),y_t')-\ell(f(x_t),y_t)\right)-\lambda\sum_{s=1}^{t-1}\ell(f(x_s),y_s)\right)\right]\right) + 2\lambda(n-t)\right\}
\end{aligned}
$$

$(e^x+e^{-x})/2 \le e^{x^2/2}$ and loss is bounded by 1 and so,

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

Step 4. (algorithm) We get a parameter free version of exponential weights algorithm,

$$
\lambda_t^* = \operatorname*{argmin}_{\lambda>0}\left\{\frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\lambda\sum_{s=1}^{t-1}\ell(f(x_s),y_s)\right)\right) + 2\lambda(n-t+1)\right\} \;\;\&\;\; q_t^* \propto \exp\left(-\lambda_t^*\sum_{s=1}^{t-1}\ell(f(x_s),y_s)\right)
$$

$$
\begin{aligned}
&\sup_{y_t} \left\{ \mathbb{E}_{\hat{y}_t \sim q_t^*} \left[\ell(\hat{y}_t, y_t) \right] + \mathbf{Rel}_n \left(x_{1:t}, y_{1:t} \right) \right\} \\
&\quad = \sup_{y_t} \left\{ \mathbb{E}_{\hat{y}_t \sim q_t^*} \left[\ell(\hat{y}_t, y_t) \right] + \inf_{\lambda > 0} \left\{ \frac{1}{\lambda} \log \left(\sum_{f \in \mathcal{F}} \exp \left(-\lambda \sum_{s=1}^{t} \ell(f(x_s), y_s) \right) \right) + 2\lambda(n-t) \right\} \right\} \\
&\quad \leq \sup_{y_t} \left\{ \mathbb{E}_{\hat{y}_t \sim q_t^*} \left[\ell(\hat{y}_t, y_t) \right] + \frac{1}{\lambda} \log \left(\sum_{f \in \mathcal{F}} \exp \left(-\lambda_t^* \sum_{s=1}^{t} \ell(f(x_s), y_s) \right) \right) \right\} + 2\lambda_t^*(n-t) \\
&\quad = \sup_{y_t} \left\{ \mathbb{E}_{\hat{y}_t \sim q_t^*} \left[\ell(\hat{y}_t, y_t) \right] + \frac{1}{\lambda} \log \left(\mathbb{E}_{f \sim q_t^*} \left[\exp(-\lambda_t^* \ell(f(x_t), y_t)) \right] \right) \right. \\
&\qquad\qquad \left. + \frac{1}{\lambda_t^*} \log \left(\sum_{f \in \mathcal{F}} \exp \left(-\lambda \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right) \right) + 2\lambda_t^*(n-t) \right\} \\
&\quad = \sup_{y_t} \left\{ \frac{1}{\lambda_t^*} \log \left(\exp \left(\lambda_t^* \mathbb{E}_{\hat{y}_t \sim q_t^*} \left[\ell(\hat{y}_t, y_t) \right] \right) \right) + \frac{1}{\lambda_t^*} \log \left(\mathbb{E}_{f \sim q_t^*} \left[\exp(-\lambda_t^* \ell(f(x_t), y_t)) \right] \right) \right. \\
&\qquad\qquad \left. + \frac{1}{\lambda_t^*} \log \left(\sum_{f \in \mathcal{F}} \exp \left(-\lambda_t^* \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right) \right) + 2\lambda_t^*(n-t) \right\} \\
&\quad \leq \frac{1}{\lambda_t^*} \log \left(\sum_{f \in \mathcal{F}} \exp \left(-\lambda_t^* \sum_{s=1}^{t-1} \ell(f(x_s), y_s) \right) \right) + 2\lambda_t^*(n-t+1) \Bigg\} = \mathbf{Rel}_n \left(x_{1:t-1}, y_{1:t-1} \right)
\end{aligned}
$$

Finally, we get the guarantee,

$$
\mathbb{E} \left[\mathrm{Reg}_n \right] \leq \tfrac{1}{n} \mathbf{Rel}_n \left(\cdot \right) = \frac{1}{n} \inf_{\lambda > 0} \left\{ \frac{1}{\lambda} \log \left(|\mathcal{F}| \right) + 2\lambda n \right\} = \sqrt{\frac{8 \log |\mathcal{F}|}{n}}
$$