

**CS281B/Stat241B. Statistical Learning Theory. Lecture
18.**

Peter Bartlett

Recall: Regularization

Regularized minimization

Consider the family of strategies of the form:

$$a_{t+1} = \arg \min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^t \ell_s(a) + R(a) \right).$$

The regularizer $R : \mathbb{R}^d \rightarrow \mathbb{R}$ is strictly convex and differentiable.

Define $\Phi_0 = R$, $\Phi_t = \Phi_{t-1} + \eta \ell_t$, so that $a_{t+1} = \arg \min_{a \in \mathcal{A}} \Phi_t(a)$.

If we replace ℓ_t by $\nabla \ell_t(a_t)$, this leads to an upper bound on regret. Thus, we can assume **linear** ℓ_t .

Recall: Regret of Regularization Methods

Theorem: For $\mathcal{A} = \mathbb{R}^d$, regularized minimization suffers regret against any $a \in \mathcal{A}$ of

$$\begin{aligned} & \sum_{t=1}^n \ell_t(a_t) - \sum_{t=1}^n \ell_t(a) \\ &= \frac{1}{\eta} \sum_{t=1}^n \left(D_{\Phi_{t-1}}(a, a_t) - D_{\Phi_t}(a, a_{t+1}) + D_{\Phi_t}(a_t, a_{t+1}) \right), \end{aligned}$$

and thus

$$\hat{L}_n \leq \inf_{a \in \mathbb{R}^d} \left(\sum_{t=1}^n \ell_t(a) + \frac{D_R(a, a_1)}{\eta} \right) + \frac{1}{\eta} \sum_{t=1}^n D_{\Phi_t}(a_t, a_{t+1}).$$

Regularization Methods: Varying η

Theorem: Define

$$a_{t+1} = \arg \min_{a \in \mathbb{R}^d} \left(\sum_{t=1}^n \underbrace{\eta_t \ell_t(a) + R(a)}_{\Phi_t(a)} \right).$$

For any $a \in \mathbb{R}^d$,

$$\hat{L}_n - \sum_{t=1}^n \ell_t(a) \leq \sum_{t=1}^n \frac{1}{\eta_t} \left(D_{\Phi_t}(a_t, a_{t+1}) + D_{\Phi_{t-1}}(a, a_t) - D_{\Phi_t}(a, a_{t+1}) \right).$$

Regularization Methods: Varying η

If we linearize the ℓ_t , we have

$$\hat{L}_n - \sum_{t=1}^n \ell_t(a) \leq \sum_{t=1}^n \frac{1}{\eta_t} (D_R(a_t, a_{t+1}) + D_R(a, a_t) - D_R(a, a_{t+1})).$$

The $D_R(a, a_t)$ terms no longer telescope. If ℓ_t are strongly convex, we can do better.

Regularization Methods: Strongly Convex Losses

Theorem: If ℓ_t is σ -strongly convex wrt R , that is, for all $a, b \in \mathbb{R}^d$,

$$\ell_t(a) \geq \ell_t(b) + \nabla \ell_t(b) \cdot (a - b) + \frac{\sigma}{2} D_R(a, b),$$

then for any $a \in \mathbb{R}^d$, this strategy with $\eta_t = \frac{2}{t\sigma}$ has regret

$$\hat{L}_n - \sum_{t=1}^n \ell_t(a) \leq \sum_{t=1}^n \frac{1}{\eta_t} D_R(a_t, a_{t+1}).$$

Strongly Convex Losses: Proof idea

$$\begin{aligned} & \sum_{t=1}^n (\ell_t(a_t) - \ell_t(a)) \\ & \leq \sum_{t=1}^n \left(\nabla \ell_t(a_t) \cdot (a_t - a) - \frac{\sigma}{2} D_R(a, a_t) \right) \\ & \leq \sum_{t=1}^n \frac{1}{\eta_t} \left(D_R(a_t, a_{t+1}) + D_R(a, a_t) - D_R(a, a_{t+1}) - \frac{\eta_t \sigma}{2} D_R(a, a_t) \right) \\ & \leq \sum_{t=1}^n \frac{1}{\eta_t} D_R(a_t, a_{t+1}) + \sum_{t=2}^n \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \frac{\sigma}{2} \right) D_R(a, a_t) \\ & \quad + \left(\frac{1}{\eta_1} - \frac{\sigma}{2} \right) D_R(a, a_1). \end{aligned}$$

And choosing $\eta_t = 2/(t\sigma)$ eliminates the second and third terms.

Strongly Convex Losses

Example: For $R(a) = \frac{1}{2}\|a\|^2$, we have

$$\hat{L}_n - L_n^* \leq \frac{1}{2} \sum_{t=1}^n \frac{1}{\eta_t} \|\eta_t \nabla \ell_t\|^2 \leq \sum_{t=1}^n \frac{G^2}{t\sigma} = O\left(\frac{G^2}{\sigma} \log n\right).$$

Key Point: When the loss is strongly convex wrt the regularizer, the regret rate can be faster; in the case of quadratic R (and ℓ_t), it is $O(\log n)$, versus $O(\sqrt{n})$.

Probabilistic Prediction Setting

Recall this probabilistic formulation of a prediction problem:

- There is a sample of size n drawn i.i.d. from an unknown probability distribution P on $\mathcal{X} \times \mathcal{Y}$:
 $(X_1, Y_1), \dots, (X_n, Y_n)$.
- Some method chooses $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$.
- It suffers regret

$$\mathbf{E}\ell(\hat{f}(X), Y) - \min_{f \in F} \mathbf{E}\ell(f(X), Y).$$

- Here, F is a class of functions from \mathcal{X} to \mathcal{Y} .

Online to Batch Conversion

- Suppose we have an online strategy that, given observations $\ell_1, \dots, \ell_{t-1}$, produces $a_t = A(\ell_1, \dots, \ell_{t-1})$.
- Can we convert this to a method that is suitable for a probabilistic setting? That is, if the ℓ_t are chosen i.i.d., can we use A 's choices a_t to come up with an $\hat{a} \in \mathcal{A}$ so that

$$\mathbf{E}l_1(\hat{a}) - \min_{a \in \mathcal{A}} \mathbf{E}l_1(a)$$

is small?

- Consider the following simple randomized method:
 1. Pick T uniformly from $\{0, \dots, n\}$.
 2. Let $\hat{a} = A(\ell_{T+1}, \dots, \ell_n)$.

Online to Batch Conversion

Theorem: If A has a regret bound of C_{n+1} for sequences of length $n + 1$, then for any stationary process generating the $\ell_1, \dots, \ell_{n+1}$, this method satisfies

$$\mathbf{E} \ell_{n+1}(\hat{a}) - \min_{a \in \mathcal{A}} \mathbf{E} \ell_{n+1}(a) \leq \frac{C_{n+1}}{n+1}.$$

(Notice that the expectation averages also over the randomness of the method.)

Online to Batch Conversion

$$\begin{aligned}\mathbf{E}l_{n+1}(\hat{a}) &= \mathbf{E}l_{n+1}(A(l_{T+1}, \dots, l_n)) \\ &= \mathbf{E} \frac{1}{n+1} \sum_{t=0}^n l_{n+1}(A(l_{t+1}, \dots, l_n)) \\ &= \mathbf{E} \frac{1}{n+1} \sum_{t=0}^n l_{n-t+1}(A(l_1, \dots, l_{n-t})) \\ &= \mathbf{E} \frac{1}{n+1} \sum_{t=1}^{n+1} l_t(A(l_1, \dots, l_{t-1})) \\ &\leq \mathbf{E} \frac{1}{n+1} \left(\min_a \sum_{t=1}^{n+1} l_t(a) + C_{n+1} \right) \\ &\leq \min_a \mathbf{E}l_t(a) + \frac{C_{n+1}}{n+1}.\end{aligned}$$

Online to Batch Conversion

Key Point:

- An online strategy with regret bound C_n can be converted to a batch method.

The regret per trial in the probabilistic setting is bounded by the regret per trial in the adversarial setting.

Optimal Regret

We have:

- a set of actions \mathcal{A} ,
- a set of loss functions \mathcal{L} .

At time t ,

- Player chooses an action a_t from \mathcal{A} .
- Adversary chooses $\ell_t : \mathcal{A} \rightarrow \mathbb{R}$ from \mathcal{L} .
- Player incurs loss $\ell_t(a_t)$.

Regret is the value of the game:

$$V_n(\mathcal{A}, \mathcal{L}) = \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_n} \sup_{\ell_n} \left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right).$$

Optimal Regret: Dual Game

Theorem: If \mathcal{A} is compact and all ℓ_t are convex, continuous functions, then

$$V_n(\mathcal{A}, \mathcal{L}) = \sup_P \mathbf{E} \left(\sum_{t=1}^n \inf_{a_t \in \mathcal{A}} \mathbf{E} [\ell_t(a_t) | \ell_1, \dots, \ell_{t-1}] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right),$$

where the supremum is over joint distributions P over sequences ℓ_1, \dots, ℓ_n in \mathcal{L}^n .

- As we'll see, this follows from a minimax theorem.
- Dual game: adversary plays first by choosing P .
- Value of the game is the difference between minimal conditional expected loss and minimal empirical loss.
- If P were i.i.d., this would be the difference between the minimal expected loss and the minimal empirical loss.

Optimal Regret: Extensions

- We can ensure convexity of the ℓ_t by allowing **mixed strategies**: replace \mathcal{A} by the set of probability distributions P on \mathcal{A} and replace $\ell(a)$ by $\mathbf{E}_{a \sim P} \ell(a)$.

Dual Game: Proof Idea

Theorem: [Sion, 1957] If \mathcal{A} is compact and for every $b \in \mathcal{B}$, $f(\cdot, b)$ is a convex-like,^a lower semi-continuous function, and for every $a \in \mathcal{A}$, $f(a, \cdot)$ is concave-like, then

$$\inf_{a \in \mathcal{A}} \sup_{b \in \mathcal{B}} f(a, b) = \sup_{b \in \mathcal{B}} \inf_{a \in \mathcal{A}} f(a, b).$$

We'll define \mathcal{B} as the set of probability distributions on \mathcal{L} and $f(a, b) = c + \mathbf{E}[\ell(a) + \phi(\ell)]$, and we'll assume that \mathcal{A} is compact and each $\ell \in \mathcal{L}$ is convex and continuous.

ℓ is convex-like [Fan, 1953]:

$$\forall a_1, a_2 \in \mathcal{A}, \alpha \in [0, 1], \exists a \in \mathcal{A}, \alpha \ell(a_1) + (1 - \alpha) \ell(a_2) \leq \ell(a).$$

Dual Game: Proof Idea

$$\begin{aligned} V_n(\mathcal{A}, \mathcal{L}) &= \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_n} \sup_{\ell_n} \left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \\ &= \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_n} \sup_{P_n} \mathbf{E} \left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right), \end{aligned}$$

because allowing mixed strategies does not help the adversary.

Dual Game: Proof Idea

$$\begin{aligned}
 V_n(\mathcal{A}, \mathcal{L}) &= \inf_{a_1} \sup_{l_1} \cdots \inf_{a_n} \sup_{l_n} \left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \\
 &= \inf_{a_1} \sup_{l_1} \cdots \inf_{a_{n-1}} \sup_{l_{n-1}} \inf_{a_n} \sup_{P_n} \mathbf{E} \left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \\
 &= \inf_{a_1} \sup_{l_1} \cdots \inf_{a_{n-1}} \sup_{l_{n-1}} \sup_{P_n} \inf_{a_n} \mathbf{E} \left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right),
 \end{aligned}$$

by Sion's generalization of von Neumann's minimax theorem.

Dual Game: Proof Idea

$$\begin{aligned}
 V_n(\mathcal{A}, \mathcal{L}) &= \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_n} \sup_{\ell_n} \left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \\
 &= \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_n} \sup_{P_n} \mathbf{E} \left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \\
 &= \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_{n-1}} \sup_{\ell_{n-1}} \sup_{P_n} \inf_{a_n} \mathbf{E} \left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \\
 &= \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_{n-1}} \sup_{P_{n-1}} \mathbf{E} \left(\sum_{t=1}^{n-1} \ell_t(a_t) + \right. \\
 &\quad \left. \sup_{P_n} \left(\inf_{a_n} \mathbf{E} [\ell_n(a_n) | \ell_1, \dots, \ell_{n-1}] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \right),
 \end{aligned}$$

splitting the sum and **allowing** the adversary a mixed strategy at round $n - 1$.

Dual Game: Proof Idea

$$\begin{aligned}
 V_n(\mathcal{A}, \mathcal{L}) &= \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_{n-1}} \sup_{P_{n-1}} \mathbf{E} \left(\sum_{t=1}^{n-1} \ell_t(a_t) + \right. \\
 &\quad \left. \sup_{P_n} \left(\inf_{a_n} \mathbf{E} [\ell_n(a_n) | \ell_1, \dots, \ell_{n-1}] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \right) \\
 &= \inf_{a_1} \sup_{\ell_1} \cdots \sup_{P_{n-1}} \inf_{a_{n-1}} \mathbf{E} \left(\sum_{t=1}^{n-1} \ell_t(a_t) + \right. \\
 &\quad \left. \sup_{P_n} \left(\inf_{a_n} \mathbf{E} [\ell_n(a_n) | \ell_1, \dots, \ell_{n-1}] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \right),
 \end{aligned}$$

applying Sion's minimax theorem again.

Dual Game: Proof Idea

$$\begin{aligned}
 V_n(\mathcal{A}, \mathcal{L}) &= \inf_{a_1} \sup_{\ell_1} \cdots \sup_{P_{n-1}} \inf_{a_{n-1}} \mathbf{E} \left(\sum_{t=1}^{n-1} \ell_t(a_t) + \right. \\
 &\quad \left. \sup_{P_n} \left(\inf_{a_n} \mathbf{E} [\ell_n(a_n) | \ell_1, \dots, \ell_{n-1}] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \right) \\
 &= \inf_{a_1} \sup_{\ell_1} \cdots \sup_{P_{n-2}} \inf_{a_{n-2}} \left(\mathbf{E} \sum_{t=1}^{n-2} \ell_t(a_t) + \right. \\
 &\quad \left. \sup_{P_{n-1}^n} \mathbf{E} \left(\sum_{t=n-1}^n \inf_{a_t} \mathbf{E} [\ell_t(a_t) | \ell_1, \dots, \ell_{t-1}] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \right) \\
 &\quad \vdots \\
 &= \sup_P \mathbf{E} \left(\sum_{t=1}^n \inf_{a_t} \mathbf{E} [\ell_t(a_t) | \ell_1, \dots, \ell_{t-1}] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right).
 \end{aligned}$$

Optimal Regret and Sequential Rademacher Averages

Theorem:

$$V_n(\mathcal{A}, \mathcal{L}) \leq 2 \sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^n \epsilon_t \ell_t(a),$$

where $\epsilon_1, \dots, \epsilon_n$ are independent Rademacher (uniform ± 1 -valued) random variables.

- Compare to the bound involving Rademacher averages in the probabilistic setting:

$$\text{excess risk} \leq c \mathbf{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(Y_t, f(X_t)) \right|.$$

- In the adversarial case, the choice of ℓ_t is deterministic, and can depend on $\epsilon_1, \dots, \epsilon_{t-1}$.

Sequential Rademacher Averages: Proof Idea

$$\begin{aligned} V_n(\mathcal{A}, \mathcal{L}) &= \sup_P \mathbf{E} \left(\sum_{t=1}^n \inf_{a_t \in \mathcal{A}} \mathbf{E} [\ell_t(a_t) | \ell_1, \dots, \ell_{t-1}] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \\ &\leq \sup_P \mathbf{E} \left(\sum_{t=1}^n \mathbf{E} [\ell_t(\hat{a}) | \ell_1, \dots, \ell_{t-1}] - \sum_{t=1}^n \ell_t(\hat{a}) \right), \end{aligned}$$

where \hat{a} minimizes $\sum_t \ell_t(a)$.

Sequential Rademacher Averages: Proof Idea

$$\begin{aligned} V_n(\mathcal{A}, \mathcal{L}) &= \sup_P \mathbf{E} \left(\sum_{t=1}^n \inf_{a_t \in \mathcal{A}} \mathbf{E} [\ell_t(a_t) | \ell_1, \dots, \ell_{t-1}] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right) \\ &\leq \sup_P \mathbf{E} \left(\sum_{t=1}^n \mathbf{E} [\ell_t(\hat{a}) | \ell_1, \dots, \ell_{t-1}] - \sum_{t=1}^n \ell_t(\hat{a}) \right) \\ &\leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^n (\mathbf{E} [\ell_t(a) | \ell_1, \dots, \ell_{t-1}] - \ell_t(a)). \end{aligned}$$

Sequential Rademacher Averages: Proof Idea

$$\begin{aligned} V_n(\mathcal{A}, \mathcal{L}) &\leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^n (\mathbf{E} [\ell_t(a) | \ell_1, \dots, \ell_{t-1}] - \ell_t(a)) \\ &= \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^n (\mathbf{E} [\ell'_t(a) | \ell_1, \dots, \ell_n] - \ell_t(a)), \end{aligned}$$

where ℓ'_t is a *tangent sequence*: conditionally independent of ℓ_t given $\ell_1, \dots, \ell_{t-1}$, with the same conditional distribution.

Sequential Rademacher Averages: Proof Idea

$$\begin{aligned} V_n(\mathcal{A}, \mathcal{L}) &\leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^n (\mathbf{E} [\ell_t(a) | \ell_1, \dots, \ell_{t-1}] - \ell_t(a)) \\ &= \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^n (\mathbf{E} [\ell'_t(a) | \ell_1, \dots, \ell_n] - \ell_t(a)) \\ &\leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^n (\ell'_t(a) - \ell_t(a)), \end{aligned}$$

moving the supremum inside the expectation.

Sequential Rademacher Averages: Proof Idea

$$\begin{aligned} V_n(\mathcal{A}, \mathcal{L}) &\leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^n (\ell'_t(a) - \ell_t(a)) \\ &= \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \left(\sum_{t=1}^{n-1} (\ell'_t(a) - \ell_t(a)) + \epsilon_n (\ell'_n(a) - \ell_n(a)) \right), \end{aligned}$$

for $\epsilon_n \in \{-1, 1\}$, since ℓ'_n has the same conditional distribution, given $\ell_1, \dots, \ell_{n-1}$, as ℓ_n .

Sequential Rademacher Averages: Proof Idea

$$\begin{aligned}
 V_n(\mathcal{A}, \mathcal{L}) &\leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^n (\ell'_t(a) - \ell_t(a)) \\
 &= \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \left(\sum_{t=1}^{n-1} (\ell'_t(a) - \ell_t(a)) + \epsilon_n (\ell'_n(a) - \ell_n(a)) \right) \\
 &= \sup_P \mathbf{E}_{\ell_1, \dots, \ell_{n-1}} \mathbf{E}_{\ell_n, \ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \left(\sum_{t=1}^{n-1} (\ell'_t(a) - \ell_t(a)) + \right. \\
 &\quad \left. \epsilon_n (\ell'_n(a) - \ell_n(a)) \right) \\
 &\leq \sup_P \mathbf{E}_{\ell_1, \dots, \ell_{n-1}} \sup_{\ell_n, \ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \left(\sum_{t=1}^{n-1} (\ell'_t(a) - \ell_t(a)) + \right. \\
 &\quad \left. \epsilon_n (\ell'_n(a) - \ell_n(a)) \right).
 \end{aligned}$$

Sequential Rademacher Averages: Proof Idea

$$\begin{aligned}
 V_n(\mathcal{A}, \mathcal{L}) &\leq \sup_P \mathbf{E}_{\ell_1, \dots, \ell_{n-1}} \mathbf{E}_{\ell_n, \ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \left(\sum_{t=1}^{n-1} (\ell'_t(a) - \ell_t(a)) + \right. \\
 &\quad \left. \epsilon_n (\ell'_n(a) - \ell_n(a)) \right) \\
 &\leq \sup_P \mathbf{E}_{\ell_1, \dots, \ell_{n-1}} \sup_{\ell_n, \ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \left(\sum_{t=1}^{n-1} (\ell'_t(a) - \ell_t(a)) + \right. \\
 &\quad \left. \epsilon_n (\ell'_n(a) - \ell_n(a)) \right) \\
 &\vdots \\
 &\leq \sup_{\ell_1, \ell'_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n, \ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \left(\sum_{t=1}^n \epsilon_t (\ell'_t(a) - \ell_t(a)) \right).
 \end{aligned}$$

Sequential Rademacher Averages: Proof Idea

$$\begin{aligned} V_n(\mathcal{A}, \mathcal{L}) &\leq \sup_{\ell_1, \ell'_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n, \ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \left(\sum_{t=1}^n \epsilon_t (\ell'_t(a) - \ell_t(a)) \right) \\ &= 2 \sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \left(\sum_{t=1}^n \epsilon_t \ell_t(a) \right), \end{aligned}$$

since the two sums are identical (ϵ_t and $-\epsilon_t$ have the same distribution).

Optimal Regret and Sequential Rademacher Averages

Theorem:

$$V_n(\mathcal{A}, \mathcal{L}) \leq 2 \sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^n \epsilon_t \ell_t(a),$$

where $\epsilon_1, \dots, \epsilon_n$ are independent Rademacher (uniform ± 1 -valued) random variables.

- Rademacher averages in probabilistic setting:

$$\text{excess risk} \leq c \mathbf{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(Y_t, f(X_t)) \right|.$$

- Sequential Rademacher averages in adversarial setting:

$$V_n(\mathcal{A}, \mathcal{L}) \leq c \sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^n \epsilon_t \ell_t(a).$$