Machine Learning Theory (CS 6783)

Lecture 19: Prediction with Arbitrary Covariates

1 Linear Betting Game With Covariates

```
For $t = 1$ to $n$
    Receive instance $x_t \in \mathcal{X}$
    Predict $\hat{y}_t \in \mathbb{R}$
    Receive label $y_t \in \{\pm 1\}$ and pay loss $\hat{y}_t \cdot y_t$
End For
```

Note that the above is a variant of the betting game where we get covariate or side information on every round and we want to perform as some benchmark ϕ that uses knowledge of this side information. You can think of this as an extension of Cover's result when we have covariates.

We say that the adaptive bound $\phi : \mathcal{X}^n \times \{\pm 1\}^n$ is achievable if there exists a strategy for the learner that ensures that:

$$\sum_{t=1}^{n} \hat{y}_t \cdot y_t \leq \phi(x_1, \ldots, x_n, y_1, \ldots, y_n) \tag{1}$$

We want to answer the question of when a performance bound ϕ is achievable and when it is achievable what the algorithm for the learner should be.

2 Meet The Trees

Definition 1. *The sequence of functions $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ with $\mathbf{X}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$ will be called an* $\mathcal{X}$-valued tree*. Here $\mathbf{X}_1 \in \mathcal{X}$ is a constant.*

We are now ready to provide the main result characterizing what ϕ's are achievable:

Lemma 1. *A necessary and sufficient condition for ϕ to be achievable in the sense of* (1) *is that*

$$\inf_{\mathbf{X}} \mathbb{E}_{\epsilon} \phi(\mathbf{X}_1, \mathbf{X}_2(\varepsilon_1), \ldots, \mathbf{X}_n(\varepsilon_{1:n-1}); \epsilon) \geq 0 \tag{2}$$

where the infimum is taken over all $\mathcal{X}$-valued tree $\mathbf{X}$.

Proof.

Proof of (1) $\Rightarrow$ (2)
To motivate the appearance of the trees, let us provide a particular way that the sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{\pm 1\}$ may evolve (i.e. a strategy of Nature). Fix a tree $\mathbf{X}$, and on round t let the Nature present the side information $x_t = \mathbf{X}_t(y_1, \ldots, y_{t-1})$ and $y_t = \varepsilon_t$, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher. This strategy is semi-oblivious, and it is clear that the left-hand side of (1) is equal to 0. Hence,

$$\mathbb{E}_\epsilon \phi(\mathbf{X}_1, \mathbf{X}_2(\varepsilon_1), \ldots, \mathbf{X}_n(\varepsilon_{1:n-1}); \epsilon) \geq 0. \tag{3}$$

Since the lower bound holds for any tree $\mathbf{X}$, can take infimum to conclude the statement.

Proof of (2) $\Rightarrow$ (1)
As before we choose the right potential for the job to prove this direction. To this end, for any t, we shall choose the potential:

$$P((x_1, y_1), \ldots, (x_t, y_t)) = \sup_{x_{t+1}} \mathbb{E}_{\epsilon_{t+1}} \sup_{x_{t+2}} \mathbb{E}_{\epsilon_{t+2}} \ldots \sup_{x_n} \mathbb{E}_{\epsilon_n} \left[-\phi(x_1, \ldots, x_n, y_1, \ldots, y_t, \epsilon_{t+1}, \ldots, \epsilon_n)\right]$$

Few observations about this potential before we proceed: **First,** note that

$$P((x_1, y_1), \ldots, (x_n, y_n)) = -\phi(x_1, \ldots, x_n, y_1, \ldots, y_n) \tag{4}$$

Second, note that

$$P(\cdot) = \sup_{x_1} \mathbb{E}_{\epsilon_1} \sup_{x_2} \mathbb{E}_{\epsilon_2} \ldots \sup_{x_n} \mathbb{E}_{\epsilon_n} \left[-\phi(x_1, \ldots, x_n, \epsilon_1, \ldots, \epsilon_n)\right]$$

However, consider the RHS of the above equation, say x_1^* was the x_1 attaining the suprema over x_1 above, we can set $X_1 = x_1^*$. Next, if $\epsilon_1 = +1$ then say conditioned on this, we can set

$$X_2(\epsilon_1) = \underset{x_2}{\operatorname{argmax}} \quad \mathbb{E}_{\epsilon_2} \ldots \sup_{x_n} \mathbb{E}_{\epsilon_n} \left[-\phi(X_1, x_2, \ldots, x_n, \epsilon_1, \ldots, \epsilon_n)\right]$$

and similarly, given a draw of $\epsilon_1, \ldots, \epsilon_{t-1}$, we can set

$$X_t(\epsilon_1, \ldots, \epsilon_{t-1}) = \underset{x_t}{\operatorname{argmax}}\ \mathbb{E}_{\epsilon_t}\ \mathbb{E}_{\epsilon_2} \ldots \sup_{x_n} \mathbb{E}_{\epsilon_n} \left[-\phi(X_1, X_2(\epsilon_1), \ldots, X_{t-1}(\epsilon_1, \epsilon_{t-2}), x_t, \ldots, x_n, \epsilon_1, \ldots, \epsilon_n)\right]$$

and so clearly we have:

$$\begin{aligned} P(\cdot) &= \sup_{x_1} \mathbb{E}_{\epsilon_1} \sup_{x_2} \mathbb{E}_{\epsilon_2} \ldots \sup_{x_n} \mathbb{E}_{\epsilon_n} \left[-\phi(x_1, \ldots, x_n, \epsilon_1, \ldots, \epsilon_n)\right] \\ &= \sup_{\mathbf{X}} \mathbb{E}_\epsilon \left[-\phi(\mathbf{X}_1, \mathbf{X}_2(\varepsilon_1), \ldots, \mathbf{X}_n(\varepsilon_{1:n-1}); \epsilon)\right] \\ &= -\inf_{\mathbf{X}} \mathbb{E}_\epsilon \left[-\phi(\mathbf{X}_1, \mathbf{X}_2(\varepsilon_1), \ldots, \mathbf{X}_n(\varepsilon_{1:n-1}); \epsilon)\right] \end{aligned}$$

But by (2) we know that the RHS above is bounded above by 0 and so we can conclude that:

$$P(\cdot) \leq 0 \tag{5}$$

Finally, observe that

$$P((x_1, y_1), \ldots, (x_t, y_t)) = \sup_{x_{t+1}} \mathbb{E}_{\epsilon_{t+1}} P((x_1, y_1), \ldots, (x_t, y_t), (x_{t+1}, \epsilon_{t+1})) \tag{6}$$

Now I claim that if we have that for any x_t:

$$\inf_{\hat{y}_t} \sup_{y_t} \{\hat{y}_t \cdot y_t + P((x_1, y_1), \ldots, (x_t, y_t))\} \leq P((x_1, y_1), \ldots, (x_{t-1}, y_{t-1})) \tag{7}$$

then we can conclude that (2) ⇒ (1). Why?

Well we want to prove that $\sum_{t=1}^n \hat{y}_t \cdot y_t - \phi(x_1, \ldots, x_n, y_1, \ldots, y_n) \leq 0$. To this end, note that from (4) we have, $\sum_{t=1}^n \hat{y}_t \cdot y_t - \phi(x_1, \ldots, x_n, y_1, \ldots, y_n) = \sum_{t=1}^n \hat{y}_t \cdot y_t + P((x_1, y_1), \ldots, (x_n, y_n))$. Now say we pick $\hat{y}_t = \operatorname*{argmin}_{\hat{y}} \sup_{y_t} \{\hat{y} \cdot y_t + P((x_1, y_1), \ldots, (x_t, y_t))\}$. Then using (7) multiple times we have,

$$\begin{aligned}
\sum_{t=1}^n \hat{y}_t \cdot y_t - \phi(x_1, \ldots, x_n, y_1, \ldots, y_n) &= \sum_{t=1}^n \hat{y}_t \cdot y_t + P((x_1, y_1), \ldots, (x_n, y_n)) \\
&= \sum_{t=1}^{n-1} \hat{y}_t \cdot y_t + \hat{y}_n y_n + P((x_1, y_1), \ldots, (x_n, y_n)) \\
&\leq \sum_{t=1}^{n-1} \hat{y}_t \cdot y_t + P((x_1, y_1), \ldots, (x_{n-1}, y_{n-1})) \\
&\leq \sum_{t=1}^{n-2} \hat{y}_t \cdot y_t + P((x_1, y_1), \ldots, (x_{n-2}, y_{n-2})) \\
&\leq \cdots \\
&\leq P(\cdot) \leq 0
\end{aligned}$$

where the last line is due to (5). Thus we have the other direction. Now the only thing left for us to do is to show that:

$$\inf_{\hat{y}_t} \sup_{y_t} \{\hat{y}_t \cdot y_t + P((x_1, y_1), \ldots, (x_t, y_t))\} \leq P((x_1, y_1), \ldots, (x_{t-1}, y_{t-1}))$$

To this end, note that:

$$\begin{aligned}
&\min_{\hat{y}_t} \max_{y_t \in \{\pm 1\}} \{\hat{y}_t \cdot y_t + P((x_1, y_1), \ldots, (x_t, y_t))\} \\
&\quad = \min_{\hat{y}_t} \max \{-\hat{y}_t + P((x_1, y_1), \ldots, (x_t, +1)), \hat{y}_t + P((x_1, y_1), \ldots, (x_t, -1))\}
\end{aligned}$$

It is easy to see that the minima above is in fact exactly $\hat{y}_t = \frac{P((x_1,y_1),\ldots,(x_t,+1)) - P((x_1,y_1),\ldots,(x_t,-1))}{2}$.

Irrespective, note that

$$
\begin{aligned}
&\min_{\hat{y}_t} \max_{y_t \in \{\pm 1\}} \left\{ \hat{y}_t \cdot y_t + P((x_1,y_1),\dots,(x_t,y_t)) \right\} \\
&\le \max \Bigg\{ -\tfrac{P((x_1,y_1),\dots,(x_t,+1)) - P((x_1,y_1),\dots,(x_t,-1))}{2} + P((x_1,y_1),\dots,(x_t,+1)), \\
&\qquad\qquad \tfrac{P((x_1,y_1),\dots,(x_t,+1)) - P((x_1,y_1),\dots,(x_t,-1))}{2} + P((x_1,y_1),\dots,(x_t,-1)) \Bigg\} \\
&= \tfrac{P((x_1,y_1),\dots,(x_t,+1)) + P((x_1,y_1),\dots,(x_t,-1))}{2} \\
&= \mathbb{E}_{\epsilon_t} P((x_1,y_1),\dots,(x_{t-1},y_{t-1}),(x_t,\epsilon_t)) \\
&\le \sup_{x_t} \mathbb{E}_{\epsilon_t} P((x_1,y_1),\dots,(x_{t-1},y_{t-1}),(x_t,\epsilon_t)) \\
&= P((x_1,y_1),\dots,(x_{t-1},y_{t-1}))
\end{aligned}
$$

Thus we have proved the claim and hence the other direction as well as was shown earlier. □

Remark 2.1. *We will sometimes write either* $\mathbf{X}_t$ *or* $\mathbf{X}_t(\epsilon)$ *instead of the more precise but longer expression* $\mathbf{X}_t(\varepsilon_1,\dots,\varepsilon_{t-1})$ *whenever this does not cause confusion.*

Example 2.1. *Let us define*

$$
\phi(x_1,\dots,x_n,y_1,\dots,y_n) = \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} y_t \cdot f(x_t) + \mathrm{Complex}_n(\mathcal{F})
$$

we want to ask the question, what is the smallest value of $\mathrm{Complex}_n(\mathcal{F})$ *such that*

$$
\inf_{\mathbf{X}} \mathbb{E}_\epsilon \phi(\mathbf{X}_1, \mathbf{X}_2(\varepsilon_1), \dots, \mathbf{X}_n(\varepsilon_{1:n-1}); \epsilon) \ge 0
$$

To this end, note that:

$$
\begin{aligned}
\inf_{\mathbf{X}} \mathbb{E}_\epsilon \phi(\mathbf{X}_1, \mathbf{X}_2(\varepsilon_1), \dots, \mathbf{X}_n(\varepsilon_{1:n-1}); \epsilon) &= \inf_{\mathbf{X}} \mathbb{E}_\epsilon \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \cdot f(\mathbf{X}_t(\epsilon_1,\dots,\epsilon_{t-1})) + \mathrm{Complex}_n(\mathcal{F}) \\
&= -\sup_{\mathbf{X}} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} (-\epsilon_t) \cdot f(\mathbf{X}_t(\epsilon_1,\dots,\epsilon_{t-1})) + \mathrm{Complex}_n(\mathcal{F}) \\
&= -\sup_{\mathbf{X}} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \cdot f(\mathbf{X}_t(\epsilon_1,\dots,\epsilon_{t-1})) + \mathrm{Complex}_n(\mathcal{F})
\end{aligned}
$$

Thus from this exercise, it clear that

$$
\mathrm{Complex}_n(\mathcal{F}) = \sup_{\mathbf{X}} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \cdot f(\mathbf{X}_t(\epsilon_1,\dots,\epsilon_{t-1}))
$$

is the right complexity term. We will refer to this complexity term as sequential Rademacher complexity which we will discuss about more in the next lecture. Note that if we had a tree $\mathbf{X}$ *above which had the same value for all its nodes on level* t*, that is if* $X_t(\epsilon_1,\dots,\epsilon_{t-1}) = x_t$ *for all* ϵ*'s then the above would be exactly the worst case statistical Rademacher Complexity. The crucial difference however is that we have an arbitrary tree that can make the sequential Rademacher complexity much larger than the statistical one in some settings.*