Adversarial online learning

Maximilian Kasy

March 2021

These slides summarize Chapter 2.1-2.2 of the following textbook:

Cesa-Bianchi, N. and Lugosi, G. (2006).

Prediction, learning, and games.
Cambridge University Press.

(To facilitate our discussion, I have changed some of the notation to be in line with Shalev-Shwartz, S. and Ben-David, S. (2014).)

Setup

Weighted average predictors

Bounding regret

Setup

- Sequential predictions at times $t = 1, 2, \ldots$
- Outcomes: $y_t \in \mathcal{Y}$.
- Predictions: $\hat{y} \in \mathcal{Y}$.
- Experts $h \in \mathcal{H}$, delivering predictions

$$\hat{y}_{h,t} \in \mathcal{Y}.$$

  ($\sim$ hypotheses / predictors).

- Any predictive features x_t are left implicit in the expert predictions.
- We assume (for today's discussion)
  1. $\mathcal{H}$ is finite,
  2. $\mathcal{Y}$ is a convex subset of a vector space.

Loss and regret

- We want to make a prediction $\hat{y}_t$, using the expert predictions $\hat{y}_{h,t}$,
- having observed $\mathcal{S}_{t-1} = (y_1, \ldots, y_{t-1})$.
- Loss at time t: $L(\hat{y}_t, y_t)$.
- Regret at time t relative to h:

$$r_{h,t} = L(\hat{y}_t, y_t) - L(\hat{y}_{h,t}, y_t).$$

- Cumulative regret at time t relative to h:

$$R_{h,t} = \sum_{s=1}^{t} r_{h,t}.$$

- Cumulative regret relative to $\mathcal{H}$:

$$R_{\mathcal{H},t} = \max_{h \in \mathcal{H}} R_{h,t}.$$

Successful learning

- Our goal: Find learning algorithms delivering $\hat{y}_t$
- such that average cumulative regret vanishes
- **for all possible realizations of** $\mathcal{S}_t = (y_1, \ldots, y_t)$:

$$\sup_{\mathcal{S}_t} \frac{1}{t} R_{\mathcal{H},t} \to 0.$$

- No probability is involved,
  this is the worst case over all possible realizations of outcomes!!
- How could that even be possible?!?
  The past carries no information about the future?!?!
  There is no stability at all over time?!?!?!

A chaotic, evil world

- **No assumption** is made about how the outcomes y_t are generated.
- We are interested in worst case behavior over all possible sequences $y_1, y_2, \ldots$

> "Imagine another set of results. The first time, the white ball drove the black ball into the pocket. The second time, the black ball bounced away. The third time, the black ball flew onto the ceiling. The fourth time, the black ball shot around the room like a frightened sparrow, finally taking refuge in your jacket pocket. The fifth time, the black ball flew away at nearly the speed of light, breaking the edge of the pool table, shooting through the wall, and leaving the Earth and the Solar System, just like Asimov once described.[13] What would you think then?"
>
> Ding watched Wang. After a long silence, Wang finally said, "This actually happened. Am I right?"

Liu Cixin, The Three Body Problem

Setup

Weighted average predictors

Bounding regret

Weighted average predictors

- We will consider weighted average predictors of the form

$$\hat{y}_t = \frac{\sum_{h \in \mathcal{H}} w_{h,t-1} \cdot \hat{y}_{h,t}}{\sum_{h \in \mathcal{H}} w_{h,t-1}},$$

- where the weights of each expert are increasing in the cumulative regret relative to that expert

$$w_{h,t} = \phi'(R_{h,t}),$$

- with ϕ nonnegative, convex, and increasing.
- This gives a larger weight to experts that performed well in the past.

Convex loss functions

Lemma 2.1

- Suppose that the loss function is convex in $\hat{y}_t$,
- and $\hat{y}_t$ is given by a weighted average predictor of this form.
- Then

$$\sup_{y_t} \sum_{h \in \mathcal{H}} r_{h,t} \cdot \phi'(R_{h,t-1}) \leq 0.$$

- *Proof*:
  - Convexity of L, Jensen's inequality,
  - weights proportional to $\phi'(R_{h,t-1})$.

Potential function

- Use boldface for vectors, with components corresponding to $h \in \mathcal{H}$.
- Potential function (a proof device):

$$\Phi(\boldsymbol{u}) := \psi\left(\sum_{h\in\mathcal{H}} \phi(u_h)\right).$$

- With this notation

$$\hat{y}_t = \frac{\langle \nabla\Phi(\boldsymbol{R}_{t-1}),\ \hat{\boldsymbol{y}}_t\rangle}{\langle \nabla\Phi(\boldsymbol{R}_{t-1}),\ \mathbf{1}\rangle}$$

- The lemma then can be rewritten as the **Blackwell condition**

$$\sup_{y_t} \langle \boldsymbol{r}_t, \nabla\Phi(\boldsymbol{R}_{t-1})\rangle \le 0.$$

- Note that $\boldsymbol{R}_t = \boldsymbol{R}_{t-1} + \boldsymbol{r}_t$.

Illustrating the Blackwell condition

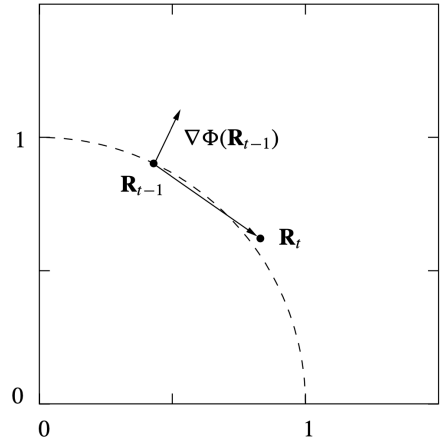

Figure 2.1. An illustration of the Blackwell condition with $N = 2$. The dashed line shows the points in regret space with potential equal to 1. The prediction at time t changed the potential from $\Phi(\mathbf{R}_{t-1}) = 1$ to $\Phi(\mathbf{R}_t) = \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t)$. Though $\Phi(\mathbf{R}_t) > \Phi(\mathbf{R}_{t-1})$, the inner product between $\mathbf{r}_t$ and the gradient $\nabla\Phi(\mathbf{R}_{t-1})$ is negative, and thus the Blackwell condition holds.

Bounding the potential

Theorem 2.1.

- Suppose that $\hat{y}_t$ satisfies the Blackwell condition.
- Then, for all t,

$$\Phi(\boldsymbol{R}_t) \leq \Phi(\mathbf{0}) + \frac{1}{2}\sum_{s=1}^{t} C(\boldsymbol{r}_t)$$

- where

$$C(\boldsymbol{r}) = \sup_{\boldsymbol{u}} \psi'\left(\sum_{h\in\mathcal{H}} \phi(u_h)\right) \sum_{h\in\mathcal{H}} \phi''(u_h) r_h^2.$$

- *Proof*:
  - Second order Taylor expansion of $\Phi(\boldsymbol{R}_t) = \Phi(\boldsymbol{R}_{t-1} + \boldsymbol{r}_t)$ in $\boldsymbol{r}_t$.
  - Bounding the first-order term using the Blackwell condition.
  - Bounding the second-order term by the supremum.
  - Telescope sum.

Exponential weighting

- Special case: Exponential weights.
- Potential (with tuning parameter η):

$$\phi(\boldsymbol{u}) = \frac{1}{\eta} \log \left(\sum_{h \in \mathcal{H}} \exp(\eta \cdot u_h) \right).$$

- Corresponding weights:

$$w_{h,t-1} = \frac{\exp\left(\eta \cdot R_{h,t-1}\right)}{\sum_{h' \in \mathcal{H}} \exp\left(\eta \cdot R_{h',t-1}\right)} = \frac{\exp\left(\eta \cdot \sum_{s=1}^{t-1} L(\hat{y}_{h,t}, y_t)\right)}{\sum_{h' \in \mathcal{H}} \exp\left(\eta \cdot \sum_{s=1}^{t-1} L(\hat{y}_{h',t}, y_t)\right)}.$$

- These weights only depend on the loss of each expert, but not on our prediction $\hat{y}_t$.
- For quadratic error loss, this is Bayesian model averaging, for normal likelihood with variance $2/\eta$, uniform prior over experts.

Bounding regret for exponential weighting

Corollary 2.2.

- Assume that L is convex in $\hat{y}$ and bounded by $[0, 1]$.
- Then, for all η and for all $\mathcal{S}_t = (y_1, \ldots, y_t)$,

$$R_{\mathcal{H},t}(\mathcal{S}_t) \leq \frac{\log(|\mathcal{H}|)}{\eta} + \frac{t\eta}{2}.$$

- For $\eta = \sqrt{2\frac{\log(|\mathcal{H}|)}{t}}$,

$$R_{\mathcal{H},t}(\mathcal{S}_t) \leq \sqrt{2t\log(|\mathcal{H}|)}.$$

Proof

- By assumption, $\phi(x) = \exp(\eta \cdot x)$, $\psi(x) = \phi^{-1}(x) = \log(x)/\eta$.
- For any estimator with weights based on a potential, and $\psi(x) = \phi^{-1}(x)$,

$$\begin{aligned}\max_{h\in\mathcal{H}} R_{h,t} &= \psi\left(\phi\left(\max_{h\in\mathcal{H}} R_{h,t}\right)\right) \\ &\leq \psi\left(\sum_{h\in\mathcal{H}} \phi\left(R_{h,t}\right)\right) = \Phi(\boldsymbol{R}_t).\end{aligned}$$

- Calculation yields $C(\boldsymbol{r}_t) \leq \eta$ (using $|r_{h,t}| \leq 1$), and $\Phi(\mathbf{0}) = \log(|\mathcal{H}|)/\eta$.
- The theorem implies

$$\begin{aligned}\Phi(\boldsymbol{R}_t) &\leq \Phi(\mathbf{0}) + \frac{1}{2}\sum_{s=1}^{t} C(\boldsymbol{r}_t) \\ &\leq \frac{\log(|\mathcal{H}|)}{\eta} + t\frac{\eta}{2}.\end{aligned}$$

Discussion

- We can do essentially as well as the best of our experts.
- No matter how the sequence y_t is generated!
- No stability or invariance in the world is assumed.
- A possible way to address the **induction problem**?
- We are guaranteed do well *if anyone can do well*.

Is this good enough?

> *The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken.*
>
> Bertrand Russell, The Problems of Philosophy.

- Should our regret bound provide consolation to the chicken?

Thank you!