Lecture 11: Agnostic Online Learnability

Han Shao

Ben-David et al. (2009)

We now return to binary classification with a nonempty hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. For a fixed labeled sequence, define

$$\mathrm{Reg}_T\big(\mathcal{A}, ((x_t, y_t))_{t=1}^T\big) = \mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^T \mathbb{1}\{\widehat{y}_t \neq y_t\}\right] - \min_{h\in\mathcal{H}} \sum_{t=1}^T \mathbb{1}\{h(x_t) \neq y_t\}.$$

For adaptive sequences, take expectation over the entire difference, as above. The current label cannot depend on the learner's current random prediction.

Definition 1 (Agnostic Online Learnability)**.** *A hypothesis class $\mathcal{H}$ is* agnostically online learnable *if there exists an algorithm $\mathcal{A}$ with a uniform regret bound $R_T = o(T)$ against every allowed online sequence.*

For finite $\mathcal{H}$, regard each hypothesis as an expert with loss

$$\ell_t(h) = \mathbb{1}\{h(x_t) \neq y_t\}.$$

After observing y_t, all these losses are available. Applying Hedge therefore gives a regret bound

$$O\left(\sqrt{T \ln |\mathcal{H}|}\right).$$

Consequently, every finite hypothesis class is agnostically online learnable.
For infinite classes, the relevant parameter is again the Littlestone dimension from the preceding lecture.

Theorem 1 (Littlestone-Dimension Regret Bounds)**.** *Let $\mathcal{H}$ have finite Littlestone dimension $d = \mathrm{Ldim}(\mathcal{H}) \geq 1$. For $T \geq \max\{d, 2\}$, every learning algorithm has worst-case expected regret at least*

$$\Omega\left(\sqrt{dT}\right).$$

There is also an online learning algorithm with regret bound

$$O\left(\sqrt{dT}\right).$$

We prove an upper bound of $O\left(\sqrt{dT \ln T}\right)$. The proofs below follow Sections 3.2 and 3.3 of Ben-David et al. (2009). The upper bound reduces the problem to Hedge by constructing finitely many experts from the Standard Optimal Algorithm (SOA). The lower bound repeats instances along a shattered tree, with independent fair labels within each block.

The Upper Bound: From SOA to a Finite Set of Experts

Fix $1 \le d \le T$. In the preceding lecture, we proved that SOA makes at most d mistakes on any sequence labeled by a hypothesis in $\mathcal{H}$. Thus, once the instances are given, a hypothesis's predictions can be reconstructed from the rounds on which SOA would make a mistake when learning that hypothesis. There are only finitely many possible sets of at most d such rounds.

Constructing the experts. Define the index set

$$\mathcal{I} = \{I \subseteq [T] : |I| \le d\}.$$

For each $I \in \mathcal{I}$, construct an expert E_I. It maintains a version space V_t^I, initialized to $V_1^I = \mathcal{H}$. On receiving x_t, if $V_t^I \neq \varnothing$, it computes

$$V_t^{I,b} = \{h \in V_t^I : h(x_t) = b\}, \qquad b \in \{0,1\},$$

and the SOA prediction

$$\widetilde{y}_t^I \in \operatorname*{argmax}_{b \in \{0,1\}} \mathrm{Ldim}(V_t^{I,b}),$$

with ties broken in favor of 0 and $\mathrm{Ldim}(\varnothing) = -1$. The expert predicts

$$e_t^I = \begin{cases} 1 - \widetilde{y}_t^I & \text{if } t \in I, \\ \widetilde{y}_t^I & \text{if } t \notin I, \end{cases}$$

and updates

$$V_{t+1}^I = \{h \in V_t^I : h(x_t) = e_t^I\}.$$

If $V_t^I = \varnothing$, define $e_t^I = 0$ and $V_{t+1}^I = \varnothing$. This convention makes every expert well-defined, even for a prescribed set I that leads to inconsistent predictions.

The update uses the expert's own prediction e_t^I, not the observed label y_t. Each expert simulates a possible realizable run of SOA, correcting its predictions on the prescribed rounds I. The actual labels are used only to evaluate expert losses and update the Hedge weights. In particular, every expert can produce its advice before the current label is revealed.

The family includes $I = \varnothing$, which corresponds to making no corrections. Although these experts need not be fixed hypotheses in $\mathcal{H}$, Hedge applies to them: its proof only requires an expert loss vector in $[0,1]^N$ on each round.

Lemma 1 (An expert matching each hypothesis)**.** *For every instance sequence $x_1, \dots, x_T$ and every $h \in \mathcal{H}$, there exists $I \in \mathcal{I}$ such that*

$$e_t^I = h(x_t) \qquad \textit{for every } t \in [T].$$

Proof. Fix h and the instance sequence. Consider a hypothetical run of SOA on

$$(x_1, h(x_1)), \dots, (x_T, h(x_T)).$$

This sequence is realizable by h, regardless of the actual labels $y_1, \dots, y_T$. Let U_t be the hypothetical SOA version space before round t, let a_t be its prediction, and define

$$I_h = \{t \in [T] : a_t \neq h(x_t)\}.$$

SOA's mistake bound gives $|I_h| \le d$, so $I_h \in \mathcal{I}$.

We prove by induction that E_{I_h} has version space $V_t^{I_h} = U_t$ and predicts $h(x_t)$ on every round. Initially both version spaces equal $\mathcal{H}$. Suppose they agree before round t. They then compute the same SOA prediction, $\widetilde{y}_t^{I_h} = a_t$, using the same tie-breaking convention. If $t \notin I_h$, this prediction already equals $h(x_t)$. If $t \in I_h$, it is the opposite binary label, and E_{I_h} flips it. In either case,

$$e_t^{I_h} = h(x_t).$$

The next version spaces therefore satisfy

$$V_{t+1}^{I_h} = \{g \in U_t : g(x_t) = h(x_t)\} = U_{t+1}.$$

This completes the induction. These version spaces never become empty, since they always contain h. □

Proof of the upper bound in theorem 1. Run Hedge on the experts $\{E_I : I \in \mathcal{I}\}$, with expert loss

$$\ell_t(I) = \mathbb{1}\{e_t^I \neq y_t\}.$$

After sampling an expert I_t from the Hedge distribution, predict $\widehat{y}_t = e_t^{I_t}$. Every expert is simulated on every round, not just the expert selected for prediction. Once y_t is revealed, all expert losses are known.

Comparing the experts with $\mathcal{H}$. For every realized labeled sequence, lemma 1 gives

$$\min_{I \in \mathcal{I}} \sum_{t=1}^{T} \ell_t(I) \leq \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\}. \tag{1}$$

The matching expert may depend on the complete sequence, which is permitted because Hedge competes with the best expert in hindsight. The family $\mathcal{I}$ and all expert rules are fixed in advance. Let

$$N = |\mathcal{I}| = \sum_{k=0}^{d} \binom{T}{k}.$$

With $\eta = \sqrt{8 \ln N / T}$,

$$\mathbb{E}_{\mathcal{A}} \left[\sum_{t=1}^{T} \mathbb{1}\{\widehat{y}_t \neq y_t\} - \min_{I \in \mathcal{I}} \sum_{t=1}^{T} \ell_t(I) \right] \leq \sqrt{\frac{T \ln N}{2}}.$$

Combining this with (1) bounds regret relative to $\mathcal{H}$ by the same quantity. This argument also applies to adaptive adversaries: the comparison holds for every realized history, and the current expert losses do not depend on the current random choice I_t. The comparator remains inside the expectation.

Counting the experts. For $1 \leq d \leq T$, the binomial-sum estimate used in Sauer's lemma gives

$$N \leq \left(\frac{eT}{d} \right)^d.$$

For completeness, set $u = d/T \in (0, 1]$. Since $u^d \leq u^k$ for $0 \leq k \leq d$,

$$u^d \sum_{k=0}^{d} \binom{T}{k} \leq \sum_{k=0}^{d} \binom{T}{k} u^k \leq (1+u)^T \leq e^{uT} = e^d.$$

Dividing by u^d proves the estimate. Consequently, the regret bound is

$$\sqrt{\frac{T}{2}\ln\left(\sum_{k=0}^{d}\binom{T}{k}\right)} \le \sqrt{\frac{dT}{2}\ln\frac{eT}{d}} = O\left(\sqrt{dT\ln T}\right), \tag{2}$$

where the last comparison uses $T \ge 2$. □

A Random-Walk Estimate for the Lower Bounds

We first prove the fluctuation estimate that will be used both in the shattered-tree construction and in the two-constant-hypothesis example.

Lemma 2 (Absolute deviation of a symmetric random walk)**.** *Let $\sigma_1, \dots, \sigma_n$ be independent uniform signs in $\{-1, +1\}$, where $n \ge 1$, and set $Z_n = \sum_{t=1}^{n} \sigma_t$. Then*

$$\sqrt{\frac{n}{3}} \le \mathbb{E}[|Z_n|] \le \sqrt{n}.$$

Equivalently, if $B_n \sim \mathrm{Bin}(n, 1/2)$, then

$$\sqrt{\frac{n}{12}} \le \mathbb{E}\left[\left|B_n - \frac{n}{2}\right|\right] \le \frac{\sqrt{n}}{2}.$$

Proof. Independence and $\mathbb{E}[\sigma_t] = 0$ give

$$\mathbb{E}[Z_n^2] = n.$$

When expanding the fourth power, a term has nonzero expectation only if every sign occurs an even number of times. Thus

$$\mathbb{E}[Z_n^4] = n + 6\binom{n}{2} = 3n^2 - 2n \le 3n^2.$$

Here the first term counts four copies of the same sign, and the second counts two copies each of two different signs.
For the upper bound, Cauchy–Schwarz yields

$$\mathbb{E}[|Z_n|] \le \sqrt{\mathbb{E}[Z_n^2]} = \sqrt{n}.$$

For the lower bound, apply Cauchy–Schwarz first to $|Z_n|^{1/2}$ and $|Z_n|^{3/2}$, and then to $|Z_n|$ and Z_n^2:

$$\begin{aligned} n^2 = \left(\mathbb{E}[Z_n^2]\right)^2 &\le \mathbb{E}[|Z_n|]\,\mathbb{E}[|Z_n|^3] \\ &\le \mathbb{E}[|Z_n|]\,\sqrt{\mathbb{E}[Z_n^2]\,\mathbb{E}[Z_n^4]} \\ &\le \sqrt{3}\,n^{3/2}\mathbb{E}[|Z_n|]. \end{aligned}$$

Dividing by $\sqrt{3}\,n^{3/2}$ gives $\mathbb{E}[|Z_n|] \ge \sqrt{n/3}$. Finally, $2B_n - n$ has the same distribution as Z_n, proving the binomial version. □

Why a Square-Root Dependence on the Horizon Is Necessary

An $\Omega(\sqrt{T})$ term is necessary even for an i.i.d. stochastic sequence. Consider just the two constant hypotheses

$$\mathcal{H} = \{h_0, h_1\}, \qquad h_0(x) = 0, \quad h_1(x) = 1.$$

Present the same fixed instance in every round, and independently draw $y_t \sim \text{Bernoulli}(1/2)$.

Proposition 1 (A stochastic square-root regret lower bound)**.** *For every possibly randomized algorithm $\mathcal{A}$ and every $T \geq 1$, the expected regret on this stochastic sequence satisfies*

$$\sqrt{\frac{T}{12}} \leq \mathbb{E}_{y_{1:T},\mathcal{A}}\left[\sum_{t=1}^{T} \mathbb{1}\{\widehat{y}_t \neq y_t\} - \min_{h\in\{h_0,h_1\}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\}\right] \leq \frac{\sqrt{T}}{2}.$$

In particular, its expected regret is $\Theta(\sqrt{T})$.

Proof. Each prediction is made without observing the current independent fair label, so

$$\mathbb{E}_{y_{1:T},\mathcal{A}}\left[\sum_{t=1}^{T} \mathbb{1}\{\widehat{y}_t \neq y_t\}\right] = \frac{T}{2}.$$

Let $N_1 = \sum_{t=1}^{T} y_t$, $N_0 = T - N_1$, and

$$Z_T = N_1 - N_0 = \sum_{t=1}^{T} (2y_t - 1).$$

The two hypotheses make N_1 and N_0 mistakes, respectively. Thus

$$\min_{h\in\{h_0,h_1\}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\} = \min\{N_0, N_1\} = \frac{T - |Z_T|}{2}.$$

Subtracting expected losses, the expected regret is exactly

$$\frac{T}{2} - \mathbb{E}_{y_{1:T}}\left[\frac{T - |Z_T|}{2}\right] = \frac{1}{2}\mathbb{E}_{y_{1:T}}[|Z_T|].$$

Since $2y_t - 1$ are independent uniform signs, lemma 2 bounds this quantity between $\sqrt{T/12}$ and $\sqrt{T}/2$. □

The distinction between a fixed comparator and the best comparator in hindsight is essential. Each constant hypothesis individually has expected loss $T/2$, but

$$\mathbb{E}\left[\min_{h\in\{h_0,h_1\}} L_T(h)\right] = \frac{T}{2} - \Theta(\sqrt{T}) < \min_{h\in\{h_0,h_1\}} \mathbb{E}[L_T(h)], \qquad L_T(h) = \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\}.$$

Regret uses the minimum inside the expectation, so these random fluctuations cannot be removed by observing that both hypotheses have the same expected loss. Averaging also gives, for each learner, a fixed label sequence with expected regret at least $\sqrt{T/12}$.

Formal Proof of the Littlestone-Dimension Lower Bound

Proof of the lower bound in theorem 1. Fix any possibly randomized learning algorithm $\mathcal{A}$ and a horizon $T \geq d$. We construct a distribution over labeled sequences for which the expected regret is at least $\sqrt{dT/24}$, and then fix one hard sequence by averaging.

Partitioning the horizon into blocks. Write

$$T = qd + r, \qquad q = \left\lfloor \frac{T}{d} \right\rfloor \geq 1, \qquad 0 \leq r < d.$$

Partition the rounds into d consecutive blocks $B_1, \ldots, B_d$. Let the first r blocks have size $q+1$ and the remaining blocks have size q. Write $n_j = |B_j|$, so $\sum_{j=1}^{d} n_j = T$.

Following a shattered tree. Fix a complete $\mathcal{H}$-shattered binary tree of depth d, with outgoing edges labeled 0 and 1. Let u_1 be its root instance. The sequence is generated block by block. In block j:

1. Present $x_t = u_j$ in every round $t \in B_j$.
2. Generate every label y_t independently from Bernoulli(1/2), independently also of $\mathcal{A}$'s randomness.
3. At the end of the block, let b_j be the majority of its labels, breaking ties in favor of 0. Follow edge b_j to select the next node u_{j+1}, when $j < d$.

The instance u_j depends only on labels in completed blocks. Thus the current label is a fresh fair coin even after conditioning on the history, the current instance, and the learner's prediction. Therefore,

$$\mathbb{E}_{y_{1:T},\mathcal{A}} \left[\sum_{t=1}^{T} \mathbb{1}\{\widehat{y}_t \neq y_t\} \right] = \frac{T}{2}. \tag{3}$$

One hypothesis realizes all the block majorities. For block j, define

$$Z_j = \sum_{t \in B_j} (2y_t - 1).$$

The numbers of 1 and 0 labels in that block are $(n_j + Z_j)/2$ and $(n_j - Z_j)/2$, respectively. A classifier predicting the block majority makes exactly

$$\frac{n_j - |Z_j|}{2}$$

mistakes there.

The chosen edges $b_1, \ldots, b_d$ form a root-to-leaf path. Since the tree is shattered, there is a single $h^\star \in \mathcal{H}$ with

$$h^\star(u_j) = b_j \qquad \text{for all } j \in [d].$$

This hypothesis achieves the majority loss in every block simultaneously. Conversely, every fixed hypothesis predicts a constant label within each block and cannot beat that block's majority. Hence, for every realization of the labels,

$$\min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\} = \frac{T}{2} - \frac{1}{2} \sum_{j=1}^{d} |Z_j|. \tag{4}$$

The comparator $h^\star$ is allowed to depend on the entire sequence: regret compares with the best fixed hypothesis *in hindsight*. It is still one hypothesis, not a different hypothesis in each block.

Accumulating the fluctuations. Within block j, Z_j is a sum of n_j independent uniform signs. Combining (3), (4), and lemma 2,

$$\mathbb{E}_{y_{1:T},\mathcal{A}}\left[\sum_{t=1}^{T}\mathbb{1}\{\widehat{y}_t \neq y_t\} - \min_{h\in\mathcal{H}}\sum_{t=1}^{T}\mathbb{1}\{h(x_t)\neq y_t\}\right]$$
$$= \frac{1}{2}\sum_{j=1}^{d}\mathbb{E}_{y_{1:T}}[|Z_j|] \geq \frac{1}{2\sqrt{3}}\sum_{j=1}^{d}\sqrt{n_j}.$$

For $T/d \geq 1$, we have $q = \lfloor T/d \rfloor \geq T/(2d)$. Since $n_j \geq q$,

$$\frac{1}{2\sqrt{3}}\sum_{j=1}^{d}\sqrt{n_j} \geq \frac{d\sqrt{q}}{2\sqrt{3}} \geq \sqrt{\frac{dT}{24}}.$$

When d divides T, all blocks have length T/d, and the same calculation gives $\sqrt{dT/12}$ directly.

Fixing a hard sequence. All random labels can be sampled before the interaction begins. The instance sequence is then determined by these labels and the fixed tree; it does not depend on any predictions made by $\mathcal{A}$. Consequently, averaging over the finitely many label sequences shows that at least one fixed labeled sequence satisfies

$$\mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^{T}\mathbb{1}\{\widehat{y}_t \neq y_t\}\right] - \min_{h\in\mathcal{H}}\sum_{t=1}^{T}\mathbb{1}\{h(x_t)\neq y_t\} \geq \sqrt{\frac{dT}{24}}.$$

This sequence is fixed before the learner makes its random choices. The lower bound therefore holds even against an oblivious adversary, and proves the $\Omega(\sqrt{dT})$ part of the theorem. □

References

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. Paper PDF.