Adaptive Online Learning

Dylan Foster, Alexander Rakhlin, and Karthik Sridharan

djfoster@cs.cornell.edu, rakhlin@wharton.upenn.edu, sridharan@cs.cornell.edu

Cornell University

Penn University of Pennsylvania

Adaptive Online Learning

Uniform Regret Bound

$$\mathbf{Reg}_n(\text{data}_{1:n}, \text{model}) \le \mathcal{B}(n)$$

Adaptive Regret Bound

$$\mathbf{Reg}_n(\text{data}_{1:n}, \text{model}) \le \mathcal{B}(\text{data}_{1:n}, \text{model})$$

- Adapt to "easy" data; retain worst-case guarantees.
- Encode **prior belief** (more likely models experience lower regret).
- Automatically tune hyperparameters (i.e. online model selection).

Questions

- Is a given adaptive bound $\mathcal{B}$ achievable?
- **VC-type** theory for adaptive online learning?

Contribution

General theory of adaptive online learning:

1. Sufficient and in many settings necessary conditions for achieving given bound $\mathcal{B}_n$.
2. Generic strategy for deriving adaptive algorithms from desired adaptive regret bounds.

Recovers a plethora of known adaptive regret bounds and enables new bounds including online PAC-Bayes and online model selection.
Works in abstract settings such as when $\mathcal{F}$ is nonparametric!

Setting

For $t = 1, \dots, n$:

- Nature provides input instance $x_t \in \mathcal{X}$
- Learner predicts $\hat{y}_t \in \mathcal{Y}$
- Nature provides label $y_t \in \mathcal{Y}$
- Learner suffers loss $\ell(\hat{y}_t, y_t)$

Regret

Measure performance through *regret* against comparator $f \in \mathcal{F}$:

$$\mathbf{Reg}_n(f) = \textstyle\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \sum_{t=1}^n \ell(f(x_t), y_t)$$

Adaptive Regret Bound

Upper bound $\mathcal{B}_n$ on $\mathbf{Reg}_n$ depending on the data, model, or both:

$$\mathbf{Reg}_n(f) \le \mathcal{B}_n(f \mid x_{1:n}, y_{1:n}).$$

Certifying Achievability

Given an adaptive regret bound $\mathcal{B}_n$, play an *adaptive online learning game* wherein nature tries to maximize the gap between the regret we experience and the proposed regret bound $\mathcal{B}_n$.

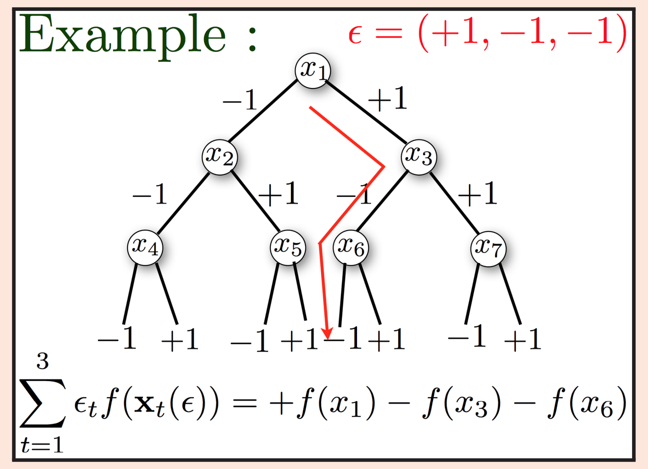
Minimax Value for Adaptive Learning

$\mathcal{B}_n$ is achievable if and only if

$$\mathcal{A}_n(\mathcal{F}, \mathcal{B}_n) \triangleq \underset{\substack{\text{Learner}\\\text{Strategy}}}{\inf} \; \underset{\substack{\text{Adversary}\\\text{Strategy}}}{\sup} \; \sup_{f \in \mathcal{F}} [\mathbf{Reg}_n(f) - \mathcal{B}_n(f \mid x_{1:n}, y_{1:n})] \le 0.$$

Definition: Tree

An $\mathcal{X}$-valued tree of depth n is a sequence $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ of mappings $\mathbf{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$.



Definition: Offset Sequential Rademacher Complexity

$$\mathcal{R}_n(\mathcal{F}, \mathcal{B}_n) \triangleq \sup_{\mathbf{x}, \mathbf{y}} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \Bigg\{ \underbrace{2 \textstyle\sum_{t=1}^n \epsilon_t f(\mathbf{x}_t(\epsilon))}_{\text{Rademacher average}} - \underbrace{\mathcal{B}_n(f \mid \mathbf{x}_{1:n}(\epsilon), \mathbf{y}_{1:n}(\epsilon))}_{\text{Offset}} \Bigg\}$$

Supremum is over $\mathcal{X}$- and $\mathcal{Y}$-valued trees. $\epsilon \in \{\pm 1\}^n$ is a sequence of Rademacher random variables.

Main Theorem

$$\mathcal{A}_n(\mathcal{F}, \mathcal{B}_n) \le \mathcal{R}_n(\mathcal{F}, \mathcal{B}_n)$$

Corollary: $\mathcal{R}_n(\mathcal{F}, \mathcal{B}_n) \le 0$ implies $\mathcal{B}_n$ is achievable.

Strategy for bounding $\mathcal{R}_n$:

1. Break class $\mathcal{F}$ into subsets via "complexity radius", then control a certain stochastic process R_i at each subset i.
2. Given a bound B_i on the typical size of R_i establish tail bound for R_i above B_i.
3. "Dilate" B_i to $B_i\theta_i$ and pass to a **maximal inequality** for final bound:

Maximal Inequality

Consider scalar-valued random variables R_i and "offsets" B_i for $i \in \mathbb{N}$. If R_i have mixed gaussian/exponential tail behavior:

One-Sided Tail Bound

$$P(R_i - B_i > \tau) \le C_1 \exp(-\tau^2/(2\sigma_i^2)) + C_2 \exp(-\tau s_i)$$

then

$$\mathbb{E} \sup_i [R_i - B_i\theta_i] \le O(\min \sigma_i + \min s_i^{-1}),$$

for

$$\theta_i B_i = O(\max\{\sigma_i \sqrt{\log \sigma_i}, s_i^{-1} \log(s_i^{-1})\} + B_i).$$

- R_i typically stands for supremum of Rademacher average over subset i.
- Tighter tail bounds for Rademacher average imply better adaptive rates.

New Regret Bounds: Generic Results

Definition: Cover

Set V of $\mathbb{R}$-valued trees is an α-cover of $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ on $\mathbf{x}$ w.r.t. ℓ_p if

$$\forall f \in \mathcal{F}, \forall \epsilon \in \{\pm 1\}^n, \exists \mathbf{v} \in V \quad \text{s.t.} \quad \sum_{t=1}^n (f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon))^p \le n\alpha^p.$$

$\mathcal{N}_p(\mathcal{F}, \alpha, \mathbf{x})$ is the smallest α-cover.

Definition: Predictable Sequence

Sequence $M_1, \dots, M_n$ where M_t depends only on $x_{1:t}$, $y_{1:t-1}$.

Predictable Sequences for Supervised Learning

Setting: Supervised learning with convex 1-Lipschitz loss.

$$\mathcal{B}_n(f \mid x_{1:n}) = \inf_\gamma \Bigg\{ K_1 \sqrt{\log n \cdot \log \mathcal{N}_2(\mathcal{F}, \gamma/2, n) \cdot \left(\textstyle\sum_{t=1}^n (f(x_t) - M_t)^2 + 1\right)} + K_2 \log n \int_{1/n}^{\gamma} \sqrt{n \log \mathcal{N}_2(\mathcal{F}, \delta, n)} d\delta + 2 \log n + 7 \Bigg\}$$

- Like [1], but predicts hypothesis behavior.
- Holds for nonparametric classes.
- Regret to Best vs. Regret to Fixed [2]: Fix a hypothesis $f^\star$, then set $M_t = f^\star(x_t)$. Specialized bound yields $O(1)$ regret against $f^\star$ and $O\big(\sqrt{n \log N}(\log(n \cdot \log N))\big)$ against arbitrary expert.

Definition: Complexity Radius

Given hypothesis class $\mathcal{F} = \bigcup_{R \ge 1} \mathcal{F}(R)$, with $\mathcal{F}(R) \subseteq \mathcal{F}(R')$ for $R \le R'$, have **complexity radius**:

$$R(f) \triangleq \inf\{R : f \in \mathcal{F}(R)\}.$$

Online Model Selection / Hyperparameter Tuning

Setting: Supervised learning with convex 1-Lipschitz loss ℓ.

$$\mathcal{B}_n(f) = \tilde{O}_n \left((\mathcal{R}_n(\mathcal{F}(2R(f))) + 1) \left(1 + \sqrt{\log \left(\frac{\log(2R(f)) \cdot \mathcal{R}_n(\mathcal{F}(2R(f)))}{\mathcal{R}_n(\mathcal{F}(1))} \right)} \right) \right)$$

- $\mathcal{R}_n(\mathcal{F}) \triangleq \mathcal{R}_n(\mathcal{F}, 0)$ (Sequential Rademacher complexity [Rakhlin-Sridharan-Tewari'11]).
- Enables unbounded hypothesis class $\mathcal{F}$. Predict as though we knew model complexity in advance at the price of logarithmic factor.
- Generalizes the unconstrained linear optimization bound [3] by taking $R(f) = \|f\|_2$. Also extends to smooth Banach spaces.

Recovering Existing Bounds

- Small loss bounds [4]
- Unconstrained linear optimization in Hilbert spaces [3]
- Regret to best fixed expert vs. regret to best [2]
- Predictable sequence bounds [Rakhlin-Sridharan'13]
- Quantile and second-order bounds for experts [6], Analogous to [Koolen-van Erven'15], [Luo-Schapire'15]

New Regret Bounds: Specific Examples

Online PAC-Bayes

Experts setting: $\mathcal{F} = \Delta_N$, $\mathcal{Y} = \{y \in \mathcal{R}^N : \|y\|_\infty \le 1\}$, $\ell(f, y) = \langle f, y \rangle$.

$$\mathcal{B}_n(f \mid y_{1:n}) = O\left(\sqrt{(\mathsf{KL}(f|\pi) + \log n) \textstyle\sum_{t=1}^n \mathbb{E}_{i \sim f} \langle e_i, y_t \rangle^2} + (\mathsf{KL}(f|\pi) + \log n) \right)$$

π is prior distribution over experts.

- Independent of the number of experts.
- Algorithm-independent.
- Depends quadratically on the expected loss of the expert we compare against.
- Addresses open question of obtaining algorithm-independent oracle-type variance bound for experts.
- Framework also recovers quantile bound [6] without log factors.

Online Linear Optimization with Spectral Norm

OLO Setting: $\mathcal{F} = \mathcal{Y} = \{f \in \mathcal{R}^d \mid \|f\|_2 \le 1\}, \ell(f, y) = \langle f, y \rangle$.

$$\mathcal{B}_n(y_{1:n}) = O\left(\sqrt{d} \log n \left(\|(\textstyle\sum_{t=1}^n y_t y_t^\top)^{1/2}\|_\sigma + 1 \right) \right)$$

Algorithms

Extend the framework of online relaxations from [Rakhlin-Shamir-Sridharan'12].

Definition: Adaptive Relaxation

For a rate $\mathcal{B}_n$ a *relaxation* $\mathbf{Rel}_n : \bigcup_{t=0}^n \mathcal{X}^t \times \mathcal{Y}^t \to \mathbb{R}$ satisfies the *initial condition*,

$$\mathbf{Rel}_n(x_{1:n}, y_{1:n}) \ge - \inf_{f \in \mathcal{F}} \{ \textstyle\sum_{t=1}^n \ell(f(x_t), y_t) + \mathcal{B}_n(f; x_{1:n}, y_{1:n}) \},$$

and the *recursive condition*,

$$\mathbf{Rel}_n(x_{1:t-1}, y_{1:t-1}) \ge \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t \in \mathcal{Y}} \mathbb{E}_{\hat{y}_t \sim q_t} [\ell(\hat{y}_t, y_t) + \mathbf{Rel}_n(x_{1:t}, y_{1:t})].$$

The relaxation's corresponding strategy is

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

which enjoys the adaptive regret bound

$$\mathbb{E}[\mathbf{Reg}_n(f)] \le \mathbb{E}[\mathcal{B}_n(f \mid x_{1:n}, y_{1:n}) + \mathbf{Rel}_n(\cdot)].$$

References

[1] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2013.

[2] Eyal Even-Dar, Michael Kearns, Yishay Mansour, and Jennifer Wortman. Regret to the best vs. regret to the average. *Machine Learning*, 72(1-2):21–37, 2008.

[3] H. Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. *Proceedings of The 27th Conference on Learning Theory*, 2014.

[4] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[5] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[6] Kamalika Chaudhuri, Yoav Freund, and Daniel J Hsu. A parameter-free hedging algorithm. In *Advances in neural information processing systems*, pages 297–305, 2009.