Online Learning and Online Convex Optimization

John Duchi

Outline

Online learning problems

The setting: given domain $X \subset \mathbb{R}^d$, we play the following game:

(1) We play a point $x_t \in X$

(2) Nature gives us a function $f_t : X \to \mathbb{R}$

(3) We suffer loss $f_t(x_t)$

Measure performance in terms of *regret* to a *fixed* predictor $x^\star$:

$$\mathrm{Reg}_T := \sum_{t=1}^{T} [f_t(x_t) - f_t(x^\star)]$$

Goal:

Classification with experts

Setting: we have input space $\mathcal{S}$ and experts

$$e_1, e_2, \ldots, e_d : \mathcal{S} \to \{-1, 1\}$$

We receive sequence $s_1, s_2, \ldots, s_T \in \mathcal{S}$ with labels $y_t \in \{-1, 1\}$
One expert is perfect. Is there a good strategy?

Online convex optimization

In online learning game, functions

$$f_t : X \to \mathbb{R} \quad \text{are convex}$$

Expert prediction in the convex setting

Question: Can we define expert prediction in a convex way?
Define loss vector

$$\ell_t = \begin{bmatrix} \mathbf{1}\left\{e_1(s_t) \neq y_t\right\} \\ \mathbf{1}\left\{e_2(s_t) \neq y_t\right\} \\ \vdots \\ \mathbf{1}\left\{e_d(s_t) \neq y_t\right\} \end{bmatrix}$$

and

$$X = \Delta_d = \left\{x \in \mathbb{R}^d_+ \mid \mathbf{1}^T x = 1\right\} \quad f_t(x) = x^\top \ell_t.$$

Online convex optimization

Idea: Let's just do (sub)gradient descent.

Initialize $x_1 \in X$, repeat:

(1) Suffer loss $f_t(x_t)$

(2) Update $g_t \in \partial f_t(x_t)$,

$$x_{t+1} = \operatorname*{argmin}_{x \in X} \left\{ \langle g_t, x \rangle + \frac{1}{2\alpha} \|x - x_t\|_2^2 \right\}$$

A regret bound

Theorem

Let $D_X^2 \geq \|x - y\|_2^2$ *for all* $x, y \in X$ *and assume* f_t *are convex and* M*-Lipschitz. Online gradient descent has regret*

$$\mathsf{Reg}_T \leq \frac{1}{2\alpha} D_X^2 + \frac{TM^2}{2}\alpha.$$

Online mirror descent

Idea: If gradient descent worked, so must mirror descent.

Initialize $x_1 \in X$, repeat:

(1) Suffer loss $f_t(x_t)$

(2) Update $g_t \in \partial f_t(x_t)$,

$$x_{t+1} = \operatorname*{argmin}_{x \in X} \left\{ \langle g_t, x \rangle + \frac{1}{\alpha_t} D_h(x, x_t) \right\}$$

Online mirror descent

Theorem

Assume that $h : X \to \mathbb{R}$ is strongly convex with respect to the norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$. If $\alpha_t = \alpha$ is fixed, then

$$\mathsf{Reg}_T \le \frac{1}{\alpha} D_h(x^\star, x_1) + \sum_{t=1}^{T} \frac{\alpha}{2} \|g_t\|_*^2$$

If α_t is non-increasing but otherwise arbitrary,

$$\mathsf{Reg}_T \le \frac{1}{\alpha_T} \sup_{x \in X} D_h(x^\star, x) + \sum_{t=1}^{T} \frac{\alpha_t}{2} \|g_t\|_*^2$$

Prediction with expert advice

Setting: expert predictions $e_j : \mathcal{S} \to \{-1, 1\}$ with loss vectors $\ell_t = [\mathbf{1}\{e_j(s_t) = y_t\}]_{j=1}^d$, simplex constraint $X = \Delta_d$, and expected loss

$$f_t(x) = x^\top \ell_t = \mathbb{E}_{j \sim x}\left[\mathbf{1}\{e_j(s_t) = y_t\}\right].$$

Use entropy divergence $h(x) = \sum_{j=1}^d x_j \log x_j$

Online to batch conversions

New twist: Suppose that the f_t are i.i.d. with $F(x) = \mathbb{E}[f_t(x)]$

Question: Does $\text{Reg}_T = o(T)$ guarantee generalization?

Theorem

Suppose that f_t are as above and $\sup_{x,y \in X}\{f_t(x) - f_t(y)\} \le B$. Then w.p. $\ge 1 - \delta$

$$F(\overline{x}_T) - F(x^\star) \le \frac{\text{Reg}_T}{T} + O(1) \cdot \sqrt{\frac{B \log \frac{1}{\delta}}{T}}$$

Generalization guarantees: proof

$$F(\overline{x}_T) - F(x^\star) \le \frac{1}{T} \sum_{t=1}^{T} [F(x_t) - F(x^\star)]$$

Strongly convex case

Suppose that sequence f_t is λ-*strongly convex*, meaning

$$f_t(y) \geq f_t(x) + \langle g_t, y - x \rangle + \frac{\lambda}{2} \|x - y\|_2^2 \quad \text{all } x, y \in X.$$

Example (Regularized learning)

Suppose that $f_t(x) = \ell_t(x) + \frac{\lambda}{2} \|x\|_2^2$, $X = \mathbb{R}^d$, $\ell_t \geq 0$ convex and M-Lipschitz.

Strongly convex regret

Theorem (Regret)

Let $f_t : X \to \mathbb{R}$ be λ-strongly convex and M-Lipschitz over X. Use stepsizes $\alpha_t = \frac{1}{\lambda t}$ in online gradient descent. Then

$$\mathrm{Reg}_T = \sum_{t=1}^{T} [f_t(x_t) - f_t(x^\star)] \le \frac{M^2}{\lambda} \log(T+1)$$

Strongly convex generalization

Theorem (Generalization)

Under the conditions of the previous theorem, if the f_t are i.i.d. with $\mathbb{E}[f_t] = F$ and λ-strongly convex, then

$$F(\overline{x}_T) - F(x^\star) \le \frac{\mathsf{Reg}_T}{T} + O(1)\left[\sqrt{\frac{M^2 \log\frac{1}{\delta}}{T} \cdot \frac{\mathsf{Reg}_T}{\lambda T}} + \frac{M^2}{\lambda}\frac{\log\frac{1}{\delta}}{T}\right]$$

Reading and bibliography

1. M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent.
   In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003

2. J. C. Duchi. Introductory lectures on stochastic convex optimization, 2016

3. E. Hazan. The convex optimization approach to regret minimization.
   In *Optimization for Machine Learning*, chapter 10. MIT Press, 2012

4. N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms.
   IEEE Transactions on Information Theory, 50(9):2050–2057, September 2004

5. N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.*
   Cambridge University Press, 2006