ORIE 7790 High Dimensional Probability and Statistics Lecture 21 - 4/28/2020

Lectures 21-22: Online Learning

Lecturer: Yudong Chen *Scribe: Billy Jin, Miaolan Xie*

Reading:

- Chapter 21 of Duchi's notes.
- Xinhua Zhang, short notes on mirror descent,
- Elad Hazan, "Introduction to Online Convex Optimization",

In these two lectures, we study online learning problems under the framework of online convex optimization. We give a few examples that fall into this framework. We then introduce a general algorithm called Online Mirror Descent for solving online convex optimization. We conclude by analyzing the regret of online mirror descent.

1 Online Convex Optimization

The setup can be described as a two-player sequential game:

- Let $W \subseteq \mathbb{R}^d$ be a *convex* parameter space.
- At each time t, player 1 (the *learner*) plays some $w_t \in W$.
- Player 2 (the *adversary*) then plays a loss function $L_t : W \to \mathbb{R}$, where L_t is convex.

Note that the learner commits to w_t **before** seeing L_t, whereas the adversary may adapt his choice of L_t to $w_1, \ldots, w_t$. The goal for the learner is to minimize regret, defined as

$$\sum_{t=1}^{T} L_t(w_t) - \sum_{t=1}^{T} L_t(w^*),$$

where $w^* := \arg\min_{w \in W} \sum_{t=1}^{T} L_t(w)$ is the best fixed decision in hindsight.

1.1 Examples

Here are some examples of problems that fall into the framework of online convex optimization.

1. **Online support vector machine**: At each time t, the learner picks a vector $w_t \in \mathbb{R}^d$. Then, a data point $(x_t, y_t) \in \mathbb{R}^d \times \{\pm 1\}$ is revealed, and the learner incurs loss $L_t(w_t)$, where $L_t(w) = \max\{1 - y_t \langle w, x_t \rangle, 0\}$. (This loss function is called the *hinge loss*.)
2. **Online logistic regression**: Same setup, except now the loss function is $L_t(w) = \log\left(1 + e^{-y_t \langle w, x_t \rangle}\right)$. (This is the *logistic loss*.)
3. **Expert prediction/adversarial bandit**: There are d experts/arms. At each time t, each expert makes a prediction (for example "I predict the stock market will go up tomorrow"). At each time t, the learner chooses a weight vector $w_t = (w_{t1}, \ldots, w_{td})$, where

$$w_{tj} = \text{weight for expert } j = \text{probability of pulling arm } j.$$

So the parameter space is $W = \Delta_d := \{w \in \mathbb{R}^d : \sum_j w_j = 1,\ w_j \geq 0\}$, which is the probability simplex in $\mathbb{R}^d$. Then losses

$$l_{tj} = \mathbb{1}\{\text{expert } j \text{ is wrong at time } t\} = \text{loss of arm } j \text{ at time } t$$

are revealed, and the learner incurs loss $L_t(w) = \langle w, l_t \rangle$. Note that $\nabla L_t(w) = l_t$.

2 Online Gradient Descent

Gradient descent extends naturally to an algorithm for online convex optimization. Online gradient descent does, at each iteration $t+1$:

$$w_{t+1} = \mathrm{Proj}_W(w_t - \eta_t g_t),$$

where η_t is the step size and $g_t \in \partial L_t(w_t)$. Note that this update is equivalent to

$$w_{t+1} = \operatorname*{arg\,min}_{w \in W} \left\{ \langle g_t, w \rangle + \frac{1}{2\eta_t} \|w - w_t\|_2^2 \right\}$$

3 Bregman Divergence

We will next see how to extend gradient descent to a more general algorithm. First, we will need to introduce the notion of Bregman divergence. Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function.

Definition 1 (Bregman Divergence)**.** *The* ***Bregman divergence*** *associated with ψ is a function $B_\psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by*

$$B_\psi(w, v) := \psi(w) - \psi(v) - \langle \nabla \psi(v), w - v \rangle$$

Remark By the convexity of ψ, the Bregman divergence B_ψ is always non-negative. One can think of $B_\psi(w, v)$ as a measure of "distance" between w and v; however, the Bregman divergence is not necessarily symmetric or satisfies the triangle inequality.

3.1 Examples

1. **Euclidean distance.** Let $\psi(w) = \frac{1}{2}\|w\|_2^2$. Then $B_\psi(w, v) = \frac{1}{2}\|w - v\|_2^2$.
2. **Mahalanobis distance.** Let $\psi(w) = \frac{1}{2} w^\top A w =: \frac{1}{2}\|w\|_A^2$, where $A \succcurlyeq 0$.
   Then $B_\psi(w, v) = \frac{1}{2}(w - v)^\top A (w - v) = \frac{1}{2}\|w - v\|_A^2$.
3. **KL-divergence.** Let $\psi(w) = \sum_{j=1}^d w_j \log w_j$ be the negative entropy. Note that ψ is convex on $\mathbb{R}_+^d$.
   Then $B_\psi(w, v) = \sum_{j=1}^d w_j \log \frac{w_j}{v_j} = D_{\mathrm{KL}}(w \parallel v)$ for all $w, v \in \Delta_d$.

4 Online Mirror Descent (OMD)

This is a generalization of gradient descent using Bregman divergences. At iteration t:

$$w_{t+1} = \operatorname*{arg\,min}_{w \in W} \left\{ \langle g_t, w \rangle + \frac{1}{\eta_t} B_\psi(w, w_t) \right\} \tag{1}$$

Remark $\langle g_t, w \rangle + \frac{1}{\eta_t} B_\psi(w, w_t)$ is convex in w. Hence this is a convex optimization problem.

4.1 Special cases of OMD

Gradient descent $\psi(w) = \frac{1}{2} \|w\|_2^2$

Exponentiated gradient descent This is online mirror descent with $W = \Delta_d$, $\psi(w) = \sum_j w_j \log w_j$, and $B_\psi(w, v) = D_{\mathrm{KL}}(w \parallel v)$. At iteration t:

$$w_{t+1} = \arg\min_{w \in W} \left\{ \langle g, w \rangle + \frac{1}{\eta_t} D_{\mathrm{KL}}(w \parallel w_t) \right\}.$$

To explicit calculate w_{t+1}, we write the Lagrangian:

$$L(w, \lambda, \tau) = \langle g, w \rangle + \frac{1}{\eta} \sum_{j=1}^{d} w_j \log \frac{w_j}{v_j} - \langle \lambda, w \rangle + \tau \left(\langle \mathbb{1}, w \rangle - 1 \right).$$

Here, $\lambda \in \mathbb{R}^d$ is the multiplier for the constraint $w \geq 0$ and $\tau \in \mathbb{R}$ is the multiplier for the constraint $\langle \mathbb{1}, w \rangle = 1$. Taking $\frac{\partial}{\partial w} L(w, \lambda, \tau) = 0$ gives

$$w_{t+1,j} = v_j \exp\left(-\eta g_j + \lambda_j \eta - \tau \eta - 1 \right) > 0.$$

Hence the constraint $w \geq 0$ is inactive, which implies $\lambda = 0$. We choose τ to normalize w, giving

$$\begin{aligned} w_{t+1} &= \left(\frac{w_{ti} \exp(-\eta_t g_{ti})}{\sum_{j=1}^{d} w_{tj} \exp\left(-\eta_t g_{tj}\right)} \right)_{i=1,\ldots,d} && (2) \\ &\propto \left(\exp\left(-\sum_{k=1}^{t} \eta_k g_{ki} \right) \right)_{i=1,\ldots,d} && (3) \\ &= \text{soft-argmin} \left\{ \sum_{k=1}^{t} \eta_k g_{ki},\ i = 1, \ldots, d \right\}. && (4) \end{aligned}$$

Remark In the context of the expert problem, g_{ki} is the loss of expert i at time k. Hence, $\sum_{k=1}^{t} g_{ki}$ is the total loss of expert i up to time t. Hence exponentiated gradient descent favors experts with low loss, but still assigns positive weight to every expert. This algorithm can thus be interpreted as a smoothed version of "follow the leader", where the weights are updated in an multiplicative fashion. (Variants of) exponentiated gradient descent is also known as **multiplicative weight update** (MWU), **follow-the-regularized-leader** (FTRL), **fictitious play** (FP), **Hedge algorithm**, and **entropic mirror descent**.

5 Analysis of Online Mirror Descent

We begin with some definitions.

Definition 2 (Strong convexity)**.** *ψ is* strongly convex *with respect to $\|\cdot\|$ if , for all v, w:*

$$\psi(w) - \psi(v) - \langle g, w - v \rangle \geq \frac{1}{2} \|w - v\|^2, \quad \textit{for all } g \in \partial \psi(v).$$

This is equivalent to $B_\psi(w, v) \geq \frac{1}{2} \|w - v\|^2$ by definition of Bregman divergence.

Example 1. Let $\psi(w) = \sum_j w_j \log w_j$ be negative entropy. Then by Pinsker's inequality, we have

$$B_\psi(w, v) = D_{\mathrm{KL}}(w \parallel v) \geq \frac{1}{2} \|w - v\|_1^2. \tag{5}$$

In other words, the negative entropy is strongly convex with respect to the ℓ_1 norm.

Definition 3 (Dual norm)**.** *The dual norm of* $\|\cdot\|$ *is the norm* $\|\cdot\|_*$ *defined by*

$$\|y\|_* = \sup_{x:\|x\|\le 1} \langle x, y \rangle .$$

Example 2. The dual norm of $\|\cdot\|_2$ is $\|\cdot\|_2$. The dual norm of $\|\cdot\|_\infty$ is $\|\cdot\|_1$. The dual norm of $\|\cdot\|_{\text{nuc}}$ (nuclear norm) is $\|\cdot\|_{\text{op}}$ (operator norm).

Theorem 1 (Regret of Online Mirror Descent)**.** *Suppose that* ψ *is strongly convex with respect to* $\|\cdot\|$ *with dual norm* $\|\cdot\|_*$*. Then online mirror descent with step size* $\eta_t \equiv \eta$ *satisfies*

$$\sum_{t=1}^{T} [L_t(w_t) - L_t(w^*)] \le \frac{1}{\eta} B_\psi(w^*, w_1) + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_*^2 .$$

Proof Recall that $w_{t+1} = \arg\min_{w\in W} \left\{ \langle g_t, w\rangle + \frac{1}{\eta} B_\psi(w, w_t) \right\}$. By the optimality condition for convex optimization (negative gradient lies in the normal cone), we have

$$\begin{aligned} 0 &\le \left\langle g_t + \frac{1}{\eta} \frac{\partial}{\partial w} B_\psi(w, w_t)\Big|_{w=w_{t+1}} ,\, w^* - w_{t+1} \right\rangle \\ &= \left\langle g_t + \frac{1}{\eta} \left(\nabla\psi(w_{t+1}) - \nabla\psi(w_t)\right) ,\, w^* - w_{t+1} \right\rangle . \end{aligned}$$

Therefore, we have

$$\begin{aligned} L_t(w_t) - L_t(w^*) &\le \langle g_t, w_t - w^* \rangle && \text{convexity of } L_t \\ &= \langle g_t, w_{t+1} - w^* \rangle + \langle g_t, w_t - w_{t+1} \rangle \\ &\le \frac{1}{\eta} \langle \nabla\psi(w_{t_1}) - \nabla\psi(w_t), w^* - w_{t+1} \rangle + \langle g_t, w_t - w_{t+1} \rangle && \text{last display equation} \\ &= \frac{1}{\eta} [B_\psi(w^*, w_t) - B_\psi(w^*, w_{t+1}) - B_\psi(w_{t+1}, w_t)] + \langle g_t, w_t - w_{t+1} \rangle , \end{aligned}$$

where the last step follows from direct calculation using definition and is sometimes known as the "three-point identity". Summing over $t = 1, \ldots, T$, the sum telescopes, and we get

$$\begin{aligned} \sum_{t=1}^{T} (L_t(w_t) - L_t(w^*)) &\le \frac{1}{\eta} [B_\psi(w^*, w_1) - B_\psi(w^*, w_{T+1}] + \sum_{t=1}^{T} \left[-\frac{1}{\eta} B_\psi(w_{t+1}, w_t) + \langle g_t, w_t - w_{t+1} \rangle \right] \\ &\le \frac{1}{\eta} B_\psi(w^*, w_1) + \sum_{t=1}^{T} \left[-\frac{1}{\eta} B_\psi(w_{t+1}, w_t) + \langle g_t, w_t - w_{t+1} \rangle \right] \end{aligned}$$

To control the last RHS term, we observe that

$$\begin{aligned} \langle g_t, w_t - w_{t+1} \rangle &\le \|g_t\|_* \|w_t - w_{t+1}\| && \text{definition of dual norm} \\ &\le \frac{\eta}{2} \|g_t\|^2 + \frac{1}{2\eta} \|w_t - w_{t+1}\|^2 && ab \le \frac{1}{2}(a^2 + b^2) \\ &\le \frac{\eta}{2} \|g_t\|_*^2 + \frac{1}{\eta} B_\psi(w_{t+1}, w_t) && \text{strong convexity of } \psi. \end{aligned}$$

Combining pieces, we obtain the desired regret bound

$$\sum_{t=1}^{T} (L_t(w_t) - L_t(w^*)) \le \frac{1}{\eta} B_\psi(w^*, w_1) + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_*^2 .$$

□

6 Applications

6.1 Online (sub)-gradient descent

Let $\psi(w) = \frac{1}{2}\|w\|_2^2$. Then ψ is strong convex with respect to $\|\cdot\|_2$, and the dual norm is $\|\cdot\|_2$. Suppose each L_t is L-Lipschitz, which implies $\|g_t\|_2 \leq L$. Then the regret bound is

$$\sum_{t=1}^{T} (L_t(w_t) - L_t(w^*)) \leq \frac{1}{2\eta}\|w^* - w_1\|_2^2 + \frac{\eta}{2} T \cdot L^2.$$

Choosing $\eta = \frac{\|w^* - w_1\|_2}{L\sqrt{T}}$ to minimize the RHS gives

$$\text{regret} \leq \|w^* - w_1\|_2 L\sqrt{T}.$$

Remark The $O(\sqrt{T})$ regret bound immediately implies a $O(\frac{1}{\sqrt{T}})$ convergence rate for the offline setting where all $L_t \equiv f$. In particular, letting $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$, we have

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T} [f(w_t) - f(w^*)] \leq \frac{\|w^* - w_1\|_2}{\sqrt{T}},$$

where the first step above is by Jensen's inequality.

6.2 Expoentiated gradient descent

Let $W = \Delta_d$, and $\psi(w) = \sum_j w_j \log w_j$ be the negative entropy. Then ψ is strongly convex with respect to $\|\cdot\|_1$, with dual norm $\|\cdot\|_\infty$. Then the regret bound is

$$\sum_{t=1}^{T} (L_t(w_t) - L_t(w^*)) \leq \frac{1}{\eta} D_{\mathrm{KL}}(w^* \parallel w_1) + \frac{\eta}{2}\sum_{t=1}^{T} \|g_t\|_\infty^2 .$$

If in addition we take the initial iterate $w_1 = (\frac{1}{d}, \ldots, \frac{1}{d})$ to be the uniform distribution, then one can verify that $D_{\mathrm{KL}}(w^* \parallel w_1) \leq \log d$. Also, set $\eta = \sqrt{\frac{\log d}{2T \max_t \|g_t\|_\infty^2}}$. Then the regret is

$$\text{regret} \leq \sqrt{T \log d \cdot \max_t \|g_t\|_\infty^2}. \tag{6}$$

Remark Compared to online gradient descent, the dependence on the gradients g_t is $\max_t \|g_t\|_\infty$ instead of $\max_t \|g_t\|_2$. Thus exponentiated gradient descent can do better than gradient descent when the gradients g_t are small in magnitude and not sparse.

6.3 Expert problem

Recall that l_{tj} is the loss of expert j at time t, and that $g_t = l_t \in \{0,1\}^d$. Thus $\|g_t\|_\infty \leq 1$. Plugging this into the bound for exponentiated gradient descent gives

$$\text{regret} \leq \sqrt{T \log d}$$

Remark This regret bound is optimal for the expert problem. In comparison, gradient descent would get $\sqrt{Td}$ regret, which has an exponentially larger dependence on the dimension d.

7 Extensions

1. We chose our step size η to be proportional to $\frac{1}{\sqrt{T}}$. This requires the time horizon to be known to the algorithm. If T is not known, one can use a varying step size $\eta_t = \frac{1}{\sqrt{t}}$ and prove essentially the same guarantees (under a slightly stronger boundedness assumption; see Duchi's notes.)

2. **Acceleration.** If more is known about the loss function L_t, then better regret bounds (in the online setting) and convergence rates (in the offline setting) can be obtained.

   - L_t is smooth (gradient is Lipschitz): We have an improvement $\sqrt{T} \to O(1)$ in regret, which translates to an improvement $\frac{1}{\sqrt{T}} \to \frac{1}{T}$ in rate.
   - L_t is strongly convex: We have an improvement $\sqrt{T} \to \log T$ in regret, and hence $\frac{1}{\sqrt{T}} \to \frac{\log T}{T}$ in rate.

   See Xinhua Zhang's notes for details.

3. So far, we assumed that we observe the losses of *all* the experts/arms, even those we did not choose/pull. This is the *full information* setting. Next week, we will look at the "bandit information" setting, where we only observe the loss of the expert/arm that we choose/pull, that is, we only see one entry of $\nabla L_t = g_t = l_t$.