**CS 7545: Machine Learning Theory** **Fall 2019**

# Lecture 11: Online Convex Optimization

*Lecturer: Jacob Abernethy* *Scribes: Liexiao Ding and Dan Li*

**Disclaimer**: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

## 11.1 Generic Framework for Online Convex Optimization

**Generalized Experts Setting** The process is shown as the following:

**Algorithm 1: Generalized Experts**

Given K $\subset \mathbb{R}^+$ bounded on convex and closed
The learner follows protocol:
**for** $t$ = *1,...,T* **do**
  Learner select $x_t \in K$
  Nature reveals loss function $f_t : K \to R^+$
**end**
Learning wants to minimize:
Regret$_t$:=$\sum_t f_t(x_t) = \min_{x \in K} \sum_{t=1}^T f_t(x)$

**Example 1 Prediction with extra advice**

$$f_t(w) = l(\sum_i \tilde{w}_i x_i^t, y^t)$$

where $w$ is normalized weight and $K = \Delta_n$

**Example 2 Online Portfolio Selection**

Let $K = \Delta_n, w^t \in \Delta_n$ be the distribution of n stocks, $r^t$ is a vector of price fluctuations

$$r^t(i) = \frac{Price_t(stock_i)}{Price_{t-1}(stock_i)}$$

The negative log wealth of best portfolio

$$\min_{w \in \Delta_n} \sum_{t=1}^T f_t(w) = -\log\left(\max_{win\Delta_n} \prod_{t=1}^T (w^t r^t)\right)$$

**Example 3 Online Regression**

Let K be a 2-Norm Ball $\Theta$ we have

$$f_t(\theta) = (\theta^\top \vec{x}_t - y_t)^2$$

where $(x_t, y_t) \in \mathbb{R}^{d+1}$ is some data point

## 11.2 Property of Online Convex Optimization

Online convex optimization combine ideas from

- Optimization
- Statistical Learning

**Claim** Online convex optimization is harder than vanilla optimization

**Goal** Find $\arg\min_{x \in K} \Phi(x)$

**Approach** Take any online convex optimization algorithm A. Initialize $x_1 \in K$, follow this protocol

**Algorithm 2:** Protocol

**for** $t$ = *1,...,T* **do**
  $x_t \leftarrow A(f_1, \ldots, f_{t-1})$
  Let $f_t(\cdot) = \Phi(\cdot)$ or $\langle \nabla\Phi(x_t), \cdot \rangle$
**end**

**Output** $\frac{1}{T}\sum_{t=1}^{T} x_t = \bar{x}_T$

**Claim** $\Phi(\bar{x}_t) - \min_{x \in K} \Phi(x^*) \leq \frac{Regret_T(A)}{T}$

**Proof**

$$\begin{aligned}
\Phi(\bar{x}_t) \leq & \frac{1}{T}\sum_{t=1}^{T} \Phi(x_t) \\
= & \frac{1}{T}\sum_{t=1}^{T} f(x_t) \\
= & \frac{1}{T}\min_{x \in K}\sum_{t=1}^{T} f_t(x) + \frac{Regret_T(A)}{T} \\
= & \min_{x \in K}\frac{1}{T}\sum_{t=1}^{T} \Phi(x) + \frac{Regret_T(A)}{T} \\
= & \min_{x \in K}\Phi(x) + \frac{Regret_T(A)}{T}
\end{aligned}$$

## 11.3 Algorithms with Online Convex Optimization

### 11.3.1 Online Gradient Descent

**Algorithm 3:** Online Gradient Descent

Initialize $x_1 \in K$
**for** $t$ = *1,...,T* **do**
  $x_{t+1} = \Pi_k(x_t - y_t \nabla f_t(x_t))$, where $\Pi_k(y) = \arg\min_{x \in K} \|y - x\|_2$
**end**

**Theorem 11.1** *Assume $f_t$ is G-Lipschitz, diam(K) = D, set*

$$y_t = \frac{D}{G\sqrt{t}}$$

*we have*

$$Regret_T(OGD) \leq \frac{3}{2} GD\sqrt{T}$$

### 11.3.2 Online batch conversion

Let X be data space and Y be label space, D represent distance over $X \cdot Y$, $\Theta \in \mathbb{R}^d$ be bounded parameter space. Set loss function

$$l : \Theta \times X \times Y \to \mathbb{R}$$

convex and lipschitz in $\Theta$. Given a sample of data, we have $(x_1, y_1), \dots, (x_n, y_n) \sim D$

### 11.3.3 Algorithm to use Online Convex Optimization for Generalization

**Algorithm 4:** Algorithm

**for** $t = 1, \dots, T$ **do**
  $\theta_t \leftarrow OCO(f_1, \dots, f_{t-1})$
  $F_t(\cdot) = l(\cdot, (x_t, y_t))$
**end**
Output: $\bar{\theta}_n = \frac{1}{n} \sum_{t=1}^{T} \theta_t$

**Theorem 11.2** *Let risk $R(\theta) := E_{x,y \sim D}[l(\theta, (x, y))]$, where $(x, y)$ is new data, then*

$$E[R(\bar{\theta}_T)] - \min_{\theta \in \Theta} R(\theta) \leq E\left[\frac{Regret_n(OCO)}{n}\right]$$