8 Convex Functions and Convex Optimization

8.1 Warmup: An Example of Online Optimization

In this section we consider an example of online optimization. Suppose you need to make some decisions with "expert" advice. However some of them are not really "experts". More precisely, our goal is to make sequential decisions over T rounds by aggregating advice from N "experts" to minimize cumulative loss, where each expert i provides a recommendation at each round and suffers loss $\ell_i^t \in [0,1]$ at round t. At each round, we select a distribution p^t over experts and incur loss $\sum_i p_i^t \ell_i^t$ in expectation. The *regret* is defined as the difference between the total expected loss and the best fixed expert's loss:

$$R_T = \sum_{t=1}^{T} \ell_{\text{alg}}^t - \min_i \sum_{t=1}^{T} \ell_i^t .$$

Since the loss ℓ_i^t is bounded in $[0,1]$, if we choose experts uniformly at random (i.e., let p^t be the uniform distribution over $[N]$) at each round, the regret can be $O(T)$. Our goal is to find a strategy so that the regret is only $o(T)$.

The Multiplicative Weights Update algorithm is an iterative method for decision-making with expert advice. It maintains weights over experts, updating them multiplicatively based on their performance to minimize regret. The algorithm can be described as follows

1. Initialize weights $w_i^1 = 1$ for all experts $i = 1, \dots, N$.
2. For each round $t = 1, \dots, T$:
   - Compute distribution $p_i^t = \frac{w_i^t}{\sum_j w_j^t}$.
   - Incur loss $\ell_{\text{alg}}^t = \sum_i p_i^t \ell_i^t$.
   - Update weights: $w_i^{t+1} = w_i^t \cdot \exp(-\eta \ell_i^t)$.

Theorem 8.1.
Taking $\eta = \sqrt{\frac{8 \ln N}{T}}$, the MWU algorithm gives $R_T \le \sqrt{\frac{T \ln N}{2}}$ after T iterations.

Remark 8.1.
To guarantee that the average regret does not exceed ε, namely,

$$\frac{1}{T} R_T = \frac{1}{T} \sum_{t=1}^{T} \ell_{\text{alg}}^t - \min_i \frac{1}{T} \sum_{t=1}^{T} \ell_i^t \le \varepsilon ,$$

it is sufficient to set $T = \Theta(\ln N / \varepsilon^2)$.

Proof. Let $W_t = \sum_i w_i^t$. Track the growth of W_t:

$$W_{t+1} = \sum_{i=1}^{N} w_i^t \cdot \exp(-\eta \ell_i^t) = W_t \sum_{i=1}^{N} p_i^t \cdot \exp(-\eta \ell_i^t) = W_t \cdot \mathbb{E}_{i \sim p_i^t}[\exp(-\eta \ell_i^t)] .$$

Recall that, from Hoeffding's lemma, for each random variable $X \in [a,b]$ with $\mathbb{E}[X] = 0$ and $\lambda > 0$, we have

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 (b-a)^2/8) ,$$

which gives

$$W_{t+1} \leq W_t \cdot \exp\Big(-\eta \cdot \mathbb{E}[\ell_i^t] + \frac{\eta^2}{8}\Big) .$$

Thus, we have

$$\ln \frac{W_{t+1}}{W_t} \leq -\eta \ell_{\mathsf{alg}}^t + \frac{\eta^2}{8} .$$

Telescoping over T rounds, it follows that

$$\ln W_{T+1} \leq \ln N - \eta \sum_{t=1}^{T} \ell_{\mathsf{alg}}^t + \frac{\eta^2 T}{8} .$$

Suppose the best expert is i^*. Note that for i^*, we obtain that

$$\ln w_{i^*}^{T+1} = -\eta \sum_{t=1}^{T} \ell_{i^*}^t \leq \ln W_{T+1}$$

Combining together all above, we conclude that

$$\sum \ell_{\mathsf{alg}}^t \leq \sum \ell_{i^*}^t + \frac{\ln N}{\eta} + \frac{\eta T}{8} .$$

Setting $\eta = \sqrt{\frac{8 \ln N}{T}}$ gives $R_T \leq \sqrt{\frac{T \ln N}{2}}$. □

To explain why this simple algorithm has such a good performance, we will introduce a viewpoint from (continuous) convex optimization.

8.2 Convex Sets

We first introduce convex sets, convex functions and convex optimization.

We begin with affine sets. Affine sets are generalization of lines. Given two points $x, y \in \mathbb{R}^n$, the line passing through x, y can be represented by

$$\ell = \{x + \theta(y-x) \mid \theta \in \mathbb{R}\} .$$

Note that $x + \theta(y-x) = (1-\theta)x + \theta y$. So we have the following definition of *affine combination*.

Definition 8.1. Affine combination and affine set
Given $x_1, \ldots, x_m \in R^n$, $\theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_m x_m$ is an affine combination of $x_1, \ldots, x_m$ if $\theta_1 + \ldots + \theta_m = 1$.
A set is affine if it is closed under affine combinations, namely, for all $m \geq 1$, for all m points $x_1, x_2, \ldots, x_m \in S$, any affine combination of $x_1, \ldots, x_m$ is still in S.

Example 8.1.
Here are some examples of affine sets:

- A line is an affine set;
- $\mathbb{R}^n$ is an affine set;
- Given $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, the hyperplane $P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{w}^\mathsf{T}\mathbf{x} + b = 0\}$ is an affine set;
- In general, given $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$, the solution set of the system of linear equations $S = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b}\}$ is an affine set.

In fact, to certify that a set S is affine, we only need to check affine combinations of any **two** points in S. Suppose we have an affine combination $\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$ for 3 points x_1, x_2, x_3. Since $\theta_1 + \theta_2 + \theta_3 = 1$, clearly there must exists two of them such that their sum is non-zero. Assume that $\theta_1 + \theta_2 \neq 0$. Then we have

$$\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 = (\theta_1 + \theta_2)\left(\frac{\theta_1}{\theta_1 + \theta_2} x_1 + \frac{\theta_2}{\theta_1 + \theta_2} x_2\right) + \theta_3 x_3 .$$

If any affine combination of two points is still in S, then $\frac{\theta_1}{\theta_1+\theta_2} x_1 + \frac{\theta_2}{\theta_1+\theta_2} x_2$ is in S and thus $\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$ is in S. For an affine combination of more than 3 points, we can rewrite it in a similar way recursively. So it suffices to verify affine combinations of 2 points.

We have shown that the solution to each linear equation is an affine set. Conversely, any affine set is also a solution set to a system of linear equations.

Proposition 8.2.
Any affine set $\subseteq \mathbb{R}^n$ is the solution set to a system of linear equations.

Proof. If S is an affine set, pick an arbitrary point $x_0 \in S$. Then we claim that the following set

$$S' = S - x_0 \triangleq \{x - x_0 \mid x \in S\}$$

is a linear space. For all $x_1, x_2 \in S'$, we have $x_1 + x_0, x_2 + x_0 \in S$ by definition. Hence, for any $a_1, a_2 \in \mathbb{R}$,

$$a_1 x_1 + a_2 x_2 + x_0 = a_1(x_1 + x_0) + a_2(x_2 + x_0) + (1 - a_1 - a_2)x_0 \in S .$$

Therefore, $a_1 x_1 + a_2 x_2 \in S'$.

Since S' can be represented as $\{x \mid Ax = 0\}$, then $S = S' + x_0$ can be represented as $\{x \mid Ax = Ax_0\}$, which is the solution set to $Ax = Ax_0$. □

Roughly speaking, *affine* can be viewed as *linear* added by some bias term. Similar to the *linear map*, we can define an *affine map* $f : \mathbb{R}^n \to \mathbb{R}^m$ by $x \mapsto Ax + b$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$. We can also define *affinely independent points* as follows.

Given $m + 1$ points $x_0, x_1, \ldots, x_m \in \mathbb{R}^n$, we say they are *affinely independent*, if there **does not** exist $\theta_0, \theta_1, \ldots, \theta_m \in \mathbb{R}$, not all zero, such that $\theta_0 + \theta_1 + \cdots + \theta_m = 0$, and

$$\theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_m x_m = 0 .$$

Equivalently, $x_0, x_1, \ldots, x_m \in \mathbb{R}^n$ are *affinely independent*, if and only if $x_1 - x_0, x_2 - x_0, \ldots, x_m - x_0$ are *linearly independent*.

Similar to the definition of lines, we can define the *segment* from x to y by

$$s = \{x + \theta(y - x) \mid \theta \in [0, 1]\}.$$

Note that the difference between lines and segments is the range of θ. Again, since $x + \theta(y - x) = (1 - \theta)x + \theta y$, we have the following definition.

Definition 8.2. Convex combination and convex set
Given $x_1, \ldots, x_m \in R^n$, $\theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_m x_m$ is a convex combination of $x_1, \ldots, x_m$ if $\theta_1 + \ldots + \theta_m = 1$, and for all $i \in [m]$, $\theta_i \geq 0$.
A set is convex if it is closed under convex combinations, namely, for all $m \geq 1$, for all m points $x_1, x_2, \ldots, x_m \in S$, any convex combination of $x_1, \ldots, x_m$ is still in S.

Again, to justify convexity, we only need to check any two points.

The *convex hull* of a set S is the set of all convex combinations of points in S, namely,

$$\operatorname{conv}(S) \triangleq \left\{ \sum_{i=1}^{m} \theta_i x_i \mid \forall\, i \in [m], \theta_i \geq 0, x_i \in S, \text{ and } \sum_{i=1}^{m} \theta_i = 1 \right\}.$$

Clearly, for any set $S \in \mathbb{R}^n$, its convex hull is a convex set.

If we would like to determine the convex hull of some set S, can we only check convex combinations of any two points? If not, how many points are sufficient?

Theorem 8.3. Carathéodory's theorem
For any $S \subseteq \mathbb{R}^n$, any point in $\operatorname{conv}(S)$ can be written as the convex combination of at most $n + 1$ points in S.

Proof. Suppose $y = \theta_1 x_1 + \cdots + \theta_m x_m$ is a convex combination of $m \geq n + 2$ points $x_1, \ldots, x_m$. We can show that y can also be written as a convex combination of $m - 1$ points among $x_1, \ldots, x_m$. Because $n + 2$ points are affinely dependent, there exists $\lambda_1, \ldots, \lambda_m$ such that $\lambda_1 + \cdots + \lambda_m = 0$, not all λ_i's are zero, and $\lambda_1 x_1 + \cdots + \lambda_m x_m = 0$. Thus, there exists $\lambda_i < 0$. For each i where $\lambda_i < 0$, let $\gamma_i = \theta_i / |\lambda_i|$, and let γ be the smallest number among them. Now we have

$$\begin{aligned} y &= \theta_1 x_1 + \cdots + \theta_m x_m + \gamma(\lambda_1 x_1 + \cdots + \lambda_m x_m) \\ &= (\theta_1 + \gamma\lambda_1) x_1 + \cdots + (\theta_m + \gamma\lambda_m) x_m. \end{aligned}$$

By definition, $\theta_i + \gamma\lambda_i \geq 0$ for all i, and there exists j such that $\theta_j + \gamma\lambda_j = 0$. Since $\sum_i \theta_i + \gamma\lambda_i = \sum_i \theta_i = 1$, we rewrite y as a convex combination of at most $m - 1$ points among $x_1, \ldots, x_m$. □

Example 8.2.
Let $\mathcal{S}^n_+$ and $\mathcal{S}^n_{++}$ denote the set of all *positive semidefinite matrices* and the set of all *positive definite matrices*, respectively, namely,

$$\begin{aligned} \mathcal{S}^n_+ &= \{A \in \mathbb{R}^{n \times n} \mid A \succeq 0\}, \\ \mathcal{S}^n_{++} &= \{A \in \mathbb{R}^{n \times n} \mid A \succ 0\}. \end{aligned}$$

Then both $\mathcal{S}^n_+$ and $\mathcal{S}^n_{++}$ are convex sets.

A fundamental property of convex sets is the separating hyperplane theorem, which states that disjoint convex sets can be *affinely separated*.