Announcements

- Lecture on 10/16 is rescheduled to 9:30AM-11AM. Location TBA.

9.1 Minimax Theorem Review

Minimax Theorem states

\[ \min_p \max_q p^T M q = \max_q \min_p p^T M q. \]

It means

\[ \exists p^* \forall q \text{ s.t. } p^*^T M q \leq V \quad \text{and} \quad \exists q^* \forall p \text{ s.t. } p M q^* \geq V \]

**Definition 9.1 (ϵ-Optimal).** The mixed strategies \( p \) and \( q \) are \( ϵ \)-optimal if there exists \( V \) such that for all \( p' \) and \( q' \),

\[ p^T M q' \leq V + \epsilon \quad \text{and} \quad p'^T M q \geq V - \epsilon \]

9.2 Approximately Solving a Linear Program

**Definition 9.2 (Linear Program).** A Linear program is a problem that can be expressed in the canonical form:

\[
\begin{align*}
\text{maximize} & \quad c^T x \\
\text{subject to} & \quad a^j^T x \leq b^j \quad \forall \ j \in I \quad \text{and} \quad x_i \geq 0 \quad \forall \ i \in [n]
\end{align*}
\]

**Lemma 9.3.** Without loss of generality, we can make the following modifications to the canonical form:

(i) Add an additional constraint that \( x \in \Delta_n \) (due to linearity)

(ii) Assume \( b_j = 0 \ \forall j \in I \), by spreading \( b_j \) into \( a^j \) using (i).

**Definition 9.4 (Linear Program).** An alternate form for linear programs is:

\[
\begin{align*}
\text{maximize} & \quad d \\
\text{subject to} & \quad a^j^T x \leq 0 \quad \forall \ j \in I, \ x \in \Delta_n, \ \text{and} \quad c^T x \geq d.
\end{align*}
\]

We will use the above definition unless noted otherwise.

**Definition 9.5 (Feasibility).** An LP is feasible if there exists \( x \in \Delta_n \) that satisfies all the constraints.

Given a feasibility checker, we can perform a binary search on \( d \) within the interval \([\min(c), \max(c)]\) (where the min and max are taken over the coordinates of \( c \)) to find an \( ϵ \)-optimal solution to LP with \( \log(1/ϵ) \) added time complexity.
Goal: Find an algorithm $A$ such that for any feasibility checker, $A$ returns $x \in \Delta_n$ such that $a_j^T x \leq \varepsilon$ for all $j \in I$, or INFEASIBLE if there is no such $x$.

Algorithm 1: EWA-LP

1. $x^{(1)} \leftarrow (\frac{1}{n}, \ldots, \frac{1}{n})$
2. for $t = 1$ to $T$
   3. $a^{(t)} \leftarrow \arg \max_{a \in [a_j : j \in I]} a^T x$
   4. if $a^{(t)} x^{(t)} \leq \varepsilon$ then
      5. return $x^{(t)}$
   else
      6. $x_i^{(t+1)} \leftarrow \frac{x_i^{(t)} \exp(-\eta a)}{\Phi(t+1)}$
    7. return INFEASIBLE

Proof. The regret bound of EWA-LP is

$$\frac{1}{T} \sum x^{(t)} a^{(t)} \leq \frac{1}{T} \text{Regret}_T$$

Suppose that LP is feasible but EWA-LP returns INFEASIBLE and $T > c^2 \log n / \varepsilon^2$. Then,

$$\varepsilon = \frac{1}{T} (T \varepsilon) \leq \frac{1}{T} \sum x^{(t)} a^{(t)} \leq \frac{1}{T} \text{Regret}_T = c \sqrt{\frac{\log n}{T}} < \varepsilon,$$

a contradiction. (The second inequality follows from the assumption that the LP is feasible).

Remark: The number of experts in EWA-LP is independent of the number of constraints.

9.3 Boosting via Minimax Duality

Setup: Let $\mathcal{X}$ be a data space (e.g. $\mathbb{R}^d$). We have a set of hypothesis $\{c : \mathcal{X} \rightarrow \{0, 1\}\}$, which contains a correct hypothesis. Let $C(x)$ be the true label $\forall x \in \mathcal{X}$. We want to find $\hat{c} \in c$ such that the error rate

$$P_{x \sim q}[\hat{c}(x) \neq C(x)]$$

is small for any distribution $q \in \Delta(\mathcal{X})$.

Weak Hypotheses: It is easy to find weak hypotheses $\mathcal{H} = \{h : \mathcal{X} \rightarrow \{0, 1\}\}$.

Example: If $\mathcal{X} \subseteq \mathbb{R}^n$, then define $H = \{h_{i,c} : i = 1, \ldots, n \text{ and } c \in \mathbb{R}\}$ where

$$h_{i,c}(x) = \begin{cases} 1 & \text{if } x_i \geq c \\ 0 & \text{otherwise} \end{cases}$$

The function $h_{i,c}$ is called a decision stump.
Weak Learning Assumption: For a positive constant $\gamma$, the weak learning assumption states:
For any distribution $q \in \Delta(\mathcal{X})$, there exists $h \in \mathcal{H}$ such that
\[
P_{x \sim q}[h(x) \neq C(x)] \leq \frac{1}{2} - \frac{\gamma}{2}
\]

Question: Is there a distribution $p$ on $\mathcal{H}$ such that the weighted majority
\[
c_p(x) = \begin{cases} 
1 & \text{if } \sum_{h \in \mathcal{H}} p(h) h(x) \geq \frac{1}{2} \\
0 & \text{otherwise}
\end{cases}
\]
achieves zero error (a.k.a. strong learning)?


Proof. Suppose $\mathcal{H}$ satisfies the weak learning assumption, and let $x_1, \ldots, x_n$ be the data. Let $M$ be a $n \times |\mathcal{H}|$ matrix such that
\[
M_{ij} = \begin{cases} 
+1 & \text{if } h_i(x_j) \neq c(x_j) \\
-1 & \text{otherwise}
\end{cases}
\]

The weak learning assumption states that for any $q \in \Delta_n$ there exists $j \in [m]$ such that
\[
P_{x \sim q}[h_j(x) \neq C(x)] \leq \frac{1}{2} - \frac{\gamma}{2}
\]
This is equivalent to
\[
q^T Me_j \leq -\gamma
\]
which in turn is equivalent to
\[
\min_{q} \max_{j} q^T Me_j \leq -\gamma
\]
By the minimax theorem, the above is true iff it's dual is. The dual
\[
\exists p \in \Delta_n \text{ s.t. } e_i^T M p \leq \gamma \ \forall \ i \in [n]
\]
is exactly strong learning. \qed

\footnote{Diagram credit: Cat Saint Croix}