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

# Lecture 8: Perceptron and Game Theory

*Lecturer: Jacob Abernethy* *Scribes: Aditi Laddha, Jiayu Chen*

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

## 8.1 Review of Prediction Setting

Given the loss function $\ell :\in [0,1] \times \{0,1\} \to [0,1]$, a pool of N experts, and an learning algorithm.

**Algorithm** Prediction with Expert Advice

1: **for** t=1$\to T$ **do**:
2: Experts predict $x_1^t \dots x_N^t \in \{0,1\}$
3: Algorithm makes prediction $\hat{y}^t \in \{0,1\}$
4: Observe $y^t \in \{0,1\}$
5: **end for**

**Definition 8.1 (Regret)** *In a prediction setting, regret is defined as the difference between the loss of the learning algorithm, and the loss of the best expert.*

$$Regret_T = \sum_{t=1}^{T} \ell(\hat{y}^t, y^t) - \min_i \sum_{t=1}^{T} \ell(x_i^t, y^t)$$

*Goal: to make $Regret_T$ small. It would be ideal if* $\frac{Regret_T}{T} = o(1)$.

## 8.2 Review of Exponential Weights Algorithm

Given a pool of N experts, each expert $i$ give prediction $x_i^t$ at each round $t$ .Given a parameter $\eta$.

**Algorithm** Exponential Weights Algorithm

1: $w_i^1 \leftarrow 1 \quad \forall i$
2: **for** t=1$\to T$ **do**:
3: Algorithm makes prediction $\hat{y} = \frac{\sum_{i=1}^{t} w_i^t x_i^t}{\sum_{i=1}^{t} w_i^t}$
4: Observe $y^t \in \{0,1\}$
5: $w_i^{t+1} = e^{-\eta \sum_{s=1}^{t} \ell(x_i^s, y^s)}$
6: **end for**

**Theorem 8.2** *Let $\mathcal{L}_T(alg)$be the accumulated losses for EWA, $\mathcal{L}_T i$ be the accumulated losses for each expert $i$, then EWA guanrantees that $\forall i$:*

$$\mathcal{L}_T(alg) \leq \frac{logN + \eta \mathcal{L}_T i}{1 - e^{\eta}}$$

**Corollary 8.3** *For 'correct' $\eta$, which means well-tuned $\eta$, we have:*

$$\frac{\mathcal{L}_T(alg) - \mathcal{L}_T i}{T} \leq \frac{logN + \sqrt{2logN\mathcal{L}_T i^*}}{T} = O(\frac{1}{\sqrt{T}})$$

*Where* $i^* = argmin_i \mathcal{L}_T i$.
*The last equation holds because the growth of* $\mathcal{L}_T i^*$ *is always slower than that of* $T$.

## 8.3 Action Setting/Hedge Setting

There are no predictions from experts, but instead the algorithm chooses actions. On each round t, there are N actions to choose from. Then the nature will reveal the associated loss. The setting is actually equivalent to previous setting.

**Algorithm** Hedge Setting

$w_i^1 \leftarrow 1 \quad \forall i$
**for** t=1$\rightarrow T$ **do**:
  Choose $\vec{p^t} \in \triangle_N$
  Observe $\vec{\ell^t} \in [0,1]^N$
  Algorithm pays $\vec{p^t}\vec{\ell^t}$
  The weight gets updated $w_i^{t+1} = e^{-\eta \sum_{s=1}^t \ell_i^t}$
**end for**

$\triangle_N$ is the set of discrete probability distribution of N choices, and $\forall i, \quad p_i \in [0,1], \sum_{i=1}^N p_i = 1$.

## 8.4 Linear Prediction Setting

First, we define sign(x) $= \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{otherwise} \end{cases}$

**Algorithm** Linear Prediction Setting

**for** t=1$\rightarrow T$ **do**:
  Observe $\vec{x}^t \in \mathbb{R}^d, \|x\|_2 \leq 1$
  Algorithm predicts $\hat{y^t} \in \{-1,1\}$
  Observe outcome $y^t \in \{-1,1\}$
**end for**

**Definition 8.4 (Linear Predictor)** *A function* $h_w(\cdot)$ *parameterized by the vector* $\vec{w} \in \mathbb{R}^d$

$$h_w(\cdot) = sign(\vec{w} \cdot \vec{x})$$

**Definition 8.5 (Perfect Linear Predictor)** *In this setting, we assume that there exists* $\vec{w}^* \in \mathbb{R}^d, \|\vec{w}^*\|^2 \leq 1$*, called a the perfect linear predictor such that*

$$sign(\vec{w}^* \cdot \vec{x}^t) = y^t, \quad \forall t$$

**Definition 8.6 (Perfect Linear Predictor with $\gamma$ margin)** *Assume for some* $\gamma > 0$*, there exists* $\vec{w}^* \in \mathbb{R}^d, \|\vec{w}^*\|^2 \leq 1$*, called a the perfect linear predictor such that*

$$(\vec{w}^* \cdot \vec{x}^t)y^t > \gamma, \quad \forall t$$

*This can be equivalently stated as there exists* $\vec{w}^* \in \mathbb{R}^d, \|\vec{w}^*\|^2 \leq \frac{1}{\gamma^2}$*, called a the perfect linear predictor with* $\gamma$ *margin, such that*

$$(\vec{w}^* \cdot \vec{x}^t)y^t > 1, \quad \forall t$$

In the Linear Prediction Setting, the assumption of the existence of a perfect linear predictor is similar to the assumption of a perfect expert in Prediction with Expert Advice. The perceptron algorithm for minimizing the number of mistakes in linear prediction setting is described below.

**Algorithm** Perceptron

```
w⃗^1 ← 0⃗ ∈ ℝ^d
for t = 1 → T do:
    Observe x⃗^t
    Algorithm predicts ŷ^t = sign(w⃗^t · x⃗^t)
    Observe y^t
    if y^t(w⃗^t x⃗^t) > 0 then
        w⃗^{t+1} = w⃗^t
    else
        w⃗^{t+1} = w⃗^t + y^t x⃗^t
    end if
end for
```

**Definition 8.7 (Hinge Loss)** *Given $\vec{w}, \vec{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}$, the Hinge Loss is defined as*

$$\ell(\vec{w}, (\vec{x}, y)) = \max\{0, -\vec{w}^\top \vec{x} y\}$$

The perceptron algorithm can be thought of as gradient descent with the loss function as Hinge loss.

$$\vec{w}^{t+1} = \vec{w}^t - \nabla \ell(\vec{w}^t, (\vec{x}^t, y^t))$$

We are using the assumption that there exists a perfect linear predictor with $\gamma$ margin in the mistake analysis.

**Lemma 8.8** *for any vectors $\vec{a}, \vec{b}$,*

$$\|\vec{a}\|^2 - \|\vec{a} - \vec{b}\|^2 = 2(\vec{a} \cdot \vec{b}) - \|\vec{b}\|^2$$

**Theorem 8.9** *Let $M_T = \sum_{t=1}^T 1[(\vec{w}^{t\top} \vec{x}^t) y^t < 0]$. Assume that there exists $\vec{w}^* \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}^+$ such that $\|\vec{w}^*\|_2 \leq \frac{1}{\gamma}$ and $y^t(\vec{w}^{*\top} \vec{x}^t) \geq 1, \forall t$. Then $M_T \leq \dfrac{1}{\gamma^2}$*

**Proof:** Let $\vec{w}^*$ satisfy the assumption. And let $\text{Mistake}(T) = \{t \in [T] : (\vec{w}^{t\top} \vec{x}^t) y^t < 0\}$.
Define $\Phi_t = \|\vec{w}^* - \vec{w}^t\|_2^2$. As $\vec{w^1} = \vec{1}$, $\Phi_1 = \|\vec{w}^*\|^2 \leq \frac{1}{\gamma^2}$. Also, $\Phi_{T+1} = \|\vec{w}^* - \vec{w}^{T+1}\|^2 \geq 0$.

$$\begin{aligned}
\frac{1}{\gamma^2} &\geq \Phi_1 - \Phi_{T+1} \\
&= \sum_{t=1}^{T} (\Phi_t - \Phi_{t+1}) \\
&= \sum_{t=1}^{T} (\|\vec{w}^* - \vec{w}^t\|^2 - \|\vec{w}^* - \vec{w}^{t+1}\|^2) \\
&= \sum_{t \in \text{Mistake}(T)} (\|\vec{w}^* - \vec{w}^t\|^2 - \|\vec{w}^* - \vec{w}^t - \vec{x}^t y^t\|^2) \\
&= \sum_{t \in \text{Mistake}(T)} (-\|\vec{x}^t y^t\|^2 + 2(\vec{w}^* - \vec{w}^t)^\top \vec{x}^t y^t) && \text{(using Lemma 8.8)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{t \in \text{Mistake}(T)} (-\|\vec{x}^t\|^2 + 2\vec{w}^{*\top}\vec{x}^t y^t - 2\vec{w}^{t\top}\vec{x}^t y^t) \\
&\geq \sum_{t \in \text{Mistake}(T)} -1 + 2 - 2\vec{w}^{t\top}\vec{x}^t y^t && (\forall t,\ \|x^t\| \leq 1 \text{ and } \vec{w}^{*\top}\vec{x}^t y^t \geq 1) \\
&\geq \sum_{t \in \text{Mistake}(T)} 1 && (\forall t \in \text{Mistake}(T),\ \vec{w}^{t\top}\vec{x}^t y^t < 0) \\
&= M_T
\end{aligned}
$$

■

## 8.5 Introduction to Game Theory

**Definition 8.10** *A two person bimatrix game is defined by matrices $M \in \mathbb{R}^{n\times m}$ and $N \in \mathbb{R}^{n\times m}$ where each player selects a distribution $\vec{p} \in \Delta_n$ and $\vec{q} \in \Delta_m$. If player 1 chooses $i \in [n]$ and player 2 chooses $j \in [m]$, the utility for player 1 is defined to be $U^1(i,j) = M_{ij}$ and the utility for player 2 is defined to be $U^2(i,j) = N_{ij}$.*

Typically, we are in the randomized setting with $U^1(\vec{p},\vec{q}) = \mathop{\mathbb{E}}_{\substack{i\sim\vec{p}\\ j\sim\vec{q}}}[M_{ij}] = \vec{p}^\top M\vec{q}$. Similarly, $U^2(\vec{p},\vec{q}) = \vec{p}^\top N\vec{q}$.

**Example (Rock-Paper-Scissor)** In this game, each player choose among Rock(1), Paper(2) or Scissor(3).Scissor wins Paper, Paper wins Rock, and Rock wins Scissor. For the outcome, 1 means win, -1 means lose, and 0 means even. Each input $x_{i,j}$ in M represents the outcome of player 1 choosing i and player 2 choosing j. Each input $x_{i,j}$ in N represents the outcome of player 2 choosing i and player 1 choosing j.

$$
M = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix} \text{ and } N = \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}
$$

If $M + N = 0$, then the game is called a zero-sum game.

**Definition 8.11 (Nash Equilibrium)** *Given $M, N \in \mathbb{R}^{n\times m}$ and $\vec{p} \in \Delta_n, \vec{q} \in \Delta_m$, $(\vec{p},\vec{q})$ is a Nash Equilibrium if*

$$
\vec{p}^\top M\vec{q} \geq \vec{\tilde{p}}^\top M\vec{q},\ \forall\vec{\tilde{p}} \in \Delta_n \text{ and } \vec{p}^\top N\vec{q} \geq \vec{p}^\top N\vec{\tilde{q}},\ \forall\vec{\tilde{q}} \in \Delta_m
$$

**Theorem 8.12 (Nash's Theorem)** *For any $M, N \in \mathbb{R}^{n\times m}$, there exists a Nash Equilibrium $(\vec{p},\vec{q})$.*

**Theorem 8.13 (Minimax Theorem)** *Let $M \in \mathbb{R}^{n\times m}$ then*

$$
\min_{\vec{p}\in\Delta_n} \max_{\vec{q}\in\Delta_m} \vec{p}^\top M\vec{q} = \max_{\vec{q}\in\Delta_m} \max_{\vec{p}\in\Delta_n} \vec{p}^\top M\vec{q}
$$