# Online Learning

Vikram M Tankasali

Machine Learning Lab, CSA Department
Indian Institute of Science.

# AGENDA

- Problem and Motivation
- Prediction with Expert Advice
- Online Convex Optimization

# Introduction

- Perceptron Algorithm.
- Sequential Prediction vs Online Prediction.

# Introduction

- Perceptron Algorithm.
- Sequential Prediction vs Online Prediction.
- Risk vs Regret.

# Introduction

- Perceptron Algorithm.
- Sequential Prediction vs Online Prediction.
- Risk vs Regret.
- Disparate areas : Information Theory(Data Compression, Gambling), Game Theory(Repeated Games), Online Algorithms .

# Introduction

- Perceptron Algorithm.
- Sequential Prediction vs Online Prediction.
- Risk vs Regret.
- Disparate areas : Information Theory(Data Compression, Gambling), Game Theory(Repeated Games), Online Algorithms .

# A Gentle Start

- **Problem** : Predicting an unknown sequence $y_1, y_2 \ldots$ of bits $y_t \in \{0, 1\}$.
  At time $t$ the forecaster makes a guess $p_t \in \{0, 1\}$ for $y_t$ based on $(f_{1,t}, \ldots, f_{N,t})$. Each $f_{i,t} \in \{0, 1\}$ is the prediction of an expert $i$. The true bit $y_t$ is revealed and forecaster suffers a loss $l(p_t, y_t)$.
- **Goal** :Bound the number of time steps $t$ in which $p_t \neq y_t$.

# A Gentle Start

- **Problem** : Predicting an unknown sequence $y_1, y_2 \ldots$ of bits $y_t \in \{0, 1\}$.
  At time $t$ the forecaster makes a guess $p_t \in \{0, 1\}$ for $y_t$ based on $(f_{1,t}, \ldots, f_{N,t})$. Each $f_{i,t} \in \{0, 1\}$ is the prediction of an expert $i$. The true bit $y_t$ is revealed and forecaster suffers a loss $l(p_t, y_t)$.
- **Goal** :Bound the number of time steps $t$ in which $p_t \neq y_t$.
- Suppose there is some expert $i$ that makes no mistakes. What will be your strategy?

# A Gentle Start

- **Problem** : Predicting an unknown sequence $y_1, y_2 \ldots$ of bits $y_t \in \{0, 1\}$.
  At time $t$ the forecaster makes a guess $p_t \in \{0, 1\}$ for $y_t$ based on $(f_{1,t}, \ldots, f_{N,t})$. Each $f_{i,t} \in \{0, 1\}$ is the prediction of an expert $i$. The true bit $y_t$ is revealed and forecaster suffers a loss $l(p_t, y_t)$.
- **Goal** :Bound the number of time steps $t$ in which $p_t \neq y_t$.
- Suppose there is some expert $i$ that makes no mistakes. What will be your strategy?
- Analysis in more general setting.

# A Gentle Start

- **Problem** : Predicting an unknown sequence $y_1, y_2 \ldots$ of bits $y_t \in \{0, 1\}$.
  At time $t$ the forecaster makes a guess $p_t \in \{0, 1\}$ for $y_t$ based on $(f_{1,t}, \ldots, f_{N,t})$. Each $f_{i,t} \in \{0, 1\}$ is the prediction of an expert $i$. The true bit $y_t$ is revealed and forecaster suffers a loss $l(p_t, y_t)$.
- **Goal** :Bound the number of time steps $t$ in which $p_t \neq y_t$.
- Suppose there is some expert $i$ that makes no mistakes. What will be your strategy?
- Analysis in more general setting.

# Protocol

- **PREDICTION WITH EXPERT ADVICE**
  Parameters: decision space $\mathbb{D}$, outcome space $\mathbb{Y}$, loss function l, set $\mathbb{E}$ of expert indices.
  For each round $t = 1, 2, \ldots$
  (1) the environment chooses the next outcome $y_t$ and the expert advice $\{f_{E,t} \in \mathbb{D} : E \in \mathbb{E}\}$; the expert advice is revealed to the forecaster;

# Protocol

- **PREDICTION WITH EXPERT ADVICE**
  Parameters: decision space $\mathbb{D}$, outcome space $\mathbb{Y}$, loss function l, set $\mathbb{E}$ of expert indices.
  For each round $t = 1, 2, \ldots$
  (1) the environment chooses the next outcome $y_t$ and the expert advice $\{f_{E,t} \in \mathbb{D} : E \in \mathbb{E}\}$; the expert advice is revealed to the forecaster;
  (2) the forecaster chooses the prediction $p_t \in \mathbb{D}$;

# Protocol

- **PREDICTION WITH EXPERT ADVICE**
  Parameters: decision space $\mathbb{D}$, outcome space $\mathbb{Y}$, loss function l, set $\mathbb{E}$ of expert indices.
  For each round $t = 1, 2, \ldots$
  (1) the environment chooses the next outcome $y_t$ and the expert advice $\{f_{E,t} \in \mathbb{D} : E \in \mathbb{E}\}$; the expert advice is revealed to the forecaster;
  (2) the forecaster chooses the prediction $p_t \in \mathbb{D}$;
  (3) the environment reveals the next outcome $y_t \in \mathbb{Y}$;

# Protocol

- **PREDICTION WITH EXPERT ADVICE**
  Parameters: decision space $\mathbb{D}$, outcome space $\mathbb{Y}$, loss function l, set $\mathbb{E}$ of expert indices.
  For each round $t = 1, 2, \ldots$
  (1) the environment chooses the next outcome $y_t$ and the expert advice $\{f_{E,t} \in \mathbb{D} : E \in \mathbb{E}\}$; the expert advice is revealed to the forecaster;
  (2) the forecaster chooses the prediction $p_t \in \mathbb{D}$;
  (3) the environment reveals the next outcome $y_t \in \mathbb{Y}$;
  (4) the forecaster incurs loss $l(p_t, y_t)$ and each expert $\mathbb{E}$ incurs loss $l(f_{E,t}, y_t)$.

# Protocol

- **PREDICTION WITH EXPERT ADVICE**
  Parameters: decision space $\mathbb{D}$, outcome space $\mathbb{Y}$, loss function l, set $\mathbb{E}$ of expert indices.
  For each round $t = 1, 2, \ldots$
  (1) the environment chooses the next outcome $y_t$ and the expert advice $\{f_{E,t} \in \mathbb{D} : E \in \mathbb{E}\}$; the expert advice is revealed to the forecaster;
  (2) the forecaster chooses the prediction $p_t \in \mathbb{D}$;
  (3) the environment reveals the next outcome $y_t \in \mathbb{Y}$;
  (4) the forecaster incurs loss $l(p_t, y_t)$ and each expert $\mathbb{E}$ incurs loss $l(f_{E,t}, y_t)$.

# GOAL

- Regret *w.r.t* Expert *E*.

$$R_{E,n} = \sum_{t=1}^{n} (l(p_t, y_t) - l(f_{E,t}, y_t)) = L_n - L_{E,n} \tag{1}$$

- GOAL .

$$\max_{i=1,\ldots,N} R_{i,n} = o(n) \quad \text{OR as n} \rightarrow \infty \quad \frac{1}{n}(L_n - \min_{i=1,\ldots,N} L_{i,n} \rightarrow 0) \tag{2}$$

# GOAL

- Regret *w.r.t* Expert *E*.

$$R_{E,n} = \sum_{t=1}^{n} (l(p_t, y_t) - l(f_{E,t}, y_t)) = L_n - L_{E,n} \tag{1}$$

- GOAL .

$$\max_{i=1,\ldots,N} R_{i,n} = o(n) \quad \text{OR as n } \rightarrow \infty \quad \frac{1}{n}(L_n - \min_{i=1,\ldots,N} L_{i,n} \rightarrow 0) \tag{2}$$

# Weighted Average Prediction I

- STRATEGY :

$$p_t = \frac{\sum_{i=1}^{N} w_{i,t-1} f_{i,t}}{\sum_{j=1}^{N} w_{j,t-1}} \tag{3}$$

  where $w_{i,t-1} \geq 0$ are weights assigned to experts at time t.
- Expert's weight is an increasing function of experts regret $R_{i,t-1} = L_{t-1} - L_{i,t-1}$.
- Let $w_{i,t-1} = \phi'(R_{i,t-1})$ where $\phi : \mathbb{R} \to \mathbb{R}$ is convex , nonnegative and increasing function.
- If the loss function $l$ is convex in its first argument then

$$\sup_{y_t \in \mathbb{Y}} \sum_{i=1}^{N} r_{i,t} \phi'(R_{i,t-1}) \leq 0 \tag{4}$$

# Weighted Average Prediction II

- **BlackWell Condition** : Let $r_t = (r_{1,t}, \ldots, r_{N,t}) \in \mathbb{R}^N$ be instantaneous regret vector. Let $\mathbf{R}_n = \sum_{t=1}^{n} r_t$.

$$\sup_{y_t \in \mathbb{Y}} \langle r_t, \nabla \Phi(\mathbf{R}_{t-1}) \rangle \leq 0 \tag{5}$$

  where $\Phi : \mathbb{R}^N \to \mathbb{R}$ is of the form $\Phi(u) = \psi(\sum_{i=1}^{n} \phi(u_i))$. Assumptions on $\Phi$ and $\psi$.

- The weighted average forecaster :

$$p_t = \frac{\sum_{i=1}^{N} \nabla \Phi(\mathbf{R}_{t-1})_i f_{i,t}}{\sum_{j=1}^{N} \nabla \Phi(\mathbf{R}_{t-1})_j} \tag{6}$$

# Weighted Average Prediction III

**Theorem**

*Assume that a forecaster satisfies the Blackwell condition for a potential* $\Phi(u) = \psi(\sum_{i=1}^{N} \phi(u_i))$*. Then for all* $n = 1, 2, \ldots,$

$$\Phi(\mathbf{R}_n) \leq \Phi(\mathbf{0}) + \frac{1}{2}\sum_{t=1}^{n} C(r_t) \tag{7}$$

*where* $C(r_t) = \sup_{u \in \mathbf{R}^N} \psi'(\sum_{i=1}^{N} \phi(u_i)) \sum_{i=1}^{N} \phi``(u_i) r_{i,t}^2$

-

# Weighted Average Prediction IV

- By Monotonicity of $\phi$ and $\psi$

$$\psi(\phi(\max_{i=1,\dots,N} R_{i,n})) = \psi(\max_{i=1,\dots,N} \phi(R_{i,n})) \leq \psi(\sum_{i=1}^{N} \phi(R_{i,n})) = \Phi(R_n) \tag{8}$$

- If $\psi$ is invertible then $\max_{i=1,\dots,n} R_{i,n} \leq \phi^{-1}(\psi^{-1}(\Phi(\mathbf{R}_n)))$

# Polynomially Weighted Average Forecaster

- $\Phi_p(\mathbf{u}) = (\sum_{i=1}^{N}(u_i)_+^p)^{\frac{2}{p}} = \|\mathbf{u}_+\|_p^2$ where $p \geq 2$
  The weights assigned to the experts are given by
  $w_{i,t-1} = \nabla\Phi_p(\mathbf{R}_{t-1})_i = \frac{2(R_{i,t-1})_+^{p-1}}{\|(\mathbf{R}_{t-1})_+\|_p^{p-2}}$

**Theorem**

*Assume that the loss function $l$ is convex in its first argument and that it takes values in $[0,1]$. Then, for any sequence $y_1, y_2, \ldots, \in \mathbb{Y}$ of outcomes and for any $n \geq 1$, the regret of the polynomially weighted average forecaster satisfies*
$L_n - \min_{i=1,\ldots,n} L_{i,n} \leq \sqrt{n(p-1)N^{\frac{2}{p}}}$

-

# Polynomially Weighted Average Forecaster

- $\Phi_p(\mathbf{u}) = (\sum_{i=1}^{N}(u_i)_+^p)^{\frac{2}{p}} = \|\mathbf{u}_+\|_p^2$ where $p \geq 2$
  The weights assigned to the experts are given by
  $w_{i,t-1} = \nabla\Phi_p(\mathbf{R}_{t-1})_i = \frac{2(R_{i,t-1})_+^{p-1}}{\|(\mathbf{R}_{t-1})_+\|_p^{p-2}}$

**Theorem**

*Assume that the loss function $l$ is convex in its first argument and that it takes values in $[0,1]$. Then, for any sequence $y_1, y_2, \ldots, \in \mathbb{Y}$ of outcomes and for any $n \geq 1$, the regret of the polynomially weighted average forecaster satisfies*
$L_n - \min_{i=1,\ldots,n} L_{i,n} \leq \sqrt{n(p-1)N^{\frac{2}{p}}}$

-

# Exponentially Weighted Average Forecaster

- $\Phi_\eta(\mathbf{u}) = \frac{1}{\eta}\ln(\sum_{i=1}^{N} e^{\eta u_i})$
- $w_{i,t-1} = \nabla\Phi_\eta(\mathbf{R}_{t-1})_i = \frac{e^{\eta R_{i,t-1}}}{\sum_{j=1}^{N} e^{\eta R_{j,t-1}}}$
- $w_{i,t} = \frac{w_{i,t-1}e^{-\eta l(f_{i,t},y_t)}}{\sum_{j=1}^{N} w_{j,t-1}e^{-\eta l(f_{j,t-1},y_t)}}$

## Theorem

*Assume that the loss function l is convex in its first argument and that it takes values in* $[0, 1]$. *Then, for any n ,* $\eta > 0$ *and any sequence* $y_1, y_2, \ldots, \in \mathbb{Y}$ *of outcomes and for any* $n \geq 1$, *the regret of the exponentially weighted average forecaster satisfies* $L_n - \min_{i=1,\ldots,n} L_{i,n} \leq \frac{\ln N}{\eta} + \frac{n\eta}{2}$

-

# Exponentially Weighted Average Forecaster

- $\Phi_\eta(\mathbf{u}) = \frac{1}{\eta}\ln(\sum_{i=1}^{N} e^{\eta u_i})$
- $w_{i,t-1} = \nabla\Phi_\eta(\mathbf{R}_{t-1})_i = \frac{e^{\eta R_{i,t-1}}}{\sum_{j=1}^{N} e^{\eta R_{j,t-1}}}$
- $w_{i,t} = \frac{w_{i,t-1} e^{-\eta l(f_{i,t}, y_t)}}{\sum_{j=1}^{N} w_{j,t-1} e^{-\eta l(f_{j,t-1}, y_t)}}$

**Theorem**

*Assume that the loss function l is convex in its first argument and that it takes values in* $[0,1]$. *Then, for any n ,* $\eta > 0$ *and any sequence* $y_1, y_2, \ldots, \in \mathbb{Y}$ *of outcomes and for any* $n \geq 1$*, the regret of the exponentially weighted average forecaster satisfies* $L_n - \min_{i=1,\ldots,n} L_{i,n} \leq \frac{\ln N}{\eta} + \frac{n\eta}{2}$

-

# Problem Formulation

- A convex repeated game is a two players game that is performed in a sequence of consecutive rounds. On round t of the repeated game, the first player chooses a vector $w_t$ from a convex set $A$. Next, the second player responds with a convex function $g_t : A \to \mathbb{R}$. Finally, the first player suffers an instantaneous loss $g_t(w_t)$. We study the game from the viewpoint of the first player.
- Different from offline convex optimization

# Problem Formulation

- A convex repeated game is a two players game that is performed in a sequence of consecutive rounds. On round t of the repeated game, the first player chooses a vector $w_t$ from a convex set $A$. Next, the second player responds with a convex function $g_t : A \to \mathbb{R}$. Finally, the first player suffers an instantaneous loss $g_t(w_t)$. We study the game from the viewpoint of the first player.
- Different from offline convex optimization
- Low Regret $R(T) = \frac{1}{T}\sum_{t=1}^{T} g_t(w_t) - \min_{w \in A} \frac{1}{T}\sum_{t=1}^{T} g_t(w)$

# Problem Formulation

- A convex repeated game is a two players game that is performed in a sequence of consecutive rounds. On round t of the repeated game, the first player chooses a vector $w_t$ from a convex set $A$. Next, the second player responds with a convex function $g_t : A \to \mathbb{R}$. Finally, the first player suffers an instantaneous loss $g_t(w_t)$. We study the game from the viewpoint of the first player.
- Different from offline convex optimization
- Low Regret $R(T) = \frac{1}{T}\sum_{t=1}^{T} g_t(w_t) - \min_{w \in A} \frac{1}{T}\sum_{t=1}^{T} g_t(w)$

# Basic

- $\lambda$ is a sub-gradient of a function $f$ at $w$ if for all $u \in A$ we have that $f(u) - f(w) \geq \langle u - w, \lambda \rangle$.
- $f : A \rightarrow \mathbb{R}$ is $\rho$-Lipschitz if for all $u, v \in A$ $|f(u) - f(v)| \leq \rho \|u - v\|$. For a convex function function $f$ if for all $x \in A$ and $v \in \partial f(x)$ then $\|v\|_* \leq \rho$.

# Basic

- $\lambda$ is a sub-gradient of a function $f$ at $w$ if for all $u \in A$ we have that $f(u) - f(w) \geq \langle u - w, \lambda \rangle$.
- $f : A \to \mathbb{R}$ is $\rho$-Lipschitz if for all $u, v \in A$ $|f(u) - f(v)| \leq \rho \|u - v\|$. For a convex function function $f$ if for all $x \in A$ and $v \in \partial f(x)$ then $\|v\|_* \leq \rho$.
- Fenchel Conjugate $f^*(\theta) = \max_x \langle x, \theta \rangle - f(x)$.

# Basic

- $\lambda$ is a sub-gradient of a function $f$ at $w$ if for all $u \in A$ we have that $f(u) - f(w) \geq \langle u - w, \lambda \rangle$.
- $f : A \to \mathbb{R}$ is $\rho$-Lipschitz if for all $u, v \in A$ $|f(u) - f(v)| \leq \rho \|u - v\|$. For a convex function function $f$ if for all $x \in A$ and $v \in \partial f(x)$ then $\|v\|_* \leq \rho$.
- Fenchel Conjugate $f^*(\theta) = \max_x \langle x, \theta \rangle - f(x)$.
- Assume that $f$ is a closed and convex function. Then $f$ is $\beta$-strongly convex w.r.t. a norm $\|.\|$ if and only if $f^*$ is $\frac{1}{\beta}$-strongly smooth w.r.t. the dual norm $\|.\|_*$.

# Basic

- $\lambda$ is a sub-gradient of a function $f$ at $w$ if for all $u \in A$ we have that $f(u) - f(w) \geq \langle u - w, \lambda \rangle$.
- $f : A \to \mathbb{R}$ is $\rho$-Lipschitz if for all $u, v \in A$ $|f(u) - f(v)| \leq \rho \|u - v\|$. For a convex function function $f$ if for all $x \in A$ and $v \in \partial f(x)$ then $\|v\|_* \leq \rho$.
- Fenchel Conjugate $f^*(\theta) = \max_x \langle x, \theta \rangle - f(x)$.
- Assume that $f$ is a closed and convex function. Then $f$ is $\beta$-strongly convex w.r.t. a norm $\|.\|$ if and only if $f^*$ is $\frac{1}{\beta}$-strongly smooth w.r.t. the dual norm $\|.\|_*$.
- If $f$ is $\beta$- strongly convex w.r.t. $\|.\|$ and $f^*(0) = 0$, then, denoting the partial sum $\sum_{j \leq i} v_j$ by $v_{1:i}$, we have, for any sequence $v_1, \ldots v_n$ and for any $u$, $\sum_{i=1}^{n} \langle v_i, u \rangle \leq f^*(v_{1:n}) \leq \sum_{i=1}^{n} \langle \nabla f^*(v_{1:i-1}), v_i \rangle + \frac{1}{2}\beta \sum_{i=1}^{n} \|v_i\|_*^2$

# Basic

- $\lambda$ is a sub-gradient of a function $f$ at $w$ if for all $u \in A$ we have that $f(u) - f(w) \geq \langle u - w, \lambda \rangle$.
- $f : A \to \mathbb{R}$ is $\rho$-Lipschitz if for all $u, v \in A$ $|f(u) - f(v)| \leq \rho \|u - v\|$. For a convex function function $f$ if for all $x \in A$ and $v \in \partial f(x)$ then $\|v\|_* \leq \rho$.
- Fenchel Conjugate $f^*(\theta) = \max_x \langle x, \theta \rangle - f(x)$.
- Assume that $f$ is a closed and convex function. Then $f$ is $\beta$-strongly convex w.r.t. a norm $\|.\|$ if and only if $f^*$ is $\frac{1}{\beta}$-strongly smooth w.r.t. the dual norm $\|.\|_*$.
- If $f$ is $\beta$- strongly convex w.r.t. $\|.\|$ and $f^*(0) = 0$, then, denoting the partial sum $\sum_{j \leq i} v_j$ by $v_{1:i}$, we have, for any sequence $v_1, \ldots v_n$ and for any $u$, $\sum_{i=1}^{n} \langle v_i, u \rangle \leq f^*(v_{1:n}) \leq \sum_{i=1}^{n} \langle \nabla f^*(v_{1:i-1}), v_i \rangle + \frac{1}{2} \beta \sum_{i=1}^{n} \|v_i\|_*^2$

# Algorithm

- Online Mirror Descent

1: Initialization: $w_1 \leftarrow \nabla f^*(0)$
2: for $t = 1$ to $T$
3: Play $w_t \in A$
4: Receive $g_t$ and pick $v_t \in \partial g_t(v_t)$
5: Update $w_{t+1} \leftarrow \nabla f^*(-\eta \sum_{s=1}^{t} v_t)$
6: end for

**Theorem**

*Suppose Algorithm is used with a function f that is $\beta$-strongly convex w.r.t. a norm $\|.\|$ on A and has $f^*(0) = 0$. Suppose the loss functions $g_t$ are convex and V-Lipschitz w.r.t. the norm $\|.\|$. Then for $\eta > 0$*

$$\sum_{t=1}^{T} g_t(w_t) - \min_{u \in A} \sum_{t=1}^{T} g_t(u) \leq \frac{\max_{u \in A} f(u)}{\eta} + \frac{\eta V^2 T}{2\beta}$$

# Algorithm

- Online Mirror Descent

1: Initialization: $w_1 \leftarrow \nabla f^*(0)$
2: for $t = 1$ to $T$
3: Play $w_t \in A$
4: Receive $g_t$ and pick $v_t \in \partial g_t(v_t)$
5: Update $w_{t+1} \leftarrow \nabla f^*(-\eta \sum_{s=1}^{t} v_t)$
6: end for

## Theorem

*Suppose Algorithm is used with a function f that is $\beta$-strongly convex w.r.t. a norm $\|.\|$ on A and has $f^*(0) = 0$. Suppose the loss functions $g_t$ are convex and V -Lipschitz w.r.t. the norm $\|.\|$. Then for $\eta > 0$*

$$\sum_{t=1}^{T} g_t(w_t) - \min_{u \in A} \sum_{t=1}^{T} g_t(u) \leq \frac{\max_{u \in A} f(u)}{\eta} + \frac{\eta V^2 T}{2\beta}$$

# Surrogate Loss and Perceptron

- In case of online classification the $0-1$ is not convex. No sublinear regret bound algorithm possible - Cover Result.
- Surrogate loss function : $l_{hinge}(w,(x,y)) = [1 - y\langle w, x\rangle]_+$ upperbounds $l_{0-1}(w,(x,y))$
- $\sum_{t=1}^{T} l_{0-1}(w_t,(x_t,y_t)) \leq \sum_{t=1}^{T} l_{hinge}(u,(x_t,y_t)) + Regret(T)$
- Apply Algorithm with $f(w) = \frac{1}{2}\|w\|_2^2$ to obtain perceptron algorithm.

# References

1. "Prediction, Learning, and Games" - Nicolo Cesa-Bianchi, Gabor Lugosi, Cambridge University Press.
2. http://www.cs.huji.ac.il/ shais/AdvancedML.html.

Questions ??