# Online learning

Note Title 2/12/2013

There are $n$ "experts"

Repeat
- Algorithm picks an expert
- Each expert recieves a payoff.

$\text{Payoff}(\text{expert } i, \text{ round } t) = V_{i,t} \in [0,1]$

Maximize total payoff.

Differences from online matching:

- Pick before $V_{i,t}$'s are revealed
- Only local constraint.

What is OPT?

- Say optimal choice on hindsight. What does OPT do?

Answer! OPT picks best expert in each round

- What is the worst scenario for ALG?
  - Deterministic & 2 experts. ALG $\to$ 1.
  - Answer!- $V_1 = 0$, $V_2 = 1$.
  - Randomized: ALG $\to$ 1 w/ $p$, 2 w/ $q$. say $p \geq q$
  - Ans:- again $v_1 = 0$ $v_2 = 1$, $E[ALG] = q \leq \frac{1}{2}$
  - Randomized & $n$ experts. ALG $\to p_1, p_2, \ldots p_n$

Ans:- $i = \arg\min_j \{p_j\}$ $V_i = 1$, $V_j = 0 \ \forall j \neq i$.

$E[ALG] \leq \frac{1}{n}$. $OPT = 1$. Can repeat this.

OPT is too powerful!

Redefine. $OPT = \max_i \left\{ \sum_{t=1}^{T} v_{i,t} \right\}$

optimal single expert in hindsight.

OPT can't change its choice for each round

Let $w_i = \sum_{t=1}^{T} v_{i,t}$. $OPT = \max_i \{ w_i \}$

Smooth approx. to max

Proof based on properties of $\phi$

$\phi(w_1, \ldots w_n) := \frac{1}{\lambda} \log \sum_i e^{\lambda w_i}$

for some $\lambda > 0$. $\vec{w} = (w_1, \ldots w_n)$

Lemma 1:- $\phi(\vec{w}) \geq OPT$

Proof. $\phi \geq \frac{1}{\lambda} \log \max_i e^{\lambda w_i} = \frac{1}{\lambda} \log e^{\lambda OPT}$

$= \cancel{\frac{1}{\lambda}} \cancel{\lambda} \cdot OPT = OPT$.

Lem 2:- $\phi \leq OPT + \frac{1}{\lambda} \log n$.

$\sum_i e^{\lambda w_i} \leq \sum_i e^{\lambda OPT} = n e^{\lambda OPT}$.

$\therefore \log \left( \sum_i e^{\lambda w_i} \right) \leq \log n + \lambda OPT$

$\therefore \phi \leq \frac{1}{\lambda} \log n + OPT$

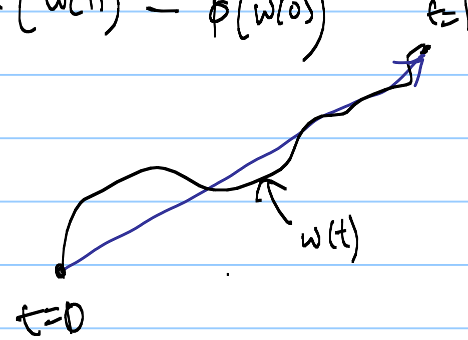
Lem 3:- $\frac{\partial \phi}{\partial w_i} = \cancel{\frac{1}{\lambda}} \frac{1}{\sum_i e^{\lambda w_i}} \cdot \cancel{\lambda} e^{\lambda w_i}$

Note:- $\sum_i \frac{\partial \phi}{\partial w_i} = \frac{\sum_i e^{\lambda w_i}}{\sum_i e^{\lambda w_i}} = 1.$

**Def:** $\vec{\nabla}\phi = \left(\frac{\partial \phi}{\partial w_1}, \frac{\partial \phi}{\partial w_2}, \dots \frac{\partial \phi}{\partial w_n}\right)$.

**Thm 1:** $\forall$ "path" in $\mathbb{R}^n$ given by $w:[0,1]$

$$\int_{t=0}^{1} \left\langle \vec{\nabla}\phi(w(t)), \frac{d\vec{w}}{dt} \right\rangle dt = \phi(w(1)) - \phi(w(0))$$



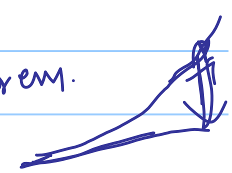

**Proof:** By chain rule, $\frac{d\phi}{dt} = \left\langle \vec{\nabla}\phi(w(t)), \frac{d\vec{w}}{dt} \right\rangle$

$$\int_{t=0}^{1} \frac{d\phi}{dt}\, dt = \phi(w(1)) - \phi(w(0))$$

**Theorem 2:** $\langle \vec{\nabla}\phi(w_0), w_1 - w_0 \rangle \geq \phi(w_1) - \phi(w_0) - \lambda \|w_1 - w_0\|_\infty$

If $w_1 \geq w_0$, co-ordinate-wise.

Discrete Analog of previous theorem.



Assume Theorem 2, define Algorithm.

$\vec{w}_0 = (0, 0, \dots, 0)$

In round $t = 1..T$

- pick expert $i$ w.p. $\nabla_i \phi(\vec{w}_{t-1})$
- $\vec{w}_t = \vec{w}_{t-1} + \vec{v}_t$

$$E[ALG] = \sum_{t=1}^{T} \langle \nabla\phi(\vec{w}_{t-1}), \vec{v}_t \rangle$$

Theorem 3:- $E[ALG] \geq OPT - \frac{1}{\lambda}\log n - \lambda T$

$$\text{Regret} = OPT - E[ALG] \leq \frac{1}{\lambda}\log n + \lambda T$$

Proof:- $E[ALG] = \sum_{t=1}^{T} \langle \nabla\phi(\vec{w_{t-1}}), \vec{v_t} \rangle$

From Thm 2, $\geq \sum_{t=1}^{T} \phi(\vec{w_t}) - \phi(\vec{w_{t-1}}) - \lambda \|\vec{v_t}\|_\infty$

$$\geq \phi(\vec{w}) - \phi(\vec{0}) - \lambda T$$

$$\geq OPT - \frac{1}{\lambda}\log n - \lambda T.$$

$\because \phi(\vec{0}) = \frac{1}{\lambda}\log \sum_i e^0 = \frac{1}{\lambda}\log n$

---

Pick $\lambda$ to minimize regret.

$$\text{Regret} = \frac{1}{\lambda}\log n + \lambda T.$$

Set $\frac{1}{\lambda}\log n = \lambda T$ i.e. $\lambda^2 = \frac{\log n}{T}$

$$\lambda = \sqrt{\frac{\log n}{T}}$$

$$\text{Regret} = 2\sqrt{T \log n}$$

---

Proof of Theorem 2:- Say $\|w_1 - w_0\|_\infty \leq 1$.

wnt $\phi(w_1) - \phi(w_0) - \langle \nabla\phi(w_0), w_1 - w_0 \rangle \leq \lambda$

$$\Leftrightarrow \frac{1}{\lambda}\log \sum_i e^{\lambda w_{1i}} - \frac{1}{\lambda}\log \sum_i e^{\lambda w_{0i}}$$

$$\leq \lambda + \frac{\sum_i e^{\lambda w_{0i}} \cdot \Delta w_i}{\sum_i e^{\lambda w_{0i}}}$$

(⇐) $$\log \frac{\sum_i e^{\lambda w_i}}{\sum_i e^{\lambda w_{0i}}} \le \lambda^2 + \frac{\sum_i e^{\lambda w_{0i}} \cdot \lambda \Delta w_i}{\sum_i e^{\lambda w_{0i}}}$$

Let $\frac{e^{\lambda w_{0i}}}{\sum_i e^{\lambda w_{0i}}} =: \mu_i$, $\lambda \Delta w_i =: \lambda_i$

(⇔) $$\log\left(\sum_i \mu_i e^{\lambda_i}\right) \le \lambda^2 + \sum_i \mu_i \lambda_i$$

(⇔) $$\sum_i \mu_i e^{\lambda_i} \le e^{\lambda^2} \cdot e^{\sum_i \mu_i \lambda_i}$$

$\left(\sum_i \mu_i = 1\right)$

$\lambda_i \le \lambda$

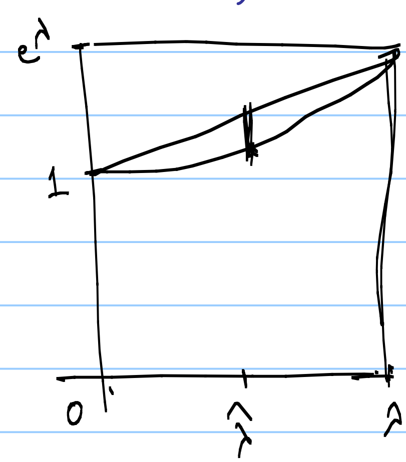
Convexity of $e^x$ ⇒

$$\sum_i \mu_i e^{\lambda_i} \ge e^{\sum_i \mu_i \lambda_i}$$

convexity of $e^x \le e^{\lambda^2}$ in $[0,\lambda]$

Claim: $\forall \ \hat{\lambda} \in [0,\lambda]$,

$$1 + \frac{e^\lambda - 1}{\lambda} \cdot \hat{\lambda} \le e^{\lambda^2} \cdot e^{\hat{\lambda}}$$




Continue proof.

By convexity of $e^x$,

$$e^{\lambda_i} \le 1 + \frac{e^\lambda - 1}{\lambda} \cdot \lambda_i$$

$$\therefore \sum_i \mu_i e^{\lambda_i} \le \sum_i \mu_i \left[1 + \frac{e^\lambda - 1}{\lambda} \cdot \lambda_i\right]$$

$$= 1 + \frac{e^\lambda - 1}{\lambda} \hat{\lambda} \quad \left[\hat{\lambda} = \sum_i \mu_i \lambda_i\right]$$

$$\le e^{\lambda^2} \cdot e^{\hat{\lambda}}$$

∎

**Proof of Claim:** If $\lambda \geq 1$, $e^{\lambda^2} \geq e^{\hat\lambda} \geq \frac{e^{\hat\lambda}-1}{\hat\lambda}$

Slope of $e^{\lambda^2} \cdot e^{\hat\lambda}$ is bigger.

Also at $\hat\lambda = 0$, is bigger.

If $\lambda < 1$, $e^{\hat\lambda} < 1 + \hat\lambda + \hat\lambda^2$.

$$\frac{e^{\hat\lambda}-1}{\hat\lambda} < 1+\hat\lambda$$

$$\therefore 1 + \hat\lambda\left(\frac{e^{\hat\lambda}-1}{\hat\lambda}\right) < 1+\hat\lambda+\lambda\hat\lambda < 1+\hat\lambda+\lambda^2$$

$$\leq e^{\lambda^2+\hat\lambda}$$

—— x ———————— x —— $\square$

## Applications:-

- Boosting in ML.

  Learn a function $f: X \to \{-1, 1\}$

  "Is this a cat?"

  Use a hypothesis class $\mathcal{H}$ of fn-s.

  $\mathcal{D}$ is a probability distribution on $X$.

**Strong learning algorithm:-** $\forall\ \mathcal{D}$, find $h \in \mathcal{H}$ s.t.

$$\Pr_{x \sim D}[h(x) = f(x)] \geq 1 - \varepsilon$$

**Weak** —— " —— —— " —— find $h \in \mathcal{H}$ s.t.

—— " —— $\geq \frac{1}{2} + \delta$

Boosting : Weak $\longrightarrow$ Strong.

fn $X$ = "training set". $\forall x \in X$, know $f(x)$

$D$ = uniform

Ada-Boost :-

- $X$ = Experts.
- Run MWU algo, Let $D_t$ be the distr$^n$ over $X$ used by MWU
- Use WL to find $h_t$ s.t $\Pr_{x \sim D_t}[h_t(x) = f(x)] \geq \frac{1}{2} + \delta$
- $\forall x \in X$ $V_{x,t} = \mathbb{1}\{h_t(x) \neq f(x)\}$

$x$ managed to fool $h_t \Rightarrow$ bigger weight on $x$.

Repeat $T$ times.

Strong-Learner :- $h(x) = \text{sign}\left\{\sum_{t=1}^{T} h_t(x)\right\}$

Majority of $h_t$'s.

$$OPT = \max_{x \in X}\left\{\sum_t V_{x,t}\right\} = \max_{x \in X}\left\{\sum_t \mathbb{1}\{h_t(x) \neq f(x)\}\right\}$$

$$E[ALG] = \sum_{t=1}^{T} \Pr_{x \in D_t}[V_{x,t}] = \sum_{t=1}^{T} \Pr_{x \sim D_t}[h_t(x) \neq f(x)]$$

$$\leq \sum_{t=1}^{T}\left(\frac{1}{2} - \delta\right) = \left(\frac{1}{2} - \delta\right)T$$

$$\Phi_T - ALh \leq \lambda T + \frac{1}{\lambda}\log n$$

$$\Phi_T \leq \left(\frac{1}{2} - \delta\right)T + \lambda T + \frac{1}{\lambda}\log n$$

Suppose $h(x) \neq f(x)$. for $\varepsilon n$ $x$'s.

Then for $\geq \frac{T}{2}$ $t$'s, $h_t(x) \neq f(x)$

i.e $\#\{x : w_x \geq T/2\} = \varepsilon n$

$$\therefore \Phi_T = \frac{1}{\lambda} \log\left(\Sigma_x e^{\lambda w_x}\right) \geq \frac{1}{\lambda} \log\left[\varepsilon n \cdot e^{\lambda \frac{T}{2}}\right]$$

$$= \frac{1}{\lambda} \log n + \frac{\log \varepsilon}{\lambda} + \frac{T}{2}$$

$$\therefore \log \varepsilon \leq (\lambda - \delta)\lambda T$$

or $\log \frac{1}{\varepsilon} \geq (\delta - \lambda)\lambda T$ $\qquad \lambda = \frac{\delta}{2}$

$$= \frac{\delta^2}{4} T$$

$$T = 4 \log{}^1/_\varepsilon / \delta^2$$

$\rbrack$