Today

Multiplicative Weight Update Method (part II)

CS5.205.1
Toolkit in TCS
- Lecture #16
(12 Apr '21)
Instructor: Prahladh Harsha

MWUM$_\varepsilon$ (Parameter: $\eta \in (0, 1/2]$)

① Initialize: $\forall i \in [n]$, $w_i^{(1)} \leftarrow 1$

② For $t \leftarrow 1$ to T

(a) Choose the probability distribution $\mathbf{p}^{(t)} = (p_1^{(t)}, \ldots, p_n^{(t)})$

where $p_i^{(t)} = \dfrac{w_i^{(t)}}{\Phi^{(t)}}$ & $\Phi^{(t)} = \sum_{i \in [n]} w_i^{(t)}$

(b) Observe the costs $M^{(t)} = (m_1^{(t)}, \ldots, m_n^{(t)}) \in [-1, 1]^n$

(c) Sample according to prob dist $p^{(t)}$

(d) Update the weights

$$w_i^{(t+1)} \leftarrow w_i^{(t)} \cdot (1 - m_i^{(t)} \cdot \varepsilon)$$

Today:

Expected Performance of $MWUM_\varepsilon$

$\ell(t) :=$ Expected loss at time step T

$$= \sum_{i=1}^{n} m_i^{(t)} \cdot p_i^{(t)} = \langle M^{(t)}, p^{(t)} \rangle$$

$L(T) :=$ Expected loss up to time T

$$= \sum_{t=1}^{T} \ell(t) = \sum_{t=1}^{T} \langle M^{(t)}, p^{(t)} \rangle.$$

Theorem: Assuming all costs $\in [-1, 1]$ & $\varepsilon \in (0, \frac{1}{2}]$, for all experts i

$$L(T) \leq \sum_{t=1}^{T} m_i^{(t)} + \varepsilon \sum_{t=1}^{T} |m_i^{(t)}| + \frac{\ln n}{\varepsilon}$$

Remarks ① If all $m_i^{(t)} \in [0, 1]$

$$L(T) \leq (1+\varepsilon) \sum_{t=1}^{T} m_i^{(t)} + \frac{\ln n}{\varepsilon}.$$

② If you knew beforehand the #steps can fix ε - learning rate appropriately

Proof: As in WM, will be using a potential function.

$$\Phi^{(t)} := \sum_{i \in [n]} w_i^{(t)}$$

$$\Phi^{(t+1)} = \sum_{i \in [n]} w_i^{(t+1)} = \sum_{i \in [n]} w_i^{(t)} \cdot \left(1 - \varepsilon m_i^{(t)}\right)$$

$$= \Phi^{(t)} - \varepsilon \, \Phi^{(t)} \sum_{i \in [n]} m_i^{(t)} \cdot p_i^{(t)}$$

$$= \Phi^{(t)} \left(1 - \varepsilon \langle M^{(t)}, p^{(t)} \rangle\right)$$

$$\leq \Phi^{(t)} \exp\left(-\varepsilon \langle M^{(t)}, p^{(t)} \rangle\right) \qquad 1 - \varepsilon x \leq e^{-\varepsilon x}$$

Hence,

$$\Phi^{(T+1)} \leq \Phi^{(1)} \exp\left(-\varepsilon \sum_{t=1}^{T} \langle M^{(t)}, p^{(t)} \rangle\right)$$

$$= n \cdot \exp\left(-\varepsilon \, L(T)\right)$$

For any expert i

$$\Phi^{(t+1)} \geq w_i^{(t+1)} = w_i^{(t)} \left(1 - \varepsilon m_i^{(t)}\right)$$

$$\Phi^{(T+1)} = w_i^{(1)} \prod_{t=1}^{T} \left(1 - \varepsilon m_i^{(t)}\right)$$

$$\geq 1 \prod_{t: m_i^{(t)} \geq 0} (1-\varepsilon)^{m_i^{(t)}} \prod_{t: m_i^{(t)} < 0} (1+\varepsilon)^{m_i^{(t)}}$$

$1 - \varepsilon x \geq (1-\varepsilon)^x$ if $x \in [0,1]$

$1 - \varepsilon x \geq (1+\varepsilon)^x$ if $x \in [-1,0]$

$$(1-\varepsilon)^{\sum_{\geq 0} m_i^{(t)}} (1+\varepsilon)^{\sum_{<0} m_i^{(t)}} \leq \Phi^{(T+1)} \leq n \cdot \exp(-\varepsilon L(T))$$

$$\ln n - \varepsilon L(T) \geq \sum_{t: \geq 0} m_i^{(t)} \ln(1-\varepsilon) + \sum m_i^{(t)} \ln(1+\varepsilon)$$

$$L(T) \leq \frac{\ln n}{\varepsilon} + \frac{1}{\varepsilon}\left(\ln\left(\frac{1}{1-\varepsilon}\right) \sum_{\geq 0} m_i^{(t)} - \ln(1+\varepsilon) \sum_{<0} m_i^{(t)} \right)$$

$$\leq \frac{\ln n}{\varepsilon} + \frac{1}{\varepsilon}\left((\varepsilon + \varepsilon^2) \sum_{\geq 0} m_i^{(t)} + (\varepsilon + \varepsilon^2) \sum_{<0} m_i^{(t)} \right)$$

$x \in [0, \frac{1}{2}]$

$\ln\left(\frac{1}{1-x}\right) \leq x + x^2$

$\ln(1+x) \geq x + x^2$

$$= \frac{\ln n}{\varepsilon} + \sum_t m_i^{(t)} + \varepsilon \sum_t |m_i^{(t)}|$$

$\boxtimes$

Theorem is true for every expert i (including the expert that makes the minimum loss)

Cor: Let P be any fixed distribution on the n experts.

$$L(T) \leq \sum_{t=1}^{T} \langle M^{(t)}, P \rangle + \varepsilon \sum_{t=1}^{T} \langle |M^{(t)}|, P \rangle + \frac{\ln n}{\varepsilon}.$$

Pf: Apply convex combination w.r.t P of previous theorem.

Hedge Algorithm

Very similar to the $MWUM_\varepsilon$

Weight Update Rule.

$MWUM_\varepsilon$: $w_i^{(t+1)} \leftarrow w_i^{(t)} (1 - \varepsilon \cdot m_i^{(t)})$

$Hedge_\varepsilon$: $w_i^{(t+1)} \leftarrow w_i^{(t)} \exp(-\varepsilon \cdot m_i^{(t)})$

Thm: (Hedge Alg). i - any expert

$$L(T) \leq \sum_{t=1}^{T} m_i^{(t)} + \varepsilon \sum_{t=1}^{T} \langle M^{(t)^2}, p^{(t)} \rangle + \frac{\ln n}{\varepsilon}.$$

$\left(\exp(-\varepsilon x) \leq 1 - \varepsilon x + \varepsilon^2 x^2\right)$
for the settings we have chosen

Alg: → Each step is producing a prob dist $p^{(T)}$ on experts.

→ Modifies $p^{(t)}$ based on experts loss at the previous step.

$\mathcal{P} = \{$ feasible probability dist on n experts $\}$

$\mathcal{P}$ = closed convex set of $[0,1]^n$

So, far $\mathcal{P}$ = all possible feasible prob distributions.

But there could be scenarios where all prob. dist not feasible.

And the alg must act only according to some $p \in \mathcal{P}$.

Qn: Does same MWUM$_\varepsilon$ work?

Ans: Possibly not, as $p^{(T)} \notin \mathcal{P}$.

We will "project" $p^{(T)}$ into $\mathcal{P}$

Use relative entropy (Kullback-Leibler divergence) to do this projection

P, Q - 2 distributions on n experts

$P = (p_1 \dots p_n)$; $Q = (q_1 \dots q_n)$

$$RE(P \| Q) = KL(P \| Q) = \sum_{i \in [n]} p_i \ln \frac{p_i}{q_i}$$

Bregman Projection:

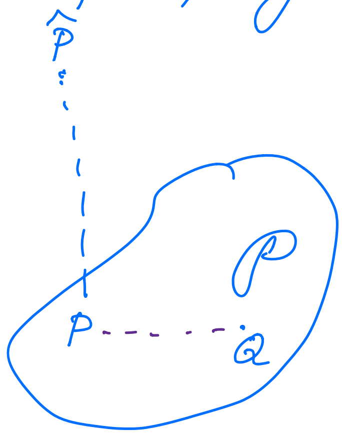
$\mathcal{P}$ - any closed convex set of prob dist.

$\hat{P}$: any prob dist ($\hat{P} \notin \mathcal{P}$)

$$Proj_{\mathcal{P}}(\hat{P}) = \operatorname*{argmin}_{P \in \mathcal{P}} RE(P \| \hat{P})$$

Convex function - can be obtained using convex optimization

Property of Projection:



$$P \stackrel{\Delta}{=} \mathrm{Proj}_{\mathcal{P}}(\hat{P})$$

Q - any prob dist in $\mathcal{P}$.

Generalized Pythagorean Inequality

$$RE(Q \| P) + RE(P \| \hat{P}) \leq RE(Q \| \hat{P})$$

Hence $RE(Q \| P) \leq RE(Q \| \hat{P})$

(because RE is non-negative)

$MWUM_\varepsilon$ (in terms of relative entropy).

① Initialize: $p^{(1)} \leftarrow$ arbitrary distribution in $\mathcal{P}$.

② For $t \leftarrow 1$ to T

(a) Observe costs $M^{(t)} = (m_1^{(t)}, \ldots, m_n^{(t)})$

(b) Sample accg to $p^{(t)}$ & act appropriately.

(c). Update $p^{(t)}$ to $p^{(t+1)}$ as follows

(i) $p^{(t)} \longrightarrow \tilde{p}^{(t+1)}$

$$\tilde{p}_i^{(t+1)} \leftarrow \frac{p_i^{(t)} (1 - \varepsilon \cdot m_i^{(t)})}{\Phi^{(t)}}$$

where $\Phi^{(t)}$ - normalization const to make $\tilde{p}^{(t+1)}$ into a dist

(New Step). (ii) $p^{(t+1)} = \mathrm{Proj}_{\mathcal{P}}(\tilde{p}^{(t+1)})$

Performance Analysis:

P - be any distribution in $\mathcal{P}$.

We will observe $RE(P \| p^{(t)})$ varies as the algorithm progresses

$$RE(p \| \hat{p}^{(t+1)}) - RE(p \| p^{(t)})$$

$$= \sum_{i \in [n]} p_i \ln \frac{p_i^{(t)}}{\hat{p}_i^{(t+1)}}$$

$$= \sum_{i \in [n]} p_i \ln \frac{\Phi^{(t)}}{1 - \varepsilon m_i^{(t)}}$$

$$\leq \sum_{i : m_i^{(t)} \geq 0} p_i \cdot m_i^{(t)} \ln \left(\frac{1}{1-\varepsilon} \right)$$

$1 - \varepsilon x \geq (1-\varepsilon)^x$ if $x \in [0,1]$

$1 - \varepsilon x \geq (1+\varepsilon)^x$ if $x \in [-1,0]$

$$+ \sum_{i : m_i^{(t)} < 0} p_i \cdot m_i^{(t)} \ln (1+\varepsilon)$$

$$+ \sum p_i \ln \Phi^{(t)}.$$

$$= \varepsilon \left(\langle M^{(t)}, p \rangle + \varepsilon \langle |M^{(t)}|, p \rangle \right)$$

$$+ \ln \Phi^{(t)}.$$

$x \in [0, \tfrac{1}{2}]$

$\ln\left(\frac{1}{1-x}\right) \leq x + x^2$

$\ln (1+x) \geq x + x^2$

$$\ln \Phi^{(t)} = \ln \left[\sum_{i \in [n]} p_i^{(t)} \left(1 - \varepsilon \cdot m_i^{(t)} \right) \right]$$

$$= \ln \left(1 - \varepsilon \sum p_i^{(t)} \cdot m_i^{(t)} \right)$$

$$= \ln \left(1 - \varepsilon \langle p^{(t)}, M^{(t)} \rangle \right)$$

$$\leq -\varepsilon \langle p^{(t)}, M^{(t)} \rangle$$

$$RE(p\,\|\,\hat{p}^{(t+1)}) - RE(p\,\|\,p^{(t)})$$
$$\leq \varepsilon\left(\langle M^{(t)}, p\rangle + \varepsilon\langle |M^{(t)}|, p\rangle\right)$$
$$- \varepsilon\langle M^{(t)}, p^{(t)}\rangle \quad \ldots (*)$$

$\langle M^{(t)}, p\rangle$ - loss of the fixed dist p

$\langle M^{(t)}, p^{(t)}\rangle$ - loss of alg at step t.

Since $p^{(t+1)} = \mathrm{Proj}_{\mathcal{P}}(\hat{p}^{(t+1)})$

$$RE(p\,\|\,p^{(t+1)}) \leq RE(p\,\|\,\hat{p}^{(t+1)})$$

$(*)$ implies

$$RE(p\,\|\,p^{(t+1)}) - RE(p\,\|\,p^{(t)})$$
$$\leq \varepsilon\left(\langle M^{(t)}, p\rangle + \varepsilon\langle |M^{(t)}|, p\rangle\right)$$
$$- \varepsilon\langle M^{(t)}, p^{(t)}\rangle$$

Rewriting & dividing by ε

$$\langle M^{(t)}, p^{(t)}\rangle \leq \langle M^{(t)}, p\rangle + \varepsilon\langle |M^{(t)}|, p\rangle$$
$$+ \frac{RE(p\,\|\,p^{(t)}) - RE(p\,\|\,p^{(t+1)})}{\varepsilon}.$$

Summing over $t \leftarrow 1$ to T

$$L(T) = \sum_{t=1}^{T} \langle M^{(t)}, p^{(t)} \rangle$$

$$\leq \sum_{t=1}^{T} \langle M^{(t)}, p \rangle + \varepsilon \sum_{t=1}^{T} \langle |M^{(t)}|, p \rangle + \frac{RE(p \| p^{(1)}) - RE(p \| p^{(T+1)})}{\varepsilon}.$$

Thm: $MWUM_\varepsilon$ (w.r.t relative entropy) has the following performance.

Let $p^{(1)}$ - any starting dist &

p - any fixed dist

$\Big\} \in \mathcal{P}.$

then

$$L(T) \leq \sum_{t=1}^{T} \langle M^{(t)}, p \rangle + \varepsilon \sum_{t=1}^{T} \langle |M^{(t)}|, p \rangle + \frac{RE(p \| p^{(1)})}{\varepsilon}.$$