An Intuitive Proof of MWU

April 10, 2023

- Recall the setting.
  - There are n experts and T days.
  - At each day $t = 1, 2, \dots, T$, the following happen *in order*:
    1. We choose a distribution $p^t = (p_1^t, \dots, p_n^t)$ of experts
    2. Then, the **adversary** reveals the **loss vector** $\ell^t = (\ell_1^t, \dots, \ell_n^t)$
    3. Our (expected) loss is $\ell_A^t = \langle p^t, \ell^t \rangle = \sum_i p_i^t \ell_i^t$.
- **Goal:** our total loss $L_A = \sum_{t=1}^T \ell_A^t$ is as good as the total loss of the best expert $L_\star = \min_i \sum_{t=1}^T \ell_i^t$.
- We want to analyze the MWU algorithm:

Algorithm 1 The Multiplicative Weights Update (MWU) algorithm

MWU(ϵ):

- $w_i^1 \leftarrow 1 \quad \forall i$
- For $t = 1, \dots, T$
  - Follows expert i with probability $p_i^t = \frac{w_i^t}{\sum_j w_j^t}$
  - After ℓ^t is revealed, $w_i^{t+1} \leftarrow w_i^t \cdot \exp(-\epsilon \ell_i^t) \quad \forall i$

Theorem 0.1. *For $0 < \epsilon \le 1/2$, MWU(ϵ) guarantees that*

$$L_A \le L_\star + \epsilon T + \frac{\ln n}{\epsilon}.$$

Theorem 0.2. *Suppose $\ell^t \in [0,1]^n$ for all t. For $0 < \epsilon \le 1/2$, MWU(ϵ) guarantees that*

$$L_A \le L_\star + \epsilon L_A + \frac{\ln n}{\epsilon}$$

so

$$L_A \le (1 + 2\epsilon)(L_\star + \frac{\ln n}{\epsilon})$$

- Most proofs in the literature[1] involved a lot of low-level calculation and the intuition is lost.

[1] https://lucatrevisan.github.io/40391/lecture11.pdf
http://www.theoryofcomputing.org/articles/v008a006/v008a006.pdf

- The proof here:
  - All low-level calculation is modularized into the soft-min function.
  - Using soft-min, high-level explanation is very intuitive
- The proof is inspired by the paper of Quanrud[2] and the blog of Zuzic[3] (they work in a more specific context of solving LPs).

1 Basic Calculus: Approximating Smooth Functions

- Let $f : \mathbb{R} \to \mathbb{R}$ be a **smooth** function.
- Suppose you know $f(x)$. Can we approximate $f(x+\delta)$? Use Taylor expansion

$$f(x+\delta) = f(x) + f'(x)\delta + \frac{1}{2}f''(x)\delta^2 + ...$$

  and, when δ is tiny, we have

$$f(x+\delta) \approx f(x) + f'(x)\delta$$

- What if $f : \mathbb{R}^n \to \mathbb{R}$?
  - The first derivative $f'(\cdot) \in \mathbb{R}^n$ is called the **gradient** of f.
  - The second derivative $f''(\cdot) \in \mathbb{R}^{n\times n}$ is called the **Hessian** of f.
- Suppose you know $f(\vec{x})$. Can we approximate $f(\vec{x}+\vec{\delta})$? Taylor expansion:

$$f(\vec{x}+\vec{\delta}) = f(\vec{x}) + \left\langle f'(\vec{x}), \vec{\delta} \right\rangle + \frac{1}{2}\left\langle \vec{\delta}, f''(\vec{x})\vec{\delta} \right\rangle + ...$$

  and, when $\|\vec{\delta}\|$ is tiny, we have

$$f(\vec{x}+\delta) \approx f(\vec{x}) + \left\langle f'(\vec{x}), \vec{\delta} \right\rangle$$

2 Overall Plan

- Let's forget the MWU algorithm you saw.
  - Suppose that we want to derive a very natural algorithm such that $L_A \le L_\star + \epsilon T + \frac{\ln n}{\epsilon}$.
  - The algorithm we derived will be MWU exactly.
- Notation:
  - $L^t = \sum_{t'=1}^{t} \ell^{t'} \in \mathbb{R}^n$ encode total loss of all experts after day t (after ℓ^t is revealed).
  - $L^t_\star = \min(L^t) := \min_i(L^t_i)$ be the total loss of the best expert after day t.
  - $L^t_A = \sum_{t'=1}^{t} \left\langle p^{t'}, \ell^{t'} \right\rangle$ is our total loss after day t.
- **Wishful hope:** whenever L^t_A increases, $L^t_\star$ increases by at least the same amount.

[2] https://epubs.siam.org/doi/abs/10.1137/1.9781611976014.11
[3] https://zuza.github.io/Intuitive-Multiplicative-Weights/

- So after T days, we have $L_A^T \le L_\star^T$.
- No regret at all. But this is clearly too good to hope for.

- Suppose, magically, there exists a function $\mathrm{smin}(L^t)$ where
  - $\mathrm{smin}(L^t) \approx \min(L^t) = L_\star^t$,
  - but $\mathrm{smin}(L^t)$ is smooth.

- **Actual plan:** whenever L_A^t increases, $\mathrm{smin}(L^t)$ increases by at least the same amount (modulo some small error).

- For each day,
  - The change in $\mathrm{smin}(L^t)$ (which is a proxy for $L_\star^t$) is

$$\approx \left\langle \mathrm{smin}'(L^{t-1}), \ell^t \right\rangle$$

    because $\mathrm{smin}(\cdot)$ is smooth.
  - By Taylor, we have

$$\mathrm{smin}(L^t) = \mathrm{smin}(L^{t-1} + \ell^t) \approx \mathrm{smin}(L^{t-1}) + \left\langle \mathrm{smin}'(L^{t-1}), \ell^t \right\rangle,$$

    * The approximation holds when ℓ^t is tiny compared to L^{t-1}.
    * We will not assume this assumption in formal proof (but there will be small error term).
  - the change in L_A^t is exactly $\left\langle p^t, \ell^t \right\rangle$ by definition:

$$L_A^t = L_A^{t-1} + \left\langle p^t, \ell^t \right\rangle$$

  - So... how should we set our distribution p^t on each day t? Clearly, we should set

$$p^t \leftarrow \mathrm{smin}'(L^{t-1})$$

    so that $\left\langle \mathrm{smin}'(L^{t-1}), \ell^t \right\rangle = \left\langle p^t, \ell^t \right\rangle$.

- So every time L_A^t increases, $\mathrm{smin}(L^t)$ increases by almost the same amount. As $L_\star^t \approx \mathrm{smin}(L^t)$, we should have

$$L_A^T \le L_\star^t + (\text{error from Taylor expansion}) + (\text{error from smin})$$

3 Soft-Max/Min: Smooth version of Max/Min

- Let $x = (x_1, \ldots, x_n)$ be a vector.
- The **soft-max** of x is $\mathrm{smax}(x) = \ln(\sum_i \exp(x_i))$.
- Why does it behave like max?
  - If $x_{i_0} \gg x_i$ for all i, $\exp(x_{i_0}) \approx \sum_i \exp(x_i)$. So $\ln(\sum_i \exp(x_i)) \approx x_{i_0}$.
  - So $\mathrm{smax}(x) \approx \max(x)$.
- The **soft-min** of x is $\mathrm{smin}(x) = -\mathrm{smax}(-x)$.

  - Note $\min(x) = -\max(-x)$.

- We will consider the *scaled* version: $\text{smin}_\epsilon(x) = -\frac{1}{\epsilon}\text{smax}(-\epsilon x)$.

Proposition 3.1. *We prove this at the end.*

- $\min(x) - \frac{\ln n}{\epsilon} \leq \text{smin}_\epsilon(x) \leq \min(x)$

- $\text{smin}'_\epsilon(x)_i = \frac{\exp(-\epsilon x_i)}{\sum_j \exp(-\epsilon x_j)}$

- *When* $\epsilon \leq 1/2$ *and* $\|\delta\|_\infty \leq 1$, $\text{smin}_\epsilon(x+\delta) \geq \text{smin}_\epsilon(x) + \langle \text{smin}'_\epsilon(x), \delta\rangle - \epsilon \frac{\sum_i \exp(-\epsilon x_i)\delta_i^2}{\sum_j \exp(-\epsilon x_j)}$

4 Deriving MWU exactly

- We use $\text{smin}_\epsilon(L^t)$ to approximate $L^t_\star$.

- Recall that we want to set p^t as the gradient of soft-min:

$$p^t \leftarrow \text{smin}'_\epsilon(L^{t-1})$$

  so

$$p^t_i = \frac{\exp(-\epsilon L_i^{t-1})}{\sum_j \exp(-\epsilon L_j^{t-1})}$$

- But what is this?
  - Recall the MWU algorithm: the weight of expert i is $w_i^{t+1} = w_i^t \cdot \exp(-\epsilon \ell_i^t)$.
  - So $w_i^t = \exp(-\epsilon L_i^{t-1})$.
  - Therefore,

$$p_i^t = \frac{w_i^t}{\sum_j w_j^t}.$$

- We just derived the MWU algorithm!

- Let's analyze it.

- We analyze how much L_A^t and $\text{smin}_\epsilon(L^t)$ change each day.
  - L_A^t increase exactly $\langle p^t, \ell^t\rangle$ by definition:

$$L_A^t - L_A^{t-1} = \langle p^t, \ell^t\rangle$$

  - How about $\text{smin}_\epsilon(L^t)$? Via Taylor's expansion, we have

$$\begin{aligned}
\text{smin}_\epsilon(L^t) - \text{smin}_\epsilon(L^{t-1}) &= \text{smin}_\epsilon(L^{t-1} + \ell^t) - \text{smin}_\epsilon(L^{t-1}) \\
&\geq \langle \text{smin}'_\epsilon(L^{t-1}), \ell^t\rangle - \epsilon \frac{\sum_i \exp(-\epsilon L_i^{t-1})(\ell_i^t)^2}{\sum_j \exp(-\epsilon L_j^{t-1})} \quad \text{as } \epsilon \leq 1/2, \|\ell^t\| \leq 1. \\
&\geq \langle p^t, \ell^t\rangle - \epsilon \langle p^t, (\ell^t)^2\rangle
\end{aligned}$$

    where $(\ell^t)^2 := (\ell_i^t)^2$ for all i.

  - That is, $\text{smin}_\epsilon(L^t)$ increases at least as much as L_A^t increases (except some error term $\epsilon \langle p^t, (\ell^t)^2 \rangle$).

- Summing over all t, the sum telescopes to

$$\begin{aligned}
\text{smin}_\epsilon(L^T) - \text{smin}_\epsilon(L^0) &\geq \sum_{t=1}^{T} \langle p^t, \ell^t \rangle - \epsilon \langle p^t, (\ell^t)^2 \rangle \\
&= L_A^T - \epsilon \sum_{t=1}^{T} \langle p^t, (\ell^t)^2 \rangle
\end{aligned}$$

- So

$$\begin{aligned}
L_A^T &\leq \text{smin}_\epsilon(L^T) - \text{smin}_\epsilon(L^0) + \epsilon \sum_{t=1}^{T} \langle p^t, (\ell^t)^2 \rangle \\
&\leq \min(L^T) + \frac{\ln n}{\epsilon} + \epsilon \sum_{t=1}^{T} \langle p^t, (\ell^t)^2 \rangle \\
&\leq L_\star^T + \frac{\ln n}{\epsilon} + \epsilon T
\end{aligned}$$

  because $\langle p^t, (\ell^t)^2 \rangle \leq 1$. This proves Theorem 0.1.

- If $\ell_i^t \in [0,1]$, then we have $(\ell_i^t)^2 \leq \ell_i^t$. So

$$\sum_{t=1}^{T} \langle p^t, (\ell^t)^2 \rangle \leq \sum_{t=1}^{T} \langle p^t, \ell^t \rangle = L_A^T$$

  and so we get

$$L_A^T \leq L_\star^T + \frac{\ln n}{\epsilon} + \epsilon L_A^T,$$

  which proves Theorem 0.2

5 Recap: How to Derive MWU

- How to design an algorithm for choosing experts with no regret property?
- **Wishful hope:** whenever L_A^t increases, $L_\star^t = \min(L^t)$ increases by at least the same amount.
  - But min is not smooth.
- **Actual plan:** whenever L_A^t increases, $\text{smin}(L^t) \approx \min(L^t)$ increases by at least the same amount (modulo some small error).
  - Since smin is smooth, we now can approximate the change of $\text{smin}(L^t)$ per day, which is just $\langle \text{smin}'_\epsilon(L^{t-1}), \ell^t \rangle$
  - Since the change of $L_A^t = \langle p^t, \ell^t \rangle$, we set $p^t = \text{smin}'_\epsilon(L^{t-1})$. So

$$p_i^t = \frac{\exp(-\epsilon L_i^{t-1})}{\sum_j \exp(-\epsilon L_j^{t-1})}.$$

- This is the same thing as saying that $p^t \sim w^t$ where $w^t = \exp(-\epsilon L_i^{t-1})$.
- To have $w^t = \exp(-\epsilon L_i^{t-1})$, we just need to multiplicatively update weights

$$w_i^{t+1} \leftarrow w_i^t \cdot \exp(-\epsilon \ell_i^t).$$

- This is the exactly what happens in MWU. Done.

6 Low-Level Calculation about Soft-max

- Let $x \in \mathbb{R}^n$. Define $\mathrm{sx}(x) = \sum_j \exp(x_j)$ for "sum-exp".

6.1 Upper/Lower Bounds

- $\mathrm{smin}_\epsilon(x) = \min(x)$ when x concentrates on only 1 entry.
- $\mathrm{smin}_\epsilon(x) = \min(x) - \frac{\ln n}{\epsilon}$ when x has uniform value on all entries.
- For other x, we have $\min(x) - \frac{\ln n}{\epsilon} \leq \mathrm{smin}_\epsilon(x) \leq \min(x)$ but I omit the proof.

6.2 Gradient

- We just compute

$$\begin{aligned}
(\mathrm{smin}'_\epsilon(x))_i &= \frac{\partial}{\partial x_i}\frac{-1}{\epsilon}(\ln \mathrm{sx}(-\epsilon x)) \\
&= \frac{-1}{\epsilon} \cdot \frac{1}{\mathrm{sx}(-\epsilon x)} \cdot \frac{\partial}{\partial x_i}\mathrm{sx}(-\epsilon x) \\
&= \frac{-1}{\epsilon} \cdot \frac{1}{\mathrm{sx}(-\epsilon x)} \cdot \exp(-\epsilon x_i) \cdot (-\epsilon) \\
&= \frac{1}{\mathrm{sx}(-\epsilon x)} \cdot \exp(-\epsilon x_i)
\end{aligned}$$

6.3 Smooth

- When $\epsilon \leq 1/2$ and $\|\delta\|_\infty \leq 1$, we have

$$\mathrm{smin}_\epsilon(x+\delta) \geq \mathrm{smin}_\epsilon(x) + \left\langle \mathrm{smin}'_\epsilon(x), \delta \right\rangle - \epsilon \frac{\sum_i \exp(-\epsilon x_i)\delta_i^2}{\mathrm{sx}(-\epsilon x)}.$$

- This is because

$$
\begin{aligned}
-\mathrm{smin}_\epsilon(x+\delta) &= \frac{1}{\epsilon}\ln\left(\sum_i \exp(-\epsilon x_i - \epsilon\delta_i)\right) \\
&= \frac{1}{\epsilon}\ln\left(\sum_i \exp(-\epsilon x_i)\exp(-\epsilon\delta_i)\right) \\
&\leq \frac{1}{\epsilon}\ln\left(\sum_i \exp(-\epsilon x_i)(1-\epsilon\delta_i + (\epsilon\delta_i)^2\right) && e^a \leq 1 + a + a^2 \forall |a| \leq 1/2. \\
&= \frac{1}{\epsilon}\ln\left(\sum_i \exp(-\epsilon x_i) - \epsilon\sum_i \exp(-\epsilon x_i)\delta_i + \epsilon^2\sum_i \exp(-\epsilon x_i)\delta_i^2\right) \\
&= \frac{1}{\epsilon}\ln\left(\mathrm{sx}(-\epsilon x)(1 - \epsilon\frac{\sum_i \exp(-\epsilon x_i)\delta_i}{\mathrm{sx}(-\epsilon x)} + \epsilon^2\frac{\sum_i \exp(-\epsilon x_i)\delta_i^2}{\mathrm{sx}(-\epsilon x)}\right) \\
&= -\mathrm{smin}_\epsilon(x) + \frac{1}{\epsilon}\ln\left(1 - \epsilon\left\langle \mathrm{smin}'_\epsilon(x), \delta\right\rangle + \epsilon^2\frac{\sum_i \exp(-\epsilon x_i)\delta_i^2}{\mathrm{sx}(-\epsilon x)}\right) \\
&\leq -\mathrm{smin}_\epsilon(x) + \frac{1}{\epsilon}\left(-\epsilon\left\langle \mathrm{smin}'_\epsilon(x), \delta\right\rangle + \epsilon^2\frac{\sum_i \exp(-\epsilon x_i)\delta_i^2}{\mathrm{sx}(-\epsilon x)}\right) && \ln(1+a) \leq a \\
&= -\mathrm{smax}_\epsilon(x) - \left\langle \mathrm{smax}'_\epsilon(x), \delta\right\rangle + \epsilon\frac{\sum_i \exp(-\epsilon x_i)\delta_i^2}{\mathrm{sx}(-\epsilon x)}
\end{aligned}
$$

Question 6.1. *Can you prove this?*

$$
\mathrm{smin}_\epsilon(x+\delta) \geq \mathrm{smin}_\epsilon(x) + \left\langle \mathrm{smin}'_\epsilon(x), \delta\right\rangle - \delta^\top \mathrm{smax}''_\epsilon(x)\delta.
$$

I don't know how to prove it.

- This is more beautiful, cleaner, and stronger than the bound we proved above.
- This is because, for any y

$$
y^\top \mathrm{smin}''_\epsilon(x) y > \epsilon\frac{\sum_i \exp(-\epsilon x_i) y_i^2}{\sum_j \exp(-\epsilon x_j)}.
$$