Qual

Yuhang Cai July 5, 2023

In this note, we consider the MAB with full information. We mainly apply the methods from (?).

An MAB is specified by a tuple $M = (A, \{l_k\}_{k=1}^K)$, where A is the set of arms and $l_k : A \to [0,1]$ is the reward function in episode $k \in [K]$. For our case we will consider the simplex measures.

$$\Delta M = \{\pi : A \to [0,1] | \sum_{a \in A} \pi(a) = 1\}.$$

And $\Delta(M, \alpha) = \{\pi | \pi \in \Delta(M), \pi(a) \geq \alpha, \forall a \in A\}$. The expected loss of any policy π at episode k can be written as

$$\sum_{a \in A} \pi(a) l_k(a) = \langle \pi, l_k \rangle.$$

Our dynamic regret is:

$$\text{Reg}_K(\pi_{1:K}^c) = \sum_{k=1}^K \langle \pi_k, l_k \rangle - \sum_{k=1}^K \langle \pi_k^c, l_k \rangle.$$

The O-REPS algorithm is:

$$\pi_{k+1} = \underset{\pi \in \Delta(M)}{\text{argmin}} \eta \langle \pi, l_k \rangle + D_\psi(\pi, \pi_k),$$

where $\eta > 0$ is step size, $\psi(\pi) = \sum_{a \in A} \pi(a) \log \pi(a)$ is the standard negative entropy. Our algorithm is:

Algorithm 1: DO-REPS for MAB

Input: step size pool $\mathcal{H}$, learning rate ε, clipping param α

1. Define $\psi(\pi) = \sum_{a \in A} \pi(a) \log \pi(a)$
2. Initialization: set $\pi_{1,i} = \arg\min_{\pi \in \Delta(M,\alpha)} \psi(\pi)$ and $p_{1,i} = 1/N, \forall i \in [N]$
3. **for** $k = 1$ *to* K **do**
4. Receive $\pi_{k,i}$ from base-learner $\mathcal{B}_i$ for $i \in [N]$
5. Compute policy $\pi_k = \sum_{i=1}^N p_{k,i} \pi_{k,i}$
6. Update the weight by $p_{k+1,i} \propto \exp\left(-\varepsilon \sum_{s=1}^k h_{s,i}\right)$ where $h_{k,i} = \langle \pi_{k,i}, \ell_k \rangle, \forall i \in [N]$
7. Each base-learner $\mathcal{B}_i$ updates prediction by $\pi_{k+1,i} = \arg\min_{\pi \in \Delta(M,\alpha)} \eta_i \langle \pi, \ell_k \rangle + D_\psi(\pi, \pi_{k,i})$

1 Main Results

Lemma 1. *Set $\pi_1 = \text{argmin}_{q \in \Delta(M,\alpha)} \psi(a)$. For any compared policies $\pi_1^c, \ldots, \pi_K^c \in \Delta(M, \alpha)$, O-REPS over a clipped space $\Delta(M, \alpha)$ ensures*

$$\sum_{k=1}^K \langle \pi_k - \pi_k^c, l_k \rangle \leq \eta K + \frac{1}{\eta}\left(\log |A| + P_K \log \frac{1}{\alpha}\right)$$

where $P_K = \sum_{k=2}^K \left\| \pi_k^c - \pi_{k-1}^c \right\|_1$ is the path-length of compared policies.

Theorem 2. *Set the clipping parameter* $\alpha = 1/K^2$, *the step size pool* $\mathcal{H} = \left\{\eta_i = 2^{i-1}\sqrt{K^{-1}\log|A|} \mid i \in [N]\right\}$, *where* $N = \left\lceil \frac{1}{2}\log\left(1 + \frac{4K\log K}{\log(|A|)}\right)\right\rceil + 1$, *and the learning rate of meta-algorithm as* $\varepsilon = \sqrt{(\log N)/K}$. *DO-REPS (Algorithm 1) satisfies*

$$\mathbb{E}\left[\mathrm{REG}_K\left(\pi^c_{1:K}\right)\right] \leq \mathcal{O}\left(\sqrt{K\left(\log|A| + P_K \log K\right)}\right)$$

where $P_K = \sum_{k=2}^{K} \|\pi^c_k - \pi^c_{k-1}\|_1$ *is the path-length of the compared policies.*

Remark. This result is much sharper than the result in (**?**). Their rate is $\tilde{O}(\sqrt{|A|K(S+1)})$, where $P_K \leq (S+1)$. However, they assume the MAB has bandit feedback.

2 Proof of Lemma 1

Proof of Lemma 1. Let $\pi'_{k+1} = \operatorname{argmin} \eta\langle \pi, l_k\rangle + D_\psi(\pi, \pi_k)$. Then, $\pi'_{k+1} = \pi_k(a)exp(-\eta l_k(a))$. Then,

$$\begin{aligned}
\sum_{k=1}^{K}\langle \pi_k - \pi_k^c, l_k\rangle &= \sum_{k=1}^{K}\langle \pi_k - \pi'_{k+1}, l_k\rangle + \langle \pi'_{k+1} - \pi_k^c\rangle \\
&\leq \sum_{k=1}^{K}\langle \pi_k - \pi'_{k+1}, l_k\rangle + \frac{1}{\eta}\sum_{k=1}^{K}\left(D_\psi(\pi_k^c, \pi_k) - D_\psi(\pi_k^c, \pi'_{k+1})\right) \\
&\leq \sum_{k=1}^{K}\langle \pi_k - \pi'_{k+1}, l_k\rangle + \frac{1}{\eta}\sum_{k=1}^{K}\left(D_\psi(\pi_k^c, \pi_k) - D_\psi(\pi_k^c, \pi_{k+1})\right),
\end{aligned}$$

where the first inequality holds due to Lemma 3 and the last one holds due to Pythagoras theorem.

For the first term, we know $1 - e^{-x} \leq x$ and we have

$$\sum_{k=1}^{K}\left\langle \pi_k - \pi'_{k+1}, l_k\right\rangle \leq \eta\sum_{k=1}^{K}\sum_{a\in A}\pi_k(a)l_k^2(a) \leq \eta\sum_{k=1}^{K}\sum_{a\in A}\pi_k(a) \leq \eta K.$$

For the lat term, we obtain:

$$\begin{aligned}
&\sum_{k=1}^{K}\left(D_\psi\left(\pi_k^c, \pi_k\right) - D_\psi\left(\pi_k^c, \pi_{k+1}\right)\right) \\
=&D_\psi\left(\pi_1^c, \pi_1\right) + \sum_{k=2}^{K}\left(D_\psi\left(\pi_k^c, \pi_k\right) - D_\psi\left(\pi_{k-1}^c, \pi_k\right)\right) \\
=&D_\psi\left(\pi_1^c, \pi_1\right) + \sum_{k=2}^{K}\sum_{a\in A}\left(\pi_k^c(a)\log\frac{\pi_k^c(a)}{\pi_k(a)} - \pi_{k-1}^c(a)\log\frac{\pi_{k-1}^c(a)}{\pi_k(a)}\right) \\
=&D_\psi\left(\pi_1^c, \pi_1\right) + \sum_{k=2}^{K}\sum_{x,a}\left(\pi_k^c(a) - \pi_{k-1}^c(a)\right)\log\frac{1}{\pi_k(a)} + \psi\left(\pi_K^c\right) - \psi\left(\pi_1^c\right) \\
\leq&\sum_{k=2}^{K}\left\|\pi_k^c - \pi_{k-1}^c\right\|_1\log\frac{1}{\alpha} + D_\psi\left(\pi_1^c, \pi_1\right) + \psi\left(\pi_K^c\right) - \psi\left(\pi_1^c\right).
\end{aligned}$$

Since π_1 minimize ψ, we have $\langle\nabla\psi(\pi_1), \pi_1^c - \pi_1\rangle \geq 0$. Thus,

$$D_\psi(\pi_1^c, \pi_1) + \psi(\pi_K^c) - \psi(\pi_1^c) \leq \psi(\pi_K^c) - \psi(\pi_1) \leq \sum_{a\in A}\pi_1(a)\log\frac{1}{\pi_1(a)} \leq \log|A|.$$

Combine them, we obtain:

$$\sum_{k=1}^{K}\langle \pi_k - \pi_k^c, l_k\rangle \leq \eta K + \frac{1}{\eta}\left(\log|A| + P_K\log\frac{1}{\alpha}\right),$$

where $P_K = \sum_{k=2}^{K}\left\|\pi_k^c - \pi_{k-1}^c\right\|_1$. □

3 Proof of Theorem 2

Proof of Thm 2. Let $\pi^u(a) = \frac{1}{|A|}, \forall a \in A$. We choose large K s.t. $\pi^u \in \Delta(M, \frac{1}{K})$. Then, we define, $u_k = (1 - \frac{1}{T})\pi_k^c + \frac{1}{T}\pi^u \in \Delta(M, \frac{1}{K^2})$. Then,

$$\begin{aligned}
\sum_{k=1}^{K} \langle \pi_k - \pi_k^c, \ell_k \rangle &= \sum_{k=1}^{K} \langle \pi_k - u_k, \ell_k \rangle + \frac{1}{K} \sum_{k=1}^{K} \langle \pi^u - \pi_k^c, \ell_k \rangle \\
&\leq \sum_{k=1}^{K} \langle \pi_k - u_k, \ell_k \rangle + 2 \\
&\leq \underbrace{\sum_{k=1}^{K} \langle \pi_k - \pi_{k,i}, \ell_k \rangle}_{\text{meta-regret}} + \underbrace{\sum_{k=1}^{K} \langle \pi_{k,i} - u_k, \ell_k \rangle}_{\text{base-regret}} + 2, \qquad (1)
\end{aligned}$$

where the last inequality holds for any index i.

Upper bound of base-regret. Since $u_k \in \Delta(M, \frac{1}{K^2}), \forall k \in [K]$. From Lemma 1, we have

$$\text{base-regret} \ \leq \eta K + \frac{\log |A| + 2\sum_{k=2}^{K} \|u_k - u_{k-1}\|_1 \log K}{\eta} \leq \eta K + \frac{\log |A| + 2P_K \log K}{\eta},$$

where $\sum_{k=2}^{K} \|u_k - u_{k-1}\|_1 \leq \sum_{k=2}^{K} \left\|\pi_k^c - \pi_{k-1}^c\right\|_1 = P_K$. The optimal step size is

$$\eta^* = \sqrt{\frac{\log |A| + 2P_K \log K}{K}}.$$

Note that $0 \leq P_K \leq 2K$, the possible range of the optimal step size is

$$\eta_{\min} = \sqrt{\frac{H \log |A|}{K}}, \text{ and } \eta_{\max} = \sqrt{\frac{H \log |A|}{K} + 4 \log K}.$$

Due to the construction of $\mathcal{H} == \left\{\eta_i = 2^{i-1}\sqrt{K^{-1} \log |A|} \mid i \in [N]\right\}$, where $N = \left\lceil \frac{1}{2} \log \left(1 + \frac{4K \log K}{\log(|A|)}\right)\right\rceil + 1$, We know

$$\eta_1 = \sqrt{\frac{\log |A|}{K}} = \eta_{\min}, \text{ and } \eta_N \geq \sqrt{\frac{\log |A|}{K} + 4 \log K} = \eta_{\max}.$$

Thus, there exists a base-learner i^* s.t $\eta_{i^*} \leq \eta^* \leq \eta_{i^*+1} = 2\eta_{i^*}$. Hence, we know

$$\begin{aligned}
\text{base-regret} \ &\leq \eta_{i^*} K + \frac{\log |A| + 2\bar{P}_K \log K}{\eta_{i^*}} \\
&\leq \eta^* K + \frac{2\left(\log(|A|) + 2P_K \log K\right)}{\eta^*} \\
&= 3\sqrt{K\left(\log |A| + 2P_K \log K\right)}, \qquad (2)
\end{aligned}$$

where the second inequality hold due to $\eta_{i^*} \leq \eta^* \leq \eta_{i^*+1} = 2\eta_i$.

Upper bound of meta-regret. Since $\pi_k = \sum \pi_{k,i} p_{k,i}$, we have

$$\text{meta-regret } = \sum_{k=1}^{K} \langle \pi_k - \pi_{k,i}, l_k \rangle = \sum_{k=1}^{K} \left\langle \sum_{i=1}^{N} p_{k,i}\pi_{k,i} - \pi_{k,i}, l_k \right\rangle = \sum_{k=1}^{K} \langle p_k - e_i, h_k \rangle,$$

where $h_{k,i} = \langle \pi_{k,i}, l_k \rangle$. It is known that the update $p_{k+1,i} \propto \exp\left(-\varepsilon \sum_{s=1}^{k} h_{s,i}\right), \forall i \in [N]$ is equal to the update $p_{k+1} = \arg\min_{p \in \Delta_N} \varepsilon \langle p, h_k \rangle + D_\psi(p, p_k)$, where $\psi(p) = \sum_{i=1}^{N} p_i \log p_i$ is the standard negative entropy. This can be further reformulated solving the unconstrained optimization problem $p'_{k+1} = \arg\min_p \varepsilon \langle p, h_k \rangle + D_\psi(p, p_k)$ at first and then projecting p'_{k+1} to the simplex Δ_N as $p_{k+1} = \arg\min_{p \in \Delta_N} D_\psi\left(p, p'_{k+1}\right)$. By standard analysis of OMD, we have

$$\begin{aligned}
\sum_{k=1}^{K} \langle p_k - e_i, h_k \rangle &\le \sum_{k=1}^{K} \langle p_k - p'_{k+1}, h_k \rangle + \sum_{k=1}^{K} \langle p'_{k+1} - e_i, h_k \rangle \\
&\le \sum_{k=1}^{K} \langle p_k - p'_{k+1}, h_k \rangle + \frac{1}{\varepsilon} \sum_{k=1}^{K} \left(D_\psi(e_i, p_k) - D_\psi\left(e_i, p'_{k+1}\right)\right) \\
&\le \sum_{k=1}^{K} \langle p_k - p'_{k+1}, h_k \rangle + \frac{1}{\varepsilon} \sum_{k=1}^{K} \left(D_\psi(e_i, p_k) - D_\psi(e_i, p_{k+1})\right) \\
&\le \sum_{k=1}^{K} \langle p_k - p'_{k+1}, h_k \rangle + \frac{1}{\varepsilon} D_\psi(e_i, p_1),
\end{aligned}$$

where the second inequality holds due to Lemma 3 and the third inequality holds due to Pythagoras theorem. Using the fact that $1 - e^{-x} \le x$ and the definition that $p_{1,i} = 1/N, h_{k,i} \le 1, \forall k \in [K], i \in [N]$, we have

$$\sum_{k=1}^{K} \langle p_k - p'_{k+1}, h_k \rangle + \frac{1}{\varepsilon} D_\psi(e_i, p_1) \le \varepsilon \sum_{k=1}^{K} \sum_{i=1}^{N} p_{k,i} h_{k,i}^2 + \frac{\ln N}{\varepsilon} \le \varepsilon K + \frac{\ln N}{\varepsilon}.$$

In particular, for the best base-learner $i^* \in [N]$, we have

$$\text{meta-regret } = \sum_{k=1}^{K} \langle \pi_k - \pi_{k,i^*}, \ell_k \rangle \le \varepsilon K + \frac{\log N}{\varepsilon} = \sqrt{K \log N}, \tag{3}$$

where the last equality holds due to the setting $\varepsilon = \sqrt{(\log N)/K}$.

Combine (1), (2) and (3), we obtain

$$\begin{aligned}
\sum_{k=1}^{K} \langle \pi_k - \pi_k^c, \ell_k \rangle &\le \text{ base-regret } + \text{ meta-regret} \\
&\le 3\sqrt{K\left(\log|A| + 2P_K \log K\right)} + \sqrt{K \log N} + 2 \\
&\le \mathcal{O}\left(\sqrt{K\left(\log(|A|) + P_K \log K\right)}\right),
\end{aligned}$$

where we used $N = \left\lceil \frac{1}{2} \log\left(1 + \frac{4K \log K}{\log(|A|)}\right) \right\rceil + 1$. □

4 Useful Lemmas

Lemma 3. *Define* $q^* = \text{argmin}_{q \in \mathcal{K}} \eta \langle q, l \rangle + D_F(q, \hat{q})$ *for some compact set* $\mathcal{K} \subset \mathbb{R}^d$*, convex and differentiable function* F*, an arbitrary point* $l \in \mathbb{R}^d$*, and a point* $\hat{q} \in \mathcal{K}$*. Then for any* $u \in \mathcal{K}$*,*

$$\langle q^* - u, l \rangle \leq \frac{1}{\eta}\big(D_F(u, \hat{q}) - D_F(u, q^*) - D_F(q^*, \hat{q})\big).$$

Proof. Since q^* is the minimal point, we know

$$\langle u - q^*, \eta l + \nabla F(q^*) - \nabla F(\hat{q}) \rangle \geq 0.$$

Hence,

$$\begin{aligned}
\langle q^* - u, l \rangle &\leq \frac{1}{\eta}\langle u - q^*, \nabla F(q^*) - \nabla F(\hat{q}) \rangle \\
&= \frac{1}{\eta}\langle u - q^*, \nabla F(q^*) \rangle - \frac{1}{\eta}\langle u - \hat{q} + \hat{q} - q^*, \nabla F(\hat{q}) \rangle \\
&= -\frac{1}{\eta}(F(u) - F(q^*) - \frac{1}{\eta}\langle u - q^*, \nabla F(q^*) \rangle) \\
&+ \frac{1}{\eta}(F(u) - F(\hat{q}) - \frac{1}{\eta}\langle u - \hat{q}, \nabla F(\hat{q}) \rangle) - \frac{1}{\eta}(F(q^*) - F(\hat{q}) - \langle q^* - \hat{q}, \nabla F(\hat{q}) \rangle) \\
&= \frac{1}{\eta}\big(D_F(u, \hat{q}) - D_F(u, q^*) - D_F(q^*, \hat{q})\big).
\end{aligned}$$

□

References