Supplementary Materials

6501-003 Reinforcement Learning (Spring 2025)

1 Inverse Gap Weighting for Multi-Armed bandits

Algorithm 1 Inverse Gap Weighting

Parameter: $\lambda > 0$.

for $t = 1, 2, \dots, T$ **do**

Let $\hat{R}_t(a) = \frac{\sum_{\tau<t} \mathbb{I}\{a_\tau = a\} r_\tau}{\sum_{\tau<t} \mathbb{I}\{a_\tau = a\}}$. // if $\sum_{\tau<t} \mathbb{I}\{a_\tau = a\} = 0$ then define $\hat{R}_t(a) = 0$

Let $b_t = \text{argmax}_{a\in\mathcal{A}} \hat{R}_t(a)$. // break ties arbitrarily

Define $\text{Gap}_t(a) = \hat{R}_t(b_t) - \hat{R}_t(a)$.

Sample a_t from distribution π_t, defined as

$$\pi_t(a) = \frac{1}{\gamma_t + \lambda \text{Gap}_t(a)}$$

where γ_t is a normalization factor that makes $\sum_{a\in\mathcal{A}} \pi_t(a) = 1$ (as discussed in the class, $\lambda_t \in [1, A]$).

Receive $r_t = R(a_t) + w_t$, where w_t is a zero-mean noise.

Theorem 1. *Inverse gap weighting (Algorithm 1) with parameter λ ensures*

$$\mathbb{E}[\textit{Regret}] \le O\left(\frac{AT}{\lambda} + \lambda \log^2 T + \sqrt{AT \log T}\right).$$

Proof. Define $N_t(a) = \sum_{\tau<t} \mathbb{I}\{a_\tau = a\}$ and $N_t^+(a) = \max\{N_t(a), 1\}$. By Hoeffding's inequality and a union bound over all $a \in \mathcal{A}$ and time t, we have

$$\left|\hat{R}_t(a) - R(a)\right| \le \sqrt{\frac{2\log(2AT/\delta)}{N_t^+(a)}} \tag{1}$$

for all $a \in \mathcal{A}$ and t with probability at least $1 - \delta$.

Suppose (1) holds. Consider the regret at round t:

$$\begin{aligned}
&R(a^\star) - R(a_t)\\
&= \left(\hat{R}_t(a^\star) - \hat{R}_t(a_t)\right) + \left(R(a^\star) - \hat{R}_t(a^\star)\right) + \left(\hat{R}_t(a_t) - R(a_t)\right)\\
&\le \left(\hat{R}_t(a^\star) - \mathbb{E}_{a\sim\pi_t}[\hat{R}_t(a)]\right) + \left(\mathbb{E}_{a\sim\pi_t}[\hat{R}_t(a)] - \hat{R}_t(a_t)\right) + \sqrt{\frac{2\log(2AT/\delta)}{N_t^+(a^\star)}} + \sqrt{\frac{2\log(2AT/\delta)}{N_t^+(a_t)}}.
\end{aligned} \tag{2}$$

We further bound the first term in (2).

$$\begin{aligned}
\hat{R}_t(a^\star) - \mathbb{E}_{a\sim\pi_t}[\hat{R}_t(a)] &= \mathbb{E}_{a\sim\pi_t}[\text{Gap}_t(a)] - \text{Gap}_t(a^\star) && \text{(by the definition of Gap}_t\text{)}\\
&= \sum_{a\in\mathcal{A}} \pi_t(a)\text{Gap}_t(a) - \text{Gap}_t(a^\star)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{a\in\mathcal{A}} \frac{\mathrm{Gap}_t(a)}{\gamma_t + \lambda \mathrm{Gap}_t(a)} - \mathrm{Gap}_t(a^\star) \\
&\le \sum_{a\in\mathcal{A}} \frac{\mathrm{Gap}_t(a)}{\lambda \mathrm{Gap}_t(a)} - \left(\frac{1}{\lambda\pi_t(a^\star)} - \frac{A}{\lambda}\right) \qquad \left(\tfrac{1}{\pi_t(a^\star)} = \gamma_t + \lambda\mathrm{Gap}_t(a^\star) \le A + \lambda \mathrm{Gap}_t(a^\star)\right) \\
&\le \frac{2A}{\lambda} - \frac{1}{\lambda\pi_t(a^\star)}. \qquad (3)
\end{aligned}
$$

Now, summing (2) over t and using (3), we get

$$
\begin{aligned}
\mathrm{Regret} &\le \sum_{t=1}^{T}\left(\frac{2A}{\lambda} - \frac{1}{\lambda\pi_t(a^\star)} + \left(\mathbb{E}_{a\sim\pi_t}[\hat{R}_t(a)] - \hat{R}_t(a_t)\right) + \sqrt{\frac{2\log(2AT/\delta)}{N_t^+(a^\star)}} + \sqrt{\frac{2\log(2AT/\delta)}{N_t^+(a_t)}}\right) \\
&= \frac{2AT}{\lambda} + \underbrace{\sum_{t=1}^{T}\left(-\frac{1}{\lambda\pi_t(a^\star)} + \sqrt{\frac{2\log(2AT/\delta)}{N_t^+(a^\star)}}\right)}_{\textbf{term}_1} + \underbrace{\sum_{t=1}^{T}\left(\mathbb{E}_{a\sim\pi_t}[\hat{R}_t(a)] - \hat{R}_t(a_t)\right)}_{\textbf{term}_2} + \underbrace{\sum_{t=1}^{T}\sqrt{\frac{2\log(2AT/\delta)}{N_t^+(a_t)}}}_{\textbf{term}_3}.
\end{aligned}
$$

Below we bound the expectation of the three terms above. First, notice that when (1) holds,

$$
\textbf{term}_1 \le \sum_{t=1}^{T} \frac{\lambda \log(2AT/\delta)\pi_t(a^\star)}{2N_t^+(a^\star)}. \qquad (\text{using } -u + \sqrt{2uv} \le \tfrac{v}{2} \text{ for } u, v > 0)
$$

Thus,

$$
\begin{aligned}
\mathbb{E}[\textbf{term}_1] &\le \mathbb{E}\left[\sum_{t=1}^{T} \frac{\lambda\log(2AT/\delta)\pi_t(a^\star)}{2N_t^+(a^\star)}\right] + \delta T \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \frac{\lambda\log(2AT/\delta)\mathbb{I}\{a_t = a^\star\}}{2N_t^+(a^\star)}\right] + \delta T \\
&= \lambda\log(2AT/\delta)\mathbb{E}\left[\frac{1}{1} + \frac{1}{1} + \frac{1}{2} + \frac{1}{3}\cdots + \frac{1}{N_T^+(a^\star)}\right] + \delta T \\
&\le O\left(\lambda \log(AT/\delta)\log T + \delta T\right).
\end{aligned}
$$

It is straightforward that

$$
\mathbb{E}[\textbf{term}_2] = 0.
$$

Finally,

$$
\begin{aligned}
\textbf{term}_3 &= \sqrt{2\log(2AT/\delta)} \sum_{a\in\mathcal{A}} \sum_{t=1}^{T} \sqrt{\frac{\mathbb{I}\{a_t = a\}}{N_t^+(a)}} \\
&= \sqrt{2\log(2AT/\delta)} \sum_{a\in\mathcal{A}} \left(\frac{1}{\sqrt{1}} + \frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{N_T^+(a)}}\right) \\
&= O\left(\sqrt{\log(AT/\delta)} \sum_{a\in\mathcal{A}} \sqrt{N_T^+(a)}\right) \\
&\le O\left(\sqrt{\log(AT/\delta)} \sqrt{A\left(\sum_{a\in\mathcal{A}} N_T^+(a)\right)}\right) \qquad (\text{by Cauchy-Schwarz inequality})
\end{aligned}
$$

$$= O\left(\sqrt{AT\log(AT/\delta)}\right).$$

Hence,

$$\mathbb{E}[\textbf{term}_3] = O\left(\sqrt{AT\log(AT/\delta)} + \delta T\right).$$

Choosing $\delta = \Theta(1/T)$ and using the assumption that $A \leq T$ (this is without loss of generality), we get

$$\mathbb{E}[\text{Regret}] \leq O\left(\frac{AT}{\lambda} + \lambda \log^2(T) + \sqrt{AT\log T}\right).$$

□