Learning with Abandonment Appendix

Sven Schmit [1] **Ramesh Johari** [2]

A. Proofs

A.1. Threshold models

Proof of Theorem 1. The proof follows from defining an appropriate dynamic program and finding the optimal policy using value iteration. We denote the state by x, denoting the best lower bound on c. In practice, if the process survives up to time t ($T > t$) the state is $x = \max_{s \le t} x_s$. Furthermore, it is convenient to use the demand, or survival, function $D(x) = 1 - F(x)$.

It is easy to see that the optimal policy is non-decreasing, so we can restrict our focus to non-decreasing policies.

The Bellman equation for the value function at state x is given by

$$V(x) = \max_{y \ge x} \frac{D(y)}{D(x)}(r(y) + \gamma V(y)). \tag{1}$$

For convenience we define the following transformation $J(x) = D(x)V(x)$ and note that we can equivalently use J to find the optimal policy. We now explicitly compute the limit of value iteration to find $J(x)$. Recall that $p(x) = r(x)(1-F(x)) = r(x)D(x)$. Start with $J_0(x) = 0$ for all x and note that the iteration takes the form

$$J_{k+1} = \max_{y \ge x} D(y)r(y) + \gamma J_k(y) = \max_{y \ge x} p(y) + \gamma J_k(y). \tag{2}$$

We prove the following two properties by induction for all $k > 0$:

1. $J_k(x) = p(x^*) \sum_{i=0}^{k-1} \gamma^i$ for all $x \le x^*$.
2. $J_k(x) < J_k(x^*)$ for all $x > x^*$.

The above is true for $k = 1$. Now assume it is true for an arbitrary k, then based on the induction assumption, it follows immediately that

$$J_{k+1}(x) = p(x^*) + \gamma J_k(x^*) \qquad \text{for all } x \le x^* \tag{3}$$

and therefore

$$J_{k+1}(x) = p(x^*) + \gamma p(x^*) \sum_{i=0}^{k-1} \gamma^i = p(x^*) \sum_{i=0}^{k} \gamma^i \tag{4}$$

as required. Furthermore,

$$J_{k+1}(x) < p(x^*) + \gamma J_k(x^*) = J_{k+1}(x^*) \qquad \text{for all } x > x^*. \tag{5}$$

The result follows from taking the limit as $k \to \infty$ and noting that for any state $x \le x^*$, it is optimal to jump to state x^* (and remain there). We also immediately see that the value of the optimal policy thus is $p(x)/\gamma$, as required. □

Proof of Proposition 2. It is immediate that the optimal policy must be constant; if the process survives $x_t = x$, then at time $t+1$ we face the same problem as at time t. So whatever action is optimal at time t, is also optimal at time $t + 1$. Let $V(x)$ denote the value of playing $x_t = x$ for all t. Then the following relation holds

$$V(x) = (1 - F(x))(r(x) + \gamma V(x)) \tag{6}$$

which leads to

$$V(x) = \frac{r(x)(1 - F(x))}{1 - \gamma(1 - F(x))}. \tag{7}$$

The result now follows immediately. □

A.2. Robustness

Proof of Proposition 3. First we consider the constant policy $x_t = x^* - y$ for all t in the noiseless case. We note that

$$r(x^* - y)D(x^* - y) \ge (r(x^*) - yL)D(x^*) \ge V(x^*) - yL \tag{8}$$

where $V(x^*)$ is the value of the optimal constant policy for the noise-free model.

Now let us consider the best possible noise model, then $\varepsilon_t = y$ for all t. But this is equivalent to the noise-free model with the threshold shifted by y. Hence, we know that a constant policy is optimal. We can bound the value

[1]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA [2]Management Science & Engineering, Stanford University, Stanford, CA, USA.. Correspondence to: Sven Schmit <schmit@stanford.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

of this model by

$$\max_x r(x)D(x-y) = \max_x r(x+y)D(x) \tag{9}$$
$$\leq \max_x (r(x)+yL)D(x) \tag{10}$$
$$= \max_x r(x)D(x) + yLD(x) \tag{11}$$
$$\leq \max_x r(x)D(x) + yL \tag{12}$$
$$= V(x^*) + yL \tag{13}$$

Hence, this implies that the constant policy $x_t = x^* - y$ is at most $\frac{2yL}{1-\gamma}$ worse than the optimal policy for the most optimistic noise model. □

Proof of Proposition 4. Let $\bar{\theta}$ be the midpoint of the η cover, $c = \frac{l+u}{2}$. Now we bound the expected value of an oracle policy, i.e. a policy that knows the true threshold θ^* as follows

$$\begin{aligned}
\mathbb{E}(v(\theta^*,\theta^*)) &\leq \frac{2\eta B}{1-\gamma} + \int_l^u v(\theta^*,\theta^*)dF_\theta \\
&\leq \frac{2\eta B}{1-\gamma} + \int_l^u v(\theta^*,\theta^*) + L|\bar{\theta}-\theta^*|dF_\theta \\
&\leq \frac{2\eta B}{1-\gamma} + \int_l^u v(\bar{\theta},\theta^*) + L\frac{u-l}{2}dF_\theta \\
&\leq \mathbb{E}(v(\bar{\theta},\theta^*)) + \frac{2\eta B}{1-\gamma} + (1-\eta)\frac{Lw}{2}
\end{aligned}$$

which completes the proof. □

A.3. Learning

Proof of Theorem 6. Due to the discretization, the proof consists of two parts. First, we show that the policy that plays the best arm i^* suffers small regret with respect to the optimal policy. Then we use the UCB regret bound to show that the learning strategy has low regret with respect to the playing arm i^*. Thus we can decompose regret into

$$\text{regret(UCB)} = \text{regret}_D + \text{regret}_U \tag{14}$$

where the first term corresponds to the discretization error and the second from the learning policy. Due to the time horizon and discounting, we write

Let x^* be the optimal strategy, i.e. it maximizes $r(x)D(x)$. Then the discretization error from playing i^*/K, by Assumption 1 is

$$\text{regret}_D \leq \frac{c_2 n}{2K^2} = \frac{c_2\sqrt{n\log n}}{2}. \tag{15}$$

Thus, the error due to the discretization is small.

Now let us bound the UCB regret with respect to action i^*/K. As Kleinberg and Leighton (2003) note, the assumption that the pulls of different arms are independent is not used in the proof. Thus we can apply Lemma 5. First, we show that the arms are sub-Gaussian. Since the rewards are bounded by 1 and independent across time, straightforward calculation shows that

$$\text{Var}\left((1-\gamma)\sum_{t=0}^{\infty}\gamma^t R_t(x)\right) = \frac{(1-\gamma)^2}{4(1-\gamma^2)} \leq \frac{1}{4}. \tag{16}$$

Then using the law of total variance, conditioning on the event $x < \theta_u$, the variance of the total obtained reward for user u, R_u, can be bounded by

$$\text{Var}(R_u) = \mathbb{E}(\text{Var}(R_u \mid \theta_u)) + \text{Var}(\mathbb{E}(R_u \mid \theta_u)) \tag{17}$$
$$= \frac{(1-F(x_u))M^2}{4} + (r(x_u))^2 F(x_u)(1-F(x_u)) \tag{18}$$
$$\leq M^2/2 \tag{19}$$

Thus we find that the reward for users is sub-Gaussian with parameter $\sigma = \frac{M^2}{2}$.

Recall the UCB regret bound

$$\text{regret(UCB)} \leq \sum_{i:\Delta_i>0}\frac{8\alpha\sigma^2}{\Delta_i}\log n + \frac{\alpha}{\alpha-2}. \tag{20}$$

We now focus on the $\sum_{i=1:\Delta_i>0}^{K}\frac{1}{\Delta_i}$ term. Let $\Delta_{(1)} \leq \Delta_{(2)} \leq \ldots \leq \Delta_{(K-1)}$ denote the ordered gaps with respect to the optimal arm. Note that for $j \geq 2$, we know $\Delta_{(j)} > c_1(\frac{j}{2K})^2$ due to Assumption 1. However, for the smallest gap, we only know $0 \leq \Delta_{(1)} \leq \frac{c_2}{K^2}$, depending how close i^*/K is to x^*. We thus obtain

$$\sum_{i=1}^{K}\frac{1}{\Delta_i} = \sum_{i=1}^{K-1}\frac{1}{\Delta_{(i)}} \tag{21}$$
$$= \frac{1}{\Delta_{(1)}} + \sum_{i\geq 2}\frac{1}{\Delta_{(j)}} \tag{22}$$
$$\leq \frac{1}{\Delta_{(1)}} + \frac{4K^2}{c_1}\sum j^{-1} \tag{23}$$
$$\leq \frac{1}{\Delta_{(1)}} + \frac{2\pi^2}{3c_1}K^2 \tag{24}$$

Thus regret is bounded by

$$\text{regret}_U \leq \frac{8\alpha\sigma^2\log n}{\Delta_{(1)}} + \frac{16\alpha\sigma^2\pi^2}{3c_1}(K-2)^2\log n + K\frac{\alpha}{\alpha-2} \tag{25}$$

However, the regret from due to playing the second best action is trivially bounded by $n\Delta_{(1)}$. Thus, we can bound the worst case when $\Delta_{(1)} = 4\sqrt{\log n/n}$. This leads to a bound of

$$\text{regret}_U \leq 2\alpha\sigma^2\sqrt{n\log n} + \frac{16\alpha\sigma^2\pi^2}{3c_1}(K-2)^2\log n + K\frac{\alpha}{\alpha-2} \tag{26}$$

since there are $K = (n/\log n)^{1/4}$ arms, we get

$$\text{regret}_u \leq 2\alpha\sigma^2\sqrt{n\log n} + \frac{16\alpha\sigma^2\pi^2}{3c_1}\sqrt{n\log n} + o(\sqrt{n\log n}) \tag{27}$$

Combining this with the bound on regret_D completes the proof. □

The regret bound for the KL-UCB algorithm is based on the following result by (Garivier & Cappé, 2011).

Lemma 1 (Theorem 2 in (Garivier & Cappé, 2011))**.** *Let $\varepsilon > 0$, and I^* denote the arm with maximal expected reward μ_{I^*}, and let I be any arm such that $\mu_I < \mu_{I^*}$. For any n, the number of times the KL-UCB algorithm chooses arm I is upper-bounded by*

$$\mathbb{E}[N_n(I)] \leq \frac{\log n}{\text{KL}(\mu_I \mid\mid \mu_{I^*})}(1+\varepsilon) + c_3\log\log n + \frac{c_4(\varepsilon)}{n^{\beta(\varepsilon)}}, \tag{28}$$

where $c_3 > 0$, and c_4 and β denote positive functions of ε.

Proof Proposition 8. The proof follows along the same lines as the proof for the UCB algorithm. For simplicity, we assume $M = 1$. For general M regret is increased by a factor M. As before, the discretization error is bounded by

$$\text{regret}_D \leq \frac{c_2 n}{2K^2} = \frac{c_2\sqrt{n\log n}}{2}. \tag{29}$$

We use Lemma 1 to bound the regret of the KL-UCB algorithm with respect to action i^*/K by noting

$$\text{regret}_U \leq \sum_{i=1}^{K-1} \Delta_{(i)}\mathbb{E}[N_n((i))] \tag{30}$$

By Pinsker's inequality $\text{KL}(\mu_I \mid\mid \mu_{I^*}) \geq 2(\mu_I - \mu_{I^*})^2$. Then we find that for $j \geq 2$,

$$\Delta_{(j)}\mathbb{E}[N_n((j))] \leq \frac{\log n}{2\Delta_{(j)}}(1+\varepsilon) + \Delta_{(j)}\left(c_3\log\log n + \frac{c_4(\varepsilon)}{n^{\beta(\varepsilon)}}\right) \tag{31}$$

Since $\Delta_{(j)} > c_1(\frac{j}{2K})^2$ due to Assumption 1, we obtain

$$\frac{\log n}{2\Delta_{(j)}}(1+\varepsilon) + \Delta_{(j)} \leq \frac{2(1+\varepsilon)\sqrt{n\log n}}{c_1 j^2} \tag{32}$$

and

$$\Delta_{(j)}\left(c_3\log\log n + \frac{c_4(\varepsilon)}{n^{\beta(\varepsilon)}}\right) \leq \frac{c_1 j^2\sqrt{\log n}}{4\sqrt{n}}\left(c_3\log\log n + \frac{c_4(\varepsilon)}{n^{\beta(\varepsilon)}}\right) \leq \frac{c_1}{4}\left(c_3\log\log n + \frac{c_4(\varepsilon)}{n^{\beta(\varepsilon)}}\right) \tag{33}$$

where the last inequality follows from $j \leq K = \sqrt{n/\log n}$. Thus we find

$$\text{regret}_U \leq \Delta_{(1)}\mathbb{E}[N_n((1))] + \frac{\pi^2(1+\varepsilon)\sqrt{n\log n}}{3c_1} + \frac{c_3}{4}\sqrt{n/\log n}\left(c_3\log\log n + \frac{c_4(\varepsilon)}{n^{\beta(\varepsilon)}}\right) \tag{34}$$

For arm (1) we note that the regret is trivially bounded by $n\Delta_{(1)}$, and this leads to worst case $\Delta_{(1)} = \sqrt{\log n/n}$. Setting $\varepsilon = 1$, we find

$$\text{regret}_U \leq \sqrt{n\log n}\left(1 + \frac{2\pi^2}{3c_1}\right) + \left(1 + \frac{c_3}{4}\sqrt{n/\log n}\right)\left(c_3\log\log n + \frac{c_4(1)}{n^{\beta(1)}}\right) \tag{35}$$

which completes the proof. □

A.4. Feedback

Proof of Lemma 9. The Bellman equation of the dynamic program for the feedback model can be written as:

$$V(l,u) = \max_{l\leq y\leq u} \frac{F(u)-F(y)}{F(u)-F(l)}(r(y) + \gamma V(y,u)) + \frac{F(y)-F(l)}{F(u)-F(l)}\gamma q V(l,y) \tag{36}$$

where l and u are the lower bounds and upper bounds on c based on the history.

Note that V is finite and therefore value iteration converges pointwise to V. We use induction on the value iterates to find the Lipschitz constant for V. Let $V_0, V_1, \ldots$ indicate the value iterates. Since $V_0(l,u) = 0$ for all states (l,u), the Lipschitz constant for V_0, denoted by $L_0 = 0$. We further claim that $L_{n+1} = L_p\frac{B}{1-\gamma} + q\gamma L_n$. Suppose this is true for $n = 1, \ldots, i-1$, then for $n = i+1$ we consider state $(l+\varepsilon, u)$ and write x^* for the optimal action in that state, and $y^* = x^* - l$. Then

$$V_{i+1}(l,u) \geq p(y^* \mid l,u)(r(x^*) + \gamma V(x^*,u)) + (1 - p(y^* \mid l,u))q\gamma V(l,x^*) \tag{37}$$

Also, $V(l,x^*) \leq V(l,u)$. Then we find

$$V_{i+1}(l+\varepsilon,u) - V_{i+1}(l,u) \leq [p(y^* \mid l+\varepsilon,u) - p(y^* \mid l,u)](r(x^*) + \gamma V_i(x^*,u)) + (1 - p(y^* \mid l+\varepsilon,u))q\gamma V_i(l+\varepsilon,x^*) - (1 - p(y^* \mid l,u))q\gamma V_i(l,x^*) \tag{38}$$

Using the Lipschitz continuity of p we can bound

$$p(y^* \mid l+\varepsilon,u) - p(y^* \mid l,u) \leq \varepsilon L_p. \tag{39}$$

Then note that

$$r(x^*) + \gamma V(x^*, u) \leq \frac{B}{1-\gamma} \tag{40}$$

and for the final two terms we note

$$\begin{aligned}(1 - p(y^* \mid l+\varepsilon, u))q\gamma V_i(l+\varepsilon, x^*) \\ - (1 - p(y^* \mid l, u)q\gamma V_i(l, x^*) \\ \leq q\gamma(V_i(l+\varepsilon, x^*) - V_i(l, x^*)) \leq q\gamma\varepsilon L_i \end{aligned} \tag{41}$$

where we use the inductive assumption. Because l, u and ε are arbitrary, we see that

$$L_n \leq \frac{L'B}{(1-q\gamma)(1-\gamma)}. \tag{42}$$

which implies V is Lipschitz. □

Proof of Proposition 10. First we note that by Lemma 8, V is Lipschitz, and we write L_v for its Lipschitz constant. Fix u, and consider a state $(u-\nu, u)$ for some $\nu > 0$. For notational convenience, for action x we write $y = x - (u - \nu)$ for the difference from the lower bound. We also use the shorthand $l = u - \nu$ and $p(y) = p(y \mid l, u)$. We can upper bound the value function by

$$V(l,u) = \max_y p(y)[r(x) + \gamma V(x,u)] + (1-p(y))q\gamma V(l,x) \tag{43}$$

$$\leq p(y)[r(l) + L_r y + \gamma V(l,u) \tag{44}$$

$$+ \gamma L_v y] + (1-p(y))q\gamma V(l,u) \tag{45}$$

$$\leq (1-\lambda(\nu)y)[r(l) + \gamma V(l,u) + Ly] \tag{46}$$

$$+ \lambda(\nu)yq\gamma V(l,u) \tag{47}$$

where we write $L = L_r + \gamma L_v$ and use the non-degeneracy of p. The derivative for the above expression with respect to y is

$$\begin{aligned}(1-2\lambda(\nu))Ly + L - \lambda(\nu)r(l) - \gamma\lambda(\nu)(1-q)V(l,u) \\ \leq (1-2\lambda(\nu))Ly + L - \lambda(\nu)r(l). \end{aligned} \tag{48}$$

Since $r(l) > 0$ for all $l \in Int\,\mathbf{X}$, for ν sufficiently small this derivative is negative for all $y \geq 0$. To complete the proof, we need this upper bound to be tight at $y = 0$, which follows immediately

$$\begin{aligned}(1-\lambda(\nu)y)[r(l) + \gamma V(l,u) + Ly] + \lambda(\nu)yq\gamma V(l,u)|_{y=0} = \\ r(l) + \gamma V(l,u) \geq \frac{r(l)}{1-\gamma}. \end{aligned} \tag{49}$$

Since r is increasing, it follows immediately that $\varepsilon(u)$ is non-decreasing in u. □

References

Garivier, A. and Cappé, O. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pp. 359–376, 2011.