Correction to "Satisficing in Multi-Armed Bandit Problems"

Paul Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard

***Abstract*—An unfortunate mistake in the proof of Theorem 8 of the above paper is corrected.**

We correct an error in the published proof of Theorem 8 of [2]. The error arises from an incorrect application of concentration inequalities. The correction follows the same structure as that published in [3, Appendix G], which corrects the proofs of performance bounds for UCL algorithms in [4] and thus Theorems 7 and 8 of [2]. For simplicity of presentation, we first state the correction and then provide the associated proof.

The heuristic value Q_i^t in [2, (27)] is

$$Q_i^t = \mu_i^t + \sigma_i^t \Phi^{-1}(1 - \alpha_t). \tag{C1}$$

To correct Theorem 8 of [2], set $\alpha_t = 1/(Kt^a)$ with $a > 4/(3(1 - \epsilon^2/16))$, $\epsilon \in (0, 4)$, and $K = \sqrt{2\pi e}$. The last part of the statement of [2, Theorem 8] should be replaced by

> "Then, the following statements hold for the satisfaction-in-mean-reward UCL algorithm with uncorrelated uninformative prior and $K = \sqrt{2\pi e}$:
>
> 1) the expected number of times a non-satisfying arm i is chosen until time T satisfies
>
> $$\mathbb{E}\left[n_i^T\right] \leq \left(\frac{8a}{\left(\Delta_i^{\mathcal{M}}\right)^2}\right) \log T + o(\log T);$$
>
> 2) the cumulative expected satisfaction-in-mean-reward regret until time T satisfies
>
> $$J_{SM} \leq \sum_{i=1}^{N} \Delta_i^{\mathcal{M}} \left(\frac{8a}{\left(\Delta_i^{\mathcal{M}}\right)^2}\right) \log T + o(\log T)."$$

For the δ-sufficing and $(\mathcal{M}, \delta)$-satisficing UCL algorithms of [2], similar corrections also hold with Q_i^t defined by (C1) and a modification to α_t. For these algorithms, the modification to α_t and its consequences can be succinctly stated by referring to the following Lemma which is a straightforward application of Theorem 2 below.

Lemma 1. *Let $\epsilon \in (0, 4)$ and define*

$$\alpha_t = 1 - \Phi\left(\sqrt{\frac{2}{1 - \epsilon^2/16} \log \frac{\log((1+\epsilon)t)}{\delta \log(1+\epsilon)}}\right). \tag{C2}$$

Then, at time t

$$\Pr\left[[2, (40)] \text{ holds}\right] = \Pr\left[\frac{\mu_i^t - m_i}{\sigma_i^t} \geq \Phi^{-1}(1 - \alpha_t)\right] \leq \delta.$$

The corrections to the four algorithms published in [2] and the corresponding corrected expressions for the performance bounds are summarized in Table I. For δ-Sufficing and $(\mathcal{M}, \delta)$-Satisficing UCL, the bounds take the form

$$f(\delta, \Delta) := \frac{8\sigma_s^2}{\Delta^2(1 - \epsilon^2/16)} \log \frac{\log((1+\epsilon)T)}{\delta \log(1+\epsilon)} + 1. \tag{C3}$$

TABLE I

SUMMARY OF THE CORRECTIONS FOR THE SATISFICING UCL ALGORITHMS. DEFINE Q_i^t BY (C1) WITH $\epsilon \in (0, 4)$, $K = \sqrt{2\pi e}$ AND SET α_t AS FOLLOWS. THE CORRECTED PERFORMANCE BOUNDS WITH f IS DEFINED BY (C3).

Algorithm	α_t	Bound
Deterministic UCL	$\alpha_t = 1/Kt^a$, $a > \frac{4}{3(1-\epsilon^2/16)}$	$\mathbb{E}\left[n_i^T\right] \leq \frac{8a\sigma_s^2}{\Delta_i^2} \log T + o(\log T)$
Satisfaction-in- -mean-reward UCL	$\alpha_t = 1/Kt^a$, $a > \frac{4}{3(1-\epsilon^2/16)}$	$\mathbb{E}\left[n_i^T\right] \leq \frac{8a}{(\Delta_i^{\mathcal{M}})^2} \log T + o(\log T)$
δ-Sufficing UCL	α_t from Equation (C2), $\delta \mapsto \delta/2$	$n_i^T \leq f(\delta/2, \Delta_i)$
$(\mathcal{M}, \delta)$-Satisficing UCL	α_t from Equation (C2), $\delta \mapsto \delta/3$	$n_i^T \leq f(\delta/3, \Delta_i^{\mathcal{M}})$

Note that with the correction, which accounts for the dependence of n_i^t on rewards accrued, the upper bound functional form (C3) is no longer independent of T. However, the dependence on T is of the form $\log \log T$, which is a very slowly increasing function of T. Therefore, in any realistic application the upper bound will effectively be constant and the qualitative result of [2] does not change.

REVISED PROOF

We employ the following concentration inequality from Garivier and Moulines [1] to fix the proof. Let $(X_t)_{t \geq 1}$ be a sequence of independent sub-Gaussian random variables with $\mathbb{E}[X_t] = \mu_t$, i. e., $\mathbb{E}[\exp(\lambda(X_t - \mu_t))] \leq \exp(\lambda^2\sigma^2/2)$ for some variance parameter $\sigma > 0$. Consider a previsible sequence $(\epsilon_t)_{t \geq 1}$ of Bernoulli variables, i.e., for all $t > 0, \epsilon_t$ is deterministically known given $\{X_\tau\}_{0 < \tau < t}$. Let

$$s^t = \sum_{s=1}^{t} X_s \epsilon_s, m^t = \sum_{s=1}^{t} \mu_s \epsilon_s, n^t = \sum_{s=1}^{t} \epsilon_s.$$

Theorem 2 ([1, Theorem 22], [3, Theorem 11])**.** *Let $(X_t)_{t \geq 1}$ be a sequence of sub-Gaussian[1] independent random variables with common variance parameter σ and let $(\epsilon_t)_{t \geq 1}$ be a previsible sequence of Bernoulli variables. Then, for all integers t and all $\delta, \epsilon > 0$,*

$$\Pr\left[\frac{s^t - m^t}{\sqrt{n^t}} > \delta\right] \leq \left\lceil \frac{\log t}{\log(1+\epsilon)} \right\rceil \exp\left(-\frac{\delta^2}{2\sigma^2}\left(1 - \frac{\epsilon^2}{16}\right)\right). \tag{C4}$$

We will also use the following lower bound for $\Phi^{-1}(1-\alpha)$, the quantile function of the normal distribution.

[1]The result in [1, Theorem 22] is stated for bounded rewards, but it extends immediately to sub-Gaussian rewards by noting that the upper bound on the moment generating function for a bounded random variable obtained using a Hoeffding inequality has the same functional form as the sub-Gaussian random variable.

Proposition 3. *For any $t \in \mathbb{N}$ and $a > 1$, the following holds:*

$$\Phi^{-1}\left(1 - \frac{1}{\sqrt{2\pi e t^a}}\right) \geq \sqrt{\nu \log t^a}, \tag{C5}$$

for any $0 < \nu \leq 1.59$.

Proof. We begin with the inequality $\Phi^{-1}(1-\alpha) > \sqrt{-\log(2\pi\alpha^2(1-\log(2\pi\alpha^2)))}$ established in [4]. It suffices to show that

$$-\log\left(\frac{1}{et^2}\left(1 - \log\left(\frac{1}{et^2}\right)\right)\right) - \nu \log t \geq 0,$$

for $\nu \in (0, 1.59]$. The left hand side of the above inequality is

$$g(t) := 1 - \log 2 + (2-\nu)\log t - \log(1 + \log t).$$

It can be verified that g admits a unique minimum at $t = e^{(\nu-1)/(2-\nu)}$ and the minimum value is $\nu - \log 2 + \log(2-\nu)$, which is positive for $0 < \nu \leq 1.59$. $\square$

In the following, we choose $\nu = 3/2$.

Correction to the proof of [2, Theorem 8]. The structure of the published proof carries through. Let i be a non-satisfying arm, i.e., $m_i < \mathcal{M}$, and recall that i^* denotes the arm with maximum mean reward. Let η be a positive integer and let $\epsilon \in (0, 4)$ and $a > 4/(3(1-\epsilon^2/16))$.

We first analyze the probability that [2, Eq. (31)] holds by applying Theorem 2. Let $\{X_\tau\}_{0<\tau<t}$ be the sequence of rewards associated with arm i, and let $(\epsilon_t)_{t\geq 1}$ equal 1 if the algorithm chooses arm i at time t. Note that, for an uncorrelated uninformative prior, $\mu_i^t = \bar{m}_i^t = s^t/n^t, \sigma_i^t = 1/\sqrt{n_i^t}, m_i = m^t/n^t$, and $n_i^t = n^t$. [2, Eq. (31)] is thus equivalent to

$$\frac{s^t}{n^t} - \frac{m^t}{n^t} \geq \frac{1}{\sqrt{n^t}}\Phi^{-1}(1-\alpha_t) \Rightarrow \frac{s^t - m^t}{\sqrt{n^t}} \geq \Phi^{-1}(1-\alpha_t).$$

Letting $\delta = \Phi^{-1}(1-\alpha_t)$ and applying (C4) yields

$$\begin{aligned}\Pr\left[\text{[2, Eq. (31)] holds}\right] &= \Pr\left[\frac{s^t - m^t}{\sqrt{n^t}} \geq \delta\right] \\ &\leq \left\lceil \frac{\log t}{\log(1+\epsilon)} \right\rceil \exp\left(-\frac{3\log t^a}{4}\left(1 - \frac{\epsilon^2}{16}\right)\right) \\ &= \left\lceil \frac{\log t}{\log(1+\epsilon)} \right\rceil t^{-\frac{3a(1-\epsilon^2)/16}{4}},\end{aligned}$$

where the second inequality follows from (C5). The same bound holds for [2, Eq. (32)].

It can be verified that for the corrected Q_i^t in equation (C1), the constant "8" in [2, Eqns. (35, 38 and 39)] will be replaced by $8a$. Following the proof in [2] with the above corrections,

$$\mathbb{E}\left[n_i^T\right] \leq \left\lceil \frac{8a}{\left(\Delta_i^{\mathcal{M}}\right)^2}\log T \right\rceil + \sum_{t=1}^{T} 3\left\lceil \frac{\log t}{\log(1+\epsilon)} \right\rceil t^{-\frac{3a(1-\epsilon^2)/16}{4}}.$$

The sum can be bounded by the integral

$$\int_1^T \left(\frac{\log t}{\log(1+\epsilon)} + 1\right) t^{-\frac{3a(1-\epsilon^2/16)}{4}} \mathrm{d}t + 1. \tag{C6}$$

It can be verified that the integral (C6) is of class $o(\log T)$ as long as the exponent $3a(1-\epsilon^2/16)/4 > 1$. Putting everything together, we have

$$\mathbb{E}\left[n_i^T\right] \leq \frac{8a\sigma_s^2}{\Delta_i^2}\log T + o(\log T).$$

The second statement follows from the definition of the cumulative expected regret. $\square$

The corrections to the proofs of [2, Theorem 10] (δ-Sufficing UCL) and [2, Theorem 11] ($(\mathcal{M}, \delta)$-Satisficing UCL) follow the same structure.

Correction to proof of [2, Theorem 10]. For the corrected α_t defined in equation (C2) with $\delta \mapsto \frac{\delta}{2}$, [2, Eq. (42)] is equivalent to

$$\Delta_i = m_{i^*} - m_i < 2C_i^t = \frac{2\sigma_s}{\sqrt{n_i^t}}\Phi^{-1}(1-\alpha_t).$$

Squaring, rearranging, and applying Equation (C2), we see that this never holds if

$$n_i^t > \frac{8\sigma_s^2}{\Delta_i^2(1-\epsilon^2/16)}\log\frac{2\log((1+\epsilon)t)}{\delta\log(1+\epsilon)} = \eta.$$

Then, Lemma 1 implies that [2, Eqns. (40, 41)] each hold with probability at most $\delta/2$. Therefore, for $n_i^t > \eta + 1 = f(\delta/2, \Delta_i)$, a non-satisfying arm is selected with probability at most δ. $\square$

Correction to proof of [2, Theorem 11]. For the corrected α_t defined in equation (C2) with $\delta \mapsto \frac{\delta}{2}$, an argument analogous to that for [2, Eq. (42)] above shows that [2, Eq. (44)] never holds for $n_i^t > \eta = f(\delta/3, \Delta_i^{\mathcal{M}}) - 1$.

Applying Lemma 1 implies that [2, Eq. (43)] holds with probability at most $\delta/3$. Similarly to the corrected proof for [2, Theorem 10] above, for $n_i^t > \eta + 1 = f(\delta/3, \Delta_i^{\mathcal{M}})$, $Q_i^t \geq Q_{i^*}^t$ with probability at most $\frac{2\delta}{3}$. Thus, a non-satisfying arm is selected with probability at most δ. $\square$

References

[1] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

[2] P. Reverdy, V. Srivastava, and N. E. Leonard. Satisficing in multi-armed bandit problems. *IEEE Transactions on Automatic Control*, 62(8):3788–3803, August 2017.

[3] P. Reverdy, V. Srivastava, and N. E. Leonard. Modeling human decision-making in generalized Gaussian multi-armed bandits. *arXiv preprint 1307.6134v5*, 2019. Available https://arxiv.org/abs/1307.6134v5.

[4] P. B. Reverdy, V. Srivastava, and N. E. Leonard. Modeling human decision making in generalized Gaussian multiarmed bandits. *Proceedings of the IEEE*, 102(4):544–571, 2014.