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 $Q_t^i$ in [2, (27)] is

$$Q_t^i = \mu_t^i + \sigma_t^i \Phi^{-1}(1 - \alpha_t). \quad (C1)$$

To correct Theorem 8 of [2], set $\alpha_t = 1/(Kt^\alpha)$ with $a = 4/(3(1 - \epsilon^2/16))$, $\epsilon \in (0, 4)$, and $K = \sqrt{2\pi \epsilon}$. 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 $K = \sqrt{2\pi \epsilon}$:

1) the expected number of times a non-satisfying arm $i$ is chosen until time $T$ satisfies

$$\mathbb{E}[n_t^i] \leq \left(\frac{8a}{(\Delta_t^M)^2}\right) \log T + o(\log T);$$

2) the cumulative expected satisfaction-in-mean-reward regret until time $T$ satisfies

$$J_{SM} \leq \sum_{i=1}^{N} \Delta_t^M \left(\frac{8a}{(\Delta_t^M)^2}\right) \log T + o(\log T).\)"

For the $\delta$-sufficing and $(\mathcal{M}, \delta)$-satisficing UCL algorithms of [2], similar corrections also hold with $Q_t^i$ defined by (C1) and a modification to $\alpha_t$. For these algorithms, the modification to $\alpha_t$ 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

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

Then, at time $t$

$$\Pr\{[2, (40)]\text{ holds}\} = \Pr\left[\frac{\mu_t^i - m_t^i}{\sigma_t^i} \geq \Phi^{-1}(1 - \alpha_t)\right] \leq \delta.$$ 

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

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

Note that with the correction, which accounts for the dependence on $n_t^i$ 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

$$s_t^i = \sum_{s=1}^{t} X_s \epsilon_s, m_t^i = \sum_{s=1}^{t} \mu_s \epsilon_s, n_t^i = \sum_{s=1}^{t} \epsilon_s.$$ 

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

$$\Pr\left[\frac{s_t^i - m_t^i}{\sqrt{n_t^i}} > \delta\right] \leq \left[\frac{\log t}{\log(1 + \epsilon)}\right] \exp\left(-\frac{\delta^2}{2\sigma^2} \left(1 - \frac{2}{1 - \epsilon^2/16}\right)\right). \quad (C4)$$

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$\delta$</th>
<th>Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic UCL</td>
<td>$\alpha_t = 1/Kt^\alpha$, $a &gt; \frac{1}{(1 - \epsilon^2/16)}$</td>
<td>$\mathbb{E}[m_t^i] \leq \frac{\sigma_t^i}{\Delta_t^M} \log T + o(\log T)$</td>
</tr>
<tr>
<td>Satisfaction-in-mean-reward UCL</td>
<td>$\alpha_t = 1/Kt^\alpha$, $a &gt; \frac{1}{(1 - \epsilon^2/16)}$</td>
<td>$\mathbb{E}[m_t^i] \leq \frac{\sigma_t^i}{\Delta_t^M} \log T + o(\log T)$</td>
</tr>
<tr>
<td>$\delta$-Sufficing UCL</td>
<td>$\alpha_t$ from Equation (C2), $\delta \to \delta/2$</td>
<td>$\mathbb{E}[m_t^i] \leq \frac{\sigma_t^i}{\Delta_t^M} \log T + o(\log T)$</td>
</tr>
<tr>
<td>$(\mathcal{M}, \delta)$-Satisficing UCL</td>
<td>$\alpha_t$ from Equation (C2), $\delta \to \delta/3$</td>
<td>$\mathbb{E}[m_t^i] \leq \frac{\sigma_t^i}{\Delta_t^M} \log T + o(\log T)$</td>
</tr>
</tbody>
</table>

1The 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:

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

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 \( 0 < \nu \leq 1.59 \).

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., \( m_i < M_i \), and recall that \( i^* \) denotes the arm with maximum mean reward. Let \( \eta \) 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 \( \{X_t\}_{t=0}^{\infty} \) be the sequence of rewards associated with arm \( i \) and let \( (\epsilon_i)_{t \geq 1} \) equal 1 if the algorithm chooses arm \( i \) at time \( t \).

Note that, for an uncorrelated uninformative prior, \( \mu_i = \bar{m}_i = s_i/n_i, \sigma_i = 1/\sqrt{n_i}, m_i = m/n_i \), and \( n_i = n \). [2, Eq. (31)] is thus equivalent to

\[
s_i - m_i \geq 1/\sqrt{n_i} \Phi^{-1}(1 - \alpha_i) \Rightarrow s_i - m_i \geq \sqrt{n} \Phi^{-1}(1 - \alpha_i).
\]

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

\[
Pr [\text{[2, Eq. (31)] holds}] = Pr\left[ \frac{s_i - m_i}{\sqrt{n_i}} \geq \delta \right] \\
\leq \left[ \frac{\log t}{\log(1+\epsilon)} \right] \exp \left( -3 \log ta^2 \left( 1 - \frac{\epsilon^2}{16} \right) \right) \\
= \left[ \frac{\log t}{\log(1+\epsilon)} \right] t^{-3a(1-\epsilon^2)/16},
\]

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

It can be verified that for the corrected \( Q'_i \) in equation (C1), the constant \( \pi \) in [2, Eqns. (35, 38 and 39)] will be replaced by \( 8a \).

Following the proof in [2] with the above corrections,

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

The sum can be bounded by the integral

\[
\int_{1}^{T} \left( \frac{\log t}{\log(1+\epsilon)} + 1 \right) t^{-3a(1-\epsilon^2)/16} dt + 1. \tag{C6}
\]

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

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

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

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

Correction to proof of [2, Theorem 10]. For the corrected \( \alpha_t \) defined in equation (C2) with \( \delta \Rightarrow \delta /2 \), [2, Eq. (42)] is equivalent to

\[
\Delta_i = m_i + m_i < 2C_i \Rightarrow \frac{2\sigma_i}{\sqrt{n_i}} \Phi^{-1}(1 - \alpha_t).
\]

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

\[
n_i^t > \frac{8a\sigma_i^2}{(\Delta_i^M)^2} \log 2 \left( 1 + \epsilon t \right) \log(1 + \epsilon) = \eta.
\]

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 \( \delta \).

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

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 \( n_i^t > \eta + 1 = f(\delta /3, \Delta_i^M) \), \( Q_i' \geq Q_i'' \), with probability at most \( \frac{25}{7} \). Thus, a non-satisfying arm is selected with probability at most \( \delta \).

REFERENCES


