Comments and Corrections

Corrections to "Satisficing in Multiarmed Bandit Problems"

Paul Reverdy, *Member, IEEE*, Vaibhav Srivastava, *Member, IEEE*, and Naomi Ehrich Leonard, *Fellow, IEEE*

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

We correct an error in the published proof of [2, Th. 8]. 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 in [2, Ths. 7 and 8]. 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 of [2, Th. 8], 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, Th. 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 nonsatisfying 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 above.

Manuscript received April 23, 2019; revised December 20, 2019; accepted February 29, 2020. Date of publication March 17, 2020; date of current version December 24, 2020.

Paul Reverdy is with the Department of Aerospace and Mechanical Engineering, University of Arizona, Tucson, AZ 85721 USA (e-mail: preverdy@arizona.edu).

Vaibhav Srivastava is with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824 USA (e-mail: vaibhav@egr.msu.edu).

Naomi Ehrich Leonard is with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: naomi@princeton.edu).

Digital Object Identifier 10.1109/TAC.2020.2981433

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	$\alpha_t = 1/Kt^a$, $a > \frac{4}{3(1-\epsilon^2/16)}$	$\mathbb{E}\left[n_i^T\right] \leq \frac{8a\sigma_i^2}{\Delta_i^2}\log T + o(\log T)$
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), $\delta \mapsto \delta/3$	$n_i^T \leq f(\delta/3, \Delta_i^{\mathcal{M}})$

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

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

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 $t > 0$, ϵ_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 (See [1, Th. 22] and [3, Th. 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.

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

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

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

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

It can be verified that g admits a unique minimum at $t = e^{(\nu-1)/(a(2-\nu))}$ 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, Th. 8]: The structure of the published proof carries through. Let i be a nonsatisfying 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)].

[1]The result in [1, Th. 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.

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

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

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. ■

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

Correction to proof of [2, Th. 10]: For the corrected α_t defined in (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 (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, eqs. (40, (41)] each hold with probability at most $\delta/2$. Therefore, for $n_i^t > \eta + 1 = f(\delta/2, \Delta_i)$, a nonsatisfying arm is selected with probability at most δ. ■

Correction to proof of [2, Th. 11]: For the corrected α_t defined in (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, Th. 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 nonsatisfying arm is selected with probability at most δ. ■

REFERENCES

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

[2] P. Reverdy, V. Srivastava, and N. E. Leonard, "Satisficing in multi-armed bandit problems," *IEEE Trans. Autom. Control*, vol. 62, no. 8, pp. 3788–3803, Aug. 2017.

[3] P. Reverdy, V. Srivastava, and N. E. Leonard, "Modeling human decision-making in generalized Gaussian multi-armed bandits," 2019, *arXiv:1307.6134v4*.

[4] P. B. Reverdy, V. Srivastava, and N. E. Leonard, "Modeling human decision making in generalized Gaussian multiarmed bandits," *Proc. IEEE*, vol. 102, no. 4, pp. 544–571, Apr. 2014.

Paul Reverdy (Member, IEEE) received the B.S. degree in engineering physics and the B.A. degree in applied mathematics from the University of California, Berkeley, CA, USA, in 2007, and the M.A. and Ph.D. degrees in mechanical and aerospace engineering from Princeton University, Princeton, NJ, USA, in 2011 and 2014, respectively.

He is currently an Assistant Professor of aerospace and mechanical engineering with the University of Arizona, Tucson, AZ, USA. From 2007 to 2009, he worked as a Research Assistant with the Federal Reserve Board of Governors, Washington, DC, USA. From 2014 to 2017, he was a Postdoctoral Fellow in electrical and systems engineering with the University of Pennsylvania, where he was affiliated with the GRASP laboratory. His research interests lie at the intersection of human and machine decision making and control, with applications in robotics, machine learning, and engineering design optimization.

Dr. Reverdy received a National Defense Science and Engineering Graduate Fellowship for graduate study and the Best Student Paper Award from the 2014 European Control Conference.

Vaibhav Srivastava (Member, IEEE) received the B.Tech. degree in mechanical engineering from the Indian Institute of Technology Bombay, Mumbai, India, in 2007, the M.S. degree in mechanical engineering, the M.A. degree in statistics, and the Ph.D. degree in mechanical engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2011 and 2012, respectively.

He is currently an Assistant Professor with the Electrical and Computer Engineering Department, Michigan State University, East Lansing, MI, USA. He served as a Lecturer and Associate Research Scholar with the Mechanical and Aerospace Engineering Department, Princeton University, Princeton, NJ, USA from 2013 to 2016. His research interests include modeling, analysis, and design of human cognition, shared autonomous systems, socio-cognitive networks, computational networks, and robotic search and surveillance missions.

Dr. Srivastava was the recipient of the Best Paper Award (as co-author) at the 2014 European Control Conference.

Naomi Ehrich Leonard (Fellow, IEEE) received the B.S.E. degree in mechanical engineering from Princeton University, Princeton, NJ, USA, in 1985 and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 1991 and 1994, respectively.

From 1985 to 1989, she worked as an Engineer in the electric power industry. She is currently the Edwin S. Wilsey Professor of mechanical and aerospace engineering and the Director of the Council on Science and Technology with Princeton University, Princeton, NJ, USA. She is also an Associated Faculty Member of Princeton University's Program in applied and computational mathematics and an affiliated Faculty Member of the Princeton Neuroscience Institute. Her research and teaching are in control and dynamical systems with current interests in coordinated control for multiagent systems, mobile sensor networks, adaptive ocean sampling, collective animal behavior, and human decision-making dynamics.