Correction regarding "Adaptive treatment assignment in experiments for policy choice"

Maximilian Kasy* Anja Sautmann†

November 10, 2021

In a comment posted on Arxiv on Sep 16, 2021, Ariu et al. (2021) point out some problems regarding the statement of item 3 of Theorem 1 in Kasy and Sautmann (2021) (KS hereafter). Ariu et al. (2021) show, based on Carpentier and Locatelli (2016), that a counter-example to this statement can be constructed. Their argument proves existence of a parameter vector such that the rate of convergence of (frequentist) expected policy regret to 0 is slower than that claimed by Theorem 1 of KS.[1]

In the present note, we first provide a corrected statement of our theorem. We next show that convergence of posterior beliefs, as in this corrected theorem, implies convergence of posterior expected policy regret. We then elaborate on Ariu et al. (2021) by discussing the issues with the original item 3 of Theorem 1 in KS.[2]

Theorem 1 (Corrected statement) *Consider exploration sampling in the setting of Section 2 in KS, with fixed wave size $N_t = N \geq 1$. Assume that the optimal policy $d^{(1)}$ is unique and that $\theta^{d^{(1)}} < 1$. As $T \to \infty$, the following holds:*

1. *The share of observations $\bar{q}_T^{d^{(1)}}$ assigned to the best treatment converges almost surely to $\rho^{d^{(1)}} \equiv \operatorname{plim}_{T\to\infty} \bar{q}_T^{d^{(1)}} = 1/2$.*
2. *The share of observations $\bar{q}_T^d$ assigned to each treatment $d \neq d^{(1)}$ converges almost surely to a non-random share ρ^d. $(\rho^1, \ldots, \rho^k)$ is such that $-\frac{1}{NT}\log(1 - p_T^{d^{(1)}}) \to^p \Gamma^*$ for some $\Gamma^* > 0$.*
3. *Under any adaptive allocation rule, $\limsup_{T\to\infty} -\frac{1}{NT}\log(1 - p_T^{d^{(1)}}) \leq \Gamma^*$ on any sample path with $\lim_{T\to\infty} \bar{q}_T^{d^{(1)}} = 1/2$.*

Item 3 is a restatement of item 2 in Theorem 1 of Russo (2020) for the Bernoulli model with uniform priors, which is proved in part 1 of Theorem 6, Supplementary Appendix I of Shang et al. (2020). The exploration sampling algorithm proposed in KS thus shares the desirable convergence properties of the top-two algorithms proposed in Russo (2016) (published as Russo 2020) and Qin et al. (2017). This includes (i) convergence of the sample shares $\bar{q}_T^d$ to the shares $\boldsymbol{\rho}_\gamma^d$, and (ii) (constrained) efficiency of the sample shares $\bar{q}_T^d$ for the convergence rate of the posterior probability $p_T^{d^{(1)}}$. Put differently, under no algorithm can we get a faster convergence of $p_T^{d^{(1)}}$, provided that the assignment share of the best arm $d^{(1)}$ converges to 1/2. Furthermore, as argued in KS, exploration sampling is well suited for the batched settings typically encountered in economic field experiments and in calibrated simulations shows large gains in terms of policy regret relative to both non-adaptive assignments and Thompson sampling. We next show that these gains in policy regret are expected, according to posterior beliefs.

*Department of Economics, Oxford University, maximilian.kasy@economics.ox.ac.uk
†World Bank, asautmann@worldbank.org

[1] This proof of existence requires that the number of treatment arms exceeds $\exp(800) \approx 10^{347}$. It is presumably possible to construct examples for a smaller number of treatment arms, however.

[2] We thank Daniel Russo and Chao Qin for helpful advice and discussions in preparing this correction.

Posterior expected policy regret Lemma 1 below shows that the convergence of posterior beliefs, as in Theorem 1 above, implies the convergence of posterior expected policy regret, $E[\Delta^{d_T^*}|\boldsymbol{m}_T, \boldsymbol{r}_T]$, with at least the same rate as $1 - p_T^{d^{(1)}}$.

In the proof of the lemma, we consider the policy with the highest posterior probability of being optimal, $\tilde{d}_T = \text{argmax}_{\,d}\, p_T^d$, as an intermediate object. This policy is in general different from $d_T^* = \text{argmax}_{\,d}\, E[\theta^d|\boldsymbol{m}_T, \boldsymbol{r}_T]$, which has the highest posterior expected welfare and therefore the lowest posterior expected regret.

Lemma 1 (Bound on posterior expected policy regret)

$$E[\Delta^{d_T^*}|\boldsymbol{m}_T, \boldsymbol{r}_T] \leq 1 - p_T^{d^{(1)}}.$$

Proof: Denote $\tilde{d}_T = \text{argmax}_{\,d}\, p_T^d$, and recall $d_T^* = \text{argmax}_{\,d}\, E[\theta^d|\boldsymbol{m}_T, \boldsymbol{r}_T]$. The latter definition implies

$$E[\Delta^{d_T^*}|\boldsymbol{m}_T, \boldsymbol{r}_T] = \min_d E[\max_{d'} \theta^{d'} - \theta^d|\boldsymbol{m}_T, \boldsymbol{r}_T] \leq E[\Delta^{\tilde{d}_T}|\boldsymbol{m}_T, \boldsymbol{r}_T].$$

Next, we can use the law of iterated expectations, and the fact that $\Delta^d = 0$ for $d = \text{argmax}_{\,d'}\, \theta^{d'}$ and $\Delta^d \leq 1$ for all other d, to decompose and bound as follows.

$$\begin{aligned}
E[\Delta^{\tilde{d}_T}|\boldsymbol{m}_T, \boldsymbol{r}_T] =& E[\Delta^{\tilde{d}_T}|\boldsymbol{m}_T, \boldsymbol{r}_T, \tilde{d}_T = \operatorname*{argmax}_{d'}\, \theta^{d'}] \cdot P(\tilde{d}_T = \operatorname*{argmax}_{d'}\, \theta^{d'}|\boldsymbol{m}_T, \boldsymbol{r}_T) \\
&+ E[\Delta^{\tilde{d}_T}|\boldsymbol{m}_T, \boldsymbol{r}_T, \tilde{d}_T \neq \operatorname*{argmax}_{d'}\, \theta^{d'}] \cdot P(\tilde{d}_T \neq \operatorname*{argmax}_{d'}\, \theta^{d'}|\boldsymbol{m}_T, \boldsymbol{r}_T) \\
\leq& P(\tilde{d}_T \neq \operatorname*{argmax}_{d'}\, \theta^{d'}|\boldsymbol{m}_T, \boldsymbol{r}_T) = 1 - p_T^{\tilde{d}_T} \leq 1 - p_T^{d^{(1)}}.
\end{aligned}$$

Throughout this proof, note that θ, Δ, and $\text{argmax}_{\,d'}\, \theta^{d'}$ are random objects with distributions determined by the posterior for θ, whereas $d^{(1)}$ in the last line of the equations above is the true optimal treatment (which is non-random). □

In interpreting this lemma, it is worth emphasizing that posterior expected policy regret, which averages over the posterior for θ, is distinct from (frequentist) expected policy regret, as defined in KS, which averages over the sampling distribution of d_T^* given the parameter θ.

Posterior expected policy regret is also distinct from hybrid objects such as $\sum_d \Delta^d \cdot p_T^d$, which is discussed in Section 6.1 of Russo (2020), as well as in the revised version of Ariu et al. (2021): Note that Δ^d is a frequentist object (defined by expectations over the sampling distribution given θ), while p_T^d is a Bayesian object (defined by the posterior distribution over θ given the observed sample).

Allowing for unbounded prior support: Updated citations for the proof Ariu et al. (2021) note that the Beta-Bernoulli model does not satisfy all the regularity conditions of Assumption 1 in Russo (2020). In particular Russo (2020) requires compact support for the prior distribution. While the uniform distribution for θ obviously has compact support, this is not the case for the implied distribution of the "natural parameter" of the Bernoulli distribution (in exponential family form), which equals $\eta^d = \log(\theta^d/(1-\theta^d))$ for each of the components d.

Fortunately, this point has been remedied by Shang et al. (2020). In particular, Theorem 6 in Appendix I of Shang et al. (2020) shows directly that claim 2 of Lemma 2 of KS extends to the case of Bernoulli outcomes and uniform priors, as considered in KS. Shang et al. (2020) also confirm the other results from Russo (2020) that were used in KS for the Bernoulli model with uniform priors; this includes Lemma 28 in Shang et al. (2020) (stating the relevant implications of Lemma 6 in KS that are used in the proof of Theorem 1 in KS), Lemma 29 in Shang et al. (2020), which maps to Lemma 5 in KS, and Lemma 30 in Shang et al. (2020), which maps to Lemma 4 in KS. The steps of the proof in KS that relate to the exploration sampling algorithm remain unchanged.[3]

[3]This proof is spelled out in greater detail in the revised version of Ariu et al. (2021).

Additional comments

The issue with the proof of item 3 of Theorem 1 in KS stems from the fact that item 1 in Lemma 2 of KS is incorrect. This Lemma connected the convergence arguments of Russo (2016) to the optimal allocation derived in Glynn and Juneja (2004). This Lemma was stated as follows: *Suppose that* $\bar{q}_T^d = m_T^d/(NT)$ *converges to* ρ^d *for all* d*, with* $\rho^{d^{(1)}} = \gamma$*. Then (1)* $\lim_{T\to\infty} -\frac{1}{NT}\log P\left(\hat{\theta}_T^d > \hat{\theta}_T^{d^{(1)}}\right) = \Gamma^d$*, and (2)* $\text{plim}_{T\to\infty} -\frac{1}{NT}\log p_T^d = \Gamma^d$*, where* $\Gamma^d = G^d(\rho^d)$ *for a function* $G^d : [0,1] \to \mathbb{R}$ *that is finitely valued, continuous, strictly increasing in* ρ^d*, and satisfies* $G^d(0) = 0$*.*

Our proof argued that the first claim follows from the arguments in Glynn and Juneja (2004), Section 2. This is not correct, for two reasons. First, the arguments of Glynn and Juneja (2004) only imply such a claim for converging non-random sequences of sample shares $\bar{q}_T^d$, but not, in general, for random sequences of shares that converge in probability. Second, the functions G^d which make item 1 correct (for non-random sequences of sample shares $\bar{q}_T^d$) are in general slightly different from the functions G^d which make item 2 correct, due to the asymmetry of KL-divergences. The remainder of this note elaborates on these two points.

Convergence of error probabilities and of posterior probabilities To understand why convergence in probability of sample shares is not enough for exponential convergence of error probabilities and policy regret, consider random sequences $\bar{q}_T^d$ of the following form:[4] Let α_T be some non-random sequence converging to 0, for instance $\alpha_T = \frac{1}{T}$. Assume that with probability $1-\alpha_T$, and independently of all potential outcomes, the shares $\bar{q}_T^d$ are equal to the fixed, non-random shares ρ^d (up to required rounding to the nearest integer); call this the event A_T. With probability α_T, in the event A_T^c, the shares $\bar{q}_T^d$ are equal to 0 for all but the worst arm $d(k)$. This sequence of sample shares converges in probability to $\boldsymbol{\rho}$. Suppose further that the true mean $\theta^{d(k)}$ of the worst arm is higher than the prior mean for the best arm $d^{(1)}$. As a consequence, by the law of large numbers, $P\left(\hat{\theta}_T^{d(k)} > \hat{\theta}_T^{d^{(1)}} | A_T^c\right) \to^p 1$.

If α_T does not go to 0 exponentially fast, the probability of making a mistaken policy choice cannot converge at an exponential rate either, since

$$P\left(\hat{\theta}_T^{d(k)} > \hat{\theta}_T^{d^{(1)}}\right) > P\left(\hat{\theta}_T^{d(k)} > \hat{\theta}_T^{d^{(1)}} | A_T^c\right) \cdot \alpha_T = (1 - o_p(1)) \cdot \alpha_T, \tag{1}$$

contradicting the first claim in Lemma 2 of KS. The same holds for (frequentist) expected policy regret.

By contrast, convergence in probability of $-\frac{1}{NT}\log p_T^d$ (or any other statistic that is a function of the data) is not affected by the vanishing events A_T^c. This holds because

$$P\left(\left|-\tfrac{1}{NT}\log p_T^d - \Gamma^d\right| < \epsilon\right) \leq P\left(\left|-\tfrac{1}{NT}\log p_T^d - \Gamma^d\right| < \epsilon \middle| A_T\right) \cdot (1-\alpha_T) + \alpha_T. \tag{2}$$

Convergence in probability is thus assured if $P\left(\left|-\frac{1}{NT}\log p_T^d - \Gamma^d\right| < \epsilon | A_T\right)$ converges to 0, since $P(A_T^c) = \alpha_T \to 0$.

Asymmetry of KL-divergences A second complication arises due to the asymmetry of KL-divergences. Consider w.l.o.g. some fixed parameter vector θ with $d^{(1)} = 1$. Let Θ^2 be the set of parameter vectors such that $d^{(1)} = 2$. The convergence of $P\left(\hat{\theta}_T \in \Theta^2\right)$ is driven by probabilities of the form $\max_{\tilde{\theta}\in\Theta^2} p_T(\tilde{\theta}|\theta)$. Here p_T is the sampling density (or probability mass function) of $(\bar{Y}_T^1 \ldots \bar{Y}_T^k)$ under the true parameter θ; $\hat{\theta}_T$ is approximately equal to $\bar{Y}_T$, and we leave the sample shares $\bar{q}_T^d$ implicit. Large-deviations arguments imply that convergence rates are governed by the KL-divergence $\min_{\tilde{\theta}\in\Theta^2} d(\theta||\tilde{\theta})$. This is the relevant divergence for Glynn and Juneja (2004).[5]

[4] This example is based on a suggestion by Daniel Russo, in private communication.

[5] Glynn and Juneja (2004) express their result in terms of Legendre transforms of the log-moment generating function, rather than KL-divergences, however.

By contrast, the convergence of the posterior probability $P_T(\theta \in \Theta^2)$ is driven by probabilities of the form $\max_{\tilde{\theta} \in \Theta^2} p_T(\theta|\tilde{\theta})$. Here p_T is the likelihood (or probability mass function) of $(\bar{Y}^1, \dots \bar{Y}^k)$ under the counterfactual parameter $\tilde{\theta}$; $\bar{Y}$ is approximately equal to the true θ in large samples. Large-deviations arguments again imply that convergence rates are governed by the KL-divergence $\min_{\tilde{\theta} \in \Theta^2} d(\tilde{\theta}||\theta)$. This is the relevant divergence for Russo 2020. The two divergences $\min_{\tilde{\theta} \in \Theta^2} d(\theta||\tilde{\theta})$ and $\min_{\tilde{\theta} \in \Theta^2} d(\tilde{\theta}||\theta)$ are equal in special cases, in particular for normally distributed outcomes with known variance. They are not equal in general, however. An interesting question for future work will be to bound the differences in the implied optimal allocations; in many settings these will be very small.

References

Ariu, K., M. Kato, J. Komiyama, and K. McAlinn (2021, September). Policy choice and best arm identification: comments on "Adaptive treatment assignment in experiments for policy choice". *arXiv:2109.08229 [cs, econ, stat]*.

Carpentier, A. and A. Locatelli (2016). Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pp. 590–604. PMLR.

Glynn, P. and S. Juneja (2004). A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004*, Volume 1. IEEE.

Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica 89*(1), 113–132.

Qin, C., D. Klabjan, and D. Russo (2017). Improving the expected improvement algorithm. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5387–5397.

Russo, D. (2016). Simple Bayesian algorithms for best-arm identification. *arXiv:1602.08448 [cs.LG]*.

Russo, D. (2020). Simple Bayesian algorithms for best-arm identification. *Operations Research 68*(6), 1625–1647.

Shang, X., R. Heide, P. Menard, E. Kaufmann, and M. Valko (2020). Fixed-confidence guarantees for Bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pp. 1823–1832. PMLR.