Supplementary Information

Pieter Libin[1,2], Timothy Verstraeten[1], Diederik M. Roijers[1],
Wenjia Wang[1], Kristof Theys[2], and Ann Nowé[1]

[1]Vrije Universiteit Brussel, Brussels, Belgium
[2]Katholieke Universiteit Leuven, Leuven, Belgium

June 29, 2019

Contents

1 Introduction

This Supplementary Information accompanies the manuscript titled "Bayesian Anytime m-top Exploration". In this document, we provide some more details on the derived posteriors in Section 2, we show additional experimental results in Section 3, we present some additional information with respect to the theoretical analyses in Section 4 and we list the figures with respect to the empirical validation of the heuristics in Section 5.

2 Posterior distributions

2.1 Truncated Normal

We consider a reward distribution $\mathcal{N}(\mu, \sigma^2)$ with known variance. We then have a conjugate prior for the mean that is Gaussian with hyper-parameters μ_0 and σ_0^2. As the means are in $[0,1]$, we choose this Gaussian prior to be truncated on said interval. To obtain an uninformative prior, we consider $\sigma_0^2 \to \infty$. This results in a uniform prior $\mathcal{U}(0,1)$ over μ:

$$\begin{aligned}\lim_{\sigma_0\to+\infty} \mathcal{TN}_{[0,1]}(\mu \mid \mu_0, \sigma_0^2) &= \lim_{\sigma_0\to+\infty} \frac{\sigma_0^{-1}\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)}{\sigma_0^{-1}\int_0^1 \exp\left(-\frac{(\mu'-\mu_0)^2}{2\sigma_0^2}\right)d\mu'} \\ &= \lim_{\sigma_0\to+\infty} \frac{1}{\int_0^1 \exp\left(\frac{(\mu-\mu_0)^2-(\mu'-\mu_0)^2}{2\sigma_0^2}\right)d\mu'} \\ &\propto 1\end{aligned} \tag{1}$$

Given rewards $\mathbf{r} = \{r_1, ..., r_n\}$ we have posterior:

$$\mu \sim \mathcal{N}(\hat{\mu}_0, \hat{\sigma}_0^2), \tag{2}$$

with,

$$\begin{aligned}\hat{\sigma}_0^2 &= \lim_{\sigma_0\to+\infty}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1} = \frac{\sigma^2}{n} \\ \hat{\mu}_0 &= \lim_{\sigma_0\to+\infty}\frac{\sigma^2}{n}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n r_i}{\sigma^2}\right) = \frac{\sum_{i=1}^n r_i}{n}\end{aligned} \tag{3}$$

2.2 Dirichlet

For a categorical reward distribution $\mathcal{C}\mathrm{at}_{\mathbf{c}}(\mathbf{p})$, the conjugate prior is a Dirichlet distribution $\mathcal{D}\mathrm{ir}_{\mathbf{c}}(\boldsymbol{\alpha})$ with prior parameter $\boldsymbol{\alpha}$. Given rewards $\mathbf{r} = \{r_1, ..., r_n\}$, we have posterior

$$\mu \sim \mathbf{c} \cdot \mathcal{D}\mathrm{ir}_{\mathbf{c}}(\boldsymbol{\alpha} + \mathbf{f}), \tag{4}$$

where $\mathbf{f}$ is a vector of frequencies at which the categories occur in $\mathbf{r}$.

We report the expectation over μ with respect to the Dirichlet posterior:

$$\mathbb{E}\big[\mathbf{c} \cdot \mathcal{D}\mathrm{ir}_{\mathbf{c}}(\boldsymbol{\alpha} + \mathbf{f})\big]. \tag{5}$$

2.3 Truncated t-distribution

We consider a Gaussian reward distribution with unknown variance and assume an uninformative Jeffreys prior $(\sigma)^{-3}$ on (μ, σ^2).
Given rewards $\mathbf{r} = \{r_1, ..., r_n\}$, this prior leads to the non-standardized t-distributed posterior, that we truncate given that we know that the arm's means are in $[0, 1]$:

$$\mu \sim \mathcal{T}_{n,[0,1]}\left(\mu_0 = \frac{\sum_{i=1}^{n} r_i}{n}, \sigma_0^2 = \frac{\sum_{i=1}^{n}(r_i - \mu_0)^2}{n^2}\right). \tag{6}$$

Given the pdf $f(\cdot)$ of a non-standardized t-distribution $\mathcal{T}_\upsilon(\mu, \sigma^2)$

$$f(x) = \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1 + \frac{(x-\mu)^2}{\upsilon\sigma^2}\right)^{-\frac{\upsilon+1}{2}}, \tag{7}$$

and cdf $F(\cdot)$, we can compute the mean of the truncated non-standardized t-distribution using this normalized definite integral:

$$\frac{\int_0^1 xf(x)dx}{F(\frac{1-\mu}{\sigma}) - F(\frac{0-\mu}{\sigma})} \tag{8}$$

From this, we can derive an analytic expression by first considering the nominator:

$$\begin{aligned}
&\int_0^1 xf(x)dx \\
&= \int_0^1 x\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1 + \frac{(x-\mu)^2}{\upsilon\sigma^2}\right)^{-\frac{\upsilon+1}{2}} dx \\
&= \int_{x=0}^{x=1} \sigma\frac{x-\mu+\mu}{\sigma}\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1 + \frac{(x-\mu)^2}{\upsilon\sigma^2}\right)^{-\frac{\upsilon+1}{2}} \frac{1}{\sigma}dx, \quad u = \frac{x-\mu}{\sigma}, du = \frac{1}{\sigma}dx \\
&= \int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} (\sigma u + \mu)\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1 + \frac{u^2}{\upsilon}\right)^{-\frac{\upsilon+1}{2}} du \\
&= \int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} \sigma uf(u)du + \int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} \mu f(u)du
\end{aligned} \tag{9}$$

Substituting this in Equation 8, we have:

$$\begin{aligned}&\frac{\int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} \sigma u f(u) du + \int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} \mu f(u) du}{F(\frac{1-\mu}{\sigma}) - F(\frac{0-\mu}{\sigma})} \\ &= \sigma \mathbb{E}\left[u \mid \frac{-\mu}{\sigma} \leq u \leq \frac{1-\mu}{\sigma}\right] + \mu\end{aligned} \tag{10}$$

3 Experimental results

In this section, we show additional results for the experiment that we conducted on the Gaussian bandit with fixed variance. In the main manuscript, we only show the results for the linear bandit with $m = 10$. Here we show all results for this experiment, i.e., the linear bandit with $m = 10$ in Figure 1, the linear bandit with $m = 50$ in Figure 2, the polynomial bandit with $m = 10$ in Figure 3 and the polynomial bandit with $m = 50$ in Figure 4.

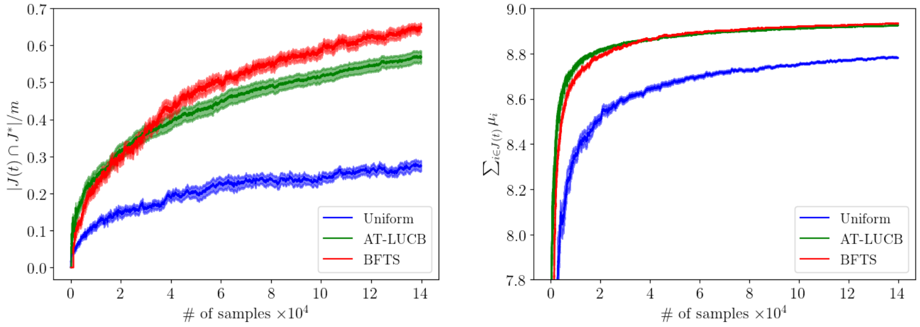

Figure 1: Results for the linear Gaussian benchmark with fixed variance ($m = 10$).

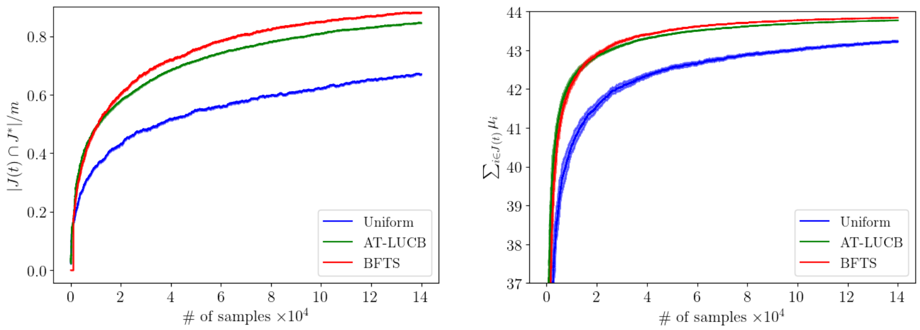

Figure 2: Results for the linear Gaussian benchmark with fixed variance ($m = 50$).

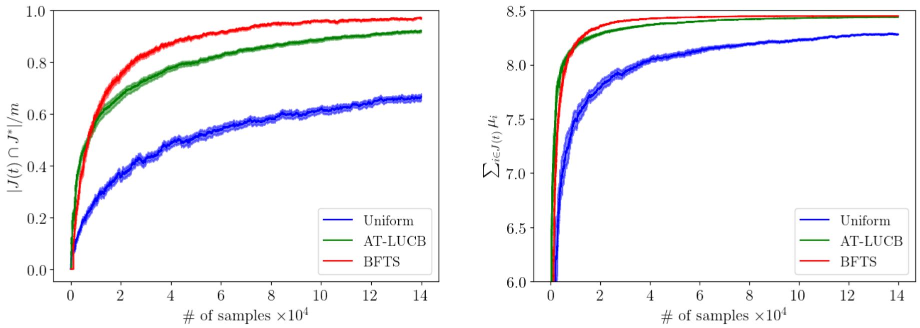

Figure 3: Results for the polynomial Gaussian benchmark with fixed variance ($m = 10$).

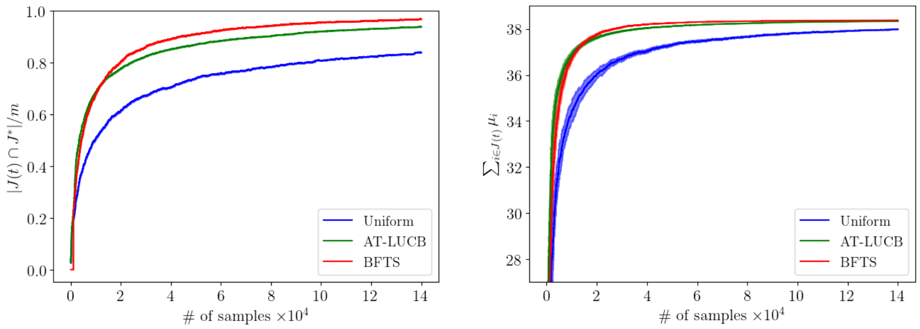

Figure 4: Results for the polynomial Gaussian benchmark with fixed variance ($m = 50$).

4 Bayesian analysis

We show the derivation for the bound in terms of A_m^{TS} as discussed in the Bayesian analysis section of the main manuscript:

$$\begin{aligned}
&P_t\left(J^* \neq J^{TS}\right) \\
&= P_t\left(\bigvee_{\rho^+} A_{\rho^+}^{TS} \in \overline{J^*}\right) \\
&\leq \sum_{\rho^+} P_t\left(A_{\rho^+}^{TS} \in \overline{J^*}\right) \\
&= \frac{\sum_{\rho^+} P_t\left(A_{\rho^+}^{TS} \in \overline{J^*}\right) \cdot m}{m} \\
&= \mathop{\mathbb{E}}_{\rho^+}\left[P_t\left(A_{\rho^+}^{TS} \in \overline{J^*}\right)\right] \cdot m \\
&\overset{(H2)}{\leq} P_t\left(A_m^{TS} \in \overline{J^*}\right) \cdot m
\end{aligned} \tag{11}$$

In the first step, we express the probability of error in terms of the arms that are ranked as optimal by Thompson sampling. In the second step, we apply a union bound. In the third and fourth step, we transform the sum to an expected value. In the final step, we apply Heuristic 2 (H2).

5 Empirical validation of the heuristics

5.1 Learning curves

In the main manuscript, we conduct an experiment to empirically validate the heuristics. We evaluate the heuristics for the same environments as in the experiments section, but with a limited number of arms ($K = 100$) and time steps (i.e., $3 \cdot 10^4$). In this section, we show the results for these additional experiments.

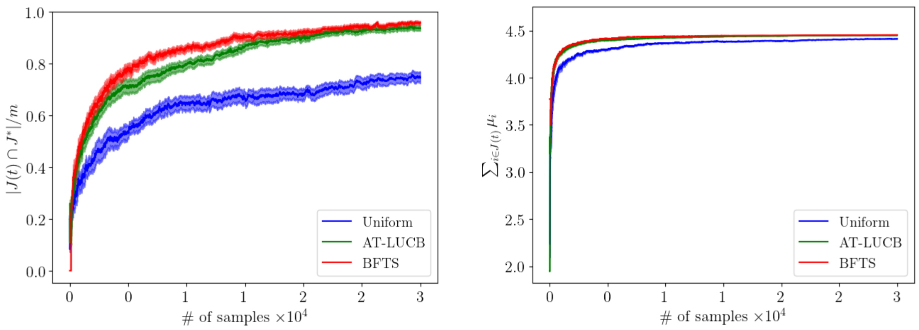

Figure 5: Results for the linear Gaussian benchmark with fixed variance ($K = 100, m = 5$).

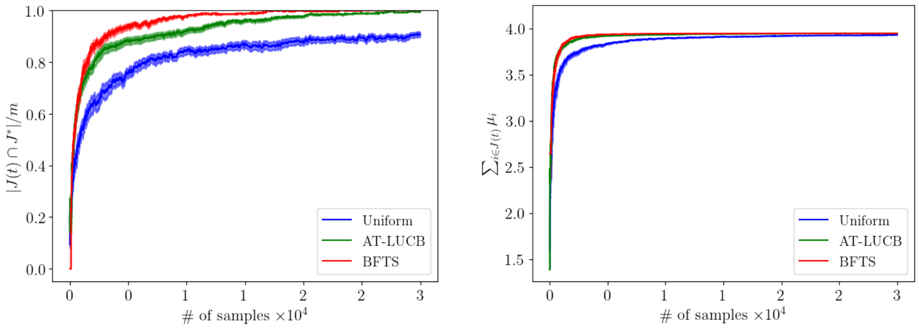

Figure 6: Results for the polynomial Gaussian benchmark with fixed variance ($K = 100, m = 5$).

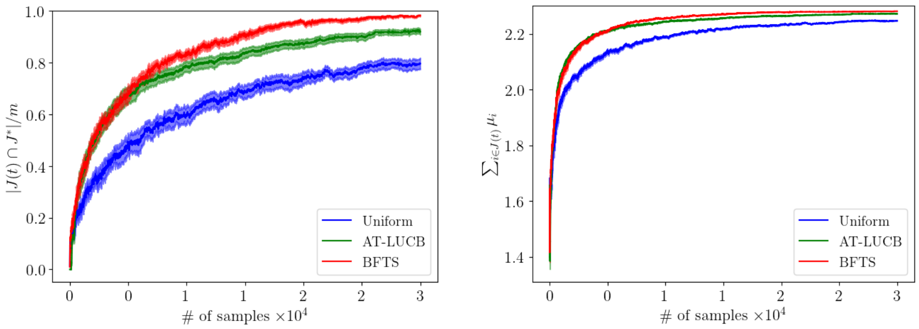

Figure 7: Results for the caption benchmark ($K = 100, m = 5$).

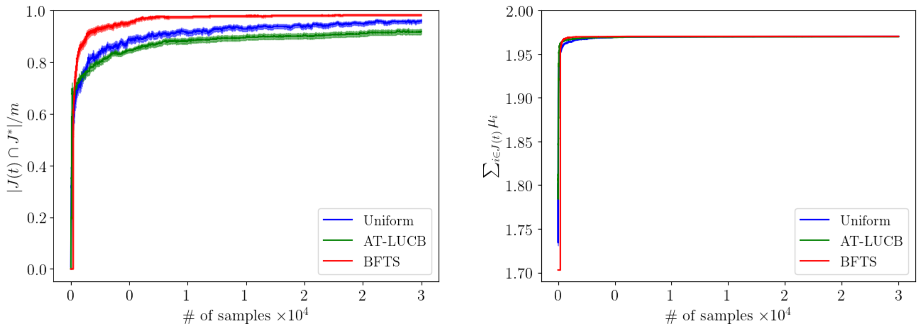

Figure 8: Results for the scaled Gaussian benchmark $(K = 100, m = 5)$.

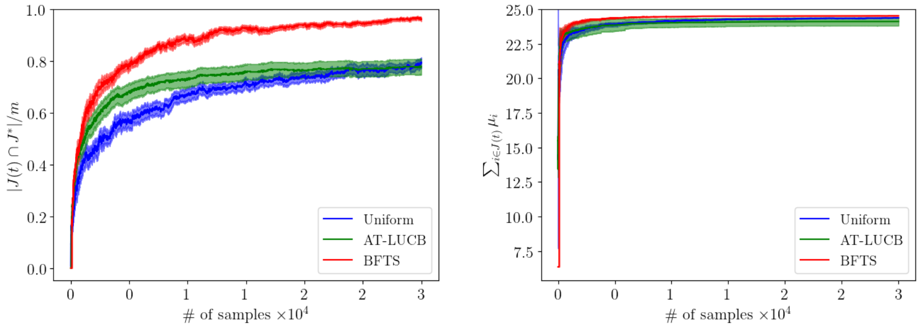

Figure 9: Results for the Poisson benchmark $(K = 100, m = 5)$.

5.2 Heuristic trajectories

In the main manuscript, we state that Heuristic 2 is to be interpreted as a zero-bounded difference in probabilities:

$$\mathop{\mathbb{E}}_{\rho^+}[P_t(A_{\rho^+}^{TS} \in \overline{J^*})] - P_t(A_m^{TS} \in \overline{J^*}) \leq 0 \tag{12}$$

We note that all failures are caused by differences that are close to zero. This can be visualized by plotting the trajectories of probability differences. We show these figures for the Caption environment (Figure 10), the Poisson environment (Figure 11) and the scaled Gaussian environment (Figure 12).

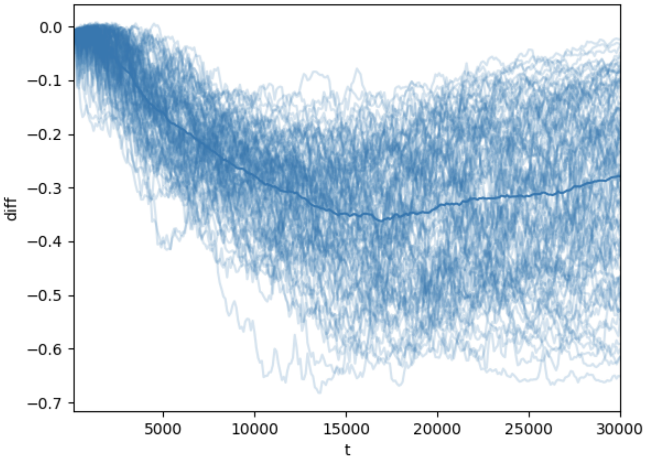

Figure 10: Trajectories of probability differences for heuristic 2 in the Caption environment (traces: light blue lines, mean: thick blue line).

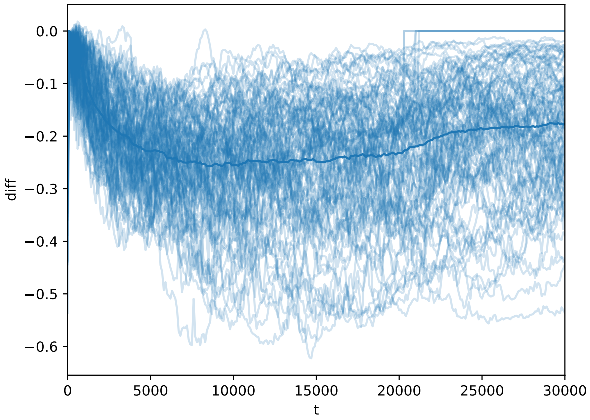

Figure 11: Trajectories of probability differences for heuristic 2 in the Poisson environment (traces: light blue lines, mean: thick blue line).

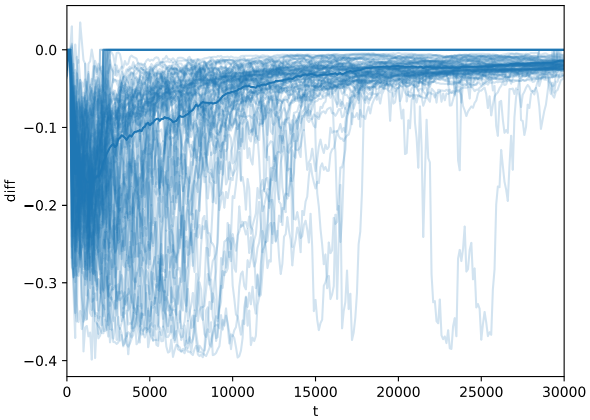

Figure 12: Trajectories of probability differences for heuristic 2 in the scaled Gaussian environment (traces: light blue lines, mean: thick blue line).