1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–8

Sequential Monte Carlo for Dynamic Softmax Bandits

Iñigo Urteaga INIGO.URTEAGA@COLUMBIA.EDU
Chris H. Wiggins CHRIS.WIGGINS@COLUMBIA.EDU
Department of Applied Physics and Applied Mathematics, Data Science Institute
Columbia University, New York City, NY 10027

Abstract

We propose a sequential Monte Carlo (SMC) approach to the dynamic softmax bandits problem, a case of multi-armed bandits (MAB) in which rewards are categorical and the bandit parameters evolve over time. We show how SMC can be combined with state-of-the-art MAB algorithms (Thompson sampling and Bayes-UCB) to attain competitive performance in dynamic softmax bandits. In the stochastic MAB setting, the reward for each action is drawn from an unknown distribution, and to make sequential optimal decisions, one must compute sufficient statistics of such distributions, e.g., expectations or upper-confidence bounds (UCB). Since closed-form expressions for these statistics of interest are analytically intractable except in special cases, we estimate them via SMC. These are accurate enough for MAB policies to operate successfully in dynamic softmax bandits.

Keywords: Multi-armed bandits, sequential Monte Carlo, dynamic softmax bandits.

1. Introduction

The multi-armed bandit (MAB) problem is named after the thought-experiment in which an agent plays a row of slot machines, i.e., one decides which arm to play next to maximize returns while simultanously learning the machine's payoffs. This setting, more formally referred to as sequential decision processes, extends to a wide range of real-world challenges that require online learning while simultaneously maximizing some notion of reward. Bayesian modeling of the MAB facilitates not only generative and interpretable modeling, but sequential and batch processing algorithm development as well. Central to Bayesian MAB algorithms (such as Thompson sampling (Russo et al., 2018) and Bayes-UCB (Kaufmann et al., 2012)) is posterior inference. One must sample from the posterior distributions and/or calculate expected rewards, which is cumbersome except for simple models (Korda et al., 2013).

In this work, we connect approximate Bayesian inference (sequential Monte Carlo (SMC) in particular) with the field of reinforcement learning and sequential decision processes. We leverage the flexibility of SMC methods (Arulampalam et al., 2002; Doucet et al., 2001; Djurić et al., 2003), which have been successful in many applications of science and engineering (Ristic et al., 2004; van Leeuwen, 2009; Ionides et al., 2006; Creal, 2012), to extend Bayesian MAB algorithms to new scenarios of interest: dynamic bandits with categorical rewards. MABs are widely used in many real-world problems, such as display advertisement and recommender systems, to handle the inherent explore-exploit tradeoff. However, assuming a stationary reward distribution hardly holds in practice, as users' preferences evolve over time; thus the need for dynamic bandits. Practical applications also often require complex

© I. Urteaga & C. H. Wiggins.

reward functions, such as categorical distributions, to model user interactions: e.g., a user ignores the recommended movie, clicks on the trailer, or watches the movie; a subscription reminder is ignored, the remind-me-later button is clicked, or the user subscribes.

Our contribution here is an SMC-based MAB method that (i) approximates the posterior softmax reward densities via random measures; and (ii) is applicable to time-varying parameter models, i.e., dynamic bandits. We study the general linear dynamical system, and provide the solution for the unknown parameter case, by combining Rao-Blackwellization and SMC. We implement SMC both for Thompson sampling and UCB-based policies in bandits with categorical contextual rewards, modeled via the softmax function.

2. Dynamic Softmax Bandits

In this paper, we study dynamic categorical bandits. Mathematically, the MAB with per-arm stochastic reward functions and dynamic parameters is modeled via the transition density of the dynamics, i.e., $\theta_{a,t} \sim p(\theta_{a,t}|\theta_{a,1:t-1})$, and the stochastic reward distribution $y_t \sim p_{a_t}(y|x_t, \theta_{a,t})$. For bandit problems where returns are not binary but categorical, and contextual information is available, the softmax function is a natural modeling choice. Given a d-dimensional context vector $x \in \mathbb{R}^d$, and per-arm parameters $\theta_a = \{\theta_{a,1}, \cdots, \theta_{a,C}\}$ for each category $c \in \{1, \cdots, C\}$, the contextual softmax reward model follows $p_a(y = c|x, \theta_a) = \exp(x^\top \theta_{a,c}) / \sum_{c'=1}^{C} \exp(x^\top \theta_{a,c'})$. In practical scenarios, one is interested in learning about the dynamic state of the world as data are collected, i.e., as the underlying parameters of the reward function evolve over time. A widely applicable framework for time-evolving bandits is the general linear model, where the reward parameters $\theta_{a,c} \in \mathbb{R}^d$ per-arm and category, follow dynamics $\theta_{a,c,t} = L_{a,c}\, \theta_{a,c,t-1} + \epsilon_{a,c}$, with noise $\epsilon_{a,c} \sim \mathcal{N}(\epsilon_{a,c}|0, \Sigma_{a,c})$, $L_{a,c} \in \mathbb{R}^{d\times d}$, and $\Sigma_{a,c} \in \mathbb{R}^{d\times d}$. When the parameters are known the transition distribution is Gaussian, $\theta_{a,c,t} \sim \mathcal{N}(\theta_{a,c,t}|L_{a,c}\, \theta_{a,c,t-1}, \Sigma_{a,c})$, and we recover the celebrated Kalman (1960) filter. For the unknown parameter case, the marginalized transition density[1] is a multivariate-t, $\theta_{a,c,t} \sim \mathcal{T}(\theta_{a,c,t}|\nu_{a,c,t}, m_{a,c,t}, R_{a,c,t})$.

A MAB policy decides which arm to play next based on the set of given contexts, played arms, and observed rewards up to time t: i.e., the history $\mathcal{H}_{1:t} = \{y_{1:t}, a_{1:t}, x_{1:t}\}$, with $y_{1:t} \equiv (y_1, \cdots, y_t)$, $a_{1:t} \equiv (a_1, \cdots, a_t)$, and $x_{1:t} \equiv (x_1, \cdots, x_t)$. In the Bayesian MAB setting, one takes into account the uncertainty on the unknown and dynamic parameters via priors and, as one interacts with the environment, the parameter posterior is updated: $p(\theta_{a,t}|\mathcal{H}_{1:t}) \propto p_{a_t}(y_t|x_t, \theta_{a,t}) p(\theta_{a,t}|\mathcal{H}_{1:t-1})$, where $p_{a_t}(y_t|x_t, \theta_{a,t})$ is the likelihood of the observed reward y_t after playing arm a_t at time t. The posterior $p(\theta_{a,t}|\mathcal{H}_{1:t})$ is necessary for both posterior sampling and confidence interval based MAB algorithms to compute sufficient statistics of the expected rewards of each arm; e.g., $\mu_{a,t} = \sum_{c=1}^{C} c \cdot p_a(y = c|x_t, \theta_{a,t})$ for the softmax model. However, $p(\theta_{a,t}|\mathcal{H}_{1:t})$ can not be computed in closed form for dynamic softmax bandits.

To that end, we implement Sequential Importance Resampling (SIR) as in Gordon et al. (1993) for the MAB problem of interest: the proposal distribution matches the assumed parameter dynamics, i.e., $\pi(\theta_{a,t}) = p(\theta_{a,t}|\theta_{a,1:t-1})$; weights are updated based on the likelihood of observed rewards, i.e., $p_a(y_t|x_t, \theta_{a,t})$; and the random measure is resampled at every time instant. In the bandit setting, one cares about the posterior density of the parameters at each time instant, i.e., the filtering density $p(\theta_{a,t}|\mathcal{H}_{1:t})$, for which there are strong theoretical

1. Details of the derivation and how to compute the sufficient statistics are provided in Appendix A .

SMC convergence guarantees (Crisan and Doucet, 2002; Chopin, 2004). Recall that the proposed SIR method approximates each per-arm categorical posterior separately. Therefore, there will be no particle degeneracy due to increased number of arms.

The propagation of parameter samples in the SIR algorithm is fundamental for the accuracy of the sequential approximation to the posterior, as well as for the performance of the SIR-based MAB policy. The increasing uncertainty of the parameter posterior encourages exploration of arms that have not been played recently, but may have evolved into new parameter spaces with exploitable reward distributions. That is, as the dynamics of unobserved arms result in broad SIR posteriors (increased uncertainty about parameters), MAB policies are more likely to explore such arm, reduce their posterior's uncertainty, and in turn, update the exploration-exploitation balance. We now describe how SIR can be used for both Thompson sampling and UCB-type policies (see Algorithm 1 for full details).

Algorithm 1 SIR for MAB

Require: A, $p(\theta_a)$, $p(\theta_{a,t}|\theta_{a,1:t-1})$, $p_a(y|x,\theta)$, M (for UCB we also require α_t)

Draw initial samples from the parameter prior $\overline{\theta}_{a,0} \sim p(\theta_a), \forall a \in A$, and $w_{a,0}^{(m)} = \frac{1}{M}$.
for $t = 0, \cdots, T-1$ **do**
  Receive context x_{t+1}
  for $a = 1, \cdots, A$ **do**
    Estimate sufficient statistics for the MAB policy, given updated $\{w_{a,t}^{(m)}\}$ and $\{\theta_{a,1:t}^{(m)}\}$
    Thompson sampling:
      Draw a sample index $s \sim \mathcal{C}at\left(w_{a,t}^{(m)}\right)$ and propagate $\theta_{a,t+1}^{(s)} \sim p(\theta_{a,t+1}|\theta_{a,1:t}^{(s)})$.
      Set $\mu_{a,t+1} = \mathbb{E}\left\{y|x_{t+1}, \theta_{a,t+1}^{(s)}\right\}$.
    Bayes-UCB:
      Draw samples $m \sim \mathcal{C}at\left(w_{a,t}^{(m)}\right)$, $m = 1, \cdots, M$ and propagate $\theta_{a,t+1}^{(m)} \sim p(\theta_{a,t+1}|\theta_{a,1:t}^{(m)})$,
      Set $\mu_{a,t+1}^{(m)} = \mathbb{E}\left\{y|x_{t+1}, \theta_{a,t+1}^{(m)}\right\}$ and compute $q_{a,t+1}(\alpha_{t+1}) = \max\{\mu | \sum_{m|\mu_{a,t+1}^{m} > \mu} w_{a,t}^{(m)} \geq \alpha_{t+1}\}$.
  end for
  Decide next action a_{t+1} to play
    Thompson sampling: $a_{t+1} = \text{argmax}_a\, \mu_{a,t+1}$
    Bayes-UCB: $a_{t+1} = \text{argmax}_a\, q_{a,t+1}(\alpha_{t+1})$
  Observe reward y_{t+1} for played arm
  Update posterior following SIR steps
    Resample $m = 1, \cdots, M$ parameters $\overline{\theta}_{a,1:t}^{(m)}$, where m is drawn with replacement according to the importance weights $w_{a,t}^{(m)}$.
    Propagate resampled parameters by drawing $\theta_{a,t+1}^{(m)} \sim p(\theta_{a,t+1}|\overline{\theta}_{a,1:t}^{(m)})$, $m = 1, \cdots, M$.
    Weight samples based on $\widetilde{w}_{a,t+1}^{(m)} \propto p(y_{t+1}|x_{t+1}, \theta_{a,t+1}^{(m)})$, $m = 1, \cdots, M$.
    Normalize weights by $w_{a,t+1}^{(m)} = \widetilde{w}_{a,t+1}^{(m)} / \sum_{m'=1}^{M} \widetilde{w}_{a,t+1}^{(m')}$, $m = 1, \cdots, M$.
end for

2.1. SIR-based Thompson Sampling

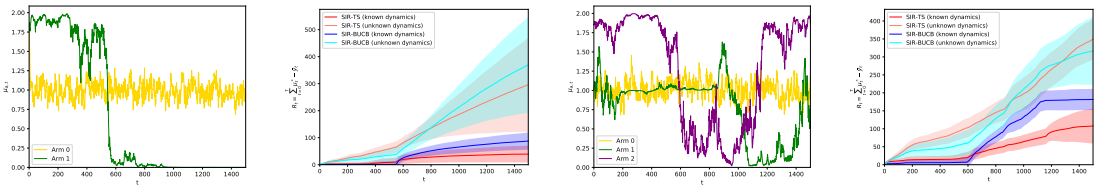
Thompson (1935) sampling is a probability matching algorithm that randomly selects an action to play according to the probability of it being optimal. Thompson sampling has been empirically proven to perform satisfactorily, and to enjoy provable optimality properties for different reward models with and without context (Agrawal and Goyal, 2012a,b; Korda et al., 2013; Russo and Roy, 2014, 2016). It is based on the computation of the probability of an arm being optimal, which is in general analytically intractable. Alternatively, Thompson sampling operates by drawing a sample parameter $\theta_{a,t}^{(s)}$ from per-arm updated posteriors $p(\theta_{a,t}|\mathcal{H}_{1:t})$, and picking the optimal arm for such sample; i.e., $a_t^* = \text{argmax}_a\, \mu_{a,t}^{(s)}$, where $\mu_{a,t}^{(s)} = \mathbb{E}_a\{y_t|x_t, \theta_{a,t}^{(s)}\}$. In Algorithm 1, we propose to draw from the SIR-based random measure instead, as it provides an accurate approximation to the true parameter posterior with high probability.

2.2. SIR-based Bayes-UCB

Bayes-UCB (Kaufmann et al., 2012) is a Bayesian approach to UCB algorithms, where Bayesian quantiles are used as proxies for confidence bounds. Kaufmann et al. (2012) have proven the asymptotic finite-time regret optimality of Bayes-UCB for Bernoulli bandits, and argue that it provides an unifying framework for UCB-based algorithms with parametric rewards. However, its application is limited to models where the quantile functions are analytically tractable. We instead compute the quantile function of interest by means of the SIR approximation to the parameter posterior. The expected reward at each round t is evaluated based on the posterior SIR samples, i.e., $\mu_{a,t}^{(m)} = \mathbb{E}_a\{y_t|x_t, \theta_{a,t}^{(m)}\}$. The quantile value $\Pr[\mu_{a,t} > q_{a,t}(\alpha_t)] = \alpha_t$ is computed by $q_{a,t}(\alpha_t) := \max\{\mu | \sum_{m|\mu_{a,t}^m > \mu} w_{a,t}^{(m)} \geq \alpha_t\}$.

3. Evaluation

We evaluate the proposed SIR-based method on dynamic bandits with softmax reward functions. We show in Figure 1 the time evolution of the expected rewards of the studied bandits, as well as the performance of the proposed methods. We observe that SIR-based Thompson sampling and Bayes-UCB are able to dynamically reach the exploitation-exploration balance (the cumulative regret plateaus in subfigures 1b and 1d). There is a regret loss incurred when the underlying dynamics are not known, as the algorithm must sequentially learn the unknown model parameters $\{L, \Sigma\}$, in order to make informed decisions. Notice the increases in regret when the parameter dynamics swap the optimal arm (subfigures 1a and 1c). Reward changes in arms impact Bayes-UCB more profoundly as time evolves, which we argue is due to the shrinking quantile value $\alpha_t \propto 1/t$ proposed by Kaufmann et al. (2012), as it is not able to capture the evolving uncertainty of the parameter posteriors. More generally, the need to determine appropriate quantile values α_t for each reward and dynamic model is a drawback for Bayes-UCB. On the contrary, Thompson sampling only relies on samples from the posterior, which SIR is able to approximate accurately enough for it to operate successfully, even in the most challenging MAB scenarios without any parameter tweaking.



(a) Expected rewards $A = 2, C = 3$. (b) Cumulative regret $A = 2, C = 3$. (c) Expected rewards $A = 3, C = 3$. (d) Cumulative regret $A = 3, C = 3$.

Figure 1: True expected reward and mean regret performance (standard deviation shown as shaded region) of the proposed SIR-based methods for dynamic softmax bandits.

4. Conclusions

In this work, we have addressed dynamic softmax bandits, by applying SMC to the multi-armed bandit setting, for both Thompson sampling and Bayes-UCB algorithms. We show that MAB policies with SIR-based approximations to the posterior attain competitive performance. We aim to extend this work to other bandit scenarios and real datasets.

References

Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. *CoRR*, abs/1209.3352, 2012a.

Shipra Agrawal and Navin Goyal. Further Optimal Regret Bounds for Thompson Sampling. *CoRR*, abs/1209.3353, 2012b.

M. Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *Signal Processing, IEEE Transactions on*, 50(2):174–188, 2 2002. ISSN 1053-587X.

José M. Bernardo and Adrian F.M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 9780470317716. doi: 10.1002/9780470316870.

Nicolas Chopin. Central Limit Theorem for Sequential Monte Carlo Methods and Its Application to Bayesian Inference. *The Annals of Statistics*, 32(6):2385–2411, 2004. ISSN 00905364.

Drew Creal. A Survey of Sequential Monte Carlo Methods for Economics and Finance. *Econometric Reviews*, 31(3):245–296, 2012.

Dan Crisan and Arnaud Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3):736–746, Mar 2002. ISSN 1053-587X. doi: 10.1109/78.984773.

Petar M. Djurić, Jayesh H. Kotecha, Jianqui Zhang, Yufei Huang, Tadesse Ghirmai, Mónica F. Bugallo, and Joaquín Míguez. Particle Filtering. *IEEE Signal Processing Magazine*, 20(5): 19–38, 9 2003.

Arnaud Doucet, Nando De Freitas, and Neil Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

Neil J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEEE Proceedings*, 140(2):107 –113, 4 1993. ISSN 0956-375X.

Edward L. Ionides, C. Bretó, and A. A. King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443, 2006.

Rudolph Emil Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

Emilie Kaufmann, Olivier Cappe, and Aurelien Garivier. On Bayesian Upper Confidence Bounds for Bandit Problems. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 592–600, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson Sampling for 1-Dimensional Exponential Family Bandits. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1448–1456. Curran Associates, Inc., 2013.

Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, 2004. ISBN 9781580538510.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1): 1–96, 2018. ISSN 1935-8237. doi: 10.1561/2200000070. URL http://dx.doi.org/10.1561/2200000070.

William R. Thompson. On the Theory of Apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935. ISSN 00029327, 10806377.

Peter Jan van Leeuwen. Particle Filtering in Geophysical Systems. *Monthly Weather Review*, 12(137):4089–4114., 2009.

Appendix A. Linearly mixing dynamic bandits

Let us consider a general linear model for the dynamics of the parameters of each arm $\theta_a \in \mathbb{R}^d$ (we omit the categorical index c for clarity of notation):

$$\theta_{a,t} = L_a \theta_{a,t-1} + \epsilon_a \ , \qquad \epsilon_a \sim \mathcal{N}\left(\epsilon_a | 0, \Sigma_a\right) \ , \tag{1}$$

where $L_a \in \mathbb{R}^{d\times d}$ and $\Sigma_a \in \mathbb{R}^{d\times d}$. One can immediately determine that, for linearly dynamic bandits with known parameters, the parameters follow

$$\theta_{a,t} \sim \mathcal{N}\left(\theta_{a,t} | L_a \theta_{a,t-1}, \Sigma_a\right) \ . \tag{2}$$

However, it is unrealistic to assume that the parameters are known in practice. We thus marginalize them out by means of the following conjugate priors for the matrix A and covariance matrix Σ (we drop the per arm subscript a for clarity)

$$\begin{aligned} p(A, \Sigma | L_0, B_0, \nu_0, V_0) &= \mathcal{NIW}\left(A, \Sigma | L_0, B_0, \nu_0, V_0\right) \\ &= p(A | L_0, B_0, \Sigma) p(\Sigma | \nu_0, V_0) \\ &= \mathcal{MN}\left(A | L_0, \Sigma, B_0\right) \mathcal{IW}\left(\Sigma | \nu_0, V_0\right) \ , \end{aligned} \tag{3}$$

where the matrix variate Gaussian distribution follows

$$\mathcal{MN}\left(A | L_0, \Sigma, B_0\right) = \frac{e^{-\frac{1}{2}\mathrm{tr}\left\{B_0^{-1}(A-L_0)^\top \Sigma^{-1}(A-L_0)\right\}}}{(2\pi)^{(d\cdot d)/2}\left|B_0\right|^{d/2}\left|\Sigma\right|^{d/2}} \ , \tag{4}$$

and the Inverse Wishart

$$\mathcal{IW}\left(\Sigma | \nu_0, V_0\right) = \frac{\left|\Sigma\right|^{-\frac{\nu_0+d+1}{2}} e^{-\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}V_0\right\}}}{2^{\frac{\nu_0 \cdot d}{2}}\left|V_0\right|^{-\frac{\nu_0}{2}}\Gamma\left(\frac{\nu_0}{2}\right)} \ . \tag{5}$$

We integrate out the unknown parameters A and Σ to derive the predictive density, i.e., the distribution of θ_t, given all the past data $\theta_{1:t}$. One can show that the resulting distribution is a multivariate t-distribution

$$\begin{aligned} f(\theta_t | \theta_{1:t-1}) &= \mathcal{T}\left(\theta_t | \nu_t, m_t, R_t\right) \\ &\propto \left|1 + \frac{1}{\nu_t}\left(\theta_t - m_t\right) R_t^{-1}\left(\theta_t - m_t\right)^\top\right|^{-\frac{\nu_t+d}{2}} \ , \end{aligned} \tag{6}$$

where ν_t denotes degrees of freedom, $m_t \in \mathbb{R}^d$ is the location parameter, and $R_t \in \mathbb{R}^{d\times d}$ represents the scale matrix (Bernardo and Smith, 2009). These follow

$$\begin{cases} \nu_t = \nu_0 + t - d \ , \\ m_t = L_{t-1}\theta_{t-1} \ , \\ R_t = \frac{V_{t-1}}{\nu_t\left(1 - \theta_{t-1}^\top (UU^\top)^{-1}\theta_{t-1}\right)} \ , \end{cases} \tag{7}$$

where the sufficient statistics of the parameters are

$$\begin{cases} B_{t-1} = \left(\Theta_{0:t-2}\Theta_{0:t-2}^\top + B_0^{-1}\right)^{-1} \ , \\ L_{t-1} = \left(\Theta_{1:t-1}\Theta_{0:t-2}^\top + A_0 B_0^{-1}\right) B_{t-1} \ , \\ V_{t-1} = \left(\Theta_{1:t-1} - L_{t-1}\Theta_{0:t-2}\right)\left(\Theta_{1:t-1} - L_{t-1}\Theta_{0:t-2}\right)^\top \\ \qquad\qquad + \left(L_{t-1} - L_0\right) B_0^{-1} \left(L_{t-1} - L_0\right)^\top + V_0 \ , \\ UU^\top = \left(\theta_{t-1}\theta_{t-1}^\top + B_{t-1}^{-1}\right) \ , \end{cases} \tag{8}$$

and we have defined the stacked parameter matrix

$$\Theta_{t_0:t_1} = [\theta_{t_0}\theta_{t_0+1}\cdots\theta_{t_1-1}\theta_{a,t_1}] \in \mathbb{R}^{d\times(t_1-t_0)} \ . \tag{9}$$

All in all, for linear dynamic bandits with unknown parameters, the per-arm parameters follow

$$\theta_{a,t} \sim \mathcal{T}\left(\theta_{a,t} | \nu_{a,t}, m_{a,t}, R_{a,t}\right) \ . \tag{10}$$