# **Supplementary File:** Generalized Contextual Bandits With Latent Features: Algorithms and Applications

Xiongxiao Xu, Hong Xie, and John C.S. Lui

***Abstract*—This supplementary file contains technical proofs to lemmas and theorems in the main paper.**

## I. Proof to Lemmas and Theorems

### A. Proof of Lemma 1

We can apply the Bayesian theorem to derive the posterior distribution as

$$p(\boldsymbol{\Psi}|\mathcal{H}_t) \propto p(\mathcal{H}_t|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma})p(\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma})$$

Note that from the independence of the prior distributions, we can derive $p(\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma})$ as

$$p(\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma}) = p(\boldsymbol{Y})p(\boldsymbol{\theta},\boldsymbol{\vartheta})p(\boldsymbol{\sigma}) = p(\boldsymbol{\theta},\boldsymbol{\vartheta})\prod_{a\in\mathcal{A}} p(\boldsymbol{y}_a)p(\sigma_a).$$

From the independence among the feedbacks or rewards, we can derive $p(\mathcal{H}_t|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma})$ as

$$p(\mathcal{H}_t|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma}) = \prod_{\tau=1}^{t-1} p(R_\tau(A_\tau)|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma})$$

$$\propto \prod_{\tau=1}^{t-1}\prod_{a\in\mathcal{A}_\tau}\left[f(g_\tau^{-1}(R_\tau(a)) - \boldsymbol{x}_a^{\mathrm{T}}\boldsymbol{\theta} - \boldsymbol{y}_a^{\mathrm{T}}\boldsymbol{\vartheta},\sigma_a)\right]^{\mathbb{1}\{A_\tau=a\}}$$

This proof is then complete. ∎

### B. Proof of Theorem 1

Given all the known model parameters $\boldsymbol{\Psi} = [\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma}]$, we define the corresponding optimal action in decision round $t$ as $A_t^*(\boldsymbol{\Psi}) \in \arg\max_{a\in\mathcal{A}_t} \bar{R}_t(a;\boldsymbol{\Psi})$. Note that in the Bayesian regret setting, the known model parameters $\boldsymbol{\Psi} = [\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma}]$ are random variables with the same probability distribution as the prior distribution $p(\boldsymbol{\Psi})$. Furthermore, the conditional probability distribution of the unknown model parameters $p(\boldsymbol{\Psi})$ given the decision history $\mathcal{H}_{t-1}$ is equivalent to the posterior distribution of $p(\boldsymbol{\Psi})$, i.e.,

$$\mathbb{P}[\boldsymbol{\Psi}|\mathcal{H}_{t-1}] = p(\boldsymbol{\Psi}|\mathcal{H}_{t-1}).$$

From the GCL-PS algorithm, i.e., Algorithm 1, the sample $\boldsymbol{\Psi}_t$ of the unknown model parameters in decision round $t$, is generated from the posterior distribution $p(\boldsymbol{\Psi}|\mathcal{H}_{t-1})$. And the action $A_t$ is obtained by $A_t \in \arg\max_{a\in\mathcal{A}_t} \bar{R}(a;\boldsymbol{\Psi}_t)$. To make the presentation clear, we denote the selected action as $A_t(\boldsymbol{\Psi}_t)$. Then we have that

$$\mathbb{P}[A_t^*(\boldsymbol{\Psi}) = a|\mathcal{H}_t] = \mathbb{P}[A_t(\boldsymbol{\Psi}_t) = a|\mathcal{H}_t], \forall a \in \mathcal{A}_t.$$

Let $U_t(a;\mathcal{H}_{t-1})$ and $L_t(a;\mathcal{H}_{t-1})$ denote an upper and lower confidence bound of $\bar{r}(a,\boldsymbol{\Psi}) \triangleq \boldsymbol{x}_a^T\boldsymbol{\theta} + \boldsymbol{y}_a^T\boldsymbol{\vartheta}$ with the decision history $\mathcal{H}_{t-1}$, which will be constructed later. Then it follows that

$$\mathbb{E}[U_t(A_t(\boldsymbol{\Psi}_t);\mathcal{H}_{t-1})] = \mathbb{E}[U_t(A_t^*(\boldsymbol{\Psi});\mathcal{H}_{t-1})]. \quad (1)$$

The $g_t$ being $\zeta_t$ Lipschitz implies the following inequality:

$$R_T^{Bay}(\mathcal{D}) \le \int \sum_{t=1}^{T} \min\left\{\Delta(\mathcal{R}), \zeta_t\left[\max_{a\in\mathcal{A}_t}\bar{r}(a,\boldsymbol{\Psi}) - \bar{r}(A_t(\Psi_t),\boldsymbol{\Psi})\right]\right\} p(\boldsymbol{\Psi})d\boldsymbol{\Psi}.$$

Then with a similar derivation as [1], we can bound the Bayesian regret as

$$R_T^{Bay}(\mathcal{D}) \le \mathbb{E}\left[\sum_{t=1}^{T}\min\{\Delta(\mathcal{R}),\zeta_t[U_t(A_t(\boldsymbol{\Psi}_t);\mathcal{H}_{t-1}) - L_t(A_t(\boldsymbol{\Psi}_t);\mathcal{H}_{t-1})]\}\right] + \Delta(\mathcal{R})T\mathbb{P}[\exists a,t,\bar{r}_t(a,\boldsymbol{\Psi}) \notin [L_t(A_t(\boldsymbol{\Psi}_t);\mathcal{H}_{t-1}),U_t(A_t(\boldsymbol{\Psi}_t);\mathcal{H}_{t-1})]].$$

Via conditioning, we can derive the right hand side of the above inequality as

$$\mathbb{E}\left[\sum_{t=1}^{T}\min\{\Delta(\mathcal{R}),\zeta_t[U_t(A_t(\boldsymbol{\Psi}_t);\mathcal{H}_{t-1}) - L_t(A_t(\boldsymbol{\Psi}_t);\mathcal{H}_{t-1})]\}\right]$$
$$= \mathbb{E}_{\boldsymbol{\Psi}\sim p(\boldsymbol{\Psi})}\left[\mathbb{E}\left[\sum_{t=1}^{T}\min\{\Delta(\mathcal{R}),\zeta_t[U_t(A_t(\boldsymbol{\Psi}_t);\mathcal{H}_{t-1}) - L_t(A_t(\boldsymbol{\Psi}_t);\mathcal{H}_{t-1})]\} \middle| \boldsymbol{\Psi}\right]\right]$$

We construct the confidence bound as

$$U_t(a;\mathcal{H}_{t-1}) = \widetilde{\boldsymbol{x}}_a^T\boldsymbol{\theta}_{t-1} + (\zeta_t\xi_a\sqrt{(d+|\mathcal{A}|)\log(T+T^2(L+1))} + \|\widetilde{\boldsymbol{\theta}}\|)\|\widetilde{\boldsymbol{x}}_a\|_{\boldsymbol{V}_{t-1}},$$
$$L_t(a;\mathcal{H}_{t-1}) = \widetilde{\boldsymbol{x}}_a^T\boldsymbol{\theta}_{t-1} - (\zeta_t\xi_a\sqrt{(d+|\mathcal{A}|)\log(T+T^2(L+1))} + \|\widetilde{\boldsymbol{\theta}}\|)\|\widetilde{\boldsymbol{x}}_a\|_{\boldsymbol{V}_{t-1}},$$

Xiongxiao Xu and Hong Xie are with Chongqing Key Laboratory of Software Theory and Technology, Chongqing University, Chongqing, China. John C.S. Lui is with Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.

where

$$\widetilde{\boldsymbol{x}}_a \triangleq \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{e}_a \end{bmatrix}, \boldsymbol{V}_t \triangleq \boldsymbol{I} + \sum_{\tau=1}^{t} \widetilde{\boldsymbol{x}}_{A_\tau} \widetilde{\boldsymbol{x}}_{A_\tau}^T,$$

$$\boldsymbol{\theta}_t \triangleq \boldsymbol{V}_t^{-1} \sum_{\tau=1}^{t} \widetilde{\boldsymbol{x}}_{A_\tau} g_\tau^{-1}(R_\tau(A_\tau)).$$

Let us define

$$\widetilde{\boldsymbol{\theta}} \triangleq \begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{y}_1^T \boldsymbol{\vartheta} \\ \vdots \\ \boldsymbol{y}_{|\mathcal{A}|}^T \boldsymbol{\vartheta} \end{bmatrix}.$$

By a similar deviation as [2] we have that with probability at least $1 - 1/T$, the following holds:

$$\begin{aligned} &|\widetilde{\boldsymbol{x}}_a^T \boldsymbol{\theta}_{t-1} - \widetilde{\boldsymbol{x}}_a^T \widetilde{\boldsymbol{\theta}}| \\ &\le (\xi_a \sqrt{(d + |\mathcal{A}|)\log(T + T^2(L+1))} + \|\widetilde{\boldsymbol{\theta}}\|) \|\widetilde{\boldsymbol{x}}_a\|_{\boldsymbol{V}_{t-1}}. \end{aligned}$$

Then we have

$$\begin{aligned} &\mathbb{P}[\exists t, a, \bar{R}_t(a, \boldsymbol{\Psi}) \notin [L_t(A_t(\boldsymbol{\Psi}_t); \mathcal{H}_{t-1}), U_t(A_t(\boldsymbol{\Psi}_t); \mathcal{H}_{t-1})] | \boldsymbol{\Psi}] \\ &\le \frac{1}{T}. \end{aligned}$$

Then by a similar deviation as [2], we have

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} \min\{\Delta(\mathcal{R}), \zeta_t [U_t(A_t(\boldsymbol{\Psi}_t); \mathcal{H}_{t-1}) - L_t(A_t(\boldsymbol{\Psi}_t); \mathcal{H}_{t-1})]\}\right] \\ &\le \left[2 \max_{\tau \le T} \zeta_\tau (\xi_{\max} \sqrt{(d + |\mathcal{A}|)\log(T + T^2(L+1))}) \right. \\ &\quad \left. + \mathbb{E}_{\boldsymbol{\Psi} \sim p(\boldsymbol{\Psi})}[\|\widetilde{\boldsymbol{\theta}}\|] + \Delta(\mathcal{R})\right] \\ &\quad \sqrt{2T(d + |\mathcal{A}|)\log\left(1 + \frac{T(L+1)}{d + |\mathcal{A}|}\right)}. \end{aligned}$$

Note that $\|\widetilde{\boldsymbol{\theta}}\| = \sqrt{\sum_{i=1}^{d} \theta^2 + \sum_{a \in \mathcal{A}} (\boldsymbol{y}_a^T \boldsymbol{\vartheta})^2}$. This proof is then complete. ∎

### *C.* Proof of Theorem 2

It suffices to show that there is an instance of our model who has a regret lower bound of $\Omega(\sqrt{T|\mathcal{A}|})$. Consider a special case of the model with $d = 0$, $\ell = 1$ and $g_t(V(A_t)) = V(A_t)$. Furthermore, consider $\mathcal{A}_t = \mathcal{A}$. Then the model reduces to the classical multi-armed bandit setting with $\mathcal{A}$ arms. It is a well known results that there is an instance of the multi-armed bandit with $\mathcal{A}$ arms such that the regret lower bound is $\Omega(\sqrt{T|\mathcal{A}|})$. Consider that the prior distribution concentrates on this instance with probability one, then we have that the Bayesian for this special case is $\Omega(\sqrt{T|\mathcal{A}|})$. This proof is then complete. ∎

### *D.* Proof of Theorem 3

The proof of this theorem by applying a result in [3]. This only involves checking the conditions of Lemma 10.11. ∎

### *E.* Proof of Theorem 4

To make the presentation clear, let $\boldsymbol{\Phi}$ denote a sample of the unknown model parameters which follows the distribution of $p_t^{(N)}(\cdot)$ (i.e., the landing probability of the MCMC in the GCL-PSMC algorithm). In fact, the action $A_t$ of the GCL-PSMC algorithm is determined by $\boldsymbol{\Phi}$. To make the presentation clear, we write $A_t$ as $A_t(\boldsymbol{\Phi})$ in the following derivation. Let $U_t(a; \mathcal{H}_{t-1})$ and $L_t(a; \mathcal{H}_{t-1})$ denote an upper and lower confidence bound of $\bar{R}_t(a; \boldsymbol{\Psi})$ with the decision history $\mathcal{H}_{t-1}$ constructed in the proof of Theorem 1. We next derive a lower bound of $\mathbb{E}[U_t(A_t(\boldsymbol{\Phi}); \mathcal{H}_{t-1}]$. First, via conditioning we have

$$\begin{aligned} &\mathbb{E}\left[\bar{R}_t(A_t^*(\boldsymbol{\Psi}); \boldsymbol{\Psi}) - \bar{R}_t(A_t(\boldsymbol{\Phi}); \boldsymbol{\Psi}) | \mathcal{H}_{t-1}\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\Psi} \sim p(\cdot|\mathcal{H}_{t-1}), \boldsymbol{\Phi} \sim p_t^{(N)}(\cdot)} \left[\bar{R}_t(A_t^*(\boldsymbol{\Psi}); \boldsymbol{\Psi}) - \bar{R}_t(A_t(\boldsymbol{\Phi}); \boldsymbol{\Psi})\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\Psi} \sim p(\cdot|\mathcal{H}_{t-1})} \left[\bar{R}_t(A_t^*(\boldsymbol{\Psi}); \boldsymbol{\Psi}) - \mathop{\mathbb{E}}_{\boldsymbol{\Phi} \sim p_t^{(N)}(\cdot)} [\bar{R}_t(A_t(\boldsymbol{\Phi}); \boldsymbol{\Psi})]\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\Psi} \sim p(\cdot|\mathcal{H}_{t-1})} \left[\bar{R}_t(A_t^*(\boldsymbol{\Psi}); \boldsymbol{\Psi}) - \mathop{\mathbb{E}}_{\boldsymbol{\Phi}' \sim p(\cdot|\mathcal{H}_{t-1})} [\bar{R}_t(A_t(\boldsymbol{\Phi}'); \boldsymbol{\Psi})]\right] \\ &+ \mathop{\mathbb{E}}_{\boldsymbol{\Psi} \sim p(\cdot|\mathcal{H}_{t-1})} \left[\mathop{\mathbb{E}}_{\boldsymbol{\Phi}' \sim p(\cdot|\mathcal{H}_{t-1})} [\bar{R}_t(A_t(\boldsymbol{\Phi}'); \boldsymbol{\Psi})] \right. \\ &\left. - \mathop{\mathbb{E}}_{\boldsymbol{\Phi} \sim p_t^{(N)}(\cdot)} [\bar{R}_t(A_t(\boldsymbol{\Phi}); \boldsymbol{\Psi})]\right] \\ &\le \mathop{\mathbb{E}}_{\boldsymbol{\Psi} \sim p(\cdot|\mathcal{H}_{t-1}), \boldsymbol{\Phi}' \sim p(\cdot|\mathcal{H}_{t-1})} \left[\bar{R}_t(A_t^*(\boldsymbol{\Psi}); \boldsymbol{\Psi}) - \bar{R}_t(A_t(\boldsymbol{\Phi}'); \boldsymbol{\Psi})\right] \\ &\quad + 2(\max_{r \in \mathcal{R}} |r|) \|p_t^{(N)}(\cdot) - p(\cdot|\mathcal{H}_{t-1})\|_{TV}. \end{aligned}$$

Then with a similar proof as Theorem 1, we have that

$$\begin{aligned} &R_T^{Bay}(\mathcal{D}_{GCL-PSMC}) \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\bar{R}_t(A_t^*(\boldsymbol{\Psi}); \boldsymbol{\Psi}) - \bar{R}_t(A_t(\boldsymbol{\Phi}); \boldsymbol{\Psi}) | \mathcal{H}_{t-1}\right]\right] \\ &\le \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\Psi} \sim p(\cdot|\mathcal{H}_{t-1}), \boldsymbol{\Phi}' \sim p(\cdot|\mathcal{H}_{t-1})} \left[\bar{R}_t(A_t^*(\boldsymbol{\Psi}); \boldsymbol{\Psi}) \right.\right. \\ &\left.\left. - \bar{R}_t(A_t(\boldsymbol{\Phi}'); \boldsymbol{\Psi})\right]\right] \\ &+ \mathbb{E}\left[\sum_{t=1}^{T} 2(\max_{r \in \mathcal{R}} |r|) \|p_t^{(N)}(\cdot) - p(\cdot|\mathcal{H}_t)\|_{TV}\right] \\ &\le R_T^{Bay}(\mathcal{D}_{GCL-PS}) + \mathbb{E}\left[\sum_{t=1}^{T} 2(\max_{r \in \mathcal{R}} |r|) \frac{\eta}{\sqrt{t}}\right] \\ &\le R_T^{Bay}(\mathcal{D}_{GCL-PS}) + 2(\max_{r \in \mathcal{R}} |r|) \sqrt{T} \eta. \end{aligned}$$

This proof is then complete. ∎

### *F.* Proof of Lemma 2

We prove this lemma by induction. When $t = 1$, it corresponds to sampling from the prior distribution. Thus, Lemma 2 trivially holds. Suppose Lemma 2 with $t$:

$$\boldsymbol{\Lambda}_{a,t}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma}) = \left(\boldsymbol{\Lambda}_a^{-1} + \frac{n_{a,t-1}}{\sigma_a^2}\boldsymbol{\vartheta}\boldsymbol{\vartheta}^T\right)^{-1},$$

$$\boldsymbol{\nu}_{a,t}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma}) = \boldsymbol{\Lambda}_{a,t}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma})\Bigg(\boldsymbol{\Lambda}_a^{-1}\boldsymbol{\nu}_a + \boldsymbol{\vartheta}\frac{1}{\sigma_a^2}\big(\sum_{\tau=1}^{t-1}\mathbb{1}_{\{A_\tau=a\}}g_\tau^{-1}(R_\tau(A_\tau)) - n_{a,t-1}\boldsymbol{\theta}^T\boldsymbol{x}_a\big)\Bigg).$$

Based on this, we next prove by induction that it also holds with $t+1$:

$$\begin{aligned}\boldsymbol{\Lambda}_{a,t+1}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma}) &= \left(\boldsymbol{\Lambda}_{a,t}^{-1}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma}) + \frac{1}{\sigma_a^2}\boldsymbol{\vartheta}\boldsymbol{\vartheta}^T\right)^{-1} \\ &= \left(\boldsymbol{\Lambda}_a^{-1} + \frac{n_{a,t-1}}{\sigma_a^2}\boldsymbol{\vartheta}\boldsymbol{\vartheta}^T + \frac{1}{\sigma_a^2}\boldsymbol{\vartheta}\boldsymbol{\vartheta}^T\right)^{-1} \\ &= \left(\boldsymbol{\Lambda}_a^{-1} + \frac{n_{a,(t+1)-1}}{\sigma_a^2}\boldsymbol{\vartheta}\boldsymbol{\vartheta}^T\right)^{-1}\end{aligned}$$

Furthermore, we have

$$\begin{aligned}&\boldsymbol{\nu}_{a,t+1}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma}) = \boldsymbol{\Lambda}_{a,t+1}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma})\Bigg(\boldsymbol{\Lambda}_{a,t}^{-1}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma})\boldsymbol{\nu}_{a,t}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma}) \\ &\quad + \boldsymbol{\vartheta}\frac{1}{\sigma_a^2}g_t^{-1}(R_t(a) - \boldsymbol{\theta}^T\boldsymbol{x}_a)\Bigg) \\ &= \boldsymbol{\Lambda}_{a,t+1}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma})\Bigg(\boldsymbol{\Lambda}_a^{-1}\boldsymbol{\nu}_a \\ &\quad + \boldsymbol{\vartheta}\frac{1}{\sigma_a^2}\big(\sum_{\tau=1}^{t-1}\mathbb{1}_{\{A_\tau=a\}}g_\tau^{-1}(R_\tau(A_\tau)) - n_{a,t-1}\boldsymbol{\theta}^T\boldsymbol{x}_a\big) \\ &\quad + \boldsymbol{\vartheta}\frac{1}{\sigma_a^2}g_t^{-1}(R_t(a) - \boldsymbol{\theta}^T\boldsymbol{x}_a)\Bigg) \\ &= \boldsymbol{\Lambda}_{a,t+1}(\boldsymbol{\theta},\boldsymbol{\vartheta},\boldsymbol{\sigma})\Bigg(\boldsymbol{\Lambda}_a^{-1}\boldsymbol{\nu}_a \\ &+ \boldsymbol{\vartheta}\frac{1}{\sigma_a^2}\big(\sum_{\tau=1}^{(t+1)-1}\mathbb{1}_{\{A_\tau=a\}}g_\tau^{-1}(R_\tau(A_\tau)) - n_{a,(t+1)-1}\boldsymbol{\theta}^T\boldsymbol{x}_a\big)\Bigg).\end{aligned}$$

Thus, the first part of Lemma 2 holds. Similarly, we can prove that the second part also holds:

$$\boldsymbol{\Sigma}_t(\boldsymbol{Y},\boldsymbol{\sigma}) = \left(\boldsymbol{\Sigma}^{-1} + \sum_{a\in\mathcal{A}}\frac{n_{a,t-1}}{\sigma_a^2}[\boldsymbol{x}_a^T,\boldsymbol{y}_a^T]^T[\boldsymbol{x}_a^T,\boldsymbol{y}_a^T]\right)^{-1},$$

$$\boldsymbol{\mu}_t(\boldsymbol{Y},\boldsymbol{\sigma}) = \boldsymbol{\Sigma}_t(\boldsymbol{Y},\boldsymbol{\sigma})\Bigg(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \sum_{a\in\mathcal{A}}[\boldsymbol{x}_a^T,\boldsymbol{y}_a^T]^T\frac{1}{\sigma_a^2}\sum_{\tau=1}^{t-1}\mathbb{1}_{\{A_\tau=a\}}g_\tau^{-1}(R_\tau(A_\tau))\Bigg).$$

The last part of Lemma 2 is a simple consequence of the Inverse Gamma distribution. This proof is then complete. ∎

## References

[1] D. Russo and B. Van Roy, “Learning to optimize via posterior sampling,” *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.
[2] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.
[3] C. Robert and G. Casella, *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.