
PAC-Bayesian Analysis of Contextual Bandits

Supplementary Material

Yevgeny Seldin^{1,4} Peter Auer² François Laviolette³ John Shawe-Taylor⁴ Ronald Ortner²

¹Max Planck Institute for Intelligent Systems, Tübingen, Germany

²Chair for Information Technology, Montanuniversität Leoben, Austria

³Département d'informatique, Université Laval, Québec, Canada

⁴Department of Computer Science, University College London, UK

seldin@tuebingen.mpg.de, {auer, ronald.ortner}@unileoben.ac.at,
francois.laviolette@ift.ulaval.ca, jst@cs.ucl.ac.uk

Abstract

This document provides supplementary material to the paper “PAC-Bayesian Analysis of Contextual Bandits”. It contains proofs of Lemmas 1, 2, and 3 from the paper and some technical details on the experiment.

1 Proof of Lemma 1

Proof. We have

$$\hat{\Delta}(\rho_t^{exp}) = \sum_s p(s) \sum_a \rho_t^{exp}(a|s) \hat{\Delta}_t(a, s).$$

The inner sum accepts the form

$$\sum_a \rho_t^{exp}(a|s) \hat{\Delta}_t(a, s) = \frac{\sum_a \hat{\Delta}_t(a, s) \tilde{\rho}_t(a) e^{\gamma_t \hat{R}_t(a, s)}}{\sum_a \tilde{\rho}_t(a) e^{\gamma_t \hat{R}_t(a, s)}} = \frac{\sum_a \hat{\Delta}_t(a, s) \tilde{\rho}_t(a) e^{-\gamma_t \hat{\Delta}_t(a, s)}}{\sum_a \tilde{\rho}_t(a) e^{-\gamma_t \hat{\Delta}_t(a, s)}},$$

where the second equality is by multiplication of nominator and denominator by $e^{-\gamma_t \hat{R}_t(a^*(s), s)}$.

The lemma follows from Lemma 6 below and the observation that $\hat{\Delta}_t(a^*(s), s) = 0$ for all s . \square

Lemma 6. *Let $x_1 = 0$ and x_2, \dots, x_n be $(n - 1)$ arbitrary numbers. Let $p(x_i)$ be a distribution over x_i -s, such that $p(x_1) = p > 0$. For any $\alpha > 0$ and $n \geq 2$:*

$$\frac{\sum_{i=1}^n p(x_i) x_i e^{-\alpha x_i}}{\sum_{i=1}^n p(x_i) e^{-\alpha x_i}} \leq \frac{1}{\alpha} \ln \frac{1}{p}.$$

Proof. By symmetry, the maximum is achieved when all x_i -s (except x_1) are equal. Let x be the common value of x_i -s. Then:

$$\frac{\sum_{i=1}^n p(x_i) x_i e^{-\alpha x_i}}{\sum_{i=1}^n p(x_i) e^{-\alpha x_i}} = \frac{(1 - p) x e^{-\alpha x}}{p + (1 - p) e^{-\alpha x}}.$$

The lemma then follows from Lemma 7. \square

Lemma 7. *For any $x \geq 0$, $0 < p \leq 1$, and $\alpha > 0$:*

$$\frac{(1 - p) x e^{-\alpha x}}{p + (1 - p) e^{-\alpha x}} \leq \frac{1}{\alpha} \ln \frac{1}{p}.$$

Proof. We apply change of variables $y = e^{-\alpha x}$. Then $x = \frac{1}{\alpha} \ln \frac{1}{y}$. By substitution:

$$\frac{(1-p)xe^{-\alpha x}}{p + (1-p)e^{-\alpha x}} = \frac{1}{\alpha} \cdot \frac{(1-p)y \ln \frac{1}{y}}{p + (1-p)y} \leq \frac{1}{\alpha} \ln \frac{1}{p},$$

where the last inequality is by Lemma 8. □

Lemma 8. For any positive y and $0 < p \leq 1$:

$$\frac{(1-p)y \ln \frac{1}{y}}{p + (1-p)y} \leq \ln \frac{1}{p}$$

Proof. By taking Taylor's expansion of $\ln z$ around $z = \frac{1}{p}$ we have:

$$\ln z \leq \ln \frac{1}{p} + p(z - \frac{1}{p}) = \ln \frac{1}{p} + pz - 1.$$

Thus:

$$\begin{aligned} \frac{(1-p)y \ln \frac{1}{y}}{p + (1-p)y} &= \frac{\frac{1-p}{p}y \ln \frac{1}{y}}{1 + \frac{1-p}{p}y} \\ &\leq \frac{\frac{1-p}{p}y(\ln \frac{1}{p} + \frac{p}{y} - 1)}{1 + \frac{1-p}{p}y} \\ &\leq \frac{\frac{1-p}{p}y \ln \frac{1}{p} + (1-p)}{1 + \frac{1-p}{p}y} \\ &\leq \frac{(\frac{1-p}{p}y + 1) \ln \frac{1}{p}}{\frac{1-p}{p}y + 1} \\ &= \ln \frac{1}{p}, \end{aligned}$$

where the last inequality follows from the fact that $1 - p \leq \ln \frac{1}{p}$. □

2 Proof of Lemma 2

Proof.

$$\begin{aligned} R(\rho) - R(\tilde{\rho}) &= \sum_s p(s) \sum_a (\rho(a|s) - \tilde{\rho}(a|s)) R(a, s) \\ &\leq \frac{1}{2} \sum_s p(s) \sum_a |\rho(a|s) - \tilde{\rho}(a|s)| \\ &= \frac{1}{2} \sum_s p(s) \sum_a |\rho(a|s) - (1 - K\varepsilon)\rho(a|s) - \varepsilon| \\ &= \frac{1}{2} \sum_s p(s) \sum_a |K\varepsilon\rho(a|s) - \varepsilon| \\ &\leq \frac{1}{2} K\varepsilon \sum_s p(s) \sum_a \rho(a|s) + \frac{1}{2} K\varepsilon \\ &= K\varepsilon. \end{aligned} \tag{1}$$

In (1) we used the fact that $0 \leq R(a, s) \leq 1$ and ρ and $\tilde{\rho}$ are probability distributions. □

3 Proof of Lemma 3

Proof.

$$\begin{aligned} V_t(a) &= \sum_{\tau=1}^t \mathbb{E}[(R_{\tau}^{h^*(S_{\tau}), S_{\tau}} - R_{\tau}^{h(S_{\tau}), S_{\tau}}] - [R(h^*) - R(h)]^2 | \mathcal{T}_{\tau-1}] \\ &= \left(\sum_{\tau=1}^t \mathbb{E}[(R_{\tau}^{h^*(S_{\tau}), S_{\tau}} - R_{\tau}^{h(S_{\tau}), S_{\tau}})^2 | \mathcal{T}_{\tau-1}] \right) - t\Delta(h)^2 \end{aligned} \quad (2)$$

$$\leq \left(\sum_{\tau=1}^t \left(\frac{\pi_{\tau}(h(S_{\tau})|S_{\tau})}{\pi_{\tau}(h(S_{\tau})|S_{\tau})^2} + \frac{\pi_{\tau}(h^*(S_{\tau})|S_{\tau})}{\pi_{\tau}(h^*(S_{\tau})|S_{\tau})^2} \right) \right) \quad (3)$$

$$\begin{aligned} &= \left(\sum_{\tau=1}^t \left(\frac{1}{\pi_{\tau}(h(S_{\tau})|S_{\tau})} + \frac{1}{\pi_{\tau}(h^*(S_{\tau})|S_{\tau})} \right) \right) \\ &\leq \frac{2t}{\varepsilon_t}, \end{aligned} \quad (4)$$

where (2) is due to the fact that $\mathbb{E}[R_{\tau}^{h(S_{\tau}), S_{\tau}} | \mathcal{T}_{\tau-1}] = R(h(S_{\tau}), S_{\tau})$, (3) is due to the fact that $R_t \leq 1$, and (4) is due to the fact that $\frac{1}{\pi_{\tau}(a|S_t)} \leq \frac{1}{\varepsilon_t}$ for all a and $1 \leq \tau \leq t$. \square

4 Experiment Details

We note that precise calculation of the mutual information $I_{\rho_t^{exp}}(S; A)$ requires calculation of the marginal distribution over actions corresponding to ρ_t^{exp} , which would require iteration through all the states and take $O(NK)$ computation time per round. The reason is that the learning rate γ_t changes at each iteration and, hence, $\rho_t^{exp}(a, s)$ changes at each iteration for all a and s . However, for the prediction we only need to know $\rho_t^{exp}(a|S_t)$ for the observed state S_t . This allows us to reduce the computation time of the algorithm to $O(K)$ operations per round. For the mutual information $I_{\rho_t^{exp}}(S; A)$ we used the running average approximation:

$$I_{\rho_t^{exp}}(S; A) = \frac{N-1}{N} I_{\rho_{t-1}^{exp}}(S; A) + \frac{1}{N} KL(\rho_t^{exp}(a|S_t) \| \tilde{\rho}_t^{exp}(a)),$$

where KL is calculated only for the observed state S_t and, therefore, the computation time is $O(K)$ operations per round. We note that since $\tilde{\rho}_t^{exp}(a)$ is not a precise marginal distribution of $\frac{1}{N} \tilde{\rho}_t^{exp}(a|s)$, the above estimate on average upper bounds the true mutual information, but, of course, is not completely precise.

Regarding the parameters of the algorithm: we took $\varepsilon_t = (Kt)^{-1/3}$, as suggested by our analysis.

In order to make the contribution of the second term in the regret decomposition comparable to the first term we should have taken

$$\begin{aligned} \gamma_t &= \frac{\ln \frac{1}{\varepsilon_{t+1}}}{1 + c_t} \sqrt{\frac{t\varepsilon_t}{2(e-2)(NI_{\rho_{t-1}^{exp}}(S; A) + K(\ln N + \ln K) + 2\ln(t+1) + \ln \frac{2m_t}{\delta})}} \\ &\leq \frac{\ln \frac{1}{\varepsilon_{t+1}}}{1 + c_t} \sqrt{\frac{t\varepsilon_t}{2(e-2)(K(\ln N + \ln K) + 2\ln(t+1) + \ln \frac{2m_t}{\delta})}}. \end{aligned}$$

However, empirically we found that it is better to set

$$\gamma_t = \frac{\ln \frac{1}{\varepsilon_{t+1}}}{1 + c_t} \sqrt{\frac{t}{2(e-2)(K(\ln N + \ln K) + 2\ln(t+1) + \ln \frac{2m_t}{\delta})}},$$

which was inspired by the tighter bound on the cumulative variance, $V_t(\rho_t^{exp}) \leq 2Kt$, which we believe to be true, but did not prove yet.