PAC-Bayesian Analysis of Counterfactual Risk in Stochastic Contextual Bandits

Junhao Wang
McGill University / Mila
junhao.wang@mail.mcgill.ca

Bogdan Mazoure
McGill University / Mila
bogdan.mazoure@mail.mcgill.ca

Gavin McCracken
McGill University / Mila
gavin.mccracken@mail.mcgill.ca

David Venuto
McGill University / Mila
david.venuto@mail.mcgill.ca

Audrey Durand
McGill University / Mila
audrey.durand@mcgill.ca

Abstract

This work tackles the off-policy evaluation problem within the contextual bandit setting, where only the action and reward recommended by the logging policy were recorded and thus available at evaluation. This kind of situation is encountered in applications where one wants to compute the optimal policy using data previously collected in an offline manner. Previous work have extended the PAC-Bayesian analysis to this setting, providing bounds on the clipped importance sampling risk estimator using a recent regularization technique known as *counterfactual risk minimization*. The contribution of this work is to tighten this existing result through the application of various PAC-Bayesian concentration inequalities: Kullback-Leibler divergence, Bernstein, and Azuma-Hœffding. This yields bounds on the empirical risk estimator that either converge at a faster rate given the amount of prior data, or that are more robust to the clipping factor.

Keywords: contextual bandits, PAC-Bayes

Acknowledgements

The authors would like to thank Vincent Luczkow and Nadeem Ward for insightful discussions.

1 Introduction

In many applications of interactive learning, one wants to leverage previously collected data in an offline manner in order to compute the optimal policy at a later stage. For instance, this is the case of recommender systems, where data may have been gathered previously under some recommendation policy (e.g. best top results) and one wants to use this data to evaluate an *alternative* recommendation policy. This is known as *off-policy* or *offline evaluation*. In this work, we will tackle the off-policy evaluation problem within the contextual bandit setting, where only the action and reward recommended by the logging policy were recorded and thus available at evaluation (Li et al., 2011).

The PAC-Bayesian analysis (Shawe-Taylor and Williamson, 1997; Shawe-Taylor et al., 1998; McAllester, 1999) has been dominantly focusing on studying supervised setting of statistical learning, where data is assumed to be independently and identically distributed (i.i.d), within the PAC (Probably Approximately Correct) learning framework (Valiant, 1984). Such analysis highlights the trade-off between the complexity of individual models from the hypothesis space and their empirical performance, with high probability guarantees on their expected performance. Seldin et al. (2011, 2012) have extended the framework to non-i.i.d setting such as bandits and reinforcement learning. Additionally, London and Sandler (2018) applied PAC-Bayesian analysis on a recent regularization technique known as *counterfactual risk minimization* (Swaminathan and Joachims, 2015) for off-policy evaluation on stochastic contextual bandits. The contribution of this work is to tighten this existing result through the application of various PAC-Bayesian concentration inequalities.

2 Contextual bandits

The stochastic contextual bandit (Langford and Zhang, 2008) is described by an arbitrary context space $\mathcal{X}$, an action space $\mathcal{K} = \{1, \dots, K\}$, and a distribution $\mathcal{D}$ over tuples (x, ρ), with $x \in \mathcal{X}$ and $\rho : \mathcal{X} \times \mathcal{K} \mapsto \mathcal{Y}$. Without loss of generality, we will assume in the following that $\mathcal{Y} = [0, 1]$. The problem can then be formulated as an episodic game where on each episode $t \in \mathbb{N}_{>0}$:

1. a context and reward function $(x_t, \rho_t \sim \mathcal{D})$ are generated from the environment;
2. the learner observes the context $x_t \in \mathcal{X}$ but does not observe the function ρ_t;
3. the learner selects an action $k_t \in \mathcal{K}$;
4. the learner observes the reward $y_t = \rho_t(x_t, k_t)$
5. the learner updates its knowledge based on this experience.

Assuming that $\mathcal{K}$ is finite, the goal of the learner is to learn a policy $\pi : \mathcal{X} \mapsto \Delta(\mathcal{K})$ for choosing actions over the probability simplex according to the context such that to maximize the expected reward $G(\pi) = \mathbb{E}_{(x_t,\rho_t)\sim\mathcal{D}}\mathbb{E}_{k_t\sim\pi(x_t)}\rho_t(x_t, k_t)$. This is equivalent to minimizing the *counterfactual risk*: $R(\pi) = 1 - G(\pi)$.

3 Off-policy evaluation

The task of off-policy policy evaluation consists in estimating either the true expected reward $G(\pi)$ or the true counterfactual risk $R(\pi)$ of an arbitrary policy π based on a n-length history $\mathcal{F}_n^{\pi_0} = \{(x_1, k_1, y_1), \dots, (x_n, k_n, y_n)\}$ generated by some policy π_0. This is often referred to as *learning from logged bandit feedback* (Li et al., 2011, 2012; Mary et al., 2014). For simplicity, we will only focus on analyzing the counterfactual risk R due to its similarity to expected reward G.

Assumption 1 (Time-invariance of π_0)**.** We assume that the logging policy π_0 is stationary, such that for every timestep $i \leqslant n$, the corresponding action $k_i \in \mathcal{F}_n^{\pi_0}$ has been sampled from the initial policy π_0.

A challenge in off-policy evaluation is to derive consistent estimators $\hat{R}(\pi, \mathcal{F}_n^{\pi_0})$ for some policy $\pi \neq \pi_0$, with low bias with regard to true counterfactual $R(\pi)$ and low variance with regard to the any history $\mathcal{F}_n^{\pi}$ for any policy π. Main methods for creating such estimators are direct modelling (Hassanpour and Greiner, 2018), importance sampling (Kearns et al., 2000; Precup, 2000), and doubly robust (Dudík et al., 2011), which combines the two previous to produce an estimator with lower variance and bias. This work focuses on the importance sampling (IS) counterfactual risk estimator

$$\hat{R}_{\mathrm{IS}}(\pi, \mathcal{F}_n^{\pi_0}) = 1 - \mathbb{E}_{(x_i,\rho_i)\sim\mathcal{D}}\left[\mathbb{E}_{k_i\sim\pi(x_i)}\left[\frac{\rho_i(x_i, k_i)}{\pi_0(x_i, k_i)}\right]\right] \approx 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(x_i, k_i)}{\pi_0(x_i, k_i)}\rho_i(x_i, k_i), \tag{1}$$

More specifically, we consider one important variant, that is the clipped importance sampling with counterfactual risk minimization objective (Hassanpour and Greiner, 2018).

Clipped importance sampling The essence of this approach is to set a lower bound on the propensity score $\pi_0(x_i, k_i)$, resulting in a clipped importance sampling risk estimator (CIS):

$$\hat{R}_{\mathrm{CIS}}(\pi, \mathcal{F}_n^{\pi_0}) = 1 - \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{\pi(x_i, k_i)}{\max\{\pi_0(x_i, k_i), p_{\min}\}}\rho_i(x_i, k_i)}_{y_i^{\mathrm{CIS}}}. \tag{2}$$

Clipping by $p_{\min}$ trades off variance for bias in the estimator. Empirical variance regularizer for the clipped importance sampling weighted reward y_i^{CIS} can be applied to lead to faster shrinking of the difference between true counterfactual risk and empirical counterfactual risk, in comparison to without such regularizer. The variance penalty term, known as counterfactual risk minimization (CRM), is based on the following generalization error bound.

Theorem 3.1 (Counterfactual Risk Minimization (Swaminathan and Joachims, 2015))**.** *Let Π denote the space of policies.*

$$R(\pi) \leqslant \hat{R}_{\mathrm{CIS}}(\pi, \mathcal{F}_n^{\pi_0}) + O\left(\sqrt{\frac{\hat{\mathbb{V}}[\hat{R}_{\mathrm{CIS}}(\pi, \mathcal{F}_n^{\pi_0})] + \mathbb{C}(\Pi)}{n}} + \frac{\mathbb{C}(\Pi)}{n}\right), \tag{3}$$

where $\mathbb{C}(\Pi) \propto \mathcal{N}_\infty(\varepsilon, \Pi)$ measures the cardinality of the minimal $\varepsilon-$covering of Π, and $\hat{\mathbb{V}}[\hat{R}_{\mathrm{CIS}}(\pi, \mathcal{F}_n^{\pi_0})]$ is the unbiased sample variance of the CIS risk estimator.

4 PAC-Bayesian Counterfactual Risk Minimization

London and Sandler (2018) provide a Bayesian perspective of CRM by applying PAC-Bayesian analysis (McAllester, 1999) on contextual bandits, in a manner similar to Seldin et al. (2011). They achieve the following CRM bound, which depends on prior distribution $\mathbb{P}$ and posterior distribution $\mathbb{Q}$ over known deterministic hypothesis space $\mathcal{H} : \mathcal{X} \to \mathcal{K}$ such that $\pi_{\mathbb{Q}}(x, k) = \mathbb{E}_{h\sim\mathbb{Q}}[\mathbb{I}\{h(x) = k\}]$ corresponds to the posterior probability that a random hypothesis h maps an action k to context x. From the PAC-Bayesian perspective, the learner, which can be seen as a Gibbs classifier, samples h from $\mathbb{Q}(\mathcal{H})$ and selects action $k = h(x)$. In the off-policy evaluation setting considered in this work, the logging policy π_0 induced by $\mathbb{P}$ generated the history $\mathcal{F}_n^{\pi_0}$, and the goal consists in estimating $R(\pi)$ using $\hat{R}_{\mathrm{CIS}}(\pi, \mathcal{F}_n^{\pi_0})$.

Theorem 4.1 (PAC-Bayesian Counterfactual Risk Minimization (London and Sandler, 2018))**.** *Let $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{K}\}$ denote a hypothesis space mapping contexts to actions and let $\mathrm{KL}(\mathbb{Q}||\mathbb{P})$ denote the Kullback-Leibler divergence between (absolutely continuous) probability measures $\mathbb{Q}, \mathbb{P}$ over the set $\mathcal{H}$. In particular, let $\mathbb{P}$ and $\mathbb{Q}$ be the prior and posterior distributions over $\mathcal{H}$, respectively. For any $n \geqslant 1, \delta \in (0, 1), p_{\min} \in (0, 1)$, with probability at least $1 - \delta$ over $\mathcal{F}_n^{\mathbb{P}}$, which is history generated by π_0 induced by $\mathbb{P}$, the following holds simultaneously for all $\mathbb{Q}$ and its corresponding induced $\pi_{\mathbb{Q}}$:*

$$R(\pi_{\mathbb{Q}}) \leqslant \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}}) + \sqrt{\frac{2(\frac{1}{p_{\min}} - 1 + \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}}))(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n}{\delta})}{p_{\min}(n-1)}} + \frac{2(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n}{\delta})}{p_{\min}(n-1)} \tag{4}$$

Note 1. When $\hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}}) = 1 - \frac{1}{p_{\min}}$, the generalization bound yields $O(\frac{1}{n})$ converging rate. In particular, minimizing $\hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\pi_{\mathbb{P}}})$ and keeping policy $\pi_{\mathbb{Q}}$ close to logging policy $\pi_{\mathbb{P}}$ through the KL minimizes $R(\pi_{\mathbb{Q}})$.

This result is obtained using the PAC-Bayesian-Hœffding inequality (McAllester, 2003). The following proposed result tightens the bound using various concentration inequalities.

4.1 Proposed result

By applying PAC-Bayes-KL inequality (Seeger, 2002), PAC-Bayes-Berstein inequality (Tolstikhin and Seldin, 2013), and PAC-Bayes-Azuma-Hœffding inequality (Seldin et al., 2012) (Theorems A.1,A.3,and A.2) on PAC-Bayesian CRM (Theorem 4.1), we obtain the following result.

Theorem 4.2 (PAC-Bayesian Counterfactual Risk Minimization Extensions)**.** *Let $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{K}\}$ denote a hypothesis space mapping contexts to actions. For any $n \geqslant 1, \delta \in (0, 1), p_{\min} \in (0, 1)$ and fixed prior, $\mathbb{P}$ on $\mathcal{H}$ and its corresponding induced* *policy π_0, with probability at least $1-\delta$ over $\mathcal{F}_n^{\mathbb{P}}$, the following bounds holds simultaneously for all $\mathbb{Q}$ with induced policy $\pi_{\mathbb{Q}}$:*

$$\textit{(KL)} \qquad R(\pi_{\mathbb{Q}}) \leqslant_{1-\delta} \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}}) + \frac{1}{p_{\min}} \sqrt{\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{n+1}{\delta}}{2n}} \tag{5}$$

$$\textit{(Bernstein)} \qquad R(\pi_{\mathbb{Q}}) \leqslant_{1-\delta} \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}}) + O\Big(\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{1}{\delta}}{n}\Big) \qquad \textit{if } \mathbb{Q} \textit{ satisfies Eq.17} \tag{6}$$

$$R(\pi_{\mathbb{Q}}) \leqslant_{1-\delta} \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}}) + O\Big(\sqrt{\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{1}{\delta}}{np_{\min}}}\Big) \qquad \textit{otherwise} \tag{7}$$

$$\textit{(Azuma-Hœffding)} \qquad R(\pi_{\mathbb{Q}}) \leqslant_{1-\delta} \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}}) + O\Big(\sqrt{\frac{p_{\min}^2 \mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{2}{\delta}}{n}}\Big). \tag{8}$$

4.2 Outline of proof

The complete proof is provided in Appendix B. Let $\mathcal{D}$ and $\mathcal{D}_n$ respectively denote the true and empirical joint distributions of context x and reward function ρ. The idea consists in constructing empirical counterfactual risk functions $r_{\mathcal{D}}$ and $r_{\mathcal{D}_n}$ such that distributions $\mathbb{P}$, $\mathbb{Q}$ over deterministic hypothesis space $\mathcal{H}$ can be applied to them:

$$r_{\mathcal{D}}(h) = \Big\langle \mathcal{D}, 1 - \mathbb{E}_{(x_i,\rho_i)\sim\mathcal{D}} \mathbb{E}_{k_i\sim\pi_{\mathbb{P}}(x_i)} \frac{\mathbb{I}\{h(c_i) = k_i\}\rho_i(x_i,k_i)}{\max\{p_{\min}, \pi_{\mathbb{P}}(x_i,k_i)\}} \Big\rangle \tag{9}$$

$$r_{\mathcal{D}_n}(h) = \Big\langle \mathcal{D}_n, 1 - \mathbb{E}_{(x_i,\rho_i)\sim\mathcal{D}} \mathbb{E}_{k_i\sim\pi_{\mathbb{P}}(x_i)} \frac{\mathbb{I}\{h(x_i) = k_i\}\rho_i(x_i,k_i)}{\max\{p_{\min}, \pi_{\mathbb{P}}(x_i,k_i)\}} \Big\rangle. \tag{10}$$

Recall that we can express the true and estimated CIS risk for $\pi_{\mathbb{Q}}$ as

$$R_{\mathrm{CIS}}(\pi_{\mathbb{Q}}) = \langle \mathbb{Q}, r_{\mathcal{D}} \rangle \qquad \text{and} \qquad \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \tau_n^{\pi_{\mathbb{P}}}) = \langle \mathbb{Q}, r_{\mathcal{D}_n} \rangle \tag{11}$$

and then use that $\dfrac{1}{\max\{p_{\min}, \pi_{\mathbb{P}}(x_i,k_i)\}} \leqslant \dfrac{1}{\pi_{\mathbb{P}}(x_i,k_i)}$ in order to obtain that $R_{\mathrm{CIS}}(\pi_{\mathbb{Q}}) \geqslant R(\pi_{\mathbb{Q}})$ and

$$R(\pi_{\mathbb{Q}}) - \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\pi_{\mathbb{P}}}) \leqslant R_{\mathrm{CIS}}(\pi_{\mathbb{Q}}) - \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\pi_{\mathbb{P}}}) = \langle \mathbb{Q}, r_{\mathcal{D}} \rangle - \langle \mathbb{Q}, r_{\mathcal{D}_n} \rangle. \tag{12}$$

Theorem 4.2 is then obtained by applying various inequalities to bound the right side of the following where $(x_i, k_i) \in \mathcal{F}_n^{\mathbb{P}}$:

$$n\langle \mathbb{Q}, r_{\mathcal{D}} - r_{\mathcal{D}_n} \rangle = \Big\langle \mathbb{Q}, \sum_{i=1}^{n} \Big[r_{\mathcal{D}} - \Big(1 - \mathbb{E}_{h\in\mathcal{H}} \frac{\mathbb{I}\{h(x_i) = k_i\}\rho(x_i,k_i)}{\max\{p_{\min}, \pi_{\mathbb{P}}(x_i,k_i)\}} \Big) \Big] \Big\rangle \tag{13}$$

and dividing by n.

4.3 Discussion

Using the PAC-Bayes-KL, PAC-Bayes-Bernstein and PAC-Bayes-Azuma-Hœffding bounds for an arbitrary martingale process allows to provide tight bounds on the true counterfactual risk of a contextual bandit. One observes that the previous result (Theorem 4.1) originally proposed by London and Sandler (2018) is $\mathcal{O}\Big(\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{n}{\delta}}{p_{\min} n}\Big)$. In comparison, the proposed result offers the following gains:

- the proposed KL-based bound (Eq. 5) tightens at a faster rate $1/\sqrt{n}$;
- the Azuma-Hœffding-based bound (Eq. 8) saves a rate $1/p_{\min}$ and tightens at a faster rate $1/\sqrt{n}$;
- the Bernstein-based bound (Eq. 6 and 7) saves a rate $1/p_{\min}$ if $\mathbb{Q}$ satisfies the condition of Eq. 17, otherwise it saves a rate $1/\sqrt{p_{\min}}$ and tightens at a faster rate $1/\sqrt{n}$.

In other words, all three bounds are more efficient than the existing result, in the sense that they either converge faster (better dependence on n) or they are less impacted by the clipping factor (better dependence on $p_{\min}$). A detailed argument outlining dominance of Eq. 5 by Theorem 4.1 under a looser condition on $\hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}})$ can be found in the appendix.

5 Conclusion

We derived three bounds on the true counterfactual risk within the contextual bandit setting based on PAC-Bayes-KL, PAC-Bayes-Bersntein and PAC-Bayes-Azuma-Hœffding inequalities, which improve upon existing results (London and Sandler, 2018). A natural line of extension would be to study the efficiency of PAC-Bayes bounds in the sequential decision making setting as applied to offline (Mandel et al., 2016) or doubly robust off-policy (Farajtabar et al., 2018) policy evaluation.

References

R. Bhatia and C. Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, 2000.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

M. Farajtabar, Y. Chow, and M. Ghavamzadeh. More robust doubly robust off-policy evaluation. *arXiv preprint arXiv:1802.03493*, 2018.

N. Hassanpour and R. Greiner. A novel evaluation methodology for assessing off-policy learning methods in contextual bandits. In *Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31*, pages 31–44. Springer, 2018.

M. J. Kearns, Y. Mansour, and A. Y. Ng. Approximate planning in large pomdps via reusable trajectories. In *Advances in Neural Information Processing Systems*, pages 1001–1007, 2000.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, 2008.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM*, 2011.

L. Li, W. Chu, J. Langford, T. Moon, and X. Wang. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, 2012.

B. London and T. Sandler. Bayesian counterfactual risk minimization. *arXiv*, abs/1806.11500, 2018.

T. Mandel, Y.-E. Liu, E. Brunskill, and Z. Popović. Offline evaluation of online reinforcement learning algorithms. In *AAAI*, 2016.

J. Mary, P. Preux, and O. Nicol. Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques. In *ICML*, 2014.

D. McAllester. Simplified pac-bayesian margin bounds. In *Learning theory and Kernel machines*, pages 203–215. Springer, 2003.

D. A. McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

D. Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

M. Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.

Y. Seldin, P. Auer, J. S. Shawe-Taylor, R. Ortner, and F. Laviolette. Pac-bayesian analysis of contextual bandits. In *NIPS*, 2011.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

J. Shawe-Taylor and R. C. Williamson. A pac analysis of a bayesian estimator. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 2–9, 1997.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.

A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, 2015.

I. O. Tolstikhin and Y. Seldin. Pac-bayes-empirical-bernstein inequality. In *NIPS*, 2013.

L. G. Valiant. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM, 1984.

A Appendix

Theorem A.1 (PAC-Bayes-KL Inequality (Seeger, 2002)). *Let $\mathcal{H}$ be a hypothesis space, let $\overline{Z}_1, \dots, \overline{Z}_n$ be a sequence of random functions, such that $\overline{Z}_i : \mathcal{H} \to [0,1]$ for $i = 1, \dots, n$. Assume $\mathbb{E}[\overline{Z}_i | \overline{Z}_1 \dots \overline{Z}_{i-1}] = \overline{b}$, where $\overline{b} : \mathcal{H} \to [0,1]$ is a deterministic function.*

Let $\overline{S}_n = \sum_{i=1}^{n} \overline{Z}_i$. Fix a prior distribution $\mathbb{P}$ over $\mathcal{H}$. Then for any $\delta \in (0,1)$, with probability greater than $1-\delta$ over $\overline{Z}_1 \dots \overline{Z}_n$, for all distributions $\mathbb{Q}$ over $\mathcal{H}$ simultaneously:

$$\mathrm{KL}\bigg(\langle \frac{1}{n}\overline{S}_n, \mathbb{Q}\rangle || \langle \overline{b}, \mathbb{Q}\rangle\bigg) \leqslant \frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{n+1}{\delta}}{n}, \tag{14}$$

which is tight if $\langle \frac{1}{n}\overline{S}_n, \mathbb{Q}\rangle$ is close to zero or one, otherwise

$$\bigg|\langle \frac{1}{n}\overline{S}_n, \mathbb{Q}\rangle - \langle \overline{b}, \mathbb{Q}\rangle\bigg| \leqslant \sqrt{\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{n+1}{\delta}}{2n}} \tag{15}$$

is tighter.

Theorem A.2 (PAC-Bayes-Azuma-Hœffding Inequality (Seldin et al., 2012)). *Let $\mathcal{H}$ be a hypothesis space, let $\overline{Z}_1, \dots, \overline{Z}_n$ be a sequence of random functions, such that $\overline{Z}_i : \mathcal{H} \to [\alpha_i, \beta_i]$ where $\alpha_i, \beta_i \in \mathbb{R}$ for $i = 1, \dots, n$ and pick $c > 1$. Let $\overline{M}_i = \sum_{j=1}^{i} \overline{Z}_j$. Fix a prior distribution $\mathbb{P}$ over $\mathcal{H}$. Then for any $\delta \in (0,1)$, with probability greater than $1-\delta$ over $\overline{Z}_1 \dots \overline{Z}_n$, for all distributions $\mathbb{Q}$ over $\mathcal{H}$ simultaneously:*

$$|\langle \mathbb{Q}, \overline{M}_n \rangle| \leqslant_{1-\delta} \frac{1+c}{2\sqrt{2}} \sqrt{\Big(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{2}{\delta} + \varepsilon(\mathbb{Q})\Big) \sum_{i=1}^{n} (\beta_i - \alpha_i)^2} \tag{16}$$

where

$$\varepsilon(\mathbb{Q}) = \frac{\ln 2}{2 \ln c}\Big(1 + \ln \Big\{\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P})}{\ln \frac{2}{\delta}}\Big\}\Big).$$

Theorem A.3 (PAC-Bayes-Bernstein Inequality (Tolstikhin and Seldin, 2013)). *Let $\mathcal{H}$ be a hypothesis space, let $\overline{Z}_1, \dots, \overline{Z}_n$ be a sequence of random functions, such that $\overline{Z}_i : \mathcal{H} \to \mathbb{R}$. Assume $\mathbb{E}[\overline{Z}_i | \overline{Z}_1 \dots \overline{Z}_{i-1}] = \overline{0}$. Thus $\forall h \in \mathcal{H}$, $\overline{Z_1}(h), \dots, \overline{Z_n}(h)$ is a martingale difference sequence. Let $\overline{M}_i = \sum_{j=1}^{i} \overline{Z}_j$ and hence $\mathbb{E}[\overline{M}_{i+1} | \overline{M}_1 \dots \overline{M}_i] = \overline{M}_i$. Then $\forall h \in \mathcal{H}$, $\overline{M}_1(h), \dots, \overline{M}_n(h)$ is a martingale. Let $\overline{V}_i : \mathcal{H} \to \mathbb{R}$ be such that $\overline{V}_i(h) = \sum_{j=1}^{i} \mathbb{E}[\overline{Z}_j(h)^2 | \overline{Z}_1(h), \dots, \overline{Z}_{j-1}(h)]$. Assume that $||\overline{Z}_i||_\infty \leqslant K \; \forall i$ with probability 1 and pick $\lambda \leqslant \frac{1}{K}$. Fix a prior distribution $\mathbb{P}$ over $\mathcal{H}$ and pick $c > 1$. Then for any $\delta \in (0,1)$, with probability greater than $1-\delta$ over $\overline{Z}_1 \dots \overline{Z}_n$, for all distributions $\mathbb{Q}$ over $\mathcal{H}$ simultaneously which satisfy:*

$$\sqrt{\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{2v}{\delta}}{(e-2)\langle \mathbb{Q}, \overline{V}_n \rangle}} \leqslant \frac{1}{K} \tag{17}$$

the following holds:

$$|\langle \mathbb{Q}, \overline{M}_n \rangle| \leqslant (1+c)\sqrt{(e-2)\langle \mathbb{Q}, \overline{V}_n \rangle (\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{2v}{\delta})} \tag{18}$$

where

$$v = \left\lceil \frac{\ln\bigg(\sqrt{\frac{(e-2)n}{\ln(\frac{2}{\delta})}}\bigg)}{\ln(c)} \right\rceil + 1,$$

and for all other $\mathbb{Q}$:

$$|\langle \mathbb{Q}, \overline{M}_n \rangle| \leqslant 2K(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln \frac{2v}{\delta}). \tag{19}$$

Theorem A.4 (Bhatia-Davis Inequality (Bhatia and Davis, 2000)). *Let $\mathbb{P}$ be a distribution with support $(m, M) \subseteq \mathbb{R}$ and $\mathbb{E}[\mathbb{P}] = \mu$. Then, the following holds:*

$$\mathbb{V}[\mathbb{P}] \leqslant (M - \mu)(\mu - m). \tag{20}$$

Tightness of PAC-Bayes-KL extension. To show that 4.1 is looser than (28) is equivalent to showing that there exists an $N > 0$ such that, for all $n > N$,

$$\sqrt{\frac{2(\frac{1}{p_{\min}} - 1 + \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}}))(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n}{\delta})}{p_{\min}(n-1)}} + \frac{2(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n}{\delta})}{p_{\min}(n-1)} - \sqrt{\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n+1}{\delta}}{p_{\min}^2 2n}} \geqslant 0.$$

Note that for $n > 1$, the second term is non-negative by properties of KL divergence and the fact that, for $\delta \in (0, 1)$ and $n \geqslant 1$, $\ln n - \ln\delta > 0$.

$$\sqrt{\frac{4n(n-1)(1 + \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}})p_{\min} - p_{\min})(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n}{\delta})}{p_{\min}^2(n-1)^2 2n}} - \sqrt{\frac{(n-1)^2(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n+1}{\delta})}{p_{\min}^2(n-1)^2 2n}}$$

Completing the difference of squares and getting rid of the common denominator yields

$$4n(n-1)(1 + \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}})p_{\min} - p_{\min})(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n}{\delta}) - (n-1)^2(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n+1}{\delta})$$

Factoring out terms dependent on sample complexity yields

$$4n(n-1)\ln n - (n-1)^2\ln(n+1) = (n-1)(3n+1)\ln n + 1 \geqslant 0$$

for $n \geqslant 1$ and $\hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\mathbb{P}}) = \frac{1}{p_{\min}}(c-1)$ where $c \geqslant \frac{(n-1)^2\ln(n+1)}{4n(n-1)\ln n}$. Therefore, picking N large enough ensures a tighter bound in the limit and completes the argument. □

B Detailed proof of Theorem 4.2

Proof. Let $\mathcal{D}$ and $\mathcal{D}_n$ respectively denote the true and empirical joint distributions of context x and reward function ρ. We construct empirical counterfactual risk functions $r_{\mathcal{D}}$ and $r_{\mathcal{D}_n}$ such that distributions $\mathbb{P}$, $\mathbb{Q}$ over deterministic hypothesis space $\mathcal{H}$ can be applied to them:

$$r_{\mathcal{D}}(h) = \left\langle \mathcal{D}, 1 - \mathbb{E}_{(x_i,\rho_i)\sim\mathcal{D}}\mathbb{E}_{k_i\sim\pi_{\mathbb{P}}(x_i)}\frac{\mathbb{I}\{h(c_i) = k_i\}\rho_i(x_i, k_i)}{\max\{p_{\min}, \pi_{\mathbb{P}}(x_i, k_i)\}} \right\rangle \tag{21}$$

$$r_{\mathcal{D}_n}(h) = \left\langle \mathcal{D}_n, 1 - \mathbb{E}_{(x_i,\rho_i)\sim\mathcal{D}}\mathbb{E}_{k_i\sim\pi_{\mathbb{P}}(x_i)}\frac{\mathbb{I}\{h(x_i) = k_i\}\rho_i(x_i, k_i)}{\max\{p_{\min}, \pi_{\mathbb{P}}(x_i, k_i)\}} \right\rangle \tag{22}$$

$$. \tag{23}$$

We can express the true and estimated CIS risk for $\pi_{\mathbb{Q}}$ as

$$R_{\mathrm{CIS}}(\pi_{\mathbb{Q}}) = \langle\, \mathbb{Q}, r_{\mathcal{D}}\rangle \qquad \text{and} \qquad \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \tau_n^{\pi_{\mathbb{P}}}) = \langle\, \mathbb{Q}, r_{\mathcal{D}_n}\rangle \tag{24}$$

then use that $\dfrac{1}{\max\{p_{\min}, \pi_{\mathbb{P}}(x_i, k_i)\}} \leqslant \dfrac{1}{\pi_{\mathbb{P}}(x_i, k_i)}$ to obtain that $R_{\mathrm{CIS}}(\pi_{\mathbb{Q}}) \geqslant R(\pi_{\mathbb{Q}})$ and

$$R(\pi_{\mathbb{Q}}) - \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\pi_{\mathbb{P}}}) \leqslant R_{\mathrm{CIS}}(\pi_{\mathbb{Q}}) - \hat{R}_{\mathrm{CIS}}(\pi_{\mathbb{Q}}, \mathcal{F}_n^{\pi_{\mathbb{P}}}) = \langle\, \mathbb{Q}, r_{\mathcal{D}}\rangle - \langle\, \mathbb{Q}, r_{\mathcal{D}_n}\rangle. \tag{25}$$

We will then apply different inequalities that bound the right side of the following with $(x_i, k_i) \in \mathcal{F}_n^{\mathbb{P}}$:

$$n\langle\, \mathbb{Q}, r_{\mathcal{D}} - r_{\mathcal{D}_n}\rangle = \left\langle \mathbb{Q}, \sum_{i=1}^{n}\left[r_{\mathcal{D}} - \left(1 - \mathbb{E}_{h\in\mathcal{H}}\frac{\mathbb{I}\{h(x_i) = k_i\}\rho(x_i, k_i)}{\max\{p_{\min}, \pi_{\mathbb{P}}(x_i, k_i)\}}\right)\right]\right\rangle. \tag{26}$$

Using PAC-Bayes-KL inequality (Theorem A.1) Scaling r_D and $1 - \frac{\mathbb{I}\{h(x_i)=k_i\}\rho(x_i,k_i)}{\max\{p_{\min},\pi_{\mathbb{P}}(x_i,k_i)\}}$ to $[0, 1]$, constant $(p_{\min} - 1)$ addition and subtraction cancel out, we have

$$\langle\, \mathbb{Q}, p_{\min}(r_{\mathcal{D}} - r_{\mathcal{D}_n})\rangle \leqslant_{1-\delta} \sqrt{\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n+1}{\delta}}{2n}} \tag{27}$$

$$\text{thus } \langle\, \mathbb{Q}, (r_{\mathcal{D}} - r_{\mathcal{D}_n})\rangle \leqslant_{1-\delta} \frac{1}{p_{\min}}\sqrt{\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{n+1}{\delta}}{2n}}. \tag{28}$$

Using PAC-Bayes-Azuma-Hœffding inequality (Theorem A.2) For any $c > 1$, we have

$$n\langle \mathbb{Q}, r_{\mathcal{D}} - r_{\mathcal{D}_n}\rangle \leqslant_{1-\delta} \frac{1+c}{2\sqrt{2}}\sqrt{\Big(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{2}{\delta} + \frac{\ln 2}{2\ln c}\Big(1 + \ln\big(\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P})}{\ln\frac{2}{\delta}}\big)\Big)\Big)\frac{n}{p_{\min}^2}}. \tag{29}$$

Using PAC-Bayes-Bernstein inequality (Theorem A.3) For any $\lambda > 0$, we have

$$n\langle \mathbb{Q}, r_{\mathcal{D}} - r_{\mathcal{D}_n}\rangle \leqslant_{1-\delta} \frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{2}{\delta}}{\lambda} + (e-2)\lambda\langle \mathbb{Q}, \overline{V}_n\rangle \tag{30}$$

$$\text{where } \overline{V}_i(h) = \sum_{j=1}^{i}\mathbb{E}[\overline{Z}_j(h)^2|\overline{Z}_1(h),\ldots,\overline{Z}_{j-1}(h)]$$

$$\text{and } \overline{Z}_i(h) = r_{\mathcal{D}} - \Big(1 - \frac{\mathbb{I}\{h(x_i) = k_i\}\rho(x_i,k_i)}{\max\{p_{\min}, \pi_{\mathbb{P}}(x_i,k_i)\}}\Big) \qquad \text{noting that } ||\overline{Z}_i|| \leqslant K = \frac{1}{p_{\min}}.$$

For $\mathbb{Q}$ that satisfies $\sqrt{\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{2v}{\delta}}{(e-2)\langle \mathbb{Q}, \overline{V}_n\rangle}} \leqslant \frac{1}{K} = p_{\min}$, we have that for any $c > 1$,

$$n\langle \mathbb{Q}, r_{\mathcal{D}} - r_{\mathcal{D}_n}\rangle \leqslant_{1-\delta} (1+c)\sqrt{(e-2)\langle \mathbb{Q}, \overline{V}_n\rangle\Big(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{2v}{\delta}\Big)} \qquad \text{where } v = \left\lceil \frac{\ln\frac{(e-2)n}{\ln\frac{2}{\delta}}}{2\ln c}\right\rceil + 1. \tag{31}$$

By using the facts that $\overline{Z}_i \in [-1, \frac{1}{p_{\min}}]$ and $\mathbb{E}[\overline{Z}_j|\overline{Z}_1,\ldots,\overline{Z}_{j-1}] = \overline{0}$ combined with the Bhatia-Davis inequality (Theorem A.4, Bhatia and Davis (2000)), we can bound

$$\mathbb{E}[\overline{Z}_n(j)^2|\overline{Z}_1(h),\ldots,\overline{Z}_{j-1}(h)] \leqslant \Big(\frac{1}{p_{\min}} - 0\Big)\Big(0 - (-1)\Big) = \frac{1}{p_{\min}} \quad \text{s.t.} \quad \overline{V}_n \leqslant \frac{n}{p_{\min}}. \tag{32}$$

Thus,

$$n\langle \mathbb{Q}, r_{\mathcal{D}} - r_{\mathcal{D}_n}\rangle \leqslant O\Big(\sqrt{\langle \mathbb{Q}, \overline{V}_n\rangle\Big(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{1}{\delta}\Big)}\Big) \leqslant O\Big(\sqrt{n\Big(\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{1}{\delta}}{p_{\min}}\Big)}\Big). \tag{33}$$

For all other $\mathbb{Q}$:

$$n\langle \mathbb{Q}, r_{\mathcal{D}} - r_{\mathcal{D}_n}\rangle \leqslant_{1-\delta} 2K\Big(\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\big(\frac{2v}{\delta}\big)\Big). \tag{34}$$

$\square$