Control Variate Diagnostics for Detecting Problems in Logged Bandit Feedback

BEN LONDON and THORSTEN JOACHIMS, Amazon Music, USA

We propose diagnostics, based on *control variates*, to detect data quality issues in logged bandit feedback data, which is of critical importance for accurate offline evaluation and training of recommendation policies. Our diagnostics can provably detect two common types of data issues: (1) when the policy that logged the data was insufficiently randomized; (2) when the logged propensity values are incorrect due to downstream filtering. We establish bounds on the false positive and false negative rates of our diagnostics, then empirically validate our approach on synthetic data.

Additional Key Words and Phrases: offline policy evaluation, contextual bandits, data quality

ACM Reference Format:

Ben London and Thorsten Joachims. 2022. Control Variate Diagnostics for Detecting Problems in Logged Bandit Feedback. In *Proceedings of CONSEQUENCES+REVEAL Workshop – RecSys 2022*. ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Offline policy evaluation and learning with logged bandit feedback [1, 2, 4–11, 13–21, 23–26] has become state-of-the-art for a wide range of contextual bandit problems, such as search, recommendation and display advertising. These methods have been widely adopted in industry [3, 12], partly because they allow us to estimate (and optimize) the impact of new recommendation policies from the safety of an offline environment, without affecting the user experience. To ensure that offline evaluation is informative, we want estimates of key metrics to be *unbiased*—meaning, in expectation, the value that we estimate offline equals the actual value when the policy is deployed online.

To achieve unbiasedness, many estimators employ an *importance weighting* scheme known as *inverse propensity scoring* (IPS). Crucially, IPS-based estimators are unbiased only when the policy that logged the data was sufficiently randomized and the logged propensities reflect the true probability of observing an action. Unfortunately, real-world applications often fail to meet these requirements; and when these failures occur, the resulting biases can lead to faulty decisions, failed online experiments, harm to user experience, and wasted time and resources.

We are therefore motivated to investigate diagnostics that can detect—and, ideally, help debug—problems in logged bandit feedback. This will instill greater confidence in offline evaluation, and improve chances for successful online experiments. Moreover, if we can detect these issues early on in the experimentation process, we will save time and effort that would have been wasted on experiments with flawed data, and shield users from potentially negative experiences.

Toward this goal, we propose diagnostics based on a statistical concept known as *control variates*. In particular, we adopt a control variate that is often used to control variance in offline policy evaluation [22]. The defining characteristic of this control variate is that, if the above assumptions hold, then its mean is exactly one. Moreover, we prove that its mean will differ from one (by a certain amount) under certain violations of the above assumptions. Having characterized the mean under good and bad conditions, we derive data-quality diagnostics based on statistical tests, and upper-bound the diagnostics' false positive and false negative rates. Finally, we empirically validate our diagnostics on synthetic data.

Manuscript submitted to ACM

2 PRELIMINARIES

We consider a setting in which we have collected a dataset of *logged bandit feedback*—such as user interactions with a recommender system. We denote this dataset by $S \triangleq (x_i, a_i, p_i, r_i)_{i=1}^n$, where $x_i \in \mathcal{X}$ is information describing the current *context* (e.g., user or time-of-day); $a_i \in \mathcal{A}$ is the *action* (a.k.a. *arm*) that was selected from a fixed[1] set $\mathcal{A}$; $p_i \in (0, 1]$ is the probability (a.k.a. *propensity*) with which a_i was selected; and $r_i \in \mathbb{R}$ is the observed *reward* (a.k.a. *utility*) in response to the selected action. We assume that contexts are i.i.d., and that actions are selected independently.

The logged actions are selected by a *logging policy*, π_0, which induces a distribution on $\mathcal{A}$ given a context. We denote the conditional distribution by $\pi_0(x)$, and the conditional probability of a given action by $\pi_0(a \mid x)$. The logging policy may be difficult or impossible to recreate in hindsight, since a recommender system is typically a complex composition of multiple, interdependent components. Thus, we will assume that the propensities of only the selected actions are logged, and that access to the logging policy, or its full conditional distribution, is not possible afterward.

A common use case for such a dataset is to evaluate a new policy, π, typically referred to as the *target policy*. Of interest is the target policy's *expected reward*, $R(\pi) \triangleq \mathbb{E}_{x,r}\, \mathbb{E}_{a\sim\pi(a\mid x)}[r(x, a)]$, where $r(x, a)$ denotes the (stochastic) reward generated by the environment for the given context and selected action. Given a fixed dataset—without knowledge of the reward distribution, or the ability to let the target policy interact with the environment—we can only estimate this quantity. Importantly, we must correct for the bias induced by the logging policy. To do so, we consider estimators that use *inverse propensity scoring* (IPS); in particular, the standard IPS reward estimator, $\hat{R}(\pi, S) \triangleq \frac{1}{n}\sum_{i=1}^n r_i \frac{\pi(a_i \mid x_i)}{p_i}$. The ratio $\pi(a_i \mid x_i)/p_i$ is often called an *importance weight*. It is straightforward to show that the IPS estimator is unbiased for any target policy π, meaning $\mathbb{E}_S[\hat{R}(\pi, S)] = R(\pi)$, when the following conditions are met:

(1) **Full support**: For any $x \in \mathcal{X}$ and $a \in \mathcal{A}$, $\pi_0(a \mid x) > 0$.
(2) **Accurate propensities**: For all $i = 1, \ldots, n$, $p_i = \pi_0(a_i \mid x_i)$, where $\pi_0(a_i \mid x_i)$ is the *true* probability of observing action a_i in context x_i.

Sadly, these assumptions often fail in practice. For instance, if the logging policy is a learned parametric model, then it is possible that the logging policy is deterministic in certain contexts, or does not sufficiently explore certain actions; meaning, it is *support deficient* (or, has *insufficient support*). Further, as the logging policy is likely just one component in a larger system—which may filter or transform the logging policy's selections—it is possible that the distribution of logged actions does not match the distribution given by the logged propensities. For example, recommender systems often have "guardrails" to prevent inappropriate recommendations, such as recommending adult content to minors. These guardrails may reject the logging policy's selections, thus altering the distribution of observed actions in ways that are not captured by the logged propensities.

When the above assumptions fail, we can no longer guarantee that the IPS estimate is unbiased. We therefore seek diagnostics to detect when the assumptions fail, which tells us when IPS estimation can be unbiased.

3 CONTROL VARIATES

Our primary tool is a statistical concept known as *control variates*. Broadly speaking, a control variate is a statistic whose expectation is a known value. In the context of offline policy evaluation, it is natural to define the control variate as the average of importance weights [22]. For a given target policy, π, we denote the control variate by

$$CV(\pi, S) \triangleq \frac{1}{n}\sum_{i=1}^n \frac{\pi(a_i \mid x_i)}{p_i}.$$

[1]For simplicity of notation, we assume that the action set is fixed, though it could also depend on the context.

When the logging policy has full support, and the logged propensities are accurate, it is straightforward to show that $\mathbb{E}_S[CV(\pi, S)] = 1$, *for any target policy* π. Thus, given a large enough dataset, we expect $CV(\pi, S) \approx 1$ for any π. When we observe otherwise—that is, when it is statistically implausible that the expected control variate is one, given the observed data—then we "reject the null hypothesis" that the data is OK. That, in a nutshell, will be our strategy.

To our knowledge, the only other work to discuss data issues in bandit feedback is by Li et al. [12]. They propose a diagnostic called the *harmonic mean test*, which uses a control variate of the form $\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}\{a_i=a^*\}}{\pi_0(a^* \mid x_i)} + \frac{\mathbb{1}\{a_i\neq a^*\}}{1-\pi_0(a^* \mid x_i)}$ for a given a^*. A key disadvantage of this diagnostic is that it requires knowledge of the logging policy; one must be able to compute $\pi_0(a^* \mid x)$ for any a^*. In contrast, our proposed diagnostics do not require the logging policy after the data is logged. It is unclear whether their diagnostic offers any advantage over ours.

4 SUMMARY OF THEORETICAL CONTRIBUTIONS

The control variate approach only works when a data issue causes the expected control variate to differ from one. This is not guaranteed for arbitrary data issues; indeed, there exist conditions under which the unbiasedness requirements (full support and accurate propensities) fail to hold, but $\mathbb{E}_S[CV(\pi, S)] = 1$ regardless (see discussion in Appendix C). Thus, the unbiasedness requirements are only *sufficient* conditions for $\mathbb{E}_S[CV(\pi, S)] = 1$, but not *necessary* conditions.

Nonetheless, we can prove that two commonplace data issues lead to $\mathbb{E}_S[CV(\pi, S)] \neq 1$ for appropriate choices of π. The first data issue that we can provably detect is when the logging policy is support deficient in some fraction of contexts. In this case, for a target policy that selects actions uniformly at random, $\pi_{\mathrm{U}}(a \mid x) \triangleq |\mathcal{A}|^{-1}$, we show that the expected control variate is strictly less than one, $\mathbb{E}_S[CV(\pi_{\mathrm{U}}, S)] < 1$. The second issue that we can provably detect is when some nonempty subset of the action set, $\mathcal{B} \subset \mathcal{A} : \mathcal{B} \neq \emptyset$, is subject to post-selection filtering (e.g., for "guardrails" or due to non-random logging issues), which causes the logged propensities to be inaccurate. In this setting, we use a target policy that always picks the same action—the "always pick a" policy, $\pi_a(a' \mid x) \triangleq \mathbb{1}\{a = a'\}$—to show that the expected control variate can be less than or greater than one, depending on whether $a \in \mathcal{B}$.

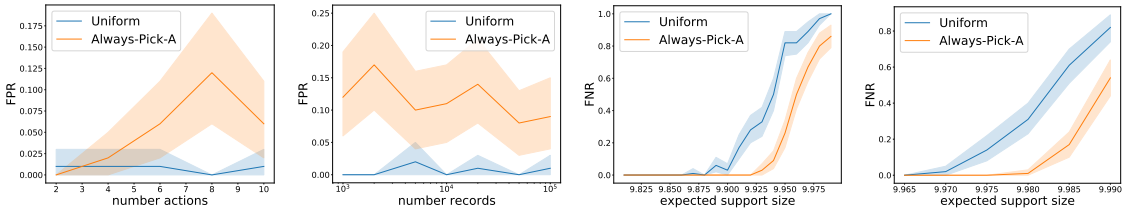
Based on these findings, we derive diagnostics using the target policies associated with our expectation bounds. As discussed in Appendix A, evaluating our diagnostics on a finite sample of data can lead to *false positives* (i.e., detecting an issue when there is none) or *false negatives* (i.e., failing to detect an actual issue). We therefore upper-bound the false positive rate (FPR) and false negative rate (FNR) of our diagnostics by appealing to concentration of the control variate around its mean. The bounds show that the FPR and FNR decrease exponentially fast as the data size grows, or as the data issues become more pronounced. Due to space limitations, we defer the details of our analysis to the appendices.

5 EXPERIMENTS

We experiment with the following two practical diagnostics, motivated by our theory. They are labeled "Uniform" and "Always-Pick-A" in reference to their respective target policies.

Diagnostic 1 (Uniform)**.** *For a given* $\delta \in (0, 1)$*, let* $\Delta > 0$ *define an approximate confidence interval, such that* $\widetilde{\Pr}_S\{|CV(\pi_U, S) - \mu| \geq \Delta\} \leq \delta$*, where* $\widetilde{\Pr}$ *indicates the approximation and* $\mu \triangleq \mathbb{E}_S[CV(\pi_U, S)]$*. If* $|CV(\pi_U, S) - 1| > \Delta$*, then reject the null hypothesis (that the data is OK).*

Diagnostic 2 (Always-Pick-A)**.** *For a given* $\delta \in (0, 1)$*, let* $\Delta_a > 0$ *define an approximate confidence interval for* $CV(\pi_a, S)$*, such that* $\widetilde{\Pr}_S\{|CV(\pi_a, S) - \mu_a| \geq \Delta_a\} \leq \delta$*, where* $\mu_a \triangleq \mathbb{E}_S[CV(\pi_a, S)]$*. If* $|CV(\pi_a, S) - 1| > \Delta_a$ *for any* $a \in \mathcal{A}$*, then reject the null hypothesis (that the data is OK).*



(a) $|\mathcal{A}|$ vs FPR (b) n vs FPR (c) FNR – Support (d) FNR – Propensities

Fig. 1. Results of our synthetic data experiments. Solid lines represent the mean of 100 trials; shaded areas indicate the 95% confidence interval. Figures 1a and 1b plot the FPR when no issue is present, whereas Figures 1c and 1d plot the FNR in the presence of insufficient support and incorrect propensities (per the distributional assumptions used in Appendix C.1), respectively.

We empirically evaluate the FPR and FNR of the above diagnostics on synthetic data. To simulate logged bandit feedback, we sample actions from a non-contextual logging policy whose action distribution is inversely proportional to an action's index in the action set; e.g., $\pi_0(a^{(1)} \mid x) \propto 1$, $\pi_0(a^{(2)} \mid x) \propto 1/2$, etc. Unless otherwise stated, the action set contains 10 actions, and each synthethic dataset contains 100,000 records. Results are averaged over 100 trials (i.e., datasets). Each diagnostic uses bootstrapping (with 1,000 resamples) to compute 99% confidence intervals for Δ or Δ_a.

False Positive Rate. We first examine the FPR of our diagnostics on "good" (i.e., full support, accurate propensities) synthetic data. Figures 1a and 1b plot the FPR as a function of the number of actions, $|\mathcal{A}|$, and number of records, n, respectively. Predictably, the FPR of Diagnostic 1 is much lower than that of Diagnostic 2. Diagnostic 2's FPR increases with $|\mathcal{A}|$, which concurs with our theory (see Proposition 4). Intuitively, this diagnostic trades precision for sensitivity. The FPR of Diagnostic 2 decreases as a function of n, while Diagnostic 1 seems largely insensitive to n in this range.

Insufficient Support. To simulate a logging policy with insufficient support, we assign zero probability to the last (i.e., tenth) action on an ϵ fraction of records. As such, the expected support size is $(1-\epsilon)\,|\mathcal{A}|+\epsilon(|\mathcal{A}|-1) = |\mathcal{A}|-\epsilon$. Figure 1c plots the FNR as a function of expected support size. The FNR is negligible until around 9.9 ($\epsilon = 0.1$), which illustrates how the expected support must be very close to full before the diagnostics have trouble detecting an issue. Bear in mind, the last action has the lowest propensity ($\propto 1/10$) to begin with, so detecting when it is missing is challenging. As expected, Diagnostic 2 is slightly more sensitive than Diagnostic 1.

Incorrect Propensities. To simulate incorrect propensities, we implement the failure mode from Appendix C.1. On an ϵ fraction of records, we assign zero probability to the first action; however, we log propensities from the *original* distribution, in which the first action had nonzero probability. Figure 1d plots the resulting FNR, again as a function of expected support size. This type of issue is more difficult to detect than insufficient support (when correct propensities are given), but the diagnostics are nonetheless able to detect issues even when the expected support is close to full. Though Diagnostic 1 is not guaranteed to work in this setting, it is nonetheless effective. However, Diagnostic 2, which was designed for this setting, is more sensitive than Diagnostic 1.

6 CONCLUSIONS AND FUTURE WORK

Our theoretical and empirical findings indicate that control variate diagnostics are accurate tools for detecting logging issues in bandit data. They can help ensure that downstream offline evaluation and learning are accurate, which in turn increases experimental velocity and avoids potential harm to the user experience. An important area for future work is to design diagnostics that can detect an even broader range of data quality issues, in other bandit settings (e.g., ranking), and that provide stronger explanations for identifying the sources of logging problems.

REFERENCES

[1] A. Agarwal, S. Basu, T. Schnabel, and T. Joachims. 2017. Effective Evaluation Using Logged Bandit Feedback from Multiple Loggers. *Knowledge Discovery and Data Mining* (2017).

[2] L. Bottou, J. Peters, J. Quiñonero-Candela, D. Charles, D. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research* 14 (2013), 3207–3260.

[3] M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti, and E. Chi. 2019. Top-K Off-Policy Correction for a REINFORCE Recommender System. In *Web Search and Data Mining*.

[4] M. Dudik, J. Langford, and L. Lihong. 2011. Doubly Robust Policy Evaluation and Learning. In *International Conference on Machine Learning*.

[5] M. Farajtabar, Y. Chow, and M. Ghavamzadeh. 2018. More Robust Doubly Robust Off-policy Evaluation. In *International Conference on Machine Learning*.

[6] A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé. 2018. Offline A/B Testing for Recommender Systems. *Web Search and Data Mining* (2018).

[7] N. Jiang and L. Li. 2016. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. In *International Conference on Machine Learning*.

[8] T. Joachims, A. Swaminathan, and M. de Rijke. 2018. Deep Learning with Logged Bandit Feedback. In *International Conference on Learning Representations*.

[9] N. Kallus. 2018. Balanced Policy Evaluation and Learning. In *Neural Information Processing Systems*.

[10] N. Kallus and A. Zhou. 2018. Policy Evaluation and Optimization with Continuous Treatments. In *Artificial Intelligence and Statistics*.

[11] J. Langford, A. Strehl, and J. Vaughan. 2008. Exploration scavenging. In *International Conference on Machine Learning*.

[12] L. Li, S. Chen, J. Kleban, and A. Gupta. 2015. Counterfactual Estimation and Optimization of Click Metrics in Search Engines: A Case Study. In *International Conference on World Wide Web*.

[13] L. Li, W. Chu, J. Langford, and X. Wang. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Web Search and Data Mining*.

[14] L. Li, R. Munos, and C. Szepesvári. 2015. Toward Minimax Off-policy Value Estimation. In *Artificial Intelligence and Statistics*.

[15] A. Liu, H. Liu, A. Anandkumar, and Y. Yue. 2019. Triply Robust Off-Policy Evaluation. *CoRR* abs/1911.05811 (2019).

[16] B. London and T. Sandler. 2019. Bayesian Counterfactual Risk Minimization. In *International Conference on Machine Learning*.

[17] Y. Ma, Y.-X. Wang, and B. Narayanaswamy. 2019. Imitation-Regularized Offline Learning. In *Artificial Intelligence and Statistics*.

[18] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *International Conference on Machine Learning*.

[19] A. Strehl, J. Langford, L. Li, and S. Kakade. 2010. Learning from Logged Implicit Exploration Data. In *Neural Information Processing Systems*.

[20] Y. Su, L. Wang, M. Santacatterina, and T. Joachims. 2019. CAB: Continuous Adaptive Blending for Policy Evaluation and Learning. In *International Conference on Machine Learning*.

[21] A. Swaminathan and T. Joachims. 2015. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *Journal of Machine Learning Research* (2015).

[22] A. Swaminathan and T. Joachims. 2015. The Self-Normalized Estimator for Counterfactual Learning. In *Neural Information Processing Systems*.

[23] P. Thomas and E. Brunskill. 2016. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In *International Conference on Machine Learning*.

[24] P. Thomas, G. Theocharous, and M. Ghavamzadeh. 2015. High-Confidence Off-Policy Evaluation. In *AAAI*.

[25] N. Vlassis, A. Bibaut, M. Dimakopoulou, and T. Jebara. 2019. On the Design of Estimators for Bandit Off-Policy Evaluation. In *International Conference on Machine Learning*.

[26] Y.-X. Wang, A. Agarwal, and M. Dudík. 2017. Optimal and Adaptive Off-policy Evaluation in Contextual Bandits. In *International Conference on Machine Learning*.

A FINITE SAMPLES AND OVERFITTING

Since we only have access to a finite sample of data, it is possible that the choice of target policy in the above diagnostic can lead to false positives (i.e., detecting a data issue when there is none) or false negatives (i.e., failing to detect an actual data issue).

For example, suppose the target policy is trained to maximize the IPS estimator: $\arg\max_{\pi \in \Pi} \hat{R}(\pi, S)$, over some predefined class of policies, Π. If Π is sufficiently expressive (such as a class of deep neural networks), then it is possible that the learned policy could simply memorize which training examples have (non)negative reward. Then, it could maximize the IPS estimator by assigning probability one to the logged actions where the reward is nonnegative, and probability zero to those where the reward is negative. If enough logged rewards are nonnegative, such a policy could result in a control variate that is significantly larger than one. For instance, if $r_i \geq 0$ for at least $n/2$ examples, and $p_i \ll 1/2$ for all i, then $CV(\pi_{\text{overfit}}, S) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}\{r_i \geq 0\}}{p_i} \gg 1$. Alternatively, if most logged rewards were negative, we could overfit the data such that $CV(\pi_{\text{overfit}}, S) \ll 1$. In either case, the control variate would indicate an issue, regardless of whether there actually was one.

A similar overfitting issue can result in failing to detect an actual data issue. Suppose the target policy overfits the logging policy's distribution such that, for every training example, $\pi_{\text{overfit}}(a_i \mid x_i) \approx p_i = \pi_0(a_i \mid x_i)$. Then, every importance weight is approximately one, and so is the control variate—regardless of whether there is a problem with the data.

For these reasons, care must be taken in which target policy is selected for the diagnostic, so as to avoid overfitting.

B DETECTING INSUFFICIENT SUPPORT

One data issue that we can provably detect is when the logging policy has insufficient support. We will use the notation $\text{supp}(\pi_0(x))$ to denote the support of π_0 conditioned on x. When π_0 is support deficient, $\text{supp}(\pi_0(x)) \subset \mathcal{A}$; hence, $|\text{supp}(\pi_0(x))| < |\mathcal{A}|$. In the following, we show that the control variate of a support deficient logging policy and a uniformly random target policy, $\pi_{\text{U}}(a \mid x) \triangleq |\mathcal{A}|^{-1}$, is strictly less than one in expectation. The proof is given in Appendix D.1.

Proposition 1. *If* $\mathbb{E}_x[|\text{supp}(\pi_0(x))|] \leq B < |\mathcal{A}|$, *then* $\mathbb{E}_S[CV(\pi_U, S)] \leq \frac{B}{|\mathcal{A}|} < 1$.

Proposition 1 tells us that insufficient support can always be detected *in expectation* by checking for $\mathbb{E}_S[CV(\pi_{\text{U}}, S)] < 1$. Of course, we cannot test the expected value, since we only have access to finite data. However, if the control variate concentrates, and the mean is bounded away from one by a sufficient amount, then we should be able to detect (with high probability) when the logging policy is support deficient. This motivates the following simple diagnostic.

Diagnostic 3. *For a given* tolerance, $\tau > 0$, *if* $CV(\pi_U, S) < 1 - \tau$, *then reject the null hypothesis (that the data has full support).*

Under modest assumptions on the logging policy—namely, that the probabilities of the (limited) support set are uniformly lower-bounded—we can upper-bound the *false positive rate* (FPR) and *false negative rate* (FNR) of the above diagnostic.[2] The proof is provided in Appendix D.2.

Proposition 2. *Let* $\kappa \triangleq \inf_{x \in \mathcal{X},\, a \in \text{supp}(\pi_0(x))} \pi_0(a \mid x)$.

[2]Note that $1 - \text{FPR}$ is the *specificity*, i.e., how often the diagnostic correctly says the data is OK; while $1 - \text{FNR}$ is the *sensitivity*, i.e., how well the diagnostic detects an issue.

- ***FPR**: If π_0 has full support (no deficiency), then for any fixed $\tau > 0$ that does not depend on the data,*

$$\Pr_S \{CV(\pi_U, S) < 1 - \tau\} \leq \exp\left(-2\kappa^2 |\mathcal{A}|^2 \tau^2 n\right).$$

- ***FNR**: If $\mathbb{E}_x[|\text{supp}(\pi_0(x))|] \leq B < |\mathcal{A}|$, then for any fixed $\tau \in (0, (1 - B/|\mathcal{A}|))$ that does not depend on the data,*

$$\Pr_S \{CV(\pi_U, S) \geq 1 - \tau\} \leq \exp\left(-2\kappa^2 \left(|\mathcal{A}| - B - \tau |\mathcal{A}|\right)^2 n\right).$$

If the logging policy's support deficiency is serious enough—meaning, B is small relative to $|\mathcal{A}|$—then the FNR decreases because the issue becomes easier to detect. And, as expected, the FNR depends on the *amount* of logged data; the FNR vanishes exponentially fast as n increases. Note that the requirement $\tau < 1 - \frac{B}{|\mathcal{A}|}$ stems from our use of Hoeffding's inequality. Intuitively, it means that the tolerance cannot be greater than the margin between the true mean and the ideal mean.

B.1 A More Practical Diagnostic.

Ideally, we would use Proposition 2 to tune τ, but the bounds depend properties of the logging policy that are, by assumption, unknown. We do not know B—if we did, we would not need to test for full support—and κ may be difficult to reason about when full support is not trivially satisfied.

We therefore propose a simple, practical diagnostic that, while sacrificing some theoretical guarantees, requires no knowledge of the distribution. Instead of specifying a threshold, τ, we can estimate a confidence interval for $CV(\pi_U, S)$, where standard techniques like the normal approximation or bootstrapping provide tighter (approximate) confidence intervals than the Hoeffding bound.[3] If the confidence interval does not contain 1, then we reject the null hypothesis that the logging policy has full support (and the propensities were logged correctly). The only parameter required for this diagnostic is the confidence, δ, which determines the false positive and false negative rates.

Diagnostic 4. *For a given $\delta \in (0, 1)$, let $\Delta > 0$ define an approximate confidence interval, such that $\widetilde{\Pr}_S \{|CV(\pi_U, S) - \mu| \geq \Delta\} \leq \delta$, where $\widetilde{\Pr}$ indicates the approximation and $\mu \triangleq \mathbb{E}_S[CV(\pi_U, S)]$. If $|CV(\pi_U, S) - 1| > \Delta$, then reject the null hypothesis (that the data is OK).*

In Section 5, we show empirically that this diagnostic is effective at detecting insufficient support, as well as other data issues.

C DETECTING INCORRECT PROPENSITIES

A second, more insidious problem is when the logged propensities do not correspond to the actual observed distribution of actions. This can happen, for instance, if the logging policy computes the wrong propensities; if the logging mechanism drops certain records not at random; or if the logging policy is part of a larger system that has the ability to filter or override the logging policy.

To distinguish between the true logging policy, π_0, and possibly incorrect propensities in the logs, we use $p(a \mid x)$ to denote the mechanism that generates a logged—and possibly faulty—propensity. Note that π_0 may denote not just a bandit policy (e.g., a softmax model), but the *entire system* responsible for selecting (and presenting) actions. Thus, any source of bias (e.g., filtering selections or dropping records) is encapsulated in π_0, such that $\pi_0(a \mid x)$ represents the *true* probability of observing action a in context x.

[3] Though we only need a one-sided confidence interval to detect insufficient support, we recommend a two-sided confidence interval to catch other problems that may manifest as $CV \gg 1$.

If we do not have the correct propensity values for IPS estimation, the reward estimates will be biased. Indeed, it can be shown that the bias of using propensities $p(a \mid x)$ when the true system propensities are $\pi_0(a \mid x)$ is given by

$$\mathop{\mathbb{E}}_{S}[\hat{R}(\pi, S)] - R(\pi) = \mathop{\mathbb{E}}_{x,r} \mathop{\mathbb{E}}_{a \sim \pi(x)} \left[r(x,a) \left(\frac{\pi_0(a \mid x)}{p(a \mid x)} - 1 \right) \right].$$

From this, it becomes clear that IPS's unbiasedness critically depends on $p(a \mid x) = \pi_0(a \mid x)$.

Rewriting the expected control variate as an expectation over the choices of the target policy, $\mathbb{E}_S[CV(\pi, S)] = \mathbb{E}_x \mathbb{E}_{a \sim \pi(x)} \left[\frac{\pi_0(a \mid x)}{p(a \mid x)} \right]$, one can see that the expected control variate can be both less than or greater than one, depending on the target policy. For instance, if π puts all of its mass on the actions for which $\pi_0(a \mid x) < p(a \mid x)$, then $\mathbb{E}_S[CV(\pi_{\mathrm{U}}, S)] < 1$; or if it puts all of its mass on actions for which $\pi_0(a \mid x) > p(a \mid x)$, then $\mathbb{E}_S[CV(\pi_{\mathrm{U}}, S)] > 1$.

If the target policy is uniform, we have that $\mathbb{E}_S[CV(\pi_{\mathrm{U}}, S)] = \mathbb{E}_x \left[\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{\pi_0(a \mid x)}{p(a \mid x)} \right]$. We still cannot guarantee that this expectation will be less than or greater than one. For example, suppose there are two actions, and that the logged propensities are always (independent of context) $p(x) = (0.1, 0.9)$. Now, suppose the logging policy was actually $\pi_0(x) = (0.01, 0.99)$; then we would have $\mathbb{E}_S[CV(\pi_{\mathrm{U}}, S)] = \frac{1}{2}\left(\frac{0.01}{0.1} + \frac{0.99}{0.9}\right) = 0.6$. However, suppose instead that $\pi_0(x) = (0.9, 0.1)$; then we would have $\mathbb{E}_S[CV(\pi_{\mathrm{U}}, S)] = \frac{1}{2}\left(\frac{0.9}{0.1} + \frac{0.1}{0.9}\right) \approx 4.56$. Therefore, it can go either way.

To complicate matters even further, there exist circumstances in which the logged propensities are wrong and the expected control variate still equals one. For example, suppose there are three actions, and that $p(x) = (0.1, 0.1, 0.8)$, while $\pi_0(x) = (0.05, 0.15, 0.8)$, both independent of the context, x. Then, $\mathbb{E}_S[CV(\pi_{\mathrm{U}}, S)] = \frac{1}{3}\left(\frac{0.05}{0.1} + \frac{0.15}{0.1} + \frac{0.8}{0.8}\right) = 1$.[4] This presents a serious challenge to our control variate approach, since it is predicated on the idea that the expected control variate equals one only when everything is OK, and this is clearly not the case. Indeed, this example points out that having full support and correct propensities is a *sufficient condition* for the expected control variate, with respect to a specific target policy, to equal one, but not a *necessary condition*.

While this is dismaying, there is yet some hope. In the following section, we identify one very practical scenario in which incorrect propensities can be detected.

C.1 Post-Selection Filtering

In this section, we will assume that π_0 and p exhibit a specific type of failure. Suppose certain actions are not allowed in certain contexts, resulting in π_0 having insufficient support for certain contexts. Further, suppose the propensity generating mechanism, p, is unaware of this issue. Thus, there may be contexts, x, for which $p(x)$ puts nonzero mass on all actions, but in reality there is some subset of actions that can never be selected; that is, $|\mathrm{supp}(p(x))| = |\mathcal{A}|$, but $|\mathrm{supp}(\pi_0(x))| < |\mathcal{A}|$. We will assume that the propensities are *proportionally* correct, given renormalization for the proper support set; meaning, for any action that can be selected, $a \in \mathrm{supp}(\pi_0(x))$, we have $\pi_0(a \mid x) \propto p(a \mid x)$.

The above scenario is all too common in practical settings. Take, for example, a music recommendation system. The system may consist of a (learned) model to select content, followed by some business logic to enforce rules (a.k.a. "guardrails") that are not easily encoded in the model or its input data. For instance, the model's selections might be filtered so that the same musical artist is not recommended multiple times in a row; or, so that explicit content is not recommended to children. The key idea is that the model is unaware that the broader system is filtering its selections; and the system is effectively limiting the support of the observed action distribution. Thus, if p represents the probability distribution of the model, and π_0 represents the distribution of the entire system, then there are some contexts for

[4] Though we have assumed that $p(x)$ and $\pi_0(x)$ are independent of x, one could construct scenarios in which they are context-dependent by assuming that they output the same distribution for two particular contexts.

which $|\text{supp}(p(x))| > |\text{supp}(\pi_0(x))|$. Further, if the system logs $p(a \mid x)$ as the propensity for the selected action, it is proportionally accurate, subject to renormalization.

To detect this failure mode, we propose a control variate diagnostic based on a different type of target policy. For a given action, $a \in \mathcal{A}$, let π_a denote the "always pick a" policy, where $\pi_a(a' \mid x) \triangleq \mathbb{1}\{a = a'\}$. If π_0 is behaving correctly—that is, if $\pi_0(x) = p(x)$ for all $x \in \mathcal{X}$, and π_0 has full support—then the control variate for π_a should be one in expectation. However, we claim that if a is a "bad" action (in the way defined above), then the expected control variate for π_a will be less than one; and if a is "good," then the expected control variate will be greater than one. The following result (proved in Appendix D.3) formalizes this claim.

Proposition 3. *Assume that the propensity distribution, p, has full support, and let $\kappa \triangleq \inf_{x \in \mathcal{X},\, a \in \mathcal{A}} p(a \mid x) > 0$. Suppose there is a nonempty set of "bad" actions, $\mathcal{B} \subset \mathcal{A} : \mathcal{B} \neq \emptyset$, such that, with probability at least $\epsilon \in (0, 1]$ over draws of x, for $a \in \mathcal{A}$, $\pi_0(a \mid x) = \frac{\mathbb{1}\{a \notin \mathcal{B}\}\, p(a \mid x)}{\sum_{a' \in \mathcal{A} \setminus \mathcal{B}} p(a' \mid x)}$; and with probability at most $1 - \epsilon$ over draws of x, $\pi_0(a \mid x) = p(a \mid x)$. Then,*

$$\text{for any } a \in \mathcal{B}, \qquad \mathop{\mathbb{E}}_{S}[CV(\pi_a, S) \mid a \in \mathcal{B}] \leq 1 - \epsilon < 1; \tag{1}$$

$$\text{and for any } a \in \mathcal{A} \setminus \mathcal{B} \text{ (i.e., } a \notin \mathcal{B}), \qquad \mathop{\mathbb{E}}_{S}[CV(\pi_a, S) \mid a \notin \mathcal{B}] \geq 1 + \frac{\epsilon\, |\mathcal{B}|}{\kappa^{-1} - |\mathcal{B}|} > 1. \tag{2}$$

It is worth pointing out that $0 < \frac{|\mathcal{B}|}{\kappa^{-1} - |\mathcal{B}|} \leq |\mathcal{A}| - 1$, which means that Equations 1 and 2 are not always symmetric. One case where they *are* symmetric is when $|\mathcal{B}| = |\mathcal{A}|/2$ (i.e., half of the actions are "bad") and p is uniformly random, meaning $\kappa = |\mathcal{A}|^{-1}$. Then, the upper bound matches the lower bound at $1 - \epsilon$. It is this potential for symmetry that prevents us from using a uniformly random target policy, which would be equivalent to averaging over all π_a.

Proposition 3 motivates the following diagnostic based on the "always pick a" target policy.

Diagnostic 5. *For a given tolerance, $\tau > 0$, if $|CV(\pi_a, S) - 1| > \tau$ for any $a \in \mathcal{A}$, then reject the null hypothesis (that the data is OK).*[5]

As before, we can bound the FPR and FNR of this diagnostic. (Proof provided in Appendix D.4.)

Proposition 4. *Let $\kappa \triangleq \inf_{x \in \mathcal{X},\, a \in \text{supp}(\pi_0(x))} \pi_0(a \mid x)$.*

- ***FPR**: If π_0 has full support and $\pi_0 = p$, then for any fixed $\tau > 0$,*

$$\Pr_{S}\{\exists a \in \mathcal{A},\ |CV(\pi_a, S) - 1| > \tau\} \leq 2\,|\mathcal{A}| \exp\left(-2\kappa^2\tau^2 n\right).$$

- ***FNR**: If π_0 and p have the properties described in Proposition 3—using the same definitions for κ, $\mathcal{B}$ and ϵ, with $\eta \triangleq \frac{|\mathcal{B}|}{\kappa^{-1} - |\mathcal{B}|}$—then for any fixed $\tau \in (0, \min\{\epsilon, \epsilon\eta\})$,*

$$\Pr_{S}\{\forall a \in \mathcal{A},\ |CV(\pi_a, S) - 1| \leq \tau\} \leq \exp\left(-2\kappa^2(\max\{\epsilon, \epsilon\eta\} - \tau)^2 n\right).$$

In practice, we would likely use confidence intervals rather than fix τ, as described in Appendix B.1. This leads to the following diagnostic, which we evaluate in Section 5.

Diagnostic 6. *For a given $\delta \in (0, 1)$, let $\Delta_a > 0$ define an approximate confidence interval for $CV(\pi_a, S)$, such that $\widetilde{\Pr}_S\{|CV(\pi_a, S) - \mu_a| \geq \Delta_a\} \leq \delta$, where $\mu_a \triangleq \mathbb{E}_S[CV(\pi_a, S)]$. If $|CV(\pi_a, S) - 1| > \Delta_a$ for any $a \in \mathcal{A}$, then reject the null hypothesis (that the data is OK).*

[5] Note that this diagnostic prioritizes sensitivity at the expense of specificity. We could alternatively require $|CV(\pi_a, S) - 1| > \tau$ for *all* $a \in \mathcal{A}$ to detect an irregularity, which would prioritize specificity over sensitivity.

D DEFERRED PROOFS

D.1 Proof of Proposition 1

Using linearity of expectation to focus on the importance weight for a single (random) interaction, we have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{S}[CV(\pi_{\mathrm{U}}, S)] &= \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{a\sim\pi_0(x)} \left[\frac{\pi_{\mathrm{U}}(a \mid x)}{\pi_0(a \mid x)}\right] \\
&= \mathop{\mathbb{E}}_{x} \left[\sum_{a\in\mathrm{supp}(\pi_0(x))} \frac{\pi_{\mathrm{U}}(a \mid x)}{\pi_0(a \mid x)} \pi_0(a \mid x)\right] \\
&= \mathop{\mathbb{E}}_{x} \left[\sum_{a\in\mathrm{supp}(\pi_0(x))} \frac{1}{|\mathcal{A}|}\right] \\
&= \frac{1}{|\mathcal{A}|} \mathop{\mathbb{E}}_{x} \left[|\mathrm{supp}(\pi_0(x))|\right] \\
&\leq \frac{B}{|\mathcal{A}|}.
\end{aligned}$$

Noting that $\frac{B}{|\mathcal{A}|} < 1$ (by definition of B) completes the proof.

D.2 Proof of Proposition 2

Throughout the proof, we use $\mu \triangleq \mathbb{E}_S[CV(\pi_{\mathrm{U}}, S)]$ to denote the expected control variate.

First, we consider the case when the data is fine (i.e., the logging policy had full support), but the diagnostic incorrectly detects a data issue (i.e., a false positive). Since there is no actual data issue, we must have $\mu = 1$. Therefore,

$$\begin{aligned}
\Pr_{S}\{CV(\pi_{\mathrm{U}}, S) < 1 - \tau\} &\leq \Pr_{S}\{CV(\pi_{\mathrm{U}}, S) \leq 1 - \tau\} \\
&= \Pr_{S}\{CV(\pi_{\mathrm{U}}, S) \leq \mu - \tau\}.
\end{aligned}$$

To upper-bound the above probability, we will use the lower tail version of Hoeffding's inequality, since $CV(\pi_{\mathrm{U}}, S)$ is essentially just an average of independent, bounded random variables. To be precise, each variable is *almost surely* bounded in the range $[0, (\kappa\,|\mathcal{A}|)^{-1}]$, since $\pi_{\mathrm{U}}(a \mid x) = |\mathcal{A}|^{-1}$ and $\pi_0(a \mid x) \geq \kappa$. Therefore, for $\tau > 0$,

$$\Pr_{S}\{CV(\pi_{\mathrm{U}}, S) - \mu \leq -\tau\} \leq \exp\left(-2\kappa^2\,|\mathcal{A}|^2\,\tau^2 n\right).$$

Next, we consider the case when the the logging policy *was* support deficient, but the diagnostic failed to detect the issue (i.e., a false negative). From Proposition 1, we have that $\mu \leq \frac{B}{|\mathcal{A}|}$. Thus, $1 - \tau - \mu \geq 1 - \tau - \frac{B}{|\mathcal{A}|}$, and

$$\begin{aligned}
\Pr_{S}\{CV(\pi_{\mathrm{U}}, S) \geq 1 - \tau\} &= \Pr_{S}\{CV(\pi_{\mathrm{U}}, S) - \mu \geq 1 - \tau - \mu\} \\
&\leq \Pr_{S}\left\{CV(\pi_{\mathrm{U}}, S) - \mu \geq 1 - \tau - \frac{B}{|\mathcal{A}|}\right\} \\
&\leq \exp\left(-2\kappa^2\,|\mathcal{A}|^2 \left(1 - \tau - \frac{B}{|\mathcal{A}|}\right)^2 n\right).
\end{aligned}$$

The last inequality follows from the upper tail version of Hoeffding's inequality, noting that $1 - \tau - \frac{B}{|\mathcal{A}|} > 0$ by our assumption that $\tau \leq 1 - \frac{B}{|\mathcal{A}|}$. To complete the proof, we simply rearrange the righthand expression.

D.3 Proof of Proposition 3

Before proceeding, let $\mathcal{X}_\epsilon \triangleq \{x : x \in \mathcal{X}, \text{supp}(\pi_0(x)) = \mathcal{B}\}$ denote the set of contexts for which the logging policy's support is limited to $\mathcal{B}$, and note that (by assumption) $\Pr_x\{x \in \mathcal{X}_\epsilon\} \geq \epsilon$.

The purpose of π_a is to isolate $\pi_0(a \mid x)$. Regardless of whether a is "good" or "bad," we have

$$\mathop{\mathbb{E}}_{S}[CV(\pi_a, S)] = \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{a' \sim \pi_0(x)} \left[\frac{\mathbb{1}\{a = a'\}}{p(a' \mid x)} \right] = \mathop{\mathbb{E}}_{x} \left[\frac{\pi_0(a \mid x)}{p(a \mid x)} \right].$$

Because we have assumed that $\mathcal{B}$ is *static* (i.e., context-independent), we have that the ratio $\pi_0(a \mid x)/p(a \mid x)$ is either at most one or at least one, depending on whether a is in $\mathcal{B}$.

Let us first consider the case where $a \in \mathcal{B}$. When $x \notin \mathcal{X}_\epsilon$, the logging policy's conditional distribution has full support on $\mathcal{A}$, and thus

$$\pi_0(a \mid x, a \in \mathcal{B}, x \notin \mathcal{X}_\epsilon) = p(a \mid x).$$

However, when $x \in \mathcal{X}_\epsilon$, we have

$$\pi_0(a \mid x, a \in \mathcal{B}, x \in \mathcal{X}_\epsilon) = 0.$$

Therefore, $\pi_0(a \mid x, a \in \mathcal{B}) = \mathbb{1}\{x \notin \mathcal{X}_\epsilon\} p(a \mid x)$, and

$$\mathop{\mathbb{E}}_{x} \left[\frac{\pi_0(a \mid x, a \in \mathcal{B})}{p(a \mid x)} \right] = \mathop{\mathbb{E}}_{x} \left[\mathbb{1}\{x \notin \mathcal{X}_\epsilon\} \frac{p(a \mid x)}{p(a \mid x)} \right] \leq 1 - \epsilon,$$

which is less than one when $\epsilon > 0$.

For the case where $a \notin \mathcal{B}$, the analysis starts the same; but when $x \in \mathcal{X}_\epsilon$ we have

$$\pi_0(a \mid x, a \notin \mathcal{B}, x \in \mathcal{X}_\epsilon) = \frac{p(a \mid x)}{Z(x)}, \quad \text{where} \quad Z(x) \triangleq \sum_{a' \in \mathcal{A} \setminus \mathcal{B}} p(a' \mid x).$$

Thus,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x} \left[\frac{\pi_0(a \mid x, a \notin \mathcal{B})}{p(a \mid x)} \right] \\
&= \mathop{\mathbb{E}}_{x} \left[\frac{\mathbb{1}\{x \notin \mathcal{X}_\epsilon\} \pi_0(a \mid x, a \notin \mathcal{B}, x \notin \mathcal{X}_\epsilon) + \mathbb{1}\{x \in \mathcal{X}_\epsilon\} \pi_0(a \mid x, a \notin \mathcal{B}, x \in \mathcal{X}_\epsilon)}{p(a \mid x)} \right] \\
&= \mathop{\mathbb{E}}_{x} \left[\frac{\mathbb{1}\{x \notin \mathcal{X}_\epsilon\} p(a \mid x) + \mathbb{1}\{x \in \mathcal{X}_\epsilon\} \frac{p(a \mid x)}{Z(x)}}{p(a \mid x)} \right] \\
&= \mathop{\mathbb{E}}_{x} \left[\mathbb{1}\{x \notin \mathcal{X}_\epsilon\} + \mathbb{1}\{x \in \mathcal{X}_\epsilon\} Z(x)^{-1} \right] \\
&= \mathop{\mathbb{E}}_{x} \left[1 + \mathbb{1}\{x \in \mathcal{X}_\epsilon\} (Z(x)^{-1} - 1) \right].
\end{aligned}$$

In the last line, we use the fact that $\mathbb{1}\{x \notin \mathcal{X}_\epsilon\} = 1 - \mathbb{1}\{x \in \mathcal{X}_\epsilon\}$. The righthand side is minimized when $Z(x)$ is maximized. Using our assumption that the logged propensities are uniformly lower-bounded by κ, we have that

$$Z(x) = 1 - \sum_{a \in \mathcal{B}} p(a \mid x) \leq 1 - \kappa |\mathcal{B}|.$$

Therefore,

$$Z(x)^{-1} - 1 \geq (1 - \kappa |\mathcal{B}|)^{-1} - 1 = \frac{\kappa |\mathcal{B}|}{1 - \kappa |\mathcal{B}|},$$

and

$$\begin{aligned}\mathop{\mathbb{E}}_{x}\left[1 + \mathbb{1}\{x \in \mathcal{X}_\epsilon\}(Z(x)^{-1} - 1)\right] &\geq 1 + \frac{\kappa\,|\mathcal{B}|}{1 - \kappa\,|\mathcal{B}|}\mathop{\mathbb{E}}_{x}[\mathbb{1}\{x \in \mathcal{X}_\epsilon\}] \\ &\geq 1 + \frac{\kappa\,|\mathcal{B}|}{1 - \kappa\,|\mathcal{B}|}\,\epsilon.\end{aligned}$$

Finally, to prove that the righthand side is greater than one, we note that $\kappa \in (0, |\mathcal{A}|^{-1}]$ and $|\mathcal{B}| \in [1, |\mathcal{A}| - 1]$, so $\kappa\,|\mathcal{B}| > 0$ and $1 > 1 - \kappa\,|\mathcal{B}| \geq 1 - \frac{|\mathcal{A}|-1}{|\mathcal{A}|} > 0$. Noting that $\epsilon > 0$ completes the proof.

D.4 Proof of Proposition 4

We start with a technical lemma, which we will use to prove the FNR bound in Proposition 4.

Lemma 1. *Assume that π_0 and p have the properties described in Proposition 3, using the same definitions for κ, $\mathcal{B}$ and ϵ, with $\eta \triangleq \frac{|\mathcal{B}|}{\kappa^{-1} - |\mathcal{B}|}$. Then, for any given $a \in \mathcal{A}$ and $\tau \in (0, \min\{\epsilon, \epsilon\eta\})$:*

$$\Pr_S\{|CV(\pi_a, S) - 1| \leq \tau\} \leq \begin{cases}\exp\left(-2\kappa^2(\epsilon - \tau)^2 n\right) & \text{if } a \in \mathcal{B};\\ \exp\left(-2\kappa^2(\epsilon\eta - \tau)^2 n\right) & \text{if } a \notin \mathcal{B}.\end{cases} \tag{3}$$

Proof. Similar to the previous proofs, we will use $\mu \triangleq \mathbb{E}_S[CV(\pi_a, S)]$ to denote the expected control variate for π_a. The proof considers two cases: either $a \in \mathcal{B}$, and $\mu < 1$; or $a \notin \mathcal{B}$, and $\mu > 1$. In either case, we show that it is unlikely for the control variate to be close to one, due to its concentration around the mean.

If $a \in \mathcal{B}$, we have (from Equation 1) that $1 - \tau - \mu \geq \epsilon - \tau$. Then,

$$\begin{aligned}\Pr_S\{|CV(\pi_a, S) - 1| \leq \tau\} &\leq \Pr_S\{CV(\pi_a, S) \geq 1 - \tau\} \\ &= \Pr_S\{CV(\pi_a, S) - \mu \geq 1 - \tau - \mu\} \\ &\leq \Pr_S\{CV(\pi_a, S) - \mu \geq \epsilon - \tau\} \\ &\leq \exp\left(-2\kappa^2(\epsilon - \tau)^2 n\right).\end{aligned}$$

The first inequality uses the fact that the interval $[1 - \tau, 1 + \tau]$ is subsumed by $[1 - \tau, \infty)$. When we apply Hoeffding's inequality at the end, we use the fact that $\pi_a(\cdot \mid \cdot)/p(\cdot \mid \cdot) \in [0, \kappa^{-1}]$ almost surely. Note that the range of τ is designed so that $\epsilon - \tau > 0$, which is a precondition of Hoeffding's inequality.

On the other hand, if $a \notin \mathcal{B}$, we have (from Equation 2) that

$$1 + \tau - \mu \leq \tau - \frac{\epsilon\,|\mathcal{B}|}{\kappa^{-1} - |\mathcal{B}|} = -(\epsilon\eta - \tau).$$

Thus,

$$\begin{aligned}\Pr_S\{|CV(\pi_a, S) - 1| \leq \tau\} &\leq \Pr_S\{CV(\pi_a, S) \leq 1 + \tau\} \\ &= \Pr_S\{CV(\pi_a, S) - \mu \leq 1 + \tau - \mu\} \\ &\leq \Pr_S\{CV(\pi_a, S) - \mu \leq -(\epsilon\eta - \tau)\} \\ &\leq \exp\left(-2\kappa^2(\epsilon\eta - \tau)^2 n\right).\end{aligned}$$

Again, the range of τ ensures that $\epsilon\eta - \tau > 0$, as required by Hoeffding's inequality. □

We can now prove Proposition 4. First, we upper-bound the FPR using a union bound over $a \in \mathcal{A}$:

$$\begin{aligned}\Pr_S\left\{\exists a \in \mathcal{A},\ |CV(\pi_a, S) - 1| > \tau\right\} &\leq \sum_{a\in\mathcal{A}} \Pr_S\left\{|CV(\pi_a, S) - 1| > \tau\right\} \\ &\leq \sum_{a\in\mathcal{A}} 2\exp\left(-2\kappa^2\tau^2 n\right).\end{aligned}$$

The last inequality follows from the two-sided Hoeffding bound, noting that $CV(\pi_a, S)$ is an average of i.i.d. random variables that are almost surely bounded in the range $[0, \kappa^{-1}]$.

To upper-bound the FNR, we start by applying a Fréchet inequality:

$$\Pr_S\left\{\forall a \in \mathcal{A},\ |CV(\pi_a, S) - 1| \leq \tau\right\} \leq \min_{a\in\mathcal{A}} \Pr_S\left\{|CV(\pi_a, S) - 1| \leq \tau\right\}.$$

Recall from Lemma 1 that $\Pr_S\left\{|CV(\pi_a, S) - 1| \leq \tau\right\}$ is upper-bounded by one of the righthand expressions in Equation 3, depending on whether $a \in \mathcal{B}$. Further, note that neither of the bounds depend on properties of a; only on properties of the logging policy's distribution. Therefore, the minimum probability (over $a \in \mathcal{A}$) is upper-bounded by the minimum of the bounds:

$$\begin{aligned}&\min_{a\in\mathcal{A}} \Pr_S\left\{|CV(\pi_a, S) - 1| \leq \tau\right\} \\ \leq\ &\min_{a\in\mathcal{A}} \mathbb{1}\{a \in \mathcal{B}\}\exp\left(-2\kappa^2(\epsilon-\tau)^2 n\right) + \mathbb{1}\{a \notin \mathcal{B}\}\exp\left(-2\kappa^2(\epsilon\eta-\tau)^2 n\right) \\ =\ &\min\left\{\exp\left(-2\kappa^2(\epsilon-\tau)^2 n\right), \exp\left(-2\kappa^2(\epsilon\eta-\tau)^2 n\right)\right\}.\end{aligned}$$

We can now reduce the expression using monotonicty of the exponent:

$$\begin{aligned}&\min\left\{\exp\left(-2\kappa^2(\epsilon-\tau)^2 n\right), \exp\left(-2\kappa^2(\epsilon\eta-\tau)^2 n\right)\right\} \\ =\ &\exp\left(\min\{-2\kappa^2(\epsilon-\tau)^2 n, -2\kappa^2(\epsilon\eta-\tau)^2 n\}\right) \\ =\ &\exp\left(-2\kappa^2\max\{|\epsilon-\tau|, |\epsilon\eta-\tau|\}^2 n\right) \\ =\ &\exp\left(-2\kappa^2(\max\{\epsilon, \epsilon\eta\}-\tau)^2 n\right).\end{aligned}$$

The last equality uses our assumption that $\tau \leq \min\{\epsilon, \epsilon\eta\}$, which guarantees $|\epsilon - \tau| = \epsilon - \tau$ and $|\epsilon\eta - \tau| = \epsilon\eta - \tau$; then, $\max\{\epsilon - \tau, \epsilon\eta - \tau\} = \max\{\epsilon, \epsilon\eta\} - \tau$.