Coverage Cliffs in Learning from Logged World Feedback

Pauline Bourigault
Imperial College London
p.bourigault22@imperial.ac.uk

Abstract

World feedback, including losses, costs, failures, latencies, health outcomes, or economic impacts measured from an environment, is a promising complement to human-preference feedback. Yet it is often logged under earlier policies, so its usefulness for a new policy is limited by off-policy coverage. We study this issue in logged contextual bandits. Instead of certifying mean loss, we ask when logged world feedback can certify upper-tail loss of a learned randomized policy. Our main result is a finite-sample PAC-Bayes certificate for posterior-averaged per-policy Value-at-Risk, obtained from posterior-uniform importance-weighted CDF control and quantile inversion. The analysis identifies a quantile-specific *coverage cliff*: if the logged feedback lacks enough weighted mass near the target quantile, the certified threshold must jump to a larger loss or become vacuous. Diagnostics show one theorem-covered continuous-outcome regime where the certificate is non-vacuous but conservative, and discrete or weak-coverage regimes where empirical improvements are not certifiable. The results highlight that logged world feedback is a coverage-limited statistical resource: it may support empirical learning without supporting a valid tail-safety certificate. Code repository

1 Introduction

Reinforcement learning from human feedback has shown that feedback can be more useful than a hand-written reward when the desired behavior is difficult to specify directly [8, 24]. Recent variants replace or augment humans with AI-generated feedback [3]. In many deployed systems, however, important feedback is neither a preference label nor a synthetic critique. It is a measured consequence of interaction with the world: latency, resource usage, financial cost, clinical outcome, safety violation, user retention, or downstream harm. These measured outcome signals are examples of world feedback.

World feedback is attractive because it is objective and scalable, but it is usually logged under previous policies. This makes the central difficulty statistical rather than only algorithmic: a new policy can be trained on logged outcome data only to the extent that the log covers the decisions and outcomes relevant to that policy. The issue is especially sharp for tail risk. A policy can have good mean performance while still placing non-negligible mass on catastrophic outcomes, and the rare outcomes that determine a quantile may be those with weakest off-policy coverage.

We use logged contextual bandits as a minimal model of world-feedback learning: a behavior policy chooses actions, the environment returns measurable outcome losses, and a new policy must be learned or certified from this logged feedback. This one-step setting underlies recommender systems, advertising, mobile health, and personalized decision making [18, 4, 30, 15]. Existing PAC-Bayes contextual-bandit work gives finite-sample guarantees for expected losses or rewards [19, 26, 1]. Distributional off-policy methods estimate target CDFs or broad return-distribution functionals [14, 7, 28], and risk-sensitive offline RL optimizes tail-sensitive criteria [33, 20, 36]. Here we focus on a

RLxF: Reinforcement Learning from World Feedback Workshop at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026

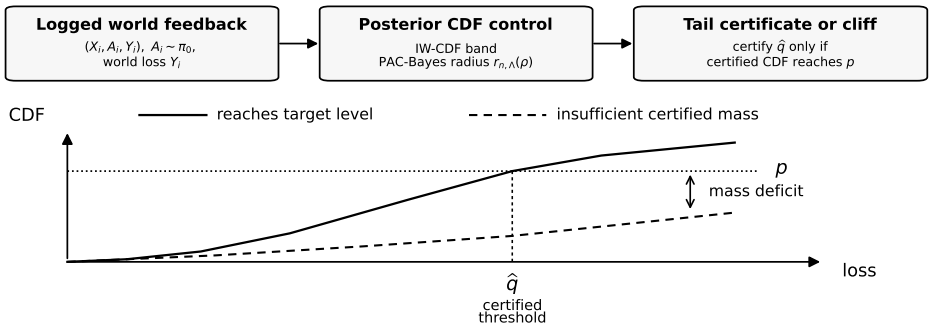

Figure 1: Overview of the paper. Logged world feedback provides outcome losses under a behavior policy. PAC-Bayes control of the importance-weighted loss CDF yields a posterior tail-risk certificate when the certified CDF reaches the target level p. If the certified CDF remains below p, the certificate has a mass deficit and must move to a larger threshold or become vacuous.

finite-sample certification question: can a posterior learned from logged world feedback be certified for upper-tail loss under a PAC-Bayes guarantee?

Contributions. We make three contributions. (i) We formulate logged world feedback as a finite-sample tail-certification problem. (ii) Under bounded importance weights and a local inverse-slope condition, we give a PAC-Bayes off-policy certificate for posterior-averaged per-policy VaR. (iii) We identify and diagnose a coverage cliff: unlike mean-risk certificates, which usually degrade gradually through variance, quantile certificates can jump or become vacuous when the weighted empirical CDF lacks enough mass near the target level. The appendix gives the exact radii, distribution-free fallback for atoms, rare-action lower bound, deterministic-center caveat, split-sample DR-CDF extension, proofs, and implementation details.

2 Tail certification from logged world feedback

Let logged data be $D_n = \{(X_i, A_i, Y_i)\}_{i=1}^n$, where X is context, A is an action chosen by a known logging policy $\pi_0(\cdot \mid X)$, and $Y \in [0, B]$ is a world-derived loss. A target policy π_θ has importance weight

$$w_\theta(x, a) = \frac{\pi_\theta(a \mid x)}{\pi_0(a \mid x)}, \qquad V_w(\theta) = \mathbb{E}_{X,A\sim\pi_0}\big[w_\theta(X, A)^2\big].$$

Let $F_\theta(y) = \mathbb{P}_{\pi_\theta}(Y \le y)$ and define the upper-tail quantile

$$Q_p(\theta) = \inf\{y : F_\theta(y) \ge p\}, \qquad p = 1 - \alpha.$$

For a prior π and posterior $\rho \ll \pi$, the certified object is

$$\mathcal{R}_p(\rho) = \mathbb{E}_{\theta\sim\rho} Q_p(\theta).$$

This is not the quantile of the Gibbs-mixture outcome distribution, because VaR is nonlinear. The appendix gives a separate conservative common-threshold certificate for the mixture quantile; the main certificate below is for $\mathcal{R}_p(\rho)$.

The estimator is based on a compactly smoothed importance-weighted CDF. With bandwidth $\varepsilon > 0$ and a compact CDF kernel H, define

$$\widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) = \frac{1}{n}\sum_{i=1}^n w_\theta(X_i, A_i) H\Big((y - Y_i)/\varepsilon\Big)$$

and the right-clipped empirical smoothed quantile

$$\widehat{\widetilde{Q}}^{\mathrm{IW}}_{n,\varepsilon,p}(\theta) = \inf\Big\{y \in [-\varepsilon, B + \varepsilon] : \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) \ge p\Big\},$$

with value $B + \varepsilon$ when the set is empty.

Theorem 1 (Informal PAC-Bayes VaR certificate; exact form in Appendix C.1)**.** *Assume* $0 \le w_\theta \le W$ *and an inverse-slope condition: for all* θ *and* $y \in [0, B]$*,*

$$Q_p(\theta) \le y + \frac{(p - F_\theta(y))_+}{m}.$$

Then, with probability at least $1 - \delta$*, simultaneously for all posteriors* $\rho \ll \pi$*,*

$$\mathbb{E}_\rho Q_p(\theta) \le \mathbb{E}_\rho \widehat{\widetilde{Q}}^{\mathrm{IW}}_{n,\varepsilon,p}(\theta) + \varepsilon + \frac{2r_{n,\Lambda}(\rho)}{m}, \tag{1}$$

where $r_{n,\Lambda}(\rho)$ *is a posterior-uniform PAC-Bayes Bernstein CDF radius of order*

$$\sqrt{\frac{\mathbb{E}_\rho V_w(\theta)\{\mathrm{KL}(\rho\|\pi) + \log(N/\delta)\}}{n}} + O\left(\frac{W\{\mathrm{KL}(\rho\|\pi) + \log(N/\delta)\}}{n}\right),$$

up to finite threshold-grid and Bernstein-parameter-grid factors.

The theorem has three important restrictions. It certifies a Gibbs posterior, not an arbitrary deterministic center. It certifies posterior-averaged per-policy VaR, not automatically the mixture quantile. Finally, the inverse-slope condition is a local mass condition around the target quantile; for atoms or unjittered zero-one losses, the valid fallback is a distribution-free level-shift certificate rather than the inverse-slope bound.

3 Coverage cliffs

The main qualitative consequence is the coverage cliff. For mean off-policy evaluation, weak overlap appears primarily as a larger variance term. For quantiles, a second failure mode appears: the weighted empirical CDF may not certify enough mass to reach level p. In that case a valid quantile certificate cannot simply pay a larger smooth penalty at the same threshold; it must move to a larger loss or to the right boundary of the support.

The formal statement in Appendix C.2 is as follows. On the PAC-Bayes CDF event, fix a posterior ρ and a threshold set $\mathcal{Y}$. If

$$\mathbb{E}_\rho \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) < p - r_{n,\Lambda}(\rho) \quad \text{for every } y \in \mathcal{Y},$$

then this CDF event alone cannot certify any threshold in $\mathcal{Y}$ as having posterior-average exceedance probability at most α. Any level-shift certificate based only on the event must choose a threshold outside $\mathcal{Y}$ or return the right-clipped boundary. This is not an artifact of the algorithm; a rare-action lower bound in the appendix shows unavoidable $\sqrt{V_w/n}/m$ dependence in a two-action subclass.

For world-feedback learning, logged outcomes do more than supply numerical losses: they determine where the learner has statistical coverage. A new policy may improve empirical tail loss while remaining uncertifiable if the log does not contain enough weighted evidence near the desired tail level.

4 Empirical diagnostics

The experiments illustrate the certification mechanism and its failure modes across theorem-aligned and weak-coverage regimes. Table 1 separates theorem-covered certification from empirical tail-risk measurements; details are in Section E. On a heteroskedastic continuous-loss bandit, the inverse-slope certificate is theorem-covered and non-vacuous for all 20 seeds, but conservative. On unjittered discrete classification-to-bandit and rare-catastrophe benchmarks, the density assumption is not the formal guarantee; the valid level-shift and mixture certificates are vacuous under these sample sizes. Those rows are therefore reported as empirical diagnostics, not as inverse-slope certificates.

Table 1: Main diagnostic summary. Lower empirical VaR is better. The first block reports the theorem-covered continuous-loss certificate. The second block reports empirical diagnostics in regimes where the inverse-slope certificate is not the formal guarantee and the valid level-shift/mixture certificates are vacuous at the tested sample sizes. "Mass def." denotes empirical IW mass deficit.

Regime	Method	Certificate status	Empirical VaR	Mass def.
Theorem-covered certificate				
Continuous	PB-OffPol-VaR	Non-vacuous density certificate; certified VaR 4.532 ± 0.019 in $20/20$ seeds	2.160	0
Empirical diagnostic regimes				
Letter	PB-Mean	Mean-risk comparator; no VaR certificate	2.344 at $p = 0.90$	0.641
Letter	PB-OffPol-VaR	Level-shift / mixture certificates vacuous	2.288 at $p = 0.90$	0.327
Rare-cat.	PB-Mean	Mean-risk comparator; no VaR certificate	1.284 at $p = 0.95$	0.120
Rare-cat.	PB-OffPol-VaR	Level-shift / mixture certificates vacuous	1.271 at $p = 0.95$	0

The diagnostics illustrate three regimes. First, the inverse-slope certificate can be executed non-vacuously in a theorem-aligned continuous-outcome setting. Second, the tail-aware objective can improve empirical VaR over a PAC-Bayes mean baseline on logged bandit tasks. Third, valid certificates can remain vacuous when the weighted CDF lacks enough mass near the target level. We therefore interpret the empirical results as evidence for the certification mechanism and its failure modes, not as a benchmark claim. In weaker-coverage regimes, model-assisted CDF methods such as direct, doubly robust, or OPRA-style baselines can be stronger practical learners.

5 Limitations and scope

The setting is contextual bandits, not full sequential RL. Logging propensities are assumed known or otherwise certified. Bounded importance weights are required. The inverse-slope theorem is appropriate for continuous losses with local mass near the target quantile, not for arbitrary atoms. Deterministic-center reporting requires an additional policy-to-CDF stability penalty. Split-sample DR-CDF certification is included in the appendix, but same-sample nuisance fitting and cross-fitting require further analysis. These restrictions are part of the guarantee rather than implementation details; weakening any of them requires additional analysis.

References

[1] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Unified PAC-Bayesian study of pessimism for offline policy learning with regularized importance sampling. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 88–109. PMLR, 2024.

[2] Hind Atbir, Farah Cherfaoui, Guillaume Metzler, Emilie Morvant, and Paul Viallard. PAC-Bayesian bounds on constrained f-entropic risk measures. *arXiv preprint arXiv:2510.11169*, 2025.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[4] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(65):3207–3260, 2013.

[5] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[6] Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Institute of Mathematical Statistics Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, 2007.

[7] Yash Chandak, Scott Niekum, Bruno Castro da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S. Thomas. Universal off-policy evaluation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[8] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[9] Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28 (1):1–47, 1975. doi: 10.1002/cpa.3160280102.

[10] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104, 2011.

[11] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014. doi: 10.1214/14-STS500.

[12] Shashank Gupta, Olivier Jeunen, Harrie Oosterhuis, and Maarten de Rijke. Optimal baseline corrections for off-policy contextual bandits. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 845–855. ACM, 2024. doi: 10.1145/3640457.3688105.

[13] Joel L. Horowitz. Bootstrap methods for median regression models. *Econometrica*, 66(6): 1327–1352, 1998. doi: 10.2307/2999619.

[14] Audrey Huang, Liu Leqi, Zachary C. Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[15] Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[16] Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 2021.

[17] Ilja Kuzborskij, Claire Vernade, András György, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 640–648. PMLR, 2021.

[18] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670. ACM, 2010. doi: 10.1145/1772690.1772758.

[19] Ben London and Ted Sandler. Bayesian counterfactual risk minimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4125–4133. PMLR, 2019.

[20] Yecheng Jason Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[21] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[22] Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes un-expected Bernstein inequality. In *Advances in Neural Information Processing Systems*, volume 32, pages 12180–12191, 2019.

[23] Zakaria Mhammedi, Benjamin Guedj, and Robert C. Williamson. PAC-Bayesian bound for the conditional value at risk. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[24] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[25] Marc Rigter, Bruno Lacerda, and Nick Hawes. RAMBO-RL: Robust adversarial model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[26] Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. PAC-Bayesian offline contextual bandits with guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29777–29799. PMLR, 2023.

[27] Otmane Sakhi, Imad Aouali, Pierre Alquier, and Nicolas Chopin. Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2566.

[28] Muhammad Faaiz Taufiq, Jean-Francois Ton, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal off-policy prediction in contextual bandits. In *Advances in Neural Information Processing Systems*, volume 35, pages 31512–31524, 2022.

[29] Muhammad Faaiz Taufiq, Arnaud Doucet, Rob Cornish, and Jean-Francois Ton. Marginal density ratio for off-policy evaluation in contextual bandits. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[30] Ambuj Tewari and Susan A. Murphy. From ads to interventions: Contextual bandits in mobile health. In James M. Rehg, Susan A. Murphy, and Santosh Kumar, editors, *Mobile Health: Sensors, Analytic Methods, and Applications*, pages 495–517. Springer International Publishing, Cham, 2017. doi: 10.1007/978-3-319-51394-2_25.

[31] Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems*, volume 26, pages 109–117, 2013.

[32] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/b13794.

[33] Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. In *International Conference on Learning Representations*, 2021.

[34] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014. doi: 10.1145/2641190.2641198.

[35] Stanislav Volgushev, Shih-Kang Chao, and Guang Cheng. Distributed inference for quantile regression processes. *Annals of Statistics*, 47(3):1634–1662, 2019. doi: 10.1214/18-AOS1730.

[36] Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with CVaR. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35864–35907. PMLR, 2023.

[37] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3589–3597. PMLR, 2017.

Supplementary material overview

Section A positions the work and states the scope of the guarantees. Section B gives the formal setup. Section C states the CDF certificate, VaR inversion, level-shift fallback, mixture-quantile certificate, coverage-cliff proposition, and rare-action lower bound. Section D gives the learning rule and extensions. Section E gives the diagnostic experiments and additional tables. Section F contains the proofs, and Section G gives implementation notes.

A Related work and certification scope

A.1 Related work

PAC-Bayes off-policy learning. PAC-Bayes off-policy bounds for contextual bandits have primarily targeted expected return or expected loss. Bayesian counterfactual risk minimization uses PAC-Bayes ideas with logged bandit feedback and importance weighting [19]; subsequent work develops PAC-Bayesian offline contextual-bandit guarantees for mixtures of decision rules and more tractable optimization [26], and unified analyses of pessimism and regularized importance sampling [1]. Our analysis inherits the Gibbs-posterior viewpoint from this literature, but changes the target functional from a scalar mean to a CDF-indexed quantile. The shift matters technically: VaR requires threshold-uniform CDF control and a separate inversion step.

Off-policy evaluation, CDF estimation, and distributional guarantees. Classical OPE uses IPS, self-normalized IPS, direct method, and doubly robust estimators to estimate expected value under a target policy [10, 11]. Minimax analyses show that overlap-dependent variance is unavoidable in agnostic contextual-bandit OPE and introduce SWITCH-type model-assisted variance reduction [37]; recent baseline-correction methods further systematize variance reduction for mean OPE/OPL [12]. Finite-sample confidence sequences and self-normalized confidence bounds give strong mean-certification tools [17, 16]. More recent distributional OPE work estimates target-policy CDFs, return-distribution functionals, or predictive intervals. OPRA gives uniform concentration for IS and DR CDF estimators and plug-in guarantees for Lipschitz risk functionals, including CVaR-like objectives, but VaR lies outside the Lipschitz-risk framework [14]. Universal Off-Policy Evaluation provides high-confidence bounds for arbitrary return-distribution parameters, including quantiles and the full return CDF, in off-policy RL [7]. Conformal off-policy prediction gives finite-sample predictive intervals under logged contextual-bandit data [28]. Marginal density-ratio methods reduce variance by focusing on outcome-distribution shift rather than policy ratios [29]. Recent logarithmic-smoothing work develops fully empirical pessimistic bounds for broad classes of importance-weighted contextual-bandit risk estimators [27]. These works are central comparators: they show that distributional and pessimistic off-policy analysis already exists, while our contribution is the PAC-Bayes posterior-uniform treatment of posterior-averaged per-policy VaR with explicit quantile-inversion and mass-deficit analysis. In principle, any valid uniform CDF band can be paired with a local-slope inversion to obtain a VaR-style certificate; the distinctive point here is the posterior-uniform PAC-Bayes learning certificate, the posterior-averaged estimand, and the explicit treatment of unnormalized-IW mass deficit.

Risk-sensitive learning and PAC-Bayes risk measures. PAC-Bayes bounds for CVaR and related risk measures show that PAC-Bayes tail-risk learning is already an established direction [23, 2]. CVaR is more regular than VaR because it can often be represented through expectations or Lipschitz variational forms; VaR, by contrast, is a quantile and can jump under small CDF perturbations near atoms or flat regions. Risk-sensitive offline and online RL studies CVaR and distributional objectives, including conservative distributional methods and minimax analyses [33, 25, 20, 36]. We focus on one-step contextual bandits with known logging propensities, where the PAC-Bayes certificate applies to a Gibbs posterior over policies.

Quantile smoothing and inversion. Smoothed quantile and smoothed CDF methods are classical in statistics [13, 35]. Their role here is as follows: compact smoothing gives a Lipschitz threshold process that can be covered by a finite grid, after which PAC-Bayes Bernstein control can be transferred to a VaR certificate through an explicit inverse-slope or level-shift argument.

Scope summary. The guarantee has an explicit scope. It applies to posterior-averaged per-policy VaR under bounded importance weights and either the inverse-slope condition or the distribution-free level-shift fallback. The inverse-slope theorem is most appropriate for continuous losses with local mass near the target quantile. For atoms or unjittered discrete losses, the level-shift certificate is the formal guarantee. The Gibbs-mixture quantile, deterministic-center deployment, and split-DR variants require the separate conditions stated in the corresponding propositions and theorem.

A.2 Scope of the guarantees

Assumptions and failure modes. The two main regularity requirements are overlap and inverse slope, and both have explicit failure modes. Overlap can be certified from known propensities by checking $\min_{i,a} \pi_0(a \mid X_i)$ and by logging the realized maximum IW ratio; if the desired policy class violates overlap, one must either restrict the policy, use clipped weights and report clipping bias, or use a different data source. The inverse-slope assumption is empirically plausible only for continuous losses with nonflat density around the target quantile. It is not appropriate for unjittered zero-one classification losses or distributions with atoms at the quantile; in those cases the distribution-free level-shift theorem is the formal guarantee. The implementation records mass deficit, empirical IW range, effective sample size, and local quantile-spacing diagnostics, and it separates density-based certificates from level-shift certificates.

The main theorem certifies the posterior-averaged per-policy VaR objective, not the quantile of the Gibbs-mixture outcome distribution. Proposition 2 gives a separate conservative route to a common-threshold mixture-quantile certificate, but it is a different estimand and does not inherit the inverse-slope rate for $\mathbb{E}_\rho Q_p(\theta)$. Proposition 4 gives a deterministic-center certificate, but only after adding a policy-to-CDF stability constant; if this constant is unavailable or too loose, deterministic-center results should remain labeled as surrogate empirical results. Because the theorem uses an unnormalized IW empirical CDF, finite-sample IW mass can fall below the target quantile level; Theorem 2 includes a mass-deficit correction, and the code reports this diagnostic. The bounded-IW assumption rules out severe propensity mismatch unless clipping or regularization is analyzed with an explicit bias term. The inverse-slope theorem applies most cleanly to continuous losses; for discrete classification-to-bandit losses, the distribution-free level-shift theorem is the correct formal certificate unless continuous jitter or a smoothed loss protocol is used. Together, these conditions specify the theorem-covered regime and separate it from empirical or surrogate use cases.

If logging propensities are estimated rather than known, the present theorem should be read conditionally on a verified propensity model. A simple deterministic reduction is possible: if $|\pi_0(a \mid x)/\hat{\pi}_0(a \mid x) - 1| \le \zeta < 1$ uniformly, then the plug-in weight $\hat{w} = \pi_\theta/\hat{\pi}_0$ has range at most $(1+\zeta)W$ and its smoothed CDF target differs from the true one by at most ζ uniformly in y and θ. One can therefore add an extra ζ CDF slack, divided by m after inverse-slope inversion, and use the enlarged range in the Bernstein term. Without such a certified propensity-error bound, learned propensities are an empirical modeling choice rather than a theorem-covered setting. Similarly, self-normalized CDFs [17] can reduce mass deficit in practice, but the ratio introduces bias and denominator dependence. A safe PAC-Bayes SNIPS-CDF theorem would need simultaneous numerator and denominator control; if the denominator is certified above $1-\tau$ and the numerator error is at most e, the induced CDF error is at most $(e+\tau)/(1-\tau)$. We leave that analysis separate from the present IW certificate.

These results should be read as finite-sample certification guarantees rather than as evidence that pure IW certification is uniformly preferable in practice. The lower bound shows that the Bernstein second-moment dependence is not an artifact of the proof, while Proposition 3 identifies an additional quantile-specific failure mode: insufficient certified CDF mass can force the VaR bound to jump. This coverage-cliff interpretation is the main conceptual distinction from mean-risk PAC-Bayes OPL. The algorithm demonstrates that the bound leads to an optimizable objective, but the experiments show that pure IW PAC-Bayes VaR is not always the best finite-sample learner. The split-sample DR-CDF extension makes the model-assisted path rigorous under conservative conditions and helps

in the rare-catastrophe stress test, but it is not a claim of uniform practical superiority. Next steps are tighter model-assisted CDF certification, including self-normalized or baseline-corrected CDF estimators, explicit cross-fitting, nuisance-error, monotonicity, and overlap-aware policy-class control. These directions could reduce vacuity, but each would need its own posterior-uniform concentration argument before being treated as theorem-covered.

A.3 Certification scope

Table 2: Certification scope. Each row states whether the corresponding reported object is covered by the PAC-Bayes VaR analysis and identifies the formal result or additional condition needed for the guarantee.

Object	Coverage	VaR-specific?	Formal support / scope
PB-OffPol-VaR	Yes	Yes	Posterior-averaged per-policy VaR; Lemma 1, Theorem 2, and Corollaries 1–2.
Level-shift VaR certificate	Yes	Yes	Distribution-free fallback without inverse-slope regularity; Theorem 3.
Gibbs-mixture quantile	Conditional	Yes	Conservative common-threshold route for a different estimand; Proposition 2.
Deterministic center	Conditional	Yes	Requires the policy-to-CDF stability condition in Assumption 3; Proposition 4.
Split-DR-PB	Conditional	Yes	Requires sample splitting and bounded CDF pilots; Theorem 5.
PB-Mean	No, comparator	No	Mean-risk PAC-Bayes baseline [19, 26, 1]; included to compare objectives, not claimed as a VaR certificate.
DM/DR/SNIPS/OPRA	No, comparator	Comparator	Empirical baselines [10, 11, 17, 14]; not covered by this work's PAC-Bayes VaR certificate.

A.4 Practical certification guidance

Practical tuning. The bandwidth ε, threshold mesh, and Bernstein grid are certificate parameters as well as optimization hyperparameters. We use three rules. First, choose ε small enough that smoothing bias is acceptable but large enough that the Newton solve and grid bracketing are stable; in continuous losses we tune over a finite predeclared grid and add the corresponding union-bound cost. Second, use a threshold mesh no coarser than the desired displayed VaR precision; the proof already accounts for the mesh through N_η. Third, use a geometric Bernstein grid inside $(0, 3/W)$, because $r_{n,\Lambda}$ is an infimum over Λ and a geometric grid is within a constant factor of the oracle scale. If M candidate triples $(\varepsilon, \mathcal{Y}, \Lambda)$ are selected after inspecting the same certification data, use the allocation in Corollary 2. In particular, without empirical-variance substitution one may replace δ by δ/M, while with the variance UCB the implementation uses $\delta/(2M)$ for the CDF event and $\delta/(2M)$ for the variance event. This finite-selection rule is conservative but prevents hidden tuning from invalidating the confidence statement. Exact IW-CDFs are monotone because H is monotone and the weights are nonnegative; prefix or isotonic projection is therefore redundant analytically for IW, but harmless for finite-grid Monte Carlo estimates when it uses only thresholds to the left of the reported threshold.

Choosing and reporting the inverse-slope parameter. The parameter m is an assumption parameter that encodes local CDF growth near the target quantile. It should be justified from domain knowledge, a known noise model, a deliberately smoothed evaluation protocol, or a separate density-lower-bound analysis, rather than tuned to minimize the reported certificate. The implementation reports local quantile-spacing and mass diagnostics because they are useful for detecting obvious violations of Assumption 2, but these diagnostics alone are not a coverage-valid estimate of m. A conservative reporting practice is to show the certificate over a small predeclared grid of plausible m values, or to use Theorem 3 when no defensible lower-slope bound is available. If m is learned from the same certification data, its lower-confidence bound must be included as an additional event in the union bound; that refinement is separate from the main theorem.

Non-vacuity accounting. A density-based certificate is non-vacuous when the right-hand side of Theorem 2 is strictly below the trivial support bound B (or a user-specified safety threshold). A level-shift certificate is non-vacuous when the returned threshold is strictly below $B + \varepsilon$. On discrete OpenML classification-to-bandit losses [34], the inverse-slope certificate is not the formal certificate unless the evaluation protocol jitters or smooths the losses; consequently, reporting an inverse-slope non-vacuity rate there would be a surrogate statistic. The supplement therefore treats

non-vacuity as a diagnostic computed from certified thresholds and empirical VaR gaps. We define the certified-to-empirical gap as $q_{\text{cert}} - \widehat{Q}_{\text{test}}$ and regard negative gaps as evidence of implementation error, sampling error, or mismatch between the certified and test estimands, since a valid upper certificate should not fall below the evaluated target.

Self-normalization and learned propensities. Self-normalized CDFs can reduce finite-sample mass deficit, but their ratio form is not covered by Lemma 1. A conservative route is to certify numerator and denominator simultaneously: if the self-normalizing denominator is at least $1-\tau$ and the numerator CDF error is at most e, then the induced CDF error is at most $(e+\tau)/(1-\tau)$. Mean-OPE variance-reduction tools such as SWITCH and optimal baseline corrections [37, 12] suggest useful CDF-estimator designs, but adapting them to VaR still requires threshold-uniform concentration followed by the inversion or level-shift step. Learned propensities require a separate uniform propensity-error certificate; without it, the theorem is conditional on the propensity model.

Proposition 1 (A sufficient stability bound for linear softmax policies)**.** *Suppose the action set has size A, the policy is linear softmax $\pi_\theta(a \mid x) = \mathrm{softmax}_a(\theta_1^\top \phi(x), \dots, \theta_A^\top \phi(x))$, and $\|\phi(x)\|_2 \le R$ almost surely. Viewing θ as the concatenated parameter matrix, Assumption 3 holds with $L_\Theta = \sqrt{A}R/2$. This bound is often loose, but it gives a checkable deterministic-center penalty for simple softmax classes.*

Proof. Deferred to Section F.12.

B Formal setup

We use losses. Let $X \in \mathcal{X}$, $A \in \mathcal{A}$ with finite $|\mathcal{A}|$, and $Y \in [0, B]$. Logged data $\mathcal{D}_n = \{(X_i, A_i, Y_i)\}_{i=1}^n$ are i.i.d. from $X \sim P_X$, $A \sim \pi_0(\cdot \mid X)$, $Y \sim P(\cdot \mid X, A)$, where the logging propensities $\pi_0(A_i \mid X_i)$ are known. A target policy π_θ has importance weight

$$w_\theta(x, a) = \frac{\pi_\theta(a \mid x)}{\pi_0(a \mid x)}, \qquad V_w(\theta) = \mathbb{E}_{P_X \pi_0}\big[w_\theta(X, A)^2\big].$$

The loss distribution under π_θ has CDF $F_\theta(y) = \mathbb{P}_{X, A\sim\pi_\theta, Y}(Y \le y)$. For $p = 1-\alpha$, define the upper-tail VaR

$$Q_p(\theta) = \inf\{y : F_\theta(y) \ge p\}.$$

For a prior π over Θ and posterior $\rho \ll \pi$, the Gibbs procedure samples $\theta \sim \rho$ and then acts according to π_θ. The quantity certified in this work is the posterior average of per-policy VaRs,

$$\mathcal{R}_p(\rho) = \mathbb{E}_{\theta\sim\rho} Q_p(\theta).$$

For comparison, the outcome distribution induced by first drawing $\theta \sim \rho$ has mixture CDF

$$\bar{F}_\rho(y) = \mathbb{E}_{\theta\sim\rho} F_\theta(y), \qquad \bar{Q}_p(\rho) = \inf\{y : \bar{F}_\rho(y) \ge p\}.$$

Because VaR is nonlinear, $\bar{Q}_p(\rho)$ and $\mathcal{R}_p(\rho)$ generally differ. Our main inverse-slope theorem certifies $\mathcal{R}_p(\rho)$; the distribution-free level-shift theorem gives a posterior-average exceedance statement and does not claim a single common-threshold certificate for $\bar{Q}_p(\rho)$ unless explicitly stated.

Assumption 1 (Bounded importance weights)**.** There exists $W < \infty$ such that $0 \le w_\theta(x, a) \le W$ for every x, a, θ.

This assumption is checkable: if the logging policy has uniform overlap $\pi_0(a \mid x) \ge \mu > 0$ for all logged actions, then Assumption 1 holds for every target policy with $W \le 1/\mu$. If overlap is enforced only after clipping, the theorem certifies the clipped target distribution and an explicit clipping-bias term must be added; we report the unclipped maximum weight, effective sample size, and clipped mass in the experiments.

Assumption 2 (Quantile inverse slope)**.** There exists $m > 0$ such that, for every θ and every $y \in [0, B]$,

$$Q_p(\theta) \le y + \frac{(p - F_\theta(y))_+}{m}. \tag{2}$$

Assumption 2 is the exact regularity needed to convert CDF error into quantile error. It is implied by a standard density condition: if F_θ is continuous at $Q_p(\theta)$ with $F_\theta(Q_p(\theta)) = p$ and has density at least m on every interval $[y, Q_p(\theta)]$ for $y \le Q_p(\theta)$, then (2) follows by integration. The condition is only needed for $y \le B$; if a smoothed empirical threshold lies above B, then $Q_p(\theta) \le B \le y$ and the inverse-slope inequality is automatic. In local-density statements we use the conservative effective slope $m = f_0/2$ after restricting to the certified local interval. Without a lower slope near the quantile, $1/\sqrt{n}$ CDF concentration cannot imply $1/\sqrt{n}$ quantile concentration. For atoms or zero-one losses, we use Theorem 3 instead of Theorem 2.

Smoothed IW-CDF. Let

$$K(u) = \frac{15}{16}(1-u^2)^2 \mathbb{1}\{|u| \le 1\}, \qquad H(u) = \int_{-\infty}^{u} K(s)ds.$$

Define

$$\widetilde{F}_{\theta,\varepsilon}(y) = \mathbb{E}_{\pi_0}\big[w_\theta(X,A)H((y-Y)/\varepsilon)\big], \quad \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) = \frac{1}{n}\sum_{i=1}^{n} w_\theta(X_i,A_i)H((y-Y_i)/\varepsilon),$$

and define the right-clipped empirical smoothed quantile

$$\widehat{\widetilde{Q}}^{\mathrm{IW}}_{n,\varepsilon,p}(\theta) = \begin{cases} \inf\{y \in [-\varepsilon, B+\varepsilon] : \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) \ge p\}, & \text{if this set is nonempty,} \\ B+\varepsilon, & \text{otherwise.} \end{cases}$$

The clipping only matters when the unnormalized empirical IW mass $n^{-1}\sum_i w_\theta(X_i,A_i)$ falls below p. Compact support gives $F_\theta(y-\varepsilon) \le \widetilde{F}_{\theta,\varepsilon}(y) \le F_\theta(y+\varepsilon)$.

C Certificate results

In this section, the formal results are arranged as a short certificate pipeline. Lemma 1 first gives posterior-uniform control of the off-policy CDF. Theorem 2 is the main VaR certificate obtained by inverting that CDF bound under Assumption 2. The remaining statements explain how the certificate is made executable, how it behaves without density assumptions, and why the overlap dependence is unavoidable.

C.1 CDF control and VaR inversion

The next lemma gives the posterior-uniform CDF deviation used throughout the work. The finite threshold grid contributes an explicit logarithmic factor. To make the Bernstein optimization simultaneous over posteriors, we optimize over a finite grid $\Lambda \subset (0, 3/W)$ of Bernstein parameters. For bracketing only, write $W_+ = \max\{W, 1\}$; the Bernstein range remains W.

Lemma 1 (PAC-Bayes Bernstein CDF deviation)**.** *Assume $0 \le w_\theta \le W$. Let $W_+ = \max\{W,1\}$, let $K_\infty = \|K\|_\infty$, set $\eta = \varepsilon/(2W_+K_\infty n)$, and let $N_\eta \le \lceil (B+2\varepsilon)/\eta \rceil + 1$ be the grid size over $[-\varepsilon, B+\varepsilon]$. Let $\Lambda \subset (0, 3/W)$ be a finite nonempty set. With probability at least $1-\delta$, simultaneously for every posterior $\rho \ll \pi$,*

$$\mathbb{E}_\rho \sup_y \big(\widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) - \widetilde{F}_{\theta,\varepsilon}(y)\big) \le r_{n,\Lambda}(\rho), \tag{3}$$

$$\mathbb{E}_\rho \sup_y \big(\widetilde{F}_{\theta,\varepsilon}(y) - \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y)\big) \le r_{n,\Lambda}(\rho), \tag{4}$$

where

$$r_{n,\Lambda}(\rho) = \inf_{\lambda \in \Lambda}\left\{\frac{L_\rho}{\lambda n} + \frac{\lambda \mathbb{E}_\rho[V_w(\theta)]}{2(1-W\lambda/3)}\right\} + \frac{1}{n}, \qquad L_\rho = \mathrm{KL}(\rho\|\pi) + \log\frac{4N_\eta|\Lambda|}{\delta}.$$

If Λ is a fine geometric grid around the oracle Bernstein parameter, this is within a constant factor of

$$\sqrt{\frac{2\mathbb{E}_\rho[V_w(\theta)]L_\rho}{n}} + O\left(\frac{WL_\rho}{n}\right) + \frac{1}{n}.$$

Proof. Deferred to Section F.1.

Theorem 2 (PAC-Bayes off-policy posterior-averaged VaR)**.** *Under Assumptions 1 and 2, with probability at least $1-\delta$, simultaneously for every posterior $\rho \ll \pi$,*

$$\mathbb{E}_\rho Q_p(\theta) \le \mathbb{E}_\rho \widehat{\widetilde{Q}}^{\mathrm{IW}}_{n,\varepsilon,p}(\theta) + \varepsilon + \frac{2r_{n,\Lambda}(\rho)}{m}. \tag{5}$$

The factor 2 is a mass-deficit correction for unnormalized IW empirical CDFs. More precisely, the proof uses one radius for CDF error at the selected empirical threshold and one radius for the right-boundary clipping deficit. If the clipping deficit $\Delta_\theta = (p - \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(B+\varepsilon))_+$ is zero ρ-almost surely, the same proof yields the sharper penalty $r_{n,\Lambda}(\rho)/m$. In the common local-density specialization, $m = f_0/2$, so the leading VaR-scale term is of order $\sqrt{\mathbb{E}_\rho V_w(\theta)/n}/f_0$ up to logarithmic factors.

Proof. Deferred to Section F.2.

Theorem 2 should be read literally as a certificate for $\mathcal{R}_p(\rho) = \mathbb{E}_\rho Q_p(\theta)$. It does not by itself imply a bound on the mixture quantile $\bar{Q}_p(\rho)$; Proposition 2 gives a separate conservative common-threshold certificate for that different estimand.

A Hoeffding version would replace the second-moment term by $W/\sqrt{n}$, which can be looser by $\sqrt{W}$ in the rare-action regime. The Bernstein/PAC-Bayes-Bernstein route follows the same variance-sensitive spirit as empirical Bernstein and PAC-Bayes Bernstein inequalities [21, 31, 22], but here it is applied to an importance-weighted smoothed CDF and then inverted into a VaR bound.

Corollary 1 (Fully empirical second-moment substitution)**.** *Assume $0 \le w_\theta \le W$ and define*

$$\widehat{V}_w(\theta) = \frac{1}{n}\sum_{i=1}^{n} w_\theta(X_i, A_i)^2.$$

With probability at least $1-\delta_v$, simultaneously for every posterior $\rho \ll \pi$,

$$\mathbb{E}_\rho V_w(\theta) \le \mathbb{E}_\rho \widehat{V}_w(\theta) + W^2\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log(1/\delta_v)}{2n}}.$$

Consequently a fully observable version of Theorem 2 is obtained by replacing $\mathbb{E}_\rho V_w(\theta)$ inside $r_{n,\Lambda}$ by the right-hand side above. If the CDF event is run at level δ_{cdf} and this variance event is run at level δ_v, the resulting empirical-variance certificate holds with probability at least $1-\delta_{\mathrm{cdf}}-\delta_v$.

Proof. Deferred to Section F.3.

Corollary 2 (Finite post-hoc tuning over certificate parameters)**.** *Let $\mathcal{T}$ be a finite, predeclared set of certificate configurations, where each element may specify a bandwidth ε, a threshold mesh, a Bernstein grid Λ, and whether Corollary 1 is used. For each $t \in \mathcal{T}$, allocate failure probabilities $\delta_{\mathrm{cdf},t}$ and, if the empirical-variance substitution is used, $\delta_{v,t}$, such that*

$$\sum_{t\in\mathcal{T}} \big(\delta_{\mathrm{cdf},t} + \delta_{v,t}\big) \le \delta,$$

with $\delta_{v,t} = 0$ for configurations not using Corollary 1. Then all displayed certificates for all configurations and all posteriors hold simultaneously with probability at least $1-\delta$. Hence one may select among these finitely many displayed certificates after inspecting the certification data without invalidating coverage. The released code uses the conservative split $\delta_{\mathrm{cdf},t} = \delta_{v,t} = \delta/(2|\mathcal{T}|)$ when the variance UCB is enabled, and $\delta_{\mathrm{cdf},t} = \delta/|\mathcal{T}|$ otherwise.

Proof. Deferred to Section F.4.

C.2 Distribution-free certificates and the coverage cliff

When Assumption 2 is inappropriate, for example because the loss has atoms, the same CDF event still yields valid but usually more conservative level-shift certificates. This subsection separates three objects that are easily conflated: policy-specific posterior-average thresholds, a common Gibbs-mixture threshold, and the empirical mass-deficit condition that causes quantile certificates to jump or become vacuous.

Theorem 3 (Distribution-free level-shift certificate)**.** *Assume only $0 \le w_\theta \le W$. On the event of Lemma 1, fix $\rho \ll \pi$ with $p + 2r_{n,\Lambda}(\rho) < 1$ and define the posterior-dependent, right-clipped threshold*

$$\widehat{q}_{\theta,\rho} = \inf\{y \in [-\varepsilon, B+\varepsilon] : \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) \ge p + 2r_{n,\Lambda}(\rho)\},$$

with $\widehat{q}_{\theta,\rho} = B + \varepsilon$ if the set is empty. Then

$$\mathbb{E}_{\theta\sim\rho}\mathbb{P}_{\pi_\theta}(Y > \widehat{q}_{\theta,\rho} + \varepsilon) \le \alpha.$$

This is a posterior-average, policy-specific-threshold safety statement: the threshold $\widehat{q}_{\theta,\rho}$ may depend on θ. It is not a single-threshold certificate for the Gibbs-mixture CDF $\bar{F}_\rho$. A deterministic-policy statement requires an additional conversion argument such as Proposition 4.

Proof. Deferred to Section F.6.

Proposition 2 (Common-threshold Gibbs-mixture quantile certificate)**.** *On the event of Lemma 1, fix a posterior $\rho \ll \pi$ and define*

$$\widehat{q}_\rho^{\,\mathrm{mix}} = \inf\left\{y \in [-\varepsilon, B+\varepsilon] : \mathbb{E}_{\theta\sim\rho}\widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) \ge p + r_{n,\Lambda}(\rho)\right\},$$

with right-boundary clipping when the set is empty. Then

$$\bar{Q}_p(\rho) \le \widehat{q}_\rho^{\,\mathrm{mix}} + \varepsilon,$$

where $\bar{Q}_p(\rho) = \inf\{y : \bar{F}_\rho(y) \ge p\}$ is the quantile of the Gibbs-mixture outcome distribution. This gives a conservative single-threshold certificate for the deployed randomized Gibbs mixture. It is different from Theorem 2: it certifies a common threshold for $\bar{F}_\rho$, but it does not bound $\mathbb{E}_\rho Q_p(\theta)$ or provide an inverse-slope rate for that posterior-average objective.

Proof. Deferred to Section F.5.

Proposition 3 (No mass, no quantile certificate)**.** *On the event of Lemma 1, fix a posterior ρ and a threshold set $\mathcal{Y} \subset [-\varepsilon, B+\varepsilon]$. Suppose that for every $y \in \mathcal{Y}$,*

$$\mathbb{E}_{\theta\sim\rho}\widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) < p - r_{n,\Lambda}(\rho).$$

Then Lemma 1 alone cannot certify any threshold in $\mathcal{Y}$ as having posterior-average exceedance probability at most α. In particular, any level-shift certificate based only on this CDF event must choose a threshold outside $\mathcal{Y}$ or return the right-clipped boundary value. Equivalently, if the weighted empirical CDF has a posterior-average mass deficit larger than the deviation radius at level p, the quantile certificate necessarily jumps to a larger grid point.

Proof. Deferred to Section F.9.

Interpretation: the coverage cliff. Proposition 3 formalizes a phenomenon that is specific to quantiles. For mean risk, weak overlap generally appears as a larger variance term. For VaR, the same overlap weakness can also prevent the empirical CDF from reaching the target level. Once the certified CDF mass falls short of p, the bound must move to a higher threshold, often producing a sharp change in the certified quantile. This is why the algorithm reports empirical mass deficit in addition to ESS and maximum IW. The split-sample DR-CDF extension in Section D.2 can partially mitigate this effect by adding model-based CDF mass for unobserved actions, but it pays a larger bounded-score constant and requires an independent nuisance pilot.

Theorem 4 (Rare-action minimax lower bound)**.** *Fix $p \in (0,1)$. Suppose $B \ge 1/(2m)$, $W \ge 2$, and $n \ge W$. There is a two-action subclass $\mathcal{P}_{m,B,W,p}$ satisfying Assumptions 1–2 for a fixed deterministic target policy such that any estimator $\widehat{Q}$ of Q_p satisfies*

$$\inf_{\widehat{Q}} \sup_{P\in\mathcal{P}_{m,B,W,p}} \mathbb{E}_P|\widehat{Q} - Q_p(P)| \ge c_p \frac{\sqrt{W}}{m\sqrt{n}},$$

where $c_p > 0$ depends only on the quantile level p; if $p \in [p_0, 1-p_0]$ then c_p can be chosen as a numerical constant depending only on p_0.

Algorithm 1 PB-OffPol-VaR

Require: logged data, prior π, posterior family ρ_ϕ, level $p = 1 - \alpha$, bandwidth ε, confidence δ, iterations T

initialize posterior parameters ϕ_0
for $t = 0, \ldots, T-1$ **do**
  sample $\theta^{(1)}, \ldots, \theta^{(M)} \sim \rho_{\phi_t}$
  for $m = 1, \ldots, M$ **do**
    compute weights $w_i^{(m)} = \pi_{\theta^{(m)}}(A_i \mid X_i)/\pi_0(A_i \mid X_i)$
    compute $\widehat{\widetilde{Q}}^{\mathrm{IW}}_{n,\varepsilon,p}(\theta^{(m)})$ by differentiable Newton or bisection on the smoothed CDF
  end for
  form $\mathcal{B}_n(\rho_{\phi_t})$ using the Monte Carlo quantile average, $\mathrm{KL}(\rho_{\phi_t} \| \pi)$, and $\widehat{V}_w$ or its UCB
  update $\phi_{t+1} \leftarrow \phi_t - \eta_t \nabla_\phi \mathcal{B}_n(\rho_{\phi_t})$
end for
return Gibbs posterior ρ_{ϕ_T}; optionally also report its deterministic center with the stability penalty in Proposition 4

Proof. Deferred to Section F.10. As in classical overlap-hardness results for contextual-bandit OPE [37], the construction logs the target action with probability $1/W$. The target importance weight is W when the rare action is observed, so $V_w = W$. A two-point perturbation of the target-action loss distribution moves the target quantile by order $1/(m\sqrt{n/W})$ while the logged experiments remain Le Cam-indistinguishable. Thus Theorem 2's $\sqrt{V_w/n}/m$ dependence is sharp up to constants and logarithmic factors in this rare-action regime.

Executable variance and tuning. Theorem 2 is stated with $V_w(\theta)$ to show the clean variance-sensitive rate. In executable certification, the implementation uses Corollary 1 to replace $\mathbb{E}_\rho V_w$ by a PAC-Bayes upper confidence bound on the empirical second moment $\mathbb{E}_\rho \widehat{V}_w$. If ε, the threshold mesh, or Λ are chosen from a finite grid after inspecting the certification data, the code uses the δ/M correction in Corollary 2. The displayed certificate total is $\widehat{q} + \varepsilon + 2r/m$; omitting the $+\varepsilon$ term would understate the formal smoothing certificate.

D Learning rule and extensions

The theory certifies a Gibbs posterior, whereas implementations usually optimize a smooth objective and often report both the posterior and a deterministic center. This section makes that translation explicit: Algorithm 1 optimizes the displayed PAC-Bayes certificate, Corollary 3 certifies the optimizer only with respect to the displayed bound, Theorem 5 gives a split-sample DR-CDF variant, and Proposition 4 gives the extra penalty needed to report a deterministic center.

D.1 Optimizing the certificate

For a Gaussian posterior family ρ_ϕ over policy parameters, PB-OffPol-VaR minimizes the certified or surrogate objective

$$\mathcal{B}_n(\rho_\phi) = \mathbb{E}_{\theta \sim \rho_\phi} \widehat{\widetilde{Q}}^{\mathrm{IW}}_{n,\varepsilon,p}(\theta) + \varepsilon + \mathrm{pen}_n(\rho_\phi), \tag{6}$$

where pen_n is the mass-safe Bernstein penalty $2r_{n,\Lambda}(\rho)/m$ from Theorem 2. In theorem-aligned mode, the expectation is estimated by Monte Carlo samples from ρ_ϕ and the empirical variance UCB is used. In fast deterministic-center mode, one evaluates the posterior mean and treats the objective as a practical surrogate unless Proposition 4's stability condition is justified.

The smoothed CDF is differentiable in the policy logits through the IW weights. The implementation uses backpropagation through an unrolled Newton solve for the smoothed quantile and includes a finite-difference gradient check.

Corollary 3 (Certified optimizer, not an oracle inequality)**.** *Let $\mathcal{R}$ be the posterior class optimized by Algorithm 1, and let*

$$\mathcal{B}_n(\rho) = \mathbb{E}_{\theta \sim \rho} \widehat{\widetilde{Q}}^{\mathrm{IW}}_{n,\varepsilon,p}(\theta) + \varepsilon + 2r_{n,\Lambda}(\rho)/m.$$

If $\widehat{\rho}$ satisfies $\mathcal{B}_n(\widehat{\rho}) \le \inf_{\rho\in\mathcal{R}} \mathcal{B}_n(\rho) + \mathrm{Opt}_T$, then, on the event of Theorem 2,

$$\mathbb{E}_{\widehat{\rho}} Q_p(\theta) \le \inf_{\rho\in\mathcal{R}} \mathcal{B}_n(\rho) + \mathrm{Opt}_T.$$

This statement is intentionally weaker than a classical oracle inequality comparing to $\inf_\rho \mathbb{E}_\rho Q_p(\theta)$. Such a comparison would require additional symmetric quantile stability or anti-concentration conditions for empirical quantiles, which are not assumed here.

Proof. Deferred to Section F.8.

D.2 Split-sample model-assisted DR-CDF extension

The empirical results in Section E show a finite-sample limitation of pure IW certification: direct-method and doubly-robust estimators [10, 11], as well as OPRA-style plugin baselines [14], can be much stronger in finite samples. We therefore add a conservative model-assisted extension, while keeping the IW-only Theorem 2 as the main result. The role of this extension is to show how the same PAC-Bayes CDF machinery can be applied to a lower-variance score under a clean independence condition, not to establish dominance over model-assisted OPE baselines. Split the logged sample into an independent nuisance sample D_{nuis} and certification sample D_{cert}. Fit a pilot smoothed CDF $\hat{m}_y(x,a) \in [0,1]$ on D_{nuis} only, freeze it, and on D_{cert} define

$$Z_{\theta,\hat{m}}(y) = \sum_a \pi_\theta(a \mid X)\hat{m}_y(X,a) + w_\theta(X,A)\{H((y-Y)/\varepsilon) - \hat{m}_y(X,A)\}.$$

Conditional on the nuisance sample and on any fixed $\hat{m}$, correct propensities give

$$\mathbb{E}_{\pi_0}[Z_{\theta,\hat{m}}(y) \mid D_{\mathrm{nuis}}] = \widetilde{F}_{\theta,\varepsilon}(y), \qquad Z_{\theta,\hat{m}}(y) \in [-W, 1+W].$$

Thus the score is unbiased for the smoothed target CDF even when the pilot is misspecified; the pilot only changes variance. The bounded range is larger than for pure IW, so the constants must use $R_{\mathrm{DR}} = 1 + 2W$.

Theorem 5 (Split-sample DR-CDF level-shift certificate)**.** *Condition on D_{nuis} and assume D_{cert} is independent of D_{nuis}, $\hat{m}_y(x,a) \in [0,1]$, correct logging propensities, and $0 \le w_\theta \le W$. The prior π may depend on D_{nuis} but not on D_{cert}; the posterior may be any distribution chosen after observing D_{cert}. Let $n_c = |D_{\mathrm{cert}}|$, let $\mathcal{Y}$ be a finite threshold grid containing the right boundary $B+\varepsilon$, set $R_{\mathrm{DR}} = 1+2W$, and define*

$$\widehat{F}^{\mathrm{DR}}_{n_c,\theta}(y) = n_c^{-1} \sum_{i\in D_{\mathrm{cert}}} Z_{i,\theta,\hat{m}}(y), \qquad S_{\mathrm{DR}}(\theta,y) = \mathbb{E}[Z_{\theta,\hat{m}}(y)^2 \mid D_{\mathrm{nuis}}].$$

For a finite grid $\Lambda_{\mathrm{DR}} \subset (0, 3/R_{\mathrm{DR}})$, define

$$r^{\mathrm{DR}}_{n_c,\mathcal{Y}}(\rho) = \inf_{\lambda\in\Lambda_{\mathrm{DR}}} \left\{ \frac{\mathrm{KL}(\rho\|\pi) + \log(2|\mathcal{Y}||\Lambda_{\mathrm{DR}}|/\delta)}{\lambda n_c} + \frac{\lambda\, \mathbb{E}_\rho \sup_{y\in\mathcal{Y}} S_{\mathrm{DR}}(\theta,y)}{2(1 - R_{\mathrm{DR}}\lambda/3)} \right\}.$$

With probability at least $1-\delta$ over D_{cert}, simultaneously for every posterior $\rho \ll \pi$ and every $y \in \mathcal{Y}$,

$$\mathbb{E}_\rho \widetilde{F}_{\theta,\varepsilon}(y) \ge \mathbb{E}_\rho \widehat{F}^{\mathrm{DR}}_{n_c,\theta}(y) - r^{\mathrm{DR}}_{n_c,\mathcal{Y}}(\rho).$$

Let $\widehat{G}^{\mathrm{DR}}_\rho(y) = \sup_{t\in\mathcal{Y}:t\le y} \mathbb{E}_\rho \widehat{F}^{\mathrm{DR}}_{n_c,\theta}(t)$ be the prefix-monotone projection on the grid. If

$$\widehat{q}^{\mathrm{DR}}_\rho = \inf\{y \in \mathcal{Y} : \widehat{G}^{\mathrm{DR}}_\rho(y) \ge p + r^{\mathrm{DR}}_{n_c,\mathcal{Y}}(\rho)\},$$

with right-boundary clipping to $B+\varepsilon$ when the set is empty, then

$$\mathbb{E}_{\theta\sim\rho} \Pr_{\pi_\theta}(Y > \widehat{q}^{\mathrm{DR}}_\rho + \varepsilon) \le \alpha$$

for grid thresholds. If one reports a continuous threshold between grid points, add the grid mesh to the displayed threshold.

Proof. Deferred to Section F.7.

Theorem 5 is a level-shift certificate: it avoids density inversion, is valid for atoms, and is rigorous only with an independently fitted bounded nuisance pilot. The implementation reports Split-DR-PB-OffPol-VaR by optimizing a differentiable DR-CDF surrogate and then recomputing the certificate with the prefix-monotone grid rule in Theorem 5. Same-sample nuisance fitting, cross-fitting, and deterministic-center deployment require additional analysis or the separate stability penalty in Proposition 4; we therefore treat them as empirical extensions rather than claims of the theorem.

D.3 Deterministic-center reporting

The certified object remains the Gibbs posterior. Reporting the posterior mean as a deterministic policy is a separate step, and it is valid only with a policy-to-CDF stability argument that links nearby policy parameters to nearby outcome CDFs.

Assumption 3 (Policy-to-CDF stability)**.** For the deterministic center $\bar{\theta}$ of a posterior ρ, there exists $L_\Theta < \infty$ such that

$$\sup_y |F_\theta(y) - F_{\bar{\theta}}(y)| \le L_\Theta \|\theta - \bar{\theta}\|_2 \quad \text{for } \rho\text{-almost every } \theta.$$

This condition is model-dependent and not automatic. For bounded features, finite actions, smooth softmax policies, and logging overlap, it can be upper-bounded from Lipschitz constants of the policy map and conditional CDFs, but the resulting constant may be too loose. We therefore do not claim a certified deterministic-center guarantee for an arbitrary neural policy class unless a numerical upper bound on L_Θ is justified; without such a bound, Proposition 4 is only a conditional conversion result.

Proposition 4 (Deterministic-center conversion)**.** *Under Assumptions 2 and 3, for any posterior ρ with center $\bar{\theta}$,*

$$Q_p(\bar{\theta}) \le \mathbb{E}_\rho Q_p(\theta) + \frac{L_\Theta}{m} \mathbb{E}_{\theta\sim\rho} \|\theta - \bar{\theta}\|_2.$$

Consequently, on the event of Theorem 2,

$$Q_p(\bar{\theta}) \le \mathbb{E}_\rho \widehat{\widetilde{Q}}^{\mathrm{IW}}_{n,\varepsilon,p}(\theta) + \varepsilon + \frac{2r_{n,\Lambda}(\rho)}{m} + \frac{L_\Theta}{m} \mathbb{E}_{\theta\sim\rho} \|\theta - \bar{\theta}\|_2.$$

For an isotropic Gaussian posterior $\rho = \mathcal{N}(\bar{\theta}, \sigma^2 I_d)$, the last expectation is at most $\sigma\sqrt{d}$.

Proof. Deferred to Section F.11.

Target interpretation on discrete losses. The inverse-slope theorem is most natural for continuous losses with nontrivial local mass around the target quantile. On atoms, ties, or unjittered classification-to-bandit losses, the formal certificate is Theorem 3 (or Theorem 5 for the split-DR construction). In that regime, the differentiable smoothed-quantile objective should be interpreted as an optimization surrogate unless an explicit jittering or smoothed-loss evaluation protocol is imposed.

E Experiment summary

The experiments illustrate when the certificates are non-vacuous and when the coverage-cliff terms dominate the bound. The evaluation separates: (i) theorem-aligned non-vacuity on continuous losses, (ii) VaR-objective behavior on discrete classification-to-bandit tasks, and (iii) coverage-cliff diagnostics when conservative certificates are vacuous. The resulting picture is useful for certification: the density certificate is executable on continuous losses, while discrete and weak-overlap settings expose where the valid level-shift and mixture certificates become vacuous.

Baselines. We compare logging, IPS and SNIPS [10, 17], DM/DR [10, 11], OPRA-style plugin and DR-CDF VaR learners [14], PB-Mean [19, 26, 1], PB-CVaR [23], PB-OffPol-VaR, clipped variants, Split-DR-PB, and a non-deployable Oracle-VaR. The external OPE and OPRA-style methods are strong practical comparators, but they are not certificates from Theorem 2.

Synthetic concentration. A lightweight two-action continuous-loss simulation with known target quantile checks the expected scaling of the IW quantile error. Across $n \in \{200, 500, 1000, 2000, 5000, 10000\}$, the log-log slope is close to $-1/2$, consistent with the inverse-slope rate in Theorem 2; see Figure 2.

Continuous-loss formal certificate check. On a heteroskedastic continuous-loss bandit with known propensities, 20 seeds, empirical variance UCBs, and finite-tuning corrections, Table 6 shows non-vacuous density certificates for all PAC-Bayes objectives. For PB-OffPol-VaR, certified VaR is 4.532 ± 0.019 versus empirical VaR 2.160, with gap 2.373, ESS about 2266, and zero mass deficit. This is the main positive certification experiment: the theorem can be executed non-vacuously, but remains conservative.

Classification-to-bandit benchmarks. Table 3 reports classification-to-bandit experiments derived from OpenML datasets [34]. It shows that PB-OffPol-VaR improves $\mathrm{VaR}_{0.90}$ over PB-Mean by about 0.057 on Letter, 0.075 on Pendigits, and 0.021 on Satimage, supporting the objective. Since the losses are unjittered and discrete, Table 7 reports the valid level-shift and mixture diagnostics; these certificates are vacuous in our sample sizes. DM/DR/SNIPS/OPRA-style baselines [10, 11, 17, 14] are often better, and Split-DR-PB is mixed.

Rare-catastrophe stress test. Table 4 shows that PB-OffPol-VaR improves $\mathrm{VaR}_{0.95}$ and $\mathrm{CVaR}_{0.90}$ over PB-Mean; PB-CVaR [23] and Split-DR-PB improve CVaR further; OPRA-style CDF methods [14] remain the strongest practical baselines. This supports the tail-aware objective while leaving model-assisted CDF methods as stronger finite-sample comparators in this stress test.

Coverage and non-vacuity diagnostics. All runs report ESS, maximum IW, second moments, empirical mass deficit, KL, finite-tuning-corrected penalties, non-vacuity indicators, and certified-to-empirical gaps. Table 7 shows vacuous valid certificates on discrete/stress-test tasks even when empirical VaR improves, which is the empirical face of the coverage cliff: the certificate refuses to certify a tail threshold without reliable weighted CDF mass at level p. Table 8 shows that on continuous losses, $\epsilon/B \in \{0.01, 0.03, 0.05, 0.10\}$ keeps the density certificate non-vacuous and changes certified VaR only mildly from about 5.44 to 5.49.

Positioning. PAC-Bayes offline contextual-bandit work mainly certifies expected loss or pessimistic IW objectives [19, 26, 1, 27]; classical OPE gives mean-risk estimators, overlap lower bounds, and confidence intervals [10, 11, 37, 17, 16]. Distributional OPE methods control CDFs, quantiles, or predictive sets in related settings [14, 7, 28, 29]. Our narrower contribution is a PAC-Bayes posterior-uniform certificate for posterior-averaged per-policy VaR with explicit inversion, mass-deficit diagnostics, and a rare-logging lower bound.

E.1 Additional experimental details

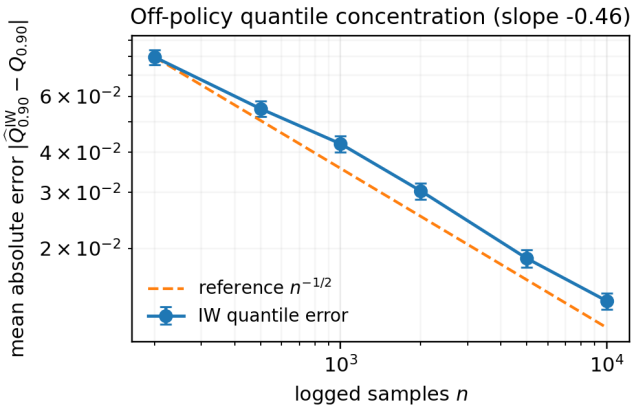

Figure 2: Synthetic concentration sanity check. In a two-action continuous-loss setting with known target quantile and target-action logging probability $1/W$, the IW empirical quantile error follows the expected $n^{-1/2}$ trend.

Table 3: OpenML classification-to-bandit results. Lower $\mathrm{VaR}_{0.90}$ is better. Entries are mean(SE) over 10 seeds. "Best" is the best non-oracle practical comparator among DM, DR, SNIPS, IPS, and OPRA-style CDF methods; it contextualizes performance and is not a PAC-Bayes certificate. "Oracle" is a non-deployable privileged lower bound that selects the true label on each test example, so its loss is zero by construction on these classification-to-bandit tasks.

Data	Log	PB-M	PB-VaR	Split	Best	Oracle
Letter	2.084(.004)	2.344(.006)	2.288(.017)	2.338(.009)	DM: 2.112(.011)	0 (priv.)
Pendigits	2.042(.005)	2.227(.006)	2.153(.014)	2.298(.019)	DR: 1.235(.061)	0 (priv.)
Satimage	2.108(.008)	2.157(.011)	2.137(.007)	2.241(.015)	DM: 1.719(.011)	0 (priv.)

Table 4: Rare-catastrophe stress test. Lower is better. Entries are mean(SE) over 20 seeds. The table has a diagnostic role: mean-oriented IPS/SNIPS have poor high-tail metrics, PB-OffPol-VaR improves over PB-Mean on high-tail metrics, Split-DR-PB improves further in this stress test, and OPRA-style CDF methods remain stronger practical baselines.

Method	$\mathrm{VaR}_{0.90}$	$\mathrm{VaR}_{0.95}$	$\mathrm{CVaR}_{0.90}$
IPS-Mean	1.818(.305)	3.445(.002)	3.356(.033)
SNIPS-Mean	1.966(.310)	3.449(.002)	3.399(.017)
PB-Mean	1.133(.001)	1.284(.003)	2.124(.011)
PB-OffPol-VaR	1.135(.000)	1.271(.001)	2.065(.003)
PB-CVaR	1.133(.000)	1.242(.001)	1.895(.007)
Split-DR-PB	1.109(.004)	1.221(.003)	1.841(.012)
OPRA-DR	0.565(.002)	0.634(.028)	0.748(.069)
OPRA-plugin	0.561(.000)	0.734(.138)	0.885(.090)
Oracle	0.397(.000)	0.478(.000)	0.498(.000)

Table 5 reports the paired tests supporting the limited empirical claims in Section E.

Table 5: Paired improvements over PB-Mean. Positive values mean lower tail risk than PB-Mean on the indicated metric. These tests support the limited empirical claim that PB-OffPol-VaR improves over the closest PAC-Bayes mean baseline, while Split-DR-PB is mixed on OpenML and helpful on the rare-catastrophe CVaR metric.

Dataset/metric	Method	Improvement	95% CI	paired t p
Letter $\mathrm{VaR}_{0.90}$	PB-OffPol-VaR	0.057	$[0.020, 0.097]$	2.41×10^{-2}
Letter $\mathrm{VaR}_{0.90}$	Split-DR-PB	0.007	$[-0.010, 0.024]$	4.89×10^{-1}
Pendigits $\mathrm{VaR}_{0.90}$	PB-OffPol-VaR	0.075	$[0.053, 0.098]$	2.37×10^{-4}
Pendigits $\mathrm{VaR}_{0.90}$	Split-DR-PB	-0.071	$[-0.101, -0.040]$	2.07×10^{-3}
Satimage $\mathrm{VaR}_{0.90}$	PB-OffPol-VaR	0.021	$[0.010, 0.033]$	8.44×10^{-3}
Satimage $\mathrm{VaR}_{0.90}$	Split-DR-PB	-0.083	$[-0.103, -0.066]$	1.69×10^{-5}
Rare-cat. $\mathrm{CVaR}_{0.90}$	PB-OffPol-VaR	0.059	$[0.043, 0.073]$	4.02×10^{-7}
Rare-cat. $\mathrm{CVaR}_{0.90}$	PB-CVaR	0.228	$[0.216, 0.241]$	1.18×10^{-18}
Rare-cat. $\mathrm{CVaR}_{0.90}$	Split-DR-PB	0.283	$[0.249, 0.316]$	1.14×10^{-12}
Rare-cat. $\mathrm{CVaR}_{0.90}$	OPRA-DR	1.375	$[1.227, 1.475]$	1.97×10^{-14}

Figure 3 reports two diagnostic plots used to audit the coverage-cliff mechanism. The plots are not an empirical universality claim; they show how the certified-to-empirical VaR gap varies with empirical IW mass deficit and realized importance-weight range in the precomputed PB-family runs. The formal statement is Proposition 3.

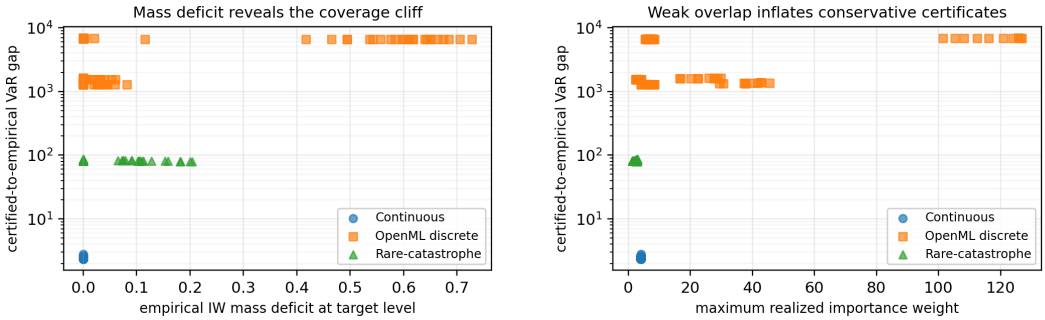

Figure 3: Certificate diagnostics from the precomputed PB-family runs. Left: empirical IW mass deficit versus certified-to-empirical VaR gap. Right: maximum realized importance weight versus the same gap. The logarithmic vertical scale makes both non-vacuous continuous certificates and vacuous discrete/weak-overlap certificates visible.

Full non-vacuity and sensitivity diagnostics. The implementation writes a per-run `results.csv` and a `nonvacuity_summary.csv` containing density-certificate, level-shift, mixture-certificate, empirical VaR, ESS, maximum IW, and mass-deficit diagnostics. Tables 6–8 report the full precomputed runs, with constant binary indicators summarized in the captions rather than repeated as table columns. Table 6 is the theorem-aligned certification experiment: the density/inverse-slope certificate from Theorem 2 is non-vacuous for all 20 seeds, although conservative. The common-threshold mixture certificate from Proposition 2 returns the right-boundary value for all 20 seeds under these constants, so the table focuses on the density certificate and overlap diagnostics. Table 7 is a limitation diagnostic for unjittered discrete OpenML tasks and the rare-catastrophe stress test. Because these losses need not satisfy Assumption 2, the formal certificates are the valid level-shift and mixture certificates; both are non-vacuous in zero seeds in these runs. The table therefore reports empirical VaR and overlap diagnostics rather than vacuous right-boundary thresholds. Table 8 varies the smoothing bandwidth on the continuous-loss benchmark; the density certificate remains non-vacuous for every seed and bandwidth, while the mixture certificate remains vacuous under the same constants.

Table 6: Full continuous-loss non-vacuity diagnostics over 20 seeds, using the empirical-variance UCB and finite-grid confidence allocation. This benchmark is aligned with the inverse-slope theorem. "Dens. nonvac." reports the number of seeds for which Theorem 2 gives a nontrivial density certificate. The common-threshold mixture certificate in Proposition 2 returns the right-boundary value for all 20 seeds under these constants.

Method	Dens. nonvac.	Cert. VaR	Emp. VaR	Gap	ESS	Def.
PB-Mean	20/20	4.675 ± 0.027	2.145	2.530	2251.5	0
PB-OffPol-VaR	20/20	4.532 ± 0.019	2.160	2.373	2266.3	0
PB-CVaR	20/20	4.595 ± 0.018	2.107	2.488	2155.7	0

Table 7: Theorem-aligned diagnostics on unjittered discrete OpenML tasks and the rare-catastrophe stress test. The valid level-shift and common-threshold mixture certificates are vacuous for every seed in these runs: $0/10$ on each OpenML row and $0/20$ on each rare-catastrophe row. The table therefore reports empirical VaR, effective sample size, mass deficit, and maximum realized importance weight instead of the corresponding right-boundary certificate values.

Dataset	Method	Emp. VaR	ESS	Def.	$w_{\max}$
Letter	PB-Mean	2.344	2193	0.641	6.7
Letter	PB-OffPol-VaR	2.288	4292	0.327	6.7
Letter	Split-DR-PB	2.338	412	0.002	116.8
Pendigits	PB-Mean	2.227	4393	0.037	6.8
Pendigits	PB-OffPol-VaR	2.153	4727	0	6.5
Pendigits	Split-DR-PB	2.298	597	0	38.6
Satimage	PB-Mean	2.157	2919	0.029	3.5
Satimage	PB-OffPol-VaR	2.137	3063	0	2.8
Satimage	Split-DR-PB	2.241	656	0	23.8
Rare-cat.	PB-Mean	1.133	5111	0.120	2.4
Rare-cat.	PB-OffPol-VaR	1.135	5921	0	1.4
Rare-cat.	Split-DR-PB	1.109	4728	0	2.8

Table 8: Bandwidth sensitivity on the continuous-loss benchmark over 10 seeds. The finite-selection correction from Corollary 2 is applied. For every displayed bandwidth, the density certificate is non-vacuous in all 10 seeds, while the common-threshold mixture certificate is vacuous in all 10 seeds under these constants. Larger smoothing increases the certified VaR mildly through the explicit $+\epsilon$ term and grid effects.

ϵ/B	Emp. VaR	Cert. VaR	Gap	ESS	$w_{\max}$
0.01	2.293 ± 0.024	5.439 ± 0.025	3.146	1697.2 ± 24.0	3.90 ± 0.02
0.03	2.285 ± 0.024	5.443 ± 0.025	3.158	1674.2 ± 20.8	3.91 ± 0.02
0.05	2.284 ± 0.024	5.455 ± 0.025	3.171	1670.9 ± 20.7	3.92 ± 0.02
0.10	2.284 ± 0.024	5.494 ± 0.025	3.211	1667.2 ± 20.7	3.92 ± 0.02

F Proofs

This section collects all formal arguments. Each proof states the probabilistic event or reduction being used before applying it.

F.1 Proof of Lemma 1

Proof idea. For a fixed threshold, parameter, sign, and Bernstein parameter, apply a bounded Bernstein exponential inequality to the smoothed IW-CDF. A finite union bound makes the event simultaneous over signs, thresholds, and Bernstein parameters; the Donsker–Varadhan change of measure then makes it simultaneous over all posteriors. A Lipschitz grid argument extends the finite-threshold bound to all thresholds.

Fix $\lambda \in \Lambda$, a grid point y_j, a sign $s \in \{-1, +1\}$, and a parameter value θ. Let

$$U_i = w_\theta(X_i, A_i) H((y_j - Y_i)/\varepsilon).$$

Then $0 \le U_i \le W$ and $\mathrm{Var}(U_i) \le \mathbb{E}U_i^2 \le V_w(\theta)$. Bernstein's exponential moment bound for bounded variables [see, e.g., 5] gives, for every $0 < \lambda < 3/W$,

$$\mathbb{E}_{\mathcal{D}_n} \exp\left\{ \lambda n s\big(\widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y_j) - \widetilde{F}_{\theta,\varepsilon}(y_j)\big) - \frac{n\lambda^2 V_w(\theta)}{2(1 - W\lambda/3)} \right\} \le 1.$$

Summing over the $2N_\eta|\Lambda|$ triples (s, j, λ) and applying Fubini and Markov's inequality with threshold $4N_\eta|\Lambda|/\delta$, we obtain an event of probability at least $1 - \delta/2$ on which, for all $\lambda \in \Lambda$,

$$\mathbb{E}_{\theta\sim\pi} \sum_{s,j} \exp\left\{ \lambda n s\big(\widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y_j) - \widetilde{F}_{\theta,\varepsilon}(y_j)\big) - \frac{n\lambda^2 V_w(\theta)}{2(1 - W\lambda/3)} \right\} \le \frac{4N_\eta|\Lambda|}{\delta}.$$

The slack from $\delta/2$ is harmless and keeps the displayed logarithm simple. On this event, for any posterior $\rho \ll \pi$, the Donsker–Varadhan variational inequality [9, 6] gives

$$\lambda n \mathbb{E}_\rho \max_{s,j} s\big(\widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y_j) - \widetilde{F}_{\theta,\varepsilon}(y_j)\big) \le \mathrm{KL}(\rho\|\pi) + \log \frac{4N_\eta|\Lambda|}{\delta} + \frac{n\lambda^2 \mathbb{E}_\rho V_w(\theta)}{2(1 - W\lambda/3)}.$$

Dividing by λn and taking the infimum over $\lambda \in \Lambda$ gives the two one-sided bounds on the finite threshold grid. Finally, $y \mapsto \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y)$ is WK_∞/ε-Lipschitz and $y \mapsto \widetilde{F}_{\theta,\varepsilon}(y)$ is at most K_∞/ε-Lipschitz because $\mathbb{E}_{\pi_0} w_\theta(X, A) = 1$. Hence both are W_+K_∞/ε-Lipschitz. Since the grid mesh is $\eta = \varepsilon/(2W_+K_\infty n)$, the continuum supremum differs from the grid supremum by at most

$$2W_+K_\infty\eta/\varepsilon = 1/n.$$

Measurability follows because the continuum supremum is the limit of suprema over refining finite rational grids and all posterior expectations are over measurable bounded functions. The event has probability at least $1 - \delta/2$, and hence at least $1 - \delta$, so the lemma follows.

F.2 Proof of Theorem 2

Proof idea. At the empirical smoothed quantile, the empirical IW-CDF has mass at least p, except when the unnormalized IW mass fails to reach p at the right boundary. The PAC-Bayes CDF event transfers this empirical mass to population smoothed CDF mass. Kernel bracketing converts smoothed CDF mass at $\widehat{q}$ into ordinary CDF mass at $\widehat{q} + \varepsilon$, and Assumption 2 converts the remaining CDF deficit into a quantile error.

For each fixed θ, compact support of K implies $F_\theta(y-\varepsilon) \le \widetilde{F}_{\theta,\varepsilon}(y) \le F_\theta(y+\varepsilon)$. Let $\widehat{q}_\theta = \widehat{\widetilde{Q}}^{\mathrm{IW}}_{n,\varepsilon,p}(\theta)$, $D_\theta = \sup_y(\widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) - \widetilde{F}_{\theta,\varepsilon}(y))$, and $\Delta_\theta = (p - \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(B + \varepsilon))_+$. By the clipped definition, $\widehat{\widetilde{F}}_{n,\theta,\varepsilon}(\widehat{q}_\theta) \ge p - \Delta_\theta$. Therefore

$$\widetilde{F}_{\theta,\varepsilon}(\widehat{q}_\theta) \ge p - \Delta_\theta - D_\theta.$$

Since $\widetilde{F}_{\theta,\varepsilon}(\widehat{q}_\theta) \le F_\theta(\widehat{q}_\theta + \varepsilon)$, set $y_\theta = \widehat{q}_\theta + \varepsilon$. If $y_\theta > B$, then $Q_p(\theta) \le B < y_\theta$ and the desired inequality is immediate. Otherwise $y_\theta \le B$ and Assumption 2 gives

$$Q_p(\theta) \le \widehat{q}_\theta + \varepsilon + (D_\theta + \Delta_\theta)/m.$$

At $B + \varepsilon$ the population smoothed CDF equals one, so $\Delta_\theta \le \sup_y(\widetilde{F}_{\theta,\varepsilon}(y) - \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y))$. Averaging over $\theta \sim \rho$ and applying both one-sided inequalities in Lemma 1 proves the displayed $2r_{n,\Lambda}(\rho)/m$ result. If $\Delta_\theta = 0$ for ρ-almost every θ, then the second one-sided inequality is not used and the same argument gives the sharper $r_{n,\Lambda}(\rho)/m$ penalty.

F.3 Proof of Corollary 1

Proof idea. The unknown second moment $V_w(\theta)$ is itself the expectation of a bounded random variable, $w_\theta(X, A)^2$. A one-sided PAC-Bayes-Hoeffding bound upper-controls its posterior average by the empirical posterior average plus a KL-dependent confidence term.

Fix θ. Since $w_\theta(X, A)^2 \in [0, W^2]$, Hoeffding's exponential inequality applied to $V_w(\theta) - \widehat{V}_w(\theta)$ and the standard PAC-Bayes change of measure give, with probability at least $1 - \delta_v$,

$$\mathbb{E}_\rho V_w(\theta) - \mathbb{E}_\rho \widehat{V}_w(\theta) \le \inf_{\lambda>0} \left\{ \frac{\mathrm{KL}(\rho\|\pi) + \log(1/\delta_v)}{\lambda} + \frac{\lambda W^4}{8n} \right\}$$

simultaneously for every $\rho \ll \pi$. Optimizing in λ gives the displayed bound. Combining this event with Lemma 1 uses a union bound: if the CDF event has failure probability δ_{cdf} and the variance event has failure probability δ_v, the empirical-variance certificate has failure probability at most $\delta_{\mathrm{cdf}} + \delta_v$.

F.4 Proof of Corollary 2

Proof idea. Every candidate certificate configuration is valid on its own event. Allocating failure probabilities across a finite set and union-bounding the events makes all displayed certificates valid simultaneously, so selecting the smallest displayed certificate after inspection does not change the total failure probability.

For each configuration $t \in \mathcal{T}$, run the CDF event and, when used, the empirical-variance event with the allocated failure probabilities $\delta_{\mathrm{cdf},t}$ and $\delta_{v,t}$. Since

$$\sum_{t\in\mathcal{T}} (\delta_{\mathrm{cdf},t} + \delta_{v,t}) \le \delta,$$

a union bound implies that all requested CDF and variance events hold simultaneously with probability at least $1-\delta$. On this simultaneous event, every displayed certificate for every configuration and every posterior is valid. Any data-dependent choice among those already-valid displayed certificates therefore preserves the same coverage statement.

F.5 Proof of Proposition 2

Proof idea. Use the lower-CDF side of Lemma 1 at a single common threshold. If the posterior-average empirical smoothed CDF clears p plus the radius, then the posterior-average population CDF clears p after the $+\varepsilon$ smoothing shift, which is exactly a Gibbs-mixture quantile certificate.

On the event of Lemma 1, the one-sided lower-CDF inequality gives, for every y,

$$\mathbb{E}_\rho \widetilde{F}_{\theta,\varepsilon}(y) \ge \mathbb{E}_\rho \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) - r_{n,\Lambda}(\rho).$$

At $y = \widehat{q}^{\mathrm{mix}}_\rho$, the right-hand side is at least p by definition, unless the clipped boundary is returned, in which case the statement is trivial because $\bar{Q}_p(\rho) \le B \le B + 2\varepsilon$. Compact support of the smoothing kernel gives $\widetilde{F}_{\theta,\varepsilon}(y) \le F_\theta(y+\varepsilon)$. Hence

$$\bar{F}_\rho(y+\varepsilon) = \mathbb{E}_\rho F_\theta(y+\varepsilon) \ge \mathbb{E}_\rho \widetilde{F}_{\theta,\varepsilon}(y) \ge p,$$

which implies $\bar{Q}_p(\rho) \le y + \varepsilon$.

F.6 Proof of Theorem 3

Proof idea. The level-shift certificate avoids density assumptions by asking the empirical CDF to exceed p by two radii. One radius pays for lower-CDF concentration at the chosen threshold and the other pays for the possible right-boundary mass deficit. Smoothing then yields an exceedance guarantee at the shifted threshold.

Let $\Delta_{\theta,\rho} = (p + 2r_{n,\Lambda}(\rho) - \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(B+\varepsilon))_+$. Since $p + 2r_{n,\Lambda}(\rho) \le 1$ and $\widetilde{F}_{\theta,\varepsilon}(B+\varepsilon) = 1$, we have $\Delta_{\theta,\rho} \le \sup_y (\widetilde{F}_{\theta,\varepsilon}(y) - \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y))$. By the clipped definition, $\widehat{\widetilde{F}}_{n,\theta,\varepsilon}(\widehat{q}_{\theta,\rho}) \ge p + 2r_{n,\Lambda}(\rho) - \Delta_{\theta,\rho}$. Combining this with Lemma 1 gives

$$\mathbb{E}_\rho \widetilde{F}_{\theta,\varepsilon}(\widehat{q}_{\theta,\rho}) \ge p + 2r_{n,\Lambda}(\rho) - r_{n,\Lambda}(\rho) - r_{n,\Lambda}(\rho) = p.$$

Since $\widetilde{F}_{\theta,\varepsilon}(y) \le F_\theta(y+\varepsilon)$, we have $\mathbb{E}_\rho F_\theta(\widehat{q}_{\theta,\rho} + \varepsilon) \ge p$, which is equivalent to the claimed posterior-average exceedance bound.

F.7 Proof of Theorem 5

Proof idea. After conditioning on the independent nuisance sample, the DR-CDF score is a bounded unbiased estimator of the smoothed target CDF for every fixed policy. The same PAC-Bayes-Bernstein argument as Lemma 1 gives a posterior-uniform lower-CDF bound on the certification sample. The prefix-monotone threshold rule then turns that CDF lower bound into a posterior-average exceedance statement.

Condition on D_{nuis} and on any prior chosen as a function of D_{nuis}. Then $\hat{m}$ and the prior are fixed when the PAC-Bayes event is formed on the independent certification sample D_{cert}. For each $y \in \mathcal{Y}$ and θ, correct propensities imply, conditionally on X and D_{nuis},

$$\begin{aligned}\mathbb{E}[Z_{\theta,\hat{m}}(y) \mid X, D_{\text{nuis}}] &= \sum_a \pi_\theta(a \mid X)\hat{m}_y(X,a) \\ &\quad + \sum_a \pi_\theta(a \mid X)\{\mathbb{E}[H((y-Y)/\varepsilon) \mid X, A=a] - \hat{m}_y(X,a)\} \\ &= \sum_a \pi_\theta(a \mid X)\mathbb{E}[H((y-Y)/\varepsilon) \mid X, A=a].\end{aligned}$$

Averaging over X gives $\mathbb{E}_{\pi_0}[Z_{\theta,\hat{m}}(y) \mid D_{\text{nuis}}] = \widetilde{F}_{\theta,\varepsilon}(y)$. The boundedness assumptions give $Z_{\theta,\hat{m}}(y) \in [-W, 1+W]$, hence the centered variable has range at most $R_{\text{DR}} = 1+2W$ and variance at most $S_{\text{DR}}(\theta, y)$. Bernstein's exponential inequality therefore gives, for every $s \in \{-1,+1\}$, $y \in \mathcal{Y}$, and $0 < \lambda < 3/R_{\text{DR}}$,

$$\mathbb{E}\exp\left\{\lambda n_c s(\widehat{F}^{\text{DR}}_{n_c,\theta}(y) - \widetilde{F}_{\theta,\varepsilon}(y)) - \frac{n_c\lambda^2 S_{\text{DR}}(\theta,y)}{2(1-R_{\text{DR}}\lambda/3)}\right\} \le 1.$$

Summing over signs, grid points, and $\lambda \in \Lambda_{\text{DR}}$, then applying Markov's inequality and the Donsker–Varadhan change of measure exactly as in Lemma 1, gives the displayed lower-CDF inequality simultaneously for all posteriors and grid points. The use of $\sup_{y\in\mathcal{Y}} S_{\text{DR}}(\theta, y)$ is a conservative simplification that makes a single posterior radius valid for all grid points.

If $\widehat{q}^{\text{DR}}_\rho$ is selected by the prefix-monotone rule, then either it is the right-clipped boundary or there exists $t \in \mathcal{Y}$ with $t \le \widehat{q}^{\text{DR}}_\rho$ and $\mathbb{E}_\rho \widehat{F}^{\text{DR}}_{n_c,\theta}(t) \ge p + r^{\text{DR}}_{n_c,\mathcal{Y}}(\rho)$. The CDF event implies $\mathbb{E}_\rho \widetilde{F}_{\theta,\varepsilon}(t) \ge p$. Since $\widetilde{F}_{\theta,\varepsilon}(t) \le F_\theta(t+\varepsilon) \le F_\theta(\widehat{q}^{\text{DR}}_\rho + \varepsilon)$, we obtain $\mathbb{E}_\rho F_\theta(\widehat{q}^{\text{DR}}_\rho + \varepsilon) \ge p$, equivalently $\mathbb{E}_\rho \Pr(Y > \widehat{q}^{\text{DR}}_\rho + \varepsilon) \le \alpha$. If the reported threshold is rounded from a continuous value to a neighboring grid point, the grid mesh must be added.

F.8 Proof of Corollary 3

Proof idea. The theorem bounds the true posterior-averaged VaR by the displayed certificate for every posterior in the optimized class. Applying the bound to the returned posterior and using the optimization suboptimality assumption proves the result.

On the event of Theorem 2, every posterior $\rho \in \mathcal{R}$ satisfies

$$\mathbb{E}_\rho Q_p(\theta) \le \mathbb{E}_\rho \widehat{\widetilde{Q}}^{\text{IW}}_{n,\varepsilon,p}(\theta) + \varepsilon + 2r_{n,\Lambda}(\rho)/m = \mathcal{B}_n(\rho).$$

Applying this inequality to $\widehat{\rho}$ and then using the optimization guarantee $\mathcal{B}_n(\widehat{\rho}) \le \inf_{\rho\in\mathcal{R}} \mathcal{B}_n(\rho) + \text{Opt}_T$ proves the result. No reverse comparison between $\mathcal{B}_n(\rho)$ and $\mathbb{E}_\rho Q_p(\theta)$ is used.

F.9 Proof of Proposition 3

Proof idea. A CDF-based quantile certificate needs certified population CDF mass at least p. If the empirical posterior-average CDF is below p by more than the deviation radius at every candidate threshold, the PAC-Bayes event can only certify population mass below p on that set.

The one-sided lower-CDF guarantee gives

$$\mathbb{E}_\rho \widetilde{F}_{\theta,\varepsilon}(y) \le \mathbb{E}_\rho \widehat{\widetilde{F}}_{n,\theta,\varepsilon}(y) + r_{n,\Lambda}(\rho) < p$$

for every $y \in \mathcal{Y}$. Therefore the PAC-Bayes CDF event does not imply posterior-average CDF mass at least p at any such threshold. Since a VaR or level-shift certificate requires certified CDF mass at level p up to the smoothing and grid slack, no threshold in $\mathcal{Y}$ can be certified from this event alone.

F.10 Proof of Theorem 4

Proof idea. Construct two bandit instances that differ only on a target action logged with probability $1/W$. Their target-action loss distributions have densities bounded below by m but have p-quantiles

separated by order $\sqrt{W/n}/m$. Because the rare branch is observed only about n/W times, the two logged experiments remain close in Hellinger distance, and Le Cam's method gives the lower bound.

We give an explicit two-point construction. Let there be one context and two actions a_- and a_+. The behavior policy satisfies

$$\pi_0(a_-) = 1/W, \qquad \pi_0(a_+) = 1 - 1/W,$$

and the target policy always selects a_-. Thus $w(a_-) = W$, $w(a_+) = 0$, and $V_w = W$. The conditional loss distribution of a_+ is identical in the two experiments and arbitrary, say a point mass at zero. Only the conditional distribution of the target action a_- differs.

Let $U \in [0, 1]$ and define a piecewise-constant function

$$\psi_p(u) = (1-p)\mathbb{1}\{u \le p\} - p\mathbb{1}\{u > p\}.$$

Then $\int_0^1 \psi_p(u)du = 0$, $|\psi_p(u)| \le 1$, and $\int_0^p \psi_p(u)du = p(1-p)$. For $0 < \gamma \le 1/2$, define two densities on $[0, 1]$ by

$$g_+(u) = 1 + \gamma\psi_p(u), \qquad g_-(u) = 1 - \gamma\psi_p(u).$$

They are valid densities and satisfy $1/2 \le g_\pm \le 3/2$. Let $Y = U/(2m)$. The corresponding loss densities on $[0, 1/(2m)] \subseteq [0, B]$ are

$$f_\pm(y) = 2m\, g_\pm(2my).$$

Hence $f_\pm(y) \ge m$ on their support, so Assumption 2 holds with constant m for both target-action distributions.

Let $q_\pm$ denote the p-quantile of Y under $f_\pm$. For the $+$ instance,

$$q_+ = \frac{p}{2m(1 + \gamma(1-p))}.$$

For the $-$ instance the p-quantile lies to the right of $p/(2m)$ and solving the piecewise-linear CDF gives

$$q_- = \frac{1}{2m}\left[p + \frac{\gamma p(1-p)}{1+\gamma p}\right].$$

Therefore, using $\gamma \le 1/2$,

$$|q_- - q_+| \ge \frac{\gamma p(1-p)}{2m}.$$

It remains to show that the logged experiments are statistically close. Let P_+ and P_- be the one-sample logged distributions. Since only the rare branch a_- differs,

$$H^2(P_+, P_-) = \frac{1}{W} H^2(g_+, g_-),$$

where H^2 denotes squared Hellinger distance between the corresponding conditional densities. Using $(\sqrt{x} - \sqrt{y})^2 \le (x-y)^2/(x+y)$ and $g_+ + g_- = 2$, we obtain

$$H^2(g_+, g_-) \le \int_0^1 \frac{(g_+(u) - g_-(u))^2}{g_+(u) + g_-(u)} du = 2\gamma^2 \int_0^1 \psi_p(u)^2 du = 2\gamma^2 p(1-p) \le \frac{\gamma^2}{2}.$$

For n i.i.d. samples, tensorization gives

$$H^2(P_+^{\otimes n}, P_-^{\otimes n}) \le nH^2(P_+, P_-) \le \frac{n\gamma^2}{2W}.$$

Choose $\gamma = c_0\sqrt{W/n}$ with $c_0 = 1/2$. Since $n \ge W$, this gives $\gamma \le 1/2$ and $H^2(P_+^{\otimes n}, P_-^{\otimes n}) \le 1/8$, so the total variation distance between the two n-sample experiments is at most $1/2$ by $\mathrm{TV} \le H$. Le Cam's two-point inequality [see, e.g., 32] yields

$$\inf_{\widehat{Q}} \sup_{P \in \{P_+, P_-\}} \mathbb{E}_P|\widehat{Q} - Q_p(P)| \ge \frac{|q_- - q_+|}{4}\big(1 - \mathrm{TV}(P_+^{\otimes n}, P_-^{\otimes n})\big) \ge c_p \frac{\sqrt{W}}{m\sqrt{n}},$$

with, for example, $c_p = p(1-p)/64$ after the conservative constants above. This proves the stated lower bound.

F.11 Proof of Proposition 4

Proof idea. At the p-quantile of a sampled policy, that policy has CDF mass at least p. Policy-to-CDF stability says the deterministic center has almost as much CDF mass at the same threshold; the inverse-slope condition converts the shortfall into an additive quantile penalty.

Fix θ and set $y = Q_p(\theta)$. Since $F_\theta(y) \geq p$, Assumption 3 gives $F_{\bar\theta}(y) \geq p - L_\Theta \|\theta - \bar\theta\|_2$. Applying Assumption 2 to the center policy $\bar\theta$ yields

$$Q_p(\bar\theta) \leq y + \frac{L_\Theta}{m}\|\theta - \bar\theta\|_2 = Q_p(\theta) + \frac{L_\Theta}{m}\|\theta - \bar\theta\|_2.$$

Averaging over $\theta \sim \rho$ proves the first claim. The second claim follows by substituting Theorem 2. For $\rho = \mathcal{N}(\bar\theta, \sigma^2 I_d)$, Jensen's inequality gives $\mathbb{E}\|\theta - \bar\theta\|_2 \leq (\mathbb{E}\|\theta - \bar\theta\|_2^2)^{1/2} = \sigma\sqrt{d}$.

F.12 Proof of Proposition 1

Proof idea. The CDF difference between two softmax policies is at most the expected total-variation distance between their action distributions. A linear-logit softmax map is Lipschitz, so bounded features convert parameter distance into CDF distance.

Let $F_a(y \mid x) = \Pr(Y \leq y \mid X = x, A = a)$. Then

$$F_\theta(y) = \mathbb{E}_X \sum_{a=1}^{A} \pi_\theta(a \mid X) F_a(y \mid X).$$

Since $0 \leq F_a(y \mid x) \leq 1$,

$$|F_\theta(y) - F_{\bar\theta}(y)| \leq \mathbb{E}_X \|\pi_\theta(\cdot \mid X) - \pi_{\bar\theta}(\cdot \mid X)\|_1.$$

The Jacobian of the softmax map is $J(z) = \mathrm{diag}(p) - pp^\top$, the covariance matrix of a categorical one-hot vector with probabilities $p = \mathrm{softmax}(z)$. Its spectral norm is at most $1/2$, so the softmax map is $1/2$-Lipschitz from ℓ_2 to ℓ_2. Therefore

$$\|\pi_\theta(\cdot \mid x) - \pi_{\bar\theta}(\cdot \mid x)\|_1 \leq \sqrt{A}\,\|\pi_\theta(\cdot \mid x) - \pi_{\bar\theta}(\cdot \mid x)\|_2 \leq \frac{\sqrt{A}}{2}\|z_\theta(x) - z_{\bar\theta}(x)\|_2 \leq \frac{\sqrt{A}R}{2}\|\theta - \bar\theta\|_2.$$

Taking the supremum over y proves Assumption 3 with the stated constant.

F.13 Proof of the checkable assumption reductions

If $\pi_0(a \mid x) \geq \mu > 0$ for all x, a, then $w_\theta(x, a) = \pi_\theta(a \mid x)/\pi_0(a \mid x) \leq 1/\mu$ because $\pi_\theta(a \mid x) \leq 1$. Thus Assumption 1 holds with $W = 1/\mu$.

For the inverse-slope reduction, suppose F_θ is continuous at $q = Q_p(\theta)$, $F_\theta(q) = p$, and F_θ has density $f_\theta(t) \geq m$ on $[y, q]$ for any $y \leq q$. If $y \geq q$, (2) is immediate. If $y < q$, then

$$p - F_\theta(y) = F_\theta(q) - F_\theta(y) = \int_y^q f_\theta(t)dt \geq m(q - y),$$

so $q \leq y + (p - F_\theta(y))/m$. This proves Assumption 2. If there is an atom at q, the equality $F_\theta(q) = p$ can fail and VaR may jump under arbitrarily small CDF perturbations; this is precisely why Theorem 3 is included.

G Implementation and reproducibility notes

The experiments use a PyTorch implementation of the compact polynomial smoothed CDF, differentiable optimization through an unrolled Newton quantile solver, and finite-difference gradient checks. The synthetic experiments use known logging propensities and controlled overlap. The classification-to-bandit experiments use OpenML-derived Letter, Pendigits, and Satimage configurations, while the rare-catastrophe stress test uses a synthetic tail-risk construction.

For theorem-aligned certification, the continuous-loss experiments recompute the displayed certificate using posterior Monte Carlo samples, the empirical-variance UCB from Corollary 1, and the finite-selection correction from Corollary 2. Reported diagnostics include empirical VaR, certified VaR

when applicable, effective sample size, maximum realized importance weight, empirical IW mass deficit, and non-vacuity indicators.

Some OpenML and fast deterministic configurations use a single posterior sample or deterministic center for computational reasons. These runs are treated as empirical diagnostics unless the corresponding posterior certificate or the stability penalty in Proposition 4 is reported. Same-sample nuisance fitting, cross-fitting, learned propensities, clipping without an explicit bias term, and deterministic-center deployment are not covered by the main theorem unless the additional conditions stated in the paper are verified.