Conservative neural posterior estimation via distributionally robust training

William Laplante*
University College London, The Alan Turing Institute
william.laplante.24@ucl.ac.uk

Yuga Hikida*
Aalto University
yuga.hikida@aalto.fi

Charita Dellaporta
University College London
h.dellaporta@ucl.ac.uk

François-Xavier Briol
University College London
f.briol@ucl.ac.uk

Ayush Bharti
Aalto University
ayush.bharti@aalto.fi

Abstract

Simulation-based inference with neural posterior estimation (NPE) often yields overconfident and unreliable posteriors under limited simulation budgets. To address this, we propose DRO-NPE, a distributionally robust approach that replaces the standard NPE objective with a worst-case loss over a Wasserstein ambiguity set. We introduce KL-based metrics for miscoverage and miscalibration, and use these to show that the DRO-NPE objective controls overfitting and reduces posterior overconfidence. Our method is tractable, parallelisable, and readily integrates with standard normalising flows. Across benchmark SBI tasks, DRO-NPE consistently improves coverage and calibration, while narrowing the gap between empirical and population NPE loss, leading to more reliable inference in low-simulation regimes.

1 Introduction

Simulation-based inference (SBI; Cranmer et al., 2020) is a powerful framework for inferring parameters of scientific models whose likelihood functions are unavailable or computationally prohibitive to evaluate, but for which simulating data is straightforward. The use of flexible neural conditional density estimators has substantially expanded the applicability of SBI to challenging problems, especially in fields such as particle physics (Brehmer, 2021), cognitive neuroscience (Fengler et al., 2021), economics (Dyer et al., 2024) and cosmology (Alsing et al., 2018; Jeffrey et al., 2021). Neural SBI methods rely on simulations from the scientific model to approximate intractable quantities such as the posterior, the likelihood, the likelihood-to-evidence ratio, or the score function; see Zammit-Mangion et al. (2024) for a recent review. In this work, we focus on the widely used neural posterior estimation (NPE) method (Papamakarios and Murray, 2016; Radev et al., 2022).

A central practical limitation of NPE is the simulation budget required to train the conditional density estimator. As many scientific simulators are expensive to run, generating a sufficiently large training set is often the main computational bottleneck. Prior work has sought to reduce simulation cost using sequential training schemes (Papamakarios and Murray, 2016; Lueckmann et al., 2017; Greenberg et al., 2019), multi-fidelity simulations (Hikida et al., 2025; Krouglova et al., 2026; Tatsuoka et al., 2025; Saoulis et al., 2025), or by exploiting heterogeneous simulation costs (Bharti et al., 2025). However, these strategies do not directly address the challenge of ensuring reliable uncertainty estimates in low-simulation regimes. In practice, the adequacy of a given simulation budget is rarely clear *a priori*, and neural SBI methods trained on limited simulations can exhibit poor uncertainty quantification, leading Hermans et al. (2022) to warn of the potential for a "trust crisis".

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Under finite simulations, NPE can yield *overconfident* posterior approximations whose credible regions contain the true parameter less often than their nominal coverage suggests. Such undercoverage is especially problematic in scientific applications, where overstated certainty can lead to misleading conclusions. By contrast, conservative posteriors are often safer as they widen uncertainty statements rather than overstating confidence. This has motivated several works on conservative uncertainty quantification in neural SBI: Delaunoy et al. (2022) proposed a regularised version of neural ratio estimation (NRE; Hermans et al., 2020) to inflate estimator uncertainty, and Delaunoy et al. (2023) extended this idea to contrastive NRE (Miller et al., 2022) and NPE. Falkiewicz et al. (2023) instead regularise NPE directly with a coverage error term. However, these methods remain largely heuristic: their regularisation terms encourage conservativeness without guarantees on uncertainty quantification error, potentially leading to overly conservative and uninformative posteriors. Instead of regularising NPE, Delaunoy et al. (2024) use Bayesian neural networks for conservative inference, at the cost of Bayesian inference over network weights. Complementarily, conformal and Neyman-inversion methods (Masserano et al., 2023; Patel et al., 2024; Cabezas et al., 2025) post-hoc calibrate credible sets with coverage guarantees, but do not directly address poor conditional density estimation.

In this paper, we view NPE through the lens of *empirical risk minimisation*. Standard NPE minimises the average loss over a finite set of simulated parameter–data pairs (the empirical risk), which approximates the expected loss (the *population risk*). When the simulation budget is limited, the empirical risk can be a poor proxy for the population risk, leading to overfitting and poorly calibrated posteriors (Hermans et al., 2022). This is illustrated in Figure 1 on Lotka–Volterra model: standard NPE (■) is concentrated away from the true parameter, while a wider reference NPE posterior trained with many more simulations (■) assigns mass around it. We make NPE conservative using *distributionally robust optimisation* (DRO; Kuhn et al., 2025), a framework for optimisation under distributional uncertainty. Our method (DRO-NPE; ■) yields a broader posterior that covers the true parameter, reflecting uncertainty from the limited simulation budget.

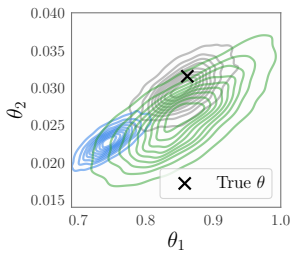


Figure 1: Lotka–Volterra posteriors in a low-simulation regime; details in Section 5.

Our contributions are as follows. First, we introduce Kullback–Leibler (KL) divergence-based uncertainty-quantification metrics (Section 3) that generalise standard notions of coverage and calibration errors (Hermans et al., 2022; Talts et al., 2018) and connect them directly to the NPE objective. Second, we propose a Wasserstein DRO objective that upper-bounds these metrics and the population NPE loss, so that, under suitable conditions, minimising it controls both posterior-fitting and uncertainty-quantification error (Theorem 4.1). We also derive a computationally cheaper regularised objective that upper-bounds the DRO objective (Proposition 4.2), thus retaining desirable theoretical properties, and verify its assumptions for standard normalising-flow architectures used in NPE (Lemma 4.3). Third, we establish excess-risk, miscoverage, and miscalibration guarantees for DRO-NPE, and compare them to those of standard NPE (Appendix B.3). Finally, we show that DRO-NPE improves uncertainty quantification and generalisation across challenging benchmarks (Section 5.1) and a real-world cosmology application (Section 5.3).

2 Background

We introduce NPE and uncertainty-quantification metrics revealing its finite-simulation failure modes. We use subscripts on distributions and CDFs to denote the random variables in question and any conditioning variables, and superscript on CDFs to indicate the underlying distribution.

Simulation-based inference with neural posterior estimation. Let $\mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$ and $\Theta \subseteq \mathbb{R}^{d_\Theta}$ denote the data and parameter spaces, and let $\mathcal{Z} := \mathcal{X} \times \Theta \subseteq \mathbb{R}^{d_\mathcal{Z}}$, where $d_\mathcal{Z} = d_\mathcal{X} + d_\Theta$. Consider a prior Π on Θ with density π and a conditional distribution $P_{\boldsymbol{x}|\boldsymbol{\theta}}$ from which one can simulate, but whose density $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, and therefore the likelihood, cannot be evaluated pointwise. Together, Π and $P_{\boldsymbol{x}|\boldsymbol{\theta}}$ induce a joint distribution $P_{\boldsymbol{x},\boldsymbol{\theta}}$, denoted by P for simplicity, with marginal $P_{\boldsymbol{x}}$ and posterior $P_{\boldsymbol{\theta}|\boldsymbol{x}}$. The goal of SBI is to approximate this posterior distribution $P_{\boldsymbol{\theta}|\boldsymbol{x}_0}$ given observed data $\boldsymbol{x}_o$.

NPE is an amortised SBI method that approximates the posterior using flexible conditional density estimators such as mixture-density networks (Bishop, 1994) and normalising flows (Papamakarios et al., 2021). Let $q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$ denote such an estimator with parameter $\phi \in \Phi \subseteq \mathbb{R}^{d_\Phi}$, and let $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ be

the corresponding distribution (we remove the explicit dependence on ϕ for notational simplicity). Assuming $P_{\boldsymbol{\theta}|\boldsymbol{x}}$ is absolutely continuous with respect to $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$, the NPE objective is obtained by minimising the expected KL divergence between $P_{\boldsymbol{\theta}|\boldsymbol{x}}$ and $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ over ϕ:

$$\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{KL}\Big(P_{\boldsymbol{\theta}|\boldsymbol{x}} \parallel Q_{\boldsymbol{\theta}|\boldsymbol{x}}\Big)\right] \overset{+K(P)}{=} \mathbb{E}_{(\boldsymbol{\theta},\boldsymbol{x})\sim P}[-\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})] := \mathcal{L}_{\mathrm{NPE}}(q_\phi; P), \tag{1}$$

where the additive constant $K(P)$ does not depend on ϕ. Note that $\mathcal{L}_{\mathrm{NPE}}(q_\phi; P)$ can be thought of as a *population risk*, as it is defined as an expectation under P, induced by the *pointwise loss* $\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}) := -\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$, where $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{\theta})$. In practice, $\mathcal{L}_{\mathrm{NPE}}(q_\phi; P)$ is approximated using independent and identically distributed (i.i.d.) samples $\boldsymbol{z}_{1:n} \sim P$, where each $\boldsymbol{z}_i$ is obtained by first sampling $\boldsymbol{\theta}_i \sim \Pi$, and then simulating $\boldsymbol{x}_i \sim P_{\boldsymbol{x}|\boldsymbol{\theta}_i}$. These samples induce the empirical measure $P_n := {}^1\!/_n \sum_{i=1}^n \delta_{\boldsymbol{z}_i}$, such that $\mathcal{L}_{\mathrm{NPE}}(q_\phi; P_n) := {}^1\!/_n \sum_{i=1}^n \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}_i)$ is the associated *empirical risk*. The conditional density model is then trained by minimising the empirical risk, $\hat{\phi}_n := \arg\min_{\phi\in\Phi} \mathcal{L}_{\mathrm{NPE}}(q_\phi; P_n)$. As a result, the estimator $q_{\hat{\phi}_n}$ is amortised: for new observed data $\boldsymbol{x}_o$, the posterior is approximated by a simple forward pass, namely $q_{\hat{\phi}_n}(\boldsymbol{\theta} \mid \boldsymbol{x}_o) \approx p(\boldsymbol{\theta} \mid \boldsymbol{x}_o)$, without requiring retraining.

Uncertainty quantification metrics in SBI. Despite its widespread success, NPE remains prone to failures in uncertainty quantification. We focus on the two most widely used diagnostics for assessing this: expected coverage of high-probability density region (HPDR) (Ward et al., 2022; Hermans et al., 2022) and simulation-based calibration (Talts et al., 2018; Säilynoja et al., 2022; Modrák et al., 2023). Other diagnostics have also been studied (Zhao et al., 2021; Linhart et al., 2022, 2023; Lemos et al., 2023; Chung et al., 2024; Dey et al., 2025), but are less common in SBI.

Definition 2.1 (Expected coverage of HPDR)**.** *The $(1-\alpha)$ HPDR induced by q at $\boldsymbol{x}$ is the region where $q(\boldsymbol{\theta} \mid \boldsymbol{x})$ is highest while containing $1-\alpha$ of its mass. Formally, $R^q_\alpha(\boldsymbol{x}) := \{\boldsymbol{\theta} \in \Theta : q(\boldsymbol{\theta} \mid \boldsymbol{x}) \geq \tau^q_\alpha(\boldsymbol{x})\}$, where $\tau^q_\alpha(\boldsymbol{x})$ is a threshold such that $Q_{\boldsymbol{\theta}|\boldsymbol{x}}(R^q_\alpha(\boldsymbol{x})) = 1-\alpha$. The expected coverage of the HPDR is then:*

$$C^P_\alpha(q) := \mathbb{E}_{\boldsymbol{z}\sim P}\left[\mathbf{1}\Big\{\boldsymbol{\theta} \in R^q_\alpha(\boldsymbol{x})\Big\}\right]. \tag{2}$$

Expected coverage of HPDR is estimated by repeatedly drawing $\boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{\theta}_i) \sim P$. For each $\boldsymbol{x}_i$, M posterior samples $\tilde{\boldsymbol{\theta}}_{i,1}, \ldots, \tilde{\boldsymbol{\theta}}_{i,M} \sim Q_{\boldsymbol{\theta}|\boldsymbol{x}_i}$ are used to evaluate the densities $q(\tilde{\boldsymbol{\theta}}_{i,m} \mid \boldsymbol{x}_i)$ for $m = 1, \ldots, M$, and the empirical α-quantile of these values approximates the threshold $\tau^q_\alpha(\boldsymbol{x}_i)$ defining the $(1-\alpha)$ HPDR. Expected coverage of HPDR is then assessed by whether the ground-truth parameter $\boldsymbol{\theta}_i$ satisfies $q(\boldsymbol{\theta}_i \mid \boldsymbol{x}_i) \geq \tau^q_\alpha(\boldsymbol{x}_i)$. We say $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ is conservative when $C^P_\alpha(q) > 1-\alpha$ and overconfident when $C^P_\alpha(q) < 1-\alpha$.

Definition 2.2 (Calibration w.r.t. S)**.** *Let $S : \Theta \times \mathcal{X} \to \mathbb{R}$ be a summary, and $Q_{S|\boldsymbol{x}}$, $P_{S|\boldsymbol{x}}$ be the distributions of $S(\boldsymbol{\theta}; \boldsymbol{x})$ under $\boldsymbol{\theta} \sim Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ and $\boldsymbol{\theta} \sim P_{\boldsymbol{\theta}|\boldsymbol{x}}$, respectively. Let the CDF of $Q_{S|\boldsymbol{x}}$ be denoted as $F^Q_{S|\boldsymbol{x}}$. We say that $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ (or equivalently q) is calibrated with respect to S if, for $S^P_{\boldsymbol{x}} \sim P_{S|\boldsymbol{x}}$, the random variable $F^Q_{S|\boldsymbol{x}}(S^P_{\boldsymbol{x}})$ is distributed as $\mathcal{U}[0,1]$ under $\boldsymbol{x} \sim P_{\boldsymbol{x}}$.*

Calibration with respect to S means that, under mild conditions on the CDF of $Q_{S|\boldsymbol{x}}$, the distribution of S under $P_{\boldsymbol{\theta}|\boldsymbol{x}}$ is the same as that of S under $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$, for $P_{\boldsymbol{x}}$-almost every $\boldsymbol{x}$. In practice, simulation-based calibration (SBC) is used to approximate calibration w.r.t. S by repeatedly drawing $(\boldsymbol{\theta}_i, \boldsymbol{x}_i) \sim P$ and computing $u_i := F^Q_{S|\boldsymbol{x}_i}\big(S(\boldsymbol{\theta}_i; \boldsymbol{x}_i)\big), i = 1, \ldots, n$. The empirical distribution of $u_{1:n}$ is then compared to $\mathcal{U}[0,1]$, for example via histograms, empirical CDFs, or summary statistics. The function S allows SBC to be applied in higher-dimensional settings. For example, $S(\boldsymbol{\theta}; \boldsymbol{x}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2$ for a reference point $\boldsymbol{\theta}_0$ is used in the *BayesFlow* library (Kühmichel et al., 2026). Note that expected coverage of HPDR is a special case of SBC when $S(\boldsymbol{\theta}; \boldsymbol{x}) = q(\boldsymbol{\theta} \mid \boldsymbol{x})$ (Delaunoy et al., 2022); see also Appendix A, Lemma A.1. Intuitively, the connection arises because coverage is simply calibration applied to a single summary evaluating whether $\boldsymbol{\theta}$ falls in the HPDR.

3 A new KL-based perspective on uncertainty quantification metrics

We consider errors in the expected coverage of HPDR and in calibration w.r.t. S, which we refer to as *miscoverage* of HPDR and *miscalibration* w.r.t. S, respectively. Superscripts on the miscoverage

and miscalibration metrics indicate the miscoverage level α and the summary statistic S, respectively. Miscoverage is often measured by the absolute deviation (Falkiewicz et al., 2023; Yao et al., 2024):

$$\Delta^{\alpha}_{\mathrm{cov}}(q;P) := \left|C^{P}_{\alpha}(q) - (1-\alpha)\right|.$$

However, this definition is blind to the prior, as the latter attains perfect expected coverage of HPDR, i.e., for any $\alpha \in (0,1)$, $C^{P}_{\alpha}(\pi) = \mathbb{E}_{\boldsymbol{z}\sim P}[\mathbf{1}\{\boldsymbol{\theta} \in R^{\pi}_{\alpha}\}] = \Pi(R^{\pi}_{\alpha}) = 1-\alpha$; therefore, $\Delta^{\alpha}_{\mathrm{cov}}(\pi;P) = 0$ for all valid α. This motivates an alternative notion of miscoverage. We adopt a KL-based approach because it removes prior blindness, links miscoverage and miscalibration to the NPE objective (Theorem 4.1), and provides a natural target for hyperparameter selection.

Definition 3.1 (KL-based miscoverage of HPDR)**.** *Fix $\alpha \in (0,1)$, and let $T_A(\boldsymbol{\theta}) := \mathbf{1}\{\boldsymbol{\theta} \in A\} \in \{0,1\}$ denote the indicator of a set $A \subseteq \Theta$. Let $(T_A)_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}}$ denote the pushforward of $P_{\boldsymbol{\theta}|\boldsymbol{x}}$ through T_A, i.e., the Bernoulli distribution of $T_A(\boldsymbol{\theta})$ when $\boldsymbol{\theta} \sim P_{\boldsymbol{\theta}|\boldsymbol{x}}$. Define $(T_A)_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ analogously. Then, KL-based miscoverage of HPDR is:*

$$\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) := \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{KL}\left((T_{R^{q}_{\alpha}(\boldsymbol{x})})_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}} \,\middle\|\, (T_{R^{q}_{\alpha}(\boldsymbol{x})})_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}\right)\right].$$

KL-based miscoverage of HPDR is weaker than the expected KL divergence between the full conditionals $P_{\boldsymbol{\theta}|\boldsymbol{x}}$ and $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ in Equation (1), since it compares only the Bernoulli distributions induced by whether $\boldsymbol{\theta}$ lies in the HPDR $R^{q}_{\alpha}(\boldsymbol{x})$. This follows from the data-processing inequality (Polyanskiy and Wu, 2025), which states that KL divergence cannot increase under pushforwards. Moreover, $\mathrm{kl}^{\alpha}_{\mathrm{cov}}$ upper-bounds $\Delta^{\alpha}_{\mathrm{cov}}$ via Pinsker's inequality: $\Delta^{\alpha}_{\mathrm{cov}}(q;P) \le \sqrt{1/2\,\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P)}$; see Lemma A.3. Small KL-based miscoverage of HPDR then implies small $\Delta^{\alpha}_{\mathrm{cov}}$, but not conversely: $\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P)$ may remain positive even when $\Delta^{\alpha}_{\mathrm{cov}}(q;P) = 0$, and, unlike $\Delta^{\alpha}_{\mathrm{cov}}$, $\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P)$ does not generally vanish for the prior. The latter would require the posterior to place the same mass as the prior on the HPDR for almost every $\boldsymbol{x}$.

The standard notion of CDF-based miscalibration quantifies deviation of the distribution of $F^{Q}_{S|\boldsymbol{x}}(S^{P}_{\boldsymbol{x}})$ from standard uniform (Talts et al., 2018; Gneiting et al., 2007). For the pushforward distributions $(F^{Q}_{S|\boldsymbol{x}})_{\#}P_{S|\boldsymbol{x}}$ and $(F^{Q}_{S|\boldsymbol{x}})_{\#}Q_{S|\boldsymbol{x}}$, where the latter is uniformly distributed under mild assumptions, this deviation is often measured via Kolmogorov–Smirnov distances or via Cramér–von Mises criteria (Rossi and Sekhposyan, 2019; Zhao et al., 2021; Falkiewicz et al., 2023). We introduce a KL-based alternative that captures deviation from uniformity and can be related to the population NPE risk.

Definition 3.2 (KL-based miscalibration w.r.t. S)**.** *For $S : \Theta \times \mathcal{X} \to \mathbb{R}$, KL-based miscalibration is:*

$$\mathrm{kl}^{S}_{\mathrm{cal}}(q;P) := \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{KL}\left((F^{Q}_{S|\boldsymbol{x}})_{\#}P_{S|\boldsymbol{x}} \,\middle\|\, (F^{Q}_{S|\boldsymbol{x}})_{\#}Q_{S|\boldsymbol{x}}\right)\right].$$

This notion is motivated by KL-based miscalibration discrepancies previously used in probabilistic regression (Utpala and Rai, 2020; Dheur and Taieb, 2023). As with $\mathrm{kl}^{\alpha}_{\mathrm{cov}}$, $\mathrm{kl}^{S}_{\mathrm{cal}}$ is weaker than the full KL divergence between $P_{\boldsymbol{\theta}|\boldsymbol{x}}$ and $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ by the data-processing inequality. The two nevertheless capture different notions of mismatch: for fixed α, $\mathrm{kl}^{\alpha}_{\mathrm{cov}}$ compares only the mass assigned to the HPDR, so $\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) = 0$ does not imply $P_{\boldsymbol{\theta}|\boldsymbol{x}} = Q_{\boldsymbol{\theta}|\boldsymbol{x}}$. By contrast, $\mathrm{kl}^{S}_{\mathrm{cal}} = 0$ implies equality of the induced distributions, and under mild conditions on $F^{Q}_{S|\boldsymbol{x}}$, equality of $P_{S|\boldsymbol{x}}$ and $Q_{S|\boldsymbol{x}}$. If S is a sufficient statistic for the model, this further implies $P_{\boldsymbol{\theta}|\boldsymbol{x}} = Q_{\boldsymbol{\theta}|\boldsymbol{x}}$. For $S(\boldsymbol{\theta};\boldsymbol{x}) = q(\boldsymbol{\theta} \mid \boldsymbol{x})$, under which coverage of HPDR is a special case of SBC, $\mathrm{kl}^{S}_{\mathrm{cal}}$ upper-bounds $\mathrm{kl}^{\alpha}_{\mathrm{cov}}$; see Lemma A.4.

4 Distributionally robust optimisation for NPE

NPE would ideally target the population risk $\mathcal{L}(q_\phi;P)$, but P is accessible only through the empirical measure P_n, which can be a poor approximation when n is not sufficiently large. In such settings, this leads to substantial performance degradation (Wang et al., 2025), and gives rise to the classical *generalisation gap* (Vapnik, 1991; Shalev-Shwartz and Ben-David, 2014). In particular, minimisation over P_n yields an overly optimistic estimate of the achievable population risk (Kuhn et al., 2019):

$$\mathbb{E}_{\boldsymbol{z}_{1:n}\overset{\text{i.i.d.}}{\sim}P}[\inf_{\phi}\mathcal{L}(q_\phi;P_n)] \le \inf_{\phi}\mathbb{E}_{\boldsymbol{z}_{1:n}\overset{\text{i.i.d.}}{\sim}P}[\mathcal{L}(q_\phi;P_n)] = \inf_{\phi}\mathcal{L}(q_\phi;P).$$

In NPE, this gap can translate into poor conditional density estimation. We therefore use distributionally robust optimisation (DRO) to minimise an upper bound on the population risk, thus accounting for distributional uncertainty (Kuhn et al., 2019; Rahimian and Mehrotra, 2022; Wang et al., 2025).

Motivating DRO-NPE. We formalise DRO-NPE as a distributionally robust objective based on the *Wasserstein* distance, defined as $W_p^p(P,Q) := \inf_{\pi\in\Pi(P,Q)} \int_{\mathcal{Z}\times\mathcal{Z}} \|\boldsymbol{z}-\boldsymbol{z}'\|^p \pi(d\boldsymbol{z}, d\boldsymbol{z}')$, where $\Pi(P,Q)$ denotes the set of all joint probability distributions of $\boldsymbol{z}$ and $\boldsymbol{z}'$ with marginals P and Q respectively, p is the order of the Wasserstein distance, and $\|\cdot\|$ is the Euclidean norm. DRO-NPE replaces the empirical NPE risk by its worst-case value over an *ambiguity set* $\mathcal{A}$ of distributions (Blanchet et al., 2024; Kuhn et al., 2025):

$$\inf_\phi \overbrace{\mathcal{L}_{\mathrm{NPE}}(q_\phi; P_n)}^{\text{empirical risk}} \xrightarrow{\text{risk aversion}} \inf_\phi \overbrace{\sup_{\tilde{P}\in\mathcal{A}} \mathcal{L}_{\mathrm{NPE}}(q_\phi; \tilde{P})}^{\text{DRO-NPE risk}}.$$

Here, we take $\mathcal{A}$ to be the p-Wasserstein ball of radius $\varepsilon \geq 0$ centred at P_n:

$$\mathcal{A}_p(P_n;\varepsilon) := \left\{\tilde{P}\in\mathcal{P}_p : W_p(\tilde{P}, P_n) \leq \varepsilon\right\}, \quad \mathcal{P}_p := \left\{\tilde{P} : \mathbb{E}_{\boldsymbol{z}\sim\tilde{P}}\left[\|\boldsymbol{z}-\boldsymbol{z}'\|^p\right] < \infty \ \ \forall \boldsymbol{z}'\in\mathcal{Z}\right\},$$

where $\mathcal{P}_p$ is the set of probability distributions with finite p-th moment. The radius $\varepsilon \geq 0$ controls the size of the ambiguity set (see Rahimian and Mehrotra, 2022; Kuhn et al., 2025): $\varepsilon = 0$ recovers the empirical risk, while larger values yield a more conservative objective by guarding against perturbations of P_n within the Wasserstein ball (Wang et al., 2025). Although many distances and divergences have been considered for defining ambiguity sets in DRO, Wasserstein balls are particularly natural here. For example, the commonly used f-divergence-based ambiguity sets (Hu and Hong, 2013; Bayraksan and Love, 2015; Husain et al., 2023) primarily re-weight the support of the empirical distribution. By contrast, Wasserstein balls are based on mass transport (Kuhn et al., 2019), and can contain distributions supported away from the observed simulations, including continuous distributions.

We now show that the population NPE risk upper bounds both $\mathrm{kl}_{\mathrm{cov}}^{\alpha}$ and $\mathrm{kl}_{\mathrm{cal}}^{S}$ by the data-processing inequality for the KL divergence (see Lemma A.2). Thus, if P were available, the NPE objective would provide a principled criterion for controlling these forms of posterior uncertainty. In finite-simulation settings, however, P is unavailable, so we instead construct a high-probability upper bound on the population risk. To do so, we rely on Wasserstein concentration results (Fournier and Guillin, 2015; Kuhn et al., 2019), stating that for a suitable radius ε decreasing with n, the true distribution P belongs to the ambiguity set $\mathcal{A}_p(P_n;\varepsilon)$ with high probability. This directly implies that the DRO-NPE objective upper-bounds the population NPE risk, and therefore also $\mathrm{kl}_{\mathrm{cov}}^{\alpha}$ and $\mathrm{kl}_{\mathrm{cal}}^{S}$; see Appendix B.1 for the proof.

Theorem 4.1. *For any $S : \Theta\times\mathcal{X}\to\mathbb{R}$ and q_ϕ:*

$$\mathrm{kl}_{\mathrm{cal}}^{S}(q_\phi;P) \overset{+K(P)}{\leq} \mathcal{L}_{\mathrm{NPE}}(q_\phi;P) \quad \textit{and} \quad \sup_{\alpha\in(0,1)} \mathrm{kl}_{\mathrm{cov}}^{\alpha}(q_\phi;P) \overset{+K(P)}{\leq} \mathcal{L}_{\mathrm{NPE}}(q_\phi;P).$$

Moreover, let $p\in[1,\infty)$ and suppose there exists $\beta > 2p$ such that $M_\beta := \mathbb{E}_{\boldsymbol{z}\sim P}[\|\boldsymbol{z}\|^\beta] < \infty$. Fix $\delta\in(0,1)$. Then there exist positive constants C_1, C_2, with $C_2 \geq 1$, depending only on $p, d_{\mathcal{Z}}, \beta, M_\beta$ such that, writing $\rho := (\frac{1}{C_1}\log\frac{2C_2}{\delta})^{1/2}$, for any q_ϕ, whenever $\varepsilon \geq \varepsilon(n,\delta) := \max\{\varepsilon_a(n,\delta), \varepsilon_b(n,\delta)\}$, where

$$\varepsilon_a(n,\delta) := \begin{cases} \rho^{1/p}\, n^{-1/(2p)}, & \text{if } d_{\mathcal{Z}} < 2p, \\ \left(2\rho\, n^{-1/2}\log\left(2+\frac{n}{\rho^2}\right)\right)^{1/p}, & \text{if } d_{\mathcal{Z}} = 2p, \\ \rho^{2/d_{\mathcal{Z}}}\, n^{-1/d_{\mathcal{Z}}}, & \text{if } d_{\mathcal{Z}} > 2p, \end{cases} \qquad \varepsilon_b(n,\delta) := \left(\frac{2C_2}{\delta}\right)^{\frac{1}{2p}} n^{-\frac{1}{2p}},$$

with probability at least $1-\delta$:

$$\mathcal{L}_{\mathrm{NPE}}(q_\phi;P) \leq \sup_{\tilde{P}\in\mathcal{A}_p(P_n;\varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_\phi;\tilde{P}).$$

While Theorem 4.1 bounds $\mathrm{kl}_{\mathrm{cal}}^{S}$ and $\mathrm{kl}_{\mathrm{cov}}^{\alpha}$ separately, they are further connected for a particular choice of summary: under mild conditions, if $S(\boldsymbol{\theta};\boldsymbol{x}) = f(q_\phi(\boldsymbol{\theta}\mid\boldsymbol{x}))$ for some strictly increasing f, then (see Lemma A.4):

$$\sup_{\alpha\in(0,1)} \mathrm{kl}_{\mathrm{cov}}^{\alpha}(q_\phi;P) \leq \mathrm{kl}_{\mathrm{cal}}^{S}(q_\phi;P),$$

so that controlling calibration also controls coverage. Furthermore, the bound on $\mathrm{kl}_{\mathrm{cal}}^{S}$ could be extended to vector-valued summaries $S : \Theta\times\mathcal{X}\to\mathbb{R}^\nu$; we leave this investigation for future work.

The moment condition on P is mild: for $p = 2$, the case most relevant in our setting (see discussion below), it reduces to P having at least a finite fourth moment, which is satisfied by many light-tailed distributions. The dependence of $\varepsilon(n, \delta)$ on $d_{\mathcal{Z}}$ reflects the curse of dimensionality inherent to Wasserstein concentration, with rates that are essentially optimal (Kuhn et al., 2019; Weed and Bach, 2019; Fournier and Guillin, 2015). In practice, however, SBI is often performed on handcrafted or learned summary statistics rather than raw observations (Radev et al., 2022; Deistler et al., 2025), which can substantially reduce the effective dimension. Alternative DRO-based upper-bounding strategies may yield faster rates in some settings (see, e.g., An and Gao, 2021; Gao, 2023) but require stronger assumptions, particularly on the loss, whereas our approach accommodates the general losses used in SBI. Next, we discuss the implementation of DRO-NPE.

A tractable objective for DRO-NPE. DRO-NPE admits two closely related formulations: a strong dual form (Equation (3)) and a regularised form (Proposition 4.2). Both provide upper bounds on the NPE population risk as in Theorem 4.1; the strong dual formulation is exact but challenging to obtain numerically, while the regularised form is computationally cheaper but approximate.

We begin with the strong dual view, which replaces the generally intractable primal Wasserstein-DRO objective with an optimisation problem over auxiliary variables. Let the class of functions satisfying $\rho-$*growth* for $\rho \geq 1$ be

$$\mathcal{G}_\rho(\mathcal{Z}) := \{f : \mathcal{Z} \to \mathbb{R} \ : \ \exists C > 0 \text{ s.t. } |f(\boldsymbol{z})| \leq C(1 + \|\boldsymbol{z}\|^\rho) \ \forall \boldsymbol{z} \in \mathcal{Z}\}.$$

Under the assumptions that for fixed q_ϕ, $\boldsymbol{z} \mapsto \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})$ is upper-semicontinuous and belongs to $\mathcal{G}_p(\mathcal{Z})$, the *strong dual formulation* is exact and finite (Gao and Kleywegt, 2023, Theorem 1):

$$\sup_{\tilde{P} \in \mathcal{A}_p(P_n; \varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_\phi; \tilde{P}) = \inf_{\lambda \geq 0} \left\{ \lambda \varepsilon^p + \frac{1}{n} \sum_{i=1}^{n} \sup_{\boldsymbol{z}' \in \mathcal{Z}} \left(\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}') - \lambda \|\boldsymbol{z}' - \boldsymbol{z}_i\|^p \right) \right\}. \tag{3}$$

Although the dual formulation retains the upper bound guarantee we seek and is more tractable than the primal Wasserstein-DRO objective, the inner supremum remains computationally demanding. (Sinha et al., 2018; Volpi et al., 2018) approximate it by gradient ascent on adversarially perturbed inputs, which requires an inner optimisation loop for every sample at each training step and relies on the smoothness of the loss for its guarantees. To avoid this, we instead focus on the case $p = 2$ and derive in Proposition 4.2 a tractable finite-sample upper bound on the dual objective based on a second-order Taylor argument and growth control of the loss. This yields a regularised objective well-suited for gradient-based optimisation. In practice, we omit the $\mathcal{O}(\varepsilon^2)$ remainder, which we expect to be negligible in most cases. The proof is in Appendix B.2.

Proposition 4.2. *Fix $q_\phi \in \mathcal{Q}$. Suppose that $\boldsymbol{z} \mapsto \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})$ admits a twice continuously differentiable extension to $\mathbb{R}^{d_{\mathcal{Z}}}$, and belongs to $\mathcal{G}_2(\mathcal{Z})$. Then, for every $\varepsilon \geq 0$,*

$$\sup_{\tilde{P} \in \mathcal{A}_2(P_n; \varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_\phi; \tilde{P}) \leq \underbrace{\mathcal{L}_{\mathrm{NPE}}(q_\phi; P_n) + \varepsilon \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}_i)\|^2 \right)^{\frac{1}{2}}}_{\text{the tractable DRO}-\text{NPE objective}} + C_\phi^n \varepsilon^2.$$

Further, if $\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \cdot)\| \in \mathcal{G}_\beta(\mathcal{Z})$, $\|\nabla^2_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \cdot)\| \in \mathcal{G}_\alpha(\mathcal{Z})$, and P is light-tailed, i.e. there exist $\eta > 0$, $a > 0$, and $M < \infty$ such that $\mathbb{E}_{\boldsymbol{z} \sim P}\left[\exp(\eta \|\boldsymbol{z}\|^a)\right] \leq M$, then, for $m := \max\{2, 2\beta, \alpha\}$ and any sequence $\varepsilon_n = o((\log n)^{-m/2a})$, the second-order term $C_\phi^n \varepsilon_n^2$ is negligible in probability.

Proposition 4.2 assumes twice continuous differentiability and growth control to upper-bound the 2-Wasserstein DRO objective by the tractable DRO-NPE objective with remainder of order $\mathcal{O}(\varepsilon^2)$. Under further light-tail and derivative-growth conditions, this remainder vanishes in probability for any ε_n decaying slightly faster than logarithmically, a mild requirement satisfied in particular by polynomial rates, such as $\varepsilon(n, \delta)$ in Theorem 4.1 with $p = 2$. This holds for fixed dimension: $d_{\mathcal{Z}}$, d_Θ and the architecture affect the constant C_ϕ^n, but not the required rate. Importantly, as shown in Lemma 4.3 below, these assumptions hold for Masked Autoregressive Flows (MAFs) (Papamakarios et al., 2017) and coupling flows, the default density estimators in the *sbi* (Tejero-Cantero et al., 2020) and *BayesFlow* (Kühmichel et al., 2026) libraries; see Appendix B.4 for a proof. By contrast, existing upper bounds for $p = 2$ Wasserstein DRO typically require global smoothness assumptions on the loss, such as Lipschitz gradients (e.g., Gao, 2023), and are therefore not directly suited to NPE with normalising flows.

Lemma 4.3. *Suppose q_ϕ is a MAF or coupling flow (formally defined in Appendix B.4), then ℓ_{NPE} satisfies the assumptions of Proposition 4.2.*

Following Proposition 4.2, in practice, we minimise the tractable DRO-NPE objective:

$$\mathcal{L}_{\mathrm{NPE}}(q_\phi; P_n) + \varepsilon\Omega(\phi; P_n), \quad \text{where} \quad \Omega(\phi; P_n) := \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}_i)\|^2 \right)^{\frac{1}{2}}.$$

As $\varepsilon \to 0$, this reduces to standard NPE. As ε increases, the regularisation term dominates, and the optimisation problem approaches $\inf_\phi \Omega(\phi; P_n) \geq 0$. If this infimum is attained at zero, then any minimiser ϕ_{min} satisfies $\nabla_{\boldsymbol{z}} \log q_{\phi_{\text{min}}}(\boldsymbol{\theta}_i \mid \boldsymbol{x}_i) = 0 \; \forall i = 1, \ldots, n$. Provided $q_{\phi_{\text{min}}}(\boldsymbol{\theta}_i \mid \boldsymbol{x}_i) > 0$, this implies $\nabla_{\boldsymbol{z}} q_{\phi_{\text{min}}}(\boldsymbol{\theta}_i \mid \boldsymbol{x}_i) = 0 \quad \forall i = 1, \ldots, n$. Thus, for large ε, the objective favours conditional densities that vary less sharply in $\boldsymbol{z}$ at the simulated pairs. This is consistent with improving expected coverage of HPDR, since overconfident densities tend to allocate excess mass around training pairs.

We further note that as n grows, for fixed ϕ and under the conditions of Proposition 4.2, $\Omega(\phi; P_n)$ converges almost surely to its finite population counterpart $\Omega(\phi; P) := (\mathbb{E}_{\boldsymbol{z}\sim P}[\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})\|^2])^{1/2}$, so that for $\varepsilon_n \to 0$ the tractable objective recovers $\mathcal{L}_{\mathrm{NPE}}(q_\phi; P)$. This argument is pointwise in ϕ; consistency of the resulting estimator with a data-dependent ε_n would require uniform convergence over ϕ, which we leave for future work.

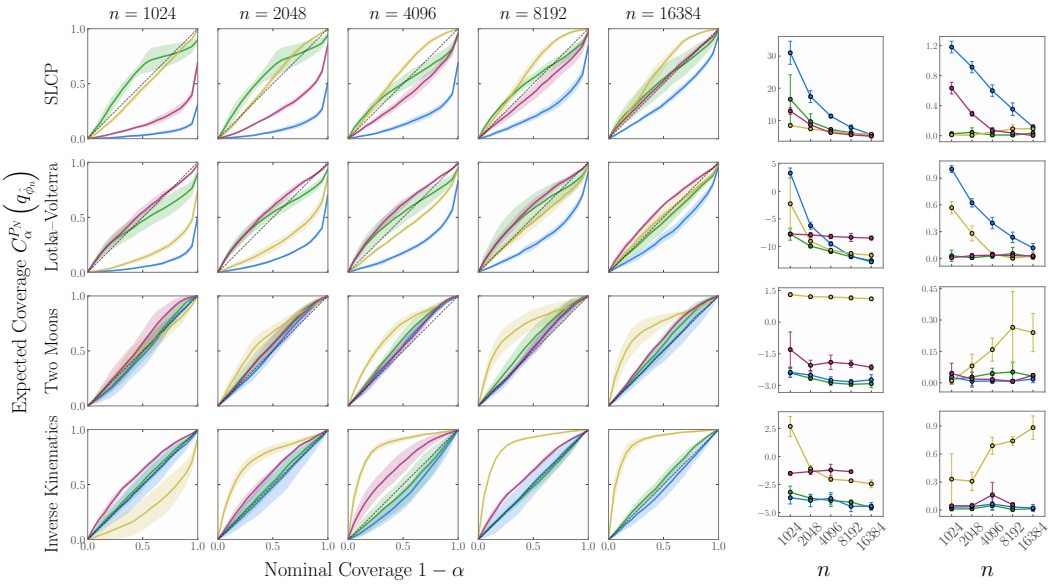
Selecting ε. We follow common DRO practice and select ε via a train-validation split (Mohajerin Esfahani and Kuhn, 2018), using KL-based miscalibration $\mathrm{kl}^S_{\mathrm{cal}}$ with $S(\boldsymbol{\theta}; \boldsymbol{x}) = q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$ as the metric. Under mild conditions, this metric upper-bounds miscoverage (Lemma A.4), so reducing it can improve coverage. It also penalises both over- and underconfidence, whereas selection via absolute miscoverage favours larger ε; see our experiments in Section 5.2. Specifically, for each candidate ε, we train DRO-NPE on the training set, estimate $\mathrm{kl}^S_{\mathrm{cal}}$ on the validation set via density-ratio estimation, and use Bayesian optimisation to select the next candidate; see Appendix C.1 for details.

Theorem 4.1 also provides a radius $\varepsilon(n, \delta)$ under which its upper bound holds with high probability, but this radius depends on constants that are not available in closed form. Using the explicit concentration bounds of Fournier (2023), one can obtain an analogue of Theorem 4.1 with a computable radius, which guarantees that the bound holds; however, this radius is far too conservative in practice. A promising alternative is to estimate $W_p(P_n, P)$ directly by bootstrapping the simulations, which satisfies the theory up to bootstrap error and is much cheaper than validation, since it requires no retraining. However, unlike the approach with $\mathrm{kl}^S_{\mathrm{cal}}$, it remains unclear how the resulting ε affects the coverage and calibration of the trained posterior, and we leave its study to future work. Both approaches are discussed further in Appendix C.2.

Further statistical guarantees for DRO-NPE. We derive statistical guarantees on the excess risk, miscalibration and miscoverage of both the exact and tractable formulations of DRO-NPE and NPE relative to the oracle estimator $\phi^\star := \arg\min_\phi \mathcal{L}_{\mathrm{NPE}}(q_\phi; P)$ (see Appendix B.3). To the best of our knowledge, comparable guarantees are not available for existing methods. In particular, under a stronger Lipschitz gradient assumption that could be satisfied in the NPE setting for bounded Θ, we derive a DRO-NPE excess risk bound that depends on $\Omega(\phi^\star; P_n)$ along with other $n-$dependent constants. In contrast, the equivalent excess risk bound for NPE additionally depends on the regulariser $\Omega(\hat{\phi}_n; P_n)$, which relies on the empirical risk minimiser. As a result, DRO-NPE can yield sharper bounds compared to NPE, when $q_{\hat{\phi}_n}$ exhibits high functional variation relative to $q_{\phi^\star}$.

5 Experiments

We evaluate DRO-NPE on a range of SBI tasks in Section 5.1, comparing it with existing NPE methods that encourage conservative posteriors. In Section 5.2, we study the effect of ε, compare alternative hyperparameter-selection metrics, and quantify the training overhead of DRO-NPE. Finally, we apply DRO-NPE to a real-world cosmology problem in Section 5.3. The code to reproduce our experiments is available at `https://github.com/yugahikida/dro-npe`.



(a) Coverage curves (b) NLPD (c) $\mathrm{kl}_{\mathrm{cal}}^{q}$

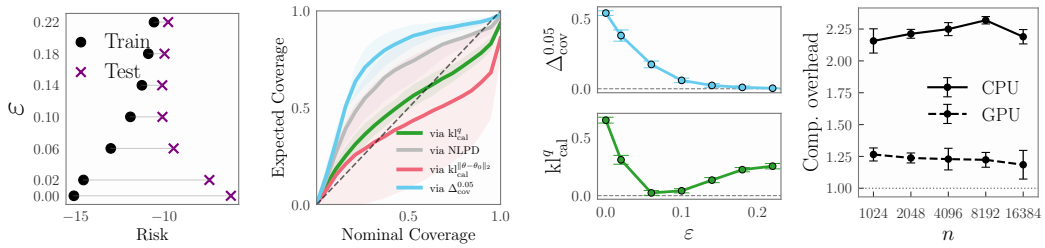
Figure 2: Benchmarking DRO-NPE (■), CR-NPE (■), Bal-NPE (■) and standard NPE (■) across four simulators and five simulation budgets. Means and standard deviations are shown over five random seeds. (a) Expected coverage curves of HPDR; the diagonal denotes perfect coverage, with curves above indicating conservativeness and curves below overconfidence. Coverage is computed at 18 nominal levels using $N = 500$ test pairs and 1000 posterior samples per test point. (b) Negative log predictive density (NLPD) on held-out test data. (c) Estimated KL-based miscalibration $\mathrm{kl}_{\mathrm{cal}}^{q}$ on held-out test data; ε is selected using a separate validation set. Some CR-NPE results are missing due to numerical instabilities during training.

5.1 Benchmarking DRO-NPE

We benchmark DRO-NPE against standard NPE, coverage-regularised NPE (CR-NPE; Falkiewicz et al., 2023), and balanced NPE (Bal-NPE; Delaunoy et al., 2023) on the following common SBI tasks from Lueckmann et al. (2021); Kühmichel et al. (2026): SLCP ($d_{\mathcal{X}} = 8$, $d_\Theta = 5$), Lotka–Volterra ($d_{\mathcal{X}} = 20$, $d_\Theta = 4$), Two Moons ($d_{\mathcal{X}} = 2$, $d_\Theta = 2$), and Inverse Kinematics ($d_{\mathcal{X}} = 2$, $d_\Theta = 4$). To isolate the effects of the training objectives, we use the same normalising-flow architecture and training protocol across methods, with the default hyperparameters of BayesFlow and Keras and without early stopping or additional regularisation. Consequently, our results may differ from those in the baseline papers. We also investigate combining each method with reduced model capacity and early stopping in ablation studies; see Appendix D.

Evaluation is based on three metrics: (i) coverage plots, which show empirical versus nominal coverage, (ii) negative log predictive density (NLPD), which is the average negative log-density assigned by the learned posterior to the true parameter on held-out test points, and (iii) our $\mathrm{kl}_{\mathrm{cal}}^{q}$ metric. We compute the expected coverage using SBC with the summary $S(\boldsymbol{\theta}, \boldsymbol{x}) = \log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$. Following Falkiewicz et al. (2023); Delaunoy et al. (2023), we report results averaged over five random seeds; see Figure 2. Details of the baselines, tasks, and implementation are in Appendix C.

NPE yields overconfident posteriors for SLCP and Lotka–Volterra. Although this effect is more pronounced in low-simulation regimes, it persists for larger n. For Two Moons and Inverse Kinematics, the averaged NPE coverage curve is closer to the diagonal, though in some cases it shows substantial variability across runs. Bal-NPE yields conservative posteriors for SLCP and Two Moons overall, but is overconfident for Lotka–Volterra and Inverse Kinematics in low-simulation regimes. A possible explanation is that the approximation $q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}) \approx p(\boldsymbol{\theta} \mid \boldsymbol{x})$ used to derive the regularisation term in Bal-NPE is inaccurate early in training, which may hinder optimisation. Interestingly, Bal-NPE yields overly conservative posteriors for Inverse Kinematics and Two Moons when the simulation budget is large. CR-NPE is conservative across almost all experiments, except for SLCP. We also



(a) (b) (c) (d)

Figure 3: Analysis of DRO-NPE on Lotka–Volterra. Means and standard deviations are shown over five random seeds. (a) training and test risk across ε. (b) Coverage curves for ε selected by $\mathrm{kl}^q_{\mathrm{cal}}$ (■), NLPD (■), $\mathrm{kl}^{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2}_{\mathrm{cal}}$ with $\boldsymbol{\theta}_0$ drawn from the prior (■), and absolute miscoverage at $\alpha = 0.05$ (■). (c) Absolute miscoverage at $\alpha = 0.05$ (■) and $\mathrm{kl}^q_{\mathrm{cal}}$ (■) across ε. (d) Relative wall-clock runtime of DRO-NPE versus standard NPE on GPU and CPU.

observe numerical instabilities at larger sample sizes, leading to missing results in some cases (such as Inverse Kinematics with $n = 16384$); see Appendix C.4 for details.

DRO-NPE yields conservative posteriors for SLCP and near-perfect coverage for Two Moons and Inverse Kinematics. In the latter two cases, standard NPE already attains near-perfect coverage, but DRO-NPE does so with lower variance across runs. The only notable failure is Lotka–Volterra at $n = 1024$, where the posterior remains slightly overconfident. This may be mitigated by selecting a larger ε. In terms of NLPD, DRO-NPE achieves the lowest values overall. This is consistent with the fact that the baselines primarily encourage conservativeness, often at the expense of predictive performance. DRO-NPE also often improves on standard NPE, which we attribute to reduced generalisation error induced by distributional robustness. Finally, DRO-NPE consistently yields low $\mathrm{kl}^q_{\mathrm{cal}}$ across simulation budgets and tasks, as expected because ε is selected to minimise this criterion.

5.2 Analysing DRO-NPE

We now study the behaviour of DRO-NPE with respect to the key design choices involved in its implementation. All experiments are conducted using the Lotka–Volterra simulator with $n = 2048$.

Generalisation gap versus ε. We first examine how ε affects the generalisation gap by comparing empirical risk on training and test data. As shown in Figure 3(a), small values of ε lead to a large generalisation gap and poor test performance, reflecting the optimistic bias of empirical risk minimisation in this setting. As ε increases, the gap narrows and test performance improves, attaining its minimum at $\varepsilon = 0.14$. Further increasing ε continues to reduce the gap but worsens test performance, as the posterior estimator becomes excessively conservative.

Choice of hyperparameter-selection metric. In Section 5.1, we select ε using the KL-based miscalibration metric $\mathrm{kl}^q_{\mathrm{cal}}$, corresponding to $S(\boldsymbol{\theta}; \boldsymbol{x}) = q(\boldsymbol{\theta} \mid \boldsymbol{x})$. We compare this choice with three alternatives: NLPD, absolute miscoverage $\Delta^{0.05}_{\mathrm{cov}}$, and $\mathrm{kl}^{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2}_{\mathrm{cal}}$, computed with $S(\boldsymbol{\theta}; \boldsymbol{x}) = \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2$ for a fixed reference point $\boldsymbol{\theta}_0$. NLPD is a natural validation metric because it equals the NPE loss and upper-bounds $\mathrm{kl}^q_{\mathrm{cal}}$, while absolute miscoverage directly measures deviation from nominal coverage. As shown in Figure 3(b), selecting ε via $\mathrm{kl}^q_{\mathrm{cal}}$ yields a coverage curve closest to the diagonal. Selection via NLPD and $\Delta^{0.05}_{\mathrm{cov}}$ produces more conservative posteriors, whereas $\mathrm{kl}^{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2}_{\mathrm{cal}}$ leads to undercoverage and higher variability across runs. Results for SLCP are in Appendix D.

To understand the behaviour of absolute miscoverage, Figure 3(c) plots $\Delta^{0.05}_{\mathrm{cov}}$ as a function of ε, together with $\mathrm{kl}^q_{\mathrm{cal}}$. We observe that $\Delta^{0.05}_{\mathrm{cov}}$ decreases nearly monotonically with ε, as larger values produce wider, more conservative posteriors and hence better coverage. This explains why selecting ε via absolute miscoverage can favour overly conservative posterior approximations. By contrast, $\mathrm{kl}^q_{\mathrm{cal}}$ is minimised at an intermediate value of ε, making it more informative for selecting ε in DRO-NPE.

Computational cost. DRO-NPE is more expensive than standard NPE, but only by a moderate constant factor. The main overhead comes from higher-order automatic differentiation through $\nabla_{\boldsymbol{z}} \log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$, so the computational cost of this regularisation term grows with the joint parameter–data dimension $d_{\mathcal{Z}}$. As shown in Figure 3(d), DRO-NPE takes roughly $2\times$ the wall-clock time of standard NPE on CPU, but only about $1.25\times$ on GPU, as the additional operations are highly parallelisable. This relative overhead varies little with n over the range considered, although it may depend on the hardware and computational setup. These timings assume fixed ε and exclude hyperparameter selection. For comparison, on Lotka–Volterra with $n = 1024$, DRO-NPE training with fixed ε takes 93 seconds on GPU, compared to 1521 seconds for the full procedure with 10 Bayesian optimisation iterations and final retraining. Further details are provided in Appendix C.6.

5.3 Application to cosmology

Finally, we evaluate DRO-NPE on the CAMELS suite (Villaescusa-Navarro et al., 2021, 2023), a state-of-the-art cosmology dataset for machine learning on simulated universes. While the original data space is 256×256 dimensional, we use their standard $d_{\mathcal{X}} = 39$ dimensional summary as data, similar to Hikida et al. (2025). One of the goals of CAMELS is to enable accurate constraints on $d_\Theta = 2$ cosmological parameters: the matter density Ω_m and the amplitude of matter fluctuations σ_8. We use 1,000 high-fidelity baryonic hydrodynamic simulations available to us, reserving 900 for training and 100 for testing. Conservative inference is particularly important in this setting as simulations are scarce, and the observables depend not only on Ω_m and σ_8, but also on nuisance astrophysical processes that can produce similar simulated data (Villaescusa-Navarro et al., 2021). Overconfident posteriors can therefore lead to overstated cosmological constraints and misleading scientific conclusions.

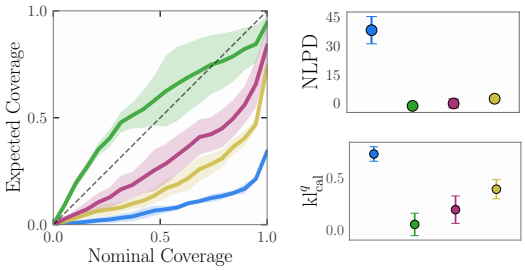


Figure 4: Cosmology results: expected coverage, NLPD, and $\mathrm{kl}^q_{\mathrm{cal}}$ for DRO-NPE (■), CR-NPE (■), Bal-NPE (■), and standard NPE (■).

Figure 4 reports expected coverage, NLPD, and $\mathrm{kl}^q_{\mathrm{cal}}$ across five independent training runs for each method on the same fixed train–test split. Standard NPE (■) is severely overconfident and has much worse NLPD than all conservative methods, while Bal-NPE (■) improves NLPD but remains strongly undercovered. CR-NPE (■) improves upon Bal-NPE across all metrics, yet its coverage remains below the diagonal for all nominal levels. DRO-NPE (■) performs best overall, with the closest coverage to the diagonal, the lowest NLPD, and the lowest $\mathrm{kl}^q_{\mathrm{cal}}$. These results suggest that DRO-NPE can provide conservative yet efficient posterior inference in challenging low-simulation settings.

6 Conclusion

We proposed DRO-NPE, a conservative formulation of NPE, to address finite-simulation failure modes including overconfidence, poor calibration, and large generalisation gaps. Our approach yields an objective that upper bounds the population NPE loss and KL-based metrics for uncertainty-quantification error, admits a tractable regularised objective, and enjoys statistical guarantees on excess risk, miscoverage, and miscalibration. Importantly, it remains compatible with standard normalising-flow architectures commonly used in SBI. However, DRO-NPE inherits limitations from DRO: it is more expensive than standard NPE, and there remains a theory–practice gap in selecting ε, since computing a theoretical radius guaranteeing the derived upper bounds is typically infeasible.

Although we focused on NPE, DRO-NPE could be extended to other conditional density estimation problems, score estimation for diffusion models, variational inference, amortised inference tasks, and other losses via analogous DRO formulations. It could also be combined with post-hoc calibration methods (Masserano et al., 2023; Patel et al., 2024; Cabezas et al., 2025) and ensembles (Hermans et al., 2022). More broadly, DRO offers one route to upper-bounding the population NPE objective; future work could investigate whether alternative frameworks, such as PAC-Bayes (Alquier, 2024; Haddouche et al., 2021), yield analogous guarantees and practical training objectives. Future work could also investigate the robustness of DRO-NPE to certain out-of-distribution shifts.

Acknowledgments

The authors thank Niall Jeffrey for help with using the CAMELS dataset. WL was supported by UCL's Centre for Doctoral Training in Data-Intensive Science and by the project 'Bayesian robustness in filtering algorithms' funded by The Alan Turing Institute. WL also acknowledges funding from the Aalto University Science Institute (AScI) Visiting Doctoral Researcher programme. YH and AB were supported by the Research Council of Finland grant no. 362534. CD and FXB were supported by the EPSRC grant [EP/Y022300/1]. CD was additionally supported by the project 'Reliable insights from scientific simulations' funded from the EPSRC grant [UKRI3030]. The authors acknowledge the computational resources provided by the Aalto Science-IT project.

References

Alquier, P. (2024). User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303. 10

Alsing, J., Wandelt, B., and Feeney, S. (2018). Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology. *Monthly Notices of the Royal Astronomical Society*, 477(3):2874–2885. 1

An, Y. and Gao, R. (2021). Generalization bounds for (Wasserstein) robust optimization. In *Advances in Neural Information Processing Systems*. 6

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43. 42

Bayraksan, G. and Love, D. K. (2015). Data-driven stochastic programming using phi-divergences. In *The operations research revolution*, pages 1–19. Informs. 5

Bharti, A., Dellaporta, C., Hikida, Y., and Briol, F.-X. (2026). Amortised and provably-robust simulation-based inference. *arXiv:2602.11325*. 29

Bharti, A., Huang, D., Kaski, S., and Briol, F.-X. (2025). Cost-aware simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, volume 258, pages 28–36. 1

Bishop, C. M. (1994). Mixture density networks. Technical Report NCRG/94/004, Neural Computing Research Group, Department of Computer Science and Applied Mathematics, Aston University. 2

Blanchet, J., Li, J., Lin, S., and Zhang, X. (2024). Distributionally robust optimization and robust statistics. *arXiv:2401.14655*. 5

Brehmer, J. (2021). Simulation-based inference in particle physics. *Nature Reviews Physics*, 3(5):305–305. 1

Cabezas, L., Santos, V. S., Ramos, T. R., Rodrigues, P. L., and Izbicki, R. (2025). CP4SBI: Local conformal calibration of credible sets in simulation-based inference. *arXiv:2508.17077*. 2, 10

Chung, Y., Char, I., and Schneider, J. (2024). Sampling-based multi-dimensional recalibration. In *International Conference on Machine Learning*, volume 235, pages 8919–8940. 3

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48). 1

Deistler, M., Boelts, J., Steinbach, P., Moss, G., Moreau, T., Gloeckler, M., Rodrigues, P. L., Linhart, J., Lappalainen, J. K., Miller, B. K., et al. (2025). Simulation-based inference: A practical guide. *arXiv:2508.12939*. 6

Delaunoy, A., Bonardeaux, M. d. l. B., Mishra-Sharma, S., and Louppe, G. (2024). Low-budget simulation-based inference with Bayesian neural networks. *arXiv:2408.15136*. 2

Delaunoy, A., Hermans, J., Rozet, F., Wehenkel, A., and Louppe, G. (2022). Towards reliable simulation-based inference with balanced neural ratio estimation. In *Advances in Neural Information Processing Systems*. 2, 3, 16, 39

Delaunoy, A., Kurt Miller, B., Forré, P., Weniger, C., and Louppe, G. (2023). Balancing simulation-based inference for conservative posteriors. In *Symposium on Advances in Approximate Bayesian Inference*. 2, 8, 39, 40

Dey, B., Zhao, D., Andrews, B. H., Newman, J. A., Izbicki, R., and Lee, A. B. (2025). Towards instance-wise calibration: local amortized diagnostics and reshaping of conditional densities (LADaR). *Machine Learning: Science and Technology*, 6(4):045058. 3

Dheur, V. and Taieb, S. B. (2023). A large-scale study of probabilistic calibration in neural network regression. In *International Conference on Machine Learning*, volume 202, pages 7813–7836. 4

Dyer, J., Cannon, P., Farmer, J. D., and Schmon, S. M. (2024). Black-box Bayesian inference for agent-based models. *Journal of Economic Dynamics and Control*, 161:104827. 1

Falkiewicz, M., Takeishi, N., Shekhzadeh, I., Wehenkel, A., Delaunoy, A., Louppe, G., and Kalousis, A. (2023). Calibrating neural simulation-based inference with differentiable coverage probability. In *Conference on Neural Information Processing Systems*. 2, 4, 8, 40, 46

Fengler, A., Govindarajan, L. N., Chen, T., and Frank, M. J. (2021). Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience. *Elife*, 10:e65074. 1

Folland, G. B. (2024). *Advanced Calculus*. Second edition. Freely available online PDF. 21

Fournier, N. (2023). Convergence of the empirical measure in expected wasserstein distance: non-asymptotic explicit bounds in $\mathbb{R}^d$. *ESAIM: Probability and Statistics*, 27:749–775. 7, 39

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738. 5, 6, 19, 24, 39

Gao, R. (2023). Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 71(6):2291–2306. 6, 24, 25, 26

Gao, R. and Kleywegt, A. (2023). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655. 6, 22

Garnett, R. (2023). *Bayesian optimization*. Cambridge University Press. 38

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268. 4

Greenberg, D., Nonnenmacher, M., and Macke, J. (2019). Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, volume 97, pages 2404–2414. 1, 41

Haddouche, M., Guedj, B., Rivasplata, O., and Shawe-Taylor, J. (2021). Pac-bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10):1330. 10

Hermans, J., Begy, V., and Louppe, G. (2020). Likelihood-free MCMC with amortized approximate ratio estimators. In *International Conference on Machine Learning*, volume 119, pages 4239–4248. 2

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., Begy, V., and Louppe, G. (2022). A trust crisis In simulation-based inference? Your posterior approximations can be unfaithful. *Transactions in Machine Learning Research*. 1, 2, 3, 10, 41

Hikida, Y., Bharti, A., Jeffrey, N., and Briol, F.-X. (2025). Multilevel neural simulation-based inference. *Conference on Neural Information Processing Systems*. 1, 10

Hu, Z. and Hong, L. J. (2013). Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9. 5

Husain, H., Nguyen, V., and van den Hengel, A. (2023). Distributionally robust Bayesian optimization with φ-divergences. In *Conference on Neural Information Processing Systems*. 5

Jeffrey, N., Alsing, J., and Lanusse, F. (2021). Likelihood-free inference with neural compression of DES SV weak lensing map statistics. *Monthly Notices of the Royal Astronomical Society*, 501(1):954–969. 1

Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*. 32

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979. 19

Krouglova, A. N., Johnson, H. R., Confavreux, B., Deistler, M., and Goncalves, P. J. (2026). Multifidelity simulation-based inference for computationally expensive simulators. In *International Conference on Learning Representations*. 1

Kruse, J., Ardizzone, L., Rother, C., and Köthe, U. (2021). Benchmarking invertible architectures on inverse problems. *arXiv:2101.10763*. 42

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs. 4, 5, 6

Kuhn, D., Shafiee, S., and Wiesemann, W. (2025). Distributionally robust optimization. *Acta Numerica*, 34:579–804. 2, 5

Kühmichel, L., Huang, J. M., Pratz, V., Arruda, J., Olischläger, H., Habermann, D., Kucharsky, S., Elsemüller, L., Mishra, A., Bracher, N., Jedhoff, S., Schmitt, M., Bürkner, P.-C., and Radev, S. T. (2026). BayesFlow 2: Multi-backend amortized Bayesian inference in python. *arXiv:2602.07098*. 3, 6, 8, 32, 40

Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L. (2023). Sampling-based accuracy testing of posterior estimators for general inference. In *International Conference on Machine Learning*, volume 202, pages 19256–19273. 3

Linhart, J., Gramfort, A., and Rodrigues, P. (2023). L-c2st: local diagnostics for posterior approximations in simulation-based inference. *Advances in Neural Information Processing Systems*, 36:56384–56410. 3

Linhart, J., Gramfort, A., and Rodrigues, P. L. C. (2022). Validation diagnostics for SBI algorithms based on normalizing flows. In *NeurIPS 2022 Machine Learning and the Physical Sciences Workshop*. 3

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv:1711.05101*. 41

Lotka, A. J. (1920). Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences*, 6(7):410–415. 41

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. (2021). Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pages 343–351. 8, 41

Lueckmann, J.-M., Gonçalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. (2017). Flexible statistical inference for mechanistic models of neural dynamics. In *Conference on Neural Information Processing Systems*. 1

Masserano, L., Dorigo, T., Izbicki, R., Kuusela, M., and Lee, A. (2023). Simulator-based inference with WALDO: Confidence regions by leveraging prediction algorithms and posterior estimators for inverse problems. In *International Conference on Artificial Intelligence and Statistics*, volume 206, pages 2960–2974. 2, 10

Miller, B. K., Weniger, C., and Forré, P. (2022). Contrastive neural ratio estimation. In *Advances in Neural Information Processing Systems*. 2

Modrák, M., Moon, A. H., Kim, S., Bürkner, P., Huurre, N., Faltejsková, K., Gelman, A., and Vehtari, A. (2023). Simulation-based calibration checking for Bayesian computation: The choice of test quantities shapes sensitivity. *Bayesian Analysis*, 20(2):461. 3

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166. 7

Papamakarios, G. and Murray, I. (2016). Fast ϵ -free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*. 1

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64. 2, 19

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*. 6, 27

Papamakarios, G., Sterratt, D., and Murray, I. (2019). Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, volume 89, pages 837–848. 41

Patel, Y., Mcnamara, D., Loper, J., Regier, J., and Tewari, A. (2024). Variational inference with coverage guarantees in simulation-based inference. In *International Conference on Machine Learning*, volume 235, pages 39861–39888. 2, 10

Polyanskiy, Y. and Wu, Y. (2025). *Information theory: From coding to learning*. Cambridge University Press. 4, 17

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., and Köthe, U. (2022). Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4):1452–1466. 1, 6, 32

Rahimian, H. and Mehrotra, S. (2022). Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85. 4, 5

Rossi, B. and Sekhposyan, T. (2019). Alternative tests for correct specification of conditional predictive densities. *Journal of Econometrics*, 208(2):638–657. 4

Säilynoja, T., Bürkner, P.-C., and Vehtari, A. (2022). Graphical test for discrete uniformity and its applications in goodness-of-fit evaluation and multiple sample comparison. *Statistics and Computing*, 32(2):32. 3

Saoulis, A. A., Piras, D., Jeffrey, N., Spurio-Mancini, A., Ferreira, A. M. G., and Joachimi, B. (2025). Transfer learning for multifidelity simulation-based inference in cosmology. *arXiv:2505.21215*. 1

Schmitt, M., Pratz, V., Köthe, U., Bürkner, P.-C., and Radev, S. T. (2024). Consistency models for scalable and fast simulation-based inference. *Advances in Neural Information Processing Systems*, 37:126908–126945. 42

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press. 4

Sinha, A., Namkoong, H., Volpi, R., and Duchi, J. (2018). Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*. 6

Sonthalia, R., Lok, J., and Rebrova, E. (2024). On regularization via early stopping for least squares regression. *arXiv:2406.04425*. 47

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv:1804.06788*. 2, 3, 4

Tatsuoka, C., Yang, M., Xiu, D., and Zhang, G. (2025). Multi-fidelity parameter estimation using conditional diffusion models. *arXiv:2504.01894*. 1

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P., Greenberg, D., and Macke, J. (2020). sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505. 6, 27

Utpala, S. and Rai, P. (2020). Quantile regularization: Towards implicit calibration of regression models. *arXiv:2002.12860*. 4

Vapnik, V. (1991). Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*. 4

Villaescusa-Navarro, F. et al. (2021). The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. *The Astrophysical Journal*, 915(1):71. 10, 42

Villaescusa-Navarro, F. et al. (2023). The CAMELS Project: Public Data Release. *The Astrophysical Journal Supplement Series*, 265(2):54. 10, 42

Volpi, R., Namkoong, H., Sener, O., Duchi, J., Murino, V., and Savarese, S. (2018). Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, volume 31. 6

Wang, S., Wang, H., Li, X., and Honorio, J. (2025). Learning against distributional uncertainty: On the trade-off between robustness and specificity. *IEEE Journal of Selected Topics in Signal Processing*. 4, 5

Ward, D., Cannon, P., Beaumont, M., Fasiolo, M., and Schmon, S. M. (2022). Robust neural posterior estimation and statistical model criticism. In *Advances in Neural Information Processing Systems*. 3

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648. 6

Yao, Y., Régaldo-Saint Blancard, B., and Domke, J. (2024). Simulation-based stacking. In *International Conference on Artificial Intelligence and Statistics*, volume 238, pages 4267–4275. 4

Zammit-Mangion, A., Sainsbury-Dale, M., and Huser, R. (2024). Neural methods for amortised parameter inference. *arxiv 2404.12484*. 1

Zeng, Y. and Lam, H. (2022). Generalization bounds with minimal dependency on hypothesis class via distributionally robust optimization. In *Advances in Neural Information Processing Systems*. 24

Zhao, D., Dalmasso, N., Izbicki, R., and Lee, A. B. (2021). Diagnostics for conditional density models and Bayesian inference algorithms. In *Uncertainty in Artificial Intelligence*, pages 1830–1840. 3, 4

Appendix

We start by presenting the theoretical results on calibration, expected coverage of HPDR, miscalibration, and miscoverage in Appendix A. Then, in Appendix B, we present theoretical results relating to the DRO-NPE method. Finally, we present implementation details and additional experiments in Appendix C and Appendix D, respectively.

A Theory: calibration and expected coverage of HPDR

In this section, we first show that the expected coverage of HPDR is a special case of calibration. Second, we present the data-processing inequality (DPI) and the chain rule for the Kullback-Leibler (KL) divergence, which lets us compare P and Q after applying pushforwards. Third, we prove that KL-based miscoverage upper-bounds its absolute value notion. Finally, we establish that KL-based miscalibration upper-bounds miscoverage.

We first prove that the expected coverage of HPDR is a special case of calibration, a result that was briefly shown in Delaunoy et al. (2022). This will require a statement more mathematical than Definition 2.2, presented below. We say that $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ (or equivalently q) is calibrated w.r.t. to S if:

$$\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[P_{S|\boldsymbol{x}}\left(F^Q_{S|\boldsymbol{x}}(S^P_{\boldsymbol{x}}) \leq u\right)\right] = u \qquad \forall u \in [0,1]. \tag{4}$$

This is equivalent to stating that $F^Q_{S|\boldsymbol{x}}(S^P_{\boldsymbol{x}})$ is distributed as $\mathcal{U}[0,1]$ under $\boldsymbol{x} \sim P_{\boldsymbol{x}}$.

Lemma A.1 (Expected coverage of HPDR as a special case of calibration)**.** *Fix q. Let $S(\boldsymbol{\theta};\boldsymbol{x}) := q(\boldsymbol{\theta} \mid \boldsymbol{x})$, and assume that the CDF of $Q_{S|\boldsymbol{x}}$ is continuous and that $F^Q_{S|\boldsymbol{x}}$ is strictly increasing. Then, for every $\alpha \in (0,1)$,*

$$C^P_\alpha(q) = 1 - \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[P_{S|\boldsymbol{x}}\left(F^Q_{S|\boldsymbol{x}}(S^P_{\boldsymbol{x}}) < \alpha\right)\right].$$

Proof. Suppose $S(\boldsymbol{\theta};\boldsymbol{x}) = q(\boldsymbol{\theta} \mid \boldsymbol{x})$. Then, notice that

$$\begin{aligned} F^Q_{S|\boldsymbol{x}}(\tau^q_\alpha(\boldsymbol{x})) &= Q_{S|\boldsymbol{x}}((-\infty, \tau^q_\alpha(\boldsymbol{x})]) \\ &= Q_{\boldsymbol{\theta}|\boldsymbol{x}}(\{S(\boldsymbol{\theta};\boldsymbol{x}) \leq \tau^q_\alpha(\boldsymbol{x})\}) \\ &= Q_{\boldsymbol{\theta}|\boldsymbol{x}}(\{q(\boldsymbol{\theta} \mid \boldsymbol{x}) \leq \tau^q_\alpha(\boldsymbol{x})\}) = \alpha, \end{aligned}$$

where the first equality is by definition of the cumulative distribution function, the second by that of the pushforward, the third by that of S, and the last by that of $\tau^q_\alpha(\boldsymbol{x})$ together with continuity of the CDF of $Q_{S|\boldsymbol{x}}$. Therefore, denoting $S^P_{\boldsymbol{x}} = S(\boldsymbol{\theta};\boldsymbol{x})$ whenever $\boldsymbol{\theta} \sim P_{\boldsymbol{\theta}|\boldsymbol{x}}$, we have that for any $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned} P_{\boldsymbol{\theta}|\boldsymbol{x}}\left(R^q_\alpha(\boldsymbol{x})\right) &= P_{\boldsymbol{\theta}|\boldsymbol{x}}\left(q(\boldsymbol{\theta} \mid \boldsymbol{x}) \geq \tau^q_\alpha(\boldsymbol{x})\right) \\ &= P_{S|\boldsymbol{x}}\left(S^P_{\boldsymbol{x}} \geq \tau^q_\alpha(\boldsymbol{x})\right) \\ &= P_{S|\boldsymbol{x}}\left(F^Q_{S|\boldsymbol{x}}(S^P_{\boldsymbol{x}}) \geq F^Q_{S|\boldsymbol{x}}(\tau^q_\alpha(\boldsymbol{x}))\right) \\ &= P_{S|\boldsymbol{x}}\left(F^Q_{S|\boldsymbol{x}}(S^P_{\boldsymbol{x}}) \geq \alpha\right). \end{aligned}$$

The first equality is the definition of the HPDR at level α. The second follows from the choice $S(\boldsymbol{\theta};\boldsymbol{x}) = q(\boldsymbol{\theta} \mid \boldsymbol{x})$. The third uses that $F^Q_{S|\boldsymbol{x}}$ is continuous and strictly increasing. The last equality uses that $F^Q_{S|\boldsymbol{x}}(\tau^q_\alpha(\boldsymbol{x})) = \alpha$, established above. Finally, we get that:

$$\begin{aligned} C^P_\alpha(q) &= \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[P_{\boldsymbol{\theta}|\boldsymbol{x}}(R^q_\alpha(\boldsymbol{x}))\right] \\ &= \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[P_{S|\boldsymbol{x}}\left(F^Q_{S|\boldsymbol{x}}(S^P_{\boldsymbol{x}}) \geq \alpha\right)\right] \\ &= 1 - \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[P_{S|\boldsymbol{x}}\left(F^Q_{S|\boldsymbol{x}}(S^P_{\boldsymbol{x}}) < \alpha\right)\right]. \end{aligned}$$

Since this identity holds for every $\alpha \in (0,1)$, it shows that expected HPDR coverage is exactly the special case of (4) obtained by taking $S(\boldsymbol{\theta};\boldsymbol{x}) = q(\boldsymbol{\theta} \mid \boldsymbol{x})$ and $u = \alpha$. □

We now define the expected Kullback-Leibler (KL) divergence, which will be used frequently in this paper. Suppose $P_{\boldsymbol{\theta}|\boldsymbol{x}}$ is absolutely continuous w.r.t. $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$, and for simplicity, that $P_{\boldsymbol{\theta}|\boldsymbol{x}}$ and $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ admit densities $p(\boldsymbol{\theta} \mid \boldsymbol{x}), q(\boldsymbol{\theta} \mid \boldsymbol{x})$. Then, the KL divergence is:

$$\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{KL}\Big(P_{\boldsymbol{\theta}|\boldsymbol{x}} \parallel Q_{\boldsymbol{\theta}|\boldsymbol{x}}\Big)\right] := \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathbb{E}_{\boldsymbol{\theta}\sim P_{\boldsymbol{\theta}|\boldsymbol{x}}}\left[\log \frac{p(\boldsymbol{\theta} \mid \boldsymbol{x})}{q(\boldsymbol{\theta} \mid \boldsymbol{x})}\right]\right].$$

We can then state the data-processing inequality and the chain rule for the KL divergence.

Lemma A.2 (Theorem 2.15 & Corollary 2.18 from Polyanskiy and Wu (2025). Chain rule and data-processing inequality for KL divergence)**.** *Suppose that for some measurable space $\mathcal{Y}$, $f : \Theta \to \mathcal{Y}$ is a measurable function. Then, for any fixed $\boldsymbol{x} \in \mathcal{X}$ and any two distributions $P_{\boldsymbol{\theta}|\boldsymbol{x}}, Q_{\boldsymbol{\theta}|\boldsymbol{x}}$, such that $P_{\boldsymbol{\theta}|\boldsymbol{x}} \ll Q_{\boldsymbol{\theta}|\boldsymbol{x}}$, i.e. $P_{\boldsymbol{\theta}|\boldsymbol{x}}$ is absolutely continuous w.r.t. $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$,*

$$\mathrm{KL}(P_{\boldsymbol{\theta}|\boldsymbol{x}} \| Q_{\boldsymbol{\theta}|\boldsymbol{x}}) = \mathrm{KL}(f_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}} \| f_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}) + \mathbb{E}_{t\sim f_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}}}\left[\mathrm{KL}(P_{\boldsymbol{\theta}|f(\boldsymbol{\theta})=t,\boldsymbol{x}} \| Q_{\boldsymbol{\theta}|f(\boldsymbol{\theta})=t,\boldsymbol{x}})\right]. \quad (5)$$

In particular,

$$\mathrm{KL}(f_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}} \| f_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}) \le \mathrm{KL}(P_{\boldsymbol{\theta}|\boldsymbol{x}} \| Q_{\boldsymbol{\theta}|\boldsymbol{x}}),$$

and equality holds if f is one-to-one, or when f is a sufficient statistic for testing $P_{\boldsymbol{\theta}|\boldsymbol{x}}$ against $Q_{\boldsymbol{\theta}|\boldsymbol{x}}$.

We now show that KL-based miscoverage controls the usual absolute coverage error. In particular, by Pinsker's inequality, small KL-based miscoverage implies small absolute miscoverage. When this upper bound is reasonably tight, reducing the former also reduces the latter. The proof uses the following elementary identity for Bernoulli distributions. Let P_1 and P_2 be Bernoulli distributions with probability parameters p_1 and p_2, respectively. Then

$$\mathrm{TV}(P_1, P_2) := \frac{1}{2}\sum_{s\in\{0,1\}} |P_1(s) - P_2(s)| = |p_1 - p_2|.$$

Lemma A.3 (KL-based miscoverage upper-bounds its absolute value notion)**.** *Define $\mathrm{kl}^{\alpha}_{\mathrm{cov}}$ as in Definition 3.1, and $\Delta^{\alpha}_{\mathrm{cov}}(q; P) := \left|C^P_{\alpha}(q) - (1-\alpha)\right|$. Then, for any $\alpha \in (0,1)$ and $q(\boldsymbol{\theta} \mid \boldsymbol{x})$,*

$$\Delta^{\alpha}_{\mathrm{cov}}(q; P) \le \sqrt{\frac{1}{2}\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q; P)}.$$

Proof. By definition of C^P_{α}, we have that

$$\Delta^{\alpha}_{\mathrm{cov}}(q; P) = |\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}[P_{\boldsymbol{\theta}|\boldsymbol{x}}(R^q_{\alpha}(\boldsymbol{x})) - (1-\alpha)]| \le \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}[|P_{\boldsymbol{\theta}|\boldsymbol{x}}(R^q_{\alpha}(\boldsymbol{x})) - (1-\alpha)|].$$

Then, for each fixed $\boldsymbol{x}$,

$$|P_{\boldsymbol{\theta}|\boldsymbol{x}}(R^q_{\alpha}(\boldsymbol{x})) - (1-\alpha)| = \mathrm{TV}\left((T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}},\, (T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}\right),$$

since $(T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}}$ and $(T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ are Bernoulli laws with parameters $P_{\boldsymbol{\theta}|\boldsymbol{x}}(R^q_{\alpha}(\boldsymbol{x}))$ and $Q_{\boldsymbol{\theta}|\boldsymbol{x}}(R^q_{\alpha}(\boldsymbol{x})) = 1-\alpha$. Therefore, applying Pinsker's inequality pointwise in $\boldsymbol{x}$ yields:

$$\begin{aligned}\Delta^{\alpha}_{\mathrm{cov}}(q; P) &\le \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{TV}\left((T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}},\, (T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}\right)\right] \\ &\le \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\sqrt{\frac{1}{2}\mathrm{KL}\left((T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}} \parallel (T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}\right)}\right].\end{aligned}$$

Finally, Jensen's inequality gives:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\sqrt{\frac{1}{2}\mathrm{KL}\left((T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}} \parallel (T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}\right)}\right] \\ &\le \sqrt{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{KL}\left((T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}} \parallel (T_{R^q_{\alpha}(\boldsymbol{x})})_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}\right)\right]} = \sqrt{\frac{1}{2}\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q; P)}.\end{aligned}$$

□

We end this section by showing that miscalibration is an upper-bound to miscoverage. This result will be used to prove Theorem 4.1 in the next section.

Lemma A.4 (Miscalibration upper-bounds miscoverage)**.** *Define*

$$\mathrm{kl}_{\mathrm{NPE}}(q;P) := \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{KL}(P_{\boldsymbol{\theta}|\boldsymbol{x}} \parallel Q_{\boldsymbol{\theta}|\boldsymbol{x}})\right] = \mathcal{L}_{\mathrm{NPE}}(q;P) + K(P).$$

Then, for any distribution Q with density q,

$$\sup_{\alpha\in(0,1)} \mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) \le \mathrm{kl}_{\mathrm{NPE}}(q;P) \qquad \textit{and} \qquad \mathrm{kl}^{S}_{\mathrm{cal}}(q;P) \le \mathrm{kl}_{\mathrm{NPE}}(q;P).$$

Further, suppose that $S(\boldsymbol{\theta};\boldsymbol{x}) = f(q(\boldsymbol{\theta} \mid \boldsymbol{x}))$ for some strictly increasing $f : [0,\infty) \to \mathbb{R}$, and $F^{Q}_{S|\boldsymbol{x}}$ is also strictly increasing. Then

$$\sup_{\alpha\in(0,1)} \mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) \le \mathrm{kl}^{S}_{\mathrm{cal}}(q;P) \le \mathrm{kl}_{\mathrm{NPE}}(q;P).$$

Proof. Showing that $\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) \le \mathrm{kl}_{\mathrm{NPE}}(q;P)$ and $\mathrm{kl}^{S}_{\mathrm{cal}}(q;P) \le \mathrm{kl}_{\mathrm{NPE}}(q;P)$ is a direct application of the data-processing inequality (DPI) from Lemma A.2. In particular, since $\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) \le \mathrm{kl}_{\mathrm{NPE}}(q;P)$ holds for any $\alpha \in (0,1)$, we can further conclude that $\sup_{\alpha\in(0,1)} \mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) \le \mathrm{kl}_{\mathrm{NPE}}(q;P)$. We then focus on showing $\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) \le \mathrm{kl}^{S}_{\mathrm{cal}}(q;P)$.

First, whenever $S(\boldsymbol{\theta};\boldsymbol{x}) = f(q(\boldsymbol{\theta} \mid \boldsymbol{x}))$ and f is strictly increasing, then

$$\begin{aligned}
(T_{R^q_\alpha(\boldsymbol{x})})_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}}(1) &= P_{\boldsymbol{\theta}|\boldsymbol{x}}(R^q_\alpha(\boldsymbol{x})) \\
&= P_{\boldsymbol{\theta}|\boldsymbol{x}}(\{q(\boldsymbol{\theta} \mid \boldsymbol{x}) \ge \tau^q_\alpha(\boldsymbol{x})\}) \\
&= P_{\boldsymbol{\theta}|\boldsymbol{x}}(\{f(q(\boldsymbol{\theta} \mid \boldsymbol{x})) \ge f(\tau^q_\alpha(\boldsymbol{x}))\}) \quad \text{since } f \text{ is strictly increasing,} \\
&= P_{S|\boldsymbol{x}}(\{s \ge f(\tau^q_\alpha(\boldsymbol{x}))\}) \quad \text{using that } S(\boldsymbol{\theta};\boldsymbol{x}) = f(q(\boldsymbol{\theta} \mid \boldsymbol{x})), \\
&= (T_{A^q_\alpha(\boldsymbol{x})})_{\#}P_{S|\boldsymbol{x}}(1),
\end{aligned}$$

where $A^q_\alpha(\boldsymbol{x}) := \{s \ge f(\tau^q_\alpha(\boldsymbol{x}))\}$, and similarly,

$$(T_{R^q_\alpha(\boldsymbol{x})})_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}}(0) = P_{S|\boldsymbol{x}}(\{s < f(\tau^q_\alpha(\boldsymbol{x}))\}) = (T_{A^q_\alpha(\boldsymbol{x})})_{\#}P_{S|\boldsymbol{x}}(0)$$

Therefore, the laws $(T_{R^q_\alpha(\boldsymbol{x})})_{\#}P_{\boldsymbol{\theta}|\boldsymbol{x}}$ and $(T_{A^q_\alpha(\boldsymbol{x})})_{\#}P_{S|\boldsymbol{x}}$ coincide. This also holds for $(T_{R^q_\alpha(\boldsymbol{x})})_{\#}Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ and $(T_{A^q_\alpha(\boldsymbol{x})})_{\#}Q_{S|\boldsymbol{x}}$. Hence, we can rewrite miscoverage as follows:

$$\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) = \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{KL}\left((T_{A^q_\alpha(\boldsymbol{x})})_{\#}P_{S|\boldsymbol{x}} \,\middle\|\, (T_{A^q_\alpha(\boldsymbol{x})})_{\#}Q_{S|\boldsymbol{x}}\right)\right].$$

But,

$$\begin{aligned}
T_{A^q_\alpha(\boldsymbol{x})}(s) &= \mathbf{1}\{s \ge f(\tau^q_\alpha(\boldsymbol{x}))\} \\
&= \mathbf{1}\{F^Q_{S|\boldsymbol{x}}(s) \ge F^Q_{S|\boldsymbol{x}}(f(\tau^q_\alpha(\boldsymbol{x})))\} \quad \text{since } F^Q_{S|\boldsymbol{x}} \text{ is strictly increasing,} \\
&= T_{B^q_\alpha(\boldsymbol{x})}(F^Q_{S|\boldsymbol{x}}(s))
\end{aligned}$$

for $B^q_\alpha(\boldsymbol{x}) := [F^Q_{S|\boldsymbol{x}}(f(\tau^q_\alpha(\boldsymbol{x}))), 1]$. Therefore,

$$(T_{A^q_\alpha(\boldsymbol{x})})_{\#}P_{S|\boldsymbol{x}} = (T_{B^q_\alpha(\boldsymbol{x})})_{\#}\left((F^Q_{S|\boldsymbol{x}})_{\#}P_{S|\boldsymbol{x}}\right),$$

and similarly for $Q_{S|\boldsymbol{x}}$. Thus,

$$\begin{aligned}
\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) &= \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{KL}\left((T_{B^q_\alpha(\boldsymbol{x})})_{\#}\left((F^Q_{S|\boldsymbol{x}})_{\#}P_{S|\boldsymbol{x}}\right) \middle\| (T_{B^q_\alpha(\boldsymbol{x})})_{\#}\left((F^Q_{S|\boldsymbol{x}})_{\#}Q_{S|\boldsymbol{x}}\right)\right)\right] \\
&\le \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{KL}\left((F^Q_{S|\boldsymbol{x}})_{\#}P_{S|\boldsymbol{x}} \middle\| (F^Q_{S|\boldsymbol{x}})_{\#}Q_{S|\boldsymbol{x}}\right)\right] \quad \text{by the DPI,} \\
&= \mathrm{kl}^{S}_{\mathrm{cal}}(q;P).
\end{aligned}$$

Since this line of reasoning holds for any $\alpha \in (0,1)$, we further conclude that

$$\sup_{\alpha\in(0,1)} \mathrm{kl}^{\alpha}_{\mathrm{cov}}(q;P) \le \mathrm{kl}^{S}_{\mathrm{cal}}(q;P).$$

□

B Theory: DRO-NPE method

In Section B.1, we adapt a Wasserstein concentration result and use it to prove Theorem 4.1. Then, in Section B.2, we specialise to the Wasserstein distance with $p = 2$, and derive a local quadratic upper bound under a growth condition and a continuous differentiability assumption, which in turn yields Proposition 4.2. In Section B.3, we further prove an excess risk type of result for miscoverage and miscalibration under stricter conditions. Finally, we verify these conditions for normalising flows (Papamakarios et al., 2021; Kobyzev et al., 2020).

B.1 Wasserstein concentration results and proof of Theorem 4.1

In this subsection, we first adapt a result from Fournier and Guillin (2015) to our setting. Then, we use this result, as well as the data-processing inequality (Lemma A.2), to prove Theorem 4.1.

Lemma B.1 (Adapting Theorem 2 from Fournier and Guillin (2015))**.** *Suppose that $p \geq 1$, and that there exists $\beta > 2p$ such that $M_\beta := \mathbb{E}_{\boldsymbol{z}\sim P}[\|\boldsymbol{z}\|^\beta] < \infty$. Then, $\forall \varsigma \in (0, \beta)$ there exist positive constants C_1, C_2, with $C_2 \geq 1$, depending only on $p, d_{\mathcal{Z}}, \beta, M_\beta, \varsigma$ such that for any $\delta \in (0, 1)$,*

$$\mathbb{P}(W_p(P_n, P) \leq \varepsilon) \geq 1 - \delta,$$

whenever $\varepsilon \geq \varepsilon(n, \delta) := \max\{\varepsilon_a(n, \delta),\, \varepsilon_b(n, \delta)\}$, where, writing $\rho := \left(\frac{1}{C_1} \log \frac{2C_2}{\delta}\right)^{1/2}$,

$$\varepsilon_a(n,\delta) := \begin{cases} \rho^{1/p}\, n^{-1/(2p)}, & \text{if } d_{\mathcal{Z}} < 2p, \\ \left(2\rho\, n^{-1/2} \log\left(2 + \dfrac{n}{\rho^2}\right)\right)^{1/p}, & \text{if } d_{\mathcal{Z}} = 2p, \\ \rho^{2/d_{\mathcal{Z}}}\, n^{-1/d_{\mathcal{Z}}}, & \text{if } d_{\mathcal{Z}} > 2p, \end{cases} \qquad \varepsilon_b(n,\delta) := \left(\frac{2C_2}{\delta}\right)^{\frac{1}{\beta-\varsigma}} n^{\frac{1}{\beta-\varsigma}-\frac{1}{p}}.$$

Proof. From Theorem 2 of Fournier and Guillin (2015), whenever P is a probability distribution on $\mathbb{R}^{d_{\mathcal{Z}}}$, if there exists $\beta > 2p$ such that $M_\beta := \mathbb{E}_{\boldsymbol{z}\sim P}[\|\boldsymbol{z}\|^\beta] < \infty$, then for all $n \geq 1, \varepsilon > 0$,

$$\mathbb{P}(W_p^p(P_n, P) \geq \varepsilon) \leq a(n, \varepsilon)\mathbf{1}_{\varepsilon \leq 1} + b(n, \varepsilon),$$

where

$$a(n,\varepsilon) = C_2 \begin{cases} \exp(-C_1 n \varepsilon^2), & \text{if } d_{\mathcal{Z}} < 2p, \\ \exp\left(-C_1 n \left(\frac{\varepsilon}{\log\left(2+\frac{1}{\varepsilon}\right)}\right)^2\right), & \text{if } d_{\mathcal{Z}} = 2p, \\ \exp\left(-C_1 n \varepsilon^{\frac{d_{\mathcal{Z}}}{p}}\right), & \text{if } d_{\mathcal{Z}} > 2p, \end{cases}$$

and

$$b(n, \varepsilon) = C_2 n (n\varepsilon)^{-\frac{\beta-\varsigma}{p}} \quad \forall \varsigma \in (0, \beta).$$

The positive constants C_1, C_2 depend only on $p, d_{\mathcal{Z}}$ and on $\beta, M_\beta, \varsigma$, and we may take $C_2 \geq 1$ without loss of generality. Therefore,

$$\mathbb{P}(W_p(P_n, P) \leq \varepsilon) \geq 1 - \mathbb{P}(W_p^p(P_n, P) \geq \varepsilon^p) \geq 1 - a(n, \varepsilon^p)\mathbf{1}_{\varepsilon^p \leq 1} - b(n, \varepsilon^p).$$

Fix $\delta \in (0, 1)$. We wish to find conditions for ε in terms of δ, n and other constants such that:

$$\mathbb{P}(W_p(P_n, P) \leq \varepsilon) \geq 1 - \delta.$$

We do so by selecting ε such that it satisfies both $a(n, \varepsilon^p)\mathbf{1}_{\varepsilon^p \leq 1} \leq \delta/2,\ b(n, \varepsilon^p) \leq \delta/2$, so that:

$$1 - a(n, \varepsilon^p)\mathbf{1}_{\varepsilon^p \leq 1} - b(n, \varepsilon^p) \geq 1 - \delta \implies \mathbb{P}(W_p(P_n, P) \leq \varepsilon) \geq 1 - \delta.$$

The bound on b does not depend on the regime of p:

$$b(n, \varepsilon^p) = C_2 n (n\varepsilon^p)^{-\frac{\beta-\varsigma}{p}} \leq \frac{\delta}{2} \iff \varepsilon \geq \varepsilon_b = \left(\frac{2C_2}{\delta}\right)^{\frac{1}{\beta-\varsigma}} n^{\frac{1}{\beta-\varsigma}-\frac{1}{p}}.$$

For the bound on a, we distinguish three cases.

Case 1: $d_{\mathcal{Z}} > 2p$

$$a(n, \varepsilon^p) = C_2 \exp(-C_1 n \varepsilon^{d_{\mathcal{Z}}}) \leq \frac{\delta}{2} \iff \varepsilon \geq \varepsilon_a = \left(\frac{1}{C_1 n} \log \frac{2C_2}{\delta} \right)^{\frac{1}{d_{\mathcal{Z}}}} = \rho^{2/d_{\mathcal{Z}}} \, n^{-1/d_{\mathcal{Z}}}.$$

Case 2: $d_{\mathcal{Z}} < 2p$

$$a(n, \varepsilon^p) = C_2 \exp(-C_1 n \varepsilon^{2p}) \leq \frac{\delta}{2} \iff \varepsilon \geq \varepsilon_a = \left(\frac{1}{C_1 n} \log \frac{2C_2}{\delta} \right)^{\frac{1}{2p}} = \rho^{1/p} \, n^{-1/(2p)}.$$

Case 3: $d_{\mathcal{Z}} = 2p$

$$a(n, \varepsilon^p) = C_2 \exp\left(-C_1 n \left(\frac{\varepsilon^p}{\log(2 + \varepsilon^{-p})} \right)^2 \right) \leq \frac{\delta}{2} \iff \frac{\varepsilon^p}{\log(2 + \varepsilon^{-p})} \geq \rho_n,$$

where $\rho_n := \rho \, n^{-1/2}$. Let $y_n := 2\rho_n \log(2 + \rho_n^{-2}) = 2\rho \, n^{-1/2} \log\left(2 + \frac{n}{\rho^2}\right)$. We show that $\varepsilon^p \geq y_n$ suffices. If $\rho_n \geq 1$, then $y_n \geq 2\rho_n \log 2 > 1$, so $\varepsilon^p > 1$ and $\mathbf{1}_{\varepsilon^p \leq 1} = 0$. If $\rho_n < 1$, since $f(y) := y/\log(2 + 1/y)$ is increasing on $(0, \infty)$, it suffices to show $f(y_n) \geq \rho_n$. As $2 \log 2 > 1 > \rho_n$, we have $y_n \geq 2\rho_n \log 2 \geq \rho_n^2$, so $1/y_n \leq \rho_n^{-2}$ and

$$\rho_n \log(2 + 1/y_n) \leq \rho_n \log(2 + \rho_n^{-2}) = y_n/2 \leq y_n,$$

i.e. $f(y_n) \geq \rho_n$. Hence $\varepsilon \geq \varepsilon_a = y_n^{1/p}$ suffices.

We then select, in all three cases,

$$\varepsilon(n, \delta) := \max \left\{ \varepsilon_a(n, \delta), \; \varepsilon_b(n, \delta) \right\}.$$

□

We are now ready to prove the Theorem.

Proof of Theorem 4.1. By Lemma A.4, we have that for any S, any q_ϕ and every $\alpha \in (0, 1)$:

$$\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q_\phi; P) \leq \mathrm{kl}^{S}_{\mathrm{cal}}(q_\phi; P) \leq \mathcal{L}_{\mathrm{NPE}}(q_\phi; P) + K(P).$$

Taking the supremum over $\alpha \in (0, 1)$ gives the first part of the theorem. Further, the assumptions of Lemma B.1 are satisfied. The lemma holds for any $\varsigma \in (0, \beta)$; for readability, we fix $\varsigma := \beta - 2p \in (0, \beta)$, so that $\beta - \varsigma = 2p$ and we recover $\varepsilon_b(n, \delta)$, as well as constants C_1, C_2 depending only on $p, d_{\mathcal{Z}}, \beta, M_\beta$, as in Theorem 4.1. We obtain that whenever $\varepsilon \geq \varepsilon(n, \delta)$,

$$\mathbb{P}(W_p(P, P_n) \leq \varepsilon) \geq 1 - \delta,$$

which implies that with probability at least $1 - \delta$, $P \in \mathcal{A}_p(P_n; \varepsilon)$. This gives that with probability at least $1 - \delta$, for any q_ϕ,

$$\mathcal{L}_{\mathrm{NPE}}(q_\phi; P) \leq \sup_{\tilde{P} \in \mathcal{A}_p(P_n; \varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_\phi; \tilde{P}).$$

Hence, putting the inequalities together, with probability at least $1 - \delta$, for any q_ϕ,

$$\mathrm{kl}^{\alpha}_{\mathrm{cov}}(q_\phi; P) \leq \mathrm{kl}^{S}_{\mathrm{cal}}(q_\phi; P) \leq \mathcal{L}_{\mathrm{NPE}}(q_\phi; P) + K(P) \leq \sup_{\tilde{P} \in \mathcal{A}_p(P_n; \varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_\phi; \tilde{P}) + K(P).$$

□

B.2 Proof of Proposition 4.2

We first show that under the assumption of $2-$growth, we can prove a local quadratic upper bound, so long as the function class is also twice continuously differentiable on $\mathbb{R}^{d_{\mathcal{Z}}}$.

Lemma B.2. *Fix $q_\phi \in Q$. Suppose that $\boldsymbol{z} \mapsto \ell(q_\phi; \boldsymbol{z}) \in \mathcal{G}_2(\mathcal{Z})$, and that it is twice continuously differentiable on $\mathbb{R}^{d_{\mathcal{Z}}}$. Then, for any fixed $\boldsymbol{z}_0 \in \mathcal{Z}$, there exists a $C_\phi(\boldsymbol{z}_0) < \infty$ which is constant in $\boldsymbol{z}$ such that*

$$\ell(q_\phi; \boldsymbol{z}) \leq \ell(q_\phi; \boldsymbol{z}_0) + \nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)^\top (\boldsymbol{z} - \boldsymbol{z}_0) + C_\phi(\boldsymbol{z}_0) \|\boldsymbol{z} - \boldsymbol{z}_0\|^2, \quad \forall \boldsymbol{z} \in \mathcal{Z}.$$

Proof. We proceed by considering two cases: when $\|\boldsymbol{z} - \boldsymbol{z}_0\| < 1$ and $\|\boldsymbol{z} - \boldsymbol{z}_0\| \geq 1$. We denote $B(\boldsymbol{z}_0, 1) := \{\boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{z}_0\| < 1\}$, and $\bar{B}(\boldsymbol{z}_0, 1)$ its closure.

Case 1: $\|\boldsymbol{z} - \boldsymbol{z}_0\| < 1$: Since $B(\boldsymbol{z}_0, 1) \subset \mathbb{R}^{d_{\boldsymbol{z}}}$, and the mapping $\boldsymbol{z} \mapsto \ell(q_\phi; \boldsymbol{z})$ is continuously differentiable on $\mathbb{R}^{d_{\boldsymbol{z}}}$, which is open and convex, Taylor's theorem (Theorem 2.68 of Folland (2024)) applies at $\boldsymbol{z}_0$ for every $\boldsymbol{z} \in B(\boldsymbol{z}_0, 1)$:

$$\ell(q_\phi; \boldsymbol{z}) = \ell(q_\phi; \boldsymbol{z}_0) + \nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)^\top (\boldsymbol{z} - \boldsymbol{z}_0) + R_1(\boldsymbol{z}_0, \boldsymbol{z} - \boldsymbol{z}_0), \quad \forall \boldsymbol{z} \in B(\boldsymbol{z}_0, 1)$$

where $R_1(\boldsymbol{z}_0, \boldsymbol{z} - \boldsymbol{z}_0)$ is the Taylor expansion remainder term. Further, since $\bar{B}(\boldsymbol{z}_0, 1)$ is compact and the Hessian $\nabla^2_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z})$ is continuous by assumption, the Hessian has finite norm: $\|\nabla^2_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z})\| \leq M(\boldsymbol{z}_0)\ \forall \boldsymbol{z} \in \bar{B}(\boldsymbol{z}_0, 1)$ for some constant $M(\boldsymbol{z}_0)$ independent of $\boldsymbol{z}$. Therefore, with Corollary 2.75 of Folland (2024), we obtain:

$$|R_1(\boldsymbol{z}_0, \boldsymbol{z} - \boldsymbol{z}_0)| \leq \frac{M(\boldsymbol{z}_0)}{2} \|\boldsymbol{z} - \boldsymbol{z}_0\|^2, \quad \forall \boldsymbol{z} \in B(\boldsymbol{z}_0, 1).$$

We then conclude:

$$\ell(q_\phi; \boldsymbol{z}) \leq \ell(q_\phi; \boldsymbol{z}_0) + \nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)^\top (\boldsymbol{z} - \boldsymbol{z}_0) + \frac{M(\boldsymbol{z}_0)}{2} \|\boldsymbol{z} - \boldsymbol{z}_0\|^2, \quad \forall \boldsymbol{z} \in B(\boldsymbol{z}_0, 1).$$

Case 2: $\|\boldsymbol{z} - \boldsymbol{z}_0\| \geq 1$: Notice that

$$-\nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)^\top (\boldsymbol{z} - \boldsymbol{z}_0) \leq \|\nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)\| \|\boldsymbol{z} - \boldsymbol{z}_0\| \leq \frac{1}{2} \|\nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)\|^2 + \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{z}_0\|^2,$$

where the first inequality is by Cauchy-Schwarz, and the second holds because whenever $a, b > 0$, $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$. Further, using the growth condition of the map $\boldsymbol{z} \mapsto \ell(q_\phi; \boldsymbol{z})$ for $\rho = 2$, for some constant $\tilde{C}_\phi$,

$$\ell(q_\phi; \boldsymbol{z}) \leq \tilde{C}_\phi (1 + \|\boldsymbol{z}_0 + \boldsymbol{z} - \boldsymbol{z}_0\|^2) \leq \tilde{C}_\phi (1 + 2\|\boldsymbol{z}_0\|^2 + 2\|\boldsymbol{z} - \boldsymbol{z}_0\|^2),$$

where the second inequality is using that $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Together, these results imply:

$$\begin{aligned}
\ell(q_\phi; \boldsymbol{z}) - \nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)^\top (\boldsymbol{z} - \boldsymbol{z}_0) &\leq \tilde{C}_\phi (1 + 2\|\boldsymbol{z}_0\|^2 + 2\|\boldsymbol{z} - \boldsymbol{z}_0\|^2) \\
&\quad + \frac{1}{2} \|\nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)\|^2 + \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{z}_0\|^2 + \ell(q_\phi; \boldsymbol{z}_0) - \ell(q_\phi; \boldsymbol{z}_0) \\
&:= K_\phi(\boldsymbol{z}_0) + \ell(q_\phi; \boldsymbol{z}_0) + \left(2\tilde{C}_\phi + \frac{1}{2}\right) \|\boldsymbol{z} - \boldsymbol{z}_0\|^2,
\end{aligned}$$

for some constant $K_\phi(\boldsymbol{z}_0)$ independent of $\boldsymbol{z}$, defined as:

$$K_\phi(\boldsymbol{z}_0) := \tilde{C}_\phi + 2\tilde{C}_\phi \|\boldsymbol{z}_0\|^2 + \frac{1}{2} \|\nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)\|^2 - \ell(q_\phi; \boldsymbol{z}_0).$$

Finally, since $\|z - \boldsymbol{z}_0\| \geq 1$, we have that

$$K_\phi(\boldsymbol{z}_0) + \left(2\tilde{C}_\phi + \frac{1}{2}\right) \|\boldsymbol{z} - \boldsymbol{z}_0\|^2 \leq \left(|K_\phi(\boldsymbol{z}_0)| + 2\tilde{C}_\phi + \frac{1}{2}\right) \|\boldsymbol{z} - \boldsymbol{z}_0\|^2 := A_\phi(\boldsymbol{z}_0) \|\boldsymbol{z} - \boldsymbol{z}_0\|^2,$$

implying

$$\ell(q_\phi; \boldsymbol{z}) \leq \ell(q_\phi; \boldsymbol{z}_0) + \nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)^\top (\boldsymbol{z} - \boldsymbol{z}_0) + A_\phi(\boldsymbol{z}_0) \|\boldsymbol{z} - \boldsymbol{z}_0\|^2, \quad \forall \boldsymbol{z} \in \{\boldsymbol{z} \in \mathcal{Z} : \|\boldsymbol{z} - \boldsymbol{z}_0\| \geq 1\}.$$

Finally, taking $C_\phi(\boldsymbol{z}_0) := \max\left\{\frac{M(\boldsymbol{z}_0)}{2}, A_\phi(\boldsymbol{z}_0)\right\}$, we get that

$$\ell(q_\phi; \boldsymbol{z}) \leq \ell(q_\phi; \boldsymbol{z}_0) + \nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_0)^\top (\boldsymbol{z} - \boldsymbol{z}_0) + C_\phi(\boldsymbol{z}_0) \|\boldsymbol{z} - \boldsymbol{z}_0\|^2 \quad \forall \boldsymbol{z} \in \mathcal{Z}.$$

□

This quadratic upper bound can then be used to solve the inner supremum in the strong dual of the Wasserstein DRO objective and prove Proposition 4.2. Note that the proof keeps the loss function more general, with ℓ_{NPE} recovered by setting $\ell \equiv \ell_{\text{NPE}}$.

Proof of Proposition 4.2. We divide this proof into two parts. First, we derive the upper bound in terms of the gradient regulariser and the C_ϕ^n remainder term. Second, we show that the remainder term remains negligible under the assumptions considered in the Proposition. We keep the proof at a general level of loss function ℓ, and do not use ℓ_{NPE} specifically.

First, recall that, as in Equation (3), under the upper semi-continuity and $2-$growth assumptions on ℓ, which hold in this setting, we have (Gao and Kleywegt, 2023):

$$\sup_{\tilde{P}\in\mathcal{A}_p(P_n;\varepsilon)} \mathbb{E}_{\boldsymbol{z}\sim\tilde{P}}[\ell(q_\phi;\boldsymbol{z})] = \inf_{\lambda\geq 0}\left\{\lambda\varepsilon^2 + \frac{1}{n}\sum_{i=1}^n \sup_{\boldsymbol{z}'\in\mathcal{Z}}\left(\ell(q_\phi;\boldsymbol{z}') - \lambda\|\boldsymbol{z}'-\boldsymbol{z}_i\|^2\right)\right\}.$$

Further, by Lemma B.2, for each i and every $\boldsymbol{z}'\in\mathcal{Z}$,

$$\ell(q_\phi;\boldsymbol{z}') \leq \ell(q_\phi;\boldsymbol{z}_i) + \nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i)^\top(\boldsymbol{z}'-\boldsymbol{z}_i) + C_\phi(\boldsymbol{z}_i)\|\boldsymbol{z}'-\boldsymbol{z}_i\|^2.$$

Hence, for every $\lambda\geq 0$,

$$\begin{aligned}\sup_{\boldsymbol{z}'\in\mathcal{Z}}\left(\ell(q_\phi;\boldsymbol{z}') - \lambda\|\boldsymbol{z}'-\boldsymbol{z}_i\|^2\right) &\leq \ell(q_\phi;\boldsymbol{z}_i) + \sup_{\boldsymbol{z}'\in\mathcal{Z}}\Big\{\nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i)^\top(\boldsymbol{z}'-\boldsymbol{z}_i) \\ &\qquad + (C_\phi(\boldsymbol{z}_i)-\lambda)\|\boldsymbol{z}'-\boldsymbol{z}_i\|^2\Big\} \\ &\leq \ell(q_\phi;\boldsymbol{z}_i) + \sup_{\boldsymbol{z}'\in\mathbb{R}^{d_{\mathcal{Z}}}}\Big\{\nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i)^\top(\boldsymbol{z}'-\boldsymbol{z}_i) \\ &\qquad + (C_\phi(\boldsymbol{z}_i)-\lambda)\|\boldsymbol{z}'-\boldsymbol{z}_i\|^2\Big\}.\end{aligned}$$

Now, restrict λ such that $\lambda > C_\phi(\boldsymbol{z}_i), \forall i = 1,\dots,n$. Then, since the objective inside the supremum is strictly concave and over $\mathbb{R}^{d_{\mathcal{Z}}}$, we get that the solution for every i to the following is the unique maximiser $\boldsymbol{z}_i^\star$:

$$\begin{aligned}0 &= \nabla_{\boldsymbol{z}'}\left(\nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i)^\top(\boldsymbol{z}'-\boldsymbol{z}_i) + (C_\phi(\boldsymbol{z}_i)-\lambda)\|\boldsymbol{z}'-\boldsymbol{z}_i\|^2\right) \\ &= \nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i) + 2(C_\phi(\boldsymbol{z}_i)-\lambda)(\boldsymbol{z}'-\boldsymbol{z}_i) \implies \boldsymbol{z}_i^\star = \boldsymbol{z}_i + \frac{\nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i)}{2(\lambda - C_\phi(\boldsymbol{z}_i))}.\end{aligned}$$

Fix $C_\phi^n := \max_{1\leq i\leq n} C_\phi(\boldsymbol{z}_i)$. Then, substituting back, we get that for every $\lambda > C_\phi^n$,

$$\begin{aligned}&\lambda\varepsilon^2 + \frac{1}{n}\sum_{i=1}^n \sup_{\boldsymbol{z}'\in\mathcal{Z}}\left(\ell(q_\phi;\boldsymbol{z}') - \lambda\|\boldsymbol{z}'-\boldsymbol{z}_i\|^2\right) \\ &\leq \frac{1}{n}\sum_{i=1}^n \ell(q_\phi;\boldsymbol{z}_i) + \lambda\varepsilon^2 + \frac{1}{4(\lambda - C_\phi^n)}\cdot\frac{1}{n}\sum_{i=1}^n\|\nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i)\|^2 \\ &= \frac{1}{n}\sum_{i=1}^n \ell(q_\phi;\boldsymbol{z}_i) + C_\phi^n\varepsilon^2 + t\varepsilon^2 + \frac{1}{4t}\cdot\frac{1}{n}\sum_{i=1}^n\|\nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i)\|^2,\end{aligned}$$

where in the last line we substitute $t = \lambda - C_\phi^n$. Now,

$$\begin{aligned}&\inf_{\lambda\geq 0}\left\{\lambda\varepsilon^2 + \frac{1}{n}\sum_{i=1}^n \sup_{\boldsymbol{z}'\in\mathcal{Z}}\left(\ell(q_\phi;\boldsymbol{z}') - \lambda\|\boldsymbol{z}'-\boldsymbol{z}_i\|^2\right)\right\} \\ &\leq \inf_{\lambda - C_\phi^n > 0}\left\{\lambda\varepsilon^2 + \frac{1}{n}\sum_{i=1}^n \sup_{\boldsymbol{z}'\in\mathcal{Z}}\left(\ell(q_\phi;\boldsymbol{z}') - \lambda\|\boldsymbol{z}'-\boldsymbol{z}_i\|^2\right)\right\} \\ &\leq \inf_{t>0}\left\{\frac{1}{n}\sum_{i=1}^n \ell(q_\phi;\boldsymbol{z}_i) + C_\phi^n\varepsilon^2 + t\varepsilon^2 + \frac{1}{4t}\cdot\frac{1}{n}\sum_{i=1}^n\|\nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i)\|^2\right\} \\ &= \frac{1}{n}\sum_{i=1}^n \ell(q_\phi;\boldsymbol{z}_i) + \varepsilon\left(\frac{1}{n}\sum_{i=1}^n\|\nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i)\|^2\right)^{1/2} + C_\phi^n\varepsilon^2,\end{aligned}$$

where in the first line we restrict the infimum to $\lambda > C_\phi^n$, in the second line we use the previously derived upper-bound and replace $t = \lambda - C_\phi^n$, and in the last line, we use that for some A, ε constant in t,

$$\inf_{t>0}\left(t\varepsilon^2 + \frac{A}{4t}\right) = \varepsilon\sqrt{A}.$$

For the remainder term, recall that

$$C_\phi^n := \max_{1\le i\le n} C_\phi(\boldsymbol{z}_i), \quad C_\phi(\boldsymbol{z}_i) := \max\left\{\frac{1}{2}\sup_{\boldsymbol{z}\in\bar{B}(\boldsymbol{z}_i,1)}\|\nabla_{\boldsymbol{z}}^2\ell(q_\phi;\boldsymbol{z})\|,\ 2\tilde{C}_\phi+\frac{1}{2}+|K_\phi(\boldsymbol{z}_i)|\right\},$$

with

$$K_\phi(\boldsymbol{z}_i) = \tilde{C}_\phi + 2\tilde{C}_\phi\|\boldsymbol{z}_i\|^2 + \frac{1}{2}\|\nabla_{\boldsymbol{z}}\ell(q_\phi;\boldsymbol{z}_i)\|^2 - \ell(q_\phi;\boldsymbol{z}_i).$$

Then, using the growth condition for $\ell, \nabla_{\boldsymbol{z}}\ell$, there exist $C_{0,\phi} < \infty$ and $C_{1,\phi} < \infty$ such that

$$K_\phi(\boldsymbol{z}_i) \le \tilde{C}_\phi + 2\tilde{C}_\phi\|\boldsymbol{z}_i\|^2 + \frac{1}{2}(C_{1,\phi}+C_{1,\phi}\|\boldsymbol{z}_i\|^\beta)^2 + C_{0,\phi}(1+\|\boldsymbol{z}_i\|^2) \le C(1+\|\boldsymbol{z}_i\|^{\max\{2,2\beta\}})$$

for some $C < \infty$ independent of $\boldsymbol{z}_i$. Further, from the growth assumption for $\nabla^2\ell$ there exists $C_{2,\phi} < \infty$ such that:

$$\frac{1}{2}\sup_{\boldsymbol{z}\in\bar{B}(\boldsymbol{z}_i,1)}\|\nabla^2\ell(q_\phi;\boldsymbol{z})\| \le \frac{1}{2}\sup_{\boldsymbol{z}\in\bar{B}(\boldsymbol{z}_i,1)} C_{2,\phi}(1+\|\boldsymbol{z}\|^\alpha) \le \frac{C_{2,\phi}}{2}(1+(1+\|\boldsymbol{z}_i\|)^\alpha) \le C'(1+\|\boldsymbol{z}_i\|^\alpha)$$

for some constant C' independent of $\boldsymbol{z}_i$. Hence, for any $\boldsymbol{z}_i$, there exists some $A_\phi < \infty$, such that

$$C_\phi(\boldsymbol{z}_i) \le A_\phi(1+\|\boldsymbol{z}_i\|^m), \quad m := \max\{2, 2\beta, \alpha\}.$$

Therefore, we obtain:

$$C_\phi^n = \max_{1\le i\le n} C_\phi(\boldsymbol{z}_i) \le A_\phi(1+\max_{1\le i\le n}\|\boldsymbol{z}_i\|^m) := \bar{C}_\phi^n.$$

Now, we turn to the light-tail assumption. Recall that $\mathbb{E}_{\boldsymbol{z}\sim P}[\exp(\eta\|\boldsymbol{z}\|^a)] \le M$. By Markov's inequality, for $t > 0$,

$$\mathbb{P}(\|\boldsymbol{z}\| > t) = \mathbb{P}(\exp(\eta\|\boldsymbol{z}\|^a) > \exp(\eta t^a)) \le \frac{\mathbb{E}_{\boldsymbol{z}\sim P}[\exp(\eta\|\boldsymbol{z}\|^a)]}{\exp(\eta t^a)} \le M\exp(-\eta t^a).$$

Hence,

$$\begin{aligned}\mathbb{P}(\max_{1\le i\le n}\|\boldsymbol{z}_i\| \le t) &= \mathbb{P}(\|\boldsymbol{z}_1\| \le t, \ldots, \|\boldsymbol{z}_n\| \le t)\\ &= 1 - \mathbb{P}(\{\|\boldsymbol{z}_1\| > t\} \cup \cdots \cup \{\|\boldsymbol{z}_n\| > t\})\\ &\ge 1 - \sum_{i=1}^n \mathbb{P}(\|\boldsymbol{z}_i\| > t) \ge 1 - nM\exp(-\eta t^a).\end{aligned}$$

Now, since $m > 0$, $\|\boldsymbol{z}\|^m$ is increasing, so we have that $\max_i\|\boldsymbol{z}_i\|^m = (\max_i\|\boldsymbol{z}_i\|)^m$. Hence,

$$\mathbb{P}\left(\max_{1\le i\le n}\|\boldsymbol{z}_i\| \le t\right) = \mathbb{P}\left((\max_{1\le i\le n}\|\boldsymbol{z}_i\|)^m \le t^m\right) = \mathbb{P}\left(\max_{1\le i\le n}\|\boldsymbol{z}_i\|^m \le t^m\right) = \mathbb{P}\left(\bar{C}_\phi^n \le A_\phi(1+t^m)\right)$$

so that

$$\mathbb{P}\left(\bar{C}_\phi^n \le A_\phi(1+t^m)\right) \ge 1 - nM\exp(-\eta t^a).$$

For some fixed $\delta > 0$, we then have that with probability at least $1-\delta$

$$C_\phi^n \le \bar{C}_\phi^n \le A_\phi\left(1+\left(\frac{1}{\eta}\log\frac{Mn}{\delta}\right)^{m/a}\right).$$

Selecting $\varepsilon_n = o((\log n)^{-m/2a})$ then yields the desired result:

$$C_\phi^n\varepsilon_n^2 \to 0 \quad \text{in probability under } P.$$

□

B.3 Excess risk, miscoverage, and miscalibration

In this section, we investigate the excess risk, miscoverage and miscalibration rates of the density estimator obtained by DRO-NPE. To obtain such theoretical results, we require a stricter assumption on the pointwise loss function ℓ: Lipschitz gradient, as stated in Assumption B.1.

Assumption B.1 (Lipschitz gradient. Gao (2023), Assumption 2)**.** *For every $q_\phi \in \mathcal{Q}$, the map $\boldsymbol{z} \mapsto \ell(q_\phi; \boldsymbol{z})$ is differentiable on $\mathcal{Z}$, and there exists $h > 0$ such that $\|\nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_1) - \nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z}_2)\| \leq h\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|$, $\forall \boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{Z}$, $\forall q_\phi \in \mathcal{Q}$.*

The first result, Theorem B.3, shows the excess miscoverage and miscalibration error rates of the DRO-NPE estimator when the objective is in its exact (or equivalently, strong dual) form relative to the oracle minimiser, which corresponds to the density estimator that minimises the population risk. The proof of Theorem B.3 relies on the excess DRO risk (or optimality gap), a quantity commonly studied in DRO (see e.g. Zeng and Lam, 2022; Gao, 2023), and uses a hybrid of Wasserstein concentration results from Fournier and Guillin (2015) and generalisation bounds from Gao (2023).

Theorem B.3 (DRO-NPE miscoverage and miscalibration rate)**.** *Define the oracle NPE and DRO minimisers as follows:*

$$\phi^\star := \arg\min_\phi \mathcal{L}_{\mathrm{NPE}}(q_\phi; P), \quad \phi_n^{\mathrm{DRO}} := \arg\min_\phi \sup_{\tilde{P} \in \mathcal{A}_2(P_n; \varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_\phi; \tilde{P}).$$

Also let

$$\begin{aligned} G^\alpha_{\mathrm{cov}}(q_\phi; P) :&= \mathrm{kl}_{\mathrm{NPE}}(q_\phi; P) - \mathrm{kl}^\alpha_{\mathrm{cov}}(q_\phi; P) \\ &= \mathbb{E}_{\boldsymbol{x} \sim P_{\boldsymbol{x}}} \left[\mathbb{E}_{t \sim (T_{R^q_\alpha(\boldsymbol{x})})_\# P_{\boldsymbol{\theta}|\boldsymbol{x}}} \left[\mathrm{KL}(P_{\boldsymbol{\theta}|T_{R^q_\alpha(\boldsymbol{x})}(\boldsymbol{\theta})=t,\boldsymbol{x}} \| Q_{\boldsymbol{\theta}|T_{R^q_\alpha(\boldsymbol{x})}(\boldsymbol{\theta})=t,\boldsymbol{x}}) \right] \right] \geq 0, \\ G^S_{\mathrm{cal}}(q_\phi; P) :&= \mathrm{kl}_{\mathrm{NPE}}(q_\phi; P) - \mathrm{kl}^S_{\mathrm{cal}}(q_\phi; P) \\ &= \mathbb{E}_{\boldsymbol{x} \sim P_{\boldsymbol{x}}} \left[\mathbb{E}_{u \sim (F^Q_{S|\boldsymbol{x}})_\# P_{S|\boldsymbol{x}}} \left[\mathrm{KL}\left(P_{\boldsymbol{\theta}|F^Q_{S|\boldsymbol{x}}(S(\boldsymbol{\theta};\boldsymbol{x}))=u,\boldsymbol{x}} \,\middle\|\, Q_{\boldsymbol{\theta}|F^Q_{S|\boldsymbol{x}}(S(\boldsymbol{\theta};\boldsymbol{x}))=u,\boldsymbol{x}} \right) \right] \right] \geq 0. \end{aligned}$$

Suppose the assumptions from Theorem 4.1 hold and pick $\varepsilon \geq \varepsilon(n, \delta)$ accordingly, for some $\delta > 0$. Further, suppose Assumption B.1 holds for $\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}) = -\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$. Then, with probability at least $1 - \delta$, for any $\alpha \in (0, 1)$,

$$\mathrm{kl}^\alpha_{\mathrm{cov}}(q_{\phi_n^{\mathrm{DRO}}}; P) - \mathrm{kl}^\alpha_{\mathrm{cov}}(q_{\phi^\star}; P) \leq 2\varepsilon\, \mathbb{E}_{\boldsymbol{z} \sim P}[\|\nabla \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z})\|^2]^{\frac{1}{2}} + 4h\varepsilon^2 + G^\alpha_{\mathrm{cov}}(q_{\phi^\star}; P) \quad (6)$$

with h as in Assumption B.1, and also:

$$\mathrm{kl}^S_{\mathrm{cal}}(q_{\phi_n^{\mathrm{DRO}}}; P) - \mathrm{kl}^S_{\mathrm{cal}}(q_{\phi^\star}; P) \leq 2\varepsilon\, \mathbb{E}_{\boldsymbol{z} \sim P}[\|\nabla \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z})\|^2]^{\frac{1}{2}} + 4h\varepsilon^2 + G^S_{\mathrm{cal}}(q_{\phi^\star}; P) \quad (7)$$

Proof. We pick $\varepsilon \geq \varepsilon(n, \delta)$, as stated in Theorem 4.1. Then, with probability at least $1 - \delta$,

$$\mathcal{L}_{\mathrm{NPE}}(q_{\phi_n^{\mathrm{DRO}}}; P) \leq \sup_{\tilde{P} \in \mathcal{A}_2(P_n; \varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_{\phi_n^{\mathrm{DRO}}}; \tilde{P}) \leq \sup_{\tilde{P} \in \mathcal{A}_2(P_n; \varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; \tilde{P}),$$

where the first inequality holds because the concentration result ensures that $P \in \mathcal{A}_2(P_n; \varepsilon)$, and the second holds by the definition of the DRO minimiser ϕ_n^{DRO}. We further conclude that with probability $1 - \delta$,

$$\sup_{\tilde{P} \in \mathcal{A}_2(P_n; \varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; \tilde{P}) \leq \sup_{\tilde{P} \in \mathcal{A}_2(P_n; 2\varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; \tilde{P}),$$

since $W_2(\tilde{P}, P) \leq W_2(\tilde{P}, P_n) + W_2(P_n, P) \leq 2\varepsilon$ by the triangle inequality, implying $\mathcal{A}_2(P_n; \varepsilon) \subseteq \mathcal{A}_2(P_n; 2\varepsilon)$.

Now, we may use Lemma 2 of Gao (2023), which states that whenever $\boldsymbol{z} \mapsto \ell(q_\phi; \boldsymbol{z})$ has an $h-$Lipschitz gradient, $\forall \varepsilon \geq 0$, for any distribution Q on $\mathcal{Z}$,

$$\left| \sup_{\tilde{P} \in \mathcal{A}_2(P_n; \varepsilon)} \mathbb{E}_{\boldsymbol{z} \sim \tilde{P}}[\ell(q_\phi; \boldsymbol{z})] - \mathbb{E}_{\boldsymbol{z} \sim Q}[\ell(q_\phi; \boldsymbol{z})] - \varepsilon\, \mathbb{E}_{\boldsymbol{z} \sim Q}\left[\|\nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z})\|^2 \right]^{\frac{1}{2}} \right| \leq \varepsilon^2 h.$$

We use this result with $\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}) := -\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$. Then, with probability at least $1 - \delta$,

$$\sup_{\tilde{P} \in \mathcal{A}_2(P_n; 2\varepsilon)} \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; \tilde{P}) - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) \leq 2\varepsilon\, \mathbb{E}_{\boldsymbol{z} \sim P}\left[\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z})\|^2 \right]^{\frac{1}{2}} + 4\varepsilon^2 h.$$

We then obtain the *DRO excess risk* as follows:

$$\mathcal{L}_{\mathrm{NPE}}(q_{\phi_n^{\mathrm{DRO}}}; P) - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) \leq 2\varepsilon\, \mathbb{E}_{\boldsymbol{z}\sim P}\left[\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z})\|^2\right]^{\frac{1}{2}} + 4\varepsilon^2 h. \tag{8}$$

Now, recall that

$$\mathrm{kl}_{\mathrm{NPE}}(q_\phi; P) := \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\mathrm{KL}\Big(P_{\boldsymbol{\theta}|\boldsymbol{x}} \parallel Q_{\boldsymbol{\theta}|\boldsymbol{x}}\Big)\right] \equiv \mathcal{L}_{\mathrm{NPE}}(q_\phi; P) + K(P).$$

By Lemma A.4, for any $\alpha \in (0,1)$ and q_ϕ, $G^\alpha_{\mathrm{cov}}(q_\phi; P) \geq 0$. Similarly, for any choice of S and q_ϕ, $G^S_{\mathrm{cal}}(q_\phi; P) \geq 0$. Therefore,

$$\begin{aligned}
&\mathrm{kl}^\alpha_{\mathrm{cov}}(q_{\phi_n^{\mathrm{DRO}}}; P) - \mathrm{kl}^\alpha_{\mathrm{cov}}(q_{\phi^\star}; P) \\
&\leq \mathrm{kl}_{\mathrm{NPE}}(q_{\phi_n^{\mathrm{DRO}}}; P) - \mathrm{kl}^\alpha_{\mathrm{cov}}(q_{\phi^\star}; P) \\
&= \mathcal{L}_{\mathrm{NPE}}(q_{\phi_n^{\mathrm{DRO}}}; P) - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) + \mathrm{kl}_{\mathrm{NPE}}(q_{\phi^\star}; P) - \mathrm{kl}^\alpha_{\mathrm{cov}}(q_{\phi^\star}; P) \\
&\leq 2\varepsilon\, \mathbb{E}_{\boldsymbol{z}\sim P}\left[\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z})\|^2\right]^{\frac{1}{2}} + 4\varepsilon^2 h + G^\alpha_{\mathrm{cov}}(q_{\phi^\star}; P),
\end{aligned}$$

and similarly for calibration,

$$\mathrm{kl}^S_{\mathrm{cal}}(q_{\phi_n^{\mathrm{DRO}}}; P) - \mathrm{kl}^S_{\mathrm{cal}}(q_{\phi^\star}; P) \leq 2\varepsilon\, \mathbb{E}_{\boldsymbol{z}\sim P}\left[\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z})\|^2\right]^{\frac{1}{2}} + 4\varepsilon^2 h + G^S_{\mathrm{cal}}(q_{\phi^\star}; P).$$

□

Next, in Lemma B.4, we derive an excess risk bound under additional assumptions for the NPE empirical risk minimiser as well as for the regularised, first-order DRO-NPE objective, so that they can be compared. The proof pattern for Lemma B.4 is almost the same for NPE and DRO-NPE, both using the generalisation bound in Corollary 3 of Gao (2023) and sharing the same remainder terms and sample-dependent DRO radius. Note that although all of the below results are in terms of $\ell_{\mathrm{NPE}}, \mathcal{L}_{\mathrm{NPE}}$, similar proof patterns can be used for more general risks and pointwise losses $\mathcal{L}, \ell$.

Before proving Lemma B.4, we first define the following NPE and DRO-NPE minimisers, and establish the additional Assumption B.2 needed for Lemma B.4.

$$\hat{\phi}_n := \arg\min_\phi \mathcal{L}_{\mathrm{NPE}}(q_\phi; P_n)$$

$$\phi_n^{\mathrm{DRO-NPE}} := \arg\min_\phi \ \mathcal{L}_{\mathrm{NPE}}(q_\phi; P_n) + \varepsilon \left(\frac{1}{n}\sum_{i=1}^n \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}_i)\|^2\right)^{\frac{1}{2}}.$$

We now state the additional assumptions needed for Lemma B.4:

Assumption B.2 (Assumptions of Corollary 3 Gao (2023))**.** *Suppose that:*

(i) *P satisfies the transportation-information inequality $T_2(\tau)$ i.e. there exists $\tau > 0$ such that*

$$W_2(P, Q) \leq \sqrt{\tau \mathrm{KL}(Q||P)} \quad \forall Q \in \mathcal{P}_2.$$

(ii) *There exists measurable functions $\kappa : \mathcal{Z} \to \mathbb{R}_+$ and $\kappa_2 : \mathcal{Z} \to \mathbb{R}_+$ and constants $\kappa_M, \kappa_L, \kappa_{2,M}, \kappa_{2,L} \geq 0$ satisfying $\kappa(\boldsymbol{z}) \leq \kappa_M + \kappa_L\|\boldsymbol{z}\|^2$ and $\kappa_2(\boldsymbol{z}) \leq \kappa_{2,M} + \kappa_{2,L}\|\boldsymbol{z}\|^2$ for all $\boldsymbol{z} \in \mathcal{Z}$, such that:*

$$|\ell(q_{\tilde{\phi}}; \boldsymbol{z}) - \ell(q_\phi; \boldsymbol{z})| \leq \kappa(\boldsymbol{z})\|\tilde{\phi} - \phi\| \quad \text{for all } \phi, \tilde{\phi} \in \Phi, \quad P - a.e. \quad \boldsymbol{z} \in \mathcal{Z}$$
$$|\nabla_{\boldsymbol{z}} \ell(q_{\tilde{\phi}}; \boldsymbol{z}) - \nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z})| \leq \kappa_2(\boldsymbol{z})\|\tilde{\phi} - \phi\| \quad \text{for all } \phi, \tilde{\phi} \in \Phi, \quad P - a.e. \quad \boldsymbol{z} \in \mathcal{Z}$$

(iii) *Let $N(\varepsilon; \Phi, \|\cdot\|)$ denote the covering number of set Φ equipped with norm $\|\cdot\|$, defined as the smallest cardinality of an ε-cover of Φ, where Φ_ε is an ε-cover of Φ if for each $\phi \in \Phi$, there exists $\tilde{\phi} \in \Phi_\varepsilon$ such that $\|\tilde{\phi} - \phi\| \leq \varepsilon$. We assume that the parameter space admits a finite covering number at scale $\frac{1}{n}$, i.e. $\log N(\frac{1}{n}; \Phi, \|\cdot\|) < \infty$.*

(iv) *Assume $\sigma = \sup_{\phi\in\Phi} \frac{\left(\mathbb{E}_{\boldsymbol{z}\sim P}\left[\|\nabla_{\boldsymbol{z}}\, \ell(q_\phi; \boldsymbol{z})\|_2^4\right]\right)^{\frac{1}{2}}}{\mathbb{E}_{\boldsymbol{z}\sim P}\left[\|\nabla_{\boldsymbol{z}}\, \ell(q_\phi; \boldsymbol{z})\|_2^2\right]} < \infty$.*

As noted by Gao (2023), Assumption B.2(i) is implied by a log-Sobolev inequality. This condition holds, for example, for Gaussian and strongly log-concave distributions, and more generally for many light-tailed distributions with Gaussian-type concentration. Assumption B.2(ii) is a Lipschitz continuity assumption in ϕ for the loss and its gradient, with a Lipschitz constant allowed to grow at most quadratically in z. Thus, changes in the parameter ϕ induce controlled changes in the loss and its gradient, up to a quadratic envelope in the data variable. For example, by the mean value theorem, (ii) holds whenever Φ is compact and convex and $\|\nabla_\phi \ell(q_\phi; z)\|$ and $\|\nabla_\phi \nabla_z \ell(q_\phi; z)\|$ grow at most quadratically in z uniformly over Φ; it can fail when Φ is unbounded, since the loss may then become arbitrarily sensitive to ϕ. Finally, Assumption B.2(iii) controls the complexity of the parameter space, and hence, together with Assumption B.2(ii), the induced loss class $\{\ell(q_\phi; \cdot) : \phi \in \Phi\}$. Since $\Phi \subseteq \mathbb{R}^{d_\Phi}$, the assumption is satisfied, for example, whenever Φ is bounded. Intuitively, (iii) requires that finitely many points suffice to place every $\phi \in \Phi$ within distance $1/n$ of one of them; for bounded $\Phi \subset \mathbb{R}^{d_\Phi}$, the logarithm of the number of points needed grows only as $d_\Phi \log n$. Finally, Assumption B.2(iv) is a uniform moment condition on the input-gradient norm, and is satisfied when $\|\nabla_{\boldsymbol{z}} \ell(q_\phi; \boldsymbol{z})\|_2$ has uniformly controlled fourth moment relative to its second moment under P, over $\phi \in \Phi$.

Having now stated the assumptions, before presenting Lemma B.4, we define a few quantities that will appear in the Lemma. Let $t > 0$ and $n > 8\sigma^2 t$. Define the sample-dependent DRO radius ε_n, the moment-based concentration remainder term ρ_n, and the covering-number, function class complexity penalty term ϱ_n as:

$$\varepsilon_n := \sqrt{\frac{\tau\big(t + \log\big(1 + \mathcal{N}(1/n; \Phi, \|\cdot\|)\big)\big)}{n}} \left(1 + \sigma\sqrt{\frac{2\big(t + \log\big(1 + \mathcal{N}(1/n; \Phi, \|\cdot\|)\big)\big)}{n}}\right),$$

$$\rho_n := \frac{2\mathbb{E}_{\boldsymbol{z}\sim P}[\kappa_2] + \sqrt{\mathbb{V}_{\boldsymbol{z}\sim P}[\kappa_2]} + \varepsilon_n\sqrt{\mathbb{E}_{\boldsymbol{z}\sim P}[\kappa_2^2] + \sqrt{\mathbb{V}_{\boldsymbol{z}\sim P}[\kappa_2^2]}}}{n}$$

$$\varrho_n := \frac{\tau(t + \log(1 + \mathcal{N}(\frac{1}{n}; \Phi, \|\cdot\|)))}{n}$$

Lemma B.4 (NPE vs DRO-NPE excess risk)**.** *Assume that Assumption B.1 and Assumption B.2 hold with ℓ_{NPE} and $\varepsilon_n, \rho_n, \varrho_n$ are defined as in the above. Then, with probability at least $1 - 2/n - 2e^{-t}$,*

$$\begin{aligned}
&\mathcal{L}_{\mathrm{NPE}}(q_{\hat{\phi}_n}; P) - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) \\
&\leq \varepsilon_n \left(\frac{1}{n}\sum_{i=1}^n \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z}_i)\|^2\right)^{\frac{1}{2}} + \left(\frac{1}{n}\sum_{i=1}^n \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\hat{\phi}_n}; \boldsymbol{z}_i)\|^2\right)^{\frac{1}{2}} + 2\rho_n + 2h\varrho_n.
\end{aligned}$$

Moreover, under the same assumptions:

$$\mathcal{L}_{\mathrm{NPE}}(q_{\phi_n^{\mathrm{DRO-NPE}}}; P) - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) \leq 2\varepsilon_n \left(\frac{1}{n}\sum_{i=1}^n \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z}_i)\|^2\right)^{\frac{1}{2}} + 2\rho_n + 2h\varrho_n$$

Proof. It follows from Corollary 3 of Gao (2023) that with probability at least $1 - \frac{2}{n} - 2e^{-t}$:

$$\mathbb{E}_{\boldsymbol{z}\sim P}[\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})] \leq \frac{1}{n}\sum_{i=1}^n \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}_i) + \varepsilon_n \left(\frac{1}{n}\sum_{i=1}^n \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}_i)\|^2\right)^{\frac{1}{2}} + \rho_n + h\varrho_n, \ \forall \phi, \tag{9}$$

Note that the same result may be applied to the function $-\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})$; this will be used below. Importantly, since the bound holds uniformly over ϕ, we may use it to handle $\hat{\phi}_n$.

Hence, applying the bound for ℓ_{NPE} and $\hat{\phi}_n$,

$$
\begin{aligned}
&\mathcal{L}_{\mathrm{NPE}}(q_{\hat{\phi}_n}; P) - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) \\
&\leq \mathcal{L}_{\mathrm{NPE}}(q_{\hat{\phi}_n}; P_n) + \varepsilon_n \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\hat{\phi}_n}; \boldsymbol{z}_i)\|^2 \right)^{\frac{1}{2}} + \rho_n + h\varrho_n - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) \\
&\leq \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P_n) + \varepsilon_n \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\hat{\phi}_n}; \boldsymbol{z}_i)\|^2 \right)^{\frac{1}{2}} + \rho_n + h\varrho_n - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) \\
&\leq \varepsilon_n \left(\left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z}_i)\|^2 \right)^{\frac{1}{2}} + \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\hat{\phi}_n}; \boldsymbol{z}_i)\|^2 \right)^{\frac{1}{2}} \right) + 2\rho_n + 2h\varrho_n,
\end{aligned}
$$

where in the first inequality we use the corollary's bound for $\phi = \hat{\phi}_n$; in the second inequality we use the fact that $\hat{\phi}_n$ minimises $\mathcal{L}_{\mathrm{NPE}}(q_\phi; P_n)$; and in the third inequality we use Equation (9), applied to the pointwise $-\ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z})$. Similarly,

$$
\begin{aligned}
&\mathcal{L}_{\mathrm{NPE}}(q_{\phi_n^{\mathrm{DRO-NPE}}}; P) - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) \\
&\leq \mathcal{L}_{\mathrm{NPE}}(q_{\phi_n^{\mathrm{DRO-NPE}}}; P_n) + \varepsilon_n \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\phi^{\mathrm{DRO-NPE}}}; \boldsymbol{z}_i)\|^2 \right)^{\frac{1}{2}} + \rho_n + h\varrho_n - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) \\
&\leq \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P_n) + \varepsilon_n \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z}_i)\|^2 \right)^{\frac{1}{2}} + \rho_n + h\varrho_n - \mathcal{L}_{\mathrm{NPE}}(q_{\phi^\star}; P) \\
&\leq 2\varepsilon_n \mathbb{E}_{\boldsymbol{z} \sim P_n} [\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_{\phi^\star}; \boldsymbol{z})\|^2]^{\frac{1}{2}} + 2\rho_n + 2h\varrho_n
\end{aligned}
$$

where we follow the same steps as before, except for the second inequality, where we use that $\phi_n^{\mathrm{DRO-NPE}}$ minimises $\mathcal{L}_{\mathrm{NPE}}(q_\phi; P_n) + \varepsilon_n \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}_i)\|^2 \right)^{\frac{1}{2}}$.

□

Lemma B.4 derives an excess risk upper bound for NPE and for the regularised objective of DRO-NPE. The two bounds are derived in similar ways and differ only in the regulariser term $\Omega(\phi; P_n)$, which can be thought of as an empirical measure of functional variation via the loss gradient. Indeed, the key distinction between the DRO and NPE bounds derived is that the DRO-NPE bound only depends on the oracle variation $\Omega(\phi^\star; P_n)$. In contrast, the NPE bound depends on both the oracle variation and the variation of the learned empirical minimiser, $\Omega(\hat{\phi}_n; P_n)$. This means that the NPE bound can be affected by the roughness of the empirical minimiser. Thus, the DRO guarantee is sharper whenever the empirical minimiser has larger variation than the oracle.

B.4 Verifying conditions for normalising flows (proof of Lemma 4.3)

In this section, we formally introduce MAFs and coupling flows and prove Lemma 4.3, showing that all assumptions of Proposition 4.2 are satisfied when q_ϕ is chosen to be one of these commonly used normalising flows.

Masked autoregressive flows. We first write the masked autoregressive flows (MAFs; Papamakarios et al. (2017)) for $q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$, implemented in the *sbi* library (Tejero-Cantero et al., 2020), defined through a single autoregressive transform $u = u(\boldsymbol{\theta}, \boldsymbol{x}) \in \mathcal{R}^{d_\Theta}$ with coordinates

$$
u_i(\boldsymbol{\theta}, \boldsymbol{x}) := \frac{\theta_i - \mu_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x})}{\sigma_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x})}, \qquad \boldsymbol{\theta}_{<i} := (\theta_1, \ldots, \theta_{i-1}).
$$

We set

$$\begin{aligned}
\mu_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) &= v_i^\top \mathbf{g}_i^{(L)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) + c_i, \\
\mathbf{g}_i^{(\ell)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) &= \tanh\big(B_i^{(\ell)} \mathbf{g}_i^{(\ell-1)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) + d_i^{(\ell)}\big), \quad \ell = 1, \dots, L, \\
\mathbf{g}_i^{(0)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) &= [\boldsymbol{\theta}_{<i}; \boldsymbol{x}],
\end{aligned}$$

$$\begin{aligned}
\sigma_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) &= \mathrm{softplus}\big(s_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x})\big), \qquad \mathrm{softplus}(t) := \log(1 + e^t), \\
s_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) &= w_i^\top \mathbf{h}_i^{(L)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) + b_i, \\
\mathbf{h}_i^{(\ell)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) &= \tanh\big(A_i^{(\ell)} \mathbf{h}_i^{(\ell-1)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) + a_i^{(\ell)}\big), \qquad \ell = 1, \dots, L. \\
\mathbf{h}_i^{(0)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) &= [\boldsymbol{\theta}_{<i}; \boldsymbol{x}],
\end{aligned}$$

for some depth L, weight matrices $\{B_i^{(\ell)}\}_{\ell=1}^L$ and $\{A_i^{(\ell)}\}_{\ell=1}^L$, biases $\{d_i^{(\ell)}\}_{\ell=1}^L$ and $\{a_i^{(\ell)}\}_{\ell=1}^L$, and final weights (v_i, c_i) and (w_i, b_i); all of these parameters are part of ϕ. Here $[\boldsymbol{\theta}_{<i}; \boldsymbol{x}]$ denotes vector concatenation. Finally, we let the base density be standard Gaussian, i.e.

$$p_0(u) = (2\pi)^{-d_\Theta/2} \exp\left(-\frac{1}{2}\|u\|^2\right).$$

Hence,

$$\begin{aligned}
\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}) &= -\frac{1}{2}\|u(\boldsymbol{\theta}, \boldsymbol{x})\|^2 - \sum_{i=1}^{d_\Theta} \log \sigma_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) - \frac{d_\Theta}{2} \log 2\pi \\
&= -\frac{1}{2} \sum_{i=1}^{d_\Theta} \left(\frac{\theta_i - \mu_i}{\sigma_i}\right)^2 - \sum_{i=1}^{d_\Theta} \log \sigma_i - \frac{d_\Theta}{2} \log 2\pi.
\end{aligned}$$

For simplicity, in what follows, we will use the notation $\mu_i := \mu_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x})$, $\sigma_i := \sigma_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x})$, $\mathbf{g}_i^{(\ell)} := \mathbf{g}_i^{(\ell)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x})$, $\mathbf{h}_i^{(\ell)} := \mathbf{h}_i^{(\ell)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x})$, and $s_i := s_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x})$.

Lemma B.5. *Let $q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$ be a masked autoregressive flow of the form described above, with tanh hidden layers, softplus scale function, and standard Gaussian base density. Let $\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}) := -\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$, $\boldsymbol{z} = (\boldsymbol{\theta}, \boldsymbol{x}) \in \mathbb{R}^{d_{\boldsymbol{z}}}$. Then the map $\boldsymbol{z} \mapsto \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})$ is twice continuously differentiable on $\mathbb{R}^{d_{\boldsymbol{z}}}$, and there exists a constant $C > 0$, depending on ϕ, such that for all $\boldsymbol{z} \in \mathbb{R}^{d_{\boldsymbol{z}}}$,*

$$|\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})| \le C(1 + \|\boldsymbol{z}\|^2).$$

Moreover, there exist constants $C_1, C_2 > 0$, depending on ϕ, such that for all $\boldsymbol{z} \in \mathbb{R}^{d_{\boldsymbol{z}}}$

$$\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})\| \le C_1(1 + \|\boldsymbol{z}\|^2), \qquad \|\nabla_{\boldsymbol{z}}^2 \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})\| \le C_2(1 + \|\boldsymbol{z}\|^2).$$

Proof. We first show that the loss ℓ_{NPE} has at most a quadratic growth rate. Recall that

$$\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}) = -\frac{1}{2} \sum_{i=1}^{d_\Theta} \left(\frac{\theta_i - \mu_i}{\sigma_i}\right)^2 - \sum_{i=1}^{d_\Theta} \log \sigma_i - \frac{d_\Theta}{2} \log 2\pi. \tag{10}$$

Since the hidden activations are tanh, all hidden-layer outputs are uniformly bounded. Hence, for each i,

1. there exists $C_{\mu,i} < \infty$ such that $|\mu_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x})| \le C_{\mu,i}$ for all $\boldsymbol{z}$;
2. $s_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x})$ is uniformly bounded, and therefore $0 < \sigma_{i,\min} \le \sigma_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) \le \sigma_{i,\max} < \infty$ for all $\boldsymbol{z}$.

Therefore,

$$\left(\frac{\theta_i - \mu_i}{\sigma_i}\right)^2 \leq \left(\frac{\theta_i - \mu_i}{\sigma_{i,\min}}\right)^2 \leq \frac{2\theta_i^2 + 2\mu_i^2}{\sigma_{i,\min}^2} \leq \frac{2\theta_i^2 + 2C_{\mu,i}^2}{\sigma_{i,\min}^2}.$$

Substituting this bound and using $|\log \sigma_i| \leq \max\{|\log \sigma_{i,\min}|, |\log \sigma_{i,\max}|\} := C_{\log\sigma}$, we get that

$$|\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})| \leq \frac{1}{2}\sum_{i=1}^{d_\Theta}\left(\frac{2\theta_i^2 + 2C_{\mu,i}^2}{\sigma_{i,\min}^2} + C_{\log\sigma}\right) + \frac{d_\Theta}{2}\log 2\pi.$$

Hence, there exists constants $A_\phi, B_\phi > 0$ such that

$$|\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})| \leq A_\phi \|\boldsymbol{\theta}\|^2 + B_\phi.$$

Since $\|\boldsymbol{\theta}\| \leq \|(\boldsymbol{\theta}, \boldsymbol{x})\| = \|\boldsymbol{z}\|$, we obtain

$$|\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})| \leq A_\phi \|\boldsymbol{z}\|^2 + B_\phi,$$

showing that the MAF with log-loss has a finite $2-$growth rate.

Next, we show that the map $\boldsymbol{z} \to \ell_{\mathrm{NPE}}$ is twice continuously differentiable on $\mathbb{R}^{d_{\mathcal{Z}}}$. Indeed, since each layer of the networks defining μ_i and s_i consists of affine transformations followed by tanh, it follows that $\mu_i, s_i \in C^\infty(\mathbb{R}^{d_{\mathcal{Z}}})$. Since $\sigma_i = \mathrm{softplus}(s_i)$ and softplus is C^∞ with strictly positive output, we have $\sigma_i \in C^\infty(\mathbb{R}^{d_{\mathcal{Z}}})$ and $\sigma_i(\boldsymbol{z}) > 0$ for all $\boldsymbol{z} \in \mathbb{R}^{d_{\mathcal{Z}}}$. Therefore, the map $\boldsymbol{z} \to \ell_{\mathrm{NPE}}$ is a composition of C^∞ functions and is thus C^∞ on $\mathbb{R}^{d_{\mathcal{Z}}}$.

We now verify the condition on the growth of gradients. Note that from the proof of Proposition 3 (see Equation 38) in Bharti et al. (2026), it follows by the same arguments after swapping $\boldsymbol{\theta}$ and $\boldsymbol{x}$ (since this proposition refers to the NLE case in which the MAF has an identical form after swapping $\boldsymbol{\theta}$ and $\boldsymbol{x}$ everywhere) we have that there exists $C_0 < \infty$ such that:

$$\|\nabla_{\boldsymbol{\theta}} \log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})\| \leq C_0(1 + \|\boldsymbol{\theta}\|^2).$$

We now verify that a similar argument holds for $\|\nabla_{\boldsymbol{x}} \log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})\|$. In particular, for each $j = 1, \ldots, d_{\mathcal{X}}$:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{x}_j}\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}) &= -\frac{1}{2}\sum_{i=1}^{d_\Theta}\frac{\partial}{\partial \boldsymbol{x}_j}\left(\frac{\theta_i - \mu_i}{\sigma_i}\right)^2 - \sum_{i=1}^{d_\Theta}\frac{\partial}{\partial \boldsymbol{x}_j}\log\sigma_i \\ &= \sum_{i=1}^{d_\Theta}\frac{\theta_i - \mu_i}{\sigma_i^2}\frac{\partial \mu_i}{\partial \boldsymbol{x}_j} + \sum_{i=1}^{d_\Theta}\frac{(\theta_i - \mu_i)^2}{\sigma_i^3}\frac{\partial \sigma_i}{\partial \boldsymbol{x}_j} - \sum_{i=1}^{d_\Theta}\frac{1}{\sigma_i}\frac{\partial \sigma_i}{\partial \boldsymbol{x}_j}.\end{aligned}$$

Following Equation (32) of Bharti et al. (2026) we have that since outputs of μ_i and s_i are affine functions of $\mathbf{h}_i^L$, there exist $C_1, C_2 < \infty$ such that:

$$|\mu_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x})| \leq C_1, \quad \left\|\frac{\partial \mu_i}{\partial \boldsymbol{x}_j}\right\| \leq C_2.$$

We now turn to the derivative of σ_i, recalling that

$$\begin{aligned}&\sigma_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) = \mathrm{softplus}\big(s_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x})\big), && s_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) = w_i^\top h_i^{(L)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) + b_i, \\ &\mathrm{softplus}'(t) = \frac{1}{1+\exp(-t)}, && \mathrm{softplus}''(t) = \frac{\exp(-t)}{(1+\exp(-t))^2},\end{aligned}$$

where, for each $\ell \in \{1, \ldots, L\}$, $h_i^{(\ell)} = \tanh\left(A_i^{(\ell)} h_i^{(\ell-1)} + a_i^{(\ell)}\right)$. Hence,

$$\frac{\partial \sigma_i}{\partial \boldsymbol{x}_j} = \frac{1}{1+\exp(-s_i)}\frac{\partial s_i}{\partial \boldsymbol{x}_j} = \frac{1}{1+\exp(-s_i)}\, w_i^\top \frac{\partial h_i^{(L)}}{\partial \boldsymbol{x}_j}.$$

Moreover, for each hidden layer,

$$\frac{\partial h_i^{(\ell)}}{\partial \boldsymbol{x}_j} = \mathrm{sech}^2\big(A_i^{(\ell)} h_i^{(\ell-1)} + a_i^{(\ell)}\big) \odot \left(A_i^{(\ell)}\frac{\partial h_i^{(\ell-1)}}{\partial \boldsymbol{x}_j}\right).$$

where $\odot$ denotes element-wise multiplication of vectors. Since i) sech^2 is uniformly bounded, ii) the network has finitely many layers with fixed finite weight matrices, and iii) $\partial h_i^{(0)}/\partial \boldsymbol{x}_j = e_j$, it follows that $\partial h_i^{(\ell)}/\partial \boldsymbol{x}_j$ is uniformly bounded on $\mathcal{X}$ for every ℓ, i, j. Therefore, both $\partial s_i/\partial \boldsymbol{x}_j$ and $\partial \sigma_i/\partial \boldsymbol{x}_j$ are uniformly bounded on $\mathbb{R}^{d_\mathcal{X}}$ and there exists $C_3 < \infty$ such that:

$$\left\| \frac{\partial \sigma_i}{\partial \boldsymbol{x}_j} \right\| \leq C_3.$$

Therefore,

$$\left| \frac{\partial}{\partial \boldsymbol{x}_j} \log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}) \right| \leq \frac{1}{\sigma_{i,\min}^3} \sum_{i=1}^{d_\Theta} (C_2 |\theta_i - \mu_i| + |\theta_i - \mu_i|^2 C_3 + C_3)$$

and since $|\mu_i| \leq C_1$ it follows that there exist $C_4, C_5 < \infty$ such that: $|\theta_i - \mu_i| \leq C_4(1 + |\theta_i|)$ and $|\theta_i - \mu_i|^2 \leq C_5(1 + |\theta_i|^2)$. Thus, there exists $C_6 < \infty$ such that

$$\left| \frac{\partial}{\partial \boldsymbol{x}_j} \log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}) \right| \leq \frac{1}{\sigma_{i,\min}^3} \sum_{i=1}^{d_\Theta} (C_2 C_4 (1 + |\theta_i| + C_3 C_5 (1 + |\theta_i|^2) + C_3) \leq C_6 (1 + \|\boldsymbol{\theta}\|^2).$$

Therefore, we have:

$$\|\nabla_{\boldsymbol{\theta}} \log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})\| \leq C_0(1 + \|\boldsymbol{\theta}\|^2), \quad \|\nabla_{\boldsymbol{x}} \log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})\| \leq C_6 \sqrt{d_\mathcal{X}} (1 + \|\boldsymbol{\theta}\|^2)$$

and hence:

$$\begin{aligned}
\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})\| = \|\nabla_{\boldsymbol{z}} (-\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}))\| &\leq \|\nabla_{\boldsymbol{\theta}} \log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})\| + \|\nabla_{\boldsymbol{x}} \log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})\| \\
&\leq C_0(1 + \|\boldsymbol{\theta}\|^2) + C_6 \sqrt{d_\mathcal{X}} (1 + \|\boldsymbol{\theta}\|^2) \\
&:= C_7 (1 + \|\boldsymbol{\theta}\|^2) \\
&\leq C_7 (1 + \|\boldsymbol{z}\|^2).
\end{aligned}$$

Lastly, we verify the growth condition for the norm of the Hessian.

Recall that

$$-\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{1}{2} \sum_{i=1}^{d_\Theta} u_i^2 + \sum_{i=1}^{d_\Theta} \log \sigma_i + \frac{d_\Theta}{2} \log 2\pi,$$

where $u_i = \frac{\theta_i - \mu_i}{\sigma_i}$. Therefore, for any $j, k \in \{1, \ldots, d_\mathcal{Z}\}$,

$$\partial_{z_j} (-\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})) = \sum_{i=1}^{d_\Theta} \left(u_i \, \partial_{z_j} u_i + \frac{1}{\sigma_i} \partial_{z_j} \sigma_i \right), \tag{11}$$

$$\partial_{z_k} \partial_{z_j} (-\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})) = \sum_{i=1}^{d_\Theta} \left(u_i \, \partial_{z_k} \partial_{z_j} u_i + (\partial_{z_k} u_i)(\partial_{z_j} u_i) - \frac{1}{\sigma_i^2} (\partial_{z_k} \sigma_i)(\partial_{z_j} \sigma_i) + \frac{1}{\sigma_i} \partial_{z_k} \partial_{z_j} \sigma_i \right). \tag{12}$$

By points 1. and 2. above we have, $|u_i| = \left| \frac{\theta_i - \mu_i}{\sigma_i} \right| \leq \frac{|\theta_i| + C_{\mu,i}}{\sigma_{i,\min}} \leq C_u(1 + |\theta_i|)$ for $C_u := \max_i(\max\{1/\sigma_{i,\min}, C_{\mu,i}/\sigma_{i,\min}\}) < \infty$. Moreover, since we have shown that both $|\partial_{z_j} \mu_i| \leq C_{\mu'}$ and $|\partial_{z_j} \sigma_i| \leq C_{\sigma'}$ where $C_{\mu'}$ and $C_{\sigma'}$ are constants independent of $\boldsymbol{z}$ and

$$\partial_{z_j} u_i = \frac{\partial_{z_j} (\theta_i - \mu_i)}{\sigma_i} - \frac{\theta_i - \mu_i}{\sigma_i^2} \partial_{z_j} \sigma_i,$$

it follows that $|\partial_{z_j} u_i| \leq C_{u'}(1 + |\theta_i|) \leq C_{u'}(1 + |\theta_i|)$ for some $C_{u'} < \infty$ depending on $\sigma_{i,\min}, C_{\mu'}$ and $C_{\sigma'}$.

We now consider the second derivative of σ_i:

$$\frac{\partial^2 \sigma_i}{\partial z_k \partial z_j} = \frac{\exp(-s_i)}{(1 + \exp(-s_i))^2} \frac{\partial s_i}{\partial z_k} \frac{\partial s_i}{\partial z_j} + \frac{1}{1 + \exp(-s_i)} \frac{\partial^2 s_i}{\partial z_k \partial z_j}. \tag{13}$$

Thus it remains to bound $\partial^2 s_i/(\partial z_k \partial z_j)$. Since

$$\frac{\partial s_i}{\partial z_j} = w_i^\top \frac{\partial h_i^{(L)}}{\partial z_j}, \qquad \frac{\partial^2 s_i}{\partial z_k \partial z_j} = w_i^\top \frac{\partial^2 h_i^{(L)}}{\partial z_k \partial z_j},$$

it suffices to bound the second derivatives of the hidden states. Differentiating once more,

$$\begin{aligned}\frac{\partial^2 h_i^{(\ell)}}{\partial z_k \partial z_j} &= \frac{\partial}{\partial z_k}\left[\operatorname{sech}^2\big(A_i^{(\ell)} h_i^{(\ell-1)} + a_i^{(\ell)}\big) \odot \left(A_i^{(\ell)} \frac{\partial h_i^{(\ell-1)}}{\partial z_j}\right)\right] \\ &= \left[-2\operatorname{sech}^2\big(A_i^{(\ell)} h_i^{(\ell-1)} + a_i^{(\ell)}\big) \tanh\big(A_i^{(\ell)} h_i^{(\ell-1)} + a_i^{(\ell)}\big)\right] \odot \left(A_i^{(\ell)} \frac{\partial h_i^{(\ell-1)}}{\partial z_k}\right) \odot \left(A_i^{(\ell)} \frac{\partial h_i^{(\ell-1)}}{\partial z_j}\right) \\ &\quad + \operatorname{sech}^2\big(A_i^{(\ell)} h_i^{(\ell-1)} + a_i^{(\ell)}\big) \odot \left(A_i^{(\ell)} \frac{\partial^2 h_i^{(\ell-1)}}{\partial z_k \partial z_j}\right).\end{aligned}$$

Now, we have that: i)$\tanh$ and sech^2 are uniformly bounded, ii) the first derivatives $\partial h_i^{(\ell-1)}/\partial z_j$ have already been shown to be bounded, iii) the flow has finitely many layers with fixed finite weight matrices, and iv)

$$\frac{\partial^2 h_i^{(0)}}{\partial z_k \partial z_j} = 0.$$

Hence, or each layer,

$$\left|\frac{\partial^2 h_i^{(\ell)}}{\partial z_k \partial z_j}\right| < \infty \qquad \forall\, \ell, i, j, k,\ \boldsymbol{z} \in \mathbb{R}^{d_{\boldsymbol{Z}}}.$$

Consequently, $\partial^2 s_i/(\partial z_k \partial z_j)$ is uniformly bounded on $\mathbb{R}^{d_{\boldsymbol{Z}}}$. Returning to (13), and using that

$$0 < \frac{\exp(-s_i)}{(1+\exp(-s_i))^2} \leq \frac{1}{4}, \qquad 0 < \frac{1}{1+\exp(-s_i)} < 1,$$

we conclude that

$$\left|\frac{\partial^2 \sigma_i}{\partial z_k \partial z_j}\right| \leq C_{\sigma''}$$

for some finite constant $C_{\sigma''}$, uniformly over $\boldsymbol{z} \in \mathbb{R}^{d_{\boldsymbol{Z}}}$. We now treat μ_i. Recall that

$$\mu_i(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) = v_i^\top g_i^{(L)}(\boldsymbol{\theta}_{<i}, \boldsymbol{x}) + c_i,$$

where the hidden layers $g_i^{(\ell)}$ are again defined through finitely many linear maps composed with $\tanh$. Therefore, using the same argument as above, but without the final softplus layer, we get that

$$\left|\frac{\partial^2 \mu_i}{\partial z_k \partial z_j}\right| \leq C_{\mu''}$$

for some finite constant $C_{\mu''}$, uniformly over $\boldsymbol{z} \in \mathbb{R}^{d_{\boldsymbol{Z}}}$. Using the bounds on the first and second derivatives of μ_i and σ_i we first have that

$$\partial_{z_j} u_i = \frac{\partial_{z_j}(\theta_i - \mu_i)}{\sigma_i} - \frac{\theta_i - \mu_i}{\sigma_i^2} \partial_{z_j} \sigma_i,$$

and hence $|\partial_{z_j} u_i| \leq C_{u'}(1 + \|\boldsymbol{\theta}\|)$ for some positive constant $C_{u'} < \infty$. Differentiating once more,

$$\begin{aligned}\partial_{z_k}\partial_{z_j} u_i = &\frac{\partial_{z_k}\partial_{z_j}(\theta_i - \mu_i)}{\sigma_i} - \frac{\partial_{z_j}(\theta_i - \mu_i)}{\sigma_i^2}\partial_{z_k}\sigma_i - \frac{\partial_{z_k}(\theta_i - \mu_i)}{\sigma_i^2}\partial_{z_j}\sigma_i \\ &- \frac{\theta_i - \mu_i}{\sigma_i^2}\partial_{z_k}\partial_{z_j}\sigma_i + 2\frac{\theta_i - \mu_i}{\sigma_i^3}(\partial_{z_k}\sigma_i)(\partial_{z_j}\sigma_i),\end{aligned}$$

so that $|\partial_{z_k}\partial_{z_j}u_i| \leq C_{u''}(1+\|\boldsymbol{\theta}\|)$ for some positive constant $C_{u''} < \infty$. Returning to the expression in Equation (12), we have that for each of the four terms:

$$\begin{aligned}
\left|u_i\partial_{z_k}\partial_{z_j}u_i\right| &\leq C_u(1+\|\boldsymbol{\theta}\|)C_{u''}(1+\|\boldsymbol{\theta}\|) \leq C_8(1+\|\boldsymbol{\theta}\|^2)\\
\left|(\partial_{z_k}u_i)(\partial_{z_j}u_i)\right| &\leq C_{u'}^2(1+\|\boldsymbol{\theta}\|)^2 \leq C_9(1+\|\boldsymbol{\theta}\|^2)\\
\left|\frac{1}{\sigma_i^2}(\partial_{z_k}\sigma_i)(\partial_{z_j}\sigma_i)\right| &\leq \sigma_{i,\min}^{-2}C_{\sigma'}^2 := C_{10}\\
\left|\frac{1}{\sigma_i}\partial_{z_k}\partial_{z_j}\sigma_i\right| &\leq \sigma_{i,\min}^{-1}C_{\sigma''} := C_{11}
\end{aligned}$$

for some finite constant $C_8, C_9, C_{10}, C_{11} < \infty$. Therefore,

$$\begin{aligned}
\left|\partial_{z_k}\partial_{z_j}(-\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}))\right| &\leq d_\Theta\left(C_8(1+\|\boldsymbol{\theta}\|^2) + C_9(1+\|\boldsymbol{\theta}\|^2) + C_{10} + C_{11}\right)\\
&\leq C_{12}(1+\|\boldsymbol{\theta}\|^2)\\
&\leq C_{12}(1+\|\boldsymbol{z}\|^2)
\end{aligned}$$

for some positive constant $C_{12} < \infty$. To conclude note that, for $\|\cdot\|_F$ denoting the Frobenius norm, it follows that:

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{z}}^2\ell_{\mathrm{NPE}}(q_\phi;\boldsymbol{z})\right\| &\leq \left\|\nabla_{\boldsymbol{z}}^2\ell_{\mathrm{NPE}}(q_\phi;\boldsymbol{z})\right\|_F\\
&= \sqrt{\sum_{j=1}^{d_\mathcal{Z}}\sum_{k=1}^{d_\mathcal{Z}}\left|\partial_{z_k}\partial_{z_j}(-\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}))\right|^2}\\
&\leq \sqrt{d_\mathcal{Z}^2C_{12}^2(1+\|\boldsymbol{z}\|^2)^2}\\
&= d_\mathcal{Z}C_{12}(1+\|\boldsymbol{z}\|^2).
\end{aligned}$$

□

Coupling flows. We next describe coupling flows (Kingma and Dhariwal, 2018; Radev et al., 2022) implemented in BayesFlow package (Kühmichel et al., 2026) for $q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$. First, the Actnorm layer, which is an invertible adaptation of batch normalisation for stable training, is applied to $\boldsymbol{\theta}$. It is followed by a permutation, multiplying by the permutation matrix $P \in \{0,1\}^{d_\Theta \times d_\Theta}$: $\tilde{\boldsymbol{\theta}} = P(\alpha \odot \boldsymbol{\theta} + \beta)$ where $\alpha, \beta \in \mathbb{R}^{d_\Theta}$ are trainable parameters, and $\odot$ denotes element-wise multiplication.

We then split the dimension into two parts, followed by affine transformation: $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{\theta}}_\mathcal{A}; \tilde{\boldsymbol{\theta}}_\mathcal{B}]$ where $\tilde{\boldsymbol{\theta}}_\mathcal{A} = \tilde{\boldsymbol{\theta}}_{1:\lfloor d_\Theta/2\rfloor}$ is the first $\lfloor d_\Theta/2\rfloor$ elements and $\tilde{\boldsymbol{\theta}}_\mathcal{B} = \tilde{\boldsymbol{\theta}}_{\lfloor d_\Theta/2\rfloor+1:d_\Theta}$ is the last $d_\Theta - \lfloor d_\Theta/2\rfloor$ elements of $\tilde{\boldsymbol{\theta}}$. An affine coupling transformation $u = u(\tilde{\boldsymbol{\theta}}, \boldsymbol{x}) \in \mathbb{R}^{d_\Theta}$ is defined by

$$u(\tilde{\boldsymbol{\theta}}_\mathcal{A}, \boldsymbol{x}) := \sigma(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x}) \odot \tilde{\boldsymbol{\theta}}_\mathcal{A} + \mu(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x}). \qquad u(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x}) := \tilde{\boldsymbol{\theta}}_\mathcal{B}.$$

Note that the forward coupling transformation is defined differently than for MAFs.

The parameters $\mu \in \mathbb{R}^{\lfloor d_\Theta/2\rfloor}$ and $\sigma \in \mathbb{R}^{\lfloor d_\Theta/2\rfloor}$ are defined in a similar way as in *MAF* in sbi package. We use the tanh activation function instead of the default function implemented in BayesFlow (mish):

$$\begin{aligned}
\sigma(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x}) &= \mathrm{softplus}\big(\sinh^{-1}(\tilde{\sigma}(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x})) + \mathbf{1}\log(e-1)\big),\\
[\mu(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x}); \tilde{\sigma}(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x})] &= W^\top\mathbf{h}^{(L)}(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x}) + b,\\
\mathbf{h}^{(\ell)}(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x}) &= \tanh\big(A^{(\ell)}\mathbf{h}^{(\ell-1)}(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x}) + a^{(\ell)}\big), \qquad \ell = 1, \ldots, L.\\
\mathbf{h}^{(0)}(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x}) &= [\tilde{\boldsymbol{\theta}}_\mathcal{B}; \boldsymbol{x}],
\end{aligned}$$

where softplus, $\sinh^{-1}$, tanh are applied element-wise. The log-determinant of the Actnorm and the affine transformation is defined by

$$\log\left|\det\frac{\partial\tilde{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}}\right| = \sum_{i=1}^{d_\Theta}\log|\alpha_i|, \qquad \log\left|\det\frac{\partial u}{\partial\tilde{\boldsymbol{\theta}}}\right| = \sum_{i=1}^{\lfloor d_\Theta/2\rfloor}\log\sigma_i(\tilde{\boldsymbol{\theta}}_\mathcal{B}, \boldsymbol{x})$$

Finally, taking the base density to be standard Gaussian, given a single coupling block, we have

$$\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}) = -\frac{1}{2}\left[\sum_{i=1}^{\lfloor d_\Theta/2 \rfloor}\Big(\sigma_i(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}, \boldsymbol{x})\tilde{\boldsymbol{\theta}}_{\mathcal{A},i} + \mu_i(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}, \boldsymbol{x})\Big)^2 + \|\tilde{\boldsymbol{\theta}}_{\mathcal{B}}\|_2^2\right] + \sum_{i=1}^{d_\Theta}\log|\alpha_i| + \sum_{i=1}^{\lfloor d_\Theta/2 \rfloor}\log\sigma_i(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}, \boldsymbol{x}) - \frac{d_\Theta}{2}\log 2\pi.$$

In practice, the sequence of the Actnorm layer, the permutation, and the mapping from $\tilde{\boldsymbol{\theta}}$ to u forms a block, which is repeated multiple times to increase the flow's expressiveness.

Lemma B.6 (Single BayesFlow coupling block)**.** *Let $\boldsymbol{z} = (\boldsymbol{\theta}, \boldsymbol{x}) \in \mathbb{R}^{d_\Theta + d_\mathcal{X}}$, and consider the BayesFlow coupling block defined above by*

$$\tilde{\boldsymbol{\theta}} = P(\alpha \odot \boldsymbol{\theta} + \beta), \qquad u_{\mathcal{A}} = \sigma(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}, \boldsymbol{x}) \odot \tilde{\boldsymbol{\theta}}_{\mathcal{A}} + \mu(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}, \boldsymbol{x}), \qquad u_{\mathcal{B}} = \tilde{\boldsymbol{\theta}}_{\mathcal{B}},$$

where μ and σ are defined through the network architecture above. Then $\boldsymbol{z} \mapsto \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})$ belongs to $\mathcal{G}_2(\mathcal{Z})$ and is of class C^2. Moreover,

$$\|\nabla_{\boldsymbol{z}}\ell_{\mathrm{NPE}}(q_\phi; \cdot)\| \in \mathcal{G}_2(\mathcal{Z}), \qquad \|\nabla^2_{\boldsymbol{z}}\ell_{\mathrm{NPE}}(q_\phi; \cdot)\| \in \mathcal{G}_2(\mathcal{Z}).$$

Proof. By construction, the ActNorm parameters satisfy $\alpha_i \neq 0$ for all i. Moreover, since the hidden activations are $\tanh$ and the output layers are affine, the maps μ and $\tilde{\sigma}$ are bounded and belong to C^2, with uniformly bounded first and second derivatives. Since

$$\sigma = \mathrm{softplus}\big(\sinh^{-1}(\tilde{\sigma}) + \mathbf{1}\log(e-1)\big),$$

the same holds for σ. As $\tilde{\sigma}$ is bounded, there exist constants

$$0 < \sigma_{\min} \leq \sigma_i(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}, \boldsymbol{x}) \leq \sigma_{\max} < \infty$$

for all i and $(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}, \boldsymbol{x})$. Hence there exist constants

$$C_\mu, C_{\mu'}, C_{\mu''}, C_{\sigma'}, C_{\sigma''} < \infty$$

such that for all i, j, k,

$$|\mu_i| \leq C_\mu, \qquad |\partial_{z_j}\mu_i| \leq C_{\mu'}, \qquad |\partial_{z_k}\partial_{z_j}\mu_i| \leq C_{\mu''},$$
$$|\partial_{z_j}\sigma_i| \leq C_{\sigma'}, \qquad |\partial_{z_k}\partial_{z_j}\sigma_i| \leq C_{\sigma''}.$$

Since P is a permutation matrix, it preserves the Euclidean norm. Hence

$$\|\tilde{\boldsymbol{\theta}}\|_2 = \|P(\alpha \odot \boldsymbol{\theta} + \beta)\|_2 = \|\alpha \odot \boldsymbol{\theta} + \beta\|_2 \leq \|\mathrm{diag}(\alpha)\|\,\|\boldsymbol{\theta}\|_2 + \|\beta\|_2 \leq C_0(1 + \|\boldsymbol{\theta}\|_2)$$

for some $C_0 < \infty$. Also, $\tilde{\boldsymbol{\theta}}$ is affine in $\boldsymbol{\theta}$, so all first derivatives $\partial_{z_j}\tilde{\boldsymbol{\theta}}_i$ are bounded and all second derivatives $\partial_{z_k}\partial_{z_j}\tilde{\boldsymbol{\theta}}_i$ vanish.

We first control the loss. Since $u_{\mathcal{B}} = \tilde{\boldsymbol{\theta}}_{\mathcal{B}}$, we have

$$\|u_{\mathcal{B}}\|_2 \leq \|\tilde{\boldsymbol{\theta}}\|_2 \leq C_0(1 + \|\boldsymbol{\theta}\|_2).$$

For each $i \leq \lfloor d_\Theta/2 \rfloor$,

$$|u_{\mathcal{A},i}| = |\sigma_i(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}, \boldsymbol{x})\tilde{\boldsymbol{\theta}}_{\mathcal{A},i} + \mu_i(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}, \boldsymbol{x})| \leq \sigma_{\max}|\tilde{\boldsymbol{\theta}}_{\mathcal{A},i}| + C_\mu \leq C_1(1 + \|\boldsymbol{\theta}\|_2)$$

for some $C_1 < \infty$. Therefore,

$$\|u\|_2^2 \leq C_2(1 + \|\boldsymbol{\theta}\|_2^2)$$

for some $C_2 < \infty$. Since $\sum_{i=1}^{d_\Theta}\log|\alpha_i|$ is constant and each $\log\sigma_i(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}, \boldsymbol{x})$ is uniformly bounded, it follows from the expression for $\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$ that

$$|\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})| = |-\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})| \leq C_3(1 + \|\boldsymbol{\theta}\|_2^2) \leq C_3(1 + \|\boldsymbol{z}\|_2^2)$$

for some $C_3 < \infty$. Hence $\ell_{\mathrm{NPE}}(q_\phi; \cdot) \in \mathcal{G}_2(\mathcal{Z})$.

We next control the gradient. For $u_{\mathcal{B}} = \tilde{\boldsymbol{\theta}}_{\mathcal{B}}$, the first derivatives are bounded and the second derivatives vanish. For $u_{\mathcal{A}}$, for each i, j,

$$\partial_{z_j} u_{\mathcal{A},i} = (\partial_{z_j}\sigma_i)\tilde{\boldsymbol{\theta}}_{\mathcal{A},i} + \sigma_i\, \partial_{z_j}\tilde{\boldsymbol{\theta}}_{\mathcal{A},i} + \partial_{z_j}\mu_i.$$

Hence

$$|\partial_{z_j} u_{\mathcal{A},i}| \le C_{\sigma'} C_0(1 + \|\boldsymbol{\theta}\|_2) + \sigma_{\max} \sup_{i,j} |\partial_{z_j}\tilde{\boldsymbol{\theta}}_{\mathcal{A},i}| + C_{\mu'} \le C_4(1 + \|\boldsymbol{\theta}\|_2)$$

for some $C_4 < \infty$. Therefore,

$$\left|\partial_{z_j} \frac{1}{2}\|u\|_2^2\right| = \left|\sum_i u_i \partial_{z_j} u_i\right| \le C_5(1 + \|\boldsymbol{\theta}\|_2^2)$$

for some $C_5 < \infty$. Moreover, since $\sigma_i \ge \sigma_{\min}$,

$$\left|\partial_{z_j} \log \sigma_i\right| = \left|\frac{\partial_{z_j}\sigma_i}{\sigma_i}\right| \le \sigma_{\min}^{-1} C_{\sigma'}.$$

Thus

$$\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})\| \le C_6(1 + \|\boldsymbol{\theta}\|_2^2) \le C_6(1 + \|\boldsymbol{z}\|_2^2)$$

for some $C_6 < \infty$, and hence $\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \cdot)\| \in \mathcal{G}_2(\mathcal{Z})$.

Finally, for the Hessian, for each i, j, k,

$$\begin{aligned}\partial_{z_k}\partial_{z_j} u_{\mathcal{A},i} = (\partial_{z_k}\partial_{z_j}\sigma_i)\tilde{\boldsymbol{\theta}}_{\mathcal{A},i} + (\partial_{z_j}\sigma_i)(\partial_{z_k}\tilde{\boldsymbol{\theta}}_{\mathcal{A},i}) + (\partial_{z_k}\sigma_i)(\partial_{z_j}\tilde{\boldsymbol{\theta}}_{\mathcal{A},i}) \\ + \sigma_i\, \partial_{z_k}\partial_{z_j}\tilde{\boldsymbol{\theta}}_{\mathcal{A},i} + \partial_{z_k}\partial_{z_j}\mu_i.\end{aligned}$$

Since $\partial_{z_k}\partial_{z_j}\tilde{\boldsymbol{\theta}}_{\mathcal{A},i} = 0$, this yields

$$|\partial_{z_k}\partial_{z_j} u_{\mathcal{A},i}| \le C_7(1 + \|\boldsymbol{\theta}\|_2)$$

for some $C_7 < \infty$. Consequently,

$$\left|\partial_{z_k}\partial_{z_j} \frac{1}{2}\|u\|_2^2\right| = \left|\sum_i \Big[(\partial_{z_k} u_i)(\partial_{z_j} u_i) + u_i\, \partial_{z_k}\partial_{z_j} u_i\Big]\right| \le C_8(1 + \|\boldsymbol{\theta}\|_2^2)$$

for some $C_8 < \infty$. Also,

$$\partial_{z_k}\partial_{z_j} \log \sigma_i = \frac{\partial_{z_k}\partial_{z_j}\sigma_i}{\sigma_i} - \frac{(\partial_{z_k}\sigma_i)(\partial_{z_j}\sigma_i)}{\sigma_i^2},$$

so

$$|\partial_{z_k}\partial_{z_j} \log \sigma_i| \le \sigma_{\min}^{-1} C_{\sigma''} + \sigma_{\min}^{-2} C_{\sigma'}^2.$$

Thus

$$\|\nabla_{\boldsymbol{z}}^2 \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})\| \le C_9(1 + \|\boldsymbol{\theta}\|_2^2) \le C_9(1 + \|\boldsymbol{z}\|_2^2)$$

for some $C_9 < \infty$, and hence $\|\nabla_{\boldsymbol{z}}^2 \ell_{\mathrm{NPE}}(q_\phi; \cdot)\| \in \mathcal{G}_2(\mathcal{Z})$. Since all components in the construction are C^2, we also have $\ell_{\mathrm{NPE}}(q_\phi; \cdot) \in C^2(\mathbb{R}^{d_\Theta + d_\mathcal{X}})$. □

Lemma B.7 (Multiple BayesFlow coupling blocks)**.** *Fix $M \ge 1$, the number of blocks, and let*

$$\boldsymbol{z}^{(0)} = \boldsymbol{z} = (\boldsymbol{\theta}, \boldsymbol{x}).$$

For $\ell = 1, \dots, M$, write $\boldsymbol{z}^{(\ell)} = (\boldsymbol{\theta}^{(\ell)}, \boldsymbol{x})$, where $\boldsymbol{\theta}^{(\ell)}$ is obtained from $\boldsymbol{\theta}^{(\ell-1)}$ as follows:

$$\tilde{\boldsymbol{\theta}}^{(\ell)} = P^{(\ell)}\big(\alpha^{(\ell)} \odot \boldsymbol{\theta}^{(\ell-1)} + \beta^{(\ell)}\big),$$

split

$$\tilde{\boldsymbol{\theta}}^{(\ell)} = \big[\tilde{\boldsymbol{\theta}}_{\mathcal{A}}^{(\ell)}; \tilde{\boldsymbol{\theta}}_{\mathcal{B}}^{(\ell)}\big],$$

and define

$$\boldsymbol{\theta}_{\mathcal{A}}^{(\ell)} = \sigma^{(\ell)}\big(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}^{(\ell)}, \boldsymbol{x}\big) \odot \tilde{\boldsymbol{\theta}}_{\mathcal{A}}^{(\ell)} + \mu^{(\ell)}\big(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}^{(\ell)}, \boldsymbol{x}\big), \qquad \boldsymbol{\theta}_{\mathcal{B}}^{(\ell)} = \tilde{\boldsymbol{\theta}}_{\mathcal{B}}^{(\ell)},$$

where each $\mu^{(\ell)}$ and $\sigma^{(\ell)}$ is defined through the BayesFlow architecture above.

Then $\boldsymbol{z} \mapsto \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})$ belongs to $\mathcal{G}_2(\mathcal{Z})$ and is of class C^2. Moreover, with

$$\beta_M := M + 1, \qquad \alpha_M := 2M,$$

we have

$$\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \cdot)\| \in \mathcal{G}_{\beta_M}(\mathcal{Z}), \qquad \|\nabla_{\boldsymbol{z}}^2 \ell_{\mathrm{NPE}}(q_\phi; \cdot)\| \in \mathcal{G}_{\alpha_M}(\mathcal{Z}).$$

Proof. For each $\ell = 1, \dots, M$, let $F^{(\ell)} : \mathcal{Z} \to \mathcal{Z}$ denote the ℓ-th block map, defined by

$$\boldsymbol{z}^{(\ell)} = F^{(\ell)}(\boldsymbol{z}^{(\ell-1)}), \qquad \boldsymbol{z}^{(0)} = \boldsymbol{z} = (\boldsymbol{\theta}, \boldsymbol{x}).$$

By Lemma B.6, each $F^{(\ell)}$ belongs to C^2, and there exist constants $a_\ell, b_\ell, c_\ell < \infty$ such that for all $\boldsymbol{z} \in \mathcal{Z}$,

$$\begin{aligned}\|F^{(\ell)}(\boldsymbol{z})\|_2 &\le a_\ell(1 + \|\boldsymbol{z}\|_2),\\ \|\nabla_{\boldsymbol{z}} F^{(\ell)}(\boldsymbol{z})\| &\le b_\ell(1 + \|\boldsymbol{z}\|_2),\\ \|\nabla^2_{\boldsymbol{z}} F^{(\ell)}(\boldsymbol{z})\| &\le c_\ell(1 + \|\boldsymbol{z}\|_2).\end{aligned}$$

Since each $P^{(\ell)}$ is a permutation matrix, it preserves Euclidean norms, so it affects only constants and not growth orders.

We first bound the iterates. Since

$$\|\boldsymbol{z}^{(\ell)}\|_2 = \|F^{(\ell)}(\boldsymbol{z}^{(\ell-1)})\|_2 \le a_\ell(1 + \|\boldsymbol{z}^{(\ell-1)}\|_2),$$

an induction on ℓ yields

$$\|\boldsymbol{z}^{(M)}\|_2 \le C_M(1 + \|\boldsymbol{z}\|_2)$$

for some $C_M < \infty$. In particular,

$$\|\boldsymbol{\theta}^{(M)}\|_2 \le C_M(1 + \|\boldsymbol{z}\|_2).$$

Now the full log-density is given by

$$\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}) = -\frac{1}{2}\|\boldsymbol{\theta}^{(M)}\|_2^2 + \sum_{\ell=1}^{M} \log \left| \det \frac{\partial \boldsymbol{\theta}^{(\ell)}}{\partial \tilde{\boldsymbol{\theta}}^{(\ell)}} \right| + \sum_{\ell=1}^{M} \log \left| \det \frac{\partial \tilde{\boldsymbol{\theta}}^{(\ell)}}{\partial \boldsymbol{\theta}^{(\ell-1)}} \right| - \frac{d_\Theta}{2} \log(2\pi).$$

The second sum is constant in $\boldsymbol{z}$, since each term equals $\sum_i \log |\alpha_i^{(\ell)}|$. The first sum equals

$$\sum_{\ell=1}^{M} \sum_{i=1}^{\lfloor d_\Theta/2 \rfloor} \log \sigma_i^{(\ell)}\big(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}^{(\ell)}, \boldsymbol{x}\big),$$

which is uniformly bounded because each $\sigma_i^{(\ell)}$ is bounded above and below away from zero by the single-block analysis. Hence

$$|\ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})| \le C'_M(1 + \|\boldsymbol{\theta}^{(M)}\|_2^2) \le C'_M(1 + \|\boldsymbol{z}\|_2^2)$$

for some $C'_M < \infty$. Therefore $\ell_{\mathrm{NPE}}(q_\phi; \cdot) \in \mathcal{G}_2(\mathcal{Z})$.

We next control the gradient. By the chain rule,

$$\nabla_{\boldsymbol{z}} \boldsymbol{z}^{(M)}(\boldsymbol{z}) = \nabla_{\boldsymbol{z}} F^{(M)}\big(\boldsymbol{z}^{(M-1)}(\boldsymbol{z})\big) \cdots \nabla_{\boldsymbol{z}} F^{(1)}\big(\boldsymbol{z}^{(0)}(\boldsymbol{z})\big).$$

Hence

$$\|\nabla_{\boldsymbol{z}} \boldsymbol{z}^{(M)}(\boldsymbol{z})\| \le \prod_{\ell=1}^{M} b_\ell\big(1 + \|\boldsymbol{z}^{(\ell-1)}(\boldsymbol{z})\|_2\big) \le C''_M(1 + \|\boldsymbol{z}\|_2)^M$$

for some $C''_M < \infty$. Since $\boldsymbol{\theta}^{(M)}$ is the first component of $\boldsymbol{z}^{(M)}$, this implies

$$\|\nabla_{\boldsymbol{z}} \boldsymbol{\theta}^{(M)}(\boldsymbol{z})\| \le C''_M(1 + \|\boldsymbol{z}\|_2)^M.$$

Therefore

$$\nabla_{\boldsymbol{z}}\Big(\tfrac{1}{2}\|\boldsymbol{\theta}^{(M)}(\boldsymbol{z})\|_2^2\Big) = \nabla_{\boldsymbol{z}} \boldsymbol{\theta}^{(M)}(\boldsymbol{z})^\top \boldsymbol{\theta}^{(M)}(\boldsymbol{z}),$$

and so

$$\Big\|\nabla_{\boldsymbol{z}}\Big(\tfrac{1}{2}\|\boldsymbol{\theta}^{(M)}(\boldsymbol{z})\|_2^2\Big)\Big\| \le \|\nabla_{\boldsymbol{z}} \boldsymbol{\theta}^{(M)}(\boldsymbol{z})\|\, \|\boldsymbol{\theta}^{(M)}(\boldsymbol{z})\|_2 \le C'''_M(1 + \|\boldsymbol{z}\|_2)^{M+1}$$

for some $C'''_M < \infty$. The derivatives of the accumulated log-determinant terms are controlled similarly, since each block contributes derivatives of $\log \sigma_i^{(\ell)}$ composed with the pre-coupling variables $(\tilde{\boldsymbol{\theta}}_{\mathcal{B}}^{(\ell)}, \boldsymbol{x})$, and the derivatives of $\log \sigma_i^{(\ell)}$ are bounded by the single-block analysis. Hence

$$\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})\| \le C'''_M(1 + \|\boldsymbol{z}\|_2)^{M+1},$$

so $\|\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \cdot)\| \in \mathcal{G}_{M+1}(\mathcal{Z})$.

Finally, we control the Hessian. Differentiating

$$\nabla_{\boldsymbol{z}} \Big(\tfrac{1}{2} \|\boldsymbol{\theta}^{(M)}(\boldsymbol{z})\|_2^2 \Big) = \nabla_{\boldsymbol{z}} \boldsymbol{\theta}^{(M)}(\boldsymbol{z})^\top \boldsymbol{\theta}^{(M)}(\boldsymbol{z})$$

gives

$$\nabla_{\boldsymbol{z}}^2 \Big(\tfrac{1}{2} \|\boldsymbol{\theta}^{(M)}(\boldsymbol{z})\|_2^2 \Big) = \big(\nabla_{\boldsymbol{z}} \boldsymbol{\theta}^{(M)}(\boldsymbol{z}) \big)^\top \nabla_{\boldsymbol{z}} \boldsymbol{\theta}^{(M)}(\boldsymbol{z}) + \sum_i \theta_i^{(M)}(\boldsymbol{z}) \, \nabla_{\boldsymbol{z}}^2 \theta_i^{(M)}(\boldsymbol{z}).$$

It therefore remains to bound $\nabla_{\boldsymbol{z}}^2 \boldsymbol{z}^{(M)}(\boldsymbol{z})$. Since $\boldsymbol{z}^{(0)}(\boldsymbol{z}) = \boldsymbol{z}$, we have

$$\nabla_{\boldsymbol{z}} \boldsymbol{z}^{(0)}(\boldsymbol{z}) = I, \qquad \nabla_{\boldsymbol{z}}^2 \boldsymbol{z}^{(0)}(\boldsymbol{z}) = 0.$$

The second-order chain rule gives the recursion

$$\nabla_{\boldsymbol{z}}^2 \boldsymbol{z}^{(\ell)}(\boldsymbol{z}) = \nabla_{\boldsymbol{z}}^2 F^{(\ell)}\big(\boldsymbol{z}^{(\ell-1)}(\boldsymbol{z})\big) \big[\nabla_{\boldsymbol{z}} \boldsymbol{z}^{(\ell-1)}(\boldsymbol{z}), \nabla_{\boldsymbol{z}} \boldsymbol{z}^{(\ell-1)}(\boldsymbol{z}) \big] + \nabla_{\boldsymbol{z}} F^{(\ell)}\big(\boldsymbol{z}^{(\ell-1)}(\boldsymbol{z})\big) \, \nabla_{\boldsymbol{z}}^2 \boldsymbol{z}^{(\ell-1)}(\boldsymbol{z}),$$

Hence

$$\|\nabla_{\boldsymbol{z}}^2 \boldsymbol{z}^{(\ell)}(\boldsymbol{z})\| \le c_\ell (1 + \|\boldsymbol{z}^{(\ell-1)}(\boldsymbol{z})\|_2) \|\nabla_{\boldsymbol{z}} \boldsymbol{z}^{(\ell-1)}(\boldsymbol{z})\|^2 + b_\ell (1 + \|\boldsymbol{z}^{(\ell-1)}(\boldsymbol{z})\|_2) \|\nabla_{\boldsymbol{z}}^2 \boldsymbol{z}^{(\ell-1)}(\boldsymbol{z})\|.$$

Using the bounds

$$\|\boldsymbol{z}^{(\ell-1)}(\boldsymbol{z})\|_2 \le C_M (1 + \|\boldsymbol{z}\|_2), \qquad \|\nabla_{\boldsymbol{z}} \boldsymbol{z}^{(\ell-1)}(\boldsymbol{z})\| \le C_M'' (1 + \|\boldsymbol{z}\|_2)^{\ell-1},$$

an induction on ℓ yields

$$\|\nabla_{\boldsymbol{z}}^2 \boldsymbol{z}^{(M)}(\boldsymbol{z})\| \le C_M'''' (1 + \|\boldsymbol{z}\|_2)^{2M-1}$$

for some $C_M'''' < \infty$. In particular,

$$\|\nabla_{\boldsymbol{z}}^2 \boldsymbol{\theta}^{(M)}(\boldsymbol{z})\| \le C_M'''' (1 + \|\boldsymbol{z}\|_2)^{2M-1}.$$

Therefore

$$\Big\| \nabla_{\boldsymbol{z}}^2 \Big(\tfrac{1}{2} \|\boldsymbol{\theta}^{(M)}(\boldsymbol{z})\|_2^2 \Big) \Big\| \le \|\nabla_{\boldsymbol{z}} \boldsymbol{\theta}^{(M)}(\boldsymbol{z})\|^2 + \|\boldsymbol{\theta}^{(M)}(\boldsymbol{z})\|_2 \, \|\nabla_{\boldsymbol{z}}^2 \boldsymbol{\theta}^{(M)}(\boldsymbol{z})\| \le C_M''''' (1 + \|\boldsymbol{z}\|_2)^{2M}$$

for some $C_M''''' < \infty$. The second derivatives of the accumulated log-determinant terms are controlled in the same way and are of no larger polynomial order. Hence

$$\|\nabla_{\boldsymbol{z}}^2 \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z})\| \le C_M''''' (1 + \|\boldsymbol{z}\|_2)^{2M},$$

so $\|\nabla_{\boldsymbol{z}}^2 \ell_{\mathrm{NPE}}(q_\phi; \cdot)\| \in \mathcal{G}_{2M}(\mathcal{Z})$. Since each $F^{(\ell)}$ is C^2, the composition is also C^2, and therefore $\ell_{\mathrm{NPE}}(q_\phi; \cdot) \in C^2(\mathcal{Z})$. □

Lemma 4.3 then follows directly from Lemmas B.5 and B.7.

C Experimental details

This section is organised as follows. Appendix C.1 outlines our hyperparameter-selection procedure using the KL-based miscalibration metric, Appendix C.3 describes the baseline methods, Appendix C.4 contains the implementation details, Appendix C.5 contains the details of the simulators used in Section 5, and Appendix C.6 discusses the computational cost.

C.1 Hyperparameter selection method for DRO-NPE

Estimating KL-based miscalibration. We show that KL-based miscalibration can be estimated by density-ratio estimation, but first introduce the notation needed for this. Recall that $P_{\boldsymbol{x}}$ is the marginal distribution obtained from the joint $P \equiv P_{\boldsymbol{\theta},\boldsymbol{x}}$, and that $F_{S|\boldsymbol{x}}^Q$ is the CDF of $Q_{S|\boldsymbol{x}}$, which is the distribution of $S(\boldsymbol{\theta}; \boldsymbol{x})$ whenever $\boldsymbol{\theta} \sim Q_{\boldsymbol{\theta}|\boldsymbol{x}}$ for some fixed $\boldsymbol{x}$. For each $\boldsymbol{x}$, we define $\widetilde{U} \mid \boldsymbol{x} \sim (F_{S|\boldsymbol{x}}^Q)_\# P_{S|\boldsymbol{x}}$. Let $P_{\widetilde{U},\boldsymbol{x}}$ denote the resulting joint distribution of $(\widetilde{U}, \boldsymbol{x})$. We also define an independent reference pair $(U, \boldsymbol{x})$, where $U \sim \mathcal{U}[0,1]$ is independent of $\boldsymbol{x} \sim P_{\boldsymbol{x}}$, so that

$P_{U,\boldsymbol{x}} = P_U \otimes P_{\boldsymbol{x}}$. We write the respective densities as $p_{U,\boldsymbol{x}}(u, \boldsymbol{x}) = p_U(u)p_{\boldsymbol{x}}(\boldsymbol{x}) = p_{\boldsymbol{x}}(\boldsymbol{x})$ since $p_U(u) = 1$ on $[0, 1]$, and $p_{\widetilde{U},\boldsymbol{x}}(u, \boldsymbol{x}) = p_{\widetilde{U}|\boldsymbol{x}}(u \mid \boldsymbol{x})p_{\boldsymbol{x}}(\boldsymbol{x})$.

Under the continuity conditions for the uniform reference, the KL-based miscalibration is

$$\mathrm{kl}^S_{\mathrm{cal}}(q_\phi; P) = \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[-h\left((F^Q_{S|\boldsymbol{x}})_{\#}P_{S|\boldsymbol{x}}\right)\right] = \int p_{\boldsymbol{x}}(\boldsymbol{x}) \int p_{\widetilde{U}|\boldsymbol{x}}(u \mid \boldsymbol{x}) \log p_{\widetilde{U}|\boldsymbol{x}}(u \mid \boldsymbol{x})\, du\, d\boldsymbol{x},$$

where $h(\mu) = -\int p(u) \log p(u)\, du$ denotes the differential entropy. Since $p_U(u) = 1$ on $[0, 1]$, we can equivalently write

$$\begin{aligned}
&\int p_{\boldsymbol{x}}(\boldsymbol{x}) \int p_{\widetilde{U}|\boldsymbol{x}}(u \mid \boldsymbol{x}) \log p_{\widetilde{U}|\boldsymbol{x}}(u \mid \boldsymbol{x})\, du\, d\boldsymbol{x} \\
&\quad = \iint p_{\widetilde{U}|\boldsymbol{x}}(u \mid \boldsymbol{x})p_{\boldsymbol{x}}(\boldsymbol{x}) \log \frac{p_{\widetilde{U}|\boldsymbol{x}}(u \mid \boldsymbol{x})p_{\boldsymbol{x}}(\boldsymbol{x})}{p_U(u)p_{\boldsymbol{x}}(\boldsymbol{x})}\, du\, d\boldsymbol{x} \\
&\quad = \mathrm{KL}\left(P_{\widetilde{U},\boldsymbol{x}} \,\middle\|\, P_{U,\boldsymbol{x}}\right).
\end{aligned}$$

Therefore, $\mathrm{kl}^S_{\mathrm{cal}}(q_\phi; P)$ can be estimated by density-ratio estimation between samples from the dependent joint $P_{\widetilde{U},\boldsymbol{x}}$ and the independent reference $P_U \otimes P_{\boldsymbol{x}}$.

To this end, we formulate a binary classification problem: samples from class $y = 1$ are drawn from $(\widetilde{U}, \boldsymbol{x}) \sim P_{\widetilde{U},\boldsymbol{x}}$, while samples from class $y = 0$ are drawn from $(U, \boldsymbol{x}) \sim P_{U,\boldsymbol{x}} = P_U \otimes P_{\boldsymbol{x}}$. With equal class priors, the Bayes-optimal classifier $D^\star : [0, 1] \times \mathcal{X} \to [0, 1]$ is given by

$$D^\star(u, \boldsymbol{x}) = \frac{p_{\widetilde{U},\boldsymbol{x}}(u, \boldsymbol{x})}{p_{\widetilde{U},\boldsymbol{x}}(u, \boldsymbol{x}) + p_{U,\boldsymbol{x}}(u, \boldsymbol{x})}.$$

Rearranging gives

$$\frac{p_{\widetilde{U}|\boldsymbol{x}}(u \mid \boldsymbol{x})p_{\boldsymbol{x}}(\boldsymbol{x})}{p_U(u)p_{\boldsymbol{x}}(\boldsymbol{x})} = \frac{p_{\widetilde{U},\boldsymbol{x}}(u, \boldsymbol{x})}{p_{U,\boldsymbol{x}}(u, \boldsymbol{x})} = \frac{D^\star(u, \boldsymbol{x})}{1 - D^\star(u, \boldsymbol{x})}.$$

Thus, the desired ratio can be recovered directly from the classifier via its odds. In practice, we train a discriminator D_ψ by minimising the binary cross-entropy loss, yielding the approximation

$$\frac{p_{\widetilde{U},\boldsymbol{x}}(u, \boldsymbol{x})}{p_{U,\boldsymbol{x}}(u, \boldsymbol{x})} \approx \frac{D_{\widehat{\psi}}(u, \boldsymbol{x})}{1 - D_{\widehat{\psi}}(u, \boldsymbol{x})},$$

where $\widehat{\psi}$ are the trained parameters. Finally, given $D_{\hat{\psi}}$, we can obtain Monte Carlo estimation of $\mathrm{kl}^S_{\mathrm{cal}}$ with sample size N as

$$\widehat{\mathrm{kl}}^S_{\mathrm{cal}}(q_\phi; P) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{D_{\hat{\psi}}(\tilde{U}_i, \boldsymbol{x}_i)}{1 - D_{\hat{\psi}}(\tilde{U}_i, \boldsymbol{x}_i)}.$$

In practice, we summarise $\boldsymbol{x}$ by the total posterior variance $\gamma_\phi(\boldsymbol{x}) = \sum_{j=1}^{d_\Theta} \mathrm{Var}_{Q_{\boldsymbol{\theta}|\boldsymbol{x}}}(\theta_j \mid \boldsymbol{x})$ and fit the logistic classifier

$$D_\psi(u, \gamma) = [1 + \exp(-(\psi_0 + \psi_1 u + \psi_2 \gamma))]^{-1}$$

We use this low-dimensional summary and a simple classifier to limit estimation complexity. This is particularly important in the small-sample regime, where D_ψ and $\widehat{\mathrm{kl}}^S_{\mathrm{cal}}$ must be estimated using an even smaller subset of the available data reserved for validation; a more complex estimator would overfit, leading to unreliable estimates of $\widehat{\mathrm{kl}}^S_{\mathrm{cal}}$ and, consequently, a poor choice of ε.

The resulting score is a practical approximation: the compression of $\boldsymbol{x}$, the restricted classifier, and finite posterior sampling can limit its sensitivity to miscalibration. Alternatively, when sufficient samples are available, more flexible models such as a neural network could be used to learn $D_\psi(u, \boldsymbol{x})$ directly from the full $\boldsymbol{x}$.

We use $M = 1000$ posterior samples per evaluation pair and $K = 5$ classifier folds. Cross-fitting separates classifier fitting from log-odds evaluation, as detailed in Algorithm 1. Note that we only

Algorithm 1 Cross-fitted estimation of KL-based miscalibration

Require: Evaluation pairs $\{(\boldsymbol{\theta}_i, \boldsymbol{x}_i)\}_{i=1}^N$, posterior density estimator q_ϕ, summary S, number of posterior samples $\mathcal{M}$, number of folds K

Ensure: Estimate $\widehat{\mathrm{kl}}_{\mathrm{cal}}^S$ of KL-based miscalibration

(a) Construct rank samples

for $i = 1, \dots, N$ **do**

  Draw $\tilde{\boldsymbol{\theta}}_i^{(1)}, \dots, \tilde{\boldsymbol{\theta}}_i^{(\mathcal{M})} \overset{\text{iid}}{\sim} q_\phi(\cdot \mid \boldsymbol{x}_i)$

  Compute

$$\widetilde{u}_i = \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} \mathbf{1}\{S(\tilde{\boldsymbol{\theta}}_i^{(m)}; \boldsymbol{x}_i) \le S(\boldsymbol{\theta}_i; \boldsymbol{x}_i)\}$$

end for

(b) Summarise observations by posterior variance

for $i = 1, \dots, N$ **do**

  Using the posterior samples, compute

$$\widehat{\gamma}_i = \sum_{j=1}^{d_\Theta} \widehat{\mathrm{Var}}\left[\tilde{\theta}_j \mid \boldsymbol{x}_i\right]$$

end for

(c) Split evaluation pairs into K folds

Partition $\{1, \dots, N\}$ into $\mathcal{I}_1, \dots, \mathcal{I}_K$

Draw independent $U_i \sim \mathcal{U}[0,1]$ for $i = 1, \dots, N$

for $k = 1, \dots, K$ **do**

(d) Fit the classifier on the other $K-1$ folds

  Fit logistic regression by binary cross-entropy to distinguish

$$\{(\widetilde{U}_i, \widehat{\gamma}_i, 1) : i \notin \mathcal{I}_k\} \quad \text{from} \quad \{(U_i, \widehat{\gamma}_i, 0) : i \notin \mathcal{I}_k\},$$

  using $D_\psi(u, \gamma) = [1 + \exp(-(\psi_0 + \psi_1 u + \psi_2 \gamma))]^{-1}$

(e) Evaluate on held-out observed-rank pairs

  for $i \in \mathcal{I}_k$ **do**

    Compute

$$\widehat{r}_i = \log \frac{D_{\widehat{\psi}^{(-k)}}(\widetilde{U}_i, \widehat{\gamma}_i)}{1 - D_{\widehat{\psi}^{(-k)}}(\widetilde{U}_i, \widehat{\gamma}_i)}$$

  end for

end for

return $\widehat{\mathrm{kl}}_{\mathrm{cal}}^S = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \widehat{r}_i$

Selecting ε. Given an estimate of $\mathrm{kl}_{\mathrm{cal}}^S$, we choose ε using a validation set. We split n samples from P into training and validation sets, reserving 10% for validation, and minimise the validation estimate of $\mathrm{kl}_{\mathrm{cal}}^S$ over $\varepsilon \in [\varepsilon_{\min}, \varepsilon_{\max}]$. For each queried value of ε, we train q_ϕ on the training set and evaluate $\mathrm{kl}_{\mathrm{cal}}^S$ on the validation set. The selected value $\varepsilon^\star$ is then used to retrain q_ϕ on the full dataset. In our experiments, we perform Bayesian optimisation (Garnett, 2023) over $\log \varepsilon$ on the interval $[\log 0.001, \log 10]$, using a budget of 10 iterations. Optimising over $\log \varepsilon$ explores each order of magnitude equally, which is appropriate because ε controls the strength of regularisation.

C.2 Alternative approaches to select ε

Selecting ε in DRO typically involves a trade-off between theoretical guarantees, empirical performance, and computational cost. In our experiments, we prioritise empirical performance by selecting ε via hold-out validation. To illustrate this trade-off, we describe two alternative approaches: an explicit theoretical radius, and a bootstrap estimate of $W_2(P_n, P)$.

Explicit theoretical radius. The radius $\varepsilon(n, \delta)$ in Theorem 4.1 relies on the concentration bounds of Fournier and Guillin (2015), whose constants are not explicit. Fournier (2023) instead provides explicit, computable constants, at the cost of a looser bound. Specifically, for P supported in a ball of radius R with respect to the supremum norm $\|\cdot\|_\infty$, Fournier (2023, Section 2.5, Table 2) gives

$$\sqrt{\mathbb{E}\big[W_2^{(\infty)}(P_n, P)^2\big]} \leq B(2, n, d_{\mathcal{Z}}),$$

where $W_2^{(\infty)}$ denotes the 2-Wasserstein distance with ground cost induced by $\|\cdot\|_\infty$, and the expectation is over $z_{1:n} \overset{\text{i.i.d.}}{\sim} P$. Since $\|z\|_2 \leq \sqrt{d_{\mathcal{Z}}}\,\|z\|_\infty$ for all $z \in \mathcal{Z}$, the Euclidean W_2 used throughout satisfies

$$\mathbb{E}\big[W_2(P_n, P)^2\big] \leq d_{\mathcal{Z}}\,\mathbb{E}\big[W_2^{(\infty)}(P_n, P)^2\big] \leq d_{\mathcal{Z}}\,B(2, n, d_{\mathcal{Z}})^2.$$

This bound holds in expectation, whereas Theorem 4.1 requires $P \in \mathcal{A}_2(P_n; \varepsilon)$ with probability at least $1 - \delta$. Applying Markov's inequality to $W_2(P_n, P)^2$ gives

$$\mathbb{P}\big(W_2(P_n, P) > \varepsilon\big) = \mathbb{P}\big(W_2(P_n, P)^2 > \varepsilon^2\big) \leq \frac{d_{\mathcal{Z}}\,B(2, n, d_{\mathcal{Z}})^2}{\varepsilon^2},$$

so that choosing

$$\varepsilon_n := \sqrt{\frac{d_{\mathcal{Z}}}{\delta}}\,B(2, n, d_{\mathcal{Z}})$$

ensures $P \in \mathcal{A}_2(P_n; \varepsilon_n)$ with probability at least $1-\delta$, and hence that the upper bound in Theorem 4.1 holds.

As an example, consider Two Moons, for which $d_{\mathcal{Z}} = 4$ and Fournier (2023, Table 2) gives

$$B(2, n, 4) = 2R\,\frac{\sqrt{0.73 \log n + 1.26}}{n^{1/4}}.$$

Two Moons is not compactly supported, owing to the Gaussian noise term $r \sim \mathcal{N}(0.1, 0.01^2)$, but its tails decay exponentially fast. We take $R = 1.12$, the radius of the smallest supremum-norm ball containing at least 95% of the mass of P; a fully rigorous treatment would instead use the bounds of Fournier (2023) for non-compactly supported distributions. With $\delta = 0.05$, this gives $\varepsilon_n \approx 8.90$ for $n = 1024$ and $\varepsilon_n \approx 5.11$ for $n = 16384$.

Bootstrapping. The second approach estimates $W_2(P_n, P)$ directly by bootstrapping the simulated data. This choice satisfies the theory up to bootstrapping error, which can be made negligible with a large number of bootstrap samples. It is also considerably cheaper than hold-out validation, since it does not require retraining the normalising flow.

Comparison. For Two Moons with $n = 1024$ and $\delta = 0.05$, the explicit theoretical radius gives $\varepsilon_n \approx 8.90$ at virtually zero cost, whereas bootstrapping gives $\varepsilon_n \approx 0.129$, averaged over five random seeds with 1000 bootstrap samples, and takes 34 seconds on average. Both values are substantially larger than those obtained via hold-out validation, which are of order 10^{-3}. However, hold-out validation is considerably more expensive, requiring over 1000 seconds in this example. Overall, hold-out validation yields substantially smaller, less conservative choices of ε, but does not guarantee that the theoretical conditions are satisfied. Nevertheless, it is commonly used in DRO, and our experiments demonstrate its effectiveness in the context of NPE.

C.3 Description of baseline methods

Balanced NPE (Delaunoy et al., 2023). The method extends the Balanced NRE method proposed by Delaunoy et al. (2022) to NPE. NRE relies on the ratio $r(\boldsymbol{x} \mid \boldsymbol{\theta}) := \frac{p(\boldsymbol{\theta}, \boldsymbol{x})}{p(\boldsymbol{\theta})p(\boldsymbol{x})}$ and the Bayes-optimal classifier $d^\star : \Theta \times \mathcal{X} \to [0, 1]$, which distinguishes samples from the joint distribution $P_{\boldsymbol{\theta},\boldsymbol{x}}$ and the product of marginals $\Pi \otimes P_{\boldsymbol{x}}$. The optimal classifier is defined as $d^\star(\boldsymbol{\theta}, \boldsymbol{x}) := \frac{p(\boldsymbol{\theta}, \boldsymbol{x})}{p(\boldsymbol{\theta}, \boldsymbol{x}) + p(\boldsymbol{\theta})p(\boldsymbol{x})}$, and relates to $r(\boldsymbol{x} \mid \boldsymbol{\theta})$ via $r(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{d^\star(\boldsymbol{\theta}, \boldsymbol{x})}{1 - d^\star(\boldsymbol{\theta}, \boldsymbol{x})}$.

In practice, the optimal classifier is not available and is estimated from finite samples, denoted $\hat{d}$. Delaunoy et al. (2023) suggest that overconfident posteriors obtained from NRE originate from

overconfidence in $\hat{d}$. To mitigate this, they introduce a regularisation term that encourages $\hat{d}$ to be more conservative than $d^\star$. Specifically, they define a balanced classifier d_b satisfying $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{x}\sim P_{\boldsymbol{\theta},\boldsymbol{x}}}[d_b(\boldsymbol{\theta},\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{x}\sim \Pi\otimes P_{\boldsymbol{x}}}[d_b(\boldsymbol{\theta},\boldsymbol{x})]=1$. They show that any balanced classifier satisfies $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{x}\sim P_{\boldsymbol{\theta},\boldsymbol{x}}}\left[\frac{d^\star(\boldsymbol{\theta},\boldsymbol{x})}{d_b(\boldsymbol{\theta},\boldsymbol{x})}\right]\geq 1$ and $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{x}\sim \Pi\otimes P_{\boldsymbol{x}}}\left[\frac{1-d^\star(\boldsymbol{\theta},\boldsymbol{x})}{1-d_b(\boldsymbol{\theta},\boldsymbol{x})}\right]\geq 1$, implying that d_b is less confident than $d^\star$ in expectation.

To impose this property in the NPE setting, Delaunoy et al. (2023) add the balancing regularisation term $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{x}\sim P_{\boldsymbol{\theta},\boldsymbol{x}}}[\hat{d}(\boldsymbol{\theta},\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{x}\sim \Pi\otimes P_{\boldsymbol{x}}}[\hat{d}(\boldsymbol{\theta},\boldsymbol{x})]-1$ on top of the standard NPE loss. This term is estimated using the approximate posterior density by noting that $r(\boldsymbol{x}\mid\boldsymbol{\theta})=\frac{p(\boldsymbol{\theta}\mid\boldsymbol{x})}{p(\boldsymbol{\theta})}\approx\frac{q_\phi(\boldsymbol{\theta}\mid\boldsymbol{x})}{p(\boldsymbol{\theta})}$ and $\hat{d}(\boldsymbol{\theta},x)\approx r(\boldsymbol{x}\mid\boldsymbol{\theta})/(1+r(\boldsymbol{x}\mid\boldsymbol{\theta}))$. The resulting regularised objective is

$$\mathcal{L}_{\text{Bal}}(q_\phi;P_{\boldsymbol{\theta},\boldsymbol{x}},\Pi\otimes P_{\boldsymbol{x}}):=\mathcal{L}_{\text{NPE}}(q;P_{\boldsymbol{\theta},\boldsymbol{x}})+\lambda\left[\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{x}\sim P_{\boldsymbol{\theta},\boldsymbol{x}}}\left[\frac{\frac{q_\phi(\boldsymbol{\theta}\mid\boldsymbol{x})}{p(\boldsymbol{\theta})}}{1+\frac{q_\phi(\boldsymbol{\theta}\mid\boldsymbol{x})}{p(\boldsymbol{\theta})}}\right]+\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{x}\sim \Pi\otimes P_{\boldsymbol{x}}}\left[\frac{\frac{q_\phi(\boldsymbol{\theta}\mid\boldsymbol{x})}{p(\boldsymbol{\theta})}}{1+\frac{q_\phi(\boldsymbol{\theta}\mid\boldsymbol{x})}{p(\boldsymbol{\theta})}}\right]-1\right],$$

where $\lambda>0$ is a regularisation parameter. In their implementation, samples from $\Pi\otimes P_{\boldsymbol{x}}$ are approximated using joint samples $\{(\boldsymbol{\theta}_i,\boldsymbol{x}_i)\}_{i=1}^n\sim P_{\boldsymbol{\theta},\boldsymbol{x}}$. While these are used for estimating terms involving the joint, samples from the marginals are constructed by cyclically shifting $\boldsymbol{\theta}$, yielding mismatched pairs $\{(\boldsymbol{\theta}_n,\boldsymbol{x}_1),(\boldsymbol{\theta}_1,\boldsymbol{x}_2),\dots,(\boldsymbol{\theta}_{n-1},\boldsymbol{x}_n)\}$.

Coverage-regularised NPE (Falkiewicz et al., 2023). This method penalises deviations of the SBC ranks, computed as in Definition 2.2, from the standard uniform distribution, using $S(\boldsymbol{\theta};\boldsymbol{x})=q_\phi(\boldsymbol{\theta}\mid\boldsymbol{x})$. Given $(\boldsymbol{\theta},\boldsymbol{x})\sim P_{\boldsymbol{\theta},\boldsymbol{x}}$, define

$$u=u(q;\boldsymbol{\theta},\boldsymbol{x})=\frac{1}{\mathcal{M}}\sum_{m=1}^{\mathcal{M}}\mathbf{1}\Big\{q_\phi\left(\tilde{\boldsymbol{\theta}}^{(m)}\mid\boldsymbol{x}\right)<q_\phi\left(\boldsymbol{\theta}\mid\boldsymbol{x}\right)\Big\}\quad\text{where}\quad\tilde{\boldsymbol{\theta}}^{(1)},\dots,\tilde{\boldsymbol{\theta}}^{(\mathcal{M})}\sim Q_{\boldsymbol{\theta}\mid\boldsymbol{x}}.$$

Let $u_{(1)},\dots,u_{(n)}$ denote the sorted sequence of $\{u_i\}_{i=1}^n$ in ascending order with $u_i=u(q_\phi;\boldsymbol{\theta}_i,\boldsymbol{x}_i)$. The empirical CDF evaluated at these points is given by $\hat{F}(u_{(i)})=\frac{i}{n}$ for $i\in\{1,\dots,n\}$. This empirical CDF is compared to the CDF of the standard uniform distribution, which satisfies $F(u_{(i)})=u_{(i)}$. The resulting calibration error is defined as the L^2 distance between F and $\hat{F}$, and in practice this is estimated through samples $u_1,\dots,u_n$: $\frac{1}{n}\sum_{i=1}^n\left(u_{(i)}-\frac{i}{n}\right)^2$. To encourage conservative approximate posteriors, the authors propose penalising only overconfident deviations. This is achieved by introducing a one-sided penalty that considers only cases where the uniform CDF dominates the empirical CDF, namely $\frac{1}{n}\sum_{i=1}^n(\max\{u_{(i)}-\frac{i}{n},0\})^2$.

To allow backpropagation with respect to ϕ, u is estimated via self-normalised importance sampling with the prior as a proposal distribution such that

$$u=\frac{\sum_{m=1}^{\mathcal{M}}\frac{q_\phi(\tilde{\boldsymbol{\theta}}^{(m)}\mid\boldsymbol{x})}{p(\tilde{\boldsymbol{\theta}}^{(m)})}\mathbf{1}\{q_\phi(\tilde{\boldsymbol{\theta}}^{(m)}\mid\boldsymbol{x})<q_\phi(\boldsymbol{\theta}\mid\boldsymbol{x})\}}{\sum_{m=1}^{\mathcal{M}}\frac{q_\phi(\tilde{\boldsymbol{\theta}}^{(m)}\mid\boldsymbol{x})}{p(\tilde{\boldsymbol{\theta}}^{(m)})}},$$

where $\tilde{\boldsymbol{\theta}}^{(1)},\dots,\tilde{\boldsymbol{\theta}}^{(\mathcal{M})}\sim\Pi$. Moreover, differentiable relaxations of both the indicator function and the sorting operation are employed. The resulting regularised objective is

$$\mathcal{L}_{\text{Cal}}(q_\phi;P_n):=\mathcal{L}_{\text{NPE}}(q_\phi;P_n)+\lambda\frac{1}{n}\sum_{i=1}^n\left(\max\left\{u_{(i)}\left(q_\phi;\boldsymbol{\theta}_{(i)},\boldsymbol{x}_{(i)}\right)-\frac{i}{n},0\right\}\right)^2.$$

C.4 Implementation details

Conditional density estimator. Across all experiments, we employ a conditional normalising flow model based on coupling layers, as implemented in the *BayesFlow* package (Kühmichel et al., 2026), for both DRO-NPE and the baseline methods. We use six coupling layers and adopt affine transformations as the coupling functions, following the default configuration. For the multi-layer perceptron used to predict the parameters of the affine transformations, we retain the default architecture except for the choice of activation function: we replace the standard activation with the tanh, which ensures that the theoretical assumptions underlying our method are satisfied.

Training procedure. The model is trained for 1000 epochs with a batch size of 64. We train the network using the AdamW optimiser (Loshchilov and Hutter, 2017) with a learning rate of 5×10^{-4}. Unless otherwise stated, we do not use a validation set for early stopping or model selection, in order to isolate the effect of the training objective. Consequently, the results of our experiments may not exactly reproduce those reported in earlier works where this effect is not explicitly disentangled. For CR-NPE, we observed numerical instabilities during prolonged training. Therefore, for this method, we train the network until instability occurs (i.e., loss value becoming *nan*) or for 1000 epochs (whichever occurs first). In case of instability, we set the final model to be the one in the last stable training epoch. Note that while this instability is also present in the original implementation, it is more pronounced in our setting, potentially due to the use of a different conditional density estimator from the original paper. Nevertheless, we retain this choice to ensure a fair and meaningful comparison across methods. For ε-selection in DRO-NPE, we use 10% of the data as validation set.

Hyperparameters for other methods. Following the recommendations in the respective works, we set $\lambda = 5$ and $M = 16$ in CR-NPE and $\lambda = 100$ in Bal-NPE.

C.5 Description of simulators

SLCP. The simulator is deliberately constructed to exhibit a simple likelihood while inducing a complex posterior, making it well-suited for benchmarking. The parameter $\boldsymbol{\theta} \in \mathbb{R}^5$ governs a two-dimensional Gaussian distribution. We draw four independent samples from this distribution and concatenate them, resulting in an observation $\boldsymbol{x} \in \mathbb{R}^8$ (Papamakarios et al., 2019; Lueckmann et al., 2021; Hermans et al., 2022).

$$
\begin{aligned}
\textbf{Prior:} \quad & \boldsymbol{\theta} \sim \mathcal{U}([-3,3]^5) \\
\textbf{Model:} \quad & \boldsymbol{x}_i \sim \mathcal{N}\left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \begin{bmatrix} \theta_3^4 & \tanh(\theta_5)\,\theta_3^2\theta_4^2 \\ \tanh(\theta_5)\,\theta_3^2\theta_4^2 & \theta_4^4 \end{bmatrix} \right), \quad i = 1, \ldots, 4 \\
& \boldsymbol{x} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_4^\top]^\top
\end{aligned}
$$

Two moons. The two moons simulator is a commonly used toy benchmark in SBI, designed to exhibit a posterior distribution with multimodality. Both the parameter and observation spaces are two-dimensional, with $\boldsymbol{\theta} \in \mathbb{R}^2$ and $\boldsymbol{x} \in \mathbb{R}^2$ (Lueckmann et al., 2021; Greenberg et al., 2019).

$$
\begin{aligned}
\textbf{Prior:} \quad & \boldsymbol{\theta} \sim \mathcal{U}([-1,1]^2), \\
\textbf{Model:} \quad & \boldsymbol{x} = \begin{bmatrix} -|\theta_1 + \theta_2|/\sqrt{2} \\ (-\theta_1 + \theta_2)/2 \end{bmatrix} + \begin{bmatrix} r\cos\alpha + 0.25 \\ r\sin\alpha \end{bmatrix}, \\
& \text{where } \alpha \sim \mathcal{U}(-\pi/2, \pi/2) \text{ and } r \sim \mathcal{N}(0.1, 0.01^2).
\end{aligned}
$$

Lotka–Volterra. The Lotka–Volterra model describes the nonlinear dynamics of interacting prey and predator populations. The objective is to infer the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^4$, which controls the species interaction dynamics, from noisy observations of both populations over time. Each population is summarised at 10 evenly spaced time points, yielding an observation $\boldsymbol{x} \in \mathbb{R}^{20}$ (Lueckmann et al., 2021; Lotka, 1920).

$$
\begin{aligned}
\textbf{Prior:} \quad & \theta_1 \sim \text{LogNormal}(-0.125, 0.5), \quad \theta_2 \sim \text{LogNormal}(-3, 0.5) \\
& \theta_3 \sim \text{LogNormal}(-0.125, 0.5), \quad \theta_4 \sim \text{LogNormal}(-3, 0.5) \\
\textbf{Model:} \quad & x_{1,i} \sim \text{LogNormal}(\log(X(t_i)), 0.1), \quad i = 1, \ldots 10 \\
& x_{2,i} \sim \text{LogNormal}(\log(Y(t_i)), 0.1), \quad i = 1, \ldots 10 \\
& \text{where } (X, Y) \text{ evolve according to} \\
& \frac{dX}{dt} = \theta_1 X - \theta_2 XY, \quad \frac{dY}{dt} = -\theta_3 Y + \theta_4 XY, \\
& \text{with time horizon } T = 20, \text{ and initial condition } (X(0), Y(0)) = (30, 1). \\
& \text{Observations are collected at 10 evenly spaced time points } t_i \in [0, T].
\end{aligned}
$$

Inverse Kinematics. The inverse kinematics benchmark models a two-dimensional, multi-jointed robotic arm whose forward mapping sends joint configurations to the arm endpoint position. The

parameter $\boldsymbol{\theta} \in \mathbb{R}^4$ specifies the configuration of the arm, consisting of an initial height θ_1 and three subsequent joint angles $\theta_2, \theta_3, \theta_4$. The resulting observation $\boldsymbol{x} \in \mathbb{R}^2$ corresponds to the two-dimensional coordinates of the arm (Kruse et al., 2021; Schmitt et al., 2024).

$$
\begin{aligned}
\textbf{Prior:} \quad & \boldsymbol{\theta} \sim \mathcal{N}\big(\mathbf{0}, \operatorname{diag}\big(\tfrac{1}{16}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}\big)\big) \\
\textbf{Model:} \quad & x_1 = l_1 \sin(\theta_2) + l_2 \sin(\theta_2 + \theta_3) + l_3 \sin(\theta_2 + \theta_3 + \theta_4) + \theta_1 \\
& x_2 = l_1 \cos(\theta_2) + l_2 \cos(\theta_2 + \theta_3) + l_3 \cos(\theta_2 + \theta_3 + \theta_4) \\
& \text{with segment lengths: } l_1 = l_2 = \frac{1}{2}, l_3 = 1.
\end{aligned}
$$

Cosmology. We use the high-fidelity hydrodynamic simulations from the CAMELS suite (Villaescusa-Navarro et al., 2021, 2023), a benchmark dataset for machine learning in astrophysics. These simulations model 25 Mpc$/h$ cosmological volumes and include complex astrophysical processes such as feedback. The inference task is to estimate the cosmological parameters $\theta = (\Omega_m, \sigma_8)$, where Ω_m is the matter density and σ_8 the amplitude of fluctuations, from mock power spectrum measurements $P(k)$.

C.6 Computational cost comparison

We now compare the computational overhead of DRO-NPE with that of the baselines. Let B be the batch size and C_ϕ the cost of one evaluation of $-\log q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$, which is determined by the architecture of q_ϕ, including its input dimension $d_{\mathcal{Z}}$. We use coupling flows in our experiments. For a coupling flow with $M \geq 1$ coupling blocks, whose subnetworks are multi-layer perceptrons (MLPs) with $L \geq 1$ hidden layers of $H \geq 1$ hidden units each (i.e. $\mathbf{h}^{(\ell)} \in \mathbb{R}^H$ for $1 \leq \ell \leq L$, with hidden layers defined analogously to Appendix B), the cost C_ϕ of one evaluation is of order $MH(d_{\mathcal{Z}} + (L-1)H)$. We first bound the cost of evaluating each training objective on a mini-batch.

Standard NPE evaluates the density once per sample, so the NPE objective costs $\mathcal{O}(BC_\phi)$. Bal-NPE additionally evaluates the density on the cyclically shifted pairs, which changes the cost only by a constant factor, and therefore also costs $\mathcal{O}(BC_\phi)$. CR-NPE evaluates the density at $\mathcal{M}$ importance samples per observation, so its cost grows as $\mathcal{O}(B\mathcal{M}C_\phi)$.

DRO-NPE additionally requires the input gradients $\nabla_{\boldsymbol{z}} \ell_{\mathrm{NPE}}(q_\phi; \boldsymbol{z}_i)$ and the regulariser $\Omega(\phi; P_B)$. Reverse-mode automatic differentiation computes the gradient of a scalar at a constant multiple of the cost of evaluating it, without a multiplicative dependence on the number of variables (Baydin et al., 2018, §3.2). Hence, the input gradients for $i = 1, \ldots, B$ cost $\mathcal{O}(BC_\phi)$, without an additional factor of $d_{\mathcal{Z}}$. Computing $\Omega(\phi; P_B)$ from them requires the squared norm of each $d_{\mathcal{Z}}$-dimensional input gradient and an average over the B samples, which costs $\mathcal{O}(Bd_{\mathcal{Z}})$. This is also $\mathcal{O}(BC_\phi)$ since C_ϕ is of order $MH(d_{\mathcal{Z}} + (L-1)H)$, which is at least $d_{\mathcal{Z}}$ because $M, H \geq 1$. Evaluating the DRO-NPE objective therefore also costs $\mathcal{O}(BC_\phi)$, as for NPE, but with a larger constant.

Finally, each objective is a scalar, so by the same reverse-mode argument, its gradient with respect to ϕ costs a constant multiple of its evaluation. A full training step therefore has the same order of cost as evaluating the objective: $\mathcal{O}(BC_\phi)$ for NPE, Bal-NPE and DRO-NPE, and $\mathcal{O}(B\mathcal{M}C_\phi)$ for CR-NPE. Constant factors and actual run times depend on the hardware and implementation, so for DRO-NPE we report measured wall-clock times (Figure 3(d)): DRO-NPE takes about $2\times$ the time of NPE on CPU and $1.25\times$ on GPU.

Cost of hyperparameter-selection. Our full DRO-NPE procedure performs 10 Bayesian optimisation (BO) iterations, consisting of (i) training of the posterior estimator q_ϕ, (ii) evaluation of KL-based miscalibration $\widehat{\mathrm{kl}}^q_{\mathrm{cal}}$ on validation set, and (iii) selection of next ε using BO. These are followed by retraining of q_ϕ on the full dataset using the selected ε. Table 1 compares this total cost with the training times of standard NPE and DRO-NPE with fixed ε. All the experiments are conducted on GPU. Including hyperparameter selection increases the total computational cost to approximately 13–16 times that of DRO-NPE training with fixed ε.

We also examine the cost of evaluating $\widehat{\mathrm{kl}}^q_{\mathrm{cal}}$ within each BO iteration. On Lotka–Volterra task, a single evaluation of $\widehat{\mathrm{kl}}^q_{\mathrm{cal}}$ takes an average of 2.64 seconds for $N = 103$ validation pairs and 5.23 seconds for $N = 205$, over five seeds. These validation sizes correspond to simulation budgets

Table 1: Computational cost on Lotka–Volterra task in seconds. DRO-NPE (total) includes 10 BO iterations and final retraining. Timings are means ± standard deviations over five random seeds.

n	NPE	DRO-NPE (fixed ε)	DRO-NPE (total)
1024	73 ± 1	93 ± 3	1521 ± 21
2048	128 ± 3	159 ± 7	2465 ± 218
4096	233 ± 4	286 ± 15	4249 ± 742
8192	437 ± 11	535 ± 21	7221 ± 350
16384	826 ± 30	976 ± 57	13138 ± 293

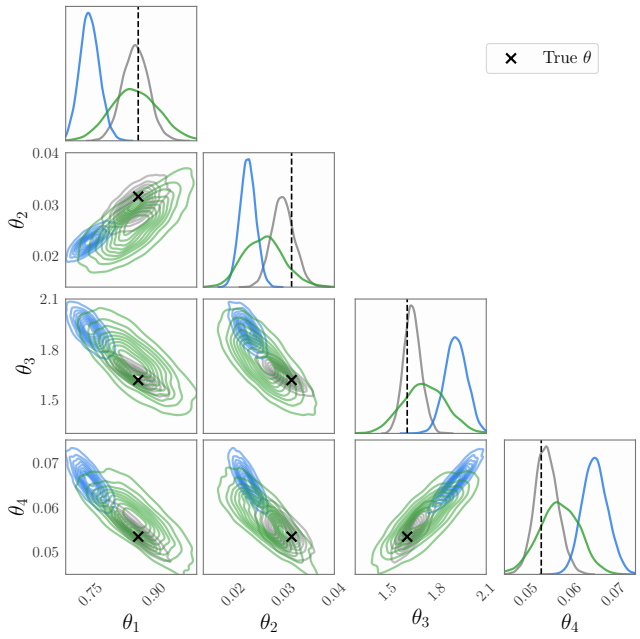


Figure 5: Posterior contour plots of Lotka–Volterra model obtained using NPE (■) and DRO-NPE (■) trained with $n = 1024$ simulations. The reference NPE posterior (■) is obtained using $n = 131072$.

$n = 1024$ and $n = 2048$, respectively. The evaluation cost is dominated by rank computation, which requires repeated density evaluations. Relative to the DRO-NPE training times reported in Table 1, evaluating $\widehat{\mathrm{kl}}^{q}_{\mathrm{cal}}$ incurs only a small additional cost.

D Additional results

In this section, we present additional results related to experiments shown in Section 5.

Additional details and results for Figure 1. The posterior contour plots of Lotka–Volterra parameters in Figure 1 were obtained using $n = 1024$ simulations for both NPE and DRO-NPE. To produce a reference posterior, we train NPE with $n = 131072$ simulations. The contour plots are computed using kernel density estimation from 5000 posterior samples. In Figure 5, we show the results for all four parameters using paired two-dimensional marginals for (θ_1, θ_2) and (θ_3, θ_4). We observe similar results for (θ_3, θ_4) as we did for (θ_1, θ_2): the NPE posterior is biased, while DRO-NPE covers the true parameter and yields a conservative posterior.

Selected values of ε. Figure 6 reports the selected value ε, which we denote as $\varepsilon^\star$, obtained by minimising the validation estimate of $\mathrm{kl}^{q}_{\mathrm{cal}}$ in the benchmarking results of Section 5.1. Across most tasks, $\varepsilon^\star$ decreases as the simulation budget grows, as expected since larger simulation budgets require less conservativeness. The selected values also vary substantially across tasks, suggesting that a single choice of ε determined solely by n is unlikely to be sufficient.

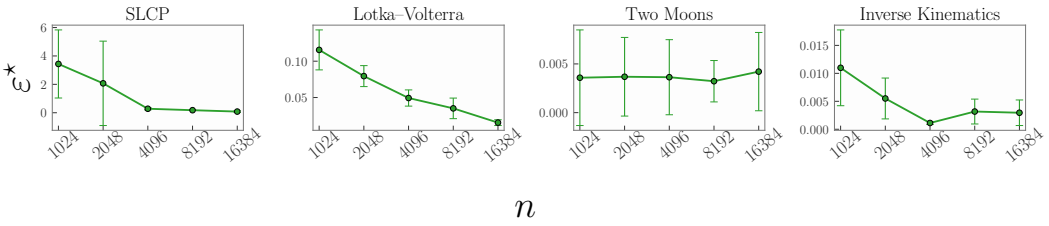
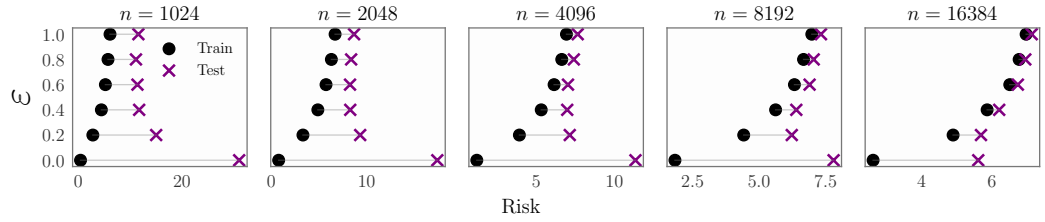
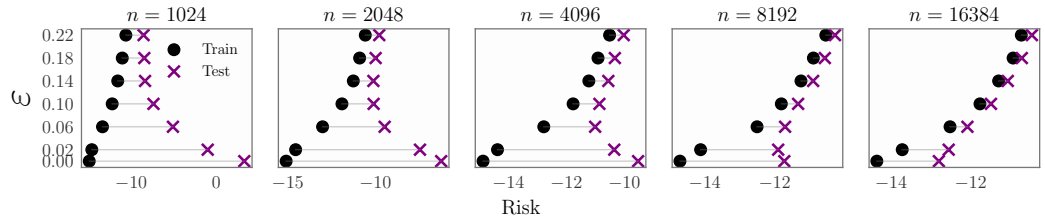


Figure 6: Selected DRO radius $\varepsilon^\star$ across benchmark tasks and simulation budgets. Values are chosen by minimising validation $\mathrm{kl}^q_{\mathrm{cal}}$; means and standard deviations are shown over five random seeds.



(a) SLCP



(b) Lotka–Volterra

Figure 7: Training and test risk across ε for SLCP and Lotka–Volterra at different simulation budgets. Increasing ε reduces the generalisation gap, but overly large values can worsen test risk. Means are shown over five random seeds.

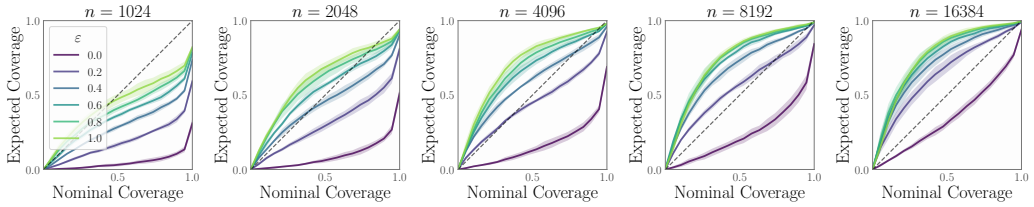
Sensitivity of DRO-NPE to ε. We provide additional sensitivity analyses for the DRO radius ε, extending the results of Section 5.2 to different simulation budgets and including the SLCP task. We report training and test risks in Figure 7, coverage curves in Figure 8, and the behaviour of absolute miscoverage and $\mathrm{kl}^q_{\mathrm{cal}}$ in Figure 9.

As shown in Figure 7, increasing ε consistently reduces the generalisation gap across tasks and simulation budgets. For small n, this reduction is accompanied by improved test risk, whereas for larger n, overly large ε can worsen test risk, indicating excessive conservativeness.

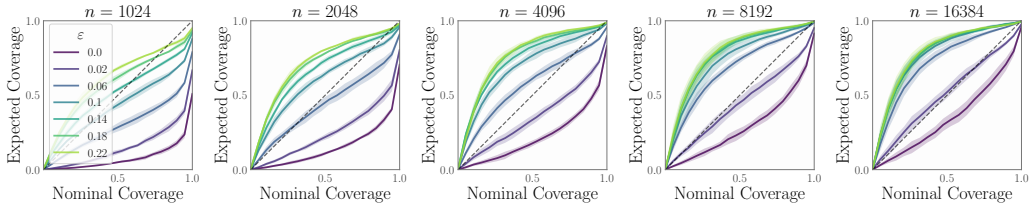
Figure 8 shows the corresponding coverage behaviour. Increasing ε generally moves coverage curves upward, making the posterior more conservative. The value of ε that yields coverage closest to the diagonal decreases as n increases, consistent with the intuition that larger simulation budgets require less regularisation.

Finally, Figure 9 compares absolute miscoverage at $\alpha = 0.05$ with $\mathrm{kl}^q_{\mathrm{cal}}$ for different ε. Absolute miscoverage decreases nearly monotonically with ε, favouring increasingly conservative posteriors. By contrast, $\mathrm{kl}^q_{\mathrm{cal}}$ is minimised at an intermediate value of ε, making it more informative for selecting a radius that balances coverage and posterior quality.

Choice of hyperparameter-selection metric. We report additional results for selecting ε using validation metrics other than $\mathrm{kl}^q_{\mathrm{cal}}$. Across SLCP and Lotka–Volterra, selecting ε via $\mathrm{kl}^q_{\mathrm{cal}}$ generally yields coverage curves closest to the diagonal while maintaining competitive NLPD. In contrast, absolute miscoverage tends to select overly conservative posteriors, whereas $\mathrm{kl}^{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|}_{\mathrm{cal}}$ often leads to

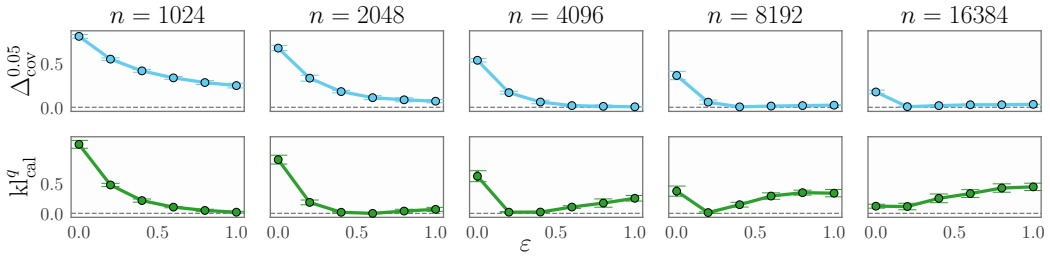


(a) SLCP

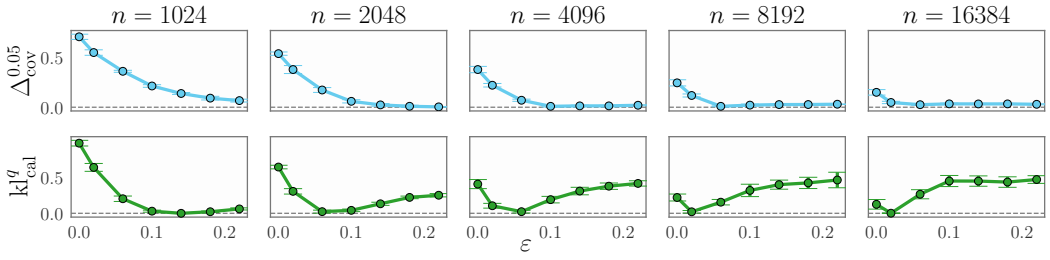


(b) Lotka–Volterra

Figure 8: Coverage curves across ε for SLCP and Lotka–Volterra at different simulation budgets. Larger ε generally yields more conservative posteriors, while the radius closest to the diagonal decreases as the simulation budget grows. Means and standard deviations are shown over five seeds.

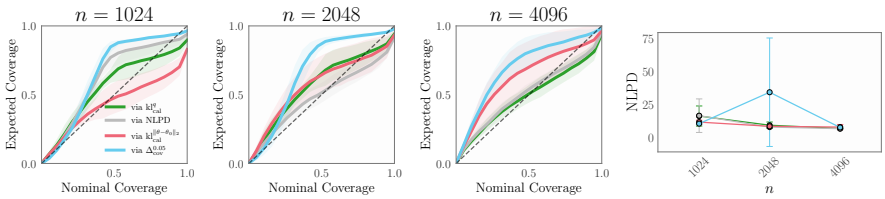


(a) SLCP

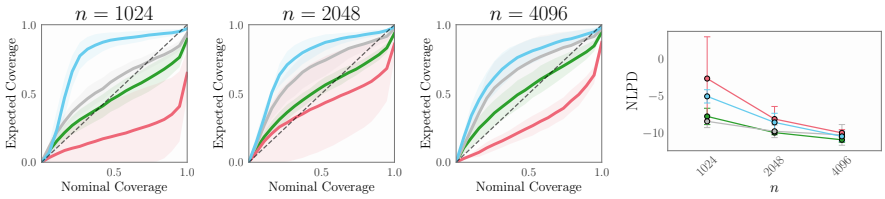


(b) Lotka–Volterra

Figure 9: Absolute miscoverage at $\alpha = 0.05$ (■) and KL-based miscalibration $\text{kl}^{q}_{\text{cal}}$ (■) across ε for SLCP and Lotka–Volterra. Means and standard deviations are shown over five random seeds.



(a) SLCP



(b) Lotka–Volterra

Figure 10: Effect of the validation metric used to select ε on SLCP and Lotka–Volterra. We compare selection by $\mathrm{kl}^q_{\mathrm{cal}}$ (■), NLPD (■), $\mathrm{kl}^{\|\theta-\theta_0\|_2}_{\mathrm{cal}}$ with θ_0 drawn from the prior (■), and absolute miscoverage at $\alpha = 0.05$ (■). Selecting by $\mathrm{kl}^q_{\mathrm{cal}}$ generally gives coverage closest to the diagonal while maintaining competitive NLPD. Means and standard deviations are shown over five seeds.

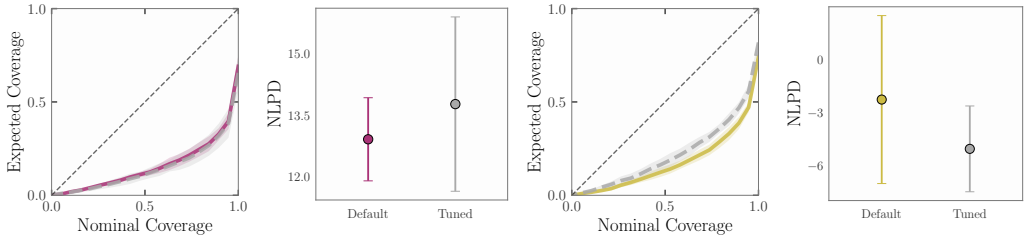


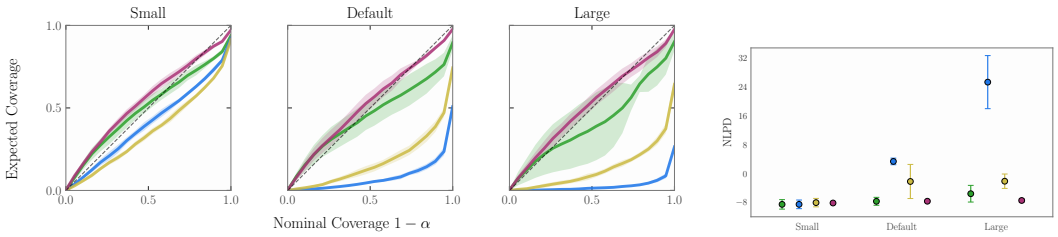
Figure 11: Effect of tuning regularisation hyperparameter using KL-based miscalibration $\mathrm{kl}^q_{\mathrm{cal}}$. Default settings of CR-NPE (■) and Bal-NPE (■) are compared with their tuned variants (■). Coverage curves and NLPD are shown with means and standard deviations over five random seeds.

undercoverage and higher variability across runs. Selection by NLPD also performs well, but tends to produce slightly more conservative coverage curves than $\mathrm{kl}^q_{\mathrm{cal}}$.

Hyperparameter tuning on baseline methods. We now use $\mathrm{kl}^q_{\mathrm{cal}}$ to tune the regularisation hyperparameter of the baseline methods in settings where they exhibit poor coverage in Section 5.1. For CR-NPE, that is Lotka–Volterra with $n = 1024$, and for Bal-NPE, that is SLCP with $n = 1024$.

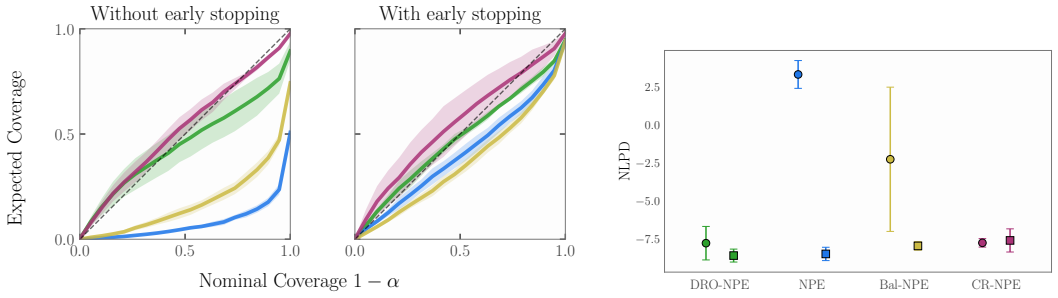
For CR-NPE, we search over $[1, 40]$, following the sensitivity range used by Falkiewicz et al. (2023). Although Falkiewicz et al. (2023) analyse sensitivity using both NLPD and the area under the coverage curve, using two metrics for model selection would require an additional weighting choice; we therefore use $\mathrm{kl}^q_{\mathrm{cal}}$ as a single validation criterion. For Bal-NPE, we search over $[100, 1000]$, since the default $\lambda = 100$ appears too small in Figure 2. The selected values are $\lambda = 29.6 \pm 15.8$ for CR-NPE and $\lambda = 415.4 \pm 303.0$ for Bal-NPE.

Figure 11 shows the results over five random seeds. For CR-NPE, tuning the hyperparameter yields the same coverage results as before, while the NLPD performance degrades slightly, both in terms of the average value and the variance across runs. This is consistent with Falkiewicz et al. (2023), who find that CR-NPE is relatively insensitive to λ. For Bal-NPE, tuning substantially improves NLPD but only modestly improves coverage, leaving the method undercovered. Overall, these results suggest that $\mathrm{kl}^q_{\mathrm{cal}}$ could potentially be useful for tuning other conservative NPE objectives, whilst highlighting that hyperparameter tuning alone does not remove the coverage limitations of these baselines.



(a) Coverage curves (b) NLPD

Figure 12: Effect of normalising-flow architecture on Lotka–Volterra with $n = 1024$. Small, Default, and Large architectures use MLP widths $(32, 32)$, $(128, 128)$, and $(512, 512)$, with 3, 6, and 12 coupling layers, respectively. We compare DRO-NPE (■), standard NPE (■), Bal-NPE (■), and CR-NPE (■). Means and standard deviations are shown over five seeds.



(a) Coverage curves (b) NLPD

Figure 13: Effect of early stopping on Lotka–Volterra with $n = 1024$ and the default architecture. We compare DRO-NPE (■), standard NPE (■), Bal-NPE (■), and CR-NPE (■) with and without early stopping. Means and standard deviations are shown over five seeds.

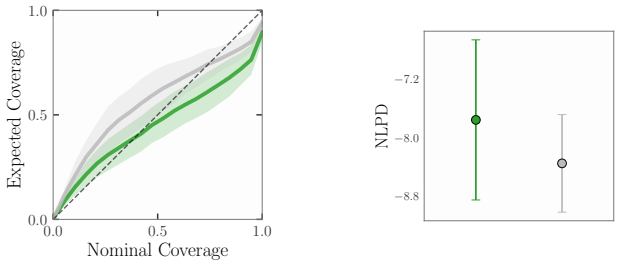
Effect of architecture. An alternative way to reduce generalisation error is to adjust the capacity of the model (normalising flow in our case) used as a posterior density estimator. We therefore investigate the effect of normalising-flow capacity on Lotka–Volterra task with $n = 1024$, using five random seeds. We compare three architectures: Small, with MLP widths $(32, 32)$ and three coupling layers; Default, with widths $(128, 128)$ and six coupling layers; and Large, with widths $(512, 512)$ and twelve coupling layers. Figure 12 reports coverage and NLPD for DRO-NPE, standard NPE, Bal-NPE, and CR-NPE.

Smaller architectures tend to yield lower NLPD and improve coverage in this setting, possibly because reduced capacity limits overfitting in the small-sample regime. Based on the mean results, DRO-NPE and CR-NPE appear less sensitive to architecture than standard NPE, whose performance deteriorates more as capacity increases.

Effect of early stopping. Another common way to reduce generalisation error and mitigate overfitting is early stopping. Although its effect on regularisation is less well understood theoretically, except in relatively simple settings such as linear models (Sonthalia et al., 2024), it may still help in practice by reducing the generalisation gap. To evaluate this empirically, we apply early stopping to all four methods on Lotka–Volterra task with $n = 1024$ and the default architecture. Figure 13 compares coverage and NLPD with and without early stopping over five random seeds.

Early stopping shifts the methods towards more conservative posteriors overall and improves NLPD for all methods except CR-NPE in this setting. The improvement is particularly pronounced for standard NPE. DRO-NPE remains comparatively well calibrated with and without early stopping, while maintaining competitive NLPD.

A more conservative variant of DRO-NPE. In our main experiments, we select ε by minimising $\mathrm{kl}^{q}_{\mathrm{cal}}$. Although this criterion yields good overall coverage, it does not explicitly encourage coverage



(a) Coverage curve (b) NLPD

Figure 14: Comparison of standard DRO-NPE, with ε selected by $\mathrm{kl}^q_{\mathrm{cal}}$ (■), against a more conservative variant with ε selected by $\mathrm{kl}^q_{\mathrm{cal}} - \gamma\tilde{\varepsilon}$ (■), using $\gamma = 0.8$. Results are shown for the Lotka–Volterra task with $n = 1024$. Means and standard deviations over five random seeds are shown.

curves to lie above the diagonal; rather, it penalises deviations from the diagonal in either direction. We therefore consider a more conservative variant of DRO-NPE, obtained by slightly modifying the selection criterion. Specifically, instead of minimising $\mathrm{kl}^q_{\mathrm{cal}}$, we minimise $\mathrm{kl}^q_{\mathrm{cal}} - \gamma\tilde{\varepsilon}$, where

$$\tilde{\varepsilon} = \frac{\log \varepsilon - \log \varepsilon_{\min}}{\log \varepsilon_{\max} - \log \varepsilon_{\min}}$$

is the log-normalised value of ε. The additional term favours larger values of ε, and hence more conservative posteriors. Here, $\gamma \geq 0$ is a user-specified hyperparameter controlling the trade-off between calibration and additional conservativeness.

We illustrate this variant in Figure 14, reporting the resulting coverage and NLPD for $\gamma = 0.8$. The modified criterion yields a more conservative posterior approximation. Interestingly, it also achieves a slightly lower NLPD, suggesting that in this case the value of ε selected using the validation set with $\mathrm{kl}^q_{\mathrm{cal}}$ alone may have been slightly smaller than optimal.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract/introduction are supported by the theoretical and empirical results in Section 4 and Section 5, respectively. Additionally, the contribution of novel uncertainty quantification metrics is in Section 3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Section 6 .

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The complete proofs of the theoretical results and the assumptions are in Appendix A and Appendix B. Some of the main assumptions are listed and discussed in Section 4 as well.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the experimental details required for reproducing the results are in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code to reproduce the experiments is attached as supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the details are in Appendix C, as well as in the attached code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide uncertainty bars across five random seeds for all the experimental results in Section 5 and Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have analysed the computational cost of our proposed method in terms of wall-clock time in Section 5.2, and also discussed memory usage in Appendix C.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All data used in this paper is either synthetically generated through simulation or obtained from open-source repositories, posing on ethical or privacy concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential positive societal impacts of our method in Section 1. We are not aware of any negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not use data or models with high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All software used in this work is open-source and has been properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Reproducible code is provided, along with instructions to replicate all experiments presented in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve any such experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve any such experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard component

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.