FUNCTION-SPACE PRIORS FOR BAYESIAN NEURAL NETWORKS VIA ANCHOR MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian neural networks typically place priors on weights, although inductive properties such as smoothness, periodicity, and geometric structure are more naturally specified in function space. Gaussian processes (GPs) provide a standard way to encode such properties, but matching a finite neural-network prior to a GP over entire functions can lead to infinite KL divergence. We instead match the distributions at a finite set of input locations, or *anchors*. We formulate anchor matching as finding a weight prior whose induced anchor distribution matches the GP while minimizing KL divergence from a base weight prior, and derive the unique solution under finite-KL conditions. Since the network-induced anchor distribution is implicit, we estimate a bounded log-density ratio from base-prior and GP samples, and use it to construct a reweighted prior with adjustable strength. We then use PAC-Bayes bounds to select the reweighting strength and training temperature over a pre-specified finite grid while preserving the risk guarantee up to a logarithmic grid-size penalty. We further characterize conditions for finite-cost anchor matching, show that exact matching can yield infinite KL divergence from Gaussian posteriors, and derive the optimal bounded reweighting and its effect on the PAC-Bayes complexity term. Overall, our framework provides a principled way to transfer structured function-space information into network weight priors while retaining compatibility with posterior inference and finite-sample risk guarantees.

1 INTRODUCTION

Bayesian neural networks (BNNs) place priors on weights (MacKay, 1992; Neal, 1996), whereas inductive properties such as smoothness and periodicity are more naturally specified in function space, for example through Gaussian processes (GPs) (Rasmussen & Williams, 2006). However, the function-space distributions induced by finite BNNs and GPs can be mutually singular, making process-level KL divergence infinite (Burt et al., 2020). Existing function-space variational methods commonly circumvent this difficulty using finite-dimensional evaluations or regularized discrepancies (Sun et al., 2019; Ma & Hernández-Lobato, 2021; Rudner et al., 2022; Wild et al., 2022; Cinquin & Bamler, 2025), incorporating functional information into the posterior objective rather than directly into the neural-network prior. We instead ask whether GP-specified functional information can be transferred directly into the weight prior of a finite BNN and how it affects posterior inference.

We formulate this transfer at finitely many input locations, or *anchors*, where the GP specifies a joint distribution over function values. Our goal is to construct a weight prior whose induced anchor distribution matches this GP target. Let $T(w) = (f_w(a_1), \dots, f_w(a_M))$ denote the network outputs at M anchors. Under the base prior p_0, these outputs induce an implicit distribution $\nu_0 = T_{\#}p_0$, while the GP specifies a Gaussian target distribution γ on the same evaluation space. Because matching γ alone does not uniquely determine a weight prior, we select the matching prior that minimizes KL divergence from p_0. Under finite-KL conditions, the unique solution is

$$\pi_\gamma(dw) = \frac{d\gamma}{d\nu_0}\big(T(w)\big)\, p_0(dw). \tag{1}$$

This information projection (Csiszár, 1975) modifies the base prior only through its anchor outputs: conditional on the anchor outputs, the distribution over weights remains unchanged. Moreover, the KL cost of this transfer is exactly the discrepancy between the GP target and the network's original anchor distribution, $\mathrm{KL}(\pi_\gamma \| p_0) = \mathrm{KL}(\gamma \| \nu_0)$ (Figure 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

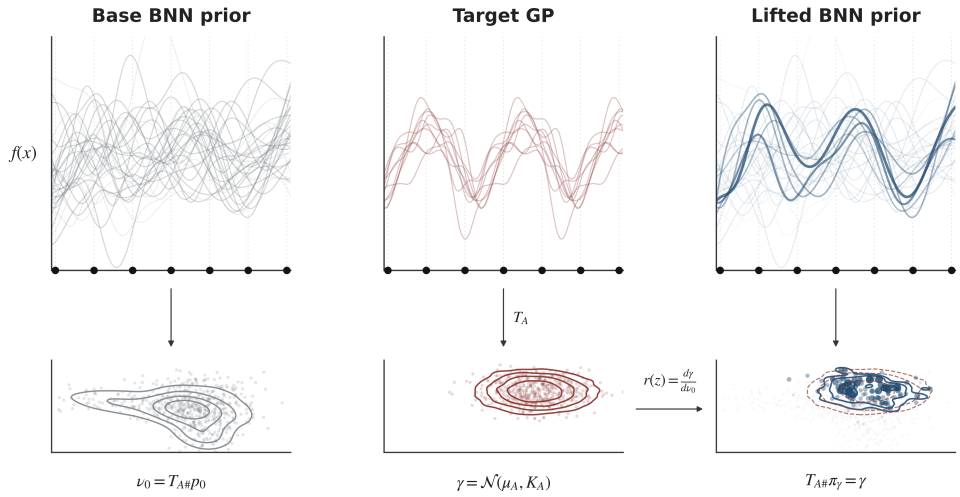

Figure 1: **Prior construction via anchor matching.** Under the finite-KL conditions, reweighting the base BNN prior by $d\gamma/d\nu_0$ matches the GP target distribution γ at the anchors while preserving the base conditional distribution of weights given the anchor outputs. The bottom panels show two-dimensional projections of the corresponding anchor distributions.

The exact construction above is not directly tractable because the network's anchor distribution is implicit, so its density ratio with the GP target cannot be evaluated directly. We therefore estimate the log density ratio from base-prior forward passes and GP samples to obtain a practical reweighting of the base prior. Exact GP matching, however, can be problematic for posterior inference. A simple ReLU example shows that the KL divergence from a Gaussian posterior to the exactly matched weight prior can be infinite (Section 3.2). We therefore constrain the estimated log density ratio to a bounded range, ensuring that any posterior with finite KL divergence from the base prior also has finite KL divergence from the reweighted prior. We further introduce a strength parameter to control the degree of reweighting toward the GP target.

We select the reweighting strength and training temperature by minimizing a PAC-Bayes bound over a pre-specified finite grid. The bound balances empirical loss against the KL divergence from the posterior to the reweighted prior (Germain et al., 2016; Guedj, 2019; Alquier, 2024). Theorem 2 shows that the risk guarantee remains valid after selecting these hyperparameters over the grid, with only a logarithmic penalty in the grid size under uniform confidence allocation.

Our theoretical analysis characterizes when the GP distribution can be matched exactly at the anchors and how prior reweighting modifies the posterior-to-prior KL term in the PAC-Bayes bound. We further derive the closest approximation to the GP target under bounded reweighting, quantifying the trade-off between anchor-level GP matching and maintaining finite posterior-to-prior KL divergence.

2 Methodology

Our goal is to encode the GP-specified distribution at the anchors in a prior over network weights. We first derive the exact minimum-KL construction and characterize its feasibility conditions (Section 2.2). Since the required log-density ratio is generally unavailable, we learn a bounded approximation from simulated network outputs and GP samples (Section 2.3). We then use the estimated log-density ratio to reweight the base prior with adjustable strength and define the corresponding posterior training objective (Section 2.4). Finally, we select the reweighting strength and training temperature over a pre-specified finite grid using a PAC-Bayes bound whose risk guarantee remains valid after selection (Section 2.5).

2.1 Setup

Let $f_w : \mathcal{X} \to \mathbb{R}$ be a neural network with parameters $w \in \Theta \subseteq \mathbb{R}^p$, and let p_0 be a tractable base prior on Θ. Fix an anchor set $A = \{a_1, \ldots, a_M\}$ and define the evaluation map

$$T(w) := \big(f_w(a_1), \ldots, f_w(a_M)\big) \in \mathbb{R}^M.$$

The base prior p_0 induces an anchor-output distribution $\nu_0 := T_{\#}p_0$, which is generally implicit but can be sampled by drawing weights from p_0 and evaluating the network at the anchors. We denote by $\gamma \in \mathcal{P}(\mathbb{R}^M)$ the desired anchor distribution. In our setting, this target is specified by a GP, $\gamma = \mathcal{N}(\mu_A, K_A)$ with $K_A \succ 0$, whose density is available in closed form.

Our goal is to *lift* this prescribed anchor distribution into weight space by constructing a prior π whose induced anchor distribution matches γ, i.e.,

$$T_{\#}\pi = \gamma.$$

2.2 THE MINIMUM-INFORMATION LIFT

The constraint $T_{\#}\pi = \gamma$ alone does not uniquely determine a weight prior: the GP target γ fixes the distribution of the anchor-output vector $z = T(w)$ but leaves the conditional distribution of weights given the anchor outputs, $\pi(\cdot \mid T = z)$, unspecified. By the chain rule for relative entropy (Lemma A1), every feasible π satisfies,

$$\mathrm{KL}(\pi\|p_0) = \underbrace{\mathrm{KL}(\gamma\|\nu_0)}_{\text{unavoidable marginal cost}} + \underbrace{\mathbb{E}_{z\sim\gamma}\,\mathrm{KL}\big(\pi(\cdot \mid T = z)\,\big\|\,p_0(\cdot \mid T = z)\big)}_{\text{additional conditional cost}}. \tag{2}$$

The first term is fixed by the requirement to match the GP target, whereas the second measures any additional change to the conditional distribution of weights given the anchor outputs. Since γ constrains only the anchor-output distribution, such conditional changes introduce structure not specified by the target. We therefore select, among all feasible priors, the one minimizing $\mathrm{KL}(\pi\|p_0)$. By (2), this minimum is attained by changing only the anchor-output distribution while preserving the base-prior conditional distribution of weights given the anchor outputs.

Proposition 1 (Minimum-information lift)**.** *Let $\gamma \in \mathcal{P}(\mathbb{R}^M)$ and consider*

$$\inf_{\pi\in\mathcal{P}(\Theta)} \mathrm{KL}(\pi\|p_0) \qquad \textit{s.t.} \quad T_{\#}\pi = \gamma. \tag{3}$$

If $\gamma \ll \nu_0$ and $\mathrm{KL}(\gamma\|\nu_0) < \infty$, then (3) *has the unique minimizer*

$$\pi_\gamma(dw) = \frac{d\gamma}{d\nu_0}\big(T(w)\big)\,p_0(dw). \tag{4}$$

Moreover, the lift preserves the base-prior conditional distribution of weights given the anchor outputs: $\pi_\gamma(\cdot \mid T = z) = p_0(\cdot \mid T = z)$ for γ-a.e. z.

The minimum KL cost in weight space is therefore exactly the KL divergence between the target and base anchor distributions:

$$\min_{\pi:\, T_{\#}\pi=\gamma} \mathrm{KL}(\pi\|p_0) = \mathrm{KL}(\pi_\gamma\|p_0) = \mathrm{KL}(\gamma\|\nu_0). \tag{5}$$

If $\gamma \not\ll \nu_0$ or $\mathrm{KL}(\gamma\|\nu_0) = \infty$, then the value of (3) *is $+\infty$.*

Proposition 1 shows that an exact lift with finite KL cost exists precisely when $\gamma \ll \nu_0$ and $\mathrm{KL}(\gamma\|\nu_0) < \infty$. Whether these conditions hold depends on the network architecture through the induced anchor distribution ν_0. We next give a sufficient condition for finite-cost lifting for networks that are affine in their final-layer parameters:

$$f_w(x) = v^\top \phi_\theta(x) + b, \qquad w = (\theta, v, b),$$

where $\phi_\theta : \mathcal{X} \to \mathbb{R}^W$ denotes the feature map induced by the preceding layers. Let P_θ denote the marginal prior distribution of θ, and define the anchor feature matrix $\Phi(\theta)$ together with the event E:

$$\Phi(\theta) = [\phi_\theta(a_i)^\top]_{i\le M} \in \mathbb{R}^{M\times W}, \qquad E = \{\mathrm{rank}\,\Phi(\theta) = M\}.$$

On E, the final-layer parameters can generate any vector of function values at the anchors. The following theorem shows that it suffices for this full-rank event to have positive prior probability to guarantee a finite-cost exact lift for any nondegenerate Gaussian target.

Theorem 1 (Feature-rank criterion for finite-cost lifting)**.** *Suppose $M \le W$, the map $\theta \mapsto \Phi(\theta)$ is Borel measurable, and p_0 factorizes over θ and (v, b) with a nondegenerate Gaussian marginal on (v, b). If $P_\theta(E) > 0$, then every nondegenerate Gaussian target γ satisfies*

$$\gamma \ll \nu_0 \qquad \textit{and} \qquad \mathrm{KL}(\gamma\|\nu_0) < \infty,$$

and hence admits an exact lift with finite KL cost. Moreover, if $P_\theta(E) = 1$, then ν_0 has an everywhere-positive Lebesgue density, so ν_0 and γ are mutually absolutely continuous.

Corollary A1 shows that under its stated conditions, non-polynomial real-analytic activations such as tanh yield $P_\theta(E) = 1$. In contrast, for the feedforward ReLU networks considered here, all units in the last hidden layer can be inactive at every anchor with positive prior probability. On this event, $\Phi(\theta) = 0$, implying $P_\theta(E) < 1$. This does not preclude finite-cost exact lifting, which requires only $P_\theta(E) > 0$. Section 3.2 examines how such degenerate ReLU configurations can affect compatibility between the exactly matched prior and Gaussian posterior approximations using a one-unit ReLU example.

2.3 Estimating the log-density ratio

Even when Theorem 1 guarantees a finite-cost exact lift, the construction in Proposition 1 cannot generally be evaluated because the network-induced anchor distribution ν_0 has no tractable density. Define the ideal log-density ratio

$$g^\star(z) := \log \frac{d\gamma}{d\nu_0}(z). \tag{6}$$

Since $e^{g^\star(z)} = d\gamma/d\nu_0(z)$, replacing the unavailable $g^\star$ by a candidate function g yields an approximate exponential reweighting. For any bounded measurable g, define the reweighted anchor distribution

$$\nu_g(dz) := \frac{e^{g(z)}}{\mathbb{E}_{\nu_0}[e^g]}\,\nu_0(dz). \tag{7}$$

We seek to learn g without evaluating the density of ν_0. The Donsker–Varadhan representation (Donsker & Varadhan, 1983; Belghazi et al., 2018) gives

$$J(g) := \mathbb{E}_\gamma[g] - \log \mathbb{E}_{\nu_0}[e^g], \qquad \mathrm{KL}(\gamma\|\nu_0) = \sup_g J(g), \qquad \mathrm{KL}(\gamma\|\nu_g) = \mathrm{KL}(\gamma\|\nu_0) - J(g),$$

where the supremum is over bounded measurable g. Thus, maximizing $J(g)$ is equivalent to minimizing $\mathrm{KL}(\gamma\|\nu_g)$.

Given N_γ samples $z_i^\gamma \sim \gamma$ from the GP target and N_0 weight samples $w_j^0 \sim p_0$ from the base prior, with corresponding anchor outputs $z_j^0 := T(w_j^0) \sim \nu_0$, we approximate $J(g)$ by the empirical objective

$$\widehat{J}(g) = \frac{1}{N_\gamma}\sum_{i=1}^{N_\gamma} g(z_i^\gamma) - \log\left(\frac{1}{N_0}\sum_{j=1}^{N_0} e^{g(z_j^0)}\right). \tag{8}$$

If both ν_0 and γ are nondegenerate Gaussian distributions, $g^\star$ is quadratic in z. For finite neural networks, however, ν_0 need not be Gaussian. We therefore use a quadratic family as a tractable approximation:

$$r_\vartheta(z) := \langle \vartheta, \psi(z)\rangle, \qquad \psi(z) := \big(1, z, (z_i z_j)_{1\le i\le j\le M}\big),$$

where r_ϑ represents a general quadratic function of z. Because r_ϑ is unbounded, its exponential need not be integrable under ν_0. We therefore use the clipped quadratic class:

$$g_\vartheta(z) := \mathrm{clip}\big(r_\vartheta(z), -T_R, T_R\big) - T_R, \qquad \mathcal{G} := \{g_\vartheta : \|\vartheta\|_\infty \le B_\infty\}, \tag{9}$$

where $\mathrm{clip}(u, a, b) := \min\{\max\{u, a\}, b\}$ and $B_\infty > 0$ bounds the coefficients in absolute value. Clipping ensures that every $g \in \mathcal{G}$ is bounded, so ν_g is well-defined and the deviation between $J(g)$ and $\widehat{J}(g)$ can be controlled. The shift by $-T_R$ affects neither ν_g nor $J(g)$, both of which are invariant to additive constants.

Appendix A.2 provides a finite-sample oracle inequality for the fitted bounded log-density ratio, separating approximation, estimation, and optimization errors; Corollary A2 gives an explicit estimation bound for the clipped quadratic class.

2.4 Bounded reweighted priors and posterior inference

We now use the estimated log-density ratio $\widehat{g}$ to construct a weight prior with adjustable reweighting strength. More generally, consider any bounded function $g : \mathbb{R}^M \to [-R_h, 0]$ with $R_h > 0$; for the learned prior, we take $g = \widehat{g}$ and $R_h = 2T_R$. Define $h(w) := g(T(w))$ and, for $\lambda \ge 0$,

$$\pi_\lambda(dw) := \frac{e^{\lambda h(w)}}{Z_\lambda}\,p_0(dw), \tag{10}$$

where $Z_\lambda := \mathbb{E}_{p_0}[e^{\lambda h}] = \mathbb{E}_{\nu_0}[e^{\lambda g}]$. The induced anchor distribution $\nu_\lambda := T_\# \pi_\lambda$ satisfies $d\nu_\lambda / d\nu_0 = e^{\lambda g}/Z_\lambda$. Here, the parameter λ controls the strength of reweighting, with $\pi_0 = p_0$. Since h is bounded and nonpositive, $Z_\lambda \in (0, 1]$, and π_λ is mutually absolutely continuous with p_0. Moreover, for any posterior ρ satisfying $\mathrm{KL}(\rho\|p_0) < \infty$,

$$|\mathrm{KL}(\rho\|\pi_\lambda) - \mathrm{KL}(\rho\|p_0)| \le \lambda R_h. \tag{11}$$

Thus, bounded reweighting preserves finiteness of the posterior-to-prior KL divergence: any posterior with finite KL divergence from p_0 also has finite KL divergence from π_λ. For the learned log-density ratio, $\lambda = 1$ yields the anchor distribution $\nu_{\widehat{g}}$ defined in (7), while varying λ controls how strongly the base prior is shifted toward this learned target. If g is replaced by the ideal log-density ratio $g^\star$, then $\lambda = 1$ recovers the exact lift π_γ whenever it has finite KL cost.

Having constructed the reweighted prior π_λ, we now use it as the prior for posterior inference. Let $r_{\mathrm{train}}(w)$ denote the empirical training loss at weights w. For $\beta > 0$, we obtain the posterior by balancing data fit against divergence from π_λ:

$$\mathbb{E}_\rho[r_{\mathrm{train}}(w)] + \frac{1}{\beta}\mathrm{KL}(\rho\|\pi_\lambda),$$

where larger β places less weight on the KL regularization. Proposition A1 gives

$$\mathrm{KL}(\rho\|\pi_\lambda) = \mathrm{KL}(\rho\|p_0) - \lambda\mathbb{E}_\rho[h] + \log Z_\lambda.$$

Since $\log Z_\lambda$ does not depend on ρ, posterior inference is equivalent to minimizing

$$\mathcal{L}_{\mathrm{train}}(\rho; \beta, \lambda) := \mathbb{E}_\rho[r_{\mathrm{train}}(w)] + \frac{1}{\beta}\mathrm{KL}(\rho\|p_0) - \frac{\lambda}{\beta}\mathbb{E}_\rho[h(w)]. \tag{12}$$

Section 2.5 selects (β, λ) from a pre-specified candidate set using a simultaneous PAC-Bayes bound.

2.5 PAC-Bayes Selection

Section 2.4 leaves the reweighting strength λ and training temperature β to be selected. Rather than fixing them in advance, we select among a pre-specified finite set of candidates using a PAC-Bayes bound that balances empirical risk against the posterior-to-prior KL divergence. Because this selection is data-dependent, a guarantee established for each fixed prior and temperature does not by itself justify choosing the pair with the smallest bound. We therefore derive a simultaneous guarantee over the entire candidate set by allocating a share of the confidence level to each candidate. The resulting bound remains valid after selection and, under uniform confidence allocation, incurs only a logarithmic penalty in the number of candidates.

Let $S = (Z_1, \ldots, Z_n)$ be the sample used for posterior inference and candidate selection. The candidate priors are constructed from auxiliary information $\mathcal{U}$, including the anchors, GP parameters, and samples used to estimate the log-density ratio. For a realization u of $\mathcal{U}$, let $\pi_j^{(u)}$ denote the jth candidate prior. Conditional on $\mathcal{U} = u$, suppose that S has distribution Q_u. For a loss ℓ, define

$$r_S(w) := \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i), \qquad R_u(w) := \mathbb{E}_{S' \sim \mathsf{Q}_u}[r_{S'}(w)].$$

For each candidate j, define the exponential-moment term

$$\Psi_j(\beta) := \log \mathbb{E}_{w \sim \pi_j^{(u)}} \mathbb{E}_{S' \sim \mathsf{Q}_u} \exp\{\beta\big(R_u(w) - r_{S'}(w)\big)\}, \tag{13}$$

where S' is an independent draw from Q_u. This term controls deviations of empirical risk below population risk. For $\ell \in [0, 1]$, $\Psi_j(\beta) \le \beta^2/(8n)$ when, conditional on $\mathcal{U} = u$, S consists of i.i.d. observations or is sampled uniformly without replacement from a fixed finite population (Corollary A3).

The following result extends the PAC-Bayes guarantee from a fixed candidate to simultaneous validity over a pre-specified finite family, thereby allowing data-dependent selection within that family.

Theorem 2 (Post-selection PAC-Bayes validity)**.** *Conditionally on $\mathcal{U} = u$, let $S \sim \mathsf{Q}_u$ and let $\{(\pi_j^{(u)}, \beta_j, \alpha_j) : j \in J\}$ be a finite family fixed before observing S, where $\alpha_j > 0$ and $\sum_{j \in J} \alpha_j = 1$. Assume $\delta \in (0, 1)$, $\beta_j > 0$, and that the exponential-moment terms in* (13) *are finite. Then, with conditional probability at least $1 - \delta$, simultaneously for every $j \in J$ and posterior ρ,*

$$\mathbb{E}_\rho[R_u(w)] \le \mathbb{E}_\rho[r_S(w)] + \frac{1}{\beta_j}\left(\mathrm{KL}(\rho\|\pi_j^{(u)}) + \log\frac{1}{\alpha_j\delta} + \Psi_j(\beta_j)\right). \tag{14}$$

In our setting, each candidate j corresponds to a choice of reweighted prior and training temperature, with the prior determined by the estimated log-density ratio and reweighting strength λ. Because Theorem 2 holds simultaneously over all candidates, the same sample S can be used both for posterior inference and for selecting the candidate with the smallest bound. Under uniform confidence allocation, $\alpha_j = 1/|J|$, the additional selection penalty is $\log |J|/\beta_j$. The proof and complete selection protocol are given in Appendix B.

3 Theoretical Analysis

We analyze how anchor-based prior reweighting affects posterior inference. We first identify how reweighting changes the posterior-to-prior KL divergence (Section 3.1), then examine the compatibility of GP matching with Gaussian posteriors (Section 3.2). Finally, we characterize the best GP approximation achievable under bounded reweighting (Section 3.3).

3.1 Anchor-space KL decomposition

We first examine how prior reweighting changes the posterior-to-prior KL divergence for a fixed posterior ρ. Let $\nu_\rho := T_{\#}\rho$ and $\nu_\lambda := T_{\#}\pi_\lambda$ denote the corresponding anchor-output distributions.

Proposition 2 (Anchor-space KL decomposition)**.** *Let $\rho \in \mathcal{P}(\Theta)$ satisfy $\mathrm{KL}(\rho\|p_0) < \infty$. For every $\lambda \geq 0$,*

$$\mathrm{KL}(\rho\|\pi_\lambda) = \mathrm{KL}(\nu_\rho\|\nu_\lambda) + G(\rho), \tag{15}$$

where $G(\rho) := \mathrm{KL}(\rho\|p_0) - \mathrm{KL}(\nu_\rho\|\nu_0) \geq 0$ is independent of λ. Consequently, the change relative to the base-prior KL is

$$\mathcal{S}(\lambda) := \mathrm{KL}(\rho\|p_0) - \mathrm{KL}(\rho\|\pi_\lambda) = \mathrm{KL}(\nu_\rho\|\nu_0) - \mathrm{KL}(\nu_\rho\|\nu_\lambda) = \lambda\mathbb{E}_\rho[h] - \log Z_\lambda. \tag{16}$$

Proposition 2 shows that, for a fixed posterior, prior reweighting changes the posterior-to-prior KL divergence only through the anchor-output distributions; the remaining term $G(\rho)$, which captures the discrepancy between the posterior and base-prior conditional distributions given the anchor outputs, is unaffected by λ. Importantly, log-density-ratio estimation aims to bring the prior anchor distribution toward the GP target γ, whereas the PAC-Bayes KL compares the posterior anchor distribution ν_ρ with the reweighted prior anchor distribution ν_λ. Closer GP matching therefore need not reduce the posterior-to-prior KL divergence. When posterior inference is performed separately under each candidate prior, ρ also depends on the prior, so $G(\rho)$ can change as well.

3.2 GP matching and Gaussian posteriors

Section 2.2 establishes conditions under which exact anchor matching can be achieved at finite KL cost from the base prior. We now examine whether the exactly matched prior is also compatible with Gaussian posterior approximations at finite posterior-to-prior KL. Consider the one-unit ReLU network

$$f_w(x) = v\,\mathrm{ReLU}(ux + c) + b, \qquad w = (u, c, v, b) \sim p_0 = \mathcal{N}(0, I_4),$$

with anchors $\{-1, 1\}$. Under p_0, the event $c < -|u|$ has probability $1/4$ and implies $f_w(-1) = f_w(1) = b$. A nondegenerate Gaussian target on $\mathbb{R}^2$ assigns zero probability to equal anchor outputs, so any prior that exactly matches this target must assign zero probability to this event. In contrast, every nondegenerate Gaussian weight posterior assigns the event positive probability and therefore has infinite KL divergence from the exactly matched prior. Thus, exact anchor matching can turn a finite posterior-to-prior KL under p_0 into an infinite one, even though an exact lift with finite $\mathrm{KL}(\pi_\gamma\|p_0)$ exists in this example (Proposition A2).

This incompatibility is not limited to exact matching or to a fixed posterior. We next consider sequences of priors π_n whose anchor distributions approach the GP target in KL, while allowing a different Gaussian posterior ρ_n for each prior. In the ReLU example above, the set of equal anchor outputs has positive probability under the base prior but zero probability under the GP target. The following theorem shows that, whenever such a set exists, increasingly accurate GP matching forces at least one of the two KL terms involving the Gaussian posterior to diverge.

Theorem 3 (Matching accuracy and Gaussian posteriors)**.** *Let p_0 be a nondegenerate Gaussian prior on $\mathbb{R}^p$. Suppose there exists a measurable set $B \subseteq \mathbb{R}^M$ such that $\gamma(B) = 0$ and $p_0(T \in B) > 0$. For any sequence of weight priors (π_n) and nondegenerate Gaussian posteriors (ρ_n),*

$$\mathrm{KL}(\gamma \| T_{\#}\pi_n) \longrightarrow 0 \quad \Longrightarrow \quad \mathrm{KL}(\rho_n \| p_0) + \mathrm{KL}(\rho_n \| \pi_n) \longrightarrow \infty. \tag{17}$$

In particular, if either KL term remains uniformly bounded, the other must diverge.

Theorem 3 applies to both mean-field and full-covariance Gaussian posteriors. It shows that, even when the Gaussian posterior is allowed to vary with the prior, increasingly accurate GP matching can cause the relevant KL terms to diverge. This motivates bounded reweighting (Section 2.4), which relaxes exact GP matching while preserving compatibility with posterior inference. In particular, for every fixed λ, (11) guarantees that $\mathrm{KL}(\rho\|\pi_\lambda) < \infty$ whenever $\mathrm{KL}(\rho\|p_0) < \infty$, keeping the posterior-to-prior KL finite for both posterior inference and PAC-Bayes risk bounds.

3.3 Optimal bounded matching

For the priors in Section 2.4, the ratio between any two reweighting factors is at most $e^{\lambda R_h}$. We now characterize the closest match to γ when this ratio is bounded by e^K, for a prescribed $K \geq 0$. Write a candidate anchor distribution as $\nu(dz) = s(z)\nu_0(dz)$, where $s : \mathbb{R}^M \to [0, \infty)$ is a measurable reweighting factor. The normalization $\mathbb{E}_{\nu_0}[s] = 1$ ensures that ν is a probability distribution. We require s to lie between a and $e^K a$ for some $a > 0$, giving the class

$$\mathcal{V}_K := \left\{ s\nu_0 : \mathbb{E}_{\nu_0}[s] = 1, \quad a \leq s \leq e^K a \ \nu_0\text{-a.e. for some } a > 0 \right\}.$$

Under the reweighting construction of Section 2.4, allowing all measurable functions $g : \mathbb{R}^M \to [-K, 0]$ generates exactly $\mathcal{V}_K$. Our prior family satisfies this restriction with $K = \lambda R_h$. The smallest achievable discrepancy from the GP target under this constraint is

$$F(K) := \inf_{\nu \in \mathcal{V}_K} \mathrm{KL}(\gamma\|\nu). \tag{18}$$

For the ReLU example in Section 3.2, $F(K) > 0$ for every finite K, so exact GP matching cannot be achieved under any finite reweighting budget. Motivated by density-ratio-constrained optimization (Fauß & Zoubir, 2018), we characterize the optimal bounded reweighting as follows.

Theorem 4 (Optimal bounded matching)**.** *Assume $\gamma \ll \nu_0$ and $\mathrm{KL}(\gamma\|\nu_0) < \infty$, and write $r = d\gamma/d\nu_0$. The optimal reweighting is obtained by clipping r between two thresholds whose ratio is e^K, chosen so that the result integrates to one. Specifically, there exists $a_K \in [e^{-K}, 1]$ such that*

$$s_K(z) = \mathrm{clip}\big(r(z), a_K, e^K a_K\big), \qquad \mathbb{E}_{\nu_0} s_K = 1. \tag{19}$$

The distribution $\nu_K^\star = s_K \nu_0$ uniquely minimizes $\mathrm{KL}(\gamma\|\nu)$ over $\mathcal{V}_K$. For the corresponding weight prior $\pi_K^\star(dw) = s_K(T(w))p_0(dw)$ and any posterior ρ with $\mathrm{KL}(\rho\|p_0) < \infty$, the change in KL divergence is bounded as follows:

$$-\log(e^K a_K) \leq \mathrm{KL}(\rho\|\pi_K^\star) - \mathrm{KL}(\rho\|p_0) \leq -\log a_K \leq K. \tag{20}$$

Theorem 4 characterizes the optimal GP approximation over all measurable bounded reweightings and provides a reference for the clipped quadratic approximation in Section 2.3.

4 Experiments

We evaluate anchor matching from prior construction to downstream prediction. We first test whether it recovers the prescribed GP distribution at the anchors, then whether this information transfers to posterior predictions in regions without observations, and finally whether it improves predictive performance on regression tasks. We also evaluate PAC-Bayes selection of the reweighting strength. Experimental details and additional results are provided in Appendix C.

4.1 Recovering the GP anchor distribution

We test whether anchor matching recovers a prescribed GP covariance more accurately than simpler modifications of the base prior. We use a one-hidden-layer ReLU network with 64 fixed hidden units and Gaussian output weights, so the induced anchor distribution is exactly Gaussian, $\nu_0 = N(0, K_0)$. Let $K_0 = \sum_{i=1}^M \kappa_i u_i u_i^\top$, $\kappa_1 \geq \cdots \geq \kappa_M$ be the eigendecomposition of K_0. We define

$$\gamma = N(0, K_A), \qquad K_A := K_0 + 4\kappa_3 \, u_3 u_3^\top,$$

which increases the variance along u_3 from κ_3 to $5\kappa_3$ while leaving all other directions unchanged. We compare three prior constructions:

- **Scalar prior scaling:** $\nu_c^{\text{scale}} = N(0, cK_0)$, $c > 0$.
- **GP-density weighting:** $\nu_\lambda^{\text{GP}}(dz) \propto \exp\Big[\lambda \max\big\{-\frac{1}{2}z^\top K_A^{-1} z,\, -T_E\big\}\Big]\, \nu_0(dz), \lambda \geq 0$.
- **Anchor matching (ours):** Reweight ν_0 using the estimated log-density ratio from Section 2.3, yielding the anchor distribution ν_λ in (10).

Table 1 reports the Gaussian W_2 distance between the target and the mean and covariance induced by each construction. Anchor matching achieves a substantially smaller distance across all five runs.

Table 1: Matching a GP target with increased variance along one principal direction. Gaussian W_2 based on the induced mean and covariance; median [min, max] over five runs; lower is better.

Scalar prior scaling	GP-density weighting	**Anchor matching (ours)**
0.643 [0.558, 0.684]	0.653 [0.562, 0.698]	**0.179 [0.131, 0.312]**

4.2 Posterior prediction in an unobserved interval

We next test whether anchor matching transfers GP information to posterior predictions. We sample a function from a Matérn-5/2 GP and observe noisy values at 64 inputs outside $[-1.5, 1.5]$, leaving the central interval unobserved. We use a two-hidden-layer ReLU network of width 32 and 12 anchors on $[-4, 4]$, and estimate the log-density ratio using the DV objective in Section 2.3. We compare posterior predictions under the base and reweighted priors with the exact GP posterior, focusing on the unobserved interval where the prior has greater influence. For both BNN priors, we use NUTS (Hoffman & Gelman, 2014) to sample from the posterior. We measure discrepancy from the GP posterior by the pointwise one-dimensional Gaussian W_2 distance, summarized over the unobserved interval by D_{GP}. Across 20 simulated datasets, mean D_{GP} decreases from 0.55 under the base prior to 0.26 with anchor matching. Figure 2 illustrates that reweighting brings posterior uncertainty closer to the GP reference in the unobserved interval, demonstrating transfer of GP information to posterior predictions. We also include fSVI (Rudner et al., 2022) and GFSVI (Cinquin & Bamler, 2025) as function-space variational baselines using the same network architecture and GP target. Both are substantially under-dispersed in the gap, illustrating the effect of the inference approximation on posterior uncertainty. Additional details and results are provided in Appendix C.2.

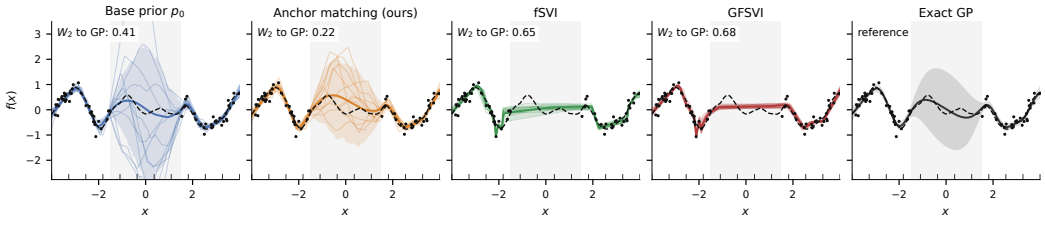

Figure 2: Posterior predictions for one simulated dataset. Dots are observations, the dashed curve is the true function, and shading marks the unobserved interval. Curves and bands show posterior means and 90% credible intervals for latent f. Panel labels report Gaussian W_2.

4.3 Downstream prediction and PAC-Bayes selection

Finally, we examine whether anchor matching improves downstream predictive performance and evaluate the risk bounds obtained after PAC-Bayes selection. We first evaluate predictive negative log-likelihood (NLL) on three UCI regression tasks: Concrete, Energy, and Yacht. We compare anchor matching with the same BNN under its base prior and with the function-space variational baselines fSVI and GFSVI. The target is a zero-mean ARD Matérn-$5/2$ GP fitted on a calibration subset and evaluated at 12 anchors (Appendix C.3). Table 2 shows that anchor matching achieves the lowest mean NLL on all three tasks, with an average rank of 1.00. Because the architecture and posterior family are unchanged from the base-prior model, this comparison isolates the effect of the function-space-informed prior.

Table 2: Predictive NLL on regression tasks. Mean ± standard deviation over three runs.

Method	Concrete	Energy	Yacht	Avg. rank
Base prior	3.022 ± 0.069	0.755 ± 0.039	0.732 ± 0.152	2.00
fSVI	3.050 ± 0.140	0.770 ± 0.099	0.770 ± 0.267	3.33
GFSVI	3.083 ± 0.129	0.756 ± 0.080	1.041 ± 0.259	3.67
Anchor matching (ours)	$\mathbf{3.016 \pm 0.063}$	$\mathbf{0.688 \pm 0.063}$	$\mathbf{0.568 \pm 0.174}$	**1.00**

We evaluate the risk bounds obtained after PAC-Bayes selection in a controlled study on Yacht and Diabetes. For each dataset, we fit a pre-specified grid of candidates indexed by the reweighting strength and training temperature, and compare PAC-Bayes selection with the base prior and validation-based selection over the same candidates (Appendix C.4). Table 3 summarizes the resulting PAC-Bayes risk bounds over seven matched splits. PAC-Bayes selection yields lower mean bounds than the base prior and validation-based selection, as expected from directly minimizing the bound. More importantly, the selected bounds are nonvacuous in all seven splits for both datasets, showing that the post-selection guarantee in Theorem 2 can also be informative in practice.

Table 3: PAC-Bayes bounds after candidate selection. Mean ± standard error over seven matched splits; lower is better. The final column reports the number of nonvacuous splits for Yacht / Diabetes.

Prior / selection	Yacht	Diabetes	Nonvacuous splits
Base prior	0.939 ± 0.049	0.506 ± 0.075	$6/7$ / $7/7$
Anchor matching / Validation	0.942 ± 0.058	0.530 ± 0.072	$6/7$ / $7/7$
Anchor matching / PAC-Bayes	$\mathbf{0.878 \pm 0.013}$	$\mathbf{0.432 \pm 0.004}$	$\mathbf{7/7}$ / $\mathbf{7/7}$

5 Related Work

Functional priors for finite networks. Functional variational inference and finite measurement-set objectives regularize posterior function values directly (Sun et al., 2019; Rudner et al., 2022; Cinquin & Bamler, 2025). Other work uses prescribed function-space targets, including GPs, to construct weight-space priors for finite neural networks (Flam-Shepherd et al., 2017; Tran et al., 2022), with ridgelet and Mercer constructions providing related connections (Matsubara et al., 2021; Alberts & Bilionis, 2026). We instead fix both the architecture and a base weight prior and minimally modify the prior to match a prescribed distribution at finite anchors. Unlike Gaussian density penalties on network evaluations (Rudner et al., 2023; Kozyrskiy et al., 2023), our target-to-base density ratio accounts for the anchor distribution already induced by the base prior.

Marginally specified priors and finite-network lifting. Replacing an induced marginal while preserving the original conditional distribution is classical in marginally specified priors (Poole & Raftery, 2000; Kessler et al., 2015; Goudie et al., 2019), including simulation-based treatment of an unavailable induced marginal (Kessler et al., 2015). We apply this principle to finite BNNs and analyze feasibility, posterior compatibility, and the effect of reweighting on the posterior-to-prior KL. We further characterize optimal GP matching under bounded reweighting.

PAC-Bayes selection. PAC-Bayes optimization and selection over pre-specified candidate families are established tools for finite-sample neural-network risk bounds (Catoni, 2007; Dziugaite & Roy, 2017; Thiemann et al., 2017; Pérez-Ortiz et al., 2021; Lotfi et al., 2022). We integrate this framework with function-space-informed prior construction to select among anchor-matched priors and characterize how reweighting affects the posterior-to-prior KL term.

6 Conclusion

We studied how function-space information can be incorporated directly into the weight prior of a finite Bayesian neural network. By restricting GP matching to finitely many anchors, we obtain the unique minimum-KL modification of a base prior that reproduces the prescribed anchor distribution. We characterize the feasibility of exact matching, its implications for posterior-to-prior KL divergence and Gaussian posterior approximations, and optimal matching under bounded reweighting. In practice, we estimate a bounded log-density ratio from base-prior and GP samples and use PAC-Bayes bounds to select the reweighting strength and training temperature. Experiments show that anchor matching better reproduces the target GP, transfers GP information to posterior predictions, and improves predictive NLL, while PAC-Bayes selection provides informative risk bounds after data-dependent selection. A natural direction for future work is to study whether increasing the number of anchors with the sample size allows the BNN posterior to achieve GP-like contraction rates.

AI USE STATEMENT

Generative AI tools assisted with writing, LaTeX editing, and reviewer-style critique. The authors are responsible for the final manuscript, mathematical claims, implementation, and reported results.

REPRODUCIBILITY STATEMENT

An anonymized implementation of the proposed method and experiments is available at `https://anonymous.4open.science/r/anchor-matching-80D2`. Appendix A contains proofs and additional theoretical details for the prior-construction and analysis results. Appendix B provides the PAC-Bayes selection proof and numerical certification protocol, and Appendix C gives the experimental details needed to reproduce the reported results.

REFERENCES

Alex Alberts and Ilias Bilionis. Bayesian neural networks with interpretable priors from Mercer kernels. *Computer Methods in Applied Mechanics and Engineering*, 455:118896, 2026. doi: 10.1016/j.cma.2026.118896.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024. doi: 10.1561/2200000100.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41, 2016. URL `https://jmlr.org/papers/v17/15-290.html`.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015. doi: 10.3150/14-bej605.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540. PMLR, 2018. URL `https://proceedings.mlr.press/v80/belghazi18a.html`.

David R. Burt, Sebastian W. Ober, Adrià Garriga-Alonso, and Mark van der Wilk. Understanding variational inference in function-space. *arXiv preprint arXiv:2011.09421*, 2020. URL `https://arxiv.org/abs/2011.09421`.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, 2007. doi: 10.1214/074921707000000391.

Tristan Cinquin and Robert Bamler. Well-defined function-space variational inference in Bayesian neural networks via regularized KL-divergence. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 752–776. PMLR, 2025. URL `https://proceedings.mlr.press/v286/cinquin25a.html`.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. IV. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983. doi: 10.1002/cpa.3160360204.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017. URL `https://auai.org/uai2017/proceedings/papers/173.pdf`.

Michael Fauß and Abdelhak M. Zoubir. On the minimization of convex functionals of probability distributions under band constraints. *IEEE Transactions on Signal Processing*, 66(6):1425–1437, 2018. doi: 10.1109/TSP.2017.2788427.

Daniel Flam-Shepherd, James Requeima, and David Duvenaud. Mapping Gaussian process priors to Bayesian neural networks. In *NIPS Workshop on Bayesian Deep Learning*, 2017. URL https://bayesiandeeplearning.org/2017/papers/65.pdf.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/84d2004bf28a2095230e8e14993d398d-Abstract.html.

Robert J. B. Goudie, Anne M. Presanis, David Lunn, Daniela De Angelis, and Lorenz Wernisch. Joining and splitting models with Markov melding. *Bayesian Analysis*, 14(1):81–109, 2019. doi: 10.1214/18-ba1104.

Robert M. Gray. *Entropy and Information Theory*. Springer, 2nd edition, 2011. doi: 10.1007/978-1-4419-7970-4.

Benjamin Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019. URL https://arxiv.org/abs/1901.05353.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

David C. Kessler, Peter D. Hoff, and David B. Dunson. Marginally specified priors for non-parametric Bayesian estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):35–58, 2015. doi: 10.1111/rssb.12059.

Bogdan Kozyrskiy, Dimitrios Milios, and Maurizio Filippone. Imposing functional priors on Bayesian neural networks. In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pp. 450–457. SCITEPRESS, 2023. doi: 10.5220/0011742900003411.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. PAC-Bayes compression bounds so tight that they can explain generalization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31459–31473, 2022. doi: 10.52202/068431-2281. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/cbeec55c50c3367024bafab2438a021b-Abstract-Conference.html.

Chao Ma and José Miguel Hernández-Lobato. Functional variational inference based on stochastic process generators. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21795–21807, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b613e70fd9f59310cf0a8d33de3f2800-Abstract.html.

David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992. doi: 10.1162/neco.1992.4.3.448.

Takuo Matsubara, Chris J. Oates, and François-Xavier Briol. The ridgelet prior: A covariance function approach to prior specification for Bayesian neural networks. *Journal of Machine Learning Research*, 22(157):1–57, 2021. URL https://jmlr.org/papers/v22/20-1300.html.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, 1996. doi: 10.1007/978-1-4612-0745-0.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021. URL https://jmlr.org/papers/v22/20-879.html.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8: 143–195, 1999. doi: 10.1017/s0962492900002919.

David Poole and Adrian E. Raftery. Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255, 2000. doi: 10.1080/01621459.2000.10474324.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. URL https://gaussianprocess.org/gpml/.

Tim G. J. Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractable function-space variational inference in Bayesian neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22686–22698, 2022. doi: 10.52202/068431-1648. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8ea50bf458f6070548b11babbe0bf89b-Abstract-Conference.html.

Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Function-space regularization in neural networks: A probabilistic perspective. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29275–29290. PMLR, 2023. URL https://proceedings.mlr.press/v202/rudner23a.html.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian neural networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.05779.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pp. 466–492. PMLR, 2017. URL https://proceedings.mlr.press/v76/thiemann17a.html.

Ba-Hien Tran, Simone Rossi, Dimitrios Milios, and Maurizio Filippone. All you need is a good functional prior for Bayesian deep learning. *Journal of Machine Learning Research*, 23(74):1–56, 2022. URL https://jmlr.org/papers/v23/20-1340.html.

Veit David Wild, Robert Hu, and Dino Sejdinovic. Generalized variational inference in function spaces: Gaussian measures meet Bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3716–3730, 2022. doi: 10.52202/068431-0269. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/18210aa6209b9adfc97b8c17c3741d95-Abstract-Conference.html.

A THEORETICAL DETAILS AND PROOFS

This appendix provides proofs and additional theoretical details following the construction and analysis in the main text. The selection theorem and its numerical evaluation are treated together in Appendix B. Throughout, $\Theta \subseteq \mathbb{R}^p$ is a Borel set and $T : \Theta \to \mathbb{R}^M$ is Borel measurable, so regular conditional distributions given T exist.

A.1 EXACT MATCHING AND THE FEATURE-RANK CONDITION

The minimum-KL construction in Section 2.2 follows from decomposing changes in the anchor distribution and the conditional distribution of weights.

Lemma A1 (Chain rule for relative entropy)**.** *Let P, Q be probability measures on a standard Borel space and let T be a Borel map into another standard Borel space. If $T_\# P \ll T_\# Q$, then*

$$\mathrm{KL}(P\|Q) = \mathrm{KL}(T_\# P\|T_\# Q) + \int \mathrm{KL}\big(P(\cdot \mid T = z)\|Q(\cdot \mid T = z)\big)\, T_\# P(dz).$$

If marginal domination fails, $\mathrm{KL}(P\|Q) = +\infty$.

Proof. Let

$$P_z := P(\cdot \mid T = z), \qquad Q_z := Q(\cdot \mid T = z).$$

Suppose first that $P \ll Q$. By disintegration and the Radon–Nikodym theorem, for $T_\# P$-almost every z,

$$P_z \ll Q_z,$$

and

$$\frac{dP}{dQ}(w) = \frac{d(T_\# P)}{d(T_\# Q)}(T(w))\frac{dP_{T(w)}}{dQ_{T(w)}}(w) \qquad P\text{-a.s.}$$

Taking logarithms, integrating with respect to P, and disintegrating the conditional term yields

$$\mathrm{KL}(P\|Q) = \mathrm{KL}(T_\# P\|T_\# Q) + \int \mathrm{KL}(P_z\|Q_z)\, T_\# P(dz).$$

Now suppose $T_\# P \ll T_\# Q$ but $P \not\ll Q$. If $P_z \ll Q_z$ for $T_\# P$-almost every z, then disintegration, together with $T_\# P \ll T_\# Q$, would imply $P \ll Q$, a contradiction. Hence $P_z \not\ll Q_z$ on a set of positive $T_\# P$-measure, so the conditional KL integral is $+\infty$, matching $\mathrm{KL}(P\|Q) = +\infty$.

Finally, if $T_\# P \not\ll T_\# Q$, there exists a measurable set B with $T_\# Q(B) = 0$ and $T_\# P(B) > 0$. Thus

$$Q(T^{-1}(B)) = 0, \qquad P(T^{-1}(B)) > 0,$$

so $P \not\ll Q$ and therefore $\mathrm{KL}(P\|Q) = +\infty$. See also Gray (2011). □

Lemma A2 (Reweighting through the evaluation map)**.** *Let $s : \mathbb{R}^M \to [0, \infty)$ be Borel measurable with $\mathbb{E}_{\nu_0}[s] = 1$. Define $s\nu_0$ by*

$$(s\nu_0)(B) := \int_B s(z)\, \nu_0(dz), \qquad B \subseteq \mathbb{R}^M \text{ Borel}.$$

Then

$$\pi(dw) = s(T(w))p_0(dw)$$

is a probability measure with anchor distribution $s\nu_0$ and satisfies

$$\pi(\cdot \mid T = z) = p_0(\cdot \mid T = z) \qquad \text{for } s\nu_0\text{-a.e. } z.$$

Proof. First,

$$\pi(\Theta) = \int s(T(w))\, p_0(dw) = \int s(z)\, \nu_0(dz) = 1,$$

so π is a probability measure. For any Borel set $B \subseteq \mathbb{R}^M$,

$$\begin{aligned} T_{\#}\pi(B) &= \pi(T^{-1}(B)) \\ &= \int_{T^{-1}(B)} s(T(w))\, p_0(dw) \\ &= \int_B s(z)\, \nu_0(dz) = (s\nu_0)(B), \end{aligned}$$

and hence $T_{\#}\pi = s\nu_0$.

It remains to identify the conditional distribution given T. Let

$$K(z, A) := p_0(A \mid T = z).$$

For any measurable $A \subseteq \Theta$ and Borel $B \subseteq \mathbb{R}^M$,

$$\begin{aligned} \int_B K(z, A)\, (s\nu_0)(dz) &= \int_B p_0(A \mid T = z)\, s(z)\, \nu_0(dz) \\ &= \int_{A \cap T^{-1}(B)} s(T(w))\, p_0(dw) \\ &= \pi\big(A \cap T^{-1}(B)\big). \end{aligned}$$

The identity above, together with the measurability of K, shows that K is a regular conditional distribution of π given T. Hence

$$\pi(\cdot \mid T = z) = p_0(\cdot \mid T = z) \qquad s\nu_0\text{-a.e. } z.$$

□

Proof of Proposition 1. Let π be any feasible prior, so that $T_{\#}\pi = \gamma$, and suppose first that $\gamma \ll \nu_0$. By Lemma A1,

$$\mathrm{KL}(\pi \| p_0) = \mathrm{KL}(\gamma \| \nu_0) + \int \mathrm{KL}\big(\pi(\cdot \mid T = z) \| p_0(\cdot \mid T = z)\big)\, \gamma(dz).$$

Hence every feasible prior satisfies

$$\mathrm{KL}(\pi \| p_0) \geq \mathrm{KL}(\gamma \| \nu_0),$$

and, when $\mathrm{KL}(\gamma \| \nu_0) < \infty$, equality holds if and only if

$$\pi(\cdot \mid T = z) = p_0(\cdot \mid T = z) \qquad \gamma\text{-a.e. } z.$$

Now suppose $\gamma \ll \nu_0$ and $\mathrm{KL}(\gamma \| \nu_0) < \infty$. Set

$$s = \frac{d\gamma}{d\nu_0}, \qquad \pi_\gamma(dw) = s(T(w)) p_0(dw).$$

Since $\mathbb{E}_{\nu_0}[s] = 1$, Lemma A2 gives

$$T_{\#}\pi_\gamma = \gamma, \qquad \pi_\gamma(\cdot \mid T = z) = p_0(\cdot \mid T = z) \quad \gamma\text{-a.e. } z.$$

Therefore π_γ is feasible and attains the lower bound,

$$\mathrm{KL}(\pi_\gamma \| p_0) = \mathrm{KL}(\gamma \| \nu_0).$$

If another feasible prior attains the same value, its conditional KL term must vanish γ-almost everywhere. It therefore has the same marginal γ and the same conditional distributions as π_γ, so disintegration implies that it equals π_γ. Thus the minimizer is unique.

Finally, if $\gamma \not\ll \nu_0$, Lemma A1 implies that every feasible prior has infinite KL divergence from p_0. If instead $\gamma \ll \nu_0$ but $\mathrm{KL}(\gamma \| \nu_0) = \infty$, the same decomposition again gives $\mathrm{KL}(\pi \| p_0) = \infty$ for every feasible π. The infimum over an empty feasible set is also $+\infty$. □

Proof of Theorem 1. Write

$$H_\theta = [\Phi(\theta), \mathbf{1}_M],$$

and let (v, b) have mean m_* and covariance $C_* \succ 0$. Conditional on θ, the anchor-output vector is Gaussian,

$$T(w) \mid \theta \sim \mathcal{N}\big(m(\theta), \Sigma(\theta)\big),$$

where

$$m(\theta) = H_\theta m_*, \qquad \Sigma(\theta) = H_\theta C_* H_\theta^\top.$$

On the event

$$E = \{\operatorname{rank} \Phi(\theta) = M\},$$

the matrix H_θ has full row rank. Since $C_* \succ 0$, it follows that

$$\Sigma(\theta) \succ 0 \qquad \text{on } E.$$

Thus, conditional on every $\theta \in E$, the anchor-output distribution has a nondegenerate Gaussian density.

To obtain a uniform lower bound on a positive-probability subset of E, define

$$E_j = E \cap \left\{ j^{-1} I \preceq \Sigma(\theta) \preceq jI,\ \|m(\theta)\| \le j \right\}.$$

Because $(E_j)_{j \ge 1}$ increases to E and $P_\theta(E) > 0$, there exists j such that $P_\theta(E_j) > 0$.

For $\theta \in E_j$, let

$$q_\theta(z) = (2\pi)^{-M/2} \det\big(\Sigma(\theta)\big)^{-1/2} \exp\left\{ -\frac{1}{2} (z - m(\theta))^\top \Sigma(\theta)^{-1} (z - m(\theta)) \right\}$$

be the conditional Gaussian density. The bounds defining E_j imply

$$\det\big(\Sigma(\theta)\big)^{-1/2} \ge j^{-M/2}$$

and

$$(z - m(\theta))^\top \Sigma(\theta)^{-1} (z - m(\theta)) \le j(\|z\| + j)^2.$$

Hence, for every $\theta \in E_j$,

$$q_\theta(z) \ge (2\pi j)^{-M/2} \exp\left\{ -\frac{j}{2} (\|z\| + j)^2 \right\}.$$

Integrating over $\theta \in E_j$, the density of the absolutely continuous component of ν_0 therefore satisfies, for Lebesgue-a.e. z,

$$p_{\mathrm{ac}}(z) \ge P_\theta(E_j) (2\pi j)^{-M/2} \exp\left\{ -\frac{j}{2} (\|z\| + j)^2 \right\} > 0.$$

Since the nondegenerate Gaussian target γ is absolutely continuous with respect to Lebesgue measure, this positivity implies

$$\gamma \ll \nu_0.$$

Moreover, the lower bound above gives

$$-\log p_{\mathrm{ac}}(z) \le C_j + \frac{j}{2} (\|z\| + j)^2$$

for a finite constant C_j. Because a Gaussian target has finite entropy and finite second moments,

$$\mathrm{KL}(\gamma \| \nu_0) < \infty.$$

Thus every nondegenerate Gaussian target admits a finite-cost exact lift by Proposition 1.

Finally, suppose $P_\theta(E) = 1$. Then ν_0 is a mixture of nondegenerate Gaussian distributions. Its density can be written as

$$p_{\mathrm{ac}}(z) = \int q_\theta(z)\, P_\theta(d\theta),$$

and satisfies $p_{\mathrm{ac}}(z) > 0$ for every z. Since p_{ac} integrates to one, it is finite Lebesgue-almost everywhere; redefining it on a Lebesgue-null set if necessary gives an everywhere finite and positive version. Therefore ν_0 is equivalent to Lebesgue measure, and hence to every nondegenerate Gaussian target γ. Thus $\nu_0 \sim \gamma$. □

The next corollary gives an almost-sure rank condition for analytic activations. For deeper networks it requires that preceding layers keep the anchor representations distinct.

Corollary A1 (Full rank for analytic hidden features)**.** *Suppose $M \le W$, the anchors are distinct, the activation is non-polynomial and real analytic, and the hidden weights and biases have a nondegenerate joint Gaussian prior. If the layers preceding the final hidden layer are almost surely injective on the anchors, then $\Phi(\theta)$ has full row rank almost surely. For one hidden layer, no preceding-layer assumption is needed.*

Proof. Condition on the preceding layers, so that their anchor representations are distinct. The span of ridge functions $\sigma(u^\top x + c)$ is dense on compact sets for a continuous non-polynomial activation (Pinkus, 1999). Their evaluation vectors at these finitely many points therefore span $\mathbb{R}^M$: otherwise a nonzero linear functional would annihilate every such vector and, by density, every vector in $\mathbb{R}^M$. Hence there exist parameters $(u_1, c_1), \dots, (u_M, c_M)$ whose evaluation vectors are linearly independent. Since $W \ge M$, assign them to M hidden units. The determinant of the resulting $M \times M$ evaluation matrix, viewed as a function of these units' weights and biases, is a real-analytic function that is not identically zero. Its zero set is therefore Lebesgue-null. The conditional Gaussian distribution of these parameters is nondegenerate, so the determinant is nonzero almost surely. Since this holds for almost every realization of the preceding layers, $\Phi(\theta)$ has full row rank almost surely. □

A.2 Learning a bounded log-density ratio

Recall that a bounded function g induces $d\nu_g/d\nu_0 = e^g/\mathbb{E}_{\nu_0} e^g$, and define the corresponding population objective by $J(g) := \mathbb{E}_\gamma g - \log \mathbb{E}_{\nu_0} e^g$ (Section 2.3). For bounded g and finite $\mathrm{KL}(\gamma\|\nu_0)$,

$$\mathrm{KL}(\gamma\|\nu_g) = \mathrm{KL}(\gamma\|\nu_0) - \mathbb{E}_\gamma g + \log \mathbb{E}_{\nu_0} e^g = \mathrm{KL}(\gamma\|\nu_0) - J(g).$$

Hence maximizing $J(g)$ over $\mathcal{G}$ is equivalent to minimizing $\mathrm{KL}(\gamma\|\nu_g)$ over the corresponding family of reweighted distributions.

Theorem A1 (Oracle inequality for bounded log-density-ratio estimation)**.** *Assume $\gamma \ll \nu_0$ and $\mathrm{KL}(\gamma\|\nu_0) < \infty$, and let $\widehat{g} \in \mathcal{G}$ satisfy $\widehat{J}(\widehat{g}) \ge \sup_{g\in\mathcal{G}} \widehat{J}(g) - \varepsilon_{\mathrm{opt}}$. Define*

$$D_N := \sup_{g\in\mathcal{G}} |J(g) - \widehat{J}(g)|.$$

For $\delta_{\mathrm{c}} \in (0,1)$, let $\Delta_N(\mathcal{G}, \delta_{\mathrm{c}})$ be any deterministic quantity such that $\Pr\{D_N \le \Delta_N(\mathcal{G}, \delta_{\mathrm{c}})\} \ge 1 - \delta_{\mathrm{c}}$, where the probability is over the samples used to fit g. Then, with probability at least $1 - \delta_{\mathrm{c}}$,

$$\mathrm{KL}\big(\gamma\|\nu_{\widehat{g}}\big) \le \underbrace{\inf_{g\in\mathcal{G}} \mathrm{KL}(\gamma\|\nu_g)}_{\textit{approximation}} + \underbrace{2\Delta_N(\mathcal{G}, \delta_{\mathrm{c}})}_{\textit{estimation}} + \underbrace{\varepsilon_{\mathrm{opt}}}_{\textit{optimization}}. \tag{21}$$

Proof. On $D_N \le \Delta_N$, every $g \in \mathcal{G}$ satisfies

$$J(\widehat{g}) \ge \widehat{J}(\widehat{g}) - \Delta_N \ge \widehat{J}(g) - \varepsilon_{\mathrm{opt}} - \Delta_N \ge J(g) - \varepsilon_{\mathrm{opt}} - 2\Delta_N.$$

Substituting the population identity and taking the infimum over g gives (21). □

For a fit that maximizes $\widehat{J}(g_\vartheta) - \Omega(\vartheta)$ to accuracy $\varepsilon_{\mathrm{opt}}$, with $\Omega \ge 0$, the same argument gives

$$\mathrm{KL}(\gamma\|\nu_{\widehat{g}}) + \Omega(\widehat{\vartheta}) \le \inf_{\vartheta} \{\mathrm{KL}(\gamma\|\nu_{g_\vartheta}) + \Omega(\vartheta)\} + 2\Delta_N + \varepsilon_{\mathrm{opt}},$$

where the infimum is over the coefficient box in (9). This covers a ridge-penalized fit without changing the function class. The optimization error is relative to the global empirical optimum; a local solver's stopping criterion does not by itself bound this term. For the clipped quadratic class in (9), the estimation term in Theorem A1 can be made explicit.

Corollary A2 (Deviation bound for clipped quadratic class)**.** *Let the class in* (9) *be fixed independently of the fitting samples from γ and ν_0, which are i.i.d. within each sample and independent of each other.*

Set $d = 1 + M + \binom{M+1}{2}$, $B = \sqrt{d}\, B_\infty$, *and* $R_h = 2T_R$. *Suppose a deterministic finite constant* S_ψ *satisfies*

$$S_\psi \geq \max_{Q \in \{\gamma, \nu_0\}} \left(\mathbb{E}_Q \|\psi(Z)\|_2^2\right)^{1/2}.$$

Then a valid choice in Theorem A1 is

$$\begin{aligned}\Delta_N = 2BS_\psi \left(N_\gamma^{-1/2} + e^{R_h} N_0^{-1/2}\right) \\ + R_h \sqrt{\frac{\log(4/\delta_c)}{2N_\gamma}} + e^{R_h} \sqrt{\frac{\log(4/\delta_c)}{2N_0}}.\end{aligned} \tag{22}$$

Proof. Clipping is 1-Lipschitz, and a common additive constant does not change expected Rademacher complexity. Thus, for either sampling distribution,

$$\mathfrak{R}_N(\mathcal{G}) \leq \frac{B}{N} \mathbb{E} \left\| \sum_{i=1}^N \varsigma_i \psi(Z_i) \right\|_2 \leq \frac{BS_\psi}{\sqrt{N}},$$

where the ς_i are independent Rademacher signs. Symmetrization and bounded differences bound the uniform deviation of the target mean by $2\mathfrak{R}_{N_\gamma}(\mathcal{G})$ plus its displayed R_h-scaled concentration term. For the base mean, $e^g \in [e^{-R_h}, 1]$; the exponential is 1-Lipschitz on $[-R_h, 0]$ and the logarithm is e^{R_h}-Lipschitz on this range of means. Its deviation is therefore at most

$$e^{R_h} \left(2\mathfrak{R}_{N_0}(\mathcal{G}) + \sqrt{\frac{\log(4/\delta_c)}{2N_0}} \right).$$

Apply each one-sided concentration bound at failure probability $\delta_c/4$ and take a union over the two signs and two samples. □

Finite fourth moments of the anchor outputs ensure that S_ψ is finite. Any preprocessing that defines the class must also be fixed independently of these fitting draws, or the result must be applied conditionally on an independent pilot sample.

A.3 Posterior KL under anchor reweighting

This subsection provides the KL identities used in Sections 2.4 and 3.1. Bounded reweighting changes the posterior-to-prior KL by a finite amount. Because the reweighting function depends only on anchor outputs, this KL difference can be expressed entirely through the corresponding anchor-output distributions.

Proposition A1 (KL under bounded reweighting)**.** *For measurable* $h : \Theta \to [-R_h, 0]$, $\lambda \geq 0$, *and* $\mathrm{KL}(\rho\|p_0) < \infty$,

$$\mathrm{KL}(\rho\|\pi_\lambda) = \mathrm{KL}(\rho\|p_0) - \lambda \mathbb{E}_\rho h + \log Z_\lambda, \qquad |\mathrm{KL}(\rho\|\pi_\lambda) - \mathrm{KL}(\rho\|p_0)| \leq \lambda R_h.$$

Proof. The density $e^{\lambda h}/Z_\lambda$ is positive with bounded logarithm. Integrating $\log(d\rho/d\pi_\lambda) = \log(d\rho/dp_0) - \lambda h + \log Z_\lambda$ proves the identity. Both $-\lambda \mathbb{E}_\rho h$ and $-\log Z_\lambda$ lie in $[0, \lambda R_h]$, giving the absolute bound (11). □

Proof of Proposition 2. Applying Lemma A1 to ρ and p_0 gives

$$\mathrm{KL}(\rho\|p_0) = \mathrm{KL}(\nu_\rho\|\nu_0) + G(\rho),$$

where

$$G(\rho) := \mathbb{E}_{\nu_\rho} \mathrm{KL}\big(\rho(\cdot \mid T = z) \| p_0(\cdot \mid T = z)\big).$$

Since $\mathrm{KL}(\rho\|p_0) < \infty$, both terms are finite.

By Lemma A2, π_λ has the same conditional distributions as p_0 given T, while $\nu_\lambda \sim \nu_0$. Since the first decomposition gives $\nu_\rho \ll \nu_0$, we also have $\nu_\rho \ll \nu_\lambda$. Therefore, applying Lemma A1 to ρ and π_λ gives

$$\mathrm{KL}(\rho\|\pi_\lambda) = \mathrm{KL}(\nu_\rho\|\nu_\lambda) + G(\rho),$$

which proves (15). Subtracting the two decompositions yields

$$\mathrm{KL}(\rho\|p_0) - \mathrm{KL}(\rho\|\pi_\lambda) = \mathrm{KL}(\nu_\rho\|\nu_0) - \mathrm{KL}(\nu_\rho\|\nu_\lambda).$$

Finally, Proposition A1 identifies this difference as $\lambda\mathbb{E}_\rho[h] - \log Z_\lambda$, giving (16). □

A.4 Exact matching and Gaussian posteriors

This subsection establishes the incompatibility discussed in Section 3.2 between Gaussian posterior approximations and priors whose anchor distributions approach the GP target. We first show that the ReLU example admits a finite-cost exact lift. We then quantify how GP matching reduces the prior probability of an output event that retains positive probability under Gaussian posteriors, and show that, even when a different Gaussian posterior ρ_n is used for each prior π_n, the two quantities $\mathrm{KL}(\rho_n\|p_0)$ and $\mathrm{KL}(\rho_n\|\pi_n)$ cannot remain simultaneously bounded.

Proposition A2 (A finite ReLU example)**.** *For $f_w(x) = v\,\mathrm{ReLU}(ux + c) + b$, $w \sim \mathcal{N}(0, I_4)$, and anchors $\{-1, 1\}$, the diagonal $B = \{(s, s) : s \in \mathbb{R}\}$ has $\nu_0(B) = 1/4$. Every nondegenerate Gaussian target γ on $\mathbb{R}^2$ satisfies $\gamma \ll \nu_0$ and $\mathrm{KL}(\gamma\|\nu_0) < \infty$.*

Proof. Outside $uv = 0$, which has probability zero, equal outputs occur precisely when both activations vanish, i.e., $c \le -|u|$. This wedge occupies one quarter of the rotation-invariant Gaussian distribution of (u, c). For feasibility, restrict to the positive-probability event $D = \{3/4 \le u \le 5/4,\ |c| \le 1/4\}$. Conditional on $(u, c) \in D$, the outputs are $(b, sv+b)$, where $s = u+c \in [1/2, 3/2]$, with covariance

$$\Sigma_s = \begin{pmatrix} 1 & 1 \\ 1 & 1+s^2 \end{pmatrix}.$$

These covariances are uniformly bounded and positive definite. Their mixture has a positive density, and convexity in the second argument gives

$$\mathrm{KL}(\gamma\|\nu_0) \le -\log p_0(D) + \mathbb{E}\big[\mathrm{KL}(\gamma\|\mathcal{N}(0, \Sigma_{u+c})) \mid D\big] < \infty.$$

Since $\gamma(B) = 0$, every exact matching prior gives $T^{-1}(B)$ zero mass. Every nondegenerate Gaussian weight posterior gives it positive mass and hence has infinite KL divergence from that prior. □

The same phenomenon extends to the feedforward ReLU networks in Section 2.2. There is an open set of weight configurations on which the biases in the final hidden layer are sufficiently negative that all final hidden activations vanish at every anchor. On this event,

$$T(w) = b\mathbf{1}_M,$$

so the anchor-output vector lies in

$$B := \{t\mathbf{1}_M : t \in \mathbb{R}\}.$$

A nondegenerate Gaussian weight prior assigns positive probability to this open parameter set. When $M > 1$, B has Lebesgue measure zero in $\mathbb{R}^M$, and hence every nondegenerate Gaussian target assigns zero probability to B.

The following bound quantifies the same event argument for an approximate match. Write $\mathrm{kl}(a\|b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$ for binary relative entropy, with its usual boundary conventions.

Proposition A3 (Event-mass bound)**.** *If $\gamma(B) = 0$, $a = \rho(T \in B) > 0$, and $\mathrm{KL}(\gamma\|T_\#\pi) \le \varepsilon < \infty$, then*

$$\mathrm{KL}(\rho\|\pi) \ge \mathrm{kl}\big(a \,\|\, \min\{a, 1 - e^{-\varepsilon}\}\big). \tag{23}$$

Proof. Set $b = \pi(T \in B)$. Data processing gives $\varepsilon \ge \mathrm{kl}(0\|b) = -\log(1-b)$, hence $b \le 1 - e^{-\varepsilon}$. Applying data processing to the weight event also gives $\mathrm{KL}(\rho\|\pi) \ge \mathrm{kl}(a\|b)$. Minimizing the latter over the allowed b proves the claim. □

Proof of Theorem 3. An invertible affine change of coordinates makes $p_0 = \mathcal{N}(0, I_p)$ without changing either KL divergence. For $\rho = \mathcal{N}(m, \Sigma)$,

$$2\mathrm{KL}(\rho\|p_0) = \|m\|^2 + \sum_{i=1}^{p}(t_i - \log t_i - 1),$$

where t_i are the eigenvalues of Σ. Thus a bound $\mathrm{KL}(\rho\|p_0) \le C$ places m in a compact set and bounds all eigenvalues away from zero and infinity. The resulting Gaussian family is compact in total variation: convergent parameters give pointwise density convergence, and Scheffé's lemma gives total-variation convergence.

Let $D = T^{-1}(B)$. The map $\rho \mapsto \rho(D)$ is continuous in total variation. Every Gaussian in this compact family is equivalent to p_0 and has $\rho(D) > 0$, so its minimum a_C is positive. Write $\varepsilon_n = \mathrm{KL}(\gamma\|T_{\#}\pi_n) \to 0$ and $b_n = \pi_n(D) \le 1 - e^{-\varepsilon_n} \to 0$. If the sum in (17) did not diverge, a subsequence would have both KL terms bounded by some $C < \infty$. Along that subsequence, with $a_n = \rho_n(D) \ge a_C$,

$$\mathrm{KL}(\rho_n\|\pi_n) \ge \mathrm{kl}(a_n\|b_n) \ge a_C \log(1/b_n) - \log 2 \longrightarrow \infty,$$

a contradiction. If $b_n = 0$, the corresponding KL is already infinite. □

Thus, under a uniform bound on $\mathrm{KL}(\rho_n\|p_0)$, convergence of the prior anchor distributions to the GP target in KL forces (ρ_n) to fail to be contiguous with respect to (π_n), since $\pi_n(D) \to 0$ while $\inf_n \rho_n(D) > 0$. The divergence of $\mathrm{KL}(\rho_n\|\pi_n)$ in Theorem 3 is a quantitative manifestation of this incompatibility.

A.5 The best approximation under bounded reweighting

This subsection proves the optimal bounded matching result in Section 3.3. We first show that the optimal reweighting is obtained by clipping $r = d\gamma/d\nu_0$ at two thresholds determined jointly by the range constraint and normalization. We then show that any finite reweighting range leaves a strictly positive approximation error whenever the base anchor distribution assigns positive mass to a set that is null under the GP target.

Proof of Theorem 4. Every measurable $g : \mathbb{R}^M \to [-K, 0]$ induces $s = e^g/\mathbb{E}_{\nu_0} e^g \in \mathcal{V}_K$. Conversely, if $a \le s \le e^K a$ and $\mathbb{E}_{\nu_0} s = 1$, then $g = \log s - \log(e^K a) \in [-K, 0]$ induces $s\nu_0$. For $K = 0$, the only feasible density is $s = 1$.

For $k = e^K > 1$, the function $Z(a) = \mathbb{E}_{\nu_0} \mathrm{clip}(r, a, ka)$ is continuous and nondecreasing on $[k^{-1}, 1]$. Since $Z(k^{-1}) \le 1 \le Z(1)$, there exists $a_K \in [k^{-1}, 1]$ such that $Z(a_K) = 1$. Write $a = a_K$ and $s^* = \mathrm{clip}(r, a, ka)$, and partition the space into $L = \{r < a\}$, $I = \{a \le r \le ka\}$, and $U = \{r > ka\}$. For another feasible normalized density s, choose $b > 0$ with $b \le s \le kb$. Then

$$\begin{aligned}\mathbb{E}_{\nu_0}\left[\frac{r}{s^*}(s - s^*)\right] &= \mathbb{E}_{\nu_0}\left[\left(\frac{r}{s^*} - 1\right)(s - s^*)\right] \\ &\le (b - a)\left\{\int_L \left(\frac{r}{a} - 1\right) d\nu_0 + k\int_U \left(\frac{r}{ka} - 1\right) d\nu_0\right\} = 0.\end{aligned}$$

The inequality uses $s \ge b$ on L and $s \le kb$ on U; the coefficient vanishes on I. The bracket equals $a^{-1}\int(r - s^*)d\nu_0 = 0$. Using $\log t \le t - 1$ yields

$$\mathrm{KL}(\gamma\|s\nu_0) - \mathrm{KL}(\gamma\|s^*\nu_0) = -\mathbb{E}_\gamma \log(s/s^*) \ge -\mathbb{E}_{\nu_0}\left[\frac{r}{s^*}(s - s^*)\right] \ge 0.$$

All quantities are finite because normalized feasible densities lie in $[k^{-1}, k]$ and the base target KL is finite.

Equality requires $s = s^*$ γ-almost everywhere. If $r > 0$ ν_0-almost everywhere, this proves uniqueness. Otherwise $\nu_0(\{r = 0\}) > 0$ and normalization forces $\nu_0(U) > 0$: without upper clipping, the positive floor would increase the total mass above one. On U, equality gives $s = s^* = ka$, so $s \le kb$ requires $b \ge a$. Thus $s \ge a = s^*$ on $\{r = 0\}$, and equality of total masses forces equality there as well. The minimizing distribution is unique, even if the threshold parameter is not.

A corresponding optimal bounded reweighting function is

$$g_K^* = \log s_K - \log(e^K a_K) \in [-K, 0].$$

Finally, $\mathrm{KL}(\rho\|\pi_K^*) - \mathrm{KL}(\rho\|p_0) = -\mathbb{E}_\rho \log s_K(T(w))$. The bounds on s_K and $a_K \geq e^{-K}$ give (20). □

Positive error at finite range. If $\gamma(B) = 0$ and $b_0 = \nu_0(B) \in (0, 1)$, then every measurable $g : \mathbb{R}^M \to [-K, 0]$ satisfies

$$\nu_g(B) \geq \frac{e^{-K} b_0}{1 - b_0 + e^{-K} b_0}.$$

Indeed, the unnormalized mass on B is at least $e^{-K} b_0$, whereas that on B^c is at most $1 - b_0$. Binary data processing therefore gives

$$F(K) \geq \log\left(1 + e^{-K} \frac{b_0}{1 - b_0}\right) > 0. \tag{24}$$

For the ReLU example, $b_0 = 1/4$ and the floor is $\log(1 + e^{-K}/3)$. The bound is attainable over distributions: if $\nu_0 = (1 - b_0)\gamma + b_0\xi$ with $\xi(B) = 1$, the reweighting function $g = -K\mathbf{1}_B$ attains equality. This does not imply that the same floor is attained by every BNN or by the fitted clipped quadratic reweighting.

B Selection and Numerical Certification

Theorem 2 provides simultaneous validity over the candidate family and all posteriors, while the reported bounds also involve expectations estimated by Monte Carlo. Proposition A4 accounts for both selection and Monte Carlo uncertainty.

B.1 The selection theorem and the sampling model

Proof of Theorem 2. Condition on $\mathcal{U} = u$. For candidate j, define $F_S(w) = \beta_j(R_u(w) - r_S(w))$. By (13) and Markov's inequality, with failure probability at most $\delta\alpha_j$,

$$\log \mathbb{E}_{\pi_j^{(u)}} e^{F_S} \leq \Psi_j(\beta_j) + \log \frac{1}{\delta\alpha_j}.$$

On this event, the change-of-measure inequality $\mathbb{E}_\rho F_S \leq \mathrm{KL}(\rho\|\pi_j^{(u)}) + \log \mathbb{E}_{\pi_j^{(u)}} e^{F_S}$ holds for every posterior ρ. A union bound over j proves (14), including for a data-selected index and posterior. Integrating over $\mathcal{U}$ preserves the failure probability. This is the conditional form of the standard PAC-Bayes argument (Alquier et al., 2016). □

Corollary A3 (Bounded loss under sampling without replacement)**.** *Let $P = \{z_1, \ldots, z_N\}$ be a fixed population and let S be a uniform sample of size n without replacement. For $\ell(w, z) \in [0, 1]$, Theorem 2 applies with*

$$R_P(w) = \frac{1}{N}\sum_{i=1}^{N} \ell(w, z_i), \qquad \Psi_j(\beta) \leq \frac{\beta^2}{8n}.$$

The same moment bound holds for i.i.d. observations, with population risk equal to expected loss under their common distribution.

Proof. For fixed w, the exponential moment of the without-replacement sample mean is no larger than its with-replacement counterpart (Hoeffding, 1963; Bardenet & Maillard, 2015). Hoeffding's lemma gives $\mathbb{E}_S e^{\beta(R_P(w) - r_S(w))} \leq e^{\beta^2/(8n)}$. Integration over the prior bounds Ψ_j. The i.i.d. case follows directly from the same lemma. □

In Appendix C.4, each run partitions the observations uniformly at random into calibration, validation, certification, and remainder sets C, V, S, and T. Conditionally on C, V, and the prior-construction randomness, S is a uniform sample without replacement from $P \setminus (C \cup V)$, so Corollary A3 applies with this population and $n = |S|$.

B.2 Evaluating the certificate

The bound involves $\mathbb{E}_q r_S$, $\mathbb{E}_q h$, and $\log Z_\lambda$. Because $\mathbb{E}_q h$ enters with coefficient $-\lambda \le 0$, we need upper bounds on the first and third quantities and a lower bound on the second. All three are expectations of bounded variables.

Lemma A3 (Empirical Bernstein bound, Maurer & Pontil 2009)**.** *For $m \ge 2$ i.i.d. variables $X_i \in [0,1]$, let $\bar{X}$ and $\widehat{V}$ be their sample mean and unbiased sample variance. With probability at least $1-\eta$,*

$$\mathbb{E}X_1 \le \bar{X} + \sqrt{\frac{2\widehat{V}\log(2/\eta)}{m}} + \frac{7\log(2/\eta)}{3(m-1)}. \tag{25}$$

A lower bound follows by applying the inequality to $1 - X_i$; other bounded ranges follow by rescaling. Bounds may be intersected with the known range.

Proposition A4 (Numerical certificate)**.** *In Theorem 2, let $\pi_j^{(u)}(dw) \propto e^{\lambda_j h_j(w)} p_{0,j}(dw)$ with $\lambda_j \ge 0$ and $h_j \in [-R_j, 0]$, and suppose $\Psi_j(\beta_j) \le \beta_j^2/(8n)$. Before S is observed, fix weights $\xi_j > 0$ with $\sum_j \xi_j \le 1$ and a split $\delta = \delta_{\mathrm{grid}} + \delta_Z + \delta_r + \delta_h$. For each j, fix a posterior q_j after observing S and before the Monte Carlo evaluation; q_j may depend on S. Use Lemma A3 to compute an upper bound $\hat{r}_j^+$ on $\mathbb{E}_{q_j} r_S$ and a lower bound $\hat{a}_j^-$ on $\mathbb{E}_{q_j} h_j$ from i.i.d. draws of q_j, at failure probabilities $\delta_r \xi_j$ and $\delta_h \xi_j$, and an upper bound $\overline{Z}_j$ on $Z_j = \mathbb{E}_{p_{0,j}} e^{\lambda_j h_j}$ from i.i.d. draws of $p_{0,j}$ independent of the samples used to fit the reweighting function, at failure probability $\delta_Z \xi_j$. Set $\overline{\log Z}_j = \log\min\{1, \overline{Z}_j\}$, and $\overline{\log Z}_j = 0$ if $\lambda_j = 0$. Then, with probability at least $1-\delta$, simultaneously for every $j \in J$,*

$$\mathbb{E}_{q_j} R_u \le B_j := \hat{r}_j^+ + \frac{1}{\beta_j}\Bigg(\mathrm{KL}(q_j \| p_{0,j}) - \lambda_j \hat{a}_j^- + \overline{\log Z}_j + \log\frac{1}{\alpha_j \delta_{\mathrm{grid}}} + \frac{\beta_j^2}{8n}\Bigg). \tag{26}$$

Hence $B_{\hat{j}}$ remains valid for an index $\hat{j}$ selected using these same evaluations. Draws and bounds may be shared across indices.

Proof. By Theorem 2, (14) holds for all j and all posteriors outside an event of probability at most δ_{grid}. Conditionally on S, the frozen posteriors, and the reweighting functions, a union bound shows that all Monte Carlo bounds hold outside an event of probability at most $\sum_j (\delta_r + \delta_h + \delta_Z)\xi_j \le \delta_r + \delta_h + \delta_Z$; this requires no independence between evaluations. On the intersection, Proposition A1 and $\lambda_j \ge 0$ give

$$\mathrm{KL}(q_j \| \pi_j^{(u)}) = \mathrm{KL}(q_j \| p_{0,j}) - \lambda_j \mathbb{E}_{q_j} h_j + \log Z_j \le \mathrm{KL}(q_j \| p_{0,j}) - \lambda_j \hat{a}_j^- + \overline{\log Z}_j,$$

and $\mathbb{E}_{q_j} r_S \le \hat{r}_j^+$. Substituting both into (14) gives (26). □

The candidate family, allocation, and selection rules used in the experiments are given in Appendix C.4.

C Experimental details and supporting results

Throughout, anchor matching fits the quadratic log-density-ratio function of Section 2.3 using the empirical Donsker–Varadhan objective. In Appendices C.2 and C.3, the fitted quadratic is subsequently clipped using simulation-based quantiles. These empirical fits are evaluated directly rather than through the finite-sample guarantee of Theorem A1.

C.1 Matching the GP mean and covariance

The hidden layer has width 64 and ReLU activation; its weights and biases are drawn from $N(0, 1.2^2)$ and $N(0, 0.5^2)$ and then fixed, and output weights follow $N(0, 1.5^2 I/64)$. With 12 evenly spaced

anchors on $[-1, 1]$ and hidden-feature matrix Φ, $K_0 = (1.5^2/64)\Phi\Phi^\top$. Each of five repetitions redraws the hidden layer and all simulation pools.

Scalar scaling searches $c \in \{0.5, 0.6, \ldots, 6.0\}$, and both reweightings search $\lambda \in \{0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 1.2\}$. GP-density weighting uses $T_E = 200$. The quadratic log-density-ratio function is fitted from 25,000 base and 50,000 GP draws with coefficient box 50 and $T_R = 40$. Reweighted means and covariances are estimated from 50,000 fresh base draws with self-normalized importance weights. For mean m and covariance K, we compute W_2 from

$$W_2^2(N(m, K), N(0, K_A)) = \|m\|^2 + \mathrm{tr}\Big(K + K_A - 2(K_A^{1/2} K K_A^{1/2})^{1/2}\Big),$$

which compares first and second moments also when the reweighted distribution is not Gaussian. For each construction, we report the minimum over its prespecified grid.

Why the simpler constructions fall short. Scalar scaling multiplies every eigenvalue of K_0 by the same factor, so it cannot raise the variance along u_3 alone. Under ν_0, the coordinates in the common eigenbasis of K_0 and K_A are independent, and the GP-density weight is even and nonincreasing in each squared coordinate. Reweighting therefore keeps cross-covariances at zero and, by Chebyshev's association inequality, increases no variance; without clipping, the covariance is $(K_0^{-1} + \lambda K_A^{-1})^{-1} \preceq K_0$. The best population distance of GP-density weighting is thus attained at $\lambda = 0$. The exact log ratio has quadratic part $\frac{1}{2} z^\top (K_0^{-1} - K_A^{-1}) z$, which raises the variance along u_3.

C.2 Posterior prediction in an unobserved interval

Data and priors. Each dataset draws a function from a zero-mean Matérn-5/2 GP with unit variance and lengthscale 1 on a 401-point grid over $[-4, 4]$, selects 64 inputs without replacement outside $[-1.5, 1.5]$, and adds Gaussian noise with standard deviation 0.2. This GP is both the anchor target and the exact reference. The network has two hidden ReLU layers of width 32 and a linear output; weights and biases have variances $\sigma_w^2/d_{\mathrm{in}}$ and σ_b^2 with $(\sigma_w, \sigma_b) = (2, 4)$, where d_{in} is the layer input width. Both weight priors use 12 evenly spaced anchors on $[-4, 4]$. We generate 20 datasets.

Reweighting function and strength. The quadratic log-density-ratio function is fitted from 25,000 base and 50,000 GP anchor-output draws with coefficient box 5; a further 25,000 base draws are held out. Using 20,000 additional GP draws and 20,000 held-out base draws, we center the fitted quadratic r at its GP median m and set, with $\widetilde{r} = r - m$,

$$H = \max\{Q_{0.9}^{\mathrm{GP}}(\widetilde{r}),\ -Q_{0.1}^{\mathrm{base}}(\widetilde{r})\}, \qquad g_R(z) = \mathrm{clip}(\widetilde{r}(z), -H, H) - H.$$

The strength minimizes the sliced 2-Wasserstein distance (SW_2; 1,500 resampled vectors, 200 projections) between the reweighted held-out base draws and GP draws over

$$\{0, 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, 1, 1.28, 2.56, 5.12\}.$$

No observed data enter the fitting, clipping, or strength selection used for prior construction.

Inference. Posteriors under both weight priors are sampled with NUTS using two chains of 400 warm-up and 400 retained iterations and target acceptance 0.85. Initial runs with anchor-output $\widehat{R} > 1.01$ or effective sample size below 200 are rerun with four chains of 400 warm-up and 800 retained iterations. All reported runs have $\widehat{R} \le 1.01$ and effective sample size at least 200; predictions use 800 draws from an initial run or 1,000 from a replacement run.

Function-space variational references. We adapt fSVI and GFSVI (Rudner et al., 2022; Cinquin & Bamler, 2025) to the same network, target GP, and anchors, which serve as context inputs X_c, with mean-field Gaussian weights. With $m_c = f_\mu(X_c)$ and $Q_c = J_c D J_c^\top$, where J_c is the output Jacobian at the variational mean and D the diagonal weight covariance, both add

$$\frac{1}{n}\mathrm{KL}(N(m_c, Q_c + \epsilon I) \,\|\, N(\mu_c, K_c + \epsilon I))$$

to the mean expected negative log-likelihood, with $\epsilon = 0$ for fSVI and $\epsilon = 10^{-4}$ for GFSVI; fSVI thus uses the target GP in place of its weight-induced function prior. The Jacobian is recomputed at

every step without being differentiated, and the likelihood and predictions use the nonlinear network. Both use the $(\sigma_w, \sigma_b) = (\sqrt{2}, 0.3)$ scale specification, 20,000 Adam steps at learning rate 3×10^{-3}, eight likelihood particles, initial variational standard deviation 0.05, and 1,000 predictive draws, on ten independently generated datasets.

Metrics and results. With $\mu(x), s(x)$ the posterior mean and standard deviation of the latent BNN output and $m_{\mathrm{GP}}(x), s_{\mathrm{GP}}(x)$ their exact-GP counterparts, we report on the gap grid G

$$D_{\mathrm{GP}} = \left[\frac{1}{|G|} \sum_{x \in G} \left\{ (\mu(x) - m_{\mathrm{GP}}(x))^2 + (s(x) - s_{\mathrm{GP}}(x))^2 \right\} \right]^{1/2}, \tag{27}$$

the root-mean-square pointwise Gaussian W_2 distance; it does not compare joint laws across inputs. Gap NLL scores noisy held-out observations in the gap under the posterior mixture $L^{-1} \sum_{\ell=1}^{L} N(y \mid f_{w_\ell}(x), 0.2^2)$, and in closed form for the GP.

Table 4: Repeated posterior comparisons: mean (standard error across datasets). SD ratio is the gap-averaged latent posterior standard deviation divided by that of the exact GP. Bold marks the best non-reference value in each metric (closest to one for SD ratio).

Prior / method	Inference	N	$D_{\mathrm{GP}} \downarrow$	Gap NLL$\downarrow$	SD ratio
Base prior	NUTS	20	.550 (.092)	1.18 (.09)	1.34
Anchor matching (ours)	NUTS	20	**.256 (.034)**	**1.09 (.11)**	**1.02**
fSVI	Variational	10	.686 (.052)	3.13 (1.26)	.33
GFSVI	Variational	10	.700 (.036)	5.03 (1.54)	.19
Exact GP	Exact	20	0	1.03 (.09)	1

Anchor matching has smaller D_{GP} than the base prior on 17 of the 20 datasets. The much smaller SD ratios of fSVI and GFSVI indicate under-dispersed mean-field uncertainty in the gap. Figures 2 and 3 use the strength selected by the prior-only criterion above.

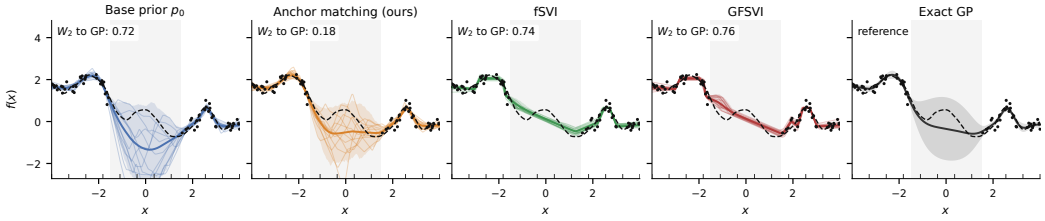

Figure 3: Another dataset under the protocol and panel layout of Figure 2.

C.3 UCI REGRESSION

Data and splits. We use Concrete, Energy (heating load), and Yacht. For each dataset, approximately $72/18/10\%$ of the observations are assigned to fitting, calibration, and test subsets, with the same splits used by all methods. Standardization uses the fitting subset. The calibration subset determines the target GP and method-specific hyperparameters but does not enter the BNN likelihood; the test set is accessed only after all selections are fixed. For the predictive comparison in Table 2, all methods use the same calibration-based model-selection principle: method-specific hyperparameters, including λ and the baseline-specific choices below, are selected by predictive NLL on the calibration subset. PAC-Bayes-based selection is studied separately in Appendix C.4.

Models and prior construction. The base-prior and anchor-matching models use the same two-hidden-layer width-16 ReLU network, whitened prior $p_0 = N(0, I)$, and rank-10-plus-diagonal Gaussian posterior. The physical weight standard deviation is $\sqrt{2/d_{\mathrm{in}}}$ and the bias standard deviation is 1. Each candidate is trained independently for 8,000 full-batch Adam steps at learning rate 3×10^{-3}, with four Monte Carlo samples per step and gradient-norm clipping at 100. The two models differ only in the prior.

The target is a zero-mean ARD Matérn-$5/2$ GP with a white-noise term, fitted on the calibration subset. The fitted noise variance is shared across all methods, while the signal kernel defines the target covariance at the anchors. The signal kernel is evaluated at 12 farthest-point anchors chosen from the fitting inputs, with jitter $10^{-6}I$. The quadratic log-density-ratio function is fitted from $25{,}000$ base and $50{,}000$ GP anchor-output draws with coefficient box $[-5, 5]$ and ℓ_2 penalty 10^{-4}. It is centered at its GP-sample median c and clipped using $H = \max\{Q^{\mathrm{GP}}_{0.9}(r-c), -Q^{\mathrm{base}}_{0.1}(r-c), 10^{-3}\}$, with $g(z) = \mathrm{clip}(r(z) - c, -H, H) - H$. We select λ from $\{0, 0.32, 1, 3.2, 10\}$.

Baselines and evaluation. We adapt fSVI and GFSVI to the same network, posterior family, fitted GP, and anchors. For fSVI, the learning rate is selected from $\{10^{-3}, 3\times10^{-3}, 10^{-2}\}$; for GFSVI, the function-space regularization parameter is selected from $\{10^{-10}, 10^{-6}, 10^{-3}\}$ at learning rate 3×10^{-3}. Predictive NLL is computed in the original target units using the shared noise variance. The base-prior, anchor-matching, and fSVI predictions use Rao–Blackwellization of the Gaussian output layer with $20{,}000$ draws for calibration and $100{,}000$ for testing; GFSVI uses its analytic linearized predictive.

C.4 Certified strength selection

Data and certified risk. This experiment is separate from Appendix C.3 and uses Yacht and Diabetes. Each run partitions the observations uniformly at random into calibration C, validation V, certification sample S, and remainder T, in proportions of approximately $20/15/50/15\%$:

Dataset	$\lvert C\rvert$	$\lvert V\rvert$	$\lvert S\rvert$	$\lvert T\rvert$
Yacht	61	46	154	47
Diabetes	88	66	221	67

The calibration set determines standardization, GP hyperparameters, and anchors; V is used for validation-based selection; S is used for posterior training and certificate-based selection; and T is used for neither training nor selection. Table 3 reports seven splits generated after the protocol was fixed. Three earlier splits were used during protocol development and are not part of the reported evaluation.

With x' and y' standardized using C, we use the bounded loss

$$\ell(w, x, y) = \min\left\{\frac{(f_w(x') - y')^2}{9},\, 1\right\}.$$

The certificate controls the Gibbs risk on $P_{\mathrm{cert}} = P \setminus (C \cup V) = S \cup T$:

$$R_{P_{\mathrm{cert}}}(q) = \mathbb{E}_{w\sim q}\left[\frac{1}{|P_{\mathrm{cert}}|}\sum_{i\in P_{\mathrm{cert}}} \ell(w, x_i, y_i)\right].$$

This is a bounded finite-population risk, not a certificate for predictive NLL. Since $\ell \in [0, 1]$, an uncapped bound below one is nonvacuous.

Model, reweighting function, and training. The network has one hidden ReLU layer of width 16 and a scalar output, with whitened prior $p_0 = N(0, I)$, corresponding to physical weight standard deviation $\sqrt{2/d_{\mathrm{in}}}$ and bias standard deviation 0.3. The posterior q is a diagonal Gaussian over all network weights.

The target is a zero-mean ARD Matérn-$5/2$ GP fitted on C, and the 12 anchors are k-means centers of the standardized calibration inputs. The quadratic log-density-ratio function is fitted in prior-whitened coordinates from $25{,}000$ base and $25{,}000$ GP draws, with coefficient box 50 and ridge penalty 10^{-4}. It is clipped with $T_R = 5$ and shifted so that the resulting weight-space reweighting function satisfies $h(w) \in [-10, 0]$.

Posterior training uses the objective in (12). Each candidate receives $10{,}000$ Adam steps with initial learning rate 3×10^{-3}, reduced by a factor 0.3 after 75% of training, using batches of at most 96, eight antithetic particles, and gradient-norm clipping at 100.

Candidates and confidence allocation. We fit posteriors for

$$\lambda \in \{0, 0.25, 1, 4, 16\}, \qquad \beta_{\mathrm{train}}/n \in \{1, 10\},$$

giving two base-prior checkpoints and eight positive-strength anchor-matching checkpoints. The temperature used to evaluate the certificate is selected separately from

$$\beta_{\mathrm{bound}}/n \in \{1, 2, 4\},$$

without retraining the posterior. Thus β_{train} and β_{bound} are distinct quantities.

The prespecified confidence registry contains the two base-prior specifications, eight anchor-matching specifications, and eight reserved GP-density-weighting specifications, each paired with the three values of β_{bound}. Hence

$$|J| = (2 + 8 + 8) \times 3 = 54, \qquad \alpha_j = \frac{1}{54}.$$

The 24 indices associated with these reserved specifications are not used in the reported comparison, and their confidence allocation is not redistributed, making the reported bounds only more conservative.

We use

$$\delta = 0.05 = 0.02_{\mathrm{grid}} + 0.01_Z + 0.01_r + 0.01_h, \qquad \xi_j = \frac{1}{54},$$

in Proposition A4. The KL term $\mathrm{KL}(q\|p_0)$ is evaluated in closed form.

Numerical evaluation and selection. All fitted posteriors and reweighting functions are fixed before numerical certificate evaluation. Normalizer bounds use $25{,}000$ independent p_0 draws, while the empirical-risk and reweighting-function bounds use $4{,}096$ i.i.d. posterior draws. For $\lambda = 0$, $\overline{\log Z} = 0$ exactly. By Proposition A4, the resulting bounds hold simultaneously over all 54 indices, so the same Monte Carlo evaluations may be used for both selection and reporting. Validation NLL is estimated from a separate posterior sample.

The base-prior rule selects β_{train} by validation NLL. Anchor matching / validation selects $(\lambda, \beta_{\mathrm{train}})$ by validation NLL. For each of these two rules, the reported certificate is the minimum over the three prespecified values of β_{bound}. Anchor matching / PAC-Bayes instead jointly selects the fitted checkpoint and β_{bound} by minimizing the valid certificate over all ten fitted checkpoints and all three bound temperatures.

Reporting. Table 3 reports the mean and standard error of the uncapped certificate over the seven reported splits. Each split carries its own certificate at confidence level 0.95; the cross-split summaries are descriptive. Because the PAC-Bayes rule directly minimizes the certificate over the prespecified candidate family, its smaller reported value should not by itself be interpreted as evidence of lower predictive risk.