A PAC-BAYESIAN VIEW OF GENERALISATION FOR PHYSICS-INFORMED MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-informed machine learning (PIML) integrates mechanistic knowledge, typically in the form of partial differential equations (PDEs), into data-driven models. Despite strong empirical performance, its statistical generalisation properties remain poorly understood, particularly in the regression setting with unbounded losses. Existing analyses rely largely on approximation or stability arguments and do not fully capture how physical structure influences generalisation from finite data. In this work, we develop a PAC-Bayesian framework for PIML that provides high-probability generalisation guarantees in the presence of potentially unbounded losses. Our analysis leverages the structure of physics-informed objectives to derive component-wise bounds whose complexity scales with the input-gradient energy of each loss, revealing a direct link between physical regularity and generalisation. We instantiate this framework under population-level regularity conditions and introduce a PAC-Bayesian calibration procedure that yields computable input-gradient complexities while explicitly controlling rare large-gradient events through a residual-tail correction. We then adopt a multi-task perspective that jointly treats data fidelity, PDE residuals, initial conditions, and boundary conditions, leading to a refined certificate with a single PAC-Bayesian complexity penalty, thereby avoiding the looseness induced by controlling each component independently. Based on the resulting certificate, we further develop a bound-aware learning procedure that jointly promotes empirical accuracy, proximity to a physics-informed prior, and low input-gradient complexity. Experiments across six PDE benchmarks demonstrate substantially tighter certificates than conservative bounded-loss and sub-Gaussian alternatives, while targeted ablations characterise the effects of posterior-training data, calibration data, gradient envelopes, and hypothesis localisation on the resulting guarantees.

1 INTRODUCTION

Physics-informed machine learning (PIML, Karniadakis et al., 2021) integrates prior scientific knowledge, typically expressed through partial differential equations (PDEs), into data-driven models. By constraining learning through physical laws, PIML has shown strong empirical performance in forward and inverse problems, scientific simulation, and hybrid modelling. However, the statistical mechanisms by which these constraints affect generalisation remain poorly understood, especially in realistic settings where the losses are unbounded and several heterogeneous data and physics objectives are optimised jointly.

Existing theoretical analyses mainly study approximation, convergence, optimisation, or Rademacher-complexity based generalisation (Shin et al., 2023; Doumèche et al., 2025b; Ryck & Mishra, 2022; Mishra & Molinaro, 2022; Jiao et al., 2022; Lu et al., 2022; Xu et al., 2025). In this work, we instead study PIML through PAC-Bayes theory (Alquier, 2024; Guedj, 2019; Hellström et al., 2025), which controls the generalisation of randomised predictors by trading empirical performance against the information complexity $D_{\mathrm{KL}}(\rho\|\pi)$ between a posterior ρ and a prior π.

Applying PAC-Bayes theory to PIML is challenging because common regression losses, such as squared data errors and differential residuals, are generally unbounded. Existing PAC-Bayes extensions therefore require additional tail or CGF control (Haddouche & Guedj, 2023; Holland, 2019;

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.

Casado et al., 2024; Rodríguez-Gálvez et al., 2024), but the resulting constants are often difficult to interpret or estimate and are only weakly connected to the physical structure of the problem.

Our key idea is to control the concentration of the original, potentially unbounded PIML losses through their input-space regularity. We relate the posterior-averaged CGF of each loss component to its population variance, and then connect this variance to the corresponding input-gradient energy. To make the resulting population quantities computable, we introduce a bounded calibration surrogate and an exact decomposition of the original gradient energy into a clipped contribution and a residual tail. This yields a computable PAC-Bayesian complexity composed of an empirical gradient term, a finite-sample correction, and an explicit gradient-tail term. The resulting certificate is posterior-dependent and can be incorporated into bound-aware optimisation. The regularity constants required by the theory are empirically assessed from calibration data, while the guarantee is stated conditionally on the corresponding population assumptions. After deriving the component-wise certificate, we further exploit the shared-model structure of the heterogeneous PIML losses through a joint PAC-Bayes-Chernoff refinement that avoids repeating KL and confidence penalties across loss components.

Our contributions are three-fold. First, we derive PAC-Bayesian generalisation bounds for PIML with potentially unbounded losses by linking loss concentration to input-gradient regularity. Second, we obtain an end-to-end computable tail-aware certificate for the original unbounded risks through bounded calibration and explicit control of rare large-gradient events, together with a practical bound-aware learning procedure. Third, we derive a joint multi-loss refinement and evaluate the resulting certificates against conventional bounded-loss and sub-Gaussian PAC-Bayes alternatives on PDE benchmarks. Our code can be found at PAC-Bayes-PIML.

2 Preliminaries

2.1 Physics-informed machine learning

A PDE with equation constraints, initial conditions (ICs), and boundary conditions (BCs) can be formulated as $\mathcal{D}[u](\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \Omega$, $\mathcal{I}[u](0, \boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \Omega_0$, and $\mathcal{B}[u](\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \partial\Omega$, where $\mathcal{D}$, $\mathcal{I}$, and $\mathcal{B}$ are the residual differential, initial, and boundary operators, respectively. The aim is to learn the target function $u : \mathbb{R}^d \to \mathbb{R}$, which maps an input coordinate $\boldsymbol{x} \in \Omega \subset \mathbb{R}^d$ to an output value $y_{\text{obs}} \in \mathbb{R}$. Here, $\boldsymbol{x}$ may contain both temporal and spatial coordinates, Ω_0 denotes the initial-time domain, and $\partial\Omega$ denotes the spatial boundary.

Our goal is to learn a parametric model $u_{\boldsymbol{\theta}}$ that approximates the target function u, with $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}$ belonging to a model class Θ. As an abuse of notation, $\boldsymbol{\theta} \in \Theta$ denotes both a model and its parameters. We further assume that $u_{\boldsymbol{\theta}}$ lies in the Sobolev space, i.e., $u_{\boldsymbol{\theta}} \in H^{p+1}(\Omega)$, where p is the highest-order derivative appearing in the PDE, and that the derivatives involved in the PDE residuals and their input gradients admit pointwise representatives, i.e., $u_{\boldsymbol{\theta}} \in C^{p+1}(\Omega)$.

Typically, learning is performed by minimising the following physics-informed empirical risk:

$$\begin{aligned}\widehat{\mathcal{R}}_{\text{PIML},\lambda}(\boldsymbol{\theta}) &= \lambda_d \widehat{\mathcal{R}}_d(\boldsymbol{\theta}) + \lambda_p \widehat{\mathcal{R}}_p(\boldsymbol{\theta}) + \lambda_{ic} \widehat{\mathcal{R}}_{ic}(\boldsymbol{\theta}) + \lambda_{bc} \widehat{\mathcal{R}}_{bc}(\boldsymbol{\theta}) \\ &= \frac{\lambda_d}{|S_d|} \sum_{(\boldsymbol{x}, y_{\text{obs}}) \in S_d} \ell_d(u_{\boldsymbol{\theta}}(\boldsymbol{x}), y_{\text{obs}}) + \frac{\lambda_p}{|S_p|} \sum_{\boldsymbol{x} \in S_p} \ell_p(\mathcal{D}[u_{\boldsymbol{\theta}}](\boldsymbol{x})) \\ &\quad + \frac{\lambda_{ic}}{|S_{ic}|} \sum_{\boldsymbol{x} \in S_{ic}} \ell_{ic}(\mathcal{I}[u_{\boldsymbol{\theta}}](\boldsymbol{x})) + \frac{\lambda_{bc}}{|S_{bc}|} \sum_{\boldsymbol{x} \in S_{bc}} \ell_{bc}(\mathcal{B}[u_{\boldsymbol{\theta}}](\boldsymbol{x})) \,. \end{aligned} \tag{1}$$

Here, $\ell_d, \ell_p, \ell_{ic}, \ell_{bc}$ denote the data-fidelity, PDE, IC, and BC losses, respectively, with the latter three corresponding to physics-based loss components. The hyperparameters $(\lambda_d, \lambda_p, \lambda_{ic}, \lambda_{bc})$ control their relative importance during optimisation.

We let S_d, S_p, S_{ic}, S_{bc} denote the corresponding sets of observations and PDE, IC, and BC collocation points. **The samples are assumed to be drawn i.i.d. within each dataset, while the datasets are mutually independent**. We allow each dataset to be generated from a different probability distribution, thereby accommodating different observation and collocation procedures. In particular, S_d is generated over $\Omega \times \mathbb{R}$, whereas the physics-based samples are generated over their corresponding physical domains.

To streamline the subsequent analysis, we write the PIML objective in terms of N_L (e.g., $N_L = 4$ in later experiments) generic loss components. The i-th loss $\ell_i(\boldsymbol{\theta}, \boldsymbol{x})$ is associated with a dataset $S_i = \{\boldsymbol{x}_{i,j}\}_{j=1}^{m_i}$ with $\boldsymbol{x}_{i,j} \overset{\text{i.i.d.}}{\sim} D_i$, and we denote the collection of datasets by $S = \{S_i\}_{i=1}^{N_L}$. Here, $\boldsymbol{x}$ denotes either an observation including its target when required, or a point from the corresponding physical domain. For the data-fidelity loss, the input gradient $\nabla_{\boldsymbol{x}} \ell_i(\boldsymbol{\theta}, \boldsymbol{x})$ is understood to be taken only with respect to the coordinate variables in Ω.

For each loss component, its population and empirical risks are defined as $\mathcal{R}_i(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x}\sim D_i}[\ell_i(\boldsymbol{\theta}, \boldsymbol{x})]$ and $\widehat{\mathcal{R}}_i(\boldsymbol{\theta}) := \frac{1}{m_i}\sum_{j=1}^{m_i} \ell_i(\boldsymbol{\theta}, \boldsymbol{x}_{i,j})$, such that $\widehat{\mathcal{R}}_{\text{PIML},\lambda}(\boldsymbol{\theta}) = \sum_{i=1}^{N_L} \lambda_i \widehat{\mathcal{R}}_i(\boldsymbol{\theta})$.

For the theoretical analysis, we assume the basic regularity conditions $\ell_i(\boldsymbol{\theta}, \boldsymbol{x}) \geq 0$, $\mathcal{R}_i(\boldsymbol{\theta}) < \infty$, and $\operatorname{ess\,inf}_{\boldsymbol{x}\sim D_i} \ell_i(\boldsymbol{\theta}, \boldsymbol{x}) = 0$ for every i and $\boldsymbol{\theta} \in \Theta$. Whenever strict convexity of the corresponding cumulant generating function is required, we additionally assume that $\mathcal{R}_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\theta}, \boldsymbol{x})$ is non-degenerate under D_i.

Training physics-informed models is challenging and commonly requires adapting the loss weights $\{\lambda_i\}$ to improve optimisation (McClenny & Braga-Neto, 2020; Wang et al., 2022). These weights may therefore vary across training runs and need not provide a meaningful definition of the target generalisation risk. We consequently study the unweighted population and empirical PIML risks $\mathcal{R}_{\text{PIML}}(\boldsymbol{\theta}) := \sum_{i=1}^{N_L} \mathcal{R}_i(\boldsymbol{\theta})$ and $\widehat{\mathcal{R}}_{\text{PIML}}(\boldsymbol{\theta}) := \sum_{i=1}^{N_L} \widehat{\mathcal{R}}_i(\boldsymbol{\theta})$, with corresponding generalisation gap $\operatorname{gen}(\boldsymbol{\theta}, S) = \mathcal{R}_{\text{PIML}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\text{PIML}}(\boldsymbol{\theta})$. This component-wise formulation serves as the basic PIML setting throughout the main analysis. Later, after developing the input-gradient certificate for the individual loss components, we additionally exploit their shared-model structure to derive a joint aggregation that avoids repeating PAC-Bayesian complexity penalties across losses.

2.2 PAC-Bayes bounds for unbounded regression losses

PAC-Bayes theory studies the generalisation of randomised predictors by considering a prior distribution $\pi \in \mathcal{M}(\Theta)$ and a data-dependent posterior distribution $\rho \in \mathcal{M}(\Theta)$ over the hypothesis class. Throughout the paper, we assume $\rho \ll \pi$ and that π is independent of the posterior-training data. The resulting guarantees typically balance the empirical performance under ρ against the information complexity $D_{\text{KL}}(\rho\|\pi)$.

A central difficulty in applying PAC-Bayes theory to PIML is that its natural regression losses, such as squared data errors and squared differential residuals, are generally unbounded. Classical bounded-loss PAC-Bayes inequalities are therefore not directly applicable unless a potentially conservative uniform bound on the loss is introduced. Recent developments extend PAC-Bayesian analysis beyond this setting by controlling the concentration behaviour of the loss rather than requiring a tight uniform bound on its range (Alquier & Guedj, 2018; Haddouche & Guedj, 2023; Haddouche et al., 2021).

A general route follows from the variational representation of the KL divergence. For a single loss ℓ, with population risk $\mathcal{R}$ and empirical risk $\widehat{\mathcal{R}}$, an oracle-style PAC-Bayes bound may contain the complexity $\frac{1}{\lambda m}\left[D_{\text{KL}}(\rho\|\pi) + \ln \frac{\Psi_{\pi,D}(\lambda)}{\delta}\right]$, where $\Psi_{\pi,D}(\lambda) := \mathbb{E}_{\boldsymbol{\theta}\sim\pi}\mathbb{E}_{S\sim D^m}\left[\exp\left(\lambda m(\mathcal{R}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}(\boldsymbol{\theta}))\right)\right]$ captures the exponential-moment behaviour of the generalisation gap (Alquier et al., 2016). Hence, obtaining a computable guarantee requires additional control of this exponential moment. Existing approaches obtain such control through assumptions on the tails or moments of the loss (Alquier & Guedj, 2018; Catoni, 2004; Germain et al., 2016), by directly constraining its cumulant generating function (CGF) (Rodríguez-Gálvez et al., 2024), or through assumptions on the underlying data distribution, often specialised to Gaussian settings (Guo et al., 2025; Shalaeva et al., 2019). Therefore, removing a uniform loss bound does not eliminate the need for information about the loss distribution; rather, it replaces range control by an appropriate distributional or moment condition.

A further practical issue is the choice of the exponential tilting parameter λ. In several PAC-Bayes bounds, λ must be selected independently of the training data, commonly through discretisation and an additional union bound. PAC-Bayes-Chernoff techniques instead combine Cramér-Chernoff arguments with generalised inverses to optimise directly over admissible $\lambda > 0$ (Casado et al., 2024). Defining the loss CGF as $\Lambda_{\boldsymbol{\theta}}(\lambda) := \log \mathbb{E}_{\boldsymbol{x}\sim D}[\exp(\lambda(\mathcal{R}(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}, \boldsymbol{x})))]$, this yields complexities of

the form $\inf_{\lambda>0}\left[\frac{D_{\mathrm{KL}}(\rho\|\pi)+\ln\frac{m}{\delta}}{\lambda(m-1)}+\frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]}{\lambda}\right]$, shifting the main challenge to obtaining a useful and computable control of the posterior-averaged CGF.

A promising route is to exploit functional inequalities and model regularity to control the CGF through input-gradient quantities (Gat et al., 2022; Casado et al., 2024). This perspective is particularly natural for PIML, where data and physics losses are defined over structured input domains and their input gradients quantify the local variation of the corresponding residuals. Motivated by this connection, we control potentially unbounded PIML losses through their input-space regularity rather than a global bound on their magnitude, and develop the calibration and tail-aware machinery needed to make the resulting quantities computable.

Nevertheless, directly using such a relation requires population quantities such as a functional-inequality coefficient and the population input-gradient energy, which are generally unavailable in practical PIML problems. Our analysis therefore follows the chain CGF $\rightarrow$ loss variance $\rightarrow$ input regularity $\rightarrow$ computable input-gradient complexity. We first derive this control separately for each original loss component, then introduce bounded auxiliary quantities only to calibrate the unknown regularity coefficient and control the population gradient energy from finite samples. The clipping operations used for these auxiliary quantities do not modify the empirical or population risks appearing in the final guarantee. Finally, after obtaining the component-wise certificate, we return to the shared structure of the PIML losses and derive a joint PAC-Bayes aggregation as an additional tightening step.

3 Input-Gradient PAC-Bayes Bounds for PIML

Our goal is to derive a computable PAC-Bayes guarantee for the original, potentially unbounded PIML risk by exploiting input regularity. Rather than bounding the loss, we relate each loss CGF to its population variance and control the latter through its input-gradient energy. We then make these population quantities computable via calibration and a tail-aware gradient decomposition. Crucially, bounded quantities enter only the regularity control; the target population and empirical risks remain those of the original, unclipped losses.

3.1 From loss concentration to input regularity

For every loss component i, its CGF is defined as $\Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda) := \log \mathbb{E}_{\boldsymbol{x}\sim D_i}[\exp(\lambda\,(\mathcal{R}_i(\boldsymbol{\theta})-\ell_i(\boldsymbol{\theta},\boldsymbol{x})))]$. Our first assumption relates its posterior average to the population variance of the corresponding loss.

Assumption 1 (Posterior-averaged CGF–variance shape). *For every loss component i, there exists $\kappa_{p,i}>0$ such that, for every posterior $\rho \ll \pi$ considered below and every admissible $\lambda>0$,* $\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda)\right] \le \frac{\kappa_{p,i}\lambda^2}{2}\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\mathrm{Var}_{\boldsymbol{x}\sim D_i}[\ell_i(\boldsymbol{\theta},\boldsymbol{x})]]$.

Assumption 1 is a population-level condition whose coefficient $\kappa_{p,i}$ cannot be certified exactly from finite samples; our guarantees are therefore conditional on its validity. Empirically, we examine the CGF-to-variance ratio across relevant λ and choose conservative $\kappa_{p,i}$ values dominating the observed ratios (Appendix C.3.1). These diagnostics support the practical plausibility of the assumption but do not certify it. Conditional on Assumption 1, the remaining data-dependent quantities are controlled by the PAC-Bayesian results below.

To connect the population variance in Assumption 1 to the input regularity of the learned model, for each loss component we define its posterior-averaged population variance-gradient coefficient as $C_i(\rho) := \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathrm{Var}_{\boldsymbol{x}\sim D_i}[\ell_i(\boldsymbol{\theta},\boldsymbol{x})]}{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D_i}[g_i(\boldsymbol{x},\boldsymbol{\theta})]}$, where $g_i(\boldsymbol{x},\boldsymbol{\theta}) := \|\nabla_{\boldsymbol{x}}\ell_i(\boldsymbol{\theta},\boldsymbol{x})\|_2^2$ denotes the input-gradient energy. Whenever the denominator is positive, this definition gives $\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathrm{Var}_{\boldsymbol{x}\sim D_i}[\ell_i(\boldsymbol{\theta},\boldsymbol{x})] = C_i(\rho)\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D_i}[g_i(\boldsymbol{x},\boldsymbol{\theta})]$. Combining this identity with Assumption 1 shows that controlling the CGF reduces to obtaining computable upper bounds on $C_i(\rho)$ and on the population input-gradient energy. The latter quantity may itself be highly variable. We therefore assume a fixed finite envelope over the localised hypothesis space considered in the analysis, whose plausibility is subsequently assessed through stress-test diagnostics.

Assumption 2 (Input-gradient admissibility). *For every loss component $i \in \{1, \dots, N_L\}$, we prescribe a finite envelope $U_i > 0$ such that $0 \le g_i(\boldsymbol{x}, \boldsymbol{\theta}) \le U_i$ for every admissible input $\boldsymbol{x}$ and every $\boldsymbol{\theta} \in \Theta_{\mathrm{loc}}$ considered in the analysis.*

Assumption 2 is a population-level condition, so our certificates are conditional on the prescribed envelopes $\{U_i\}_{i=1}^{N_L}$. We fix U_i prior to posterior learning and stress-test the condition over available data and models from the localised hypothesis region; these diagnostics (see Appendix C.3.2) provide empirical evidence, not uniform certification. To avoid using the potentially conservative U_i as the concentration scale, for $G_i \in (0, U_i)$ we decompose $g_i = g_{i,G_i} + r_{i,G_i}$, where $g_{i,G_i} := \min\{g_i, G_i\}$ and $r_{i,G_i} := (g_i - G_i)_+$. This permits concentration at the smaller scale G_i and mitigates dependence on U_i by confining it to the residual-gradient correction.

3.2 Computable calibration of the population variance-gradient coefficient

The coefficient $C_i(\rho)$ depends on the population variance and input-gradient energy of the original loss and cannot generally be evaluated from finite samples. We therefore introduce bounded auxiliary quantities solely for its calibration. To simplify notation, we first state the result for a generic loss ℓ under a distribution D, and subsequently apply it component-wise to PIML.

For prescribed calibration levels $K, G > 0$, define $\ell_K(\boldsymbol{x}, \boldsymbol{\theta}) := \min\{\ell(\boldsymbol{x}, \boldsymbol{\theta}), K\}$ and $g_{K,G}(\boldsymbol{x}, \boldsymbol{\theta}) := \min\{\|\nabla_{\boldsymbol{x}} \ell_K(\boldsymbol{x}, \boldsymbol{\theta})\|_2^2, G\}$. Hence, $0 \le \ell_K \le K$ and $0 \le g_{K,G} \le G$. We define the corresponding bounded calibration coefficient as $C_{K,G}(\rho) := \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathrm{Var}_{\boldsymbol{x}\sim D}[\ell_K(\boldsymbol{x},\boldsymbol{\theta})]}{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D}[g_{K,G}(\boldsymbol{x},\boldsymbol{\theta})]}$.

Assumption 3 (Calibration dominance). *The calibration levels K and G are chosen such that, for every posterior $\rho \ll \pi$ considered in the analysis, $C(\rho) \le C_{K,G}(\rho)$.*

Assumption 3 formalises the role of (K, G) as conservative calibration levels rather than modifications of the target loss. Its interpretation is particularly transparent when K is sufficiently large that loss clipping is inactive on the relevant population. Indeed, if $\ell_K(\boldsymbol{x}, \boldsymbol{\theta}) = \ell(\boldsymbol{x}, \boldsymbol{\theta})$ almost surely under $D \times \rho$, their population variances coincide, whereas $g_{K,G}(\boldsymbol{x}, \boldsymbol{\theta}) \le g(\boldsymbol{x}, \boldsymbol{\theta})$ almost surely, and the calibration-dominance relation follows immediately. More generally, we choose K sufficiently large so that clipping is relatively rare, and regard Assumption 3 as the corresponding population-level regularity condition (see Appendix C.3.3).

Now, let $S^{\mathrm{cal}} = (X_1, \dots, X_{M_c}) \sim D^{M_c}$ be a calibration set independent of the posterior-training data. Define the second-order U-statistic $\widehat{V}_{K,S}(\boldsymbol{\theta}) := \frac{1}{M_c(M_c-1)} \sum_{1\le j<k\le M_c} \left(\ell_K(X_j, \boldsymbol{\theta}) - \ell_K(X_k, \boldsymbol{\theta})\right)^2$ and the empirical clipped-gradient energy $\widehat{J}_{K,G,S}(\boldsymbol{\theta}) := \frac{1}{M_c} \sum_{j=1}^{M_c} g_{K,G}(X_j, \boldsymbol{\theta})$. Then $\mathbb{E}_{S^{\mathrm{cal}}}[\widehat{V}_{K,S}(\boldsymbol{\theta})] = \mathrm{Var}_{\boldsymbol{x}\sim D}[\ell_K(\boldsymbol{x}, \boldsymbol{\theta})]$. Although $\widehat{V}_{K,S}$ contains $\binom{M_c}{2}$ pairs, its standard second-order U-statistic reduction has effective sample size $m_c := \lfloor M_c/2 \rfloor$. Moreover, $2\widehat{V}_{K,S}(\boldsymbol{\theta})/K^2 \in [0,1]$ and $\widehat{J}_{K,G,S}(\boldsymbol{\theta})/G \in [0,1]$.

For $q \in [0,1]$ and $\varepsilon \ge 0$, define $\mathrm{kl}_+^{-1}(q, \varepsilon) := \sup\{p \in [q,1] : \mathrm{kl}(q\|p) \le \varepsilon\}$ and $\mathrm{kl}_-^{-1}(q, \varepsilon) := \inf\{p \in [0,q] : \mathrm{kl}(q\|p) \le \varepsilon\}$, where $\mathrm{kl}(q\|p) := q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}$. For any $\delta_C \in (0,1)$ and $\rho \ll \pi$, further define $\varepsilon_V(\rho, \delta_C) := \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log \frac{4\sqrt{m_c}}{\delta_C}}{m_c}$ and $\varepsilon_J(\rho, \delta_C) := \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log \frac{4\sqrt{M_c}}{\delta_C}}{M_c}$.

Theorem 4 (PAC-Bayes calibration of the population variance-gradient coefficient). *Suppose that Assumption 3 holds. For any prior $\pi \in \mathcal{M}(\Theta)$ independent of the calibration sample and any confidence level $\delta_C \in (0,1)$, with probability at least $1 - \delta_C$ over S^{cal}, simultaneously for all $\rho \ll \pi$, we have $C(\rho) \le C_{K,G}(\rho) \le \overline{C}_{K,G}(\rho, \delta_C)$, where*

$$\overline{C}_{K,G}(\rho, \delta_C) := \frac{\frac{K^2}{2} kl_+^{-1}\left(\frac{2\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{V}_{K,S}(\boldsymbol{\theta})]}{K^2}, \varepsilon_V(\rho, \delta_C)\right)}{G kl_-^{-1}\left(\frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{J}_{K,G,S}(\boldsymbol{\theta})]}{G}, \varepsilon_J(\rho, \delta_C)\right)}. \tag{2}$$

The bound is finite whenever the lower confidence bound in the denominator is strictly positive.

Theorem 4 therefore uses bounded clipping only as a calibration device: the resulting confidence bound controls the population variance-gradient coefficient of the original, unclipped loss. In particular, neither K nor ℓ_K enters the target population risk.

Corollary 5 (Population CGF control by the input-gradient). *For each loss i, let (K_i, G_i) denote its prescribed calibration levels and suppose that Assumption 3 holds component-wise. Applying Theorem 4 with confidence δ_C/N_L and taking a union bound gives, with probability at least $1-\delta_C$, simultaneously for every i and every $\rho \ll \pi$, $C_i(\rho) \le \overline{C}_{i,K_i,G_i}\left(\rho, \frac{\delta_C}{N_L}\right)$. For convenience, define $\overline{C}_i(\rho,\delta_C) := \kappa_{p,i}\overline{C}_{i,K_i,G_i}\left(\rho, \frac{\delta_C}{N_L}\right)$. Combining this result with Assumption 1 gives*

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\Lambda^{(i)}_{\boldsymbol{\theta}}(\lambda)\right] \le \frac{\overline{C}_i(\rho,\delta_C)\lambda^2}{2}\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D_i}[g_i(\boldsymbol{x},\boldsymbol{\theta})], \tag{3}$$

simultaneously for every i, every $\rho \ll \pi$, and every admissible $\lambda > 0$.

After calibration, the remaining obstacle to a computable PAC-Bayes bound is the population input-gradient energy of each original loss.

3.3 Component-wise input-gradient generalisation bound

We control the population input-gradient energy in equation 3 using the decomposition $g_i = g_{i,G_i} + r_{i,G_i}$, where $0 \le g_{i,G_i} \le G_i$ and $0 \le r_{i,G_i} \le H_i := U_i - G_i$. We apply PAC-Bayes concentration to the bounded component and a binary-kl inequality to the normalised residual. For every component i, define the empirical input-gradient contribution $\widehat{\mathcal{G}}_i(\rho,\delta_C) := \overline{C}_i(\rho,\delta_C)\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\sum_{j=1}^{m_i} g_i(\boldsymbol{x}_{i,j},\boldsymbol{\theta})$ and the finite-sample correction $\Delta_{i,G_i}(\rho,\delta_C,\delta_J) := m_i\overline{C}_i(\rho,\delta_C)G_i\sqrt{\frac{D_{\mathrm{KL}}(\rho\|\pi)+\ln\frac{N_L m_i}{\delta_J}}{2(m_i-1)}}$. For the residual contribution, define $\widehat{q}_{i,G_i}(\rho) := \frac{1}{m_i}\sum_{j=1}^{m_i}\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\frac{r_{i,G_i}(\boldsymbol{x}_{i,j},\boldsymbol{\theta})}{H_i}\right]$ and $\mathcal{T}_{i,G_i}(\rho,\delta_i) := H_i\mathrm{kl}_+^{-1}\left(\widehat{q}_{i,G_i}(\rho), \frac{D_{\mathrm{KL}}(\rho\|\pi)+\log\frac{2\sqrt{m_i}}{\delta_i}}{m_i}\right)$. Note that in practice, the gradient envelope is usually large enough such that $\widehat{q}_{i,G_i}(\rho) \ll 1$, hence the term $\mathcal{T}_{i,G_i}(\rho,\delta_i)$ strongly depends on the KL term and acts as a regulariser. The resulting component-wise input-gradient complexity is

$$\mathcal{G}_{i,G_i}(\rho,\delta_C,\delta_J,\delta_T) := \widehat{\mathcal{G}}_i(\rho,\delta_C) + \underbrace{\Delta_{i,G_i}(\rho,\delta_C,\delta_J) + m_i\overline{C}_i(\rho,\delta_C)\mathcal{T}_{i,G_i}\left(\rho,\tfrac{\delta_T}{N_L}\right)}_{\mathrm{Reg}(\rho,\pi)}. \tag{4}$$

Theorem 6 (Component-wise PIML bound with input-gradient complexity). *Suppose that Assumptions 1, 2, and 3 hold. Then, with probability at least $1-\delta$ over the calibration and posterior-training samples, simultaneously for every $\rho \ll \pi$ and every loss component i,*

$$\mathcal{R}_i(\rho) \le \widehat{\mathcal{R}}_i(\rho) + \frac{1}{m_i}\sqrt{2\,\mathcal{G}_{i,G_i}\left(\rho,\frac{\delta}{3},\frac{\delta}{6},\frac{\delta}{6}\right)K_i^{\mathrm{PB}}\left(\rho,\pi,\frac{\delta}{3N_L}\right)}, \tag{5}$$

where $\mathcal{R}_i(\rho) := \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\mathcal{R}_i(\boldsymbol{\theta})]$, $\widehat{\mathcal{R}}_i(\rho) := \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{\mathcal{R}}_i(\boldsymbol{\theta})]$, and $K_i^{\mathrm{PB}}(\rho,\pi,\delta_i) := \frac{m_i\left(D_{\mathrm{KL}}(\rho\|\pi)+\ln\frac{m_i}{\delta_i}\right)}{m_i-1}$.

Theorem 6 is the building block of our PIML analysis. It controls the loss through the posterior-average input-gradient $\widehat{\mathcal{G}}_i(\rho,\delta_C)$, involving the input gradients of both the model $u_{\boldsymbol{\theta}}$ and its physical residual operators, which are well-defined as $u_{\boldsymbol{\theta}} \in C^{p+1}(\Omega)$ (Section 2.1). The terms $\mathrm{Reg}(\rho,\pi)$ and $K_i^{\mathrm{PB}}(\rho,\pi,\delta_i)$ act as regularisers, favoring posteriors close to the physics-informed prior. Summing the above component-wise guarantees via a union bound yields the following result.

Corollary 7 (Task-wise PIML union bound). *Under the conditions of Theorem 6, with probability at least $1-\delta$ over the calibration and posterior-training samples, simultaneously for every $\rho \ll \pi$,*

$$\mathcal{R}_{\mathrm{PIML}}(\rho) \le \widehat{\mathcal{R}}_{\mathrm{PIML}}(\rho) + \sum_{i=1}^{N_L}\frac{1}{m_i}\sqrt{2\,\mathcal{G}_{i,G_i}(\rho)K_i^{\mathrm{PB}}(\rho,\pi)}, \tag{6}$$

where the confidence arguments are allocated with total failure probability at most δ.

This bound already provides an end-to-end computable certificate for the PIML risk using the proposed input-gradient complexity. However, it applies the PAC-Bayes-Chernoff argument independently to each loss component and consequently repeats the KL and confidence penalties across the N_L terms. We next exploit their shared-model structure to obtain a joint sample-weighted certificate with a single PAC-Bayesian complexity penalty.

3.4 Joint multi-task refinement

The component-wise analysis above does not exploit that all data and physics losses are evaluated on the same randomised predictor. PIML can therefore be viewed as a particular multi-task problem in which the loss components share the same learnable model and differ only through the fixed operators defining the data, PDE, IC, and BC residuals. This structure allows all $M := \sum_{i=1}^{N_L} m_i$ samples to be treated jointly. To this end, define the sample-weighted population and empirical risks as $\mathcal{R}^{\mathcal{S}}_{\mathrm{PIML}}(\boldsymbol{\theta}) := \frac{1}{M}\sum_{i=1}^{N_L} m_i \mathcal{R}_i(\boldsymbol{\theta})$ and $\widehat{\mathcal{R}}^{\mathcal{S}}_{\mathrm{PIML}}(\boldsymbol{\theta}) := \frac{1}{M}\sum_{i=1}^{N_L}\sum_{j=1}^{m_i} \ell_i(\boldsymbol{\theta}, \boldsymbol{x}_{i,j})$, with corresponding generalisation gap $\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S) := \mathcal{R}^{\mathcal{S}}_{\mathrm{PIML}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}^{\mathcal{S}}_{\mathrm{PIML}}(\boldsymbol{\theta})$. Further define the sample-weighted CGF $\Lambda^{\mathcal{S}}_{\boldsymbol{\theta}}(\lambda) := \frac{1}{M}\sum_{i=1}^{N_L} m_i \Lambda^{(i)}_{\boldsymbol{\theta}}(\lambda)$.

Theorem 8 (Joint PAC-Bayes-Chernoff bound). *For any $\delta \in (0,1)$ and any prior $\pi \in \mathcal{M}(\Theta)$ independent of the posterior-training samples, with probability at least $1-\delta$ over S, simultaneously for all $\rho \ll \pi$,*

$$\mathrm{gen}^{\mathcal{S}}(\rho, S) \le \inf_{\lambda>0}\left\{ \frac{D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{M}{\delta}}{\lambda(M-1)} + \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\Lambda^{\mathcal{S}}_{\boldsymbol{\theta}}(\lambda)]}{\lambda} \right\}. \tag{7}$$

Unlike the task-wise construction in Corollary 7, Theorem 8 introduces only one KL and one confidence penalty, while the task-specific concentration behaviour is aggregated through $\Lambda^{\mathcal{S}}_{\boldsymbol{\theta}}$. The input-gradient machinery developed above applies directly because $\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\Lambda^{\mathcal{S}}_{\boldsymbol{\theta}}(\lambda)] \le \frac{\lambda^2}{2M}\sum_{i=1}^{N_L} m_i \overline{C}_i(\rho,\delta_C)\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D_i}[g_i(\boldsymbol{x},\boldsymbol{\theta})]$.

Define the joint computable input-gradient complexity as $\mathcal{G}^{\mathcal{S}}_{\mathbf{G}}(\rho,\delta_C,\delta_J,\delta_T) := \widehat{\mathcal{G}}^{\mathcal{S}}(\rho,\delta_C) + \Delta^{\mathcal{S}}_{\mathbf{G}}(\rho,\delta_C,\delta_J) + \mathcal{T}^{\mathcal{S}}_{\mathbf{G}}(\rho,\delta_C,\delta_T)$, where $\widehat{\mathcal{G}}^{\mathcal{S}}(\rho,\delta_C) := \sum_{i=1}^{N_L}\overline{C}_i(\rho,\delta_C)\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\sum_{j=1}^{m_i} g_i(\boldsymbol{x}_{i,j},\boldsymbol{\theta})$, $\Delta^{\mathcal{S}}_{\mathbf{G}}(\rho,\delta_C,\delta_J) := \sum_{i=1}^{N_L} m_i\overline{C}_i(\rho,\delta_C)G_i\sqrt{\frac{D_{\mathrm{KL}}(\rho\|\pi)+\ln\frac{N_L m_i}{\delta_J}}{2(m_i-1)}}$, and $\mathcal{T}^{\mathcal{S}}_{\mathbf{G}}(\rho,\delta_C,\delta_T) := \sum_{i=1}^{N_L} m_i\overline{C}_i(\rho,\delta_C)\mathcal{T}_{i,G_i}\left(\rho,\frac{\delta_T}{N_L}\right)$.

Theorem 9 (Joint PIML bound with input-gradient complexity). *Suppose that Assumptions 1, 2, and 3 hold for every loss component. Then, with probability at least $1-\delta$ over the calibration and posterior-training samples, simultaneously for every $\rho \ll \pi$,*

$$\mathcal{R}^{\mathcal{S}}_{\mathrm{PIML}}(\rho) \le \widehat{\mathcal{R}}^{\mathcal{S}}_{\mathrm{PIML}}(\rho) + \frac{1}{M}\sqrt{2\,\mathcal{G}^{\mathcal{S}}_{\mathbf{G}}\left(\rho,\frac{\delta}{3},\frac{\delta}{6},\frac{\delta}{6}\right)K\left(\rho,\pi,\frac{\delta}{3}\right)}, \tag{8}$$

where $K(\rho,\pi,\delta_G) := \frac{M\left(D_{\mathrm{KL}}(\rho\|\pi)+\ln\frac{M}{\delta_G}\right)}{M-1}$.

Theorem 9 is the joint counterpart of Corollary 7. Both bounds use the same calibrated input-gradient control for the original PIML losses, while the joint construction aggregates the component-wise CGFs before the PAC-Bayes-Chernoff optimisation and therefore avoids repeating the KL and confidence penalties.

Corollary 10 (Joint PIML bounds for balanced and data-scarce regimes). *Under the conditions of Theorem 9, consider $\mathcal{R}_{\mathrm{PIML}}(\rho) := \sum_{i=1}^{N_L}\mathcal{R}_i(\rho)$. If $m_i = m$ for all components, multiplying equation 8 by N_L yields a valid certificate for $\mathcal{R}_{\mathrm{PIML}}(\rho)$. If instead $m_d < m_{\mathrm{phy}}$, with $m_i = m_{\mathrm{phy}}$ for all physical components, applying Theorem 6 to $\mathcal{R}_d$ and Theorem 9 jointly to the physical block yields $\mathcal{R}_{\mathrm{PIML}}(\rho) \le U_d(\rho,\delta_d) + U_{\mathrm{phy}}(\rho,\delta_{\mathrm{phy}})$ with probability at least $1-\delta$, for any $\delta_d + \delta_{\mathrm{phy}} = \delta$, where the joint physical bound is multiplied by $N_{\mathrm{phy}} := |\mathcal{I}_{\mathrm{phy}}|$.*

4 BOUND-AWARE ALGORITHM

We now turn the proposed certificate into a practical bound-aware learning procedure. To facilitate the assessment of the regularity assumptions, we localise the hypothesis space around a physics-informed reference model $\boldsymbol{\theta}_\pi$ as $\Theta_{\mathrm{loc}} := \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_\pi\|_2 \leq R_{\mathrm{loc}}\}$, and consider Gaussian prior and posterior distributions truncated to this neighbourhood, $\pi = \mathcal{N}_{\Theta_{\mathrm{loc}}}(\boldsymbol{\theta}_\pi, \sigma_\pi^2 I)$ and $\rho = \mathcal{N}_{\Theta_{\mathrm{loc}}}(\boldsymbol{\theta}_\rho, \sigma_\rho^2 I)$. The prior mean is pretrained on an independent dataset S^{prior} using only the physics losses, after which π is fixed. The same independent data are used to determine conservative loss, gradient, and envelope scales (K_i, G_i, U_i), while an independent posterior set S^{post} provides the surrogate population variance-gradient coefficients $\{\overline{C}_i^{\mathrm{train}}\}_{i=1}^{N_L}$ for training. Details of the localisation, calibration, and empirical assessment of Assumptions 1-3 are given in Appendix B.

Starting from $\boldsymbol{\theta}_\rho = \boldsymbol{\theta}_\pi$, we optimise the posterior using a stochastic surrogate of the bound. At each update, a model $\boldsymbol{\theta} \sim \rho$ is sampled through reparameterisation to estimate the sample-weighted empirical PIML risk $\widehat{R}^B_{\mathrm{PIML}}(\boldsymbol{\theta})$ and the input-gradient complexity $\mathcal{G}^B_{\mathbf{G},\mathrm{tr}}(\boldsymbol{\theta}; \pi)$ on mini-batches $B = \{B_i\}_{i=1}^{N_L}$. Our surrogate training objective is

$$\mathcal{L}^B_{\mathrm{BA}}(\boldsymbol{\theta}, \rho) := \widehat{R}^B_{\mathrm{PIML}}(\boldsymbol{\theta}) + \frac{1}{M}\sqrt{2\mathcal{G}^B_{\mathbf{G},\mathrm{tr}}(\boldsymbol{\theta}; \rho, \pi) K_{\mathrm{tr}}(\rho, \pi)}, \qquad \boldsymbol{\theta} \sim \rho, \tag{9}$$

where $\mathcal{G}^B_{\mathbf{G},\mathrm{tr}}$ contains the clipped-gradient, finite-sample, and gradient-tail contributions derived in Section 3, while K_{tr} contains the exact $D_{\mathrm{KL}}(\rho\|\pi)$ complexity of the current posterior. Hence, posterior learning jointly promotes empirical accuracy, proximity to the physics-informed prior, and low input-gradient complexity. After optimisation, the posterior-dependent calibration quantities are recomputed on S^{cal} and the theoretical certificate is evaluated using the full expressions of Section 3.

Balanced and unbalanced regimes. Training and evaluation follows Corollary 10 (joint bound) or Corollary 7 (union bound). In the balanced regime $m_i = m$ of joint bound, we optimise the stochastic objective in equation 9, multiplied by N_L to recover the equally weighted PIML risk. In the data-scarce regime $m_d < m_{\mathrm{phy}}$, we instead optimise the sum of the component-wise observational objective and the joint physical objective, with $\delta_d = \delta/N_L$ and $\delta_{\mathrm{phy}} = \delta - \delta_d$.

5 EXPERIMENT

We evaluate the proposed PAC-Bayesian framework on six PDE benchmarks: Wave-1D, Convection, Burgers, Heat-2D MS, Helmholtz and Flow Mixing-2D. For each physical loss, collocation samples are split into disjoint prior-training, posterior-training, test, and calibration sets. We consider two observational-data regimes: *balanced* (10k/5k posterior/calibration samples for Burgers and 20k/10k otherwise) and *data-scarce*, using only 100 posterior-training and 100 independent calibration observations while keeping the full physics-based collocation sets. **The prior is trained exclusively on physics-based losses**. See Appendix C.1, C.2 for details. For convenience, **Ours-joint** denotes our multi-task certificate (Corollary 10), while **Ours-union** stands for the component-wise construction (Corollary 7). We compare against **Bounded-U**, inspired by (Zakerinia & Lampert, 2025), using a fast-rate multi-task bound with a conservative uniform loss envelope, and **Sub-Gaussian**, using sub-Gaussian control through the same gradient envelope.

Table 1 shows that our input-gradient certificates are consistently tighter than both baselines, while maintaining comparable or lower test risks, indicating that bound reduction does not generally come at the expense of predictive performance. In the balanced regime, *Ours-joint* yields bounds ranging from 6.26×10^{-3} (Heat-2D MS) to 9.16 (Helmholtz). The gain is largest on Heat-2D MS, where the bound is over an order of magnitude tighter than Sub-Gaussian and nearly two orders tighter than Bounded-U. Joint and union bounds are comparable on Wave-1D and Helmholtz, while joint control yields modest gains on Convection and Burgers and clearer improvements on Flow Mixing and Heat-2D MS, tightening the latter from 9.60×10^{-3} to 6.26×10^{-3}. In the data-scarce regime, certificates deteriorate despite largely stable test risks, reflecting the statistical cost of limited observational evidence; nevertheless, our bounds remain substantially tighter than both baselines across all problems. Table 2 further shows that the certificates tighten systematically with posterior sample size. On Heat-2D MS, *Ours-joint* decreases from 1.73 with only 10 observations to 6.26×10^{-3} with $20k$ observations, remaining tighter than both baselines throughout. Joint and union bound are

nearly indistinguishable under severe data scarcity but separate as the sample size grows, reaching 6.26×10^{-3} versus 9.60×10^{-3} at 20k samples. This highlights both the benefit of additional observational evidence and, once finite-sample effects diminish, joint multi-loss control. See Appendix C.4 for complete results and detailed analysis of all benchmarks. Furthermore, additional ablations in Appendix C.5 examine the impact of the observational calibration sample size, the gradient envelope and the localisation radius to the bound, showing the importance of increasing calibration data, and identifying the large gradient envelope as the main source of inflation.

Table 1: PAC-Bayes bounds and empirical risks across PDE benchmarks and data regimes.

Problem	Data regime	Quantity	Ours-joint	Ours-union	Bounded-U	Sub-Gaussian
Wave-1D	Balanced	Test risk	(4.30±0.52)e−1	(4.30±0.52)e−1	(4.43±0.53)e−1	(4.61±0.54)e−1
		Bound	**(6.46±0.75)e−1**	**(6.46±0.75)e−1**	(1.40±0.09)e0	(8.05±0.73)e0
	Data-scarce	Test risk	(4.31±0.52)e−1	(4.31±0.52)e−1	(4.68±0.54)e−1	(4.61±0.54)e−1
		Bound	**(6.57±0.77)e−1**	**(6.57±0.77)e−1**	(6.61±0.01)e1	(8.08±0.74)e0
Convection	Balanced	Test risk	(1.91±0.23)e−2	(1.92±0.23)e−2	(2.11±0.35)e−2	(2.12±0.35)e−2
		Bound	**(3.12±0.43)e−2**	(3.18±0.44)e−2	(5.21±0.12)e−1	(4.62±0.55)e−1
	Data-scarce	Test risk	(2.06±0.31)e−2	(2.06±0.31)e−2	(2.16±0.38)e−2	(2.13±0.36)e−2
		Bound	**(7.77±0.72)e−2**	(7.81±0.73)e−2	(6.52±0.00)e1	(6.14±0.60)e−1
Burgers	Balanced	Test risk	(3.22±1.01)e−3	(3.22±1.01)e−3	(3.24±1.02)e−3	(3.24±1.02)e−3
		Bound	**(6.01±0.44)e−1**	(6.14±0.44)e−1	(8.51±0.09)e−1	(1.75±0.14)e1
	Data-scarce	Test risk	(3.23±1.02)e−3	(3.23±1.02)e−3	(3.25±1.03)e−3	(3.24±1.02)e−3
		Bound	**(1.89±0.58)e0**	(1.90±0.58)e0	(6.55±0.00)e1	(2.18±0.26)e1
Heat-2D MS	Balanced	Test risk	(2.05±0.54)e−3	(2.10±0.57)e−3	(2.23±0.66)e−3	(2.22±0.66)e−3
		Bound	**(6.26±1.67)e−3**	(9.60±2.33)e−3	(4.46±0.04)e−1	(1.76±0.49)e−1
	Data-scarce	Test risk	(2.24±0.68)e−3	(2.24±0.68)e−3	(2.24±0.68)e−3	(2.24±0.68)e−3
		Bound	**(5.74±0.78)e−1**	(5.76±0.78)e−1	(6.51±0.00)e1	(2.08±0.29)e0
Helmholtz	Balanced	Test risk	(4.62±0.48)e0	(4.62±0.48)e0	(4.62±0.48)e0	(4.73±0.48)e0
		Bound	**(9.16±0.76)e0**	(9.17±0.76)e0	(1.08±0.07)e1	(1.78±0.15)e2
	Data-scarce	Test risk	(4.62±0.48)e0	(4.62±0.48)e0	(4.73±0.48)e0	(4.73±0.48)e0
		Bound	**(9.23±0.76)e0**	**(9.23±0.76)e0**	(3.34±0.01)e2	(1.78±0.15)e2
Flow Mixing	Balanced	Test risk	(3.22±0.23)e−3	(3.22±0.23)e−3	(3.22±0.23)e−3	(3.22±0.23)e−3
		Bound	**(3.98±0.22)e−2**	(5.58±0.33)e−2	(4.51±0.01)e−1	(1.44±0.09)e0
	Data-scarce	Test risk	(3.22±0.23)e−3	(3.22±0.23)e−3	(3.22±0.23)e−3	(3.22±0.23)e−3
		Bound	**(1.21±0.16)e1**	**(1.21±0.16)e1**	(6.51±0.00)e1	(4.19±0.53)e1

Table 2: Effect of posterior observational sample size on Heat-2D MS bounds.

$m_d^{\text{posterior}}$	Ours-Joint	Ours-Union	Bounded-U	Sub-Gaussian
10	**(1.73±0.29)e0**	**(1.73±0.29)e0**	(4.25±0.00)e2	(2.43±0.41)e0
100	**(2.43±0.41)e−1**	(2.45±0.41)e−1	(6.51±0.00)e1	(9.59±1.63)e−1
1000	**(3.45±0.60)e−2**	(3.64±0.63)e−2	(8.26±0.00)e0	(4.42±0.84)e−1
10000	**(9.20±2.08)e−3**	(1.12±0.25)e−2	(1.35±0.00)e0	(2.58±0.60)e−1
20000	**(6.26±1.67)e−3**	(9.60±2.33)e−3	(4.46±0.04)e−1	(1.76±0.49)e−1

6 Conclusion

We introduced a PAC-Bayesian framework for physics-informed machine learning that connects physical regularity to generalisation through empirically accessible input-gradient quantities. We instantiated this framework through population-level regularity conditions and a PAC-Bayes calibration procedure, yielding data-dependent input-gradient complexities while controlling rare large-gradient events via a bounded residual correction. Adopting a multi-task perspective further yields a refined joint PIML certificate with a single KL complexity. Beyond the theory, we proposed a bound-aware learning algorithm and showed across six PDE benchmarks that our certificates are consistently tighter than bounded-envelope and sub-Gaussian alternatives, retain a substantial advantage under observational data scarcity, and tighten systematically with sample size. Our results further demonstrate the benefit of joint multi-task control while identifying the gradient tail as a key source that inflates the bounds. A more detailed discussion of the limitations is provided in Appendix D. Promising directions include sharper residual-gradient control, improved data-dependent calibration, and extensions to more complex physical systems.

AI USE STATEMENT

Large language models (LLMs) were used to enhance the clarity and conciseness of the manuscript, including suggesting synonyms, refining phrasing, and clarifying selected sections. They were also employed to assist with code refactoring and visualization of some figures. All conceptual contributions, experiment design, analysis, and the writing of the original content were carried out by the authors.

ETHICS STATEMENT

This research adheres to the ICLR Code of Ethics. All experiments use publicly available benchmarks and environments and involve no human subjects, personally identifiable information, or sensitive data. The contributions are methodological and are not directly deployed in safety-critical or socially sensitive settings. We report our results transparently and acknowledge the limitations of our work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide code PAC-Bayes-PIML for the baseline implementation, model architecture, dataset generation (including the generated data), data preprocessing, the full bound-aware procedure, and benchmark scripts with fixed random seeds. We also include the complete proof of the theoretical results along with a detailed description of the bound-aware learning algorithm in the Appendix.

REFERENCES

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 01 2024.

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5), 2018. ISSN 1573-0565. doi: 10.1007/s10994-017-5690-0.

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.

Daniel Berend and Tamir Tassa. Improved bounds on bell numbers and on moments of sums of random variables. *Probability and Mathematical Statistics*, 30(2):185–205, 2010. ISSN 0208-4147.

Ioar Casado, Luis A. Ortega, Aritz Pérez, and Andres R Masegosa. PAC-Bayes-Chernoff bounds for unbounded losses. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Olivier Catoni. *Statistical learning theory and stochastic optimization. Ecole d'été de probabilités de Saint-Flour XXXI-2001*. Springer, 2004. Collection : Lecture notes in mathematics n°1851.

Junwoo Cho, Seungtae Nam, Hyunmo Yang, Seok-Bae Yun, Youngjoon Hong, and Eunbyung Park. Separable physics-informed neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Monroe David Donsker and Sathamangalam Ranga Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Nathan Doumèche, Francis Bach, Gérard Biau, and Claire Boyer. Physics-informed machine learning as a kernel method. In *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2023, Edmonton, Canada*, Proceedings of Machine Learning Research, pp. 1399–1450. PMLR, 2024.

Nathan Doumèche, Francis Bach, Gérard Biau, and Claire Boyer. Physics-informed kernel learning. *J. Mach. Learn. Res.*, 26:124:1–124:39, 2025a.

Nathan Doumèche, Gérard Biau, and Claire Boyer. On the convergence of PINNs. *Bernoulli*, 31(3): 2127–2151, 2025b.

Itai Gat, Yossi Adi, Alex Schwing, and Tamir Hazan. On the importance of gradient norm in PAC-bayesian bounds. In *Annual Conference on Neural Information Processing Systems*, 2022.

Pascal Germain, Alexandre Lacasse, Mario Marchand, Sara Shanian, and François Laviolette. From pac-bayes bounds to kl regularization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

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. Curran Associates, Inc., 2016.

Benjamin Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the second congress of the French Mathematical Society*, volume 33, 2019.

Ruixin Guo, Ruoming Jin, Xinyu Li, and Yang Zhou. PAC-Bayes bounds for multivariate linear regression and linear autoencoders. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Maxime Haddouche and Benjamin Guedj. PAC-Bayes generalisation bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. PAC-Bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10), 2021. ISSN 1099-4300. doi: 10.3390/e23101330.

Maxime Haddouche, Paul Viallard, Umut Şimşekli, and Benjamin Guedj. A PAC-Bayesian Link Between Generalisation and Flat Minima. In *ALT 2025 - 36th International Conference on Algorithmic Learning Theory*, pp. 1–31, Milan, Italy, February 2025.

Zhongkai Hao, Jiachen Yao, Chang Su, Hang Su, Ziao Wang, Fanzhi Lu, Zeyu Xia, Yichi Zhang, Songming Liu, Lu Lu, and Jun Zhu. Pinnacle: A comprehensive benchmark of physics-informed neural networks for solving pdes. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 76721–76774. Curran Associates, Inc., 2024. doi: 10.52202/079017-2442.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. *Generalization Bounds: Perspectives from Information Theory and PAC-Bayes*. Foundations and Trends in Machine Learning, 2025.

Matthew Holland. PAC-Bayes under potentially heavy tails. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yuling Jiao, Yanming Lai, Dingwei Li, Xiliang Lu, Fengru Wang, Yang Wang, and Jerry Zhijian Yang. A rate of convergence of physics informed neural networks for the linear second order elliptic PDEs. *Communications in Computational Physics*, 31(4):1272–1295, 2022.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6), 5 2021. ISSN 2522-5820.

Yiping Lu, Haoxuan Chen, Jianfeng Lu, Lexing Ying, and Jose H. Blanchet. Machine learning for elliptic PDEs: Fast rate generalization bound, neural scaling law and minimax optimality. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Andreas Maurer. A note on the pac bayesian theorem. *ArXiv*, cs.LG/0411099, 2004.

Levi McClenny and Ulisses Braga-Neto. Self-adaptive physics-informed neural networks using a soft attention mechanism. *J. Comput. Phys.*, 474:111722, 2020.

Siddhartha Mishra and Roberto Molinaro. Estimates on the generalization error of physics-informed neural networks for approximating a class of inverse problems for PDEs. *IMA Journal of Numerical Analysis*, 42(2):981–1022, 04 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity. *Journal of Machine Learning Research*, 25(110):1–43, 2024.

Tim De Ryck and Siddhartha Mishra. Generic bounds on the approximation error for physics-informed (and) operator learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Anna Scampicchio, Leonardo Felipe Toso, Rahel Rickenbach, James Anderson, and Melanie Zeilinger. Physics-informed learning under mixing: How physical knowledge speeds up learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Vera Shalaeva, Alireza Fakhrizadeh Esfahani, Pascal Germain, and Mihály Petreczky. Improved PAC-Bayesian bounds for linear regression. In *AAAI Conference on Artificial Intelligence*, 2019.

Yeonjong Shin, Zhongqiang Zhang, and George Em Karniadakis. Error estimates of residual minimization using neural networks for linear PDEs. *Journal of Machine Learning for Modeling and Computing*, 4(4), 2023.

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A general framework for the practical disintegration of pac-bayesian bounds. *Machine Learning*, 113:519 – 604, 2024.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2021.110768.

Xianliang Xu, Ye Li, and Zhongyi Huang. Refined generalization analysis of the deep Ritz method and physics-informed neural networks. In *Proceeding of the International Conference on Machine Learning (ICML)*, 2025.

Hossein Zakerinia and Christoph Lampert. Fast rate bounds for multi-task and meta-learning with different sample sizes. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

APPENDIX TABLE OF CONTENTS

A Complete Proofs of Theoretical Results **14**
A.1 Auxiliary lemmas 14
A.2 Proof of Theorem 8 15
A.3 Proof of Theorem 4 19
A.4 Proof of Theorem 6 21
A.5 Proof of Theorem 9 24

B Details of the Bound-Aware Algorithm **27**

C Details of Experimental Setting and Results **29**
C.1 Benchmark 29
C.2 Implementation Details 33
C.3 Empirical assessment of the population-level assumptions 34
C.3.1 Empirical Assessment of the Assumption 1 34
C.3.2 Empirical Assessment of the Assumption 2 35
C.3.3 Empirical Assessment of the Assumption 3 39
C.4 Full Main Results 40
C.5 Ablation Study 42
C.5.1 Impact of the posterior's training sample size 42
C.5.2 Impact of the calibration's sample size 45
C.5.3 Impact of the gradient envelope 46
C.5.4 Impact of the localisation radius 46

D Limitations **47**

E Related works **47**

A Complete Proofs of Theoretical Results

A.1 Auxiliary lemmas

Lemma 11 (General PAC-Bayesian Bound). *For any distribution D on $\mathcal{X}$, for any hypothesis set Θ, for any prior $\pi \in \mathcal{M}(\Theta)$, for any measurable function $\varphi : \Theta \times \mathcal{X}^m \to \mathbb{R}$, for any posterior $\rho \in \mathcal{M}(\Theta)$ we have:*

$$\mathop{\mathbb{P}}_{S\sim D^m}\left[\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\varphi(\boldsymbol{\theta}, S) \le \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathbb{E}_{S'\sim D^m}\mathbb{E}_{\boldsymbol{\theta}'\sim\pi}e^{\varphi(\boldsymbol{\theta}', S')}\right)\right] \ge 1-\delta. \tag{10}$$

Proof. By the Donsker-Varadhan variational formula (Donsker & Varadhan, 1975),

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\varphi(\boldsymbol{\theta}, S) \le D_{\mathrm{KL}}(\rho\|\pi) + \ln \mathbb{E}_{\boldsymbol{\theta}\sim\pi}e^{\varphi(\boldsymbol{\theta}, S)}. \tag{11}$$

By Markov's inequality, with probability at least $1-\delta$ over $S \sim D^m$,

$$\mathbb{E}_{\boldsymbol{\theta}\sim\pi}e^{\varphi(\boldsymbol{\theta}, S)} \le \frac{1}{\delta}\mathbb{E}_{S'\sim D^m}\mathbb{E}_{\boldsymbol{\theta}\sim\pi}e^{\varphi(\boldsymbol{\theta}, S')}. \tag{12}$$

Substituting this inequality into equation 11 yields the result. □

Lemma 12 (Convex comparison with a binomial variable (Berend & Tassa, 2010)**).** *Let $X_1, \ldots, X_t$ be independent random variables satisfying $0 \le X_j \le 1$ almost surely, and let*

$$X := \sum_{j=1}^{t} X_j, \qquad \mu := \mathbb{E}[X].$$

Let $Y \sim \mathrm{Bin}(t, \mu/t)$. Then, for any convex function f,

$$\mathbb{E}[f(X)] \le \mathbb{E}[f(Y)].$$

Lemma 13 (Exponential moment of the binary-kl divergence (Maurer, 2004)**).** *Let $Y \sim \mathrm{Bin}(t, \mu/t)$. Then*

$$\mathbb{E}\left[\exp\left(t\, kl\left(\frac{Y}{t}\,\middle\|\,\frac{\mu}{t}\right)\right)\right] \le 2\sqrt{t},$$

where

$$kl(q\|p) := q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$$

denotes the binary Kullback–Leibler divergence.

Lemma 14 (PAC-Bayes binary-kl bound for independent bounded losses). *Let π be a prior distribution over Θ independent of the sample S. For each $\boldsymbol{\theta} \in \Theta$, let $X_1(\boldsymbol{\theta}), \ldots, X_m(\boldsymbol{\theta})$ be independent random variables satisfying $0 \le X_j(\boldsymbol{\theta}) \le 1$ almost surely, and define*

$$\widehat{\mathcal{R}}_S(\boldsymbol{\theta}) := \frac{1}{m}\sum_{j=1}^{m} X_j(\boldsymbol{\theta}), \qquad \mathcal{R}(\boldsymbol{\theta}) := \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[X_j(\boldsymbol{\theta})].$$

Then, with probability at least $1-\delta$ over S, simultaneously for every posterior ρ on Θ,

$$kl\Big(\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{\mathcal{R}}_S(\boldsymbol{\theta})]\,\Big\|\,\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\mathcal{R}(\boldsymbol{\theta})]\Big) \le \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log\frac{2\sqrt{m}}{\delta}}{m}.$$

Proof. Fix $\boldsymbol{\theta} \in \Theta$ and define

$$X(\boldsymbol{\theta}) := \sum_{j=1}^{m} X_j(\boldsymbol{\theta}), \qquad \mu(\boldsymbol{\theta}) := \mathbb{E}[X(\boldsymbol{\theta})] = m\mathcal{R}(\boldsymbol{\theta}).$$

Consider

$$f_{\boldsymbol{\theta}}(x) := \exp\Big(m\,\mathrm{kl}\Big(\frac{x}{m}\,\Big\|\,\mathcal{R}(\boldsymbol{\theta})\Big)\Big).$$

Since $q \mapsto \mathrm{kl}(q\|p)$ is convex for fixed p, and the exponential is convex and non-decreasing, $f_{\boldsymbol{\theta}}$ is convex. Lemma 12 therefore yields, for $Y \sim \mathrm{Bin}(m, \mathcal{R}(\boldsymbol{\theta}))$,

$$\mathbb{E}_S \exp\Big(m\,\mathrm{kl}\Big(\widehat{\mathcal{R}}_S(\boldsymbol{\theta}) \,\Big\|\, \mathcal{R}(\boldsymbol{\theta})\Big)\Big) \le \mathbb{E}_Y \exp\left(m\,\mathrm{kl}\left(\frac{Y}{m} \,\middle\|\, \mathcal{R}(\boldsymbol{\theta})\right)\right).$$

Applying Lemma 13,

$$\mathbb{E}_S \exp\Big(m\,\mathrm{kl}\Big(\widehat{\mathcal{R}}_S(\boldsymbol{\theta}) \,\Big\|\, \mathcal{R}(\boldsymbol{\theta})\Big)\Big) \le 2\sqrt{m}.$$

Taking expectation with respect to the prior and using Tonelli's theorem gives

$$\mathbb{E}_S \mathbb{E}_{\boldsymbol{\theta}\sim\pi} \exp\Big(m\,\mathrm{kl}\Big(\widehat{\mathcal{R}}_S(\boldsymbol{\theta}) \,\Big\|\, \mathcal{R}(\boldsymbol{\theta})\Big)\Big) \le 2\sqrt{m}.$$

Hence, by Markov's inequality, with probability at least $1-\delta$,

$$\mathbb{E}_{\boldsymbol{\theta}\sim\pi} \exp\Big(m\,\mathrm{kl}\Big(\widehat{\mathcal{R}}_S(\boldsymbol{\theta}) \,\Big\|\, \mathcal{R}(\boldsymbol{\theta})\Big)\Big) \le \frac{2\sqrt{m}}{\delta}.$$

On this event, the Donsker–Varadhan change-of-measure inequality implies that, for every posterior ρ,

$$m\,\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\mathrm{kl}\Big(\widehat{\mathcal{R}}_S(\boldsymbol{\theta}) \,\Big\|\, \mathcal{R}(\boldsymbol{\theta})\Big)\right] \le D_{\mathrm{KL}}(\rho\|\pi) + \log\frac{2\sqrt{m}}{\delta}.$$

Finally, since binary kl is jointly convex, Jensen's inequality gives

$$\mathrm{kl}\Big(\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{\mathcal{R}}_S(\boldsymbol{\theta})] \,\Big\|\, \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\mathcal{R}(\boldsymbol{\theta})]\Big) \le \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\mathrm{kl}\Big(\widehat{\mathcal{R}}_S(\boldsymbol{\theta}) \,\Big\|\, \mathcal{R}(\boldsymbol{\theta})\Big)\right],$$

which proves

$$\mathrm{kl}\Big(\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{\mathcal{R}}_S(\boldsymbol{\theta})] \,\Big\|\, \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\mathcal{R}(\boldsymbol{\theta})]\Big) \le \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log\frac{2\sqrt{m}}{\delta}}{m}.$$

□

A.2 Proof of Theorem 8

Lemma 15 (Finiteness and differentiability of the CGFs and Cramér transforms). *For each loss function i, let $\mathcal{R}_i(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x}\sim D_i}[\ell_i(\boldsymbol{\theta}, \boldsymbol{x})] < \infty$ with $\ell_i(\boldsymbol{\theta}, \boldsymbol{x}) \ge 0$, and define the centered CGF*

$$\Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda) := \log \mathbb{E}_{\boldsymbol{x}\sim D_i} \exp\big(\lambda\big(\mathcal{R}_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\theta}, \boldsymbol{x})\big)\big), \qquad \lambda \ge 0,$$

with effective domain $D_{\Lambda_{\boldsymbol{\theta}}^{(i)}} := \{\lambda \in \mathbb{R}_+ : \Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda) < \infty\}$, and Cramér transform

$$\Lambda_{\boldsymbol{\theta}}^{(i)\star}(a) := \sup_{\lambda\ge 0}\Big\{\lambda a - \Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda)\Big\}.$$

Then $D_{\Lambda_{\boldsymbol{\theta}}^{(i)}} = [0, \infty)$ and, provided $\mathcal{R}_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\theta}, \boldsymbol{x})$ is non-degenerate, $\Lambda_{\boldsymbol{\theta}}^{(i)}$ is strictly convex and C^∞ on $(0, \infty)$. Moreover, letting

$$B_{\boldsymbol{\theta}}^{(i)} := \operatorname*{ess\,sup}_{\boldsymbol{x}\sim D_i}\big(\mathcal{R}_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\theta}, \boldsymbol{x})\big) = \mathcal{R}_i(\boldsymbol{\theta}) - \operatorname*{ess\,inf}_{\boldsymbol{x}\sim D_i} \ell_i(\boldsymbol{\theta}, \boldsymbol{x}),$$

we have $\Lambda_{\boldsymbol{\theta}}^{(i)\star}(a) < \infty$ for all $a \in [0, B_{\boldsymbol{\theta}}^{(i)})$, and $\Lambda_{\boldsymbol{\theta}}^{(i)\star}$ is differentiable on $(0, B_{\boldsymbol{\theta}}^{(i)})$. In particular, if $\operatorname{ess\,inf}_{\boldsymbol{x}\sim D_i} \ell_i(\boldsymbol{\theta}, \boldsymbol{x}) = 0$, these properties hold on $[0, \mathcal{R}_i(\boldsymbol{\theta}))$ and $(0, \mathcal{R}_i(\boldsymbol{\theta}))$, respectively.

Proof. Fix i and $\boldsymbol{\theta}$. Since $\ell_i(\boldsymbol{\theta}, \boldsymbol{x}) \ge 0$, for every $\lambda \ge 0$,

$$\Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda) = \log \mathbb{E}_{\boldsymbol{x}\sim D_i} e^{\lambda(\mathcal{R}_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\theta}, \boldsymbol{x}))} \le \lambda \mathcal{R}_i(\boldsymbol{\theta}) < \infty.$$

Hence $D_{\Lambda_{\boldsymbol{\theta}}^{(i)}} = [0, \infty)$. Standard properties of CGFs imply that $\Lambda_{\boldsymbol{\theta}}^{(i)} \in C^\infty((0, \infty))$ and is strictly convex whenever $\mathcal{R}_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\theta}, \boldsymbol{x})$ is non-degenerate.

Since $\Lambda_{\boldsymbol{\theta}}^{(i)}(0) = 0$ and

$$\Lambda_{\boldsymbol{\theta}}^{(i)\prime}(0) = \mathbb{E}_{\boldsymbol{x}\sim D_i}[\mathcal{R}_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\theta}, \boldsymbol{x})] = 0,$$

strict convexity implies that $\Lambda_{\boldsymbol{\theta}}^{(i)\prime}$ is strictly increasing on $(0, \infty)$, with

$$\inf_{\lambda>0} \Lambda_{\boldsymbol{\theta}}^{(i)\prime}(\lambda) = 0, \qquad \lim_{\lambda\to\infty} \Lambda_{\boldsymbol{\theta}}^{(i)\prime}(\lambda) = B_{\boldsymbol{\theta}}^{(i)}.$$

Thus $\Lambda_{\boldsymbol{\theta}}^{(i)\prime}$ maps $(0, \infty)$ bijectively onto $(0, B_{\boldsymbol{\theta}}^{(i)})$. Therefore, for every $a \in (0, B_{\boldsymbol{\theta}}^{(i)})$, there is a unique $\lambda_a > 0$ such that $\Lambda_{\boldsymbol{\theta}}^{(i)\prime}(\lambda_a) = a$. Since $\lambda \mapsto \lambda a - \Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda)$ is strictly concave, λ_a is its unique maximiser and

$$\Lambda_{\boldsymbol{\theta}}^{(i)\star}(a) = \lambda_a a - \Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda_a) < \infty.$$

Moreover, $\Lambda_{\boldsymbol{\theta}}^{(i)\star}(0) = 0$ since $\Lambda_{\boldsymbol{\theta}}^{(i)}(0) = 0$ and Jensen's inequality gives $\Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda) \geq 0$. Hence $\Lambda_{\boldsymbol{\theta}}^{(i)\star}$ is finite on $[0, B_{\boldsymbol{\theta}}^{(i)})$.

Finally, uniqueness of the maximiser and standard convex duality imply that $\Lambda_{\boldsymbol{\theta}}^{(i)\star}$ is differentiable on $(0, B_{\boldsymbol{\theta}}^{(i)})$, with

$$\Lambda_{\boldsymbol{\theta}}^{(i)\star\prime}(a) = \lambda_a = \left(\Lambda_{\boldsymbol{\theta}}^{(i)\prime}\right)^{-1}(a).$$

Since $\operatorname{ess\,inf}_{\boldsymbol{x}\sim D_i} \ell_i(\boldsymbol{\theta}, \boldsymbol{x}) = 0$ (see Section 2.1, we obtain $B_{\boldsymbol{\theta}}^{(i)} = \mathcal{R}_i(\boldsymbol{\theta})$, which gives the stated particular case. □

Remark 16. *The proof also gives* $0 \leq \Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda) \leq \lambda \mathcal{R}_i(\boldsymbol{\theta})$ *for every admissible* $\lambda \geq 0$. *Consequently,* $\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\mathcal{R}_i(\boldsymbol{\theta})] < \infty$ *is sufficient to guarantee*

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda)\right] \leq \lambda \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\mathcal{R}_i(\boldsymbol{\theta})] < \infty, \qquad \lambda \geq 0.$$

Proposition 17 (Multi-task Cramér–Chernoff). *Let* $M = \sum_{i=1}^{N_L} m_i$ *and define the sample-weighted CGF and its Cramér transform as*

$$\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda) := \frac{1}{M} \sum_{i=1}^{N_L} m_i \Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda), \qquad \Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(a) := \sup_{\lambda\geq 0} \left\{\lambda a - \Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda)\right\}.$$

For any $\boldsymbol{\theta} \in \Theta$ *and* $a \in \mathbb{R}$,

$$\mathop{\mathbb{P}}_{S}\left[\operatorname{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S) \geq a\right] \leq e^{-M\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(a)}.$$

Moreover, by Lemma 15, $\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}$ *is finite and* C^∞ *on* $(0, \infty)$ *and, provided at least one positively weighted loss is non-degenerate, it is strictly convex. Defining*

$$B_{\boldsymbol{\theta}}^{\mathcal{S}} := \frac{1}{M} \sum_{i=1}^{N_L} m_i B_{\boldsymbol{\theta}}^{(i)},$$

we further have $\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(a) < \infty$ *for* $a \in [0, B_{\boldsymbol{\theta}}^{\mathcal{S}})$, *and* $\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}$ *is differentiable and strictly increasing on* $(0, B_{\boldsymbol{\theta}}^{\mathcal{S}})$.

Proof. Applying Markov's inequality for any $\lambda > 0$ and using the i.i.d. assumption (see Section 2.1),

$$\begin{aligned}
\mathop{\mathbb{P}}_{S}\left[\operatorname{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S) \geq a\right] &\leq e^{-M\lambda a} \mathbb{E}_S\left[e^{M\lambda \operatorname{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S)}\right] \\
&= e^{-M\lambda a} \mathbb{E}_S\left[\exp\left(\lambda \sum_{i=1}^{N_L} \sum_{j=1}^{m_i} \left(\mathcal{R}_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\theta}, \boldsymbol{x}_{i,j})\right)\right)\right] \\
&= e^{-M\lambda a} \prod_{i=1}^{N_L} \prod_{j=1}^{m_i} \mathbb{E}_{\boldsymbol{x}_{i,j}\sim D_i} \exp\left(\lambda\left(\mathcal{R}_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\theta}, \boldsymbol{x}_{i,j})\right)\right) \\
&= e^{-M\lambda a} e^{\sum_{i=1}^{N_L} m_i \Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda)} = e^{-M(\lambda a - \Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda))}.
\end{aligned}$$

Taking the infimum over $\lambda > 0$ yields

$$\mathop{\mathbb{P}}_{S}\big[\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S) \geq a\big] \leq e^{-M \sup_{\lambda \geq 0}\{\lambda a - \Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda)\}} = e^{-M\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(a)}.$$

For the regularity statement, $\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}$ is a positive weighted sum of the single-loss CGFs in Lemma 15; hence it is finite and C^∞ on $(0, \infty)$, and strictly convex whenever at least one component with $m_i > 0$ is non-degenerate. Moreover,

$$\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\prime}(0) = 0, \qquad \lim_{\lambda \to \infty} \Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\prime}(\lambda) = \frac{1}{M}\sum_{i=1}^{N_L} m_i B_{\boldsymbol{\theta}}^{(i)} = B_{\boldsymbol{\theta}}^{\mathcal{S}}.$$

The same convex-duality argument as in Lemma 15 therefore gives the stated finiteness, differentiability, and monotonicity properties of $\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}$. □

Lemma 18 (Exponential domination of the sample-weighted Cramér transform). *For every $\boldsymbol{\theta} \in \Theta$ and $c \geq 0$,*

$$\mathop{\mathbb{P}}_{S}\big[M\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}\big(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S)\big) \geq c\big] \leq e^{-c} = \mathop{\mathbb{P}}_{X \sim \mathrm{Exp}(1)}[X \geq c].$$

Proof. Fix $\boldsymbol{\theta}$ and $c \geq 0$. By Proposition 17, $\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}$ is continuous and strictly increasing on $(0, B_{\boldsymbol{\theta}}^{\mathcal{S}})$. Moreover,

$$\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S) \leq B_{\boldsymbol{\theta}}^{\mathcal{S}} \qquad \text{a.s.}$$

We distinguish three cases.

Case 1: $c/M < \sup_{a \in (0, B_{\boldsymbol{\theta}}^{\mathcal{S}})} \Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(a)$. Using the generalised inverse of $\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}$ and Proposition 17,

$$\begin{aligned}\mathop{\mathbb{P}}_{S}\big[M\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}\big(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S)\big) \geq c\big] &\leq \mathop{\mathbb{P}}_{S}\Big[\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S) \geq \big(\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}\big)^{-1}(c/M)\Big] \\ &\leq \exp\Big[-M\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}\Big(\big(\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}\big)^{-1}(c/M)\Big)\Big] = e^{-c}.\end{aligned}$$

Case 2: $c/M = \sup_{a \in (0, B_{\boldsymbol{\theta}}^{\mathcal{S}})} \Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(a)$. Since $\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(a) < c/M$ for every $a < B_{\boldsymbol{\theta}}^{\mathcal{S}}$, the event can only occur at $\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S) = B_{\boldsymbol{\theta}}^{\mathcal{S}}$. Hence, by Proposition 17 and monotonicity,

$$\mathop{\mathbb{P}}_{S}\big[M\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}\big(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S)\big) \geq c\big] \leq e^{-M\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(B_{\boldsymbol{\theta}}^{\mathcal{S}})} \leq e^{-c},$$

with the first term interpreted as zero when $\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(B_{\boldsymbol{\theta}}^{\mathcal{S}}) = \infty$.

Case 3: $c/M > \sup_{a \in (0, B_{\boldsymbol{\theta}}^{\mathcal{S}})} \Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(a)$. Again, the event can only occur at the boundary $B_{\boldsymbol{\theta}}^{\mathcal{S}}$. If $\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(B_{\boldsymbol{\theta}}^{\mathcal{S}}) < c/M$, the event is empty; otherwise Proposition 17 gives

$$\mathop{\mathbb{P}}_{S}\big[M\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}\big(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S)\big) \geq c\big] \leq e^{-M\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(B_{\boldsymbol{\theta}}^{\mathcal{S}})} \leq e^{-c}.$$

Combining the three cases proves the result. □

Theorem 8 (Joint PAC-Bayes-Chernoff bound). *For any $\delta \in (0, 1)$ and any prior $\pi \in \mathcal{M}(\Theta)$ independent of the posterior-training samples, with probability at least $1 - \delta$ over S, simultaneously for all $\rho \ll \pi$,*

$$\mathrm{gen}^{\mathcal{S}}(\rho, S) \leq \inf_{\lambda > 0}\left\{\frac{D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{M}{\delta}}{\lambda(M-1)} + \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda)]}{\lambda}\right\}. \tag{7}$$

Proof. Apply Lemma 11 with $\varphi(\boldsymbol{\theta}, S) = M_1\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S))$, where $0 < M_1 < M$. It holds with probability at least $1 - \delta$ that

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho} M_1\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S)) \leq D_{\mathrm{KL}}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathbb{E}_S\mathbb{E}_{\boldsymbol{\theta}\sim\pi} e^{M_1\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta}, S))}\right). \tag{13}$$

Since the prior π is independent of the training samples S, Tonelli's theorem allows us to swap the expectations and then control $\mathbb{E}_S\left[e^{M_1\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta},S))}\right]$ for every fixed $\boldsymbol{\theta} \in \Theta$. In particular, Lemma 18 gives

$$\mathop{\mathbb{P}}_{S}\left[M\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta},S)) \geq a\right] \leq \mathop{\mathbb{P}}_{X\sim\mathrm{Exp}(1)}[X \geq a]. \tag{14}$$

Since $X \sim \mathrm{Exp}(1)$ implies $(M_1/M)X \sim \mathrm{Exp}(M/M_1)$, we obtain

$$\mathop{\mathbb{P}}_{S}\left[M_1\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta},S)) \geq a\right] \leq \mathop{\mathbb{P}}_{X\sim\mathrm{Exp}(M/M_1)}[X \geq a]. \tag{15}$$

Since $X \sim \mathrm{Exp}(M/M_1)$ implies $e^X \sim \mathrm{Pareto}(M/M_1, 1)$, for every $t > 1$ we have

$$\mathop{\mathbb{P}}_{S}\left[e^{M_1\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta},S))} \geq t\right] \leq \mathop{\mathbb{P}}_{X\sim\mathrm{Pareto}(M/M_1,1)}[X \geq t]. \tag{16}$$

Let $Z := e^{M_1\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta},S))}$. Since $Z \geq 1$, its expectation can be written as $\mathbb{E}[Z] = 1 + \int_1^\infty \mathbb{P}(Z \geq t)\, dt$. Hence, using equation 16,

$$\begin{aligned}
\mathbb{E}_S\left[e^{M_1\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta},S))}\right] &= 1 + \int_1^\infty \mathop{\mathbb{P}}_{S}\left[e^{M_1\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta},S))} \geq t\right] dt \\
&\leq 1 + \int_1^\infty \mathop{\mathbb{P}}_{X\sim\mathrm{Pareto}(M/M_1,1)}[X \geq t] dt \\
&= \mathbb{E}_{X\sim\mathrm{Pareto}(M/M_1,1)}[X] = \frac{\frac{M}{M_1}}{\frac{M}{M_1} - 1} = \frac{M}{M - M_1}.
\end{aligned} \tag{17}$$

Moreover,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}\sim\rho} M_1\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta},S)) &= \mathbb{E}_{\boldsymbol{\theta}\sim\rho} M_1 \sup_{\lambda\geq 0}\left\{\lambda \mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta},S) - \Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda)\right\} \\
&\geq M_1 \sup_{\lambda\geq 0} \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\lambda \mathrm{gen}^{\mathcal{S}}(\boldsymbol{\theta},S) - \Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda)\right] \\
&= M_1 \sup_{\lambda\geq 0}\left\{\lambda \mathrm{gen}^{\mathcal{S}}(\rho,S) - \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda)\right]\right\} =: M_1\Lambda_{\rho}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\rho,S)),
\end{aligned} \tag{18}$$

where

$$\Lambda_{\rho}^{\mathcal{S}}(\lambda) := \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda)\right], \qquad \Lambda_{\rho}^{\mathcal{S}\star}(a) := \sup_{\lambda\geq 0}\left\{\lambda a - \Lambda_{\rho}^{\mathcal{S}}(\lambda)\right\}.$$

Plugging equation 17 and equation 18 into equation 13, we have

$$\mathop{\mathbb{P}}_{S}\left[M_1\Lambda_{\rho}^{\mathcal{S}\star}(\mathrm{gen}^{\mathcal{S}}(\rho,S)) \leq D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln\frac{M}{M - M_1}\right] \geq 1 - \delta.$$

Choosing $M_1 = M - 1$ and using the definition of $\Lambda_{\rho}^{\mathcal{S}\star}$, we obtain, with probability at least $1 - \delta$,

$$(M-1)\sup_{\lambda\geq 0}\left\{\lambda \mathrm{gen}^{\mathcal{S}}(\rho,S) - \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda)\right]\right\} \leq D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{M}{\delta}.$$

Hence, for every $\lambda > 0$,

$$\mathrm{gen}^{\mathcal{S}}(\rho,S) \leq \frac{D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{M}{\delta}}{\lambda(M-1)} + \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda)\right]}{\lambda},$$

and therefore

$$\mathop{\mathbb{P}}_{S}\left[\mathrm{gen}^{\mathcal{S}}(\rho,S) \leq \inf_{\lambda>0}\left\{\frac{D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{M}{\delta}}{\lambda(M-1)} + \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\Lambda_{\boldsymbol{\theta}}^{\mathcal{S}}(\lambda)\right]}{\lambda}\right\}\right] \geq 1 - \delta.$$

□

A.3 Proof of Theorem 4

Theorem 4 (PAC-Bayes calibration of the population variance-gradient coefficient). *Suppose that Assumption 3 holds. For any prior $\pi \in \mathcal{M}(\Theta)$ independent of the calibration sample and any confidence level $\delta_C \in (0,1)$, with probability at least $1-\delta_C$ over S^{cal}, simultaneously for all $\rho \ll \pi$, we have $C(\rho) \le C_{K,G}(\rho) \le \overline{C}_{K,G}(\rho, \delta_C)$, where*

$$\overline{C}_{K,G}(\rho,\delta_C) := \frac{\frac{K^2}{2} kl_+^{-1}\left(\frac{2\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{V}_{K,S}(\boldsymbol{\theta})]}{K^2}, \varepsilon_V(\rho,\delta_C)\right)}{G kl_-^{-1}\left(\frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{J}_{K,G,S}(\boldsymbol{\theta})]}{G}, \varepsilon_J(\rho,\delta_C)\right)}. \tag{2}$$

The bound is finite whenever the lower confidence bound in the denominator is strictly positive.

Proof. The proof separately controls the numerator and denominator of the bounded calibration coefficient $C_{K,G}(\rho) = \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathrm{Var}_{\boldsymbol{x}\sim D}[\ell_K(\boldsymbol{x},\boldsymbol{\theta})]}{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D}[g_{K,G}(\boldsymbol{x},\boldsymbol{\theta})]}$. We obtain a PAC-Bayes upper confidence bound on the posterior-averaged variance in the numerator and a PAC-Bayes lower confidence bound on the posterior-averaged clipped-gradient energy in the denominator. Their ratio gives $\overline{C}_{K,G}(\rho,\delta_C)$, while Assumption 3 transfers this control to the population variance-gradient coefficient $C(\rho)$ of the original, unclipped loss.

Step 1: PAC-Bayes–kl upper bound on the calibration variance. For a fixed $\boldsymbol{\theta}$, define the symmetric kernel $h_{K,\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{x}') := \frac{(\ell_K(\boldsymbol{x},\boldsymbol{\theta})-\ell_K(\boldsymbol{x}',\boldsymbol{\theta}))^2}{K^2}$. Since $0 \le \ell_K \le K$, we have $0 \le h_{K,\boldsymbol{\theta}} \le 1$. If $\boldsymbol{x},\boldsymbol{x}' \overset{\text{i.i.d.}}{\sim} D$, then

$$\begin{aligned} H_K(\boldsymbol{\theta}) &:= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim D}[h_{K,\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{x}')] \\ &= \frac{1}{K^2}\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim D}\left[(\ell_K(\boldsymbol{x},\boldsymbol{\theta}) - \ell_K(\boldsymbol{x}',\boldsymbol{\theta}))^2\right] \\ &= \frac{2}{K^2}\mathrm{Var}_{\boldsymbol{x}\sim D}[\ell_K(\boldsymbol{x},\boldsymbol{\theta})], \end{aligned} \tag{19}$$

where the last equality follows from the standard identity $\mathbb{E}[(X-X')^2] = 2\mathrm{Var}(X)$ for independent copies X, X'.

The empirical counterpart associated with the calibration sample $S^{\mathrm{cal}} = (\boldsymbol{x}_1,\dots,\boldsymbol{x}_{M_c})$ is the second-order U-statistic

$$\begin{aligned} H_{K,S}(\boldsymbol{\theta}) &:= \frac{2}{M_c(M_c-1)} \sum_{1\le j<k\le M_c} h_{K,\boldsymbol{\theta}}(\boldsymbol{x}_j,\boldsymbol{x}_k) \\ &= \frac{2}{K^2}\widehat{V}_{K,S}(\boldsymbol{\theta}), \end{aligned} \tag{20}$$

where $\widehat{V}_{K,S}$ is the U-statistic defined in Section 3.2. In particular, $H_{K,S}(\boldsymbol{\theta}) \in [0,1]$ and $\mathbb{E}_{S^{\mathrm{cal}}}[H_{K,S}(\boldsymbol{\theta})] = H_K(\boldsymbol{\theta})$.

The terms entering $H_{K,S}$ are not mutually independent because the same calibration observation appears in several pairs. The permutation argument provides a convenient way to reduce this dependent average over all pairs to averages over disjoint, and hence independent pairs. Let $m_c := \lfloor M_c/2 \rfloor$. For any permutation τ of $\{1,\dots,M_c\}$, reorder the calibration observations according to τ and group the first $2m_c$ observations into the consecutive pairs $(\boldsymbol{x}_{\tau(1)},\boldsymbol{x}_{\tau(2)}),(\boldsymbol{x}_{\tau(3)},\boldsymbol{x}_{\tau(4)}),\dots,(\boldsymbol{x}_{\tau(2m_c-1)},\boldsymbol{x}_{\tau(2m_c)})$. If M_c is odd, the single remaining observation is simply unused for that pairing. Define the corresponding paired average as $H_{K,S,\tau}(\boldsymbol{\theta}) := \frac{1}{m_c}\sum_{j=1}^{m_c} h_{K,\boldsymbol{\theta}}\left(\boldsymbol{x}_{\tau(2j-1)},\boldsymbol{x}_{\tau(2j)}\right)$. For each fixed permutation τ, no calibration observation occurs in more than one of these m_c pairs. Since the original observations $\boldsymbol{x}_1,\dots,\boldsymbol{x}_{M_c}$ are i.i.d., these disjoint pairs are mutually independent under $S^{\mathrm{cal}} \sim D^{M_c}$. Thus, $H_{K,S,\tau}(\boldsymbol{\theta})$ is an ordinary empirical average of m_c independent $[0,1]$-valued random variables, each having expectation $H_K(\boldsymbol{\theta})$.

The key point is that this pairing is only an analytical device: we do not replace the complete U-statistic by one particular pairing. Instead, we average over all permutations. Because the kernel is

symmetric, every unordered pair $(\boldsymbol{x}_a, \boldsymbol{x}_b)$, $a < b$, has the same probability of occupying one of the m_c paired positions under a uniformly random permutation. Consequently, averaging $H_{K,S,\tau}(\boldsymbol{\theta})$ uniformly over τ assigns the same weight to every unordered pair and exactly recovers the complete U-statistic,

$$H_{K,S}(\boldsymbol{\theta}) = \mathbb{E}_\tau[H_{K,S,\tau}(\boldsymbol{\theta})], \tag{21}$$

where the expectation is taken uniformly over permutations. Intuitively, each fixed permutation selects a collection of independent pairs from the calibration sample, while averaging over all such selections restores the original average over all $\binom{M_c}{2}$ pairs. This is why the complete U-statistic can retain all pairwise information while its concentration is controlled using the effective sample size $m_c = \lfloor M_c/2 \rfloor$ associated with the maximum number of disjoint pairs.

For every fixed τ, Lemma 14, applied to the m_c independent $[0,1]$-valued paired observations, gives the PAC-Bayes binary-kl control corresponding to $H_{K,S,\tau}$. The same result extends to the complete U-statistic through equation 21: since binary kl is jointly convex,

$$\begin{aligned} &\mathrm{kl}\left(\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[H_{K,S}(\boldsymbol{\theta})] \| \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[H_K(\boldsymbol{\theta})]\right) \\ &= \mathrm{kl}\left(\mathbb{E}_\tau \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[H_{K,S,\tau}(\boldsymbol{\theta})] \| \mathbb{E}_\tau \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[H_K(\boldsymbol{\theta})]\right) \\ &\le \mathbb{E}_\tau \mathrm{kl}\left(\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[H_{K,S,\tau}(\boldsymbol{\theta})] \| \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[H_K(\boldsymbol{\theta})]\right). \end{aligned} \tag{22}$$

Equivalently, the usual permutation reduction for U-statistics allows Lemma 14 to be applied with effective sample size m_c. Hence, with probability at least $1-\delta_V$ over S^{cal}, simultaneously for every $\rho \ll \pi$,

$$\mathrm{kl}\left(\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[H_{K,S}(\boldsymbol{\theta})] \| \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[H_K(\boldsymbol{\theta})]\right) \le \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log \frac{2\sqrt{m_c}}{\delta_V}}{m_c}. \tag{23}$$

We allocate half of the calibration failure probability to this event, i.e., $\delta_V = \delta_C/2$. By definition, $\varepsilon_V(\rho, \delta_C) = \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log \frac{4\sqrt{m_c}}{\delta_C}}{m_c}$, so equation 23 becomes

$$\mathrm{kl}\left(\mathbb{E}_\rho[H_{K,S}] \| \mathbb{E}_\rho[H_K]\right) \le \varepsilon_V(\rho, \delta_C).$$

The definition of kl_+^{-1} therefore implies

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[H_K(\boldsymbol{\theta})] \le \mathrm{kl}_+^{-1}\left(\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[H_{K,S}(\boldsymbol{\theta})], \varepsilon_V(\rho, \delta_C)\right).$$

Substituting equation 19 and equation 20 yields

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathrm{Var}_{\boldsymbol{x}\sim D}[\ell_K(\boldsymbol{x}, \boldsymbol{\theta})] \le \frac{K^2}{2}\mathrm{kl}_+^{-1}\left(\frac{2\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{V}_{K,S}(\boldsymbol{\theta})]}{K^2}, \varepsilon_V(\rho, \delta_C)\right). \tag{24}$$

This provides the required upper confidence bound on the numerator of $C_{K,G}(\rho)$.

Step 2: PAC-Bayes–kl lower bound on the calibration gradient energy. We now control the denominator. Since $g_{K,G}(\boldsymbol{x}, \boldsymbol{\theta}) = \min\{\|\nabla_{\boldsymbol{x}} \ell_K(\boldsymbol{x}, \boldsymbol{\theta})\|_2^2, G\}$, we have $0 \le g_{K,G}(\boldsymbol{x}, \boldsymbol{\theta})/G \le 1$. For fixed $\boldsymbol{\theta}$, define the empirical clipped-gradient energy $\widehat{J}_{K,G,S}(\boldsymbol{\theta}) := \frac{1}{M_c}\sum_{j=1}^{M_c} g_{K,G}(\boldsymbol{x}_j, \boldsymbol{\theta})$ and its population counterpart $J_{K,G}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x}\sim D}[g_{K,G}(\boldsymbol{x}, \boldsymbol{\theta})]$. Consequently,

$$\frac{\widehat{J}_{K,G,S}(\boldsymbol{\theta})}{G} = \frac{1}{M_c}\sum_{j=1}^{M_c} \frac{g_{K,G}(\boldsymbol{x}_j, \boldsymbol{\theta})}{G}, \qquad \frac{J_{K,G}(\boldsymbol{\theta})}{G} = \mathbb{E}_{\boldsymbol{x}\sim D}\left[\frac{g_{K,G}(\boldsymbol{x}, \boldsymbol{\theta})}{G}\right],$$

and both quantities belong to $[0,1]$.

Unlike the U-statistic in Step 1, the empirical gradient energy is an ordinary average of the M_c i.i.d. calibration observations. We may therefore apply Lemma 14 directly with sample size M_c. With probability at least $1-\delta_J$ over S^{cal}, simultaneously for every $\rho \ll \pi$,

$$\mathrm{kl}\left(\frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{J}_{K,G,S}(\boldsymbol{\theta})]}{G} \middle\| \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[J_{K,G}(\boldsymbol{\theta})]}{G}\right) \le \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log \frac{2\sqrt{M_c}}{\delta_J}}{M_c}. \tag{25}$$

We allocate the remaining half of the calibration failure probability to this event, $\delta_J = \delta_C/2$. Since $\varepsilon_J(\rho, \delta_C) = \frac{D_{\mathrm{KL}}(\rho\|\pi)+\log\frac{4\sqrt{M_c}}{\delta_C}}{M_c}$, equation 25 gives

$$\mathrm{kl}\left(\frac{\mathbb{E}_\rho[\widehat{J}_{K,G,S}]}{G}\middle\|\frac{\mathbb{E}_\rho[J_{K,G}]}{G}\right) \le \varepsilon_J(\rho, \delta_C).$$

By definition of the lower inverse kl_-^{-1},

$$\frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[J_{K,G}(\boldsymbol{\theta})]}{G} \ge \mathrm{kl}_-^{-1}\left(\frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{J}_{K,G,S}(\boldsymbol{\theta})]}{G}, \varepsilon_J(\rho, \delta_C)\right).$$

Therefore,

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D}[g_{K,G}(\boldsymbol{x}, \boldsymbol{\theta})] \ge G\mathrm{kl}_-^{-1}\left(\frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{J}_{K,G,S}(\boldsymbol{\theta})]}{G}, \varepsilon_J(\rho, \delta_C)\right). \tag{26}$$

This is the required lower confidence bound on the denominator of $C_{K,G}(\rho)$.

Step 3: upper bound on the bounded calibration coefficient. Let $\mathcal{E}_V$ and $\mathcal{E}_J$ denote the events in equation 24 and equation 26. Since $\mathbb{P}(\mathcal{E}_V) \ge 1 - \delta_C/2$ and $\mathbb{P}(\mathcal{E}_J) \ge 1 - \delta_C/2$, a union bound gives $\mathbb{P}(\mathcal{E}_V \cap \mathcal{E}_J) \ge 1 - \delta_C$. Both events are uniform over $\rho \ll \pi$, and therefore their intersection is also uniform over all admissible posteriors.

On $\mathcal{E}_V \cap \mathcal{E}_J$, whenever the lower confidence bound in equation 26 is strictly positive, we may divide the upper bound on the numerator by the lower bound on the denominator. Recalling $C_{K,G}(\rho) := \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathrm{Var}_{\boldsymbol{x}\sim D}[\ell_K(\boldsymbol{x},\boldsymbol{\theta})]}{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D}[g_{K,G}(\boldsymbol{x},\boldsymbol{\theta})]}$, we obtain

$$\begin{aligned} C_{K,G}(\rho) &\le \frac{\frac{K^2}{2}\mathrm{kl}_+^{-1}\left(\frac{2\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{V}_{K,S}(\boldsymbol{\theta})]}{K^2}, \varepsilon_V(\rho, \delta_C)\right)}{G\mathrm{kl}_-^{-1}\left(\frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{J}_{K,G,S}(\boldsymbol{\theta})]}{G}, \varepsilon_J(\rho, \delta_C)\right)} \\ &= \overline{C}_{K,G}(\rho, \delta_C). \end{aligned} \tag{27}$$

Step 4: transfer to the population variance-gradient coefficient of the original loss. The preceding argument controls $C_{K,G}(\rho)$, which is defined using the bounded calibration quantities ℓ_K and $g_{K,G}$. The quantity required by the subsequent generalisation analysis is instead $C(\rho) = \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathrm{Var}_{\boldsymbol{x}\sim D}[\ell(\boldsymbol{x},\boldsymbol{\theta})]}{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D}[g(\boldsymbol{x},\boldsymbol{\theta})]}$, which is associated with the original, unclipped loss. Assumption 3 provides precisely the required link, namely $C(\rho) \le C_{K,G}(\rho)$. Therefore, on the same event of probability at least $1 - \delta_C$, simultaneously for every $\rho \ll \pi$,

$$C(\rho) \le C_{K,G}(\rho) \le \overline{C}_{K,G}(\rho, \delta_C).$$

This proves Theorem 4. □

A.4 Proof of Theorem 6

Theorem 6 (Component-wise PIML bound with input-gradient complexity). *Suppose that Assumptions 1, 2, and 3 hold. Then, with probability at least $1 - \delta$ over the calibration and posterior-training samples, simultaneously for every $\rho \ll \pi$ and every loss component i,*

$$\mathcal{R}_i(\rho) \le \widehat{\mathcal{R}}_i(\rho) + \frac{1}{m_i}\sqrt{2\,\mathcal{G}_{i,G_i}\left(\rho, \frac{\delta}{3}, \frac{\delta}{6}, \frac{\delta}{6}\right) K_i^{\mathrm{PB}}\left(\rho, \pi, \frac{\delta}{3N_L}\right)}, \tag{5}$$

where $\mathcal{R}_i(\rho) := \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\mathcal{R}_i(\boldsymbol{\theta})]$, $\widehat{\mathcal{R}}_i(\rho) := \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\widehat{\mathcal{R}}_i(\boldsymbol{\theta})]$, *and* $K_i^{\mathrm{PB}}(\rho, \pi, \delta_i) := \frac{m_i\left(D_{\mathrm{KL}}(\rho\|\pi)+\ln\frac{m_i}{\delta_i}\right)}{m_i-1}$.

Proof. Fix a loss component $i \in \{1, \dots, N_L\}$. We prove the result through four ingredients: calibration of the population variance-gradient coefficient, concentration of the bounded part of the input-gradient energy, binary-kl control of the residual gradient contribution, and a PAC-Bayes–Chernoff bound for the original loss. The corresponding failure probabilities are denoted by $\delta_C, \delta_J, \delta_T, \delta_G \in (0,1)$ and are chosen at the end of the proof such that $\delta_C + \delta_J + \delta_T + \delta_G = \delta$.

Step 1: calibration of the CGF–gradient relation. By Corollary 5, there exists an event $\mathcal{E}_C$ with $\mathbb{P}(\mathcal{E}_C) \geq 1 - \delta_C$ such that, simultaneously for every component i and every $\rho \ll \pi$, for every admissible $\lambda > 0$, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda)] \leq \frac{\overline{C}_i(\rho, \delta_C)\lambda^2}{2} \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D_i}[g_i(\boldsymbol{x}, \boldsymbol{\theta})], \tag{28}$$

where $\overline{C}_i(\rho, \delta_C) := \kappa_{p,i}\overline{C}_{i,K_i,G_i}\left(\rho, \frac{\delta_C}{N_L}\right)$.

Thus, conditional on the calibration event, it remains to upper-bound the population input-gradient energy of the original loss.

Step 2: exact decomposition of the population input-gradient energy. For the prescribed threshold $G_i \in (0, U_i)$, recall that $g_i = g_{i,G_i} + r_{i,G_i}$ with $g_{i,G_i} := \min\{g_i, G_i\}$ and $r_{i,G_i} := (g_i - G_i)_+$. Assumption 2 implies $0 \leq g_i \leq U_i$, hence $0 \leq g_{i,G_i} \leq G_i$ and $0 \leq r_{i,G_i} \leq H_i := U_i - G_i$. Consequently,

$$\mathbb{E}_\rho\mathbb{E}_{D_i}[g_i] = \mathbb{E}_\rho\mathbb{E}_{D_i}[g_{i,G_i}] + \mathbb{E}_\rho\mathbb{E}_{D_i}[r_{i,G_i}]. \tag{29}$$

We control the two population terms separately.

Step 2a: bounded-gradient concentration. Since $g_{i,G_i}(\boldsymbol{x}, \boldsymbol{\theta}) \in [0, G_i]$, Hoeffding's lemma gives, for every fixed $\boldsymbol{\theta}$ and every $\eta > 0$,

$$\mathbb{E}_{\boldsymbol{x}\sim D_i} \exp\left(\eta\left[\mathbb{E}_{\boldsymbol{x}\sim D_i} g_{i,G_i}(\boldsymbol{x}, \boldsymbol{\theta}) - g_{i,G_i}(\boldsymbol{x}, \boldsymbol{\theta})\right]\right) \leq \exp\left(\frac{\eta^2 G_i^2}{8}\right).$$

Applying the single-loss PAC-Bayes–Chernoff argument to the bounded random variable g_{i,G_i} with confidence δ_J/N_L therefore gives, on an event $\mathcal{E}_{J,i}$ of probability at least $1-\delta_J/N_L$, simultaneously for every $\rho \ll \pi$,

$$\mathbb{E}_\rho\mathbb{E}_{D_i}[g_{i,G_i}] - \frac{1}{m_i}\mathbb{E}_\rho \sum_{j=1}^{m_i} g_{i,G_i}(\boldsymbol{x}_{i,j}, \boldsymbol{\theta}) \leq \frac{D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{N_L m_i}{\delta_J}}{\eta(m_i - 1)} + \frac{\eta G_i^2}{8}. \tag{30}$$

The right-hand side has the form $a/\eta + b\eta$, with $a = \frac{D_{\mathrm{KL}}(\rho\|\pi)+\ln\frac{N_L m_i}{\delta_J}}{m_i - 1}$ and $b = G_i^2/8$. Its minimum over $\eta > 0$ is $2\sqrt{ab}$, reached at $\eta = \sqrt{a/b}$. Hence,

$$\mathbb{E}_\rho\mathbb{E}_{D_i}[g_{i,G_i}] \leq \frac{1}{m_i}\mathbb{E}_\rho \sum_{j=1}^{m_i} g_{i,G_i}(\boldsymbol{x}_{i,j}, \boldsymbol{\theta}) + G_i\sqrt{\frac{D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{N_L m_i}{\delta_J}}{2(m_i - 1)}}. \tag{31}$$

At this point, the concentration argument necessarily involves g_{i,G_i} because its bounded range $[0, G_i]$ is what permits the use of Hoeffding's lemma. However, the empirical term in equation 31 can be upper-bounded by the empirical input-gradient energy of the original loss. Indeed, $g_{i,G_i} \leq g_i$ pointwise, and therefore

$$\frac{1}{m_i}\mathbb{E}_\rho \sum_{j=1}^{m_i} g_{i,G_i}(\boldsymbol{x}_{i,j}, \boldsymbol{\theta}) \leq \frac{1}{m_i}\mathbb{E}_\rho \sum_{j=1}^{m_i} g_i(\boldsymbol{x}_{i,j}, \boldsymbol{\theta}) =: \widehat{J}_i(\rho). \tag{32}$$

Combining equation 31 and equation 32, multiplying by $m_i\overline{C}_i(\rho, \delta_C)$, and defining $\widehat{\mathcal{G}}_i(\rho, \delta_C) := \overline{C}_i(\rho, \delta_C)\mathbb{E}_\rho \sum_{j=1}^{m_i} g_i(\boldsymbol{x}_{i,j}, \boldsymbol{\theta})$, and $\Delta_{i,G_i}(\rho, \delta_C, \delta_J) := m_i\overline{C}_i(\rho, \delta_C)G_i\sqrt{\frac{D_{\mathrm{KL}}(\rho\|\pi)+\ln\frac{N_L m_i}{\delta_J}}{2(m_i-1)}}$, gives

$$m_i\overline{C}_i(\rho, \delta_C)\mathbb{E}_\rho\mathbb{E}_{D_i}[g_{i,G_i}] \leq \widehat{\mathcal{G}}_i(\rho, \delta_C) + \Delta_{i,G_i}(\rho, \delta_C, \delta_J). \tag{33}$$

Taking a union bound over the N_L events $\{\mathcal{E}_{J,i}\}_{i=1}^{N_L}$ gives an event $\mathcal{E}_J$ with $\mathbb{P}(\mathcal{E}_J) \geq 1 - \delta_J$ on which equation 33 holds simultaneously for every component and every posterior.

Step 2b: binary-kl control of the residual gradient. The residual term cannot be discarded because the original population gradient energy satisfies equation 29. Since $0 \le r_{i,G_i} \le H_i$, define the normalised residual $Z_{i,G_i}(\boldsymbol{x},\boldsymbol{\theta}) := r_{i,G_i}(\boldsymbol{x},\boldsymbol{\theta})/H_i \in [0,1]$. Its empirical and population posterior averages are $\widehat{q}_{i,G_i}(\rho) := \frac{1}{m_i}\sum_{j=1}^{m_i} \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[Z_{i,G_i}(\boldsymbol{x}_{i,j},\boldsymbol{\theta})]$ and $q_{i,G_i}(\rho) := \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D_i}[Z_{i,G_i}(\boldsymbol{x},\boldsymbol{\theta})]$. By construction,

$$\mathbb{E}_\rho \mathbb{E}_{D_i}[r_{i,G_i}] = H_i q_{i,G_i}(\rho). \tag{34}$$

Applying the PAC-Bayes binary-kl inequality to $Z_{i,G_i} \in [0,1]$ with confidence δ_T/N_L gives an event $\mathcal{E}_{T,i}$ of probability at least $1-\delta_T/N_L$ such that, simultaneously for every $\rho \ll \pi$,

$$\mathrm{kl}\left(\widehat{q}_{i,G_i}(\rho) \| q_{i,G_i}(\rho)\right) \le \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log\frac{2N_L\sqrt{m_i}}{\delta_T}}{m_i}. \tag{35}$$

By the definition of the upper inverse kl_+^{-1},

$$q_{i,G_i}(\rho) \le \mathrm{kl}_+^{-1}\left(\widehat{q}_{i,G_i}(\rho), \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log\frac{2N_L\sqrt{m_i}}{\delta_T}}{m_i}\right).$$

Therefore,

$$\mathbb{E}_\rho \mathbb{E}_{D_i}[r_{i,G_i}] \le H_i \mathrm{kl}_+^{-1}\left(\widehat{q}_{i,G_i}(\rho), \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log\frac{2N_L\sqrt{m_i}}{\delta_T}}{m_i}\right) = \mathcal{T}_{i,G_i}\left(\rho, \frac{\delta_T}{N_L}\right). \tag{36}$$

Multiplying by $m_i\overline{C}_i(\rho,\delta_C)$ gives

$$m_i\overline{C}_i(\rho,\delta_C)\mathbb{E}_\rho \mathbb{E}_{D_i}[r_{i,G_i}] \le m_i\overline{C}_i(\rho,\delta_C)\mathcal{T}_{i,G_i}\left(\rho, \frac{\delta_T}{N_L}\right). \tag{37}$$

Again, a union bound over the N_L residual events yields an event $\mathcal{E}_T$ satisfying $\mathbb{P}(\mathcal{E}_T) \ge 1-\delta_T$.

Step 2c: computable upper bound on the original population gradient energy. Combining equation 29 with equation 33 and equation 37, on $\mathcal{E}_C \cap \mathcal{E}_J \cap \mathcal{E}_T$ we have, simultaneously for every component i and every $\rho \ll \pi$,

$$\begin{aligned} m_i\overline{C}_i(\rho,\delta_C)\mathbb{E}_\rho \mathbb{E}_{D_i}[g_i] &= m_i\overline{C}_i(\rho,\delta_C)\mathbb{E}_\rho \mathbb{E}_{D_i}[g_{i,G_i}] + m_i\overline{C}_i(\rho,\delta_C)\mathbb{E}_\rho \mathbb{E}_{D_i}[r_{i,G_i}] \\ &\le \widehat{\mathcal{G}}_i(\rho,\delta_C) + \Delta_{i,G_i}(\rho,\delta_C,\delta_J) + m_i\overline{C}_i(\rho,\delta_C)\mathcal{T}_{i,G_i}\left(\rho,\frac{\delta_T}{N_L}\right) \\ &= \mathcal{G}_{i,G_i}(\rho,\delta_C,\delta_J,\delta_T). \end{aligned} \tag{38}$$

Substituting equation 38 into equation 28 gives

$$\mathbb{E}_\rho[\Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda)] \le \frac{\lambda^2}{2m_i}\mathcal{G}_{i,G_i}(\rho,\delta_C,\delta_J,\delta_T). \tag{39}$$

Step 3: PAC-Bayes–Chernoff control of the original loss. We now apply the single-loss PAC-Bayes–Chernoff inequality directly to the original, unclipped loss ℓ_i. With confidence δ_G/N_L, there exists an event $\mathcal{E}_{G,i}$ of probability at least $1-\delta_G/N_L$ such that, simultaneously for every $\rho \ll \pi$,

$$\mathcal{R}_i(\rho) - \widehat{\mathcal{R}}_i(\rho) \le \inf_{\lambda>0}\left\{\frac{D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{N_L m_i}{\delta_G}}{\lambda(m_i-1)} + \frac{\mathbb{E}_\rho[\Lambda_{\boldsymbol{\theta}}^{(i)}(\lambda)]}{\lambda}\right\}. \tag{40}$$

Taking a union bound over the N_L losses gives an event $\mathcal{E}_G$ with $\mathbb{P}(\mathcal{E}_G) \ge 1-\delta_G$.

On the intersection $\mathcal{E}_C \cap \mathcal{E}_J \cap \mathcal{E}_T \cap \mathcal{E}_G$, substituting equation 39 into equation 40 yields

$$\mathcal{R}_i(\rho) - \widehat{\mathcal{R}}_i(\rho) \le \inf_{\lambda>0}\left\{\frac{D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{N_L m_i}{\delta_G}}{\lambda(m_i-1)} + \frac{\lambda}{2m_i}\mathcal{G}_{i,G_i}(\rho,\delta_C,\delta_J,\delta_T)\right\}.$$

Using $\inf_{\lambda>0}\{a/\lambda + b\lambda\} = 2\sqrt{ab}$ gives

$$\mathcal{R}_i(\rho) - \widehat{\mathcal{R}}_i(\rho) \le \sqrt{\frac{2\mathcal{G}_{i,G_i}(\rho,\delta_C,\delta_J,\delta_T)\left(D_{\mathrm{KL}}(\rho\|\pi) + \ln\frac{N_L m_i}{\delta_G}\right)}{m_i(m_i-1)}}$$
$$= \frac{1}{m_i}\sqrt{2\mathcal{G}_{i,G_i}(\rho,\delta_C,\delta_J,\delta_T)K_i^{\mathrm{PB}}\left(\rho,\pi,\frac{\delta_G}{N_L}\right)}, \tag{41}$$

where $K_i^{\mathrm{PB}}(\rho,\pi,\delta_i) := \frac{m_i\left(D_{\mathrm{KL}}(\rho\|\pi)+\ln\frac{m_i}{\delta_i}\right)}{m_i-1}$.

Step 4: confidence allocation. The events $\mathcal{E}_C, \mathcal{E}_J, \mathcal{E}_T, \mathcal{E}_G$ have respective failure probabilities at most $\delta_C, \delta_J, \delta_T, \delta_G$. Hence, by a union bound, their intersection has probability at least $1-(\delta_C + \delta_J + \delta_T + \delta_G)$. Choosing $\delta_C = \delta_G = \delta/3$ and $\delta_J = \delta_T = \delta/6$ gives total failure probability δ. Therefore, with probability at least $1-\delta$, simultaneously for every component i and every $\rho \ll \pi$,

$$\mathcal{R}_i(\rho) \le \widehat{\mathcal{R}}_i(\rho) + \frac{1}{m_i}\sqrt{2\,\mathcal{G}_{i,G_i}\left(\rho,\frac{\delta}{3},\frac{\delta}{6},\frac{\delta}{6}\right)K_i^{\mathrm{PB}}\left(\rho,\pi,\frac{\delta}{3N_L}\right)},$$

which proves the theorem. □

A.5 Proof of Theorem 9

Theorem 9 (Joint PIML bound with input-gradient complexity). *Suppose that Assumptions 1, 2, and 3 hold for every loss component. Then, with probability at least $1-\delta$ over the calibration and posterior-training samples, simultaneously for every $\rho \ll \pi$,*

$$\mathcal{R}^{\mathcal{S}}_{\mathrm{PIML}}(\rho) \le \widehat{\mathcal{R}}^{\mathcal{S}}_{\mathrm{PIML}}(\rho) + \frac{1}{M}\sqrt{2\,\mathcal{G}^{\mathcal{S}}_{\mathbf{G}}\left(\rho,\frac{\delta}{3},\frac{\delta}{6},\frac{\delta}{6}\right)K\left(\rho,\pi,\frac{\delta}{3}\right)}, \tag{8}$$

where $K(\rho,\pi,\delta_G) := \frac{M\left(D_{\mathrm{KL}}(\rho\|\pi)+\ln\frac{M}{\delta_G}\right)}{M-1}$.

Proof. The proof follows the same calibrated input-gradient control established in Theorem 6, but differs in the final PAC-Bayes step: instead of applying the Chernoff argument independently to every loss component, we first aggregate their CGFs and obtain a single sample-weighted PAC-Bayesian complexity. We use the confidence allocation $\delta_C = \delta_G = \delta/3$ and $\delta_J = \delta_T = \delta/6$.

Step 1: simultaneous calibration of the loss components. As in the proof of Theorem 6, applying Corollary 5 gives an event $\mathcal{E}_C$ satisfying $\mathbb{P}(\mathcal{E}_C) \ge 1-\delta_C$ such that, simultaneously for every i and every $\rho \ll \pi$,

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\Lambda^{(i)}_{\boldsymbol{\theta}}(\lambda)] \le \frac{\overline{C}_i(\rho,\delta_C)\lambda^2}{2}\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D_i}[g_i(\boldsymbol{x},\boldsymbol{\theta})]. \tag{42}$$

Define the population sample-weighted input-gradient complexity

$$\mathcal{G}^{\mathcal{S}}(\rho,\delta_C) := \sum_{i=1}^{N_L} m_i\overline{C}_i(\rho,\delta_C)\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D_i}[g_i(\boldsymbol{x},\boldsymbol{\theta})]. \tag{43}$$

Since $\Lambda^{\mathcal{S}}_{\boldsymbol{\theta}}(\lambda) = \frac{1}{M}\sum_{i=1}^{N_L} m_i\Lambda^{(i)}_{\boldsymbol{\theta}}(\lambda)$, equation 42 gives, on $\mathcal{E}_C$,

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\Lambda^{\mathcal{S}}_{\boldsymbol{\theta}}(\lambda)] = \frac{1}{M}\sum_{i=1}^{N_L} m_i\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\Lambda^{(i)}_{\boldsymbol{\theta}}(\lambda)]$$
$$\le \frac{\lambda^2}{2M}\sum_{i=1}^{N_L} m_i\overline{C}_i(\rho,\delta_C)\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\mathbb{E}_{\boldsymbol{x}\sim D_i}[g_i(\boldsymbol{x},\boldsymbol{\theta})]$$
$$= \frac{\lambda^2}{2M}\mathcal{G}^{\mathcal{S}}(\rho,\delta_C). \tag{44}$$

Step 2: joint PAC-Bayes–Chernoff control of the original PIML losses. Apply Theorem 8 to the original losses $\{\ell_i\}_{i=1}^{N_L}$ with failure probability δ_G. There exists an event $\mathcal{E}_G$ satisfying $\mathbb{P}(\mathcal{E}_G) \geq 1 - \delta_G$ such that, simultaneously for every $\rho \ll \pi$,

$$\mathrm{gen}^{\mathcal{S}}(\rho, S) \leq \inf_{\lambda>0} \left\{ \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log \frac{M}{\delta_G}}{\lambda(M-1)} + \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\Lambda^{\mathcal{S}}_{\boldsymbol{\theta}}(\lambda)]}{\lambda} \right\}. \tag{45}$$

Combining equation 44 and equation 45, on $\mathcal{E}_C \cap \mathcal{E}_G$,

$$\mathrm{gen}^{\mathcal{S}}(\rho, S) \leq \inf_{\lambda>0} \left\{ \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log \frac{M}{\delta_G}}{\lambda(M-1)} + \frac{\lambda}{2M}\mathcal{G}^{\mathcal{S}}(\rho, \delta_C) \right\}.$$

Optimising over $\lambda > 0$ gives

$$\mathcal{R}^{\mathcal{S}}_{\mathrm{PIML}}(\rho) - \widehat{\mathcal{R}}^{\mathcal{S}}_{\mathrm{PIML}}(\rho) \leq \frac{1}{M}\sqrt{2\mathcal{G}^{\mathcal{S}}(\rho, \delta_C)K(\rho, \pi, \delta_G)}, \tag{46}$$

where $K(\rho, \pi, \delta_G) := \frac{M\left(D_{\mathrm{KL}}(\rho\|\pi) + \log \frac{M}{\delta_G}\right)}{M-1}$. It therefore remains only to obtain a computable upper bound on $\mathcal{G}^{\mathcal{S}}(\rho, \delta_C)$.

Step 3: decomposition of the population input-gradient complexity. For every component i, $g_i = g_{i,G_i} + r_{i,G_i}$, hence

$$\begin{aligned}\mathcal{G}^{\mathcal{S}}(\rho, \delta_C) = &\sum_{i=1}^{N_L} m_i \overline{C}_i(\rho, \delta_C)\mathbb{E}_\rho \mathbb{E}_{D_i}[g_{i,G_i}] \\ &+ \sum_{i=1}^{N_L} m_i \overline{C}_i(\rho, \delta_C)\mathbb{E}_\rho \mathbb{E}_{D_i}[r_{i,G_i}].\end{aligned} \tag{47}$$

We control the two sums on the right-hand side using the component-wise concentration arguments already established above.

Step 3a: empirical control of the bounded-gradient contribution. For every i, the argument leading to equation 31 gives, with confidence δ_J/N_L,

$$\mathbb{E}_\rho \mathbb{E}_{D_i}[g_{i,G_i}] \leq \frac{1}{m_i}\mathbb{E}_\rho \sum_{j=1}^{m_i} g_{i,G_i}(\boldsymbol{x}_{i,j}, \boldsymbol{\theta}) + G_i\sqrt{\frac{D_{\mathrm{KL}}(\rho\|\pi) + \log \frac{N_L m_i}{\delta_J}}{2(m_i - 1)}}. \tag{48}$$

Taking a union bound over all components gives an event $\mathcal{E}_J$ with probability at least $1 - \delta_J$ on which equation 48 holds simultaneously for every i and every $\rho \ll \pi$.

As before, $g_{i,G_i} \leq g_i$ pointwise. Therefore,

$$\mathbb{E}_\rho \sum_{j=1}^{m_i} g_{i,G_i}(\boldsymbol{x}_{i,j}, \boldsymbol{\theta}) \leq \mathbb{E}_\rho \sum_{j=1}^{m_i} g_i(\boldsymbol{x}_{i,j}, \boldsymbol{\theta}).$$

Multiplying equation 48 by $m_i\overline{C}_i(\rho, \delta_C)$ and summing over i gives

$$\begin{aligned}\sum_{i=1}^{N_L} m_i \overline{C}_i(\rho, \delta_C)\mathbb{E}_\rho \mathbb{E}_{D_i}[g_{i,G_i}] &\leq \sum_{i=1}^{N_L} \overline{C}_i(\rho, \delta_C)\mathbb{E}_\rho \sum_{j=1}^{m_i} g_i(\boldsymbol{x}_{i,j}, \boldsymbol{\theta}) \\ &\quad + \sum_{i=1}^{N_L} m_i \overline{C}_i(\rho, \delta_C) G_i \sqrt{\frac{D_{\mathrm{KL}}(\rho\|\pi) + \log \frac{N_L m_i}{\delta_J}}{2(m_i - 1)}} \\ &= \widehat{\mathcal{G}}^{\mathcal{S}}(\rho, \delta_C) + \Delta^{\mathcal{S}}_{\mathbf{G}}(\rho, \delta_C, \delta_J).\end{aligned} \tag{49}$$

Hence the empirical part of the final joint complexity is expressed directly through the original input-gradient energies, although bounded concentration is still obtained through g_{i,G_i}.

Step 3b: control of the residual-gradient contribution. For each i, define $Z_{i,G_i} := r_{i,G_i}/H_i \in [0,1]$. From the binary-kl argument in the proof of Theorem 6, with confidence δ_T/N_L,

$$\mathbb{E}_\rho \mathbb{E}_{D_i}[r_{i,G_i}] \le \mathcal{T}_{i,G_i}\left(\rho, \frac{\delta_T}{N_L}\right).$$

A union bound over the N_L components therefore gives an event $\mathcal{E}_T$ satisfying $\mathbb{P}(\mathcal{E}_T) \ge 1-\delta_T$ such that, simultaneously for every i and every $\rho \ll \pi$,

$$\mathbb{E}_\rho \mathbb{E}_{D_i}[r_{i,G_i}] \le \mathcal{T}_{i,G_i}\left(\rho, \frac{\delta_T}{N_L}\right). \tag{50}$$

Multiplying by $m_i \overline{C}_i(\rho, \delta_C)$ and summing gives

$$\sum_{i=1}^{N_L} m_i \overline{C}_i(\rho, \delta_C) \mathbb{E}_\rho \mathbb{E}_{D_i}[r_{i,G_i}] \le \mathcal{T}^{\mathcal{S}}_{\mathbf{G}}(\rho, \delta_C, \delta_T). \tag{51}$$

Step 3c: computable joint gradient complexity. Combining equation 47, equation 49, and equation 51, on $\mathcal{E}_C \cap \mathcal{E}_J \cap \mathcal{E}_T$,

$$\begin{aligned} \mathcal{G}^{\mathcal{S}}(\rho, \delta_C) &\le \widehat{\mathcal{G}}^{\mathcal{S}}(\rho, \delta_C) + \Delta^{\mathcal{S}}_{\mathbf{G}}(\rho, \delta_C, \delta_J) + \mathcal{T}^{\mathcal{S}}_{\mathbf{G}}(\rho, \delta_C, \delta_T) \\ &= \mathcal{G}^{\mathcal{S}}_{\mathbf{G}}(\rho, \delta_C, \delta_J, \delta_T). \end{aligned} \tag{52}$$

Step 4: combination of the high-probability events. The four events $\mathcal{E}_C, \mathcal{E}_G, \mathcal{E}_J, \mathcal{E}_T$ have failure probabilities at most $\delta_C, \delta_G, \delta_J, \delta_T$, respectively. Since $\delta_C = \delta_G = \delta/3$ and $\delta_J = \delta_T = \delta/6$, their total failure probability is at most $\delta_C + \delta_G + \delta_J + \delta_T = \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{6} + \frac{\delta}{6} = \delta$. Hence, by a union bound, all preceding inequalities hold simultaneously with probability at least $1-\delta$.

Substituting equation 52 into equation 46 gives

$$\mathcal{R}^{\mathcal{S}}_{\mathrm{PIML}}(\rho) \le \widehat{\mathcal{R}}^{\mathcal{S}}_{\mathrm{PIML}}(\rho) + \frac{1}{M}\sqrt{2\,\mathcal{G}^{\mathcal{S}}_{\mathbf{G}}(\rho, \delta_C, \delta_J, \delta_T) K(\rho, \pi, \delta_G)}.$$

Finally, inserting $\delta_C = \delta_G = \delta/3$ and $\delta_J = \delta_T = \delta/6$ yields

$$\mathcal{R}^{\mathcal{S}}_{\mathrm{PIML}}(\rho) \le \widehat{\mathcal{R}}^{\mathcal{S}}_{\mathrm{PIML}}(\rho) + \frac{1}{M}\sqrt{2\,\mathcal{G}^{\mathcal{S}}_{\mathbf{G}}\left(\rho, \frac{\delta}{3}, \frac{\delta}{6}, \frac{\delta}{6}\right) K\left(\rho, \pi, \frac{\delta}{3}\right)},$$

which is equation 8 and completes the proof. □

B Details of the Bound-Aware Algorithm

This section provides further details of the bound-aware learning procedure introduced in Section 4 and summarised in Algorithm 1. The procedure consists of three stages: learning and localising a physics-informed prior, calibrating the quantities entering the bound and optimising the posterior, and certifying the learned posterior. We particularly detail the practical selection of the localisation radius, the calibration thresholds (K_i, G_i), and the gradient envelope U_i.

Data splitting and prior learning. We use three mutually independent datasets S^{prior}, S^{cal}, and S^{post} for prior learning and prior-side diagnostics, final calibration, and posterior learning, respectively. Starting from $\boldsymbol{\theta}_0$, **the prior mean $\boldsymbol{\theta}_\pi$ is first trained on S^{prior} using physics-only information** (i.e., by fixing $\lambda_d = 0$ for $\widehat{\mathcal{R}}_{\text{PIML},\lambda}(\boldsymbol{\theta})$), after which the prior π is fixed. Importantly, S^{cal} remains untouched throughout prior construction and posterior learning and is accessed only after the learned posterior has been fixed. This disjoint construction preserves the independence required by the calibration theorem, while the dependence of the posterior on S^{post} is handled by the PAC-Bayes complexity.

We localise the hypothesis space around $\boldsymbol{\theta}_\pi$ using

$$R_{\text{loc}} = r_{\text{loc}} \sigma \sqrt{d_\theta}, \tag{53}$$

where σ denotes the Gaussian standard deviation and d_θ the number of trainable parameters. We use $r_{\text{loc}} = 3$ by default, corresponding to a 3σ radial scale after accounting for the typical $\sqrt{d_\theta}$ dependence of the Euclidean norm of an isotropic Gaussian perturbation.

Reference loss and gradient scales. Since $\boldsymbol{\theta}_\pi$ is already well trained, evaluating the loss and its input-gradient energy only at the prior mean may underestimate the values encountered over the localised hypothesis region. We therefore sample models on the boundary $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\pi\|_2 = R_{\text{loc}}$ and evaluate each loss component over S^{prior}, which contains a disjoint subset of 100 observational samples for evaluating the data-fidelity loss. This yields component-wise reference scales $K_{\text{ref},i}$ and $G_{\text{ref},i}$.

The calibration thresholds are then parameterised as

$$K_i = K_{\text{factor}} K_{\text{ref},i}, \qquad G_i = G_{\text{factor}} G_{\text{ref},i}. \tag{54}$$

We fix $G_{\text{factor}} = 1$ and vary only K_{factor} when assessing the calibration-dominance condition in Assumption 3. This choice is motivated by the role of G_i in the denominator of the bounded calibration coefficient: fixing it at the observed reference gradient scale provides a positive and numerically stable denominator, while adjusting K_i is sufficient to empirically assess the required dominance relation. Before posterior learning, this procedure is performed under the fixed prior π on S^{prior} to select a training value $K^{\text{train}}_{\text{factor}}$. After posterior learning, it is repeated on the same S^{prior} under the fixed learned posterior ρ to select the posterior-specific certification value $K^{\text{cert}}_{\text{factor}}(\rho)$. The reference scales $K_{\text{ref},i}$ and $G_{\text{ref},i}$ remain fixed throughout.

Importantly, both K_i and G_i are used only in the calibration construction of Theorem 4. In particular, K_i clips the loss entering the empirical calibration variance, while G_i clips the corresponding calibration input-gradient energy. Neither threshold clips the original loss or the empirical input-gradient complexity used for posterior learning and final certification.

Gradient envelope. Assumption 2 additionally requires a population envelope U_i for the original input-gradient energy. We parameterise it relative to the same reference gradient scale as

$$U_i = U_{G,\text{factor}} G_{\text{ref},i}, \tag{55}$$

with $U_{G,\text{factor}} \gg 1$. In our experiments, $U_{G,\text{factor}}$ is chosen conservatively enough, typically at least 10^4, and the resulting envelopes are stress-tested using inputs and models sampled throughout the localised parameter region.

This way, the roles of G_i and U_i should be distinguished. The former is a comparatively small clipping threshold used only to construct the bounded calibration surrogate, whereas the latter is a conservative envelope required by the population input-gradient assumption and the finite-sample

concentration argument. In particular, the empirical contribution appearing in the proposed bound is evaluated using the original input-gradient energy g_i, rather than its clipped counterpart g_{i,G_i}. Hence, the final complexity retains the observed gradient magnitude directly, while U_i enters only through the corresponding concentration correction.

Overall, the three scales used in the analysis are

$$K_i = K_{\mathrm{factor}} K_{\mathrm{ref},i}, \qquad G_i = G_{\mathrm{ref},i}, \qquad U_i = U_{G,\mathrm{factor}} G_{\mathrm{ref},i}, \tag{56}$$

where K_i and G_i determine the bounded calibration coefficient and U_i controls the concentration of the original input-gradient energy. The localisation factor r_{loc} determines the parameter-space region over which the corresponding reference scales are assessed. Here $K_{\mathrm{factor}} = K_{\mathrm{factor}}^{\mathrm{train}}$ during posterior learning and $K_{\mathrm{factor}} = K_{\mathrm{factor}}^{\mathrm{cert}}(\rho)$ during final posterior-specific calibration.

Calibration of the regularity quantities. Once π, $(K_{\mathrm{ref},i}, G_{\mathrm{ref},i}, G_i, U_i)_{i=1}^{N_L}$, and the localisation region are fixed, the CGF-to-variance coefficients $\{\kappa_{p,i}\}_{i=1}^{N_L}$, the input-gradient envelope, and the calibration-dominance ratios are empirically assessed on S^{prior}. The coefficients $\{\kappa_{p,i}\}_{i=1}^{N_L}$ are fixed before posterior learning and are subsequently kept unchanged for final certification. As discussed in Section 4, these diagnostics provide empirical evidence for the corresponding population assumptions, while the final theoretical guarantee remains conditional on their validity.

For posterior learning, S^{post} is used to estimate surrogate coefficients $\{\overline{C}_i^{\mathrm{train}}\}_{i=1}^{N_L}$ using $K_{\mathrm{factor}}^{\mathrm{train}}$. These coefficients guide optimisation only and do not enter the final certificate. After learning, ρ is fixed, S^{prior} is reused to select $K_{\mathrm{factor}}^{\mathrm{cert}}(\rho)$, and only then is the untouched S^{cal} used to obtain the final posterior-specific coefficients $\{\overline{C}_i^{\mathrm{cert}}(\rho)\}_{i=1}^{N_L}$ through Theorem 4.

Bound-aware posterior learning. The posterior is initialised at the learned prior mean, $\boldsymbol{\theta}_\rho \leftarrow \boldsymbol{\theta}_\pi$, and subsequently optimised on S^{post} using the set of surrogate coefficients $\{\overline{C}_i^{\mathrm{train}}\}_{i=1}^{N_L}$ estimated from the same posterior-training split under the fixed prior π. At each iteration, mini-batches $B = \{B_i\}_{i=1}^{N_L}$ are sampled and a single model $\boldsymbol{\theta} \sim \rho$ is drawn through reparameterisation. The same realisation is used to evaluate the empirical PIML risk $\widehat{R}_{\mathrm{PIML}}^{B}(\boldsymbol{\theta})$ and the empirical input-gradient complexity. In accordance with Theorem 9, the latter is constructed from the original, unclipped input-gradient energies $g_i(\boldsymbol{x}, \boldsymbol{\theta})$, together with their finite-sample corrections based on the prescribed envelopes U_i.

In contrast, the PAC-Bayes complexity is a distribution-level quantity and is evaluated directly from the current posterior parameters through $D_{\mathrm{KL}}(\rho\|\pi)$. The resulting stochastic objective therefore combines a single-sample Monte Carlo realisation of the empirical risk and input-gradient complexity with the exact KL complexity of the current posterior. This construction directly reflects the final certificate: posterior learning favors low empirical PIML risk, small observed input-gradient energy, and limited displacement from the physics-informed prior. Importantly, neither the original PIML losses nor their input-gradient energies are clipped during posterior optimisation. The clipping thresholds (K_i, G_i) are confined to the calibration procedure used to estimate the population variance-gradient coefficients.

Posterior calibration and certification. After optimisation, the learned posterior ρ is fixed. We first reuse S^{prior} to select the posterior-specific $K_{\mathrm{factor}}^{\mathrm{cert}}(\rho)$ and empirically assess the calibration-dominance condition under ρ, while keeping the previously fixed reference scales $(K_{\mathrm{ref},i}, G_{\mathrm{ref},i})$, gradient thresholds G_i, envelopes U_i, and coefficients $\kappa_{p,i}$ unchanged. The posterior-dependent coefficients $\{\overline{C}_i^{\mathrm{cert}}(\rho)\}_{i=1}^{N_L}$ are then evaluated on the untouched S^{cal} using Theorem 4. The posterior-averaged empirical losses and original input-gradient energies are estimated on S^{post} using draws $\boldsymbol{\theta} \sim \rho$, and the final certificate is computed according to Theorem 6 and Corollary 7, or Theorem 9 and Corollary 10. Thus, S^{prior} determines the physics-informed prior, reference scales, fixed prior-side quantities, and the training and certification values of K_{factor} through their respective diagnostics; S^{post} determines the surrogate calibration quantities and the learned posterior and supplies the empirical terms entering the PAC-Bayes bound; and the independent S^{cal} is reserved exclusively for the final posterior-specific calibration. The dependence of ρ on S^{post} is handled by the PAC-Bayes KL complexity, while keeping S^{cal} untouched until the posterior is fixed preserves the independence required by the final calibration theorem.

Algorithm 1 Bound-aware PAC-Bayes learning and certification

Require: Prior data S^{prior}; calibration data S^{cal}; posterior data S^{post}; initial parameters $\boldsymbol{\theta}_0$; standard deviation σ; confidence δ; localisation factor r_{loc}; gradient factor $U_{G,\text{factor}}$; iterations $N_T, N_{T'}$.

Step 1: prior learning and prior-side setup

$\boldsymbol{\theta}_\pi \leftarrow \boldsymbol{\theta}_0$
for $t = 1$ to N_T **do**
  Draw mini-batches from S^{prior} and update $\boldsymbol{\theta}_\pi$ to minimise $\widehat{\mathcal{R}}_{\text{PIML},\lambda}(\boldsymbol{\theta})$ (equation 1)
end for
Fix π and set $R_{\text{loc}} = r_{\text{loc}}\sigma\sqrt{d_\theta}$
Sample models on $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\pi\|_2 = R_{\text{loc}}$ and estimate $(K_{\text{ref},i}, G_{\text{ref},i})_{i=1}^{N_L}$ from S^{prior}
Fix $G_{\text{factor}} = 1$, set $G_i = G_{\text{ref},i}$ and $U_i = U_{G,\text{factor}} G_{\text{ref},i}$, and empirically assess the input-gradient envelopes
Set each $\kappa_{p,i}$ to a conservative upper envelope of the empirical CGF-to-variance ratios evaluated on S^{prior}, and keep it fixed thereafter
Select $K^{\text{train}}_{\text{factor}}$ on S^{prior} under π according to the empirical diagnostics for Assumption 3, and set $K_i^{\text{train}} = K^{\text{train}}_{\text{factor}} K_{\text{ref},i}$

Step 2: bound-aware posterior learning

Estimate the surrogate coefficients $\{\overline{C}_i^{\text{train}}\}_{i=1}^{N_L}$ from S^{post} using $(K_i^{\text{train}}, G_i)_{i=1}^{N_L}$
$\boldsymbol{\theta}_\rho \leftarrow \boldsymbol{\theta}_\pi$
for $t = 1$ to $N_{T'}$ **do**
  Draw mini-batches $B = \{B_i\}_{i=1}^{N_L}$ from S^{post}
  Draw a reparameterised model $\boldsymbol{\theta} \sim \rho$
  Evaluate $\widehat{R}^B_{\text{PIML}}(\boldsymbol{\theta})$ and the empirical original input-gradient complexity using the same $\boldsymbol{\theta}$
  Evaluate the corresponding finite-sample gradient correction using $(U_i)_{i=1}^{N_L}$
  Evaluate $K_{\text{tr}}(\rho, \pi)$, including $D_{\text{KL}}(\rho\|\pi)$, from the current distribution parameters
  Update $\boldsymbol{\theta}_\rho$ by minimising $\mathcal{L}^B_{\text{BA}}(\boldsymbol{\theta}, \rho)$ in equation 9
end for

Step 3: posterior-specific calibration and certification

Fix the learned posterior ρ
Select $K^{\text{cert}}_{\text{factor}}(\rho)$ on S^{prior} under the fixed posterior ρ according to the empirical diagnostics for Assumption 3, and set $K_i^{\text{cert}} = K^{\text{cert}}_{\text{factor}}(\rho) K_{\text{ref},i}$
Estimate $\{\overline{C}_i^{\text{cert}}(\rho)\}_{i=1}^{N_L}$ from the untouched S^{cal} using $(K_i^{\text{cert}}, G_i)_{i=1}^{N_L}$
Estimate the posterior-averaged empirical risks and original input-gradient energies on S^{post} using draws $\boldsymbol{\theta} \sim \rho$
Evaluate the final certificate according to Theorem 6 and Corollary 7, or Theorem 9 and Corollary 10
return learned posterior ρ and its PAC-Bayes certificate

C Details of Experimental Setting and Results

C.1 Benchmark

1D-Wave. This problem considers a one-dimensional hyperbolic PDE that is commonly used as a benchmark for wave propagation phenomena. The governing equation is [1]

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} - 4\frac{\partial^2 u}{\partial x^2} &= 0, && x \in (0,1),\ t \in (0,1), \\ u(0,x) &= \sin(\pi x) + \frac{1}{2}\sin(\beta\pi x), && x \in [0,1], \\ \frac{\partial u(0,x)}{\partial t} &= 0, && x \in [0,1], \\ u(t,0) &= u(t,1) = 0, && t \in [0,1]. \end{aligned} \tag{57}$$

[1] For clarity, π denotes the numeric constant $\pi \approx 3.14$ in this section, to avoid confusion with the prior distribution in the bounds.

The corresponding analytic solution is

$$u(t,x) = \sin(\pi x)\cos(2\pi t) + \frac{1}{2}\sin(\beta\pi x)\cos(2\beta\pi t). \tag{58}$$

In our experiments, we set $\beta = 4$. Despite its relatively simple form, this problem poses challenges for physics-informed neural networks (PINNs) due to the presence of multiple frequency components in the solution. For simplicity, in the experimental parts, we denote the three physical loss terms as follows:

$$\begin{aligned}
\ell_p(\boldsymbol{\theta}, \mathbf{x}) &= \left(\frac{\partial^2 u_{\boldsymbol{\theta}}}{\partial t^2} - 4\frac{\partial^2 u_{\boldsymbol{\theta}}}{\partial x^2}\right)^2, \\
\ell_{ic}(\boldsymbol{\theta}, \mathbf{x}) &= \left(u_{\boldsymbol{\theta}}(0,x) - (\sin(\pi x) + \frac{1}{2}\sin(\beta\pi x))\right)^2, \\
\ell_{bc}(\boldsymbol{\theta}, \mathbf{x}) &= \frac{1}{2}\left[(u_{\boldsymbol{\theta}}(t,0))^2 + (u_{\boldsymbol{\theta}}(t,1))^2\right].
\end{aligned} \tag{59}$$

Convection. This problem considers a hyperbolic PDE that serves as a standard benchmark for transport-dominated dynamics, also known as advection. The governing equation is

$$\begin{aligned}
\frac{\partial u}{\partial t} + \beta\frac{\partial u}{\partial x} &= 0, && x \in (0, 2\pi),\ t \in (0,1), \\
u(0,x) &= \sin(x), && x \in [0, 2\pi], \\
u(t,0) &= u(t,2\pi), && t \in [0,1].
\end{aligned} \tag{60}$$

The corresponding analytic solution is

$$u(t,x) = \sin(x - \beta t). \tag{61}$$

In our experiments, we set $\beta = 50$. Despite the simple closed-form expression, this problem remains challenging for physics-informed neural networks (PINNs) due to the high-frequency and rapidly propagating solution features. We consider the following physical losses in the experiments:

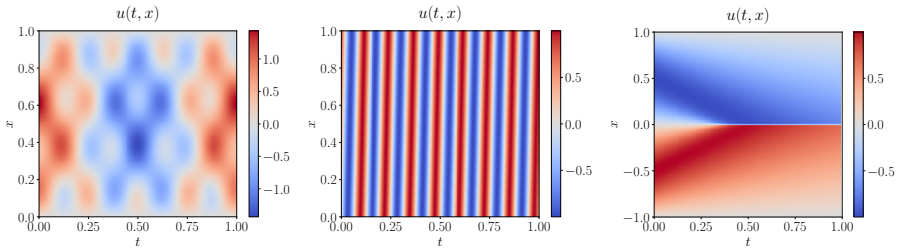
$$\begin{aligned}
\ell_p(\boldsymbol{\theta}, \mathbf{x}) &= \left(\frac{\partial u_{\boldsymbol{\theta}}}{\partial t} + \beta\frac{\partial u_{\boldsymbol{\theta}}}{\partial x}\right)^2, \\
\ell_{ic}(\boldsymbol{\theta}, \mathbf{x}) &= \left(u_{\boldsymbol{\theta}}(0,x) - \sin(x)\right)^2, \\
\ell_{bc}(\boldsymbol{\theta}, \mathbf{x}) &= \left(u_{\boldsymbol{\theta}}(t,0) - u_{\boldsymbol{\theta}}(t,2\pi)\right)^2.
\end{aligned} \tag{62}$$

Burgers (PINNacle, Hao et al., 2024). The Burgers equation is a fundamental nonlinear PDE that serves as a simplified model for various physical phenomena, including fluid dynamics, gas dynamics, traffic flow, and acoustics. It combines nonlinear convection with diffusion, making it an important benchmark for studying the interaction between nonlinear wave propagation and dissipative effects. The one-dimensional viscous Burgers equation is commonly expressed as

$$\begin{aligned}
\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} &= \nu\frac{\partial^2 u}{\partial x^2}, \\
u(0,x) &= -\sin(\pi x), \\
u(t,-1) &= u(t,1) = 0,
\end{aligned} \tag{63}$$

where $\nu = \frac{0.01}{\pi}$ is the kinematic viscosity coefficient. We consider the following physical losses in the experiments:

$$\begin{aligned}
\ell_p(\boldsymbol{\theta}, \mathbf{x}) &= \left(\frac{\partial u_{\boldsymbol{\theta}}}{\partial t} + u_{\boldsymbol{\theta}}\frac{\partial u_{\boldsymbol{\theta}}}{\partial x} - \nu\frac{\partial^2 u_{\boldsymbol{\theta}}}{\partial x^2}\right)^2, \\
\ell_{ic}(\boldsymbol{\theta}, \mathbf{x}) &= \left(u_{\boldsymbol{\theta}}(0,x) + \sin(\pi x)\right)^2, \\
\ell_{bc}(\boldsymbol{\theta}, \mathbf{x}) &= \frac{1}{2}\left[(u_{\boldsymbol{\theta}}(t,-1))^2 + (u_{\boldsymbol{\theta}}(t,1))^2\right].
\end{aligned} \tag{64}$$



(a) 1D-Wave (b) Convection (c) Burgers

Figure 1: Visualisation of the solution $u(t, x)$ for 1D-Wave, Convection and Burgers.

Figure 1 shows the solution map of the Burgers' problem, with a remarkable discontinuity around $x = 0$. This highly contrast physical structure make it hard to precisely predict the outcome in this region, thus causing very large gradients that inflate the bounds significantly.

Heat-2D Multi-Scale (MS) (based on PINNacle (Hao et al., 2024)). The two-dimensional heat equation is a fundamental parabolic PDE that describes the diffusion of heat over time in a spatial domain. The Heat-2D MS problem considers anisotropic diffusion with substantially different diffusion coefficients along the two spatial dimensions. Moreover, the high-frequency oscillations along the x-direction introduce a multiscale spatial structure, making this problem a challenging benchmark for physics-informed learning. The Heat-2D MS equation is expressed as

$$\begin{aligned} \frac{\partial u}{\partial t} &= a_x \frac{\partial^2 u}{\partial x^2} + a_y \frac{\partial^2 u}{\partial y^2}, \\ u(0, x, y) &= \sin(20\pi x)\sin(\pi y), \\ u(t, 0, y) &= u(t, 1, y) = 0, \\ u(t, x, 0) &= u(t, x, 1) = 0, \end{aligned} \tag{65}$$

where $a_x = \frac{1}{(500\pi)^2}$ and $a_y = \frac{1}{\pi^2}$ are the diffusion coefficients. The analytical solution is given by $u(t, x, y) = \sin(20\pi x)\sin(\pi y)\exp\left[-\left(a_x(20\pi)^2 + a_y\pi^2\right)t\right]$, where $(x, y) \in [0, 1]^2$ and $t \in [0, 5]$ (see Figure 2). We consider the following physical losses in the experiments:

$$\begin{aligned} \ell_p(\boldsymbol{\theta}, \mathbf{x}) &= \left(\frac{\partial u_{\boldsymbol{\theta}}}{\partial t} - a_x \frac{\partial^2 u_{\boldsymbol{\theta}}}{\partial x^2} - a_y \frac{\partial^2 u_{\boldsymbol{\theta}}}{\partial y^2}\right)^2, \\ \ell_{ic}(\boldsymbol{\theta}, \mathbf{x}) &= \left(u_{\boldsymbol{\theta}}(0, x, y) - \sin(20\pi x)\sin(\pi y)\right)^2, \\ \ell_{bc}(\boldsymbol{\theta}, \mathbf{x}) &= \frac{1}{4}\left[\left(u_{\boldsymbol{\theta}}(t, 0, y)\right)^2 + \left(u_{\boldsymbol{\theta}}(t, 1, y)\right)^2 + \left(u_{\boldsymbol{\theta}}(t, x, 0)\right)^2 + \left(u_{\boldsymbol{\theta}}(t, x, 1)\right)^2\right]. \end{aligned} \tag{66}$$

Helmholtz (SPINN, Cho et al., 2023). To further evaluate the proposed bounds in a higher-dimensional setting, we consider the three-dimensional Helmholtz equation

$$\nabla^2 u + k^2 u = f(\mathbf{x}), \qquad \mathbf{x} = (x, y, z) \in [0, 1]^3, \tag{67}$$

with homogeneous Dirichlet boundary conditions. We use the manufactured exact solution $u(x, y, z) = \sin(A_1\pi x)\sin(A_2\pi y)\sin(A_3\pi z)$, illustrated by Fig. 3, where the corresponding source term is

$$f(x, y, z) = \left[k^2 - \pi^2(A_1^2 + A_2^2 + A_3^2)\right] u(x, y, z). \tag{68}$$

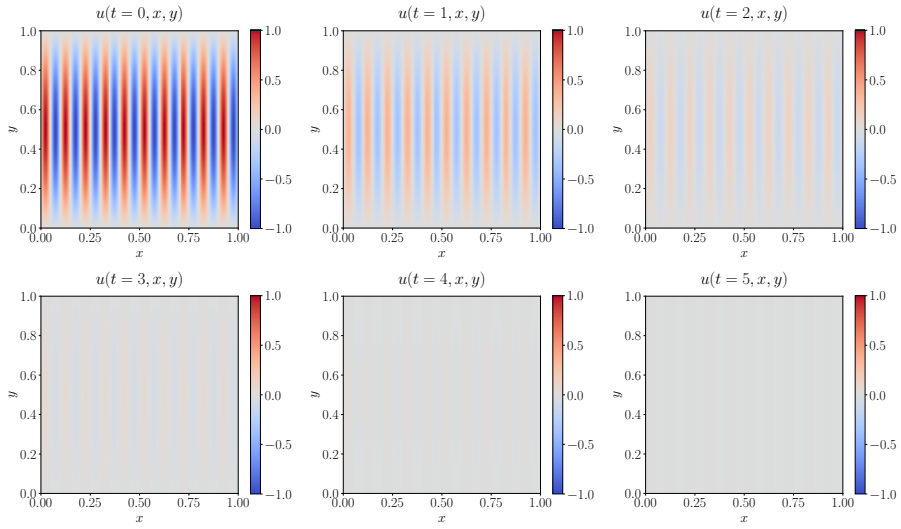


Figure 2: Visualisation of the solution $u(t, x, y)$ for the Heat-2D MS.

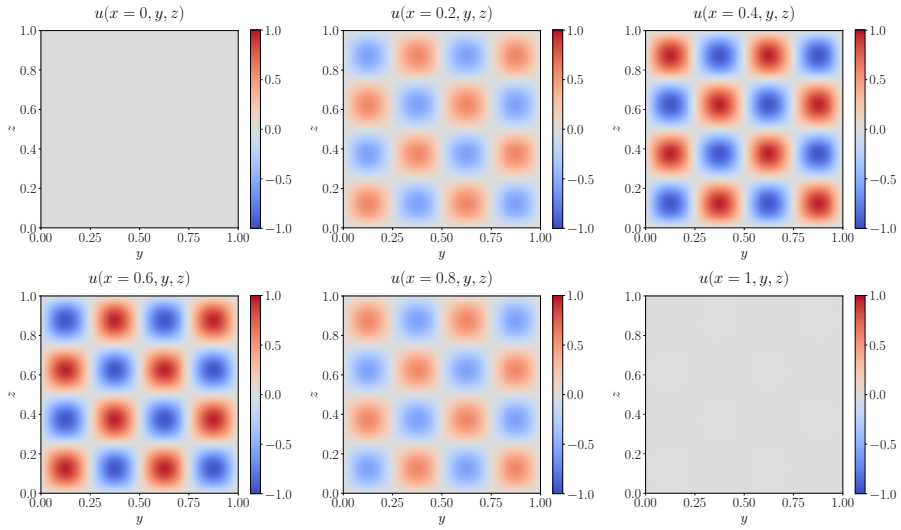


Figure 3: Visualisation of the solution $u(x, y, z)$ for the Helmholtz.

In our experiments, we set $A_1 = A_2 = A_3 = 4$ and $k = 1$. We consider the following physical losses in the experiments:

$$\begin{aligned}
\ell_p(\boldsymbol{\theta}, \mathbf{x}) &= \left(\frac{\partial^2 u_{\boldsymbol{\theta}}}{\partial x^2} + \frac{\partial^2 u_{\boldsymbol{\theta}}}{\partial y^2} + \frac{\partial^2 u_{\boldsymbol{\theta}}}{\partial z^2} + k u_{\boldsymbol{\theta}} - f(x, y, z) \right)^2, \\
\ell_{ic}(\boldsymbol{\theta}, \mathbf{x}) &= \left(u_{\boldsymbol{\theta}}(0, y, z) \right)^2, \\
\ell_{bc}(\boldsymbol{\theta}, \mathbf{x}) &= \frac{1}{5} \Big[\left(u_{\boldsymbol{\theta}}(1, y, z) \right)^2 + \left(u_{\boldsymbol{\theta}}(x, 0, z) \right)^2 + \left(u_{\boldsymbol{\theta}}(x, 1, z) \right)^2 \\
&\qquad + \left(u_{\boldsymbol{\theta}}(x, y, 0) \right)^2 + \left(u_{\boldsymbol{\theta}}(x, y, 1) \right)^2 \Big].
\end{aligned} \tag{69}$$

Flow mixing-2D (SPINN, Cho et al., 2023). This is a time-dependent transport equation that describes the mixing of two fluids at an interface under a spatially-varying rotational velocity field.

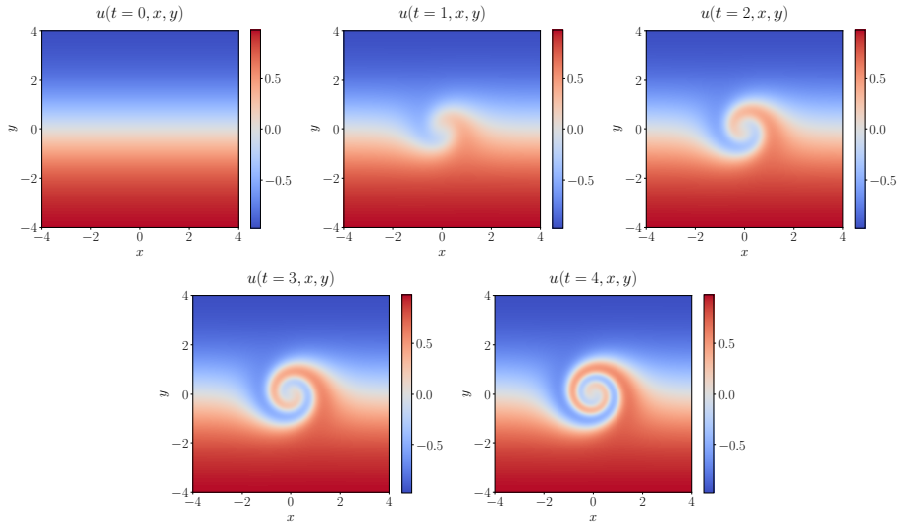


Figure 4: Visualisation of the solution $u(t, x, y)$ for the Flow Mixing 2D.

The problem provides a challenging benchmark due to the nonlinear spatial structure of the velocity field and the progressive deformation of the fluid interface over time. Following the setting in SPINN, the governing equation is expressed as:

$$\begin{aligned} \frac{\partial u}{\partial t} + a(x,y)\frac{\partial u}{\partial x} + b(x,y)\frac{\partial u}{\partial y} = 0, \qquad & (t,x,y) \in [0,4] \times [-4,4]^2, \\ u(0,x,y) = -\tanh\left(\frac{y}{2}\right), & \\ u(t,x,y) = u_{\text{sol}}(t,x,y), \qquad & (x,y) \in \partial[-4,4]^2, \end{aligned} \tag{70}$$

where $a(x,y) = -\frac{v_t}{v_{t,\max}}\frac{y}{r}$; $b(x,y) = \frac{v_t}{v_{t,\max}}\frac{x}{r}$; $v_t = \operatorname{sech}^2(r)\tanh(r)$; with $r = \sqrt{x^2+y^2}$. We set $v_{t,\max} = 0.385$. The initial and boundary conditions are obtained from the analytical solution

$$u_{\text{sol}}(t,x,y) = -\tanh\left(\frac{y}{2}\cos(\omega t) - \frac{x}{2}\sin(\omega t)\right), \qquad \omega = \frac{1}{r}\frac{v_t}{v_{t,\max}}. \tag{71}$$

Fig. 4 illustrates the solution map at different time instant. The apparent singularity of ω at $r = 0$ is removable, with $\omega(0,0) = 1/v_{t,\max}$. We consider the following physical losses in the experiments:

$$\begin{aligned} \ell_p(\boldsymbol{\theta}, \mathbf{x}) &= \left(\frac{\partial u_{\boldsymbol{\theta}}}{\partial t} + a(x,y)\frac{\partial u_{\boldsymbol{\theta}}}{\partial x} + b(x,y)\frac{\partial u_{\boldsymbol{\theta}}}{\partial y}\right)^2, \\ \ell_{ic}(\boldsymbol{\theta}, \mathbf{x}) &= \left(u_{\boldsymbol{\theta}}(0,x,y) + \tanh\left(\frac{y}{2}\right)\right)^2, \\ \ell_{bc}(\boldsymbol{\theta}, \mathbf{x}) &= \frac{1}{4}\Big[\big(u_{\boldsymbol{\theta}}(t,-4,y) - u_{\text{sol}}(t,-4,y)\big)^2 + \big(u_{\boldsymbol{\theta}}(t,4,y) - u_{\text{sol}}(t,4,y)\big)^2 \\ &\qquad + \big(u_{\boldsymbol{\theta}}(t,x,-4) - u_{\text{sol}}(t,x,-4)\big)^2 + \big(u_{\boldsymbol{\theta}}(t,x,4) - u_{\text{sol}}(t,x,4)\big)^2\Big] \end{aligned} \tag{72}$$

C.2 Implementation Details

Data setting. Throughout the experiments, our primary objective is to numerically validate the proposed PAC-Bayes bounds rather than to pursue state-of-the-art predictive performance. For problems admitting closed-form solutions (i.e., Wave-1D, Convection, Heat-2D MS, Helmholtz and Flow Mixing), we uniformly sample 70k collocation points from the domains associated with each physical loss component, namely, Ω for ℓ_p, Ω_0 for ℓ_i, and $\partial\Omega$ for ℓ_b, in either the 2D (t, x) or 3D $((t, x, y)$ or $(x, y, z))$ domain. These points are partitioned into prior, posterior, test, and calibration sets of

10k, 20k, 30k, and 10k samples, respectively. For the data-fidelity loss ℓ_d, we use the same posterior, test, and calibration sets, while reserving only 100 samples for the prior set to estimate the reference loss, gradient quantities and K_{factor} (see Section B). For Burgers, the ground-truth solution is available only at 25,600 discrete points rather than in closed form. We therefore partition these observational data into prior, posterior, test, and calibration sets of 100, 10k, 10k, and 5.5k samples, respectively. For the physical collocation points of Burgers, we instead allocate 5k, 10k, 15k, and 5k samples to the prior, posterior, test, and calibration sets, respectively, such that the posterior set has the same size as the observational posterior set in the balanced setting. In the unbalanced setting, we vary the number of observational posterior samples by selecting a fixed subset consisting of the first $m_d^{\text{posterior}}$ samples of the corresponding posterior set. Similarly, when varying the number of observational calibration samples, we select the first $m_d^{\text{calibration}}$ samples of the calibration set. For each benchmark, all datasets and their partitions are generated once and reused across experimental configurations to ensure a consistent evaluation protocol.

Model and training setting. We use a standard PINN with three hidden layers of width 256 and a scalar output, with a $\tanh$ activation function applied after each hidden layer. The resulting network contains approximately $d_{\boldsymbol{\theta}} \approx 1.3 \times 10^5$ trainable parameters. We normalise the model inputs using the mean and standard deviation determined by the corresponding ranges of the domain Ω. In Step 1, the prior model is initialised from a uniform distribution scaled by $1/\sqrt{\text{fan}_{\text{in}}}$, where fan_{in} denotes the number of input connections to the neurons in the corresponding layer. The prior is then trained with the NTK scheme (Wang et al., 2022) for $N_T = 20\text{k}$ iterations for Wave-1D, Convection and Flow Mixing, and $N_T = 30\text{k}$ iterations for the rest. For each random seed, we train a single prior model, which is subsequently used both to estimate the constants entering the bounds and to initialise the posterior optimisation for all bounds in Step 2. Consequently, all bounds are evaluated using the same estimated constants and posterior initialisation, ensuring a consistent comparison.

In Step 2, we optimise each bound with the confidence parameter fixed at $\delta = 0.1$ and the variance fixed at $\sigma^2 = 10^{-6}$ for both the prior distribution π and posterior distribution ρ. The parameters of the posterior distribution are optimised for $N_{T'} = 5\text{k}$ iterations. We use the Adam optimiser with a batch size of 300 in both steps. The learning rate is initialised at 10^{-3} in Step 1 and 10^{-7} in Step 2, with an exponential decay factor of 0.95 applied every 1,000 iterations. The two partial derivatives of the kl_+^{-1} are defined based on (Viallard et al., 2024). All experiments are implemented in PyTorch (Paszke et al., 2019) and NumPy. For both calibration and certification of the PAC-Bayes bounds, the expected empirical risks are estimated via Monte Carlo sampling using 1,000 models drawn from the posterior distribution. Each experiment is repeated with five independent random seeds, with each seed fixed throughout the entire training, calibration, and certification pipeline. The reported variability therefore reflects the stochasticity of the complete procedure, rather than that of an individual stage in isolation. All experiments are conducted on a single NVIDIA RTX 5090 GPU.

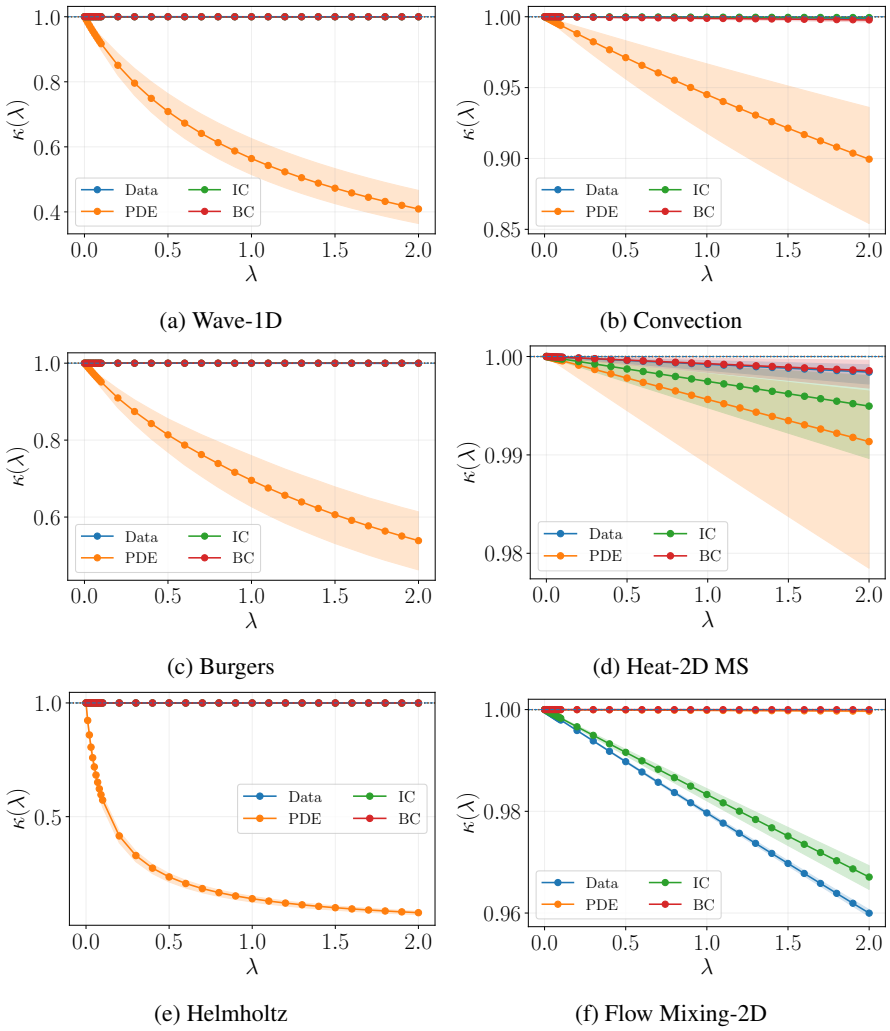
C.3 Empirical assessment of the population-level assumptions

Assumptions 1–3 impose population-level regularity conditions that cannot in general be verified uniformly from finite samples. Our certificates are therefore conditional on their validity. In practice, we assess their plausibility using data independent of posterior training and models sampled throughout the localised hypothesis region Θ_{loc}. Specifically, we examine the CGF-to-variance ratios underlying Assumption 1, the observed input-gradient energies relative to the prescribed envelopes U_i in Assumption 2, and the ratio $C(\rho)/C_{K,G}(\rho)$ in Assumption 3. These diagnostics are used to select problem-dependent constants and to stress-test the assumptions over the empirically relevant region, but should not be interpreted as uniform certification of the corresponding population conditions. Moreover, all constants are selected using S^{prior}; S^{cal} is used only for held-out diagnostic assessment and final PAC-Bayes calibration.

C.3.1 Empirical Assessment of the Assumption 1

We evaluate the empirical ratio $\kappa_i(\lambda)$ over the considered range of λ using models sampled throughout Θ_{loc}. For each seed-specific trained prior and loss component, we set $\kappa_{p,i}$ to the maximum observed ratio over the considered models and λ-values. In almost all cases, the maximum is attained as $\lambda \to 0^+$, where the ratio approaches 1, in which case we conservatively set $\kappa_{p,i} = 1$. Only a small number of seed–loss combinations require values slightly larger than one. Figure 5 reports the mean

empirical ratio across random seeds together with the corresponding minimum–maximum range. The data, IC, and BC components remain particularly close to one, whereas the PDE component generally decreases more noticeably with λ and exhibits greater variability across seeds.



(a) Wave-1D (b) Convection

(c) Burgers (d) Heat-2D MS

(e) Helmholtz (f) Flow Mixing-2D

Figure 5: Empirical CGF-to-variance ratio $\kappa_i(\lambda)$ as a function of λ for each loss component.

C.3.2 Empirical Assessment of the Assumption 2

We empirically stress-test the prescribed gradient envelopes U_i using calibration inputs and models sampled throughout Θ_{loc}. Figure 6–11 shows typical distributions of the observed input-gradient energies, together with their mean, reference threshold G_i, maximum observed value, and prescribed envelope U_i. In general, the data, IC, and BC components exhibit strongly concentrated gradient distributions, whereas the PDE component displays a substantially broader and heavier-tailed distribution spanning several orders of magnitude. Nevertheless, the maximum observed gradient remains below U_i for every component and across the considered models. In particular, the PDE component provides the most stringent empirical stress test, with its observed maximum approaching the envelope more closely than the typical gradient scale. These results support the choice $U_i = U_{G,\text{factor}} G_{\text{ref},i}$ with $U_{G,\text{factor}} \gg 1$, while illustrating why U_i is used only as a concentration envelope rather than as a representative scale of the empirical input-gradient complexity.

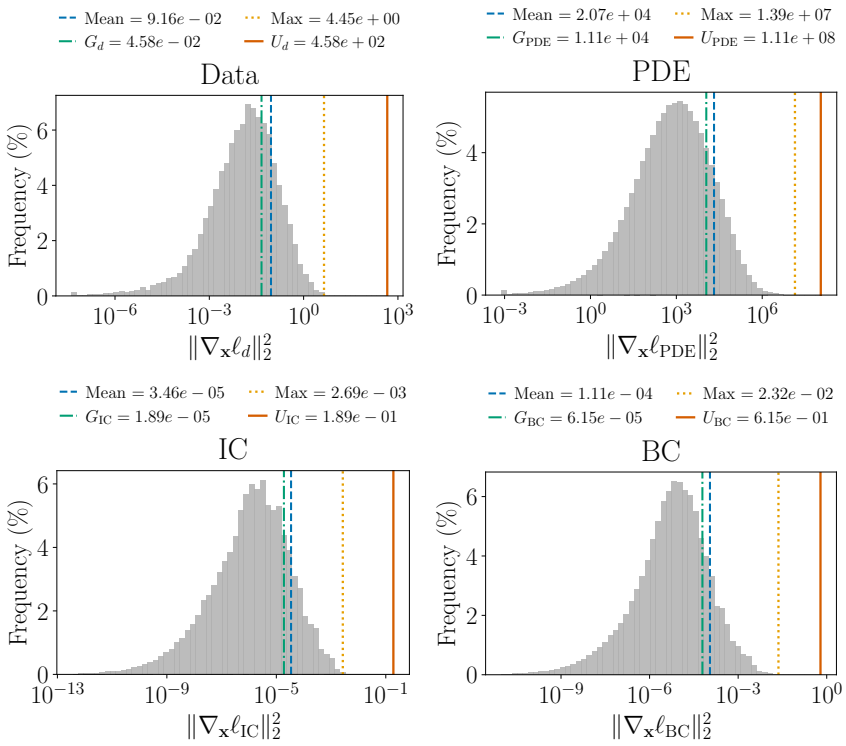


Figure 6: Empirical input-gradient energy distributions for Wave-1D, with a common scaling factor of $U_{G,\text{factor}} = 10^4$ applied to all loss components.

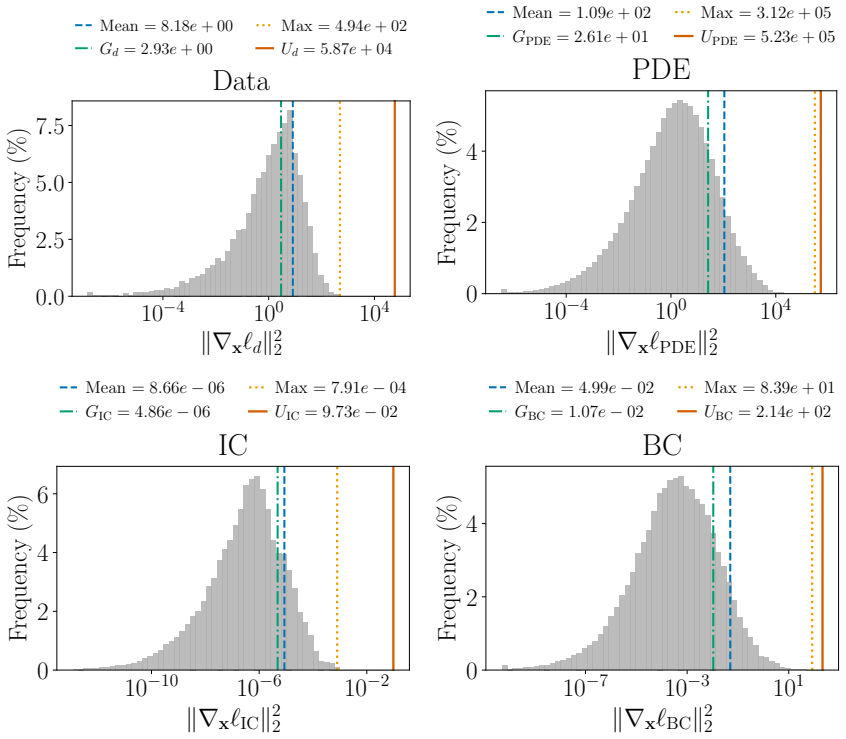


Figure 7: Empirical input-gradient energy distributions for Convection, with a common scaling factor of $U_{G,\text{factor}} = 2 \times 10^4$ applied to all loss components.

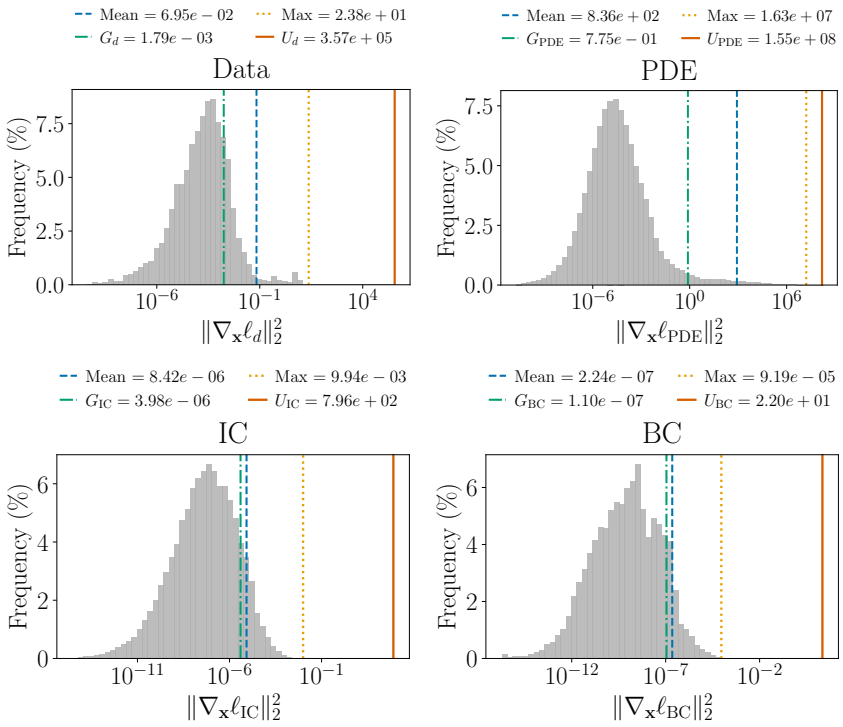


Figure 8: Empirical input-gradient energy distributions for Burgers, with a common scaling factor of $U_{G,\text{factor}} = 2 \times 10^8$ applied to all loss components.

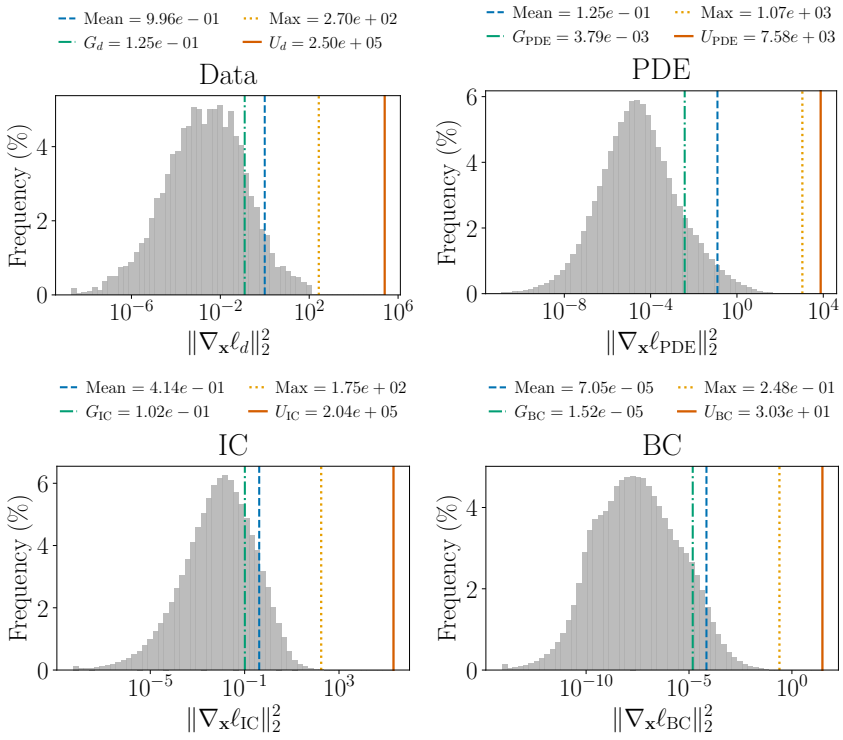


Figure 9: Empirical input-gradient energy distributions for Heat-2D MS, with a common scaling factor of $U_{G,\text{factor}} = 2 \times 10^6$ applied to all loss components.

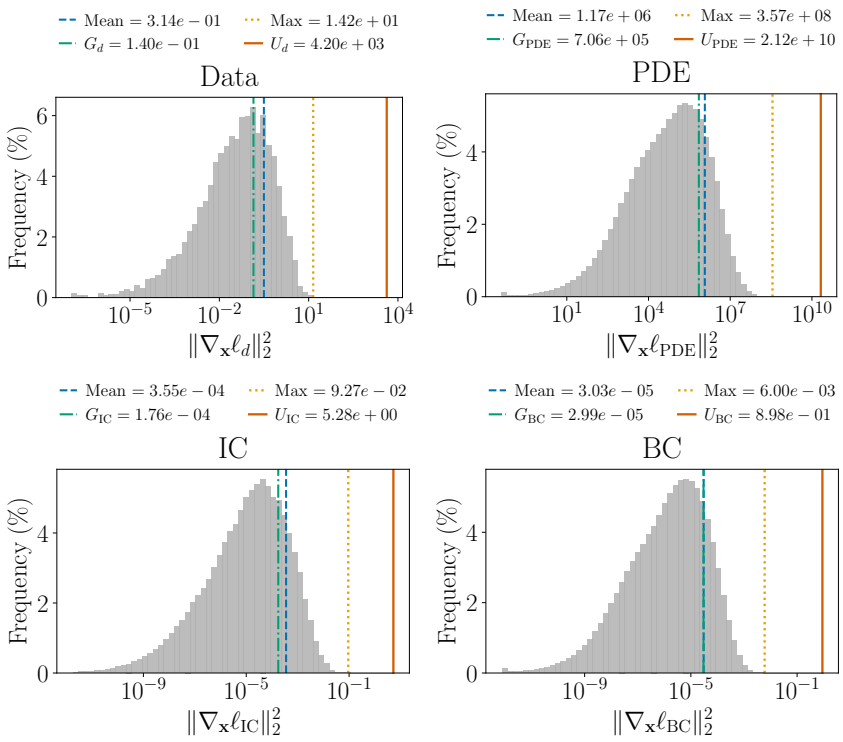


Figure 10: Empirical input-gradient energy distributions for Helmholtz, with a common scaling factor of $U_{G,\text{factor}} = 3 \times 10^4$ applied to all loss components.

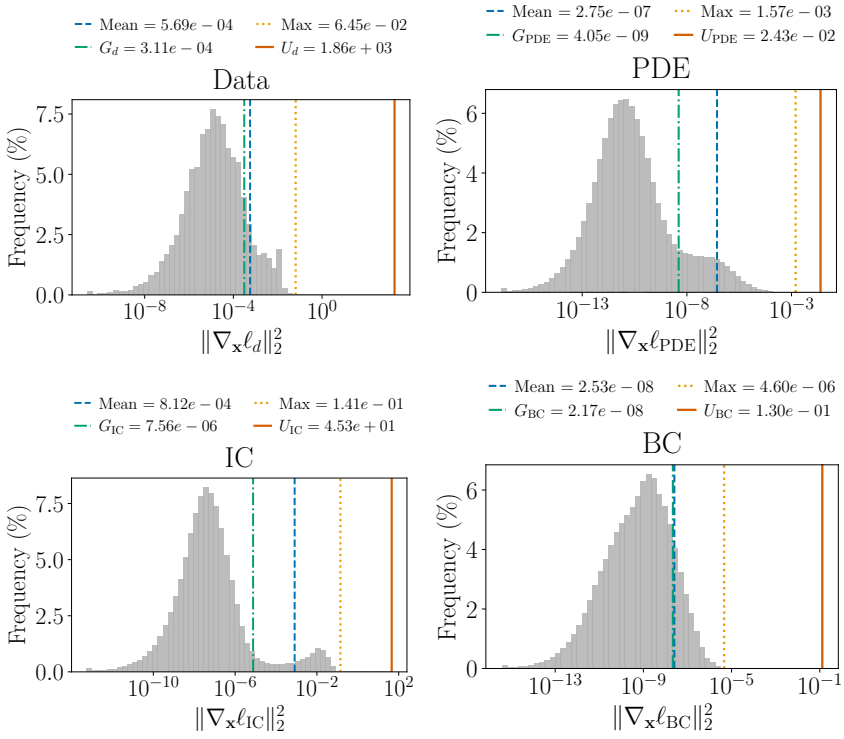
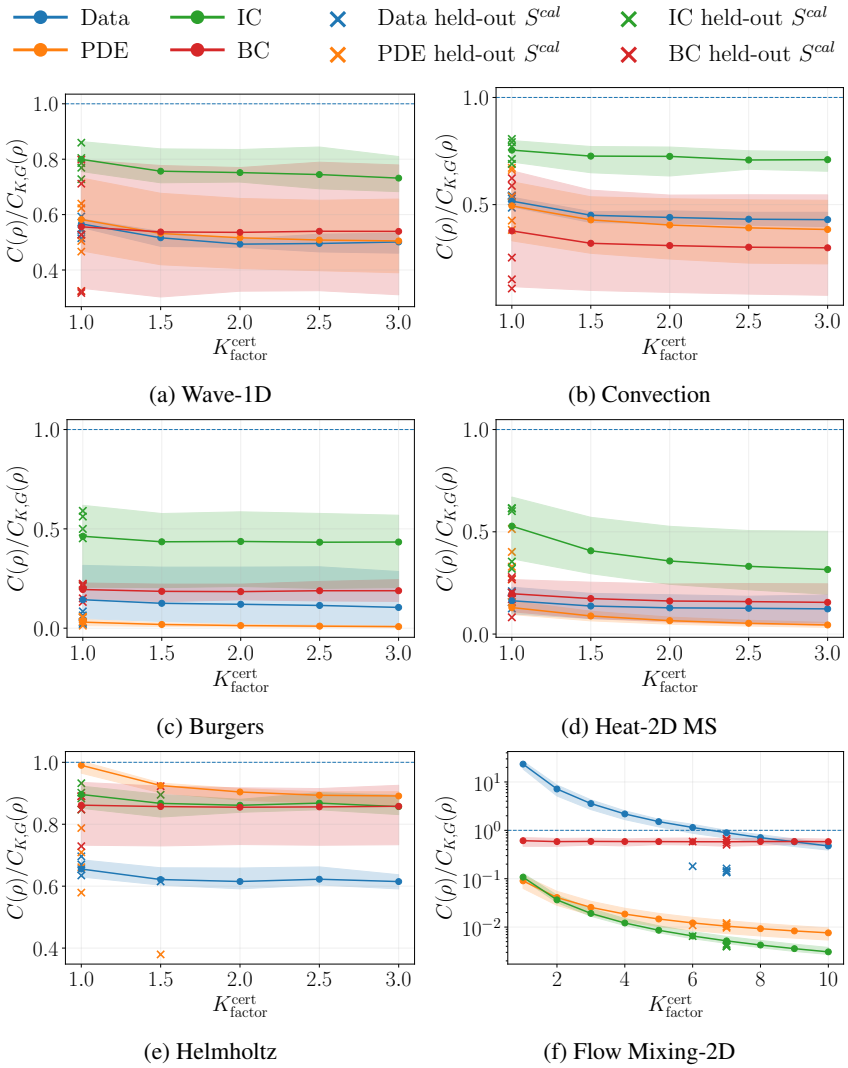


Figure 11: Empirical input-gradient energy distributions for Flow Mixing-2D, with a common scaling factor of $U_{G,\text{factor}} = 6 \times 10^6$ applied to all loss components.

C.3.3 Empirical Assessment of the Assumption 3

As described in Section B, we set $K_i = K_{\text{factor}} K_{\text{ref},i}$ and fix $G_i = G_{\text{ref},i}$, selecting K_{factor} by empirically assessing the calibration-dominance condition. Specifically, we evaluate the empirical ratio $C_S(\rho)/C_{S,K,G}(\rho)$ over a range of $K^{\text{cert}}_{\text{factor}}$ values knowing that assumption 3 requires this ratio to remain below one. The $K^{\text{cert}}_{\text{factor}}$ is chosen from the prior set S^{prior} as the smallest value where such ratio is below 1 for all losses, then we assess the ratio at that value on the held-out dataset S^{calib} to see whether it also holds on unseen data. Figure 12 reports the mean ratio across random seeds, together with the corresponding minimum-maximum range. With the exception of Flow Mixing, the empirical ratios remain below one for all loss components over the entire range considered and decrease consistently as K_{factor} increases. This behavior is expected: as the clipping thresholds increase, the clipped losses progressively approach their unclipped counterparts, increasing the numerator of $C_{K,G}(\rho)$, while its denominator remains clipped at the same threshold, hence resulting in larger $C_{K,G}(\rho)$ and consequently smaller ratio $C(\rho)/C_{K,G}(\rho)$. It is noteworthy mentioning that the S^{calib}-evaluated ratios at the chosen K_{factor} values are consistently smaller than 1, indicating that the chosen values allows the assumption to be satisfied on unseen data as well.



(a) Wave-1D (b) Convection

(c) Burgers (d) Heat-2D MS

(e) Helmholtz (f) Flow Mixing-2D

Figure 12: The empirical calibration-dominance ratio $C(\rho)/C_{K,G}(\rho)$ as a function of the $K^{\text{cert}}_{\text{factor}}$ evaluated on S^{prior}, and validated on held-out S^{calib}.

C.4 Full Main Results

Compared PAC-Bayes bounds. We evaluate both certificates developed in Section 3 and compare them with two conventional PAC-Bayes alternatives constructed under the same learned prior and posterior distributions and the same overall confidence level. We denote the component-wise certificate of Theorem 6 by *ours-union* and its joint multi-task refinement in Theorem 9 by *ours-joint*. Both proposed bounds rely on the same calibrated input-gradient regularity and control the original, potentially unbounded PIML losses through the empirical input-gradient energies. They both arise from the same theoretical framework, with the latter exploiting the joint multi-task structure to avoid repeatedly paying the PAC-Bayesian complexity across the loss components.

A broad numerical comparison with existing PAC-Bayes bounds is inherently challenging because different results rely on different assumptions on the loss tails, variance, or regularity. Moreover, some bounds involve problem-dependent constants for which a directly computable finite-sample estimation procedure is not specified, so their numerical evaluation would require additional assumptions or researcher-chosen approximations that can materially affect the resulting certificate. We therefore focus on representative baselines whose required quantities can be instantiated under a common experimental protocol, allowing the reported certificates to be compared on equal computational footing. In this respect, our work also contributes to the relatively limited literature on end-to-end computable PAC-Bayes certificates for physics-informed neural networks with potentially unbounded losses: beyond deriving the theoretical bound, we provide an explicit calibration procedure for the problem-dependent quantities entering the final certificate. For comparison, we additionally consider two PAC-Bayes baselines. The first baseline, denoted by *bounded-u*, is a conventional bounded-loss PAC-Bayes certificate adapted to the same sample-centric multi-task construction as our joint bound. For each component, we assume a conservative loss envelope $0 \le \ell_i \le U_{\ell,i}$. Throughout the experiments, we set $U_{\ell,i} = 10^3$ for all components and benchmarks, except for Helmholtz, for which we use $U_{\ell,i} = 5 \times 10^3$ to accommodate its substantially larger PDE losses ($> 10^3$). These ranges are chosen conservatively relative to the loss scales observed over the localised hypothesis region, typically exceeding the observed loss values by approximately one order of magnitude. For a jointly treated block $\mathcal{I}$, we define the common envelope $U_{\mathcal{I}} := \max_{i\in\mathcal{I}} U_{\ell,i}$, such that $\ell_i/U_{\mathcal{I}} \in [0,1]$ for every $i \in \mathcal{I}$. Letting $M_{\mathcal{I}} := \sum_{i\in\mathcal{I}} m_i$, the corresponding sample-weighted empirical normalised risk is

$$\widehat{q}_{\mathcal{I}}(\rho) := \frac{1}{M_{\mathcal{I}} U_{\mathcal{I}}} \sum_{i\in\mathcal{I}} m_i \widehat{\mathcal{R}}_i(\rho).$$

According to (Zakerinia & Lampert, 2025), applying the binary-kl PAC-Bayes inequality jointly to the $M_{\mathcal{I}}$ samples gives, with confidence at least $1-\delta_{\mathcal{I}}$,

$$\mathcal{R}^{\mathcal{S}}_{\mathcal{I}}(\rho) \le U_{\mathcal{I}} \mathrm{kl}_+^{-1}\left(\widehat{q}_{\mathcal{I}}(\rho), \frac{D_{\mathrm{KL}}(\rho\|\pi) + \log\frac{2\sqrt{M_{\mathcal{I}}}}{\delta_{\mathcal{I}}}}{M_{\mathcal{I}}}\right).$$

When the components within $\mathcal{I}$ have equal sample sizes, multiplying this sample-weighted bound by $|\mathcal{I}|$ recovers their equally weighted population risk. Accordingly, in the balanced regime we apply a single binary-kl event jointly to all data and physics components. In the data-scarce regime, we bound the observational component separately and apply one joint sample-centric bound to the physical block containing the PDE, IC, and BC losses, with confidence budgets $\delta_d = \delta/4$ and $\delta_{\mathrm{phy}} = 3\delta/4$. Thus, *bounded-u* follows the same multi-task aggregation used by our joint construction, but does not exploit input-gradient regularity and instead controls the losses directly through conservative uniform envelopes.

The second baseline, denoted by *sub-gaussian*, similarly follows the sample-centric multi-task construction but retains the calibrated relation between loss concentration and input-gradient regularity. From Corollary 5, with probability at least $1-\delta_C$, each component satisfies

$$\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\Lambda^{(i)}_{\boldsymbol{\theta}}(\lambda)] \le \frac{\lambda^2}{2}\overline{C}_i(\rho,\delta_C)U_{G,i},$$

corresponding to the conservative variance proxy

$$\sigma_i^2(\rho,\delta_C) := \overline{C}_i(\rho,\delta_C)U_{G,i}.$$

For a jointly treated block $\mathcal{I}$, these component-specific proxies are aggregated according to their sample sizes as

$$\mathcal{G}_{U,\mathcal{I}}(\rho, \delta_C) := \sum_{i \in \mathcal{I}} m_i \sigma_i^2(\rho, \delta_C) = \sum_{i \in \mathcal{I}} m_i \overline{C}_i(\rho, \delta_C) U_{G,i}.$$

Substituting this quantity into the joint PAC-Bayes-Chernoff argument and optimising over $\lambda > 0$ yields the sample-weighted complexity

$$\frac{1}{M_{\mathcal{I}}} \sqrt{2\, \mathcal{G}_{U,\mathcal{I}}(\rho, \delta_C) K_{\mathcal{I}}(\rho, \pi, \delta_G)}.$$

As for *bounded-u*, the balanced regime treats all components jointly and rescales the resulting sample-weighted certificate to the equally weighted PIML risk. In the data-scarce regime, the observational loss is treated separately while the physical components are aggregated into a single joint block, using $\delta_d = \delta/4$ and $\delta_{\text{phy}} = 3\delta/4$. Within each block, the calibration event receives $\delta_C = \delta_{\mathcal{I}}/3$, while the remaining $2\delta_{\mathcal{I}}/3$ is allocated to the PAC-Bayes generalisation event since no empirical-gradient concentration event is required. Hence, *sub-gaussian* shares both the calibrated regularity information and the sample-centric multi-task structure of *ours-joint*, but replaces the posterior-averaged empirical input-gradient energy by its conservative envelope $U_{G,i}$.

Consequently, the four certificates provide a progressively controlled comparison. In particular, *ours-union* is the component-wise certificate of Theorem 6, while *ours-joint* is its sample-centric multi-task refinement in Theorem 9. The *sub-gaussian* baseline retains this joint sample-centric structure and the calibrated input-gradient regularity, but removes the empirical adaptivity of the gradient complexity. Finally, *bounded-u* retains the joint sample-centric treatment but removes the input-gradient regularity altogether, relying instead on conservative loss envelopes. The resulting comparisons therefore distinguish the gain from the joint refinement, the gain from using empirical rather than worst-case input-gradient information, and the gain from exploiting input-gradient regularity in the first place.

Overall benchmark results. Tables 3–6 report the complete PAC-Bayes results over the six benchmark problems. In the balanced regime, *ours-joint* yields bounds of 6.46×10^{-1}, 3.12×10^{-2}, 6.01×10^{-1}, 6.26×10^{-3}, 9.16, and 3.98×10^{-2} on Wave-1D, Convection, Burgers, Heat-2D MS, Helmholtz, and Flow Mixing, respectively. These are consistently tighter than the bounded-U and sub-Gaussian alternatives, often by one or more orders of magnitude. The improvement is particularly noticeable on Heat-2D MS and Flow Mixing, where the joint certificates are respectively about $71\times$ and $11\times$ tighter than bounded-U, and $28\times$ and $36\times$ tighter than the sub-Gaussian baseline. The proposed posteriors also generally attain comparable or lower empirical risks, indicating that the tighter certificates are not obtained by sacrificing predictive performance.

The comparison between *ours-joint* and *ours-union* isolates the benefit of the joint multi-task refinement. The two are nearly identical on Wave-1D, Convection, Burgers, and Helmholtz, while clearer gains appear on Heat-2D MS and Flow Mixing where the observational loss's contribution to the total risk is considerable ($\frac{3.92\times10^{-4}}{2.05\times10^{-3}} \approx 20\%$ for the former and $\frac{1.65\times10^{-3}}{3.09\times10^{-3}} \approx 53\%$ for the latter). In the balanced regime, the bound decreases from 9.60×10^{-3} to 6.26×10^{-3} on Heat-2D MS and from 5.58×10^{-2} to 3.98×10^{-2} on Flow Mixing, reductions of approximately 35% and 29%, respectively. On Heat-2D MS, this tightening is accompanied by a slightly lower test risk (2.05×10^{-3} versus 2.09×10^{-3}).

An important factor governing absolute certificate tightness is the regularity of the underlying solution. Since our complexity depends explicitly on input-gradient energies, problems with sharp or highly varying solution structures naturally incur larger complexity. This is visible for Burgers, whose shock-like region produces a substantially larger gap between empirical risk and certificate than smoother problems such as Wave-1D or Heat-2D MS. Thus, the certificate reflects not only predictive error but also the local regularity of the learned solution.

The data-scarce regime provides a more stringent assessment, restricting the observational component to only 100 posterior-training and 100 independent calibration samples. Predictive performance remains remarkably stable: for example, the joint test risk changes from 2.05×10^{-3} to 2.24×10^{-3} on Heat-2D MS and from 3.22×10^{-3} to 3.23×10^{-3} on Burgers. The certificates nevertheless increase to 5.74×10^{-1} and 1.89, respectively, reflecting the reduced statistical evidence available

Table 3: PAC-Bayes benchmark results for Wave-1D in the balanced and data-scarce regimes. The balanced regime uses the full available datasets for posterior learning and calibration, whereas the data-scarce regime restricts the observational component to 100 samples for posterior learning and 100 independent samples for calibration, while retaining the full physical datasets. For clarity, we highlight the value with blue for smallest empirical risks and bound values.

Data regime	Quantity	Ours-joint	Ours-union	Bounded-U	Sub-Gaussian
Balanced	Train risk ($\widehat{\mathcal{R}}_d$)	(1.51±0.18)e−4	(1.51±0.18)e−4	(1.53±0.20)e−4	(1.55±0.21)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(4.30±0.52)e−1	(4.30±0.52)e−1	(4.44±0.53)e−1	(4.62±0.54)e−1
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(1.19±0.11)e−4	(1.19±0.11)e−4	(1.21±0.12)e−4	(1.23±0.14)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(1.25±0.12)e−4	(1.25±0.12)e−4	(1.27±0.12)e−4	(1.28±0.13)e−4
	Train risk (total)	(4.31±0.52)e−1	(4.31±0.52)e−1	(4.44±0.53)e−1	(4.62±0.54)e−1
	Test risk ($\widehat{\mathcal{R}}_d$)	(1.51±0.18)e−4	(1.51±0.18)e−4	(1.53±0.20)e−4	(1.55±0.21)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(4.29±0.52)e−1	(4.29±0.52)e−1	(4.43±0.53)e−1	(4.61±0.54)e−1
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(1.19±0.11)e−4	(1.19±0.11)e−4	(1.21±0.12)e−4	(1.23±0.14)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(1.25±0.11)e−4	(1.25±0.11)e−4	(1.27±0.12)e−4	(1.28±0.13)e−4
	Test risk (total)	(4.30±0.52)e−1	(4.30±0.52)e−1	(4.43±0.53)e−1	(4.61±0.54)e−1
	KL($\rho\|\pi$)	(8.59±1.67)e−1	(8.58±1.67)e−1	(2.31±0.81)e−1	(1.50±0.57)e−2
	Bound	(6.46±0.75)e−1	(6.46±0.75)e−1	(1.40±0.09)e0	(8.05±0.73)e0
	Bound w/o tail	(4.80±0.63)e−1	(4.80±0.63)e−1	-	-
Data-scarce	Train risk ($\widehat{\mathcal{R}}_d$)	(1.58±0.16)e−4	(1.58±0.16)e−4	(1.61±0.17)e−4	(1.61±0.17)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(4.31±0.52)e−1	(4.31±0.52)e−1	(4.69±0.54)e−1	(4.62±0.54)e−1
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(1.20±0.11)e−4	(1.20±0.11)e−4	(1.23±0.14)e−4	(1.23±0.14)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(1.25±0.12)e−4	(1.25±0.12)e−4	(1.29±0.13)e−4	(1.28±0.13)e−4
	Train risk (total)	(4.31±0.52)e−1	(4.31±0.52)e−1	(4.69±0.54)e−1	(4.62±0.54)e−1
	Test risk ($\widehat{\mathcal{R}}_d$)	(1.51±0.19)e−4	(1.51±0.19)e−4	(1.55±0.21)e−4	(1.55±0.21)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(4.30±0.52)e−1	(4.30±0.52)e−1	(4.68±0.54)e−1	(4.61±0.54)e−1
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(1.20±0.11)e−4	(1.20±0.11)e−4	(1.24±0.14)e−4	(1.23±0.14)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(1.25±0.11)e−4	(1.25±0.11)e−4	(1.28±0.13)e−4	(1.28±0.13)e−4
	Test risk (total)	(4.31±0.52)e−1	(4.31±0.52)e−1	(4.68±0.54)e−1	(4.61±0.54)e−1
	KL($\rho\|\pi$)	(7.94±1.44)e−1	(7.94±1.44)e−1	(1.09±0.53)e−4	(1.45±0.54)e−2
	Bound	(6.57±0.77)e−1	(6.57±0.77)e−1	(6.61±0.01)e1	(8.08±0.74)e0
	Bound w/o tail	(4.81±0.63)e−1	(4.81±0.63)e−1	-	-

for both posterior learning and calibration. This separation illustrates that low empirical risk alone does not imply a tight finite-sample certificate.

Despite this deterioration, the proposed certificates remain substantially tighter than the conservative baselines. In the data-scarce regime, *ours-joint* gives 7.77×10^{-2}, 1.89, 5.74×10^{-1}, and 9.23 on Convection, Burgers, Heat-2D MS, and Helmholtz, compared with 65.2, 65.5, 65.1, and 334 for bounded-U, and 6.14×10^{-1}, 21.8, 2.08, and 178 for the sub-Gaussian baseline. Flow Mixing is more strongly affected by data scarcity, but its joint bound of 12.1 remains below both bounded-U (65.1) and sub-Gaussian (41.9). Thus, the advantage of the input-gradient construction persists even when observational information is severely limited.

Finally, the oracle "bound w/o tail" results identify the residual-tail correction as the main source of deterioration under data scarcity. In the balanced regime, removing this contribution reduces the joint bounds to 4.80×10^{-1}, 2.13×10^{-2}, 3.60×10^{-2}, 2.32×10^{-3}, 5.34, and 7.98×10^{-3} across the six problems. Under data scarcity, the corresponding values remain only 4.81×10^{-1}, 2.41×10^{-2}, 4.50×10^{-2}, 4.19×10^{-3}, 5.34, and 7.21×10^{-2}, despite much larger full certificates on several problems. This shows that the empirical input-gradient component remains informative with limited data, while the rigorous residual-gradient control becomes the dominant source of conservatism, motivating sharper finite-sample control of this term. The "bound w/o tail" therefore also provides an oracle indication of the certificate tightness that could be approached under substantially sharper control of the residual gradient tail.

C.5 Ablation Study

C.5.1 Impact of the posterior's training sample size

We investigate the effect of the observational posterior-training sample size $m_d^{\mathrm{posterior}}$ on Heat-2D MS by varying it from 10 to the full 20k samples, while retaining the complete calibration set and all physical samples, thereby isolating the effect of posterior-training data scarcity. As shown in Fig. 13,

Table 4: PAC-Bayes benchmark results for Convection in the balanced and data-scarce regimes.

Data regime	Quantity	Ours-joint	Ours-union	Bounded-U	Sub-Gaussian
Balanced	Train risk ($\widehat{\mathcal{R}}_d$)	(4.90±0.62)e−4	(4.91±0.63)e−4	(5.30±0.92)e−4	(5.30±0.92)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(1.77±0.21)e−2	(1.77±0.21)e−2	(1.96±0.32)e−2	(1.96±0.33)e−2
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(4.23±0.96)e−4	(4.24±0.97)e−4	(4.69±1.54)e−4	(4.70±1.54)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(3.89±0.58)e−4	(3.89±0.58)e−4	(4.08±0.77)e−4	(4.08±0.77)e−4
	Train risk (total)	(1.90±0.23)e−2	(1.91±0.23)e−2	(2.10±0.34)e−2	(2.10±0.35)e−2
	Test risk ($\widehat{\mathcal{R}}_d$)	(4.92±0.62)e−4	(4.93±0.63)e−4	(5.32±0.91)e−4	(5.32±0.92)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(1.78±0.22)e−2	(1.79±0.22)e−2	(1.97±0.33)e−2	(1.98±0.33)e−2
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(4.24±0.96)e−4	(4.25±0.97)e−4	(4.70±1.53)e−4	(4.70±1.54)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(3.87±0.58)e−4	(3.87±0.58)e−4	(4.07±0.76)e−4	(4.07±0.76)e−4
	Test risk (total)	(1.91±0.23)e−2	(1.92±0.23)e−2	(2.11±0.35)e−2	(2.12±0.35)e−2
	$\mathrm{KL}(\rho\|\pi)$	(8.79±5.92)e−1	(8.39±5.63)e−1	(1.51±0.97)e−2	(1.45±0.83)e−2
	Bound	(3.12±0.43)e−2	(3.18±0.44)e−2	(5.21±0.12)e−1	(4.62±0.55)e−1
	Bound w/o tail	(2.13±0.31)e−2	(2.15±0.32)e−2	-	-
Data-scarce	Train risk ($\widehat{\mathcal{R}}_d$)	(5.34±0.85)e−4	(5.34±0.85)e−4	(5.45±0.93)e−4	(5.42±0.91)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(1.90±0.28)e−2	(1.90±0.28)e−2	(2.00±0.35)e−2	(1.97±0.33)e−2
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(4.59±1.38)e−4	(4.59±1.38)e−4	(4.78±1.68)e−4	(4.72±1.58)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(4.04±0.72)e−4	(4.04±0.72)e−4	(4.12±0.80)e−4	(4.09±0.78)e−4
	Train risk (total)	(2.04±0.30)e−2	(2.04±0.30)e−2	(2.15±0.38)e−2	(2.12±0.36)e−2
	Test risk ($\widehat{\mathcal{R}}_d$)	(5.25±0.85)e−4	(5.25±0.85)e−4	(5.37±0.95)e−4	(5.33±0.93)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(1.92±0.29)e−2	(1.92±0.29)e−2	(2.02±0.36)e−2	(1.99±0.34)e−2
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(4.59±1.38)e−4	(4.59±1.38)e−4	(4.79±1.68)e−4	(4.73±1.58)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(4.02±0.71)e−4	(4.02±0.71)e−4	(4.10±0.80)e−4	(4.08±0.77)e−4
	Test risk (total)	(2.06±0.31)e−2	(2.06±0.31)e−2	(2.16±0.38)e−2	(2.13±0.36)e−2
	$\mathrm{KL}(\rho\|\pi)$	(8.81±6.35)e−2	(8.74±6.30)e−2	(6.57±0.92)e−6	(6.92±4.03)e−3
	Bound	(7.77±0.72)e−2	(7.81±0.73)e−2	(6.52±0.00)e1	(6.14±0.60)e−1
	Bound w/o tail	(2.41±0.40)e−2	(2.42±0.41)e−2	-	-

Table 5: PAC-Bayes benchmark results for Burgers in the balanced and data-scarce regimes.

Data regime	Quantity	Ours-joint	Ours-union	Bounded-U	Sub-Gaussian
Balanced	Train risk ($\widehat{\mathcal{R}}_d$)	(9.11±3.71)e−5	(9.11±3.71)e−5	(9.12±3.91)e−5	(9.10±3.70)e−5
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(3.21±1.55)e−3	(3.21±1.55)e−3	(3.23±1.57)e−3	(3.23±1.57)e−3
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(4.67±0.54)e−5	(4.67±0.54)e−5	(4.58±0.57)e−5	(4.67±0.54)e−5
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(4.87±0.94)e−5	(4.87±0.94)e−5	(4.89±0.79)e−5	(4.88±0.94)e−5
	Train risk (total)	(3.40±1.57)e−3	(3.40±1.57)e−3	(3.41±1.58)e−3	(3.42±1.59)e−3
	Test risk ($\widehat{\mathcal{R}}_d$)	(9.59±4.53)e−5	(9.59±4.53)e−5	(9.60±4.75)e−5	(9.58±4.53)e−5
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(3.03±0.99)e−3	(3.03±0.99)e−3	(3.05±0.99)e−3	(3.05±1.00)e−3
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(4.66±0.54)e−5	(4.66±0.54)e−5	(4.58±0.57)e−5	(4.66±0.54)e−5
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(4.87±0.94)e−5	(4.87±0.94)e−5	(4.89±0.79)e−5	(4.88±0.94)e−5
	Test risk (total)	(3.22±1.01)e−3	(3.22±1.01)e−3	(3.24±1.02)e−3	(3.24±1.02)e−3
	$\mathrm{KL}(\rho\|\pi)$	(1.27±1.43)e−3	(1.23±1.39)e−3	(4.49±2.69)e−4	(3.48±1.29)e−6
	Bound	(6.01±0.44)e−1	(6.14±0.44)e−1	(8.51±0.09)e−1	(1.75±0.14)e1
	Bound w/o tail	(3.60±2.37)e−2	(3.62±2.38)e−2	-	-
Data-scarce	Train risk ($\widehat{\mathcal{R}}_d$)	(2.08±1.45)e−4	(2.08±1.45)e−4	(2.08±1.47)e−4	(2.08±1.45)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(3.23±1.56)e−3	(3.23±1.56)e−3	(3.24±1.57)e−3	(3.23±1.57)e−3
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(4.66±0.54)e−5	(4.66±0.54)e−5	(4.58±0.57)e−5	(4.67±0.54)e−5
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(4.88±0.94)e−5	(4.88±0.94)e−5	(4.89±0.78)e−5	(4.88±0.94)e−5
	Train risk (total)	(3.53±1.69)e−3	(3.53±1.69)e−3	(3.55±1.70)e−3	(3.54±1.70)e−3
	Test risk ($\widehat{\mathcal{R}}_d$)	(9.58±4.53)e−5	(9.58±4.53)e−5	(9.60±4.75)e−5	(9.58±4.53)e−5
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(3.04±0.99)e−3	(3.04±0.99)e−3	(3.06±1.00)e−3	(3.05±1.00)e−3
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(4.66±0.54)e−5	(4.66±0.54)e−5	(4.58±0.58)e−5	(4.66±0.54)e−5
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(4.88±0.94)e−5	(4.88±0.94)e−5	(4.89±0.78)e−5	(4.88±0.94)e−5
	Test risk (total)	(3.23±1.02)e−3	(3.23±1.02)e−3	(3.25±1.03)e−3	(3.24±1.02)e−3
	$\mathrm{KL}(\rho\|\pi)$	(1.45±1.84)e−4	(1.44±1.83)e−4	(1.98±0.35)e−6	(2.15±0.49)e−6
	Bound	(1.89±0.58)e0	(1.90±0.58)e0	(6.55±0.00)e1	(2.18±0.26)e1
	Bound w/o tail	(4.50±2.73)e−2	(4.50±2.73)e−2	-	-

all certificates become tighter as $m_d^{\mathrm{posterior}}$ increases, with *ours-joint* and *ours-union* improving by more than two orders of magnitude and remaining substantially tighter than the conservative baselines. The gradient-tail contribution accounts for much of the looseness at very small $m_d^{\mathrm{posterior}}$ and rapidly decreases with additional samples, while the total test risk reduces only moderately (but still more notable on ours bounds). This contrast shows that limited observational data have

Table 6: PAC-Bayes benchmark results for Heat-2D MS in the balanced and data-scarce regimes.

Data regime	Quantity	Ours-joint	Ours-union	Bounded-U	Sub-Gaussian
Balanced	Train risk ($\widehat{\mathcal{R}}_d$)	(3.92±0.96)e−4	(3.98±0.98)e−4	(4.13±1.04)e−4	(4.12±1.04)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(3.69±0.84)e−4	(3.73±0.83)e−4	(3.81±0.83)e−4	(3.81±0.83)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(9.33±3.55)e−4	(9.64±3.80)e−4	(1.05±0.47)e−3	(1.05±0.47)e−3
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(3.52±0.96)e−4	(3.58±0.97)e−4	(3.70±1.02)e−4	(3.69±1.01)e−4
	Train risk (total)	(2.05±0.55)e−3	(2.09±0.58)e−3	(2.22±0.67)e−3	(2.21±0.67)e−3
	Test risk ($\widehat{\mathcal{R}}_d$)	(3.90±0.96)e−4	(3.96±0.98)e−4	(4.10±1.04)e−4	(4.09±1.04)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(3.76±0.66)e−4	(3.80±0.68)e−4	(3.88±0.74)e−4	(3.88±0.74)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(9.41±3.59)e−4	(9.72±3.84)e−4	(1.06±0.47)e−3	(1.06±0.47)e−3
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(3.48±0.95)e−4	(3.53±0.96)e−4	(3.66±1.01)e−4	(3.65±1.01)e−4
	Test risk (total)	(2.05±0.54)e−3	(2.10±0.57)e−3	(2.23±0.66)e−3	(2.22±0.66)e−3
	$\mathrm{KL}(\rho\|\pi)$	(1.91±1.13)e−1	(9.07±5.86)e−2	(8.87±8.11)e−4	(1.58±1.02)e−3
	Bound	(6.26±1.67)e−3	(9.60±2.33)e−3	(4.46±0.04)e−1	(1.76±0.49)e−1
	Bound w/o tail	(2.32±0.60)e−3	(2.56±0.70)e−3	-	-
Data-scarce	Train risk ($\widehat{\mathcal{R}}_d$)	(3.77±0.85)e−4	(3.77±0.85)e−4	(3.77±0.85)e−4	(3.77±0.85)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(3.82±0.83)e−4	(3.82±0.83)e−4	(3.82±0.83)e−4	(3.82±0.83)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(1.06±0.48)e−3	(1.06±0.48)e−3	(1.07±0.48)e−3	(1.07±0.48)e−3
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(3.71±1.02)e−4	(3.71±1.02)e−4	(3.71±1.02)e−4	(3.71±1.02)e−4
	Train risk (total)	(2.19±0.66)e−3	(2.19±0.66)e−3	(2.20±0.67)e−3	(2.19±0.67)e−3
	Test risk ($\widehat{\mathcal{R}}_d$)	(4.11±1.05)e−4	(4.11±1.05)e−4	(4.12±1.05)e−4	(4.11±1.05)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(3.89±0.75)e−4	(3.89±0.75)e−4	(3.89±0.75)e−4	(3.89±0.75)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(1.07±0.48)e−3	(1.07±0.48)e−3	(1.07±0.49)e−3	(1.07±0.48)e−3
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(3.67±1.02)e−4	(3.67±1.02)e−4	(3.67±1.02)e−4	(3.67±1.02)e−4
	Test risk (total)	(2.24±0.68)e−3	(2.24±0.68)e−3	(2.24±0.68)e−3	(2.24±0.68)e−3
	$\mathrm{KL}(\rho\|\pi)$	(3.85±2.73)e−5	(3.83±2.67)e−5	(3.09±0.34)e−6	(1.65±1.07)e−5
	Bound	(5.74±0.78)e−1	(5.76±0.78)e−1	(6.51±0.00)e1	(2.08±0.29)e0
	Bound w/o tail	(4.19±1.17)e−3	(4.32±1.23)e−3	-	-

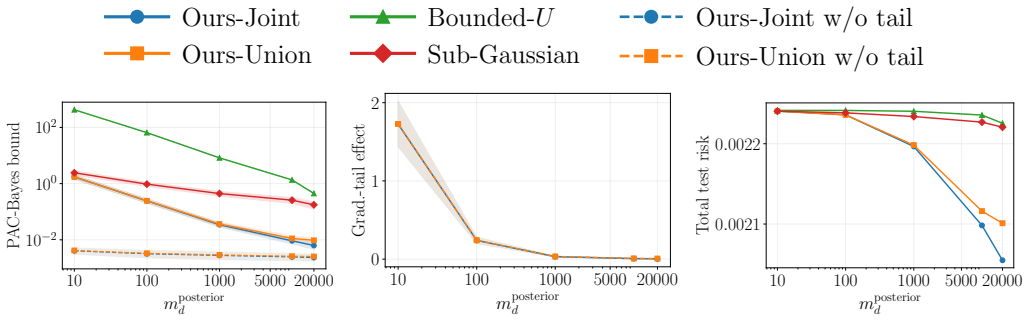
Table 7: PAC-Bayes benchmark results for Helmholtz in the balanced and data-scarce regimes.

Data regime	Quantity	Ours-joint	Ours-union	Bounded-U	Sub-Gaussian
Balanced	Train risk ($\widehat{\mathcal{R}}_d$)	(3.65±0.30)e−4	(3.65±0.30)e−4	(3.65±0.30)e−4	(3.67±0.31)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(4.66±0.48)e0	(4.66±0.48)e0	(4.66±0.48)e0	(4.77±0.48)e0
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(3.64±0.29)e−4	(3.64±0.29)e−4	(3.64±0.29)e−4	(3.65±0.30)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(4.29±0.38)e−4	(4.29±0.38)e−4	(4.29±0.37)e−4	(4.31±0.38)e−4
	Train risk (total)	(4.66±0.48)e0	(4.66±0.48)e0	(4.66±0.48)e0	(4.77±0.48)e0
	Test risk ($\widehat{\mathcal{R}}_d$)	(3.65±0.30)e−4	(3.65±0.30)e−4	(3.65±0.30)e−4	(3.67±0.30)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(4.62±0.48)e0	(4.62±0.48)e0	(4.62±0.48)e0	(4.72±0.48)e0
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(3.64±0.29)e−4	(3.64±0.29)e−4	(3.64±0.29)e−4	(3.65±0.30)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(4.28±0.37)e−4	(4.28±0.37)e−4	(4.28±0.37)e−4	(4.30±0.37)e−4
	Test risk (total)	(4.62±0.48)e0	(4.62±0.48)e0	(4.62±0.48)e0	(4.73±0.48)e0
	$\mathrm{KL}(\rho\|\pi)$	(1.77±0.48)e−1	(1.77±0.48)e−1	(1.87±0.42)e−1	(1.70±0.42)e−3
	Bound	(9.16±0.76)e0	(9.17±0.76)e0	(1.08±0.07)e1	(1.78±0.15)e2
	Bound w/o tail	(5.34±0.51)e0	(5.34±0.51)e0	-	-
Data-scarce	Train risk ($\widehat{\mathcal{R}}_d$)	(3.58±0.31)e−4	(3.58±0.31)e−4	(3.59±0.31)e−4	(3.59±0.31)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(4.66±0.48)e0	(4.66±0.48)e0	(4.78±0.48)e0	(4.77±0.48)e0
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(3.64±0.29)e−4	(3.64±0.29)e−4	(3.65±0.30)e−4	(3.65±0.30)e−4
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(4.29±0.37)e−4	(4.29±0.37)e−4	(4.31±0.38)e−4	(4.31±0.38)e−4
	Train risk (total)	(4.67±0.48)e0	(4.67±0.48)e0	(4.78±0.48)e0	(4.77±0.48)e0
	Test risk ($\widehat{\mathcal{R}}_d$)	(3.65±0.30)e−4	(3.65±0.30)e−4	(3.67±0.30)e−4	(3.67±0.30)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(4.62±0.48)e0	(4.62±0.48)e0	(4.73±0.48)e0	(4.72±0.48)e0
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(3.64±0.29)e−4	(3.64±0.29)e−4	(3.65±0.30)e−4	(3.65±0.30)e−4
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(4.28±0.37)e−4	(4.28±0.37)e−4	(4.31±0.37)e−4	(4.30±0.37)e−4
	Test risk (total)	(4.62±0.48)e0	(4.62±0.48)e0	(4.73±0.48)e0	(4.73±0.48)e0
	$\mathrm{KL}(\rho\|\pi)$	(1.71±0.43)e−1	(1.71±0.43)e−1	(1.76±0.26)e−4	(1.75±0.32)e−3
	Bound	(9.23±0.76)e0	(9.23±0.76)e0	(3.34±0.01)e2	(1.78±0.15)e2
	Bound w/o tail	(5.34±0.51)e0	(5.34±0.51)e0	-	-

a much stronger effect on the tightness of the certificate than on predictive accuracy: the physics-informed training can still yield an accurate solution, while the small observational sample leads to substantially larger finite-sample generalisation penalties.

Table 8: PAC-Bayes benchmark results for Flow Mixing-2D in the balanced and data-scarce regimes.

Data regime	Quantity	Ours-joint	Ours-union	Bounded-U	Sub-Gaussian
Balanced	Train risk ($\widehat{\mathcal{R}}_d$)	(1.65±0.12)e−3	(1.65±0.12)e−3	(1.65±0.12)e−3	(1.65±0.12)e−3
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(1.13±0.24)e−5	(1.13±0.24)e−5	(1.13±0.24)e−5	(1.13±0.24)e−5
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(1.34±0.10)e−3	(1.34±0.10)e−3	(1.34±0.10)e−3	(1.34±0.10)e−3
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(8.95±1.26)e−5	(9.01±1.30)e−5	(9.09±1.36)e−5	(9.12±1.39)e−5
	Train risk (total)	(3.09±0.22)e−3	(3.09±0.22)e−3	(3.10±0.22)e−3	(3.10±0.22)e−3
	Test risk ($\widehat{\mathcal{R}}_d$)	(1.74±0.12)e−3	(1.74±0.12)e−3	(1.74±0.12)e−3	(1.74±0.12)e−3
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(1.16±0.24)e−5	(1.16±0.24)e−5	(1.16±0.24)e−5	(1.16±0.24)e−5
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(1.38±0.11)e−3	(1.38±0.11)e−3	(1.38±0.11)e−3	(1.38±0.11)e−3
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(8.96±1.26)e−5	(9.01±1.30)e−5	(9.09±1.36)e−5	(9.13±1.39)e−5
	Test risk (total)	(3.22±0.23)e−3	(3.22±0.23)e−3	(3.22±0.23)e−3	(3.22±0.23)e−3
	KL($\rho\|\pi$)	(1.05±0.78)e−3	(5.77±4.13)e−4	(6.54±4.58)e−5	(5.74±4.42)e−6
	Bound	(3.98±0.22)e−2	(5.58±0.33)e−2	(4.51±0.01)e−1	(1.44±0.09)e0
	Bound w/o tail	(7.98±0.31)e−3	(8.89±0.38)e−3	-	-
Data-scarce	Train risk ($\widehat{\mathcal{R}}_d$)	(1.47±0.17)e−3	(1.47±0.17)e−3	(1.47±0.17)e−3	(1.47±0.17)e−3
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(1.13±0.24)e−5	(1.13±0.24)e−5	(1.13±0.24)e−5	(1.13±0.24)e−5
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(1.34±0.10)e−3	(1.34±0.10)e−3	(1.34±0.10)e−3	(1.34±0.10)e−3
	Train risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(9.13±1.39)e−5	(9.13±1.39)e−5	(9.14±1.39)e−5	(9.14±1.39)e−5
	Train risk (total)	(2.91±0.27)e−3	(2.91±0.27)e−3	(2.91±0.27)e−3	(2.91±0.27)e−3
	Test risk ($\widehat{\mathcal{R}}_d$)	(1.74±0.12)e−3	(1.74±0.12)e−3	(1.74±0.12)e−3	(1.74±0.12)e−3
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{PDE}}$)	(1.16±0.24)e−5	(1.16±0.24)e−5	(1.16±0.24)e−5	(1.16±0.24)e−5
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{IC}}$)	(1.38±0.11)e−3	(1.38±0.11)e−3	(1.38±0.11)e−3	(1.38±0.11)e−3
	Test risk ($\widehat{\mathcal{R}}_{\mathrm{BC}}$)	(9.14±1.40)e−5	(9.14±1.40)e−5	(9.14±1.40)e−5	(9.14±1.40)e−5
	Test risk (total)	(3.22±0.23)e−3	(3.22±0.23)e−3	(3.22±0.23)e−3	(3.22±0.23)e−3
	KL($\rho\|\pi$)	(3.13±1.06)e−6	(3.15±1.05)e−6	(3.37±1.69)e−6	(3.50±1.55)e−6
	Bound	(1.21±0.16)e1	(1.21±0.16)e1	(6.51±0.00)e1	(4.19±0.53)e1
	Bound w/o tail	(7.21±0.57)e−2	(7.22±0.57)e−2	-	-

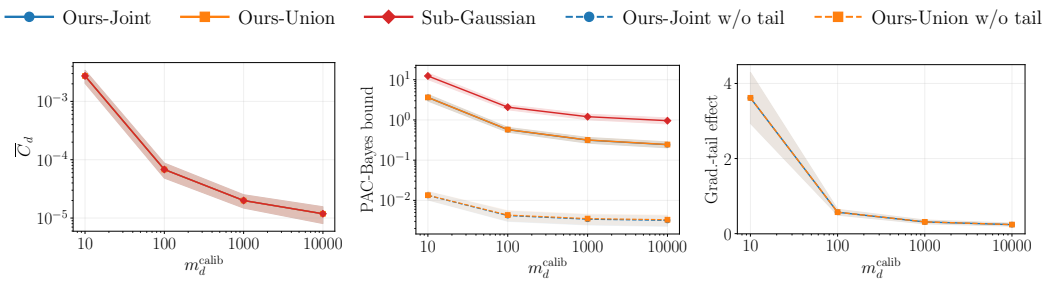


(a) Bound value (b) Gradient tail effect (c) Total test risk

Figure 13: Impact of the number of observational posterior training samples $m_d^{\mathrm{posterior}}$.

C.5.2 Impact of the calibration's sample size

We next study the effect of the observational calibration sample size m_d^{calib} on Heat-2D MS by varying it from 10 to the full 10k samples, while fixing $m_d^{\mathrm{posterior}} = 100$ and retaining all physical samples. For this ablation, the calibrated coefficient $\overline{C}_d$ is shared by all compared input-gradient bounds, allowing us to isolate the effect of calibration quality independently of posterior-training sample size. As shown in Fig. 14, increasing m_d^{calib} substantially decreases $\overline{C}_d$, particularly in the small-sample regime, and leads to a corresponding tightening of all certificates. The improvement is most pronounced between 10 and the first few hundred calibration samples and gradually saturates thereafter, consistently with the convergence of $\overline{C}_d$ towards a stable scale. The gradient tail's contribution also follows the same variation. The similar trend across the proposed and sub-Gaussian bounds further confirms that the improvement originates primarily from the increasingly informative (tighter) calibration coefficient rather than from differences in their posterior-training complexities.

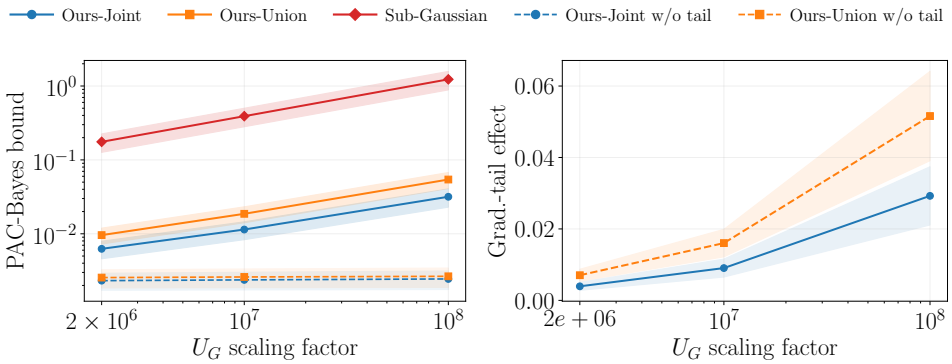


(a) Calibrated $\overline{C}_d$ value (b) Bound value (c) Gradient tail effect

Figure 14: Impact of the number of observational calibration sample size m_d^{calib}.

C.5.3 Impact of the gradient envelope

We further assess the sensitivity to the prescribed gradient envelope by varying $U_{G,\text{factor}} \in \{2 \times 10^6, 10^7, 10^8\}$ on Heat-2D MS, while retaining the full posterior-training and calibration datasets and fixing all remaining hyperparameters. As shown in Fig. 15, increasing $U_{G,\text{factor}}$ progressively loosens all certificates, with a particularly strong deterioration of the sub-Gaussian bound, which directly uses $U_{G,i}$ to control the complete input-gradient energy. For our bounds, this dependence arises primarily through the residual gradient-tail correction: the oracle certificates without this contribution remain nearly unchanged as $U_{G,\text{factor}}$ increases, whereas the tail effect grows notably with the prescribed envelope. This confirms the benefit of restricting the conservative gradient envelope to the residual-tail control rather than using it as the concentration scale of the entire input-gradient complexity, while also motivating the use of the smallest conservative envelope supported by the empirical diagnostics.

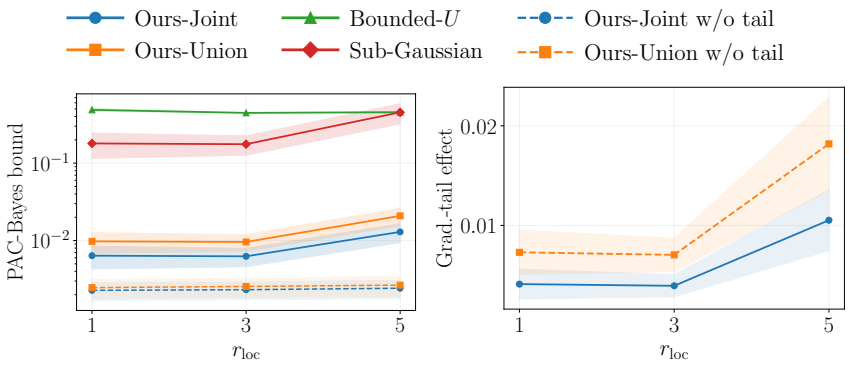


(a) Bound value (b) Gradient tail effect

Figure 15: Impact of the gradient enveloppe through U_G's scaling factor ($U_{G,\text{factor}}$).

C.5.4 Impact of the localisation radius

We finally investigate the effect of the localisation radius $R_{\text{loc}} = r_{\text{loc}}\sigma\sqrt{d_\theta}$ on Heat-2D MS by varying $r_{\text{loc}} \in \{1, 3, 5\}$ while retaining the full posterior-training and calibration datasets. As shown in Fig. 16, increasing the localisation radius substantially reduces $D_{\text{KL}}(\rho\|\pi)$, but simultaneously enlarges the proposed certificates through the input-gradient complexity, as models farther from the well-trained region generally exhibit larger input-gradient magnitudes. One can see that the gradient-tail contribution increasing markedly over the considered range, in particular from 3 to 5. Accordingly, the oracle bounds without the tail remain nearly unchanged, whereas the complete *ours-joint*, *ours-union* and sub-Gaussian certificates progressively increase with r_{loc}. In contrast, Bounded-U is comparatively insensitive to the localisation radius because it does not depend on the observed input-gradient magnitudes.



(a) Bound value (b) Gradient tail effect

Figure 16: Impact of the truncation radius through the radius factor r_{loc}.

D LIMITATIONS

This paper presents a PAC-Bayesian analysis for physics-informed machine learning, providing both generalisation guarantees and practical algorithms supported by extensive experiments on six benchmarks. Despite promising results, including substantially tighter certificates than the considered baselines and intuitive empirical observations, several limitations remain. Regarding the theoretical analysis, the use of a gradient envelope to control the residual gradient contribution can still lead to deteriorated bounds, particularly in data-scarce regimes and when the underlying solution is highly irregular. Therefore, deriving sharper finite-sample control of the residual gradient represents a promising direction for further tightening the resulting certificates. In addition, developing a calibration procedure for Theorem 4 that avoids the use of a union bound could be particularly advantageous in data-scarce scenario. Besides, it remains unclear how the incorporation of physical constraints affects the generalisation behavior of the data-fidelity loss, which constitutes an important direction for future theoretical investigation. From an experimental perspective, an important direction is to investigate how the proposed bounds scale to more complex multi-output problems (e.g., the Navier-Stokes equations), where several potentially high-dimensional and coupled PDE residual terms must be handled simultaneously. Another promising direction for mitigating the gradient-envelope issue is to enforce input-gradient regularity by construction, for instance through Lipschitz-constrained neural networks, and to investigate whether such architectures can yield tighter certificates in an end-to-end manner, from learning a physics-informed prior under these constraints to directly minimising the final PIML bound.

E RELATED WORKS

Generalisation of PIML. Despite tremendous empirical success, the generalisation of physics-informed machine learning (PIML) methods remains an open question that requires further study. A common line of work focuses on the convergence of physics informed models to the PDE solution (Shin et al., 2023; Doumèche et al., 2025b). In particular, (Doumèche et al., 2025b) addresses overfitting of PINNs by introducing a Sobolev regularisation term to enforce estimator regularity and show consistency and strong convergence properties. Other works derive approximation error bounds under stability assumptions on the underlying PDEs (Ryck & Mishra, 2022; Mishra & Molinaro, 2022). Several papers study generalisation via (local) Rademacher complexity (Jiao et al., 2022; Lu et al., 2022; Xu et al., 2025); notably, (Xu et al., 2025) obtains generalisation bounds for second order elliptic PDEs by adopting a multi-task perspective and assuming the underlying solutions lie in Barron or Sobolev spaces. (Doumèche et al., 2024; 2025a) treat the PIML problem as a kernel regression task and use Fourier methods to construct tractable estimators with quantified convergence rates. (Scampicchio et al., 2026) views physical loss terms as a form of regularisation in addition to data-fidelity loss and shows that, under a knowledge alignment assumption (i.e., the regulariser is approximately zero at the true solution), the estimator converges at a faster rate. Together, these works highlight the importance of encoding prior knowledge about the physical

system—via regularity, stability, or data-physics alignment—to establish generalisation guarantees for approximation error or the total PIML loss. In the same vein, we connect generalisation, data, and physical regularity through a PAC-Bayesian framework that yields computable input-gradient complexity terms, and empirically show that the resulting certificates can be meaningfully evaluated across PDE benchmarks and remain substantially tighter than representative alternatives under observational data scarcity.

PAC-Bayesian analysis for PIML. PAC-Bayes theory has emerged as a powerful framework for analysing modern machine learning models through data-dependent probabilistic guarantees that balance empirical risk and model complexity via information-theoretic divergences (Alquier, 2024; Guedj, 2019; Hellström et al., 2025). While early PAC-Bayesian analyses primarily focused on bounded-loss classification settings, more recent works have extended the framework to regression and heavy-tailed losses through higher-order moment control (Haddouche & Guedj, 2023; Holland, 2019), assumptions on the CGF (Casado et al., 2024; Rodríguez-Gálvez et al., 2024), and structural assumptions on the data distribution (Shalaeva et al., 2019; Guo et al., 2025). These advances are particularly relevant for PIML, where losses induced by PDE residuals or physical constraints are typically unbounded and may exhibit complex tail behaviour. Moreover, PAC-Bayes naturally accommodates prior physical knowledge through structured priors and constrained hypothesis spaces. More recently, model-dependent analyses have connected generalisation to regularity measures based on parameter norms (Germain et al., 2009), parameter gradients (Haddouche et al., 2025), and input gradients (Gat et al., 2022; Casado et al., 2024). In particular, input-gradient information provides a natural connection between statistical generalisation and the regularity of learned functions, and has been incorporated into PAC-Bayesian bounds under different concentration assumptions (Gat et al., 2022; Casado et al., 2024). Building on this broader perspective, we develop a PIML-specific framework based on a population CGF-to-gradient control, together with a PAC-Bayes calibration procedure and a tail-aware decomposition that yields computable input-gradient complexities without bounding the original losses. We further exploit the multi-loss structure of PIML to derive a refined joint certificate with a single KL complexity, and introduce a bound-aware learning procedure for optimising the resulting guarantees. To the best of our knowledge, our work is among the first to provide an end-to-end PAC-Bayesian framework for both learning and certifying PIML models in the presence of potentially unbounded losses.