STUDENT-T PROCESSES AS INFINITE-WIDTH LIMITS OF POSTERIOR BAYESIAN NEURAL NETWORKS

FRANCESCO CAPORALI*, STEFANO FAVARO†, AND DARIO TREVISAN‡

Abstract. The asymptotic properties of Bayesian Neural Networks (BNNs) have been extensively studied, particularly regarding their approximations by Gaussian processes in the infinite-width limit. We extend these results by showing that posterior BNNs can be approximated by Student-t processes, which offer greater flexibility in modeling uncertainty. Specifically, we show that, if the parameters of a BNN follow a Gaussian prior distribution, and the variance of both the last hidden layer and the Gaussian likelihood function follows an Inverse-Gamma prior distribution, then the resulting posterior BNN converges to a Student-t process in the infinite-width limit. Our proof leverages the Wasserstein metric to establish control over the convergence rate of the Student-t process approximation.

1. Introduction

Bayesian neural networks (BNNs), composed of multiple layers of interconnected neurons, have become a powerful tool in modern machine learning, enabling the modeling of complex data structures while quantifying predictive uncertainty [Nea96]. Unlike neural networks (NNs), BNNs offer a solid probabilistic framework where model parameters are treated as random variables with associated probability distributions. In particular, such a framework allows for the incorporation of both prior knowledge and observed data through the prior distribution and likelihood function, respectively.

1.1. **Background and motivation.** The theoretical study of BNNs dates back to the foundational work of Neal [Nea96], which, inspired by Bayesian principles, showed that wide shallow BNNs converge to Gaussian processes if initialized with independent Gaussian parameters. This result was later extended to deep BNNs [Mat+18; Lee+18; BT24; Fav+24] as well as to alternative architectures [Nov+20; Yan21], strengthening the connection between deep learning and Gaussian processes in machine learning [RW06].
Building on this foundation, significant effort has been devoted to analyzing the posterior behavior of BNNs. Notably, several studies have examined their exact infinite-width limiting posterior distribution, establishing its asymptotic convergence to a Gaussian process [Hro+20; Tre23]. Parallel research has explored approximate posterior inference methods, including Variational Inference (VI) [Blu+15] and Monte Carlo Markov Chain (MCMC) sampling [Izm+21; PFP24], providing an empirical validation of these theoretical results.
Despite the significant progresses in the study of posterior BNNs, existing work typically assumes a fixed variance for the Gaussian prior on the network parameters, a simplification that limits substantially the diversity of posterior behaviors that BNNs can capture. In this paper, we address this critical limitation by introducing a more flexible model in which the variance itself follows a prior distribution.

Key words and phrases. Bayesian neural network, infinite-width limit, posterior distribution, Student-t processes, Wasserstein distance.
Workshop on the Theory of AI for Scientific Computing (TASC 2025), 38th Annual Conference on Learning Theory (COLT 2025).
* Dept. of Operations Research and Financial Engineering, Princeton University `fc4978@princeton.edu`
† Dept. of Economics and Statistics, University of Torino and Collegio Carlo Alberto, `stefano.favaro@unito.it`
‡ Dept. of Mathematics, University of Pisa, `dario.trevisan@unipi.it`

1.2. **Our contribution.** Our main contribution is to show that relaxing the fixed-variance assumption in BNNs by using an Inverse-Gamma prior leads to a novel limiting behavior, while preserving the classic prior introduced by Neal [Nea96]. Specifically, we prove that while the prior distribution of a BNN converges to a Gaussian process in the infinite-width limit, the marginal posterior distribution converges to a Student-t process. The relevance of this result is twofold. First, it provides a new representation of Student-t processes, which have been widely studied and applied in machine learning and statistics [SWG14; TW18; SD23]. Second, it suggests that modifying the prior distributions can yield a broader class of limiting elliptical processes [FKN90; Ban+20], opening new research directions on the asymptotic behavior of posterior BNNs. This insight shows that a careful selection of prior distributions can enhance model flexibility and uncertainty quantification, offering practical benefits in Bayesian deep learning.
Our approach relies on optimal transport tools, specifically the Wasserstein metric, in order to establish convergence rates and gain control over the distances of the distributions under analysis. Building on prior works [BT24; Tre23], we extend the framework to a hierarchical Gaussian-Inverse-Gamma model. In this model, while the prior and likelihood are still assumed to follow multivariate Gaussian distributions with diagonal covariance, the variance of both the last hidden layer and the likelihood function is modeled using an Inverse-Gamma distribution.

1.3. **Outline.** In Section 2, we introduce the notation and key tools used in the proof of our main result. Section 3 presents the main result of the paper, while Section 4 reports experimental results that serve as a sanity check for the developed theory.

2. Preliminaries

To clarify the discussion, we refer the reader to Appendices A.1 to A.3 for a review of tensors, random variables, and optimal transport tools.

2.1. **Wasserstein distance.** For a finite set S, denote by μ and ν two probability measures defined on $(\mathbb{R}^S, \|\cdot\|)$ with finite moment of order p, for some $p \geq 1$. The p-Wasserstein distance between μ and ν is defined as

$$\mathcal{W}_p(\mu, \nu) := \inf \left\{ \mathbb{E}\left[\|\boldsymbol{x} - \boldsymbol{y}\|^p\right]^{1/p} \mid \boldsymbol{x}, \boldsymbol{y} \text{ r.v.s with } \mathbb{P}_{\boldsymbol{x}} = \mu, \mathbb{P}_{\boldsymbol{y}} = \nu \right\},$$

where the infimum is taken over all the random variables $(\boldsymbol{x}, \boldsymbol{y})$, jointly defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with marginal laws μ and ν. The random variable $(\boldsymbol{x}, \boldsymbol{y})$ is referred to as a coupling of (μ, ν) and its law, γ, is called transport plan. We introduce the following abuse of notation: if $\boldsymbol{x} \sim \mu$ and $\boldsymbol{y} \sim \nu$, $\mathcal{W}_p(\boldsymbol{x}, \boldsymbol{y}) := \mathcal{W}_p(\mu, \nu)$.

Theorem 2.1 (Theorem 6.9 of Villani [Vil08])**.** *Given* $(\boldsymbol{x}_n)_{n=1}^{\infty}$*,* $\boldsymbol{x}$ *random variables, then* $\lim_{n\to\infty} \mathcal{W}_p(\boldsymbol{x}_n, \boldsymbol{x}) = 0$ *if and only if* $\boldsymbol{x}_n \xrightarrow{law} \boldsymbol{x}$*, and* $\lim_{n\to\infty} \mathbb{E}[\|\boldsymbol{x}_n\|^p] = \mathbb{E}[\|\boldsymbol{x}\|^p]$.

2.2. **BNNs.** Consider a supervised learning framework in a regression setting, with a given training dataset $\mathcal{D} := \{(\boldsymbol{x}_{\mathcal{D},i}, \boldsymbol{y}_{\mathcal{D},i})\}_{i=1}^k$, i.e.,

$$\boldsymbol{x}_{\mathcal{D}} := \sum_{i=1}^k \boldsymbol{x}_{\mathcal{D},i} \otimes \boldsymbol{e}_i \in \mathbb{R}^{d_{\text{in}} \times k} \quad \text{and} \quad \boldsymbol{y}_{\mathcal{D}} := \sum_{i=1}^k \boldsymbol{y}_{\mathcal{D},i} \otimes \boldsymbol{e}_i \in \mathbb{R}^{d_{\text{out}} \times k}.$$

Definition 2.2 (Fully connected feed-forward NN)**.** A fully connected feed-forward NN is defined through an architecture $\boldsymbol{\alpha} := (\boldsymbol{n}, \boldsymbol{\varphi})$ with:

(1) $\boldsymbol{n}$ denoting the sizes of the $L+1$ layers[1] (with $L \geq 2$)

$$\boldsymbol{n} := \big(n_0(=: d_{\text{in}}), n_1, \ldots, n_{L-1}, n_L(=: d_{\text{out}})\big),\ n_l \in \mathbb{N}_{>0},\ \forall\, l = 0, \ldots, L;$$

(2) $\boldsymbol{\varphi}$ denoting the L activation functions (applied component-wise)

$$\boldsymbol{\varphi} := (\varphi_1, \ldots, \varphi_L),\ \varphi_l : \mathbb{R} \to \mathbb{R},\ \forall\, l \in [L],\ \text{with } \varphi_1(x) = x, \forall\, x \in \mathbb{R}.$$

[1] An input layer, $L-1$ hidden layers and an output layer.

In particular, $\forall\, \boldsymbol{x}_0 \in \mathbb{R}^{d_{\text{in}}}$, the NN is defined as $f(\boldsymbol{x}_0) \coloneqq f^{(L)}(\boldsymbol{x}_0)$ with

$$\begin{aligned} & f^{(l)} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{n_l},\ \forall\, l = 0, \dots, L, \\ & f^{(0)}(\boldsymbol{x}_0) = \boldsymbol{x}_0, \quad f^{(l)}(\boldsymbol{x}_0) = \boldsymbol{W}^{(l)} \varphi_l \left(f^{(l-1)}(\boldsymbol{x}_0) \right) + \boldsymbol{b}^{(l)} \text{ for } l \in [L], \end{aligned} \tag{1}$$

where, for any $l \in [L]$, $\boldsymbol{W}^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ and $\boldsymbol{b}^{(l)} \in \mathbb{R}^{n_l}$ denote weight matrices and bias vectors, respectively.

BNNs exploit the power of the Bayes' rule within this supervised learning framework. By defining $\boldsymbol{\theta} \in \mathbb{R}^t$ (with $t = \sum_{l=1}^{L} n_l(n_{l-1}+1)$) the flattened concatenation of all parameters of the NN (both weights and biases), it is possible to apply Bayes' theorem to describe the posterior distribution of a BNN:

$$p_{\boldsymbol{\theta}|\mathcal{D}}(\boldsymbol{\theta}) = \frac{p_{\mathcal{D}|\boldsymbol{\theta}}(\mathcal{D})\, p_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p_{\mathcal{D}}(\mathcal{D})} \propto p_{\boldsymbol{x}_\mathcal{D}, \boldsymbol{y}_\mathcal{D}|\boldsymbol{\theta}}(\boldsymbol{x}_\mathcal{D}, \boldsymbol{y}_\mathcal{D})\, p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto p_{\boldsymbol{y}_\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{x}_\mathcal{D}}(\boldsymbol{y}_\mathcal{D})\, p_{\boldsymbol{\theta}}(\boldsymbol{\theta}),$$

where we assumed that $\boldsymbol{x}_\mathcal{D}$ is independent of $\boldsymbol{\theta}$ and all the random variables admit a density with respect to the Lebesgue measure. In particular, if $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_\mathcal{D}) \coloneqq p_{\boldsymbol{y}_\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{x}_\mathcal{D}}(\boldsymbol{y}_\mathcal{D})$ is the likelihood function then

$$p_{\boldsymbol{\theta}|\mathcal{D}}(\boldsymbol{\theta}) \propto \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_\mathcal{D})\, p_{\boldsymbol{\theta}}(\boldsymbol{\theta}). \tag{2}$$

Definition 2.3 (BNN)**.** Let f be a NN (see eq. (1)) with architecture $\boldsymbol{\alpha}$. In order to define a BNN we have to put a prior distribution over $\boldsymbol{\theta}$ and a likelihood function $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_\mathcal{D})$ for $\boldsymbol{\theta}$ associated to the training set $\mathcal{D}$.

Remark 2.4. Bayes' theorem, and a related notion of posterior measure, can be naturally built without the necessity of density functions. Let $\boldsymbol{\theta} : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^S$ random variable, S finite set, and an evidence $\mathcal{D}$ with density $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_\mathcal{D})$, we define the posterior measure of $\boldsymbol{\theta}$ as

$$\mathbb{P}_{\boldsymbol{\theta}|\mathcal{D}} \coloneqq \frac{\mathcal{L}(\cdot; \boldsymbol{y}_\mathcal{D})}{\int_{\mathbb{R}^S} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_\mathcal{D}) d\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \mathbb{P}_{\boldsymbol{\theta}}, \tag{3}$$

where, given ν measure on $(\mathbb{R}^S, \mathcal{B}(\mathbb{R}^S))$ (S finite set), and $f : \mathbb{R}^S \to \mathbb{R}$, we use the notation $f\nu$ to denote the measure on $(\mathbb{R}^S, \mathcal{B}(\mathbb{R}^S))$ absolutely continuous with respect to ν ($f\nu \ll \nu$), with density f, i.e., $\forall\, A \in \mathcal{B}(\mathbb{R}^S)$, $f\nu(A) = \int_A f(\boldsymbol{u}) d\nu(\boldsymbol{u})$. For the sake of simplicity we always work with densities (as in eq. (2)) if they are available, and we swap to the measure theoretic definition in eq. (3) otherwise.

The prior distribution considered on the parameters is the Gaussian independent prior [Nea96]. Specifically, given the vector of variances $\boldsymbol{\sigma} \coloneqq \left((\sigma^2_{\boldsymbol{W}^{(l)}}, \sigma^2_{\boldsymbol{b}^{(l)}}) \right)_{l=1}^{L} \in (\mathbb{R}^+ \times \mathbb{R}^+)^L$, we assume that

$$\boldsymbol{W}^{(l)} \sim \mathcal{N}\left(\boldsymbol{0}_{n_l \times n_{l-1}}, \sigma^2_{\boldsymbol{W}^{(l)}}/n_{l-1}\, \boldsymbol{I}_{n_l \times n_{l-1}} \right),\ \boldsymbol{b}^{(l)} \sim \mathcal{N}\left(\boldsymbol{0}_{n_l}, \sigma^2_{\boldsymbol{b}^{(l)}} \boldsymbol{I}_{n_l} \right),\ \text{for } l \in [L]. \tag{4}$$

The focus of this paper is to study the distribution that the posterior measure of $\boldsymbol{\theta}$ induces on f, which requires to investigate the behavior of the induced prior. In particular, we need to retrace the well-known results which state that the asymptotic distribution of $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \coloneqq \sum_{i=1}^{m} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \otimes \boldsymbol{e}_i$ converges to the neural network Gaussian process (NNGP), where $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^{m}$, $m \in \mathbb{N}_{>0}$, is a generic input set.

Remark 2.5. In order to have a simpler description of the subsequent theory it is convenient to write the layers of the BNNs in a compact form: $f^{(0)}_{\boldsymbol{\theta}} : \mathbb{R}^{d_{\text{in}} \times m} \to \mathbb{R}^{d_{\text{in}} \times m}$, $f^{(0)}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{x}$ and for every $l \in [L]$, $f^{(l)}_{\boldsymbol{\theta}} : \mathbb{R}^{n_{l-1} \times m} \to \mathbb{R}^{n_l \times m}$,

$$\begin{aligned} f^{(l)}_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \sum_{i=1}^{m} f^{(l)}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \otimes \boldsymbol{e}_i = \sum_{i=1}^{m} \boldsymbol{W}^{(l)} \varphi_l \left(f^{(l-1)}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \right) \otimes \boldsymbol{e}_i + \boldsymbol{b}^{(l)} \otimes \boldsymbol{1}_m = \\ &= \left(\boldsymbol{W}^{(l)} \otimes \boldsymbol{I}_m \right) \varphi_l \left(f^{(l-1)}_{\boldsymbol{\theta}}(\boldsymbol{x}) \right) + \boldsymbol{b}^{(l)} \otimes \boldsymbol{1}_m, \end{aligned}$$

where $\boldsymbol{W}^{(l)} \otimes \boldsymbol{I}_m \in \mathbb{R}^{(n_l \times n_{l-1}) \times (m \times m)}$ should be thought as an element of $\mathbb{R}^{(n_l \times m) \times (n_{l-1} \times m)}$. We define $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \coloneqq f^{(L)}_{\boldsymbol{\theta}}(\boldsymbol{x})$.

2.3. NNGP.

Definition 2.6. Let $H = (H(\boldsymbol{x}_0))_{\boldsymbol{x}_0 \in \mathbb{R}^{d_{\text{in}}}}$ be a stochastic process such that for any $\boldsymbol{x}_0 \in \mathbb{R}^{d_{\text{in}}}$, $H(\boldsymbol{x}_0)$ is a random vector with values in $\left(\mathbb{R}^{d_{\text{out}}}, \mathcal{B}\left(\mathbb{R}^{d_{\text{out}}}\right)\right)$. Then H is said to be a Gaussian process with mean function $\boldsymbol{M} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ and covariance kernel $\boldsymbol{H} : \mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$, if for any $m > 0$, given $\boldsymbol{x} = (\boldsymbol{x}_i)_{i=1}^m \in \mathbb{R}^{d_{\text{in}} \times m}$,

$$H(\boldsymbol{x}) \coloneqq (H(\boldsymbol{x}_1), \dots, H(\boldsymbol{x}_m)) \sim \mathcal{N}\left(\boldsymbol{M}(\boldsymbol{x}), \boldsymbol{H}(\boldsymbol{x})\right),$$

where

$$\boldsymbol{M}(\boldsymbol{x}) = (\boldsymbol{M}(\boldsymbol{x}_1), \dots, \boldsymbol{M}(\boldsymbol{x}_m)) \in \mathbb{R}^{d_{\text{out}} \times m},$$

and $\boldsymbol{H}(\boldsymbol{x}) \in \mathbb{R}^{(d_{\text{out}} \times m) \times (d_{\text{out}} \times m)}$ can be viewed as a block matrix with $m \times m$ blocks such that, for any $i, j \in [m]^2$, the block (i, j) of $\boldsymbol{H}(\boldsymbol{x})$ is $(\boldsymbol{H}(\boldsymbol{x}))_{i,j} = \boldsymbol{H}(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$. For such H we set $H \sim \mathcal{GP}\left(\boldsymbol{M}, \boldsymbol{H}\right)$.

Remark 2.7. In analogy with Definition 2.6 it is possible to define Student-t processes replacing the condition on the distribution of $H(\boldsymbol{x})$ with a multivariate Student-t with a degrees of freedom: $H(\boldsymbol{x}) \sim t_a\left(\boldsymbol{M}(\boldsymbol{x}), \boldsymbol{H}(\boldsymbol{x})\right)$, $a > 0$.

Following Matthews et al. [Mat+18] and Lee et al. [Lee+18], we report below the laws of the random matrices $(G^{(l)}(\boldsymbol{x}))_{l=1}^L$ associated with the infinite-width limits of the L hidden layers of a BNN, evaluated on the input set $\boldsymbol{x}$, deriving general expressions for their corresponding Gaussian processes, $(G^{(l)})_{l=1}^L$:

$$\begin{aligned} &G^{(0)}(\boldsymbol{x}) \coloneqq \boldsymbol{x} \in \mathbb{R}^{d_{\text{in}} \times m} \text{ constant r.v.}, \\ &G^{(l)}(\boldsymbol{x}) \sim \mathcal{N}\left(\boldsymbol{0}_{n_l \times m}, \boldsymbol{I}_{n_l} \otimes \boldsymbol{K}^{(l)}(\boldsymbol{x})\right), \text{ with } \boldsymbol{K}^{(l)}(\boldsymbol{x}) \coloneqq \left(\boldsymbol{K}^{(l)}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)_{i,j \in [m] \times [m]}, \end{aligned} \tag{5}$$

and, $\forall\, \boldsymbol{x}_0, \boldsymbol{x}_0' \in \mathbb{R}^{d_{\text{in}}}$, $\forall\, l = 2, \dots, L$,

$$\begin{aligned} \boldsymbol{K}^{(1)}(\boldsymbol{x}_0, \boldsymbol{x}_0') &\coloneqq \sigma^2_{\boldsymbol{W}^{(1)}}(\boldsymbol{x}_0^T \boldsymbol{x}_0')/d_{\text{in}} + \sigma^2_{\boldsymbol{b}^{(1)}}, \\ \boldsymbol{K}^{(l)}(\boldsymbol{x}_0, \boldsymbol{x}_0') &\coloneqq \sigma^2_{\boldsymbol{W}^{(l)}} \mathbb{E}\left[\varphi_l\left(G^{(l-1)}(\boldsymbol{x}_0)_1\right) \varphi_l\left(G^{(l-1)}(\boldsymbol{x}_0')_1\right)\right] + \sigma^2_{\boldsymbol{b}^{(l)}}. \end{aligned}$$

Note that, $\forall\, l \in [L]$, $\boldsymbol{I}_{n_l} \otimes \boldsymbol{K}^{(l)}(\boldsymbol{x}) \in \mathbb{R}^{(n_l \times n_l) \times (m \times m)}$ should be thought reshaped, as elements of $\mathbb{R}^{(n_l \times m) \times (n_l \times m)}$. From eq. (5) it is possible to define the asymptotic Gaussian processes of each hidden layer, $l \in [L]$, as

$$G^{(l)} = (G^{(l)}(\boldsymbol{x}_0))_{\boldsymbol{x}_0 \in \mathbb{R}^{d_{\text{in}}}} \quad \text{and} \quad G^{(l)} \sim \mathcal{GP}\left(\boldsymbol{0}, \boldsymbol{I}_{n_l} \otimes \boldsymbol{K}^{(l)}\right). \tag{6}$$

In analogy with the notation introduced for $f_{\boldsymbol{\theta}}$ we define $G(\boldsymbol{x}) \coloneqq G^{(L)}(\boldsymbol{x})$ and $\boldsymbol{K}(\boldsymbol{x}) \coloneqq \boldsymbol{K}^{(L)}(\boldsymbol{x})$. We refer to $G = (G(\boldsymbol{x}_0))_{\boldsymbol{x}_0 \in \mathbb{R}^{d_{\text{in}}}}$ as the NNGP associated to a BNN with architecture $\boldsymbol{\alpha}$ and vector of variances $\boldsymbol{\sigma}$.

2.4. Quantitative CLT for prior BNNs.

Theorem 2.8 (Basteri and Trevisan; Trevisan [BT24; Tre23])**.** *Let $f_{\boldsymbol{\theta}}$ BNN, with architecture $\boldsymbol{\alpha} = (\boldsymbol{n}, \boldsymbol{\varphi})$, $\boldsymbol{\varphi}$ collection of Lipschitz activation functions, a prior on $\boldsymbol{\theta}$ as in eq. (4), $\left(G^{(l)}\right)_{l=1}^L$ Gaussian processes as in eq. (6) and an input set $\boldsymbol{x} \in \mathbb{R}^{d_{\text{in}} \times m}$. Then, $\forall\, l \in [L]$ exists a constant $c > 0$ independent of $(n_j)_{j=1}^l$, such that,*

$$\mathcal{W}_p\left(f_{\boldsymbol{\theta}}^{(l)}(\boldsymbol{x}), G^{(l)}(\boldsymbol{x})\right) \le c\sqrt{n_l} \sum_{j=1}^{l-1} \frac{1}{\sqrt{n_j}}. \tag{7}$$

The constant c in eq. (7) in general depends on the input set $\boldsymbol{x}$, that must be finite. We remark that quantitative functional bounds, i.e., for infinitely many inputs, have been also established by Favaro et al. [Fav+24].

Remark 2.9. An immediate consequence of Theorem 2.8 (achievable applying Theorem 2.1) is that, by letting n_l grow to ∞, the process $f_{\boldsymbol{\theta}}^{(l)}$ associated with the l-th hidden layer of a BNN evaluated on $\boldsymbol{x}$ converges in distribution to the NNGP's component $G^{(l)}(\boldsymbol{x})$: given $n_{min} := \min_{j=1,\dots,l-1} n_j$,

$$f_{\boldsymbol{\theta}}^{(l)}(\boldsymbol{x}) \xrightarrow[n_{min}\to\infty]{law} \mathcal{N}\left(\mathbf{0}_{n_l\times m}, \boldsymbol{I}_{n_l} \otimes \boldsymbol{K}^{(l)}(\boldsymbol{x})\right).$$

Therefore, Theorem 2.8 yields a quantitative version of what has already been proved by Matthews et al.; Lee et al. [Mat+18; Lee+18].

3. Student-t approximation of posterior BNNs

Our goal is to extend the closeness result between the *induced* prior distribution on the BNN, $f_{\boldsymbol{\theta}}$, and the corresponding NNGP, G, established in Theorem 2.8, to their respective *induced* posterior distributions. In particular, the main result of this section, Theorem 3.4, provides a posterior counterpart of Theorem 2.8.

3.1. **Law of posterior NNGP.** It is useful to start by introducing the hierarchical model applied to the NNGP. In particular, by assuming

$$\begin{gathered} \sigma^2_{\boldsymbol{W}^{(l)}}, \sigma^2_{\boldsymbol{b}^{(l)}} \text{ constants}, \forall\, l = 1,\dots,L-1, \\ \sigma^2 := \sigma^2_{\boldsymbol{W}^{(L)}} = \sigma^2_{\boldsymbol{b}^{(L)}} = \sigma^2_{\boldsymbol{y}_{\mathcal{D}}} \quad \text{and} \quad \sigma^2 \sim \mathcal{IG}(a,b), \text{ with } a,b>0, \end{gathered} \tag{8}$$

we have

$$\begin{aligned} G(\boldsymbol{x}_{\mathcal{D}}) \,|\, \sigma^2 &\sim \mathcal{N}\left(\mathbf{0}_{n_L\times k}, \sigma^2 \boldsymbol{I}_{n_L} \otimes \boldsymbol{K}'(\boldsymbol{x}_{\mathcal{D}})\right), \\ \sigma^2 &\sim \mathcal{IG}(a,b), \text{ Inverse-Gamma with } a,b>0, \\ \boldsymbol{y}_{\mathcal{D}} \,|\, G(\boldsymbol{x}_{\mathcal{D}}), \sigma^2 &\sim \mathcal{N}\left(G(\boldsymbol{x}_{\mathcal{D}}), \sigma^2 \boldsymbol{I}_{n_L\times k}\right), \end{aligned} \tag{9}$$

with

$$\boldsymbol{K}'(\boldsymbol{x}_{\mathcal{D}}) = \mathbb{E}\left[\varphi_L\left(G^{(L-1)}(\boldsymbol{x}_{\mathcal{D}})\right)^T \varphi_L\left(G^{(L-1)}(\boldsymbol{x}_{\mathcal{D}})\right)\right]/n_L + \mathbf{1}_{k\times k}, \tag{10}$$

rescaled NNGP kernel independent of σ^2. Therefore, observing that

$$p_{G(\boldsymbol{x}_{\mathcal{D}}),\sigma^2|\mathcal{D}}(\boldsymbol{z},\sigma^2) \propto p_{\boldsymbol{y}_{\mathcal{D}}|G(\boldsymbol{x}_{\mathcal{D}}),\sigma^2}(\boldsymbol{y}_{\mathcal{D}})\, p_{G(\boldsymbol{x}_{\mathcal{D}})|\sigma^2}(\boldsymbol{z})\, p_{\sigma^2}(\sigma^2),$$

assuming $\boldsymbol{K}'(\boldsymbol{x}_{\mathcal{D}})$ to be invertible, $n_L = 1$, and applying standard tricks (see Appendix B), we obtain that

$$\begin{aligned} G(\boldsymbol{x}_{\mathcal{D}}) \,|\, \sigma^2, \mathcal{D} &\sim \mathcal{N}\left(\boldsymbol{y}_{\mathcal{D}}\boldsymbol{M}^{-1}, \sigma^2 \boldsymbol{M}^{-1}\right), \\ \sigma^2 \,|\, \mathcal{D} &\sim \mathcal{IG}\left(a + \frac{k}{2}, b + \frac{1}{2}\left(\boldsymbol{y}_{\mathcal{D}}\left(\boldsymbol{I}_k - \boldsymbol{M}^{-1}\right)(\boldsymbol{y}_{\mathcal{D}})^T\right)\right). \end{aligned}$$

Hence,

$$G(\boldsymbol{x}_{\mathcal{D}}) \,|\, \mathcal{D} \sim t_{2a+k}\left(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}}\right),$$

with

$$\begin{aligned} \boldsymbol{M} &:= \boldsymbol{I}_k + \boldsymbol{K}'(\boldsymbol{x}_{\mathcal{D}})^{-1}, \\ \boldsymbol{\mu}_{\text{post}} &:= \boldsymbol{y}_{\mathcal{D}}\boldsymbol{M}^{-1}, \\ \boldsymbol{\Sigma}_{\text{post}} &:= \left(b + \frac{1}{2}\left(\boldsymbol{y}_{\mathcal{D}}\left(\boldsymbol{I}_k - \boldsymbol{M}^{-1}\right)(\boldsymbol{y}_{\mathcal{D}})^T\right)\right)\frac{2}{2a+k}\boldsymbol{M}^{-1}. \end{aligned} \tag{11}$$

Remark 3.1. In a completely analogous way, it is possible to show that, given an input test set $\boldsymbol{x}_{\mathcal{T}} \in \mathbb{R}^{n_0\times k'}$,

$$\begin{aligned} G(\boldsymbol{x}_{\mathcal{T}}) \,|\, \mathcal{D} &\sim t_{2a+k}\left(\boldsymbol{\mu}'_{\text{post}}, \left(b + \frac{1}{2}\left(\boldsymbol{y}_{\mathcal{D}}\left(\boldsymbol{I}_k - \boldsymbol{M}^{-1}\right)(\boldsymbol{y}_{\mathcal{D}})^T\right)\right)\frac{2}{2a+k}\boldsymbol{\Sigma}'_{\text{post}}\right), \text{ with} \\ \boldsymbol{\mu}'_{\text{post}} &:= \boldsymbol{K}'(\boldsymbol{x}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{D}})\left(\boldsymbol{K}'(\boldsymbol{x}_{\mathcal{D}}) + \sigma^2 I_k\right)^{-1}\boldsymbol{y}_{\mathcal{D}}, \\ \boldsymbol{\Sigma}'_{\text{post}} &:= \boldsymbol{K}'(\boldsymbol{x}_{\mathcal{T}}) - \boldsymbol{K}'(\boldsymbol{x}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{D}})\left(\boldsymbol{K}'(\boldsymbol{x}_{\mathcal{D}}) + \sigma^2 I_k\right)^{-1}\boldsymbol{K}'(\boldsymbol{x}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{D}}). \end{aligned} \tag{12}$$

Indeed, following the strategy adopted by Rasmussen and Williams [RW06, eqs. (2.22) to (2.24)], we can observe that

$$G(\boldsymbol{x}_{\mathcal{T}}) \,|\, \sigma^2, \mathcal{D} \sim \mathcal{N}\left(\boldsymbol{\mu}'_{\text{post}}, \boldsymbol{\Sigma}'_{\text{post}}\right),$$
$$\sigma^2 \,|\, \mathcal{D} \sim \mathcal{IG}\left(a + \frac{k}{2}, b + \frac{1}{2}\left(\boldsymbol{y}_{\mathcal{D}}\left(\boldsymbol{I}_k - \boldsymbol{M}^{-1}\right)(\boldsymbol{y}_{\mathcal{D}})^T\right)\right),$$

which in turn implies eq. (12) by means of Lemma B.1.

3.2. **Posterior BNNs.** We define $\widetilde{\mu} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x})$, $\mu \sim G(\boldsymbol{x})$, with $\boldsymbol{x} = (\boldsymbol{x}_{\mathcal{D}}, \boldsymbol{x}_{\mathcal{T}}) \in \mathbb{R}^{n_0 \times (k+k')}$, fixed input set which extends the input training set with a possible input test set, and omit the dependence on $\boldsymbol{y}_{\mathcal{D}}$ in the Gaussian likelihood $\mathcal{L}$[2], where

$$\mathcal{L}: \mathbb{R}^{n_L \times k} \times \mathbb{R}^+ \to \mathbb{R},\ \mathcal{L}(\boldsymbol{z}, s) = \frac{1}{(2\pi s)^{n_L k/2}} \exp\left(-\frac{1}{2s}\left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F^2\right). \tag{13}$$

The objective is to bound the 1-Wasserstein distance between the marginal posterior of the BNN, $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, and the marginal posterior of the NNGP evaluated on the input set, $G(\boldsymbol{x})$. The latter can be found integrating with respect to s the prior measures μ and $\widetilde{\mu}$, both multiplied by the prior density of the variance $p_{\sigma^2}(s)$ and the likelihood $\mathcal{L}(\cdot, s)$. In formulas, we aim to find an upper bound for $\mathcal{W}_1\left(\widetilde{\mu}_{\text{post}}, \mu_{\text{post}}\right)$, with $\mu_{\text{post}} \sim G(\boldsymbol{x}) \,|\, \mathcal{D}$, $\widetilde{\mu}_{\text{post}} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \mathcal{D}$, probability measures defined as in the following Definition 3.3.

Remark 3.2. The likelihood function can can be extended to the space $\mathbb{R}^{n_L \times (k+k')}$, by artificially making it depend on the test input set while disregarding its contribution. Consequently, the entire reasoning developed below extends naturally to this more general case through a straightforward change of variables. However, to maintain a simpler notation and ensure a coherent presentation, we state and prove our main result under the framework introduced in Section 3.1, i.e., assuming $\boldsymbol{x} = \boldsymbol{x}_{\mathcal{D}}$.

Definition 3.3. Given a BNN as in Definition 2.3, we assume to have a hierarchical model as the one described in eq. (9) for the NNGP, and an analogous model for the BNN (i.e., Gaussian prior on $\boldsymbol{\theta}$ as in eq. (4), prior on the variance σ^2 as in eq. (8) and a likelihood as in eq. (13)). Then, for any $A \in \mathcal{B}\left(\mathbb{R}^{n_L \times k}\right)$, we define the posterior measures as follows: given $\mu \sim G(\boldsymbol{x})$ and $\widetilde{\mu} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x})$,

$$\mu_{\text{post}}(A) := \int_A \int_{\mathbb{R}^+} \frac{1}{I} \mathcal{L}(\boldsymbol{z}, s) p_{\sigma^2}(s) ds \mu(d\boldsymbol{z}),\ I := \int_{\mathbb{R}^{n_L \times k}} \int_{\mathbb{R}^+} \mathcal{L}(\boldsymbol{z}, s) p_{\sigma^2}(s) ds \mu(d\boldsymbol{z}),$$
$$\widetilde{\mu}_{\text{post}}(A) := \int_A \int_{\mathbb{R}^+} \frac{1}{\widetilde{I}} \mathcal{L}(\boldsymbol{z}, s) p_{\sigma^2}(s) ds \widetilde{\mu}(d\boldsymbol{z}),\ \widetilde{I} := \int_{\mathbb{R}^{n_L \times k}} \int_{\mathbb{R}^+} \mathcal{L}(\boldsymbol{z}, s) p_{\sigma^2}(s) ds \widetilde{\mu}(d\boldsymbol{z}).$$

3.3. **Main result.** Building on Definition 3.3, the main result of this work can be summarized in the following Theorem 3.4.

Theorem 3.4. *Let $f_{\boldsymbol{\theta}}$, G, $\boldsymbol{x}$ and $\boldsymbol{y}_{\mathcal{D}}$ as above, $\mathcal{L}$ density of a $\mathcal{N}\left(\boldsymbol{z}, \sigma^2_{y_{\mathcal{D}}} \boldsymbol{I}_{n_L \times k}\right)$. Assume a common variance for the last hidden layer of the BNN and the likelihood, distributed as an Inverse-Gamma*

$$\sigma^2 := \sigma^2_{\boldsymbol{W}^{(L)}} = \sigma^2_{\boldsymbol{b}^{(L)}} = \sigma^2_{y_{\mathcal{D}}},\ \sigma^2 \sim \mathcal{IG}(a, b),$$

with

$$a > \frac{1}{2},\ b > \left(1 + \frac{\varepsilon + 2}{2\varepsilon + 2}\right) \left\|\boldsymbol{y}_{\mathcal{D}}\right\|_F^2,\ \textit{for any } \varepsilon < 1/\left\|\boldsymbol{K}'(\boldsymbol{x})\right\|_{\text{op}}. \tag{14}$$

Then, there exists a constant $c > 0$, independent of $(n_l)_{l=1}^{L-1}$, such that

$$\mathcal{W}_1\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \mathcal{D}, G(\boldsymbol{x}) \,|\, \mathcal{D}\right) \leq \frac{c}{\sqrt{n_{min}}}.$$

Sketch of the proof. The idea is to show the convergence of $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \mathcal{D}$ to $G(\boldsymbol{x}) \,|\, \mathcal{D}$ through the following steps:

[2] Now $\mathcal{L}$ depends also on $s := \sigma^2$, which is no more a parameter.

(1) partially retracing the strategy introduced by Trevisan [Tre23], we first prove that there exist some constants $c > 0$ independent of $(n_l)_{l=1}^{L-1}$ and σ^2, and a function $h : \mathbb{R}^+ \to \mathbb{R}^+$ independent on $(n_l)_{l=1}^{L-1}$ as well, such that

$$\mathcal{W}_1\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) \mid (\mathcal{D}, \sigma^2), G(\boldsymbol{x}) \mid (\mathcal{D}, \sigma^2)\right) \leq h(\sigma^2)\frac{c_0}{\sqrt{n_{min}}}; \tag{15}$$

(2) we try to apply the convexity property of 1-Wasserstein distance in the following Proposition 3.5 (proof in Appendix A.3) to the families of probabilities $\left(\mathbb{P}_{f_{\boldsymbol{\theta}}(\boldsymbol{x}) \mid (\mathcal{D}, \sigma^2)}\right)_{\sigma^2 \in \mathbb{R}^+}$ and $\left(\mathbb{P}_{G(\boldsymbol{x}) \mid (\mathcal{D}, \sigma^2)}\right)_{\sigma^2 \in \mathbb{R}^+}$, which would lead to the thesis.

Proposition 3.5. *Let us consider two Markov kernels $(\mu(s))_{s \in \mathbb{R}^+}$, $(\widetilde{\mu}(s))_{s \in \mathbb{R}^+}$ with source $\mathbb{R}^+$ and target $\mathbb{R}^T$ and a measure ν on $\mathbb{R}^+$ (T finite set). Defining the probability measures on $\mathbb{R}^T$ such that, $\forall\, B \in \mathcal{B}\left(\mathbb{R}^T\right)$,*

$$\mu(B) := \int_{\mathbb{R}^+} \mu(s)(B) d\nu(s), \quad \widetilde{\mu}(B) := \int_{\mathbb{R}^+} \widetilde{\mu}(s)(B) d\nu(s),$$

the following convexity property for the distance $\mathcal{W}_1$ holds:

$$\mathcal{W}_1\left(\mu, \widetilde{\mu}\right) \leq \int_{\mathbb{R}^+} \mathcal{W}_1\left(\mu(s), \widetilde{\mu}(s)\right) d\nu(s). \tag{16}$$

Unfortunately the second step is not easy as it could seem since Proposition 3.5 requires two families of probability measures (in particular Markov kernels) integrated with respect to the same measure ν, which is not exactly our setting. Let us describe the issue before approaching the solution. For this purpose it is useful to introduce

$$I_{\sigma^2}(s) := \int_{\mathbb{R}^{n_L \times k}} \mathcal{L}(\boldsymbol{z}, s)\mu(d\boldsymbol{z}), \quad \widetilde{I}_{\sigma^2}(s) := \int_{\mathbb{R}^{n_L \times k}} \mathcal{L}(\boldsymbol{z}, s)\widetilde{\mu}(d\boldsymbol{z}),$$

which allow us to write, $\forall\, A \in \mathcal{B}\left(\mathbb{R}^{n_L \times k}\right)$,

$$\begin{aligned} \mu_{\text{post}}(A) &= \int_{\mathbb{R}^+} \mu_{\sigma^2}(s)(A)\frac{I_{\sigma^2}(s)}{I} p_{\sigma^2}(s) ds, \text{ with } \mu_{\sigma^2}(s)(A) := \int_A \frac{\mathcal{L}(\boldsymbol{z}, s)\mu(d\boldsymbol{z})}{I_{\sigma^2}(s)}, \\ \widetilde{\mu}_{\text{post}}(A) &= \int_{\mathbb{R}^+} \widetilde{\mu}_{\sigma^2}(s)(A)\frac{\widetilde{I}_{\sigma^2}(s)}{\widetilde{I}} p_{\sigma^2}(s) ds, \text{ with } \widetilde{\mu}_{\sigma^2}(s)(A) := \int_A \frac{\mathcal{L}(\boldsymbol{z}, s)\widetilde{\mu}(d\boldsymbol{z})}{\widetilde{I}_{\sigma^2}(s)}. \end{aligned} \tag{17}$$

Eventually, we can note that the two families $(\mu_{\sigma^2}(s))_{s \in \mathbb{R}^+}$ and $(\widetilde{\mu}_{\sigma^2}(s))_{s \in \mathbb{R}^S}$, which coincide respectively with $\left(\mathbb{P}_{G(\boldsymbol{x}) \mid (\mathcal{D}, \sigma^2)}\right)_{\sigma^2 \in \mathbb{R}^+}$ and $\left(\mathbb{P}_{f_{\boldsymbol{\theta}}(\boldsymbol{x}) \mid (\mathcal{D}, \sigma^2)}\right)_{\sigma^2 \in \mathbb{R}^+}$, are integrated with respect to different measures. Hence, it is clear that to apply eq. (16) it is necessary to use the triangle inequality,

$$\mathcal{W}_1\left(\mu_{\text{post}}, \widetilde{\mu}_{\text{post}}\right) \leq \mathcal{W}_1\left(\mu_{\text{post}}, \bar{\mu}\right) + \mathcal{W}_1\left(\bar{\mu}, \widetilde{\mu}_{\text{post}}\right), \tag{18}$$

where we inserted a third measure,

$$\bar{\mu}(A) := \int_{\mathbb{R}^+} \widetilde{\mu}_{\sigma^2}(s)(A)\frac{I_{\sigma^2}(s)}{I} p_{\sigma^2}(s) ds, \ \forall\, A \in \mathcal{B}\left(\mathbb{R}^{n_L \times k}\right), \tag{19}$$

specifically constructed to satisfy the hypothesis of the convexity property.
Now, to conclude, we just need to control both the terms on the right-hand side of eq. (18).

1^{st} term. We just apply the aforementioned convexity property getting

$$\mathcal{W}_1\left(\mu_{\text{post}}, \bar{\mu}\right) \leq \int_{\mathbb{R}^+} \mathcal{W}_1\left(\mu_{\sigma^2}(s), \widetilde{\mu}_{\sigma^2}(s)\right) \frac{I_{\sigma^2}(s)}{I} p_{\sigma^2}(s) ds.$$

Therefore by eq. (15) we get

$$\mathcal{W}_1\left(\mu_{\text{post}}, \bar{\mu}\right) \leq \frac{c_0}{\sqrt{n_{min}}} \int_{\mathbb{R}^+} h(s)\frac{I_{\sigma^2}(s)}{I} p_{\sigma^2}(s) ds \leq \frac{c_1}{\sqrt{n_{min}}},$$

where, in order to bound the last integral it is necessary to introduce a constraint on a and b, as in eq. (14).

2^{nd} term. We exploit the following technical Lemma 3.6, in combination with several bounds on the first moments of the considered probability measures.

Lemma 3.6. *Let μ, ν be measures on $\mathbb{R}^S$, then, denoting with $|\mu - \nu|$ the total variation measure, it holds*

$$\mathcal{W}_1(\mu, \nu) \leq \int_{\mathbb{R}^S} \|\boldsymbol{u}\| \, d|\mu - \nu|(\boldsymbol{u}).$$

□

Remark 3.7. All the details of the sketched part of the previous proof can be found in Appendix C: a formal statement and a proof of eq. (15) can be found in Appendix C.1; additionally, the bounds for the two terms found using the triangle inequality (eq. (18)) can be found respectively in Appendix C.2 and Appendix C.3. In Figure 1, we include a chart illustrating the dependencies among the results that lead to the proof of Theorem 3.4.

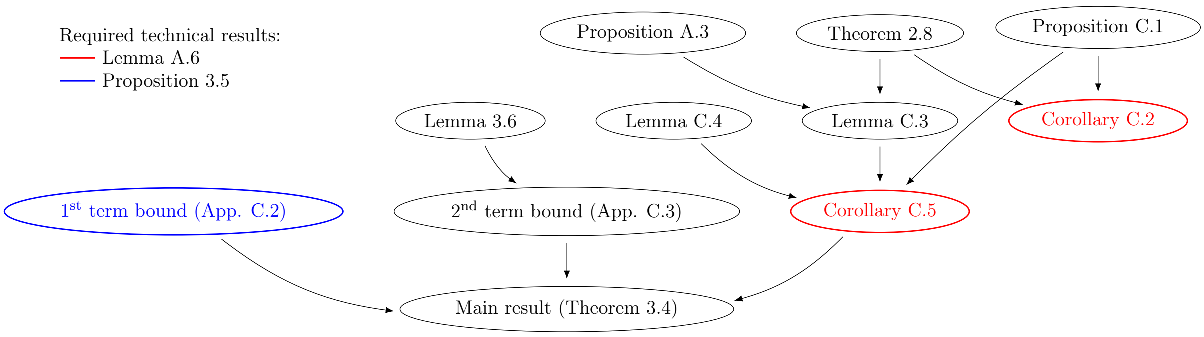

FIGURE 1. Dependency of results for the proof of Theorem 3.4.

By exploiting the connection between $\mathcal{W}_1$ and weak convergence, together with what was observed in Section 3.1, Theorem 3.4 leads us to a characterization of the asymptotic behavior of the exact posterior law of a BNN trained following the Gaussian-Inverse-Gamma model, showing convergence to a Student-t process in the infinite-width limit.

Corollary 3.8. *Under the same assumptions of Theorem 3.4 and $n_L = 1$, the posterior of the BNN $f_{\boldsymbol{\theta}}$ with Gaussian-Inverse-Gamma prior and Gaussian likelihood, evaluated in the input set $\boldsymbol{x}$, converges in law to a multivariate Student-t variable with $2a + k$ degrees of freedom:*

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \mathcal{D} \xrightarrow[n_{min}\to\infty]{law} t_{2a+k}\left(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}}\right),$$

with M, $\boldsymbol{\mu}_{\text{post}}$ and $\boldsymbol{\Sigma}_{\text{post}}$ as in eq. (11).

4. Simulations

We present a procedure to sample from the posterior distribution of a BNN, ensuring consistency between the theoretical results and practical implementations. We consider a hierarchical Bayesian model, where we place a prior on both the network parameters, $\boldsymbol{\theta} = (\boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}, \dots, \boldsymbol{W}^{(L)}, \boldsymbol{b}^{(L)})$, and the variance σ^2:

$$\begin{aligned}
&\boldsymbol{W}^{(l)} \sim \mathcal{N}\left(\boldsymbol{0}_{n_l \times n_{l-1}}, \sigma^2_{\boldsymbol{W}^{(l)}}/n_{l-1}\, \boldsymbol{I}_{n_l \times n_{l-1}}\right), && \boldsymbol{b}^{(l)} \sim \mathcal{N}\left(\boldsymbol{0}_{n_l}, \sigma^2_{\boldsymbol{b}^{(l)}} \boldsymbol{I}_{n_l}\right), \text{ for } l \in [L-1], \\
&\boldsymbol{W}^{(L)} \,|\, \sigma^2 \sim \mathcal{N}\left(\boldsymbol{0}_{n_L \times n_{L-1}}, \sigma^2/n_{L-1}\, \boldsymbol{I}_{n_L \times n_{L-1}}\right), && \boldsymbol{b}^{(L)} \,|\, \sigma^2 \sim \mathcal{N}\left(\boldsymbol{0}_{n_L}, \sigma^2 \boldsymbol{I}_{n_L}\right), \\
&\sigma^2 \sim \mathcal{IG}(a, b), && \boldsymbol{y}_{\mathcal{D}} \,|\, \boldsymbol{\theta}, \boldsymbol{x}_{\mathcal{D}}, \sigma^2 \sim \mathcal{N}\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathcal{D}}), \sigma^2 \boldsymbol{I}_{n_L \times k}\right).
\end{aligned} \tag{20}$$

The Monte Carlo sampling strategy is summarized in Algorithm 1. In particular, the idea is to sample from $\boldsymbol{\theta}, \sigma^2 \,|\, \mathcal{D}$ using a Gibbs sampling scheme. To implement line 5 we rely on MCMC methods, which have been widely studied in the literature of Bayesian optimization for BNNs. Several versions of such strategies are implemented in the `Python` library `Pyro` [Bin+19; Pas+19], and among these, we applied the No-U-turn sampler [HG11]. Whereas, in order to sample from

Algorithm 1: TRAINING OF A BNN UNDER GAUSSIAN-INVERSE-GAMMA PRIOR

Input: $\boldsymbol{\alpha}$ [architecture], $\boldsymbol{\sigma}$ [BNN's variances], (a, b) [Inverse-Gamma parameters], f [function]

1. **build** training set $\mathcal{D}$ starting from a reference function f
2. **build** test set $\mathcal{T}$ based on a fine partitioning of the domain of f
3. **initialize** $\sigma^2_{(0)}$
4. **for** $i = 0, \dots m$ **do**
5.   **sample** $\boldsymbol{\theta}_{(i+1)} \,|\, \sigma^2, \mathcal{D}$ using `NUTS`
6.   **sample** $\sigma^2_{(i+1)} \,|\, \boldsymbol{\theta}_{(i)}, \mathcal{D}$ from $p_{\sigma^2 | \boldsymbol{\theta}_{(i+1)}, \mathcal{D}}(\sigma^2)$
7. **return** $f_{\boldsymbol{\theta}_{(m)}}(\mathcal{T}) \,|\, \mathcal{D}$

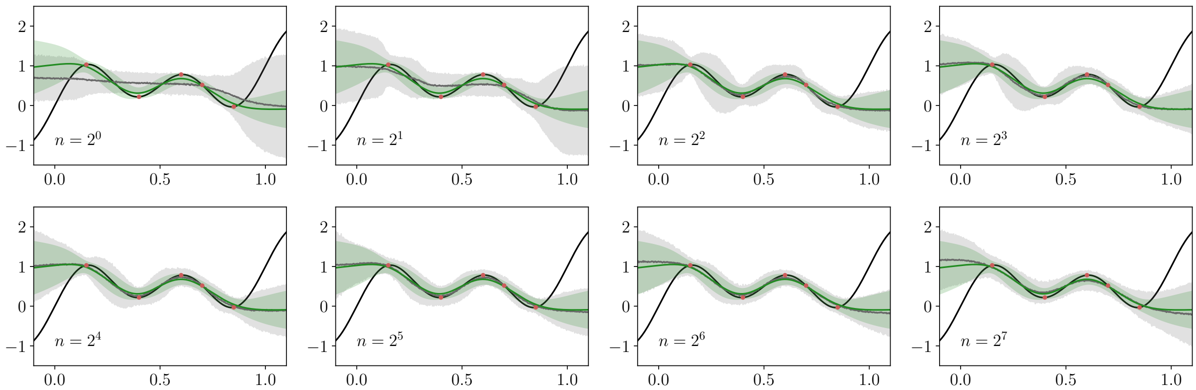

FIGURE 2. Sequence of posterior BNNs, $(f_{\boldsymbol{\theta}_n} \,|\, \mathcal{D})_n$ (in gray), converging to the corresponding posterior Student-t process, $G \,|\, \mathcal{D}$ (in green), in the infinite-width limit. Given $\mathcal{D}$ (in red), training set, we sampled 100 values from both $G \,|\, \mathcal{D}$ and $f_{\boldsymbol{\theta}_n} \,|\, \mathcal{D}$ for each width $n \in \{2^0, \dots, 2^7\}$, following Remark 3.1 and Algorithm 1, respectively. The networks used have 2 hidden layers, `erf` activations and parameter variances set to 5. Additionally, the hyperparameters (a, b) are set to $(3, 2)$.

the marginal posterior of the variance (line 6) we exploit the strategy adopted by Ding et al. [DEM22, Appendix D], since it is possible to show that it follows a positive Generalized Inverse Gaussian distribution (referred as $\mathcal{GIN}^+$ by the authors). Such a derivation, as well as additional implementation details, can be found in Appendix D.
We report in Figure 2 the comparison between a sequence of BNNs trained using the strategy discussed above and the limiting Student-t process discussed in Appendix B (see Remark 3.1). As the (all equal) widths of the hidden layers increase, the models' output distributions become closer. We also consider the case in which we replace the Gaussian-Inverse-Gamma prior with the classical Gaussian prior with fixed variance (see Figure 3). In this case the limiting process to which the sequence of BNNs converges is simply the posterior NNGP (see Hron et al.; Trevisan [Hro+20; Tre23]).

Remark 4.1. We can observe that in Figure 3, i.e., under Gaussian prior, the convergence is much faster and more precise compared to the Gaussian-Inverse-Gamma prior case (Figure 2). This behavior, while likely influenced by our specific sampling procedure for the posterior BNNs, is also consistent with theoretical expectations. Indeed, although the theoretical convergence rates of the limiting processes are identical for both the Gaussian prior and the Gaussian-Inverse-Gamma prior cases (see Corollary C.2 and Theorem 3.4), the associated multiplicative constants differ significantly in magnitude. Therefore, given a common fixed width, different distances between the posterior BNNs and their limiting processes are to be expected.

We conclude by comparing the asymptotic processes under both frameworks. The posterior Student-t process models the variance of the data more accurately compared to the posterior Gaussian process. This is an expected behavior, as the Gaussian-Inverse-Gamma model

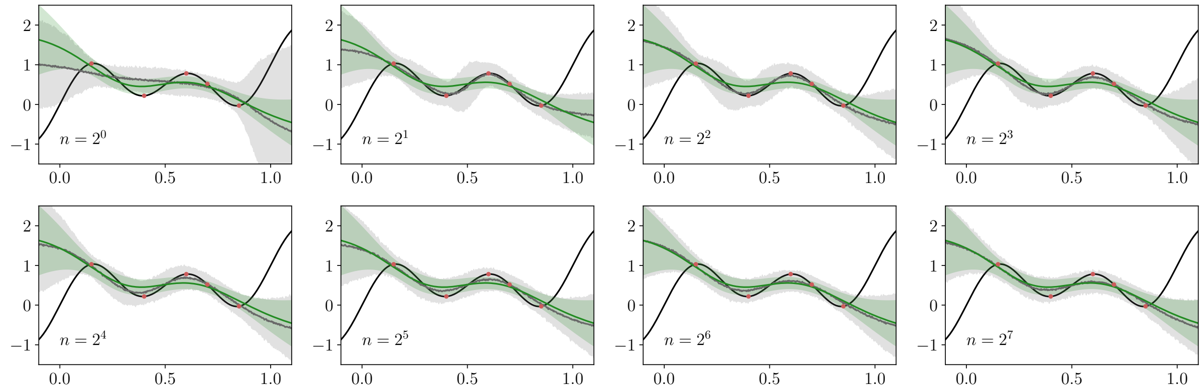

FIGURE 3. Sequence of posterior BNNs, $(f_{\boldsymbol{\theta}_n} \mid \mathcal{D})_n$ (in gray), converging to the corresponding posterior Gaussian process, $G \mid \mathcal{D}$ (in green), in the infinite-width limit. Given $\mathcal{D}$ (in red), training set, we sampled 100 values from both $G \mid \mathcal{D}$ and $f_{\boldsymbol{\theta}_n} \mid \mathcal{D}$ for each width $n \in \{2^0, \dots, 2^7\}$. The sampling was performed following Rasmussen and Williams [RW06, eqs. (2.22)-(2.24)] for $G \mid \mathcal{D}$ and the built-in `NUTS` algorithm in `Pyro` for $f_{\boldsymbol{\theta}_n} \mid \mathcal{D}$. The networks used have 2 hidden layers, `erf` activations, parameter variances set to 2, and likelihood variance set to 0.1.

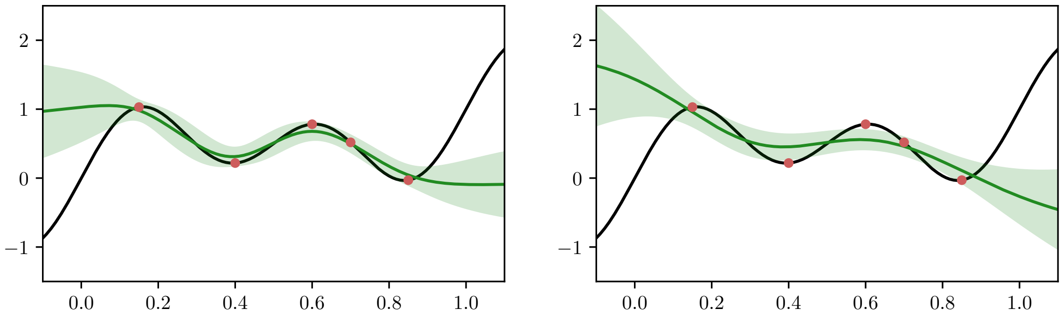

FIGURE 4. Posterior Student-t process (on the right) and posterior Gaussian process (on the left). We followed the same strategy and used the same parameters introduced to generate Figures 2 and 3.

explicitly estimates the data variance during the Bayesian learning, whereas no such estimation is performed when we use a Gaussian prior. This final result also highlights that using a Gaussian-Inverse-Gamma prior provides a more accurate representation of the data, particularly in scenarios in which the dataset is relatively small.

REFERENCES

[Ban+20] Maria Bankestad, Jens Sjolund, Jalil Taghia, and Thomas Schon. "The elliptical processes: a family of fat-tailed stochastic processes". In: *Preprint arXiv:2003.07201* (2020) (cit. on p. 2).

[Bin+19] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. "Pyro: Deep Universal Probabilistic Programming". In: *Journal of Machine Learning Research* 20.28 (2019), pp. 1–6 (cit. on p. 8).

[Blu+15] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. "Weight Uncertainty in Neural Networks". In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. 2015 (cit. on p. 1).

[BS09] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2009 (cit. on p. 18).

[BT24] Andrea Basteri and Dario Trevisan. "Quantitative Gaussian approximation of randomly initialized deep neural networks". In: *Machine Learning* 113.9 (2024), pp. 6373–6393 (cit. on pp. 1, 2, 4).

[DEM22] Cheng Ding, Juan Estrada, and Santiago Montoya-Blandón. "Bayesian Inference of Network Formation Models with Payoff Externalities". In: *Preprint* (2022) (cit. on pp. 9, 29, 30).

[Fav+24] Stefano Favaro, Boris Hanin, Domenico Marinucci, Ivan Nourdin, and Giovanni Peccati. "Quantitative CLTs in Deep Neural Networks". In: *Preprint arXiv:2307.06092* (2024) (cit. on pp. 1, 4).

[FKN90] Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC, 1990 (cit. on p. 2).

[HG11] Matthew D. Hoffman and Andrew Gelman. "The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15 (2011), pp. 1593–1623 (cit. on p. 8).

[Hro+20] Jiri Hron, Yasaman Bahri, Roman Novak, Jeffrey Pennington, and Jascha Sohl-Dickstein. "Exact posterior distributions of wide Bayesian neural networks". In: *Workshop on Uncertainty and Robustness in Deep Learning, ICML*. 2020 (cit. on pp. 1, 9).

[Izm+21] Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. "What Are Bayesian Neural Network Posteriors Really Like?" In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021 (cit. on p. 1).

[Lee+18] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Narain Sohl-Dickstein. "Deep Neural Networks as Gaussian Processes". In: *International Conference on Learning Representations (ICLR)*. 2018 (cit. on pp. 1, 4, 5).

[Mat+18] Alexander Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. "Gaussian Process Behaviour in Wide Deep Neural Networks". In: *International Conference on Learning Representations (ICLR)*. 2018 (cit. on pp. 1, 4, 5).

[Nea96] Radford M. Neal. *Bayesian learning for Neural Networks*. Lecture Notes in Statistics. Springer New York, 1996 (cit. on pp. 1–3).

[Nov+20] Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. "Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes". In: *International Conference on Learning Representations (ICLR)*. 2020 (cit. on p. 1).

[Pas+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach 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: *Conference on Neural Information Processing Systems (NeurIPS)*. 2019 (cit. on p. 8).

[PFP24] Lucia Pezzetti, Stefano Favaro, and Stefano Peluchetti. "Function-space MCMC for Bayesian wide neural networks". In: *Preprint arXiv:2408.14325* (2024) (cit. on p. 1).

[Rud87] Walter Rudin. *Real and Complex Analysis*. Mathematics series. McGraw-Hill, 1987 (cit. on p. 26).

[RW06] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006 (cit. on pp. 1, 6, 10).

[SD23] Jeremy Sellier and Petros Dellaportas. "Bayesian online change point detection with Hilbert space approximate Student-t process". In: *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 2023 (cit. on p. 2).

[SWG14] Amar Shah, Andrew Wilson, and Zoubin Ghahramani. "Student-t Processes as Alternatives to Gaussian Processes". In: *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2014 (cit. on p. 2).

[Tre23] Dario Trevisan. "Wide Deep Neural Networks with Gaussian Weights are Very Close to Gaussian Processes". In: *Preprint arXiv:2312.11737* (2023) (cit. on pp. 1, 2, 4, 7, 9, 16, 20).

[TW18] Brendan D. Tracey and David Wolpert. "Upgrading from Gaussian Processes to Student's-T Processes". In: *Non-Deterministic Approaches Conference, American Institute of Aeronautics and Astronautics (AIAA)*. 2018 (cit. on p. 2).

[Vil08] Cédric Villani. *Optimal transport: Old and New*. Springer Berlin, Heidelberg, 2008 (cit. on pp. 2, 16, 26).

[Yan21] Greg Yang. "Tensor Programs I: Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021 (cit. on p. 1).

Contents

Appendix A. Notation

We summarize below the preliminary tools and notations used throughout the article. Proofs of standard results are omitted, with references provided. For clarity, the content is organized into thematic sections.

A.1. **Tensors.** Given S finite set (e.g., $[n]$ or $[n_1] \times \cdots \times [n_k]$, for some $k \geq 2$, where $[n] = \{1, \ldots, n\}$) we denote with $\mathbb{R}^S$ the vector space of real valued functions $\boldsymbol{v} : S \to \mathbb{R}$ (column vectors which generalizes to multidimensional tensors). We adopt the notation $\boldsymbol{v}_s := \boldsymbol{v}(s)$, $\forall\, s \in S$, and we introduce the following conventions: $\boldsymbol{e}_s$ is the s^{th} vector of the canonical base ($\boldsymbol{e}_s(s) = 1$ and $\boldsymbol{e}_s(r) = 0$, $\forall\, r \in S \setminus \{s\}$), $\mathbf{1}_S := \sum_{s \in S} \boldsymbol{e}_s$ and $\mathbf{0}_S$ is the constant null vector. When S is used as a subscript/superscript we also simplify the notation: $[n_1] \times \cdots \times [n_k]$ becomes just $n_1 \times \cdots \times n_k$.
Given S, T two finite sets, we can see the vector space $\mathbb{R}^{S \times T}$ as the space of linear transformations $\boldsymbol{A} : \mathbb{R}^T \to \mathbb{R}^S$, defining

$$\forall\, \boldsymbol{v} \in \mathbb{R}^T,\ \boldsymbol{A}\boldsymbol{v} := \boldsymbol{A}(\boldsymbol{v}) = \sum_{s \in S} \left(\sum_{t \in T} \boldsymbol{A}_{s,t} \boldsymbol{v}_t \right) \boldsymbol{e}_s.$$

If $S = [n]$ and $T = [m]$ then $\boldsymbol{A}$ can be represented as a standard matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, where $\boldsymbol{A}_{i,j} := (\boldsymbol{A}(\boldsymbol{e}_j))(i)$, $\forall\, i \in [n], j \in [m]$. Analogously if $S = [n_1] \times [n_2]$ and $T = [m_1] \times [m_2]$ it is possible to represent $\boldsymbol{A}$ as a 4-dimensional tensor such that $\boldsymbol{A}_{i,j,k,l} := \boldsymbol{A}(\boldsymbol{e}_{(k,l)})((i,j))$, $\forall\, (i,j) \in [n_1] \times [n_2], (k,l) \in [m_1] \times [m_2]$.
If $S = T$, we introduce the notation Sym^S for the set of symmetric linear transformations in $\mathbb{R}^{S \times S}$, i.e., $\boldsymbol{A} \in \mathrm{Sym}^S$ if and only if $\boldsymbol{A} \in \mathbb{R}^{S \times S}$, $\boldsymbol{A} = \boldsymbol{A}^T$. Moreover, we denote with Sym^S_+ the

subset of Sym^S composed by symmetric, positive definite matrices, where a matrix $\boldsymbol{A}$ is said to be positive definite if $\forall\, \boldsymbol{v} \in \mathbb{R}^S \setminus \{0\}$, $\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} > 0$.

We define the outer product (or tensor product) between $\boldsymbol{v} \in \mathbb{R}^S$ and $\boldsymbol{w} \in \mathbb{R}^T$ as $\boldsymbol{v} \otimes \boldsymbol{w} \in \mathbb{R}^{S\times T}$ with $(\boldsymbol{v} \otimes \boldsymbol{w})_{s,t} := \boldsymbol{v}_s \boldsymbol{w}_t$, $\forall\, s \in S, t \in T$, denoting $\boldsymbol{v} \otimes \boldsymbol{v}$ as $\boldsymbol{v}^{\otimes 2}$. We also define the identity map $\boldsymbol{I}_S : \mathbb{R}^S \to \mathbb{R}^S$ as $\boldsymbol{I}_S := \sum_{s\in S} \boldsymbol{e}_s^{\otimes 2}$, observing that, $\forall\, \boldsymbol{v} \in \mathbb{R}^S$, $\boldsymbol{I}_S \boldsymbol{v} = \sum_{s\in S}(\sum_{t\in S}(\boldsymbol{I}_S)_{s,t}\boldsymbol{v}_t)\boldsymbol{e}_s = = \sum_{s\in S} \boldsymbol{v}_s \boldsymbol{e}_s = \boldsymbol{v}$. If $S = [n]$, $T = [m]$ then $\boldsymbol{v} \otimes \boldsymbol{w} = \boldsymbol{v}\boldsymbol{w}^T$.
Given a generic pair of elements in a real vector space, $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^S$, we define the standard scalar product between them as

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle := \mathrm{tr}\,(\boldsymbol{v} \otimes \boldsymbol{w}) = \sum_{s\in S} \boldsymbol{v}_s \boldsymbol{w}_s.$$

Analogously we can define the Euclidean norm (or 2-norm) induced by the scalar product as

$$\|\boldsymbol{v}\| = \langle \boldsymbol{v}, \boldsymbol{v} \rangle^{1/2} = \left(\sum_{s\in S} \boldsymbol{v}_s^2\right)^{1/2}. \tag{21}$$

If $S = [n]$, $n \in \mathbb{N}_{>0}$, we also define the p-norm of $\boldsymbol{v}$, $p \geq 1$, as $\|\boldsymbol{v}\|_p := (\sum_{i=1}^n \boldsymbol{v}_i^p)^{1/p}$. If nothing is specified we always consider $\mathbb{R}^n$ as a normed spaced with the Euclidean norm defined in eq. (21). In the matrix case, $S = [n] \times [m]$, this norm is usually referred as Frobenius norm, therefore for an improved readability we denote it as $\|\cdot\|_F$.
Given an operator $\boldsymbol{A} \in \mathrm{Sym}^S$, with $S = [n]$, we define the operator norm as

$$\|\boldsymbol{A}\|_{\mathrm{op}} := \sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2 = \max\{\lambda \,|\, \lambda \in \mathrm{Sp}(\boldsymbol{A})\}.$$

Lemma A.1. *For every $\boldsymbol{x}, \boldsymbol{y} \in (\mathbb{R}^S, \|\cdot\|)$ real vector space with the Euclidean norm, the following inequalities hold:*

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 \leq 2(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2), \tag{22}$$

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 \geq \frac{\varepsilon}{\varepsilon+1}\|\boldsymbol{x}\|^2 - \varepsilon\|\boldsymbol{y}\|^2, \forall\, \varepsilon \in \mathbb{R}^+. \tag{23}$$

Proof. We prove both the statements using the triangle inequality and its inverse: $\forall\, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^S$, $\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$ and $\|\boldsymbol{u} - \boldsymbol{v}\| \geq \|\boldsymbol{u}\| - \|\boldsymbol{v}\|$.
Let us start observing that $\forall\, a, b, \varepsilon > 0$, we have

$$(a+b)^2 = a^2 + b^2 + \varepsilon 2a\frac{b}{\varepsilon} \leq a^2 + b^2 + \varepsilon\left(a^2 + \frac{b^2}{\varepsilon^2}\right) = a^2(1+\varepsilon) + b^2\left(1 + \frac{1}{\varepsilon}\right).$$

With $\varepsilon = 1$ we get $(a+b)^2 \leq 2(a^2+b^2)$, and therefore by triangle inequality follows the first result,

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 \leq (\|\boldsymbol{x}\| + \|\boldsymbol{y}\|)^2 \leq 2\left(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2\right).$$

For the second inequality we use the reverse triangle inequality,

$$\|\boldsymbol{x}\|^2 \leq (\|\boldsymbol{x} - \boldsymbol{y}\| + \|\boldsymbol{y}\|)^2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|^2(1+\varepsilon) + \|\boldsymbol{y}\|\left(1 + \frac{1}{\varepsilon}\right),$$

which implies

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 \geq \|\boldsymbol{x}\|^2 \frac{1}{1+\varepsilon} - \|\boldsymbol{y}\|^2 \frac{1}{\varepsilon}.$$

Now by simply substituting $\varepsilon' = \varepsilon^{-1}$ in place of ε we get $(1 + (\varepsilon')^{-1})^{-1} = \varepsilon'/(\varepsilon'+1)$ and so the thesis. □

A.2. **Random variables.** Given S finite set, a random variable $\boldsymbol{x}$ with values in $\mathbb{R}^S$ is a measurable map

$$\boldsymbol{x} : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^S, \mathcal{B}\left(\mathbb{R}^S\right)).$$

It is adopted the same notation for deterministic tensors introduced in Appendix A.1 and random tensors, i.e., random variable with values in tensor spaces. We denote with $\mathbb{P}_{\boldsymbol{x}}$ the distribution (or law) of $\boldsymbol{x}$,

$$\mathbb{P}_{\boldsymbol{x}} := \mathbb{P}\left(\boldsymbol{x}^{-1}(A)\right), \forall A \in \mathcal{B}\left(\mathbb{R}^S\right).$$

We write $\boldsymbol{x} \sim \boldsymbol{y}$ if two random variables share the same distribution.

Given $\boldsymbol{x}$ random variable with values on $\mathbb{R}^S$ we define its mean value, or first moment, and its variance[3], or second moment of $\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]$, respectively as

$$\mathbb{E}[\boldsymbol{x}] := (\mathbb{E}[\boldsymbol{x}_s])_{s\in S} = \left(\int_\Omega \boldsymbol{x}_s(\omega) d\mathbb{P}(\omega)\right)_{s\in S} \in \mathbb{R}^S,$$

and

$$\operatorname{Var}(\boldsymbol{x}) := \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^{\otimes 2}\right] \in \operatorname{Sym}_+^S.$$

Given $p \geq 1$, we define the Lebesgue norm of order p of $\boldsymbol{x}$ as

$$\|\boldsymbol{x}\|_{L^p} := \mathbb{E}\left[\|\boldsymbol{x}\|^p\right]^{1/p} \in \mathbb{R}.$$

Recalling that $\|\boldsymbol{x}\|^2 = \operatorname{tr}\left(\boldsymbol{x}^{\otimes 2}\right)$, thanks to the linearity of the integral, it is possible to exchange trace and expectation so that we can write

$$\|\boldsymbol{x}\|_{L^2}^2 = \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{x}^{\otimes 2}\right)\right] = \operatorname{tr}\left(\mathbb{E}\left[\boldsymbol{x}^{\otimes 2}\right]\right).$$

A.2.1. *Gaussian random variables.* Given $\boldsymbol{\mu} \in \mathbb{R}^S$, $\boldsymbol{\Sigma} \in \operatorname{Sym}_+^S$, we denote with $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the distribution of a Gaussian random variable $\boldsymbol{x} : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^S, \mathcal{B}\left(\mathbb{R}^S\right))$, such that

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu} \quad \text{and} \quad \operatorname{Var}(\boldsymbol{x}) = \mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{\mu})^{\otimes 2}\right] = \boldsymbol{\Sigma}.$$

Remark A.2. Let $\left(\mathbb{R}^S, \|\cdot\|\right)$ be a normed space with a Euclidean norm and $\boldsymbol{\mu} = \mathbf{0}_S$. Then,

$$\mathbb{E}\left[\|\boldsymbol{x}\|\right]^2 \leq \mathbb{E}\left[\|\boldsymbol{x}\|^2\right] = \mathbb{E}\left[\sum_{s\in S} \boldsymbol{x}_s^2\right] = \sum_{s\in S} \mathbb{E}\left[\boldsymbol{x}_s^2\right] = \operatorname{tr}(\boldsymbol{\Sigma}).$$

A.2.2. *Inverse-Gamma random variables.* A random variable s, with values in $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$, is said to be Inverse-Gamma distributed with parameters a and b in $\mathbb{R}^+$ (denoted as $s \sim \mathcal{IG}(a, b)$), if $\mathbb{P}_s$ admits a density with respect to the Lebesgue measure λ^+ on $\mathbb{R}^+$, and in particular

$$p_s(s) := \frac{d\mathbb{P}_s}{d\lambda^+}(s) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{s}\right)^{a+1} \exp\left(-\frac{b}{s}\right).$$

As the name suggests, and a simple change of variables shows, one can equivalently say that the variable $1/s$ is Gamma distributed with shape and rate parameters (a, b).

A.2.3. *Multivariate Student-t random variables.* Given $k \in \mathbb{N}_{>0}$, a random variable $\boldsymbol{z}$, with values in $\left(\mathbb{R}^k, \mathcal{B}\left(\mathbb{R}^k\right)\right)$, is said to be k-dimensional Student-t distributed with ν degrees of freedom, location $\boldsymbol{\mu} \in \mathbb{R}^k$ and scale $\boldsymbol{\Sigma} \in \operatorname{Sym}_+^k$ (denoted as $\boldsymbol{z} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), if $\mathbb{P}_{\boldsymbol{z}}$ admits a density with respect to the Lebesgue measure λ^k on $\mathbb{R}^k$, and in particular

$$p_{\boldsymbol{z}}(\boldsymbol{z}) := \frac{d\mathbb{P}_{\boldsymbol{z}}}{d\lambda^k}(\boldsymbol{z}) = \frac{\Gamma((\nu+k)/2)}{\Gamma(\nu/2)(\nu\pi)^{k/2}\det(\boldsymbol{\Sigma})^{1/2}} \left(1 + \frac{1}{\nu}(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})\right)^{-\frac{\nu+k}{2}}.$$

[3]Often referred as covariance if $|S| > 1$.

A.3. **Wasserstein distance.** We first recall two well-known properties (see e.g., Villani [Vil08]) of the Wasserstein metric, together with the main two technical tools exploited in the demonstration of our core result: Proposition 3.5 and Lemma A.6.

Proposition A.3. *Given two random variables $\boldsymbol{x}$, $\boldsymbol{y}$ with values in $\mathbb{R}^S$ and a positive constant a, $\forall\, p \geq 1$, it holds*

$$\mathcal{W}_p\left(a\boldsymbol{x}, a\boldsymbol{y}\right) = a\mathcal{W}_p\left(\boldsymbol{x}, \boldsymbol{y}\right). \tag{24}$$

Theorem A.4 (Kantorovich duality for $\mathcal{W}_1$)**.** *Given T finite set, μ and $\widetilde{\mu}$ probability measures on $\mathbb{R}^T$, then we have*

$$\mathcal{W}_1\left(\mu, \widetilde{\mu}\right) = \sup_{\substack{f:\mathbb{R}^T \to \mathbb{R}, \\ \mathrm{Lip}(f) \leq 1}} \left(\int_{\mathbb{R}^T} f(\boldsymbol{x}) d\mu(\boldsymbol{x}) - \int_{\mathbb{R}^T} f(\boldsymbol{x}) d\widetilde{\mu}(\boldsymbol{x}) \right). \tag{25}$$

The supremum above is computed over all the functions $f : \mathbb{R}^T \to \mathbb{R}$ that are Lipschitz continuous, with Lipschitz constant $\mathrm{Lip}\,(f) \leq 1$. Notice that one can further restrict to functions f such that $f(0) = 0$ since adding constants to f does not change the difference of the integrals.

Before we state and prove Proposition 3.5, let us recall the notion of Markov kernel.

Definition A.5 (Markov kernel)**.** Let us consider $(E, \mathcal{E})$, $(F, \mathcal{F})$ measurable spaces, a Markov kernel with source $(E, \mathcal{E})$ and target $(F, \mathcal{F})$ is a map $K_\mu : E \times \mathcal{F} \to [0, 1]$, such that

- $\forall\, B \in \mathcal{F}$, the map $s \to K_\mu(s, B)$ for $s \in E$ is measurable from $(E, \mathcal{E})$ to $([0, 1], \mathcal{B}\left([0, 1]\right))$;
- $\forall\, s \in E$, the map $B \to K_\mu(s, B)$ for $B \in \mathcal{F}$ is a probability measure on $(F, \mathcal{F})$.

For any fixed $s \in E$, we denote $\mu(s) \coloneqq K_\mu(s, \cdot)$ and the Markov kernel as $K_\mu = (\mu(s))_{s \in E}$.

Proof of Proposition 3.5. Consider $f : \mathbb{R}^T \to \mathbb{R}$ with $\mathrm{Lip}\,(f) \leq 1$. Then by eq. (25) with the measures $\mu(s)$ and $\widetilde{\mu}(s)$ it follows

$$\int_{\mathbb{R}^T} f(\boldsymbol{x}) d\mu(s)(\boldsymbol{x}) - \int_{\mathbb{R}^T} f(\boldsymbol{x}) d\widetilde{\mu}(s)(\boldsymbol{x}) \leq \mathcal{W}_1\left(\mu(s), \widetilde{\mu}(s)\right). \tag{26}$$

Integrating both sides in eq. (26) with respect to ν yields

$$\int_{\mathbb{R}^T} f(\boldsymbol{x}) d\mu(\boldsymbol{x}) - \int_{\mathbb{R}^T} f(\boldsymbol{x}) d\widetilde{\mu}(\boldsymbol{x}) \leq \int_{\mathbb{R}^+} \mathcal{W}_1\left(\mu(s), \widetilde{\mu}(s)\right) d\nu(s).$$

We conclude by taking the supremum over the possible f's and again by Kantorovich duality eq. (25)). □

The following Lemma A.6 shows that if two *prior* distributions are close with respect to the Wasserstein metric and the (common) Likelihood is sufficiently regular, then also the *posterior* distributions will be close, in a quantitative way.

We use the notation

$$m_p(\mu) \coloneqq \int_{\mathbb{R}^S} \|\boldsymbol{z}\|^p \, d\mu(\boldsymbol{z}).$$

for the moment of order $p \geq 1$ of a measure μ.

Lemma A.6 (Lemma 5.1 of Trevisan [Tre23])**.** *Let μ, $\widetilde{\mu}$ be probability measures on $(\mathbb{R}^S, \|\cdot\|)$ for some finite set S and finite moments of order $p \geq 1$. Fix $g : \mathbb{R}^S \to \mathbb{R}^+$ be a uniformly bounded (by $\|g\|_\infty$) Lipschitz continuous map (with constant $\mathrm{Lip}\,(g)$), such that*

$$\mu(g) \coloneqq \int_{\mathbb{R}^S} g(\boldsymbol{z}) d\mu(\boldsymbol{z}) > 0 \quad \textit{and} \quad \widetilde{\mu}(g) > 0.$$

Defining the probability measures $\mu_g \ll \mu$ and $\widetilde{\mu}_g \ll \widetilde{\mu}$, with respective densities $\frac{d\mu_g}{d\mu} \coloneqq \frac{g}{\mu(g)}$ and $\frac{d\widetilde{\mu}_g}{d\widetilde{\mu}} \coloneqq \frac{g}{\widetilde{\mu}(g)}$, it holds

$$\mathcal{W}_1\left(\widetilde{\mu}_g, \mu_g\right) \leq \frac{1}{\mu(g)} \left(\mathrm{Lip}\,(g)\, m_{p/(p-1)}(\mu) + \left(1 + \frac{m_1(\mu)\,\mathrm{Lip}\,(g)}{\widetilde{\mu}(g)}\right) \|g\|_\infty \right) \mathcal{W}_p\left(\widetilde{\mu}, \mu\right). \tag{27}$$

Appendix B. Posterior NNGP

Given the Bayesian framework presented in eqs. (8) to (10), in order to get the posterior distribution of $G(\boldsymbol{x}_\mathcal{D}) \mid \mathcal{D}$ we write explicitly all the densities and apply the Bayes rule. In particular, assuming $\boldsymbol{K}'(\boldsymbol{x}_\mathcal{D}) \in \mathrm{Sym}_+^k$ invertible (we already know that is symmetric positive semi-definite), flattening all the random matrices by columns and defining

$$\boldsymbol{y}_\mathrm{f} := \mathrm{flat}\left(\boldsymbol{y}_\mathcal{D}\right), \quad \boldsymbol{z}_\mathrm{f} := \mathrm{flat}\left(\boldsymbol{z}\right),$$

with $\boldsymbol{z} := G(\boldsymbol{x}_\mathcal{D}) \in \mathbb{R}^{n_L \times k}$, we get

$$\begin{aligned}
p_{G(\boldsymbol{x}_\mathcal{D})|\sigma^2}(\boldsymbol{z}) &= \frac{1}{\left((2\pi\sigma^2)^{n_L k} \det\left(\boldsymbol{K}'(\boldsymbol{x}_\mathcal{D}) \otimes_K \boldsymbol{I}_{n_L}\right)\right)^{1/2}} \cdot \\
&\quad \cdot \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{z}_\mathrm{f}^T \left(\boldsymbol{K}'(\boldsymbol{x}_\mathcal{D}) \otimes_K \boldsymbol{I}_{n_L}\right)^{-1} \boldsymbol{z}_\mathrm{f}\right), \\
p_{\sigma^2}(\sigma^2) &= \frac{b^a}{\Gamma(a)} \frac{1}{(\sigma^2)^{a+1}} \exp\left(-\frac{b}{\sigma^2}\right), \\
p_{\boldsymbol{y}_\mathcal{D}|G(\boldsymbol{x}_\mathcal{D}),\sigma^2}(\boldsymbol{y}_\mathcal{D}) &= \frac{1}{(2\pi\sigma^2)^{n_L k/2}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{y}_\mathrm{f} - \boldsymbol{z}_\mathrm{f})^T (\boldsymbol{y}_\mathrm{f} - \boldsymbol{z}_\mathrm{f})\right).
\end{aligned}$$

with $\otimes_K$ representing the Kronecker product.
By performing explicit computation we retrieve the posterior distribution of $G(\boldsymbol{x}_\mathcal{D}), \sigma^2 \mid \mathcal{D}$ as

$$\begin{aligned}
p_{G(\boldsymbol{x}_\mathcal{D}),\sigma^2|\mathcal{D}}(\boldsymbol{z}, \sigma^2) &\propto p_{\boldsymbol{y}_\mathcal{D}|G(\boldsymbol{x}_\mathcal{D}),\sigma^2}(\boldsymbol{y}_\mathcal{D})\, p_{G(\boldsymbol{x}_\mathcal{D})|\sigma^2}(\boldsymbol{z})\, p_{\sigma^2}(\sigma^2) \propto \\
&\propto \frac{1}{(\sigma^2)^{n_L k/2}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{y}_\mathrm{f} - \boldsymbol{z}_\mathrm{f})^T (\boldsymbol{y}_\mathrm{f} - \boldsymbol{z}_\mathrm{f})\right) \cdot \\
&\quad \cdot \frac{1}{(\sigma^2)^{n_L k/2} \sqrt{\det\left(\boldsymbol{K}'(\boldsymbol{x}_\mathcal{D})\right)}} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{z}_\mathrm{f}^T \left(\boldsymbol{K}'(\boldsymbol{x}_\mathcal{D}) \otimes_K \boldsymbol{I}_{n_L}\right)^{-1} \boldsymbol{z}_\mathrm{f}\right) \cdot \\
&\quad \cdot \frac{1}{(\sigma^2)^{a+1}} \exp\left(-\frac{b}{\sigma^2}\right).
\end{aligned}$$

Defining $\boldsymbol{N} := \boldsymbol{K}'(\boldsymbol{x}_\mathcal{D}) \otimes_K \boldsymbol{I}_{n_1} \in \mathrm{Sym}_+^{n_1 k}$, $\boldsymbol{M} := \boldsymbol{I}_{n_1 k} + \boldsymbol{N}^{-1} \in \mathrm{Sym}_+^{n_1 k}$, through simple manipulations of the exponent we get

$$\begin{aligned}
(\boldsymbol{y}_\mathrm{f} - \boldsymbol{z}_\mathrm{f})^T (\boldsymbol{y}_\mathrm{f} - \boldsymbol{z}_\mathrm{f}) + \boldsymbol{z}_\mathrm{f}^T \boldsymbol{N}^{-1} \boldsymbol{z}_\mathrm{f} &= \|\boldsymbol{y}_\mathrm{f}\|_2^2 + \|\boldsymbol{z}_\mathrm{f}\|_2^2 - 2\boldsymbol{y}_\mathrm{f}^T \boldsymbol{z}_\mathrm{f} + \boldsymbol{z}_\mathrm{f}^T \boldsymbol{N}^{-1} \boldsymbol{z}_\mathrm{f} = \\
&= \boldsymbol{z}_\mathrm{f}^T \left(\boldsymbol{I}_{n_L k} + \boldsymbol{N}^{-1}\right) \boldsymbol{z}_\mathrm{f} - 2\boldsymbol{y}_\mathrm{f}^T \boldsymbol{z}_\mathrm{f} + \boldsymbol{y}_\mathrm{f}^T \boldsymbol{y}_\mathrm{f} \pm \boldsymbol{y}_\mathrm{f}^T \left(\boldsymbol{I}_{n_L k} + \boldsymbol{N}^{-1}\right)^{-1} \boldsymbol{y}_\mathrm{f} = \\
&= \boldsymbol{z}_\mathrm{f}^T \boldsymbol{M} \boldsymbol{z}_\mathrm{f} - 2\boldsymbol{y}_\mathrm{f}^T \boldsymbol{z}_\mathrm{f} + \boldsymbol{y}_\mathrm{f}^T \boldsymbol{y}_\mathrm{f} \pm \boldsymbol{y}_\mathrm{f}^T \boldsymbol{M}^{-1} \boldsymbol{y}_\mathrm{f} = \\
&= \left(\boldsymbol{z}_\mathrm{f} - \boldsymbol{M}^{-1}\boldsymbol{y}_\mathrm{f}\right)^T \boldsymbol{M} \left(\boldsymbol{z}_\mathrm{f} - \boldsymbol{M}^{-1}\boldsymbol{y}_\mathrm{f}\right) + \boldsymbol{y}_\mathrm{f}^T \left(\boldsymbol{I}_{n_L k} - \boldsymbol{M}^{-1}\right) \boldsymbol{y}_\mathrm{f}.
\end{aligned}$$

Substituting and multiplying for the constant term $\sqrt{\det\left(\boldsymbol{M}^{-1}\right)}$ we obtain

$$\begin{aligned}
p_{G(\boldsymbol{x}_\mathcal{D}),\sigma^2|\mathcal{D}}(\boldsymbol{z}, \sigma^2) &\propto \frac{1}{(\sigma^2)^{n_L k/2} \sqrt{\det\left(\boldsymbol{M}^{-1}\right)}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{z}_\mathrm{f} - \boldsymbol{M}^{-1}\boldsymbol{y}_\mathrm{f})^T \boldsymbol{M} (\boldsymbol{z}_\mathrm{f} - \boldsymbol{M}^{-1}\boldsymbol{y}_\mathrm{f})\right) \cdot \\
&\quad \cdot \frac{1}{(\sigma^2)^{(a+n_L k/2)+1}} \exp\left(-\frac{1}{\sigma^2} \left(b + \frac{1}{2} \left(\boldsymbol{y}_\mathrm{f}^T \left(\boldsymbol{I}_{n_L k} - \boldsymbol{M}^{-1}\right) \boldsymbol{y}_\mathrm{f}\right)\right)\right).
\end{aligned} \tag{28}$$

From eq. (28) it is possible to identify two kernels: one associable with a Gaussian density and the other with an Inverse-Gamma density,

$$\begin{aligned}
\mathrm{flat}\left(G(\boldsymbol{x}_\mathcal{D})\right) \mid \sigma^2, \mathcal{D} &\sim \mathcal{N}\left(\boldsymbol{M}^{-1}\, \mathrm{flat}\left(\boldsymbol{y}_\mathcal{D}\right), \sigma^2 \boldsymbol{M}^{-1}\right), \\
\sigma^2 \mid \mathcal{D} &\sim \mathcal{IG}\left(a + \frac{n_L k}{2}, b + \frac{1}{2}\left(\mathrm{flat}\left(\boldsymbol{y}_\mathcal{D}\right)^T \left(\boldsymbol{I}_{n_1 k} - \boldsymbol{M}^{-1}\right) \mathrm{flat}\left(\boldsymbol{y}_\mathcal{D}\right)\right)\right).
\end{aligned}$$

This result allows us to apply the following Lemma B.1 (see Bernardo and Smith [BS09]) and state that the induced posterior distribution, $G(\boldsymbol{x}_{\mathcal{D}}) \,|\, \mathcal{D}$ is a $(n_L \times k)$-dimensional Student-t:

$$\text{flat}\left(G(\boldsymbol{x}_{\mathcal{D}})\right) \,|\, \mathcal{D} \sim t_{2a+n_L k}\left(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}}\right),$$

with

$$\begin{aligned}
\boldsymbol{M} &\coloneqq \boldsymbol{I}_{n_L k} + \left(\boldsymbol{K}'(\boldsymbol{x}) \otimes_K \boldsymbol{I}_{n_L}\right)^{-1}, \\
\boldsymbol{\mu}_{\text{post}} &\coloneqq \boldsymbol{M}^{-1}\,\text{flat}\left(\boldsymbol{y}_{\mathcal{D}}\right), \\
\boldsymbol{\Sigma}_{\text{post}} &\coloneqq \left(b + \frac{1}{2}\left(\text{flat}\left(\boldsymbol{y}_{\mathcal{D}}\right)^T\left(\boldsymbol{I}_{n_L k} - \boldsymbol{M}^{-1}\right)\text{flat}\left(\boldsymbol{y}_{\mathcal{D}}\right)\right)\right)\frac{2}{2a + n_L k}\boldsymbol{M}^{-1}.
\end{aligned}$$

Lemma B.1. *Let $k \in \mathbb{N}_{>0}$, and $(\boldsymbol{z}, \sigma^2)$ Gaussian-Inverse-Gamma (Gaussian-IG) distributed, i.e.,*

$$\begin{aligned}
&\boldsymbol{z} \,|\, \sigma^2 \sim \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Lambda}\right) \ \text{with } \boldsymbol{\mu} \in \mathbb{R}^k,\ \boldsymbol{\Lambda} \in \text{Sym}_+^k \ \text{and} \\
&\sigma^2 \sim \mathcal{IG}\left(\alpha, \beta\right) \ \text{with } \alpha, \beta > 0,
\end{aligned}$$

then $\boldsymbol{z}$ is distributed as a k-dimensional Student-t with 2α degrees of freedom, $\boldsymbol{z} \sim t_{2\alpha}\left(\boldsymbol{\mu}, \frac{\beta}{\alpha}\boldsymbol{\Lambda}\right)$.

Proof. We know, by hypothesis that $(\boldsymbol{z}, \sigma^2)$ is such that

$$\begin{aligned}
p_{(\boldsymbol{z},\sigma^2)}(\boldsymbol{z}, \sigma^2) = &\frac{1}{(2\pi)^{\frac{k}{2}}\left(\sigma^2\right)^{\frac{k}{2}}\sqrt{\det\left(\boldsymbol{\Lambda}\right)}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{z} - \boldsymbol{\mu})^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})\right)\cdot \\
&\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left(-\frac{\beta}{\sigma^2}\right).
\end{aligned}$$

Marginalizing over the variance σ^2 we get

$$\begin{aligned}
p_{\boldsymbol{z}}(\boldsymbol{z}) &= \int_0^\infty p_{(\boldsymbol{z},\sigma^2)}(\boldsymbol{z}, \sigma^2)d\sigma^2 = \\
&\propto \int_0^\infty \exp\left(-\frac{1}{2\sigma^2}\left(2\beta + (\boldsymbol{z} - \boldsymbol{\mu})^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})\right)\right)\left(\frac{1}{\sigma^2}\right)^{\alpha+\frac{k}{2}+1} d\sigma^2.
\end{aligned} \tag{29}$$

Setting $a = \alpha + \frac{k}{2}$, $b = \frac{2\beta+(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})}{2}$, $s = \sigma^2$ one can rewrite the last line of eq. (29) as

$$\begin{aligned}
p_{\boldsymbol{z}}(\boldsymbol{z}) &\propto \int_0^\infty s^{-(a+1)}\exp\left(-\frac{b}{s}\right)ds = \int_\infty^0\left(\frac{t}{b}\right)^{a+1}e^{-t}\left(-\frac{b}{t^2}\right)dt = \\
&= \int_0^\infty b^{-a}t^{a-1}e^{-t}dt = \Gamma(a)b^{-a} \propto \left(\frac{2\beta + (\boldsymbol{z} - \boldsymbol{\mu})^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})}{2}\right)^{-\left(\alpha+\frac{k}{2}\right)} \propto \\
&\propto \left(1 + \frac{1}{2\alpha}(\boldsymbol{z} - \boldsymbol{\mu})^T\left(\frac{\beta}{\alpha}\boldsymbol{\Lambda}\right)^{-1}(\boldsymbol{z} - \boldsymbol{\mu})\right)^{-\frac{2\alpha+k}{2}},
\end{aligned} \tag{30}$$

where in the first equality of eq. (30) we performed the change of variable $t = \frac{b}{s}$. In the final form of $p_{\boldsymbol{z}}(\boldsymbol{z})$ it is possible to recognize the kernel of a k-dimensional Student-t, $\boldsymbol{z} \sim t_{2\alpha}\left(\boldsymbol{\mu}, \frac{\beta}{\alpha}\boldsymbol{\Lambda}\right)$. □

Appendix C. Proof of the main result

C.1. Distance between marginal posterior of BNNs and NNGP.

Proposition C.1. *Let $\mathcal{L}$ be a Gaussian likelihood, $\mathcal{L} \sim \mathcal{N}\left(\boldsymbol{z}, \sigma_{\boldsymbol{y}_{\mathcal{D}}}^2\boldsymbol{I}_{n_L\times k}\right)$, it holds*

$$\|\mathcal{L}\|_\infty = \frac{1}{\left(2\pi\sigma_{\boldsymbol{y}_{\mathcal{D}}}^2\right)^{n_L k/2}} \quad \text{and} \quad \text{Lip}\left(\mathcal{L}\right) = \frac{e^{-1/2}}{\sqrt{\sigma_{\boldsymbol{y}_{\mathcal{D}}}^2}}\frac{1}{\left(2\pi\sigma_{\boldsymbol{y}_{\mathcal{D}}}^2\right)^{n_L k/2}}.$$

Proof. We first rewrite the map in a more compact form in terms of the Frobenius norm of $(\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z})$: $\forall\, \boldsymbol{z} \in \mathbb{R}^{n_L \times k}$,

$$\mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}}) = \frac{1}{\left(2\pi\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}\right)^{n_L k/2}} \exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}} \left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F^2\right).$$

Then for the uniform norm it is sufficient to recall that $\mathcal{L}$ is a bell-shaped map with maximum in the mean point. Therefore, it is immediate that

$$\left\|\mathcal{L}\right\|_\infty = \mathcal{L}(\boldsymbol{y}_{\mathcal{D}}; \boldsymbol{y}_{\mathcal{D}}) = \frac{1}{\left(2\pi\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}\right)^{n_L k/2}}.$$

For the identification of the Lipschitz constant it is necessary to recall that, as a consequence of the Mean Value Theorem, given a map $g : \Omega \to \mathbb{R}$ with Ω open convex subset of $\mathbb{R}^S$, S finite set, if $\sup_{\boldsymbol{z}\in\Omega} \|\partial/\partial\boldsymbol{z}\, g(\boldsymbol{z})\| \leq L$, then $\mathcal{L}$ is L-Lipschitz. Hence, our objective is to identify the value of $\sup_{\boldsymbol{z}\in\mathbb{R}^{n_L\times k}} \|\partial/\partial\boldsymbol{z}\, \mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}})\|_F$.
Let us define $c = \frac{1}{\left(2\pi\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}\right)^{n_L k/2}}$, then

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{z}}\mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}}) &= c\,\frac{\partial}{\partial\boldsymbol{z}} \exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}} \left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F^2\right) = \\
&= c\,\frac{\partial}{\partial\left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F^2} \exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}} \left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F^2\right) \frac{\partial}{\partial(\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z})} \left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F^2 \frac{\partial}{\partial\boldsymbol{z}}(\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}) = \\
&= c\left(-\frac{1}{2\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}} \exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}} \left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F^2\right)\right) 2(\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z})\left(-\boldsymbol{I}_{n_L\times k}\right) = \frac{\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}}{\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}}\mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}}).
\end{aligned}$$

To find the supremum of $h : \mathbb{R}^{n_L\times k} \to \mathbb{R}$,

$$h(\boldsymbol{z}) := \left\|\frac{\partial}{\partial\boldsymbol{z}}\mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}})\right\|_F = \frac{\left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F}{\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}}\mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}}),\ \forall\, \boldsymbol{z} \in \mathbb{R}^{n_L\times k},$$

we first have to notice that h is a positive real valued map, and that its first derivative has zeros in every $\boldsymbol{z}_0$ such that $\left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}_0\right\|_F^2 = \sigma^2_{\boldsymbol{y}_{\mathcal{D}}}$, which, as a consequence, are critical points. Indeed, for every $\boldsymbol{z} \neq \boldsymbol{y}_{\mathcal{D}}$,

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{z}}h(\boldsymbol{z}) &= \frac{1}{\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}}\left(\frac{\partial}{\partial\boldsymbol{z}}\mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}}) \left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F + \mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}})\frac{\partial}{\partial\boldsymbol{z}}\left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F\right) = \\
&= \frac{1}{\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}}\left(\frac{\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}}{\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}}\mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}}) \left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F - \mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}})\frac{\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}}{\left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F}\right) = \\
&= \frac{1}{\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}}\mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}})\frac{\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}}{\left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F}\left(\frac{\left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F^2}{\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}} - 1\right),
\end{aligned}$$

which is null if and only if $\left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F^2 / \sigma^2_{\boldsymbol{y}_{\mathcal{D}}} = 1$.
For any such $\boldsymbol{z}_0$ we would have that

$$h(\boldsymbol{z}_0) = \mathcal{L}\left(\boldsymbol{z}_0; \boldsymbol{y}_{\mathcal{D}}\right) = \frac{1}{\sqrt{\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}}}\frac{e^{-1/2}}{\left(2\pi\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}\right)^{n_L k/2}}.$$

It is also possible to define

$$l : \mathbb{R}^+ \to \mathbb{R},\ l(x) := \frac{x}{\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}}\frac{1}{\left(2\pi\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}\right)^{n_L k/2}} \exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{y}_{\mathcal{D}}}}x^2\right),\ \forall\, x \in \mathbb{R}^+,$$

which, by construction, has the following property:

$$\forall\, \boldsymbol{z} \in \mathbb{R}^{n_L\times k},\ l(\left\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{z}\right\|_F) = h(\boldsymbol{z}).$$

From the definition it is easy to see that $\sup_{x\in\mathbb{R}^+} l(x) = h(\boldsymbol{z}_0)$, just computing its first two derivatives l' and l''. Defining $d = \sigma_{\boldsymbol{y}_\mathcal{D}}^{-2}\left(2\pi\sigma_{\boldsymbol{y}_\mathcal{D}}^2\right)^{-n_L k/2}$ positive constant we have

$$\begin{aligned} l'(x) &= d\exp\left(-\frac{1}{2\sigma_{\boldsymbol{y}_\mathcal{D}}^2}x^2\right)\left(1 - x^2\frac{1}{\sigma_{\boldsymbol{y}_\mathcal{D}}^2}\right), \\ l''(x) &= -d\exp\left(-\frac{1}{2\sigma_{\boldsymbol{y}_\mathcal{D}}^2}x^2\right)\frac{1}{\sigma_{\boldsymbol{y}_\mathcal{D}}^2}x\left(\left(1 - x^2\frac{1}{\sigma_{\boldsymbol{y}_\mathcal{D}}^2}\right) + 2\right) = \\ &= -d\exp\left(-\frac{1}{2\sigma_{\boldsymbol{y}_\mathcal{D}}^2}x^2\right)\frac{1}{\sigma_{\boldsymbol{y}_\mathcal{D}}^2}x\left(3 - \frac{x^2}{\sigma_{\boldsymbol{y}_\mathcal{D}}^2}\right), \end{aligned}$$

and so

$$l'(x) = 0 \iff x^2 = \sigma_{\boldsymbol{y}_\mathcal{D}}^2,\ l''(x) \leq 0 \iff x^2 \leq 3\sigma_{\boldsymbol{y}_\mathcal{D}}^2.$$

This implies that in $x = \sqrt{\sigma_{\boldsymbol{y}_\mathcal{D}}^2}$ there is a local maximum and observing the asymptotic behavior, $\lim_{x\to+\infty} l(x) = 0$, we know that it is the unique global maximum. Finally, evaluating l in $\sqrt{\sigma_{\boldsymbol{y}_\mathcal{D}}^2}$ we get

$$l\left(\sqrt{\sigma_{\boldsymbol{y}_\mathcal{D}}^2}\right) = \frac{1}{\sqrt{\sigma_{\boldsymbol{y}_\mathcal{D}}^2}}\frac{e^{-1/2}}{\left(2\pi\sigma_{\boldsymbol{y}_\mathcal{D}}^2\right)^{n_L k/2}}.$$

Therefore, it follows that the critical points $\boldsymbol{z}_0$ are global maxima for h and so

$$\sup_{\boldsymbol{z}\in\mathbb{R}^{n_L\times k}}\left\|\frac{\partial}{\partial\boldsymbol{z}}\mathcal{L}(\boldsymbol{z};\boldsymbol{y}_\mathcal{D})\right\|_F = h(\boldsymbol{z}_0).$$

□

For the sake of completeness we report below a more detailed version of the Corollary 5.3 of Trevisan [Tre23], which we use as a starting point for the subsequent results.

Corollary C.2. *Given $f_{\boldsymbol{\theta}}$ BNN, with architecture $\boldsymbol{\alpha} = (\boldsymbol{n}, \boldsymbol{\varphi})$, $\boldsymbol{\varphi}$ collection of Lipschitz activation functions, prior distribution on $\boldsymbol{\theta}$ as in eq. (4), G Gaussian process as in eq. (6), $\boldsymbol{x}$, $\boldsymbol{y}_\mathcal{D}$ as in Section 3.2, and a Gaussian likelihood function $\mathcal{L} \sim \mathcal{N}\left(\boldsymbol{z}, \sigma_{\boldsymbol{y}_\mathcal{D}}^2\boldsymbol{I}_{n_L\times k}\right)$, exists a constant*

$$c\left(\mathcal{D}, \boldsymbol{\varphi}, \boldsymbol{\sigma}, \sigma_{\boldsymbol{y}_\mathcal{D}}^2, n_L\right) > 0,\ \textit{independent of}\ (n_l)_{l=1}^{L-1},$$

such that,

$$\mathcal{W}_1\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \mathcal{D}, G(\boldsymbol{x}) \,|\, \mathcal{D}\right) \leq c\frac{1}{\sqrt{n_{min}}},$$

for all $n_{min} := \min_{l=1,\dots,L-1} n_l$ sufficiently large.

Proof. Let $\widetilde{\mu}$ the law of the induced prior distribution of a BNN, $\widetilde{\mu} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x})$, and μ be the law of the associated NNGP, $\mu \sim G(\boldsymbol{x})$, probability measures on $\left(\mathbb{R}^{n_L\times k}, \|\cdot\|\right)$.
Let $g := \mathcal{L} : \mathbb{R}^{n_L\times k} \to \mathbb{R}$, bounded Lipschitz map. The posterior distributions $\mathbb{P}_{f_{\boldsymbol{\theta}}(\boldsymbol{x})|\mathcal{D}}$ and $\mathbb{P}_{G(\boldsymbol{x})|\mathcal{D}}$ are, by construction, respectively equal to $\widetilde{\mu}_g$ and μ_g as they are defined in Lemma A.6, therefore, by a direct application with $p = 2$, we get the following rewriting of eq. (27):

$$\mathcal{W}_1\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \mathcal{D}, G(\boldsymbol{x}) \,|\, \mathcal{D}\right) \leq \frac{1}{p^{(1)}}\left(\mathrm{Lip}\left(\mathcal{L}\right)p^{(3)} + \left(1 + \frac{p^{(4)}\,\mathrm{Lip}\left(\mathcal{L}\right)}{p^{(2)}}\right)\|\mathcal{L}\|_\infty\right)\mathcal{W}_2\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}), G(\boldsymbol{x})\right), \tag{31}$$

where $\mathrm{Lip}\left(\mathcal{L}\right), \|\mathcal{L}\|_\infty$ are positive constants depending on $k, \sigma_{\boldsymbol{y}_\mathcal{D}}^2, n_L$, as showed in Proposition C.1, and

$$\begin{gathered} p^{(1)} = \mathbb{E}_{\boldsymbol{z}\sim G(\boldsymbol{x})}\left[\mathcal{L}(\boldsymbol{z};\boldsymbol{y}_\mathcal{D})\right],\ p^{(2)} = \mathbb{E}_{\boldsymbol{z}\sim f_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\mathcal{L}(\boldsymbol{z};\boldsymbol{y}_\mathcal{D})\right], \\ p^{(3)} = \mathbb{E}_{\boldsymbol{z}\sim G(\boldsymbol{x})}\left[\|\boldsymbol{z}\|_F^2\right],\ p^{(4)} = \mathbb{E}_{\boldsymbol{z}\sim G(\boldsymbol{x})}\left[\|\boldsymbol{z}\|_F\right]. \end{gathered}$$

It is immediate that $p^{(3)}$ and $p^{(4)}$ are finite indeed they are the second and the first moment of a multivariate centered Gaussian, $G(\boldsymbol{x}) \sim \mathcal{N}\left(\mathbf{0}_{n_L\times k}, \boldsymbol{I}_{n_L}\otimes \boldsymbol{K}(\boldsymbol{x})\right)$, respectively. Therefore, as we observed in Remark A.2, it holds

$$(p^{(4)})^2 \le p^{(3)} = \operatorname{tr}\left(\boldsymbol{I}_{n_L}\otimes \boldsymbol{K}(\boldsymbol{x})\right) = n_L \operatorname{tr}\left(\boldsymbol{K}(\boldsymbol{x})\right) \le n_L k \max_{i\in[k]} \boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_i) \le$$
$$\le n_L k \left(\sigma^2_{\boldsymbol{W}^{(L)}} \max_{i\in[k]} \mathbb{E}\left[\left\|\varphi_L\left(G^{(L-1)}(\boldsymbol{x}_i)\right)\right\|_2^2\right]/n_L + \sigma^2_{\boldsymbol{b}^{(L)}}\right). \tag{32}$$

So both the terms can be bounded by expressions only dependent on $\boldsymbol{x},\boldsymbol{\varphi},\boldsymbol{\sigma}$ and n_L.
For the remaining terms we still exploit the normality of $G(\boldsymbol{x})$ (see eq. (5)). Due to the boundedness and positivity of $\mathcal{L}:\mathbb{R}^{d_{\text{out}}\times k}\to(0,\|\mathcal{L}\|_\infty)$, it is clear that $p^{(1)}\in(0,\infty)$, indeed we are integrating $\mathcal{L}$ with respect to a strictly positive probability measure: $0 < p^{(1)}(\mathcal{D},\boldsymbol{\varphi},\boldsymbol{\sigma},\sigma^2_{\boldsymbol{y}_\mathcal{D}}, n_L) = \mathbb{E}\left[\mathcal{L}(G(\boldsymbol{x});\boldsymbol{y}_\mathcal{D})\right] \le \|\mathcal{L}\|_\infty$.
Moreover, it is possible to notice that considering $\boldsymbol{v}\sim G(\boldsymbol{x})$, $\boldsymbol{w}\sim f_{\boldsymbol{\theta}}(\boldsymbol{x})$,

$$|p^{(1)}-p^{(2)}| = \left|\mathbb{E}\left[\mathcal{L}(\boldsymbol{v};\boldsymbol{y}_\mathcal{D})-\mathcal{L}(\boldsymbol{w};\boldsymbol{y}_\mathcal{D})\right]\right| \le \left\|\mathcal{L}(\boldsymbol{v};\boldsymbol{y}_\mathcal{D})-\mathcal{L}(\boldsymbol{w};\boldsymbol{y}_\mathcal{D})\right\|_{L^1} \le \operatorname{Lip}(\mathcal{L})\left\|\boldsymbol{v}-\boldsymbol{w}\right\|_{L^1},$$

and taking the infimum over the couplings $(\boldsymbol{v},\boldsymbol{w})$ we get

$$|p^{(1)}-p^{(2)}| \le \operatorname{Lip}(\mathcal{L})\,\mathcal{W}_1\left(G(\boldsymbol{x}), f_{\boldsymbol{\theta}}(\boldsymbol{x})\right) \le \operatorname{Lip}(\mathcal{L})\, c_1 \frac{1}{\sqrt{n_{min}}}, \tag{33}$$

where for the last inequality we used a compact version of the result in Theorem 2.8. Recalling that c_1 and $\operatorname{Lip}(\mathcal{L})$ depend only on $\boldsymbol{x},\boldsymbol{\varphi},\boldsymbol{\sigma},\sigma^2_{\boldsymbol{y}_\mathcal{D}}$ and n_L, eq. (33) implies that, for n_{min} sufficiently large, $p^{(2)}$ is also strictly positive.
To conclude it is sufficient to note that eq. (31), together with Theorem 2.8 and the observations made about $(p^{(i)})_{i=1}^4$ lead to

$$\mathcal{W}_1\left(f_{\boldsymbol{\theta}}(\boldsymbol{x})\,|\,\mathcal{D}, G(\boldsymbol{x})\,|\,\mathcal{D}\right) \le c_2 \mathcal{W}_2\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}), G(\boldsymbol{x})\right) \le c_2 c_3 \sqrt{n_L}\sum_{l=1}^{L}\frac{1}{\sqrt{n_k}} \le c\frac{1}{\sqrt{n_{min}}},$$

where c_2 and c_3 depends on $\mathcal{D},\boldsymbol{\varphi},\boldsymbol{\sigma},\sigma^2_{\boldsymbol{y}_\mathcal{D}}, n_L$. ☐

Lemma C.3. *Given $f_{\boldsymbol{\theta}}$, G and $\boldsymbol{x}$ as in Theorem 2.8, $p\ge 1$, and assuming $\sigma^2 := \sigma^2_{\boldsymbol{W}^{(L)}} = \sigma^2_{\boldsymbol{b}^{(L)}} = \sigma^2_{\boldsymbol{y}_\mathcal{D}}$, there exists a constant*

$$c\left(p,\boldsymbol{x},\boldsymbol{\varphi},(\boldsymbol{\sigma}_l)_{l=1}^{L-1}\right) > 0,\ \textit{independent of } (n_l)_{l=1}^{L}, \sigma^2,$$

such that

$$\mathcal{W}_p\left(f_{\boldsymbol{\theta}}(\boldsymbol{x})\,|\,\sigma^2, G(\boldsymbol{x})\,|\,\sigma^2\right) \le \sigma c\sqrt{n_L}\sum_{l=1}^{L-1}\frac{1}{\sqrt{n_l}},$$

Proof. The proof is straightforward if we observe that, by construction (see eqs. (1) and (5)) $f_{\boldsymbol{\theta}} = \sigma f_{\boldsymbol{\theta}'}$ and $G=\sigma G'$, with

$$\boldsymbol{\theta}' \text{ such that } \boldsymbol{\sigma}' = \left(\left((\sigma^2_{\boldsymbol{W}^{(l)}},\sigma^2_{\boldsymbol{b}^{(l)}})\right)_{l=1}^{L-1}, (1,1)\right), \tag{34}$$

and G' built with weights and bias variances as in $\boldsymbol{\sigma}'$.
Indeed, applying eq. (24) and Theorem 2.8 we get

$$\mathcal{W}_p\left(f_{\boldsymbol{\theta}}(\boldsymbol{x})\,|\,\sigma^2, G(\boldsymbol{x})\,|\,\sigma^2\right) = \sigma\mathcal{W}_p\left(f'_{\boldsymbol{\theta}}(\boldsymbol{x}), G'(\boldsymbol{x})\right) \le \sigma c\sqrt{n_L}\sum_{l=1}^{L-1}\frac{1}{\sqrt{n_L}},$$

with c independent of σ^2. ☐

Lemma C.4. *Given $f_{\boldsymbol{\theta}}$, G, $\boldsymbol{x}$, $\boldsymbol{y}_\mathcal{D}$ and $\mathcal{L}$ as in Corollary C.2, assuming $\sigma^2 := \sigma^2_{\boldsymbol{W}^{(L)}} = \sigma^2_{\boldsymbol{b}^{(L)}} = \sigma^2_{\boldsymbol{y}_\mathcal{D}}$, $\boldsymbol{K}'(\boldsymbol{x})\in\operatorname{Sym}_+^k$ (rescaled NNGP kernel), and*

$$\boldsymbol{v}\sim G(\boldsymbol{x})\,|\,\sigma^2,\ \boldsymbol{w}\sim f_{\boldsymbol{\theta}}(\boldsymbol{x})\,|\,\sigma^2,$$

$\forall\, \varepsilon < 1/\left\|\boldsymbol{K}'(\boldsymbol{x})\right\|_{\mathrm{op}}$ *it holds*

$$c_1 \cdot \left(\sigma^2\right)^{-n_L k/2} \exp\left(-\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{\sigma^2}\right) \leq \mathbb{E}\left[\mathcal{L}(\boldsymbol{v};\boldsymbol{y}_{\mathcal{D}})\right] \leq c_2 \cdot \left(\sigma^2\right)^{-n_L k/2} \exp\left(-\frac{1}{2\sigma^2}\frac{\varepsilon}{\varepsilon+1}\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right),$$

$$c_3 \exp\left(-\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{\sigma^2}\right) \leq \mathbb{E}\left[\mathcal{L}(\boldsymbol{w};\boldsymbol{y}_{\mathcal{D}})\right] \leq c_4 \cdot \left(\sigma^2\right)^{-n_L k/2},$$

where $n_{min} = \min_{l=1,\dots,L-1} n_l$ *is sufficiently large and the constants depend on* $\boldsymbol{x}, \boldsymbol{\varphi}$, $(\boldsymbol{\sigma}_l)_{l=1}^{L-1}, n_L$.

Proof. We separately discuss the four bounds.

Bounds related to $G(\boldsymbol{x})$. It is possible to rewrite $\mathbb{E}\left[\mathcal{L}(\boldsymbol{v};\boldsymbol{y}_{\mathcal{D}})\right]$ as follows:

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\boldsymbol{v};\boldsymbol{y}_{\mathcal{D}})\right] &\propto \mathbb{E}\left[\left(\sigma^2\right)^{-n_L k/2}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}_{\mathcal{D}}-\boldsymbol{v}\|_F^2\right)\right] = \\ &= \left(\sigma^2\right)^{-n_L k/2}\int_{\mathbb{R}^{n_L\times k}}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}_{\mathcal{D}}-\boldsymbol{v}\|_F^2\right)d\mathbb{P}_{G(\boldsymbol{x})|\sigma^2}(\boldsymbol{v}). \end{aligned} \tag{35}$$

Starting with the density of the Gaussian variable $G(\boldsymbol{x})$,

$$\begin{aligned} p_{G(\boldsymbol{x})}(\boldsymbol{z}) = &\frac{1}{\left((2\pi)^{n_L k}\det\left(\boldsymbol{K}(\boldsymbol{x})\otimes_K \boldsymbol{I}_{n_L}\right)\right)^{1/2}}\cdot \\ &\cdot\exp\left(-\frac{1}{2}\operatorname{flat}(\boldsymbol{z})^T\left(\boldsymbol{K}(\boldsymbol{x})\otimes_K\boldsymbol{I}_{n_L}\right)^{-1}\operatorname{flat}(\boldsymbol{z})\right),\end{aligned}$$

and recalling that, $\forall\, \boldsymbol{A}\in\mathbb{R}^{n\times n}, \boldsymbol{B}\in\mathbb{R}^{k\times k}$, then $(\boldsymbol{A}\otimes_K\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1}\otimes_K\boldsymbol{B}^{-1}$, $\det(\boldsymbol{A}\otimes_K\boldsymbol{B}) = = \det(\boldsymbol{A})^k\det(\boldsymbol{B})^n$, and $\boldsymbol{K}(\boldsymbol{x}) = \sigma^2\boldsymbol{K}'(\boldsymbol{x})$, with $\boldsymbol{K}'(\boldsymbol{x})$ as in eq. (10), it is easy to see that

$$p_{G(\boldsymbol{x})|\sigma^2}(\boldsymbol{v}) = \frac{1}{(2\pi\sigma^2)^{n_L k/2}}\frac{1}{\det\left(\boldsymbol{K}'(\boldsymbol{x})\right)^{n_L/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i,j\in[k]\times[k]}\left(\boldsymbol{K}'(\boldsymbol{x})^{-1}\right)_{i,j}\boldsymbol{v}_i^T\boldsymbol{v}_j\right). \tag{36}$$

Lower bound. Substituting eq. (36) in eq. (35) we get

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\boldsymbol{v};\boldsymbol{y}_{\mathcal{D}})\right] \propto &\left(\sigma^2\right)^{-n_L k}\int_{\mathbb{R}^{n_L\times k}}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}_{\mathcal{D}}-\boldsymbol{v}\|_F^2\right)\cdot \\ &\cdot\exp\left(-\frac{1}{2\sigma^2}\sum_{i,j\in[k]\times[k]}\left(\boldsymbol{K}'(\boldsymbol{x})^{-1}\right)_{i,j}\boldsymbol{v}_i^T\boldsymbol{v}_j\right)d\boldsymbol{v} \geq \end{aligned} \tag{37}$$

$$\begin{aligned}&\geq \left(\sigma^2\right)^{-n_L k}\exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right)\left(\sigma^2\right)^{n_L k/2}\cdot \\ &\cdot\int_{\mathbb{R}^{n_L\times k}}\exp\left(-\frac{1}{2}\left(\sum_{i,j\in[k]\times[k]}\left(\boldsymbol{K}'(\boldsymbol{x})^{-1}+2\cdot\boldsymbol{I}_k\right)_{i,j}\boldsymbol{u}_i^T\boldsymbol{u}_j\right)\right)d\boldsymbol{u},\end{aligned} \tag{38}$$

where the inequality follows applying $\|\boldsymbol{y}_{\mathcal{D}}-\boldsymbol{v}\|_F^2 \leq 2(\|\boldsymbol{y}_{\mathcal{D}}\|_F^2+\|\boldsymbol{v}\|_F^2)$ (see eq. (22)) and performing the change of variable $\boldsymbol{u} := \boldsymbol{v}/(\sigma^2)^{1/2}$. The conclusion for the lower bound follows easily observing that $\boldsymbol{K}'(\boldsymbol{x})$ is positive definite, so it holds

$$\left(\boldsymbol{K}'(\boldsymbol{x})^{-1}+2\cdot\boldsymbol{I}_k\right)\otimes_K\boldsymbol{I}_{n_L}\in\mathrm{Sym}_+^{n_L k},$$

and therefore given $\boldsymbol{u}_{\mathrm{f}} = \operatorname{flat}(\boldsymbol{u})$ the integral in eq. (38) is equal to a constant depending on $\boldsymbol{x}, \boldsymbol{\varphi}, (\boldsymbol{\sigma}_l)_{l=1}^{L-1}, n_L$:

$$\begin{aligned}&\int_{\mathbb{R}^{n_L\times k}}\exp\left(-\frac{1}{2}(\boldsymbol{u}_{\mathrm{f}}^T\left(\left(\boldsymbol{K}'(\boldsymbol{x})^{-1}+2\cdot\boldsymbol{I}_k\right)\otimes_K\boldsymbol{I}_{n_L}\right)\boldsymbol{u}_{\mathrm{f}})\right)d\boldsymbol{u}_{\mathrm{f}} = \\ &= \left((2\pi)^k\det\left(\boldsymbol{K}'(\boldsymbol{x})^{-1}+2\cdot\boldsymbol{I}_k\right)\right)^{n_L/2}.\end{aligned} \tag{39}$$

Upper bound. For the upper bound the procedure is similar. Starting from the result in eq. (37), applying the inequality $\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{v}\|_F^2 \geq \frac{\varepsilon}{1+\varepsilon}\|\boldsymbol{y}_{\mathcal{D}}\|_F^2 - \varepsilon\|\boldsymbol{v}\|_F^2$, for a fixed $\varepsilon > 0$ (see eq. (23)) and performing again the change of variable $\boldsymbol{u} := \boldsymbol{v}/(\sigma^2)^{1/2}$ we get

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{v};\boldsymbol{y}_{\mathcal{D}})\right] \leq c \cdot \left(\sigma^2\right)^{-n_L k} \exp\left(-\frac{1}{2\sigma^2}\frac{\varepsilon}{\varepsilon+1}\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right)\left(\sigma^2\right)^{n_L k/2}\cdot$$
$$\cdot \int_{\mathbb{R}^{n_L\times k}} \exp\left(-\frac{1}{2}\sum_{i,j\in[k]\times[k]}\left(\boldsymbol{K}'(\boldsymbol{x})^{-1} - \varepsilon \boldsymbol{I}_k\right)_{i,j}\boldsymbol{v}_i^T\boldsymbol{v}_j\right)d\boldsymbol{v}.$$

Now, in order to have a convergent integral it is sufficient to impose the matrix $\boldsymbol{K}'(\boldsymbol{x})^{-1} - \varepsilon\boldsymbol{I}_k$ to be positive definite. Indeed, in that case, we could conclude as before. However, this condition is obviously verified

$$\forall\,\varepsilon < \min\left\{\lambda \,|\, \lambda \in \operatorname{Sp}\left(\boldsymbol{K}'(\boldsymbol{x})^{-1}\right)\right\} = \max\left\{\lambda \,|\, \lambda \in \operatorname{Sp}\left(\boldsymbol{K}'(\boldsymbol{x})\right)\right\}^{-1} = 1/\left\|\boldsymbol{K}'(\boldsymbol{x})\right\|_{\text{op}},$$

indeed for such ε one has that

$$\operatorname{Sp}\left(\boldsymbol{K}'(\boldsymbol{x})^{-1} - \varepsilon\boldsymbol{I}_k\right) = \operatorname{Sp}\left(\boldsymbol{K}'(\boldsymbol{x})^{-1}\right) - \varepsilon \subset \mathbb{R}^+.$$

Bounds related to $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. As for the Gaussian process we can write

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{w};\boldsymbol{y}_{\mathcal{D}})\right] \propto \left(\sigma^2\right)^{-n_L k/2}\int_{\mathbb{R}^{n_L\times k}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{w}\|_F^2\right) d\mathbb{P}_{f_{\boldsymbol{\theta}}(\boldsymbol{x})|\sigma^2}(\boldsymbol{w}). \tag{40}$$

We also recall the definitions of $f_{\boldsymbol{\theta}'} = (\sigma^2)^{-1/2} f_{\boldsymbol{\theta}}$ and $G' = (\sigma^2)^{-1/2}G$, respectively the rescaled BNN and NNGP, random processes independent on σ^2, as in eq. (34).

Lower bound. Applying the same inequality and the same change of variable used in the lower bound related to $G(\boldsymbol{x})$ we get

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{w};\boldsymbol{y}_{\mathcal{D}})\right] \geq c\exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right)\int_{\mathbb{R}^{n_L\times k}} \exp\left(-\|\boldsymbol{u}\|_F^2\right) d\mathbb{P}_{f'_{\boldsymbol{\theta}}(\boldsymbol{x})}(\boldsymbol{u}) =$$
$$= c\exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right)\mathbb{E}_{\boldsymbol{u}\sim f'_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[e^{-\|\boldsymbol{u}\|_F^2}\right]. \tag{41}$$

Now, we can exploit the fact that we know how to integrate $e^{-\|\cdot\|_F^2}$ with respect to the measure $\mathbb{P}_{G'(\boldsymbol{x})}$ (we already computed this integral up to a constant depending on the usual parameters, in eq. (39)) to obtain analogous results for the mean value in eq. (41). As in eq. (33), exploiting Theorem 2.8, and observing that

$$\operatorname{Lip}\left(e^{-\|\cdot\|_F^2}\right) = \max_{\boldsymbol{u}\in\mathbb{R}^{n_L\times k}}\left\|\frac{\partial}{\partial\boldsymbol{u}}e^{-\|\boldsymbol{u}\|_F^2}\right\|_F = \max_{\boldsymbol{u}\in\mathbb{R}^{n_L\times k}}\left\|-2\boldsymbol{u}e^{-\|\boldsymbol{u}\|_F^2}\right\|_F = \sqrt{\frac{2}{e}},$$

it is possible to write the following upper bound to the difference of the mean values of $e^{-\|\cdot\|_F^2}$ with respect to the laws of $G'(\boldsymbol{x})$ and $f_{\boldsymbol{\theta}'}(\boldsymbol{x})$:

$$\left|\mathbb{E}_{\boldsymbol{u}\sim G'(\boldsymbol{x})}\left[e^{-\|\boldsymbol{u}\|_F^2}\right] - \mathbb{E}_{\boldsymbol{u}\sim f'_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[e^{-\|\boldsymbol{u}\|_F^2}\right]\right| \leq \operatorname{Lip}\left(e^{-\|\cdot\|_F^2}\right)\mathcal{W}_1\left(G'(\boldsymbol{x}), f_{\boldsymbol{\theta}'}(\boldsymbol{x})\right) \leq \frac{c}{\sqrt{n_{min}}}. \tag{42}$$

Therefore, assuming

$$\sqrt{n_{min}} \geq 2c/\mathbb{E}_{\boldsymbol{u}\sim G'(\boldsymbol{x})}\left[e^{-\|\boldsymbol{u}\|_F^2}\right],$$

we get

$$\mathbb{E}_{\boldsymbol{u}\sim f'_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[e^{-\|\boldsymbol{u}\|_F^2}\right] \geq \mathbb{E}_{\boldsymbol{u}\sim G'(\boldsymbol{x})}\left[e^{-\|\boldsymbol{u}\|_F^2}\right]/2,$$

and therefore

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{w};\boldsymbol{y}_{\mathcal{D}})\right] \geq c\exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right).$$

Upper bound. The upper bound can easily be obtained observing that the negative exponential in eq. (40) is smaller than 1 and therefore its integral is smaller than 1 as well[4]. □

It is now possible to state and prove the following Corollary C.5, which is a version of Corollary C.2 in which the dependence on σ^2 is explicit, under the assumptions of BNN built using the hierarchical model defined in eq. (9).

Corollary C.5. *Given $f_{\boldsymbol{\theta}}$, G, $\boldsymbol{x}$, $\boldsymbol{y}_{\mathcal{D}}$ and $\mathcal{L}$ as in Corollary C.2, and assuming $\sigma^2 := \sigma^2_{\boldsymbol{W}^{(L)}} = = \sigma^2_{\boldsymbol{b}^{(L)}} = \sigma^2_{\boldsymbol{y}_{\mathcal{D}}}$ exist some constants*

$$c_i\left(\mathcal{D}, \boldsymbol{\varphi}, (\boldsymbol{\sigma}_l)_{l=1}^{L-1}, n_L\right) > 0,\ i = 0, \dots, 4,\ \textit{independent of}\ (n_l)_{l=1}^{L-1}, \sigma^2,$$

such that,

$$\mathcal{W}_1\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, (\mathcal{D}, \sigma^2), G(\boldsymbol{x}) \,|\, (\mathcal{D}, \sigma^2)\right) \le h(\sigma^2) \frac{c_0}{\sqrt{n_{min}}},$$

with

$$h(\sigma^2) = c_1 \left(\sigma^2\right)^{1/2} \exp\left(\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{\sigma^2}\right)\left(c_2 + c_3\left(\sigma^2\right)^{1/2} + c_4\left(\sigma^2\right)^{-n_L k/2} \exp\left(\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{\sigma^2}\right)\right).$$

Proof. In analogy with the proof of Corollary C.2, the idea is to apply Lemma A.6 with

$$\widetilde{\mu} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \sigma^2,\ \mu \sim G(\boldsymbol{x}) \,|\, \sigma^2,\ g := \mathcal{L}(\boldsymbol{z}; \boldsymbol{y}_{\mathcal{D}}, \sigma^2),$$

so that

$$\widetilde{\mu}_g = \frac{g}{\widetilde{\mu}(g)} \widetilde{\mu} = \mathbb{P}_{f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, (\sigma^2, \mathcal{D})} \quad \text{and} \quad \mu_g = \frac{g}{\mu(g)} \mu = \mathbb{P}_{G(\boldsymbol{x}) \,|\, (\sigma^2, \mathcal{D})}.$$

By a direct application of eq. (27) we get

$$\begin{aligned} &\mathcal{W}_1\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, (\sigma^2, \mathcal{D}), G(\boldsymbol{x}) \,|\, (\sigma^2, \mathcal{D})\right) \le \\ &\le \frac{1}{p^{(1)}}\left(\operatorname{Lip}(\mathcal{L})\, p^{(3)} + \left(1 + \frac{p^{(4)} \operatorname{Lip}(\mathcal{L})}{p^{(2)}}\right)\|\mathcal{L}\|_\infty\right) \mathcal{W}_2\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \sigma^2, G(\boldsymbol{x}) \,|\, \sigma^2\right), \end{aligned} \tag{43}$$

where $\|\mathcal{L}\|_\infty$, $\operatorname{Lip}(\mathcal{L})$ are reported explicitly in Proposition C.1, whereas $1/p^{(1)}$, $1/p^{(2)}$ and $p^{(3)}$, $p^{(4)}$ are upper bounded respectively in Lemma C.4 and eq. (32): all the constants depends on $\boldsymbol{x}, \boldsymbol{\varphi}, (\boldsymbol{\sigma}_l)_{l=1}^{L-1}$ and n_L

$$\begin{gathered} \|\mathcal{L}\|_\infty = c\left(\sigma^2\right)^{-n_L k/2},\ \operatorname{Lip}(\mathcal{L}) = c\left(\sigma^2\right)^{-n_L k/2 - 1/2}, \\ p^{(1)} \ge c\left(\sigma^2\right)^{-n_L k/2} \exp\left(-\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{\sigma^2}\right),\ p^{(2)} \ge c \exp\left(-\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{\sigma^2}\right), \\ p^{(3)} \le c\,\sigma^2,\ p^{(4)} \le c\left(\sigma^2\right)^{1/2}. \end{gathered} \tag{44}$$

Hence, substituting the results in eq. (44) inside eq. (43) and applying Lemma C.3 we obtain

$$\mathcal{W}_1\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, (\sigma^2, \mathcal{D}), G(\boldsymbol{x}) \,|\, (\sigma^2, \mathcal{D})\right) \le h(\sigma^2) \frac{c}{\sqrt{n_{min}}},\ \text{with}$$

$$\begin{aligned} h(\sigma^2) &= \left(\sigma^2\right)^{1/2}\left(\sigma^2\right)^{n_L k/2} \exp\left(\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{\sigma^2}\right)\Bigg[c_1\left(\sigma^2\right)^{-n_L k/2 + 1/2} + \\ &\quad + \left(1 + c_2\left(\sigma^2\right)^{-n_L k/2} \exp\left(\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{\sigma^2}\right)\right) c_3\left(\sigma^2\right)^{-n_L k/2}\Bigg] \le \\ &\le c_1\left(\sigma^2\right)^{1/2} \exp\left(\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{\sigma^2}\right)\left(c_2 + c_3\left(\sigma^2\right)^{1/2} + c_4\left(\sigma^2\right)^{-n_L k/2} \exp\left(\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{\sigma^2}\right)\right). \end{aligned}$$

□

[4]To find a sharper upper bound reproducing the result just showed for the lower bound is not trivial. We cannot apply an analogue of eq. (42) because the map $e^{\frac{\varepsilon}{2}\|\cdot\|_F^2}$ is not Lipschitz.

C.2. **1^{st} term bound.** As we already mentioned, in order to control the first term we need first to apply the convexity property presented in eq. (16) to μ_{post} and $\bar{\mu}$ defined respectively in eqs. (17) and (19). Doing that we would have

$$\mathcal{W}_1\left(\mu_{\text{post}}, \bar{\mu}\right) \leq \int_{\mathbb{R}^+} \mathcal{W}_1\left(\mu_{\sigma^2}(s), \widetilde{\mu}_{\sigma^2}(s)\right) \frac{I_{\sigma^2}(s)}{I} p_{\sigma^2}(s) ds \tag{45}$$

where the probability measure ν in eq. (16) here is simply

$$\nu \ll \lambda^+, \text{ with } \frac{d\nu}{d\lambda^+}(s) = \frac{I_{\sigma^2}(s)}{I} p_{\sigma^2}(s), \ \lambda^+ \text{ Lebesgue measure on } \mathbb{R}^+.$$

Hence, in order to get eq. (45) we just need to prove that both $(\mu_{\sigma^2}(s))_{s\in\mathbb{R}^+}$ and $(\widetilde{\mu}_{\sigma^2}(s))_{s\in\mathbb{R}^S}$ are Markov kernels with source $\mathbb{R}^+$ and target $\mathbb{R}^{n_L\times k}$.
We first observe that $\forall\, s \in \mathbb{R}^+$, both $\mu_{\sigma^2}(s)$ and $\widetilde{\mu}_{\sigma^2}(s)$ are probability measures, which follows from $\mu_{\sigma^2}(s)(\mathbb{R}^{n_L\times k}) = \widetilde{\mu}_{\sigma^2}(s)(\mathbb{R}^{n_L\times k}) = 1$ and applying dominated convergence.
It remains only to check that for any $B \in \mathcal{B}\left(\mathbb{R}^{n_L\times k}\right)$, $\mu_{\sigma^2}(\cdot)(B)$ and $\widetilde{\mu}_{\sigma^2}(\cdot)(B)$, are measurable from $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ to $([0,1], \mathcal{B}([0,1]))$, which is again easy to observe: the maps

$$s \to \int_B \mathcal{L}(\boldsymbol{z}, s)\mu(d\boldsymbol{z}), \quad s \to \int_B \mathcal{L}(\boldsymbol{z}, s)\widetilde{\mu}(d\boldsymbol{z}), \quad s \to I_{\sigma^2}(s), \quad s \to \widetilde{I}_{\sigma^2}(s)$$

are continuous because Lebesgue integrals of the map $s \to \mathcal{L}(\boldsymbol{z}, s)$ with respect to $\boldsymbol{z}$, integration variable of a probability measure, and therefore we have the thesis.

So, the final bound on the first term can be found explicitly observing that the two probability measures $\mu_{\sigma^2}(s)$ and $\widetilde{\mu}_{\sigma^2}(s)$ parametrized by s, coincide with the laws of $G(\boldsymbol{x}) \,|\, (\mathcal{D}, \sigma^2 = s)$ and $\boldsymbol{f_\theta}(\boldsymbol{x}) \,|\, (\mathcal{D}, \sigma^2 = s)$. By Corollary C.5,

$$\mathcal{W}_1\left(\mu_{\sigma^2}(s), \widetilde{\mu}_{\sigma^2}(s)\right) \leq h(s)\frac{c}{\sqrt{n_{min}}},$$

with c and h as in the statement of the result used, which implies

$$\mathcal{W}_1\left(\mu_{\text{post}}, \bar{\mu}\right) \leq \frac{c}{\sqrt{n_{min}}} \int_{\mathbb{R}^+} h(s)\frac{I_{\sigma^2}(s)}{I} p_{\sigma^2}(s) ds. \tag{46}$$

In Lemma C.4 we already computed the bounds for $I_{\sigma^2}(s)$, therefore we can also bound $I_{\sigma^2}(s)/I$: observing

$$I_{\sigma^2}(s) = \mathbb{E}_{\boldsymbol{z}\sim G(\boldsymbol{x})|\sigma^2=s}\left[\mathcal{L}(\boldsymbol{z}, s)\right] \quad \text{and} \quad I = \int_{\mathbb{R}^+} I_{\sigma^2}(s)p_{\sigma^2}(s)ds,$$

it is easy to check that, fixed $\varepsilon < (\lambda_+)^{-1}$, with $\lambda_+ := \max\{\lambda \,|\, \lambda \in \operatorname{Sp}(\boldsymbol{K}'(\boldsymbol{x}))\}$, we have

$$\begin{aligned} I_{\sigma^2}(s) &\leq c\, s^{-n_L k/2} \exp\left(-\frac{1}{s}\frac{\varepsilon}{2(\varepsilon+1)}\|\boldsymbol{y}_\mathcal{D}\|_F^2\right), \text{ and} \\ I &\geq c\int_{\mathbb{R}^+} s^{-n_L k/2} \exp\left(-\frac{\|\boldsymbol{y}_\mathcal{D}\|_F^2}{s}\right) s^{-a-1}\exp\left(-\frac{b}{s}\right) ds = \\ &= c\int_{\mathbb{R}^+} s^{n_L k/2}\exp\left(-s\,\|\boldsymbol{y}_\mathcal{D}\|_F^2\right) s^{a+1}\exp\left(-sb\right)s^{-2}ds = \\ &= c\int_{\mathbb{R}^+} s^{n_L k/2+a-1}\exp\left(-s(\|\boldsymbol{y}_\mathcal{D}\|_F^2 + b)\right) ds = c\,\Gamma\left(n_L k/2 + a\right). \end{aligned} \tag{47}$$

Hence, the bound in eq. (46) can be further simplified bounding the following integral:

$$\begin{aligned}
&\int_{\mathbb{R}^+} h(s)\frac{I_{\sigma^2}(s)}{I}p_{\sigma^2}(s)ds \leq \\
&\quad \leq c\int_{\mathbb{R}^+}\left(c_1 s^{1/2}\exp\left(\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{s}\right)\left(c_2 + c_3 s^{1/2} + c_4 s^{-n_L k/2}\exp\left(\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{s}\right)\right)\right)\cdot \\
&\qquad \cdot s^{-n_L k/2}\exp\left(-\frac{1}{s}\frac{\varepsilon}{2(\varepsilon+1)}\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right)\cdot s^{-a-1}\exp\left(-\frac{b}{s}\right)ds = \\
&\quad = c_1\int_{\mathbb{R}^+} s^{a+n_L k/2-1/2-1}\exp\left(-s\left(b+\left(\frac{\varepsilon}{2\varepsilon+2}-1\right)\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right)\right)ds + \\
&\qquad + c_2\int_{\mathbb{R}^+} s^{a+n_L k/2-1-1}\exp\left(-s\left(b+\left(\frac{\varepsilon}{2\varepsilon+2}-1\right)\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right)\right)ds + \\
&\qquad + c_3\int_{\mathbb{R}^+} s^{a+n_L k-1/2-1}\exp\left(-s\left(b+\left(\frac{\varepsilon}{2\varepsilon+2}-2\right)\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right)\right)ds = \\
&\quad = c_1\Gamma\left(a+n_L k/2-1/2\right) + c_2\Gamma\left(a+n_L k/2-1\right) + c_3\Gamma\left(a+n_L k-1/2\right),
\end{aligned}$$

under the assumptions

$$a > \frac{1}{2} \quad \text{and} \quad b > \left(1+\frac{\varepsilon+2}{2\varepsilon+2}\right)\|\boldsymbol{y}_{\mathcal{D}}\|_F^2.$$

Therefore, it holds

$$\mathcal{W}_1\left(\mu_{\text{post}}, \bar{\mu}\right) \leq \frac{c}{\sqrt{n_{min}}}.$$

C.3. **2^nd term bound.** Finally, to control the second term the idea is to apply the Theorem 6.15 of Villani [Vil08], which in our case is simplified to Lemma 3.6. We use the following characterization of the total variation measure (see Section 6.1 of Rudin [Rud87]): $\forall\, A \in \mathcal{B}\left(\mathbb{R}^S\right)$,

$$|\mu-\nu|(A) = \sup_{(A_i)_{i=1}^{\infty},\ \bigsqcup_{i=1}^{\infty} A_i = A}\ \sum_{i=1}^{\infty}|(\mu-\nu)(A_i)|.$$

Now the problem is that we do not know how to measure maps with the finite measure $|\bar{\mu}-\widetilde{\mu}_{\text{post}}|$, but we know how to bound it.
Indeed, introducing the following notation to improve the readability,

$$k(s) := \frac{I_{\sigma^2}(s)}{I} - \frac{\widetilde{I}_{\sigma^2}(s)}{\widetilde{I}},$$

we have

$$\bar{\mu} - \widetilde{\mu}_{\text{post}} = \int_{\mathbb{R}^+} k(s)\widetilde{\mu}_{\sigma^2}(s)p_{\sigma^2}(s)ds,$$

and therefore $\forall\, A \in \mathcal{B}\left(\mathbb{R}^{n_L\times k}\right)$,

$$\begin{aligned}
|\bar{\mu}-\widetilde{\mu}_{\text{post}}|(A) &= \sup_{(A_i)_{i=1}^{\infty},\ \bigsqcup_{i=1}^{\infty} A_i = A}\ \sum_{i=1}^{\infty}\left|\int_{\mathbb{R}^+} k(s)\widetilde{\mu}_{\sigma^2}(s)(A_i)p_{\sigma^2}(s)ds\right| \leq \\
&\leq \sup_{(A_i)_{i=1}^{\infty},\ \bigsqcup_{i=1}^{\infty} A_i = A}\ \sum_{i=1}^{\infty}\int_{\mathbb{R}^+}|k(s)|\widetilde{\mu}_{\sigma^2}(s)(A_i)p_{\sigma^2}(s)ds = \\
&= \sup_{(A_i)_{i=1}^{\infty},\ \bigsqcup_{i=1}^{\infty} A_i = A}\ \lim_{j\to\infty}\int_{\mathbb{R}^+}|k(s)|\widetilde{\mu}_{\sigma^2}(s)\left(\bigsqcup_{i=1}^{j} A_i\right)p_{\sigma^2}(s)ds.
\end{aligned}$$

It is easy to observe that it is possible to apply Dominated Convergence Theorem: $\forall\, j \in \mathbb{N}_{>0}$,

$$\left||k(s)|\widetilde{\mu}_{\sigma^2}(s)\left(\bigsqcup_{i=1}^{j} A_i\right)p_{\sigma^2}(s)\right| \leq |k(s)|p_{\sigma^2}(s) \quad \text{and} \quad \int_{\mathbb{R}^+}|k(s)|p_{\sigma^2}(s) \leq 2.$$

Therefore,

$$|\bar{\mu} - \widetilde{\mu}_{\text{post}}|(A) \leq \sup_{(A_i)_{i=1}^{\infty},\, \bigsqcup_{i=1}^{\infty} A_i = A} \int_{\mathbb{R}^+} |k(s)| \widetilde{\mu}_{\sigma^2}(s) \left(\bigsqcup_{i=1}^{\infty} A_i \right) p_{\sigma^2}(s) ds =$$
$$= \int_{\mathbb{R}^+} |k(s)| \widetilde{\mu}_{\sigma^2}(s)(A) p_{\sigma^2}(s) ds =: \nu(A).$$

It is easy to observe that $\nu : \mathcal{B}\left(\mathbb{R}^{n_L \times k}\right) \to \mathbb{R}^+$ is a finite measure, indeed:

- $\nu(\emptyset) = 0$, $\nu(\mathbb{R}^{n_L \times k}) = \int_{\mathbb{R}^+} |k(s)| p_{\sigma^2}(s) ds \leq 2$;
- given $A = \bigsqcup_{i=1}^{\infty} A_i$, again using dominated convergence we have

$$\sum_{i=1}^{\infty} \nu(A_i) = \sum_{i=1}^{\infty} \int_{\mathbb{R}^+} |k(s)| \widetilde{\mu}_{\sigma^2}(A_i) p_{\sigma^2}(s) ds = \int_{\mathbb{R}^+} |k(s)| \widetilde{\mu}_{\sigma^2} \left(\bigsqcup_{i=1}^{\infty} A_i \right) p_{\sigma^2}(s) ds =$$
$$= \int_{\mathbb{R}^+} |k(s)| \widetilde{\mu}_{\sigma^2}(A) p_{\sigma^2}(s) ds = \nu(A).$$

Applying Lemma 3.6 and observing that if $\forall\, A \in \mathcal{B}\left(\mathbb{R}^{n_L \times k}\right)$, $|\bar{\mu} - \widetilde{\mu}_{\text{post}}|(A) \leq \nu(A)$ then $\forall\, f$ measurable from $\mathbb{R}^{n_L \times k}$ to $\mathbb{R}^+$, $|\bar{\mu} - \widetilde{\mu}_{\text{post}}|(f) \leq \nu(f)$, we get

$$\mathcal{W}_1\left(\bar{\mu}, \widetilde{\mu}_{\text{post}}\right) \leq \int_{\mathbb{R}^{n_L \times k}} \|\boldsymbol{z}\|_F \, d|\bar{\mu} - \widetilde{\mu}_{\text{post}}|(\boldsymbol{z}) \leq \int_{\mathbb{R}^{n_L \times k}} \|\boldsymbol{z}\|_F \, d\nu(\boldsymbol{z}) =$$
$$= \int_{\mathbb{R}^+} |k(s)| p_{\sigma^2}(s) \int_{\mathbb{R}^{n_L \times k}} \|\boldsymbol{z}\|_F \, \widetilde{\mu}_{\sigma^2}(s)(d\boldsymbol{z}) ds =$$
$$= \int_{\mathbb{R}^+} |k(s)| p_{\sigma^2}(s) \frac{1}{\widetilde{I}_{\sigma^2}(s)} \int_{\mathbb{R}^{n_L \times k}} \|\boldsymbol{z}\|_F \, \mathcal{L}(\boldsymbol{z}, s) \widetilde{\mu}(d\boldsymbol{z}) ds, \tag{48}$$

where the identity from the 1[st] to the 2[nd] line follows by Fubini's Theorem.
The inner integral is easy to compute bounding the likelihood $\mathcal{L}$ in terms of the variable $\boldsymbol{z}$, already computed in Proposition C.1: $\|\mathcal{L}\|_{\infty} = c\, s^{-n_L k/2}$. Using this result we obtain

$$\int_{\mathbb{R}^{n_L \times k}} \|\boldsymbol{z}\|_F \, \mathcal{L}(\boldsymbol{z}, s) \widetilde{\mu}(d\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{z} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \sigma^2 = s} \left[\|\boldsymbol{z}\|_F \, \mathcal{L}(\boldsymbol{z}, s) \right] \leq c\, s^{-n_L k/2} \mathbb{E}_{\boldsymbol{z} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \sigma^2 = s} \left[\|\boldsymbol{z}\|_F \right].$$

Now the procedure to compute the first moment of the distribution of $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \sigma^2 = s$ is analogous to the one used in the proof of Lemma C.4 for the lower bound related to $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. First we recall that $f_{\boldsymbol{\theta}'} = s^{-1/2} f_{\boldsymbol{\theta} \,|\, \sigma^2 = s}$ and also $G' = s^{-1/2}(G \,|\, \sigma^2 = s)$. Then we apply the change of variable $\boldsymbol{u} = \boldsymbol{z}/s^{1/2}$, and we get

$$\mathbb{E}_{\boldsymbol{z} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \sigma^2 = s} \left[\|\boldsymbol{z}\|_F \right] = \int_{\mathbb{R}^{n_L \times k}} \|\boldsymbol{z}\|_F \, d\mathbb{P}_{f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \sigma^2 = s}(\boldsymbol{z}) =$$
$$= \int_{\mathbb{R}^{n_L \times k}} \left\| \boldsymbol{u} s^{1/2} \right\|_F s^{n_L k/2} d\mathbb{P}_{f_{\boldsymbol{\theta}'}(\boldsymbol{x})}(\boldsymbol{u}) =$$
$$= s^{n_L k/2 + 1/2} \mathbb{E}_{\boldsymbol{u} \sim f_{\boldsymbol{\theta}'}(\boldsymbol{x})} \left[\|\boldsymbol{u}\|_F \right].$$

Now that we removed the dependence in terms of s it is possible to bound the moment of the rescaled BNN using the moment of the rescaled NNGP, which can be upper bounded as in eq. (32):

$$\mathbb{E}_{\boldsymbol{u} \sim G'(\boldsymbol{x})} \left[\|\boldsymbol{u}\|_F \right] \leq (n_L k)^{\frac{1}{2}} \cdot \left(\max_{i \in [k]} \mathbb{E} \left[\left\| \varphi_L \left(G^{(L-1)}(\boldsymbol{x}_i) \right) \right\|_2^2 \right] / n_L + 1 \right)^{\frac{1}{2}},$$

with a right-hand side that is just a constant term depending on $\boldsymbol{x}, \boldsymbol{\varphi}, (\boldsymbol{\sigma}_l)_{l=1}^{L-1}, n_L$.
In order to do so we replicate an analogue of eq. (42): thanks to the triangle inequality we know $\operatorname{Lip}\left(\|\cdot\|_F\right) = 1$, and applying Theorem 2.8 we get

$$\left| \mathbb{E}_{\boldsymbol{u} \sim G'(\boldsymbol{x})} \left[\|\boldsymbol{u}\|_F \right] - \mathbb{E}_{\boldsymbol{u} \sim f'_{\boldsymbol{\theta}}(\boldsymbol{x})} \left[\|\boldsymbol{u}\|_F \right] \right| \leq \operatorname{Lip}\left(\|\cdot\|_F\right) \mathcal{W}_1\left(G'(\boldsymbol{x}), f_{\boldsymbol{\theta}'}(\boldsymbol{x})\right) \leq \frac{c}{\sqrt{n_{min}}}.$$

Finally, assuming $\sqrt{n_{min}} \geq 2c/\mathbb{E}_{\boldsymbol{u}\sim G'(\boldsymbol{x})}\left[\|\boldsymbol{u}\|_F\right]$, we derive

$$\mathbb{E}_{\boldsymbol{u}\sim f'_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\|\boldsymbol{u}\|_F\right] \leq \frac{3}{2}\mathbb{E}_{\boldsymbol{u}\sim G'(\boldsymbol{x})}\left[\|\boldsymbol{u}\|_F\right],$$

which implies

$$\int_{\mathbb{R}^{n_L\times k}} \|\boldsymbol{z}\|_F \,\mathcal{L}(\boldsymbol{z},s)\widetilde{\mu}(d\boldsymbol{z}) \leq cs^{1/2}.$$

Restarting from the result in eq. (48) and using the lower bound related to $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ in Lemma C.4 we get

$$\begin{aligned}\mathcal{W}_1\left(\bar{\mu},\widetilde{\mu}_{\text{post}}\right) &\leq c\int_{\mathbb{R}+}|k(s)|\frac{s^{1/2}}{\widetilde{I}_{\sigma^2}(s)}p_{\sigma^2}(s)ds \leq \\ &\leq c\int_{\mathbb{R}+}|k(s)|\exp\left(\frac{\|\boldsymbol{y}_{\mathcal{D}}\|_F^2}{s}\right)s^{1/2}s^{-a-1}\exp\left(-\frac{b}{s}\right)ds.\end{aligned}$$

Now it is sufficient to show

$$|k(s)| = \left|\frac{I_{\sigma^2}(s)}{I} - \frac{\widetilde{I}_{\sigma^2}(s)}{\widetilde{I}}\right| \leq s^{-n_Lk/2}\frac{c}{\sqrt{n_{min}}} \tag{49}$$

to have the thesis. Indeed, we would have

$$\begin{aligned}\mathcal{W}_1\left(\bar{\mu},\widetilde{\mu}_{\text{post}}\right) &\leq \frac{c}{\sqrt{n_{min}}}\int_{\mathbb{R}+}s^{-n_Lk/2-a+1/2-1}\exp\left(-\frac{1}{s}\left(b-\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right)\right)ds = \\ &= \frac{c}{\sqrt{n_{min}}}\Gamma\left(n_Lk/2-1/2+a\right) = \frac{c}{\sqrt{n_{min}}},\end{aligned}$$

under the assumption $b > \|\boldsymbol{y}_{\mathcal{D}}\|_F^2$.

Let us thus prove eq. (49). We begin by observing that we already know how to bound the absolute difference $\left|I_{\sigma^2}(s)-\widetilde{I}_{\sigma^2}(s)\right|$, using the same arguments applied in the proof of Lemma C.4:

$$\begin{aligned}\left|I_{\sigma^2}(s)-\widetilde{I}_{\sigma^2}(s)\right| &\leq \text{Lip}\left(\mathcal{L}(\cdot,s)\right)\mathcal{W}_1\left(G(\boldsymbol{x}),f_{\boldsymbol{\theta}}(\boldsymbol{x})\right) \leq cs^{-n_Lk/2-1/2}\cdot s^{1/2}\frac{c}{\sqrt{n_{min}}} \leq \\ &\leq s^{-n_Lk/2}\frac{c}{\sqrt{n_{min}}}.\end{aligned} \tag{50}$$

Therefore, it remains to show that $I^{-1} \leq c$ and $\widetilde{I}^{-1} \geq c$ for some c depending only on the usual parameters. Indeed, it would yield

$$\left|\frac{I_{\sigma^2}(s)}{I} - \frac{\widetilde{I}_{\sigma^2}(s)}{\widetilde{I}}\right| \leq c\left|I_{\sigma^2}(s)-\widetilde{I}_{\sigma^2}(s)\right| \leq s^{-n_Lk/2}\frac{c}{\sqrt{n_{min}}}.$$

We already saw $I \geq c$ in eq. (47), but in order to find an upper bound for $\widetilde{I}$ we also need an upper bound for I. The procedure to get it is analogous to the one used in eq. (47): starting from Lemma C.4 we bound the negative exponential with 1, and we are done,

$$\begin{aligned}I &\leq c\int_{\mathbb{R}+}s^{-n_Lk/2}\exp\left(-\frac{1}{s}\frac{\varepsilon}{2(\varepsilon+1)}\|\boldsymbol{y}_{\mathcal{D}}\|_F^2\right)p_{\sigma^2}(s)ds \leq c\int_{\mathbb{R}+}s^{-n_Lk/2-a-1}\exp\left(-\frac{b}{s}\right)ds = \\ &= c\Gamma\left(n_Lk/2+a\right).\end{aligned}$$

To prove $\widetilde{I} \leq c$ it is sufficient to observe that

$$\begin{aligned}\left|I-\widetilde{I}\right| &= \left|\int_{\mathbb{R}+}\left(I_{\sigma^2}(s)-\widetilde{I}_{\sigma^2}(s)\right)p_{\sigma^2}(s)ds\right| \leq \int_{\mathbb{R}+}\left|I_{\sigma^2}(s)-\widetilde{I}_{\sigma^2}(s)\right|p_{\sigma^2}(s)ds \leq \\ &\leq \frac{c}{\sqrt{n_{min}}}\int_{\mathbb{R}+}s^{-n_Lk/2-a-1}\exp\left(-\frac{b}{s}\right)ds = \frac{c}{\sqrt{n_{min}}},\end{aligned}$$

where from the 1st to the 2nd line we used the inequality in eq. (50). Hence, considering n_{min} sufficiently large to have $\sqrt{n_{min}} > 2c/I$, it follows $|I-\widetilde{I}| \leq I/2$ which implies $\widetilde{I} \leq 3/2\, I \leq c$.

Appendix D. Simulations details

Starting from the hierarchical model in eq. (20), to sample from the posterior BNN using Algorithm 1, we only need to explicitly define a method for sampling from $\sigma^2 \mid \boldsymbol{\theta}, \mathcal{D}$. To achieve this, it is sufficient to retrieve the kernel of its density, which can be written explicitly. Recalling that $\boldsymbol{x}_\mathcal{D}$ is assumed independent of $\boldsymbol{\theta}$ and is also obviously independent of σ^2, it holds

$$
\begin{aligned}
p_{\sigma^2|\boldsymbol{\theta},\boldsymbol{x}_\mathcal{D},\boldsymbol{y}_\mathcal{D}}(\sigma^2) = \frac{p_{\sigma^2,\boldsymbol{\theta},\boldsymbol{x}_\mathcal{D},\boldsymbol{y}_\mathcal{D}}(\sigma^2,\boldsymbol{\theta},\boldsymbol{y}_\mathcal{D},\boldsymbol{x}_\mathcal{D})}{p_{\boldsymbol{\theta},\boldsymbol{x}_\mathcal{D},\boldsymbol{y}_\mathcal{D}}(\boldsymbol{\theta},\boldsymbol{x}_\mathcal{D},\boldsymbol{y}_\mathcal{D})} \propto p_{\sigma^2}(\sigma^2)\, p_{\boldsymbol{\theta}\,|\,\sigma^2}(\boldsymbol{\theta})\, p_{\boldsymbol{x}_\mathcal{D}}(\boldsymbol{x}_\mathcal{D})\, p_{\boldsymbol{y}_\mathcal{D}|\boldsymbol{\theta},\boldsymbol{x}_\mathcal{D},\sigma^2}(\boldsymbol{y}_\mathcal{D}) \propto \\
\propto p_{\sigma^2}(\sigma^2)\, p_{\boldsymbol{W}^{(L)}\,|\,\sigma^2}\left(\boldsymbol{W}^{(L)}\right)\, p_{\boldsymbol{b}^{(L)}\,|\,\sigma^2}\left(\boldsymbol{b}^{(L)}\right)\, p_{\boldsymbol{y}_\mathcal{D}|\boldsymbol{\theta},\boldsymbol{x}_\mathcal{D},\sigma^2}(\boldsymbol{y}_\mathcal{D}).
\end{aligned}
\tag{51}
$$

Exploiting

$$
f_{\boldsymbol{\theta}}(\boldsymbol{x}_\mathcal{D}) = \left(\sigma^2\right)^{1/2}\left(\frac{\boldsymbol{W}^{(L)}}{\left(\sigma^2\right)^{1/2}}\varphi\left(f_{\boldsymbol{\theta}}^{(L-1)}(\boldsymbol{x}_\mathcal{D})\right) + \frac{\boldsymbol{b}^{(L)}}{\left(\sigma^2\right)^{1/2}}\right) =: \left(\sigma^2\right)^{1/2} f_{\boldsymbol{\theta}'}(\boldsymbol{x}_\mathcal{D}),
$$

with $f_{\boldsymbol{\theta}'}(\boldsymbol{x}_\mathcal{D})$ independent of σ^2 (as in eq. (34)), we have also

$$
\begin{aligned}
p_{\sigma^2}(\sigma^2) &\propto \left(\sigma^2\right)^{-(a+1)} \exp\left(-\frac{b}{\sigma^2}\right), \\
p_{\boldsymbol{W}^{(L)}\,|\,\sigma^2}(\boldsymbol{W}^{(L)}) &\propto \left(\sigma^2\right)^{-n_L n_{L-1}/2} \exp\left(-\frac{n_{L-1}}{2\sigma^2}\left\|\boldsymbol{W}^{(L)}\right\|_F^2\right), \\
p_{\boldsymbol{b}^{(L)}\,|\,\sigma^2}(\boldsymbol{b}^{(L)}) &\propto \left(\sigma^2\right)^{-n_L/2} \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{b}^{(L)}\right\|_F^2\right), \\
p_{\boldsymbol{y}_\mathcal{D}|\boldsymbol{\theta},\sigma^2}(\boldsymbol{y}_\mathcal{D}) &\propto \left(\sigma^2\right)^{-n_L k/2} \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{y}_\mathcal{D} - f_{\boldsymbol{\theta}}(\boldsymbol{x}_\mathcal{D})\right\|_F^2\right) = \\
&= \left(\sigma^2\right)^{-n_L k/2} \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{y}_\mathcal{D}\right\|_F^2\right) \cdot \exp\left(-\frac{1}{2}\left\|f_{\boldsymbol{\theta}'}(\boldsymbol{x}_\mathcal{D})\right\|_F^2\right) \cdot \\
&\quad \cdot \exp\left(+\frac{1}{\left(\sigma^2\right)^{1/2}} \operatorname{flat}(\boldsymbol{y}_\mathcal{D})^T \operatorname{flat}(f_{\boldsymbol{\theta}'}(\boldsymbol{x}_\mathcal{D}))\right).
\end{aligned}
\tag{52}
$$

Hence, substituting the identities reported in eq. (52) inside eq. (51) we get

$$
p_{\sigma^2|\boldsymbol{\theta},\mathcal{D}}(\sigma^2) \propto \left(\sigma^2\right)^{-(a'+1)} \exp\left(-\frac{b'}{\sigma^2}\right) \exp\left(+\frac{c'}{\left(\sigma^2\right)^{1/2}}\right), \tag{53}
$$

with $a', b' \in \mathbb{R}^+$, $c' \in \mathbb{R}$ such that

$$
\begin{aligned}
a' &:= a + (n_{L-1} + k + 1)n_L/2, \\
b' &:= b + \frac{1}{2}\left(n_{L-1}\left\|\boldsymbol{W}^{(L)}\right\|_F^2 + \left\|\boldsymbol{b}^{(L)}\right\|_F^2 + \left\|\boldsymbol{y}_\mathcal{D}\right\|_F^2\right), \\
c' &:= \operatorname{flat}(\boldsymbol{y}_\mathcal{D})^T \operatorname{flat}(f_{\boldsymbol{\theta}'}(\boldsymbol{x}_\mathcal{D})).
\end{aligned}
$$

As mentioned in Section 4, we have used the companion library of Ding et al. [DEM22] to sample from the density in eq. (53) in the `Python` implementation of Algorithm 1. This has been possible because, given x random variable with density

$$
p_x(x) \propto x^{-(a+1)} \exp\left(-\frac{b}{x}\right) \exp\left(\frac{c}{x^{1/2}}\right) \mathbb{1}_{\{x \geq 0\}},
$$

the normalization constant is

$$\int_0^{+\infty} p_x(x) = \int_0^{+\infty} x^{-(a+1)} \exp\left(-\frac{b}{x}\right) \exp\left(\frac{c}{x^{1/2}}\right) dx =$$
$$= 2\int_0^{+\infty} y^{2a+1} \exp\left(-by^2\right) \exp\left(cy\right) dy =$$
$$= 2(2b)^{-a}\Gamma(2a) \exp\left(\frac{c^2}{8b}\right) D_{-2a}\left(\frac{-c}{(2b)^{1/2}}\right),$$

with D being the parabolic cylinder function. Therefore, given $z \sim \mathcal{GIN}^+(2a+1, c/2b, \sqrt{1/2b})$ (as defined by Ding et al. [DEM22, Appendix D]), applying the transformation $y \coloneqq x^2$, we get a random variable with the same distribution as our target x: $x \sim y$.

Remark D.1. The simulations closely follow the theoretical framework developed in this work. However, during the sampling of the posterior Student-t process and BNNs, it is performed a rescaling of σ^2. This adjustment is applied where σ^2 is used as the variance of $\boldsymbol{y}_{\mathcal{D}} \mid \boldsymbol{\theta}, \boldsymbol{x}_{\mathcal{D}}, \sigma^2$, in order to address numerical stability issues encountered during the sampling process described in Algorithm 1.