On Uncertainty Quantification for Near-Bayes Optimal Algorithms

Ziyu Wang
University of Oxford
wzy196@gmail.com

Chris Holmes
University of Oxford
cholmes@stats.ox.ac.uk

Abstract

Bayesian modelling allows for the quantification of predictive uncertainty which is crucial in safety-critical applications. Yet for many machine learning (ML) algorithms, it is difficult to construct or implement their Bayesian counterpart. In this work we present a promising approach to address this challenge, based on the hypothesis that commonly used ML algorithms are efficient across a wide variety of tasks and may thus be *near Bayes-optimal* w.r.t. an unknown task distribution. We prove that it is possible to recover the Bayesian posterior defined by the task distribution, which is unknown but optimal in this setting, by building a *martingale posterior* using the algorithm. We further propose a practical uncertainty quantification method that apply to general ML algorithms. Experiments based on a variety of non-NN and NN algorithms demonstrate the efficacy of our method.

1 Introduction

Bayesian modelling represents an important approach that enables favourable predictive performance in the small-sample regime and allows for the quantification of predictive uncertainty which is vital for high-stakes applications. Yet for many machine learning (ML) algorithms it can be difficult to derive or implement their Bayesian counterpart. For example, neural network (NN)-based algorithms often rely on implicit regularisation mechanisms (Zhang et al., 2021) which are hard to replicate with explicitly constructed Bayesian priors (Razin and Cohen, 2020); approximate inference in overparameterised models can be challenging (Sun et al., 2018; Wang et al., 2018); and when algorithms are offered as a black-box service (e.g., OpenAI, 2023), adapting them to Bayesian principles becomes impossible.

How can we bring back the benefits of the Bayesian paradigm without being limited by its traditional constraints? In this work we present a promising approach towards addressing this issue, based on the following **basic postulation**: the ML algorithm of interest has competitive average-case performance on hypothetical datasets, or *tasks*, sampled from an *unknown task distribution* π, and our present task can be viewed as a random sample from the same π. Formally, suppose the algorithm $\mathcal{A}$ maps a training dataset $z_{1:n}$ to a parameter estimate $\mathcal{A}(z_{1:n})$; we require it to satisfy an inequality of the following form,

$$\mathbb{E}_{\theta_0\sim\pi}\mathbb{E}_{(z_{1:n},z_*)\sim\mathbb{P}_{\theta_0}}\ell(\mathcal{A}(z_{1:n}),z_*) \le \inf_{\mathcal{A}'}\mathbb{E}_{\theta_0\sim\pi}\mathbb{E}_{(z_{1:n},z_*)\sim p_{\theta_0}}\ell(\mathcal{A}'(z_{1:n}),z_*) + \epsilon_n. \quad (1)$$

In the above, θ_0 is a parameter that determines the data generating process p_{θ_0} in an ML dataset, or a task; $(z_{1:n}, z_*)$ denote the training and test samples; $\ell(\theta, z)$ denotes the loss function; and $\mathcal{A}'$ ranges over all possible algorithms that maps $z_{1:n}$ to an $\mathcal{A}'(z_{1:n}) \approx \theta_0$.

To understand this postulation, imagine a practitioner working on a new image classification dataset. To understand the suitability of a certain algorithm $\mathcal{A}$ (which may be a combination of an NN model, optimisation and validation protocols), it is natural for them to start by reviewing past literature that evaluated $\mathcal{A}$ on datasets deemed similar to the present one. At a high level, the past and present tasks can be viewed as independent samples from the unknown distribution π, and promising reports from past literature provide evidence that (1) holds with a small ϵ_n. The practitioner may then commit to the algorithm with the smallest ϵ_n, within their other constraints. As another type of example, condition (1) is also relevant in *multi-task learning* scenarios, where it often appears as an assumption of the algorithm (e.g., Pentina and Lampert, 2014; Riou et al., 2023). Foundation models (Bommasani et al., 2021) that are pretrained on a diverse mix of datasets can also be viewed as optimised for (1), with a distribution π designed to align with the characteristics of the downstream task of interest.

Algorithms that satisfy (1) are *near-Bayes optimal*, as knowledge of the Bayesian posterior defined by π would enable the minimisation of (1) (Ferguson, 1967). As exemplified above, there are many practical scenarios where it is more reasonable to assume knowledge of a near-optimal $\mathcal{A}$ than that of a *correctly specified* π. Yet the challenge of uncertainty quantification with such an $\mathcal{A}$ remains: as an example, for parametric models and regular choices of π maximum likelihood estimation provides a near-Bayes optimal algorithm for parameter estimation (Van der Vaart, 2000), but it does not provide any uncertainty estimate which is especially needed in the small-sample regime.

To address this issue, we build on the ideas of Fong et al. (2021) and study *martingale posteriors* (MPs), defined as the distribution of parameter estimates obtained by first using $\mathcal{A}$ to generate a synthetic dataset, and then applying $\mathcal{A}$ to the combined sample of real and synthetic data (see Sec. 2 for a review). We prove that when the algorithm defines an approximate martingale, satisfies (1) and additional technical conditions, the resulted MP will provide a good approximation for the Bayesian posterior defined by π in a Wasserstein distance. Such results allow us to draw from the benefits of the latter without requiring explicit knowledge of π. As we will discuss in detail, our results also improves the understanding of MPs, by covering a wider range of algorithms and the pre-asymptotic ($n < \dim \theta$) regime. In the latter aspect it also demonstrates advantages over traditional approaches such as bootstrap aggregation (Breiman, 1996).

As a further contribution, we present MP-inspired algorithms based on sequential minimisation of the empirical risk. Our method is related to the classical nonparametric and parametric bootstrap methods but demonstrates advantages over both approaches. It can further be extended to NN models using a modified maximum likelihood objective. We evaluate the proposed method empirically on a variety of tasks including hyperparameter learning for Gaussian process models, classification with boosting tree and stacking algorithms, and conditional density estimation with diffusion models, where it consistently outperforms standard ensemble methods such as deep ensemble (Lakshminarayanan et al., 2017) and bootstrapping.

The rest of the paper is structured as follows: in Sec. 2 we review the background on Bayesian models and martingale posteriors which motivates the present work. Sec. 3 presents our theoretical results; Sec. 4 presents the proposed method, which is evaluated in Sec. 5. We provide concluding remarks in Sec. 6. For space reasons, discussion of related work is deferred to App. C.

2 Background

Notations. We adopt the following notations throughout the paper: $(\cdot)_{m:n}$ denotes a range of subscripts (e.g., $z_{m:n} = (z_m, z_{m+1}, \ldots, z_n)$). $\lesssim, \gtrsim, \asymp$ denote (in)equality up to a multiplicative constant. $\sim$ denotes asymptotic equivalence.

Bayesian modelling. Suppose we are given i.i.d. samples $\{z_i\}_{i=1}^n$ from an unknown distribution p_{θ_0} and wish to learn a $\hat{p}_n \approx p_{\theta_0}$. In Bayesian modelling we specify a parameter space Θ, a likelihood function $p(z \mid \theta)$ and a prior π over Θ. We can then compute (or approximate) the posterior $\pi(d\theta \mid z_{1:n}) \propto \pi(d\theta) \prod_{i=1}^n p(z_i \mid \theta)$. The posterior defines the predictive distribution $\pi(z_{n+1} \in \cdot \mid z_{1:n}) = \int \pi(d\theta \mid z_{1:n}) p(z \in \cdot \mid \theta)$ that provides the learned approximation for p_{θ_0}. It also quantifies the uncertainty in the prediction process through the variation in $\pi(\cdot \mid z_{1:n})$.

When π is correctly specified, predictors derived from the posterior generally enjoy good theoretical guarantees. One way to understand their benefits is through their ability to minimise various average-case losses where data is sampled from the prior predictive distribution: for instance, the posterior predictive density minimises the log loss $\mathcal{L}_{\log}(\hat{f}_n) := \mathbb{E}_{\theta_0 \sim \pi} \mathbb{E}_{(z_{1:n}, z_{n+1}) \sim p_{\theta_0}} \log \hat{f}_n(z_{n+1}; z_{1:n})$. As the loss functional is defined w.r.t. training and test data $(z_{1:n}, z_{n+1})$ sampled from the prior predictive distribution, such statements are only relevant when the prior π is correctly specified to model the true data distribution.

All Bayesian models are correctly specified for some tasks, but they do not necessarily cover the present task at hand. In many cases, specifying models based on vague subjective beliefs or computational considerations can lead to disappointing performance. As mentioned in Sec. 1, Bayesian NN models may constitute such an example. Here we note that the specification of NN priors is non-trivial: for instance, the convenient $\mathcal{N}(0, \alpha I)$ prior can lead to undesirable consequences, despite its apparent connection to standard ℓ_2 regularisation (Fortuin et al., 2021; Tran et al., 2022). Such issues—coupled with challenges in inference—motivate the search of alternative methods for uncertainty quantification.

Martingale posteriors. We review the framework of martingale posteriors (Fong et al., 2021) which will serve as the basis of our work. In a nutshell, the idea of Fong et al. (2021) can be described as follows: suppose we have the observations $z_{1:n}$ and a suitable collection of predictive distributions $\{p_j(z_{j+1}|z_{1:j}) : j \geq n\}$, then we can sample from $\{p_j\}$ recursively to impute the missing observations $\hat{z}_{j+1} \sim p_j(\cdot \mid z_{1:n}, \hat{z}_{n+1:j}),\ n \leq j < N$, and obtain a random $\widehat{\theta}_N$ estimated on the combined sample $\{z_{1:n}, \hat{z}_{n+1:N}\}$. The randomness of $\widehat{\theta}_N$ comes from the randomness of the imputations. Thus, when $N \to \infty$ is large, it reflects our uncertainty about the true data distribution p_{θ_0}; and when θ_0 is identifiable, this is the only relevant source of uncertainty about θ_0, i.e., the *epistemic uncertainty* (Der Kiureghian and Ditlevsen, 2009). This line of reasoning suggests that the distribution of $\widehat{\theta}_N \mid z_{1:n}$ can fulfil a similar role as the Bayesian posterior in quantifying the epistemic uncertainty. The framework is justified in part through the fact that it generalises Bayesian posteriors, as we explain shortly.

To have a principled construction, we need assumptions on the predictive distributions $\{p_j\}$. In this work we are primarily interested in scenarios where $\{p_j\}$ is defined by an algorithm $\mathcal{A}$ that maps any collection of j observations $z_{1:j}$ to a (deterministic or stochastic) parameter $\widehat{\theta}_j$. Formally,

$$p_j(z_{j+1} = \cdot \mid z_{1:j}) := \mathbb{E}_{\widehat{\theta}_j \sim \mathcal{A}(z_{1:j})} p_{\widehat{\theta}_j}(\cdot). \tag{2}$$

Restricting to such scenarios, a sufficient condition for the resulted $\widehat{\theta}_N$ to have a well-defined limit is for the conditional mean $\{\mathbb{E}(\widehat{\theta}_j \mid z_{1:n}, \hat{z}_{n+1:j})\}_{j=n}^{\infty}$ to form a bounded martingale: it then follows from Doob's theorem (Doob, 1949) that $\widehat{\theta}_N$ converges a.s. to some $\widehat{\theta}_\infty$. The distribution $\widehat{\theta}_\infty \mid z_{1:n}$ is thus called a *martingale posterior* (MP).

We now demonstrate that the above construction recovers the Bayesian posterior if we use the posterior predictive distributions as $\{p_j\}$, under the assumption that the posterior mean, $\bar{\theta}^B_j$, is bounded under a certain norm $\|\cdot\|$. In such cases, the conditional mean $\mathbb{E}(\widehat{\theta}_j \mid z_{1:j}) = \bar{\theta}^B_j$, and we have $\mathbb{E}(\bar{\theta}^B_{j+1} \mid \mathcal{F}_j) = \iint \theta\, \pi(d\theta \mid z_{1:j+1}) \pi(dz_{j+1} \mid z_{1:j}) = \bar{\theta}^B_j$ for all $j \geq n$. Thus, Doob's theorem applies and $\widehat{\theta}_N \to \widehat{\theta}_\infty$ a.s. for some $\widehat{\theta}_\infty$. By de Finetti's representation theorem, we can see that $\widehat{\theta}_\infty \mid \mathcal{F}_n$ will distribute as $\pi(\cdot \mid z_{1:n})$ if θ_0 is identifiable. In other words, the MP is now equivalent to the Bayesian posterior.

Remark 2.1 (supervised learning). The above formulation can be extended to cover supervised learning tasks where $z_i = (x_i, y_i)$ and θ parameterises the distribution $p(y \mid x)$, if we sample x_{j+1} in (2) from an external distribution independent of θ (e.g., a generative model, the empirical distribution defined by $x_{1:j}$, or unlabelled data if available).

Remark 2.2 (identifiability and norm). θ_0 will not be identifiable in overparameterised ML models if we use standard choices of $\|\cdot\|$ (e.g., Euclidean norm for NN parameters). However, the framework can still apply if we can determine suitable semi-norms over Θ, or replace the notion of parameter with equivalence classes of parameters that define the same *prediction function* (Sun et al., 2018) which in turn defines the likelihood. For instance, if the prediction function is determined by a linear map of a (transformed) parameter, as in the wide NN model in Jacot et al. (2018), we can use that linear map to define a semi-norm.

Martingales for machine learning? The MP framework relieves the requirement for an explicitly and correctly specified prior, as long as the user can express their prior knowledge in the form of an algorithm $\mathcal{A}$. Nonetheless, there is still the requirement that $\mathcal{A}$ define a martingale. Past works have explored various choices of $\mathcal{A}$, including nonparametric resampling and copula-based algorithms (Fong et al., 2021) and purpose-built NN models that satisfy this requirement (Lee et al., 2022; Ghalebikesabi et al., 2023). Yet it is unclear how common ML algorithms, such as approximate empirical risk minimisation (ERM) on general NN models, can be adapted for this purpose. In this work we bridge this gap, building on the observation that online gradient descent (GD) defines a valid MP (Holmes and Walker, 2023): for

$$\widehat{\theta}_{j+1} := \widehat{\theta}_j + \eta_j \nabla_\theta \log p_{\widehat{\theta}_j}(\widehat{z}_{j+1}), \ \text{ where } \widehat{z}_{j+1} \sim p_{\widehat{\theta}_j}, \tag{3}$$

we have $\mathbb{E}(\widehat{\theta}_{j+1} \mid z_{1:j}) = \widehat{\theta}_j$. Our starting point is the observation that a natural gradient variant of (3) enjoys desirable properties and connects to sequential maximum likelihood estimation (MLE) (Sec. 3.2.1). We will show how the latter perspective allows us to derive algorithms for high-dimensional models (Sec. 3.2.2) and, from a methodological point of view, DNN models (Sec. 4).

Another unaddressed question is how general MPs can be justified theoretically, beyond the somewhat vague belief that the imputations from a suitable $\mathcal{A}$ may "approximate the missing data well". While previous works established consistency for specific MPs (Fong et al., 2021; Holmes and Walker, 2023), consistency results do not justify the uncertainty estimates from the MPs: they only guarantee the MP credible sets converges to the true parameter θ_0, but do not imply such sets include θ_0 in any finite-sample scenario. Moreover, the intuition that imputations approximate the missing data well is somewhat flawed in the small-sample regime, in which case our estimate $\mathcal{A}(z_{1:n})$ is still a poor approximation to θ_0; yet it is in this regime where uncertainty quantification is most needed. In the next section we address this question, starting from the basic postulation (1).

3 Martingale Posteriors with Near-Optimal Algorithms

In this section we present our theoretical contribution. We will state our result formally in Sec. 3.1; it can be informally summarised as follows: for algorithms that define approximate MPs, satisfy stability conditions and is "efficient" on a task distribution π in the sense of (1), the resulted MP will be close to the Bayesian posterior defined by π in a Wasserstein distance. It follows that the MP will provide useful uncertainty estimates on new tasks sampled from π, which is valuable when explicit knowledge of π is not available and thus cannot be used to construct the Bayesian posterior.

As discussed in Sec. 1, our conceptual setup covers generic ML algorithms such as approximate MLE on DNN models: they are generally considered efficient on a variety of tasks that, loosely speaking, may represent samples from π; and based on prior knowledge, a practitioner may assume the present task also falls into this category. While our theorem will not cover deep models, we illustrate in Sec. 3.2 how it justifies similar algorithms on examples that cover high-dimensional, overparameterised models and the small-sample regime. The examples will provide valuable insight to the algorithm's behaviour in more complex settings.

3.1 Setup and Main Result

Additional notations and setup. Our theoretical analysis applies to simplified scenarios that nonetheless capture interesting aspects of ML applications. We first restrict to *online algorithms*, where the MP can be defined through

$$\widehat{\theta}_{j+1} := \widehat{\mathrm{Alg}}_{j+1}(\widehat{\theta}_j, \widehat{z}_{j+1}), \ \text{ where } \widehat{z}_{j+1} \sim p_{\widehat{\theta}_j},$$

and $\{\widehat{\mathrm{Alg}}_j : \Theta \times \mathcal{Z} \mapsto \Theta\}$ is a sequence of measurable functions. This covers the GD algorithm (3), which will serve as an important example to motivate our assumptions. We also assume the existence of a (semi-)norm $\|\cdot\|$ that, informally speaking, measures relevant differences between parameters (see Rem. 2.2).

Let us define $\bar{\theta}^B_j := \mathbb{E}_{\theta \sim \pi(\theta | z_{1:n}, z^B_{n+1:j})} \theta$, $z^B_{j+1} \sim \pi(z_{j+1} | z_{1:n}, z^B_{n+1:j})$ so that $\lim_{N\to\infty} \bar{\theta}^B_N$ exists and distributes as the posterior $\pi(\theta | z_{1:n})$ assuming boundedness (see Sec. 2). Define $\widehat{\Delta}_j(\theta, z) := \widehat{\mathrm{Alg}}_j(\theta, z) - \theta$, $\Delta^B_j := \bar{\theta}^B_j - \bar{\theta}^B_{j-1}$. We use $\mathbb{E}_\pi$ to denote the expectation w.r.t. the prior predictive distribution: for all $j \in \mathbb{N}$, $g : \mathcal{Z}^{\otimes j} \to \mathbb{R}$, $\mathbb{E}_\pi g(z_{1:j}) = \mathbb{E}_{\theta_0 \sim \pi, z_{1:j} \sim p_{\theta_0}} g(z_{1:j})$. Define the expected posterior contraction rate

$$\bar{\varepsilon}^2_{B,j} := \mathbb{E}_{\theta_0 \sim \pi, z_{1:j} \overset{iid}{\sim} p_{\theta_0}} \mathbb{E}_{\theta_{p,j} \sim \pi(\cdot | z_{1:j})} \|\theta_{p,j} - \bar{\theta}^B_j\|^2,$$

where with slight abuse of notation we use $\bar{\theta}_j^B$ to refer to the posterior mean w.r.t. $z_{1:j}$. Note that in the above probability space, $\theta_{p,j}$ and θ_0 are conditionally i.i.d. given $z_{1:j}$, so $\bar{\varepsilon}_{B,j}$ also equals an expected error rate for the estimator $\bar{\theta}_j^B$, which minimises the expected error. We hence define the average "excess error" incurred by our $\widehat{\mathrm{Alg}}_j$

$$\breve{\varepsilon}_{ex,j}^2 := \mathbb{E}_{\theta_0 \sim \pi, z_{1:j} \overset{iid}{\sim} p_{\theta_0}} (\|\breve{\theta}_j - \theta_0\|^2 - \|\bar{\theta}_j^B - \theta_0\|^2),$$

where $\breve{\theta}_j := \widehat{\mathrm{Alg}}_j(\breve{\theta}_{j-1}, z_j)$ is defined by applying $\widehat{\mathrm{Alg}}_j$ to the same set of $z_{1:j}$. (Note the subtrahend equals $\bar{\varepsilon}_{B,j}^2$.)

Assumptions. We now introduce the assumptions. First we require $\widehat{\mathrm{Alg}}_j$ to define an approximate martingale:

Assumption 3.1 (approximate martingale)**.** *There exists $\delta > 0$ s.t. for all $j \geq n$ and $\theta \in \Theta$, we have*

$$\|\mathbb{E}_{\hat{z} \sim p_\theta} \widehat{\Delta}_j(\theta, \hat{z})\|^2 \leq j^{-2(1+\delta)} \bar{\varepsilon}_{B,j}^2. \tag{4}$$

Now we introduce our first assumption on stability. For the GD algorithm (3), its condition (i) merely requires $\nabla_\theta \log p_\theta(z)$ to be Lipschitz continuous w.r.t. θ and z. Condition (ii) below relates the norm $\|\cdot\|$ to a 2-Wasserstein distance.

Assumption 3.2 (stability I)**.** *There exist a norm $\|\cdot\|_z$ over $\mathcal{Z}$, $\iota > 0, L_1, L_2 > 0$ and $\eta_j \leq j^{-(1+\iota)/2}$ s.t.* ***(i)*** *for all $j \geq n$, the following inequalities hold for all $\theta, \theta' \in \Theta, z, z' \in \mathcal{Z}$:*

$$\|\widehat{\Delta}_j(\theta, z) - \widehat{\Delta}_j(\theta', z)\|^2 \leq \eta_j^2 L_1^2 \|\theta - \theta'\|^2, \quad \|\widehat{\Delta}_j(\theta, z) - \widehat{\Delta}_j(\theta, z')\|^2 \leq \eta_j^2 L_2^2 \|z - z'\|_z^2.$$

(ii) *one of the following holds:*

$$W_{2,z}^2(p_\theta, p_{\theta'}) \leq C_\Theta \|\theta - \theta'\|^2, \tag{5}$$

$$\text{or} \quad W_{2,z}^2(p_\theta, p_{\theta'}) \leq C_\Theta \|\theta - \theta'\|, \;\; \eta_j \leq j^{-(3+\iota)/4}, \tag{5'}$$

where $W_{2,z}$ denotes the 2-Wasserstein distance under $\|\cdot\|_z$.

The following assumption characterises efficiency. Its first inequality merely requires the excess error to have a higher order. All examples in Sec. 3.2 will satisfy $\eta_j \asymp j^{-1}$, in which case its second inequality is also satisfied if $\breve{\varepsilon}_{ex,j}^2 \lesssim j^{-s} \bar{\varepsilon}_{B,j}^2$ for an arbitrarily small $s > 0$.

Assumption 3.3 (efficiency)**.** *There exist $s \in (0, \min\{\delta, \iota\})$ and a sequence $\{\nu_l\} \to 0$ s.t. for all $l \geq n$, we have*

$$\breve{\varepsilon}_{ex,n}^2 = o_n(\bar{\varepsilon}_{B,n}^2), \quad \sum_{j=l}^{\infty} j^{1+s} \eta_j^2 \breve{\varepsilon}_{ex,j}^2 \leq \nu_l \bar{\varepsilon}_{B,l}^2.$$

The following is a further condition on stability. It has appeared in previous work studying the convergence of similar GD algorithms (Moulines and Bach, 2011, H6).

Assumption 3.4 (stability II)**.** *There exist $C_{\mathcal{A}}, C'_{\mathcal{A}} \geq 0$, $\{H_{\theta,j} \in \mathbb{R}^{d \times d} : \theta \in \Theta, j \in \mathbb{N}\}$ s.t. for all $\theta, \theta' \in \Theta$, we have*

$$\|(\mathbb{E}_{z' \sim \mathbb{P}_{\theta'}} \widehat{\mathrm{Alg}}_{j+1}(\theta, z') - \theta) - \eta_j H_{\theta,j}(\theta' - \theta)\| \leq C_{\mathcal{A}} \|\theta' - \theta\|^2.$$

Moreover, we have $\sup_{j \geq n, \theta \in \Theta} \|H_{\theta,j}\|_{op}^2 \leq C'_{\mathcal{A}} < \infty$.

Finally, we introduce a set of conditions that are typically trivial in the large-sample regime (e.g., $\bar{\varepsilon}_{B,n}^2 \asymp d/n$, $d \leq n^{\iota - s}$). It can also hold in the pre-asymptotic regime if $C_{\mathcal{A}}$ is small, as we will see in Sec. 3.2.1.

Assumption 3.5 (miscellaneous conditions)**.**

(i) *For all $j \geq n$ we have $\breve{\varepsilon}_{ex,j} \leq 1$, $\bar{\varepsilon}_{B,j} \geq j^{-1}$.*

(ii) $\lim_{j \to \infty} \bar{\varepsilon}_{B,j} = 0$. *$\{\breve{\varepsilon}_{ex,j}\}$ is non-increasing.*

(iii) $C_{\mathcal{A}} \sum_{j \geq n} j^{1+s} \eta_j^2 \bar{\varepsilon}_{B,j}^4 \leq \nu_n \bar{\varepsilon}_{B,n}^2$.

Main result. Our main result is the following:

Theorem 3.1 (proof in App. A.1)**.** *Let $\pi_n, \hat{p}_{mp,n}$ denote the Bayesian posterior and the (approximate) MP defined by $z_{1:n}$, and $W_{2,\theta}$ be the 2-Wasserstein distance w.r.t. $\|\cdot\|$. Under Asm. 3.1-3.5, there exists some $C > 0$ determined by $(C_\Theta, C_\mathcal{A}, C'_\mathcal{A}, L_1, L_2)$ s.t. for $\chi_n = C/(sn^s) \to 0$ we have*

$$\mathbb{E}_\pi W^2_{2,\theta}(\pi_n, \hat{p}_{mp,n}) \leq e^{\chi_n}((\chi_n + \nu_n)\bar{\varepsilon}^2_{B,n} + \breve{\varepsilon}^2_{ex,n}) \tag{6}$$

$$= o_n(\bar{\varepsilon}^2_{B,n}). \tag{7}$$

Theorem 3.1 provides an average-case upper bound on the 2-Wasserstein distance between the MP and the Bayesian posterior. Such Wasserstein distance bounds justify the use of MP credible sets to approximate their Bayesian counterpart, which is desirable in our setting as π is assumed to be correct but unknown. To see how W_2 bounds link Bayesian and MP credible sets, observe that by the Chebyshev inequality, any MP-credible sets A with nominal level $1 - \gamma$ can be enlarged by $\Delta\epsilon_{n,t} := t_n^{-1/2} W_{2,\theta}(\pi_n, \hat{p}_{mp,n})$, so that the resulted set $\{\theta \in \Theta : \exists\theta' \in A \; s.t. \; \|\theta' - \theta\| \leq \Delta\epsilon_{n,t}\}$ has a Bayesian posterior mass $\geq 1 - \gamma - t_n$. The radii of Bayesian credible balls generally have the order of the posterior contraction rate; it follows from (7) that $\mathbb{E}_\pi \Delta\epsilon^2_{n,t}$ can have a higher order than the expectation of the squared radii, which is $\mathcal{O}(\bar{\varepsilon}^2_{B,j})$, and this holds even for some $t_n \to 0$. In this sense, the modification is asymptotically negligible.

3.2 Examples

3.2.1 Exponential Family Models and Sequential MLE

Let $\bar{p}_\eta(z) \propto e^{\eta^\top T(z) - A(\eta)}$ denote an exponential family model with natural parameter η, and $\theta(\eta) := \mathbb{E}_{z \sim \bar{p}_\eta} T(z)$ denote the corresponding mean parameter. Then we have $\theta = \nabla_\eta A$, and we can use $p_\theta := \bar{p}_{(\nabla A)^{-1}(\theta)}$ to denote the model distribution corresponding to θ.

Consider an algorithm that, for any set of n observations $\{z_i\}_{i=1}^n$, computes the maximum likelihood estimate $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n T(z_i)$. The algorithm can be expressed as

$$\widehat{\mathrm{Alg}}_j(\widehat{\theta}_{j-1}, z_j) := \widehat{\theta}_{j-1} + j^{-1}(T(z_j) - \widehat{\theta}_{j-1}). \tag{8}$$

Note that (8) is equivalent to the *natural gradient* algorithm with step-size j^{-1} (Amari, 2016) and thus generalises (3).

We choose π to be a conjugate prior, defined by the following density in the space of natural parameters, $\bar{\pi}(\eta) \propto e^{\eta^\top \theta_\pi - \alpha A(\eta)}$. ($\alpha > 0, \theta_\pi \in \mathbb{R}^d$ are its hyperparameters.) We consider any π s.t. $n + \alpha > 2$, $\alpha = O(1)$ is not too large and $\bar{\varepsilon}^2_{B,j} < \infty$, and assume the function T is L-Lipschitz. It then follows that

- Asm. 3.1 holds for all δ as (8) defines an exact martingale. Asm. 3.2 (i) holds for $\eta_j = (j + 1)^{-1}, L_1 = 1, L_2 = L$. Asm. 3.4 holds for $H_{\theta,j} = I, C_\mathcal{A} = 0, C'_\mathcal{A} = 1$.
- Asm. 3.3 holds with any $s \in (0, 1)$ and $\nu_l \leq 2\alpha l^{-1+s}$, because we have $\breve{\varepsilon}^2_{ex,j} \leq 2\alpha j^{-1}\bar{\varepsilon}^2_{B,j}$ for all $j \geq n$; see App. A.2.1 for its proof.
- Asm. 3.5 holds when $\bar{\varepsilon}^2_{B,n} \leq n/2\alpha$. For d-dimensional models where $\bar{\varepsilon}^2_{B,n} \lesssim d/n$, it suffices to have $n \gtrsim \sqrt{d}$.

Validation of Asm. 3.2 (ii) is more challenging due to a somewhat lack of understanding of Wasserstein distance properties for exponential family models. We first note that it can be established on a case-by-case basis by studying optimal transport plans; in this way we can verify that the Gaussian model $p_\theta = \mathcal{N}(\theta, \Sigma_0)$, and the exponential model $p_\theta = \mathrm{Exp}(\theta)$ satisfy its (5) with $C_\Theta = O(\|\Sigma_0\|^{-1}_{op})$ and $C_\Theta = 1$, respectively, and the Bernoulli model satisfies (5') with $C_\Theta = 8$. Another scenario where a similar version of (5') hold is when $\sup_{z \in \mathcal{Z}} \|T(z)\|$ is bounded and the eigenvalues of the Fisher information matrices are bounded from both sides. See App. A.2.1 for a detailed discussion.

When the assumption holds, our theorem will apply and provide a bound of $W^2_{2,\theta}(\hat{p}_{mp,n}, \pi_n) \lesssim \alpha^{1/2} n^{-1/2} \bar{\varepsilon}^2_{B,n}$. Note that this applies to the pre-asymptotic regime of $\sqrt{d} \lesssim n \lesssim d$, even though the estimation error is $\|\hat{\theta}_n - \theta_0\| \gtrsim 1$.

3.2.2 Regularised Algorithms in High Dimensions

The above example involves unregularised MLE, which is known to have poor performance in various high-dimensional problems. We now present a high-dimensional example where a regularised variant of the MP enjoys good guarantees established by our theorem. We will also discuss how the example connects to Gaussian processes (GPs).

A linear-Gaussian inverse problem. Let $(\|\cdot\|_{\mathcal{H}}, \|\cdot\|_{\mathcal{Z}})$ be two Hilbert norms for θ and z, respectively, and $A : \mathcal{H} \to \mathcal{Z}$ be a Hilbert-Schmidt operator. Suppose the observations are generated by $z_i \mid \theta \sim \mathcal{N}_{\mathcal{Z}}(A\theta, I)$ where $\mathcal{N}_{\mathcal{Z}}$ denotes the shifted iso-normal process on $\mathcal{Z}$.[1] We define our MP using preconditioned GD:

$$\widehat{\mathrm{Alg}}_j(\theta, z) := \theta + \eta_j G_j \nabla \log p(z; \theta) \tag{9}$$

where $\eta_j = j^{-1}$, $G_j = (A^\top A + j^{-1} I)^{-1}$, and compare with the Bayesian posterior determined by $\pi = \mathcal{N}_{\mathcal{H}}(0, I)$.

The above setup is closely related to the classical inverse problems defined by Gaussian white noise (Cavalier, 2008). Following a convention in that literature, we assume the singular values $s_i(A) \asymp i^{-\beta}$ for some $\beta > 1/2$, and quantify the difference between the MP and posterior through the norm $\|\theta - \theta'\| = \|(A^\top A)^{\alpha/2}(\theta - \theta')\|_{\mathcal{H}}$, where $\alpha \in \mathbb{R}$ is a problem parameter. With $\alpha = 1$, the problem can be viewed as regression in a Sobolev space. See App. A.2.2 for details.

It then follows that (see proofs in Appendix A.2.2)

- Asm. 3.1, 3.2 and 3.4 holds for the above η_j, all $\delta > 0$, and $L_1 = L_2 = C_\Theta = C'_{\mathcal{A}} = 1, C_{\mathcal{A}} = 0$. Asm. 3.5 always when, e.g., $\alpha = 1$.
- Asm. 3.3 holds for all s and $\{\nu_j\}$, because $\breve{\varepsilon}_{ex,j} = 0$.

Our theorem thus applies and gives a bound of $\mathcal{O}(\bar{\varepsilon}^2_{B,n}/n)$.

We note that $\{\widehat{\mathrm{Alg}}_j\}$ will produce the same output as the posterior mean if we apply it to π-generated data. However, the result is still non-trivial, because the model samples used to define the MP, $\{\widehat{z}_j\}_{j=n+1}^{\infty}$, is different from $\{z_j^B\}$: the latter comes from the Bayesian predictive distribution, which is defined by a mixture of parameters—the full posterior—as opposed to merely the posterior mean estimate. It is thus interesting that the ratio between the Wasserstein distance and $\bar{\varepsilon}_{B,j}$ remains bounded by a dimension-free factor.

Connections to GP regression. The above example connects to GP regression through its connection to certain nonparametric inverse problem that is asymptotically equivalent to regression (Cavalier, 2008). We can also observe that if we choose $\mathcal{H}$ to be a reproducing kernel Hilbert space defined by a kernel k_x, the above prior π will reduce to the standard GP prior defined by k_x, and the operator $A : \mathcal{H} \ni f \mapsto (f(x_1), \ldots, f(x_n))$ is Hilbert-Schmidt; hence, the derivations should apply to GP regression.

We refer readers to App. A.2.2 for a detailed discussion along the above lines, where we also note that the MP defined by (9) can be used for GP inference. However, the following algorithm provides a more practical alternative:

$$\widehat{\theta}_{j+1} := \operatorname*{arg\,min}_{\theta \in \mathcal{H}} \Big(\sum_{i=1}^{j} (f_{\widehat{\theta}_j}(x_i) - f_\theta(x_i))^2 + (f_\theta(\hat{x}_{j+1}) - \hat{y}_{j+1})^2 + \frac{1}{n} \|\theta - \widehat{\theta}_j\|^2_{\mathcal{H}} \Big), \tag{10}$$

where f_θ refers to the regression function defined by $\theta \in \mathcal{H}$, $\hat{x}_{j+1} \sim \mathrm{Uniform}(\{x_i\}_{i=1}^j)$ (see Rem. 2.1), $\hat{y}_{j+1} \sim p(\hat{y}_{j+1} \mid f(X) = \widehat{\theta}_j, \hat{x}_{j+1})$, and with a slight abuse of notation we use (x_i, y_i) to refer to the i-th (real or synthetic) observation received by the algorithm. As we discuss in App. A.2.2, the algorithm (10) is based on the same principle of iterative maximum-a-posteriori estimation as (9).

Similar to some previous works for GP inference (which we review in App. C), we can implement (10) by using random feature approximations for $\mathcal{H}$; we can also apply the resulted algorithm to

[1] See van der Vaart et al. (2008, p. 207). In particular, if $\|\cdot\|_{\mathcal{Z}}$ is a Euclidean norm, $\mathcal{N}_{\mathcal{Z}}$ will be the standard Gaussian distribution.

overparameterised random feature models which represent a simplified theoretical model for DNNs (Lee et al., 2019). We refrain from a full analysis in light of the rich literature on this topic, but refer readers to App. D.1 where we validate (10) empirically. From our perspective, (10) is particularly interesting because as we explain shortly, its first term corresponds to a *function-space Bregman divergence* of the log likelihood loss, which motivates the use of similar algorithms in broader scenarios.

4 MP-Inspired Uncertainty for General ML Algorithms

In Sec. 3.2 we have illustrated the efficacy of MP-based uncertainty quantification (UQ) based on a sequential MLE algorithm and its regularised variant. Such results suggest that similar procedures should be broadly applicable, even to models beyond the scope of our analysis. In this section we discuss the implementation of such MP-inspired procedures.

From MLE/MP to an "iterative parametric bootstrap" scheme. All MP procedures in Sec. 3.2 have the following structure: at each iteration j, *(i)* sample $\widehat{z}_{j+1} \sim p_{\widehat{\theta}_j}$; *(ii)* let $\widehat{\theta}_{j+1}$ be the (regularised) MLE on all past samples $\{z_{1:n}, \widehat{z}_{n+1:j+1}\}$. It is thus natural to generalise the algorithm as follows:

Algorithm 1 MP-inspired uncertainty quantification

```
1. Initialisation: $D_n := z_{1:n}, \widehat{\theta}_n \leftarrow \mathcal{A}_0(D_n)$
2. for $j \leftarrow n, n+1, \ldots, n + \lfloor N/\Delta n \rfloor$
    (a) Sample $\widehat{z}_{n_j:n_j+\Delta n} \sim p_{\widehat{\theta}_j}$; $D_{j+1} \leftarrow D_j \cup \widehat{z}_{n_j:n_j+\Delta n}$
    (b) $\widehat{\theta}_{j+1} \leftarrow \mathcal{A}(D_{j+1}; \widehat{\theta}_j)$
3. Repeat 1–2 for $K$ times, possibly in parallel; use the resulted $\{\widehat{\theta}^{(k)}_{n+\lfloor N/\Delta n \rfloor}\}_{k=1}^{K}$ to form an ensemble predictor
```

In the above, $(\mathcal{A}_0(D), \mathcal{A}(D; \theta))$ denote a general parameter estimation algorithm that approximately optimise a regularised empirical risk (through e.g., MLE or MAP estimation). Our analysis loosely suggests that any algorithm may be used if it is efficient in the sense of (1). To accelerate convergence, we allow optimisation to resume from the previous-iteration optima θ which can be close to the new one. Compared with the previous examples, we also modify the procedure to process $\Delta n > 1$ samples at each iteration.

The above procedure has a form similar to *parametric bootstrap* (Efron, 2012), which at each round k draws n samples $\{\hat{z}^{(k)}_{pb,j}\}_{j=1}^{n}$ from the initial $p_{\widehat{\theta}_n}$ and computes a parameter sample $\hat{\theta}^{(k)}_{pb} := \mathcal{A}(\{\hat{z}^{(k)}_{pb,j}\})$. Our procedure can be viewed as an iterative variant of the above scheme, but we also retain the original dataset $\{z_{1:n}\}$. With a choice of $\Delta n < n$ and $N > n$, we may hope to achieve better performance. This is suggested by our previous analysis which may become applicable at $\Delta n = 1, N \to \infty$, and we will also support this claim with additional examples and experiments.

A modified objective for DNNs. Many ML algorithms can be viewed as optimising a (regularised) empirical risk, in which case they can be directly plugged into Alg. 1. However, DNN-based algorithms represent a notable exception: in the estimation of DNN models *early stopping* can play a crucial role. It is thus inadvisable to apply Alg. 1 directly to DNNs using standard learning objectives such as log likelihood: if an old parameter $\widehat{\theta}_j$ does not reach the optima of its respective objective, when processing the new samples $\{\widehat{z}_{n_j:n_j+\Delta n}\}$, the loss for the old samples D_j will continue to be reduced as well, which is undesirable as it cancels the effect of early stopping. To address this issue we adopt the modified objective in Bae et al. (2022). Concretely, suppose the original objective for $\widehat{\theta}_{j+1}$ has the form of $\sum_{z \in D_{j+1}} \ell(f(z, \theta), z)$, where $f(z, \theta)$ denotes the output from the DNN model;

we adopt the following modified objective for $\widehat{\theta}_{j+1}$:

$$\sum_{z\in D_j} \ell_B(f(z,\theta), z; f(z;\widehat{\theta}_j)) + \sum_{l=n_j}^{n_j+\Delta n} \ell(f(\widehat{z}_l,\theta), \widehat{z}_l), \tag{11}$$

where $\ell_B(f, z; \bar{f}) := \ell(f, z) - \ell(\bar{f}, z) - \nabla_f \ell(\bar{f}, z)(f - \bar{f})$ defines a "function-space" Bregman divergence for the original ℓ. Hence, as long as $\ell(f(z,\theta), z)$ is convex w.r.t. the function value $f(z,\theta)$, the first term of (11) is always minimised by the old parameter $\widehat{\theta}_j$, thus retaining the regularisation effect of its optimisation non-convergence. As a concrete example, when ℓ is the squared loss for regression, the above objective will have the form of (10) (minus the regularisation term). The objective can be augmented with explicit regularisation if desired. UQ for DNNs can then be implemented by plugging the resulted estimation algorithm $\mathcal{A}$ into Alg. 1. We discuss implementation details in App. B.

Comparison to conventional bootstrap. The proposed method is broadly similar to bootstrap aggregation methods: both build an ensemble of model parameters by estimating on perturbed versions of the training set, where perturbation is implemented through either the addition or resampling of samples. We will demonstrate the superior empirical performance of our method in Sec. 5; here we present simplified examples which may provide additional insight.

Example 4.1 (comparison to nonparametric bootstrap)**.** *Suppose the training data* $z_{1:n} \sim \mathcal{N}(\theta_0, I)$ *with* $d := \dim z_i$ *satisfying* $n \ll d \ll n^2$*. Let our Alg. 1 be defined with* $\Delta n = 1, N \gg n$ *and the sequential MLE algorithm as* $\mathcal{A}$*. It follows by Sec. 3.2.1 that* $p(\hat{\theta}_N \mid z_{1:n}) = \mathcal{N}(\hat{\theta}_n, \Sigma_n)$ *where* $\Sigma_n \sim I$*. Note how this distribution quantifies a non-trivial amount of uncertainty in the* $(d-n)$*-dimensional null space of the empirical covariance* $\frac{1}{n}\sum_{i=1}^n (z_i - \bar{z}_i)(z_i - \bar{z}_i)^\top$*. In contrast, the sampling distribution of nonparametric bootstrap will have no variation in this subspace, falsely indicating complete confidence in the subspace where the data does not provide any information at all.*

Example 4.2 (comparison to parametric bootstrap)**.** *Consider a two-dimensional dataset generated as follows:* $z_{i,1} \sim \mathrm{Bern}(1-\epsilon), z_{i,2}|z_{i,1} \sim \mathcal{N}(\theta_{z_1=z_{i,1}}, 1)$*. With* $n \sim \epsilon^{-1}/2$ *the expected number of observations with* $z_{i,1} = 0$ *is* 0.5*, so there should be significant uncertainty about* $\theta_{z_1=0}$*. However, parametric bootstrap may underestimate the uncertainty: the probability of a resampled dataset* $D_n^{(k)}$ *containing no samples with* $z_{i,1} = 0$ *is* $(1-\epsilon^{-1})^n \sim e^{-1/2}$*, in which case there may not be any meaningful variation in the respective estimate,* $\hat{\theta}^{(k)}_{z_1=0}$*, e.g., if the estimation algorithm applies a small regularisation on* $|\theta_{z_1=0}|$*. However, our method with* $N \gg n$ *will eventually update all* $\hat{\theta}^{(k)}$ *with probability* $1 - (1-\epsilon^{-1})^N \to 1$*.*

The above examples are clearly oversimplified. In practice, the initialisation randomness in optimisation will also contribute to the uncertainty estimates and may help narrow the gap between these procedures, especially on DNN models (Lakshminarayanan et al., 2017). Still, the examples illustrated how our method may have a more direct impact on the final uncertainty estimates, especially in aspects which the training data is not informative about. It may thus offer additional improvements over conventional ensemble approaches that rely solely on initialisation randomness.

5 Experiments

In this section we evaluate the proposed method empirically, across a variety of tasks involving NN and non-NN models.

Hyperparameter learning for GPs. We first evaluate the proposed method on a GP hyperparameter learning task, using empirical Bayes (`EB`) as the base estimation algorithm. EB is a standard approach for this problem, but on problems with fewer observations it can suffer from overfitting and local optima issues (Williams and Rasmussen, 2006). Thus, we investigate the possibility of alleviating overfitting using the proposed method (`IPB`), and compare it with nonparametric bootstrap (`BS`) and a vanilla ensemble method (`Ens`) that aggregates approximate local optima obtained by starting from random initialisations.

We adopt GP models with a Matérn-3/2 kernel and a Gaussian likelihood, and optimise a vector-valued bandwidth hyperparameter (as in Automatic Relevance Detection; Neal, 1996) and the likelihood variance. We evaluate on 9 UCI regression datasets frequently used in recent works on GP and BNN models (e.g., Salimbeni and Deisenroth, 2017; Sun et al., 2018; Dutordoir et al., 2020), and subsample $n \in \{75, 300\}$ data points for training. We report the following metrics: root mean-squared error (RMSE), negative log predictive density (NLPD) and continuous ranked probability score (CRPS). All experiments are repeated on 50 random train/test splits. For space reasons, we defer full details and results to App. D.2, and report the average rank of each method in Table 1. We can see that the proposed method (IPB) attains the best average rank w.r.t. all metrics.

Table 1: GP experiment: average rank across all UCI datasets for each metric. Boldface denotes the best method.

Metric	$n = 75$				$n = 300$			
	EB	BS	Ens	IPB	EB	BS	Ens	IPB
RMSE	3.1	2.7	2.4	**1.4**	2.9	3.0	2.0	**1.1**
NLPD	3.0	2.0	2.6	**1.6**	2.7	3.0	2.2	**1.1**
CRPS	3.0	2.3	2.6	**1.4**	2.7	3.3	2.1	**1.1**

Classification with boosting tree and stacking algorithms. We now turn to classification tasks and illustrate our method using two predictive algorithms: *(i)* gradient boosting decision trees (*GBDT*s, Friedman, 2001), and *(ii) stacking* (Wolpert, 1992; Caruana et al., 2004) based on a variety of tree and DNN models. Both are highly competitive approaches that outperform deep learning methods (Grinsztajn et al., 2022; Shwartz-Ziv and Armon, 2022), yet they do not have a natural Bayesian counterpart. Our method fills in this important gap and provides a means to mitigate overfitting and conduct uncertainty quantification based on Bayesian principles.

We adopt the implementations of XGBoost (Chen and Guestrin, 2016) and AutoGluon (Erickson et al., 2020) for the two base algorithms, and evaluate on 30 OpenML (Bischl et al., 2017) datasets as chosen by Hollmann et al. (2022). For each algorithm, we apply our method (`IPB`) and compare with bootstrap aggregation (`BS`) and the base algorithm without additional aggregation. Our Alg. 1 is implemented by sampling $\hat{x}_{n+i}$ from the empirical distribution of all past inputs. All hyperparameters for the base algorithms, and $(\Delta n, N)$ in our algorithm, are determined using log likelihood on a validation set. Full details are deferred to App. D.3.

Table 2 reports the average test accuracy and negative log likelihood (NLL), which are computed using 10 random splits for each dataset. We can see that for both choices of base algorithms, our method outperforms over the base algorithm as well as its bagging variant. The improvement of the likelihood metric is particularly notable and is expected for methods that better account for the predictive uncertainty. Full results are deferred to App. D.3, where we further show that the improvement is consistent across all datasets, and that our method produces *informative uncertainty estimates for the feature importance scores* from GDBT.

Table 2: Classification experiment: average test metrics and average ranks across 30 OpenML datasets. Boldface indicates the best result within each group of methods. Ranks are calculated by sorting across all six methods.

Metric	GBDT			Stacking		
	(Base)	+ BS	+ IPB	(Base)	+ BS	+ IPB
NLL	0.215	0.207	**0.200**	0.215	0.190	**0.185**
NLL Rank	4.77	4.33	**3.20**	3.60	3.03	**2.07**
Accuracy	90.4	90.7	**90.9**	91.0	91.3	**91.5**
Acc. Rank	4.87	4.43	**3.23**	3.50	2.50	**2.47**

Interventional density estimation with diffusion models. Finally, we present a set of NN-based experiments which concern the estimation of interventional distributions (Pearl, 2009) given a causal graph. Such a task can be seen as conditional density estimation but involves distribution shifts induced by the intervention. Recent works demonstrated the efficacy of variational auto-encoder

(Sánchez-Martin et al., 2022), flow (Khemakhem et al., 2021) and diffusion models (Chao et al., 2023) on this task. We are interested in whether our algorithm could lead to further improvements by accounting for predictive uncertainty, which can be especially relevant here due to the presence of distribution shift.

We instantiate Alg. 1 using a collection of diffusion models following Chao et al. (2023), and the modified objective (11). We use a fully-connected NN model and determine hyperparameters using the training objective evaluated on an (in-distribution) validation set. We evaluate on two sets of datasets: *(i)* 8 synthetic datasets in Chao et al. (2023); *(ii)* a set of real-world fMRI datasets constructed by Khemakhem et al. (2021). In both cases we repeat all experiments 30 times, using independently sampled train/validation splits and initialisation for NN parameters. We note that this deviates from Khemakhem et al. (2021) who appear to have only averaged over patients (i.e., different datasets) but not NN initialisation. See App. D.4 for full details.

For the synthetic datasets, we compute the maximum mean discrepancy (MMD) w.r.t. the ground truth on a grid of queries, following Chao et al. (2023). We compare with alternative ensemble methods applied to the same model: aggregation with parametric (`PB`) and nonparametric (`BS`) bootstrap, deep ensemble (`Ens`), and the method of He et al. (2020, `NTKGP`). Among the baselines, `Ens` has demonstrated strong performance in previous benchmarks (e.g., Gustafsson et al., 2020; Ovadia et al., 2019), and `NTKGP` is motivated from a setup similar to our Sec. 3.2.2 (see App. C for details). Table 3 reports the average rank of the MMD metric across all datasets; we can see that the proposed method (IPB) achieves the best overall performance. Full results and additional discussions are deferred to App. D.4, where we also evaluate the credible/confidence intervals (for interventional expectation functions) produced by all methods, and find our method generally provides the best coverage, followed by the two bagging baselines.

Table 3: Interventional density estimation: average rank across all synthetic datasets. Boldface indicates the best result.

n	PB	Ens.	NTKGP	BS	IPB
100	3.6	1.9	5.0	3.1	**1.0**
1000	4.0	1.9	5.0	2.4	**1.2**

On the fMRI datasets, we report the median absolute error following Khemakhem et al. (2021); Chao et al. (2023), as well as the CRPS which puts more emphasis on the estimation quality for the entire distribution. We compare with the flow-based method of Khemakhem et al. (2021) and the two baselines evaluated therein, as well as the same diffusion model combined with deep ensemble (`D+Ens`) and nonparametric bootstrap (`D+BS`). As shown in Table 4, our method (`D+IPB`) achieves the best overall performance.

Table 4: Results for the fMRI datasets. Boldface indicates the best result ($p < 0.05$ in a Z test).

Metric	Linear	ANM	Flow	D + Ens	D + BS	D + IPB
CRPS	.738$\pm$.10	.551$\pm$.01	.546$\pm$.02	.520$\pm$.00	**.518**$\pm$.00	**.518**$\pm$.00
Abs. Err	.658$\pm$.03	.655$\pm$.01	**.605**$\pm$.02	.609$\pm$.01	.611$\pm$.01	**.604**$\pm$.00

6 Conclusion

We studied uncertainty quantification using general ML algorithms, starting from the postulation that commonly used algorithms should be near-Bayes optimal on an unknown task distribution. We proved in simplified settings that it is possible to recover the unknown but optimal Bayesian posterior through a martingale posterior, and proposed a novel method which is applicable across NN and non-NN models. Experiments confirmed the efficacy of the method.

Our work is clearly not without limitations: the theoretical results do not cover real-world applications such as those involving practical NN models, and the experiments largely focuses on small-sample datasets. Nonetheless, our work demonstrates the potential of the cross-task perspective in Bayesian modelling, and we hope that it may inspire further investigation into this problem.

References

Aitchison, L. (2020a). A statistical theory of cold posteriors in deep neural networks. *arXiv preprint arXiv:2008.05912*.

Aitchison, L. (2020b). Why bigger is not always better: on finite and infinite neural networks. In *International Conference on Machine Learning*, pages 156–164. PMLR.

Amari, S.-i. (2016). *Information geometry and its applications*, volume 194. Springer.

Bae, J., Ng, N., Lo, A., Ghassemi, M., and Grosse, R. (2022). If influence functions are the answer, then what is the question? arXiv:2209.05364 [cs, stat].

Bischl, B., Casalicchio, G., Feurer, M., Gijsbers, P., Hutter, F., Lang, M., Mantovani, R. G., van Rijn, J. N., and Vanschoren, J. (2017). Openml benchmarking suites. *arXiv preprint arXiv:1708.03731*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24:123–140.

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398.

Burt, D., Rasmussen, C. E., and Van Der Wilk, M. (2019). Rates of convergence for sparse variational Gaussian process regression. In *International Conference on Machine Learning*, pages 862–871. PMLR.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, page 18.

Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004.

Chao, P., Blöbaum, P., and Kasiviswanathan, S. P. (2023). Interventional and counterfactual inference with diffusion models. *arXiv preprint arXiv:2302.00860*.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems*, 32.

D'Angelo, F. and Fortuin, V. (2021). Repulsive deep ensembles are Bayesian. *Advances in Neural Information Processing Systems*, 34:3451–3465.

DeepMind, e. a. (2020). The DeepMind JAX Ecosystem.

Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pages 23–27.

Dusenberry, M., Jerfel, G., Wen, Y., Ma, Y., Snoek, J., Heller, K., Lakshminarayanan, B., and Tran, D. (2020). Efficient and scalable Bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pages 2782–2792. PMLR.

Dutordoir, V., Durrande, N., and Hensman, J. (2020). Sparse Gaussian processes with spherical harmonic features. In *International Conference on Machine Learning*, pages 2793–2802. PMLR.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The annals of applied statistics*, 6(4):1971.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. (2020). Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*.

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach*. Academic press.

Fong, E., Holmes, C., and Walker, S. G. (2021). Martingale posterior distributions. arXiv:2103.15671 [math, stat].

Fort, S., Hu, H., and Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

Fortuin, V., Garriga-Alonso, A., Ober, S. W., Wenzel, F., Ratsch, G., Turner, R. E., van der Wilk, M., and Aitchison, L. (2021). Bayesian neural network priors revisited. In *International Conference on Learning Representations*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Ghalebikesabi, S., Holmes, C. C., Fong, E., and Lehmann, B. (2023). Quasi-Bayesian nonparametric density estimation via autoregressive predictive updates. In *Uncertainty in Artificial Intelligence*, pages 658–668. PMLR.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520.

Gustafsson, F. K., Danelljan, M., and Schon, T. B. (2020). Evaluating scalable Bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 318–319.

He, B., Lakshminarayanan, B., and Teh, Y. W. (2020). Bayesian deep ensembles via the neural tangent kernel. *arXiv preprint arXiv:2007.05864*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. (2022). Tabpfn: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*.

Holmes, C. C. and Walker, S. G. (2023). Statistical inference with exchangeability and martingales. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2247):20220143.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. G. (2021). What are Bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Jeffreys, H. (1939). *The theory of probability*. Oxford University Press.

Khemakhem, I., Monti, R., Leech, R., and Hyvarinen, A. (2021). Causal autoregressive flows. In *International conference on artificial intelligence and statistics*, pages 3520–3528. PMLR.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lam, H. and Wang, Z. (2023). Resampling stochastic gradient descent cheaply for efficient uncertainty quantification. arXiv:2310.11065 [cs, stat].

Lee, H., Yun, E., Nam, G., Fong, E., and Lee, J. (2022). Martingale posterior neural processes. In *The Eleventh International Conference on Learning Representations*.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32.

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., and Li, J. (2019). Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24.

Nabarro, S., Ganev, S., Garriga-Alonso, A., Fortuin, V., van der Wilk, M., and Aitchison, L. (2022). Data augmentation in Bayesian neural networks and the cold posterior effect. In *Uncertainty in Artificial Intelligence*, pages 1434–1444. PMLR.

Neal, R. (1996). Bayesian learning for neural networks. *Lecture Notes in Statistics*.

Nieman, D., Szabo, B., and Van Zanten, H. (2022). Contraction rates for sparse variational approximations in Gaussian process regression. *The Journal of Machine Learning Research*, 23(1):9289–9314.

OpenAI (2023). Fine-tuning service. `https://platform.openai.com/docs/guides/fine-tuning/`.

Osband, I., Aslanides, J., and Cassirer, A. (2018). Randomized prior functions for deep reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8626–8638.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Pearce, T., Leibfried, F., and Brintrup, A. (2020). Uncertainty in neural networks: Approximately Bayesian ensembling. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 234–244. PMLR.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pentina, A. and Lampert, C. (2014). A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pages 991–999. PMLR.

Razin, N. and Cohen, N. (2020). Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187.

Riou, C., Alquier, P., and Chérief-Abdellatif, B.-E. (2023). Bayes meets Bernstein at the meta level: an analysis of fast rates in meta-learning with PAC-Bayes. *arXiv preprint arXiv:2302.11709*.

Salimbeni, H. and Deisenroth, M. (2017). Doubly stochastic variational inference for deep Gaussian processes. *Advances in neural information processing systems*, 30.

Sánchez-Martin, P., Rateike, M., and Valera, I. (2022). Vaca: designing variational graph autoencoders for causal queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8159–8168.

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

Snelson, E. L. (2008). *Flexible and efficient Gaussian process models for machine learning*. University of London, University College London (United Kingdom).

Steinwart, I. (2019). Convergence types and rates in generic karhunen-loeve expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395.

Sun, S., Zhang, G., Shi, J., and Grosse, R. (2018). Functional variational Bayesian neural networks. In *International Conference on Learning Representations*.

Syversveen, A. R. (1998). Noninformative Bayesian priors. interpretation and problems with construction and applications. *Preprint statistics*, 3(3):1–11.

Tran, B.-H., Rossi, S., Milios, D., and Filippone, M. (2022). All you need is a good functional prior for Bayesian deep learning. *The Journal of Machine Learning Research*, 23(1):3210–3265.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

van der Vaart, A. W., van Zanten, J. H., et al. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections*, 3:200–222.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wang, Z., Ren, T., Zhu, J., and Zhang, B. (2018). Function space particle optimization for Bayesian neural networks. In *International Conference on Learning Representations*.

Wen, Y., Tran, D., and Ba, J. (2020). BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*.

Wenzel, F., Roth, K., Veeling, B. S., Świątkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. (2020). How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Xu, A. and Raginsky, M. (2022). Minimum excess risk in Bayesian learning. *IEEE Transactions on Information Theory*, 68(12):7935–7955.

Yao, Y., Vehtari, A., and Gelman, A. (2022). Stacking for non-mixing Bayesian computations: The curse and blessing of multimodal posteriors. *The Journal of Machine Learning Research*, 23(1):3426–3471.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, 23(4):550–560.

A Deferred Proofs

In the proofs we adopt the following notations: for all $j \geq n$, let $\mathcal{F}_j$ be the σ-algebra generated by "all observations up to iteration j"; this includes $\{z_{1:n}, \widehat{z}_{n+1:j}, z^B_{n+1:j}\}$ as well as the samples $\{\breve{z}_{n+1:j}\}$ which will be defined shortly below. Define $\mathbb{E}_j := \mathbb{E}(\cdot \mid \mathcal{F}_j)$. We will also make frequent use of the inequality

$$\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle \delta^{1/2}a, \delta^{-1/2}b\rangle \leq (1+\delta)\|a\|^2 + (1+\delta^{-1})\|b\|^2, \tag{12}$$

which holds for all Hilbert norms, a, b and $\delta > 0$. This implies, in particular, $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$. It also follows that, for any $\{\mathcal{F}_j\}$-adapted $\{a_j\}$ and any $\{b_j\}$,

$$\begin{aligned}\mathbb{E}_j\|a_j + b_j\|^2 &= \|a_j\|^2 + \mathbb{E}_j\|b_j\|^2 + 2\langle \delta^{1/2}a_j, \mathbb{E}_j\delta^{-1/2}b_j\rangle \\ &\leq (1+\delta)\|a_j\|^2 + \mathbb{E}_j\|b_j\|^2 + \delta^{-1}\|\mathbb{E}_j b_j\|^2. \end{aligned}\tag{13}$$

A.1 Proof for Theorem 3.1

By the first inequality in assumption 3.3 it suffices to prove (6). By definitions, it suffices to create a coupled version of $(\widehat{\theta}_\infty, \bar{\theta}^B_\infty) \mid \mathcal{F}_n$ so that $\mathbb{E}_\pi\|\widehat{\theta}_\infty - \bar{\theta}^B_\infty\|^2$ is bounded by the RHS of (6). And since Asm. 3.5 (ii) and 3.3 imply that $\lim_{j\to\infty}\breve{\varepsilon}_{ex,j} = 0$, we have $\lim_{j\to\infty}\mathbb{E}_\pi\|\breve{\theta}_j - \bar{\theta}^B_\infty\|^2 = 0$, and

$$\mathbb{E}_\pi\|\breve{\theta}_\infty - \widehat{\theta}_\infty\|^2 = \mathbb{E}_\pi\|\bar{\theta}^B_\infty - \widehat{\theta}_\infty\|^2, \tag{14}$$

and it suffices to bound the LHS. We will construct a sequence of couplings between $\{\widehat{z}_{j+1}\}$ and $\{z^B_{j+1}\}$ so that the LHS is bounded as claimed. For this purpose, we will introduce an additional r.v. $\breve{z}_{j+1}$ s.t. $\mathbb{P}(\breve{z}_{j+1} \in \cdot \mid \mathcal{F}_j) = \mathbb{P}_{\breve{\theta}_j}$, and define the joint distribution $\mathbb{P}(\breve{z}_{j+1}, \widehat{z}_{j+1}, z^B_{j+1} \mid \mathcal{F}_j) = \mathbb{P}(\breve{z}_{j+1} \mid \mathcal{F}_j)\mathbb{P}(\widehat{z}_{j+1} \mid \breve{z}_{j+1}, \mathcal{F}_j)\mathbb{P}(z^B_{j+1} \mid \breve{z}_{j+1}, \mathcal{F}_j)$ with the last two terms determined by various optimal transport plans.

Let $s > 0$ be defined in assumption 3.3. Consider the decomposition

$$\begin{aligned}&\mathbb{E}_j\|\widehat{\theta}_{j+1} - \breve{\theta}_{j+1}\|^2 \\ &= \mathbb{E}_j\|\widehat{\theta}_j + \widehat{\Delta}_j(\widehat{\theta}_j, \widehat{z}_{j+1}) - (\breve{\theta}_j + \widehat{\Delta}_j(\breve{\theta}_j, \breve{z}_{j+1}) - \widehat{\Delta}_j(\breve{\theta}_j, \breve{z}_{j+1}) + \widehat{\Delta}_j(\breve{\theta}_j, z^B_{j+1}))\|^2 \\ &\overset{(13)}{\leq} (1 + j^{-(1+s)})\mathbb{E}_j\|\widehat{\theta}_j - \breve{\theta}_j - (\widehat{\Delta}_j(\breve{\theta}_j, \breve{z}_{j+1}) - \widehat{\Delta}_j(\breve{\theta}_j, z^B_{j+1}))\|^2 \\ &\qquad + \mathbb{E}_j\|\widehat{\Delta}_j(\widehat{\theta}_j, \widehat{z}_{j+1}) - \widehat{\Delta}_j(\breve{\theta}_j, \breve{z}_{j+1})\|^2 + j^{1+s}(2\|\mathbb{E}_j\widehat{\Delta}_j(\widehat{\theta}_j, \widehat{z}_{j+1})\|^2 + 2\|\mathbb{E}_j\widehat{\Delta}_j(\breve{\theta}_j, \breve{z}_{j+1})\|^2) \\ &=: (1 + j^{-(1+s)})A_j + B_j + j^{1+s}C_j. \end{aligned}\tag{15}$$

We will bound the three terms in turn.

For C_j, we note that by Asm. 3.3, Asm. 3.1 also holds for $\delta = s$, and thus we have

$$j^{1+s}C_j \leq 2j^{-(1+s)}\bar{\varepsilon}^2_{B,j}. \tag{16}$$

For B_j, first note that by assumption 3.2 (i) we have

$$\begin{aligned}B_j &\leq 2(\mathbb{E}_j\|\widehat{\Delta}_j(\widehat{\theta}_j, \widehat{z}_{j+1}) - \widehat{\Delta}_j(\breve{\theta}_j, \widehat{z}_{j+1})\|^2 + \mathbb{E}_j\|\widehat{\Delta}_j(\breve{\theta}_j, \widehat{z}_{j+1}) - \widehat{\Delta}_j(\breve{\theta}_j, \breve{z}_{j+1})\|^2) \\ &\leq 2\eta_j^2(L_1^2\|\widehat{\theta}_j - \breve{\theta}_j\|^2 + L_2^2\mathbb{E}_j\|\widehat{z}_{j+1} - \breve{z}_{j+1}\|_z^2). \end{aligned}\tag{17}$$

Let $\mathbb{P}(\widehat{z}_{j+1} \mid \mathcal{F}_j, \breve{z}_{j+1})$ be defined by the optimal transport plan that minimises the transport cost above. Then if (5) holds, the above will be bounded by $2\eta_j^2(L_1^2 + L_2^2 C_\Theta)\|\widehat{\theta}_j - \breve{\theta}_j\|^2$, and we have $\eta_j \leq j^{-(1+\iota)/2}$; otherwise (5') must hold, and we have $j^{1/4}\eta_j \leq j^{-(1+\iota)/2}$, and

$$\begin{aligned}2\eta_j^2 L_2^2\mathbb{E}_j\|\widehat{z}_{j+1} - \breve{z}_{j+1}\|_z^2 &\leq 2\eta_j^2 L_2^2 C_\Theta\|\breve{\theta}_j - \widehat{\theta}_j\| = L_2^2 C_\Theta \cdot 2j^{-1/2}(j^{1/4}\eta_j)\cdot(j^{1/4}\eta_j)\|\breve{\theta}_j - \widehat{\theta}_j\| \\ &\leq L_2^2 C_\Theta \cdot \big((j^{-1/2}(j^{1/4}\eta_j))^2 + (j^{1/4}\eta_j\|\breve{\theta}_j - \widehat{\theta}_j\|)^2\big) \\ &= L_2^2 C_\Theta \cdot (j^{1/4}\eta_j)^2\big(\|\breve{\theta}_j - \widehat{\theta}_j\|^2 + j^{-1}\big) \\ &\leq L_2^2 C_\Theta \cdot (j^{1/4}\eta_j)^2(\|\breve{\theta}_j - \widehat{\theta}_j\|^2 + \bar{\varepsilon}^2_{B,j}), \end{aligned}$$

where the last inequality follows from Assumption 3.5. Define $\eta'_j := j^{-(1+\iota)/2}$, then in both cases we have

$$2\eta_j^2 L_2^2 \mathbb{E}_j \|\widehat{z}_{j+1} - \breve{z}_{j+1}\|_z^2 \le L_2^2 C_\Theta \eta_j'^2 (\|\breve{\theta}_j - \widehat{\theta}_j\|^2 + \bar{\varepsilon}_{B,j}^2). \tag{18}$$

Plugging back to (17) we have

$$B_j \le 2\eta_j'^2 (L_1^2 + L_2^2 C_\Theta)(\|\widehat{\theta}_j - \breve{\theta}_j\|^2 + \bar{\varepsilon}_{B,j}^2). \tag{19}$$

For A_j, we first use (13) to bound it as

$$\begin{aligned} A_j \le \mathbb{E}_j((1 + j^{-(1+s)})\|\widehat{\theta}_j - \breve{\theta}_j\|^2 + \|\widehat{\Delta}_j(\breve{\theta}_j, \breve{z}_{j+1}) - \widehat{\Delta}_j(\breve{\theta}_j, z_{j+1}^B)\|^2) \\ + j^{1+s}\|\mathbb{E}_j(\widehat{\Delta}_j(\breve{\theta}_j, \breve{z}_{j+1}) - \widehat{\Delta}_j(\breve{\theta}_j, z_{j+1}^B))\|^2 \\ \le (1 + j^{-(1+s)})\|\widehat{\theta}_j - \breve{\theta}_j\|^2 + \mathbb{E}_j\|\widehat{\Delta}_j(\breve{\theta}_j, \breve{z}_{j+1}) - \widehat{\Delta}_j(\breve{\theta}_j, z_{j+1}^B)\|^2 + 2j^{1+s}C_j + \\ 2j^{1+s}\|\mathbb{E}_j\widehat{\Delta}_j(\breve{\theta}_j, z_{j+1}^B)\|^2. \end{aligned} \tag{20}$$

We now bound the second and last terms above. For the second term we introduce our coupling between $(\breve{z}_{j+1}, z_{j+1}^B) \mid \mathcal{F}_j$ as follows. Recall the conditional distribution $z_{j+1}^B \mid \mathcal{F}_j$ can be represented as $\theta \sim \pi(\theta \mid \mathcal{F}_j),\ z_{j+1}^B \sim \mathbb{P}_\theta$; we thus define $\mathbb{P}(z_{j+1}^B \mid \mathcal{F}_j, \widehat{z}_{j+1})$ through

$$\theta \sim \pi(\theta \mid \mathcal{F}_j), \ \ z_{j+1}^B \mid (\theta, \breve{z}_{j+1}) \sim \Gamma_{\mathbb{P}_{\breve{\theta}_j} \to \mathbb{P}_\theta}(\cdot \mid \breve{z}_{j+1}), \tag{21}$$

where $\Gamma_{P \to Q}$ denotes the conditional probability derived from the optimal transport plan from P to Q. Clearly this preserves both marginal distributions as required, and we have

$$\begin{aligned} \mathbb{E}_j\|\widehat{\Delta}_j(\breve{\theta}_j, \breve{z}_{j+1}) - \widehat{\Delta}_j(\breve{\theta}_j, z_{j+1}^B)\|^2 &\le \eta_j^2 L_2^2 \mathbb{E}_j \|\breve{z}_{j+1} - z_{j+1}^B\|_z^2 && \text{(Asm. 3.2 (i))} \\ &\overset{(21)}{\le} \eta_j^2 L_2^2 \mathbb{E}_{\theta \sim \pi(\cdot|\mathcal{F}_j)} W_2^2(\mathbb{P}_{\breve{\theta}_j}, \mathbb{P}_\theta). \end{aligned}$$

Repeating the proof for (18) we find the above is bounded as

$$\begin{aligned} \eta_j^2 L_2^2 \mathbb{E}_{\theta \sim \pi(\cdot|\mathcal{F}_j)} W_2^2(\mathbb{P}_{\breve{\theta}_{j+1}}, \mathbb{P}_\theta) &\le L_2^2 C_\Theta \eta_j'^2 (\mathbb{E}_{\theta \sim \pi(\cdot|\mathcal{F}_j)}\|\breve{\theta}_j - \theta\|^2 + \bar{\varepsilon}_{B,j}^2) \\ &= L_2^2 C_\Theta \eta_j'^2 (\breve{\varepsilon}_{ex,j}^2 + 2\bar{\varepsilon}_{B,j}^2), \end{aligned} \tag{22}$$

where the last line follows from the fact that $\theta \mid \mathcal{F}_j \overset{d}{=} \bar{\theta}_\infty^B \mid \mathcal{F}_j$. Now, turning to the last term of (20), we have

$$\begin{aligned} &\|\mathbb{E}_j\widehat{\Delta}_j(\breve{\theta}_j, z_{j+1}^B)\|^2 \\ &\overset{(21)}{=} \|\mathbb{E}_{\theta \sim \pi(\cdot|\mathcal{F}_j)}\mathbb{E}_{z \sim \mathbb{P}_\theta}\widehat{\Delta}_j(\breve{\theta}_j, z)\|^2 \\ &= \|\mathbb{E}_{\theta|\mathcal{F}_j}\mathbb{E}_{z|\theta}(\widehat{\Delta}_j(\breve{\theta}_j, z) - \eta_j H_{\breve{\theta}_j}(\theta - \breve{\theta}_j) + \eta_j H_{\breve{\theta}_j}(\theta - \breve{\theta}_j))\|^2 \\ &\le 2\|\mathbb{E}_{\theta|\mathcal{F}_j}\mathbb{E}_{z|\theta}(\widehat{\Delta}_j(\breve{\theta}_j, z) - \eta_j H_{\breve{\theta}_j}(\theta - \breve{\theta}_j))\|^2 + 2\|\mathbb{E}_{\theta|\mathcal{F}_j}\eta_j H_{\breve{\theta}_j}(\theta - \breve{\theta}_j)\|^2 \\ &\le 2(\mathbb{E}_{\theta|\mathcal{F}_j}\|\mathbb{E}_{z \sim \mathbb{P}_\theta}\widehat{\Delta}_j(\breve{\theta}_j, z) - H_{\breve{\theta}_j}(\theta - \breve{\theta}_j)\|)^2 + 2\|\eta_j H_{\breve{\theta}_j}(\bar{\theta}_j^B - \breve{\theta}_j)\|^2 \\ &\le 2(\mathbb{E}_{\theta|\mathcal{F}_j} C_\mathcal{A}\Theta\eta_j\|\breve{\theta}_j - \theta\|^2)^2 + 2C'_\mathcal{A}\eta_j^2\breve{\varepsilon}_{ex,j}^2 && \text{(Asm. 3.4)} \\ &= 2\eta_j^2 C_\mathcal{A}^2(\breve{\varepsilon}_{ex,j}^2 + \bar{\varepsilon}_{B,j}^2)^2 + 2C'_\mathcal{A}\eta_j^2\breve{\varepsilon}_{ex,j}^2 \le 4\eta_j^2(C_\mathcal{A}'^2\breve{\varepsilon}_{ex,j}^2 + C_\mathcal{A}\bar{\varepsilon}_{B,j}^4). && \text{(Asm. 3.5 (i))} \end{aligned} \tag{23}$$

Plugging (23) and (22) into (20), we have

$$\begin{aligned} A_j \le (1 + j^{-(1+s)})\|\widehat{\theta}_j - \breve{\theta}_j\|^2 + \eta_j'^2 L_2^2 C_\Theta(\breve{\varepsilon}_{ex,j}^2 + 2\bar{\varepsilon}_{B,j}^2) \\ + 8j^{1+s}\eta_j^2(C'_\mathcal{A}\breve{\varepsilon}_{ex,j}^2 + C_\mathcal{A}\bar{\varepsilon}_{B,j}^4) + 2j^{1+s}C_j \\ \le (1 + j^{-(1+s)})\|\widehat{\theta}_j - \breve{\theta}_j\|^2 + C'_\Theta(\eta_j'^2\bar{\varepsilon}_{B,j}^2 + j^{1+s}\eta_j^2(\breve{\varepsilon}_{ex,j}^2 + C_\mathcal{A}\bar{\varepsilon}_{B,j}^4)) + 2j^{1+s}C_j, \end{aligned} \tag{24}$$

where the constant C'_Θ is determined by L_1, L_2, C_Θ and $C'_\mathcal{A}$. Plugging (24), (19) and (16) into (15) and taking expectation, we find

$$\begin{aligned} \mathbb{E}_\pi\|\widehat{\theta}_{j+1} - \breve{\theta}_{j+1}\|^2 \le (1 + 2j^{-(1+s)} + \eta_j'^2 C'_\Theta)\mathbb{E}_\pi\|\widehat{\theta}_j - \breve{\theta}_j\|^2 \\ + 3C'_\Theta(\eta_j'^2\bar{\varepsilon}_{B,j}^2 + j^{1+s}\eta_j^2(\breve{\varepsilon}_{ex,j}^2 + C_\mathcal{A}\bar{\varepsilon}_{B,j}^4)) + 8j^{-(1+s)}\bar{\varepsilon}_{B,j}^2. \end{aligned}$$

Define $\Delta\chi_j := 2j^{-(1+s)} + C'_\Theta \eta_j'^2, \chi_l := \sum_{j=l}^\infty \Delta\chi_j$. Then $\chi_l < \infty$ and we have

$$
\begin{aligned}
&\mathbb{E}_\pi \|\widehat{\theta}_{j+1} - \breve{\theta}_{j+1}\|^2 \\
&\leq e^{\Delta\chi_j} \mathbb{E}_\pi \|\widehat{\theta}_j - \breve{\theta}_j\|^2 + 3C'_\Theta(\eta_j'^2 \bar{\varepsilon}_{B,j}^2 + j^{1+s}\eta_j^2(\breve{\varepsilon}_{ex,j}^2 + C_{\mathcal{A}}\bar{\varepsilon}_{B,j}^4)) + 8j^{-(1+s)}\bar{\varepsilon}_{B,j}^2, \\
&\mathbb{E}_\pi \|\widehat{\theta}_N - \breve{\theta}_N\|^2 \\
&\leq e^{\chi_n} \left(\mathbb{E}_n \|\widehat{\theta}_n - \breve{\theta}_n\|^2 + \sum_{j=n}^N 3C'_\Theta(\eta_j'^2 \bar{\varepsilon}_{B,j}^2 + j^{1+s}\eta_j^2(\breve{\varepsilon}_{ex,j}^2 + C_{\mathcal{A}}\bar{\varepsilon}_{B,j}^4)) + 8j^{-(1+s)}\bar{\varepsilon}_{B,j}^2 \right) \\
&\leq e^{\chi_n}(\mathbb{E}_n \|\widehat{\theta}_n - \breve{\theta}_n\|^2 + C(\chi_n + \nu_n)\bar{\varepsilon}_{B,n}^2),
\end{aligned}
$$

where the last inequality follows by Asm. 3.3, 3.5 (iii) and the constant C is determined by C'_Θ. This completes the proof. □

A.2 Deferred Proofs in Section 3.2

A.2.1 Proof for the claims in Section 3.2.1

Claim A.1. *In the setting of Sec. 3.2.1 we have* $\breve{\varepsilon}_{ex,j}^2 \leq 2\alpha j^{-1}\bar{\varepsilon}_{B,j}^2$.

Proof. It follows by our choice of π that

$$
\bar{\theta}_j^B = \frac{j\breve{\theta}_j + \theta_\pi}{j+\alpha} = \bar{\theta}_{j-1}^B + \frac{1}{j+\alpha}(z_j^B - \bar{\theta}_{j-1}^B).
$$

To bound $\bar{\varepsilon}_{B,j}$ we use the above representation, and the fact that $\{\bar{\theta}_j^B\}$ define a martingale; it follows that

$$
\bar{\varepsilon}_{B,j}^2 = \mathbb{E}_\pi \|\bar{\theta}_j^B - \bar{\theta}_\infty^B\|^2 = \sum_{k=j}^\infty \mathbb{E}_\pi \|\bar{\theta}_k^B - \bar{\theta}_{k+1}^B\|^2 = \sum_{k=j}^\infty \mathbb{E}_\pi \frac{\|T(z_{k+1}^B) - \bar{\theta}_k^B\|^2}{(k+\alpha)^2}.
$$

Observe that $\mathbb{P}(z_{k+1}^B \in dz \mid \bar{\theta}_k^B) = \int \mathbb{P}_{\tilde{\theta}_k}(dz)\pi_{k,\bar{\theta}_k^B}(d\tilde{\theta}_k)$, where $\pi_{k,\bar{\theta}_k^B}(d\theta) = \pi(\theta \mid z_{\leq k}^B)$ is the posterior measure, and is *determined by* the posterior mean $\bar{\theta}_k^B$: the posterior for natural parameter is $\pi(\eta \mid z_{\leq k}^B) \propto \exp((k+\alpha)\eta^\top \bar{\theta}_k^B - (k+\alpha)A(\eta))$, and $\pi(\theta \mid z_{\leq k}^B)$ is merely its pushforward by ∇A. Therefore, we have $z_{k+1}^B \perp\!\!\!\perp \bar{\theta}_k^B \mid \tilde{\theta}_k$, and

$$
\begin{aligned}
\mathbb{E}\|T(z_{k+1}^B) - \bar{\theta}_k^B\|^2 &= \mathbb{E}\|T(z_{k+1}^B) - \tilde{\theta}_k\|^2 + \mathbb{E}\|\tilde{\theta}_k - \bar{\theta}_k^B\|^2 + \mathbb{E}\langle (T(z_{k+1}^B) - \tilde{\theta}_k \mid \tilde{\theta}_k, \cancel{\bar{\theta}_k^B}), \tilde{\theta}_k - \bar{\theta}_k^B \rangle \\
&\overset{(i)}{=} \mathbb{E}\|T(z_{k+1}^B) - \tilde{\theta}_k\|^2 + \mathbb{E}\|\tilde{\theta}_k - \bar{\theta}_k^B\|^2 \\
&\geq \mathbb{E}\|T(z_{k+1}^B) - \tilde{\theta}_k\|^2 \\
&\overset{(ii)}{=} \mathbb{E}_{\theta\sim\pi, z\sim\mathbb{P}_\theta} \|T(z) - \theta\|^2 =: V_\pi.
\end{aligned}
$$

In the above, (i) holds because $\tilde{\theta}_k$ is the mean parameter for z_{k+1}^B, and (ii) holds because the marginal distributions for all posterior samples $\tilde{\theta}_k$ equal the prior. Plugging back, we find

$$
\bar{\varepsilon}_{B,j}^2 \geq \sum_{k=j}^\infty \frac{V_\pi}{(k+\alpha)^2} \geq \frac{1}{j+\alpha} V_\pi.
$$

For $\breve{\varepsilon}_{ex,j}$, we have

$$
\begin{aligned}
\mathbb{E}_\pi \|\widehat{\theta}_j - \theta_0\|^2 &= \mathbb{E}_{\theta\sim\pi, z_{1:j}\sim\mathbb{P}_\theta^{\otimes j}} (\mathbb{E}(\|\widehat{\theta}_j - \theta\|^2 \mid \theta)) \\
&= \mathbb{E}_{\theta\sim\pi, z_{1:j}\sim\mathbb{P}_\theta^{\otimes j}} \left(\mathbb{E}\left(\left\| \frac{1}{j} \sum_{k=1}^j T(z_k) - \theta \right\|^2 \middle| \theta \right) \right) \\
&= \mathbb{E}_{\theta\sim\pi, z\sim\mathbb{P}_\theta} \frac{\|T(z) - \theta\|^2}{j} = \frac{1}{j} V_\pi,
\end{aligned}
$$

where the last equality follows from conditional independence. It thus follows that

$$\breve{\varepsilon}^2_{ex,j} \le \frac{\alpha}{j(j+\alpha-1)} V_\pi \le \frac{\alpha}{j} \cdot \frac{j+\alpha}{j+\alpha-1} \bar{\varepsilon}^2_{B,j} \le \frac{2\alpha}{j} \bar{\varepsilon}^2_{B,j}.$$

This completes the proof. □

Claim A.2. *Let F_θ denote the Fisher information matrix for p_θ. In the setting of Sec. 3.2.1, Theorem 3.1 holds if* $(\|T\|_\infty := \sup_{z\in\mathcal{Z}} \|T(z)\|, \sup_\theta \lambda_{max}(F_\theta), \sup_\theta \lambda^{-1}_{min}(F_\theta))$ *are all bounded.*

Proof for Claim A.2. Observe that Theorem 3.1 will continue to hold if we replace all occurrences of z with $T(z)$ (and the norm $\|\cdot\|_z$ with $\|\cdot\|$) in its proofs and assumptions: this is because both the MP and the Bayesian posterior only depend on z through $T(z)$. Therefore, to prove the claim it suffices to establish Assumption 3.2 (ii)–or Eq. (5')–after the replacement. The equation holds because

$$\begin{aligned}
& W_2^2(T_\# p_\theta, T_\# p_{\theta'}) \\
& \le 2 \sup_{z,z'<\infty} \|T(z) - T(z')\|^2 D_{TV}(T_\# p_\theta, T_\# p_{\theta'}) && \text{(Villani, 2009, Theorem 6.15)} \\
& \le 8\|T\|_\infty^2 D_{TV}(p_\theta, p_{\theta'}) \\
& \le 8\|T\|_\infty^2 \sqrt{\mathrm{KL}(p_\theta, p_{\theta'})/2} && \text{(Pinsker's inequality)} \\
& = 8\|T\|_\infty^2 \sqrt{A(\eta') - A(\eta) - \nabla A(\eta)^\top(\eta' - \eta)} \\
& \le 4\sqrt{2}\|T\|_\infty^2 (\sup_{\tilde\eta} \|\nabla^2 A(\tilde\eta)\|_{op})^{1/2} \|\eta - \eta'\| \\
& \le 4\sqrt{2}\|T\|_\infty^2 \sup_{\tilde\eta} \|\nabla^2 A(\tilde\eta)\|_{op}^{1/2} (\sup_{\tilde\eta'} \|(\nabla^2 A(\tilde\eta'))^{-1}\|_{op}) \|\theta - \theta'\|.
\end{aligned}$$

In the above, $\eta = (\nabla A)^{-1}(\theta), \eta' = (\nabla A)^{-1}(\theta')$ are the respective natural parameters, $T_\#$ denotes the pushforward measure, the LHS is the replaced LHS of (5'), and the coefficients in the RHS are bounded by assumptions, in particular because $\nabla^2 A(\eta) = F_\theta^{-1}$. This completes the proof. □

We note that it should be possible to replace the uniform boundedness conditions with their local counterparts (that only holds in a neighbourhood of θ_0); the resulted conditions can be used to establish a conditional version of the theorem (which can be easily proved by adapting the existing proof). We omit the discussion for brevity.

Finally, we substantiate on the claims about specific exponential family models: for Gaussian model (5) holds because the transport plan is $z \mapsto z + \theta' - \theta$; for $\{Exp(\theta)\}$ (5) holds by considering the transport plan $z \mapsto \frac{\theta'}{\theta} z$. For the Bernoulli model we can establish (5') using the first two inequalities in the above proof.

A.2.2 Deferred proofs and additional discussion for Section 3.2.2

Connection to nonparametric inverse problems and regression. Section 3.2.2 is closely connected to the following inverse problem:

$$\bar{z}_n = A\theta_0 + n^{-1/2} W, \quad \text{where } W \sim \mathcal{N}_{\mathcal{Z}}(0, I). \tag{25}$$

Indeed, we can recover the above problem by setting $\bar{z}_n := \frac{1}{n}\sum_{i=1}^n z_i$. The latter is the classical (nonparametric) linear inverse problem; see Cavalier (2008) for a review. Strictly speaking, our setup is different from (25) as we observe $\{z_i\}$, but *the difference is irrelevant* to our discussion, since we can verify that both the MP and the Bayesian posterior only depend on $\{z_i\}$ through $\bar{z}_n$ and are thus applicable to (25).

When $\alpha = 1$, the problem can be equivalently stated as $\bar{z}_n = \theta_0' + n^{-1/2} W$ where $\theta_0' := A\theta_0$; and the norm of interest becomes $\|\hat\theta - \theta_0\| = \|A\hat\theta - \theta_0'\|_{\mathcal{Z}}$. This is the signal-in-white noise problem which is asymptotically equivalent to regression (Brown and Low, 1996). The prior π for θ corresponds to the GP[2] prior $\pi' := \mathcal{N}_{\mathcal{Z}}(0, AA^\top)$ for θ'. Such priors are "infinitesimally weaker" than assuming

[2] see van der Vaart et al. (2008) for a definition of GPs in Hilbert spaces.

θ_0' to live in $S^{2\beta-1} := \{\theta' = \sum_i i^{-(2\beta-1)/2} a_i \psi_i$ for some $\{a_i\} \in \ell_2(\mathbb{N})\}$ where $\{\psi_i\}$ denotes the left singular vectors of A, as $\theta' \sim \pi'$ will fall into $S^{2\beta-1-\epsilon}$ a.s. for all $\epsilon > 0$ (van der Vaart et al., 2008). The spaces $S^{(\cdot)}$ are known as *Sobolev classes* (see e.g., Cavalier, 2008) and can recover the L_2-Sobolev spaces for suitable choices of β and $\{\psi_i\}$.

Inapplicability of MLE / natural gradient. For both (25) and the data generating process in Section 3.2.2, the MLE $\hat{\theta}_n$ satisfies $A\hat{\theta}_n = \bar{z}_n = \frac{1}{n}\sum_{i=1}^n z_i$. When $\alpha = 1$, the estimation error $\|\hat{\theta}_n - \theta_0\|$ thus equals the *dimensionality* of $\mathcal{Z}$, and is unbounded if the dimensionality is so; the same applies to the natural gradient algorithm with $\eta_j = j^{-1}$ due to its exact equivalence to MLE in this scenario. In contrast, the Bayesian estimator have a bounded error (see (26) below) due to its regularisation effect.

Validating the assumptions for the linear-Gaussian MP. Observe that the posterior equals

$$\pi(\theta \mid z_{\le j}) = \mathcal{N}(\theta \mid \hat{\Sigma}_j^{-1} A^\top \bar{z}_j, (j\hat{\Sigma}_j)^{-1}),$$

where $\hat{\Sigma}_j := A^\top A + j^{-1} I,\ \bar{z}_j := \frac{1}{j}\left(\sum_{i=1}^n z_i + \sum_{i=n+1}^j z_i^B\right)$, and $A^\top$ denotes the adjoint. And we have

$$\bar{\varepsilon}_{B,j}^2 = \mathrm{Tr}((A^\top A)^\alpha (j\hat{\Sigma}_j)^{-1}) = \sum_{i=1}^\infty \frac{s_i^{2\alpha}}{j s_i^2 + 1} \asymp j^{-1} + j^{-\alpha} m_j, \tag{26}$$

where $m_j := \max\{m \in \mathbb{N} : s_m^2 \ge j^{-1}\} \asymp j^{1/2\beta}$. We have introduced the Hilbert spaces $\mathcal{H}, \mathcal{Z}$ and defined the parameter norm $\|\theta\| := \|(A^\top A)^{\alpha/2}\theta\|_{\mathcal{H}} =: \|S\theta\|_{\mathcal{H}}$. In instantiating the theorem we will set the data norm as $\|z\|_z := \|(AA^\top)^{(\alpha-1)/2} z\|_{\mathcal{Z}}$.

We now verify the assumptions in turn.

1. Assumption 3.1 holds for all $\delta > 0$ because $\widehat{\mathrm{Alg}}_j$ defines an exact martingale.
2. Assumption 3.2 holds because for its (i), we have

   $$\begin{aligned}\|\widehat{\Delta}_j(\theta, z) - \widehat{\Delta}_j(\theta', z)\|^2 &= \|S(\widehat{\Delta}_j(\theta, z) - \widehat{\Delta}_j(\theta', z))\|_{\mathcal{H}}^2 \\ &= \|j^{-1} g_j(A^\top A) A^\top A S(\theta - \theta')\|_{\mathcal{H}}^2 \le j^{-2}\|\theta - \theta'\|^2, \\ \|\widehat{\Delta}_j(\theta, z) - \widehat{\Delta}_j(\theta, z')\|^2 &= \|S \cdot j^{-1} g_j(A^\top A) A^\top (z - z')\|_{\mathcal{H}}^2 \\ &\le j^{-2}\|(A^\top A) g_j(A^\top A)\|_{op}^2 \|(AA^\top)^{(\alpha-1)/2}(z - z')\|_{\mathcal{Z}}^2 \le j^{-2}\|z - z'\|_z^2.\end{aligned}$$

   And for its condition (ii),

   $$W_2^2(p_\theta, p_{\theta'}; \|\cdot\|) = \|A\theta - A\theta'\|^2 = \|\theta - \theta'\|^2.$$

3. To verify assumption 3.3 we first prove that

   $$\widehat{\Delta}_j(\bar{\theta}_j^B, z_{j+1}^B) = \Delta_j^B.$$

   This is because there exist independent rvs $e_i \sim \mathcal{N}(0, \sigma^2 I),\ \Delta e_i \sim \mathcal{N}(0, j^{-1} A\hat{\Sigma}_j^{-1} A^\top)$ s.t. for $\bar{e}_i := e_i + \Delta e_i$, we can have

   $$\Delta_j^B = \hat{\Sigma}_j^{-1} A^\top \left(\frac{j-1}{j}\bar{z}_{j-1} + \frac{1}{j}(A\bar{\theta}_j^B + \bar{e}_j)\right) - \hat{\Sigma}_{j-1}^{-1} A^\top \bar{z}_{j-1} = j^{-1}\hat{\Sigma}_j^{-1} A^\top \bar{e}_j = \widehat{\Delta}_j(\bar{\theta}_j^B, z_{j+1}^B).$$

   Since we also have $\breve{\theta}_n = \bar{\theta}_n^B$, it follows by induction that $\breve{\theta}_j = \bar{\theta}_j^B$ for all $j \ge n$. Thus, $\breve{\varepsilon}_{ex,j} \equiv 0$, and the assumption holds for $\nu_l \equiv 0$.
4. Assumption 3.4 holds for $C_{\mathcal{A}} = 0, C'_{\mathcal{A}} = 1$ and $\eta_j = j^{-1}$ because

   $$\mathbb{E}_{z' \sim \mathbb{P}_{\theta'}} \widehat{\Delta}_j(\theta, z') = j^{-1} \underbrace{g_j(A^\top A) A^\top A}_{=: H_{\theta,j}}(\theta' - \theta).$$

5. Assumption 3.5 holds when $\alpha = 1$ since $\bar{\varepsilon}_{B,j}^2 \asymp j^{-1+1/2\beta}$. It also holds for a range of α depending on the value of β.

(Non-asymptotic) connections to GP regression. Consider a GP model with input space $\mathcal{X}$, prior $\pi_{gp} = \mathcal{GP}(0, k)$ and likelihood $p(y \mid f(x)) = \mathcal{N}(f(x), 1)$. Let $\bar{\mathcal{H}}$ be the reproducing kernel Hilbert space (RKHS) defined by k, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the training data, and $K := (k(x_i, x_j))_{ij} \in \mathbb{R}^{n\times n}$ be the Gram matrix. Introduce the notations $f(X) := (f(x_1); \ldots; f(x_n)) \in \mathbb{R}^n$ and $Y := (y_1; \ldots; y_n) \in \mathbb{R}^n$. Let $\mathcal{H} \subset \bar{\mathcal{H}}$ be the subspace spanned by $\{k(x_i, \cdot)\}_{i=1}^n$ with the inherited norm. Then we can identify the projection of any $f \in \bar{\mathcal{H}}$ onto $\mathcal{H}$ with $f(X)$, and its norm satisfies $\|f(X)\|_{\mathcal{H}}^2 = f(X)^\top K^{-1} f(X)$. Let $\mathcal{Z} = \mathbb{R}^n$ be equipped with the Euclidean norm. We substitute the remaining quantities in section 3.2.2 as follows:

$$\theta = f(X), \quad A\theta = \frac{1}{\sqrt{n}} f(X), \quad \frac{1}{n}\sum_{i=1}^n z_i = \frac{1}{\sqrt{n}} Y.$$

Then it is clear that θ follows the prior π and the conditional distribution $\frac{1}{n}\sum_{i=1}^n z_i \mid \theta$ equals that defined by the likelihood in section 3.2.2, and we can readily verify that the posterior in Sec. 3.2.2 for $\theta = f(X)$ equals the GP marginal posterior. Following section 3.2.2, we can consider an MP defined by (9) and $\widehat{z}_j \sim \mathcal{N}(\widehat{\theta}_j, n^{-1} I)$, which provides a high-quality approximation to the GP marginal posterior.

As noted above, on $\{z_j\}$ sampled from the prior predictive distribution (9) has a behaviour equivalent to sequential posterior mean estimation which, for linear-Gaussian Bayesian models, is equivalent to sequential maximum-a-posteriori (MAP) estimation. Based on the same idea of sequential MAP estimation we can derive the update rule (10) for GP regression. Note that (10) and (9) are not an exact match because the GP MAP also depends on the sampled $\hat{x}_j$. (If we continue the analogy above, (10) can be viewed as an MAP in a Bayesian model where we impute at all n input locations simultaneously in each iteration, and scale the resulted log likelihood by $1/\sqrt{n}$.) Nonetheless, we expect their behaviour to be similar. A separate analysis for (10) may be possible, but we forego this discussion given the rich literature on GP inference. Instead, we refer readers to Appendix D.1 for an empirical evaluation for (10).

Remark A.1. The above discussion restricted to the marginal posterior $f(X) \mid (X, Y)$ and does not cover predictive uncertainty in out-of-distribution (OOD) regions. We note that for models that define continuous prediction functions, the uncertainty for $f(X)$ always translates to some uncertainty in OOD regions due to the continuity constraint; the MP will also provides additional uncertainty if we sample $\hat{x}_j$ from the OOD regions. However, an equally important source of OOD uncertainty is from the model's *initialisation randomness*, which can be fully characterised in the GP example above.

To see this, consider an MP defined by (10) and the choice of $\hat{x}_{j+1} \sim \mathrm{Unif}\{x_{1:n}, \hat{x}_{n+1:j}\}$. We claim that the resulted algorithm will fully retain the initialisation randomness for uncertainty in OOD regions. Formally, for any $f \in \bar{\mathcal{H}}$, or an interpolating RKHS which cover all GP samples (Steinwart, 2019), and any $x_* \in \mathcal{X}$, we can decompose $f(x_*) = f_{\|}(x_*) + f_{\perp}(x_*)$ by projecting $f =: f_{\|} + f_{\perp}$ into $\mathcal{H}$ and its orthogonal complement. Then the GP posterior for $f_{\|}$ and $f_{\perp}$ are then independent, and the latter is equivalent to the prior; this is because the likelihood is independent of $f_{\perp}$. The MP update admits a similar factorisation for the same reason, and thus any initialisation randomness will be retained in the MP, and an exact match to the GP posterior can be possible if we initialise based on the GP prior.

B Implementation Details for Algorithm 1

Choices of Δn and N. If the base algorithm is "correctly specified" for the problem as hypothesised, we should ideally choose Δn and N to match the exact martingale posterior ($\Delta n = 1, N \to \infty$) as close as possible, but computational constraints may prevent an exact match. A larger Δn or a smaller N generally leads to an underestimation of uncertainty.

We note that no adjustment is needed if, as in many applications, the goal is merely to improve predictive performance by better accounting for epistemic uncertainty, since the algorithm can still account for a substantial proportion of the uncertainty; and similar underestimation issues may also emerge in the applications of approximate Bayesian inference to complex models, when due to computational constraints we cannot recover the exact posterior. Nonetheless, for the construction of credible sets, we provide a rule of thumb to compensate for this effect by analysing simplified settings. Specifically, consider the natural GD algorithm

$$\widehat{\theta}_{j+1} := \widehat{\theta}_j + (j+1)^{-1} F_{\widehat{\theta}_j}^{-1} \nabla_\theta \log p_{\widehat{\theta}_j}(\widehat{z}_{j+1}), \tag{27}$$

where F_θ denotes the Fisher information matrix. Suppose $n/\Delta_n \in \mathbb{N}$ for simplicity, then the covariance of the parameter ensemble from Algorithm 1 is

$$\sum_{j'=n/\Delta_n}^{\infty} \frac{\Delta_n}{((j'+1)\Delta_n)^2} F_{\widehat{\theta}_j}^{-1} \approx \sum_{j'=n/\Delta_n}^{\infty} \frac{\Delta_n}{((j'+1)\Delta_n)^2} F_{\theta_0}^{-1} \sim \left(\frac{1}{n+\Delta_n} - \frac{1}{N+\Delta_n}\right) F_{\theta_0}^{-1}. \tag{28}$$

The exact MP has covariance $\sim n^{-1} F_{\theta_0}^{-1}$, so to match the exact MP it suffices to inflate the covariance by a factor $\sim \frac{\Delta n}{n} + \frac{n}{N}$. The same inflation applies to credible sets for linear functionals of the parameter which, for linear-in-parameter regression models, include pointwise credible intervals for the true regression function. Note that the same adjustment applies to any GD algorithms with a step-size of $\eta_j \sim j^{-1}$, which is generally related with sequential ERM algorithms (and thus Alg. 1) as shown in Section 3.2. And the above discussion is relevant in a deep learning context if we consider ultrawide NNs (Lee et al., 2019).

In reality, we expect the adjustment to produce conservative credible sets for NN-based algorithms, since it also (unnecessarily) inflates the initialisation randomness. However, the scale of the adjustment is generally small, and together with the unadjusted credible sets they can provide a two-sided bound for the predictive uncertainty.

In our experiments we adopt $N \asymp n \asymp \Delta n$ where the ratios $(N/n, n/\Delta n)$ are in the range of $[1, 10]$, and determine the adjustment scale by explicitly numerical approximation of the ratio between the coefficient of (28) and n^{-1}. For base algorithms that are potentially misspecified we determine the ratio through cross validation.

Early stopping for NN-based algorithms. While the objective (11) always prevent overfitting to past samples, we still need to determine the number of optimisation iterations for the new samples $\widehat{z}_{n_j:n_j+\Delta n}$. In our experiments we use a simple strategy: we use a validation set to determine the number of iterations L for estimation on the n real samples, and optimise for $L\Delta n/n$ iterations when "finetuning" on (each group of) Δn synthetic samples. Other optimisation hyperparameters are also kept consistent across the initial estimation and finetuning.

C Related Work

Our work is motivated by challenges of designing and implementing Bayesian counterparts for ML methods. As discussed in Section 2, NN methods may constitute an important example, due in part to the challenges in inference and prior specification. Another issue is the choice of likelihood: applications in computer vision and natural language processing often involve loss functions that do not have a likelihood interpretation (Lin et al., 2017; Li et al., 2019), and even when a likelihood-based objective leads to efficient point predictors, its suitability for Bayesian NNs can still be debatable if the application involves human-annotated datasets (Aitchison, 2020a) or data augmentation (Nabarro et al., 2022).[3] Compared with the general success of non-Bayesian deep learning approaches, these issues indicate that in typical deep learning applications, it is often easier to express the "prior knowledge" about what method is best suited for a given problem through algorithms, rather than through explicitly defined Bayesian models.

Our work provides an efficient ensemble method for uncertainty quantification. Many ensemble methods have been proposed for NN models (see e.g., Liu and Wang, 2016; Lakshminarayanan et al., 2017; Osband et al., 2018; Wang et al., 2018; D'Angelo and Fortuin, 2021, to name a few). Our method stands out for its applicability beyond NN models, while it also retains advantages over the bootstrap aggregation method—known for a similar trait—by more effectively leveraging the parametric model when it is available (Example 4.1). Restricting to NN models, however, our method could be combined with the low-rank ensemble methods (Wen et al., 2020; Dusenberry et al., 2020) to improve its scalability; it may also be interesting to analyse an "ensemble of ensembles", as in Yao et al. (2022), which may also be beneficial in our setting.

The GP example in Section 3.2.2 is connected to the ensemble algorithms in Osband et al. (2018); Pearce et al. (2020); He et al. (2020), which are designed for DNNs but motivated from the same

[3] See also the works of Wenzel et al. (2020); Izmailov et al. (2021) who reported performance issues with Bayesian NNs (with Gaussian priors) in the presence of data augmentation.

GP regression setting. As observed in He et al. (2020), the GP example is relevant in a deep learning context given the connection between ultrawide NNs and GPs (Lee et al., 2019). While GP regression serves as an interesting motivating example, the ultrawide NNs in that literature represent an oversimplified model (Chizat et al., 2019) and should not be viewed as a "correct prior" for NNs (Aitchison, 2020b). Yet to ensure a close match to the GP posterior, those ensemble methods involve design choices that may not be generally beneficial, such as an ℓ_2 regularisation with a fixed n^{-1} scaling. Our method is motivated from a more general perspective, but we also compare with He et al. (2020) empirically. We also note that the specific problem of (conjugate) GP inference is by now well-understood; there exist algorithms with good statistical and computational guarantees (Burt et al., 2019; Nieman et al., 2022).

From a theoretical perspective, our result is related to the work of Efron (2012) who connected parametric bootstrap to a specific Bayesian posterior defined by the Jeffreys prior (Jeffreys, 1939). However, the Jeffreys prior has counterintuitive behaviours for multidimensional ($d > 1$) models (see e.g., Syversveen, 1998) and cannot be defined for infinite-dimensional models such as our Section 3.2.2. There is also a literature on statistical inference with bootstrap resampling and gradient descent methods (see Lam and Wang 2023 and references therein), which studies similar but different algorithms to the example (3). Such works have the different goal of recovering the sampling distribution for regular parametric models ($d < \infty$ does not grow w.r.t. n), which is not relevant beyond that setting (see Appendix A.2.2). Our requirement on the "excess error" $\breve{\varepsilon}_{ex,j}$, which is defined using a Hilbert norm, may be generalisable to a condition on the mutual information; this is somewhat reminiscent of the development in Xu and Raginsky (2022). However, this is not straightforward as the martingale condition still requires a norm.

D Experiment Details and Full Results

D.1 Toy Experiment: Gaussian Process Regression

We first illustrate the proposed method on a 1-dimensional GP regression task, to understand its behaviour and complement the GP discussion in Section 3.2.2.

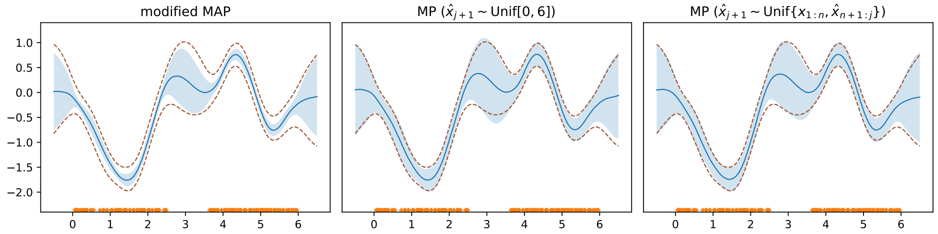

Figure 1: GP example: visualisation of the approximate MP defined by Eq. (10), compared with the ensemble predictors defined by a modified MAP estimator with similar initialisation randomness (Eq. (29)). Solid line and shade indicate the mean estimate and 80% pointwise credible intervals (CIs) for the true regression function. Dashed line indicates the 80% CIs from the exact posterior. Dots at bottom indicate the location of training inputs.

Experiment setup. We instantiate Algorithm 1 using (10) as the estimation algorithm, with random Fourier approximation for the RKHS. We adopt the Snelson dataset (Snelson, 2008) and remove the samples with input within the $[0.4, 0.6]$ quantile to create an out-of-distribution region for visualisation. We adopt a Matérn-3/2 kernel with bandwidth 1 approximated with 200 random Fourier features, and specify a Gaussian likelihood with variance $\sigma^2 = 0.64$. We set $N = 6n, \Delta n = 0.1n$ in Algorithm 1, and consider two choices for $\hat{x}_j$: (i) uniform sampling from $[0, 6]$, and (ii) nonparametric resampling as in Remark A.1. We compare with an ensemble of the following modified MAP predictor:

$$\hat{f}_n := \arg\min_f \sum_{i=1}^n (f(x_i) - y_i)^2 + \frac{\sigma^2}{n} \|f - \tilde{f}_0\|^2_{\mathcal{H}}, \quad \text{where } \tilde{f}_0 \sim \mathcal{GP}(0, k_x), \tag{29}$$

and k_x denotes the Matérn kernel. The random $\tilde{f}_0$ provides a source of initialisation randomness which is also needed for the MP to match the exact Bayesian posterior in out-of-distribution regions

(Remark A.1). More broadly, (29) is analogous to the deep ensemble method (Lakshminarayanan et al., 2017) in which the predictive uncertainty is also derived solely from initialisation randomness, and the comparison between (10) and (29) may provide insights for the more general case.

Results and discussion. Figure 1 visualises the predictive uncertainty from the MP, the modified MAP ensemble, and the exact posterior. We can see that the MP produces a close match to the GP posterior, as expected in Section 3.2.2; and the results are highly consistent across the two choices of samplers for $\hat{x}_j$. In contrast, (29) underestimates uncertainty, especially in in-distribution regions. While conjugate GP inference is a well-studied problem, the above result suggests that in more general scenarios, the uncertainty derived from our method may also have a more desirable behaviour than that from methods relying solely on initialisation randomness. We will observe such results in the DNN experiments in Appendix D.4.

D.2 Hyperparameter learning for Gaussian processes

Setup details. To implement Algorithm 1, we sample $\hat{x}_{n+i}$ from a kernel density estimate and $\hat{y}_{n+i} \mid \hat{x}_{n+i}$ from the GP's marginal predictive distribution, and use $\Delta n = 0.25n, N = 4n$. In preliminary experiments we find that a larger choice of N or a smaller choice of Δn appears to lead to diminishing improvements for performance; thus we adopt this choice for simplicity. For all methods, we implement the base empirical Bayes algorithm with the L-BFGS-B optimiser (Zhu et al., 1997) using a step-size of 0.05 and 1600 iterations, and build an ensemble of $K = 16$ predictors.

The hyperparameter learning process has a high variation across randomly sampled training sets due to the small sample sizes. Therefore, we use Wilcoxon signed-rank tests to check for statistically significant improvement, and account for ties in computing the ranks for Table 1, by defining the rank of each method as the number of methods that significantly outperform it as determined by the Wilcoxon test.

Full results and discussion. Full results are shown in Table 5. As we can see, our method consistently improves upon the EB baseline and is competitive against the other ensemble approaches. Nonparametric bootstrap also demonstrates competitive performance with $n = 75$, but generally underperforms the EB baseline when $n = 300$. It is possible that the distribution of parameter estimates from bootstrap has a very high variation, which may be only beneficial when overfitting is severe. We note that the performance difference is often small compared to the standard deviation, but the improvement over baselines is consistent as evidenced by the Wilcoxon test.

D.3 Classification with boosting tree and stacking algorithms

Deferred setup details. We evaluate on the 30 datasets from the OpenML CC18 benchmark (Bischl et al., 2017) with $n \leq 2000, \dim x \leq 100, \dim y \leq 10$. In all experiments we adopt a 60-20-20 split for train/validation/test, and determine the hyperparameters for the base algorithm using the log loss on validation set. We implement our method by refitting a predictor from scratch at each iteration; in other words, in Algorithm 1 we define both $\mathcal{A}_0(D_{j+1}; \hat{\theta}_j)$ and $\mathcal{A}_0(D_n)$ as the predictor resulted by applying the base algorithm to the respective dataset.

For the GDBT algorithm, we adopt the implementation from XGBoost and conduct search for the following hyperparameters: tree depth $D \in \{4, 5, 6, 7\}$, number of boosting iterations $L \in \{50, 100, 200\}$ and learning rate $\eta \in \{10, 30, 100\}/L$. We also conduct early stopping using the validation set with a tolerance of 10 rounds. For the instantiations of our method and bagging, we build an ensemble of 50 predictors; for our method, we determine $\Delta n \in \{0.125n, 0.25n, n\}, N \in \{n, 3n\}$ based on the same validation loss.

For stacking, we use the default implementation in AutoGluon (`TabularPredictor(eval_metric="log_loss") .fit`), which determines the hyperparameters for the individual models based on pre-defined rules and uses the validation set to estimate a linear stacking model following Caruana et al. (2004). As the stacking algorithm is more computation intensive, we build an ensemble of 20 predictors for our method and bagging, and set $\Delta n = N = n$ for our method.

Table 5: Full results for the GP experiment: mean and standard deviation for all test metrics. Boldface indicates the best result ($p < 0.05$ in a Wilcoxon signed-rank test).

Dataset	RMSE				NLPD				CRPS			
	Emp. Bayes	Bootstrap	Ensemble	Proposed	Emp. Bayes	Bootstrap	Ensemble	Proposed	Emp. Bayes	Bootstrap	Ensemble	Proposed
$n = 75$												
Boston	4.47 ±0.93	**4.39** ±0.79	4.53 ±0.89	4.49 ±0.86	3.28 ±0.42	**2.72** ±0.14	3.19 ±0.38	3.17 ±0.37	2.31 ±0.38	**2.16** ±0.27	2.30 ±0.34	2.28 ±0.33
Concrete	8.16 ±0.94	8.21 ±0.79	8.16 ±0.89	**8.10** ±0.86	3.66 ±14.40	**3.47** ±0.08	3.55 ±3.42	3.54 ±1.69	4.38 ±0.50	**4.44** ±0.38	4.38 ±0.49	**4.31** ±0.47
Energy	1.27 ±0.28	1.47 ±0.19	1.27 ±0.27	**1.27** ±0.25	1.26 ±0.28	1.60 ±0.14	**1.25** ±0.25	**1.24** ±0.24	0.55 ±0.12	0.73 ±0.09	0.55 ±0.12	**0.55** ±0.10
Kin8nm	**0.19** ±0.02	0.19 ±0.01	**0.19** ±0.02	**0.19** ±0.02	**−0.23** ±0.14	**−0.23** ±0.06	**−0.23** ±0.12	**−0.22** ±0.13	**0.11** ±0.01	0.11 ±0.01	**0.11** ±0.01	**0.11** ±0.01
Naval	0.01 ±0.00	0.01 ±0.00	0.00 ±0.00	**0.00** ±0.00	**−5.05** ±0.12	−4.06 ±0.12	−4.99 ±0.11	**−5.03** ±0.13	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	**0.00** ±0.00
Power	4.54 ±0.22	5.07 ±0.42	4.54 ±0.22	**4.54** ±0.19	2.94 ±0.05	3.10 ±0.07	2.94 ±0.05	**2.94** ±0.04	2.50 ±0.12	2.86 ±0.22	2.50 ±0.12	**2.49** ±0.10
Protein	6.03 ±0.35	**5.76** ±0.14	5.92 ±0.32	5.92 ±0.32	3.36 ±0.35	**3.17** ±0.05	3.22 ±0.22	3.22 ±0.22	3.50 ±0.23	**3.31** ±0.09	3.38 ±0.21	3.37 ±0.21
Winered	0.76 ±0.04	**0.71** ±0.03	0.75 ±0.04	0.74 ±0.04	1.30 ±2.25	**1.08** ±0.07	1.19 ±0.27	1.18 ±0.25	0.43 ±0.03	**0.39** ±0.02	0.42 ±0.03	0.42 ±0.03
Winewhite	0.87 ±0.04	**0.81** ±0.03	0.85 ±0.04	0.84 ±0.05	1.49 ±0.28	**1.22** ±0.06	1.40 ±0.26	1.36 ±0.28	0.50 ±0.03	**0.45** ±0.02	0.48 ±0.03	0.48 ±0.03
$n = 300$												
Boston	3.22 ±0.52	**3.19** ±0.43	3.19 ±0.50	**3.17** ±0.48	2.55 ±0.16	**2.42** ±0.09	2.54 ±0.16	2.52 ±0.14	1.61 ±0.18	1.62 ±0.13	1.60 ±0.17	**1.58** ±0.16
Concrete	6.51 ±0.41	6.77 ±0.61	6.51 ±0.41	**6.47** ±0.42	3.24 ±0.13	**3.22** ±11.24	3.24 ±0.13	**3.22** ±0.12	3.42 ±0.21	3.53 ±0.29	3.42 ±0.21	**3.40** ±0.21
Energy	0.60 ±0.14	0.69 ±0.13	**0.58** ±0.15	**0.57** ±0.15	0.77 ±0.12	0.90 ±0.07	**0.71** ±0.10	**0.70** ±0.11	0.29 ±0.03	0.34 ±0.03	0.28 ±0.04	**0.28** ±0.04
Kin8nm	**0.12** ±0.00	0.13 ±0.00	**0.12** ±0.00	**0.12** ±0.00	**−0.69** ±0.03	−0.62 ±0.02	**−0.69** ±0.03	**−0.69** ±0.03	**0.07** ±0.00	0.07 ±0.00	**0.07** ±0.00	**0.07** ±0.00
Naval	0.00 ±0.00	0.00 ±0.00	**0.00** ±0.00	**0.00** ±0.00	−7.00 ±0.04	−6.49 ±0.04	−7.01 ±0.04	**−7.01** ±0.04	**0.00** ±0.00	0.00 ±0.00	**0.00** ±0.00	**0.00** ±0.00
Power	4.31 ±0.10	4.73 ±0.17	4.31 ±0.10	**4.30** ±0.10	2.88 ±0.03	3.04 ±0.04	2.88 ±0.03	**2.88** ±0.03	2.36 ±0.04	2.68 ±0.09	2.36 ±0.04	**2.36** ±0.04
Protein	5.18 ±0.16	5.36 ±0.10	5.15 ±0.14	**5.14** ±0.14	3.07 ±0.03	**3.07** ±0.02	3.06 ±0.04	**3.06** ±0.04	2.93 ±0.08	3.02 ±0.07	2.92 ±0.07	**2.91** ±0.07
Winered	0.71 ±0.04	**0.67** ±0.03	0.70 ±0.05	**0.69** ±0.05	**0.87** ±0.15	0.98 ±0.05	0.94 ±0.12	**0.93** ±0.11	0.38 ±0.02	**0.37** ±0.02	0.37 ±0.03	**0.37** ±0.03
Winewhite	0.79 ±0.02	**0.76** ±0.02	0.78 ±0.03	0.78 ±0.03	1.13 ±0.04	1.12 ±0.03	1.10 ±0.04	**1.09** ±0.04	0.44 ±0.01	**0.42** ±0.01	0.43 ±0.01	0.43 ±0.01

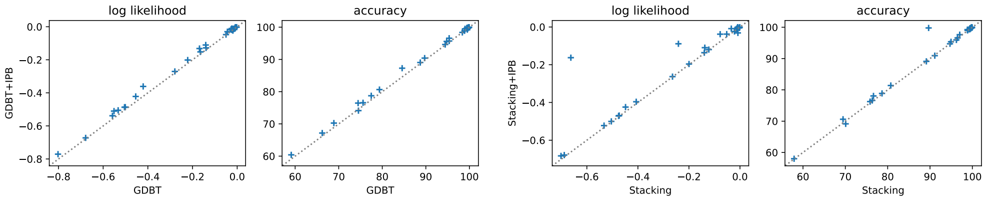

Figure 2: Classification experiment: scatter plot of the test metrics (for each dataset averaged over 10 random splits; higher is better) for the base algorithm vs the proposed method.

Additional results. Table 6–7 report the full test metrics on all 30 datasets; for each baseline method we further conducts a Wilcoxon test to compare its distribution of loss metrics (for each dataset, averaged over 10 random splits) against that of the proposed method, and report the p-value in the respective table. As we can see, except for the test accuracy of the stacking+bagging baseline, our method always leads to a statistically significant improvement ($p < 0.05$).

We note that for stacking, the AutoGluon library recommends a more sophisticated multi-level algorithm (corresponding to `.fit(presets="best_quality")`) for the best predictive performance. We evaluated that algorithm under identical conditions, and found it to perform better than our chosen base algorithm but worse than bagging and our method applied to the latter (average accuracy 91.1%, NLL 0.198 in the setting of Table 2). As the algorithm also has a significantly higher computational cost, we refrain from testing our method with it, although we expect a similar improvement in performance if our method were applied.

Figure 3 visualises the uncertainty estimates for the information gain-based feature importance scores, obtained using our method on the UCI adult dataset. As we can see, the correlation structure of the approximate MP is informative about feature dependencies; for example, the strong negative correlation between "marital status" and "relationship" indicates that these two features are interchangeable for prediction.

D.4 Interventional density estimation

Setup details. For the base estimation algorithm, we adopt a fully-connected NN model with 128 hidden units in each layer, and determine the other hyperparameters in the following range: (i) number of hidden layers $D \in \{2, 3, 4\}$, (ii) learning rate $\eta \in \{0.1, 0.5, 1, 5\} \times 10^{-3}$, (iii) training iterations $L \in \{2, 4, 8\} \times 1000$, and (iv) activation function from {swish, selu, tanh}. The hyperparameters are determined by evaluating the training objective on an in-distribution validation set, on the `chain-na` dataset from Chao et al. (2023). We use the AdamW optimiser (Loshchilov and Hutter, 2019) with default hyperparameters in Optax (DeepMind, 2020). For our method, we instantiate the proximal Bregman objective (11) using the weighted score matching loss in Ho et al. (2020), and set $\Delta n = 0.1n, N = 6n$: beyond this range, a larger value of N leads to diminishing improvement, and the results appear somewhat insensitive to the choice of Δn. Other implementation details are discussed in Appendix B.

On the synthetic datasets, we consider two evaluation setups:

- Following Chao et al. (2023) we evaluate distributional estimates for $\mathbb{P}(x_{\mathrm{desc}(i)} \mid \mathrm{do}(x_i = x))$, where $\mathrm{desc}(i)$ denotes the descendents of node i in the causal graph and x ranges over a uniform grid of the $[0.1, 0.9]$ quantile. We report the maximum mean discrepancy for in this setup.
- We present a more direct evaluation of the uncertainty estimates, by evaluating the average coverage of pointwise credible intervals for the mean outcome $\mathbb{E}(x_d \mid \mathrm{do}(x_{1:d-1} = \cdot))$ and the L_2 distance between the estimated CDF and ground truth. The latter is equivalent to CRPS and is thus a meaningful surrogate for forecasting error. The value for $x_{1:d-1}$ is determined by varying one of the variables on a uniform grid and fixing the others to $\{-0.5, 0, 0.5\}$, consecutively.

On the fMRI dataset, we report the median of absolute error following Khemakhem et al. (2021); Chao et al. (2023) and the CRPS. Our setup, where we average over random seeds (which determine

Table 6: Classification experiment: average negative log likelihood across random train/test splits in each dataset.

Dataset	GBDT			Stacking		
	(Base)	+ BS	+ IPB	(Base)	+ BS	+ IPB
banknote-authentication	.002±.00	.003±.00	.003±.00	.009±.01	.002±.00	.001±.00
blood-transfusion-service-center	.504±.03	.486±.02	.487±.02	.473±.02	.470±.02	.469±.03
breast-w	.139±.02	.129±.02	.128±.02	.138±.03	.103±.01	.110±.02
mfeat-karhunen	.012±.00	.022±.00	.012±.00	.008±.01	.086±.04	.031±.03
mfeat-morphological	.018±.01	.021±.00	.014±.00	.014±.01	.030±.02	.009±.00
eucalyptus	.802±.04	.786±.03	.771±.03	.689±.05	.704±.04	.679±.05
mfeat-zernike	.017±.01	.021±.00	.012±.00	.241±.43	.059±.03	.089±.13
cmc	.028±.01	.018±.00	.016±.00	.019±.01	.022±.01	.020±.01
credit-approval	.169±.03	.159±.02	.132±.02	.122±.03	.125±.02	.120±.03
vowel	.533±.02	.506±.02	.505±.02	.504±.03	.501±.02	.500±.02
credit-g	.011±.00	.018±.00	.010±.00	.003±.00	.004±.00	.005±.00
analcatdata_authorship	.044±.03	.045±.02	.030±.01	.052±.04	.029±.01	.038±.02
balance-scale	.421±.06	.362±.03	.361±.03	.663±.64	.137±.04	.163±.05
analcatdata_dmft	.501±.02	.490±.01	.487±.01	.476±.02	.472±.01	.471±.02
diabetes	.222±.02	.205±.01	.201±.01	.200±.01	.200±.01	.196±.01
pc4	.279±.02	.270±.02	.270±.02	.264±.02	.264±.02	.263±.02
pc3	.019±.01	.022±.01	.024±.01	.078±.10	.026±.01	.038±.02
kc2	.016±.01	.014±.01	.014±.01	.021±.02	.021±.01	.024±.02
pc1	.009±.01	.003±.00	.003±.00	.001±.00	.001±.00	.001±.00
tic-tac-toe	.551±.03	.536±.02	.510±.02	.448±.02	.450±.02	.424±.02
vehicle	.141±.03	.117±.03	.110±.03	.119±.05	.094±.03	.095±.04
wdbc	.025±.02	.015±.01	.011±.00	.034±.03	.027±.02	.009±.01
qsar-biodeg	.558±.02	.543±.01	.538±.01	.533±.02	.524±.01	.523±.01
dresses-sales	.678±.01	.672±.01	.672±.01	.701±.02	.683±.02	.683±.02
mfeat-fourier	.025±.01	.027±.00	.019±.00	.010±.01	.031±.02	.010±.00
MiceProtein	.023±.01	.025±.00	.011±.00	.008±.01	.020±.01	.002±.00
steel-plates-fault	.021±.01	.024±.01	.016±.00	.010±.01	.020±.01	.005±.00
climate-model-simulation-crashes	.165±.04	.158±.03	.152±.03	.140±.03	.139±.02	.136±.03
car	.050±.01	.072±.01	.048±.01	.028±.01	.047±.01	.025±.01
cylinder-bands	.454±.05	.429±.02	.422±.03	.407±.05	.407±.03	.396±.04
Wilcoxon p-value vs IPB	3.1e-08	6e-07	-	2.2e-06	0.029	-

Table 7: Classification experiment: average test accuracy across random train/test splits in each dataset.

Dataset	GBDT			Stacking		
	(Base)	+ BS	+ IPB	(Base)	+ BS	+ IPB
banknote-authentication	99.9±0.1	99.9±0.1	100.0±0.0	99.9±0.1	100.0±0.1	100.0±0.0
blood-transfusion-service-center	77.5±2.1	79.0±1.6	78.7±1.7	78.7±1.1	78.7±1.2	78.9±1.6
breast-w	95.4±0.6	95.9±0.9	95.7±0.6	96.5±0.7	96.6±0.4	96.6±0.5
mfeat-karhunen	99.9±0.1	99.8±0.1	99.9±0.1	99.8±0.1	99.9±0.1	100.0±0.1
mfeat-morphological	99.4±0.4	99.6±0.2	99.8±0.2	99.6±0.2	99.7±0.2	99.8±0.2
eucalyptus	66.2±2.2	65.9±2.0	67.2±2.0	69.5±2.9	69.1±2.9	70.6±2.4
mfeat-zernike	99.7±0.2	99.7±0.2	99.8±0.2	89.7±18.5	99.9±0.1	99.8±0.1
cmc	99.0±0.3	99.5±0.1	99.5±0.2	99.5±0.2	99.6±0.2	99.4±0.2
credit-approval	94.8±0.8	95.0±0.4	95.6±0.8	96.2±1.1	95.7±0.7	95.9±1.1
vowel	74.5±2.1	75.5±1.7	76.5±1.9	75.8±1.8	75.0±1.8	76.2±2.0
credit-g	99.8±0.2	99.8±0.2	99.9±0.1	99.9±0.1	99.9±0.1	99.9±0.1
analcatdata_authorship	98.9±0.6	98.7±0.7	98.9±0.6	98.9±0.6	99.1±0.4	99.0±0.5
balance-scale	84.6±1.9	89.2±1.9	87.3±1.7	95.0±1.2	94.8±0.9	95.4±0.9
analcatdata_dmft	75.6±1.9	75.6±2.3	76.6±2.5	76.4±1.5	76.6±2.1	76.6±1.6
diabetes	89.7±1.0	90.4±0.9	90.4±1.0	91.1±1.0	90.8±1.0	90.9±0.9
pc4	88.7±1.2	89.2±1.1	89.0±1.2	89.1±1.1	89.1±1.0	89.1±1.0
pc3	99.5±0.3	99.6±0.3	99.2±0.6	99.2±0.6	99.4±0.5	99.3±0.5
kc2	99.6±0.2	99.6±0.3	99.6±0.2	99.6±0.3	99.7±0.2	99.6±0.2
pc1	99.9±0.2	99.9±0.2	99.9±0.2	99.9±0.1	100.0±0.0	99.9±0.1
tic-tac-toe	74.5±1.7	74.0±1.5	74.1±1.3	76.6±1.3	77.1±0.9	78.1±1.2
vehicle	95.4±1.3	95.7±1.2	96.6±1.1	97.0±0.7	97.3±0.6	97.6±0.5
wdbc	99.5±0.3	99.5±0.3	99.7±0.2	99.4±0.3	99.8±0.2	99.7±0.2
qsar-biodeg	68.9±1.4	69.3±1.6	70.3±1.5	70.1±1.5	69.8±1.5	69.1±1.6
dresses-sales	59.1±1.7	60.4±2.0	60.4±1.9	57.9±2.7	57.8±2.6	58.0±2.9
mfeat-fourier	99.5±0.2	99.6±0.2	99.7±0.2	99.6±0.1	99.8±0.1	99.7±0.1
MiceProtein	99.7±0.2	99.5±0.3	99.9±0.1	99.8±0.1	100.0±0.0	100.0±0.0
steel-plates-fault	99.5±0.2	99.7±0.2	99.8±0.1	99.8±0.1	99.9±0.1	99.9±0.1
climate-model-simulation-crashes	94.4±1.5	94.3±1.4	94.6±1.3	94.7±1.4	94.4±1.6	94.7±1.4
car	98.4±0.4	97.6±0.4	98.4±0.3	98.8±0.5	98.3±0.5	99.2±0.4
cylinder-bands	79.4±2.3	79.7±2.4	80.6±1.7	80.7±1.6	81.1±2.0	81.4±1.4
Wilcoxon p-value vs IPB	5e-05	0.0047	-	0.0011	0.056	-

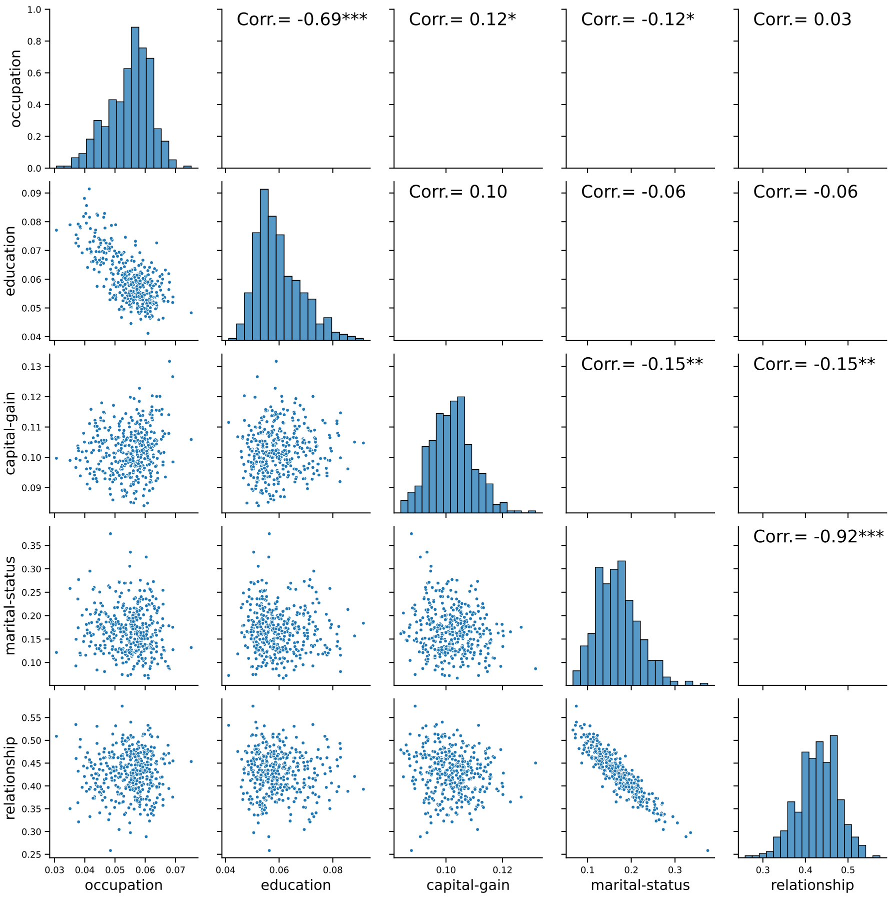

Figure 3: Classification experiment: approximate MP for the GDBT feature importance scores and their pairwise correlations. Plotted are the top 5 features in the UCI adult dataset.

the initialisation and train/validation split), appears different from Khemakhem et al. (2021), and we can exactly match their reported results using a single (default) seed set in their codebase. Nonetheless, the results remain statistically consistent.

Full results and discussion. Full results for the synthetic experiments are shown in Table 8 (in the setting of Table 3 and Chao et al. (2023)) and Table 9–10 (for the evaluation of uncertainty). As we can see, our method attains the best overall performance for both prediction and uncertainty quantification. The vanilla ensemble method achieves the best predictive performance across baselines, which is consistent with previous reports (Fort et al., 2019; Gorishniy et al., 2021). NTKGP is generally uncompetitive; even through the method is applied to the same DNN models, it is possible that the ultrawide NN perspective which motivated their design choices is less applicable to diffusion models which utilise multi-output NNs. The predictive performance of PB is uncompetitive possibly related to its discard of real data. For uncertainty quantification, however, both bootstrap baselines demonstrate better performance than the other baselines, although our method still achieves better performance. Note that due to the distribution shift we cannot expect the coverage of credible intervals to match their exact nominal level.

Table 8: Interventional density estimation: full results in the setting of Table 3. Reported is the estimate and 95% CI for the $100 \times \mathrm{MMD}^2$ metric across 30 trials. Boldface indicates the best result ($p < 0.05$ in a Wilcoxon signed-rank test).

Method	chain-na	chain-nonlin	diamond-na	diamond-nonlin	triangle-na	triangle-nonlin	y-na	y-nonlin
$N = 100$								
PB	31.75±4.25	8.40±1.37	13.84±1.50	18.86±3.90	29.59±4.58	20.77±4.98	10.35±0.92	7.36±0.98
Ens	27.40±3.55	**6.72**±1.02	11.87±1.43	15.40±3.18	25.28±3.85	18.55±4.83	9.42±0.93	**6.54**±0.77
NTKGP	47.80±0.87	11.96±1.43	31.45±1.12	51.96±2.25	38.92±1.67	42.52±2.45	19.97±1.22	22.10±1.63
BS	30.30±3.36	**6.83**±1.05	12.81±1.40	19.88±3.54	28.21±4.81	23.09±5.21	11.54±1.73	6.76±0.78
IPB	**19.94**±2.35	**6.31**±0.87	**8.74**±0.92	**9.64**±1.38	**16.35**±1.42	**10.02**±1.77	**8.14**±0.76	**6.56**±0.93
$N = 1000$								
PB	9.28±0.69	2.63±0.22	3.52±0.32	4.02±0.43	5.98±0.43	3.42±0.27	3.35±0.30	2.62±0.23
Ens	7.45±0.60	**2.42**±0.17	**2.85**±0.22	3.49±0.36	4.84±0.35	3.13±0.21	**2.84**±0.27	2.40±0.17
NTKGP	21.55±0.39	2.83±0.20	8.03±0.20	11.64±0.38	12.42±0.37	6.13±0.25	5.39±0.24	3.85±0.22
BS	8.58±0.68	**2.31**±0.15	3.15±0.28	3.67±0.39	5.80±0.42	3.08±0.26	3.05±0.26	**2.31**±0.13
IPB	**6.25**±0.46	2.58±0.20	**2.78**±0.15	**3.22**±0.37	**4.23**±0.31	**2.79**±0.19	**2.98**±0.21	**2.27**±0.13

Table 9: Interventional density estimation experiment: additional results for quality of uncertainty estimates, when $n = 100$. Reported are the estimate and 95% CI for the mean of each test metric. Boldface indicates the best result ($p < 0.05$ in a Wilcoxon signed-rank test).

Method	chain-na	chain-nonlin	diamond-na	diamond-nonlin	triangle-na	triangle-nonlin	y-na	y-nonlin
CDF L_2								
PB	0.023±0.004	0.008±0.002	0.041±0.005	0.068±0.010	0.073±0.008	0.049±0.010	0.013±0.002	0.012±0.003
Ens	0.019±0.003	**0.007**±0.002	0.035±0.002	0.078±0.011	0.075±0.008	0.049±0.010	0.013±0.002	**0.009**±0.002
NTKGP	0.039±0.002	0.009±0.002	0.053±0.002	0.098±0.005	0.086±0.006	0.082±0.006	0.027±0.002	0.028±0.003
BS	0.022±0.004	**0.006**±0.001	0.037±0.004	0.083±0.011	0.076±0.008	0.058±0.010	0.015±0.003	**0.010**±0.001
IPB	**0.013**±0.002	**0.006**±0.001	**0.028**±0.002	**0.054**±0.010	**0.059**±0.006	**0.035**±0.007	**0.010**±0.001	**0.011**±0.002
Average coverage of 90% CI								
PB	0.960±0.017	0.731±0.109	0.762±0.090	0.637±0.073	0.506±0.065	0.640±0.068	0.750±0.095	0.806±0.088
Ens	0.388±0.101	0.334±0.090	0.231±0.046	0.244±0.043	0.181±0.025	0.271±0.049	0.304±0.082	0.345±0.085
NTKGP	0.388±0.094	0.412±0.095	0.256±0.044	0.152±0.024	0.151±0.015	0.158±0.017	0.182±0.045	0.265±0.075
BS	0.861±0.075	0.806±0.080	0.762±0.081	0.572±0.074	0.511±0.068	0.638±0.075	0.801±0.062	0.798±0.094
IPB	0.966±0.009	0.865±0.048	0.934±0.028	0.915±0.031	0.796±0.047	0.930±0.028	0.833±0.066	0.804±0.070
Average width of 90% CI								
PB	0.216±0.014	0.339±0.033	0.205±0.020	0.382±0.035	0.587±0.080	0.442±0.050	0.431±0.032	0.308±0.016
Ens	0.069±0.007	0.109±0.006	0.049±0.003	0.120±0.012	0.173±0.018	0.138±0.012	0.164±0.010	0.087±0.005
NTKGP	0.117±0.004	0.133±0.002	0.110±0.003	0.173±0.005	0.176±0.010	0.186±0.009	0.199±0.011	0.140±0.005
BS	0.199±0.012	0.339±0.015	0.176±0.013	0.402±0.021	0.615±0.061	0.502±0.029	0.488±0.024	0.323±0.020
IPB	0.168±0.007	0.338±0.012	0.208±0.016	0.768±0.046	1.043±0.126	0.785±0.074	0.459±0.021	0.268±0.009

Table 10: Interventional density estimation experiment: additional results for quality of uncertainty estimates, when $n = 1000$. Reported are the estimate and 95% CI for the mean of each test metric. For CDF L_2, boldface indicates the best result ($p < 0.05$ in a Wilcoxon signed-rank test).

Method	chain-na	chain-nonlin	diamond-na	diamond-nonlin	triangle-na	triangle-nonlin	y-na	y-nonlin
CDF L_2								
PB	0.006±0.001	0.001±0.000	0.017±0.001	0.041±0.005	0.029±0.002	0.010±0.002	0.004±0.000	0.002±0.001
Ens	0.004±0.000	**0.001**±0.000	0.016±0.001	0.043±0.005	**0.026**±0.002	0.011±0.002	0.003±0.000	**0.002**±0.000
NTKGP	0.014±0.000	0.001±0.000	0.027±0.001	0.047±0.004	0.035±0.002	0.016±0.001	0.007±0.000	0.004±0.000
BS	0.005±0.001	**0.001**±0.000	0.017±0.001	0.043±0.006	**0.027**±0.002	0.011±0.001	0.004±0.000	**0.002**±0.000
IPB	**0.004**±0.001	0.001±0.000	**0.014**±0.001	**0.031**±0.005	**0.027**±0.002	**0.008**±0.002	**0.003**±0.000	**0.002**±0.000
Average coverage of 90% CI								
PB	0.870±0.064	0.901±0.054	0.908±0.041	0.746±0.051	0.701±0.052	0.878±0.037	0.958±0.027	0.947±0.032
Ens	0.633±0.085	0.712±0.089	0.522±0.055	0.347±0.061	0.331±0.045	0.408±0.044	0.716±0.062	0.679±0.073
NTKGP	0.654±0.108	0.709±0.086	0.539±0.056	0.254±0.038	0.159±0.020	0.377±0.022	0.683±0.072	0.687±0.072
BS	0.963±0.021	0.989±0.008	0.848±0.049	0.662±0.070	0.624±0.056	0.758±0.044	0.935±0.032	0.937±0.029
IPB	0.927±0.029	0.884±0.042	0.927±0.029	0.838±0.050	0.670±0.048	0.891±0.029	0.876±0.050	0.890±0.044
Average width of 90% CI								
PB	0.090±0.002	0.182±0.005	0.082±0.003	0.217±0.014	0.600±0.041	0.263±0.016	0.265±0.006	0.152±0.004
Ens	0.045±0.001	0.097±0.001	0.032±0.001	0.069±0.002	0.139±0.006	0.073±0.004	0.144±0.003	0.069±0.001
NTKGP	0.060±0.001	0.104±0.001	0.050±0.000	0.081±0.002	0.128±0.004	0.091±0.001	0.150±0.003	0.085±0.001
BS	0.082±0.002	0.159±0.002	0.067±0.001	0.173±0.006	0.434±0.023	0.175±0.004	0.220±0.004	0.126±0.002
IPB	0.072±0.001	0.153±0.002	0.063±0.001	0.230±0.010	0.450±0.021	0.235±0.005	0.219±0.005	0.117±0.002