Bayes in action in deep learning and dictionary learning

Julyan Arbel, Hong-Phuong Dang, Clement Elvira, Cédric Herzet, Zacharie Naulet, Mariia Vladimirova

ESAIM: PROCEEDINGS AND SURVEYS, November 2023, Vol. 74, p. 90-107
Editors: Marguerite Zani, Clément Marteau - Journées MAS 2020, Fédération Siméon Poisson (Orléans-Tours),

BAYES IN ACTION IN DEEP LEARNING AND DICTIONARY LEARNING

Julyan Arbel[1], Hong-Phuong Dang[2], Clement Elvira[3], Cedric Herzet[4], Zacharie Naulet[5] and Mariia Vladimirova[1]

Abstract. This article summarizes some recent works and associated challenges in the field of Bayesian statistics that were presented during the *Journées MAS* 2020. The goal of the session was to give an overview of the many aspects of Bayesian statistics investigated by young researchers of the community.

Résumé. Cet article résume quelques travaux récents et leurs défis dans le domaine de la Statistique Bayésienne. Ces travaux ont été présentés aux *Journées MAS* en 2020. Le but de la session était de donner un aperçu de tous les aspects de la Statistique Bayésienne investigués par de jeunes chercheuses et chercheurs de la communauté.

1. Introduction

On Thursday, 26th of August 2021 the "Bayesian statistics" session was held at the Journées MAS 2020 in Orléans. This article describes some of the works presented during this session.

The Bayesian interpretation of probabilities stipulates that probabilities must be understood as a measure of the degree of belief in the event. The main philosophy of Bayesian inference is based on this idea. In short, a Bayesian statistician starts with an initial belief on the object of interest (the prior belief) and updates this belief on the basis of the observations at hand (the posterior belief). The prior belief may be based on anterior studies, knowledge, or personal belief. This differs from the *frequentist* paradigm, which is based on the other main interpretation of probabilities, in which the probability of an event is viewed as the limiting value of its relative frequency after an infinite repetition of independent trials.

Formally, a Bayesian model is the joint distribution Π of a random variable (X, θ) taking values in $\mathcal{X} \times \Theta$, where X is the observation and $\theta \in \Theta$ the "random" parameter. That is conditional on θ it is assumed that X has law $\mathbb{P}_\theta$. Prior information about θ is incorporated in the model via the marginal distribution $\nu(\cdot) \coloneqq \Pi(\mathcal{X} \times \cdot)$, the so-called *prior distribution*. Having observed $X = x$, Bayesian inference is made on the basis of the law of $\theta \mid X = x$, written here $\Pi(\theta \in \cdot \mid X = x)$, known as the *posterior distribution*. In contrast, a frequentist model is a family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ of distributions over $\mathcal{X}$, with the usual modeling assumption that there exists $\theta_0 \in \Theta$ unknown such that X as law $\mathbb{P}_{\theta_0}$. This last modeling assumption is not required for Bayesian inference, and the most subjective Bayesians reject its validity [1].

Besides the philosophical differences, it is of interest to understand what the Bayesian framework can offer to a pragmatic statistician. Arguably one of the most appealing features of the Bayesian framework is the ability

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[2] LaTIM, INSERM-UMR1101, *Univ. de Bretagne Occidentale*, Brest, & ECAM Rennes, Louis de Broglie, Bruz, France
[3] IETR UMR CNRS 6164, CentraleSupelec Rennes Campus 35576 Cesson Sévigné, France
[4] Inria centre Rennes - Bretagne Atlantique, Rennes, France
[5] Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France.

Article published online by EDP Sciences and available at https://www.esaim-proc.org or https://doi.org/10.1051/proc/202374090

to incorporate prior knowledge in the model in a most natural manner via ν. We can also mention that the approach gives a systematic and natural way to handle uncertainty quantification through credible sets that are sets C_x such that $\Pi(\theta \in C_x \mid X = x) \geq 1 - \alpha$, where $1 - \alpha$ is a credibility level. In nonparametric statistics, it is also often the case that Bayes' estimators are intrinsically adaptive to the smoothness of the true parameter – assuming it exists – while frequentist adaptative procedures can be tricky to build [2, Chapter 10]. Somewhat anecdotally, even the most frequentist statisticians may find some comfort in the use of Bayes' estimators, using them as a powerful tool to construct admissible (even minimax) frequentist procedures [3].

Yet, as appealing as the Bayesian framework can seem, there is still a lot of research going on to understand and address its limitations. We may establish the following taxonomy of fruitful research topics in Bayesian statistics. We believe these are the current main research areas in the community, but we do not exclude the possibility that this list may be biased by the authors' centers of interest.

(A) *Design of prior distributions and complex models with meaningful interpretation.* As many modern applications involve complex data and/or sampling mechanisms, it is often challenging to design interpretable and tractable Bayesian models. We note that this issue is not specific to the Bayesian paradigm. Yet, it is an important field of research in the community and deserved to be mentioned here.

(B) *Efficient posterior computations.* The posterior distribution is rarely available in closed-form, neither numerically computable. The traditional way to overcome this difficulty is to obtain representative samples from the posterior $\Pi(\theta \in \cdot \mid X = x)$. A powerful and popular technique involves building a Markov chain whose stationary distribution is the posterior [4]. However, this method is intrinsically sequential and often does not scale well with a large number of data. Many efforts have been put in recent years to try and solve this issue. Among the most popular solutions are methods based on subsampling the data [5], or building *coresets* [6], or optimizing for the best approximant of the posterior distribution in a given family [7], or splitting the data and perform computation in parallel [8].

(C) *Frequentist validation of Bayes procedures.* Of interest to the pragmatic Bayesian are the frequentist properties of Bayes estimates. That is, under the frequentist modeling assumption that X has law $\mathbb{P}_{\theta_0}$ for a fixed "true" $\theta_0 \in \Theta$, what is the behaviour of $\Pi(\theta \in \cdot \mid X)$ (as a measure-valued random variable). Questions of interest regard its concentration on neighborhoods of θ_0, the rate of this contraction, the limiting distribution shape, etc. Although these questions are rather well understood for parametric models [9, Chapter 10], they are much more challenging in nonparametric models [2].

(D) *Bayesian Machine Learning.* The impressive successes of Machine Learning (ML) algorithms in recent years have attracted the attention of many statisticians. This is especially the case of *learning* algorithms, which consists on learning, either from labeled data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ (*supervised learning* setting), or from unlabeled data $X_1, \ldots, X_n \in \mathcal{X}$ (*unsupervised learning* setting), a function $f : \mathcal{X} \to \mathcal{Y}$ that maps features to labels. The popularity of learning in the statistical community is certainly due to the emergence of Statistical Learning Theory [10, 11], a popular framework for theoretical analysis of learning algorithms. Since the goal of learning is reminiscent to prediction, without surprise many statistical model-based approaches (including Bayesian!) have been proposed over the last decade to design learning algorithms.

The goal of the "Bayesian statistics" session was to present works by young researchers in the field, with the aim to cover as much as possible the whole spectrum of topics described above. Let us briefly summarize the content of the session, listing the talks in their order of appearance. The first talk was given by Maxime Vono and was concerned with the research topic (B). He presented a generic method to perform distributed (*ie.* parallel) posterior computations. The second talk was given by Hong-Phuong Dang and was concerned with topics (A), (B) and (D). She has presented the advantages of a Bayesian approach for image denoising, together with an algorithm for efficient posterior sampling. The third talk, given by Thibault Randrianarisoa, was concerned with topic (C). He presented frequentist guarantees for point-estimation and confidence sets in density estimation for a Bayesian model based on *Pólya trees*. The last talk, given by Mariia Vladimirova, was concerned with topics (A) and (D). She explained how we can understand and interpret the internals of somewhat complicated models such as Bayesian neural networks.

This article presents in more detail some of the works that have been presented during the session, organised and chaired by Zacharie Naulet who wrote this introduction. The focus of the article is put on topic (D), in particular on two popular Bayesian supervised learning algorithms that are Bayesian neural networks and sparse linear models. Section 2, written by Mariia Vladimirova and Julyan Arbel, focuses on distributional properties of Bayesian neural networks. Section 3, written by Hong-Phuong Dang, Clément Elivra, and Cédric Herzet, describes an approximate posterior sampling scheme for dictionary learning.

Notations

We use the following notations throughout the paper. Matrices and vectors are respectively denoted by uppercase (*e.g.*, $\mathbf{W}$, $\mathbf{D}$) and lowercase (*e.g.*, $\mathbf{g}$, $\mathbf{h}$, $\mathbf{d}$) bold letters. $\mathbf{I}_L$ is the identity matrix of dimension L. The nth column of matrix $\mathbf{D}$ is written as $\mathbf{d}_n$ and element (k, n) as d_{kn}. $\odot$ denotes element-wise product between two vectors. $\|\cdot\|_F$ is the Frobenius norm of a matrix; $\|\cdot\|_0$ returns the number of nonzero elements in its argument. $\mathbb{I}\{\cdot\}$ is the indicator function which is equal to 1 when the statement between braces is true and to 0 otherwise. The symbol $\propto$ refers to equality up to a normalization constant. Finally, the notations $\mathcal{N}$, $\mathcal{IG}$, Ber, Beta, and $\mathrm{GWT}_{\mathbb{R}}$ stand respectively for the Normal, Inverse Gamma, Bernoulli, Beta, and generalized Weibull-tail distributions. We use the following notations for neural networks. The number of hidden layers, called *depth*, is denoted L. Each layer following the input layer consists of units which are linear combinations of previous layer units transformed by a function (oftentimes nonlinear), referred to as the *activation function* and denoted by $\phi : \mathbb{R} \to \mathbb{R}$. Given an input $\mathbf{x} \in \mathbb{R}^N$ (for instance an image made of N pixels) the ℓ-th hidden layer, $\ell = 1, \dots, L$, consists of two vectors of size denoted by H_ℓ (called the *width* of layer). The vector of units before application of the non-linearity is called *pre-activation*, and is denoted by $\mathbf{g}^{(\ell)}(\mathbf{x}) = \mathbf{g}^{(\ell)} = (g_1^{(\ell)}, \dots, g_{H_\ell}^{(\ell)})$, while the vector obtained after element-wise application of ϕ is called *post-activation* and is denoted by $\mathbf{h}^{(\ell)}(\mathbf{x}) = \mathbf{h}^{(\ell)} = (h_1^{(\ell)}, \dots, h_{H_\ell}^{(\ell)})$. More specifically, these vectors are defined as

$$\mathbf{g}^{(\ell)}(\mathbf{x}) = \mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)}(\mathbf{x}), \quad \mathbf{h}^{(\ell)}(\mathbf{x}) = \phi(\mathbf{g}^{(\ell)}(\mathbf{x})), \tag{1}$$

where $\mathbf{W}^{(\ell)}$ is a weight matrix of dimension $H_\ell \times H_{\ell-1}$ including a bias vector, with the convention that $H_0 = N$, the input dimension.

2. Bayesian neural networks distributional properties

Neural networks suffer from numerous limitations, despite many advances and high interest in the deep learning field. Their adoption in practical and safety-critical applications is still restrained [12]. To overcome limitations, many researchers work on understanding the mechanisms behind deep learning models and developing new tools. For instance, the influence of different architecture and training procedure on outputs and their better description can help with the choice of a proper model for a given problem and, in general, with transparency and trustworthiness. In the same way, it is essential to have well-calibrated uncertainty for believing the prediction outputs [13,14]. Uncertainty estimates can also serve as a means of transparency as they inform when the model does not know the correct prediction [15].

Bayesian inference is considered as one of the solutions to trustworthiness as it allows to provide some uncertainty for the outputs. Instead of taking into account a single answer of a model, Bayesian methods allow considering an entire distribution of answers. Bayesian neural networks achieve good performance while providing an uncertainty quantification of their outputs. The Bayesian approach *does not open the black box* of neural network-based models; however, it gives a new perspective to studying internal mechanisms of neural networks. For recent reviews, see [16–18].

We suggest studying the internal mechanisms in Bayesian neural networks. More specifically, we focus on prior distributions at the unit level. The flagship result is that function-space priors converge to a Gaussian process when the layers' width tends to infinity. We extend this result to finite-width Bayesian neural networks by providing a characterization of the marginal prior distribution of the units. We provide an accurate

characterization of hidden units tails through sub-Weibull and Weibull-tail descriptions. The obtained results illustrate the heavy-tailed nature of hidden units in deep layers for different weight priors. We believe that these characterizations help to understand the internal mechanisms of neural networks and to suggest model improvements.

2.1. **Sub-Weibull hidden units**

The recent work by Bibi et al. [19] provides the expression of the first two moments of the output units of a one-layer neural network. Obtaining moments is a preliminary first step to characterizing a whole distribution. However, the methodology of [19] is also limited to one hidden layer neural networks.

Later work focuses on moments of hidden units and shows that any order moments are finite under mild assumptions on the activation function. More specifically, the *sub-Weibull* property of distributions is shown, conjecturing that hidden units are heavier-tailed with going *deeper* in the network [20, 21]. This is in contrast with their GP limit which is obtained when going *wider*. To describe this result, we start with the formal definition of a Sub-Weibull random variable:

Definition 2.1 (Sub-Weibull random variable)**.** *A random variable X satisfying for all $x > 0$ and for some $\theta > 0$*

$$\mathbb{P}(|X| \geq x) \leq a \exp\left(-x^{1/\theta}\right), \tag{2}$$

is called a sub-Weibull random variable with so-called tail parameter θ, which is denoted by $X \sim \mathrm{subW}(\theta)$.

Sub-Weibull distributions are characterized by tails lighter than (or equally light as) Weibull distributions; in the same way as sub-Gaussian or sub-exponential distributions correspond to distributions with tails lighter than Gaussian and exponential distributions, respectively. Sub-Weibull distributions are parameterized by a positive tail index θ and are equivalent to sub-Gaussian for $\theta = 1/2$ and sub-exponential for $\theta = 1$.

Given some input $\mathbf{x}$, such prior distribution induces by forward propagation (1) a prior distribution on the pre-nonlinearities and post-nonlinearities, whose *tail properties* are the focus of this section. To this aim, the activation function ϕ is required to span at least half of the real line as follows. We introduce an extended version of the activation function assumption from Matthews et al. [22]:

Definition 2.2 (Extended envelope property)**.** *An activation function $\phi : \mathbb{R} \to \mathbb{R}$ is said to obey the extended envelope property if there exist $c_1, c_2 \geq 0$, $d_1, d_2 > 0$ such that the following inequalities hold*

$$\begin{aligned} &|\phi(u)| \geq c_1 + d_1|u| \quad \text{for all } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-, \\ &|\phi(u)| \leq c_2 + d_2|u| \quad \text{for all } u \in \mathbb{R}. \end{aligned} \tag{3}$$

The interpretation of this property is that ϕ must shoot to infinity at least in one direction ($\mathbb{R}_+$ or $\mathbb{R}_-$, at least linearly (first line of (3)), and also at most linearly (second line of (3)). Of course, compactly supported nonlinearities such as sigmoid and tanh do not satisfy the extended envelope property but the majority of other nonlinearities do, including ReLU, ELU, PReLU, and SeLU.

Theorem 2.1 (Vladimirova et al. [20])**.** *Consider a Bayesian neural network with centered Gaussian priors and with activation function ϕ satisfying the extended envelope condition of Definition 2.2. Then conditional on the input $\mathbf{x}$, the marginal prior distribution induced by forward propagation (1) on any unit (pre- or post-activation) of the ℓ-th hidden layer is sub-Weibull with tail parameter $\theta = \ell/2$. That is for any $1 \leq \ell \leq L$, and for any $1 \leq m \leq H_\ell$,*

$$u_m^{(\ell)} \sim \mathrm{subW}(\ell/2),$$

where a subW *distribution is defined in Definition 2.1, and $u_m^{(\ell)}$ is either a pre-activation $g_m^{(\ell)}$ or a post-activation $h_m^{(\ell)}$.*

2.2. Weibull-tail hidden units

Further, this result is improved by showing that hidden units are *Weibull-tail* distributed. Weibull-tail distributions are characterized in a different manner than sub-Weibull distributions, not based on moments but on a precise description of their tails. Denote by $F_X(\cdot)$ and $\overline{F}_X(\cdot)$, respectively, the cumulative distribution function and survival function of some random variable X.

Definition 2.3 (Generalized Weibull-tail on $\mathbb{R}$)**.** *A random variable X is generalized Weibull-tail on $\mathbb{R}$ with tail parameter $\beta > 0$ if both its right and left tails are upper and lower bounded by some Weibull-tail functions with tail parameter β:*

$$\begin{aligned} \mathrm{e}^{-x^\beta l_1^r(x)} \leq & \overline{F}_X(x) \leq \mathrm{e}^{-x^\beta l_2^r(x)}, \qquad \text{for } x > 0 \text{ and } x \text{ large enough,} \\ \mathrm{e}^{-|x|^\beta l_1^l(|x|)} \leq & F_X(x) \leq \mathrm{e}^{-|x|^\beta l_2^l(|x|)}, \quad \text{for } x < 0 \text{ and } -x \text{ large enough,} \end{aligned}$$

where l_1^r, l_2^r, l_1^l and l_2^l are slowly-varying functions. We note $X \sim \mathrm{GWT}_{\mathbb{R}}(\beta)$.

This tail description reveals the difference between hidden units' distributional properties in finite- and infinite-width Bayesian neural networks, since hidden units are generalized Weibull-tail with a tail parameter depending on those of the weights:

Theorem 2.2 (Vladimirova et al. [23])**.** *Consider a Bayesian neural network as described in Equation (1) with ReLU activation function. Let ℓ-th layer weights be independent symmetric generalized Weibull-tail on $\mathbb{R}$ with tail parameter $\beta_w^{(\ell)}$. Then conditional on the input $\mathbf{x}$, the marginal prior distribution induced by forward propagation (1) on any pre-activation is generalized Weibull-tail on $\mathbb{R}$: for any $1 \leq \ell \leq L$, and for any $1 \leq m \leq H_\ell$,*

$$g_m^{(\ell)} \sim \mathrm{GWT}_{\mathbb{R}}(\beta^{(\ell)}),$$

with tail parameter $\beta^{(\ell)}$ such that $\frac{1}{\beta^{(\ell)}} = \frac{1}{\beta_w^{(1)}} + \cdots + \frac{1}{\beta_w^{(\ell)}}$, where a $\mathrm{GWT}_{\mathbb{R}}$ distribution is defined in Definition 2.3.

Note that the most popular case of weight prior, iid Gaussian [24], corresponds to $\mathrm{GWT}_{\mathbb{R}}(2)$ weights. This leads to units of layer ℓ which are $\mathrm{GWT}_{\mathbb{R}}(\frac{2}{\ell})$.

2.3. Generalized Weibull-tail vs sub-Weibull properties

Some of the commonly used techniques to study the tail behavior is to consider probability tail bounds such as sub-Gaussian, sub-exponential, or their generalization to sub-Weibull distributions [21, 25]. The sub-Weibull property of Definition 2.1 for a non-negative random variable ensures the existence of the moment-generating function as well as bounds on moments. In contrast, the Weibull-tail property of Definition 2.3 characterizes the survival or density functions without a hand on moments.

While tail parameters in Definition 2.1 and Definition 2.3 of sub-Weibull and generalized Weibull-tail properties play different roles, there exist connections. Notice that for any constants $a, b, \beta > 0$, function $l(x) = b - \frac{\log a}{x^\beta} \geq 0$ is slowly-varying and $a\mathrm{e}^{-bx^\beta} = \mathrm{e}^{-x^\beta l(x)}$. It means that if a random variable X is sub-Weibull with parameter $\theta = 1/\beta > 0$, satisfying Equation (2), then the survival function of X is upper-bounded by a Weibull-tail function with tail parameter β and slowly-varying function $l(x) = 1$. Conversely, let X be a generalized Weibull-tail random variable with tail parameter β and denote by l the slowly-varying function associated with its definition. Then by Proposition 1.3.6 in [26], for any $\alpha > 0$, we have $x^\alpha l(x) \to \infty$ and $x^{-\alpha} l(x) \to 0$ when $x \to \infty$. So for any β_1, β_2 such that $0 < \beta_1 < \beta < \beta_2$, there exist $a_1, a_2 > 0$ such that

$$a_1 \mathrm{e}^{-x^{\beta_2}} \leq \overline{F}_X(x) = \mathrm{e}^{-x^\beta l(x)} \leq a_2 \mathrm{e}^{-x^{\beta_1}}.$$

In other words, $\mathrm{GWT}_{\mathbb{R}_+}(\beta) \subset \mathrm{SubW}(1/\beta_1)$ and $\mathrm{GWT}_{\mathbb{R}_+}(\beta) \not\subset \mathrm{SubW}(1/\beta_2)$, as illustrated on Figure 1.

Theorem 2.1 shows that hidden units of Bayesian neural networks with iid Gaussian priors are sub-Weibull with tail parameter proportional to the hidden layer number, that is $\theta = \frac{\ell}{2}$. It means that the unit distributions

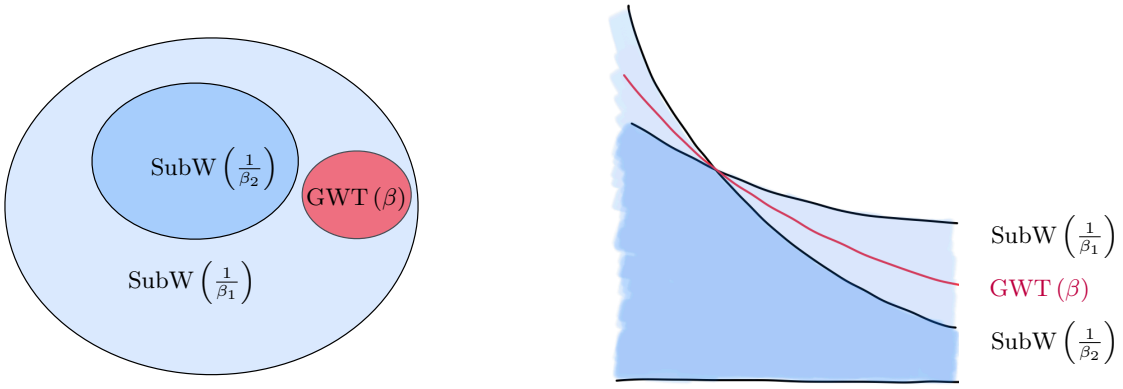

FIGURE 1. Relation between sub-Weibull and generalized Weibull-tail characteristics.

of hidden layer ℓ can be upper-bounded by some Weibull distributions $ae^{-x^{2/\ell}}$ for all ℓ. For larger tail parameter θ, Weibull distribution $ae^{-x^{1/\theta}}$ is heavier-tailed but being sub-Weibull does not guarantee the heaviness of the tails. However, this upper bound is optimal in the sense that it is achieved for neural networks with one hidden unit per layer.

From Theorem 2.2, for neural networks with independent Gaussian weights, hidden units of ℓ-th layer are generalized Weibull-tail with tail parameter $\beta = 1/\theta = 2/\ell$ so they have upper and lower bounds of the form $e^{-x^{2/\ell}l(x)}$ up to a constant where l is some slowly-varying function. Therefore, it proves that hidden units are heavier-tailed when going deeper for any finite number of hidden units per layer.

2.4. **Related literature and discussion of the results**

Regularization interpretation. It is well-known that performing a maximum a posterior (MAP) estimation in a Bayesian context is akin to employing a penalized maximum likelihood estimation (MLE), where the role of the penalty is played by the negative log-prior. According to this lens, [20] show that the sub-Weibull prior obtained in Theorem 2.1 induces a different regularization at the level of the units $\mathbf{U}$ than at the weights level $\mathbf{W}$. As summarized in Table 1, the negative log-prior for some Gaussian prior is nothing but an $\mathcal{L}^2$ penalty (also called weight-decay in the deep learning community), while the negative log-prior for sub-Weibull priors of tail parameter $\ell/2$ may take the more elaborate form of $\mathcal{L}^{2/\ell}$ penalties.

Layer	Penalty on $\mathbf{W}$	Approximate penalty on $\mathbf{U}$	
1	$\|\mathbf{W}^{(1)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(1)}\|_2^2$	$\mathcal{L}^2$ (weight decay)
2	$\|\mathbf{W}^{(2)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(2)}\|$	$\mathcal{L}^1$ (Lasso)
ℓ	$\|\mathbf{W}^{(\ell)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(\ell)}\|_{2/\ell}^{2/\ell}$	$\mathcal{L}^{2/\ell}$

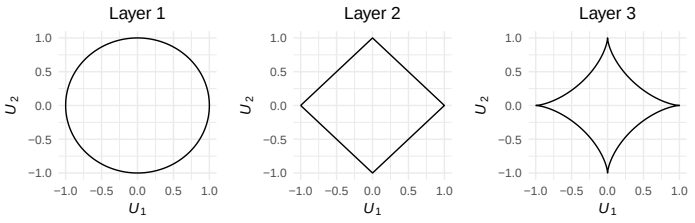

TABLE 1. Left: Comparison of Bayesian neural network penalties on weights $\mathbf{W}$ and units $\mathbf{U}$ for varying layer depths. Right: Graphical representation of the unit penalties implied at the unit level for varying layer depths.

Gaussian pre-activations. The study of the neural networks' distributional properties through Bayesian analysis, where the weights are assigned some prior distribution, has attracted a lot of attention in recent years. One of the main results in the field states that Bayesian deep neural network units converge in distribution to a Gaussian process when the layers' width goes to infinity. Originally stated by [27] for one-hidden-layer

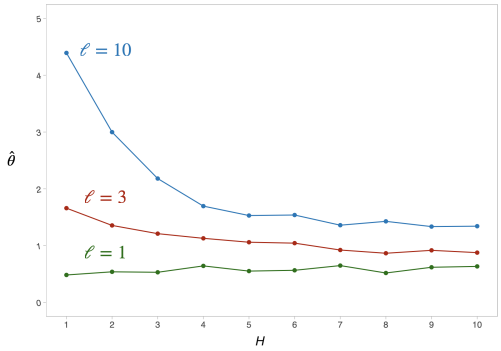

FIGURE 2. Tail index estimator $\hat{\theta}$ for units at layers $\ell \in \{1, 3, 10\}$ as a function of the width $H \in \{1, 2, \ldots, 10\}$.

neural networks, this result was recently shown to carry over to deep neural networks by [22, 28]. In contrast, Theorem 2.2 shows that the *non-asymptotic* (i.e. for finite-width neural networks) prior distribution of units from the ℓ-th layer induced by some prior on the weights gets heavier when going deeper in the network. This result puts into perspective the infinite-width Gaussian process property which might be far from holding for real-world, often very deep, neural networks. To illustrate this point, we conducted the following experiment. We simulated Bayesian fully connected neural networks according to standard Gaussian weights, with varying depth $\ell \in \{1, 3, 10\}$ and varying width (but fixed for each architecture) $H = H_\ell \in \{1, 2, \ldots, 10\}$. We propagated those random weights to units conditional on a fixed (once for all randomly sampled) input. For every layer, we then computed the tail index estimator $\hat{\theta}$ proposed in [21] for the tail parameter θ appearing in Equation 2.1. We can see in Figure 2 that the theoretical result of Theorem 2.1 that states that $\theta = \ell/2$ is well in line with the estimates obtained with networks of width $H = 1$. When the width increases, the estimates for θ tend to decrease, narrowing the gap to the lower bound of $1/2$ corresponding to a Gaussian distribution. A better understanding of this Gaussian is important since it is assumed to hold in a number of subsequent works, for instance, relative to information propagation in the neural network, as described below.

Information propagation and Edge of Chaos. An active line of research focuses on the propagation of deterministic inputs in neural networks [29–31]. These works build upon the limiting Gaussian process property of neural networks, in order to devise efficient initialization rules for neural networks. The main idea is to explore the covariance between pre-activations for two given different data points. [29] and [30] obtain recurrence relations under the assumption of Gaussian initialization and Gaussian pre-activations. They conclude that there is a critical line, so-called *Edge of Chaos*, separating signal propagation into two regions. The first one is an ordered phase in which all inputs end up asymptotically correlated. The second one is a chaotic phase in which all inputs end up asymptotically independent. To propagate the information deeper in a neural network, one should choose Gaussian prior variances corresponding to the separating line Edge of Chaos. [31] show that the smoothness of the activation function also plays an important role. Since this line of works considers Gaussian priors not only on the weights but also on the pre-activations, it is closely related to a wide regime where the number of hidden units per layer tends to infinity. Given that hidden units are heavier-tailed with depth, we argue that the impact of the Gaussian pre-activations assumption on the Edge of Chaos should be better understood. This issue is further investigated in [32].

Cold posterior effect and priors. It was recently empirically found that Gaussian priors in neural networks lead to the *cold posterior effect* in which a tempered "cold" posterior, obtained by exponentiating the posterior to some power greater than one, performs better than an untempered one [33]. The performed Bayesian inference is considered sub-optimal due to the need for cold posteriors, and the model is deemed misspecified. From that angle, [33] suggest that Gaussian priors might not be a good choice for Bayesian neural networks. In some works, data augmentation is argued to be the main reason for this effect [34, 35]

as the increased amount of observed data naturally leads to higher posterior contraction [34]. At the same time, even considering the data augmentation for some models, the cold posterior effect is still present. [35] hypothesize that using an appropriate prior incorporating knowledge of data augmentation might provide a solution. Moreover, heavy-tailed priors have been shown to mitigate the cold posterior effect [36]. According to Theorem 2.2, heavier-tailed priors lead to even heavier-tailed induced priors in function-space. Thus, the heavy-tail property of distributions in function-space might be a highly beneficial feature. [36] also proposed correlated priors for convolutional neural networks since trained weights are empirically strongly correlated, see also [37]. Correlated priors improve overall performance but do not alleviate the cold posterior effect. Another research direction to understanding the cold posterior effect is through the lens of generalization bounds such as PAC-Bayesian ones [38]. It is argued that discussions of the cold posterior effect should take into account that approximate Bayesian inference does not readily provide guarantees of performance on out-of-sample data, which are better described through generalization bounds.

2.5. **Conclusion**

Bayesian inference is one of the solutions to trustworthiness as it provides a framework to describe some uncertainty for the outputs. Instead of taking into account a single answer of a model, Bayesian methods allow considering an entire distribution of answers. Bayesian neural networks achieve good performance while providing an uncertainty quantification of their outputs. Though the Bayesian approach *does not open the black box* of neural network-based models, it opens a new perspective to study the internal mechanisms. We discussed some recent advances in describing Bayesian neural networks at the level of units. We hope that these results help to understand better the underlying working flow.

3. Small-variance asymptotics approximation of Gibbs sampling for dictionary learning

3.1. **Problem statement**

Most digital acquisition devices have traits that make measurements subject to noise. A standard signal processing task therefore consists of recovering some unknown signal $\mathbf{s} \in \mathbb{R}^L$ from its noisy observation

$$\mathbf{x} = \mathbf{s} + \mathbf{b} \tag{4}$$

where $\mathbf{b}$ represents some corrupting noise. It is well-known that reducing the noise variance is only possible if some *prior* information on $\mathbf{s}$ is known, *e.g.*, $\mathbf{s} \in \mathcal{S}_{\text{target}} \subset \mathbb{R}^L$ [39, Ch. 11]. Since such knowledge is rarely available in practice, many denoising algorithms leverage the construction of a surrogate model $\mathcal{S}_{\text{model}} \supseteq \mathcal{S}_{\text{target}}$ from a set of collected data, say

$$\mathbf{x}_n = \mathbf{s}_n + \mathbf{b}_n \quad n = 1, \dots, N. \tag{5}$$

In this paper, we consider the so-called "s-sparse" model where each element of $\{\mathbf{s}_n\}_{n=1}^N$ is assumed to stem from a noisy linear combination of a few columns of some matrix $\mathbf{D} \in \mathbb{R}^{L \times K}$, that is

$$\mathcal{S}_{\text{model}} = \{\mathbf{s} = \mathbf{D}\mathbf{w} + \boldsymbol{\epsilon} \mid \|\mathbf{w}\|_0 \leq s, \|\boldsymbol{\epsilon}\|_2 \leq \delta\}, \quad s \in \mathbb{N}^*, \, \delta \in \mathbb{R}_+. \tag{6}$$

During the last decade, model (6) has attracted the attention of many researchers. Its success mostly revolves around the combination of two ingredients. First, many natural signals have been shown to lie close to a sparse model, see *e.g.*, [39, Sec. 9.3] for a discussion in image processing. Second, a remarkable amount of algorithms (along with their theoretical analyses) have been proposed in the literature [40] to obtain sparse representations.

In particular, many of them have been shown to be robust to additive noise and stable with respect to sparsity defect, see *e.g.*, [40, Th. 4.19].

From the point of view of our denoising problem, a "good" sparse model should satisfy several (often contradictory) objectives. On the one hand, robustness against noise requires $\mathcal{S}_{\text{model}}$ not to "spread" too much in $\mathbb{R}^L$. This implies that s, K and δ should be chosen as small as possible. On the other hand, $\mathcal{S}_{\text{model}}$ must obviously contain the set of target vectors $\mathcal{S}_{\text{target}}$. Hence, s, K and δ should be chosen sufficiently large to obey this requirement. Lastly, the value of s, K and δ is also constrained from a "learning" point of view: since $\mathcal{S}_{\text{model}}$ has to be learned from a *finite* number of (noisy) samples $\{\mathbf{x}_n\}_{n=1}^N$, the number of degrees of freedom of $\mathcal{S}_{\text{model}}$ should not be too large to avoid overfitting and allow for generalization [41].

Many approaches to build sparse models satisfying these requirements have been proposed in the literature. The problem of finding a "good" matrix $\mathbf{D}$ from a set of data points $\{\mathbf{x}_n\}_{n=1}^N$ is often known as *"dictionary learning"*. Most methods addressing dictionary learning leverage the seminal work by Olshausen & Field [42] and are based on the resolution of an optimization problem where some constraints on s and δ are included explicitly or implicitly, see *e.g.*, [43–47]. The most popular example of such an approach is probably the "K-SVD" algorithm proposed by Aharon *et al.* in [43] where the authors introduced a dictionary learning algorithm searching (heuristically) a solution of the following problem:

$$\min_{\mathbf{D}\in\mathbb{R}^{L\times K},\{\mathbf{w}_n\in\mathbb{R}^K\}_{n=1}^N} \sum_{n=1}^{N} \frac{1}{2}\|\mathbf{x}_n - \mathbf{D}\mathbf{w}_n\|_F^2 + \lambda\|\mathbf{w}_n\|_0 \tag{7}$$

for some $\lambda > 0$. Here the trade-off between sparsity level and modeling/observation noise is implicitly chosen via the tuning parameter λ. The choice of λ and K has to be done manually using, *e.g.*, cross-validation [41]. This operation may be computationally cumbersome since problem (7) needs then to be solved for each new value of K. Moreover, the number and the range of values to test for K often remain a heuristic choice. A few works have elaborated on the K-SVD approach to propose adaptive dictionary learning methods that infer the size of the dictionary within the optimization process, see [48–51]. Nevertheless, these procedures still call for important parameter tuning. For example, most of them need to know the noise or the sparsity level.

Another well-known family of dictionary learning procedures are Bayesian methods, see *e.g.*, [52–55]. Unlike "optimization-based" methods, Bayesian procedures can naturally include the evaluation of model parameters within the estimation process. For example, in [52–54] the parameters of a Bernoulli-Gamma-Gaussian model and the variance of the observation noise were embedded within the dictionary learning problem. In [55], Dang *et al.* went one step further and incorporated the size of the dictionary (K) into the learning task by using an "Indian-Buffet-Process" (IBP) prior. In these papers, a tractable implementation solving the joint "dictionary-parameters" estimation problem is usually achieved by resorting to Monte-Carlo sampling algorithms. Although these procedures are known to converge asymptotically (in the number of samples) to exact a posteriori estimates, their main drawback stands in their computational complexity since a large number of iterations are often needed to attain the target sampling distribution.

Many solutions have been proposed in the literature to overcome the high-computational cost of Monte-Carlo methods, *e.g.*, stochastic methods, sequential Monte Carlo, particle Markov chain Monte-Carlo, stochastic variational inference or variational Bayes methods. In this paper we focus on the so-called "small-variance asymptotics" (SVA) approximation of the Gibbs sampler proposed in [56] and further extended in [57–64]. We emphasize that exploiting this type of approximation in the context of dictionary learning can lead to procedures combining the advantages of "optimization-based" and "Bayesian-based" methods, namely reasonable computational cost and integrated estimation of the model parameters. The material of this paper leverages contributions [65–67].

Target	Distribution family	Parameters definition	
$\mathbf{d}_k \mid \cdot$	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\boldsymbol{\Sigma} = \left(L + \sigma_{\mathbf{b}}^{-2} \sum_{n=1}^{N} (w_{kn})^2\right)^{-1} \mathbf{I}_L$	$\boldsymbol{\mu} = \sigma_{\mathbf{b}}^{-2} \boldsymbol{\Sigma} \sum_{n=1}^{N} w_{kn} \left(\mathbf{x}_n - \sum_{j \neq k}^{K} \mathbf{d}_j w_{jn}\right)$
$c_{kn} \mid z_{kn} = 1, \cdot$	$\mathcal{N}(\mu, \Sigma)$	$\Sigma = \left(\sigma_{\mathbf{b}}^{-2} \|\mathbf{d}_k\|_2^2 + \sigma_{\mathbf{c}}^{-2}\right)^{-1}$	$\mu = \sigma_{\mathbf{b}}^{-2} \Sigma \mathbf{d}_k^{\mathrm{T}} \left(\mathbf{x}_n - \sum_{j \neq k}^{K} \mathbf{d}_j w_{jn}\right)$
$c_{kn} \mid z_{kn} = 0, \cdot$	$\mathcal{N}(\mu, \Sigma)$	$\Sigma = \sigma_{\mathbf{c}}^2$	$\mu = 0$
$\sigma_{\mathbf{b}}^2 \mid \cdot$	$\mathcal{IG}(\alpha, \beta)$	$\alpha = e_0 + \frac{NL}{2}$	$\beta = f_0 + \frac{1}{2} \|\mathbf{X} - \mathbf{DW}\|_F^2$
$\sigma_{\mathbf{c}}^2 \mid \cdot$	$\mathcal{IG}(\alpha, \beta)$	$\alpha = c_0 + \frac{KN}{2}$	$\beta = d_0 + \frac{1}{2} \|\mathbf{C}\|_F^2$

TABLE 2. Expressions of some a posteriori conditional probabilities related to model (8)-(12). Shorthand notation "$X \mid \cdot$" refers to random variable X conditioned to all the other variables of the problem.

3.2. **Bayesian model**

This section introduces the Bayesian model used for our dictionary learning problem. We assume that each element of the training set $\{\mathbf{x}_n\}_{n=1}^N$ is an independent realization of the following model

$$\mathbf{x} = \mathbf{D}\mathbf{w} + \mathbf{b} \tag{8}$$

corresponding (to some extent) to the sparse model (6) with $\delta = 0$,[1] and where

$$\begin{array}{rcll} \mathbf{b} & \sim & \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{b}}^2 \mathbf{I}_L) & \\ \mathbf{d}_k & \sim & \mathcal{N}(\mathbf{0}, L^{-1} \mathbf{I}_L) & \forall 1 \leq k \leq K \end{array} \tag{9}$$

and $\mathbf{w}$ obeys some "sparsity-enforcing" distribution.

In this paper, we consider the following option:

$$\mathbf{w} = \mathbf{c} \odot \mathbf{z} \qquad \text{where} \qquad \mathbf{c} \sim \mathcal{N}(0, \sigma_{\mathbf{c}}^2 \mathbf{I}_K) \tag{10}$$

and $\mathrm{p}(\mathbf{z})$ is a distribution on $\{0,1\}^K$ which favors sparsity. We will consider two particular choices for $\mathrm{p}(\mathbf{z})$ herefafter, namely a Bernoulli distribution (Section 3.3.1) and an "Indian-Buffet Process" (Section 3.3.2). We finally assume that variances $\sigma_{\mathbf{b}}^2$ and $\sigma_{\mathbf{c}}^2$ are distributed as

$$\sigma_{\mathbf{b}}^2 \sim \mathcal{IG}(e_0, f_0) \tag{11}$$

$$\sigma_{\mathbf{c}}^2 \sim \mathcal{IG}(c_0, d_0) \tag{12}$$

for some positive hyper-parameters c_0, d_0, e_0, f_0. In the sequel, we will use the notations $\mathbf{X}$, $\mathbf{W}$, $\mathbf{C}$, $\mathbf{Z}$ to denote matrices whose columns correspond to realizations $\{\mathbf{x}_n\}_{n=1}^N$, $\{\mathbf{w}_n\}_{n=1}^N$, $\{\mathbf{c}_n\}_{n=1}^N$ and $\{\mathbf{z}_n\}_{n=1}^N$, respectively. As a final remark, we note that the variance of each $\mathbf{d}_k$ in (9) has been fixed to L^{-1} to address the multiplicative factor indeterminacy in the pair $(\mathbf{D}, \mathbf{W})$.[2]

[1] The choice $\delta = 0$ is arbitrary and follows from the fact that the model defect $\boldsymbol{\epsilon}$ in (6) and the observation noise $\mathbf{b}$ in (4) cannot be disambiguated.

[2] Setting the variance of each $\mathbf{d}_k$ proportional to L^{-1} implies that $\mathbb{E}\left[\mathbf{d}_k^{\mathrm{T}} \mathbf{d}_k\right] = 1$, thus corresponding to a form of "soft normalization".

3.3. **Small-variance approximation of Gibbs sampling**

A standard approach to compute a point estimate from some a posteriori distribution is to resort to Monte-Carlo approximation and Gibbs sampling [4]. The former consists of approximating the expectation of a random quantity by the arithmetic mean of a set of realizations; the latter is a powerful tool to drawn samples from the distribution of interest. Gibbs sampling is a "Markov chain Monte Carlo" method: it generates a sequence of iterates which can be shown to asymptotically converge to realizations of the target distribution. The sequence of iterates is generated by sequentially sampling realizations of the problem's random variables (here $\sigma_{\mathbf{b}}^2$, $\sigma_{\mathbf{c}}^2$, elements of $\mathbf{D}$, $\mathbf{C}$, $\mathbf{Z}$ and potential hyperparameters) from their conditional a posteriori distributions.

Some conditional distributions associated to $\sigma_{\mathbf{b}}^2$, $\sigma_{\mathbf{c}}^2$, $\mathbf{D}$, $\mathbf{C}$ are summarized in Table 2. The conditional distribution of $\mathbf{Z}$ will be detailed in Sections 3.3.1 and 3.3.2 for two particular choices of $\mathrm{p}(\mathbf{Z})$. We note that, as far as $\sigma_{\mathbf{b}}^2$, $\sigma_{\mathbf{c}}^2$, $\mathbf{D}$, $\mathbf{C}$ are concerned, the conditional probabilities involved in the implementation of the Gibbs sampler correspond to distributions which can be efficiently sampled by many well-known methods of the literature, see *e.g.*, [4, Chapter 2].

Unfortunately, despite this desirable feature, solving dictionary learning problems via Gibbs sampling often entails a computational complexity superior to "optimization-based" methods by orders of magnitude. In this paper we consider a particular approximation of the Gibbs sampler, known as "small-variance asymptotics" (SVA), to decrease the computational cost of this method. SVA approximation has been proposed in [56] and consists of approximating (some of) the problem's conditional distributions by their limit expression as the noise variance $\sigma_{\mathbf{b}}^2$ tends to zero. More specifically, SVA methods commonly build on the following two key ingredients:[3]

i) A Gibbs procedure designed to sample the target posterior probabilities (that is $\mathrm{p}(\mathbf{C}, \mathbf{Z}, \mathbf{D}, \sigma_{\mathbf{b}}^2, \sigma_{\mathbf{c}}^2 | \mathbf{X})$ in our image-denoising framework). The definition of this sampler involves in particular the specification of conditional probabilities on (subsets of) variables $\mathbf{C}, \mathbf{Z}, \mathbf{D}, \sigma_{\mathbf{b}}^2, \sigma_{\mathbf{c}}^2$.

ii) The approximation of the above conditional probabilities by single-mass discrete probabilities. The name "SVA" comes from the fact that the construction of the single-mass approximation is based on the asymptotic behavior of the target conditional probability as $\sigma_{\mathbf{b}}^2$ tends to zero (and for some particular choice for the scaling of hyper-parameters).

We note that, upon the approximation mentioned in *ii)*, the sampling of the conditional probabilities appearing in the Gibbs procedure reduces to deterministic (sometimes closed-form) updates of the variables $\mathbf{C}, \mathbf{Z}, \mathbf{D}, \sigma_{\mathbf{b}}^2, \sigma_{\mathbf{c}}^2$. Hence, in a nutshell, SVA methods can be seen as approximations of Gibbs samplers in which the sampling of the conditional probabilities are replaced by deterministic updates. Although there exist (to the best of our knowledge) no theoretical guarantees on the quality of the posterior estimates obtained from SVA approximations, this framework has been shown in many contributions [56–58,68] to drastically reduce the computational complexity of standard Gibbs samplers while achieving nice empirical estimation performance.

In rest of this section, we illustrate how the above two keys ingredients particularize to our image-denoising setups. We consider the cases where $\mathrm{p}(\mathbf{Z})$ corresponds to a Beta-Bernoulli distribution in Section 3.3.1 and an "Indian-Buffet Process" in Section 3.3.2.

3.3.1. *Beta-Bernoulli model*

In this section, we consider the case where

$$\mathbf{z}_n \sim \prod_{k=1}^{K} \mathrm{Ber}(\pi_k) \quad \text{with} \quad \boldsymbol{\pi} \sim \prod_{k=1}^{K} \mathrm{Beta}(a_0, b_0), \quad a_0, b_0 > 0. \tag{13}$$

[3] We refer the reader to the original paper [56] for a detailed explanation of the rationale behind this method.

We see that the number K of atoms is fixed in advance in model (13). Nevertheless, the probability π_k of activation of each atom is left as a degree of freedom to the estimator. The learning procedure can therefore possibly "decrease" the number of atoms in the dictionary by setting some activation parameters to zero.

The conditional posterior probabilities of z_{kn} and π_k take the form:

$$z_{kn} \mid \cdot \sim \mathrm{Ber}\Big(\frac{p_{kn}}{p_{kn}+1-\pi_k}\Big) \quad \text{where} \quad p_{kn} = \pi_k \exp\Big[\frac{-1}{2\sigma_{\mathbf{b}}^2}\Big(c_{kn}^2\|\mathbf{d}_k\|_2^2 - 2c_{kn}\mathbf{d}_k^{\mathrm{T}}(\mathbf{x}_n - \sum_{j\neq k}^{K}\mathbf{d}_j w_{jn})\Big)\Big] \tag{14}$$

$$\pi_k \mid \cdot \sim \mathrm{Beta}\Big(a_0 + \sum_{n=1}^{N} z_{kn}, b_0 + N - \sum_{n=1}^{N} z_{kn}\Big). \tag{15}$$

SVA approximation consists of considering the limit form of (some of) the conditional probabilities in Table 2 and (14)-(15) for $\sigma_{\mathbf{b}}^2 \to 0$ in the Gibbs sampler. To avoid degeneracy of the problem at stake, the activation probabilities are moreover parameterized as a particular function of the noise variance. More specifically, we let

$$\pi_k = \exp\Big(-\tfrac{\lambda_k}{2\sigma_{\mathbf{b}}^2}\Big) \text{ with } \lambda_k > 0, \tag{16}$$

so that $\pi_k \to 0$ as $\sigma_{\mathbf{b}}^2 \to 0$. Considering the parameter of the Bernoulli distribution in (14), we thus easily find that

$$\lim_{\sigma_{\mathbf{b}}^2\to 0}\Big(\tfrac{p_{kn}}{p_{kn}+1-\pi_k}\Big) = \begin{cases} 0 & \text{if } \rho_{kn} > 0 \\ 1 & \text{otherwise} \end{cases} \tag{17}$$

where $\rho_{kn} \triangleq c_{kn}^2\|\mathbf{d}_k\|_2^2 - c_{kn}\mathbf{d}_k^{\mathrm{T}}(\mathbf{x}_n - \sum_{j\neq k}^{K}\mathbf{d}_j w_{jn}) + \lambda_k$. Hence, under hypothesis (16) and in the small-variance limit, the realizations of the a posteriori conditional probability on z_{kn} obey the following deterministic rule:

$$z_{kn} = \begin{cases} 1 & \text{if } \rho_{kn} < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

This update is tantamount to defining the Gibbs sampling update as the mode of the conditional probabilities. A similar approach (not detailed here) can be followed for the distributions specified in Table 2. When the distribution is Gaussian, the SVA approximation thus corresponds to the limit of the mean as $\sigma_{\mathbf{b}}^2 \to 0$, see [67].

3.3.2. *Indian-Buffet-Process model*

Let us now assume that the elements of $\{\mathbf{z}_n\}_{n=1}^N$ correspond to binary sequences ($K = \infty$). With a slight abuse of notation, we will stick to matrix notations and let $\mathbf{Z}$ denote the column-wise concatenation of sequences $\{\mathbf{z}_n\}_{n=1}^N$ and z_{kn} the kth element of sequence $\mathbf{z}_n$. We consider the following model on $\mathbf{Z}$ (better known as "two-parameter Indian-Buffet[4] Process" (IBP) in the literature):

$$\mathrm{p}(\mathbf{Z}|\alpha,\xi) \propto \frac{(\alpha\xi)^{K_+}}{\gamma(\mathbf{Z})} \prod_{k=1}^{K_+} \beta(\sum_n z_{kn}, N - \sum_n z_{kn} + \xi) \prod_{k>K_+,n} \mathbb{I}\{z_{kn} = 0\} \tag{19}$$

[4]This name conceals a culinary metaphor which was initially used to explain the construction of the IBP, see [69].

where $\alpha > 0, \xi > 0$ are two model parameters,[5] $K_+ \triangleq \| \sum_{n=1}^N \mathbf{z}_n \|_0$ (that is, the number of atoms that participate to the reconstruction of *at least* one observation), β is the beta function and $\gamma(\mathbf{Z}) \neq 0$ is a function which basically depends on the "structure" of $\mathbf{Z}$ (but unimportant for our exposition here), see [70].

As $K = \infty$, the number of atoms considered in our model is theoretically infinite. However, only atoms corresponding to "not all-zero columns" of $\mathbf{Z}$ (referred to as active atoms) contribute to the reconstruction of the observations. A careful examination of model (19) indicates that the IBP promotes realizations of $\mathbf{Z}$ with finite numbers of active atoms (K_+). In that sense, we say that the size of the dictionary is left as a degree of freedom in the IBP model.

Let us concentrate on the SVA approximation of the following conditional probability

$$\mathrm{p}(\mathbf{C}, \mathbf{Z} | \mathbf{X}, \mathbf{D}, \sigma_{\mathbf{b}}^2, \sigma_{\mathbf{c}}^2, \alpha, \xi) \propto \mathrm{p}(\mathbf{X} | \mathbf{D}, \mathbf{C}, \mathbf{Z}, \sigma_{\mathbf{b}}^2)\ \mathrm{p}(\mathbf{Z} | \alpha, \xi)\ \mathrm{p}(\mathbf{C} | \sigma_S^2). \tag{20}$$

Similarly to Section 3.3.1, parametrization of α, ξ as functions of $\sigma_{\mathbf{b}}^2$ must be considered to avoid degeneracy of the conditional probability when $\sigma_{\mathbf{b}}^2 \to 0$. In particular, letting

$$\alpha = \exp\left(\frac{\sigma_{\mathbf{b}}^2}{2\lambda_1} - \frac{\lambda_1}{\sigma_{\mathbf{b}}^2} \right), \qquad \xi = \exp\left(\frac{\lambda_2}{\sigma_{\mathbf{b}}^2} - \frac{\sigma_{\mathbf{b}}^2}{2\lambda_2} \right) \qquad \text{for } \lambda_1 > \lambda_2 > 0, \tag{21}$$

we have that (20) concentrates on its mode(s) when $\sigma_{\mathbf{b}}^2 \to 0$. Moreover, the latter correspond to the minimizers of the following function:

$$\sum_{n=1}^N \left(\frac{1}{2} \left\| \mathbf{x}_n - \sum_{k=1}^{K_+} \mathbf{d}_k (c_{kn} \odot z_{kn}) \right\|_2^2 + \lambda_2 \sum_k z_{kn} \right) + (\lambda_1 - \lambda_2) K_+. \tag{22}$$

We note that, for a given value of K_+, the minimization over the first K_+ components[6] of $\mathbf{c}_n$ and $\mathbf{z}_n$ reduces to a standard ℓ_0-penalized sparse representation problem over a dictionary of K_+ atoms. This problem has been extensively studied in the literature during the last decades and many heuristic methods exist to solve it (exactly or to good accuracy) in many setups, see *e.g.*, [40].

As a final remark, let us note that (22) may admit infinitely many minimizers since for all n and $k \leq K_+$ such that $z_{kn} = 0$, all choices of the coefficient c_{kn} lead to the same value of the objective function. To resolve this ambiguity, we proceed to an extra sampling step according to the posterior probability

$$\lim_{\sigma_{\mathbf{b}}^2 \to 0} \mathrm{p}(\mathbf{C} | \mathbf{X}, \mathbf{D}, \mathbf{Z}, \sigma_{\mathbf{b}}^2, \sigma_{\mathbf{c}}^2, \alpha, \xi) \tag{23}$$

after solving (22). An SVA analysis of (23) indicates that the value of c_{kn} remains unchanged if $z_{kn} = 1$ and that c_{kn} has to be drawn according to some normal distribution otherwise (see Table 2).

3.4. **Numerical experiments**

This section reports the main conclusions of several numerical experiments in image processing conducted in [65–67]. More precisely, a standard method in image processing to assess the relevance of different dictionary learning approaches is to compare their denoising performance. As far as our simulation setup is concerned, the

[5] Whose behavior can be understood as follows: the "mass" parameter α controls the total number of atoms participating to the reconstruction of each observation; the "concentration" parameter ξ controls the "frequency" at which each atom is used to reconstruct the observations.

[6] The other components are equal to zero by definition of K_+.

FIGURE 3. Denoising results in the high-noise scenario obtained by using the IBP model. From left to right: noisy, denoised, original images and learned dictionary.

two models presented in this paper lead to denoising performance comparable to the state of the art. Moreover, they benefit from some properties of Bayesian techniques while inducing a computational cost similar to "optimization-based" methods. Figure 3 displays typical denoising results obtained by our proposed procedures. The rest of the section is organized as follows. We describe below our simulation setup. The performance of the dictionary learning procedures obtained from the two considered models are discussed in Sections 3.4.1 and 3.4.2, respectively.

Our simulation setup is as follows. A set of 5 (standard) images of size 512×512 is considered, namely "`barbara`", "`hill`", "`mandrill`", "`lena`" and "`peppers`". Each image is corrupted with an additive noise whose entries are i.i.d. realizations of a zero-mean Gaussian with standard deviation $\sigma_{\texttt{img}}$. We consider the cases $\sigma_{\texttt{img}} = 25$ and $\sigma_{\texttt{img}} = 40$, respectively referred to as "low-noise scenario" and "high-noise scenario" hereafter. Each image is then decomposed into a set of 8×8 overlapping patches, resulting in $N = 16129$ (corresponding to 50% overlapping) training vectors $\{\mathbf{x}_n\}_{n=1}^N \subset \mathbb{R}^{64}$. We then apply the Bayesian model described in Section 3.2 to all patches. More precisely, we interpret each patch $\mathbf{x}_n$ as a deteriorated version of some noise-free patch which obeys the sparse model (6). For the two models, the estimator of the nth (denoised) patch is thus defined as $\widehat{\mathbf{D}}(\widehat{\mathbf{c}}_n \odot \widehat{\mathbf{z}}_n)$ where the "hat" symbol refers to the SVA approximation of a MAP estimator. More details can be found in [65, 67]. The reconstructed image is eventually obtained by merging all estimated patches. Since the latter have been designed so that each pixel is involved in many patches, the value of one pixel of the reconstructed image is obtained by averaging its value over all patches where it was involved. The denoising performance is finally obtained by evaluating a peak signal-to-noise ratio (PSNR) between the original image and the estimated one. Hence, the higher the PSNR, the better the denoising performance.

3.4.1. *Results for the Beta-Bernoulli model*

In this section, we report numerical results obtained with the SVA approximation of the Gibbs sampler presented in Section 3.3.1. We refer to this method as "BBG-SVA" hereafter.

As mentioned earlier, the size of the dictionary (namely K) is a parameter of the model for BBG-SVA. Figure 4 shows the evolution of the PSNR obtained with BBG-SVA seen as a function of K for the two considered noise levels (*i.e.*, $\sigma_{\texttt{img}} = 25$ and $\sigma_{\texttt{img}} = 40$) and several images. For all considered setups, we observe that the denoising performance stabilizes for some value of K. Such a behavior has already been observed in [55] and suggests the existence of an "optimal" size that depends on both the image and the noise level. Besides, overestimating K seems to only affect the computational cost of the method. For these reasons, we focus on the value of $K = 300$ in the next experiment. Interestingly, this finding is also coherent with typical setups in image processing where a dictionary of size $K = 256$ or 512 atoms is typically learned [43, 54, 71].

An empirical comparison of the denoising performance of BBG-SVA with state-of-the-art dictionary learning approaches has been carried out in [67, Table 1] for $K = 300$. We have observed that BBG-SVA achieves denoising performance similar to its competitors for a running time of the same order of magnitude. This

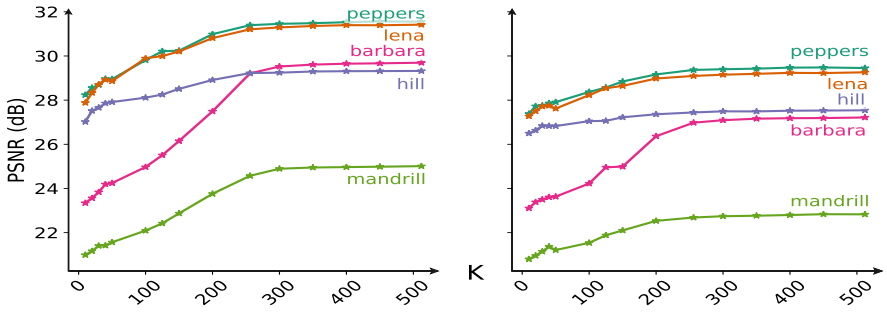

FIGURE 4. Evolution of BBG-SVA performances (PSNR of the reconstructed image) seen as function of the dictionary size K for several images. Left and right images correspond to the low-noise and high-noise scenarii, respectively.

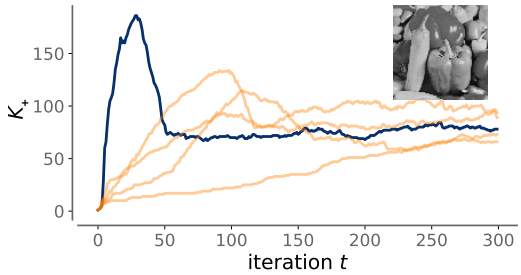

FIGURE 5. Evolution of the dictionary size K_+ across iterations for several couples of (λ_1, λ_2) in the low-noise scenario. The orange lines stand for discarded couples and the blue curve is the retained one.

indicates that BBG-SVA is able to automatically adjust the values of the hyper-parameters $\{\lambda_k\}_{k=1}^K$ in (18) and still features the flexibility of Bayesian approaches despite the SVA approximation.

3.4.2. *Results for the Indian-Buffet-Process model*

In this section, we illustrate the performance of the SVA approximation of the Gibbs sampler presented in Section 3.3.2. We refer to this method as "IBPDL-SVA" hereafter. In our simulations, we use OMP –a standard greedy procedure, see [40]– to address (22). Details about the implementation of the addition/removal of atoms (that is the update of K_+) can be found in [65].

In constrast to BBG-SVA, IBPDL-SVA includes the size of the dictionary as a variable of the problem. Such a feature has nevertheless been obtained at the cost of two additional hyperparameters in the model, namely λ_1 and λ_2 (see (21)). In our simulations, these two parameters are tuned by resorting to cross-validation on one image and reused for the others.

Figure 5 shows the evolution of the effective size K_+ of the dictionary along the iterations of the procedure for several couples of (λ_1, λ_2) and in the low-noise scenario. The "`peppers`" image is considered in this simulation. The blue curve denotes the couple (λ_1, λ_2) chosen by cross-validation while the orange ones correspond to the rejected couples. Interestingly, one observes some form of stability with respect to the choice of λ_1 and λ_2. More specifically, all the curves tend to stabilize around a common value of K_+ for all choices of λ_1 and λ_2.

An empirical comparison of the denoising performance of IBPDL-SVA with state-of-the-art dictionary learning approaches is available in [65,66]. Two conclusions can be drawn from these experimental results. First, the denoising performance of IBPDL-SVA is similar to that of other state-of-the-art dictionary learning procedures while automatically estimating the size of the dictionary and the noise level corrupting the data. Second, a

comparison of the complexity of IBPDL-SVA and the Gibbs sampler proposed in [55],[7] reveals that the former enables gain in terms of computational time of (at least) one order of magnitude with respect to the latter. More precisely, we have observed that Gibbs sampler in [55] performs in average 30 iterations per hours while our proposed methods reaches 150 iterations in 30 minutes[8].

3.5. **Conclusion**

Here we studied the "small-variance asymptotics" approximation of the Gibbs sampler related to two Bayesian models for dictionary learning. Our analysis leverages a carefully-designed coupling between the parameters of the model and the (estimated) noise variance. For the two models, the resulting method gathers both the flexibility of Bayesian modeling and the numerical efficiency of optimization methods. The relevance of the proposed approach was assessed on an image denoising application. The performance of the proposed approach was shown to be comparable with that of existing supervised methods, while automatically tuning the size of the dictionary and the level of the corrupting noise.

References

[1] P. Diaconis and D. Freedman, "On the consistency of Bayes estimates," *The Annals of Statistics*, pp. 1–26, 1986.

[2] S. Ghosal and A. Van der Vaart, *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, 2017, vol. 44.

[3] C. Robert, *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

[4] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, ser. Springer Texts in Statistics. Springer New York, 2013.

[5] R. Bardenet, A. Doucet, and C. C. Holmes, "On markov chain monte carlo methods for tall data," *Journal of Machine Learning Research*, vol. 18, no. 47, 2017.

[6] J. Huggins, T. Campbell, and T. Broderick, "Coresets for scalable bayesian logistic regression," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[7] J. Paisley, D. Blei, and M. Jordan, "Variational bayesian inference with stochastic search," *arXiv preprint arXiv:1206.6430*, 2012.

[8] B. Szabo and H. van Zanten, "Distributed function estimation: Adaptation using minimal communication," *Mathematical Statistics and Learning*, 2022.

[9] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[10] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.

[11] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[12] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[13] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, "A survey of uncertainty in deep neural networks," *arXiv preprint arXiv:2107.03342*, 2021.

[14] J. Antoran, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, "Getting a CLUE: A method for explaining uncertainty estimates," in *International Conference on Learning Representations*, 2021.

[15] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo *et al.*, "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty," in *AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[16] L. V. Jospin, W. Buntine, F. Boussaid, H. Laga, and M. Bennamoun, "Hands-on Bayesian neural networks–a tutorial for deep learning users," *arXiv preprint arXiv:2007.06823*, 2020.

[17] V. Fortuin, "Priors in Bayesian deep learning: A review," *International Statistical Review*, 2022.

[18] J. Arbel, K. Pitas, and M. Vladimirova, "A primer on Bayesian neural networks: review and debates," *Preprint*, 2022.

[19] A. Bibi, M. Alfadly, and B. Ghanem, "Analytic expressions for probabilistic moments of PL-DNN with Gaussian input," in *Computer Vision and Pattern Recognition*, 2018.

[20] M. Vladimirova, J. Verbeek, P. Mesejo, and J. Arbel, "Understanding priors in Bayesian neural networks at the unit level," in *International Conference on Machine Learning*, 2019.

[7] which leverages the same model as IBPDL-SVA.

[8] for a similar implementation and the same machine. Such a figure also varies from one image and corrupting noise level to another.

[21] M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel, "Sub-Weibull distributions: Generalizing sub-Gaussian and sub-exponential properties to heavier tailed distributions," *Stat*, vol. 9, no. 1, p. e318, 2020.
[22] A. G. d. G. Matthews, M. Rowland, J. Hron, R. E. Turner, and Z. Ghahramani, "Gaussian process behaviour in wide deep neural networks," in *International Conference on Learning Representations*, 2018.
[23] M. Vladimirova, J. Arbel, and S. Girard, "Bayesian neural network unit priors and generalized Weibull-tail property," in *Asian Conference on Machine Learning*, 2021.
[24] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 1996.
[25] A. K. Kuchibhotla and A. Chakrabortty, "Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression," *arXiv preprint arXiv:1804.02605*, 2018.
[26] N. H. Bingham, C. M. Goldie, and J. L. Teugels, *Regular variation*. Cambridge University Press, 1989.
[27] R. M. Neal, "Bayesian learning via stochastic dynamics," in *Advances in neural information processing systems*, 1993.
[28] J. Lee, J. Sohl-Dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, "Deep neural networks as Gaussian processes," in *International Conference on Learning Representations*, 2018.
[29] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, "Exponential expressivity in deep neural networks through transient chaos," in *International Conference on Neural Information Processing Systems*, 2016.
[30] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, "Deep information propagation," in *International Conference on Learning Representations*, 2017.
[31] S. Hayou, A. Doucet, and J. Rousseau, "On the impact of the activation function on deep neural networks training," in *International Conference on Machine Learning*, 2019.
[32] P. Wolinski and J. Arbel, "Gaussian Pre-Activations in a Neural Network: Myth or Reality?" *Arxiv Preprint*, 2022.
[33] F. Wenzel, K. Roth, B. Veeling, J. Swiatkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin, "How good is the Bayes posterior in deep neural networks really?" in *International Conference on Machine Learning*, 2020.
[34] P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson, "What are Bayesian neural network posteriors really like?" in *International Conference on Machine Learning*, 2021.
[35] S. Nabarro, S. Ganev, A. Garriga-Alonso, V. Fortuin, M. van der Wilk, and L. Aitchison, "Data augmentation in Bayesian neural networks and the cold posterior effect," *arXiv preprint arXiv:2106.05586*, 2021.
[36] V. Fortuin, A. Garriga-Alonso, F. Wenzel, G. Rätsch, R. Turner, M. van der Wilk, and L. Aitchison, "Bayesian neural network priors revisited," *Symposium on Advances in Approximate Bayesian Inference*, 2021.
[37] M. Vladimirova, J. Arbel, and S. Girard, "Dependence between Bayesian neural network units," in *Bayesian Deep Learning Workshop*, 2021.
[38] K. Pitas and J. Arbel, "Cold Posteriors through PAC-Bayes," *Submitted*, 2022.
[39] S. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 2009.
[40] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer New York, 2013.
[41] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014.
[42] B. Olshausen and D. Field, "Natural image statistics and efficient coding," in *Network: Computation in Neural Systems*, vol. 7, 1996, pp. 333–339.
[43] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, 2006.
[44] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, vol. 17, no. 1, pp. 53–69, 2008.
[45] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 2272–2279.
[46] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.
[47] N. Rao and F. Porikli, "A clustering approach to optimize online dictionary learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 1293–1296.
[48] R. Mazhar and P. Gader, "EK-SVD: Optimized dictionary design for sparse representations," in *Proceedings of the International Conference on Pattern Recognition*, 2008, pp. 1–4.
[49] J. Feng, L. Song, X. Yang, and W. Zhang, "Sub clustering K-SVD: Size variable dictionary learning for sparse representations," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 2149–2152.
[50] C. Rusu and B. Dumitrescu, "Stagewise K-SVD to design efficient dictionaries for sparse representations," *IEEE Signal Processing Letters*, vol. 19, no. 10, pp. 631–634, Oct. 2012.
[51] M. Marsousi, K. Abhari, P. Babyn, and J. Alirezaie, "An adaptive approach to learn overcomplete dictionaries with efficient numbers of elements," *IEEE Transactions on Signal Processing*, 2014.
[52] Q. Cheng, R. Chen, and T. Li, "Simultaneous wavelet estimation and deconvolution of reflection seismic signals," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 34, pp. 377–384, 1996.
[53] D. Ge, J. Idier, and E. Le Carpentier, "Enhanced sampling schemes for MCMC based blind Bernoulli-Gaussian deconvolution," *Signal Processing*, vol. 91, no. 4, pp. 759–772, Apr. 2011.

[54] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images," *IEEE Transactions on Image Processing*, 2012.
[55] H.-P. Dang and P. Chainais, "Indian buffet process dictionary learning : algorithms and applications to image processing," *International Journal of Approximate Reasoning*, 2017.
[56] K. Jiang, B. Kulis, and M. I. Jordan, "Small-variance asymptotics for exponential family Dirichlet process mixture models," in *Proceedings of Advances in neural information processing systems*, 2012.
[57] B. Kulis and M. I. Jordan, "Revisiting K-Means: New algorithms via Bayesian nonparametrics," in *Proceedings of the International Conference on Machine Learning (ICML)*, Madison, WI, USA, 2012, pp. 1131–1138.
[58] T. Broderick, B. Kulis, and M. Jordan, "MAD-Bayes: MAP-based Asymptotic Derivations from Bayes," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.
[59] A. Roychowdhury, K. Jiang, and B. Kulis, "Small-variance asymptotics for hidden Markov models," in *Proceedings of Advances in neural information processing systems*, vol. 26, 2013.
[60] Y. Wang and J. Zhu, "Small-variance asymptotics for Dirichlet process mixtures of SVMs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014, pp. 2135–2141.
[61] J. H. Huggins, K. Narasimhan, A. Saeedi, and V. K. Mansinghka, "JUMP-Means: Small-variance asymptotics for Markov jump processes," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015, pp. 693–701.
[62] Y. Wang and J. Zhu, "DP-space: Bayesian nonparametric subspace clustering with small-variance asymptotics," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 37, 2015, pp. 862–870.
[63] M. Pereyra and S. McLaughlin, "Fast unsupervised Bayesian image segmentation with adaptive spatial regularisation," *IEEE Transactions on Image Processing*, 2017.
[64] T. Campbell, B. Kulis, and J. How, "Dynamic clustering algorithms via small-variance analysis of Markov chain mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 06, pp. 1338–1352, 2019.
[65] C. Elvira, H.-P. Dang, and P. Chainais, "Small variance asymptotics and Bayesian nonparametrics for dictionary learning," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2018.
[66] H.-P. Dang, C. Elvira, and P. Chainais, "Vers une méthode d'optimisation non paramétrique pour l'apprentissage de dictionnaire en utilisant small-variance asymptotics pour modèle probabiliste," in *Proceedings of the Conférence sur l'Apprentissage automatique (CAP)*, 2018.
[67] H.-P. Dang and C. Elvira, "Parameter-free small variance asymptotics for dictionary learning," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2019.
[68] M. Pereyra and S. McLaughliny, "Small-variance asymptotics of hidden potts-mrfs: Application to fast bayesian image segmentation," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1597–1601.
[69] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," in *Proceedings of Advances in neural information processing systems*, 2006.
[70] Z. Ghahramani, T. L. Griffiths, and P. Sollich, "Bayesian nonparametric latent feature models," *Bayesian Statistic*, 2007.
[71] I. Tosic and P. Frossard, "Dictionary learning: What is the right representation for my signal," *IEEE Signal Processing Magazine*, 2011.