Variational Bayesian dropout: pitfalls and fixes

Jiri Hron [1] **Alexander G. de G. Matthews** [1] **Zoubin Ghahramani** [1 2]

Abstract

Dropout, a stochastic regularisation technique for training of neural networks, has recently been reinterpreted as a specific type of approximate inference algorithm for Bayesian neural networks. The main contribution of the reinterpretation is in providing a theoretical framework useful for analysing and extending the algorithm. We show that the proposed framework suffers from several issues; from undefined or pathological behaviour of the true posterior related to use of improper priors, to an ill-defined variational objective due to singularity of the approximating distribution relative to the true posterior. Our analysis of the improper log uniform prior used in variational Gaussian dropout suggests the pathologies are generally irredeemable, and that the algorithm still works only because the variational formulation annuls some of the pathologies. To address the singularity issue, we proffer Quasi-KL (QKL) divergence, a new approximate inference objective for approximation of high-dimensional distributions. We show that motivations for variational Bernoulli dropout based on discretisation and noise have QKL as a limit. Properties of QKL are studied both theoretically and on a simple practical example which shows that the QKL-optimal approximation of a full rank Gaussian with a degenerate one naturally leads to the Principal Component Analysis solution.

1. Introduction

Srivastava et al. (2014) proposed dropout as a cheap way of preventing Neural Networks (NN) from overfitting. This work was rather impactful and sparked large interest in studying and extending the algorithm. One strand of this research lead to reinterpretation of dropout as a form of approximate Bayesian variational inference (Kingma et al., 2015; Gal & Ghahramani, 2016; Gal, 2016).

There are two main reasons for attempting reinterpretation of an existing method: 1) providing a principled interpretation of the empirical behaviour; 2) extending the method based on the acquired insights. Variational Bayesian dropout has been arguably successful in meeting the latter criterion (Kingma et al., 2015; Gal, 2016; Molchanov et al., 2017). This paper thus focuses on the former by studying the theoretical soundness of variational Bayesian dropout and the implications for interpretation of the empirical results.

The first main contribution of our work is identification of two main sources of issues in current variational Bayesian dropout theory:

(a) use of improper or pathological prior distributions;

(b) singularity of the approximate posterior distribution.

As we describe in Section 3, the log uniform prior introduced in (Kingma et al., 2015) generally does not induce a proper posterior, and thus the reported sparsification (Molchanov et al., 2017) cannot be explained by the standard Bayesian and the related minimum description length (MDL) arguments. In this sense, sparsification via variational inference with log uniform prior falls into the same category of non-Bayesian approaches as, for example, Lasso (Tibshirani, 1996). Specifically, the approximate uncertainty estimates do not have the usual interpretation, and the model may exhibit overfitting. Consequently, we study the objective from a non-Bayesian perspective, proving that the optimised objective is impervious to some of the described pathologies due to the properties of the variational formulation itself, which might explain why the algorithm can still provide good empirical results.[1]

Section 4 shows how mismatch between support of the approximate and the true posterior renders application of the standard Variational Inference (VI) impossible by making the Kullback-Leibler (KL) divergence undefined. As the second main contribution, we address this issue by proving that the remedies to this problem proposed in (Gal & Ghahramani, 2016; Gal, 2016) are special cases of a broader

[1]Department of Engineering, University of Cambridge, Cambridge, United Kingdom [2]Uber AI Labs, San Francisco, California, USA. Correspondence to: Jiri Hron <jh2084@cam.ac.uk>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1]An earlier version of this work was published in (Hron et al., 2017).

class of limiting constructions leading to a unique objective which we name Quasi-KL (QKL) divergence.

Section 5 provides initial discussion of QKL's properties, uses those to suggest an explanation for the empirically observed difficulty in tuning hyperparameters of the true model (e.g. Gal (2016, p. 119)), and demonstrates the potential of QKL on an illustrative example where we try to approximate a full rank Gaussian distribution with a degenerate one using QKL, only to arrive at the well known Principal Component Analysis (PCA) algorithm.

2. Background

Assume we have a discriminative probabilistic model $y \,|\, x, \boldsymbol{W} \sim \mathrm{P}(y \,|\, x, \boldsymbol{W})$ where (x, y) is a single input-output pair, and $\boldsymbol{W}$ is the set of model parameters generated from a prior distribution $\mathrm{P}(\boldsymbol{W})$. In Bayesian inference, we usually observe a set of data points $(\boldsymbol{X}, \boldsymbol{Y}) = \{(x_n, y_n)\}_{n=1}^{\mathrm{N}}$ and aim to infer the posterior $p(\boldsymbol{W} \,|\, \boldsymbol{X}, \boldsymbol{Y}) \propto p(\boldsymbol{W}) \prod_n p(y_n \,|\, x_n, \boldsymbol{W})$,[2] which can be subsequently used to obtain the posterior predictive density $p(\boldsymbol{Y}' \,|\, \boldsymbol{X}', \boldsymbol{X}, \boldsymbol{Y}) = \int p(\boldsymbol{Y}' \,|\, \boldsymbol{X}', \boldsymbol{W}) p(\boldsymbol{W} \,|\, \boldsymbol{X}, \boldsymbol{Y}) \mathrm{d}\boldsymbol{W}$. If $p(y \,|\, x, \boldsymbol{W})$ is a complicated function of $\boldsymbol{W}$ like a neural network, both tasks often become computationally infeasible and thus we need to turn to approximations.

Variational inference approximates the posterior distribution over a set of latent variables $\boldsymbol{W}$ by maximising the evidence lower bound (ELBO),

$$\mathcal{L}(q) = \mathop{\mathbb{E}}_{\mathrm{Q}(\boldsymbol{W})} [\log p(\boldsymbol{Y} \,|\, \boldsymbol{X}, \boldsymbol{W})] - \mathrm{KL}\left(\mathrm{Q}(\boldsymbol{W}) \| \,\mathrm{P}(\boldsymbol{W})\right),$$

with respect to (w.r.t.) an approximate posterior $\mathrm{Q}(\boldsymbol{W})$. If $\mathrm{Q}(\boldsymbol{W})$ is parametrised by ψ and the ELBO is differentiable w.r.t. ψ, VI turns inference into optimisation. We can then approximate the density of posterior predictive distribution using $q(\boldsymbol{Y}' \,|\, \boldsymbol{X}', \boldsymbol{X}, \boldsymbol{Y}) = \int p(\boldsymbol{Y}' \,|\, \boldsymbol{X}', \boldsymbol{W}) q(\boldsymbol{W}) \mathrm{d}\boldsymbol{W}$, usually by Monte Carlo integration.

A particular discriminative probabilistic model is a Bayesian neural network (BNN). BNN differs from a standard NN by assuming a prior over the weights $\boldsymbol{W}$. One of the main advantages of BNNs over standard NNs is that the posterior predictive distribution can be used to quantify uncertainty when predicting on previously unseen data $(\boldsymbol{X}', \boldsymbol{Y}')$. However, there are at least two challenges in doing so:

1) difficulty of reasoning about choice of the prior $\mathrm{P}(\boldsymbol{W})$;

2) intractability of posterior inference.

For a subset of architectures and priors, Item 1 can be addressed by studying limit behaviour of increasingly large networks (see, for example, (Neal, 1996; Matthews et al., 2018)); in other cases, sensibility of $\mathrm{P}(\boldsymbol{W})$ must be assessed individually. Item 2 necessitates approximate inference – a particular type of approximation related to dropout, the topic of this paper, is described below.

Dropout (Srivastava et al., 2014) was originally proposed as a regularisation technique for NNs. The idea is to multiply inputs of a particular layer by a random noise variable which should prevent co-adaptation of individual neurons and thus provide more robust predictions. This is equivalent to multiplying the rows of the subsequent weight matrix by the same random variable. The two proposed noise distributions were Bernoulli(p) and Gaussian $\mathcal{N}(1, \alpha)$.

Bernoulli and Gaussian dropout were later respectively reinterpreted by Gal & Ghahramani (2016) and Kingma et al. (2015) as performing VI in a BNN. In both cases, the approximate posterior is chosen to factorise either over rows or individual entries of the weight matrices. The prior usually factorises in the same way, mostly to simplify calculation of $\mathrm{KL}\left(\mathrm{Q}(\boldsymbol{W}) \| \,\mathrm{P}(\boldsymbol{W})\right)$. It is the choice of the prior and its interaction with the approximating posterior family that is studied in the rest of this paper.

3. Improper and pathological posteriors

Both Gal & Ghahramani (2016) and Kingma et al. (2015) propose using a prior distribution factorised over individual weights $w \in \boldsymbol{W}$. While the former opts for a zero mean Gaussian distribution, Kingma et al. (2015) choose to construct a prior for which $\mathrm{KL}\left(\mathrm{Q}(\boldsymbol{W}) \| \,\mathrm{P}(\boldsymbol{W})\right)$ is independent of the mean parameters $\boldsymbol{\theta}$ of their approximate posterior $q(w) = \phi_{\theta, \alpha\theta^2}(w)$, $w \in \boldsymbol{W}, \theta \in \boldsymbol{\theta}$, where ϕ_{μ, σ^2} is the density function of $\mathcal{N}(\mu, \sigma^2)$. The decision to pursue such independence is motivated by the desire to obtain an algorithm that has no weight shrinkage – that is to say one where Gaussian dropout is the sole regularisation method. Indeed, the authors show that the log uniform prior $p(w) \coloneqq \mathrm{C}/|w|$ is the only one where $\mathrm{KL}\left(\mathrm{Q}(\boldsymbol{W}) \| \,\mathrm{P}(\boldsymbol{W})\right)$ has this mean parameter independence property. The log uniform prior is equivalent to a uniform prior on $\log|w|$. It is an improper prior (Kingma et al., 2015, p. 12) which means that there is no $\mathrm{C} \in \mathbb{R}$ for which $p(w)$ is a valid probability density.

Improper priors can sometimes lead to proper posteriors (e.g. normal Jeffreys prior for Gaussian likelihood with unknown mean and variance parameters) if C is treated as a positive finite constant and the usual formula for computation of posterior density is applied. We show this is generally not the case for the log uniform prior, and that any remedies in the form of proper priors that are in some sense close to the log uniform (such as uniform priors over floating point numbers) will lead to severely pathological inferences.

[2] Throughout the paper, $\mathrm{P}(\boldsymbol{W})$ refers to the distribution and $p(\boldsymbol{W})$ to its density function. Analogously for other distributions.

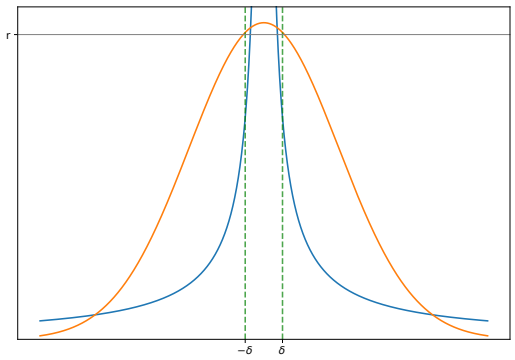

Figure 1. Illustration of Proposition 1. Blue is the prior, orange the likelihood, and green shows a particular neighbourhood of $w = 0$ where the likelihood is greater than $r > 0$ (such neighbourhood exists by the continuity). Integral of the likelihood over $(-\delta, \delta)$ w.r.t. $\mathrm{P}(w)$ diverges because the likelihood can be lower bounded by $r > 0$ and the prior assigns infinite mass to this neighbourhood.

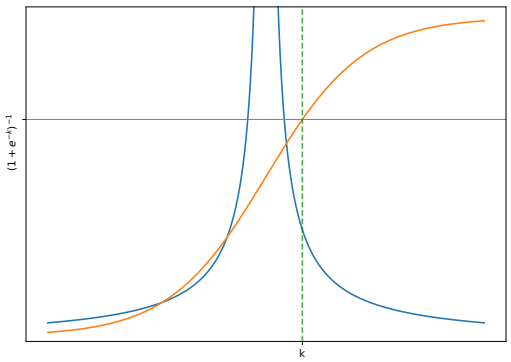

Figure 2. Visualisation of the infinite tail mass example. Blue is the prior, orange the sigmoid likelihood, and green shows the lower bound of the $[k, \infty)$ interval. The sigmoid function is greater than zero for any $k > 0$. The integral of the likelihood over $[k, \infty)$ w.r.t. $\mathrm{P}(w)$ can thus again be lower bounded by a diverging integral.

3.1. Pathologies of the log uniform prior

For any proper posterior density, the normaliser $\mathrm{Z} = \int_{\mathbb{R}^{\mathrm{D}}} p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) p(\boldsymbol{W}) \mathrm{d}\boldsymbol{W}$ has to be finite (D denotes the total number of weights). We will now show that this requirement is generally not satisfied for the log uniform prior combined with commonly used likelihood functions.

Proposition 1. *Assume the log uniform prior is used and that there exists some $w \in \boldsymbol{W}$ such that the likelihood function at $w = 0$ is continuous in w and non-zero. Then the posterior is improper.*

All proofs can be found in the appendix. Notice that standard architectures with activations like rectified linear or sigmoid, and Gaussian or Categorical likelihood satisfy the above assumptions, and thus the posterior distribution for non-degenerate datasets will generally be improper. See Figure 1 for a visualisation of this case.

Furthermore, the pathologies are not limited to the region near $w = 0$, but can also arise in the tails (Figure 2). As an example, we will consider a single variable Bayesian logistic regression problem $p(y \mid x, w) = 1/(1 + \exp(-xw))$, and again use the log uniform prior for w. For simplicity, assume that we have observed $(x = 1, y = 1)$ and wish to infer the posterior distribution. To show that the right tail has infinite mass, we integrate over $[k, \infty), k > 0$,

$$\int_{[k,\infty)} p(w) p(y \mid x, w) \mathrm{d}w = \int_{[k,\infty)} \frac{\mathrm{C}}{|w|} \frac{1}{1 + \exp(-w)} \mathrm{d}w$$
$$> \int_{[k,\infty)} \frac{\mathrm{C}}{|w|} \frac{1}{1 + \exp(-k)} \mathrm{d}w = \frac{\mathrm{C} \cdot (\infty - \log k)}{1 + \exp(-k)} = \infty \,.$$

Equivalently, we could have obtained infinite mass in the left tail, for example by taking the observation to be $(x = -1, y = 1)$. Because the sigmoid function is continuous and equal to $1/2$ at $w = 0$, the posterior also has infinite mass around the origin, exemplifying both of the discussed degeneracies. The normalising constant is of course still infinite and thus the posterior is again improper.

The practical implication of these pathologies is that even tasks as simple as MAP estimation (Proposition 1 implies unbounded posterior density) or posterior mean estimation will fail as the target is undefined. In general, improper posteriors lead to undefined or incoherent inferences. The above shows that this is the case for the log uniform prior combined with BNNs and related models, making Bayesian inference, exact and approximate, ill-posed.

3.2. Pathologies of the truncated log uniform prior

Neklyudov et al. (2017) proposed to swap the log uniform prior on $(-\infty, \infty)$ for a distribution that is uniform on a sufficiently wide bounded interval in the $\log|w|$ space (will be referred to as *the log space* from now on), i.e. $p(\log|w|) = 1/(b-a)\mathbb{I}_{[a,b]}(w), a < b$ where $\mathbb{I}_A$ is the indicator function of the set A. This prior can be used in place of the log uniform if the induced posteriors in some sense converge to a well-defined limit for any dataset as $[a, b]$ gets wider. If this is not the case, choice of $[a, b]$ becomes a prior assumption and must be justified as such because different choices will lead to sometimes considerably different inferences. We now show that posteriors generally do not converge for the truncated log uniform prior and discuss some of the related pathologies of the induced exact posterior.

To illustrate the considerable effect the choice of $[a, b]$ might have, we return to the example of posterior inference in a logistic regression model $p(y \mid x, w) = 1/(1 + \mathrm{e}^{-xw})$ after observing $(x = 1, y = 1)$, using the prior $p_n(w) =$

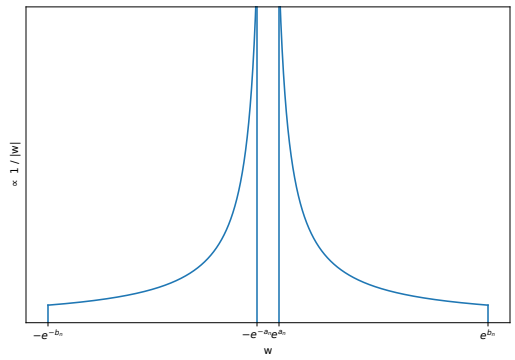

Figure 3. A truncated log uniform prior transformed to the original space. Notice that the support gap around the origin narrows as $a_n \to -\infty$, and the tail support expands as $b_n \to \infty$ which yields the more pathological inferences the wider $[a_n, b_n]$ gets.

$\mathbb{I}_{I_n}(w)\,\mathrm{C}_n/|w|$ where $I_n = [-\mathrm{e}^{b_n}, -\mathrm{e}^{a_n}] \cup [\mathrm{e}^{a_n}, \mathrm{e}^{b_n}]$ (i.e. the appropriate transformation of the closed interval $[a_n, b_n]$ from *the log space* – see Figure 3). We exemplify the sensitivity of the posterior distribution to the choice of the $(I_n)_{n\in\mathbb{N}}$ sequence by studying the limiting behaviour of the posterior mean and variance. Using the definition of $\mathbb{I}_{I_n}(w)$ and symmetry, the normaliser of the posterior is,

$$\begin{aligned}\mathrm{Z}_n &= \int_{-\mathrm{e}^{b_n}}^{-\mathrm{e}^{a_n}} \frac{1}{|w|}\frac{1}{1+\mathrm{e}^{-w}}\mathrm{d}w + \int_{\mathrm{e}^{a_n}}^{\mathrm{e}^{b_n}} \frac{1}{|w|}\frac{1}{1+\mathrm{e}^{-w}}\mathrm{d}w \\ &= \int_{\mathrm{e}^{a_n}}^{\mathrm{e}^{b_n}} \frac{1}{|w|}\frac{1+\mathrm{e}^{w}}{1+\mathrm{e}^{w}}\mathrm{d}w = b_n - a_n\,.\end{aligned}$$

Similar ideas can be used to derive the first two moments,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathrm{P}_n}(w) &= \frac{\int_{\mathrm{e}^{a_n}}^{\mathrm{e}^{b_n}} \frac{1}{1+\mathrm{e}^{-w}}\mathrm{d}w - \int_{-\mathrm{e}^{b_n}}^{-\mathrm{e}^{a_n}} \frac{1}{1+\mathrm{e}^{-w}}\mathrm{d}w}{b_n - a_n} \\ &= \frac{h(\mathrm{e}^{b_n}) + h(-\mathrm{e}^{b_n}) - h(\mathrm{e}^{a_n}) - h(-\mathrm{e}^{a_n})}{b_n - a_n}\,, \quad (1)\end{aligned}$$

$$\mathop{\mathbb{E}}_{\mathrm{P}_n}(w^2) = \int_{\mathrm{e}^{a_n}}^{\mathrm{e}^{b_n}} \frac{|w|}{b_n - a_n}\frac{1+\mathrm{e}^{w}}{1+\mathrm{e}^{w}}\mathrm{d}w = \frac{\mathrm{e}^{2b_n} - \mathrm{e}^{2a_n}}{2(b_n - a_n)}\,, \quad (2)$$

where $h(x) \coloneqq \log(1+\mathrm{e}^{x})$, and P_n stands for $\mathrm{P}_n(w \mid x, y)$. To understand sensitivity of the posterior mean to the choice of $(I_n)_{n\in\mathbb{N}}$, we now construct sequences which respectively lead to convergence of the mean to zero, an arbitrary positive constant, and infinity.[3] To emphasise this is not specific to the posterior mean, we show that the variance might equally well be zero, infinite, or undefined.

To get $\lim_{n\to\infty} \mathbb{E}_{\mathrm{P}_n}(w) = 0$, notice that for a fixed b_n, the second term in Equation (1) tends to $\log(4)/\infty = 0$. Hence we can make the posterior mean converge to zero by making the first term also tend to zero; a way to achieve this is setting $b_n = \log(\log|a_n|)$, which tends to infinity as $a_n \to \infty$. The limit of Equation (2) for the same sequence, and thus the variance, tends to zero as well.

For $\lim_{n\to\infty} \mathbb{E}_{\mathrm{P}_n}(w) = c > 0$, we again focus on the first term in Equation (1) as the second term tends to zero for any increasing sequence $I_n \nearrow \mathbb{R}$. Simple algebra shows that for any diverging sequence $b_n \to \infty$, taking $a_n = b_n - \mathrm{e}^{b_n}/c$ yields the desired result. The same sequence leads to infinite second moment and thus to infinite variance.

Finally, a choice which results in infinite mean and thus undefined variance is setting $a_n = -b_n$, for which the mean grows as e^{b_n}/b_n. We would like to point out that this symmetric growth of a_n with b_n is of particular interest as it corresponds to changing between different precisions of the float format representation on the computer as considered in Kingma et al. (2015, Appendix A).

3.3. Variational Gaussian dropout as penalised maximum likelihood

We have established that optimisation of the ELBO implied by a BNN with log uniform prior over its weights cannot generally be interpreted as a form of approximate Bayesian inference. Nevertheless, the reported empirical results suggest that the objective might possess reasonable properties. We thus investigate if and how the pathologies of the true posterior translate into the variational objective as used in (Kingma et al., 2015; Molchanov et al., 2017).

Firstly, we derive a new expression for $\mathrm{KL}\left(\mathrm{Q}(w)\|\,\mathrm{P}(w)\right)$, and for its derivative w.r.t. the variational parameters, which will help us with further analysis.

Proposition 2. *Let $q(w) = \phi_{\mu,\sigma^2}(w)$, and $p(w) = \mathrm{C}/|w|$. Denote $u \coloneqq \mu^2/(2\sigma^2)$. Then,*

$$\begin{aligned}&\mathrm{KL}\left(\mathrm{Q}(w)\|\,\mathrm{P}(w)\right) \\ &= \textit{const.} + \frac{1}{2}\left(\log 2 + \mathrm{e}^{-u}\sum_{k=0}^{\infty}\frac{u^k}{k!}\psi(1/2+k)\right) \quad (3)\end{aligned}$$

$$= \textit{const.} - \frac{1}{2}\left.\frac{\partial \mathrm{M}(a;1/2;-u)}{\partial a}\right|_{a=0}\,, \quad (4)$$

where $\psi(x)$ denotes the digamma function, and $\mathrm{M}(a;b;z)$ the Kummer's function of the first kind.

We can obtain gradients w.r.t. μ and σ^2 using,

$$\nabla_u \mathrm{KL}\left(\mathrm{Q}(w)\|\,\mathrm{P}(w)\right) = \begin{cases} 1 & u = 0 \\ \frac{D_+(\sqrt{u})}{\sqrt{u}} & u > 0 \end{cases}\,, \quad (5)$$

and the chain rule; $D_+(x)$ is the Dawson integral. The derivative is continuous in u on $[0, \infty)$.

[3] It would be equally possible to get convergence to an arbitrary negative constant, and negative infinity if the observation was $(x = -1, y = 1)$.

Before proceeding, we note that Equation (5) is sufficient to implement first order gradient-based optimisation, and thus can be used to replace the approximations used in (Kingma et al., 2015; Molchanov et al., 2017). Note that numerically accurate implementations of the $D_+(x)$ exist in many programming languages (e.g. (Johnson, 2012)).

In VI literature, the term $\mathrm{KL}\left(\mathrm{Q}(w) \| \mathrm{P}(w)\right)$ is often interpreted as a regulariser, constraining $\mathrm{Q}(w)$ from concentrating at the maximum likelihood estimate which would be optimal w.r.t. the other term $\mathbb{E}_{\mathrm{Q}(\boldsymbol{W})}[\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W})]$ in the ELBO. It is thus natural to ask what effect this term has on the variational parameters. Noticing that only the infinite sum in Equation (3) depends on these parameters, and that the first summand is always equal to $\psi(1/2)$, we can focus on terms corresponding to $k \geq 1$. Because $\psi(1/2 + k) > 0, \forall k \geq 1$, all summands are non-negative. Hence the penalty will be minimised if $\mu^2/(2\sigma^2) = 0$, i.e. when $\mu = 0$ and/or $\sigma^2 \to \infty$; Corollary 3 is sufficient to establish that this minimum is unique.

Corollary 3. *Under assumptions of Proposition 2,* $\mathrm{KL}\left(\mathrm{Q}(w) \| \mathrm{P}(w)\right)$ *is strictly increasing for* $u \in [0, \infty)$.

Sections 3.1 and 3.2 suggests the pathological behaviour is non-trivial to remove unless we replace the (truncated) log uniform prior.[4] An alternative route is to interpret optimisation of the variational objective from above as a type of penalised maximum likelihood estimation.

Proposition 2 and Corollary 3 suggest that the variational formulation cancels the pathologies of the true posterior distribution which both invalidates the Bayesian interpretation, but also means that the algorithm may perform well in terms of accuracy and other metrics of interest. Since the $\mathrm{KL}\left(\mathrm{Q}(\boldsymbol{W}) \| \mathrm{P}(\boldsymbol{W})\right)$ regulariser will force the mean parameters to be small, and the variances to be large, and the $\mathbb{E}_{\mathrm{Q}(\boldsymbol{W})}[\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W})]$ will generally push the parameters towards the maximum likelihood solution, the resulting fit might have desirable properties if the right balance between the two is struck. As the Bayesian interpretation no longer applies, the balance can be freely manipulated by reweighing the KL by any positive constant. The strict page limit and desire to discuss the singularity issue lead us to leave exploration of this direction to future work.

4. Approximating distribution singularities

Both the Bernoulli and Gaussian dropout can be seen as members of a larger family of algorithms where individual layer inputs are perturbed by elementwise i.i.d. random noise. This is equivalent to multiplying the corresponding row $\boldsymbol{w}_i$ of the subsequent weight matrix by the same noise variable. One could thus define $\boldsymbol{w}_i = s_i \boldsymbol{\theta}_i$, $s_i \sim \mathrm{Q}(s_i)$, $\mathrm{Q}(s_i)$ being an arbitrary distribution, and treat the induced distribution over $\boldsymbol{w}_i$ as an approximate posterior $\mathrm{Q}(\boldsymbol{w}_i)$.

[4] Louizos et al. (2017) made promising progress there.

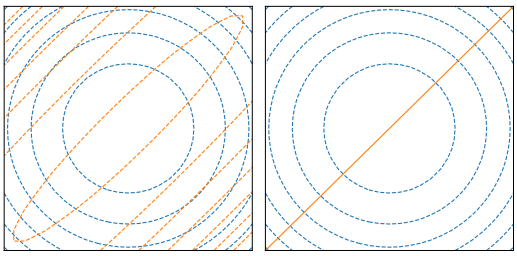

Figure 4. Illustration of approximating distribution singularities. On the left, blue is the standard and orange a correlated Gaussian density. *Null sets*, are (Borel) sets with zero measure under a distribution. Since both distributions have the same null sets, they are absolutely continuous w.r.t. each other. On the right, orange now represents a degenerate Gaussian supported on a line. Blue assigns zero probability to the line whereas orange assigns all of its mass; orange assigns probability zero to any set excluding the line but blue does not. Hence neither is absolutely continuous w.r.t. the other, and thus KL-divergence is undefined.

An issue with this approach is that it leads to undefined $\mathrm{KL}\left(\mathrm{Q}(\boldsymbol{W}) \| \mathrm{P}(\boldsymbol{W} \mid \boldsymbol{X}, \boldsymbol{Y})\right)$ whenever the prior assigns zero mass to the individual directions defined by $\boldsymbol{\theta}$. To understand why, note that $\mathrm{KL}\left(\mathrm{Q}(\boldsymbol{W}) \| \mathrm{P}(\boldsymbol{W} \mid \boldsymbol{X}, \boldsymbol{Y})\right)$ is defined only if $\mathrm{Q}(\boldsymbol{W})$ is *absolutely continuous* w.r.t. $\mathrm{P}(\boldsymbol{W} \mid \boldsymbol{X}, \boldsymbol{Y})$ which means that whenever $\mathrm{P}(\boldsymbol{W} \mid \boldsymbol{X}, \boldsymbol{Y})$ assigns probability zero to a particular set, $\mathrm{Q}(\boldsymbol{W})$ does so too. The right-hand side plot in Figure 4 shows a simple example of the case where neither distribution is absolutely continuous w.r.t. the other: the blue Gaussian assigns zero mass to any set with Lebesgue measure zero, such as the line along which the orange distribution places all its mass, and thus the orange Gaussian distribution is not absolutely continuous w.r.t. the blue one. This example is relevant to our problem from above, where $\mathrm{Q}(\boldsymbol{w}_i)$ always assigns all its mass to along the direction defined by the vector $\boldsymbol{\theta}_i$. For more details, see for example (Matthews, 2016, Section 2.1). When a measure is not absolutely continuous w.r.t. another measure, it can be shown to have a so called *singular* component relative to that measure, which we use as a shorthand for referring to this issue. Consequences for variational Bayesian interpretations of dropout are discussed next.

4.1. Implications for Bayesian dropout interpretations

Section 3.2 in (Kingma et al., 2015) proposes to use a shared Gaussian random variable for whole rows of the posterior weight matrices. Specifically $s_i \sim \mathcal{N}(1, \alpha)$ is substituted for $\mathrm{Q}(s_i)$ in the generic algorithm described in the previous section. We call such behaviour in the context of variational inference an *approximating distribution singularity*. The singularity has two possible negative consequences.

First, if only the s_i scalars are treated as random variables, $\boldsymbol{\theta}$ become parameters of the discriminative model instead of the variational distribution. Optimisation of the ELBO will yield a valid Bayesian posterior approximation for the s_i. The lack of regularisation of $\boldsymbol{\theta}$ might lead to significant overfitting though, as $\boldsymbol{\theta}$ represent all weights in the BNN.

Second, if the fully factorised log uniform prior is used as before, then the directions defined by $\boldsymbol{\theta}$ constitute a measure zero subspace of $\mathbb{R}^{\mathrm{D}}$, and thus the $\mathrm{KL}\left(\mathrm{Q}(\boldsymbol{W})\|\,\mathrm{P}(\boldsymbol{W})\right)$ and consequently $\mathrm{KL}\left(\mathrm{Q}(\boldsymbol{W})\|\,\mathrm{P}(\boldsymbol{W}\mid\boldsymbol{X},\boldsymbol{Y})\right)$ are undefined for any configuration of $\boldsymbol{\theta}$. This is an instance of the issue described in the previous section. As a consequence, standard variational inference with this approximating family and target posterior is impossible.

A similar problem is encountered in (Gal & Ghahramani, 2016; Gal, 2016). The approximate posterior is defined as $\mathrm{Q}(\boldsymbol{w}_i) = p\,\delta_{\mathbf{0}} + (1-p)\,\delta_{\boldsymbol{\theta}_i}$ for each row in every weight matrix. The assumed prior is a product of independent non-degenerate Gaussian distributions which by definition assigns non-zero mass only to sets of positive Lebesgue measure. Again, the approximate posterior is not absolutely continuous w.r.t. the prior and thus the KL is undefined.

To address this issue, Gal & Ghahramani (2016) propose to replace the Dirac deltas in $\mathrm{Q}(\boldsymbol{w}_i)$ by Gaussian distributions with small but non-zero noise (we call this the *convolutional approach*). As an alternative, Gal (2016) proposes to instead discretise the Gaussian prior and the approximate posterior so both assign positive mass only to a shared finite set of values. Because the discretised Gaussian assigns non-zero mass to all points in the set, the approximate posterior is absolutely continuous w.r.t. this prior (we refer to this as the *discretisation approach*).

Strictly speaking, the two approaches cannot be equivalent because the corresponding random variables take values in distinct measurable spaces ($\mathbb{R}^{\mathrm{D}}$ and a discrete grid respectively). Notwithstanding, both approaches are claimed to lead to the same optima for the variational parameters.[5] The suggested method for addressing this discrepancy is to introduce a *continuous relaxation* (Gal, 2016, p. 119) of the optimisation problem for the discrete case. The precise details of this relaxation are not given. One could define it as the relaxation that satisfied the required *KL-condition* (Gal, 2016, Appendix A), but there is of course then a risk of a circular argument. Putting these intuitive arguments on a firmer footing is one motivation for what follows here.

In the light of Section 3.2, it is natural to ask whether either of the proposed approaches will tend to a stable objective as the added noise shrinks to zero, and the discretisation becomes increasingly refined, respectively for the convolutional and discretisation approaches. Theorem 4 provides an affirmative answer by proving that both approaches lead to the same limit under reasonable assumptions.[6]

Theorem 4. *Let* Q, P *be Borel probability measures on* $\mathbb{R}^{\mathrm{D}}$, P *with a continuous density* p *w.r.t. the* D*-dimensional Lebesgue measure, and* Q *supported on an at most countable measurable set* $S \subset \mathbb{Q}^{\mathrm{D}}$*, with density* q *w.r.t. the counting measure on* $\mathbb{Q}^{\mathrm{D}}$*. If* S *is infinite, further assume that* $\operatorname{diam}(S) < \infty$*, i.e.* $\sup_{x,y\in S}\|x-y\|_2 < \infty$*.*

Then there exists a sequence $(s^{(n)}) \subset \mathbb{R}$ *independent of* Q *and* P *s.t. the limit for both the sequences of convolved and discretised distributions* $\{(\mathrm{Q}^{(n)}, \mathrm{P}^{(n)})\}_{n\in\mathbb{N}}$,[7]

$$\lim_{n\to\infty}\left\{\mathrm{KL}\left(\mathrm{Q}^{(n)}\|\,\mathrm{P}^{(n)}\right) - s^{(n)}\right\} = \mathop{\mathbb{E}}_{\mathrm{Q}}\left(\log\tfrac{q}{p}\right), \tag{6}$$

given the perturbation noise is Gaussian and eventually shrinks to zero, and that the discretisation creates ever finer grid with equally sized cells as $n \to \infty$*. The sequence* $(s^{(n)})$ *tends to 0 if* $\mathrm{Q} \ll \mathrm{P}$ *and to infinity otherwise.*

The right-hand side (r.h.s.) of Equation (6) satisfies Gal's *KL condition*, i.e. it leads to the same optimisation problem and thus unifies the convolutional and discretisation approach.

Unlike in (Gal, 2016, Appendix A), our derivation does not make an extraneous assumption on the distribution over any function of the $\boldsymbol{\theta}$ parameters nor does it require that the expectation of $\|\boldsymbol{\theta}_i\|_2^2$ grows without bounds with $\dim(\boldsymbol{\theta}_i)$. Neither of these two assumptions is sure to hold in practice as $\boldsymbol{\theta}$ are being optimised, and $\boldsymbol{\theta}_i$ in any modern (B)NN is initially scaled by $\sqrt{\dim(S)}$ exactly to achieve approximately constant Euclidean norm irrespective of the dimension.

We explored whether Equation (6) holds more generally. Theorem 5 extends the convolutional approach to a considerably larger class of approximating distributions.

Theorem 5. *Let* Q, P *be Borel probability measures on* $\mathbb{R}^{\mathrm{D}}$*,* P *with a bounded continuous density* p *w.r.t. the Lebesgue measure on* $\mathbb{R}^{\mathrm{D}}$*, and* Q *supported on a measurable linear manifold* $S \subset \mathbb{R}^{\mathrm{D}}$ *of (Hamel) dimension* K_S*. Assume* Q *has a continuous bounded density* q *w.r.t. the Lebesgue measure on* S*, where the continuity is w.r.t. the trace topology.*

Then there exists a sequence $(s^{(n)}) \subset \mathbb{R}$ *dependent only on* K_S *s.t. the following holds for the convolutional approach,*

$$\lim_{n\to\infty}\left\{\mathrm{KL}\left(\mathrm{Q}^{(n)}\|\,\mathrm{P}^{(n)}\right) - s^{(n)}_{\mathrm{K}_S}\right\} = \mathop{\mathbb{E}}_{\mathrm{Q}}\left(\log\tfrac{q}{p}\right), \tag{7}$$

given the perturbation noise is Gaussian and eventually shrinks to zero. The sequence $(s^{(n)}_{\mathrm{K}_S})$ *tends to 0 if* $\mathrm{Q} \ll \mathrm{P}$ *and to infinity otherwise.*

[5] Modulo the Euclidean distance to a closest point in the finite set for the discretisation approach.

[6] We state only the most important assumptions in Theorems 4 and 5. **Please see the appendix for the full set of assumptions**.

[7] $\mathrm{P}^{(n)} = \mathrm{P}\ , \forall n \in \mathbb{N}$, in the convolutional case.

A result related to Theorem 5 for the discretisation approach can be derived under assumptions similar to Theorem 4 with one important difference: $(s^{(n)}_{\mathrm{K}_S})$, if it exists, is affected not only by K_S, but also by the orientation of S in $\mathbb{R}^{\mathrm{D}}$. This is because the dominating Lebesgue measure is different for each affine subspace S and thus, unlike in the countable support case, q cannot be defined w.r.t. a single dominating measure. Implicit in Theorems 4 and 5 is that the same constant can be subtracted from $\mathrm{KL}\left(\mathrm{Q}^{(n)} \| \, \mathrm{P}^{(n)}\right)$ for all distributions Q with the same type of support. Hence if we are optimising over a family of singular approximating distributions, the sequence $(s^{(n)})$ (resp. $(s^{(n)}_{\mathrm{K}_S})$) does not need to change between updates to obtain the desired limit.

Before moving to Section 5 which discusses some of the merits of using Equations (6) and (7) as an objective for approximate Bayesian inference, let us make two comments.

First, taking the limit makes the decision about size of perturbation or coarseness of the discretisation unnecessary. The sequences used do not cause the same instability problems discussed in Section 3.2 because the true posterior is well-defined even in the limit, which we assume in saying that P is a probability measure. The main open question is thus whether optimisation of the r.h.s. of Equation (6) will yield a sensible approximation of this posterior.

Second, if there is a family of approximate posterior distributions $\mathcal{Q}$ parametrised by $\psi \in \Psi$, the equality,

$$\operatorname*{argmin}_{\psi \in \Psi} \mathop{\mathbb{E}}_{\mathrm{Q}_\psi} \left(\log \tfrac{q_\psi}{p}\right) = \lim_{n \to \infty} \operatorname*{argmin}_{\psi \in \Psi} \mathrm{KL}\left(\mathrm{Q}^{(n)}_\psi \| \, \mathrm{P}^{(n)}\right), \tag{8}$$

need not hold unless stricter conditions are assumed. Equation (8) is of interest in cases when $\mathrm{KL}\left(\mathrm{Q}^{(n)}_\psi \| \, \mathrm{P}^{(n)}\right)$ has some desirable properties (e.g. good predictive performance) which we would like to preserve. However, this is not the case for variational Bernoulli dropout as the objective being used by Gal & Ghahramani (2016) is, in terms of gradients w.r.t. the variational parameters, identical to the limit.

Furthermore, we can view both the discretisation and convolutional approaches as mere alternative vehicles to derive the same quasi discrepancy measure (cf. Section 5). If this quasi discrepancy possesses favourable properties, the precise details of optima attained along the sequence might be less important. One benefit of this view is in avoiding arguments like the previously mentioned *continuous relaxation* (Gal, 2016, p. 119).

5. Quasi-KL divergence

The r.h.s. of Equations (6) and (7) is markedly similar to the formula for standard KL divergence. We now make this link explicit. If $\mathrm{Z}_{\mathrm{P}_S} := \int_S p \, \mathrm{d}m_S < \infty$, m_S being either the counting or the Lebesgue measure dominating measure for q, we can the probability density $p_S := p/\mathrm{Z}_{\mathrm{P}_S}$, and denote the corresponding distribution on $(S, \mathcal{B}_S)$ by P_S. We term Equation (9) the *Quasi-KL* (QKL) divergence,

$$\mathrm{QKL}\left(\mathrm{Q} \| \, \mathrm{P}\right) := \mathop{\mathbb{E}}_{\mathrm{Q}} \left(\log \tfrac{q}{p}\right) = \mathrm{KL}\left(\mathrm{Q} \| \, \mathrm{P}_S\right) - \log \mathrm{Z}_{\mathrm{P}_S}. \tag{9}$$

Taking Equation (9) as a loss function says that we would like to find such a Q for which the KL divergence between Q and P_S is as small as possible, while making sure that the corresponding set S runs through high density regions of P, preventing Q from collapsing to subspaces where p is easily approximated by q but takes low values. Since p is continuous (c.f. Theorem 4), values of p roughly indicate how much mass P assigns to the region where S is placed.

Standard KL divergence and QKL are equivalent when $\mathrm{Q} \ll \mathrm{P}$ and the two distributions have the same support. QKL is not a proper statistical divergence though, as it is lower bounded by $-\log \mathrm{Z}_{\mathrm{P}_S}$ instead of zero. The non-negativity could have been satisfied by defining QKL as $\mathrm{KL}\left(\mathrm{Q} \| \, \mathrm{P}_S\right)$, dropping the $\log \mathrm{Z}_{\mathrm{P}_S}$ term. However, this would mean losing the above discussed effect of forcing S to lie in a relatively high density region of P, and also the motivation of being a limit of the two sequences considered in Theorem 4.

Nevertheless, QKL inherits some of the attractive properties of KL divergence: the density p need only be known up to a constant, the reparameterisation trick (Kingma & Welling, 2014) and analogical approaches for discrete random variables (Maddison et al., 2017; Jang et al., 2017; Tucker et al., 2017) still apply, and stochastic optimisation and integral approximation techniques can be deployed if desired.

On a more cautionary note, we emphasise that $\mathbb{E}_{\mathrm{Q}}(\log \frac{p}{q})$ is upper bounded by $\log \mathrm{Z}_{\mathrm{P}_S}$ and not the log marginal likelihood as is the case for standard KL use in VI. Hence optimisation of this objective w.r.t. hyperparameters of P need not work very well, since the resulting estimates could be biased towards regions where the variational family performs best.[8] This might explain why prior hyperparameters usually have to be found by validation error based grid search (Gal, 2016, e.g. p. 119) instead of ELBO optimisation as is common in the sparse Gaussian Process literature (Titsias, 2009).

Whether and when is QKL an attractive alternative to the more computationally expensive but proper statistical discrepancy measures which are capable of handling singular distributions (e.g. Wasserstein distances) is beyond the scope of this paper. To provide basic intuition of whether QKL might be a sensible objective for inference, Section 5.1 focuses on a simple practical example that yields a well known algorithm as the optimal solution to QKL optimisation, and exemplifies some of the above discussed behaviour.

[8] A similar issue for KL was observed by Turner et al. (2010).

5.1. QKL and Principal Component Analysis

Proposition 6 is an application of Theorem 5:

Proposition 6. *Assume* $\mathrm{P} = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ *a (strictly) positive definite matrix of rank* D*, with a degenerate Gaussian* $\mathrm{Q} = \mathcal{N}(\mathbf{0}, \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^{\mathrm{T}})$*, where* $\boldsymbol{A}$ *is a* $\mathrm{D} \times \mathrm{K}$ *matrix with orthonormal columns, and* $\boldsymbol{V}$ *is a* $\mathrm{K} \times \mathrm{K}$ *(strictly) positive definite diagonal matrix. Then,*

$$\mathrm{QKL}\left(\mathrm{Q} \| \mathrm{P}\right) = \mathrm{c} - \frac{1}{2}\sum_{k=1}^{\mathrm{K}} \log \boldsymbol{V}_{kk} + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{V}\right)$$

where c *is constant w.r.t.* $\boldsymbol{A}, \boldsymbol{V}$*. The optimal solution* $\boldsymbol{A}, \boldsymbol{V}$ *is to set columns of* $\boldsymbol{A}$ *to the top* K *eigenvectors of* $\boldsymbol{\Sigma}$ *and the diagonal of* $\boldsymbol{V}$ *to the corresponding eigenvalues.*[9]

Proposition 6 shows that the QKL-optimal way to approximate a full rank Gaussian with a degenerate one is to perform PCA on the covariance matrix. The result is intuitively satisfying as PCA preserves the directions of highest variance; S was thus indeed forced to align with the highest density regions under P as suggested in Section 5. See Figure 5 for a visualisation of this behaviour. Proposition 7 presents a variation of the result of Tipping & Bishop (1999), showing that Equation (8) can hold in practice.

Proposition 7. *Assume similar conditions as in Proposition 6, except* Q *will now be replaced with a series of distributions convolved with Gaussian noise:* $\mathrm{Q}^{(n)} = \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{(n)}\boldsymbol{V}^{(n)}(\boldsymbol{A}^{(n)})^{\mathrm{T}} + \tau^{(n)}\boldsymbol{I})$*. Given* $\tau^{(n)} \downarrow 0$ *as* $n \to 0$ *and the obvious constraints on* $\boldsymbol{A}^{(n)}, \boldsymbol{V}^{(n)}$*, Equation (8) holds in the sense of shrinking Euclidean/Frobenius norm between* $\{\boldsymbol{A}^{(n)}, \boldsymbol{V}^{(n)}\}$ *and the PCA solution.*

It is necessary to mention that both the QKL from Proposition 6 and any of the yet unconverged KL divergences in Proposition 7 have $\binom{\mathrm{D}}{\mathrm{K}}$ local optima for any combination of the eigenvectors which might lead to potentially problematic behaviour of gradient based optimisation.

6. Conclusion

The original intent behind dropout was to provide a simple yet effective regulariser for neural networks. The main value of the subsequent reinterpretation as a form of approximate Bayesian VI thus arguably lies in providing a principled theoretical framework which can explain the empirical behaviour, and guide extensions to the method. We have shown the current theory behind variational Bayesian dropout to have issues stemming from two main sources: 1) use of improper or pathological priors; 2) singular approximating distributions relative to the true posterior.

[9] We have assumed both Gaussians are zero mean to simplify the notation. Analogical results holds in the more general case.

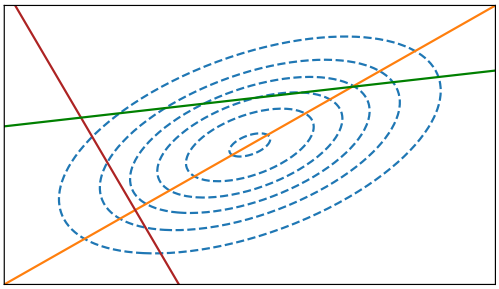

Figure 5. Visualisation of the relationship between QKL minimisation and PCA. The target in this example is the blue two dimensional Gaussian distribution. The approximating family is the set of all Gaussian distributions concentrated on a line, which would be problematic with conventional VI (c.f. Section 4). For all of the linear subspaces shown by the coloured lines the KL term on the right hand side of Equation (9) can be made zero by a suitable choice of the normal mean and variance. The remaining term $-\log \mathrm{Z}_{\mathrm{P}_S}$ therefore dictates the choice of subspace. The orange line is optimal aligning with the largest eigenvalue PCA solution.

The former issue pertains to the improper log uniform prior in variational Gaussian dropout. We proved its use leads to irremediably pathological behaviour of the true posterior, and consequently studied properties of the optimisation objective from a non-Bayesian perspective, arguing it is set up in such a way that cancels some of the pathologies and can thus still provide good empirical results, albeit not because of the Bayesian or the related MDL arguments.

The singular approximating distribution issue is relevant to both the Bernoulli and Gaussian dropout by making standard VI impossible due to an undefined objective. We have shown that the proposed remedies in (Gal & Ghahramani, 2016; Gal, 2016) can be made rigorous and are special cases of a broader class of limiting constructions leading to a unique objective which we termed quasi-KL divergence. We presented initial observations about QKL's properties, suggested an explanation for the empirical difficulty of obtaining hyperparameter estimates in dropout-based approximate inference, and motivated future exploration of QKL by showing it naturally yields PCA when approximating a full rank Gaussian with a degenerate one.

As use of improper priors and singular distributions is not isolated to the variational Bayesian dropout literature, we hope our work will contribute to avoiding similar pitfalls in future. Since it relaxes the standard KL assumptions, QKL will need further careful study in subsequent work. Nevertheless, based on our observations from Section 5 and the previously reported empirical results of variational Bayesian dropout, we believe QKL inspires a promising future research direction with potential to obtain a general framework for the design of computationally cheap optimisation-based approximate inference algorithms.

Acknowledgements

We would like to thank Matej Balog, Diederik P. Kingma, Dmitry Molchanov, Mark Rowland, Richard E. Turner, and the anonymous reviewers for helpful conversations and valuable comments. Jiri Hron holds a Nokia CASE Studentship. Alexander Matthews and Zoubin Ghahramani acknowledge the support of EPSRC Grant EP/N014162/1 and EPSRC Grant EP/N510129/1 (The Alan Turing Institute).

References

Gal, Y. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

Hron, J., Matthews, Alexander G. de G., and Ghahramani, Z. Variational Gaussian Dropout is not Bayesian. In *Second workshop on Bayesian Deep Learning (NIPS 2017)*. 2017.

Jang, E., Gu, S., and Poole, B. Categorical Reparameterization with Gumbel-Softmax. 2017. URL https://arxiv.org/abs/1611.01144.

Johnson, S. G. Faddeeva Package. http://ab-initio.mit.edu/wiki/index.php/Faddeeva_Package, 2012.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*, April 2014.

Kingma, D. P., Salimans, T., and Welling, M. Variational Dropout and the Local Reparameterization Trick. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2575–2583. Curran Associates, Inc., 2015.

Louizos, C., Ullrich, K., and Welling, M. Bayesian compression for deep learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3290–3300. Curran Associates, Inc., 2017.

Maddison, C. J., Mnih, A., and Teh, Y. W. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations (ICLR)*, 2017.

Matthews, Alexander G. de G. *Scalable Gaussian process inference using variational methods*. PhD thesis, University of Cambridge, 2016.

Matthews, Alexander G. de G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. Gaussian Process Behaviour in Wide Deep Neural Networks. *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1-nGgWC-.

Molchanov, D., Ashukha, A., and Vetrov, D. Variational Dropout Sparsifies Deep Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2498–2507. PMLR, 2017.

Neal, R. M. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

Neklyudov, K., Molchanov, D., Ashukha, A., and Vetrov, D. P. Structured Bayesian Pruning via Log-Normal Multiplicative Noise. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6778–6787. Curran Associates, Inc., 2017.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Tipping, M. E. and Bishop, C. M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Titsias, M. K. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Tucker, G., Mnih, A., Maddison, C. J., Lawson, J., and Sohl-Dickstein, J. REBAR: Low-variance, unbiased gradient estimates for discrete latent variable models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2624–2633. Curran Associates, Inc., 2017.

Turner, R. E., Berkes, P., and Sahani, M. Two problems with variational expectation maximisation for time-series models. *Inference and Estimation in Probabilistic Time-Series Models*, 2010.