NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

as a manuscript

Molchanov Dmitry

DOUBLY STOCHASTIC VARIATIONAL INFERENCE WITH SEMI-IMPLICIT AND IMPROPER DISTRIBUTIONS

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow — 2022

The PhD Dissertation was prepared at National Research University Higher School of Economics.

Academic Supervisor: Dmitry P. Vetrov, Candidate of Sciences, National Research University Higher School of Economics.

1 Introduction

1.1 Topic of the thesis

Doubly stochastic variational inference [1; 2] is one of the main tools in modern Bayesian deep learning. This work extends doubly stochastic variational inference to new classes of models. The first class are the models based on variational dropout [3]. This work refines the variational lower bound proposed in the original work, removes the imposed limitations on the parameter space and exposes a number of new counter-intuitive properties of the variational dropout model. Further, we extend doubly stochastic variational inference to a broad class of semi-implicit distributions [4]. The proposed doubly semi-implicit variational inference algorithm defines a proper variational lower bound that is suitable for semi-implicit posterior approximations and prior distributions. This work also investigates the properties and potential applications of the proposed procedure.

Actuality of the work. Machine learning allows to obtain high-quality solutions to ill-posed or loosely defined problems. The field of machine learning provides a set of tools that allow to automatically build complex algorithms based on data and domain knowledge. The versatility and effectiveness of machine learning models made them ubiquitous in a large variety of academic and industrial applications such as automated decision making, computer vision, analyzing, manipulating and generating natural data such as images or speech, and many others [5]. Before the discussion of the contributions we first need to set up the context of this work.

This work focuses on parametric machine learning models. Such models are typically defined by a parametric function or an algorithm that uses the inputs and the parameters of the model, and transforms them into the outputs of the model. For example, in a classification problem, the inputs might contain feature representations of objects, the parameters might contain the weights of a neural network, and the outputs might be the estimated probabilities for each class. To train such a model means to find the value of the parameters that achieve the best possible value of the chosen objective function given the training data. The objective function is typically constructed from a data term, which determines how well the data fits the particular set of parameters, and regularization terms that promote other desirable properties of the solution, such as sparsity or smoothness. Direct computation and optimization of these objective functions is often prohibitively computationally expensive, so different techniques such as Monte Carlo sampling and stochastic gradient descent

are used to estimate and optimize these objective functions approximately. In this work we study an important family of such objective functions, the variational evidence lower bounds [1; 6], and propose new ways to estimate them and their gradients, lifting some of the common restrictions on the variational lower bounds and the models they are applied to.

This work heavily relies on the Bayesian approach to machine learning [7]. In the Bayesian formalism the model is defined as a set of probability distributions over the involved variables, and the structure that ties them all together. These distributions usually include the likelihood, or the probabilistic distribution of the labels given the input and the model parameters. This typically corresponds to the data term of the objective function in conventional machine learning models. They also include the prior distribution of the parameters of the model, incorporating our domain knowledge, prior data or other biases that we might want to introduce. This typically corresponds to the regularization term of conventional models. Together the likelihood and the prior define the joint distribution over the labels and the parameters of the model given the input data. Instead of training, the Bayesian formalism dictates us to perform Bayesian inference, i.e. to obtain the posterior distribution of the model parameters given the training data. The posterior distribution incorporates all relevant information from the training data and represents our uncertainty about the model parameters. Then, in order to obtain the predictions for a new test object, one needs to average the predictions of the model over the posterior distribution. In theory, this approach has a number of benefits over conventional machine learning models. Conventional machine learning techniques typically result in a single model, usually corresponding to a maximum likelihood estimate or a maximum a posteriori estimate of a probabilistic machine learning model. The Bayesian approach, however, provides us with an infinitely large weighted ensemble of models, defined by the posterior distribution, and ensemble models are known for their improved robustness and better prediction performance. It also provides a way to incorporate domain knowledge or other biases in a principled way by defining a prior distribution over the model parameters. The posterior distribution can also serve as a compressed representation of the training data and can be refined when new data arrives without retraining the model from scratch and without suffering from catastrophic forgetting like conventional machine learning methods [8; 9]. This process is known as Bayesian incremental learning, which simply means using the obtained posterior distribution as a prior distribution while performing inference with a new set of data.

Unfortunately, exact Bayesian inference is only possible with a very limited set of models. For example, when the model is defined using a deep neural networks with a billion of parameters, performing Bayesian inference would involve the computation of an intractable billion-dimensional integral. Thus, modern Bayesian methods, especially applied to deep neural networks, rely on various approximate inference techniques. There are two main approaches to Bayesian deep learning. One approach uses modern gradient-based MCMC techniques such as the stochastic gradient Langevin dynamics (SGLD) [10] and its extensions [11; 12] to obtain samples from the posterior distribution, bypassing the need to construct the distribution itself. Another approach is based on stochastic variational inference [1; 13], where the posterior distribution is approximated by a simple parametric family of distributions. This approximation is carried out by recasting the inference problem as an optimization problem that has a similar complexity and structure as the training process of conventional models. Both approaches have their benefits and shortcomings. Modern gradient-based MCMC techniques suffer from highly correlated samples and, consequentially, low effective sample size. They are less suited for some applications, e.g. it is not clear how to reuse the posterior samples for Bayesian incremental learning. They also have a high memory footprint, essentially requiring to store numerous instances of trained models. However, the resulting sample-based approximation of the posterior distribution is typically more accurate than the parametric approximation obtained by variational inference techniques. On the other hand, variational inference techniques provide a compressed representation of the posterior distribution and allow to obtain infinitely many samples on demand without retraining the model. The constructed variational distribution can be reused as a prior distribution, allowing for approximate Bayesian incremental learning. However, the predictive performance heavily depends on the richness of the approximation family, while variational inference is only practical with simple approximations. For example, the fully-factorized Gaussian distribution remains one of the most popular approximation families. Both MCMC-based and VI-based approaches have the same structure as conventional training algorithms. They typically require using specific objective functions and injecting specific random noise either during the forward pass (in case of VI) or during the backward pass (in case of MCMC). Existing conventional models can often be easily modified to undergo Bayesian treatment, and the approximate inference techniques can benefit from the rich selection of tricks developed by the deep learning community to aid with training of deep neural networks.

Many existing deep learning techniques employ some kind of parameter, activation or gradient noise during training as a heuristic to reduce overfitting. For example, dropout [14] and its variants introduce Bernoulli or Gaussian multiplicative noise on the parameter or activation level. Batch normalization [15] implicitly introduces noise in the activations by adding a dependency on a random selection of objects in the minibatch. Given the form of the noise, it is possible to reverse engineer these techniques and show that they actually perform variational inference with a specific kind of noise in a model with a specific prior distribution. This recasting as Bayesian inference then provides us with a number of consequences. First of all, instead of obtaining a single model we now obtain an approximate posterior distribution, so we can perform posterior averaging during test-time evaluation. This provides a powerful insight: any stochasticity used during training can be averaged out during testing, typically resulting in better robustness and predictive performance. Secondly, some of the hyperparameters, e.g. the dropout rate, now become variational parameters, meaning that we can and should optimize the objective w.r.t. them [3]. Because we can now employ gradient optimization to tune these parameters, we are not limited by the complexity of cross-validation and can, for example, find a separate optimal dropout rate for each layer or even for each single weight. This way the Bayesian treatment can give us a powerful mechanism of automatic hyperparameter tuning. Finally, since we now understand the true nature of the process, we can make changes to it, choosing a different prior distribution, a different posterior approximation or tweaking the approximate inference technique.

Among other things, doubly stochastic variational inference relies on the calculation of the Kullback-Leibler divergence between the approximate posterior and the prior distribution. It makes the available selection of prior and approximate posterior distributions limited. Variational dropout, one of the most wide-acclaimed examples of Bayesian neural networks, uses the improper log-uniform prior. It is the only prior distribution that has the properties, desired by the authors, and the KL divergence between their approximate posterior (a fully factorized Gaussian) and this prior is intractable. Therefore the authors use a polynomial approximation that is only accurate at a limited range of variational parameters, heavily limiting the already crude posterior approximation. In this work we lift these limitations by proposing a different approximation that is accurate on the full range of variational parameters. After these limitations have been lifted, we have conducted a broad study of the variational dropout model at the full range of its variational parameters, and have discovered that

variational dropout sparsifies deep neural networks, allowing for high levels of model compression. We have also discovered that by limiting the variational approximation in a different way we can get rid of a class of local optima and obtain variance networks, a model with zero-mean variance-only latent feature representations that can provide diverse samples from the approximate posterior, allowing for effective posterior averaging and resulting in a highly robust ensemble.

Doubly stochastic variational inference is typically limited to explicit distributions, or distributions with a closed-form expression for probabilistic density. If the approximate posterior and the prior are explicit, and the approximate posterior is reparameterizable, it is possible to estimate the value and the gradients of the KL divergence. However, the family of explicit reparameterizable distributions is still fairly limited. It this work, we extend doubly stochastic variational inference to work with semi-implicit approximate posteriors and priors. Semi-implicit distributions [4] are a broad family of reparameterizable distributions that generally do not have closed-form expressions for density. Semi-implicit distributions are defined as infinite mixtures of explicit distributions with an arbitrary mixture distribution and can approximate any implicit distribution to a given precision. They can rely on universal approximators such as deep neural networks and in theory can approximate any given distribution. We present doubly semi-implicit variational inference, a new family of variational lower bounds for models with semi-implicit approximate posteriors and priors. The proposed bounds are asymptotically exact and can be used to estimate the variational lower bound up to a given precision. Among other advanced methods for variational inference, DSIVI has a number of advantages. Unlike unbiased implicit variational inference (UIVI [16]) and operator variational inference (OPVI [17]), it supports both semi-implicit approximate posteriors and semi-implicit priors. Unlike density ratio estimation techniques (DRE, e.g. adversarial variational Bayes [18]) and kernel implicit variational inference (KIVI [19]), DSIVI optimizes a proper variational lower bound, whereas DRE techniques and KIVI optimize a biased surrogate with no reliable way to estimate the bias. DSIVI has less restrictions on the mixing distribution, which has to be explicit in UIVI and hierarchical variational inference (HVI [20; 21]). Since the approximate posterior and the prior distributions in DSIVI lie in the same general family, the resulting semi-implicit approximate posterior can be reused as a prior distribution in Bayesian incremental learning. Finally, even when the KL divergence can be estimated directly, DSIVI bounds can be useful to reduce the required complexity. For example, it is known that

the aggregated posterior distribution is the optimal prior for the variational autoencoder [1]. However, the complexity of obtaining a single estimate of the KL divergence between the approximate posterior and the prior is $O(N)$, N being the size of the training set, which is prohibitive for stochastic gradient descent. By using DSIVI bounds we can reduce this complexity to $O(K)$, where K is the number of samples used by DSIVI, allowing to get a trade-off between the computational complexity and the tightness of the obtained bound. This improves upon VampPrior [22] by having the same computational complexity, less optimizable parameters, less hyperparameters, and better resulting quality.

The goal of this work is the expansion of the toolset, available for Bayesian deep learning practitioners. The proposed extensions are expected to enable new applications of deep Bayesian models as well as improve upon existing models.

1.2 Key results and conclusions

The novelty of this work can be summarized in the following contributions:

1. A way to estimate the variational dropout objective with no restrictions on variational parameters, leading to the discovery of two new modes of operation of the variational dropout model.
2. A way to train and assess models that can robustly encode features as zero-mean distributions, leading to improved sample diversity and model robustness.
3. A new algorithm of variational inference that is applicable to semi-implicit posterior approximations and prior distributions, allows for implicit mixing distributions and optimizes a proper evidence lower bound.

Theoretical and practical significance. The obtained results have allowed to discover new properties of the variational dropout model, resulting in a practical way to compress and accelerate deep neural networks. This was the first successful application of Bayesian methods to the compression of modern deep neural networks and has spawned a line of works on Bayesian compression for deep learning models. This work also expands the variational inference toolset, lifting some of the common restrictions on the models and the choice of the posterior approximation. We also provide a principled way to reduce the complexity of variational inference with the aggregated posterior

prior distribution in variational autoencoders, improving upon the previously used VampPrior technique.

Methodology and research methods. This work uses the toolset of deep learning and Bayesian deep learning. The numerical experiments and visualizations have been performed using Python frameworks Numpy, Theano, Lasagne, PyTorch, Pandas and MatPlotLib.

Reliability of the declared results is supported by a clear presentation of the used algorithms, experiment setups, proofs of theorems. The source code used to perform the experiments have been made publicly accessible.

Main provisions for the defense:

1. The way to estimate the variational dropout objective at the full range of variational parameters.
2. The variance network model with zero-mean variance-only embeddings and a way to train the variational dropout model to converge to a variance network.
3. The doubly semi-implicit variational inference algorithm that extends doubly stochastic variational inference to semi-implicit posterior approximations and prior distributions.

Personal contribution into the main provisions for the defense. All stated theoretical results are obtained by the author. The author has formulated and proved all included theorems. The code for the experiments and visualization, the technical setup and the text of the papers are results of the collaboration between all coauthors of papers.

Publications and probation of the work

First-tier publications.

1. *Dmitry Molchanov, Arsenii Ashukha, Dmitry Vetrov* Variational Dropout Sparsifies Deep Neural Networks // In Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70:2498-2507, 2017. CORE rank A* conference.
2. *Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, Dmitry Vetrov* Doubly Semi-Implicit Variational Inference // In Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 89:2593-2602, 2019. CORE Rank A conference.

Other publications.

1. *Kirill Neklyudov, Dmitry Molchanov, Arsenii Ashukha, Dmitry Vetrov* Variance Networks: When Expectation Does Not Meet Your Expectations // International Conference on Learning Representations (ICLR), 2019. Indexed by SCOPUS.
2. *Iuliia Molchanova, Dmitry Molchanov, Novi Quadrianto, Dmitry Vetrov* Structured Semi-Implicit Variational Inference // 2nd Symposium on Advances in Approximate Bayesian Inference, 2019. Best industrial paper run-up award.

Reports at conferences and seminars.

1. Research Seminar on Bayesian Methods in Machine Learning, Moscow, 02 September 2016. Topic: "Variational dropout for deep neural networks and linear models".
2. Russian Supercomputing Days, Moscow, 26 September 2016. "Variational dropout for deep neural networks and linear models".
3. The 34th International Conference on Machine Learning, Sydney, Australia, 09 August 2017. "Variational dropout sparsifies deep neural networks".
4. Research Seminar on Bayesian Methods in Machine Learning, Moscow, 11 May 2018. Topic: "Variance networks".
5. Research Seminar on Bayesian Methods in Machine Learning, Moscow, 14 September 2018. Topic: "Variational inference with implicit distributions".
6. The 22nd International Conference on Artificial Intelligence and Statistics, Okinawa, Japan, 16 April 2019. Topic: "Doubly Semi-Implicit Variational Inference"
7. International conference on Learning Representations, New Orleans, USA, 09 May 2019. Topic: "Variance Networks: When Expectation Does Not Meet Your Expectations".
8. 2nd Symposium on Advances in Approximate Bayesian Inference, Vancouver, Canada, 08 December 2019. Topic: "Structured Semi-Implicit Variational Inference".

Volume and structure of the work. The thesis contains an introduction, contents of publications and a conclusion. The full volume of the thesis is 78 pages.

2 Content of the work

2.1 Variational Dropout Sparisfies Deep Neural Networks

The first chapter describes the variational dropout model and a new way to estimate the variational dropout objective. Variational dropout [3] is a Bayesian generalization of Gaussian dropout that allows to automatically tune dropout rates during training with no need for cross-validation. It also results in a Gaussian approximation to the posterior distribution of a certain probabilistic model, allowing to perform posterior averaging. The originally proposed variational dropout objective was only computed for a limited range of Gaussian dropout rates. In this work, we proposed a way to compute the variational dropout objective in the full range of Gaussian dropout rates, which have allowed to discover two new modes of operation of the variational dropout model, namely the sparsity inducing mode and the zero-mean variance-only mode.

First we will introduce the necessary notation. The probabilistic models used in this chapter are defined with the likelihood function $p(y \mid x, w)$ and the prior distribution $p(w)$, where x denotes the features of an object, y denotes the target vector and w denotes the parameters of the model, e.g. the weight matrices. Then, the posterior distribution is defined using the Bayes theorem:

$$p(w \mid X_{train}, Y_{train}) = \frac{p(Y_{train} \mid X_{train}, w)p(w)}{\int p(Y_{train} \mid X_{train}, w)p(w)dw}, \tag{1}$$

where (X_{train}, Y_{train}) is the training dataset. For deep neural networks this distribution cannot be computed directly, so a parametric approximation $q_\phi(w)$ is used instead. The parameters ϕ of this approximation, called variational parameters, can be found by minimizing the KL divergence between the approximation and the true posterior distribution:

$$\mathrm{KL}(q_\phi(w) \,\|\, p(w \mid X_{train}, Y_{train})) \to \min_\phi \tag{2}$$

This optimization problem uses the intractable posterior distribution in its definition. However, it is equivalent to the following optimization problem, suitable for standard stochastic gradient-based optimization techniques.

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(w)} \log p(Y_{train} \mid X_{train}, w) - \mathrm{KL}(q_\phi(w) \,\|\, p(w)) \to \max_\phi \tag{3}$$

This objective function is called the evidence lower bound (ELBO) since it is a lower bound on the evidence $\log p(Y_{train} \mid X_{train})$ of the model.

In this work we consider reparameterizable [1] distributions $q_\phi(w)$. It means that there exists a non-parametric distribution $p(\epsilon)$ and a so-called reparameterizing function $g(\phi, \epsilon)$ such that the random variable $\tilde{w} = g(\phi, \epsilon)$ follows the distribution $q_\phi(w)$. Such distributions allow us to estimate the ELBO objective and its gradients:

$$L(\phi) = \frac{N}{M} \sum_{i=1}^{M} \log p(y_i \,|\, x_i, g(\epsilon, \phi)) - \mathrm{KL}(q_\phi(w) \,\|\, p(w)) \simeq \mathcal{L}(\phi), \tag{4}$$

$$\nabla L(\phi) \simeq \nabla \mathcal{L}(\phi), \tag{5}$$

where N and M are the training set and mini-batch size, (x_i, y_i) is a random training object, $L(\phi)$ and $\nabla L(\phi)$ are unbiased mini-batch estimates of the ELBO objective and its gradients respectively, suitable for modern stochastic gradient optimization techniques.

The process of optimizing the objective (3) is called doubly stochastic variational inference (DSVI). It results in a parametric approximation $q_{\phi^*}(w)$ of the posterior distribution. This distribution can then be used to perform approximate posterior averaging for new objects (x_{new}, y_{new}):

$$\begin{gathered} p(y_{new} \,|\, x_{new}, X_{train}, Y_{train}) = \\ = \int p(y_{new} \,|\, x_{new}, w) p(w \,|\, X_{train}, Y_{train}) dw \approx \\ \approx \int p(y_{new} \,|\, x_{new}, w) q_{\phi^*}(w) dw \simeq \frac{1}{K} \sum_{k=1}^{K} p(y_{new} \,|\, x_{new}, g(\epsilon_k, \phi^*)). \end{gathered} \tag{6}$$

It can also be reused as a new prior distribution for further steps of Bayesian incremental learning [9].

Dropout [14; 28] is a regularization technique often used for learning deep neural networks. It consists of multiplying the inputs or outputs of each layer by multiplicative noise. Different variants of dropout introduce this noise in different parts of the model (e.g. layer inputs, outputs or layer weights [29]), and use different noise distributions. One popular choice is Bernoulli distribution with probability $1 - p$, where p is called the dropout rate. Gaussian distribution with mean 1 and corresponding variance $\alpha = p/(1-p)$ is another popular choice, known as Gaussian dropout.

The variational dropout model [3] is defined by a fully factorized Gaussian posterior approximation $q_\phi(w) = \prod_i \mathcal{N}(w_i \,|\, \mu_i, \alpha\mu_i^2)$ and a log-uniform prior distribution $p(w) \propto \prod_i \frac{1}{|w_i|}$. It turns out that the ELBO objective (3)

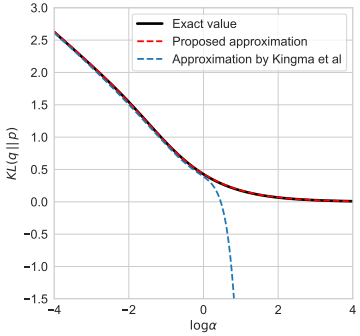


Figure 1: Different approximations of the KL divergence: unlike the proposed approximation, the original one [3] is tight only for $\alpha \leq 1$.

Network	Method	Error %	Sparsity per Layer %	$\frac{\lvert\mathbf{W}\rvert}{\lvert\mathbf{W}_{\neq 0}\rvert}$
LeNet-300-100	Original	1.64		1
	Pruning	1.59	92.0 – 91.0 – 74.0	12
	DNS	1.99	98.2 – 98.2 – 94.5	56
	SWS	1.94		23
	(ours) Sparse VD	1.92	98.9 – 97.2 – 62.0	**68**
LeNet-5-Caffe	Original	0.80		1
	Pruning	0.77	34 – 88 – 92.0 – 81	12
	DNS	0.91	86 – 97 – 99.3 – 96	111
	SWS	0.97		200
	(ours) Sparse VD	0.75	67 – 98 – 99.8 – 95	**280**

Table 1: Comparison of different sparsity-inducing techniques (Pruning [23; 24], DNS [25], SWS [26]) on LeNet [27] architectures. Our method provides the highest level of sparsity with a similar accuracy.

exactly corresponds to the Gaussian dropout model and objective. The KL-divergence term is a function of the dropout rate α and is constant in the neural networks weight μ, and can be omitted if the dropout rates are kept constant, recovering the Gaussian dropout procedure. However, in principle, the variational dropout objective allows to perform stochastic gradient optimization of the dropout rates as well, and even allows to define a separate dropout rate for each weight of a neural network.

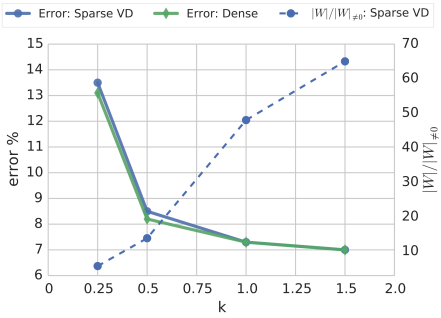
The computation of the variational dropout objective involves the computation of the intractable KL term:

$$\mathrm{KL}(\mathcal{N}(w_i \,|\, \mu_i, \alpha\mu_i^2) \,\|\, \mathrm{LogU}(w_i)) \propto -\frac{1}{2}\log\alpha + \mathbb{E}_{\mathcal{N}(\epsilon_i \,|\, 1,\alpha)} \log|\epsilon_i| \qquad (7)$$

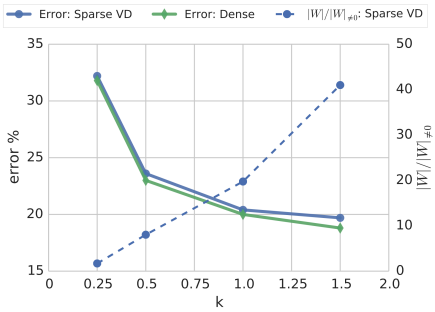
This expression has been originally approximated by a 3rd degree polynomial in α [3]. However, this approximation is only accurate for $0 < \alpha \leq 1$ and the whole training procedure does not work for the noise variances larger than 1. In this work we propose a different approximation that captures the asymptotic behavior of the KL divergence and is accurate on the full range of α:

$$\begin{gathered}\mathrm{KL}(\mathcal{N}(w_i \,|\, \mu_i, \alpha\mu_i^2) \,\|\, \mathrm{LogU}(w_i)) \approx \\ \approx k_1\sigma(k_2 + k_3\log\alpha_i)) - 0.5\log(1+\alpha_i^{-1}), \text{ where} \\ k_1 = 0.63576 \quad k_2 = 1.87320 \quad k_3 = 1.48695\end{gathered} \qquad (8)$$

The comparison between the true values of the KL divergence, the original approximation and the proposed approximation can be seen in Figure 1.



(a) Results on the CIFAR-10 dataset



(b) Results on the CIFAR-100 dataset

Figure 2: Accuracy and sparsity level for VGG-like architectures of different sizes. The number of neurons and filters scales as k. Dense networks were trained with Binary Dropout, and Sparse VD networks were trained with Sparse Variational Dropout on all layers. The overall sparsity level, achieved by our method, is reported as a dashed line. The accuracy drop is negligible in most cases, and the sparsity level is high, especially in larger networks.

We have observed that in many cases the optimal value of dropout rates α approaches infinity. We have found that such extreme values of dropout rates do not destabilize training and enable two new properties, or modes of operation, of variational dropout model. One of such modes are sparse neural networks. We discuss the other mode, variance networks, in the next chapter.

When we tune a separate value of dropout rate α for each weight of a neural network, we observe that most dropout rates α_i go to infinity, while the corresponding means μ_i and variances $\alpha_i \mu_i^2$ go to zero. This means that the corresponding marginal distribution of the approximate posterior has collapsed to a delta-function centered at zero, and the corresponding weight can be removed from the model, resulting in extremely sparse weight matrices. Using this technique we have been able to compress convolutional deep neural networks up to 68 times on VGG-like CIFAR networks [30] and up to 280 times on LeNet architectures [27] with a negligible accuracy degradation. The results are presented in Table 1 and Figure 2.

2.2 Variance Networks: When Expectation Does Not Meet Your Expectation

In the previous chapter, we have introduced the variational dropout model, proposed a way to compute its objective on the full range of the Gaussian dropout rates, and have shown the sparsity-inducing properties of variational dropout. In this chapter, we continue to investigate the variational

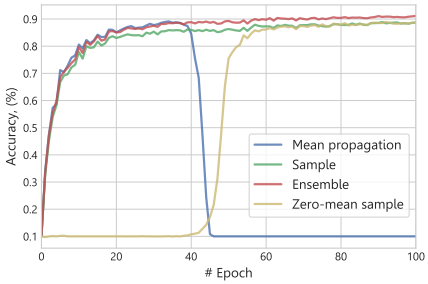


Figure 3: CIFAR-10 test set accuracy for a VGG-like neural network [30] with a layer-wise parameterization with weights replaced by their expected values (deterministic), sampled from the variational distribution (sample), the test-time averaging (ensemble) and zero-mean approximation accuracies.

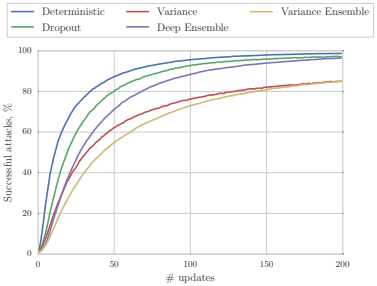


Figure 4: Results on iterative fast sign adversarial attacks [31] for a VGG-like architecture on CIFAR-10 dataset. For each iteration we report the successful attack rate. Plain variance networks outperform deep ensembles [32], and deep ensembles of variance networks have the lowest rate of successful attacks.

dropout model and show how the change of parameterization can remove a class of local optima, revealing variance networks. Variance networks are a new mode of operation of variational dropout and related models. It results in a zero-mean posterior approximation that stores the latent feature representations in zero-mean variance-only embeddings, resulting in a high sample diversity, and resulting in a highly robust ensemble after posterior averaging.

The variational dropout objective promotes large values of dropout rates, pushing many of them to infinity in practice [33]. This introduces large amounts of noise that degrades the predictive performance, so the neural network chooses to remove the noisy weights by pushing their means to zero in such a way that their variances $\alpha_i \mu_i^2$ go to zero as well. This allows the model to satisfy the itch for large values of dropout rates without compromising the predictive performance. However, we can prevent this behavior by introducing parameter sharing and tying some of the dropout rates together. For example, if all the dropout rates for a particular layer are the same, and pushed to infinity, the model cannot collapse the posterior to a delta function without severe performance degradation. It turns out that this scenario still leads to infinite values of alpha, however the neural network experiences a phase transition, as illustrated in Figure 3. Before the phase transition the approximate posterior uses the mean to describe the values of the neural networks weights, as indicated by the classification accuracy, achieved by the mean of the approximate posterior. After the phase transition, the mean of the posterior no longer represents the weights and can be zeroed out, and the variances describe the weights in-

stead. This results in a zero-mean posterior approximation, and, consequently, zero-mean feature representations. Notice that the sample quality does not degrade, and the overall training process is stable. We call this mode of operation variance networks.

To show how the large values of alpha lead to a zero-mean posterior approximation, we have derived an upper bound on the maximum mean discrepancy [34] between a non-zero-mean posterior and a zero-mean posterior. This provides an upper bound on the change in the predictions of a neural network after zeroing out the mean of the approximate posterior.

Theorem 1. *Assume that* $\alpha_t \longrightarrow +\infty$ *as* $t \longrightarrow +\infty$. *Then the Gaussian dropout posterior* $q_t(w) = \prod_{i=1}^{D} \mathcal{N}(w_i \,|\, \mu_{t,i}, \alpha_t \mu_{t,i}^2)$ *becomes indistinguishable from its zero-centered approximation* $q_t^0(w) = \prod_{i=1}^{D} \mathcal{N}(w_i \,|\, 0, \alpha_t \mu_{t,i}^2)$ *in terms of Maximum Mean Discrepancy:*

$$\mathrm{MMD}(q_t^0(w) \,\|\, q_t(w)) \leq \sqrt{\frac{2D}{\pi}} \cdot \frac{1}{\sqrt{\alpha_t}} \tag{9}$$

$$\lim_{t \to +\infty} \mathrm{MMD}(q_t^0(w) \,\|\, q_t(w)) = 0 \tag{10}$$

We have tested the variational dropout model with different parameterizations, representing different levels of dropout rate sharing:

	zero-mean	layer-wise	neuron-wise	weight-wise	additive
$q(w_{ij})$	$\mathcal{N}(0, \sigma_{ij}^2)$	$\mathcal{N}(\mu_{ij}, \alpha \mu_{ij}^2)$	$\mathcal{N}(\mu_{ij}, \alpha_j \mu_{ij}^2)$	$\mathcal{N}(\mu_{ij}, \alpha_{ij} \mu_{ij}^2)$	$\mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$

(11)

Here the additive and weight-wise parameterizations are equivalent, and other parameterizations are their less flexible special cases. The values of the evidence lower bound and its parts, as well as the resulting accuracy of the models, can be found in Table 2. Note that by removing local optima, less flexible parameterizations converge to variance networks, resulting in a better value of the evidence lower bound.

We have also investigated different prior distributions. We have found that the Student's t-distribution prior with a small degree of freedom and the automatic relevance determination prior result in the same model behavior, making them suitable both for model sparsification and for training variance networks.

The variance-only embeddings promote the diversity of the samples, as indicated by a relatively low single-sample accuracy and a high multi-sample accuracy in Figure 3. We show that this results in a highly robust ensemble that

Table 2: Variational lower bound (ELBO), its decomposition into the data term and the KL term, and test set accuracy for different parameterizations. The test-time averaging accuracy is roughly the same for all procedures, but a clear phase transition is only achieved in layer-wise and neuron-wise parameterizations.

Metric		Parameterization				
		zero-mean	layer	neuron	weight	additive
Evidence Lower Bound	$\mathcal{L}(\phi)$	$\mathbf{-4.0}$	$\mathbf{-17.4}$	$\mathbf{-31.4}$	-602.6	-227.9
Data term	$\mathbb{E}_q \log p(T \mid X, W)$	-4.0	-15.8	-17.0	-33.8	-31.2
Regularizer term	$\mathrm{KL}(q \,\|\, p)$	0.0	1.7	14.4	568.8	196.7
Mean propagation acc. (%)	$\hat{y} = \arg\max_t p(t \mid x, \mathbb{E}_q W)$	11.3	11.3	11.3	96.6	99.2
Test-time averaging acc. (%)	$\hat{y} = \arg\max_t \mathbb{E}_q p(t \mid x, W)$	99.4	99.2	99.2	99.4	99.2

improves upon the deep ensembles in the defense against adversarial attacks, as shown in Figure 4.

2.3 Doubly Semi-Implicit Variational Inference

In the previous chapters we have considered doubly stochastic variational inference with fully factorized Gaussian posterior approximations. Now, we will present doubly semi-implicit variational inference, DSIVI, extending doubly stochastic variational inference to work with semi-implicit posterior approximations and prior distributions. This allows to use universal approximators like deep neural networks to construct the posterior approximation, capturing the complex multimodal nature of the posterior distribution, and allows to refine them using the Bayesian incremental learning framework.

Semi-implicit distributions [4] are defined as mixtures of explicit conditional distributions:

$$q_\phi(z) = \int q_\phi(z \mid \psi) q_\phi(\psi) d\psi. \tag{12}$$

The conditional distribution $q_\phi(z \mid \psi)$ is explicit and reparameterizable, and the mixing variables ψ follow an implicit reparameterizable distribution $q_\phi(\psi)$, resulting in an implicit marginal distribution $q_\phi(z)$. In order to sample from such a distribution, one may sample a mixing variable $\hat{\psi} \sim q_\phi(\psi)$, and then sample the main variable $\hat{z} \sim q_\phi(z \mid \hat{\psi})$. Any implicit distribution can be represented in a similar form, $q_\phi(z) = \int \delta(z - z') q_\phi(z')\, dz'$, and then closely approximated by a semi-implicit distribution using, for example, a Gaussian conditional distribution $q_\phi(z) \approx \int \mathcal{N}(z \mid z', \sigma^2) q_\phi(z')\, dz'$ with a sufficiently small variance σ^2.

For the first time, we present doubly semi-implicit variational inference (DSIVI) bound, a proper variational lower bound that can be estimated for semi-implicit posterior approximations and semi-implicit prior distributions.

It defines a lower bound on the following ELBO objective:

$$\begin{gathered}\mathcal{L} = \mathbb{E}_{q_\phi(z)} \log p(x \mid z) - \mathrm{KL}(q_\phi(z) \,\|\, p_\theta(z)), \text{ where} \\ q_\phi(z) = \int q_\phi(z \mid \psi) q_\phi(\psi) d\psi, \\ p_\theta(z) = \int p_\theta(z \mid \zeta) p_\theta(\zeta) d\zeta.\end{gathered} \tag{13}$$

The following expression defines the DSIVI lower bound for the DSVI bound (13):

$$\begin{gathered}\underline{\underline{\mathcal{L}}}^{q,p}_{K_1,K_2} = \mathbb{E}_{q_\phi(z)} \log p(x \mid z) - \\ - \mathbb{E}_{\psi^{0..K_1} \sim q_\phi(\psi)} \mathbb{E}_{q_\phi(z \mid \psi^0)} \log \frac{1}{K_1 + 1} \sum_{k=0}^{K_1} q_\phi(z \mid \psi^k) \\ + \mathbb{E}_{\zeta^{1..K_2} \sim p_\theta(\zeta)} \mathbb{E}_{q_\phi(z)} \log \frac{1}{K_2} \sum_{k=1}^{K_2} p_\theta(z \mid \zeta^k),\end{gathered} \tag{14}$$

Theorem 2. *The DSIVI bound* (14) *monotonically increases in* K_1 *and* K_2*, and converges to the evidence lower bound* $\mathcal{L}$ (13) *from below as* K_1 *and* K_2 *approach infinity:*

$$\underline{\underline{\mathcal{L}}}^{q,p}_{K_1,K_2} \leq \underline{\underline{\mathcal{L}}}^{q,p}_{K_1+1,K_2}, \qquad \underline{\underline{\mathcal{L}}}^{q,p}_{K_1,K_2} \leq \underline{\underline{\mathcal{L}}}^{q,p}_{K_1,K_2+1}, \qquad \underline{\underline{\mathcal{L}}}^{q,p}_{K_1,K_2} \leq \mathcal{L}, \tag{15}$$

$$\lim_{K_1,K_2 \to +\infty} \underline{\underline{\mathcal{L}}}^{q,p}_{K_1,K_2} = \mathcal{L}. \tag{16}$$

Notice that the complexity of estimation of the data term is the same as in conventional DSVI. The computation of the KL divergence term requires $K_1 + 1$ samples from the mixing distribution $q_\phi(\psi)$ of the semi-implicit posterior approximation and K_2 samples from the mixing distribution $p_\theta(\zeta)$ of the prior distribution. In most cases this sampling can utilize batching and can be efficiently performed using modern deep learning hardware, without a significant overhead to the training of corresponding non-semi-implicit models. The number of samples K_1 and K_2 can be chosen depending on the complexity of the involved semi-implicit distributions and on the available computational budget. More complex mixing distributions typically require higher amounts of samples. We compare DSIVI to other methods for implicit variational inference [18; 19] on a controlled synthetic task. The results are presented in Figure 5.

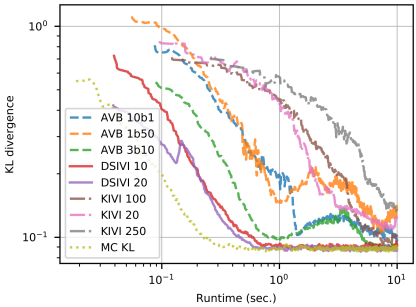


Figure 5: Comparison of different techniques for implicit VI. "KIVI K" corresponds to KIVI with K MC samples; "DSIVI K" corresponds to DSIVI with $K_1 = K_2 = K$; "AVB MbK" corresponds to AVB with M updates of discriminator per one update of ϕ and K MC samples to estimate the discriminator's gradients. "MC KL" corresponds to direct stochastic minimization of the KL divergence.

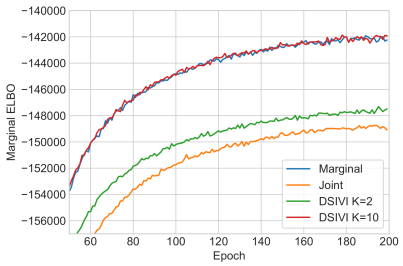


Figure 6: Variational inference for a Bayesian neural networks with a hierarchical prior (Gaussian prior with a Gamma hyperprior on the prior precision). Models are trained using different variational objectives. The estimates of the marginal evidence lower bound are presented in this plot.

One kind of semi-implicit distributions is particularly common. Bayesian models often involve hierarchical prior distributions and posterior approximations [2; 19; 35–38]. For example, a discriminative Bayesian model with a hierarchical prior can be defined as follows:

$$p(t,w,\alpha \mid x) = p(t \mid x, w)p(w \mid \alpha)p(\alpha). \tag{17}$$

One common way to perform inference in such a model is to perform joint inference over the main variables and the mixing variables, recovering a joint posterior approximation $q_\phi(w,\alpha)$ [19; 37; 38]. However since the likelihood function only depends on the parameters w, we are not necessarily interested in the mixing variable α, and would like to marginalize it out, resulting in a model with a semi-implicit prior and a semi-implicit posterior approximation. This gives us two alternative objectives to optimize, the joint bound and the marginal bound:

$$\mathcal{L}^{joint}(\phi) = \mathbb{E}_{q_\phi(w,\alpha)} \log \frac{p(t \mid x, w)p(w \mid \alpha)p(\alpha)}{q_\phi(w,\alpha)}, \tag{18}$$

$$\mathcal{L}^{marginal}(\phi) = \mathbb{E}_{q_\phi(w)} \log \frac{p(t \mid x, w)p(w)}{q_\phi(w)}. \tag{19}$$

Next, we show that the joint objective introduces an additional bias that is difficult to control, and the point of optimum of the marginal objectives produces a better fit to the original posterior of the model.

Theorem 3. *Let ϕ_j and ϕ_m maximize $\mathcal{L}^{joint}$ and $\mathcal{L}^{marginal}$ correspondingly. Then*

$$\mathrm{KL}(q_{\phi_m}(w) \,\|\, p(w \,|\, X_{tr}, T_{tr})) \leq \mathrm{KL}(q_{\phi_j}(w) \,\|\, p(w \,|\, X_{tr}, T_{tr})). \tag{20}$$

We stress that when it is feasible to marginalize, it is beneficial to marginalize, while DSIVI provides the necessary tools to perform such marginalization and giving the control over the introduced additional inference gap. This finding is also supported by our experiments (see Figure 6).

DSIVI bounds can also be useful in explicit models defined as large mixtures. In this case, DSIVI bounds can be used to reduce the computational complexity of estimating the densities of such mixtures. One example is the training of variational autoencoders [1] with the aggregated posterior priors. The aggregated posterior prior of a variational autoencoder is the optimal prior distribution in terms of the evidence lower bound, and is defined as the mixture of approximate posteriors, conditioned on all the training points:

$$p^*(z) = \frac{1}{N} \sum_{n=1}^{N} q_\phi(z \,|\, x_n). \tag{21}$$

One popular approximation to this prior is a variational approximation called VampPrior [22] that involves training of a number of inducing points u_k:

$$p^{Vamp}(z) = \frac{1}{K} \sum_{k=1}^{K} q_\phi(z \,|\, u_k). \tag{22}$$

However, we can now represent the aggregated posterior prior as a semi-implicit distribution with a discrete uniform mixing distribution over the inputs, and apply the DSIVI bound for such prior. This simple modification allows to outperform the VampPrior models, resulting in higher values of test log-likelihood on different VAE architectures [1; 20] (see Table 3 for details).

As with most approximate inference techniques, the performance of DSIVI may quickly degrade as the number of dimensions increases. The nature of DSIVI bound is to approximate the infinite mixture with a finite mixture and derive the bound from there. However, complex multidimensional distributions may require an exponentially large amount of samples to obtain a good

Table 3: We compare VampPrior with its semi-implicit modifications, DSIVI. We report the the IWAE objective for VampPrior-data, and the corresponding lower bound for DSIVI-based methods). Only the prior distribution is semi-implicit.

Method	LL
VAE+VampPrior-data	-85.05
VAE+VampPrior	-82.38
VAE+DSIVI (K=500)	≥ -83.02
VAE+DSIVI (K=5000)	$\geq \mathbf{-82.16}$
HVAE+VampPrior-data	-81.71
HVAE+VampPrior	-81.24
HVAE+DSIVI (K=5000)	$\geq \mathbf{-81.09}$

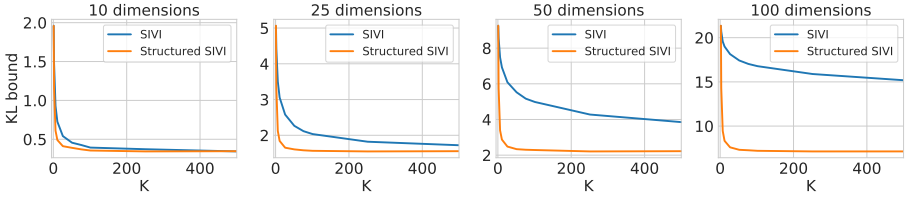


Figure 7: SIVI and SSIVI KL bounds for an autoregressive semi-implicit model and a synthetic multi-dimensional distribution. As expected, SSIVI always outperforms SIVI, and the gap increases with the number of dimensions.

approximation. To mitigate this problem, we show that by using the inherent structure in the model, it is possible to significantly improve upon the original DSIVI bound.

Consider a structured posterior approximation:

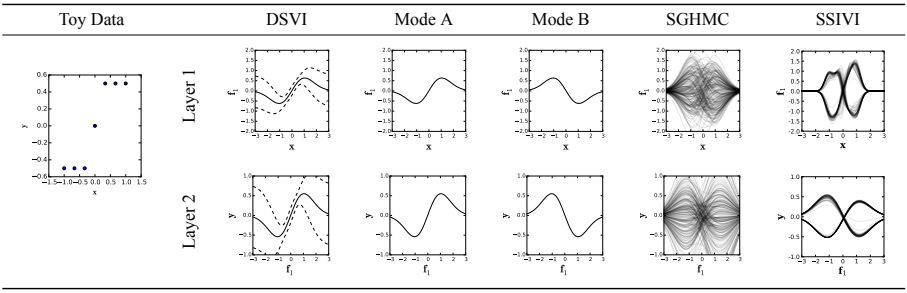
$$q_\phi(z) = q_\phi(z_1) \prod_{i=2}^{d} q_\phi(z_i \mid z_{1..i-1}), \tag{23}$$

$$q_\phi(z_i \mid z_{1..i-1}) = \int q_\phi(z_i \mid z_{1..i-1}, \epsilon_i) q(\epsilon_i) d\epsilon_i. \tag{24}$$

For such a distribution the DSIVI bound can be applied in two different ways. One way would be to bound the entropy of the distribution as a whole, resulting in $\underline{\mathcal{H}}_K^{\text{SIVI}}[q_\phi(z)]$. Another way would be to first decompose the entropy into a sum of entropies of lower-dimensional distributions, and then obtain a separate bound for each entropy, resulting in $\underline{\mathcal{H}}_K^{\text{SSIVI}}[q_\phi(z)]$.

$$\mathcal{H}[q_\phi(z)] \geq \underline{\mathcal{H}}_K^{\text{SIVI}}[q_\phi(z)] = -\mathbb{E}_{\epsilon^{0..K}} \mathbb{E}_{z \mid \epsilon^0} \log \frac{1}{K+1} \sum_{k=0}^{K} q_\phi(z \mid \epsilon^k). \tag{25}$$

Figure 8: Visualisation of layers in a two-layer DGP [39]. All columns except "SSIVI" are taken from [40]. They show the mean and the standard deviation of the variational posterior under DSVI [41], two MAP solutions under Mode A and Mode B, and the posterior function samples under SGHMC and SSIVI.



$$\mathcal{H}[q_\phi(z)] \geq \underline{\mathcal{H}}_K^{\text{SSIVI}}[q_\phi(z)] =$$
$$= -\sum_{i=1}^{d} \mathbb{E}_{z_{1..i-1}} \mathbb{E}_{\epsilon_i^{0..K}} \mathbb{E}_{z_i \mid z_{1..i-1}, \epsilon_i^0} \log \frac{1}{K+1} \sum_{k=0}^{K} q_\phi(z_i \mid z_{1..i-1}, \epsilon_i^k). \tag{26}$$

We have shown that a structured bound introduces a much lower inference gap than the original bound, essentially using an exponentially large mixture of $(K+1)^d$ distributions, placed on a d-dimensional grid, instead of a mixture of $K+1$ distributions.

Theorem 4. *For a structured semi-implicit distribution* (23)*, the following inequalities hold:*

$$\mathcal{H}[q_\phi(z)] \geq \underline{\mathcal{H}}_K^{\text{SSIVI}}[q_\phi(z)] \geq \underline{\mathcal{H}}_K^{\text{SIVI}}[q_\phi(z)]. \tag{27}$$

In practice, this makes a large difference, as shown in Figure 7.

Finally, to show the generality of DSIVI, we apply it to a complex deep probabilistic model, a deep Gaussian process [39], and show that, unlike plain doubly stochastic variational inference [41], it can successfully recover the multimodal nature of the posterior distribution [40]. See Figure 8 for details.

Conclusion

The main results of the work can be summarized as follows.

1. A new way to estimate the variational objective of variational dropout is proposed. This has lifted the limitations on the space of variational

parameters of the model and allowed to discover the sparsity inducing properties of the model, resulting in a practical method for compressing deep neural networks. This method has been applied to deep convolutional networks, and extended to other popular architectures in consecutive works.

2. Several parameterizations of the variational dropout model have been proposed to remove a specific class of local optima, causing a phase transition during the training and ultimately leading to a new mode of operation of neural networks. The resulting model, variance networks, uses a zero-mean posterior approximation to represent the object features as zero-mean distributions, resulting in high sample diversity and robustness to adversarial attacks after performing approximate posterior averaging. The discovered phase transition is not limited to variational dropout and can be observed in other models with related prior distributions, such as the automatic relevance determination prior and the Student's t-distribution prior. This study highlights the importance of parameterization of variational inference objectives, and shows how different parameterizations can lead to different properties of the resulting model.
3. We propose and study doubly semi-implicit variational inference, a new variational inference algorithm. It provides a way to estimate a proper variational lower bound for models with semi-implicit posterior approximations and prior distributions. DSIVI provides an asymptotically exact lower bound on the standard evidence lower bound with two hyperparameters that allow to trade-off the computational complexity and the introduced inference gap. We apply DSIVI to obtain a practical algorithm for training variational autoencoders with optimal priors, to recover a multimodal posterior of a deep Gaussian process, and study it on several synthetic tasks. We also highlight the importance of marginalization in models with hierarchical priors. DSIVI significantly expands the toolset, available for Bayesian deep learning practitioners, allows to perform inference with a broader range of distributions, and, unlike most advanced modern Bayesian deep learning techniques, supports Bayesian incremental learning.

Bibliography

1. *Kingma D. P.*, *Welling M.* Auto-encoding variational bayes // ICLR. — 2014.
2. *Titsias M.*, *Lázaro-Gredilla M.* Doubly stochastic variational Bayes for non-conjugate inference // International Conference on Machine Learning. — 2014. — P. 1971–1979.
3. *Kingma D. P.*, *Salimans T.*, *Welling M.* Variational dropout and the local reparameterization trick // Advances in Neural Information Processing Systems. — 2015. — P. 2575–2583.
4. *Yin M.*, *Zhou M.* Semi-Implicit Variational Inference // Proceedings of the 35th International Conference on Machine Learning. Vol. 80. — PMLR, 2018. — P. 5660–5669. — URL: http://proceedings.mlr.press/v80/yin18b.html.
5. *Goodfellow I.*, *Bengio Y.*, *Courville A.* Deep Learning. — MIT Press, 2016. — http://www.deeplearningbook.org.
6. *Hoffman M. D.*, *Blei D. M.*, *Wang C.*, *Paisley J.* Stochastic Variational Inference // Journal of Machine Learning Research. — 2013. — Vol. 14. — P. 1303–1347.
7. *Bishop C. M.* Pattern recognition and machine learning. — springer, 2006.
8. *Kirkpatrick J.* [et al.]. Overcoming catastrophic forgetting in neural networks // Proceedings of the National Academy of Sciences. — 2017. — Vol. 114, no. 13. — P. 3521–3526. — eprint: https://www.pnas.org/content/114/13/3521.full.pdf. — URL: https://www.pnas.org/content/114/13/3521.
9. *Kochurov M.*, *Garipov T.*, *Podoprikhin D.*, *Molchanov D.*, *Ashukha A.*, *Vetrov D.* Bayesian Incremental Learning for Deep Neural Networks. — 2018. — URL: https://openreview.net/forum?id=ByZzFPJDG.
10. *Welling M.*, *Teh Y. W.* Bayesian learning via stochastic gradient Langevin dynamics // Proceedings of the 28th International Conference on Machine Learning (ICML-11). — 2011. — P. 681–688.
11. *Chen T.*, *Fox E.*, *Guestrin C.* Stochastic gradient hamiltonian monte carlo // International Conference on Machine Learning. — 2014. — P. 1683–1691.
12. *Zhang R.*, *Li C.*, *Zhang J.*, *Chen C.*, *Wilson A. G.* Cyclical Stochastic Gradient MCMC for Bayesian Deep Learning // arXiv preprint arXiv:1902.03932. — 2019.
13. *Hoffman M. D.*, *Blei D. M.*, *Wang C.*, *Paisley J.* Stochastic variational inference // The Journal of Machine Learning Research. — 2013. — Vol. 14, no. 1. — P. 1303–1347.
14. *Srivastava N.*, *Hinton G. E.*, *Krizhevsky A.*, *Sutskever I.*, *Salakhutdinov R.* Dropout: a simple way to prevent neural networks from overfitting. // Journal of Machine Learning Research. — 2014. — Vol. 15, no. 1. — P. 1929–1958.

15. *Ioffe S.*, *Szegedy C.* Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift // CoRR. — 2015. — Vol. abs/1502.03167. — arXiv: 1502.03167.

16. *Titsias M. K.*, *Ruiz F.* Unbiased implicit variational inference // The 22nd International Conference on Artificial Intelligence and Statistics. — PMLR. 2019. — P. 167–176.

17. *Ranganath R.*, *Tran D.*, *Altosaar J.*, *Blei D.* Operator variational inference // Advances in Neural Information Processing Systems. — 2016. — P. 496–504.

18. *Mescheder L.*, *Nowozin S.*, *Geiger A.* Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks // arXiv preprint arXiv:1701.04722. — 2017.

19. *Shi J.*, *Sun S.*, *Zhu J.* Kernel implicit variational inference // arXiv preprint arXiv:1705.10119. — 2017.

20. *Ranganath R.*, *Tran D.*, *Blei D.* Hierarchical variational models // International Conference on Machine Learning. — 2016. — P. 324–333.

21. *Louizos C.*, *Welling M.* Multiplicative Normalizing Flows for Variational Bayesian Neural Networks // arXiv preprint arXiv:1703.01961. — 2017.

22. *Tomczak J. M.*, *Welling M.* VAE with a VampPrior // arXiv preprint arXiv:1705.07120. — 2017.

23. *Han S.*, *Pool J.*, *Tran J.*, *Dally W.* Learning both weights and connections for efficient neural network // Advances in Neural Information Processing Systems. — 2015. — P. 1135–1143.

24. *Han S.*, *Mao H.*, *Dally W. J.* Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding // arXiv preprint arXiv:1510.00149. — 2015.

25. *Guo Y.*, *Yao A.*, *Chen Y.* Dynamic network surgery for efficient dnns // Advances In Neural Information Processing Systems. — 2016. — P. 1379–1387.

26. *Ullrich K.*, *Meeds E.*, *Welling M.* Soft Weight-Sharing for Neural Network Compression // arXiv preprint arXiv:1702.04008. — 2017.

27. *LeCun Y.*, *Bottou L.*, *Bengio Y.*, *Haffner P.* Gradient-based learning applied to document recognition // Proceedings of the IEEE. — 1998. — Vol. 86, no. 11. — P. 2278–2324.

28. *Gal Y.*, *Ghahramani Z.* Dropout as a Bayesian approximation: Representing model uncertainty in deep learning // arXiv preprint arXiv:1506.02142. — 2015. — Vol. 2.

29. *Wan L.*, *Zeiler M.*, *Zhang S.*, *Cun Y. L.*, *Fergus R.* Regularization of neural networks using dropconnect // Proceedings of the 30th International Conference on Machine Learning (ICML-13). — 2013. — P. 1058–1066.

30. *Zagoruyko S.* 92.45 on CIFAR-10 in Torch. — 2015. — URL: http://torch.ch/blog/2015/07/30/cifar.html.

31. *Goodfellow I. J.*, *Shlens J.*, *Szegedy C.* Explaining and harnessing adversarial examples // arXiv preprint arXiv:1412.6572. — 2014.

32. *Lakshminarayanan B.*, *Pritzel A.*, *Blundell C.* Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles // Advances in Neural Information Processing Systems 30 / ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett. — Curran Associates, Inc., 2017. — P. 6405–6416.

33. *Molchanov D.*, *Ashukha A.*, *Vetrov D.* Variational Dropout Sparsifies Deep Neural Networks // Proceedings of the 34th International Conference on Machine Learning. Vol. 70 / ed. by D. Precup, Y. W. Teh. — PMLR, 06–11 Aug/2017. — P. 2498–2507. — (Proceedings of Machine Learning Research). — URL: http://proceedings.mlr.press/v70/molchanov17a.html.

34. *Gretton A.*, *Borgwardt K. M.*, *Rasch M. J.*, *Schölkopf B.*, *Smola A.* A kernel two-sample test // Journal of Machine Learning Research. — 2012. — Vol. 13, Mar. — P. 723–773.

35. *Neal R. M.* Bayesian learning for neural networks. Vol. 118. — 1995.

36. *Tipping M.* Sparse Bayesian Learning and the Relevance Vector Machine. — 2000.

37. *Hernández-Lobato J. M.*, *Adams R.* Probabilistic backpropagation for scalable learning of bayesian neural networks // International Conference on Machine Learning. — 2015. — P. 1861–1869.

38. *Louizos C.*, *Ullrich K.*, *Welling M.* Bayesian compression for deep learning // Advances in Neural Information Processing Systems. — 2017. — P. 3288–3298.

39. *Damianou A.*, *Lawrence N.* Deep gaussian processes // Artificial Intelligence and Statistics. — 2013. — P. 207–215.

40. *Havasi M.*, *Hernández-Lobato J. M.*, *Murillo-Fuentes J. J.* Inference in Deep Gaussian Processes using Stochastic Gradient Hamiltonian Monte Carlo // Advances in Neural Information Processing Systems. — 2018. — P. 7517–7527.

41. *Salimbeni H.*, *Deisenroth M.* Doubly stochastic variational inference for deep Gaussian processes // Advances in Neural Information Processing Systems. — 2017. — P. 4588–4599.