

Data Augmentation for Bayesian Deep Learning

Yuxi Wang*, Nicholas Polson†, and Vadim O. Sokolov‡

Abstract. Deep Learning (DL) methods have emerged as one of the most powerful tools for functional approximation and prediction. While the representation properties of DL have been well studied, uncertainty quantification remains challenging and largely unexplored. Data augmentation techniques are a natural approach to provide uncertainty quantification and to incorporate stochastic Monte Carlo search into stochastic gradient descent (SGD) methods. The purpose of our paper is to show that training DL architectures with data augmentation leads to efficiency gains. We use the theory of scale mixtures of normals to derive data augmentation strategies for deep learning. This allows variants of the expectation-maximization and MCMC algorithms to be brought to bear on these high dimensional nonlinear deep learning models. To demonstrate our methodology, we develop data augmentation algorithms for a variety of commonly used activation functions: logit, ReLU, leaky ReLU and SVM. Our methodology is compared to traditional stochastic gradient descent with back-propagation. Our optimization procedure leads to a version of iteratively re-weighted least squares and can be implemented at scale with accelerated linear algebra methods providing substantial improvement in speed. We illustrate our methodology on a number of standard datasets. Finally, we conclude with directions for future research.

Keywords: deep learning, data augmentation, MCMC, back-propagation, SGD.

1 Introduction

Deep neural networks (DNNs) have become a central tool for Artificial Intelligence (AI) applications such as, image processing (ImageNet, Krizhevsky et al. (2012)), object recognition (ResNet, He et al. (2016)) and game intelligence (AlphaGoZero, Silver et al. (2016)). The approximability (Poggio et al., 2017; Bauer and Kohler, 2019) and rate of convergence of deep learning, either in the frequentist fashion (Schmidt-Hieber, 2020) or from a Bayesian predictive point of view (Polson and Rockova, 2018; Wang and Rockova, 2020), have been well-explored and understood. Fan et al. (2021) provides a selective overview of deep learning. However, training deep learners is challenging due to the high dimensional search space and the non-convex objective function. Deep neural networks have also suffered from issues such as local traps, miscalibration and overfitting. Various efforts have been made to improve the generalization performance and many of their roots lie in Bayesian modeling. For example, Dropout (Wager et al., 2013) is commonly used and can be viewed as a deterministic ridge ℓ_2 regularization. Sparsity structure via spike-and-slab priors (Polson and Rockova, 2018) on weights

arXiv: 1903.09668

*Booth School of Business, University of Chicago, Chicago, IL, yuxi.wang@chicagobooth.edu

†Booth School of Business, University of Chicago, Chicago, IL, ngp@chicagobooth.edu

‡Volgenau School of Engineering, George Mason University, Fairfax, VA, vsokolov@gmu.edu

helps DNNs adapt to smoothness and avoid overfitting. Rezende et al. (2014) propose stochastic back-propagation through the use of latent Gaussian variables.

In this paper, following the spirit of hierarchical Bayesian modeling, we develop data augmentation strategies for deep learning with a complete data likelihood function equivalent to weighted least squares regression. By using the theory of mean-variance mixtures of Gaussians, our latent variable representation brings all of the conditionally linear model theory to deep learning. For example, it allows for the straightforward specification of uncertainty at each layer of deep learning and for a wide range of regularization penalties. Our method applies to commonly used activation functions such as ReLU, leaky ReLU, logit (see also Gan et al. (2015)), and provides a general framework for training and inference in DNNs. It inherits the advantages and disadvantages of data augmentation schemes. For *approximation* methods like Expectation-Maximization (EM) and Minorize-Maximization (MM), they are stable as they increase the objective but can be slow in the neighborhood of the maximum point even with acceleration methods such as Nesterov acceleration available and the performance is highly dependent on the properties of the objective function. Stochastic *exploratory* methods like MCMC have the main advantage of addressing uncertainty quantification (UQ) and are stable in the sense they require no tuning. Hyper-parameter estimation is immediately available using traditional Bayesian methods. DA augments the objective function with extra hidden units which allow for efficient step size selection for the gradient descent search. In some of the applications, data augmentation methods can be formulated in terms of complete data sufficient statistics, a considerable advantage when dealing with large datasets where most of the computational expense comes from repeatedly iterating over the data. By combining the MCMC methods with the J-copies trick (Jacquier et al., 2007), we can move faster towards posterior mode and avoid local maxima. Traditional methods for training deep learning models such as stochastic gradient descent (SGD) have none of the above advantages. We also note that we exploit the advantages of SGD and accelerated linear algebra methods when we implement our weighted least squares regression step.

Data augmentation strategies are commonplace in statistical algorithms and accelerated convergence (Nesterov, 1983; Green, 1984) is available. Our goal is to show similar efficiency improvements for deep learning. Our work builds on Deng et al. (2019) who provide adaptive empirical Bayes methods. In particular, we show how to implement standard activation functions, including ReLU (Polson and Rockova, 2018), logistic (Zhou et al., 2012; Hernández-Lobato and Adams, 2015) and SVM (Mallick et al., 2005) activation functions and provide specific data augmentation strategies and algorithms. The core subroutine of the resulting algorithms solves a least squares problem. Scalable linear algebra libraries such as Compute Unified Device Architecture (CUDA) and accelerated linear algebra (XLA) are available for implementation. To illustrate our approach, empirically we experiment with two benchmark datasets using Pólya-Gamma data augmentation for logit activation functions. For the deep architecture embedded in our approach, we adopt deep ReLU networks. Deep networks are able to achieve the same level of approximation accuracy with exponentially fewer parameters for compositional functions (Mhaskar et al., 2017). Poggio et al. (2017) further show how deep networks can avoid the curse of dimensionality. The ReLU function is favored due to its ability to avoid vanishing gradients and its expressibility and inherent sparsity. Ap-

proximation properties of deep ReLU networks have been developed in Montufar et al. (2014), Telgarsky (2017), and Liang and Srikant (2017). Yarotsky (2017) and Schmidt-Hieber (2020) show that deep ReLU networks can yield a rate-optimal approximation of smooth functions of an arbitrary order. Polson and Rockova (2018) provide posterior rates of convergence for sparse deep learning.

There is another active area of research that revives traditional statistical models with the computational power of DL (Bhadra et al., 2021). Examples include Gaussian Process models (Higdon et al., 2008; Gramacy and Lee, 2008), Generalized Linear Models (GLM) and Generalized Linear Mixed Models (GLMM) (Tran et al., 2020) and Partial Least Squares (PLS) (Polson et al., 2021). Our method benefits from the computation efficiency and flexibility of expression of the deep neural network. In addition, our work builds on the sampling optimization literature (Pincus, 1968, 1970) which now uses MCMC methods. Other examples include Ma et al. (2019) who study that sampling can be faster than optimization and Neelakantan et al. (2017) showing that gradient noise can improve learning for very deep networks. Gan et al. (2015) implements data augmentation inside learning deep sigmoid belief networks. Neal (2011) and Chen et al. (2014) provide Hamiltonian Monte Carlo (HMC) algorithms for MCMC. Duan et al. (2018) proposes a family of calibrated data-augmentation algorithms to increase the effective sample size.

The rest of our paper is outlined as follows. Section 2 provides the general setting of deep neural networks and shows how DA can be integrated into deep learning using the duality between Bayesian simulation and optimization. Section 3 describes our data augmentation (DA) schemes and two approaches to implement them. Section 4 provides applications to Gaussian regression, support vector machines and logistic regression using Pólya-Gamma augmentation (Polson et al., 2013). Section 5 provides the experiments of DA on both regression and classification problems. Section 6 concludes with directions for future research.

2 Bayesian Deep Learning

In deep learning we wish to recover a multivariate predictive map $f_{\theta}(\cdot)$ denoted by

$$\mathbf{y} = f_{\theta}(\mathbf{x}),$$

where $\mathbf{y} = (y_1, \dots, y_n)', y_i \in \mathbb{R}$ denotes a univariate output and $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)', \mathbf{x}_i \in \mathbb{R}^p$ a high-dimensional set of inputs. Using training data of input-output pairs $\{y_i, \mathbf{x}_i\}_{i=1}^n$ that generalizes well out-of-sample, the goal is to provide a predictive rule for a new input variable \mathbf{x}_*

$$y_* = f_{\hat{\theta}}(\mathbf{x}_*),$$

where $\hat{\theta}$ is estimated from training data typically using SGD. The interest in deep learners lies in their ability to perform better than the additive rule for those interpolation or prediction problems. Other statistical alternatives include Gaussian processes but they often have difficulty in handling higher dimensions.

Deep learners use compositions (Kolmogorov, 1957; Vitushkin, 1964) of ridge functions rather than additive functions that are commonplace in statistical applications.

With $L \in \mathbb{N}$ we denote the number of hidden layers and with $p_l \in \mathbb{N}$ the number of neurons at the l^{th} layer. Setting $p_{L+1} = p, p_0 = p_1 = 1$, we denote with $\mathbf{p} = (p_0, p_1, \dots, p_{L+1}) \in \mathbb{N}^{L+2}$ the vector of neuron counts for the entire network. Imagine composing L layers, a deep predictor then takes the form

$$\mathbf{y} = f_{\boldsymbol{\theta}}(\mathbf{x}) = (f_{W_0, b_0} \circ f_{W_1, b_1} \circ \dots \circ f_{W_L, b_L})(\mathbf{x}), \quad (2.1)$$

where $b_l \in \mathbb{R}^{p_l}$ is a shift vector, $W_l \in \mathbb{R}^{p_{l-1} \times p_l}$ is a weight matrix that links neurons between $(l-1)^{th}$ and l^{th} layers and $f_{W_l, b_l}(x) = f_l(W_l x + b_l)$ is a semi-affine function. We denote with $\boldsymbol{\theta} = \{(W_0, b_0), (W_1, b_1), \dots, (W_L, b_L)\}$ as the stacked parameters. We can rewrite the compositions in (2.1) with a set of latent variables $Z = (Z_1, Z_2, \dots, Z_L)'$ as

$$\begin{aligned} \mathbf{y} &= f_0(Z_1 W_0 + b_0), \\ Z_l &= f_l(Z_{l+1} W_l + b_l), \quad l = 1, \dots, L, \\ Z_{L+1} &= \mathbf{x}, \end{aligned} \quad (2.2)$$

where $Z_l \in \mathbb{R}^{n \times p_l}$ is the matrix of hidden nodes in l -th layer. We only consider the case $p = 1$ and $Z_1 \in \mathbb{R}^n$ in our work. We provide discussion on extensions to cases $p > 1$ for some of our applications in Section 4.

2.1 Bayesian Simulation and Regularization Duality

The problem of deep learning regularization (Polson and Sokolov, 2017) is to find a set of parameters $\boldsymbol{\theta}$ which minimizes a combination of a negative log-likelihood $\ell(\mathbf{y}, f_{\boldsymbol{\theta}}(\mathbf{x}))$ and a penalty function $\phi(\boldsymbol{\theta})$ defined by

$$\hat{\boldsymbol{\theta}} := \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^n \ell(y_i, f_{\boldsymbol{\theta}}(\mathbf{x}_i)) + \lambda \sum_{j=1}^{\#\boldsymbol{\theta}} \phi(\theta_j), \quad (2.3)$$

where λ controls regularization and $\#\boldsymbol{\theta}$ denotes the number of parameters in $\boldsymbol{\theta}$.

When the function $f_{\boldsymbol{\theta}}(\mathbf{x})$ is a deep learner defined as (2.1), we can specify different amount of penalty λ_l and form of regularization function $\phi_l(\cdot)$ for each layer. Then the objective function can be written as

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\boldsymbol{\theta}}(\mathbf{x}_i)) + \sum_{l=0}^L \lambda_l \phi_l(W_l, b_l). \quad (2.4)$$

Commonly used regularization techniques for deep learners include L^2 (weight decay), spike-and-slab regularization (Polson and Rockova, 2018) and dropout (Wager et al., 2013), which can also be viewed as a variant of L^2 -regularization.

As such the optimization problem in (2.4) of training a deep learner $f_{\boldsymbol{\theta}}(\cdot)$ involves a highly nonlinear objective function. Stochastic gradient descent (SGD) is a popular tool based on back-propagation (a.k.a. the chain rule), but it often suffers from local traps and overfitting due to the non-convex nature of the problem. We propose data

augmentation techniques which can be seamlessly applied in this context and provide efficiency gains. This is achieved via the hierarchical duality between optimization with regularization and finding the *maximum a posteriori* (MAP) estimate (Polson and Scott, 2011), as described in the following lemma.

Lemma 2.1. *The regularization problem*

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\boldsymbol{\theta}}(x_i)) + \sum_{l=0}^L \lambda_l \phi_l(W_l, b_l) \right\}$$

is equivalent to finding the Bayesian MAP estimator defined by

$$\arg \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} | \mathbf{y}) = \arg \max_{\boldsymbol{\theta}} \exp \left\{ -\frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\boldsymbol{\theta}}(x_i)) - \sum_{l=0}^L \lambda_l \phi_l(W_l, b_l) \right\},$$

which corresponds to the mode of a posterior distribution characterized as

$$\begin{aligned} p(\boldsymbol{\theta} | \mathbf{y}) &= p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) / p(\mathbf{y}), \\ p(\mathbf{y} | \boldsymbol{\theta}) &\propto \exp \left\{ -\sum_{i=1}^n \ell(y_i, f_{\boldsymbol{\theta}}(x_i)) \right\}, \quad p(\boldsymbol{\theta}) \propto \exp \left\{ -\sum_{l=0}^L \lambda_l \phi_l(W_l, b_l) \right\}. \end{aligned}$$

Here $p(\boldsymbol{\theta})$ can be interpreted as a prior probability distribution and the log-prior as the regularization penalty.

2.2 A Stochastic Top Layer

By exploiting the duality from Lemma 2.1, we wish to use a Bayesian framework to add stochastic layers – so as to fully account for the uncertainty in estimating the predictive rule $f_{\boldsymbol{\theta}}(\cdot)$. Thus, we convert the sequence of composite functions in the deep learner specified in (2.2) to a stochastic version given by

$$\begin{aligned} \mathbf{y} | Z_1 &\sim p(\mathbf{y} | Z_1), \\ Z_l &\sim N(f_l(W_l Z_{l+1} + b_l), \tau_l^2), \quad l = 1, 2, \dots, L, \\ Z_{L+1} &= \mathbf{x}. \end{aligned} \tag{2.5}$$

Now the hidden variables $Z = (Z_1, \dots, Z_L)'$ can be viewed as data augmentation variables and hence will also allow the contribution of fast scalable algorithms for inference and prediction.

For the ease of computation, we only replace the top layer of the DNN with a stochastic layer. We denote network structure below the top layer with $\mathbf{B} = \{(W_1, b_1), \dots, (W_L, b_L)\}$, and the network structure can be rewritten as

$$\mathbf{y} = f_0(Z_1 W_0 + b_0), \quad Z_1 = f_{\mathbf{B}}(\mathbf{x}),$$

where the function $f_0(Z_1 W_0 + b_0)$ is the top layer structure and function $f_{\mathbf{B}}(\mathbf{x})$ is the network architecture below the top layer. Considering the objective function in (2.4), we implement the solutions with a two-step iterative search. At iteration t , we have

1. DA-update for the top layer W_0, b_0 as the MAP estimator of the distribution

$$\begin{aligned} p(W_0, b_0 \mid Z_1^{(t)}, \mathbf{y}) &\propto p(\mathbf{y}, Z_1^{(t)} \mid W_0, b_0) p(W_0, b_0) \\ &\propto \exp \left\{ -\frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\theta}(x_i) \mid \mathbf{B}^{(t)}) + \lambda_0 \phi_0(W_0, b_0) \right\}. \end{aligned} \quad (2.6)$$

2. SGD-update for the deep architecture \mathbf{B}

$$\begin{aligned} \mathbf{B}^{(t+1)} &= \arg \min_{\mathbf{B}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\theta}(x_i) \mid (W_0, b_0)^{(t+1)}) + \sum_{l=1}^L \lambda_l \phi_l(W_l, b_l) \\ &= \arg \min_{\mathbf{B}} \frac{1}{n} \sum_{i=1}^n \ell(Z_1^{(t)}, f_{\mathbf{B}}(x_i)) + \sum_{l=1}^L \lambda_l \phi_l(W_l, b_l). \end{aligned}$$

3. Sample $Z_1^{(t+1)}$ from a normal distribution $\mathcal{N}(\mu_z^{(t)}, \sigma_z^{(t)})$ where $\mu_z^{(t)}$ and $\sigma_z^{(t)}$ are determined jointly by $\{\theta^{(t)}, \mathbf{x}, \mathbf{y}\}$.

The main contribution of our work comes from two aspects: (1) we update top layer weights $\{W_0, b_0\}$ conditional on \mathbf{B} as in (2.6), which is also equivalent to conditioning on Z_1 , with data augmentation techniques as later shown in Section 3; (2) the latent variables Z_1 is sampled from a normal distribution rather than optimized by gradient descent methods. Z_1 serves as a bridge that connects a weighted L^2 -norm model f_0 and a deep learner $f_{\mathbf{B}}$. Commonly used activation functions $\{f_l\}_{l=1}^L$ are linear affine functions, rectified linear units (ReLU), sigmoid, hyperbolic tangent (tanh), and etc. We illustrate our methods with a deep ReLU network, i.e., $\{f_l\}_{l=1}^L$ are ReLU functions, due to its expressibility and inherent sparsity. In the next section, we introduce our data augmentation strategies and show how the stochastic layers can be achieved via data augmentation.

3 Data Augmentation for Deep Learning

Data augmentation introduces a vector of auxiliary variables, denoted by $\omega = (\omega_1, \dots, \omega_n)'$ with $\omega_i \in \mathbb{R}$, such that the posterior can be written as

$$p(\theta \mid \mathbf{y}) = E_{\omega} [p(\theta, \omega \mid \mathbf{y})],$$

where the augmented auxiliary distribution, $p(\theta, \omega \mid \mathbf{y})$ factorizes nicely into complete conditionals $p(\theta \mid \omega, \mathbf{y})$ and $p(\omega \mid \theta, \mathbf{y})$. A crucial ingredient is that $p(\theta \mid \omega, \mathbf{y})$ is easily managed typically via conditional Gaussians.

Data augmentation tricks allow us to express the likelihood as an expectation of a weighted L^2 -norm. Specifically, we write

$$\begin{aligned} \exp \left\{ -\ell(\mathbf{y}, f_{\theta}(\mathbf{x})) \right\} &= E_{\omega} \left\{ \exp \left(-Q(\mathbf{y} \mid f_{\theta}(\mathbf{x}), \omega) \right) \right\} \\ &= \int_0^{\infty} \exp \left(-Q(\mathbf{y} \mid f_{\theta}(\mathbf{x}), \omega) \right) p(\omega) d\omega, \end{aligned}$$

where $p(\boldsymbol{\omega})$ is the prior on the auxiliary variables $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)'$ and the function $Q(\mathbf{y} | f_{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\omega})$ is designed to be a quadratic form, given the data augmentation variables $\boldsymbol{\omega}$. The function $f_{\boldsymbol{\theta}}(\mathbf{x}) = (f_0 \circ \dots \circ f_L)(\mathbf{x})$ is a deep learner.

Table 1 shows that standard activation functions such as ReLU, logit, lasso and check can be expressed in the form of (3.1). Commonly used activation functions for deep learning, with an appropriate stochastic assumptions for w (for notation of simplicity, we derive the standard form for the single observation case) can be expressed as

$$\begin{aligned} \exp(-\max(1-x, 0)) &= E_{\omega} \left\{ \frac{1}{\sqrt{2\pi\omega}} \exp\left(-\frac{1}{2\omega}(x-1-\omega)^2\right) \right\}, \quad \text{where } \omega \sim \mathcal{GIG}(1, 0, 0), \\ \exp(-\log(1+e^x)) &= E_{\omega} \left\{ \exp\left(-\frac{1}{2}\omega x^2\right) \right\}, \quad \text{where } \omega \sim \mathcal{PG}(1, 0), \\ \exp(-|x|) &= E_{\omega} \left\{ \frac{1}{\sqrt{2\pi\omega}} \exp\left(-\frac{1}{2\omega}x^2\right) \right\}, \quad \text{where } \omega \sim \mathcal{E}\left(\frac{1}{2}\right). \end{aligned}$$

Here \mathcal{GIG} denotes the Generalized Inverse Gaussian distribution, \mathcal{PG} represents the Pólya Gamma distribution (Polson et al., 2013), and \mathcal{E} represents the exponential distribution.

$l(W, b)$	$Q(W, b, \omega)$	$p(\omega)$
ReLU: $\max(1 - z_i, 0)$	$\int_0^\infty \frac{1}{\sqrt{2\pi c \lambda}} \exp\left\{-\frac{(x+a\lambda)^2}{2c\lambda}\right\} d\lambda = \frac{1}{a} \exp\left(-\frac{2\max(ax, 0)}{c}\right)$	$\mathcal{GIG}(1, 0, 0)$
Logit: $\log(1 + e^{z_i})$	$\frac{1}{2^b} e^{(a-b/2)\psi} \int_0^\infty e^{-\omega\psi^2/2} p(\omega) d\omega = \frac{(e^\psi)^a}{(1+e^\psi)^b}$	$\mathcal{PG}(b, 0)$
Lasso: $ z_i $	$\int_0^\infty \frac{1}{\sqrt{2\pi c \lambda}} \exp\left\{-\frac{x^2}{2c\lambda}\right\} e^{-\frac{1}{2}\lambda} d\lambda = \frac{1}{c} \exp\left(-\frac{ x }{c}\right)$	$\mathcal{E}(\frac{1}{2})$
Check: $ z_i + (2\tau - 1)z_i$	$\int_0^\infty \frac{1}{\sqrt{2\pi c \lambda}} \exp\left\{-\frac{(x+(2\tau-1)\lambda)^2}{2c^2\lambda}\right\} e^{-2\tau(1-\tau)\lambda} d\lambda = \frac{1}{c} \exp\left(-\frac{2}{c}\rho_\tau(x)\right)$	$\mathcal{GIG}(1, 0, 2\sqrt{\tau-\tau^2})$

Table 1: Data Augmentation Strategies. Here $\rho_\tau(x) = \frac{1}{2}|x| + (\tau - \frac{1}{2})x$ is the check function.

Using the data augmentation strategies, the objectives are represented as mixtures of Gaussians. DA can perform such an optimization with only the use of a sequence of iteratively re-weighted L^2 -norms. This allows us to use XLA techniques to accelerate the training.

Remark 3.1. *The log-posterior is optimized given the training data, $\{y_i, \mathbf{x}_i\}_{i=1}^n$. Deep learning possesses the key property that $\nabla_{\boldsymbol{\theta}} \log p(\mathbf{y} | \boldsymbol{\theta}, \mathbf{x})$ is computationally inexpensive to evaluate using tensor methods for very complicated architectures and fast implementation on large datasets. One caveat is that the posterior is highly multi-modal and providing good hyperparameter tuning can be expensive. This is clearly a fruitful area of research for state-of-the-art stochastic MCMC algorithms to provide more efficient algorithms. For shallow architectures, the alternating direction method of multipliers (ADMM) is an efficient solution to the optimization problem.*

Similarly we can represent the regularization penalty $\exp(-\lambda\phi(\boldsymbol{\theta}))$ in data augmentation form. Hence, we can then replace the optimization problem in (2.4) with

$$\hat{\boldsymbol{\theta}} := \arg \max_{\boldsymbol{\theta}} E_{\boldsymbol{\omega}} \left[\exp \left(-\frac{1}{n} \sum_{i=1}^n Q(y_i | f_{\boldsymbol{\theta}}(\mathbf{x}_i), \boldsymbol{\omega}) - \sum_{l=0}^L \lambda_l \phi_l(W_l, b_l) \right) \right], \quad (3.1)$$

using the duality in Lemma 2.1.

There are two approaches to Monte Carlo optimization which could handle our data augmentation (Geyer, 1996), missing data methods like Expectation-Maximization (EM) algorithms or stochastic search methods like Markov Chain Monte Carlo (MCMC). The first approach is based on a probabilistic *approximation* of the objective function (3.1) and is less concerned with exploring Θ . The second type is more *exploratory* which aims to optimize the objective function by visiting the entire range of Θ and is less tied to the properties of the function.

For EM algorithms, we consider constructing a surrogate optimization problem which has the same solution to (3.1) (Lange et al., 2000). Specifically, we define a new objective function as

$$H(\boldsymbol{\theta}) = E_{\boldsymbol{\omega}|\boldsymbol{\theta}} \left[\exp \left(-\frac{1}{n} \sum_{i=1}^n Q(y_i | f_{\boldsymbol{\theta}}(\mathbf{x}_i), \boldsymbol{\omega}) - \sum_{l=0}^L \lambda_l \phi_l(W_l, b_l) \right) \right],$$

which is a concave function to be maximized. A natural choice of the surrogate function can be constructed using Jensen's inequality as

$$G(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) = -E_{\boldsymbol{\omega}|\boldsymbol{\theta}^{(t)}} \left[\frac{1}{n} \sum_{i=1}^n Q(y_i | f_{\boldsymbol{\theta}}(\mathbf{x}_i), \boldsymbol{\omega}) + \sum_{l=0}^L \lambda_l \phi_l(W_l, b_l) \right],$$

where each ω_i is drawn from conditional distribution $p(\omega_i | \boldsymbol{\theta}) \propto p(\omega_i, \boldsymbol{\theta})$ and the minorization is satisfied as

$$\log H(\boldsymbol{\theta}) \geq G(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}).$$

Maximizing $G(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$ drives $H(\boldsymbol{\theta})$ uphill. The ascent property of the EM algorithm relies on the nonnegativity of the Kullback-Leibler divergence of two conditional probability densities (Hunter and Lange, 2004; Lange, 2013a). The EM algorithm enjoys the numerical stability as it steadily increases the likelihood without wildly overshooting or undershooting. It simplifies the optimization problem by (1) avoiding large matrix inversion; (2) linearizing the objective function; (3) separating the variables of the optimization problem (Lange, 2013b). In Section 4.3 we show how Pólya-Gamma augmentation leads to an EM algorithm for logistic regression.

The exploratory alternative to solve (3.1) is stochastic search methods such as MCMC. The data augmentation strategies enable us to sample from the joint posterior

$$p(\boldsymbol{\theta} | \mathbf{y}) \propto \exp \left(-\frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\boldsymbol{\theta}}(\mathbf{x}_i)) - \sum_{l=0}^L \lambda_l \phi_l(W_l, b_l) \right)$$

$$\begin{aligned}
&= E_{\boldsymbol{\omega}} \left[\exp \left(-\frac{1}{n} \sum_{i=1}^n Q(y_i | f_{\boldsymbol{\theta}}(\mathbf{x}_i), \boldsymbol{\omega}) - \sum_{l=0}^L \lambda_l \phi_l(W_l, b_l) \right) \right] \\
&= \int_0^\infty \exp \left(-\frac{1}{n} \sum_{i=1}^n Q(y_i | f_{\boldsymbol{\theta}}(\mathbf{x}_i), \boldsymbol{\omega}) \right) p(\boldsymbol{\omega}) p(\boldsymbol{\theta}) d\boldsymbol{\omega},
\end{aligned}$$

where the prior is related to the regularization penalty, via

$$p(\boldsymbol{\theta}) \propto \exp \left(-\sum_{l=0}^L \lambda_l \phi_l(W_l, b_l) \right).$$

Hence, we can provide an MCMC algorithm in the augmented space $(\boldsymbol{\theta}, \boldsymbol{\omega})$ and simulate from the joint posterior distribution, denoted by $p(\boldsymbol{\theta}, \boldsymbol{\omega} | \mathbf{y})$, namely

$$p(\boldsymbol{\theta}, \boldsymbol{\omega} | \mathbf{y}) \propto \exp \left(-Q(\mathbf{y} | f_{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\omega}) \right) p(\boldsymbol{\theta}) p(\boldsymbol{\omega}).$$

A sequence can be simulated using MCMC Gibbs conditionals,

$$\begin{aligned}
p(\boldsymbol{\theta}^{(t)} | \boldsymbol{\omega}^{(t)}, \mathbf{y}) &\propto \exp \left(-Q(\mathbf{y} | f_{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\omega}^{(t)}) \right) p(\boldsymbol{\theta}), \\
p(\boldsymbol{\omega}^{(t+1)} | \boldsymbol{\theta}^{(t)}, \mathbf{y}) &\propto \exp \left(-Q(\mathbf{y} | f_{\boldsymbol{\theta}^{(t)}}(\mathbf{x}), \boldsymbol{\omega}) \right) p(\boldsymbol{\omega}).
\end{aligned}$$

Then we recover stochastic draws $\boldsymbol{\theta}^{(t)} \sim p(\boldsymbol{\theta} | \mathbf{y})$ from the marginal posterior. These draws can be used in prediction to account for *predictive uncertainty*, namely

$$p(y_* | f(\mathbf{x}_*)) = \int p(y_* | \boldsymbol{\theta}, f_{\boldsymbol{\theta}}(\mathbf{x}_*)) p(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta} \approx \frac{1}{T} \sum_{t=1}^T p(y_* | \boldsymbol{\theta}^{(t)}, f_{\boldsymbol{\theta}^{(t)}}(\mathbf{x}_*)). \quad (3.2)$$

As $Q(\mathbf{y} | f_{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\omega})$ is conditionally quadratic, the update step for $\boldsymbol{\theta} | \boldsymbol{\omega}, \mathbf{y}$ can be achieved using SGD or a weighted L^2 -norm – the weights $\boldsymbol{\omega}$ are adaptive and provide an automatic choice of the learning rate, thus avoiding backtracking which can be computationally expensive. And the performance of MCMC search is less tied to the statistical properties (i.e. convexity or concavity) of the objective function. We provide examples of how Gaussian regression and SVMs can be implemented in Section 4.1 and Section 4.2.

3.1 MCMC with J-copies

The MCMC methods offer a full description of the objective function (3.1) over the entire space Θ . Inspired by the simulated annealing algorithm (Metropolis et al., 1953), we introduce a scaling factor J to allow for faster moves on the surface of (3.1) to maximize. It also helps avoiding the trapping attraction of local maxima. In addition, the corresponding posterior is connected to the Boltzmann distribution, whose density is prescribed by the energy potential $f(\boldsymbol{\theta})$ and temperature J as

$$\pi_J(\boldsymbol{\theta}) = \exp \{ -Jf(\boldsymbol{\theta}) \} / Z_J \text{ for } \boldsymbol{\theta} \in \Theta, \quad (3.3)$$

where $Z_J = \int_{\Theta} \exp \{ -Jf(\boldsymbol{\theta}) \} d\boldsymbol{\theta}$ is an appropriate normalizing constant.

To simulate the posterior mode without evaluating the likelihood directly (Jacquier et al., 2007), we sample J independent copies of hidden variable Z_1 . Denoted the copies with Z_1^1, \dots, Z_1^J , we sample them simultaneously and independently from the posterior distribution

$$Z_1^j | \boldsymbol{\theta}, \mathbf{x}, \mathbf{y} \stackrel{iid}{\sim} \mathcal{N}(\mu_z, \sigma_z^2), \quad j = 1, \dots, J,$$

where μ_z, σ_z are determined by $\{\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}\}$. And we stack the J copies as

$$\mathbf{y}^{(S)} = \begin{bmatrix} \mathbf{y} \\ \mathbf{y} \\ \mathbf{y} \\ \vdots \\ \mathbf{y} \end{bmatrix}, \quad Z_1^{(S)} = \begin{bmatrix} Z_1^1 \\ Z_1^2 \\ Z_1^3 \\ \vdots \\ Z_1^J \end{bmatrix}, \quad f_{\mathbf{B}}(\mathbf{x}^{(S)}) = \begin{bmatrix} f_{\mathbf{B}}(\mathbf{x}) \\ f_{\mathbf{B}}(\mathbf{x}) \\ f_{\mathbf{B}}(\mathbf{x}) \\ \vdots \\ f_{\mathbf{B}}(\mathbf{x}) \end{bmatrix}, \quad (3.4)$$

where $\mathbf{y}^{(S)}$, $Z_1^{(S)}$ and $f_{\mathbf{B}}(\mathbf{x}^{(S)})$ are $(n \times J)$ -dimensional vectors. We use $Z_1^{(S)}$ to amplify the information in \mathbf{y} , which is especially useful in the finite sample problems. Figure 1 illustrates our network architecture.

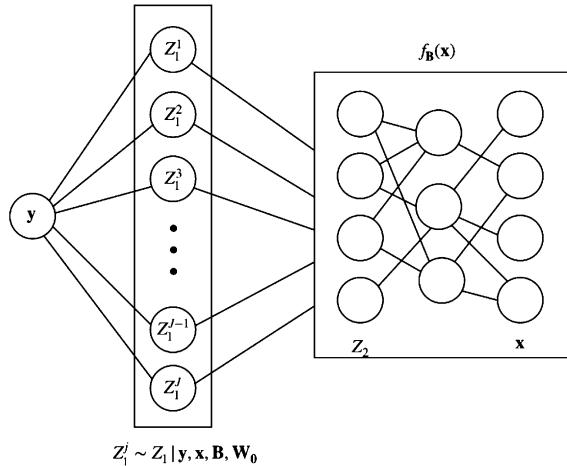


Figure 1: J -copies Network Architecture.

With the stacked system, the joint distribution of the parameters $\boldsymbol{\theta}$ and the augmented hidden variables $Z_1^{(S)}$ given data \mathbf{y}, \mathbf{x} can be written as

$$\pi_J(\boldsymbol{\theta}, Z_1^{(S)} | \mathbf{x}, \mathbf{y}) \propto \prod_{j=1}^J p(\mathbf{y} | \boldsymbol{\theta}, Z_1^j) p(Z_1^j | \boldsymbol{\theta}, \mathbf{x}, \mathbf{y}) p(\boldsymbol{\theta}).$$

Hence, the marginal joint posterior

$$p(\boldsymbol{\theta} | \mathbf{x}, \mathbf{y}) = \int \pi_J(\boldsymbol{\theta}, Z_1^{(S)} | \mathbf{x}, \mathbf{y}) dZ_1^{(S)}$$

concentrates on the density proportional to $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\theta})^J p(\boldsymbol{\theta})$ and provides us with a simulation solution to finding the MAP estimator (Pincus, 1968, 1970).

Another alternative to simulate from the posterior mode is Hamiltonian Monte Carlo (Neal, 2011), which is a modification of Metropolis-Hastings (MH) sampler. Adding an additional momentum variable $\boldsymbol{\nu}$ to the Boltzmann distribution in (3.3), and generating draws from joint distribution

$$\pi_J(\boldsymbol{\theta}, \boldsymbol{\omega}) \propto \exp(-Jf(\boldsymbol{\theta}) - (1/2)\boldsymbol{\nu}^T M^{-1} \boldsymbol{\nu}),$$

where M is a mass matrix. Chen et al. (2014) adopt this approach in a deep learning setting.

3.2 Connection to Diffusion Theory

An alternative to the MCMC algorithm can be derived from diffusion theory (Phillips and Smith, 1996). For example, we can approximate the random walk Metropolis-Hastings algorithm with the Langevin diffusion L_t defined by the stochastic differential equation $dL_t = dB_t + \frac{1}{2}\nabla \log f(L_t)dt$, where B_t is the standard Brownian motion. More specifically, let $d := |\boldsymbol{\theta}|$, we write the random walk like transition as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \frac{\sigma^2}{2}\nabla \log f(\boldsymbol{\theta}^{(t)}) + \sigma\boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}_d(0, I_d)$ and σ^2 corresponds to the discretization size.

This can also be derived by taking a second-order approximation of $\log(f)$, namely

$$\begin{aligned} \log f(\boldsymbol{\theta}^{(t+1)}) &= \log f(\boldsymbol{\theta}^{(t)}) + (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})' \nabla \log f(\boldsymbol{\theta}^{(t)}) \\ &\quad - \frac{1}{2}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})' H(\boldsymbol{\theta}^{(t)}) (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}), \end{aligned}$$

where $H(\boldsymbol{\theta}^{(t)}) = -\nabla^2 \log f(\boldsymbol{\theta}^{(t)})$ is the Hessian matrix. By taking exponential transformation on both sides, the random walk type approximation to $f(\boldsymbol{\theta}^{(t+1)})$ is

$$\begin{aligned} f(\boldsymbol{\theta}^{(t+1)}) &\propto \exp\left\{(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})' \nabla \log f(\boldsymbol{\theta}^{(t)}) - \frac{1}{2}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})' H(\boldsymbol{\theta}^{(t)}) (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}^{(t+1)} - \tilde{\boldsymbol{\theta}}^{(t)})' H(\boldsymbol{\theta}^{(t)}) (\boldsymbol{\theta}^{(t+1)} - \tilde{\boldsymbol{\theta}}^{(t)})\right\}, \end{aligned}$$

where $\tilde{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}^{(t)} + H^{-1}(\boldsymbol{\theta}^{(t)})\nabla \log f(\boldsymbol{\theta}^{(t)})$. If we simplify this approximation by replacing $H(\boldsymbol{\theta}^{(t)})$ with $\sigma^{-2}I_p$, the Taylor approximation leads to updating step as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \sigma^2 \nabla \log f(\boldsymbol{\theta}^{(t)}) + \sigma\boldsymbol{\epsilon}_t.$$

Roberts and Rosenthal (1998) give further discussion on the choice of σ that would yield an acceptance rate of 0.574 to achieve optimal convergence rate.

Mandt et al. (2017) show that SGD can be interpreted as a multivariate Ornstein-Uhlenbeck process

$$d\boldsymbol{\theta}^{(t)} = -\eta A\boldsymbol{\theta}^{(t)}dt + \eta\sqrt{\frac{C}{S}}dW^{(t)},$$

here η is the constant learning rate, A is the symmetric Hessian matrix at the optimum and $\frac{C}{S}$ is the covariance of the mini-batch (of size S) gradient noise, which is assumed to be approximately constant near the local optimum of the loss. They also provide results on discrete-time dynamics on other Stochastic Gradient MCMC algorithms, such as Stochastic Gradient Langevin dynamics (SGLD) by Welling and Teh (2011) and Stochastic Gradient Fisher Scoring by Ahn et al. (2012).

Combing their results and the Langevin dynamics of MCMC algorithms, we can write the approximation of our DA-DL updating scheme as

$$\begin{aligned} \begin{bmatrix} W_0 \\ b_0 \end{bmatrix}^{(t+1)} &= \begin{bmatrix} W_0 \\ b_0 \end{bmatrix}^{(t)} + \sigma^2 \nabla \log f_0(Z_1^{(t)} W_0^{(t)} + b_0^{(t)}) + \sigma \boldsymbol{\epsilon}_{0t}, \\ \boldsymbol{B}^{(t+1)} &= \boldsymbol{B}^{(t)} - \eta \nabla^2 f_{\boldsymbol{B}^*}(\boldsymbol{x}) \boldsymbol{B}^{(t)} + \frac{C}{\sqrt{S}} \eta \boldsymbol{\epsilon}_{\boldsymbol{B}t}. \end{aligned}$$

Similar adaptive dynamics are also observed in other methods. Geman and Hwang (1986) show the convergence of the annealing process using Langevin equations. Slice sampling (Neal, 2003) adaptively chooses the step size based on the local properties of the density function. By constructing local quadratic approximations, it could adapt to the dependencies between variables. Murray et al. (2010) further propose elliptical slice sampling that operates on the ellipse of states.

4 Applications

To illustrate our methodology, we provide three examples: (1) a standard Gaussian regression model with squared loss; (2) a binary classification model under the support vector machine framework; (3) a logistic regression model paired with a Pólya mixing distribution. For the Gaussian regression and SVM models, we implement with J -copies stacking strategy to provide full posterior modes.

Before diving into the examples, we introduce the notations we use throughout this section. We continue to denote the output with $\boldsymbol{y} = (y_1, \dots, y_n)', y_i \in \mathbb{R}$, the input with $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)', \boldsymbol{x}_i \in \mathbb{R}^p$, the latent variable of the top layer with $Z_1 = (z_{1,1}, \dots, z_{1,n})', z_{1,i} \in \mathbb{R}$ and the stacked version as in (3.4). We introduce stochastic noises $\boldsymbol{\epsilon}_0 = (\epsilon_{0,1}, \dots, \epsilon_{0,n})'$ in the top layer and $\boldsymbol{\epsilon}_z = (\epsilon_{z,1}, \dots, \epsilon_{z,n})'$ in the second layer, where $\epsilon_{0,i} \stackrel{iid}{\sim} \mathcal{N}(0, \tau_0^2)$ and $\epsilon_{z,i} \stackrel{iid}{\sim} \mathcal{N}(0, \tau_z^2)$. The scale parameters τ_0 and τ_z are pre-specified and determine the level of randomness or uncertainty for the DA-update and SGD-update respectively. We use η to denote the learning rate used in the SGD updates and T is number of training epochs. We use $\|\cdot\|$ to denote ℓ_2 -norm such that $\|\boldsymbol{y}\| = \sqrt{\sum_{i=1}^n y_i^2}$ and the matrix-type norm as $\|\boldsymbol{y}\|_\Sigma = \sqrt{\boldsymbol{y}^T \Sigma \boldsymbol{y}}$.

Our models differ from standard deep learning models and some newly proposed Bayesian approaches in the adoption of stochastic noises ϵ_0 and ϵ_z . It distinguishes our model from other deterministic neural networks. By letting ϵ_z follow a spiky distribution that puts most of its mass around zero, we can control the estimation approximating to posterior mode instead of posterior mean. The randomness allows us to adopt a stacked system and make the best use of data especially when the dataset is small.

4.1 Gaussian Regression

We consider the regression model as

$$\begin{aligned} y_i &= z_{1,i}W_0 + b_0 + \epsilon_{0,i}, & \text{where } y_i \in (-\infty, \infty), \epsilon_{0,i} &\stackrel{i.i.d.}{\sim} \mathcal{N}(0, \tau_0^2), \\ z_{1,i} &= f_{\mathbf{B}}(x_i) + \epsilon_{z,i}, & \text{where } \epsilon_{z,i} &\stackrel{i.i.d.}{\sim} \mathcal{N}(0, \tau_z^2). \end{aligned}$$

The posterior updates are given by

$$\hat{W}_0 = \text{Cov}(Z_1, \mathbf{y})/\text{Var}(Z_1), \quad (4.1)$$

$$\hat{b}_0 = \bar{\mathbf{y}} - W_0 \bar{Z}_1, \quad (4.2)$$

$$p(Z_1 | \mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = C_z \exp \left\{ -\frac{1}{2\tau_0^2} \|\mathbf{y} - Z_1 W_0 - b_0\|^2 - \frac{1}{2\tau_z^2} \|Z_1 - f_{\mathbf{B}}(\mathbf{x})\|^2 \right\},$$

where $\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n y_i$ and C_z is a normalizing constant. The latent variable Z_1 is drawn from following normal distribution $Z_1 \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ with the mean and variance specified as

$$\mu_Z = \frac{\tau_z^2 W_0 (\mathbf{y} - b_0) + \tau_0^2 f_{\mathbf{B}}(\mathbf{x})}{W_0^2 \tau_z^2 + \tau_0^2}, \quad \sigma_Z^2 = \frac{\tau_0^2 \tau_z^2}{W_0^2 \tau_z^2 + \tau_0^2}. \quad (4.3)$$

The J copies of Z_1 are simulated and stacked as

$$Z_1^j \stackrel{iid}{\sim} \mathcal{N}(\mu_Z, \sigma_Z^2), \quad Z_1^{(S)} = (Z_1^1, \dots, Z_1^J)'.$$

The updating scheme for this Gaussian regression is summarized in Algorithm 1.

The model can also be generalized to multivariate y . Let y_i be a q -dimensional vector, we denote each dimension as y_{ik} , $k = 1, \dots, q$, and the model is written as

$$\begin{aligned} y_{ik} &= z_{1,i}W_{0k} + b_{0k} + \epsilon_{0,ik}, & \text{where } y_{ik} \in (-\infty, \infty), \epsilon_{0,ik} &\stackrel{iid}{\sim} \mathcal{N}(0, \tau_0^2), \\ z_{1,i} &= f_{\mathbf{B}}(x_i) + \epsilon_{z,i}, & \text{where } \epsilon_{z,i} &\stackrel{iid}{\sim} \mathcal{N}(0, \tau_z^2), \end{aligned}$$

where $W_0 = (W_{01}, \dots, W_{0q})'$ is now a q -dimensional vector with W_{0k} computed similarly to (4.1), $b_0 = (b_{01}, \dots, b_{0q})'$ is also q -dimensional with b_{0k} calculated as (4.2). The posterior update for Z_1 becomes

$$p(Z_1 | \mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = C_z \exp \left\{ -\frac{1}{2\tau_0^2} \sum_{k=1}^q \|y_k - Z_1 W_{0k} - b_{0k}\|^2 - \frac{1}{2\tau_z^2} \|Z_1 - f_{\mathbf{B}}(\mathbf{x})\|^2 \right\},$$

which is a multivariate normal distribution with the mean and variance as

$$\mu_Z = \frac{\tau_z^2 \sum_{k=1}^q W_{0k} (y_k - b_{0k}) + \tau_0^2 f_{\mathbf{B}}(\mathbf{x})}{\tau_z^2 \sum_{k=1}^q W_{0k}^2 + \tau_0^2}, \quad \sigma_Z^2 = \frac{\tau_0^2 \tau_z^2}{\tau_z^2 \sum_{j=k}^q W_{0k}^2 + \tau_0^2}.$$

Algorithm 1 Data Augmentation with J -copies for Gaussian Regression (DA-GR).

1: Initialize $\mathbf{B}^{(0)}, W_0^{(0)}, b_0^{(0)}$
2: **for** epoch $t = 1, \dots, T$ **do**
3: Update the weights in the top layer with $\{\mathbf{y}^{(S)}, Z_1^{(t,S)}\}$
 $W_0^{(t)} = \text{Cov}(Z_1^{(t,S)}, \mathbf{y}^{(S)}) / \text{Var}(Z_1^{(t,S)})$
 $b_0^{(t)} = \bar{\mathbf{y}}^{(S)} - W_0^{(t)} \bar{Z}_1^{(t,S)}$
4: Update the deep learner $f_{\mathbf{B}}$ with $\{Z_1^{(t,S)}, \mathbf{x}^{(S)}\}$
 $\mathbf{B}^{(t)} = \mathbf{B}^{(t-1)} - \eta \nabla f_{\mathbf{B}^{(t-1)}}(\mathbf{x}^{(S)} \mid Z_1^{(t,S)})$ ▷ SGD
5: Update $Z_1^{(S)}$ jointly from deep learner $f_{\mathbf{B}}$ and sampling layer f_0
 $Z_1^{j,(t+1)} \mid W_0^{(t)}, b_0^{(t)}, \mathbf{y}, f_{\mathbf{B}^{(t)}}(\mathbf{x}) \stackrel{iid}{\sim} \mathcal{N}(\mu_z^{(t)}, \sigma_z^{(t)2}), j = 1, \dots, J$
6: **return** $\hat{\mathbf{y}} = W_0^{(T)} f_{\mathbf{B}^{(T)}}(\mathbf{x}) + b_0^{(T)}$

4.2 Support Vector Machines (SVMs)

Support vector machines require data augmentation for rectified linear unit (ReLU) activation functions. Polson and Scott (2011) and Mallick et al. (2005) write the support vector machine model as

$$\mathbf{y} = Z_1 W_0 + \boldsymbol{\lambda} + \sqrt{\boldsymbol{\lambda}} \boldsymbol{\epsilon}_0, \text{ where } \boldsymbol{\lambda} \sim p(\boldsymbol{\lambda}),$$

where $p(\boldsymbol{\lambda})$ follows a flat uniform prior. The augmentation variable $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)'$ can be regarded as slacks admitting fuzzy boundaries between classes.

By incorporating the augmentation variable $\boldsymbol{\lambda}$, the ReLU deep learning model can be written as

$$\begin{aligned} y_i &= z_{1,i} W_0 + \lambda_i + \sqrt{\lambda_i} \epsilon_{0,i}, & \text{where } y_i \in \{-1, 1\}, \epsilon_{0,i} &\stackrel{i.i.d.}{\sim} \mathcal{N}(0, \tau_0^2), \\ z_{1,i} &= f_{\mathbf{B}}(\mathbf{x}_i) + \epsilon_{z,i}, & \text{where } \epsilon_{z,i} &\stackrel{i.i.d.}{\sim} \mathcal{N}(0, \tau_z^2). \end{aligned}$$

From a probabilistic perspective, the likelihood function for the output \mathbf{y} is given by

$$\begin{aligned} p(y_i \mid W_0, z_{1,i}) &\propto \exp \left\{ -\frac{2}{\tau_0^2} \max(1 - y_i z_{1,i} W_0, 0) \right\} \\ &\propto \int_0^\infty \frac{1}{\tau_0 \sqrt{2\pi \lambda_i}} \exp \left(-\frac{1}{2\tau_0^2} \frac{(1 + \lambda_i - y_i z_{1,i} W_0)^2}{\lambda_i} \right) d\lambda_i. \end{aligned}$$

Derived from this augmented likelihood function, the conditional updates are

$$\begin{aligned} W_0 \mid \mathbf{y}, Z_1, \boldsymbol{\lambda} &\propto \left[\prod_{i=1}^n \frac{1}{\tau_0 \sqrt{\lambda_i}} \right] \left[\exp \left\{ -\frac{1}{2\tau_0^2} \sum_{i=1}^n \frac{(1 + \lambda_i - y_i z_{1,i} W_0)^2}{\lambda_i} \right\} \right], \\ Z_1 \mid \mathbf{y}, \mathbf{x}, W_0, \mathbf{B} &\propto \exp \left\{ -\frac{1}{2\tau_0^2} \|\mathbf{y} - Z_1 W_0\|_{\Lambda^{-1}}^2 - \frac{1}{2\tau_z^2} \|Z_1 - f_{\mathbf{B}}(\mathbf{x})\|^2 \right\}, \end{aligned}$$

where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ is the diagonal matrix of the augmentation variables.

In order to generate the latent variables, we use conditional Gibbs sampling as

$$\lambda_i^{-1} \mid W_0, y_i, z_{1,i} \sim \mathcal{IG}(|1 - y_i z_{1,i} W_0|^{-1}, \tau_0^{-2}), \quad (4.4)$$

$$W_0 \mid \mathbf{y}, Z_1, \lambda \sim \mathcal{N}(\mu_w, \sigma_w^2), \quad (4.5)$$

$$Z_1 \mid \mathbf{y}, \mathbf{x}, W_0, \mathbf{B} \sim \mathcal{N}(\mu_z, \sigma_z^2) \quad (4.6)$$

with the means and variances given by

$$\begin{aligned} \mu_w &= \frac{\sum_{i=1}^n y_i z_{1,i} \frac{1+\lambda_i}{\lambda_i}}{\tau_0^2 \sum_{i=1}^n \frac{y_i^2 z_{1,i}^2}{\lambda_i}}, & \sigma_w^2 &= \frac{1}{\tau_0^2 \sum_{i=1}^n \frac{y_i^2 z_{1,i}^2}{\lambda_i}}, \\ \mu_z &= \frac{W_0 \tau_z^2 \mathbf{y} + \tau_0^2 f_{\mathbf{B}}(\mathbf{x}) \Lambda \mathbf{1}}{W_0 \tau_z^2 + \tau_0^2 \Lambda \mathbf{1}}, & \sigma_z^2 &= \frac{\tau_0^2 \tau_z^2 \Lambda \mathbf{1}}{W_0^2 \tau_z^2 + \tau_0^2 \Lambda \mathbf{1}}, \end{aligned}$$

where \mathcal{IG} denotes the Inverse Gaussian distribution and $\mathbf{1} = (1, \dots, 1)'$ is a n-dimensional unit vector.

The J -copies strategy can also be adopted here. Z_1^j and λ^j needs to be sampled independently for $j = 1, \dots, J$. Algorithm 2 summarizes the updating scheme with J -copies for SVMs.

Algorithm 2 Data Augmentation with J -copies for SVM (DA-SVM).

- 1: Initialize $\mathbf{B}^{(0)}, W_0^{(0)}, \lambda^{(0)}$
- 2: **for** epoch $t = 1, \dots, T$ **do**
- 3: Update the weights and slack variables in the top layer with $\{\mathbf{y}^{(S)}, Z_1^{(t,S)}\}$
 $\{\lambda^{(t,S)}\}^{-1} \mid W_0^{(t-1)}, \mathbf{y}^{(S)}, Z_1^{(t,S)} \sim \mathcal{IG}(|1 - \mathbf{y}^{(S)} Z_1^{(t,S)} W_0^{(t-1)}|^{-1}, \frac{1}{\tau_0^2})$
 $W_0^{(t)} \mid \mathbf{y}^{(S)}, Z_1^{(t,S)}, \lambda^{(t,S)} \sim \mathcal{N}(\mu_w^{(t)}, \sigma_w^{(t)2})$
- 4: Update the deep learner $f_{\mathbf{B}}$ with $\{Z_1^{(t,S)}, \mathbf{x}^{(S)}\}$
 $\mathbf{B}^{(t)} = \mathbf{B}^{(t-1)} - \eta \nabla f_{\mathbf{B}^{(t-1)}}(\mathbf{x}^{(S)} \mid Z_1^{(t,S)}) \quad \triangleright \text{SGD}$
- 5: Update $Z_1^{(S)}$ jointly from the deep learner $f_{\mathbf{B}}$ and the sampling layer W_0
 $Z_1^{j,(t+1)} \mid W_0^{(t)}, \lambda^{j,(t)}, \mathbf{y}, f_{\mathbf{B}^{(t)}}(\mathbf{x}) \stackrel{iid}{\sim} \mathcal{N}(\mu_z^{(t)}, \sigma_z^{(t)2}), j = 1, \dots, J$
- 6: **return** $\hat{y} = \begin{cases} 1, & \text{if } W_0^{(T)} f_{\mathbf{B}^{(T)}}(\mathbf{x}) > 0 \\ -1, & \text{otherwise.} \end{cases}$

4.3 Logistic Regression

The aim of this example is to show how EM algorithm can be implemented via a weighted L^2 -norm in deep learning. Adopting the logistic regression model from Polson and Scott (2013), we focus on the penalization of W_0 , with parameter optimization given by

$$\hat{W}_0 = \arg \min_{W_0} \left[\frac{1}{n} \sum_{i=1}^n \log \left(1 + \exp \left(- y_i f_{\mathbf{B}}^{DL}(\mathbf{x}_i) W_0 \right) \right) + \phi(W_0 \mid \tau) \right].$$

The outcomes y_i are coded as ± 1 , and τ is assumed fixed.

For likelihood function ℓ and regularization penalty ϕ , we assume

$$p(y_i | \sigma) \propto \int_0^\infty \frac{\sqrt{\omega_i}}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{\omega_i}{2\sigma^2} (y_i f_{\mathbf{B}}(\mathbf{x}_i) W_0 - \frac{1}{2\omega_i})^2 \right\} p(\omega_i) d\omega_i, \quad (4.7)$$

$$p(W_0 | \tau) = \int_0^\infty \frac{\sqrt{\lambda}}{\sqrt{2\pi}\tau} \exp \left\{ -\frac{\lambda}{2\tau^2} (W_0 - \mu_W - \kappa_W \lambda^{-1})^2 \right\} p(\lambda) d\lambda, \quad (4.8)$$

where μ_W, κ_W are pre-specified terms controlling the prior of the penalty term and λ is endowed with a Pólya distribution prior $P(\lambda)$. Let ω_i^{-1} have a Pólya distribution with $\alpha = 1, \kappa = 1/2$, the following three updates will generate a sequence of estimates that converges to a stationary point of posterior

$$W_0^{(t+1)} = (\tau^{-2} \Lambda^{(t)} + \mathbf{x}_*^T \Omega^{(t)} \mathbf{x}_*)^{-1} \left(\frac{1}{2} \mathbf{x}_*^T \mathbf{1} \right),$$

$$\omega_i^{(t+1)} = \frac{1}{z_i^{(t+1)}} \left(\frac{e^{z_i^{(t+1)}}}{1 + e^{z_i^{(t+1)}}} - \frac{1}{2} \right), \quad \lambda^{(t+1)} = \frac{\kappa_W + \tau^2 \phi'(W_0^{(t)} | \tau)}{W_0^{(t)} - \mu_W},$$

where $z_i^{(t)} = y_i z_{1,i}^T W_0^{(t)} = y_i \text{logit}(\hat{y}_i)$, \mathbf{x}_* is a matrix with rows $x_i^* = y_i z_{1,i}$, $\Omega = \text{diag}(\omega_1, \dots, \omega_n)$ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ are diagonal matrices. \mathbf{x}_* can be written as $\mathbf{x}_* = \text{diag}(\mathbf{y}) Z_1$, $\phi'(\cdot)$ denotes the derivative of standard normal density function.

In the non-penalized case, with $\lambda_i = 0$ for every i , the updates can be simplified as weighted least squares

$$W_0^{(t+1)} = (Z_1^{(t)} \text{diag}(\mathbf{y}) \Omega^{(t)} \text{diag}(\mathbf{y}) Z_1^{(t)})^{-1} \left(\frac{1}{2} \mathbf{y}^T Z_1^{(t)} \right),$$

$$\omega_i^{(t+1)} = \frac{1}{z_i^{(t+1)}} \left(\frac{e^{z_i^{(t+1)}}}{1 + e^{z_i^{(t+1)}}} - \frac{1}{2} \right).$$

We focus on the non-penalized binary classification case and Algorithm 3 summarizes our approach. Further generalizations are available. For example, a ridge-regression penalty, along with the generalized double-pareto prior (Armagan et al., 2013) can be implemented by adding a sample-wise L^2 -regularizer. A multinomial generalization of this model can be found in Polson and Scott (2013).

5 Experiments

We illustrate the performance of our methods on both synthetic and real datasets, compared to the deep ReLU networks without the data augmentation layer. We refer to the latter as DL in our results. We denote the data augmented gaussian regression in Algorithm 1 as DA-GR, the SVM implementation in Algorithm 2 as DA-SVM and the logistic regression in Algorithm 3 as DA-logit. For appropriate comparison, we adopt the same network structures, such as the number of layers, the number of hidden nodes, and regularizations like dropout rates, for DL and our methods. The differences between our methods and DL are that (1) the top layer weights W_0, b_0 of DL are updated via

Algorithm 3 Data Augmentation for Logistic Regression (DA-logit).

```

1: Initialize  $W_0^{(0)}, b_0^{(0)}, \mathbf{B}^{(0)}$ 
2: for epoch  $t = 1, \dots, T$  do
3:   Retrieve the input and output of the top layer
    $Z_1^{(t)} = f_{\mathbf{B}^{(t-1)}}(\mathbf{x})$  ▷ input
    $\mathbf{y}^{(t)} = \text{sigmoid}(W_0^{(t-1)}Z_1^{(t)} + b_0^{(t-1)})$  ▷ output
4:   Calculate the sample-wise weights
    $\mathbf{z}^{(t)} = \mathbf{y} \cdot \text{logit}(\mathbf{y}^{(t)})$  ▷ transformed responses
    $\boldsymbol{\omega}^{(t)} = \frac{1}{\mathbf{z}^{(t)}}(\text{sigmoid}(\mathbf{z}^{(t)}) - \frac{1}{2})$  ▷ weights
5:   Update the entire deep learner  $f_{\boldsymbol{\theta}}$  with  $\{\mathbf{y}, \mathbf{x}\}$ 
    $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} - \eta \nabla f_{\boldsymbol{\theta}^{(t-1)}}(\mathbf{x} \mid \mathbf{y}, \text{sample\_weights} = \boldsymbol{\omega}^{(t)})$  ▷ SGD
6: return  $\hat{\mathbf{y}} = \begin{cases} 1, & \text{if } f_{\boldsymbol{\theta}^{(T)}}(\mathbf{x}) > \frac{1}{2} \\ -1, & \text{otherwise.} \end{cases}$ 

```

SGD optimization, while the weights W_0, b_0 of our methods are updated via MCMC or EM; (2) for binary classification, DA-logit and DL adopt a sigmoid activation function in the top layer to produce a binary output, while DA-SVM uses a linear function in the top layer and the augmented sampling layer transforms the continuous value into a binary output. For all experiments, the datasets are partitioned into 70% training and 30% testing randomly. For the optimization we use a modification of the SGD algorithm, the Adaptive moment estimation (Adam, Kingma and Ba (2015)) algorithm. The Adam algorithm combines the estimate of the stochastic gradient with the earlier estimate of the gradient, and scales this using an estimate of the second moment of the unit-level gradient. We have also explored Root Mean Squared propagation (RMSprop Tieleman and Hinton (2012)) optimizer and we observe similar decreases in regression or classification errors.

To illustrate how the choice of J could affect the speed of convergence, we include different implementations of DA-GR and DA-SVM with $J = 2, 5, 10$. We have explored different sampling noise variance τ_0, τ_z , but the choices, in general, do not affect the results significantly.

5.1 Friedman Data

The benchmark (Friedman, 1991) setup uses a regression of the form

$$y_i = 10 \sin(\pi x_{i1} x_{i2}) + 20(x_{i3} - 0.5)^2 + 10x_{i4} + 5x_{i5} + \epsilon_i, \quad \text{with } \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and only the first 5 covariates are predictive of y_i . We run the experiments with $n = 100, 1000$ and $p = 10, 50, 100, 1000$ to explore the performance in both low dimensional and high dimensional scenarios. We implement both one-layer ($L = 1$) and two-layer ($L = 2$) ReLU networks with 64 hidden units in each layer. For DA-GR model, we let $\tau_0 = 0.1, \tau_z = 1$. The experiments are repeated 50 times with different random seeds.

Figure 2 reports the three quartiles of the out-of-sample squared errors (MSEs). The top row is the performance of the one-layer networks and the bottom row is the performance of the two-layer networks. The two-layer networks perform better and converge faster. For DA-GR, when $J = 5$ or $J = 10$, it converges significantly faster and the prediction errors are also smaller. When $J = 2$, the performance of DA-GR is relatively similar to the deep learning model with only SGD updates. This is due to the fact that DA-GR with J -copies learns the posterior mode which is equivalent to the minimization point of the objective function, and it concentrates on the mode faster when J becomes larger.

The computation costs of DA are higher as shown in Figure 3. This is not entirely unexpected since we introduce sampling steps. When J increases, the computation costs also increase slightly. Given the improvement in convergence speed and prediction errors, our data augmentation strategies are still worthwhile even with some extra computation costs. In addition, for each epoch, we can draw the sample-wise posteriors in parallel and the gap between the computation time can be further mitigated.

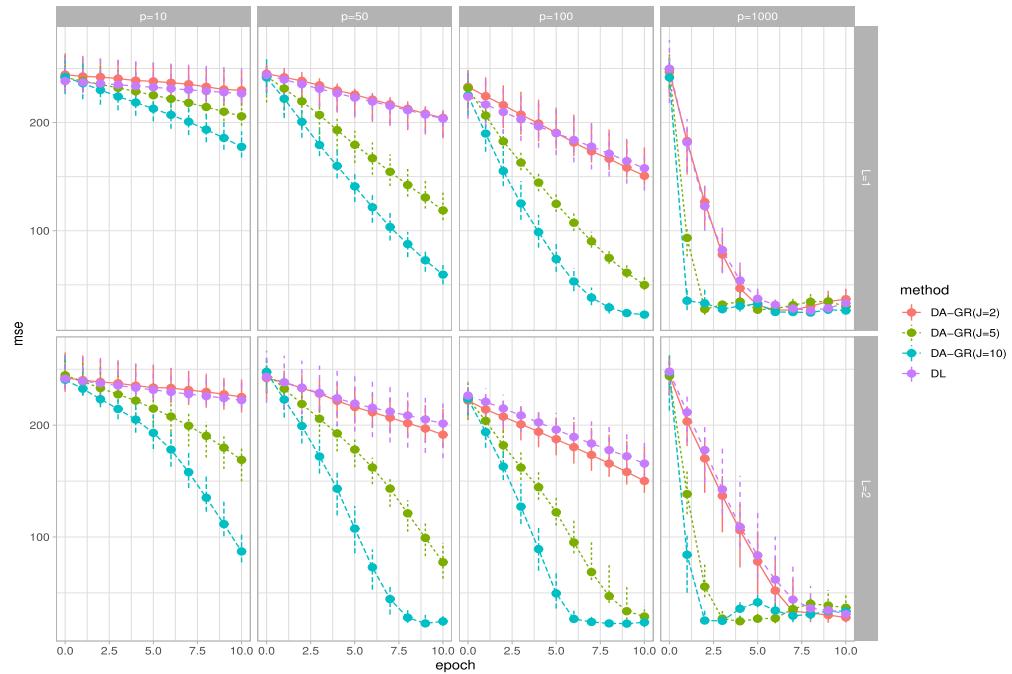


Figure 2: Quartiles of out-of-sample MSEs under the Friedman Setup. We explore cases where $n = 1000$ and $p = 10, 50, 100, 1000$. The tests are repeated 50 times. The medians of out-of-sample MSEs after training for 1 to 10 epochs are plotted with lines and the vertical bars mark the 25 % and 75% quantiles of the MSEs. DA-GR refers to DA Gaussian regression shown in Algorithm 1 and DL stands for the ReLU networks without the data-augmentation layer.

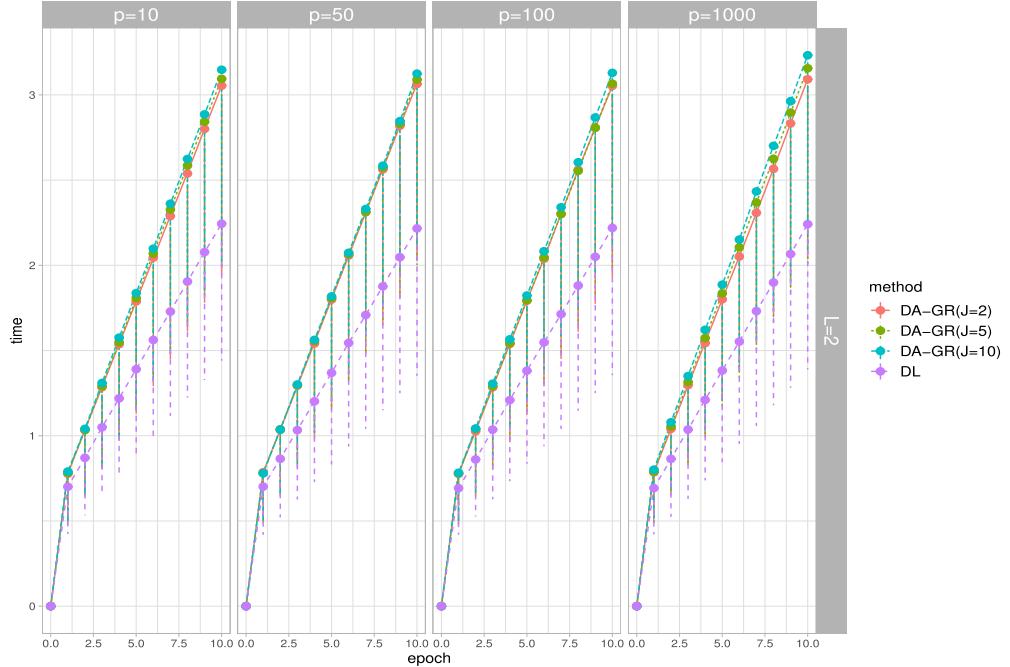


Figure 3: Computation Time under the Friedman setup. The setups are $n = 1000$ and $p = 10, 50, 100, 1000$. The averaged time (over 50 repetitions) for computing 1 to 10 epochs is plotted with lines and the vertical bars mark the 25% and 75% quantiles of the computation time collected. We only include one figure of computation time comparison here since the scale is relatively the same for all cases.

5.2 Boston Housing Data

Another classical regression benchmark dataset is the Boston Housing dataset,¹ see, for example, Hernández-Lobato and Adams (2015). The data contains $n = 506$ observations with 13 features. To show the robustness of DA, we repeat the experiment 20 times with different training subsets. We adopt the ReLU networks with one hidden layer of 64 units and set the dropout rate to be 0.5. For the DA-GR model, we let $\tau_0 = 0.1, \tau_Z = 1$.

Figure 4 shows the prediction errors of all methods. DA-GR with $J = 10$ performs significantly better than the others, in terms of both prediction errors and convergence rates. Meanwhile, DA-GR with $J = 2$ behaves similarly to SGD at the beginning, but it converges significantly faster than SGD after a few epochs. This again, shows that with the J -copies strategy, our method helps the optimization converge at a faster speed, and injecting the noise helps the model generalize well out-of-sample.

¹<https://archive.ics.uci.edu/ml/machine-learning-databases/housing/>

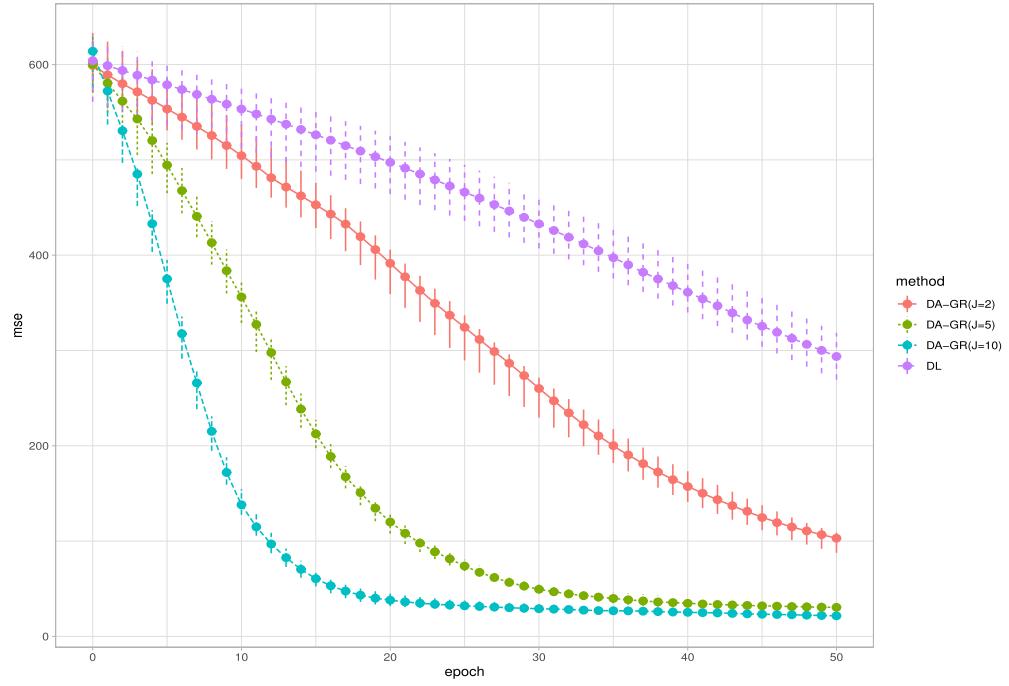


Figure 4: Out-of-sample MSEs for the Boston Housing dataset. The experiment is repeated 20 times with different training subsampling. The medians of MSEs after training for 1 to 50 epochs are provided, with the vertical bars marking the 25% and 75% quantiles of the errors. DA-GR refers to the data augmentation strategy in Algorithm 1 and DL stands for the ReLU networks without the data-augmentation layer.

5.3 Wine Quality Data Set

The Wine Quality Data Set² contains 4898 observations with 11 features. The output wine rating is an integer variable ranging from 0 to 10 (the observed range in the data is from 3 to 9). The frequency of each rating is reported in Table 2.

rating	3	4	5	6	7	8	9
frequency	20	163	1457	2198	880	175	5

Table 2: Frequencies of Different Wine Ratings.

The most frequent ratings are 5 and 6. Since we focus on binary classification problems, we provide two types of classifications, both of which have relatively balanced categories: (1) wine with a rating of 5 or 6 (Test 1); (2) wine with a rating of ≤ 5 or

²P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, ‘Wine Quality Data Set’, UCI Machine Learning Repository.

> 5 (Test 2). We use the same network architectures adopted in Friedman's example with $\tau_0 = \tau_z = 0.1$.

Figure 5 provides results for the two types of binary classifications. In both cases, DA-SVM performs better than DA-logit and DL. The advantage of large J is still significant and helps converge especially in the early phase. DA-logit outperforms DL in Test 1 when the network is shallow ($L = 1$), while in other cases performs similarly to DL.

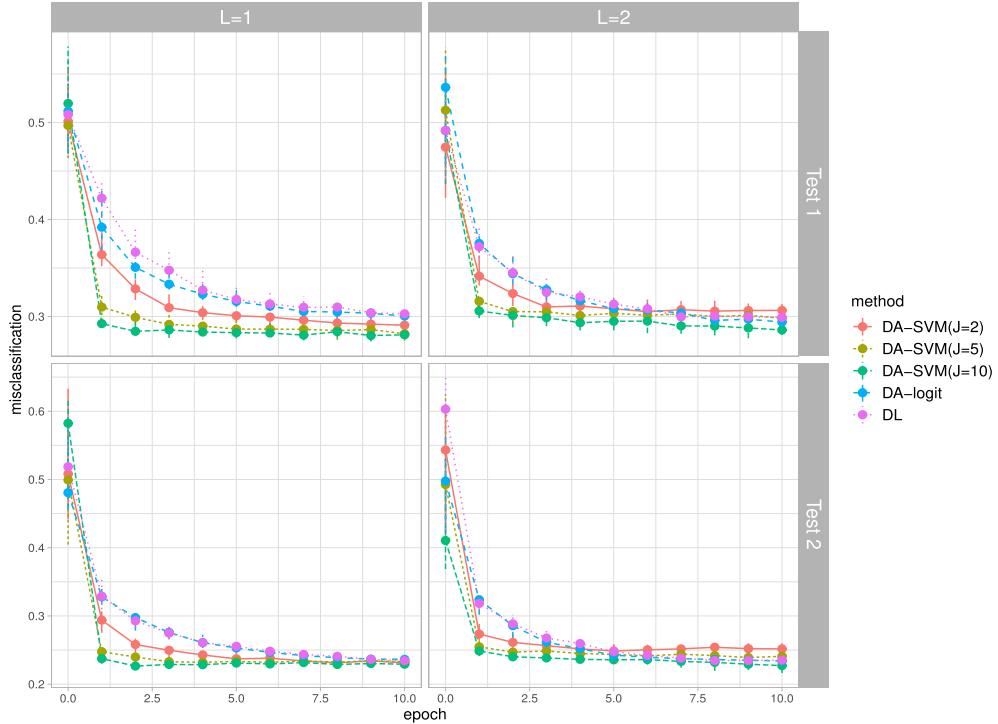


Figure 5: Binary Classifications on the Wine Quality dataset. Two types of binary classifications are considered here. The experiment is repeated 20 times with different training subsampling. We compare the misclassification rates of DA-SVM in Algorithm 2 with $J = 2, 5, 10$, DA-logit in Algorithm 3 and the ReLU networks without the data augmentation layer (DL), after training for 1 to 10 epochs.

5.4 Airbnb Data Set

The Airbnb Kaggle competition³ provides a more challenging application with 21 3451 observations in total, and classified by destination into 12 classes: 10 most popular countries, other and no destination found (NDF), where *other* corresponds to any other country which is not among the top 10 and NDF corresponds to situations that no

³<https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings>

booking was made. The countries are denoted with their standard codes, as ‘AU’ for Australia, ‘CA’ for Canada, ‘DE’ for Germany, ‘ES’ for Spain, ‘FR’ for France, ‘UK’ for United Kingdom, ‘IT’ for Italy, ‘NL’ for Netherlands, ‘PT’ for Portugal, ‘US’ for United States. Table 3 reports the percentage of each class. We follow the preprocessing steps in Polson and Sokolov (2017). The list of variables contains information from the sessions records (number of sessions, summary statistics of action types, device types and session duration), and user tables such as gender, language, affiliate provider etc. All categorical variables are converted to binary dummies, which leads to 661 features in total. For the neural network architecture, we use a two-layer ReLU network with 64 hidden units on each layer and set the dropout rate to be 0.3. For the SVM model, we let $\tau_0 = \tau_z = 0.1$.

	AU	CA	DE	ES	FR	UK	IT	NDF	NL	PT	US	other
% obs	0.25	0.67	0.50	1.05	2.35	1.09	1.33	58.35	0.36	0.10	29.22	4.73

Table 3: Percentage of Each Class (#obs = 213451).

Our goal is to test the binary classification models on this dataset. We consider two types of binary responses, both of which have relatively balanced amounts of observations in each category.

1. Spain (1.05%) vs United Kingdom (1.09%)
2. United Kingdom (1.09%) vs Italy (1.33%)

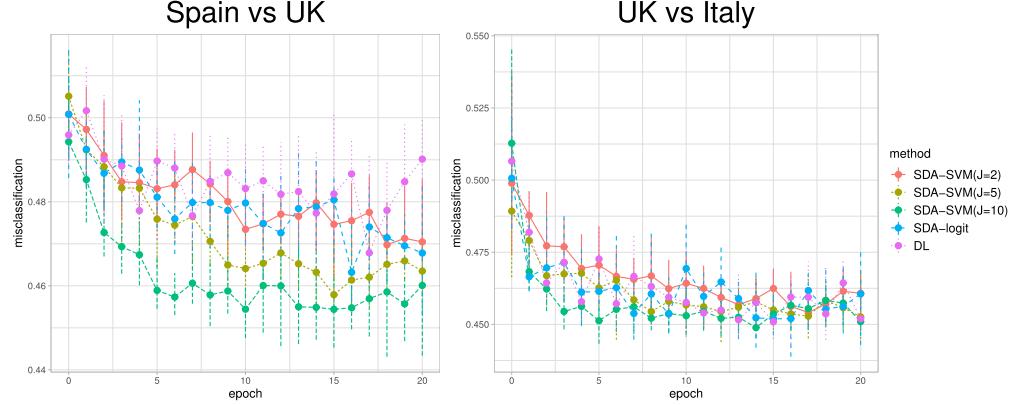


Figure 6: Binary Classifications on the Airbnb Booking Dataset. Two types of binary classifications are considered here. The experiment is repeated 20 times with different training subsampling. We compare the misclassification rates of DA-SVM in Algorithm 2 with $J = 2, 5, 10$, DA-logit in Algorithm 3 and the ReLU networks without the data augmentation layer, after training for 1 to 20 epochs.

Figure 6 demonstrates the binary classifications for Spain versus UK and UK versus Italy. For both cases, the out-of-sample misclassification rates are not small and the fluctuations over epochs are big, suggesting that a better model structure may be needed.

However, we still observe that DA-SVM with $J = 5$ or $J = 10$ has smaller classification errors over epochs and the out-of-sample errors decrease faster during earlier phase of training.

5.5 Summary of Experiment Results

From the above examples, we observe that DA-logit which is implemented under the EM principle does not show an obvious advantage over the vanilla neural network. It shows some improvements on the convergence speed when the network is shallow in the Wine Quality dataset case as in Figure 5. This could be partially due to the fact that we did not apply regularization on the DA layer for our logit implementation. More importantly, the performance of the EM algorithm is contingent on the statistical properties of the objective function. Although the surrogate function is constructed via only the top layer whose quadratic form ensures concavity, the property of the objective function as a whole becomes complicated when the deep network architecture is more complex. Since our method also inherits the negative side of EM and MM algorithms, convergence to the global maximum is not guaranteed in the absence of concavity. However, this observation could open the possibility of future research where we can combine the EM algorithms with shape-constrained neural networks (Gupta et al., 2020).

On the contrary, the MCMC methods with the J-copies strategy significantly improve the prediction errors and convergence speed of the neural networks for both regression and classification problems. And the advantages become more outstanding when J is larger. The phenomenon suggests that the stochastic exploratory methods are preferable when the statistical property of the objective function is unknown or too complex. And the J -copies scheme largely relieves the problem of being trapped into local modes.

One concern of using MCMC methods is the extra computation costs induced by the sampling steps. In our current version where $p_1 = 1$, the sample-wise sampling steps can be computed in parallel. If one wishes to introduce a higher dimension latent variable Z_1 such that $p_1 > 1$, the computation costs will increase as it may involve sampling from multivariate distributions. In that case, fast sampling implementation such as Bhattacharya et al. (2016) is recommended to speed up the process.

6 Discussion

Various regularization methods have been deployed in neural networks to prevent overfitting, such as early stopping, weight decay, dropout (Hinton et al., 2012), gradient noise (Neelakantan et al., 2017). Bayesian strategies tackle the regularization problem by proposing probability structures on the weights. We show that data augmentation strategies are available for many standard activation functions (ReLU, SVM, logit) used in deep learning.

Using MCMC provides a natural stochastic search mechanism that avoids procedures such as back-tracking and provides full descriptions of the objective function over the

entire range Θ . Training deep neural networks thus benefits from additional hidden stochastic augmentation units (a.k.a. data augmentation). Uncertainty can be injected into the network through the probabilistic distributions on only one or two layers, permitting more variability of the network. When more data are observed, the level of uncertainty decreases as more information is learned and the network becomes more deterministic. We also exploit the duality between *maximum a posteriori* estimation and optimization. We provide a J -copies stacking scheme to speed up the convergence to posterior mode and avoid trapping attraction of the local modes. Concerning efficiency, DA provides a natural framework to convert the objective function into weighted least squares and is straightforward to implement with the current deep learning training process.

Our three motivational examples illustrated the advantages of data augmentation. Our work has the potential to be generalized to many other data augmentation schemes and different regularization priors. Probabilistic structures on more units and layers are also possible to allow for more uncertainty.

Our DA-DL methods enjoy the benefits of both worlds. On one hand, with the data augmentation on top, it is robust to random weight initialization. Although we still need to specify the learning rates for the deep architecture, the top layer can learn adaptively and the entire network becomes less sensitive to the choice of learning rate. On the other hand, the fast SGD updates from the deep architecture largely alleviate the computation concerns compared to a fully Bayesian hierarchical model.

There are many directions to future research, including adding more sampling layers so the model could accommodate more randomness and flexibility, and using weighted Bayesian bootstrap (Newton et al., 2021) to approximate the unweighted posteriors by assigning random weight to each observation and penalty. Uncertainty quantification for prediction is also possible. Although we focus on the training aspect of deep learning, one can collect posterior draws $\theta^{(t)}$ from the MCMC procedure when the training process converges. Using (3.2), we can construct predictive intervals and conduct inference.

References

Ahn, S., Balan, A. K., and Welling, M. (2012). “Bayesian posterior sampling via stochastic gradient fisher scoring.” In *Proceedings of the 29th International Conference on Machine Learning*, 1591–1598. [12](#)

Armagan, A., Dunson, D. B., and Lee, J. (2013). “Generalized double Pareto shrinkage.” *Statistica Sinica*, 23(1): 119. [MR3076161](#). [16](#)

Bauer, B. and Kohler, M. (2019). “On deep learning as a remedy for the curse of dimensionality in nonparametric regression.” *The Annals of Statistics*, 47(4): 2261–2285. [MR3953451](#). doi: <https://doi.org/10.1214/18-AOS1747>. [1](#)

Bhadra, A., Datta, J., Polson, N., Sokolov, V., and Xu, J. (2021). “Merging two cultures: deep and statistical learning.” *arXiv preprint arXiv:2110.11561*. [3](#)

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). “Fast sampling with

Gaussian scale mixture priors in high-dimensional regression.” *Biometrika*, asw042. [MR3620452](#). doi: <https://doi.org/10.1093/biomet/asw042>. 23

Chen, T., Fox, E., and Guestrin, C. (2014). “Stochastic gradient Hamiltonian Monte Carlo.” In *International Conference on Machine Learning*, 1683–1691. 3, 11

Deng, W., Zhang, X., Liang, F., and Lin, G. (2019). “An adaptive empirical Bayesian method for sparse deep learning.” In *Advances in Neural Information Processing Systems*, 5563–5573. 2

Duan, L. L., Johndrow, J. E., and Dunson, D. B. (2018). “Scaling up data augmentation MCMC via calibration.” *Journal of Machine Learning Research*, 19(1): 2575–2608. [MR3899766](#). 3

Fan, J., Ma, C., and Zhong, Y. (2021). “A selective overview of deep learning.” *Statistical Science*, 36(2): 264–290. [MR4255194](#). doi: <https://doi.org/10.1214/20-sts783>. 1

Friedman, J. H. (1991). “Multivariate adaptive regression splines.” *The Annals of Statistics*, 19(1): 1–67. [MR1091842](#). doi: <https://doi.org/10.1214/aos/1176347963>. 17

Gan, Z., Henao, R., Carlson, D., and Carin, L. (2015). “Learning deep sigmoid belief networks with data augmentation.” In *Artificial Intelligence and Statistics*, 268–276. 2, 3

Geman, S. and Hwang, C.-R. (1986). “Diffusions for global optimization.” *SIAM Journal on Control and Optimization*, 24(5): 1031–1043. [MR0854068](#). doi: <https://doi.org/10.1137/0324060>. 12

Geyer, C. J. (1996). “Estimation and optimization of functions.” In *Markov chain Monte Carlo in practice*, 241–258. Chapman and Hall. 8

Gramacy, R. B. and Lee, H. K. H. (2008). “Bayesian treed Gaussian process models with an application to computer modeling.” *Journal of the American Statistical Association*, 103(483): 1119–1130. [MR2528830](#). doi: <https://doi.org/10.1198/016214508000000689>. 3

Green, P. J. (1984). “Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives.” *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2): 149–170. [MR0781879](#). 2

Gupta, M., Louidor, E., Mangylov, O., Morioka, N., Narayan, T., and Zhao, S. (2020). “Multidimensional shape constraints.” In *International Conference on Machine Learning*, 3918–3928. PMLR. 23

He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Deep residual learning for image recognition.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. 1

Hernández-Lobato, J. M. and Adams, R. (2015). “Probabilistic backpropagation for scalable learning of Bayesian neural networks.” In *International Conference on Machine Learning*, 1861–1869. 2, 19

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). “Computer model calibration using high-dimensional output.” *Journal of the American Statistical Association*, 103(482): 570–583. [MR2523994](#). doi: <https://doi.org/10.1198/016214507000000888>. 3

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). “Improving neural networks by preventing co-adaptation of feature detectors.” *arXiv preprint arXiv:1207.0580*. 23

Hunter, D. R. and Lange, K. (2004). “A tutorial on MM algorithms.” *The American Statistician*, 58(1): 30–37. [MR2055509](#). doi: <https://doi.org/10.1198/0003130042836>. 8

Jacquier, E., Johannes, M., and Polson, N. (2007). “MCMC maximum likelihood for latent state models.” *Journal of Econometrics*, 137(2): 615–640. [MR2354958](#). doi: <https://doi.org/10.1016/j.jeconom.2005.11.017>. 2, 10

Kingma, D. P. and Ba, J. (2015). “Adam: A method for stochastic optimization.” 17

Kolmogorov, A. N. (1957). “On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition.” In *Doklady Akademii Nauk*, volume 114, 953–956. Russian Academy of Sciences. [MR0111809](#). 3

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). “Imagenet classification with deep convolutional neural networks.” *Advances in Neural Information Processing Systems*, 25: 1097–1105. 1

Lange, K. (2013a). “The MM algorithm.” In *Optimization*, 185–219. Springer. 8

Lange, K. (2013b). *Optimization*, volume 95. Springer Science & Business Media. 8

Lange, K., Hunter, D. R., and Yang, I. (2000). “Optimization transfer using surrogate objective functions.” *Journal of Computational and Graphical Statistics*, 9(1): 1–20. [MR1819865](#). doi: <https://doi.org/10.2307/1390605>. 8

Liang, S. and Srikant, R. (2017). “Why deep neural networks for function approximation?” In *International Conference on Learning Representations*. [MR3796894](#). doi: <https://doi.org/10.1109/tnnls.2017.2665555>. 3

Ma, Y.-A., Chen, Y., Jin, C., Flammarion, N., and Jordan, M. I. (2019). “Sampling can be faster than optimization.” *Proceedings of the National Academy of Sciences*, 116(42): 20881–20885. [MR4025861](#). doi: <https://doi.org/10.1073/pnas.1820003116>. 3

Mallick, B. K., Ghosh, D., and Ghosh, M. (2005). “Bayesian classification of tumours by using gene expression data.” *Journal of the Royal Statistical Society: Series B*, 67(2): 219–234. [MR2137322](#). doi: <https://doi.org/10.1111/j.1467-9868.2005.00498.x>. 2, 14

Mandt, S., Hoffman, M. D., and Blei, D. M. (2017). “Stochastic gradient descent as approximate Bayesian inference.” *Journal of Machine Learning Research*, 18(1): 4873–4907. [MR3763768](#). 12

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). “Equation of state calculations by fast computing machines.” *Journal of Chemical Physics*, 21(6): 1087–1092. [9](#)

Mhaskar, H., Liao, Q., and Poggio, T. A. (2017). “When and why are deep networks better than shallow ones?” In *Proceedings of the 31th Conference on Artificial Intelligence*, 2343–2349. [2](#)

Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). “On the number of linear regions of deep neural networks.” In *Advances in Neural Information Processing Systems*, 2924–2932. [3](#)

Murray, I., Adams, R., and MacKay, D. (2010). “Elliptical slice sampling.” In *Proceedings of the thirteenth International Conference on Artificial Intelligence and Statistics*, 541–548. [12](#)

Neal, R. M. (2003). “Slice sampling.” *The Annals of Statistics*, 705–741. [MR1994729](#).
doi: <https://doi.org/10.1214/aos/1056562461>. [12](#)

Neal, R. M. (2011). “MCMC using Hamiltonian dynamics.” *Handbook of Markov Chain Monte Carlo*, 2(11): 2. [MR3185067](#). [3, 11](#)

Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., and Martens, J. (2017). “Adding gradient noise improves learning for very deep networks.” *International Conference on Learning Representations*. [3, 23](#)

Nesterov, Y. (1983). “A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$.” In *Doklady AN USSR*, volume 269, 543–547. [MR0701288](#). [2](#)

Newton, M. A., Polson, N. G., and Xu, J. (2021). “Weighted Bayesian bootstrap for scalable posterior distributions.” *Canadian Journal of Statistics*, 49(2): 421–437. [MR4267927](#). doi: <https://doi.org/10.1002/cjs.11570>. [24](#)

Phillips, D. B. and Smith, A. F. (1996). “Bayesian model comparison via jump diffusions.” *Markov Chain Monte Carlo in Practice*, 215: 239. [MR1397970](#). [11](#)

Pincus, M. (1968). “A closed form solution of certain programming problems.” *Operations Research*, 16(3): 690–694. [MR0234725](#). doi: <https://doi.org/10.1287/opre.16.3.690>. [3, 11](#)

Pincus, M. (1970). “A Monte Carlo Method for the approximate solution of certain types of constrained optimization problems.” *Operations Research*, 18(6): 1225–1228. [MR0272152](#). doi: <https://doi.org/10.1287/opre.18.6.1225>. [3, 11](#)

Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., and Liao, Q. (2017). “Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review.” *International Journal of Automation and Computing*, 14(5): 503–519. [MR3564936](#). doi: <https://doi.org/10.1142/S0219530516400042>. [1, 2](#)

Polson, N., Sokolov, V., and Xu, J. (2021). “Deep learning partial least squares.” *arXiv preprint arXiv:2106.14085*. [3](#)

Polson, N. G. and Rockova, V. (2018). “Posterior concentration for sparse deep learning.” In *Advances in Neural Information Processing Systems*, 938–949. [1](#), [2](#), [3](#), [4](#)

Polson, N. G. and Scott, J. G. (2013). “Data augmentation for Non-Gaussian regression models using variance-mean mixtures.” *Biometrika*, 100(2): 459–471. [MR3068446](#). doi: <https://doi.org/10.1093/biomet/ass081>. [15](#), [16](#)

Polson, N. G., Scott, J. G., and Windle, J. (2013). “Bayesian inference for logistic models using Pólya–Gamma latent variables.” *Journal of the American Statistical Association*, 108(504): 1339–1349. [MR3174712](#). doi: <https://doi.org/10.1080/01621459.2013.829001>. [3](#), [7](#)

Polson, N. G. and Scott, S. L. (2011). “Data augmentation for Support Vector Machines.” *Bayesian Analysis*, 6(1): 1–23. [MR2781803](#). doi: <https://doi.org/10.1214/11-BA601>. [5](#), [14](#)

Polson, N. G. and Sokolov, V. (2017). “Deep learning: a Bayesian perspective.” *Bayesian Analysis*, 12(4): 1275–1304. [MR3724986](#). doi: <https://doi.org/10.1214/17-BA1082>. [4](#), [22](#)

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). “Stochastic backpropagation and approximate inference in deep generative models.” In *International Conference on Machine Learning*, 1278–1286. [2](#)

Roberts, G. O. and Rosenthal, J. S. (1998). “Optimal scaling of discrete approximations to Langevin diffusions.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1): 255–268. [MR1625691](#). doi: <https://doi.org/10.1111/1467-9868.00123>. [11](#)

Schmidt-Hieber, J. (2020). “Nonparametric regression using deep neural networks with ReLU activation function.” *The Annals of Statistics*, 48(4): 1875–1897. [MR4134774](#). doi: <https://doi.org/10.1214/19-AOS1875>. [1](#), [3](#)

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. (2016). “Mastering the game of Go with deep neural networks and tree search.” *Nature*, 529(7587): 484–489. [1](#)

Telgarsky, M. (2017). “Neural networks and rational functions.” In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 3387–3393. JMLR.org. [3](#)

Tieleman, T. and Hinton, G. (2012). “Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning.” *COURSERA Neural Networks Mach. Learn.* [17](#)

Tran, N., M-N sand Nguyen, Nott, D., and Kohn, R. (2020). “Bayesian deep net GLM and GLMM.” *Journal of Computational and Graphical Statistics*, 29(1): 97–113. [MR4085867](#). doi: <https://doi.org/10.1080/10618600.2019.1637747>. [3](#)

Vitushkin, A. (1964). “Proof of the existence of analytic functions of several complex variables which are not representable by linear superpositions of continuously differ-

entiable functions of fewer variables.” *Soviet Mathematics*, 5: 793–796. [MR0165138](#). [3](#)

Wager, S., Wang, S., and Liang, P. S. (2013). “Dropout training as adaptive regularization.” In *Advances in Neural Information Processing Systems*, 351–359. [1](#), [4](#)

Wang, Y. and Rockova, V. (2020). “Uncertainty quantification for sparse deep learning.” In *Artificial Intelligence and Statistics*. [1](#)

Welling, M. and Teh, Y. W. (2011). “Bayesian learning via stochastic gradient Langevin dynamics.” In *Proceedings of the 28th International Conference on Machine Learning*, 681–688. [12](#)

Yarotsky, D. (2017). “Error bounds for approximations with deep ReLU networks.” *Neural Networks*, 94: 103–114. [3](#)

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012). “Beta-negative binomial process and Poisson factor analysis.” In *Artificial Intelligence and Statistics*, 1462–1471. [2](#)