5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–23

Variational Bayesian Last Layers

Anonymous Authors
Anonymous Institution

1. Introduction

Calibrated uncertainty quantification is essential for safe decision-making with learned models. However, many methods for improving uncertainty quantification in deep learning (including Bayesian methods) have seen limited application due to their complexity over standard deep learning. In this work, we investigate variational learning of Bayesian last layer (BLL) neural networks. In contrast to previous Bayesian deep learning formulations, BLL models consider only the uncertainty over the output layer of the network. While this is relatively restrictive in representational ability, it enables desirable features such as exact inference under certain distributional assumptions and exact likelihood computation.

To train these BLL models, we derive deterministic lower bounds on the marginal likelihood. In contrast to standard variational training strategies, our bounds can be computed without sampling and may be trained with standard mini-batch training strategies. Moreover, we show that all resulting training loss terms may be computed with complexity no greater than standard neural network training complexity. Thus, we show that our variational Bayesian last layer model yields some approximate Bayesian inference effectively for free.

2. Bayesian Last Layer Neural Networks

We first introduce Bayesian last layer models which maintain a non-point posterior only for the last layer in a neural network. These models correspond to Bayesian (linear or logistic) regression or Bayesian Gaussian discriminant analysis (for each of the three models we present, respectively) with learned features.

2.1. Regression

The canonical BLL model for the regression case[1] is

$$\boldsymbol{y}_t = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_t, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}_t \tag{1}$$

where $\boldsymbol{\phi} : \mathbb{R}^{N_x} \times \Theta \to \mathbb{R}^{N_\phi}$ is a vector of neural network features; the input $\boldsymbol{\theta} \in \Theta$ is the weights of the neural network[2]. We will typically write $\boldsymbol{\phi}_t := \boldsymbol{\phi}(\boldsymbol{x}_t, \boldsymbol{\theta})$ for notational convenience. The subscript $t \in \mathbb{N}$ is used to index the data. The term $\boldsymbol{\varepsilon}_t$ is assumed to be

1. We present a scalar output version of the regression case, and defer the multivariate output case to the appendix.
2. We discuss only scalar-output regression in the body of the paper; we defer details on the multivariate case to the appendix.

© A. Authors.

normally distributed with zero mean and covariance Σ, and these noise terms are i.i.d. across realizations.

We specify a Gaussian prior $p(\boldsymbol{w}) = \mathcal{N}(\bar{\boldsymbol{w}}_0, S_0)$, assumed independent of the noise $\boldsymbol{\varepsilon}_t$. Posterior inference in the BLL model is analytically tractable for a fixed set of features. The marginal likelihood may be computed either via direct computation or by iterating over the dataset. Fixing a distribution over $\boldsymbol{w}$ of the form $\mathcal{N}(\bar{\boldsymbol{w}}, S)$ (and writing $\boldsymbol{\eta} = (\bar{\boldsymbol{w}}, S)$) the posterior predictive is

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}) = \mathcal{N}(\bar{\boldsymbol{w}}^\top \boldsymbol{\phi}_t, \boldsymbol{\phi}_t^\top S \boldsymbol{\phi}_t + \Sigma). \tag{2}$$

2.2. Discriminative Classification

In this subsection we introduce a BLL model that closely matches standard classification neural networks. We assume N_y classes, and we aim to predict the correct class given an input $\boldsymbol{x}_t$. Our model takes the form

$$\boldsymbol{z}_t = W \boldsymbol{\phi}(\boldsymbol{x}_t, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}_t \tag{3}$$

$$p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, W, \boldsymbol{\theta}) = \mathrm{softmax}(\boldsymbol{z}_t) \tag{4}$$

where $\boldsymbol{z}_t \in \mathbb{R}^{N_y}$ are the logits. These are also interpreted as unnormalized joint data-label log likelihoods (Grathwohl et al., 2020), where

$$\boldsymbol{z}_t = \log p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid W, \boldsymbol{\theta}) - Z(W, \boldsymbol{\theta}) \tag{5}$$

where $Z(W, \boldsymbol{\theta})$ is a normalizing constant, independent of the data. The term $\boldsymbol{\varepsilon}_t \in \mathbb{R}^{N_y}$ is a zero-mean Gaussian noise term with variance Σ. Typically in logistic regression this noise term is ignored, although it has seen use to model label noise (Collier et al., 2021). We include it to unify the presentation, and the variance can be assumed zero as necessary.

As in the regression case, we specify a Gaussian prior for W. In contrast with the regression setting, exact inference and computation of the posterior predictive is not analytically tractable in this model. We refer to this model—consisting of multinominal Bayesian logistic regression on learned neural network features—as discriminative classification, due to logistic regression being a classical discriminative learning algorithm.

2.3. Generative Classification

The second classification model we consider is the *generative classification* model, so-called due to its similarity to classical generative models such as Gaussian discriminant analysis. We again assume deterministic input features parameterized by a neural network, which we write as $\boldsymbol{\phi}_t$. We will assume that in this feature space, the features associated with each class are normally distributed. Placing a normal prior on the means of these feature distributions and a (conjugate) Dirichlet prior on class probabilities, we have a model of the form

$$\boldsymbol{\rho} \sim \mathrm{Dir}(\boldsymbol{\alpha}_0) \qquad \bar{\boldsymbol{\phi}}[\boldsymbol{y}_t] \sim \mathcal{N}(\boldsymbol{\mu}_0[\boldsymbol{y}_t], S_0[\boldsymbol{y}_t]) \tag{6}$$

$$\boldsymbol{y}_t \sim \mathrm{Cat}(\boldsymbol{\rho}) \qquad \boldsymbol{\phi}_t \mid \boldsymbol{y}_t \sim \mathcal{N}(\bar{\boldsymbol{\phi}}[\boldsymbol{y}_t], \Sigma). \tag{7}$$

In this model, $\boldsymbol{\mu}_0[\boldsymbol{y}_t] \in \mathbb{R}^{N_\phi}$ and $S_0[\boldsymbol{y}_t] \in \mathbb{R}^{N_\phi \times N_\phi}$ are the prior mean and covariance over $\bar{\boldsymbol{\phi}}[\boldsymbol{y}_t] \in \mathbb{R}^{N_\phi}$, the mean embedding for each class[3]. We use the square bracket notation to

3. With some abuse of notation, the term $\boldsymbol{y}_t$ may refer both to an integer class index and a one-hot encoding, depending on the context.

index the statistics for each class; thus, we parameterize $\boldsymbol{\mu}_0[k]$ and $S_0[k]$ for $k = 1, \ldots, N_y$. The terms $\boldsymbol{\rho} \in \mathcal{P}_{N_y}$ correspond to class probabilities, and where $\mathcal{P}_{N_y}$ denotes the probability simplex embedded in $\mathbb{R}^{N_y}$. These class probabilities are in turn used in the categorical distribution over the class. We will write $\bar{\boldsymbol{\phi}} := \{\bar{\boldsymbol{\phi}}[1], \ldots, \bar{\boldsymbol{\phi}}[N_y]\}$.

For a distribution over model parameters

$$p(\boldsymbol{\rho}, \bar{\boldsymbol{\phi}} \mid \boldsymbol{\eta}) = \mathrm{Dir}(\boldsymbol{\alpha}) \prod_{k=1}^{N_y} \mathcal{N}(\boldsymbol{\mu}[k], S[k]) \tag{8}$$

for which we write $\boldsymbol{\eta} = \{\boldsymbol{\alpha}, \boldsymbol{\mu}, S\}$, we have

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\mu}[\boldsymbol{y}], C[\boldsymbol{y}]) \tag{9}$$

$$p(\boldsymbol{y} = k \mid \boldsymbol{\eta}) = \frac{\boldsymbol{\alpha}[k]}{\sum_{k'=1}^{N_y} \boldsymbol{\alpha}[k']} \tag{10}$$

via analytical marginalization, and where $C[\boldsymbol{y}] := \Sigma + S[\boldsymbol{y}]$. To compute the predictive over class labels, we apply Bayes' rule, yielding

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\eta}) = \mathrm{softmax}_{\boldsymbol{y}}(\log p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\eta}) + \log p(\boldsymbol{y} \mid \boldsymbol{\eta})). \tag{11}$$

Here,

$$\log p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\eta}) = -\frac{1}{2}((\boldsymbol{\phi} - \boldsymbol{\mu}[\boldsymbol{y}])^\top C[\boldsymbol{y}]^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu}[\boldsymbol{y}]) + \mathrm{logdet} C[\boldsymbol{y}] + c) \tag{12}$$

where c is a constant, shared for all classes, that may be ignored due to the shift-invariance of the softmax. Grouping the log determinant term with the class prior yields a bias term. Instead of a linear transformation of the input features to obtain a class logit, we instead have a quadratic transformation. This formulation is a strict generalization of standard classifier architectures (Harrison, 2021), in which we have quadratic decision regions as opposed to linear ones.

2.4. Inference and Training in BLL Models

BLL models have seen growing popularity in recent years, despite a trend toward increasingly deep models (Snoek et al., 2015; Azizzadenesheli et al., 2018; Harrison et al., 2018; Weber et al., 2018; Riquelme et al., 2018; Harrison et al., 2020; Ober and Rasmussen, 2019; Kristiadi et al., 2020; Thakur et al., 2020; Watson et al., 2020, 2021; Daxberger et al., 2021; Willes et al., 2022; Sharma et al., 2022; Schwöbel et al., 2022; Zhang et al., 2021). Exact marginalization enables computationally efficient treatment of uncertainty, as well as resulting in lower-variance training objectives compared to sampling-based Bayesian models. The standard objective for training in BLL models is the (log) marginal likelihood (Harrison et al., 2018), via gradient descent on

$$\frac{1}{T} \log p(Y \mid X, \boldsymbol{\theta}) \tag{13}$$

where X, Y denote stacked data across t. We include a factor of $1/T$ to enable better comparison with standard, non-Bayesian, training pipelines (typically based on average loss over mini-batches) and across dataset sizes. This training objective can be problematic,

however: gradient computation requires computing the full marginal likelihood, and mini-batches do not yield unbiased gradient estimators as in standard training with an arbitrary loss function. Moreover, due to the flexibility of neural network features, a full marginal likelihood training objective can result in substantial over-concentration of the approximate posterior (Thakur et al., 2020).

3. Deterministic Stochastic Variational Inference for BLL Networks

In this section we will turn to stochastic variational inference (Hoffman et al., 2013) to compute approximate posteriors, using the paramterizations of the BLL models as variational models. In particular, we aim to jointly compute an approximate last layer posterior and optimize network weights by maximizing lower bounds on marginal likelihood. As such, we will avoid distributional assumptions made in the previous section. We write the parameters as $\boldsymbol{\xi}$ and aim to find an approximate posterior $q(\boldsymbol{\xi} \mid \boldsymbol{\eta})$ parameterized by $\boldsymbol{\eta}$. Concretely, throughout this section we will develop bounds of the form

$$\frac{1}{T}\log p(Y \mid X, \boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) - \frac{1}{T}\mathrm{KL}(q(\boldsymbol{\xi} \mid \boldsymbol{\eta}) || p(\boldsymbol{\xi})) \tag{14}$$

where $\mathcal{L}$ is architecture dependent and developed in the remainder of this section.

3.1. Regression

We consider the log marginal likelihood $\log p(Y \mid X, \boldsymbol{\theta})$, with marginalized parameters $\boldsymbol{\xi} = \{\boldsymbol{w}\}$, and have the following lower bound.

Theorem 1 *Let $q(\boldsymbol{\xi} \mid \boldsymbol{\eta}) = \mathcal{N}(\bar{\boldsymbol{w}}, S)$ denote the variational posterior for the BLL model defined in Section 2.1. Then, (14) holds with*

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) = \frac{1}{T}\sum_{t=1}^{T}\left(\log \mathcal{N}(\boldsymbol{y}_t \mid \bar{\boldsymbol{w}}^\top \boldsymbol{\phi}_t, \Sigma) - \frac{1}{2}\boldsymbol{\phi}_t^\top S \boldsymbol{\phi}_t \Sigma^{-1}\right). \tag{15}$$

The proof for this result and all others is available in Appendix C. When $q(\boldsymbol{\xi} \mid \boldsymbol{\eta}) = p(\boldsymbol{\xi} \mid Y, X)$ and distributional assumptions are satisfied, this lower bound is tight (this may be shown by direct substitution). This correspondence between the variational and true posterior for appropriately-chosen variational families is well known—see Knoblauch et al. (2019) for a thorough discussion. We note that a similar objective for regression models was developed in Watson et al. (2021).

3.2. Discriminative Classification

In the discriminative classification case, the parameters are $\boldsymbol{\xi} = \{W\}$. We will assume a diagonal covariance matrix Σ, and write $\sigma_i^2 := \Sigma_{ii}$. We will fix a mean field variational posterior of the form $q(W \mid \boldsymbol{\eta}) = \prod_{k=1}^{N_y} q(\boldsymbol{w}[k] \mid \boldsymbol{\eta}) = \prod_{k=1}^{N_y} q(\bar{\boldsymbol{w}}[k], S[k])$, where $\boldsymbol{w}[k]$ denotes the k'th row of W. This factorization retains dense covariances for each class, but sacrifices cross-class covariances. While we only present this factorized variational posterior, a similar training objective may be derived with a fully dense variational posterior. Under the mean field variational posterior, we have the following bound on the marginal likelihood.

Theorem 2 *Let $q(W \mid \boldsymbol{\eta}) = \prod_{k=1}^{N_y} \mathcal{N}(\bar{\boldsymbol{w}}[k], S[k])$ denote the variational posterior for the discriminative classification model defined in Section 2.2. Then, (14) holds with*

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) = \frac{1}{T} \sum_{t=1}^{T} \left(\boldsymbol{y}_t^\top \bar{W} \boldsymbol{\phi}_t - \mathrm{LSE}_k \left[\bar{\boldsymbol{w}}[k]^\top \boldsymbol{\phi}_t + \frac{1}{2} (\boldsymbol{\phi}_t^\top S[k] \boldsymbol{\phi}_t + \sigma_k^2) \right] \right) \tag{16}$$

Here, $\mathrm{LSE}(\cdot)$ denotes the log-sum-exp function. In contrast to the regression case, this lower bound is a lower bound on the standard ELBO (due to two applications of Jensen's inequality) and the bound is not tight. We have reduced variance (which would be induced by sampling logit values before the softmax in standard SVI (Ovadia et al., 2019)) for bias due to this lower bound.

3.3. Generative Classification

In the generative classification case, the parameters are $\boldsymbol{\xi} = \{\bar{\boldsymbol{\phi}}, \boldsymbol{\rho}\}$. In this setting, the Dirichlet posterior over class probabilities $p(\boldsymbol{\rho} \mid Y)$ can be computed exactly with one pass over the data by simply counting class occurrences. We therefore only consider a variational posterior of the form $q(\boldsymbol{\xi} \mid \boldsymbol{\eta}, Y) = q(\bar{\boldsymbol{\phi}} \mid \boldsymbol{\eta})$ for the class embeddings, where $q(\bar{\boldsymbol{\phi}} \mid \boldsymbol{\eta}) = \prod_{k=1}^{N_y} \mathcal{N}(\boldsymbol{\mu}[k], S[k])$. This yields the following lower bound.

Theorem 3 *Let $q(\bar{\boldsymbol{\phi}} \mid \boldsymbol{\eta}) = \prod_{k=1}^{N_y} \mathcal{N}(\boldsymbol{\mu}[k], S[k])$ denote the variational posterior over class embeddings for the generative classification model defined in Section 2.3. Let $p(\boldsymbol{\rho} \mid Y) = Dir(\boldsymbol{\alpha}_T)$ denote the exact Dirichlet posterior over class probabilities, with $\boldsymbol{\alpha}_T$ denoting the Dirichlet posterior concentration parameters. Then, (14) holds with*

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) = \frac{1}{T} \sum_{t=1}^{T} \big(& \log \mathcal{N}(\boldsymbol{\phi}_t \mid \boldsymbol{\mu}[\boldsymbol{y}_t], \Sigma) - \frac{1}{2} \mathrm{tr}(\Sigma^{-1} S[\boldsymbol{y}_t]) + \log \boldsymbol{\alpha}_T[\boldsymbol{y}_t] \\ & - \mathrm{LSE}_k[\log \mathcal{N}(\boldsymbol{\phi}_t \mid \boldsymbol{\mu}[k], \Sigma + S[k]) + \log \boldsymbol{\alpha}_T[k]] \big) \end{aligned} \tag{17}$$

This training objective is again a lower bound on the ELBO, and is not tight. The first Dirichlet term (in the upper line) vanishes in gradient computation, but the second term inside the log-sum-exp function does not. In the case that the posterior concentration parameters are equal for all classes (as in the case of a balanced dataset), the concentration parameter can be pulled out of the $\mathrm{LSE}(\cdot)$ (due to the equivariance of log-sum-exp under shifts) and can be ignored.

3.4. Training Variational BLL Models

We propose two methods to learn the features in VBLL models (details available in Appendix D). First, we can jointly optimize the last layer variational posterior together with MAP estimation of the features. While one may expect this to result in substantial over-concentration for weak feature priors, in practice we observe that stochastic regularization due to mini-batch optimization prevents overconcentration. Second, we can combine last layer SVI with variational feature learning (Blundell et al., 2015), corresponding to approximate collapsed VI (Teh et al., 2006).

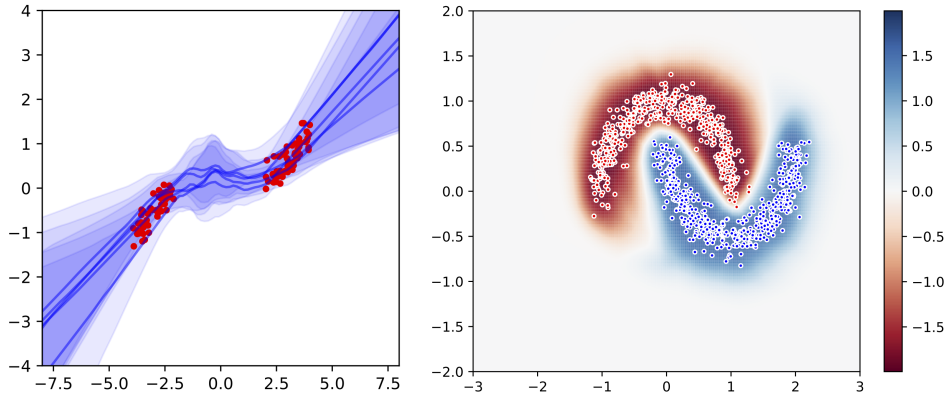

Figure 1: **Left**: a VBLL regression model with BNN features trained on 50 data points generated from a re-scale cubic with Gaussian noise. The plot shows the 95% predictive credible region under the variational posterior for several sampled feature weights. **Right**: Visualizing (re-scaled) $p(\boldsymbol{x} \mid \boldsymbol{y} = 1) - p(x \mid \boldsymbol{y} = 0)$ from the G-VBLL model on the half moon dataset, showing good sensitivity to Euclidean distance and sensible embedding densities.

	Accuracy (↑)	ECE (↓)	NLL (↓)	SVHN AUC (↑)	CIFAR-100 AUC (↑)
DNN	95.8	0.028	0.183	0.946	0.893
SNGP	95.7	0.017	**0.149**	0.960	**0.902**
D-VBLL	**96.2**	0.023	0.167	**0.963**	0.896
G-VBLL	**96.2**	0.022	0.178	0.944	0.816
BNN	95.9	0.032	0.318	0.958	0.869
D-VBLL BNN	**96.2**	0.023	0.167	0.805	0.748
G-VBLL BNN	**96.2**	**0.009**	0.213	0.925	0.796

Table 1: Results for Wide ResNet-28-10 on CIFAR-10.

4. Experiments and Conclusion

In Figure 1, we show predictions on simple datasets. The left figure shows a regression VBLL model with variational features trained on the function $f(x) = cx^3$, with training data shown in red. The right figure shows the generative classification model (G-VBLL) on the half-moon dataset. In particular, we visualize the feature density for each class.

To quantitatively evaluate performance of VBLL models, we train the discriminative (D-VBLL) and generative classification models on the CIFAR-10 image classification task. We baseline these models with a deterministic neural network (DNN), a Bayesian neural network (BNN, Blundell et al. (2015)) and SNGP (Liu et al., 2022). All experiments utilize a Wide ResNet-28-10 backbone architecture. We report accuracy, negative log-likelihood (NLL) and expected calibration error (ECE) for the CIFAR-10 test set and evaluate out-of-distribution (OOD) detection performance using the Street View House Numbers (SVHN) Netzer et al. (2011) and CIFAR-100 test sets as out-of-distribution data.

Table 1 summarizes the CIFAR-10 results. D-VBLL and G-VBLL report strong accuracy performance and competitive metrics for both ECE and NLL. D-VBLL also demonstrates strong OOD detection ability, reporting the highest AUC score for the SVHN dataset.

References

Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through Bayesian deep q-networks. *arXiv:1802.04412*, 2018.

David Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 1947.

David M Blei and John D Lafferty. A correlated topic model of science. *The annals of applied statistics*, 2007.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning (ICML)*, 2015.

George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. Correlated input-dependent label noise in large-scale image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless Bayesian deep learning. *Neural Information Processing Systems (NeurIPS)*, 2021.

Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In *International Conference on Machine Learning (ICML)*, 2020.

Sebastian Farquhar, Michael A Osborne, and Yarin Gal. Radial bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning. In *Artificial Intelligence and Statistics (AISTATS)*, 2020.

Vincent Fortuin. Priors in bayesian deep learning: A review. *International Statistical Review*, 2022.

Vincent Fortuin, Adrià Garriga-Alonso, Sebastian W Ober, Florian Wenzel, Gunnar Rätsch, Richard E Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. *arXiv:2102.06571*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Seymour Geisser. Bayesian estimation in multivariate analysis. *The Annals of Mathematical Statistics*, 1965.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations (ICLR)*, 2020.

James Harrison. *Uncertainty and efficiency in adaptive robot learning and control*. PhD thesis, Stanford University, 2021.

James Harrison, Apoorva Sharma, and Marco Pavone. Meta-learning priors for efficient online Bayesian regression. *Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2018.

James Harrison, Apoorva Sharma, Chelsea Finn, and Marco Pavone. Continuous meta-learning without tasks. *Neural Information Processing Systems (NeurIPS)*, 2020.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv:1610.02136*, 2016.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *Uncertainty in Artificial Intelligence (UAI)*, 2013.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In *International Conference on Machine Learning (ICML)*, 2021.

David Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic Control*, 1973.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*, 2014.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. *arXiv preprint arXiv:1904.02063*, 2019.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being Bayesian, even just a bit, fixes overconfidence in relu networks. In *International Conference on Machine Learning (ICML)*, 2020.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Learnable uncertainty under laplace approximations. In *Uncertainty in Artificial Intelligence (UAI)*, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Neural Information Processing Systems (NeurIPS)*, 2017.

Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When gaussian process meets big data: A review of scalable gps. *IEEE transactions on neural networks and learning systems*, 2020.

Jeremiah Zhe Liu, Shreyas Padhy, Jie Ren, Zi Lin, Yeming Wen, Ghassen Jerfel, Zack Nado, Jasper Snoek, Dustin Tran, and Balaji Lakshminarayanan. A simple approach to improve single-model deep uncertainty via distance-awareness. *arXiv preprint arXiv:2205.00403*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 1992.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 2020.

Eric Thomas Nalisnick. *On priors for Bayesian neural networks*. University of California, Irvine, 2018.

Radford M Neal. *Bayesian learning for neural networks*. Springer, 1995.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Sebastian W Ober and Carl Edward Rasmussen. Benchmarking the neural linear model for regression. *arXiv preprint arXiv:1912.08416*, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Tim Pearce, Russell Tsuchida, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. Expressive priors in Bayesian neural networks: Kernel combinations and periodic functions. In *Uncertainty in Artificial Intelligence (UAI)*, 2020.

C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*. Springer, 1992.

Carl Edward Rasmussen. *Gaussian processes in machine learning*. Springer, 2004.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv:2106.09022*, 2021.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep Bayesian bandits showdown. In *International Conference on Learning Representations (ICLR)*, 2018.

Pola Schwöbel, Martin Jørgensen, Sebastian W Ober, and Mark Van Der Wilk. Last layer marginal likelihood for invariance learning. In *Artificial Intelligence and Statistics (AISTATS)*, 2022.

Mrinank Sharma, Sebastian Farquhar, Eric Nalisnick, and Tom Rainforth. Do Bayesian neural networks need to be fully stochastic? *arXiv preprint arXiv:2211.06291*, 2022.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Neural Information Processing Systems (NeurIPS)*, 2017.

Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Prabhat, and Ryan Adams. Scalable Bayesian optimization using deep neural networks. *International Conference on Machine Learning (ICML)*, 2015.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. *arXiv:1903.05779*, 2019.

Yee Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Neural Information Processing Systems (NeurIPS)*, 2006.

Sujay Thakur, Cooper Lorsung, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. Uncertainty-aware (una) bases for Bayesian regression using multi-headed auxiliary networks. *arXiv preprint arXiv:2006.11695*, 2020.

George C Tiao and Arnold Zellner. On the bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1964.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics (AISTATS)*, 2009.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning (ICML)*, 2020.

Joe Watson, Jihao Andreas Lin, Pascal Klink, and Jan Peters. Neural linear models with functional gaussian process priors. In *Advances in Approximate Bayesian Inference (AABI)*, 2020.

Joe Watson, Jihao Andreas Lin, Pascal Klink, Joni Pajarinen, and Jan Peters. Latent derivative Bayesian last layer networks. In *Artificial Intelligence and Statistics (AISTATS)*, 2021.

Noah Weber, Janez Starc, Arpit Mittal, Roi Blanco, and Lluís Màrquez. Optimizing over a Bayesian last layer. In *NeurIPS workshop on Bayesian Deep Learning*, 2018.

John Willes, James Harrison, Ali Harakeh, Chelsea Finn, Marco Pavone, and Steven Waslander. Bayesian embeddings for few-shot open world recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2022.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Andrew G Wilson, Zhiting Hu, Russ R Salakhutdinov, and Eric P Xing. Stochastic variational deep kernel learning. In *Neural Information Processing Systems (NeurIPS)*, 2016a.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial Intelligence and Statistics (AISTATS)*, 2016b.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Xueting Zhang, Debin Meng, Henry Gouk, and Timothy M Hospedales. Shallow Bayesian meta learning for real-world few-shot recognition. In *International Conference on Computer Vision (ECCV)*, 2021.

Appendix A. Related Work and Discussion

Bayesian methods capable of flexible nonlinear learning have been a topic of active study for the last several decades. Historically, early interest in Bayesian neural networks (MacKay, 1992; Neal, 1995) diminished as Gaussian processes rose to prominence (Rasmussen, 2004). In recent years, however, there has been growing interest in methods capable of learning expressive features, effectively quantifying uncertainty, and training efficiently on large datasets. Variational methods have seen particular attention in both neural networks (Blundell et al., 2015; Ovadia et al., 2019) and GPs (Hensman et al., 2013; Titsias, 2009; Liu et al., 2020) due to their flexibility and their ability to produce mini-batch gradient estimation training schemes.

While a wide range of work has aimed to produce more performant approximate Bayesian methods (including more expressive prior and posterior representations (Fortuin et al., 2021; Izmailov et al., 2021; Sun et al., 2019; Wilson and Izmailov, 2020)), they have still seen limited application, often due to the increased computational expense of these methods (Lakshminarayanan et al., 2017; Dusenberry et al., 2020). While some approaches to Bayesian neural network learning has focused on improved inferential quality, other lines of work have focused on designing comparatively inexpensive approximate Bayesian methods. Indeed, simple strategies such as Bayesian dropout (Gal and Ghahramani, 2016) and stochastic weight averaging (Maddox et al., 2019) have seen much wider use than more expressive methods due to their simplicity.

One of the simplest Bayesian models is the BLL model that is the focus of this paper, which enables single-pass, often deterministic uncertainty prediction. This model has gained prominence through the lens of deep kernel learning (Wilson et al., 2016b,a; Watson et al., 2020; Liu et al., 2022) and within few-shot learning (Harrison et al., 2018, 2020; Harrison, 2021; Watson et al., 2021; Zhang et al., 2021). Deep kernel learning aims to augment standard neural network kernels with neural network inputs. This approach allows control of the behavior of uncertainty, particularly as a function of Euclidean distance (Liu et al., 2022). While stochastic variational inference has been applied to these models (Wilson et al., 2016a), efficient and deterministic mini-batch methods have not been a major focus. Moreover, classification in these models is typically relies on sampling logits before passing

these through softmax functions, increasing variance (Ovadia et al., 2019; Kristiadi et al., 2020, 2021), or on Laplace approximation (Liu et al., 2022).

Within few-shot learning, exact conjugacy of the Bayesian linear regression model (Harrison et al., 2018) and Bayesian GDA (Harrison et al., 2020; Zhang et al., 2021; Snell et al., 2017) has been exploited for efficient few-shot adaptation. These models have (in addition to (Van Amersfoort et al., 2020) among others) shown the strong performance of GDA-based/radial basis function networks, especially on problems such as out of distribution detection, which we further highlight in this work. However, training these models (as well as the DKL methods discussed previously) relies on direct computation of the marginal likelihood. In contrast to prior work on DKL and few-shot learning, our approach achieves efficient and deterministic training and prediction through our variational objectives and through similarly exploiting conjugacy, and thus the added complexity compared to standard neural network models is minimal.

Appendix B. The Multivariate Regression Model

In the multivariate regression case, we consider a model of the form

$$\boldsymbol{y}_t = W\boldsymbol{\phi}_t + \boldsymbol{\varepsilon}_t \tag{18}$$

and place a matrix normal (Tiao and Zellner, 1964; Geisser, 1965) prior on W, with $W \sim \mathcal{MN}(\bar{W}_0, I, S_0)$. For a discussion of the matrix normal distribution, we refer the reader to (Box and Tiao, 2011).

Given the matrix normal prior and the above model, the posterior is also matrix normal. We thus fix a matrix normal variational posterior. In Appendix C.2, we obtain an ELBO of the form

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) = \frac{1}{T}\sum_{t=1}^{T}\left(\log \mathcal{N}(\boldsymbol{y}_t \mid \bar{W}^\top \boldsymbol{\phi}_t, \Sigma) - \frac{1}{2}\boldsymbol{\phi}_t^\top S \boldsymbol{\phi}_t \mathrm{tr}(\Sigma^{-1})\right). \tag{19}$$

for $\boldsymbol{\eta} = \{\bar{W}, S\}$, and we use this as a training objective.

For a parameter distribution $\mathcal{MN}(\bar{W}, I, S)$, prediction in this model is analytically tractable and is

$$p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\eta}, \boldsymbol{\theta}) = \mathcal{N}(\bar{W}\boldsymbol{\phi}_t, \boldsymbol{\phi}_t^\top S \boldsymbol{\phi}_t I + \Sigma). \tag{20}$$

Appendix C. Proofs and Further Theoretical Results

C.1. Helper Results

Our first result builds on results from the variational Gaussian process literature (Titsias, 2009; Hensman et al., 2013).

Lemma 4 *Let $q(\boldsymbol{\mu}) = \mathcal{N}(\bar{\boldsymbol{\mu}}, S)$ and $p(\boldsymbol{y} \mid X, \boldsymbol{\mu}) = \mathcal{N}(X\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{y} \in \mathbb{R}^N, \bar{\boldsymbol{\mu}}, \boldsymbol{\mu} \in \mathbb{R}^M$, $X \in \mathbb{R}^{N\times M}$, and $S, \Sigma \in \mathbb{R}^{M\times M}$. Then*

$$\mathbb{E}_{q(\boldsymbol{\mu})}[\log p(\boldsymbol{y} \mid X, \boldsymbol{\mu})] = \log p(\boldsymbol{y} \mid X, \bar{\boldsymbol{\mu}}) - \frac{1}{2}\mathrm{tr}(\Sigma^{-1} X S X^\top). \tag{21}$$

Proof We have

$$\mathbb{E}_{q(\boldsymbol{\mu})}[\log p(\boldsymbol{y} \mid X\boldsymbol{\mu})] = -\frac{1}{2}\mathbb{E}_{q(\boldsymbol{\mu})}[c + \operatorname{logdet}(2\pi\Sigma) + (\boldsymbol{y} - X\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{y} - X\boldsymbol{\mu})] \tag{22}$$

$$= -\frac{1}{2}\left(c + \operatorname{logdet}(2\pi\Sigma) + \mathbb{E}_{q(\boldsymbol{\mu})}[(\boldsymbol{y} - X\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{y} - X\boldsymbol{\mu})]\right) \tag{23}$$

$$= -\frac{1}{2}\left(c + \operatorname{logdet}(2\pi\Sigma) + (\boldsymbol{y} - X\bar{\boldsymbol{\mu}})^\top \Sigma^{-1}(\boldsymbol{y} - X\bar{\boldsymbol{\mu}}) + \operatorname{tr}(\Sigma^{-1}XSX^\top)\right) \tag{24}$$

where $c \in \mathbb{R}$ is a constant and where the last line follows from the fact that $\boldsymbol{y} - X\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{y} - X\bar{\boldsymbol{\mu}}, XSX^\top)$. The first two terms form the desired log density. ■

Based on this result, we can state a straightforward corollary for generative classification.

Corollary 5 *Let $q(\boldsymbol{\mu}) = \mathcal{N}(\bar{\boldsymbol{\mu}}, S)$ and $p(\boldsymbol{y} \mid \boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{y}, \bar{\boldsymbol{\mu}}, \boldsymbol{\mu} \in \mathbb{R}^N$, $S, \Sigma \in \mathbb{R}^{N\times N}$. Then*

$$\mathbb{E}_{q(\boldsymbol{\mu})}[\log p(\boldsymbol{y} \mid \boldsymbol{\mu})] = \log p(\boldsymbol{y} \mid \bar{\boldsymbol{\mu}}) - \frac{1}{2}\operatorname{tr}(\Sigma^{-1}S). \tag{25}$$

Proof This result follows from Lemma 4 by simply choosing $X = I$. ■

We can also present a variant for multivariate classification.

Corollary 6 *Let $q(W) = \mathcal{MN}(\bar{W}, I, S)$ and $p(\boldsymbol{y} \mid \boldsymbol{x}, W) = \mathcal{N}(W\boldsymbol{x}, \Sigma)$ with $\boldsymbol{y} \in \mathbb{R}^M, \bar{W}, W \in \mathbb{R}^{M\times N}$; $x \in \mathbb{R}^N$; $S \in \mathbb{R}^{N\times N}$; and $\Sigma \in \mathbb{R}^{M\times M}$. Then*

$$\mathbb{E}_{q(W)}[\log p(\boldsymbol{y} \mid \boldsymbol{x}, W)] = \log p(\boldsymbol{y} \mid \boldsymbol{x}, \bar{W}) - \frac{1}{2}\boldsymbol{x}^\top S\boldsymbol{x}\operatorname{tr}(\Sigma^{-1}). \tag{26}$$

Proof Our proof closely follows that of Lemma 4. Expanding the likelihood in the expectation, we have

$$\mathbb{E}_{q(W)}[\log p(\boldsymbol{y} \mid \boldsymbol{x}, W)] = \log p(\boldsymbol{y} \mid \boldsymbol{x}, \bar{W}) - \frac{1}{2}\mathbb{E}_W[\boldsymbol{x}^\top (W - \bar{W})^\top \Sigma^{-1}(W - \bar{W})\boldsymbol{x}] \tag{27}$$

Leveraging the matrix normal identity

$$\mathbb{E}_{W\sim\mathcal{MN}(\bar{W},V,U)}[W^\top AW] = U\operatorname{tr}(A^\top V) + \bar{W}^\top A\bar{W} \tag{28}$$

and the fact that $W - \bar{W} \sim \mathcal{MN}(0, I, S)$, we have

$$\mathbb{E}[(W - \bar{W})^\top \Sigma^{-1}(W - \bar{W})] = S\operatorname{tr}(\Sigma^{-1}) \tag{29}$$

which completes the proof. ■

Lemma 7 *Let $p(\boldsymbol{x} \mid \bar{\boldsymbol{x}}) = \mathcal{N}(\bar{\boldsymbol{x}}, \Sigma)$, and let $\bar{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{\mu}, S)$. Then,*

$$\mathbb{E}_{\bar{\boldsymbol{x}}}[p(\boldsymbol{x} \mid \bar{\boldsymbol{x}})] = \mathcal{N}(\boldsymbol{\mu}, \Sigma + S). \tag{30}$$

Proof We build upon (Jacobson, 1973) and note

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},S)}[\exp(-\frac{1}{2}\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{x})] = \sqrt{\frac{\det(S^{-1})}{\det(S^{-1}+\Sigma^{-1})}}\exp(-\frac{1}{2}\boldsymbol{\mu}^\top S^{-1}(S-(\Sigma^{-1}+S^{-1})^{-1})S^{-1}\boldsymbol{\mu}). \tag{31}$$

Note, by Woodbury's identity

$$S^{-1}(S-(\Sigma^{-1}+S^{-1})^{-1})S^{-1} = (S+\Sigma)^{-1} \tag{32}$$

Let $\boldsymbol{z} := \boldsymbol{x} - \bar{\boldsymbol{x}}$, then $\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{x}-\boldsymbol{\mu}, S)$. We then have

$$\mathbb{E}[p(\boldsymbol{x} \mid \bar{\boldsymbol{x}})] = \mathbb{E}[\exp(-\frac{1}{2}\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|^2_{\Sigma^{-1}} + \frac{1}{2}\text{logdet}(2\pi\Sigma^{-1}))] \tag{33}$$

$$= \mathbb{E}[\exp(-\frac{1}{2}\boldsymbol{z}^\top\Sigma^{-1}\boldsymbol{z})]]\exp(\frac{1}{2}\text{logdet}(2\pi\Sigma^{-1})) \tag{34}$$

For the expectation we apply (31). We simplify the determinant term of (31) as

$$\sqrt{\frac{\det(S^{-1})}{\det(S^{-1}+\Sigma^{-1})}} = \exp(-\frac{1}{2}\text{logdet}(I+S\Sigma^{-1})) \tag{35}$$

Combining, we have

$$\mathbb{E}[\exp(-\frac{1}{2}\boldsymbol{z}^\top\Sigma^{-1}\boldsymbol{z})]] = \exp(-\frac{1}{2}(\|\boldsymbol{x}-\boldsymbol{\mu}\|_{(S+\Sigma)^{-1}} + \text{logdet}(I+S\Sigma^{-1})) \tag{36}$$

We have two log determinant terms, from (34) and the above. We can combine them as

$$\frac{1}{2}\text{logdet}(2\pi\Sigma^{-1}) - \frac{1}{2}\text{logdet}(I+S\Sigma^{-1}) = -\frac{1}{2}(\text{logdet}(\frac{1}{2\pi}\Sigma) + \text{logdet}(I+S\Sigma^{-1})) \tag{37}$$

$$= -\frac{1}{2}\text{logdet}((\frac{1}{2\pi}\Sigma)(I+S\Sigma^{-1})) \tag{38}$$

$$= -\frac{1}{2}\text{logdet}(\frac{1}{2\pi}\Sigma + \frac{1}{2\pi}S) \tag{39}$$

Combining all terms completes the proof. ■

C.2. Proof of Theorem 1

Theorem 1 *Let $q(\boldsymbol{\xi} \mid \boldsymbol{\eta}) = \mathcal{N}(\bar{\boldsymbol{w}}, S)$ denote the variational posterior for the BLL model defined in Section 2.1. Then, (14) holds with*

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) = \frac{1}{T}\sum_{t=1}^{T}\left(\log\mathcal{N}(\boldsymbol{y}_t \mid \bar{\boldsymbol{w}}^\top\boldsymbol{\phi}_t, \Sigma) - \frac{1}{2}\boldsymbol{\phi}_t^\top S\boldsymbol{\phi}_t\Sigma^{-1}\right). \tag{40}$$

Proof First,

$$\log p(Y \mid X, \boldsymbol{\theta}) = \log \mathbb{E}_{p(\boldsymbol{\xi})}[p(Y \mid X, \boldsymbol{\xi}, \boldsymbol{\theta})] \tag{41}$$

$$= \log \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[p(Y \mid X, \boldsymbol{\xi}, \boldsymbol{\theta}) \frac{p(\boldsymbol{\xi}}{q(\boldsymbol{\xi} \mid \boldsymbol{\eta})}] \tag{42}$$

$$\geq \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(Y \mid X, \boldsymbol{\xi}, \boldsymbol{\theta})] - \mathrm{KL}(q(\boldsymbol{\xi} \mid \boldsymbol{\eta})||p(\boldsymbol{\xi})) \tag{43}$$

$$= \sum_{t=1}^{T} \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\xi}, \boldsymbol{\theta})] - \mathrm{KL}(q(\boldsymbol{\xi} \mid \boldsymbol{\eta})||p(\boldsymbol{\xi})). \tag{44}$$

Note that the first term in the last line is the log of a Normal distribution. Applying Lemma 5, we have

$$\mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\xi}, \boldsymbol{\theta})] = \log p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\xi}, \boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\phi}_t^\top S \boldsymbol{\phi}_t \Sigma^{-1} \tag{45}$$

which completes the proof. ■

We can also state the following corollary for the multivariate case.

Corollary 8 *Let $q(\boldsymbol{\xi} \mid \boldsymbol{\eta}) = \mathcal{MN}(\bar{W}, I, S)$ denote the variational posterior for the multivariate BLL model defined in Appendix B. Then, (14) holds with*

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) = \frac{1}{T} \sum_{t=1}^{T} \left(\log \mathcal{N}(\boldsymbol{y}_t \mid \bar{W}\boldsymbol{\phi}_t, \Sigma) - \frac{1}{2}\boldsymbol{\phi}_t^\top S \boldsymbol{\phi}_t \mathrm{tr}(\Sigma^{-1}) \right). \tag{46}$$

Proof The proof follows the proof of Theorem 1, applying Corollary 6 instead of Lemma 5. ■

C.3. Proof of Theorem 2

Theorem 2 *Let $q(W \mid \boldsymbol{\eta}) = \prod_{k=1}^{N_y} \mathcal{N}(\bar{\boldsymbol{w}}[k], S[k])$ denote the variational posterior for the discriminative classification model defined in Section 2.2. Then, (14) holds with*

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) = \frac{1}{T} \sum_{t=1}^{T} \left(\boldsymbol{y}_t^\top \bar{W} \boldsymbol{\phi}_t - \mathrm{LSE}_k \left[\bar{\boldsymbol{w}}[k]^\top \boldsymbol{\phi}_t + \frac{1}{2}(\boldsymbol{\phi}_t^\top S[k] \boldsymbol{\phi}_t + \sigma_k^2) \right] \right) \tag{47}$$

Proof We construct an ELBO via

$$\log p(Y \mid X, \boldsymbol{\theta}) = \log \mathbb{E}_{p(\boldsymbol{\xi})}[p(Y \mid X, \boldsymbol{\theta}, \boldsymbol{\xi})] \tag{48}$$

$$\geq \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(Y \mid X, \boldsymbol{\theta}, \boldsymbol{\xi})] - \mathrm{KL}(q(\boldsymbol{\xi} \mid \boldsymbol{\eta})||p(\boldsymbol{\xi})) \tag{49}$$

$$= \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\sum_{t=1}^{T} \log \mathrm{softmax}_y(\log p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{\theta}, \boldsymbol{\xi}))] - \mathrm{KL}(q(\boldsymbol{\xi} \mid \boldsymbol{\eta})||p(\boldsymbol{\xi})) \tag{50}$$

Expanding the log-softmax term, we have

$$\mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}\left[\sum_t \log \text{softmax}_{\boldsymbol{y}_t}(\log p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{\theta}, \boldsymbol{\xi}))\right] = \tag{51}$$
$$\sum_t \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{\theta}, \boldsymbol{\xi}))] - \sum_t \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\text{LSE}_{\boldsymbol{y}_t}[\log p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{\theta}, \boldsymbol{\xi})].$$

As previously, under the variational posterior these likelihoods factorize across the data. The first term may be directly evaluated, yielding

$$\mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{\theta}, \boldsymbol{\xi}))] = \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[W]\boldsymbol{\phi}_t. \tag{52}$$

The second term (containing the log-sum-exp) can not be computed exactly, and so we will bound this term for both the discriminative and generative classifiers. Via Jensen's inequality, we have

$$-\mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\text{LSE}_{\boldsymbol{y}_t}[\log p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{\theta}, \boldsymbol{\xi})] \geq -\log \sum_{\boldsymbol{y}_t} \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\exp(\log p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{\theta}, \boldsymbol{\xi}))] \tag{53}$$

In the case of the discriminative model, we follow (Blei and Lafferty, 2007) and note that for each row i

$$\mathbb{E}_{\boldsymbol{w}_i \sim \mathcal{N}(\bar{\boldsymbol{w}}_i, S_i)}[\exp(\boldsymbol{w}_i^\top \boldsymbol{\phi}_t + \boldsymbol{\varepsilon}_t[i])] = \exp(\bar{\boldsymbol{w}}_i^\top \boldsymbol{\phi}_t + \frac{1}{2}(\boldsymbol{\phi}_t^\top S_i \boldsymbol{\phi}_t + \sigma_i^2)) \tag{54}$$

which relies on assumed independence of rows of W (although relaxation of this assumption is possible). Combining these results yields a lower bound on the ELBO, which is itself a lower bound on the marginal likelihood. ■

C.4. Proof of Theorem 3

Theorem 3 *Let $q(\bar{\boldsymbol{\phi}} \mid \boldsymbol{\eta}) = \prod_{k=1}^{N_y} \mathcal{N}(\boldsymbol{\mu}[k], S[k])$ denote the variational posterior over class embeddings for the generative classification model defined in Section 2.3. Let $p(\boldsymbol{\rho} \mid Y) = Dir(\boldsymbol{\alpha}_T)$ denote the exact Dirichlet posterior over class probabilities, with $\boldsymbol{\alpha}_T$ denoting the Dirichlet posterior concentration parameters. Then, (14) holds with*

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) = \frac{1}{T}\sum_{t=1}^{T}(\log \mathcal{N}(\boldsymbol{\phi}_t \mid \boldsymbol{\mu}[\boldsymbol{y}_t], \Sigma) - \frac{1}{2}\text{tr}(\Sigma^{-1}S[\boldsymbol{y}_t]) + \log \boldsymbol{\alpha}_T[\boldsymbol{y}_t] \tag{55}$$
$$- \text{LSE}_k[\log \mathcal{N}(\boldsymbol{\phi}_t \mid \boldsymbol{\mu}[k], \Sigma + S[k]) + \log \boldsymbol{\alpha}_T[k]])$$

Proof Note that

$$\log p(Y \mid X, \boldsymbol{\theta}) \geq \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(Y \mid X, \boldsymbol{\xi}, \boldsymbol{\theta})] - \text{KL}(q(\boldsymbol{\xi} \mid \boldsymbol{\eta})||p(\boldsymbol{\xi})) \tag{56}$$

where

$$\mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(Y \mid X, \boldsymbol{\xi}, \boldsymbol{\theta})] = \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(X \mid Y, \boldsymbol{\theta}, \boldsymbol{\xi}) - \log p(X \mid \boldsymbol{\theta}, \boldsymbol{\xi})] + \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(Y \mid \boldsymbol{\xi})] \tag{57}$$

All of these terms factorize over the data, as previously. We first note that for the last term,

$$\log \mathbb{E}_{\boldsymbol{\rho}}[p(\boldsymbol{y}_t \mid \boldsymbol{\theta}, \boldsymbol{\rho})] = \log \boldsymbol{\alpha}[\boldsymbol{y}_t] - \log \sum_{\boldsymbol{y}} \boldsymbol{\alpha}[\boldsymbol{y}]. \tag{58}$$

via standard Dirichlet-Categorical marginalization, and where $\boldsymbol{\alpha}$ correspond to posterior Dirichlet concentration parameters. The first term in (57) is the embedding likelihood; we can compute this expectation of the log likelihood via Corollary 5.

The second term in (57) is less straight-forward. Note that

$$\mathbb{E}[\log p(\boldsymbol{x}_t \mid \boldsymbol{\theta}, \boldsymbol{\xi})] = \mathbb{E}[\log \sum_{\boldsymbol{y}} p(\boldsymbol{x}_t \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi})] \tag{59}$$

which can be written as a log-sum-exp of log joint likelihood. We will again apply Jensen's to exchange the log and sum, and note

$$\begin{aligned}
-\mathbb{E}[\log p(\boldsymbol{x}_t \mid \boldsymbol{\theta}, \boldsymbol{\xi})] &= -\mathbb{E}[\log \sum_{\boldsymbol{y}} p(\boldsymbol{x}_t \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi})] && (60)\\
&\geq -\log \mathbb{E}[\sum_{\boldsymbol{y}} p(\boldsymbol{x}_t \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi})] && (61)\\
&= -\log \sum_{\boldsymbol{y}} \mathbb{E}_{\bar{\phi}[\boldsymbol{y}]}[p(\boldsymbol{x}_t \mid \boldsymbol{y}, \boldsymbol{\theta}, \bar{\phi}[\boldsymbol{y}])] \mathbb{E}_{\boldsymbol{\rho}}[p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\rho})] && (62)\\
&= -\mathrm{LSE}_{\boldsymbol{y}}[\log \mathbb{E}_{\boldsymbol{\rho}}[p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\rho})] + \log \mathbb{E}_{\bar{\phi}[\boldsymbol{y}]}[p(\boldsymbol{x}_t \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\xi})]] && (63)
\end{aligned}$$

where the second line follows from Jensen's, and the third line follows from the structure of the variational posterior. The first term in (63) corresponds to (58). The second term in (58) (the sum over concentration parameters) is equivalent for all classes $\boldsymbol{y}$, and thus can be pulled out of the log-sum-exp (due to the equivariance of this function under shifts) where it cancels the same third term in (57).

To compute the second expectation in (63), we apply Lemma 7. Combining all terms completes the proof. ■

Appendix D. Algorithmic Details

In this section we present concrete details on training VBLL models. We first describe the procedure for MAP estimation and variational learning of features, as described in the paper body. We then discuss prior choice, describe the resultant regularization terms, and describe prediction within these models.

D.1. Feature Point Estimation

We propose to train our models via joint variational inference for the last layer and MAP estimation of network weights (and noise covariance), yielding optimization problem

$$\boldsymbol{\theta}^*, \boldsymbol{\eta}^*, \Sigma^* = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma} \left\{ \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) + \frac{1}{T}(\log p(\boldsymbol{\theta}) + \log p(\Sigma) - \mathrm{KL}(q(\boldsymbol{\xi} \mid \boldsymbol{\eta}) || p(\boldsymbol{\xi}))) \right\}. \tag{64}$$

We will write the three terms on the RHS (scaled by $1/T$) as $\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma)$. Reasonable priors for neural network weights have been discussed in several papers (Blundell et al., 2015; Pearce et al., 2020; Fortuin, 2022; Farquhar et al., 2020; Watson et al., 2020; Dusenberry et al., 2020; Nalisnick, 2018). In this work, we use simple isotropic Gaussian priors which yields a weight decay regularizer. We use an inverse Wishart prior on the noise covariance (or a product of inverse Gamma priors for diagonal covariances). While variational inference for the noise covariance is possible, we choose point estimation to simplify the model. This optimization problem is solved via stochastic gradient descent on mini-batch objectives.

For prediction with VBLL models, we predict directly using the variational posterior, exploiting the conjugate prediction results described in Section 2. For all three VBLL models, training objective computation and prediction can be reduced from cubic to quadratic complexity (in the last layer input width) by careful parameterization and computation. The assumptions required to achieve quadratic complexity for the first two models are minor. However, for the generative classification model, diagonal covariances must be assumed. We discuss complexity in the next section.

D.2. Collapsed Variational Inference for Bayesian Neural Networks

Stochastic variational approximations to the posterior over the network weights have previously been used for Bayesian learning (Blundell et al., 2015). In this section, we aim to compute a variational posterior $q(\boldsymbol{\theta})$, following the same SVI methodology as discussed previously. Whereas our approaches developed in the previous section were deterministic, SVI for all network weights is not possible via deterministic marginalization. Thus, computing

$$\mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})] \tag{65}$$

is typically approximated using Monte Carlo methods. In Blundell et al. (2015), the authors turn to the reparameterization gradient estimator (Kingma and Welling, 2014; Mohamed et al., 2020) to enable the computation of the (Monte Carlo estimator of the) gradient with respect to the parameters of the variational posterior. We could take a similar strategy for both $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$, turning to sampling-based approximation. However, this sampling scheme yields both noisy gradient estimates and is expensive, as each sample corresponds to a full network evaluation. Our approach will instead marginalize the last layer and sample (some of) the other layers. This corresponds to Rao-Blackwellization (Rao, 1992; Blackwell, 1947) of the variational lower bound estimator, yielding lower variance gradient estimates.

We will choose a mean field posterior that factorizes over the parameters and weights, $q(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \boldsymbol{\eta}) = q(\boldsymbol{\xi} \mid \boldsymbol{\eta}_{\boldsymbol{\xi}}) q(\boldsymbol{\theta} \mid \boldsymbol{\eta}_{\boldsymbol{\theta}})$. Note that further mean field approximations for $q(\boldsymbol{\theta} \mid \boldsymbol{\eta})$ are typically employed. For example, (Blundell et al., 2015) factorize the posterior over all weights in the neural network. Given this posterior approximation, we have

$$\log p(Y \mid X) \geq \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\eta}_{\boldsymbol{\theta}}) q(\boldsymbol{\xi}|\boldsymbol{\eta}_{\boldsymbol{\xi}})}[\log p(Y \mid X, \boldsymbol{\xi}, \boldsymbol{\theta})] - \mathrm{KL}(q(\boldsymbol{\xi} \mid \boldsymbol{\eta}_{\boldsymbol{\xi}}) || p(\boldsymbol{\xi})) - \mathrm{KL}(q(\boldsymbol{\theta} \mid \boldsymbol{\eta}_{\boldsymbol{\theta}}) || p(\boldsymbol{\theta})) \tag{66}$$

under the assumption that the prior $p(\boldsymbol{\xi}, \boldsymbol{\theta}) = p(\boldsymbol{\xi}) p(\boldsymbol{\theta})$ and thus

$$\frac{1}{T} \log p(Y \mid X) \geq \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\eta}_{\boldsymbol{\theta}})}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}_{\boldsymbol{\xi}})] - \frac{1}{T} \mathrm{KL}(q(\boldsymbol{\xi} \mid \boldsymbol{\eta}_{\boldsymbol{\xi}}) || p(\boldsymbol{\xi})) - \frac{1}{T} \mathrm{KL}(q(\boldsymbol{\theta} \mid \boldsymbol{\eta}_{\boldsymbol{\theta}}) || p(\boldsymbol{\theta})) \tag{67}$$

for the lower bounds $\mathcal{L}$ developed in Section 3. Thus, algorithmically, we first compute the inner expectation and then approximate the outer expectation with a sampling-based estimator.

D.3. Weight Priors

We place a prior on all trainable terms: the (neural network) feature weights as well as the noise covariance Σ. In the regression and generative classification case, it is possible to perform conjugate inference for the noise covariance in addition to means. Due to the algorithmic complexity of this approach, we instead propose to perform MAP estimation of Σ for these models. We use a standard inverse-Wishart prior for the noise covariance. Ignoring terms that vanish in gradient computation, we have likelihood

$$\log p(\Sigma) = \frac{\nu + N + 1}{2} \mathrm{logdet} \Sigma^{-1} - \frac{1}{2} \mathrm{tr}(M \Sigma^{-1}) \tag{68}$$

where Σ is $N \times N$, $\nu > N - 1$ are the degrees of freedom and M is the scale matrix. The terms ν, M are hyperparameters that are fixed.

For the feature weights $\boldsymbol{\theta}$ we will typically use a simple isotropic Gaussian prior, which corresponds to l_2 regularization. Numerous regularization strategies have been proposed for controlling the kernel generated by neural network features (Watson et al., 2021; Fortuin, 2022), and we view this as necessary follow-up work but beyond the scope of this paper.

D.4. Training

We now present our full training approach for the regression and classification settings. Generally, we will minimize the lower bounds we developed for each model. We note that $\mathcal{L}$ is a sum over data; following (Blundell et al., 2015), we compute an (unbiased) estimator $\hat{\mathcal{L}}$ for this term with mini-batches.

The factorization of the ELBO over the data implies a mini-batch estimator for the gradient. Note that

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\xi}, \boldsymbol{\theta})] = \mathbb{E}_t \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\xi}, \boldsymbol{\theta})] \tag{69}$$

where the outer expectation is with respect to a uniform distribution over t. Note that this also holds for lower bound on the data likelihood, in the case of classification. We can construct a randomized estimator for this expectation based on sub-sampling the data, in our case in mini-batches. For a mini-batch of B datapoints, this yields an estimator for the ELBO of the form

$$\hat{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\eta}, \Sigma) = \frac{1}{B} \sum_{t=1}^{B} \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta})}[\log p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\xi}, \boldsymbol{\theta}, \Sigma)]. \tag{70}$$

Where in the classification case, the first term may be replaced with a lower bound. This re-scaling of the KL divergence was also discussed by Blundell et al. (2015). Note that in the limit of infinite training data ($T \to \infty$) the weight on the KL term goes to zero.

We have trained VBLL models with both momentum SGD and AdamW (Loshchilov and Hutter, 2017). While both work effectively, they result in substantially different uncertainty

representations far from the data, and generally recommend users several optimizers. The interaction of VBLLs with the stochastic regularization associated with different optimizers is an important direction of future work. Practically, gradient clipping was necessary to stabilize late training, especially in the regression case. As the noise variance concentrates, gradient magnitude is highly sensitive to small perturbations to features, which can be rapidly destabilizing; gradient clipping was necessary and sufficient to prevent this destabilization. Beyond these details, training VBLL models did not differ from training normal models.

D.5. Prediction and Monitoring

After training, we are left with learned network weights $\boldsymbol{\theta}^*$, noise covariance Σ^*, and variational posterior parameters $\boldsymbol{\eta}^*$. To make predictions, there are two options. In the case of the regression and generative classification model, we may discard the variational posterior and leverage exact conjugacy. Under (Gaussian) distributional assumptions, exact posteriors may be computed with fixed features.

Exact posteriors may be badly calibrated due to violation of distributional assumptions. Instead, we may make predictions under the variational posterior directly, under the assumption that $q(\boldsymbol{\xi} \mid \boldsymbol{\eta}^*) \approx p(\boldsymbol{\xi} \mid X, Y)$, yielding

$$p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, D) \approx \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta}^*)}[p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, \boldsymbol{\xi}, \boldsymbol{\theta}^*, \Sigma^*)] \tag{71}$$

where $(\boldsymbol{x}_*, \boldsymbol{y}_*)$ denote a test point. For the discriminative classification model, only prediction under the variational posterior is possible, and in this model, sampling or a lower bound may be used. In the regression and generative classification case, conjugacy with respect to the variational posteriors may be exploited to yield a closed-form predictive density (via (2) and (11)).

The generative classification case provides predicted class probability biases (the predicted probability of seeing a particular class before observing a label) through the Dirichlet posterior. In cases where a system designer believes there is likely to exist distributional shift between the training data and the evaluation conditions, predictions may be directly controlled by modifying this Dirichlet posterior. We discuss this in more detail in the experiments.

For the Bayesian feature treatment, prediction can be done by sampling features and computing mixture distributions, yielding

$$p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, D) \approx \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\eta}^*)}\mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta}^*)}[p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, \boldsymbol{\xi}, \boldsymbol{\theta}, \Sigma^*)] \tag{72}$$

$$\approx \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta}^*)}[p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, \boldsymbol{\xi}, \boldsymbol{\theta}_k, \Sigma^*)] \tag{73}$$

for $\boldsymbol{\theta}_k$ sampled i.i.d. from the variational posterior. We suppress dependence on Σ in the above, but in our experiments perform MAP estimation of this term. In the regression case, this averaging is straightforward. For the classification cases, we can average pre-softmax or post-softmax. For example, in the case of generative classification, both

$$p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, D) \approx \frac{1}{K}\sum_{k=1}^{K} \mathrm{softmax}_{\boldsymbol{y}}(\log p(\boldsymbol{y} \mid D) + \log \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta}^*)}[p(\boldsymbol{x}_* \mid \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{\theta}_k)]) \tag{74}$$

and

$$p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, D) \approx \mathrm{softmax}_{\boldsymbol{y}}(\log p(\boldsymbol{y} \mid D) + \log \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{\eta}^*)}[p(\boldsymbol{x}_* \mid \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{\theta}_k)]) \tag{75}$$

are valid Monte Carlo estimators for the predictive density, and the same holds for the discriminative classifier. In practice, we typically use the former (in which we directly average the post-softmax samples) due to the relative implementation simplicity, although the latter is necessary for some forms of out of distribution detection. Note that in the latter estimator,

$$\log \frac{1}{K} \sum_{k} \boldsymbol{x}_k = \mathrm{LSE}_k(\log \boldsymbol{x}_k) - \log K \tag{76}$$

for generic $\boldsymbol{x}_k$ and $\log K$ vanishes in the softmax and my therefore be ignored, and where the use of log-sum-exp improves numerical stability.

D.6. Out of Distribution Detection

A desirable feature of robust deep learning models is the ability to distinguish between in-distribution and out-of-distribution (OoD) data. We use several metrics for OoD detection with VBLL models. For the discriminative VBLL, we follow previous work (Liu et al., 2022) and use the maximum softmax probability (Hendrycks and Gimpel, 2016) for an OoD measure. This is computed by sampling from the distribution over logits and passing these samples through the softmax, where they are averaged.

For the generative classification model, we use the feature density

$$p(\boldsymbol{x}_* \mid D) \approx \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}[\mathrm{LSE}_{\boldsymbol{y}}(\log p(\boldsymbol{x}_* \mid \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{\theta}))] \tag{77}$$

as an OoD measure. In the above, the expectation are with respect to the variational posteriors; for the MAP estimation case, this corresponds to direct evaluation. Inspired by Ren et al. (2019, 2021), we subtract the log density under the feature prior as a normalization strategy, which slightly improves performance.

Appendix E. Parameterization and Complexity

In this section, we discuss how to parameterize each of the terms appearing in each type of VBLL. In each model, we use a "mixed" parameterization—in contrast to the standard parameterization or natural parameterization. More precisely, we will parameterize the inverse noise covariance Σ^{-1} and the covariance of the variational posterior S via Cholesky factorizations, and directly parameterize means $\bar{W}, \boldsymbol{\mu}$. In our (limited) comparisons of the performance of different parameterizations, we found that our mixed parameterization performed equivalently (if slightly better) to the standard parameterization, and both performed better than natural parameterization. Interestingly, this stands in contrast to standard practice in variational Gaussian process learning (Hensman et al., 2013), in which authors frequently aim to derive natural gradient optimization algorithms.

We will show that for each VBLL model, under a set of reasonable assumptions, complexity is at worst quadratic in the last layer width and linear in the output dimension. These complexity results enable use of VBLL models on problems with high input dimensionality and high output dimensionality. Moreover, our mini-batch gradient estimation training

objective results in (standard) linear complexity of gradient estimation in batch size, enabling training on much larger datasets than is possible with standard marginal likelihood objectives.

E.1. Regression

Our analysis will focus on the multivariate case, for which the univariate outputs are a special case. We directly parameterize the mean $\bar{W} \in \mathbb{R}^{N_y \times N_\phi}$. The covariances are parameterized via Cholesky decomposition to guarantee positive semi-definiteness; in particular we parameterize $\Sigma^{-1} = LL^\top$ and $S = PP^\top$. Each of these Cholesky matrices must be lower triangular with positive diagonals, and so they are parameterized in automatic differentiation packages as

$$L = \texttt{tril}(L_d) + \texttt{diag}(\exp(\boldsymbol{m})) \tag{78}$$

for matrix L_d and vector $\boldsymbol{m}$, and where `tril` denotes the triangular-lower function.

Given these parameterization, we show the complexity of each operation required for training is at most quadratic in N_ϕ. The training objective has two terms: the log Gaussian density and the trace term. For the log density, we have

$$\boldsymbol{e}_t^\top \Sigma^{-1} \boldsymbol{e}_t = \boldsymbol{e}_t^\top LL^\top \boldsymbol{e}_t \tag{79}$$

for $\boldsymbol{e}_t = \boldsymbol{y} - \bar{W}\boldsymbol{\phi}_t$. The term $L^\top \boldsymbol{e}_t$ can be computed in $\mathcal{O}(N_y^2)$ time. The second term is $\boldsymbol{\phi}_t^\top S \boldsymbol{\phi}_t \mathrm{tr}(\Sigma^{-1})$, for which $\boldsymbol{\phi}_t^\top S \boldsymbol{\phi}_t$ can be computed in $\mathcal{O}(N_\phi^2)$ time, and the trace term

$$\mathrm{tr}(\Sigma^{-1}) = \mathrm{tr}(LL^\top) = \|L\|_F^2 \tag{80}$$

which can be computed in $\mathcal{O}(N_y^2)$ time via squaring and summing the elements of L.

The remaining terms are the KL penalty on the variational posterior, and the inverse Wishart prior on the noise covariance. Fixing a prior $\mathcal{MN}(\bar{W}_0, I, S_0)$, the KL penalty for the multivariate regression case is (ignoring constants)

$$KL(q(\boldsymbol{\xi} \mid \boldsymbol{\eta}) || p(\boldsymbol{\xi})) = \frac{1}{2}(\mathrm{tr}((\bar{W} - \bar{W}_0)^\top (\bar{W} - \bar{W}_0) S_0^{-1}) + N_y \mathrm{tr}(S_0^{-1} S) + N_y \log \frac{\det S_0}{\det S}) \tag{81}$$

We will fix an isotropic prior, $S_0 = sI$ for $s > 0$. Thus, the first term is

$$\mathrm{tr}((\bar{W} - \bar{W}_0)^\top (\bar{W} - \bar{W}_0) S_0^{-1}) = \frac{1}{s} \|\bar{W} - \bar{W}_0\|_F^2 \tag{82}$$

yielding complexity $\mathcal{O}(N_y N_\phi)$, and the second term is

$$N_y \mathrm{tr}(S_0^{-1} S) = \frac{N_y}{s} \mathrm{tr}(S) \tag{83}$$

where the trace can again be computed as the squared Frobenius norm of the Cholesky factor of P, for complexity $\mathcal{O}(N_\phi^2)$. The last term is

$$N_y \log \frac{\det S_0}{\det S} = N_y N_\phi \log s - N_y \mathrm{logdet} S \tag{84}$$

where $\mathrm{logdet} S = 2\mathrm{logdet}(P)$ which is equal to the sum of the log diagonal elements, which can be computed in $\mathcal{O}(N_\phi)$

Finally, we have the inverse Wishart noise covariance prior, which has terms $\text{logdet}\Sigma^{-1}$ and $\text{tr}(M\Sigma^{-1})$ for scale matrix M. The log determinant term may be computed as previously, with complexity $\mathcal{O}(N_y)$. Choosing scale matrix $M = mI$, we have $\text{tr}(M\Sigma^{-1}) = m\text{tr}(\Sigma^{-1})$ which again is $\mathcal{O}(N_y^2)$. Summing all of this up, we have the total complexity of VBLL computations as $\mathcal{O}(N_y^2 + N_\phi^2)$, which is equivalent to the complexity of standard matrix multiplication; thus, there is effectively zero added computational expense from the VBLL model compared to a standard network. The reader may easily verify that complexity of prediction is no greater than the training complexity in the regression model.

E.2. Classification

The complexity for the discriminative classification model follows from the regression model. We use the same parameterization, although we turn to a diagonal noise covariance Σ. The computation of the KL penalty is identical to the regression case. The only difference is that $\boldsymbol{\phi}_t^\top S[\boldsymbol{y}]\boldsymbol{\phi}_t$ must be computed for all classes $\boldsymbol{y}$, yielding complexity $\mathcal{O}(N_\phi^2 N_y)$. This term dominates the complexity of this model; however, further factorization of the covariance is straightforward and can reduce the practical complexity. To predict in these models, sampling realizations of the last layer must be done to sample logits. This sampling is straightforward to do using the Cholesky factorization of the covariance, and has quadratic complexity.

For the generative classification model, we are limited by the $\Sigma + S[\boldsymbol{y}]$ term in the log-sum-exp. As far as we are aware, there is no (practical) way to compute this term with quadratic complexity, or otherwise inexpensively compute this log density. Thus, in this paper we restrict Σ and S to diagonal matrices, which results in linear complexity in N_ϕ for all operations in loss computation. Thus, under this approximate posterior, the complexity of the full training loss computation is $\mathcal{O}(N_\phi N_y)$, which is equivalent to standard neural network models. This covariance structure is relatively restrictive, and improvements may results from sparse covariance structures.

Appendix F. Experimental Details

All CIFAR-10 experiments utilize the Wide ResNet-28-10 (WRN-28-10) backbone network architecture. Hyper-parameters are similar to those proposed by Zagoruyko and Komodakis (2016). Unlike the original implementation of WRN, we do not employ Nesterov momentum and we fully decay an initial learning rate of 0.1 according to a Cosine Annealing schedule instead of a stepped decay schedule. We train each model for 200 epochs and use the standard WRN data-augmentations.

The BNN backbone-based models are primarily deterministic and only implement a single final Bayesian linear layer with a prior distribution of $\mathcal{N}(0, 0.01)$. Each BNN-based model samples 10 predictions for each test data point. This operation is relatively cheap when compared to a fully stochastic network because the intermediate features are cached prior to the final Bayesian linear layer sampling and computation.

The DNN, BNN, D-VBLL and D-VBLL BNN models use maximum softmax probability as an OOD measure. The G-VBLL and G-VBLL BNN models use a normalized feature density. Refer to Section D.6 for additional details. The CIFAR-10 test set is used as in-distribution data and the CIFAR-100 and SVHN test sets are used as out-of-distribution data for their respective experiments. In-distribution data normalization is used in both cases.