Analytically Tractable Inference in Neural Networks – An Alternative to Backpropagation

Luong Ha Nguyen and James-A. Goulet
Department of Civil, Geologic and Mining Engineering
POLYTECHNIQUE MONTREAL, CANADA

Abstract

Until now, neural networks have been predominantly relying on backpropagation and gradient descent as the inference engine in order to learn a neural network's parameters. This is primarily because closed-form Bayesian inference for neural networks has been considered to be intractable. This short paper outlines a new analytical method for performing *tractable approximate Gaussian inference* (TAGI) in Bayesian neural networks. The method enables the analytical inference of the posterior mean vector and diagonal covariance matrix for weights and biases. One key aspect is that the method matches or exceeds the state-of-the-art performance while having the same computational complexity as current methods relying on the gradient backpropagation, i.e., linear complexity with respect to the number of parameters in the network. In addition to reducing in the number of hyperparameters due to the absence of gradient-based optimization, it enables unprecedented features such as the propagation of uncertainty from the input of a network up to its output, and it allows inferring the value of hidden states, inputs, as well as latent variables. In this paper, we present benchmark results demonstrating the performance of TAGI on deep architectures and showcases some of the new avenues it enables

1 Introduction

Until now, neural networks have been predominantly relying on backpropagation [22] and gradient descent as the inference engine in order to learn a neural network's parameters. This is primarily because closed-form Bayesian inference for neural networks has been considered to be intractable [7]. Several approximate inference methods for Bayesian Neural Networks (BNN) have been proposed, e.g., Laplace approximation [15], Hamiltonian Monte Carlo sampling [17], variational inference [11, 1], and Monte Carlo dropout [6]. All the recent methods [12, 10, 2, 14, 20, 25, 5] that are either based on moment matching, variational approaches, or dropout, share a common aspect; the inference of parameters is still treated as an optimization problem relying on the gradient backpropagation.

This short paper outlines a new analytical method for performing *tractable approximate Gaussian inference* (TAGI) [8] in BNNs. The method enables the analytical inference of the posterior mean vector and diagonal covariance matrix for weights and biases. One key aspect is that the method matches or exceeds the state-of-the-art performance while having the same computational complexity as current methods relying on the gradient backpropagation, i.e., linear complexity with respect to the number of parameters in the network. Performing Bayesian inference in neural networks enables several key features, such as the quantification of epistemic uncertainty associated with model parameters, the online estimation of parameters, and a reduction in the number of hyperparameters due to the absence of gradient-based optimization. Moreover, the analytical framework proposed also enables unprecedented features such as the propagation of uncertainty from the input of a net-

Bayesian Deep Learning workshop, NeurIPS 2021.

work up to its output, and it allows inferring the value of hidden states, inputs, as well as latent variables.

The first part covers the theoretical foundation and working principles of the analytically tractable uncertainty propagation in neural networks, as well as the parameter and hidden state inference. Then, the second part will go through benchmarks demonstrating the superiority of the approach on supervised, unsupervised, and reinforcement learning tasks. In addition, we will showcase how TAGI can be applied to reinforcement learning problems such as the Atari game environment. Finally, the last part will present how we can leverage the analytic inference capabilities of our approach to enable novel applications of neural networks such as closed-form direct adversarial attacks, and the usage of a neural network as a generic black-box optimization method.

2 Tractable Approximate Gaussian Inference (TAGI)

TAGI [8] assumes that the joint distribution between the observations and a neural network's parameters is approximated by a multivariate Gaussian distribution,

$$f\begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{y}\end{pmatrix} = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\theta}\\ \boldsymbol{y}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu_\theta}\\ \boldsymbol{\mu_Y}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma_\theta} & \boldsymbol{\Sigma_{Y\theta}^\intercal}\\ \boldsymbol{\Sigma_{Y\theta}} & \boldsymbol{\Sigma_Y}\end{bmatrix}\right),\ f(\boldsymbol{\theta}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu_{\theta|y}}, \boldsymbol{\Sigma_{\theta|y}}) \begin{cases}\boldsymbol{\mu_{\theta|y}} = \boldsymbol{\mu_\theta} + \boldsymbol{\Sigma_{Y\theta}^\intercal}\boldsymbol{\Sigma_Y^{-1}}(\boldsymbol{y} - \boldsymbol{\mu_Y})\\ \boldsymbol{\Sigma_{\theta|y}} = \boldsymbol{\Sigma_\theta} - \boldsymbol{\Sigma_{Y\theta}^\intercal}\boldsymbol{\Sigma_Y^{-1}}\boldsymbol{\Sigma_{Y\theta}}\end{cases}$$

so that the parameter inference can build upon the Gaussian conditional equations describing the probability density function (PDF) of $\boldsymbol{\theta}$ conditional on observations $\boldsymbol{y}$.

The approach is inherently divided in two steps; first propagate uncertainties through the network in order to obtain the joint PDFs between the quantities to be updated (i.e., neural network's parameters and hidden state units) and the observations, and then update these quantities. The first key operation to be considered is the propagation of uncertainty from the activation units $\boldsymbol{A}^{(j)} \sim \mathcal{N}(\boldsymbol{\mu_A^{(j)}}, \boldsymbol{\Sigma_A^{(j)}})$ of a hidden layer j to a hidden unit $Z_i^{(j+1)}$ on the subsequent layer $j+1$, $Z_i^{(j+1)} = \sum_{k=1}^{\mathtt{A}} W_{i,k}^{(j)} A_k^{(j)} + B_i^{(j)}$, where $W_{i,k}^{(j)}$ are weights and $B_i^{(j)}$ bias parameters that are modelled by Gaussian random variables. In order to maintain the analytical tractability of uncertainty propagation through hidden layers, TAGI approximates the product of any pair of weight and activation unit by a Gaussian random variable $WA \approx \mathcal{N}(\mu_{WA}, \sigma^2_{WA})$, for which the exact moments can be computed analytically using the *Gaussian multiplicative approximation* (GMA) [8]. The second key operation is the propagation of uncertainty through non-linear activation functions $A_i^{(j+1)} = \phi(Z_i^{(j+1)})$, where, in order to maintain the analytical tractability, TAGI locally linearize $\phi(\cdot)$ at the expected value of the hidden units $\mu_{Z_i}^{(j+1)}$. Maintaining the computational tractability of uncertainty propagation through hidden layers requires assuming diagonal covariance structures for hidden units among a same layer $\boldsymbol{\Sigma_Z^{(j)}}$, and for the parameters $\boldsymbol{\Sigma_\theta}$.

The update step, i.e., Gaussian conditional inference, is performed using a recursive layer-wise procedure; Using the short-hand notation $\{\boldsymbol{\theta}^+, \boldsymbol{Z}^+\} \equiv \{\boldsymbol{\theta}^{(j+1)}, \boldsymbol{Z}^{(j+1)}\}$ and $\{\boldsymbol{\theta}, \boldsymbol{Z}\} \equiv \{\boldsymbol{\theta}^{(j)}, \boldsymbol{Z}^{(j)}\}$, the posteriors for the parameters and hidden states are computed following

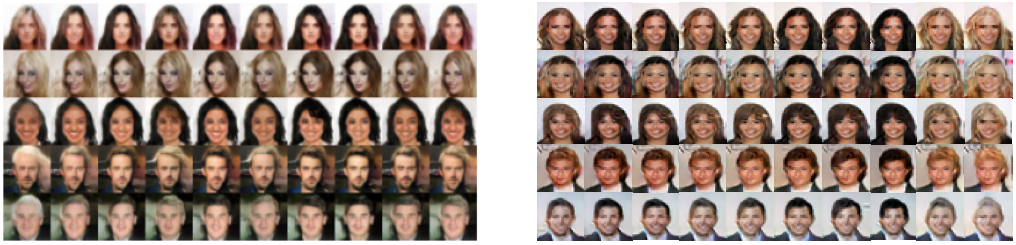
$$\begin{aligned}
f(\boldsymbol{z}|\boldsymbol{y}) &= \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu_{Z|y}}, \boldsymbol{\Sigma_{Z|y}}) & f(\boldsymbol{\theta}|\boldsymbol{y}) &= \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu_{\theta|y}}, \boldsymbol{\Sigma_{\theta|y}})\\
\boldsymbol{\mu_{Z|y}} &= \boldsymbol{\mu_Z} + \mathbf{J_Z}\left(\boldsymbol{\mu_{Z^+|y}} - \boldsymbol{\mu_{Z^+}}\right) & \boldsymbol{\mu_{\theta|y}} &= \boldsymbol{\mu_\theta} + \mathbf{J_\theta}\left(\boldsymbol{\mu_{Z^+|y}} - \boldsymbol{\mu_{Z^+}}\right)\\
\boldsymbol{\Sigma_{Z|y}} &= \boldsymbol{\Sigma_Z} + \mathbf{J_Z}\left(\boldsymbol{\Sigma_{Z^+|y}} - \boldsymbol{\Sigma_{Z^+}}\right)\mathbf{J_Z^\intercal} & \boldsymbol{\Sigma_{\theta|y}} &= \boldsymbol{\Sigma_\theta} + \mathbf{J_\theta}\left(\boldsymbol{\Sigma_{Z^+|y}} - \boldsymbol{\Sigma_{Z^+}}\right)\mathbf{J_\theta^\intercal}\\
\mathbf{J_Z} &= \boldsymbol{\Sigma_{ZZ^+}}\boldsymbol{\Sigma_{Z^+}^{-1}}, & \mathbf{J_\theta} &= \boldsymbol{\Sigma_{\theta Z^+}}\boldsymbol{\Sigma_{Z^+}^{-1}}.
\end{aligned} \tag{1}$$

Note that the layer-wise recursive procedure defined in equations 1 only requires the storage of the joint prior PDFs for pairs of subsequent hidden layers and pairs of hidden layers and the parameters directly connecting into them. This allows maintaining the computational tractability of the uncertainty propagation and inference steps, which scale linearly with respect to the number of weight parameters. As we will see in the next section, the applicability of TAGI extends beyond feedforward neural network (FNN) to convolutional (CNN) and generative architectures (GAN), as well as discrete- and continuous-action reinforcement learning problems.

3 Benchmarks & New Avenues

This section presents benchmark results demonstrating the performance of TAGI on deep architectures and showcases some of the new avenues it enables. In addition to the early experiments

conducted on FNN for regression and classification problems [8], we have recently showed that TAGI outperforms deterministic and Bayesian CNN networks trained using backpropagation [19]. For the classification task on the CIFAR-10 images, using a Resnet18 [9], TAGI leads to an error



(a) Backprop. (100 epochs) (b) TAGI (50 epochs)

Figure 1: Latent space for hair color from (a) a 4.1M parameters backprop-trained network [4] with (b) a 0.7M parameters TAGI-trained network.

rate of 13.8%, while analogous networks trained with deterministic backpropagation leads to 14.0%, MC-dropout 17.2%, and VOGN 15.7%[21]. Moreover, TAGI only used one-third as many epochs as the other approaches.

While experimenting with infoGANs [4], we showed [19] that TAGI was able to generate comparable images while using half as many epochs and a network that is six times smaller than the analogous architecture relying on backpropagation. Figure 1 presents a comparison of such images.

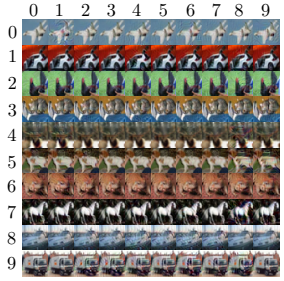


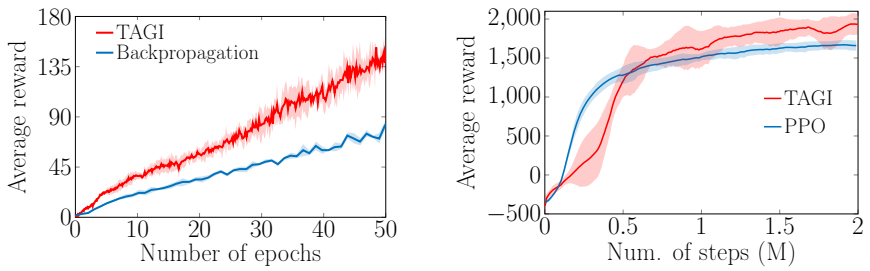
Figure 2: $> 99\%$ effective adversarial attacks for different target labels $\tilde{y}$ in each columns.

In addition to enabling the inference and propagation of uncertainty in FNNs, CNNs and GANs architecture, TAGI introduces novel possibilities due to its capacity to infer not only a network's parameters, but also any of its hidden states and input variables. With TAGI, the generation of adversarial-attack images can be done analytically, without relying on an optimization process; The prior knowledge $\{\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{X}}\}$ for the target image is propagated forward through the network by following the procedure presented in §2. Then, the correct label is replaced by the target label $\tilde{y}$ that is chosen for the attack. After performing the inference step, the image defined by its updated mean vector $\boldsymbol{\mu}_{\boldsymbol{X}|\tilde{y}}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{X}|\tilde{y}}$ is now modified in order to trigger the class $\tilde{y}$. In order to improve the quality of the attack, the process is repeated recursively over multiple iterations, where the inferred values $\{\boldsymbol{\mu}^{(i)}_{\boldsymbol{X}|\tilde{y}}, \boldsymbol{\Sigma}^{(i)}_{\boldsymbol{X}|\tilde{y}}\}$ at iteration i are used as the prior's hyperparameters at the next iteration $i+1$. Figure 2 present examples of $> 99\%$ effective TAGI-generated (i.e., optimization-free) adversarial attacks for CIFAR-10 images [19].

For reinforcement learning, TAGI natively enables dealing with the exploration/exploitation tradeoff using Thompson sampling [23]. Figure 3a compares the average reward over 100 episodes for three runs obtained for a TAGI deep Q-network [18], with the results from Mnih et al. [16] for the Breakout Atari game. Figure 3b displays a similar comparison for the Half Cheetah MuJoCo environment [24, 3], where TAGI employs simultaneously a policy and a value network in order to handle continuous actions [13], without relying on gradient backpropagation.

4 Conclusion

TAGI's performance on various experiments challenges the common belief that large-scale neural networks can only be trained by relying on gradient backpropagation. We have shown through



(a) Atari|Breakout (b) MuJoCo|HalfCheetah

Figure 3: Comparison of the reward from three runs for (a) the Breakout Atari game and (b) the Half Cheetah V2 MuJoCo environment for TAGI-based and Backprop-based RL frameworks.

[18, 13, 19, 8] that this current paradigm is no longer the only alternative as TAGI can be used to learn the parameters of complex networks in an analytically tractable manner, without relying on gradient-based optimization. In addition, TAGI requires fewer hyperparameters across different tasks. The applications presented are only a subset from the variety of problems that can take advantage of analytical inference, either through the adaptation of existing architectures or through the development of new ones.

Acknowledgements

The first author was financially supported by research grants from Hydro-Quebec, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] D. Barber and C. M. Bishop. Ensemble learning in bayesian neural networks. *NATO ASI Series F Computer and Systems Sciences*, 168:215–238, 1998.

[2] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

[3] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[4] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv:1606.03657*, 2016.

[5] S. Farquhar, L. Smith, and Y. Gal. Liberty or depth: Deep bayesian neural nets do not need complex weight posterior approximations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4346–4357, 2020.

[6] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016.

[8] J.-A. Goulet, L. H. Nguyen, and S. Amiri. Tractable approximate Gaussian inference for Bayesian neural networks. *Journal of Machine Learning Research*, (20-1009), 2021 (accepted for publication).

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

[11] G. E. Hinton and D. Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13. ACM, 1993.

[12] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, 2015.

[13] Nguyen L.H. and Goulet J-A. Analytically tractable hidden-states inference in Bayesian neural networks. *arXiv preprint arXiv:2107.03759*, 2021.

[14] C. Louizos and M. Welling. Structured and efficient variational deep learning with matrix Gaussian posteriors. In *International Conference on Machine Learning*, pages 1708–1716, 2016.

[15] D.J.C. MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[16] V. Mnih, Adria P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

[17] R. M. Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.

[18] L. H. Nguyen and J.-A. Goulet. Analytically tractable Bayesian deep Q-learning. *arXiv preprint arXiv:2106.11086*, 2021.

[19] L. H. Nguyen and J-A. Goulet. Analytically tractable inference in deep neural networks. *arXiv preprint arXiv:2103.05461*, 2021.

[20] K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota, and M. E. Khan. Practical deep learning with Bayesian principles. In *Advances in Neural Information Processing Systems*, 2019.

[21] K. Osawa, S. Swaroop, M. E. E. Khan, A. Jain, R. Eschenhagen, R. E. Turner, and R. Yokota. Practical deep learning with bayesian principles. In *Advances in neural information processing systems*, pages 4287–4299, 2019.

[22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533—536, 1986.

[23] M. Strens. A Bayesian framework for reinforcement learning. In *ICML*, volume 2000, pages 943–950, 2000.

[24] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[25] A. Wu, S. Nowozin, E. Meeds, R. E. Turner, J. M. Hernández-Lobato, and A. L. Gaunt. Deterministic variational inference for robust Bayesian neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.