Submission and Formatting Instructions for International Conference on Machine Learning (ICML 2018)

Aeiau Zzzz [*1] **Bauiu C. Yyyy** [*1 2] **Cieua Vvvvv** [2] **Iaesut Saoeu** [3] **Fiuea Rrrr** [1] **Tateu H. Yasehe** [3 1 2] **Aaoeu Iasoh** [2] **Buiui Eueu** [3] **Aeuia Zzzz** [3] **Bieea C. Yyyy** [1 2] **Teoau Xxxx** [3] **Eee Pppp** [3]

Abstract

This document provides a basic paper template and submission guidelines. Abstracts must be a single paragraph, ideally between 4–6 sentences long. Gross violations will trigger corrections at the camera-ready phase.

1. Introduction

In recent years, deep learning has achieved remarkable success across a wide range of classification tasks, including image recognition, natural language processing, and medical diagnosis (**?**). Despite their high predictive accuracy, traditional neural networks often operate as black boxes, providing point estimates without quantifying the uncertainty associated with their predictions. This limitation poses significant challenges, especially in safety-critical applications such as autonomous driving, healthcare, and financial decision-making, where understanding the confidence of predictions is crucial for reliable and trustworthy outcomes (**?**). Uncertainty estimation in neural networks addresses this challenge by providing measures of confidence alongside predictions. Bayesian Neural Networks (BNNs) offer a principled framework for uncertainty estimation by treating model parameters as random variables and capturing their posterior distribution (**?**). By integrating over the uncertainty in the model parameters, BNNs can provide more robust predictions and better generalize to unseen data. However, exact Bayesian inference in neural networks is intractable due to the high dimensionality and complexity of the parameter space, necessitating the development of approximate inference methods. Several approaches have been proposed to approximate the posterior distribution in BNNs, each with its own set of advantages and limitations. Sampling-based methods like Stochastic Gradient Langevin Dynamics (SGLD) () and Hamiltonian Monte Carlo (HMC) (**?**) provide accurate posterior approximations but are often computationally expensive and challenging to scale to large networks. On the other hand, approximation techniques such as Monte Carlo Dropout (MC Dropout) (**?**) and Deep Ensembles (**?**) offer more computationally feasible alternatives but may lack the theoretical rigor and flexibility to capture complex posterior distributions fully. Variational Inference (VI) methods, including Bayesian Backpropagation (BBP) (**?**), strike a balance between scalability and accuracy but are constrained by the choice of variational families, which may oversimplify the true posterior. Despite these advancements, existing methods often face trade-offs between computational efficiency, scalability, and the fidelity of uncertainty estimates. In particular, many approaches rely on assumptions such as Gaussianity in the parameter or output space, which may not hold in practice, especially for complex and high-dimensional data distributions. Additionally, the computational overhead associated with sampling-based methods limits their applicability to large-scale classification tasks, prompting the need for more efficient inference techniques.

In this work, we implement a novel method that leverages Tractable Approximate Gaussian Inference (TAGI) to provide efficient and scalable uncertainty estimation in Bayesian Deep Networks. TAGI offers an to sampling-based approaches by maintaining and updating approximate Gaussian distributions over network parameters, enabling the propagation of uncertainty through deep learning models seamlessly. Furthermore, we propose a probabilistic softmax formulation that translates the uncertainty from the logit space to the output space, allowing for accurate and computationally efficient uncertainty estimates in classification tasks.

The focus of our project can be summarized as follows:

- **Tractable Approximate Gaussian Inference (TAGI):** We re-implement the toy example proposed in the original TAGI paper (**?**).
- **Probabilistic Softmax for Uncertainty Estimation:**

[*]Equal contribution [1]Department of Computation, University of Torontoland, Torontoland, Canada [2]Googol ShallowMind, New London, Michigan, USA [3]School of Computation, University of Edenborrow, Edenborrow, United Kingdom. Correspondence to: Cieua Vvvvv <c.vvvvv@googol.com>, Eee Pppp <ep@eden.co.uk>.

IFT 6269 project report - Université de Montréal

We provide an implementation of the probabilistic softmax formulation porposed by the authors of TAGI, along with an evaluation on datasets such as MNIST and CIFAR-10.

- **Ablation Study:** We conduct an ablation study to evaluate the effectiveness of TAGI-based inference and the probabilistic softmax formulation in uncertainty estimation for classification tasks.

2. Related Work

Bayesian Neural Networks (BNNs) provide a principled framework for uncertainty estimation by capturing a posterior distribution over model parameters. Various methods, including sampling-based techniques and approximations, have been proposed to estimate uncertainty in classification tasks efficiently. Below, we review prominent methods alongside a recent advance using analytical approximations.

2.1. Monte Carlo Dropout (MC Dropout)

MC Dropout, introduced by (**?**), approximates Bayesian inference by using dropout during both training and inference. Multiple stochastic forward passes are performed at inference by sampling different dropout masks, and the outputs are aggregated to estimate the predictive distribution. The predictive mean is computed as:

$$\hat{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}, \mathbf{w}^{(t)}),$$

where $\mathbf{w}^{(t)}$ are the sampled weights with dropout and T is the number of stochastic passes. This approach is computationally efficient and easy to implement but is limited by the reliance on dropout regularization, which might not fully capture posterior uncertainty.

2.2. Stochastic Gradient Langevin Dynamics (SGLD)

SGLD, proposed by (**?**), combines stochastic gradient descent (SGD) with Langevin dynamics to sample from the posterior distribution of weights. The weight updates include Gaussian noise:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \mathcal{L}(\mathbf{w}_t) + \sqrt{2\eta}\epsilon_t,$$

where η is the learning rate, $\mathcal{L}$ is the loss function, and $\epsilon_t \sim \mathcal{N}(0, I)$ is Gaussian noise. By incorporating noise into the gradient updates, SGLD provides samples that approximate the posterior. However, the method is more computationally intensive compared to approximation techniques like MC Dropout and can be sensitive to hyperparameters such as the step size.

2.3. Hamiltonian Monte Carlo (HMC)

HMC (**?**) extends traditional Markov Chain Monte Carlo (MCMC) by using Hamiltonian dynamics to sample efficiently from the posterior. It introduces momentum variables and explores the posterior landscape using the Hamiltonian:

$$\mathcal{H}(\mathbf{w}, \mathbf{p}) = -\log p(\mathbf{w}|\mathcal{D}) + \frac{\|\mathbf{p}\|^2}{2},$$

where $\mathbf{p}$ is the momentum. HMC provides accurate posterior samples, but its computational cost is prohibitive for high-dimensional neural network parameters. Additionally, it requires careful tuning of the step size and number of leapfrog steps.

2.4. Bayesian Backpropagation (BBP)

Bayesian Backpropagation (BBP), also known as Bayes by Backprop (**?**), uses variational inference to approximate the posterior distribution over weights. A variational distribution $q(\mathbf{w})$, typically Gaussian, is optimized to minimize the Kullback-Leibler (KL) divergence from the true posterior. The variational posterior is parameterized as:

$$q(\mathbf{w}|\boldsymbol{\theta}) = \prod_i \mathcal{N}(w_i; \mu_i, \sigma_i^2),$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ are learnable parameters. Inference involves sampling weights from $q(\mathbf{w})$ and averaging multiple forward passes. While BBP provides a principled approach to Bayesian inference, the quality of the approximation depends on the chosen variational family, which might not fully capture the true posterior.

2.5. Deep Ensembles

Deep Ensembles (**?**) estimate uncertainty by training multiple neural networks with different random initializations. The ensemble predictions are averaged to compute the predictive distribution:

$$\hat{\mathbf{y}} = \frac{1}{M}\sum_{m=1}^{M} f(\mathbf{x}; \mathbf{w}_m),$$

where $\mathbf{w}_m$ are the weights of the m^{th} ensemble member. The variance across the ensemble outputs represents model uncertainty. This method is robust and often outperforms other techniques in practice. However, it is computationally expensive due to the need to train multiple models independently.

2.6. Variational Inference (VI)

Variational Inference (**?**) approximates the posterior distribution $p(\mathbf{w}|\mathcal{D})$ with a simpler variational distribution $q(\mathbf{w})$

by minimizing the KL divergence:

$$\mathrm{KL}(q(\mathbf{w})\|p(\mathbf{w}|\mathcal{D})) = \int q(\mathbf{w}) \log \frac{q(\mathbf{w})}{p(\mathbf{w}|\mathcal{D})}\, d\mathbf{w}.$$

Common variational families include mean-field Gaussians, which are computationally efficient but may not fully capture the complexity of the true posterior. Variational inference is scalable and adaptable but often trades accuracy for efficiency.

2.7. Fast Predictive Uncertainty for Classification with Bayesian Deep Networks

(**?**) propose a method that revisits the Laplace Bridge technique to approximate the softmax output distribution using a Dirichlet distribution. The method maps Gaussian distributions in logit space $\mathcal{N}(\mu, \Sigma)$ to Dirichlet distributions in output space $\mathrm{Dir}(\boldsymbol{\alpha})$ via the analytical mapping:

$$\alpha_k = \exp\left(\mu_k + \frac{\sigma_k^2}{2}\right), \quad \text{for } k = 1, \ldots, K,$$

where μ_k and σ_k^2 are the mean and variance of the logit for class k , and α_k is the parameter of the resulting Dirichlet distribution. This approach avoids sampling and significantly reduces computational costs. It scales effectively to large datasets like ImageNet and complex architectures like DenseNet. While the method is efficient, its limitations include reliance on Gaussian assumptions in logit space and potential inaccuracies in modeling highly non-Gaussian posteriors. Additionally, its performance on small or imbalanced datasets is underexplored.

Each method offers unique strengths and weaknesses. MC Dropout and Deep Ensembles are practical and straightforward but may lack the theoretical rigor of sampling-based methods like SGLD and HMC. Variational inference approaches, such as BBP, balance scalability and accuracy but depend heavily on the choice of variational family. Sampling-based methods like SGLD and HMC provide accurate posterior approximations but are computationally expensive. The Dirichlet-based approach by Hobbhahn et al. introduces computational efficiency and scalability but trades off some flexibility in capturing complex posterior distributions.

In the next section, we introduce a novel method that combines the efficiency of analytical approximations with the theoretical grounding of Bayesian inference. The method relies on Tractable Approximate Gaussian Inference (TAGI) as a learning paradigm that allows passing uncertainty through deep learning models, and the formulation of a probabilistic softmax that further allows passing that uncertainty to the output layer.

3. Methodology

3.1. Tractable approximate gaussian inference (TAGI)

Goulet et al. (**?**) proposed the Tractable Approximate Gaussian Inference (TAGI) method, which is a Bayesian approach to neural networks. Compared to traditional NNs, this method allows to (1) optimize parameters analytically, (2) treat uncertainties from input to output layers and provide density forecasts.
Consider a fully connected NN presented in Figure **??**, the formulations of a TAGI-NN, which is similar to that of a traditional NN except that parameters and variables are assumed to be normally distributed, are defined by the following deterministic equations:

$$\begin{aligned}
\boldsymbol{z}^{(1)} &= \boldsymbol{w}^{(0)}\boldsymbol{x} + \boldsymbol{b}^{(0)}, \\
\boldsymbol{a}^{(i)} &= \tilde{\sigma}(\boldsymbol{z}^{(i)}) \quad (i = 1 : \mathtt{L}), \\
\boldsymbol{z}^{(i)} &= \boldsymbol{w}^{(i)}\boldsymbol{a}^{(i-1)} + \boldsymbol{b}^{(i)} \quad (i = 2 : \mathtt{L}), \\
\boldsymbol{z}^{(\mathtt{O})} &= \boldsymbol{w}^{(\mathtt{L})}\boldsymbol{a}^{(\mathtt{L})} + \boldsymbol{w}^{(\mathtt{L})} \\
\boldsymbol{y} &= \boldsymbol{z}^{(\mathtt{O})} + \boldsymbol{v}.
\end{aligned}$$

where $\boldsymbol{w}, \boldsymbol{b}$ are parameters, $\mathtt{L}$ is the number of hidden layers, $\boldsymbol{v}$ are the observation noise, $\boldsymbol{z}$ are hidden units, $\boldsymbol{y}$ are outputs, and $\tilde{\sigma}(\cdot)$ is the linearized activation function.

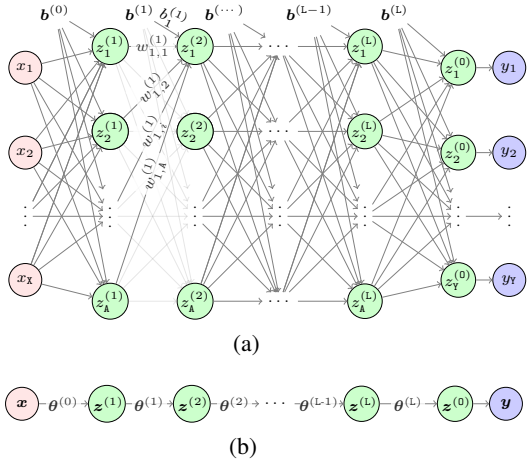

Figure 1. (a) A neural network representation. (b) A compacted NN representation, reproduced from (**?**).

In Equation **??**, the product of two Gaussian random variables is not Gaussian, despite of this, the authors proposed using the *Gaussian Multiplication Approximation* (GMA) to approximate the PDF of any product term $X_i X_j$ given by

$$\mathbb{E}\left[X_1 X_2\right] = \mu_1 \mu_2 + \mathrm{cov}\left(X_1, X_2\right)$$

$$\mathrm{cov}\left(X_3, X_1 X_2\right) = \mathrm{cov}\left(X_1, X_3\right) \mu_2 + \mathrm{cov}\left(X_2, X_3\right) \mu_1$$

$$\begin{aligned}\operatorname{cov}(X_1X_2, X_3X_4) = &\operatorname{cov}(X_1, X_3)\operatorname{cov}(X_2, X_4)\\ &+ \operatorname{cov}(X_1, X_4)\operatorname{cov}(X_2, X_3)\\ &+ \operatorname{cov}(X_1, X_3)\mu_2\mu_4\\ &+ \operatorname{cov}(X_1, X_4)\mu_2\mu_3\\ &+ \operatorname{cov}(X_2, X_3)\mu_1\mu_4\\ &+ \operatorname{cov}(X_2, X_4)\mu_1\mu_3\end{aligned}$$

$$\begin{aligned}\operatorname{var}(X_1X_2) = &\sigma_1^2\sigma_2^2 + \operatorname{cov}(X_1, X_2)^2\\ &+ 2\operatorname{cov}(X_1, X_2)\mu_1\mu_2\\ &+ \sigma_1^2\mu_2^2 + \sigma_2^2\mu_1^2\end{aligned}$$

where μ, σ are respectively the mean and standard deviation of a random variable. Unlike a traditional NN where parameters are updated using backpropagation, model parameters in a TAGI-NN can be obtained analytically by a layer-wise scheme as depicted in Figure **??**. First, given the

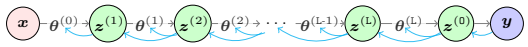

Figure 2. Recursive layer-wise inference, reproduced from (**?**). Cyan arrows indicate inference directions.

observations $\boldsymbol{y}$, the output layer is updated as

$$\begin{aligned}f\left(\boldsymbol{z}^{(\mathtt{0})} \mid \boldsymbol{y}\right) &= \mathcal{N}\left(\boldsymbol{z}^{(\mathtt{0})}; \boldsymbol{\mu}_{\boldsymbol{Z}^{(\mathtt{0})}|\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{Z}^{(\mathtt{0})}|\boldsymbol{y}}\right)\\ \boldsymbol{\mu}_{\boldsymbol{Z}^{(\mathtt{0})}|\boldsymbol{y}} &= \boldsymbol{\mu}_{\boldsymbol{Z}^{(\mathtt{0})}} + \boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Z}^{(\mathtt{0})}}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{Y}}\right)\\ \boldsymbol{\Sigma}_{\boldsymbol{Z}^{(\mathtt{0})}|\boldsymbol{y}} &= \boldsymbol{\Sigma}_{\boldsymbol{Z}^{(\mathtt{0})}} - \boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Z}^{(\mathtt{0})}}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Z}^{(\mathtt{0})}}.\end{aligned}$$

Next, the hidden units and parameters of the layer i^{th} are updated using a layer-wise procedure as

$$\begin{aligned}f(\boldsymbol{z} \mid \boldsymbol{y}) &= \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}_{\boldsymbol{Z}|\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{Z}|\boldsymbol{y}}\right)\\ \boldsymbol{\mu}_{\boldsymbol{Z}|\boldsymbol{y}} &= \boldsymbol{\mu}_{\boldsymbol{Z}} + \mathbf{J}_{\boldsymbol{Z}}\left(\boldsymbol{\mu}_{\boldsymbol{Z}^{+}|\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{Z}^{+}}\right)\\ \boldsymbol{\Sigma}_{\boldsymbol{Z}|\boldsymbol{y}} &= \boldsymbol{\Sigma}_{\boldsymbol{Z}} + \mathbf{J}_{\boldsymbol{Z}}\left(\boldsymbol{\Sigma}_{\boldsymbol{Z}^{+}|\boldsymbol{y}} - \boldsymbol{\Sigma}_{\boldsymbol{Z}^{+}}\right)\mathbf{J}_{\boldsymbol{Z}}^{\top}\\ \mathbf{J}_{\boldsymbol{Z}} &= \boldsymbol{\Sigma}_{\boldsymbol{Z}\boldsymbol{Z}^{+}}\boldsymbol{\Sigma}_{\boldsymbol{Z}^{+}}^{-1},\\ f(\boldsymbol{\theta} \mid \boldsymbol{y}) &= \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}}\right)\\ \boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{y}} &= \boldsymbol{\mu}_{\boldsymbol{\theta}} + \mathbf{J}_{\boldsymbol{\theta}}\left(\boldsymbol{\mu}_{\boldsymbol{Z}^{+}|\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{Z}^{+}}\right)\\ \boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}} &= \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \mathbf{J}_{\boldsymbol{\theta}}\left(\boldsymbol{\Sigma}_{\boldsymbol{Z}^{+}|\boldsymbol{y}} - \boldsymbol{\Sigma}_{\boldsymbol{Z}^{+}}\right)\mathbf{J}_{\boldsymbol{\theta}}^{\top}\\ \mathbf{J}_{\boldsymbol{\theta}} &= \boldsymbol{\Sigma}_{\boldsymbol{\theta}Z^{+}}\boldsymbol{\Sigma}_{\boldsymbol{Z}^{+}}^{-1},\end{aligned}$$

where $\{\boldsymbol{\theta}^{+}, \boldsymbol{Z}^{+}\}$ is the short-hand notation of $\left\{\boldsymbol{\theta}^{(j+1)}, \boldsymbol{Z}^{(j+1)}\right\}$, and $\{\boldsymbol{\theta}, \boldsymbol{Z}\}$ is the short-hand notation of $\left\{\boldsymbol{\theta}^{(j)}, \boldsymbol{Z}^{(j)}\right\}$. While this section only introduces the TAGI method for a simple feedforward neural network, this method is applicable for any NN architectures that are already developed. This is because the difference between TAGI-NN and traditional NN models is the parameter update mechanism but the architecture is unchanged.

I NEED TO FURTHER ADD WHY THE ASSUMPTIONS HOLD.

4. Probabilistic Softmax

In this section we will present how the current classification in TAGI is performed, we will present a probabilistic softmax formulation and a new Remax activation.

4.1. Classification with TAGI

TAGI addresses classification tasks through a hierarchical binary tree structure, where each class is uniquely represented as a path in the tree. The tree's height is given by $\mathtt{H} = \lceil \log_2(\mathtt{K}) \rceil$, and the number of hidden states is $\mathtt{Y} = \mathtt{K} - 1$, provided $\log_2(\mathtt{K}) \in \mathbb{Z}^{+}$.

For instance, when $\mathtt{K} = 8$, the hierarchical decomposition results in $\mathtt{H} = 3$ layers, where each class $y_{\mathcal{C}}^{(\mathtt{C})}$ is uniquely identified by a binary sequence $\mathcal{C} = \{j, k, l\} \in \{0, 1\}^3$, corresponding to the indices of the path in the tree.

This approach reformulates the classification task as a regression problem, eliminating the need for a softmax activation function. However, the current implementation has limitations in capturing epistemic uncertainty effectively, as it enforces constraints on the probabilities of each hidden state to remain within the interval $[0, 1]$. Although a solution to this issue has been proposed by the authors, it has not yet been implemented.

Additionally, directly applying the softmax function is impractical in this context because leveraging uncertainty requires a probabilistic representation rather than a deterministic one. This work aims to address these challenges by implementing and evaluating the probabilistic softmax framework previously proposed by the authors, which is expected to overcome the limitations of the current implementation.

4.2. Closed-form softmax using a log-transformation

In neural networks, the softmax function is commonly defined as:

$$\operatorname{softmax}(\boldsymbol{z}, i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}. \tag{1}$$

While effective in standard implementations, this function poses challenges in the context of Tractable Approximate Gaussian Inference (TAGI) (**?**). Specifically, performing this operation within TAGI requires dividing two Gaussian-distributed variables. However, the result of such a division does not follow a Gaussian distribution, which disrupts the probabilistic consistency of the framework.

To address this limitation, the authors of TAGI (**?**) proposed a novel approach to adapt the softmax function for probabilistic contexts where z_i follows a Gaussian distribution. The method leverages a log-normal transformation to ensure mathematical tractability. By matching the moments of the neural network's output in the log-normalized space, the division inherent to the softmax function can be refor-

mulated as a subtraction between log-normal distributions. Subsequently, the moments of the resulting subtraction are remapped to a Gaussian distribution, preserving the probabilistic framework.This adaptation enables the application of softmax within TAGI while maintaining the Gaussian assumptions.

For hidden units $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu_Z}, \boldsymbol{\sigma_Z^2} \cdot \mathbf{I})$ such that $\boldsymbol{z} \in \mathbb{R}^{\mathsf{z}}$, then taking the exponential function of z_i leads to a lognormal distribution

$$\begin{aligned} E = \exp(Z) &\sim \ln\mathcal{N}(e; \mu_Z, \sigma_Z) \\ \mu_E &= \exp\left(\mu_Z + \tfrac{1}{2}\sigma_Z^2\right) \\ \sigma_E^2 &= \exp\left(2\mu_Z + \sigma_Z^2\right) \cdot \left(\exp(\sigma_Z^2) - 1\right), \end{aligned} \tag{2}$$

where the covariance between Z_i and E_i is

$$\operatorname{cov}(Z, E) = \sigma_Z^2 \cdot \mu_E. \tag{3}$$

Given that $Z_i \perp\!\!\!\perp Z_j, \forall i \neq j$, the sum on the softmax's denominator $\tilde{E}$ is modelled by a Gaussian R.V.,

$$\begin{aligned} \textstyle\sum_j E_j = \tilde{E} &\sim \mathcal{N}(\tilde{e}; \mu_{\tilde{E}}, \sigma_{\tilde{E}}^2) \\ \mu_{\tilde{E}} &= \textstyle\sum_j \mu_{E,j} \\ \sigma_{\tilde{E}}^2 &= \textstyle\sum_j \sigma_{E,j}^2. \end{aligned} \tag{4}$$

Given the independence hypothesis between hidden units the covariance and coefficient of correlation between the softmax's numerator and denominator are only influenced by the presence of a same variable E_i in both terms so that,

$$\operatorname{cov}\big(E_i, \tilde{E}\big) = \sigma_{E_i}^2 \tag{5}$$

For a lognormal random variable X, the relations between the R.V.'s moments and the moment in the log-transformed space λ and ζ are

$$X \sim \ln\mathcal{N}(x; \lambda, \zeta) \begin{cases} \zeta &= \sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)} \\ \lambda &= \ln\mu_X - \frac{1}{2}\zeta^2. \end{cases} \tag{6}$$

Using these relations, we can obtain the softmax's denominator in the log-transformed space so that we can replace the division by a substraction,

$$\begin{aligned} \ln\tilde{E} &\sim \ln\mathcal{N}(\ln e; \mu_{\ln E}, \sigma_{\ln E}) \\ \sigma_{\ln E}^2 &= \ln\left(1 + \frac{\sigma_{\tilde{E}}^2}{\mu_{\tilde{E}}^2}\right) \\ \mu_{\ln E} &= \ln\mu_{\tilde{E}} - \tfrac{1}{2}\sigma_{\ln E}^2. \end{aligned} \tag{7}$$

The covariance between the softmax's numerator and denominator in the log-transformed space is

$$\operatorname{cov}\big(\ln E_i, \ln\tilde{E}\big) = \ln\left(1 + \operatorname{cov}\big(E_i, \tilde{E}\big) \cdot \tfrac{1}{\mu_{E,i}} \cdot \tfrac{1}{\mu_{\tilde{E}}}\right). \tag{8}$$

The i-th attention unit in the log-space $\check{A}_i = \ln E_i - \ln\tilde{E}$ defines a lognormal R.V. $A_i = \exp(\check{A}_i)$ so that

$$\begin{aligned} A_i &\sim \ln\mathcal{N}(\check{a}_i; \mu_{\check{A},i}, \sigma_{\check{A},i}) \\ \mu_{\check{A},i} &= \mu_{\ln E,i} - \mu_{\ln\tilde{E}} \\ \sigma_{\check{A},i}^2 &= \sigma_{\ln E,i}^2 + \sigma_{\ln\tilde{E}}^2 - 2\operatorname{cov}\big(\ln E_i, \ln\tilde{E}\big). \end{aligned} \tag{9}$$

The covariance between $\check{A}_i$ and $\ln E_i$ is given by

$$\operatorname{cov}(\check{A}_i, Z_i) = \sigma_{\ln E,i}^2 - \operatorname{cov}\big(\ln E_i, \ln\tilde{E}\big). \tag{10}$$

The moments of the lognormal variable A_i are

$$\begin{aligned} \mu_{A,i} &= \exp(\mu_{\check{A},i} + \tfrac{1}{2}\sigma_{\check{A},i}^2) \\ \sigma_{A,i}^2 &= \mu_{A,i}^2\big(\exp(\sigma_{\check{A},i}^2) - 1\big) \end{aligned} \tag{11}$$

In order to employ attention units with Gaussian inference, we use the approximation that A_i is Gaussian with

$$A_i \sim \mathcal{N}(a_i; \mu_{A,i}, \sigma_{A,i}^2),$$

with the assumption that $A_i \perp\!\!\!\perp A_j, \forall i \neq j$. The regression coefficient $\beta_{\check{A}|A}$ expressing the linear relationship between $\tilde{A}_i$ and A_i is given by

$$\beta_{A|\check{A}} = \frac{\operatorname{cov}(\check{A}_i, A_i)}{\sigma_{\check{A},i}^2}, \tag{12}$$

so that the covariance between A_i and Z_i is

$$\begin{aligned} \operatorname{cov}(A_i, Z_i) &= \tfrac{\operatorname{cov}(\check{A}_i, A_i)}{\sigma_{\check{A},i}^2} \cdot \operatorname{cov}(\check{A}_i, Z_i) \\ &= \exp(\mu_{\check{A},i} + \tfrac{1}{2}\sigma_{\check{A},i}^2) \cdot \operatorname{cov}(\check{A}_i, Z_i). \end{aligned} \tag{13}$$

4.3. Remax

The authors of the formulation Goulet and Nguyen observed through experimentation and comparison of the probabilistic softmax with results obtained via Monte Carlo sampling that high variance values ($\sigma^2 \approx 1$) can lead to exploding gradients. To mitigate this issue, they proposed replacing the exponential function with a novel activation function, termed the *Mixture Rectified Linear activation Unit* (Mixture ReLU).

The concept behind the *Mixture Rectified Linear activation Unit* is to propagate the hidden unit's uncertainty $Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ through a ReLU $\phi_{\mathrm{R}}(z) = \max(0, z)$ by using a Gaussian mixture between a truncated Gaussian and a zero-valued component. The expected value of the

Gaussian PDF of $\tilde{Z}$ truncated at $z \geq 0$ are

$$\begin{aligned} \phi_{\mathrm{R}}(z) &= \max(0, z) \\ \mu_A &= \mu_Z \cdot \Phi\left(\frac{\mu_Z}{\sigma_Z}\right) + \sigma_Z \cdot \varphi\left(\frac{\mu_Z}{\sigma_Z}\right) \\ \sigma_A^2 &= -\mu_A^2 + 2\mu_A\mu_Z - \mu_Z\sigma_Z\phi\left(\frac{\mu_Z}{\sigma_Z}\right) \\ &\quad + \left(\sigma_Z^2 - \mu_Z^2\right)\Phi\left(\frac{\mu_Z}{\sigma_Z}\right) \\ \mathrm{cov}(A, Z) &= \sigma_Z^2\Phi\left(\frac{\mu_Z}{\sigma_Z}\right) \end{aligned} \tag{14}$$

Then, the remax activation function is defined as

$$\mathrm{remax}(\boldsymbol{z}, i) = \frac{\mathrm{mrelu}(z_i)}{\sum_j \mathrm{mrelu}(z_j)}. \tag{15}$$

All the calculations remain the same as for the probabilistic softmax but instead of $E = \exp(Z)$, we will have

$$\boldsymbol{M} \sim \mathcal{N}(\boldsymbol{m}; \boldsymbol{\mu}_{\boldsymbol{M}}, \boldsymbol{\sigma}_{\boldsymbol{M}}^2 \cdot \mathbf{I}) = \mathrm{mrelu}(\boldsymbol{Z}).$$

Then, the convariance between A_i and Z_i is

$$\begin{aligned} \mathrm{cov}(A_i, Z_i) &= \tfrac{\mathrm{cov}(\check{A}_i, A_i)}{\sigma^2_{\check{A},i}} \cdot \tfrac{\mathrm{cov}(\check{A}_i, M_i)}{\sigma^2_{M,i}} \cdot \mathrm{cov}(M_i, Z_i) \\ &= \exp(\mu_{\check{A},i} + \tfrac{1}{2}\sigma^2_{\check{A},i}) \\ &\quad \cdot \tfrac{\mathrm{cov}(\check{A}_i, M_i) \cdot \mathrm{cov}(M_i, Z_i)}{\sigma^2_{M,i}}. \end{aligned} \tag{16}$$

5. Remax results

6. Electronic Submission

Submission to ICML 2018 will be entirely electronic, via a web site (not email). Information about the submission process and LaTeX templates are available on the conference web site at:

`http://icml.cc/2018/`

The guidelines below will be enforced for initial submissions and camera-ready copies. Here is a brief summary:

- Submissions must be in PDF.
- The maximum paper length is **8 pages excluding references and acknowledgements, and 10 pages including references and acknowledgements** (pages 9 and 10 must contain only references and acknowledgements).
- **Do not include author information or acknowledgements** in your initial submission.
- Your paper should be in **10 point Times font**.
- Make sure your PDF file only uses Type-1 fonts.
- Place figure captions *under* the figure (and omit titles from inside the graphic file itself). Place table captions *over* the table.
- References must include page numbers whenever possible and be as complete as possible. Place multiple citations in chronological order.
- Do not alter the style template; in particular, do not compress the paper format by reducing the vertical spaces.
- Keep your abstract brief and self-contained, one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase. The title should have content words capitalized.

6.1. Submitting Papers

Paper Deadline: The deadline for paper submission that is advertised on the conference website is strict. If your full, anonymized, submission does not reach us on time, it will not be considered for publication. There is no separate abstract submission.

Anonymous Submission: ICML uses double-blind review: no identifying author information may appear on the title page or in the paper itself. Section **??** gives further details.

Simultaneous Submission: ICML will not accept any paper which, at the time of submission, is under review for another conference or has already been published. This policy also applies to papers that overlap substantially in technical content with conference papers under review or previously published. ICML submissions must not be submitted to other conferences during ICML's review period. Authors may submit to ICML substantially different versions of journal papers that are currently under review by the journal, but not yet accepted at the time of submission. Informal publications, such as technical reports or papers in workshop proceedings which do not appear in print, do not fall under these restrictions.

Authors must provide their manuscripts in **PDF** format. Furthermore, please make sure that files contain only embedded Type-1 fonts (e.g., using the program `pdffonts` in linux or using File/DocumentProperties/Fonts in Acrobat). Other fonts (like Type-3) might come from graphics files imported into the document.

Authors using **Word** must convert their document to PDF. Most of the latest versions of Word have the facility to do this automatically. Submissions will not be accepted in Word format or any format other than PDF. Really. We're not joking. Don't send Word.

Those who use **LaTeX** should avoid including Type-3 fonts. Those using `latex` and `dvips` may need the following

two commands:

```
dvips -Ppdf -tletter -G0 -o paper.ps paper.dvi
ps2pdf paper.ps
```

It is a zero following the "-G", which tells dvips to use the config.pdf file. Newer TEX distributions don't always need this option.

Using `pdflatex` rather than `latex`, often gives better results. This program avoids the Type-3 font problem, and supports more advanced features in the `microtype` package.

Graphics files should be a reasonable size, and included from an appropriate format. Use vector formats (.eps/.pdf) for plots, lossless bitmap formats (.png) for raster graphics with sharp lines, and jpeg for photo-like images.

The style file uses the `hyperref` package to make clickable links in documents. If this causes problems for you, add `nohyperref` as one of the options to the `icml2018` usepackage statement.

6.2. Submitting Final Camera-Ready Copy

The final versions of papers accepted for publication should follow the same format and naming convention as initial submissions, except that author information (names and affiliations) should be given. See Section **??** for formatting instructions.

The footnote, "Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute." must be modified to "*Proceedings of the 35^{th} International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s)."

For those using the **LaTeX** style file, this change (and others) is handled automatically by simply changing `\usepackage{icml2018}` to

```
\usepackage[accepted]{icml2018}
```

Authors using **Word** must edit the footnote on the first page of the document themselves.

Camera-ready copies should have the title of the paper as running head on each page except the first one. The running title consists of a single line centered above a horizontal rule which is 1 point thick. The running head should be centered, bold and in 9 point type. The rule should be 10 points above the main text. For those using the **LaTeX** style file, the original title is automatically set as running head using the `fancyhdr` package which is included in the ICML 2018 style file package. In case that the original title exceeds the size restrictions, a shorter form can be supplied by using

```
\icmltitlerunning{...}
```

just before `\begin{document}`. Authors using **Word** must edit the header of the document themselves.

7. Format of the Paper

All submissions must follow the specified format.

7.1. Length and Dimensions

Papers must not exceed eight (8) pages, including all figures, tables, and appendices, but excluding references and acknowledgements. When references and acknowledgements are included, the paper must not exceed ten (10) pages. Acknowledgements should be limited to grants and people who contributed to the paper. Any submission that exceeds this page limit, or that diverges significantly from the specified format, will be rejected without review.

The text of the paper should be formatted in two columns, with an overall width of 6.75 inches, height of 9.0 inches, and 0.25 inches between the columns. The left margin should be 0.75 inches and the top margin 1.0 inch (2.54 cm). The right and bottom margins will depend on whether you print on US letter or A4 paper, but all final versions must be produced for US letter size.

The paper body should be set in 10 point type with a vertical spacing of 11 points. Please use Times typeface throughout the text.

7.2. Title

The paper title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

7.3. Author Information for Submission

ICML uses double-blind review, so author information must not appear. If you are using LaTeX and the `icml2018.sty` file, use `\icmlauthor{...}` to specify authors and `\icmlaffiliation{...}` to specify affiliations. (Read the TeX code used to produce this document for an example usage.) The author information will not be printed unless `accepted` is passed as an argument to the style file. Submissions that include the author information will not be reviewed.

7.3.1. Self-Citations

If you are citing published papers for which you are an author, refer to yourself in the third person. In particular, do not use phrases that reveal your identity (e.g., "in previous work (**?**), we have shown ...").

Do not anonymize citations in the reference section. The only exception are manuscripts that are not yet published (e.g., under submission). If you choose to refer to such unpublished manuscripts , anonymized copies have to be submitted as Supplementary Material via CMT. However, keep in mind that an ICML paper should be self contained and should contain sufficient detail for the reviewers to evaluate the work. In particular, reviewers are not required to look at the Supplementary Material when writing their review.

7.3.2. Camera-Ready Author Information

If a paper is accepted, a final camera-ready copy must be prepared. For camera-ready papers, author information should start 0.3 inches below the bottom rule surrounding the title. The authors' names should appear in 10 point bold type, in a row, separated by white space, and centered. Author names should not be broken across lines. Unbolded superscripted numbers, starting 1, should be used to refer to affiliations.

Affiliations should be numbered in the order of appearance. A single footnote block of text should be used to list all the affiliations. (Academic affiliations should list Department, University, City, State/Region, Country. Similarly for industrial affiliations.)

Each distinct affiliations should be listed once. If an author has multiple affiliations, multiple superscripts should be placed after the name, separated by thin spaces. If the authors would like to highlight equal contribution by multiple first authors, those authors should have an asterisk placed after their name in superscript, and the term "*Equal contribution" should be placed in the footnote block ahead of the list of affiliations. A list of corresponding authors and their emails (in the format Full Name <email@domain.com>) can follow the list of affiliations. Ideally only one or two names should be listed.

A sample file with author names is included in the ICML2018 style file package. Turn on the `[accepted]` option to the stylefile to see the names rendered. All of the guidelines above are implemented by the LaTeX style file.

7.4. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase.

7.5. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

7.5.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

7.5.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

7.6. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should

[1] Footnotes should be complete sentences.

[2] Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

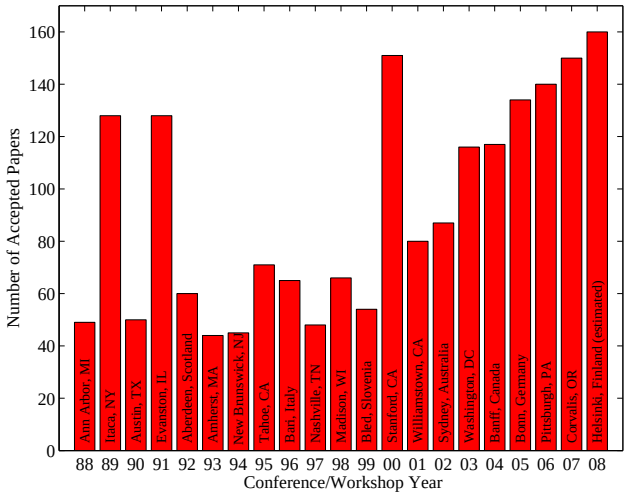

Figure 3. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true.$
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9± 0.2	96.7± 0.2	√
CLEVELAND	83.3± 0.6	80.0± 0.6	×
GLASS2	61.9± 1.4	83.8± 0.7	√
CREDIT	74.8± 0.5	78.3± 0.6	
HORSE	73.3± 0.9	69.7± 1.0	×
META	67.1± 0.6	76.5± 0.5	√
PIMA	75.1± 0.6	73.9± 0.5	
VEHICLE	44.9± 0.6	61.5± 0.4	√

serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure **??**. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

7.7. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm **??** shows an example.

7.8. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table **??**. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

7.9. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2018.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (**?**). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (**?**). List multiple references separated by semicolons (**???**). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (**?**).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section **??** for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the refer-

ences, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (**?**), conference publications (**?**), book chapters (**?**), books (**?**), edited volumes (**?**), technical reports (**?**), and dissertations (**?**).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

Please put some effort into making references complete, presentable, and consistent. If using bibtex, please protect capital letters of names and abbreviations in titles, for example, use {B}ayesian or {L}ipschitz in your .bib file.

7.10. Software and Data

We strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, do not include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look at this material when writing their review.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

A. Do *not* have an appendix here

Do not put content after the references. Put anything that you might normally include after the references in a separate supplementary file.

We recommend that you build supplementary material in a separate document. If you must create one PDF and cut it up, please be careful to use a tool that doesn't alter the margins, and that doesn't aggressively rewrite the PDF file. pdftk usually works fine.

Please do not use Apple's preview to cut off supplementary material. In previous years it has altered margins, and created headaches at the camera-ready stage.