# Dynamic Tempering (need a better name) for Asymptotic Uniform Confidence in Bayesian ReLU Networks

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

ReLU networks, in conjunction with maximum a posteriori estimation, produce overconfidence predictions far away from the training data. Recently, it has been shown that cheap Gaussian-approximated Bayesian methods can mitigate this issue to some degree. However, asymptotically, the confidence of these Bayesian ReLU networks only converges to a constant, which can be loose depending on the network. We propose a simple heuristic, based on the tempering framework, that is applied on the distribution over logits of these networks, where we let the temperature parameter depends on the input. This method gives a stronger asymptotic confidence guarantee compared to the previous work: Far away from the training data, the confidence decays proportionally to the distance and uniform at the limit. Meanwhile, we also show this method does not change the decision boundary of the underlying network and the confidence near the training data is still close its original value.

## 1 New Stuff added here for easy reference

Consider WLOG, a three-layer binary ReLU classification network:

$$f_\theta(x) := w_2^\top r(W_1 r(W_0 x + b_0) + b_1) + b_2$$

where

$$\theta := \{W_0, W_1, w_2, b_0, b_1, b_2\} \in \mathbb{R}^P \,,$$

with

$$W_0 \in \mathbb{R}^{N_1 \times N_0}, W_1 \in \mathbb{R}^{N_2 \times N_1}, w_2 \in \mathbb{R}^{N_2} \,,$$
$$b_0 \in \mathbb{R}^{N_1}, b_1 \in \mathbb{R}^{N_2}, b_2 \in \mathbb{R} \,,$$

and $r$ is the component-wise ReLU function. Given a dataset $\mathcal{D}$, we consider the loss function $\ell : \mathbb{R}^P \to \mathbb{R}$.

The idea is to add additional *inactive* ReLU units, as follows. Define $\tilde{N_1} := N_1 + M_1$ and $\tilde{N_2} := N_2 + M_2$ for some $M_1, M_2 \in \mathbb{N}$. Then we define

$$\tilde{\theta} := \{\tilde{W}_0, \tilde{W}_1, \tilde{w}_2, \tilde{b}_0, \tilde{b}_1, b_2\} \in \mathbb{R}^{\tilde{P}} \,,$$

with

$$\tilde{W}_0 \in \mathbb{R}^{\tilde{N_1} \times N_0}, \tilde{W}_1 \in \mathbb{R}^{\tilde{N_2} \times \tilde{N_1}}, \tilde{w}_2 \in \mathbb{R}^{\tilde{N_2}} \,,$$
$$\tilde{b}_0 \in \mathbb{R}^{\tilde{N_1}}, \tilde{b}_1 \in \mathbb{R}^{\tilde{N_2}}, b_2 \in \mathbb{R} \,,$$

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

where for each $i \neq 0$, we append additional $M_{i-1}$ columns containing zeros to $W_i$. In other words, we have that $(W_i)_{jk} = 0$ if $k > N_{i-1}$. Furthermore, the (regularized) loss function is now $\tilde{\ell} : \mathbb{R}^{\tilde{P}} \to \mathbb{R}$.

We can confirm that, under this construction, for any $x$, we have that $f_{\tilde{\theta}}(x) = f_\theta(x)$. Moreover, for each $i \neq 2$ the gradient of the last $M_{i+1}$ rows of $\tilde{W}_i$ is zero. Thus, we can choose these particular components to "augment" the uncertainty of the classifier, without changing the prediction.

**Proposition 1** (Uncertainty Guarantee)**.** *Suppose $f_\theta, f_{\tilde{\theta}}$ are as defined above and last-layer Laplace approximations are used to obtain the respective posterior $\mathcal{N}(w_2 \mid \mu, \Sigma), \mathcal{N}(\tilde{w}_2 \mid \tilde{\mu}, \tilde{\Sigma})$. Then, for any input $x$ and any $M \in \mathbb{N}$ additional ReLUQ units, the approximate-Gaussian variance of $f_{\tilde{\theta}}(x)$ is strictly larger than that of $f_\theta(x)$.*

*Proof sketch.* The variance of $f_\theta(x)$ is $\phi(x)^\top \Sigma \phi(x)$. Via an eigendecomposition, it can be written as $\sum_i^P \lambda_i(\Sigma)(Q\phi(x))_i^2 =: v$. Using the same argument, the variance of $f_{\tilde{\theta}}(x)$ is given by $\sum_i^{\tilde{P}} \lambda_i(\tilde{\Sigma})(\tilde{Q}\tilde{\phi}(x))_i^2 =: \tilde{v}$ where $\tilde{P} > P$. Note that $\Sigma, \tilde{\Sigma}$ are SPD, thus the summands are all positive. Therefore $\tilde{v} > v$, because $\tilde{v}$ has more summand than $v$. □

### 1.1 Possible geometric interpretation

The graph of the loss $\ell$ on Euclidean space forms an embedded submanifold of codimension one, i.e. a hypersurface, in $\mathbb{R}^P \times \mathbb{R} \simeq \mathbb{R}^{P+1}$. The the graph of $\tilde{\ell}$ is a hypersurface in the higher dimensional space $\mathbb{R}^{\tilde{P}+1}$, but restricted to the subset $\mathbb{R}^{P+1} \subset \mathbb{R}^{\tilde{P}+1}$, it is exactly the loss landscape of $\ell$.

Suppose we use a Laplace approximation with posterior mean $\mu$ (a local mode of the loss landscape). For each $x$, the variance $v(x)$ of the prediction $f_\theta(x)$ can be approximated with the quadratic form $g(x)^\top H^{-1} g(x)$, where $g(x)$ is the gradient of $f_\theta$ w.r.t. $\theta$ at $\mu$ and $H$ is the Hessian of the loss w.r.t. $\theta$ at $\mu$. This means that, in the eigenbasis $Q$ of $H^{-1}$, we can write the variance as

$$v(x) \approx \sum_{i=1}^{P} \frac{1}{\lambda_i(H)} (Qg(x))_i^2 \,.$$

The key point here is that since $\mu$ is a stationary point, $\lambda_i(H)$ is the *principal curvatures*, since the Hessian matrix $H$ in this case is the same as the second fundamental form's matrix representation. The principal curvatures are extrinsic, thus depend on the embedding of the submanifold. We can we think that ReLUQ is a choice of embedding of the original loss landscape onto $\mathbb{R}^{\tilde{P}+1}$. The goal of this method is therefore to find such embedding, modifying the principal curvatures, s.t. the variance has the desired properties.

## 2 Introduction

Meaningful uncertainty estimates on the predictions of deep networks are important in mission-critical applications such as self-driving vehicles. Without them, deep networks are susceptible to out-of-distribution data since they are known to be overconfident: except around the decision boundary, their predictive confidence is arbitrarily close to one. Specifically for ReLU networks, this overconfidence problem has been studied recently by Nguyen et al. [2015], Hein et al. [2019]. It can be shown that far away from the training data, the prediction of ReLU networks converges to one. One can therefore cannot trust the confidence estimate that ReLU networks output. This is problematic since this means that these networks cannot even detect such obvious out-of-distribution points.

Recently, it has been shown that by employing Bayesian methods on ReLU networks, this issue can be mitigated. Specifically, Kristiadi et al. [2020] showed that applying Gaussian-approximated Bayesian methods onto pre-trained ReLU networks, e.g. via a simple Laplace approximation, could make their confidence asymptotically constant. However, their result does not necessarily imply that the confidence is asymptotically uniform, which is the ideal outcome. Furthermore, their bound depends on the spectrum of the covariance matrix of the approximate posterior. Thus, depending on the network, this bound can be loose.

In this paper, we propose a simple heuristic method to fix the aforementioned problems in Gaussian-approximated Bayesian ReLU networks. Our method is based on the tempering technique that is ubiquitously used in Bayesian inference [Wenzel et al., 2020, etc.]. However, instead of tempering the posterior with a constant temperature, we temper the *distribution over logits* with a temperature that depends on the input. We therefore call our method "dynamic predictive tempering" or DPT in short. We show that DPT yields (i) confidence arbitrarily close to $0.5$ far away from the training data nad (ii) uniform confidence in the limit. Furthermore, we show that, DPT does not change the underlying decision boundary of vanilla Bayesian prediction, thus maintain the original accuracy of the network.

We begin this paper with an exposition to the overconfidence problem of ReLU networks and how can Bayesian methods mitigate this in Section 3. In Section 4 we introduce DPT and analyze its properties. Related work will be discussed in Section 5 while empirical results in Section 6.

# 3 Bayesian ReLU networks

We call a function $f : \mathbb{R}^n \to \mathbb{R}^k$ piecewise affine if there is a decomposition of the input space $\mathbb{R}^n$ by a finite set of polytopes $\{P_1, \dots, P_r\}$ such that the restriction $f|_{P_i}$ is affine for each $i = 1 \dots r$. These polytopes are referred to as the linear regions of $f$. ReLU networks are neural networks that results in piecewise affine function on the input space [Arora et al., 2018]. This class of networks includes any neural network that is composed by fully-connected and convolution layers, along with ReLU or leaky-ReLU activation function and max- or average-pooling. Commonly, these networks are trained via a maximum a posteriori (MAP) estimation. Thus in the rest of this paper, unless stated otherwise, by ReLU networks we refer to MAP-trained ReLU networks.

ReLU networks are arguably the most widely used deep architecture due to its high accuracy. However, they are not suitable for answering any problem beyond decision problems that requires accurate uncertainty estimates, since they tend to be overconfident. Hein et al. [2019] showed that, given a ReLU network $f_\theta : \mathbb{R}^n \to \mathbb{R}^k$, for almost any input point $x \in \mathbb{R}^n$, as $\alpha > 0$ goes to infinity, the confidence

$$\max_{i=\{1,\dots,k\}} \operatorname{softmax}(f_\theta(\alpha x), i) := \max_{i=\{1,\dots,k\}} \frac{\exp f_\theta^i(\alpha x)}{\sum_{j=1}^k \exp f_\theta^j(\alpha x)}$$

of $\alpha x$ goes to one. In most realistic problems, even a finite $\alpha$ could already yield such obvious out-of-distribution data, for example in natural image classification, where the in-distribution data is fully contained in the unit box $[0, 1]^n \subset \mathbb{R}^n$. This problem shows that the uncertainty estimates of ReLU networks cannot be trusted and are useless for detecting out-of-distribution data far away from the training data.

In what follows, we focus on the binary classification case due to the lack of analytic approximation of the integral of softmax function w.r.t. a Gaussian measure. Note that, in this case, the confidence is defined by

$$\operatorname{conf}(x) := \max_{i\in\{0,1\}} \sigma(f_\theta(x)) = \sigma(|f_\theta(x)|)\,. \tag{1}$$

The second equality follows from Proposition 123 in the appendix. We will revisit the multi-class case in the empirical analysis.

Recently, Kristiadi et al. [2020] argued that this overconfidence problem is due to the usage of point estimate for making predictions in a ReLU network. Approximate Bayesian methods with Gaussian posterior are therefore proposed to mitigate this problem. Let $f_\theta : \mathbb{R}^n \to \mathbb{R}$ be a binary ReLU classifier, $p(y = 1|x, \theta) := \sigma(f_\theta(x))$ be the likelihood [1], and $p(\theta|\mathcal{D}) \approx \mathcal{N}(\theta|\mu, \Sigma)$ be the approximate posterior, obtained via e.g. a Laplace approximation or variational Bayes. Given an input $x$, the distribution over the logits $f_{\boldsymbol{\theta}}(x)$ can further be approximated via a linearization $f_\theta(x)$, seen as a function of $\theta$ [MacKay, 1995]. That is, if we let $g(x) := \nabla f_\theta(x)|_\mu$ to be the gradient of $f_\theta(x)$ w.r.t. $\theta$ at $\mu$, we have the following first-order Taylor's expansion:

$$f_\theta(x) \approx f_\mu(x) + g(x)^\top(\theta - \mu)\,,$$

and thus we have

$$p(f_\theta(x)|x, \mathcal{D}) \approx \mathcal{N}(f_\theta(x)|f_\mu(x), g(x)^\top \Sigma g(x))\,. \tag{2}$$

[1] $\sigma(z) := 1/(1 + \exp(-z))$ is the logistic function.

The predictive distribution is given by the following integral

$$p(y = 1|x, \mathcal{D}) = \int \sigma(f_\theta(x)) p(f_\theta(x)|x, \mathcal{D})\, df_\theta(x)\,, \tag{3}$$

which can be approximated via the convolution of probit function [MacKay, 1992], which gives us

$$p(y = 1|x, \mathcal{D}) \approx \sigma\left(\frac{f_\mu(x)}{(1 + \pi/8\, g(x)^\top \Sigma g(x))^{\frac{1}{2}}}\right) =: \sigma(z(x))\,. \tag{4}$$

Using this framework, Kristiadi et al. [2020] analytically showed that the marginalization in (3) gives for any input $x$, confidence that asymptotically converges to a constant. Furthermore, they show empirically that for popular ReLU networks like LeNets or ResNets, the confidence already converges in finite distance regime, i.e. when $\alpha < \infty$, in both binary and multi-class cases.

# 4 Dynamic tempering

The analysis of Kristiadi et al. [2020] reveals that, popular approximate Bayesian methods could mitigate the overconfidence problem of ReLU networks to *some degree* since their proofs disregard the distance between an input $x$ and the training dataset $\mathcal{D}$. Furthermore, the confidence is only asymptotically constant, depending on the minimum eigenvalue of the posterior covariance $\Sigma$. In some networks, it might be the case that this eigenvalue is too small such that the bound becomes loose. Based on this problem, we formulate the following two desiderata: (i) far away from the training data, the confidence should be $\epsilon$-close to $0.5$, and (ii) it should be uniform in the limit, not merely a constant. Fulfilling these desiderata—thus "fixing" Bayesian methods—is our principal focus in this section.

To that end, we propose a simple heuristic based on the tempering framework. However, whereas usually the tempering is done to the posterior $p(\theta|\mathcal{D})$ via some constant parameter $t$, we propose to temper the distribution over the logits (2) given an input $x$ with a "dynamic" parameter $t(x)$ that depends on $x$ instead. That is, we follow the construction of the Bayesian ReLU networks in the previous section, but we use the following tempered distribution instead of (2):

$$\begin{aligned} p_t(f_\theta(x)|x, \mathcal{D}) &:\propto p(f_\theta(x)|x, \mathcal{D})^{t(x)} \\ &\propto \exp\left(-\frac{1}{2\, g(x)^\top \Sigma g(x)} |f_\theta(x), f_\mu(x)|^2\right)^{t(x)} \\ &= \mathcal{N}(f_\theta(x)|f_\mu(x), t(x)^{-1} \operatorname{var}(f_\theta(x))) \end{aligned} \tag{5}$$

where $t$ is a function $\mathbb{R}^n \to \mathbb{R}$ and we have used $\operatorname{var}(f_\theta(x)) := g(x)^\top \Sigma g(x)$ (cf. (2)). It is easy to see that (4) now becomes

$$p_t(y = 1|x, \mathcal{D}) \approx \sigma\left(\frac{f_\mu(x)}{(1 + \pi/8\, t(x)^{-1}\, g(x)^\top \Sigma g(x))^{\frac{1}{2}}}\right) =: \sigma(z_t(x))\,. \tag{6}$$

Note that the difference between (3) and (6) amounts to the additional scalar term $1/t(x)$ in the denominator. One can then substitute $\sigma(f_\theta(x))$ in (1) with (6) to obtain the tempered Bayesian confidence:

$$\operatorname{conf}_t(x) := \sigma(|z_t(x)|)\,. \tag{7}$$

How can we define $t$ such that we fulfill our desiderata? Since we would like to capture the distance between an input to the training dataset, one option is to use the information given by the data density $p(x)$ approximated by the training data $\mathcal{D}$. Let $t(x) \propto p(x)$ be proportional to $p(x)$. Then, it is clear that the value $t(x)$ will be high if $x$ is close to $\mathcal{D}$, since $\mathcal{D}$ are assumed to be sampled from $p(x)$. Conversely, if $x$ is far away from $\mathcal{D}$, then $t(x)$ will be close to zero since it lies on the low-density region of $p(x)$.

More concretely, let us assume that the density $p(x)$ is approximated with the kernel density estimator (KDE) with a Gaussian kernel with length-scale $\ell$:

$$p(x) \approx \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\sqrt{2\pi\ell^2}} \exp\left(-\frac{1}{2\ell^2} d(x, x_i)^2\right)\,,$$

where $d(x, x_i) := \|x - x_i\|_2$ is the Euclidean metric. We can then define $t : \mathbb{R}^n \to (0, \infty)$ as

$$t(x) := \sum_{i=1}^{m} \exp\left(-\frac{1}{2\ell^2} d(x, x_i)^2\right), \tag{8}$$

since the only requirement is that $t \propto p$. This construction has the following properties.

**Lemma 1.** *Let $t$ be defined in* (8)*, and $x \in \mathbb{R}^n$ be arbitrary. Then, it holds that* $\lim_{\delta\to\infty} t(\delta x) = 0$.

*Proof.* First, we note that $\lim_{\delta\to\infty} d(\delta x, x_i)^2 = \infty$ for each $i = 1, \ldots, m$ and that the exponent is continuous everywhere. Therefore we can write

$$\begin{aligned}
\lim_{\delta\to\infty} t(\delta x) &= \lim_{\delta\to\infty} \sum_{i=1}^{m} \exp\left(-\frac{1}{2\ell^2} d(x, x_i)^2\right) \\
&= \sum_{i=1}^{m} \exp\left(-\frac{1}{2\ell^2} \lim_{\delta\to\infty} d(x, x_i)^2\right) \\
&= \sum_{i=1}^{m} \exp(-\infty) \\
&= \sum_{i=1}^{m} 0 \\
&= 0\,,
\end{aligned}$$

thus ending the proof. □

**Lemma 2** (Hein et al., 2019)**.** *Let $\{P_i\}_{i=1}^{r}$ be the set of linear regions associated to the binary ReLU classifier $f_\theta : \mathbb{R}^n \to \mathbb{R}$. For any $x \in \mathbb{R}^n$ there exists $\beta > 0$ and $j \in \{1, \ldots, r\}$ such that $\delta x \in P_j$ for all $\delta \geq \beta$. Furthermore, the restriction $f_\theta|_{P_j}$ of $f_\theta$ to $P_j$ can be written as the affine function:*

$$f_\theta|_{P_j}(\hat{x}) = w(\theta)^\top \hat{x} + b(\theta)\,,$$

*for all $\hat{x} \in P_j$, where $w(\theta) \in \mathbb{R}^n$ and $b(\theta) \in \mathbb{R}$ are some vector and scalar that depend on $\theta$.* □

We shall refer to the polytope $P_j$ given by Lemma 2 as the *outer polytope of $f_\theta$ w.r.t. $x$*. We are now ready to state our main theorem, which states that given a ReLU network $f_\theta$ and any test point $x$, at the corresponding outer polytope, the confidence is asymptotically low. Furthermore, the confidence is uniform at the limit.

The following assumes that we don't have biases. Need to show similar results with biases.

**Theorem 1.** *Let $f_\theta : \mathbb{R}^n \to \mathbb{R}$ be a binary ReLU classifier without bias vectors, $\mathcal{N}(\theta|\mu, \Sigma)$ be the posterior, and let $t : \mathbb{R}^n \to (0, \infty)$ be defined in* (8)*. Then for any test point $x \in \mathbb{R}^n$ and arbitrary $\epsilon > 0$, in the outer polytope of $f_\theta$ w.r.t. $x$, there exists $\delta$ such that if*

$$\min_{i=1,\ldots,m} d(\delta x, x_i) \geq \sqrt{-\ell^2 \, \log \frac{(-\log(\frac{1}{\epsilon+0.5} - 1))^2 \operatorname{var}(f_\theta(x))}{m \, |f_\mu(x)|^2}}\,,$$

*then it holds that*

$$\operatorname{conf}_t(\delta x) \leq \frac{1}{2} + \epsilon\,.$$

*Furthermore, in the limit with $\delta \to \infty$, the uniform confidence* $\operatorname{conf}(\delta x) = 0.5$ *is attained.*

*Proof.* By Lemma 2, at the outer polytope of $f_\theta$ w.r.t. $x$, say $P$, we can write the restriction of $f_\theta$ to $P$ as

$$f_\theta|_P(\delta x) = w(\theta)^\top(\delta x) = \delta w(\theta)^\top x = \delta f_\theta(x)\,,$$

where $\delta \geq \beta$. Furthermore the gradient of $f_\theta(\delta x)$ w.r.t. $\theta$ (evaluated at $\mu$) can be written as follows:

$$g(x) = \nabla f_\theta(\delta x) = J^\top(\delta x) = \delta J^\top x\,,$$

where the $n \times p$ matrix $J$ is the Jacobian $\partial w / \partial \theta$ at $\mu$. The term inside the logistic in (7) is therefore

$$\begin{aligned} |z(\delta x)| &= \frac{|\delta f_\mu(x)|}{(1 + t(\delta x)^{-1}\, \pi/8\, \delta^2 (J^\top x)^\top \Sigma (J^\top x))^{\frac{1}{2}}} \\ &\leq \frac{\delta\, |f_\mu(x)|}{\delta\, (t(\delta x)^{-1}\, \pi/8\, \mathrm{var}(f_\theta(x)))^{\frac{1}{2}}} \\ &= \frac{\sqrt{t(\delta x)} |f_\mu(x)|}{\sqrt{\mathrm{var}(f_\theta(x))}}\,. \end{aligned} \tag{9}$$

Let $d(\delta x, \mathcal{D}) := \min_{i=1,\dots,m} d(\delta x, x_i)$. Notice that

$$t(\delta x) = \sum_{i=1}^{m} \exp\left(-\frac{1}{\ell^2} d(\delta x, x_i)^2\right) \leq m \exp\left(-\frac{1}{\ell^2} d(\delta x, \mathcal{D})^2\right)\,.$$

Let $\alpha := -\log(\frac{1}{\epsilon + 0.5} - 1)$. Substituting the previous inequality into (9) and using the hypothesis on $d(\delta x, \mathcal{D})$ yields

$$\begin{aligned} |z(\delta x)| &\leq \frac{\sqrt{m \exp\left(-\frac{1}{\ell^2} d(\delta x, \mathcal{D})^2\right)} |f_\mu(x)|}{\sqrt{\mathrm{var}(f_\theta(x))}} \\ &\leq \frac{\sqrt{m \exp\left(-\frac{1}{\ell^2}(-\ell^2)\, \log \frac{\alpha^2\, \mathrm{var}(f_\theta(x))}{m\, |f_\mu(x)|^2}\right)} |f_\mu(x)|}{\sqrt{\mathrm{var}(f_\theta(x))}} \\ &= \frac{\sqrt{\frac{\alpha^2\, \mathrm{var}(f_\theta(x))}{|f_\mu(x)|^2}} |f_\mu(x)|}{\sqrt{\mathrm{var}(f_\theta(x))}} \\ &= \frac{\alpha \sqrt{\mathrm{var}(f_\theta(x))}\, |f_\mu(x)|}{|f_\mu(x)| \sqrt{\mathrm{var}(f_\theta(x))}} \\ &= \alpha\,. \end{aligned}$$

Since the logistic function is monotonically increasing, it holds that

$$\begin{aligned} \mathrm{conf}_t(\delta x) &= \sigma(|z(\delta x)|) \leq \sigma(\alpha) \\ &= \frac{1}{1 + \exp(-\alpha)} = \frac{1}{1 + \exp\left(\log(\frac{1}{\epsilon + 0.5} - 1)\right)} \\ &= \frac{1}{\frac{1}{\epsilon + 0.5}} = \frac{1}{2} + \epsilon\,, \end{aligned}$$

which is the first desired result.

For the second result, Lemma 1 tells us that $\lim_{\delta \to \infty} t(\delta x) = 0$ and thus $\lim_{\delta \to \infty} t(\delta x)^{-1} = \infty$. Therefore, by the definition of confidence (7) and the continuity of the logistic function at zero, it holds that

$$\begin{aligned} \lim_{\delta \to \infty} \mathrm{conf}_t(\delta x) &= \lim_{\delta \to \infty} \sigma\left(\frac{\delta |f_\mu(x)|}{(1 + t(\delta x)^{-1}\, \pi/8\, \delta^2 \mathrm{var}(f_\theta(x)))^{\frac{1}{2}}}\right) \\ &= \lim_{\delta \to \infty} \sigma\left(\frac{|f_\mu(x)|}{(\delta^{-2} + t(\delta x)^{-1}\, \pi/8\, \mathrm{var}(f_\theta(x)))^{\frac{1}{2}}}\right) \\ &= \sigma\left(\frac{|f_\mu(x)|}{(\lim_{\delta \to \infty} \delta^{-2} + \lim_{\delta \to \infty} t(\delta x)^{-1}\, \pi/8\, \mathrm{var}(f_\theta(x)))^{\frac{1}{2}}}\right) \\ &= \sigma\left(\frac{|f_\mu(x)|}{(\lim_{\delta \to \infty} t(\delta x)^{-1}\, \pi/8\, \mathrm{var}(f_\theta(x)))^{\frac{1}{2}}}\right) \\ &= \sigma(0) \\ &= \frac{1}{2}\,, \end{aligned}$$

which concludes the proof. □

For ReLU networks with bias, I can only say things in the limit. A fully rigorous proof for the asymptotic statement is surprisingly hard.

**Theorem 2.** *Let $f_\theta : \mathbb{R}^n \to \mathbb{R}$ be a binary ReLU classifier, $\mathcal{N}(\theta|\mu, \Sigma)$ be the posterior, and let $t : \mathbb{R}^n \to (0, \infty)$ be defined in* (8)*. For any test point $x \in \mathbb{R}^n$*

$$\lim_{\delta \to \infty} \operatorname{conf}_t(\delta x) = \frac{1}{2}\,.$$

*Proof.* By Lemma 2, at the outer polytope of $f_\theta$ w.r.t. $x$, say $P$, we can write the restriction of $f_\theta$ to $P$ as

$$f_\theta|_P(\delta x) = w(\theta)^\top(\delta x) + b(\theta) = \delta(w(\theta)^\top x + \delta^{-1}\, b(\theta))\,,$$

where $\delta \geq \beta$. Furthermore the gradient of $f_\theta(\delta x)$ w.r.t. $\theta$ (evaluated at $\mu$) can be written as follows:

$$g(x) = \nabla f_\theta|_P(\delta x) = J^\top(\delta x) + v = \delta(J^\top x + \delta^{-1} v)\,,$$

where the $n \times p$ matrix $J$ is the Jacobian $\partial w / \partial \theta$ at $\mu$ and $v \in \mathbb{R}^p$ is the gradient $\nabla b$ of $b$ w.r.t. $\theta$ at $\mu$. Now, Lemma 1 tells us that $\lim_{\delta\to\infty} t(\delta x) = 0$ and thus $\lim_{\delta\to\infty} t(\delta x)^{-1} = \infty$. Therefore, by the definition of confidence (7) and the continuity of the logistic function at zero, it holds that

$$\begin{aligned}
\lim_{\delta \to \infty} \operatorname{conf}_t(\delta x) &= \lim_{\delta \to \infty} \sigma\left(\frac{|f_\mu|_P(\delta x)|}{(1 + t(\delta x)^{-1}\, \pi/8\, g(\delta x)^\top \Sigma g(\delta x))^{\frac{1}{2}}}\right) \\
&= \lim_{\delta \to \infty} \sigma\left(\frac{|\delta(w(\theta)^\top x + \delta^{-1}\, b(\theta))|}{(1 + t(\delta x)^{-1}\, \pi/8\, (\delta(J^\top x + \delta^{-1} v))^\top \Sigma(\delta(J^\top x + \delta^{-1} v)))^{\frac{1}{2}}}\right) \\
&= \lim_{\delta \to \infty} \sigma\left(\frac{\delta|w(\theta)^\top x + \delta^{-1}\, b(\theta)|}{(1 + t(\delta x)^{-1}\, \pi/8\, \delta^2\, (J^\top x + \delta^{-1} v)^\top \Sigma(J^\top x + \delta^{-1} v))^{\frac{1}{2}}}\right) \\
&= \lim_{\delta \to \infty} \sigma\left(\frac{|w(\theta)^\top x + \delta^{-1}\, b(\theta)|}{(\delta^{-2} + t(\delta x)^{-1}\, \pi/8\, (J^\top x + \delta^{-1} v)^\top \Sigma(J^\top x + \delta^{-1} v))^{\frac{1}{2}}}\right) \\
&= \sigma\left(\frac{\lim_{\delta\to\infty} |w(\theta)^\top x + \delta^{-1}\, b(\theta)|}{(\lim_{\delta\to\infty} \delta^{-2} + \lim_{\delta\to\infty} t(\delta x)^{-1}\, \pi/8\, (J^\top x + \delta^{-1} v)^\top \Sigma(J^\top x + \delta^{-1} v))^{\frac{1}{2}}}\right) \\
&= \sigma\left(\frac{|w(\theta)^\top x|}{(\lim_{\delta\to\infty} t(\delta x)^{-1}\, \pi/8\, (J^\top x + \delta^{-1} v)^\top \Sigma(J^\top x + \delta^{-1} v))^{\frac{1}{2}}}\right) \\
&= \sigma(0) \\
&= \frac{1}{2}\,,
\end{aligned}$$

which concludes the proof. □

**Proposition 2.** *Let $f_\theta : \mathbb{R}^n \to \mathbb{R}$ be a binary ReLU classifier, $\mathcal{N}(\theta|\mu, \Sigma)$ be the posterior, and let $t : \mathbb{R}^n \to (0, \infty)$ be defined in* (8)*. Let the functions $z$ and $z_t$ be defined as in* (4) *and* (6)*, respectively. Then for any $x \in \mathbb{R}^n$,*

$$\sigma(z_t(x)) = 0.5 \iff \sigma(z(x)) = 0.5 \tag{10}$$

*Proof.* It suffices to show that $z_t(x) = 0 \iff z(x) = 0$ since $\sigma^{-1}(0.5) = 0$. The keys of this proof are the observation that the codomain of $t$ is the set of positive real numbers, along with the fact that $z_t$ is just $z$ with the additional term $1/t(x)$ in its denominator. In particular, notice that the numerator of $z$ and $z_t$ is the same.

($\Longrightarrow$) Suppose $z_t(x) = 0$. The denominator of $z_t$ is positive since $\pi/8 > 0$, the matrix $\Sigma$ is positive definite, and $t(x) \in (0, \infty)$ by definition. Thus its numerator must be zero, which implies that $z(x)$ must also be zero.

($\Longleftarrow$) Suppose $z(x)$ is zero. Using the same argument, the denominator of $z$ is positive and thus its numerator must be zero, which implies that $z_t(x)$ must also be zero. □

## 5 Related work

## 6 Experiments

## 7 Conclusion

# Broader Impact

Authors are required to include a statement of the broader impact of their work, including its ethical aspects and future societal consequences. Authors should discuss both positive and negative outcomes, if any. For instance, authors should discuss a) who may benefit from this research, b) who may be put at disadvantage from this research, c) what are the consequences of failure of the system, and d) whether the task/method leverages biases in the data. If authors believe this is not applicable to them, authors can simply state this.

Use unnumbered first level headings for this section, which should go at the end of the paper. **Note that this section does not count towards the eight pages of content that are allowed.**

# References


Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, 2019.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks. *arXiv preprint arXiv:2002.10118*, 2020.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How Good is the Bayes Posterior in Deep Neural Networks Really? *arXiv preprint arXiv:2002.02405*, 2020.

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *ICLR*, 2018.

David JC MacKay. Probable networks and plausible predictions—a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 1995.

David JC MacKay. The evidence framework applied to classification networks. *Neural computation*, 1992.