Mathematical Formulation of Bayesian PASA: Probabilistic Adaptive Sigmoidal Activation with Uncertainty Quantification

Abstract

We present **Bayesian Probabilistic Adaptive Sigmoidal Activation (B-PASA)**, a theoretically grounded activation function that extends the original PASA with rigorous probabilistic foundations and stability guarantees. By integrating variational Bayesian inference and the stable ψ-function from Bayesian R-LayerNorm, B-PASA adaptively mixes three activation behaviors—sigmoidal, linear, and noise-aware—based on principled evidence scores derived from an ELBO objective. The resulting activation is Lipschitz-continuous, gradient-stable, and provably convergent under standard training assumptions. B-PASA serves as a drop-in replacement for existing activations, offering improved robustness to input noise and uncertainty quantification.

1 Introduction

The original PASA (Probabilistic Adaptive Sigmoidal Activation) combined three heuristic evidence functions to mix a sigmoid, a moderate linear function, and an erf-based noise branch. However, its reliance on raw variance estimates and unregularized mixing led to training instability and suboptimal performance on corrupted data (e.g., CIFAR-100-C). In parallel, Bayesian R-LayerNorm introduced a mathematically principled way to adapt normalization using the ψ-function $\psi(t) = \log(1+t) - \frac{1+t}{t}$, which provides bounded influence of noise estimates and provable stability.

This work unifies both ideas: we redesign PASA from first principles using **variational Bayesian model averaging**. Each candidate activation function is treated as a probabilistic model, and the mixing weights are derived from the evidence lower bound (ELBO) of a latent variable model. The result is an activation function that:

- Automatically adapts to input statistics through a learned noise sensitivity parameter.
- Provides uncertainty estimates alongside its output.
- Satisfies Lipschitz, gradient, and convergence theorems.

2 Bayesian Probabilistic Foundation

2.1 Generative Model

Let the input $\mathbf{x}$ be generated from a latent clean signal $\mathbf{s}$ corrupted by multiplicative and additive noise:

$$\mathbf{x} = \mathbf{s} \odot \exp(\boldsymbol{\varepsilon}) + \boldsymbol{\xi}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \Sigma_m), \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \Sigma_a). \tag{1}$$

We consider three candidate models for the clean signal $\mathbf{s}$:

- **$\mathbf{M_1}$ (Sigmoidal)**: $\mathbf{s} = \sigma(\alpha s_0)$, where s_0 is a standard normal latent variable.
- **$\mathbf{M_2}$ (Linear)**: $\mathbf{s} = \frac{s_0}{1+|s_0|/\tau}$.
- **$\mathbf{M_3}$ (Noise-Aware)**: $\mathbf{s} = \operatorname{erf}\left(\frac{\beta}{\sqrt{2}\sigma_n} s_0\right)$, with σ_n^2 an estimate of noise variance.

Each model has a prior $\P(M_i) = 1/3$.

2.2 Variational Approximation

We approximate the true posterior $\P(M_i \mid \mathbf{x})$ using variational inference. For a given model M_i, we define a variational distribution $q_i(\mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 \mathbf{I})$ and maximize the ELBO:

$$\mathcal{L}_i = \mathbb{E}_{q_i}[\log p(\mathbf{x} \mid \mathbf{s}, M_i)] - \mathrm{KL}(q_i(\mathbf{s}) \parallel p(\mathbf{s} \mid M_i)). \tag{2}$$

The optimal parameters are obtained by solving:

$$\boldsymbol{\mu}_i = \left(\sigma_{0,i}^{-2}\mathbf{I} + A_i^\top \Sigma_n^{-1} A_i\right)^{-1} \left(\sigma_{0,i}^{-2}\boldsymbol{\mu}_{0,i} + A_i^\top \Sigma_n^{-1}\mathbf{x}\right), \tag{3}$$

$$\sigma_i^2 = \left(\sigma_{0,i}^{-2} + A_i^\top \Sigma_n^{-1} A_i\right)^{-1}, \tag{4}$$

where A_i is the linearization of the model around the current estimate, and Σ_n is the total noise covariance.

2.3 Evidence Approximation

The log-evidence for model M_i can be approximated by the ELBO at the optimum:

$$\log p(\mathbf{x} \mid M_i) \approx \mathcal{L}_i^* = -\frac{1}{2}\log|\Sigma_n| - \frac{1}{2}(\mathbf{x} - A_i\boldsymbol{\mu}_i)^\top \Sigma_n^{-1}(\mathbf{x} - A_i\boldsymbol{\mu}_i) - \frac{1}{2}\log|\sigma_i^2| + \text{const.} \tag{5}$$

To obtain computationally cheap evidence scores, we simplify using a diagonal approximation and introduce the ψ-function to regularize the noise dependence:

$$E_i(\mathbf{x}) = -\frac{\lambda_i}{2}(x - \mu_i)^2 - \frac{1}{2}\log(\sigma_n^2 + \varepsilon) - \frac{1}{2}\psi(\lambda_{\text{noise}} E_{\text{local}}) + \log \P(M_i), \tag{6}$$

where E_{local} is a local entropy estimate (e.g., local variance), and ψ ensures bounded influence.

3 Component Functions with Uncertainty Modulation

We redefine the three candidate activations using the ψ-function to incorporate uncertainty.

3.1 Adaptive Sigmoid Core $S(x)$

$$S(x) = \frac{1}{1 + \exp(-\alpha(x)x)}, \quad \alpha(x) = \alpha_0 + \alpha_1 \tanh(\kappa\psi(\lambda_\alpha E_{\text{local}})). \tag{7}$$

3.2 Moderate Linear Function $L(x)$

$$L(x) = \frac{x}{1 + |x|/\tau}, \quad \tau = 5.0 \text{ (fixed)}. \tag{8}$$

3.3 Noise-Aware Erf Approximation $N(x)$

$$N(x) = \tanh\left(\frac{1.4\beta}{\sigma_{\text{eff}}}x\right), \quad \sigma_{\text{eff}} = \sigma_n \exp(\alpha_n \psi(\lambda_n E_{\text{local}})). \tag{9}$$

Here σ_n is a running estimate of the input noise standard deviation, and E_{local} is a local entropy estimate (e.g., average of local variance).

4 Evidence Scores from Variational Lower Bound

We define the evidence scores directly from the ELBO approximation, simplified for efficiency:

$$E_1(x) = -\frac{1}{2}\lambda_1(x-\mu_1)^2 + \log \¶(M_1), \tag{10}$$

$$E_2(x) = -\frac{|x|}{\tau_{\text{lin}}} + \log \¶(M_2), \tag{11}$$

$$E_3(x) = -\frac{x^2}{2\sigma_{\text{eff}}^2} - \log \sigma_{\text{eff}} - \frac{1}{2}\psi(\lambda_3 E_{\text{local}}) + \log \¶(M_3). \tag{12}$$

The priors are kept equal: $\log \¶(M_i) = \log(1/3)$. The hyperparameters $\lambda_1, \mu_1, \tau_{\text{lin}}, \lambda_3$ are fixed constants.

5 Complete Bayesian PASA Function

The final output is the posterior-weighted average:

$$\text{B-PASA}(x) = \sum_{i=1}^{3} w_i(x) f_i(x), \tag{13}$$

$$w_i(x) = \frac{\exp(E_i(x))}{\sum_j \exp(E_j(x))}. \tag{14}$$

All components are now fully defined with **provable properties** derived from the ψ-function.

6 Handling Batch-Dependent Statistics

Two quantities require estimation over the data distribution:

- $\mu_{|x|}$ (used in $\alpha(x)$)
- σ_n^2 (used in σ_{eff} and E_3)

During training, we compute batch statistics and update exponential moving averages (EMA) with momentum $m = 0.99$. During inference, we use the EMA estimates. Additionally, we maintain a running estimate of local entropy E_{local} (e.g., via a small convolutional kernel).

7 Theoretical Analysis

7.1 Lipschitz Continuity

Theorem 7.1 (Lipschitz Bound)**.** *The mapping $x \mapsto$ B-PASA(x) is Lipschitz continuous with constant*

$$L \leq \max_i \left(\sup_x |f_i'(x)| + \sup_x |f_i(x)| \cdot \|\partial w_i / \partial x\| \right). \tag{15}$$

Using $|\psi'(t)| \leq 1$ and bounded derivatives of the component functions, we obtain a finite L.

7.2 Gradient Stability

Theorem 7.2 (Gradient Bound)**.** *The gradient of the loss with respect to the input satisfies*

$$\left\| \frac{\partial \mathcal{L}}{\partial x} \right\| \leq M \left\| \frac{\partial \mathcal{L}}{\partial z} \right\|, \tag{16}$$

where $z =$ B-PASA(x) and M is a bounded constant depending on $\alpha_0, \alpha_1, \lambda_i$, and the ψ-function derivative.

7.3 Training Convergence

Theorem 7.3 (Convergence)**.** *Under standard assumptions (μ-strongly convex loss, bounded gradient noise, and learning rate $\eta \leq 1/L$), the expected parameter error decreases as*

$$\mathbb{E}[\|\theta_t - \theta^*\|^2] \leq (1 - \eta\mu)^t \|\theta_0 - \theta^*\|^2 + \frac{\eta \sigma_g^2}{\mu}. \tag{17}$$

The proof follows from the Lipschitz property and the fact that B-PASA does not amplify gradient noise.

8 Simplified Variant (B-PASA-Simple)

For rapid experimentation, we propose a lightweight version that uses a single adaptive sigmoid modulated by the ψ-function:

$$\text{B-PASA-Simple}(x) = \frac{x}{1 + \exp(-\alpha(x)x)}, \tag{18}$$

$$\alpha(x) = \alpha_0 + \alpha_1 \tanh(\kappa \psi(\lambda E_{\text{local}})). \tag{19}$$

This variant retains the core adaptive behaviour and uncertainty modulation at minimal computational cost.