Laplpace Bridge

Ignashin Igor

Bayesian multimodeling
Department of Intelligent Systems, MIPT

December 2024

Laplace approximation

$$p(\theta|D) := \frac{1}{Z} p(D|\theta)p(\theta) \quad \}$$

$$p(D|\theta)p(\theta) =: h(\theta) \quad \}$$

$$Z = \int \exp(\log h(\theta)) d\theta \quad \} = \int p(D|\theta)p(\theta)d\theta$$

$$\theta_{\mathsf{MAP}} := \arg\max_{\theta} \underbrace{\log p(\theta|D)}_{} = \arg\max_{\theta} \log h(\theta)$$

$$\log h(\theta) \approx \log h(\theta_{\mathsf{MAP}}) - \frac{1}{2}(\theta - \theta_{\mathsf{MAP}})^{\top}\Lambda(\theta - \theta_{\mathsf{MAP}}) \quad \} \quad \theta_{map}$$

$$\Lambda := -\nabla^2 \log h(\theta)\big|_{\theta=\theta_{\mathsf{MAP}}}$$

$$Z \approx \exp(\log h(\theta_{\mathsf{MAP}})) \int \exp\left(-\frac{1}{2}(\theta - \underbrace{\theta_{\mathsf{MAP}})^{\top}\Lambda(\theta - \theta_{\mathsf{MAP}}}_{})\right) d\theta =$$

$$= h(\theta_{\mathsf{MAP}}) \left(\frac{(2\pi)^{d/2}}{(\det \Lambda)^{1/2}}\right)$$

$$p(\theta|D) = \frac{1}{Z} h(\theta) \approx \frac{(\det \Lambda)^{-1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(\theta - \theta_{\mathsf{MAP}})^{\top}\Lambda(\theta - \theta_{\mathsf{MAP}})\right), \quad \}$$

Which we can immediately identify as the Gaussian density $\mathcal{N}(\theta|\theta_{\mathsf{MAP}}, \Sigma)$ with mean θ_{MAP} and covariance matrix $\Sigma := \Lambda^{-1}$.

Dirichle distribution

The Dirichlet distribution, which has the density function:

$$\mathrm{Dir}(\pi|\alpha) := \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\underbrace{\overbrace{\alpha_k}^{<1}-1}_{\wedge\, 0}}, \quad \nwarrow \pi_k = 0 \tag{1}$$

Is defined on the probability simplex and can be “multimodal” in the sense that the distribution diverges in the k-corner of the simplex when $\alpha_k < 1$.

This implies the incorrectness of applying the Laplace approximation to this distribution, because the approximation is unimodal.

Change of variable

However, MacKay [1998] noted that both can be elegantly corrected by changing the variable.
Consider the K-dimensional variable $\pi \sim \text{Dir}(\pi|\alpha)$ defined as the softmax of $z \in \mathbb{R}^K$:

$$z \longrightarrow \underbrace{\pi}_{} \sim \text{Dir}(\pi|\alpha)$$

$$\underbrace{\pi_k(z) := \frac{\exp(z_k)}{\sum_{l=1}^{K} \exp(z_l)},}_{} \quad \text{for all } k = 1, \ldots, K.$$

We will call z the logit of π. When expressed as a function of z, the density of the Dirichlet in π has to be multiplied by the absolute value of the determinant of the Jacobian:

$$z = \text{softmax}^{-1}(\pi)$$

$$J = \det\left(\frac{\partial \pi}{\partial z}\right) = \underbrace{\prod_k \pi_k(z_k).}_{}$$

Laplace Bridge

After multiplying by the Jacobian:

$\pi \longrightarrow z$

$\pi = \pi(z)$

$\uparrow \quad \uparrow$ softmax

$$p(z) = \mathrm{Dir}_z(\pi(z)|\alpha) := \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k(z)^{\alpha_k}.$$

This density of z, the Dirichlet distribution in the softmax basis, can now be accurately approximated by a Gaussian through a Laplace approximation.

Analytic map from the parameter $\alpha \in \mathbb{R}_+^K$ to the parameters of the Gaussian ($\mu \in \mathbb{R}^K$ and symmetric positive definite $\Sigma \in \mathbb{R}^{K\times K}$), given by:

$\pi \longleftarrow z$

$\uparrow \mathrm{Dir}(\alpha) \quad \nwarrow \mathcal{N}(\mu, \Sigma)$

$$\left\{\begin{aligned} \mu_k &= \log \alpha_k - \frac{1}{K} \sum_{l=1}^{K} \log \alpha_l \\ \Sigma_{k\ell} &= \delta_{k\ell} \left(\frac{1}{\alpha_k} - \frac{1}{K} \left(\frac{1}{\alpha_k} + \frac{1}{\alpha_\ell} - \frac{1}{K} \sum_{u=1}^{K} \frac{1}{\alpha_u} \right) \right). \end{aligned}\right.$$

$\pi \longrightarrow z$

$\mathrm{Dir}(\alpha) \quad \mathcal{N}(\mu, \Sigma)$

Inverse map

$$\mathrm{Dir}(\alpha) \rightleftarrows \mathcal{N}(\mu,\Sigma) \qquad z \sim \mathcal{N}(\mu,\Sigma) \xrightarrow{\text{softmax}} \pi$$

A pseudo-inverse of this map was provided as a side result in Hennig et al. [2012]. It maps the Gaussian parameters to those of the Dirichlet as

$$\alpha_k = \frac{1}{\Sigma_{kk}} \left(1 - \frac{2}{K} + e^{\mu_k} \sum_{l=1}^{K} e^{-\mu_l} \right) \tag{1}$$

This equation ignores off-diagonal elements of Σ.

This and the previous 2 equations will be called the Laplace Bridge.

For Bayesian deep learning, only this equation is used, which displays values from μ, Σ to α.

Despite the fact that LB implies a decrease in the expressiveness of the distribution, however, the display is still quite accurate.

Visualization Laplace bridge

$z = \ln \frac{\pi}{1-\pi}$

$\pi \rightarrow [z, 1]$

$\mathcal{N}(\mu, \sigma) \rightarrow \mathrm{Dir}(\alpha)$

$\alpha =$

$z \rightarrow \pi$

$\nabla^2 \log P(\ )$

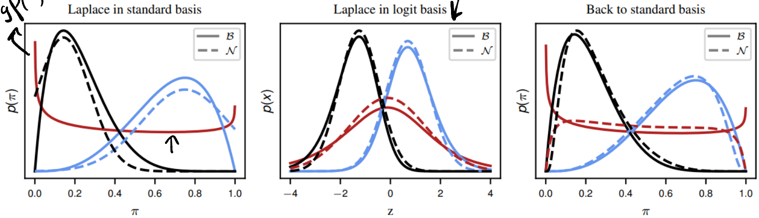


Figure 2: (Adapted from Hennig et al. [2012]). Visualization of the Laplace Bridge for the Beta distribution (1D special case of the Dirichlet) for three sets of parameters. **Left:** "Generic" Laplace approximations of standard Beta distributions by Gaussians. Note that the Beta Distribution (red) does not have a valid approximation because its Hessian is not positive semi-definite. **Middle:** Laplace approximation to the same distributions after basis transformation through the softmax (4). The transformation makes the distributions "more Gaussian" (i.e. uni-modal, bell-shaped, with support on the real line), thus making the Laplace approximation more accurate. **Right:** The same Beta distributions, with the back-transformation of the Laplace approximations from the middle figure to the simplex, yielding an improved approximate distribution. In contrast to the left-most image, the dashed lines now actually are probability densities (they integrate to 1 on the simplex).

Softmax in NN

$$\begin{tikzcd}
 & \sim \mathcal{N}(\mu,\Sigma) & \mathrm{Dir}(\alpha) \arrow[d] \\
\theta \arrow[r] & z(\theta) \arrow[r] & \pi(\theta)
\end{tikzcd}$$

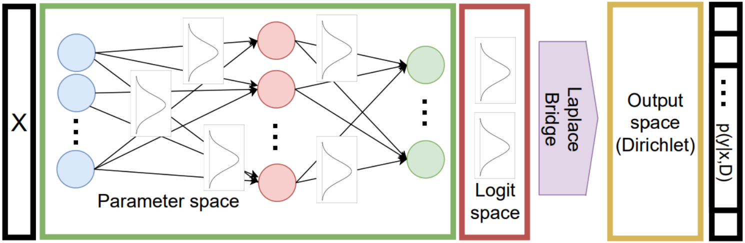


Figure 1: High-level sketch of the Laplace Bridge for BNNs. $p(y|x, D)$ denotes the marginalized softmax output, i.e. the mean of the Dirichlet.

Laplace Approximation in BNN

$$z = \underbrace{W_L \cdot \underbrace{\phi(x)}}\rightarrow \pi$$

The Laplace Bridge can be applied to any NN setup that maps from a Gaussian to probabilities by using the softmax.
Consider a last-layer Laplace approximation of the network

$$q(z|x) \approx \mathcal{N}\left(z\middle|\mu_{W(L)}\phi(x), \phi(x)^T\Sigma_{W(L)}\phi(x)\right), \qquad (8)$$

where $\phi(x)$ the output of the first $L-1$ layers, $\mu_{W(l)}$ is the maximum a posteriori (MAP) estimate for the weights of the last layer, $\Sigma_{W(l)}$ is the inverse of the negative loss Hessian w.r.t. $W(l)$, given by $\Sigma_{W(L)} = -(\nabla^2_{W(L)}L)^{-1}$ around the MAP estimate $W(L)$.

Laplace Bridge in BNN

Given a dataset $D := \{(x_i, t_i)\}_{i=1}^{D}$ and a prior $p(\theta)$, let the posterior over the parameter θ of an L-layer network f_θ be defined as:

$$\underbrace{p(\theta|D)}_{} \propto p(\theta)p(D|\theta) = \underbrace{p(\theta)}_{\mathcal{N}(0,\sigma^2 I)} \prod_{(x,t)\in D} \underbrace{p(y = t|\theta, x)}_{\text{Cat}}, \tag{40}$$

Then we can get an approximation of the posterior $p(\theta|D)$ by fitting a Gaussian $\mathcal{N}(\theta|\mu_\theta, \Sigma_\theta)$ where:

$$\mu_\theta = \theta_{\text{MAP}},$$

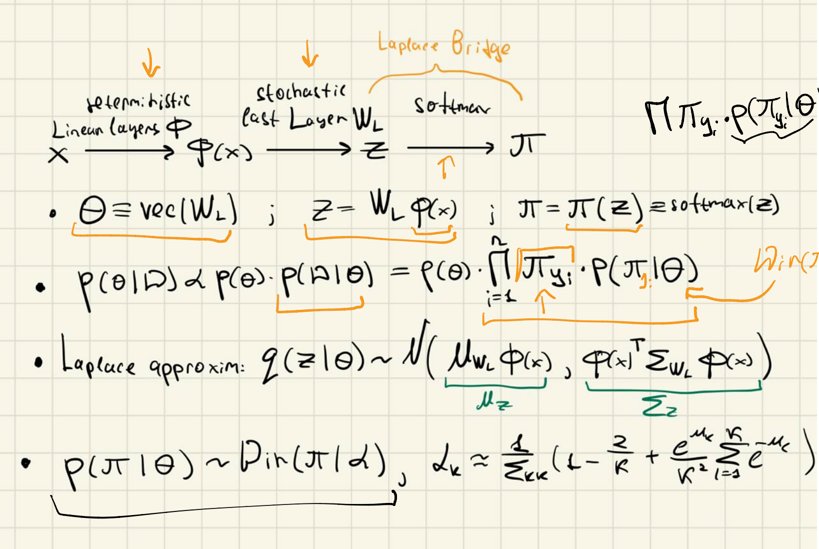
$$\Sigma_\theta = \left(-\nabla^2_{|\theta_{\text{MAP}}} \log p(\theta|D)\right)^{-1} =: H_\theta^{-1}.$$

That is, we fit a Gaussian centered at the mode θ_{MAP} of $p(\theta|D)$ with the covariance determined by the curvature at that point.

For example, prior $p(\theta)$ is a zero-mean isotropic Gaussian $\mathcal{N}(\theta|0, \sigma^2 I)$ and the likelihood function is the Categorical density:

$$p(D|\theta) = \prod_{(x,t)\in D} \text{Cat}(y = t|\text{softmax}(f_\theta(x))).$$

Notes



$$x \xrightarrow[\text{Linear layers } \phi]{\text{deterministic}} \phi(x) \xrightarrow[\text{last Layer } W_L]{\text{stochastic}} z \overbrace{\xrightarrow{\text{softmax}}}^{\text{Laplace Bridge}} \pi \qquad \prod \pi_{y_i} \cdot \underbrace{p(\pi_{y_i}|\theta)}$$

- $\underbrace{\theta \equiv \mathrm{vec}(W_L)}$; $\underbrace{z = W_L \underbrace{\phi(x)}}$; $\pi = \underbrace{\pi(z)} = \mathrm{softmax}(z)$
- $p(\theta|D) \propto p(\theta) \cdot \underbrace{p(D|\theta)} = p(\theta) \cdot \underbrace{\prod_{i=1}^{n} \boxed{\pi_{y_i}} \cdot p(\pi_{y_i}|\theta)}$ $\mathrm{Dir}(\pi|\alpha)$
- Laplace approxim: $q(z|\theta) \sim \mathcal{N}\left(\underbrace{\mu_{W_L}\phi(x)}_{\mu_z}, \underbrace{\phi(x)^T \Sigma_{W_L} \phi(x)}_{\Sigma_z}\right)$
- $\underbrace{p(\pi|\theta) \sim \mathrm{Dir}(\pi|\alpha)}$, $\alpha_k \approx \frac{1}{\Sigma_{kk}}\left(1 - \frac{2}{K} + \frac{e^{\mu_k}}{K^2}\sum_{l=1}^{K} e^{-\mu_l}\right)$

Limitations of LB

$\underset{\downarrow}{\mu, \Sigma}$ $\quad \underbrace{\mathcal{N}(\quad)} \rightarrow \underbrace{Dir(\quad)}$

Two limitations of the LB:

- First, the LB assumes that the random variable of the Gaussian sums to zero due to the difference in degrees of freedom between Dirichlet and Gaussian

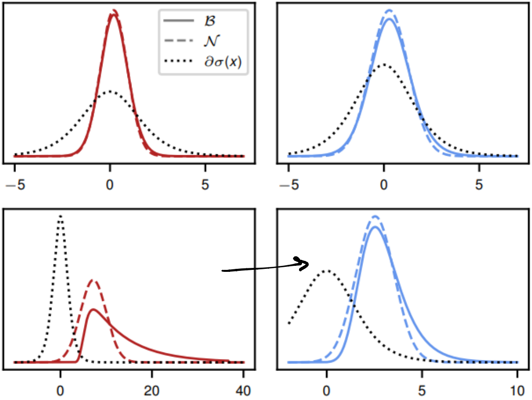
Thus, we have to add a correction that projects from any arbitrary Gaussian to one that fulfills this constraint.

$$\mathcal{N}\left(x \middle| \mu - \frac{\Sigma_{11}^T \mu}{1^T \Sigma_1}, \Sigma - \frac{\Sigma_{11} \Sigma_1^T}{1^T \Sigma_1}\right) \Big\} \; \bar{\mu}, \bar{\Sigma}$$

where 1 is the one-vector of size K.

Limitations of LB

- Second, the softmax-Dirichlet distribution is asymmetric.



Second correction

Autors propose an additional correction for practical purposes:

$$c = v_{\mathrm{mean}}(\Sigma) \cdot \frac{1}{\sqrt{K/2}}$$

$$\mu_0 = \frac{\mu}{\sqrt{c}}$$

$$\Sigma_0 = \frac{\Sigma}{c}$$

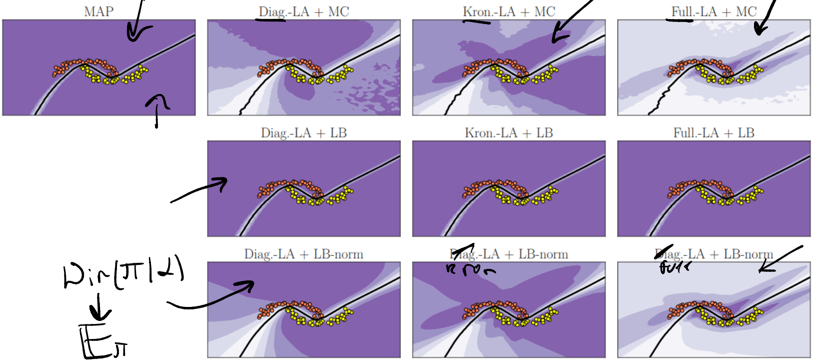
where $v_{\mathrm{mean}}(\Sigma)$ denotes the mean variance of Σ, defined as $v_{\mathrm{mean}}(\Sigma) = \sum_i \Sigma_{ii}$.
The factor of $\sqrt{\frac{1}{K/2}}$ is added because authors founds that higher dimensionalities require less correction.
This normalization, which can be understood as "pulling back" the distribution into a space where it is symmetric, has higher approximation quality.
This correction is applied after the zero-sum constraint correction.

Experiments

θ_{map}

LA

$\theta \sim \mathcal{N}(\mu, \Sigma) \longrightarrow \pi_\theta(x,y) \quad \mathbb{E}_\theta \, \pi_\theta(x,y)$



$\mathrm{Dir}(\pi \mid \alpha)$

$\downarrow$

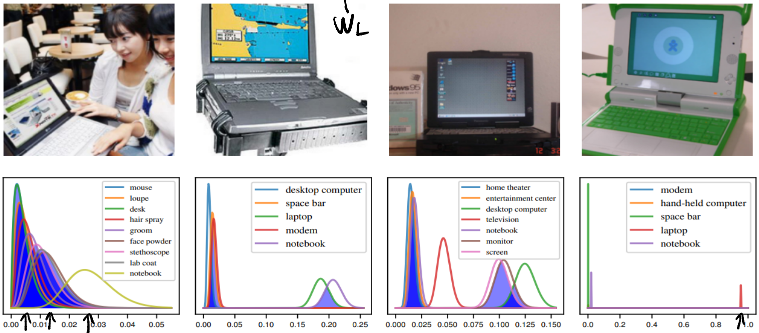
$\mathbb{E}_\pi$

Figure 3: **Left column:** vanilla MAP estimate which is overconfident. **Top row:** mean of softmax applied to Gaussian samples. **Middle row:** mean of the vanilla LB. **Bottom row:** mean of the corrected LB. The vanilla LB yields overconfident prediction far from the data. Our proposed correction fixes this issue, making the LB's approximation close to MC.

Inference

$$x \rightarrow \underset{p(y|\pi)\cdot p(\pi|\theta,x)}{P(y|x,\theta)} \nearrow Dir(\pi|\alpha) \qquad \underset{\uparrow W_L}{\theta} \sim N(\mu,\Sigma) \Rightarrow z \sim N(\mu_z,\Sigma_z) \Rightarrow \pi \sim Dir(\pi|\alpha)$$

$\alpha(\mu_\Sigma)$



laptop $\pi = 0{,}9$

Figure 6: **Upper row:** images from the "laptop" class of ImageNet. **Bottom row:** Beta marginals of the top-k predictions for the respective image. In the first column, the overlap between the marginal of all classes is large, signifying high uncertainty, i.e. the prediction is "do not know". In the second column, "notebook" and "laptop" have confident, yet overlapping marginal densities and therefore yield a top-2 prediction: "either notebook or laptop". In the third column "desktop computer", "screen" and "monitor" have overlapping marginal densities, yielding a top-3 estimate. The last case shows a top-1 estimate: the network is confident that "laptop" is the only correct label.

Conclusion

- Laplace Bridge analytically maps the marginal Gaussian prediction on logits to the Dirichlet distribution over softmax vectors.
- Preservation of predictive uncertainty.
- Significantly reduces the cost of predicting the posterior distribution in the testing phase while minimizing the cost increase in the training phase.
- The vanilla Laplace Bridge method has some limitations, for which the author has proposed a simple correction that outperforms alternative approximations of the softmax integral.