SAND2022-3773C

Robust Initialization of Variational Inference through Global Optimization and Laplace Approximations

Wyatt Bridgman [1] Mohammad Khalil [1]

Abstract

Variational inference (VI) with a mean-field approximation can sometimes be too restrictive but VI with high-fidelity surrogate posteriors, such as Gaussian mixture models (GMMs) can be computationally prohibitive due to the increase in number of optimization parameters. We propose a strategy for constructing a GMM approximation to an intractable Bayesian posterior using global optimization and Laplace approximations. We show how this can be used as an efficient initialization strategy for VI or as an alternative approximation method.

Introduction

A frequent problem arising in statistical modeling is the analysis of intractable density kernels. These are positive functions $\phi(\mathbf{x})$ such that the constant $Z = \int \phi(\mathbf{x})\,\mathrm{d}\mathbf{x}$ cannot be computed to obtain a normalized probability density $p(\mathbf{x}) = \phi(\mathbf{x})/Z$. Such intractable kernels are encountered in Bayesian inference where a distribution over latent variables $\mathbf{z}$ is inferred from observed variables $\mathbf{x}$ through a joint distribution $p(\mathbf{x}, \mathbf{z})$. The posterior distribution over latent variables is given by Bayes rule

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})}{\int p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}}$$

where the marginal likelihood $\int p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}$ defines the normalization constant that cannot be evaluated in closed form. An alternative to widely-used sampling-based approaches is to construct an approximation of an intractable density from a simpler parametric family. Variational inference seeks an approximation $q_{\boldsymbol{\theta}}(\mathbf{z}) \in \mathcal{F}_{\boldsymbol{\theta}}$ in some parametric family $\mathcal{F}_{\boldsymbol{\theta}}$ by minimizing an error measure with respect to $\boldsymbol{\theta}$ such as Kullback-Liebler (KL) divergence

$$q_{\boldsymbol{\theta}}(\mathbf{z}) = \min_{q_{\boldsymbol{\theta}} \in \mathcal{F}_{\boldsymbol{\theta}}} D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x}))$$

This recasts approximate inference as an optimization problem allowing for techniques like gradient descent to be applied using gradient estimators of the Evidence Lower Bound (ELBO) cost function

$$D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z})) - \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})}[\log p(\mathbf{x} \mid \mathbf{z})]$$

which is equivalent to minimizing the KL-divergence. This approach is often used to train Bayesian machine learning models.

Challenges and motivation

Several challenges with VI motivate our GMM approximation method:

- **Non-convexity of the ELBO:** Minimizing the KL-divergence or, equivalently, the ELBO represents a non-convex optimization problem that may exhibit multiple local minima.
- **Capturing multimodality:** For certain applications it is important to capture multiple modes of the posterior with the VI approximation.
- **Scalability:** Poor scalability with VI is encountered using high-fidelity GMM approximations $q_{\theta}(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k); \mathbf{z} \in \mathbb{R}^d$. Due to the large number of parameters associated to the covariance matrix of each component $\dim(\boldsymbol{\Sigma}) = d(d+1)/2$, VI with a GMM involves optimization in a parameter space whose dimensionality grows like $\mathcal{O}(d^2)$ making it intractable for large ML models.

Mixture model approximation procedure

We seek an approximation $q_{\boldsymbol{\theta}}(\mathbf{z})$ to $p(\mathbf{z} \mid \mathbf{x})$ in the form of a Gaussian mixture model

$$q_{\boldsymbol{\theta}}(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\boldsymbol{\theta}$ denotes the set of parameters $\boldsymbol{\theta} = \{\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$.

Global optimization Compute means as local minima of the cost function $-\log \phi(\mathbf{z})$ through a global optimization procedure relying on multiple local searches. Local searches initialized from samples of the prior $p(\mathbf{z})$ or using a low-discrepancy Sobol samples. The global optimization stage results in a set of local minima $\mathbf{z}_1^*, \ldots, \mathbf{z}_K^*$ taken as the centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ of a Gaussian mixture model with K components.

Local Laplace Approximations To estimate the covariance matrix of each component, we employ the Laplace approximation such that

$$\boldsymbol{\Sigma}_i \approx \left(\mathbf{H}_f(\boldsymbol{\mu}_i)\right)^{-1}$$

where $f(\mathbf{z}) = -\log \phi(\mathbf{z})$ and $\mathbf{H}_f(\mathbf{z})$ denotes the Hessian of f evaluated at $\mathbf{z}$. Low-rank Hessian approximations can be used in high-dimensional settings to retain efficiency.

Determining distinct modes We let the null hypothesis be that $\mathbf{x}^*$ belongs to component k,i.e., $x^* \sim \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Letting $D_M(\mathbf{x}^*, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$ be the Mahalanobis distance between $\mathbf{x}^*$ and the local Gaussian distribution, then we have

$$P(D_M(\mathbf{x}^*, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) \geq d \mid H_0) = 1 - \chi^2(d, n)$$

Setting a threshold for the p-value provides a criterion for a new Gaussian component to be distinct from those already discovered.

Computing the weights Solve the constrained least squares problem

$$\arg\min_{\boldsymbol{\pi}} \sum_{i=1}^{N} \left\{ \phi(\mathbf{z}_i) - \sum_{k=1}^{K} \tilde{\pi}_k \mathcal{N}(\mathbf{z}_i \mid \boldsymbol{\mu}_k \boldsymbol{\Sigma}_k) \right\} \quad \text{s.t. } ; \tilde{\pi}_k \geq 0$$

for the unnormalized weights $\tilde{\pi}_1, \ldots, \tilde{\pi}_K$. Letting $Z = \sum_{k=1}^{K} \tilde{\pi}_k$, we can from the normalized approximation to $p(\mathbf{z})$ as $q_{\boldsymbol{\theta}}(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_k \boldsymbol{\Sigma}_k)$ where $\pi_k = \tilde{\pi}_k / Z$.

Global sensitivity analysis

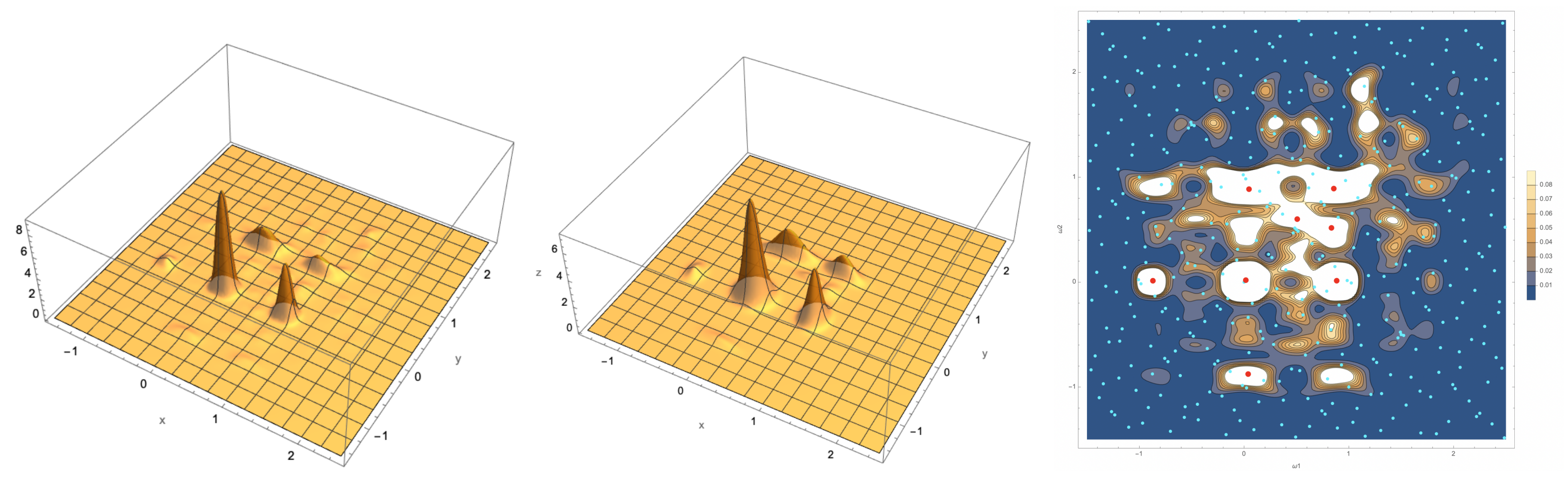
Variance-based sensitivity analysis allows decomposition of a function's $f(X_1, \ldots, X_k)$ variance as $V(f) = \sum_i V(f_i) + \sum_i \sum_{j>i} V(f_{ij}) + \cdots$ which provides a global measure of how each of the k input factors contributes to its variability. Typically this is applied to a nonlinear model function but here take a nonstandard approach of applying it to an approximation procedure to obtain a global measure of robustness over a space of applications with f taken to be the approximation error. We define as input factors the following parameters which control aspects of the true posterior we are trying to approximate with a GMM:

Parameter	Description	Distribution	S	S_T
d	Dimension	$\mathcal{U}\{8, 9, 10\}$	0.17 ± 10^{-3}	$\mathbf{0.65} \pm 10^{-2}$
K	No. of components	$\mathcal{U}\{3, 4\}$	0.13 ± 10^{-3}	0.30 ± 10^{-3}
d_π	Weight decay	$\mathcal{U}[1.3, 2]$	0.17 ± 10^{-2}	$\mathbf{0.37} \pm 10^{-2}$
c	Corr. coefficient	$\mathcal{U}[0.1, 0.7]$	0.00 ± 10^{-9}	$\mathbf{0.65} \pm 10^{-2}$
λ	Component overlap	$\mathcal{U}[10^{-4}, 10^{-2}]$	0.00 ± 10^{-9}	0.02 ± 10^{-4}

The resulting first and total order sensitivity indices S, S_T are listed in the final two columns and reflect how each parameter affects the approximation accuracy by alone and through interactions with other factors. The indices reveal that interactions between parameters creating modes with small basins of attraction provide the most significant effect on the approximation accuracy.

Low-dimensional example

We can construct a low-dimensional, multimodal posterior in the context of a simple nonlinear regression problem where we fit a model $f(\mathbf{x}; \omega_1, \omega_2) = \cos(2\pi\omega_1 x_1)\cos(2\pi\omega_2 x_2)$ to data $\mathcal{D} = \{(\mathbf{x}_i, y_i = f(\mathbf{x}_i, \tilde{\omega}_1, \tilde{\omega}_2) + \epsilon)\}_{i=1}^{N_s}$ sampled from the same function at unknown true frequencies $\tilde{\omega}_1, \tilde{\omega}_2$. Posterior defined by likelihood $p(\mathcal{D} \mid \omega_1, \omega_2) = \prod_{i=1}^{N_s} \mathcal{N}(y_i \mid f(\mathbf{x}_i, \omega_1, \omega_2), \sigma)$ with a Gaussian prior.



(a) True posterior (b) GMM approximation (c) Multistart locations

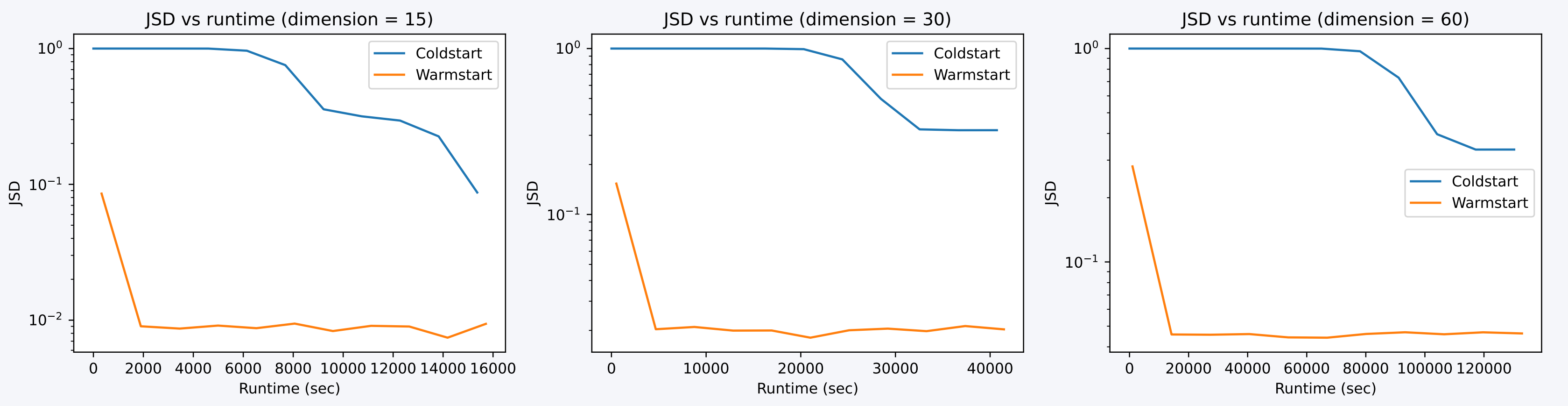
Above, we see the true 2D posterior, the GMM approximation from our algorithm, and an example contour plot of the multistart locations.

High-dimensional scalability study

We study the scalability of VI coupled with the GMM initialization procedure by carrying out VI on a high-dimensional, synthetic posterior displaying non-Gaussian trends intended to emulate a realistic modeling application. To achieve this behavior, we employ the Sinh-arcsinh distribution induced by a nonlinear transformation $Y = l + \sigma F(Z)$ of a standard normal random variable Z

$$F(Z) = \frac{\sinh((\operatorname{arcsinh}(Z) + s)t)}{2\sinh(\operatorname{arcsinh}(Z)t)}$$

where l, σ represent the location and scale, respectively, and s, t control the skewness and tail behavior. This allows for the construction of a multimodal synthetic posterior with controllable non-Gaussian trends. We then compare scalability between randomly initialized VI (cold-start) with VI initialized using our GMM approximation (warm-start) for randomly generated synthetic posteriors of dimensions 15, 30, and 60.



The figure above shows the Jenson-Shannon divergence (JSD), scaled to lie in the interval $[0, 1]$, between the GMM surrogate posterior and the true posterior as a function of total CPU runtime. JSD can be thought of as a symmetric and normalized version of KL-divergence. VI with warm-start sees at a six-fold acceleration in convergence as well as a lower final JSD approximation error.

Conclusion

Our approach for approximating intractable posteriors with GMMs through global optimization and Laplace approximations can be used to improve the scalability of VI for high-fidelity mixture distributions and may serve as a more efficient alternative to VI in certain applications.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

(1) Sandia National Laboratories