Robust initialization of variational inference through global optimization and Laplace approximations

Wyatt Bridgman[1], Reese Jones[1], Mohammad Khalil[1]

[1]*Sandia National Laboratories, Livermore, CA*

Conference on Mathematics of Data Science

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. **Sand No. ???**

September 21, 2022

boratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LL Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE

Context: novel probabilistic strategies for transfer learning

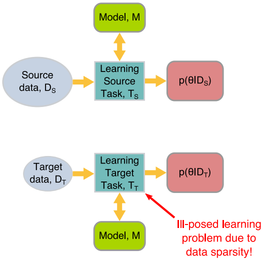
Challenge: Many Sandia mission domains defined by a lack of reliable data, preventing use of many machine learning techniques for predictive modeling:

- Expensive computer simulations.
- Prohibitive data acquisition cost.
- Limited access to classified/sensitive data.

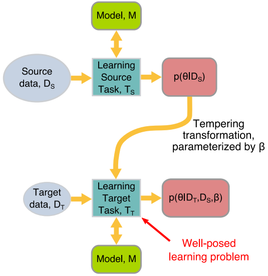
Goal: *Enhance the trust in ML within noisy and sparse data settings.*

Requirement: **High-fidelity**, **closed-form** approximations of parameter PDFs (as opposed to just samples) for approximation of multimodal target likelihood $p(\boldsymbol{\theta} \mid \mathcal{D}_t)$.

Traditional TL



Probabilistic TL



Approximation of posteriors in Bayesian inference

Bayesian inference: probabilistic model $p(\mathbf{x}, \mathbf{z})$ with $\mathbf{x}$,$\mathbf{z}$ observed, latent variables. Some structure specified on the joint distribution e.g. $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$. Here we take $\mathbf{x}$ to be a fixed set of data $\mathcal{D}$. Baye's rule gives:

$$p(\mathbf{z} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{z})p(\mathbf{z})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \mathbf{z})p(\mathbf{z})}{\int p(\mathcal{D} \mid \mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}}; \quad \underbrace{\int p(\mathcal{D} \mid \mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}}_{\text{difficult to compute}} \tag{1}$$

Sampling strategies for $p(\mathbf{z} \mid \mathcal{D})$ include:

- **MCMC:** Asymptotically exact but computationally expensive. Difficulties with multimodal distributions.
- **Dropout:** Adds UQ to neural networks via random perturbations of the weights. Less costly than Variational Inference.

Variational Inference (VI)

Sampling strategies often suffer from scalability issues, we also need **closed form** approximations for the TL framework.

Variational Inference

Approximate posterior $p(\mathbf{z} \mid \mathcal{D})$ by $q_{\boldsymbol{\theta}}(\mathbf{z}) \in \mathcal{F}_{\boldsymbol{\theta}}$ in some family $\mathcal{F}_{\boldsymbol{\theta}}$ of PDFs by minimizing error measure such as KL-divergence:

$$q_{\boldsymbol{\theta}}(\mathbf{z}) = \min_{q_{\boldsymbol{\theta}} \in \mathcal{F}_{\boldsymbol{\theta}}} D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathcal{D})) \tag{2}$$

where $\boldsymbol{\theta}$ are the variational parameters to be optimized.

- **Pros:** Obtain closed form approximation $q_{\boldsymbol{\theta}}(\mathbf{z})$ whose fidelity is determined by choice of family, e.g., whether a single Gaussian or mixture. Can be scalable to large NN models depending on choice of $\mathcal{F}_{\boldsymbol{\theta}}$.
- **Cons:** Can underestimate variance, suffers from optimization pitfalls due to nonconvexity of objectives.

Case: Variational Bayes

- Take $q(\mathbf{z}) = \prod_{i=1}^{M} q(\mathbf{z}_i)$, i.e., $q(\mathbf{z})$ is from the space $\mathcal{F}$ of PDFs that factor over partition of latent variables into $\mathbf{z} = \mathbf{z}_1, \ldots, \mathbf{z}_M$.
- Can be shown that optimal solution of functional $\min_{q(\mathbf{z}) \in \mathcal{F}} D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathcal{D}))$ is given by

$$q_j^*(\mathbf{z}_j) = \frac{e^{\mathbb{E}_{i \neq j}[\log p(\mathbf{z}, \mathcal{D})]}}{\int e^{\mathbb{E}_{i \neq j}[\log p(\mathbf{z}, \mathcal{D})]} \, \mathrm{d}\mathbf{z}_j} \tag{3}$$

- Can sometimes determine $q_j^*(\mathbf{z}_j)$ to be known PDF whose parameters satisfy set of simultaneous nonlinear equations $\rightarrow$ solve iteratively.
- For ML models $\mathbb{E}_{i \neq j}[\log p(\mathbf{z}, \mathcal{D})]$ hard to compute as $p(\mathbf{z}, \mathcal{D})$ parameterized by nonlinear NN $\rightarrow$ would have to solve set of integral equations.

VI via gradient-based optimization

- The KL-divergence plus the Evidence Lower Bound (ELBO) $\mathcal{L}_{\boldsymbol{\theta}}$ differ by a constant

$$\underbrace{\log p(\mathcal{D})}_{\text{const. w.r.t. } \boldsymbol{\theta}} = \underbrace{\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D})}_{\text{ELBO}} + D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathcal{D})) \tag{4}$$

  so that we can minimize $D_{\mathrm{KL}}(\parallel)$ by minimizing

$$-\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D}) = \underbrace{D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\mathbf{z}) \parallel p(\mathbf{z}))}_{\text{KL-div from prior}} - \underbrace{\mathbb{E}_{q_{\boldsymbol{\theta}}}\left[\log p(\mathcal{D} \mid \mathbf{z})\right]}_{\text{expected data fit}} \tag{5}$$

- If $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_j)\}_{i=1}^{N_s}$, likelihood given by IID Gaussian over model predictions $f_{\mathbf{z}}(\mathbf{x}_i)$ with noise σ, then expected data fit like a stochastic mean-squared error (MSE):

$$-\mathbb{E}_{q_{\boldsymbol{\theta}}}\left[\log p(\mathcal{D} \mid \mathbf{z})\right] = \frac{1}{2\sigma}\mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\sum_{s=1}^{N_s} ||\mathbf{y}_i - f_{\mathbf{z}}(\mathbf{x}_i)||^2\right] \tag{6}$$

Example: ELBO for linear NN

- NN given by $\mathrm{NN}_{\mathbf{W}}(\mathbf{x}) = \mathbf{W}\mathbf{x}$, $\mathbf{W} \in \mathbb{R}^{n \times n}$ where

$$q_{(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)} = \mathcal{N}(\mathbf{W} \mid \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q), \quad p(\mathbf{W}) = \mathcal{N}(\mathbf{W} \mid \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$

then the ELBO objective function has form

$$\underbrace{\frac{1}{\sigma^2} \operatorname{tr}\{(\mathbf{Y} - \boldsymbol{\mu}_q \mathbf{X})^T (\mathbf{Y} - \boldsymbol{\mu}_q \mathbf{X})\}}_{\text{least squares in } \boldsymbol{\mu}_q} + \underbrace{(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)}_{\text{regularization } \boldsymbol{\mu}_q \to \boldsymbol{\mu}_p}$$

$$+ \quad \underbrace{\log\det(\boldsymbol{\Sigma}_q^{-1} \boldsymbol{\Sigma}_p) + \operatorname{tr}(\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_q)}_{\boldsymbol{\Sigma}_q \to \boldsymbol{\Sigma}_p} + \underbrace{\frac{1}{\sigma^2} \operatorname{tr}\{\mathbf{V}\mathbf{X}\mathbf{X}^T\}}_{\boldsymbol{\Sigma}_q \to \mathbf{0}}$$

which takes the form of least squares in means $\boldsymbol{\mu}_q$ with quadratic regularization. Variance $\boldsymbol{\Sigma}_q$ balanced between prior $\boldsymbol{\Sigma}_p$ and $\mathbf{0}$.

Minimizing ELBO for nonlinear model

- Minimize $D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathcal{D}))$ via gradient descent requires $\nabla_{\boldsymbol{\theta}}(-\mathcal{L}_{\boldsymbol{\theta}})$.
- **Score function / black-box VI:**

$$\nabla_{\boldsymbol{\theta}}(-\mathcal{L}_{\boldsymbol{\theta}}) = \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z})}\left[(\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})) \log p(\mathcal{D} \mid \mathbf{z})\right]$$

- **Reparametrization gradients:** Express $\mathbf{z} = t(\epsilon, \boldsymbol{\theta})$, $\boldsymbol{\epsilon} \sim p(\epsilon)$ then gradient and expectation commute

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log p(\mathcal{D} \mid \mathbf{z})\right] = \mathbb{E}_{p(\epsilon)}[\underbrace{\nabla_{\mathbf{z}} \log p(\mathcal{D} \mid \mathbf{z})}_{\text{backprop. gradient}} \nabla_{\boldsymbol{\theta}} \mathbf{z}]$$

Lower variance than score method but reparametrization more difficult for complex distributions like GMMs (Graves, 2016),(Fignurov 2018).

Challenges with VI for high-fidelity distributions

- VI doesn't scale well with high-fidelity posterior approximations such as Gaussian mixture models or even full covariance Gaussians.
- ELBO is nonconvex, optimizers can find poor local minima (Kingma, Welling 2019).
- Some approaches to address this issues include annealing (Bowman, 2016),(Sonderby et. al., 2006) and good initialization strategies (Rossi et. al 2019).
- Growing body of literature suggesting Laplace approximations (LAs) perform well in a variety of ML/UQ tasks:

$$p(\mathbf{z} \mid \mathcal{D}) \approx \frac{1}{Z_g} \exp\left[-\frac{1}{2}(\mathbf{z} - \mathbf{z}_{\text{MAP}})^T \mathbf{\Sigma}^{-1}(\mathbf{z} - \mathbf{z}_{\text{MAP}})\right]$$

where $\mathbf{\Sigma} = -\mathbf{H}^{-1}_{\log \phi}(\mathbf{z}_{\text{MAP}})$, $\mathbf{z}_{\text{MAP}}$ is maximum a posteriori estimate

Global optimization and LA

Proposed approach: *Approximate multimodal PDF with global optimization and LAs. Can be used to initialize VI or, possibly, as an alternative approximation strategy.*

Outline of proposed method:

Unnormalized posterior distribution $\tilde{p}(\mathbf{z})$

- Global optimization carried out on $\tilde{p}$ to find modes $\mathbf{z}_1^*, \ldots, \mathbf{z}_K^*$ taken as centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ of Gaussian components.
- Laplace approximation formed at each mode:

$$\mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k = \mathbf{H}^{-1}_{-\log \tilde{p}}(\boldsymbol{\mu}_i))$$

- Fit the weights via constrained least squares:

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

then $\int \tilde{p}(\mathbf{z})\, d\mathbf{z} \approx \sum_k \tilde{\pi}_k$.

Scalability: VI vs. global opt. & LA

VI with Gaussian Mixture Models

- VI with $q_{\boldsymbol{\theta}}(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \mathbf{z} \in \mathbb{R}^d$ has variational parameters

$$\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$$

  so that $\boldsymbol{\theta} \in \mathbb{R}^{K+Kd+K(d+d^2)/2} \longrightarrow$ grows like $\mathcal{O}(d^2)$.
- Loss function nonconvex means multiple optimization runs are needed to avoid poor local minima.

Global opt. & LA

- Carry out many local optimizations in smaller parameter space $\mathbb{R}^d$ instead of several expensive optimization runs in larger variational parameter space $\mathbb{R}^{\mathcal{O}(d^2)}$.
- Enhance scalability with low-rank Hessian approximations.
- A variety of global optimization techniques can be used such as MLSL which purport to reduce number of local searches needed.

Robustness of approach via global sensitivity

Variance based global sensitivity analysis:

$$
\begin{aligned}
f &= f_0 + \sum_i f_i(X_i) + \sum_i \sum_{j>i} f_{ij}(X_i, X_j) + \cdots \\
\mathbb{V}(f) &= \mathbb{V}(f_0) + \sum_i \mathbb{V}(f_i) + \sum_i \sum_{j>i} \mathbb{V}(f_{ij}) + \cdots
\end{aligned}
$$

Use sensitivity analysis over ensemble of synthetic tests on GMMs to study how performance $f(d, K, \omega, c, \lambda)$ varies as a function of

Parameter	Description	Distribution
d	Dimension	$\mathcal{U}\{2, 10\}$
K	Number of mixture components	$\mathcal{U}\{2, 4\}$
ω	Exponential decay factor across weights	$\mathcal{U}[1, 2]$
c	Correlation coefficient	$\mathcal{U}[0, 0.7]$
λ	Maximum overlap between components	$\mathcal{U}[10^{-4}, 10^{-2}]$

Robustness of approach via global sensitivity

- ω controls spread of component sizes, λ controls the overlap between components measured by Dice metric.
- Accuracy measured by $D_{\mathsf{JSD}}(\mathcal{G}(\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \parallel \mathcal{G}(\hat{\pi}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}))$, Jenson-Shannon divergence between true, approximate GMMs. Obtained by "symmetrizing" KL-divergence, bounded.

Global sensitivity results

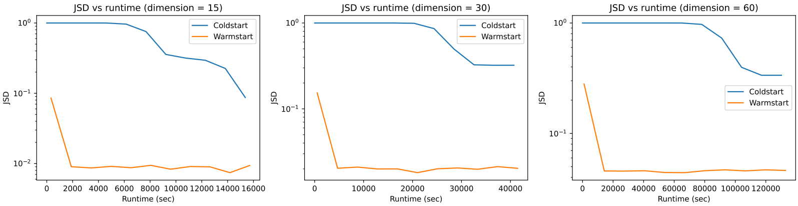
Parameter	Distribution	S	S_T
d, dim.	$\mathcal{U}\{8, 9, 10\}$	0.17 ± 10^{-3}	$\mathbf{0.65} \pm 10^{-2}$
K, no. components	$\mathcal{U}\{3, 4\}$	0.13 ± 10^{-3}	0.30 ± 10^{-3}
ω, weight decay	$\mathcal{U}[1.3, 2]$	0.17 ± 10^{-2}	$\mathbf{0.37} \pm 10^{-2}$
c, corr.	$\mathcal{U}[0.1, 0.7]$	0 ± 10^{-9}	$\mathbf{0.65} \pm 10^{-2}$
λ, overlap	$\mathcal{U}[10^{-4}, 10^{-2}]$	0 ± 10^{-9}	0.02 ± 10^{-4}

Conclusion: *Interaction between factors which increase difficulty of global optimization have the greatest effect.*

Scalability of global optimization, LA method

Can we improve the scalability of VI with high-fidelity GMM surrogate posteriors using the GMM approximation scheme?

- Carry out scalability analysis in high dimensional setting on toy problems with non-Gaussian trends.
- Cold start (randomly init. VI) versus warm start (GMM init.)
- Generate non-Gaussian mixture models by applying nonlinear transformation $Y = I + \sigma F(Z, s, t)$ on standard normal r.v. Z where s, t control skewness and tail behavior.



Conclusion: *Using GMM approx. procedure improves scalability and achievable accuracy.*

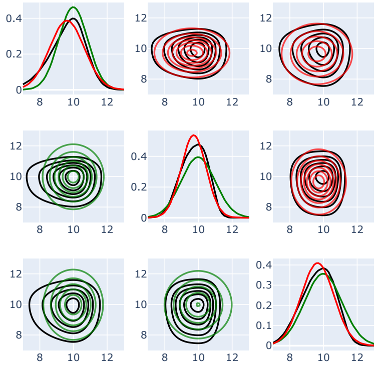
How does the Laplace approximation compare to VI?

Laplace approximation

- Captures peak and local geometry.
- Approximation away from peak worsens with increasing non-Gaussian trends.

VI-refined approximation

- refines support of modes to lie within high-probability regions of true distribution.
- Doesn't capture peaks as closely.

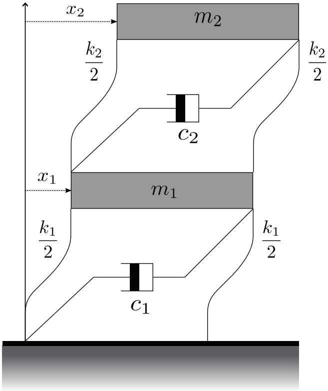


Marginals between 3 variables of 15-dim. distribution. **Black:** true, **Green:** LA, **Red:** VI.

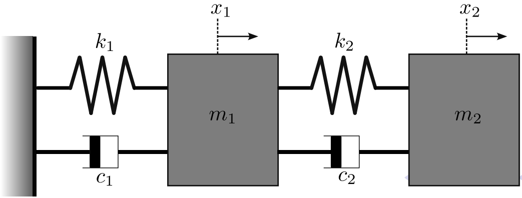
How do these approx. differences translate into **predictions?**

Structural dynamics problem

- Two-story shear frame structure.
- Subject 2nd floor to initial displacement.
- Inverse problem of determining damping coefficients c_1, c_2 while observing only the first floor's motion.
- Can obtain multimodal posterior over c_1,c_2.



Equivalent to mass-spring-damper system:



Structural dynamics problem

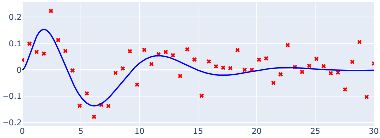
- Equations of motion:

$$\frac{d}{dt}\begin{bmatrix}\mathbf{x}\\\mathbf{v}\end{bmatrix} = \begin{bmatrix}\mathbf{0} & \mathbf{I}\\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{C}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\\mathbf{v}\end{bmatrix}$$

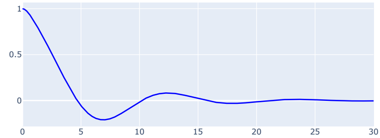
- Log likelihood from matrix exponential

$$\frac{1}{\sigma^2}\sum_{i=1}^{N_D}(y_i - \mathbf{H}\exp(\mathbf{A}(c_1, c_2)t_i)\bar{\mathbf{x}}_0)^2$$

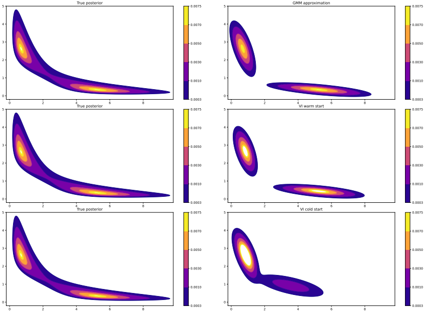
- Noisy observations of first floor



- Second floor displacement

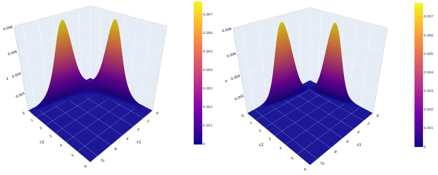


Posterior and GMM approximation

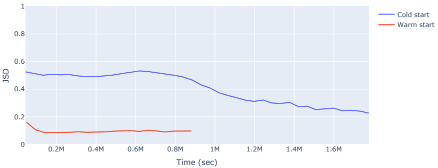


Contour plots of:

- **(Top)**: GMM approximation
- **(Middle)**: GMM approx. refined with VI
- **(Bottom)**: Example of randomly initialized VI solution. Gets stuck in local min.

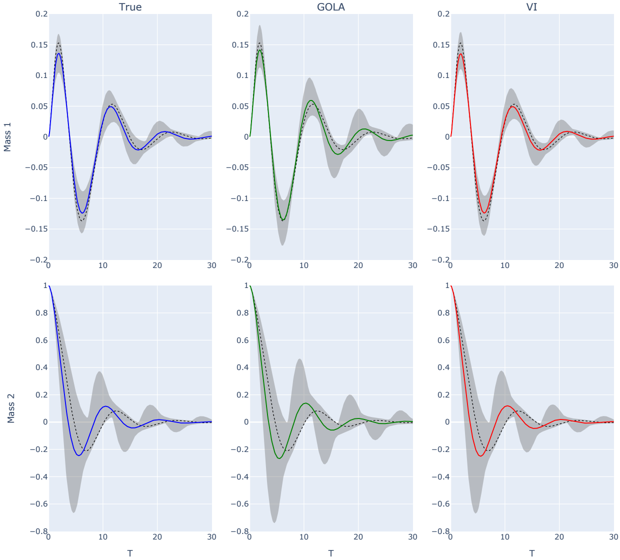


True multimodal posterior **(left)** and two-component GMM approximation **(right)**.



JS-divergence vs wall-clock time for warm-start, cold-start.

Posterior pushforward



Posterior pushforward of True (left), Global opt., LA (middle), VI-refined (right)

The End

Thank you!