Energetic Variational Inference

Lulu Kang

Illinois Institute of Technology

April 6, 2022. Illinois TechArgonne CELS Research Seminar

About Me

Bio

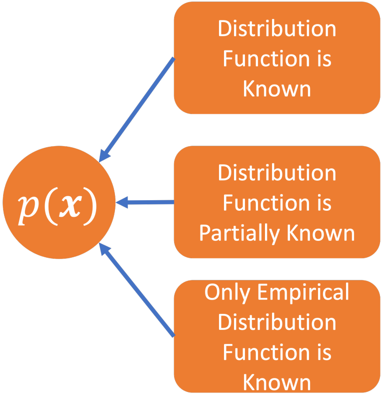
- Education
  - B.S. in Mathematics, Nanjing University, China.
  - M.S. in Operations Research, Georgia Institute of Technology.
  - Ph.D. in Industrial Engineering, Georgia Institute of Technology.
- Academic Appointment
  - Associate Professor in Applied Mathematics
  - Associate Director of Master of in Data Science, 2013-2022.
  - Director of B.S. in Data Science (new in Fall 2022), 2022-

Research Interests

Statistics	Optimization	Collaboration
• Causal Inference • Statistical Design and Analysis of Experiments • Statistical Learning • Uncertainty Quantification • Bayesian Statistics	• Optimization methods in statistics • Machine Learning • Applications in other domains	• Material Sciences • Chemistry • Mechanical Engineering • Health care

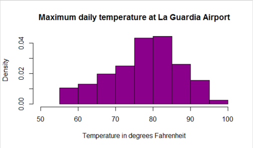
Energetic Variational Inference

One major problem in many fields: generate samples from $p(x)$



e.g. $p(x) = 0.5\mathrm{N}(x|1, 0.5^2) + 0.5N(x|3, 2^2)$

e.g. $p(x) \propto e^{-(0.1x_1^2 + (x_2 - \sin \pi x_1)^2)}$



Typical problems

- Bayesian inference: generate samples from the posterior distributions, which are usually not any known ones.
- Space filling design: the target distribution is uniform.
- Quadrature: $I = \int_\Omega f(\boldsymbol{x}) \mathrm{d}\mu(\boldsymbol{x}) \approx \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)$ where $\boldsymbol{x}_i \sim^{iid} \mu(\boldsymbol{x})$.
- Generative learning: based on existing samples, generate prediction or make classification on the queries or generate new samples.
- ...

Variational Inference in a Nutshell

- $D(f_1||f_2)$ is a discrepancy or divergence measure, that measuring the difference between any two distributions, f_1 and f_2. Examples include KL-divergence, f-divergence, maximum mean discrepancy (MMD) also known as kernel discrepancy,...
- Variational Inference answers the question of how to minimize $D(f||f^*)$ such that minimal solution f would be as close as possible to the target distribution f^*.

Formal Introduction of Variational Inference

- Variational inference seeks to find an approximation $q^*(\boldsymbol{\theta})$ to approximate the target $p(\boldsymbol{\theta}|\mathcal{D})$.

$$q^*(\boldsymbol{\theta}) = \arg \min_{q(\boldsymbol{\theta}) \in \mathcal{Q}} D\left(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta}|\mathcal{D})\right).$$

- $\mathcal{Q}$ is a user specified family of distributions where the approximation density is in. The complexity of $\mathcal{Q}$ determines the complexity of the optimization.
- Two questions:
  - how good is the approximation?
  - how to solve this minimization problem?

Possible Divergence

- $D\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})\right)$ can be KL-divergence, which measures the difference between two probability density functions.

$$\mathsf{KL}(p_1(\boldsymbol{x})||p_2(\boldsymbol{x})) = \int p_1(\boldsymbol{x})\log\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}d\boldsymbol{x} = \mathbb{E}_{\boldsymbol{x}\sim q_1}\left(\log\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right).$$

- $D\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})\right)$ can be maximum mean discrepancy (MMD), or *kernel discprepancy*.

$$\begin{aligned}\mathsf{MMD}^2(\mathcal{H},\nu_1,\nu_2) &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\nu_1}[K(\boldsymbol{x},\boldsymbol{x}')] \\ &- 2\mathbb{E}_{\boldsymbol{x}\sim\nu_1,\boldsymbol{y}\sim\nu_2}[K(\boldsymbol{x},\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{y}\sim\nu_2,\boldsymbol{y}'\sim\nu_2}[K(\boldsymbol{y},\boldsymbol{y}')]\end{aligned}$$

- Many others.

Variational Inference: Pros and Cons

- Advantage of VI: it becomes an optimization problem and can be used to for large datasets and to explore many models; faster computation due to some existing stochastic optimization algorithms.
- Disadvantage of VI: there is no convergence guarantee that the approximated density converges to the targeted density function as the algorithm iterates.
- There are many variational inference methods:
  - Mean-field: classic, simple, but limited.
  - Stein Variational Gradient Descent: take advantage of the connection between derivative of KL divergence and stein operator; use limited number of particles to approximate the stein operator and find a series of mapping to map the original distribution of the particles to a distribution that is closest to the target distribution.
  - Other particle-based variational inference methods.

Energetic Variational Inference: Flow Maps

Minimize a chosen divergence through flow maps

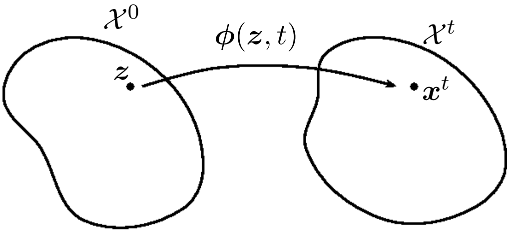


Figure: A schematic of a flow map $\phi(\boldsymbol{z}, t)$. For t fixed, $\phi(\boldsymbol{z}, t)$ maps $\mathcal{X}^0$ to $\mathcal{X}^t$. For $\boldsymbol{z}$ fixed, $\phi(\boldsymbol{z}, t)$ is the trajectory of a particle with initial position $\boldsymbol{z}$.

Energetic Variational Inference: Continuous Version

In an isothermal closed physics system, an *energy dissipation law*, is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{K}+\mathcal{F})[\phi] = -2\mathcal{D}[\phi, \phi_t], \tag{1}$$

describes how the total energy of the system decreases with time, which is a consequence of the First and Second Law of thermodynamics.

- $\mathcal{K}$ is the kinetic energy, we usually set it to be zero.
- $\mathcal{F}$ is the Helmholtz free energy $\Rightarrow$ Let $\mathcal{F}$ be the divergence measure.
- $-2\mathcal{D} \leq 0$ is the rate of energy dissipation $\Rightarrow$ specifies the mechanism of minimizing the divergence.
- ϕ is the state variable of the system.
- ϕ_t is the derivative of ϕ with respect to time.

Energetic Variational Inference: Discrete Version

In a particle-based variational inference, the time-dependent probability density $\rho(\boldsymbol{x}, t)$ is approximated by an empirical measure defined by a set of sample points $\{\boldsymbol{x}_i(t)\}$ (or particles)

$$\rho(\boldsymbol{x}, t) \approx \rho_N^t(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_i(t)), \tag{2}$$

where $\boldsymbol{x}_i(t) = \boldsymbol{\phi}(\boldsymbol{x}_i(0), t)$. Instead of computing $\boldsymbol{\phi}(\boldsymbol{z}, t)$ explicitly at each time-step, only $\boldsymbol{x}_i(t)$ is computed.

Energetic Variational Inference: Discrete Version

Discrete energy-dissipation law

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_h(\{\boldsymbol{x}_i(t)\}_{i=1}^N) = -2\mathcal{D}_h(\{\boldsymbol{x}_i(t)\}_{i=1}^N, \{\boldsymbol{x}_i'(t)\}_{i=1}^N), \tag{3}$$

which can be obtained by inserting the empirical approximation (1) into the continuous energy-dissipation law with a suitable *kernel regularization*.

The EVI Framework: Algorithms

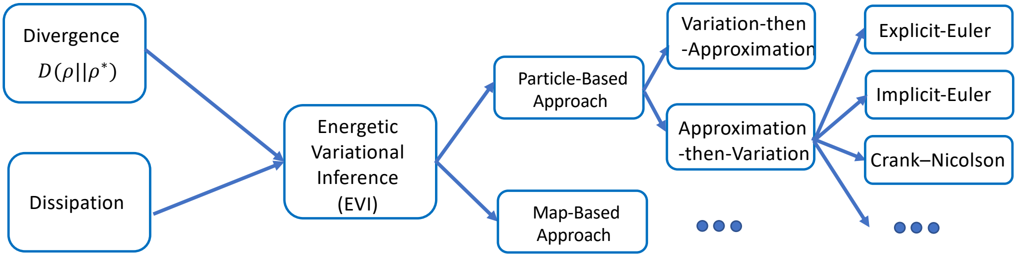


Figure: The flowchart of the proposed EVI framework and how to create its variations.

The EVI Framework: Foundations

Convergence issues for the continuous time t involve four versions of $\rho(\boldsymbol{x})$.

1. $\rho(\boldsymbol{x}, t)$–the continuous version of the density evolving at time t;
2. $\rho^{\infty}(\boldsymbol{x})$–the limiting density function of $\rho(\boldsymbol{x}, t)$ as $t \to \infty$;
3. $\rho_N(\boldsymbol{x}, t)$–the discretized version of $\rho(\boldsymbol{x}, t)$ by N particles evolving at time t;
4. (4) $\rho_N^{\infty}(\boldsymbol{x})$–the limiting density function of $\rho_N(\boldsymbol{x}, t)$ as $t \to \infty$.

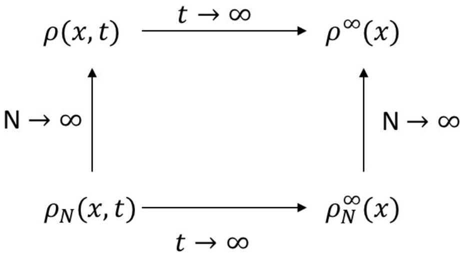


Figure: Different convergence

Examples

KL-Divergence: Toy Examples

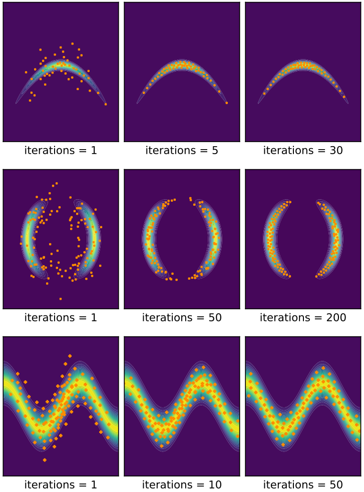


Figure: The particles obtained by EVI-Im algorithm approximating three target distributions plotted as contours.

KL-Divergence: Star-Shape Mixture Gaussian

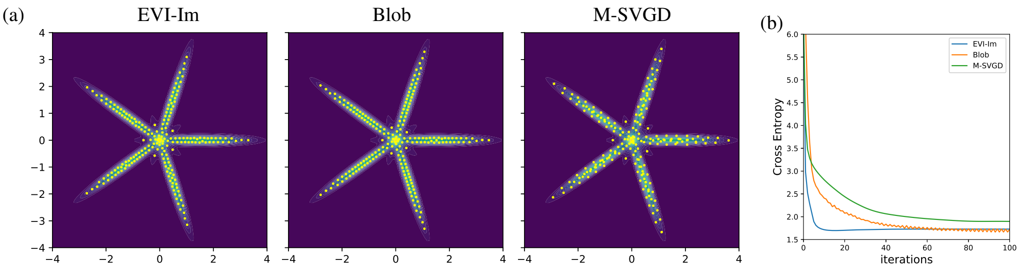


Figure: (a) Particles obtained by various methods [200 particles]: EVI-Im after 20 iterations, Blob method and matrix-valued SVGD both after 1000 iterations; (b) cross-entropy v.s. number of iterations of the three methods.

KL-Divergence: Gaussian Mixture Gaussian

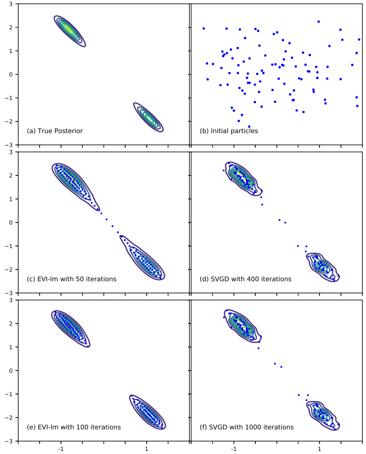


Figure: Comparison of EVI-Im and the classic SVGD ($\text{lr} = 1$) at different iterations in Example 2.

KL-Divergence: Bayesian Logistic Regression

A small data set SPLICE (1,000 training entries, 60 features).

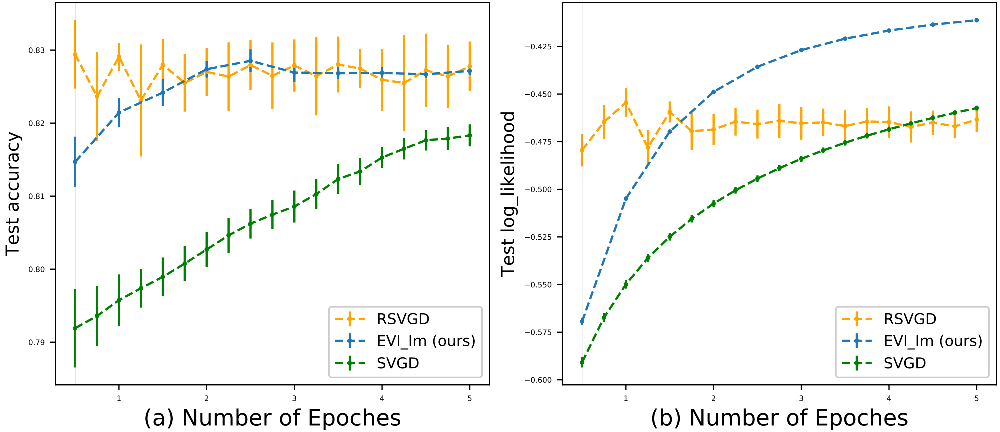


Figure: Example 3. The test accuracy and log-likelihood of the training data (20 simulations) returned by EVI-Im, RSVGD, and SVGD methods.

KL-Divergence Bayesian Logistic Regression

A large data set Covertype ($S = 581,012$ data entries and 54 features).

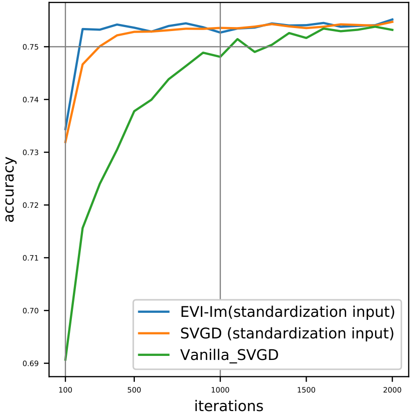


Figure: The accuracy of 20 simulations for Bayesian logistic regression on Covertype dataset using different methods.

MMD: Toy Examples

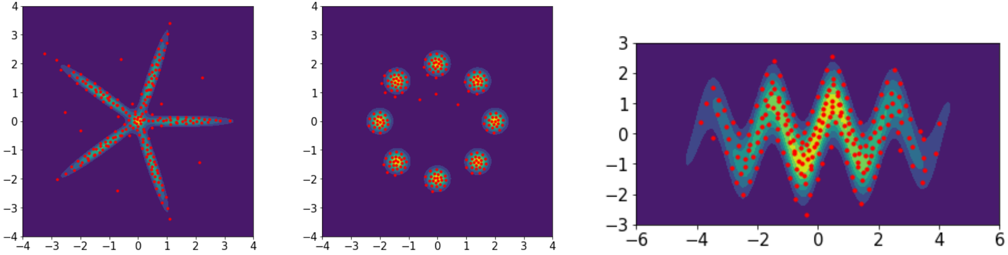


Figure: Low-discrepancy points for three target distributions.

MMD: Numerical Integration

$$I = \int_{\mathbb{R}^d} \cos(\|\boldsymbol{x}\|_2) \exp(-\|\boldsymbol{x}\|_2^2) \mathrm{d}\boldsymbol{x}.$$

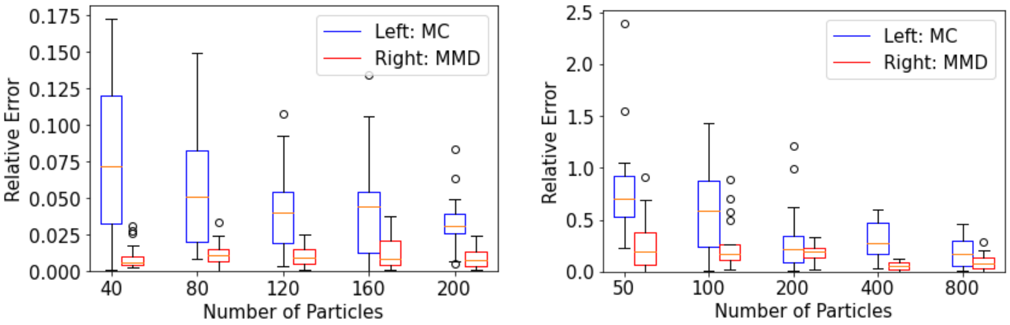


Figure: The relative error of 20 repetitions, left is 2D case right is 5D case.

MMD: Generative Learning

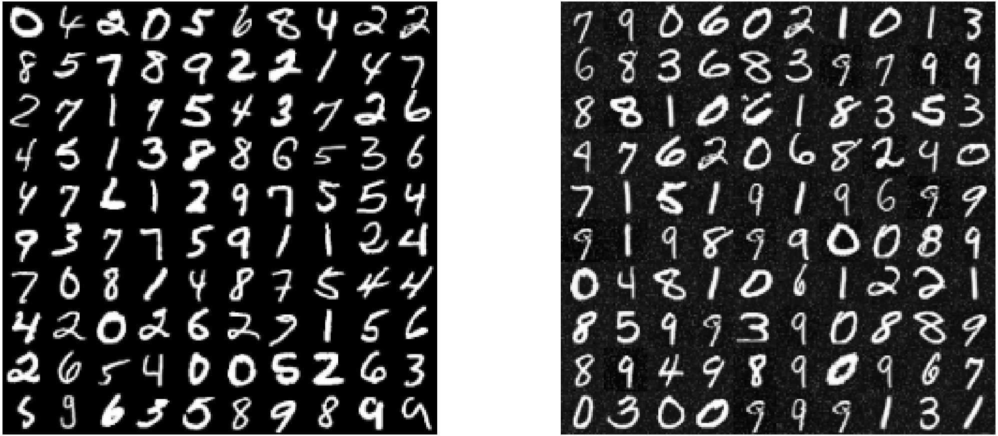


Figure: MNIST example: training samples (left) and generated samples from EVI-MMD (right).

Conclusion

Collaborators and Students

This project is supported by NSF.

- Lulu Kang, PI.
- Chun Liu, Chair of Applied Mathematics, Co-PI.
- Yiwei Wang, Assistant Visiting Professor, Co-PI.
- Jiuhai Chen, graduated in Spring 2020 with M.S. in Applied Mathematics.
- Yindong Chen, current Ph.D. student.
- Yuanxing Cheng, current Ph.D. student.
- Kaylee Rosendahl, current 2nd year AMATH undergraduate student.

Seek collaborators

We are seeking collaborators

- scientists and engineers who need statistical and machine learning solutions to advance their research, and who can provide us domain expertise and application background;
- mathematicians, statisticians, or machine learning experts who want to work on the methodological and theoretical sides.

Contact

Thank you!

Email: lkang2@iit.edu

https://sites.google.com/iit.edu/lulukang/