Stat 238, Fall 2025
Project Minis

Alexander Strang
Due: by **11:00 am** Tuesday, May 13th, 2025

Mini Project 10: Variational Inference

Policies

- This is an individual project.
- All submissions must be properly type-set and sourced. This means providing formal citations and a complete bibliography when sources are used.
- Project minis should be uploaded to Gradescope under the appropriate mini-assignment.
- Please export any notebooks as a single pdf and merge all components into one file.

Prompts

This is a reading project.

Bayesian inference is typically performed via approximate posterior sampling. However, there are many situations where sampling is inefficient, especially if we want to perform high-dimensional inference. Variational inference substitutes sampling with optimization. Instead of drawing samples from the posterior (or an approximation to the posterior produced by running an MCMC scheme), variational inference searches for the closest approximation to the target posterior among all distributions in an (often infinite-dimensional) family of tractable distributions.

Please read:

1. Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. ”Variational inference: A review for statisticians.” *Journal of the American statistical Association* 112, no. 518 (2017): 859-877.
2. Liu, Qiang, and Dilin Wang. ”Stein variational gradient descent: A general purpose Bayesian inference algorithm.” *Advances in neural information processing systems* 29 (2016).

Then write a 4 - 5 page essay that summarizes and comments on your reading. It should, at minimum, address the discussion prompts outlined below. For the second paper it may help to read:

- Papamakarios, George, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. "Normalizing flows for probabilistic modeling and inference." *Journal of Machine Learning Research* 22, no. 57 (2021): 1-64.
- Wang, Yiwei, Jiuhai Chen, Chun Liu, and Lulu Kang. "Particle-based energetic variational inference." *Statistics and Computing* 31 (2021): 1-17.

(a) (5 points) Compare variational inference methods and MCMC procedures. Identify pros and cons for each framework, and propose at least one example setting where you prefer each over the other.

(b) (3 points) Variational inference typically adopts the KL divergence between the variational distribution, q, and the target as its objective, p_*. Explain why we adopt $D_{\text{KL}}(q||p_*)$ rather than $D_{\text{KL}}(p_*||q)$). Explain how your interpretation of the objective would (or would not) change, if you exchanged the order of the distributions inside the KL divergence. What biases do you expect this to induce in the variational solution?

(c) (5 points) Clearly summarize the CAVI algorithm for mean-field inference in conditionally conjugate models. In particular, show that each step in the iterative procedure is a descent step on the KL divergence. Try to relate the procedure to another method from the course (e.g. coordinate ascent, expectation-maximization).

(d) (2 points) Discuss the biases induced by the mean-field assumption, and, given these biases, how you would use the variational distribution produced by optimizing over the mean field class.

(e) (5 points) Clearly summarize the SVGD algorithm for variational inference. Then, answer the following questions:

- What is the variational family that SVGD optimizes over?[1]
- Give an intuitive description of the particle dynamics specified by the SVGD algorithm. Compare this algorithm to another ensemble algorithm from the class. What component of the algorithm ensures that the particles spread out? What is responsible for the variance in samples in related MCMC methods?
- Are you convinced by the "median-trick" for selecting the kernel bandwidth?[2] Propose an alternative procedure for adaptively selecting the bandwidth.

(f) (5 points) SVGD, like other particle flow based methods, uses a flow map to transport particles drawn from a reference distribution (often normal). The transport map is chosen so that, after the transformation, the original distribution is as close as possible to the target. This is implemented by at each step moving a set of reference particles as if they obeyed a time inhomogeneous ODE. The vector field specifying the ODE is chosen to instantaneously decrease the KL as quickly as possible. Show that the functional gradient of the KL divergence with respect to a velocity field transporting $q(t)$ to $q(t + dt)$ can be calculated using only computationally available quantites (the unnormalized target, log densities, partial derivatives of log densities).

[1] This is a tricky question. It may help to start by asking, how would I sample from the distribution produced at the end of SVGD? The variational family will be all distributions that can be specified by this sampling procedure.

[2] Be skeptical here.