An Exploration of Variational Inference

STA 602 Final Project Report - Evan Glas

Abstract

Variational inference (VI) is a technique in Bayesian statistics for approximating and evaluating posterior distributions. VI poses posterior estimation as an optimization problem. The goal of VI is to find a distribution, $q(\theta)$, as "close" as possible to a true posterior, $p(\theta)$. Compared to other posterior estimation methods, VI demonstrates a series of advantages, including computational efficiency and relative speed of convergence. Across multiple decades, researchers have successfully applied VI towards a broad range of problems. Today, VI remains a fundamental role in the Bayesian statistical toolkit as demonstrated by its persistent presence/influence in current literature. The goal of my final project was twofold: to gain a theoretical understanding of variational inference, and to then leverage this understanding by applying VI to a real example. This report is to serve as documentation for both pieces of my project. First, I cover the mathematical underpinnings of variational inference. I then include a brief discussion of prominent applications of VI. Finally, I implement VI to fit a Gaussian mixture model.

Background

The term "variational" in VI stems from "calculus of variations", or the field of mathematics concerned with the optimization of *functionals*. Whereas *functions* describe a mapping between sets, *functionals* describe mappings of *functions*. For example, one could describe a *function* $f(x) = x$. One example of a *functional* would be a definite integral over that function, $\int_a^b f(x)dx$. Variational calculus involves the optimization of *functionals* as its objective is to find some proposal posterior distribution, a function over some parameter space, as close to the true posterior as possible.

Function	Functional
$f: x \to y$	$g: f \to y$

Variational inference seeks to address a central problem in Bayesian statistics, being the estimation of posterior distributions. Suppose we are interested in the posterior distribution $p(\theta|X)$ given observations X. Bayes rule gives the following:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \tag{1}$$

However, the denominator of the above expression, $p(X)$, is often unknown. We may compute $p(X)$ by marginalizing over θ (2), however, this calculation may become intractable as the dimension of θ grows or when θ or $X|\theta$ take on complex distributions.

$$p(X) = \int_{\theta} p(X|\theta)p(\theta)d\theta \tag{2}$$

Bayesian statistics offers multiple solutions to avoid this problem, including the use of conjugate distributions or sampling methods to estimate $p(\theta|X)$. VI takes a unique approach by treating posterior estimation as an optimization problem. The goal of VI is to find a proposal posterior distribution, $q(\theta)$, as "close" to the true posterior $p(\theta|X)$ as possible. There are several ways to define "close", however, a common choice is the *Kullback-Leibler (KL) Divergence*, $D_{KL}(P||Q)$, between p and q as defined below:

$$D_{KL}(P||Q) = \int_{\theta} p(\theta|X) \log \frac{p(\theta|X)}{q(\theta)} d\theta \tag{3}$$

However, the above equation, being an expectation with respect to p, would require knowing $p(\theta|X)$, the target distribution. As an alternative, one may apply the *reverse KL Divergence*, $D_{KL}(Q||P)$:

$$D_{KL}(Q||P) = \int_{\theta} q(\theta) \log \frac{q(\theta)}{p(\theta|X)} d\theta \tag{4}$$

VI seeks to minimize this measure of dissimilarity. This gives the following objective function:

$$\min_{q} D_{KL}(Q||P) \tag{5}$$

Evidence Lower Bound (ELBO)

We may expand the above expression for $D_{KL}(Q||P) = \int_{\theta} q(\theta) \log \frac{q(\theta)}{p(\theta|X)} d\theta$ as follows:

$$\begin{aligned} D_{KL}(Q||P) &= \mathbb{E}_q(\log q(\theta)) - \mathbb{E}_q(\log p(\theta|X)) \\ &= \mathbb{E}_q(\log q(\theta)) - \mathbb{E}_q(\log p(\theta, X)) + \mathbb{E}_q(\log p(X)) \\ &= \mathbb{E}_q(\log q(\theta)) - \mathbb{E}_q(\log p(\theta, X)) + \log p(X) \end{aligned} \tag{6}$$

Noting that $D_{KL}(Q||P) \geq 0$ and $\log p(X)$ is constant with respect to q, we may reformulate (5) to maximize a quantity known as the evidence lower bound (ELBO).

$$D_{KL}(Q||P) \geq 0 \Rightarrow 0 \geq \mathbb{E}_q(\log q(\theta)) - \mathbb{E}_q(\log p(\theta, X)) + \log p(X)$$

$$\Rightarrow \log p(X) \geq \mathbb{E}_q\left(\log \frac{p(\theta, X)}{q(\theta)}\right) = ELBO \tag{7}$$

The ELBO is named as such as it provides a minimum value for the logarithm of evidence, $p(X)$. This bound is tight if and only if $D_{KL}(Q||P) = 0$, or when Q and P are the same.

Mean-field Approximation

Without further constraint, $q(\theta)$ may take on any arbitrary form. A constraint, the mean-field approximation, is to assert that each of the $\theta_i's$ are conditionally independent in the posterior

distribution. That is, we set $q(\theta) = \prod_{i=1}^{N} q_i(\theta_i)$. The mean-field approximation may greatly reduce the complexity of optimization, facilitating iterative updates to q by maximizing the ELBO with respect to each q_i individually. As described below, the coordinate ascent variational inference algorithm (CAVI) leverages the mean-field approximation to derive an explicit update step for q_i.

Coordinate Ascent Variational Inference (CAVI)

Coordinate Ascent Variational Inference (CAVI) is an optimization technique to find the optimal proposal distribution q. The objective of CAVI is to maximize the ELBO and thereby minimize the KL divergence between q and $p(\theta|X)$. CAVI works by updating each q_i one at a time while holding the remaining $q_{\neg i}$ fixed (the $\neg$ symbol is used to denote "not"). At each update of q_i, CAVI maximizes the ELBO with respect to q_i alone. CAVI repeatedly cycles through the q_i's completing updates until approximate convergence (as indicated by the value of the ELBO). To derive the update step of CAVI, we may begin by manipulating the expression for the ELBO as follows:

$$ELBO = \mathbb{E}_q\left(\log\frac{p(\theta,X)}{q(\theta)}\right) = \int_\theta q(\theta)\log\frac{p(\theta,X)}{q(\theta)}\,d\theta$$

We may then leverage the mean-field approximation to express the ELBO in terms of an expectation with respect to one q_i and the remaining $q_{\neg i}$.

$$= \int_{\theta_i}\int_{\theta_{\neg i}} q(\theta_1)q(\theta_2)\dots q(\theta_n)\log\frac{p(\theta_i|X,\theta_{\neg i})p(X,\theta_{\neg i})}{q(\theta_i)q(\theta_{\neg i})}\,d\theta_{\neg i}d\theta_i$$

$$= \int_{\theta_i} q(\theta_i)\int_{\theta_{\neg i}} q_{\theta_{\neg i}}(\theta_{\neg i})\left(\log\frac{p(\theta_i|X,\theta_{\neg i})}{q(\theta_i)} + \log\frac{p(X,\theta_{\neg i})}{q(\theta_{\neg i})}\right)d\theta_{\neg i}d\theta_i$$

$$= \int_{\theta_i} q(\theta_i)\left(\int_{\theta_{\neg i}} q_{\theta_{\neg i}}(\theta_{\neg i})\left(\log\frac{p(\theta_i|X,\theta_{\neg i})}{q(\theta_i)}\right)d\theta_{\neg i} + c\right)d\theta_i$$

$$= \int_{\theta_i} q(\theta_i)\left(\int_{\theta_{\neg i}} q_{\theta_{\neg i}}(\theta_{\neg i})\log p(\theta_i|X,\theta_{\neg i})\,d\theta_{\neg i} - \log q(\theta_i)\right)d\theta_i + c$$

$$= \int_{\theta_i} q(\theta_i)\log\frac{\exp\left(\mathbb{E}_{q_{\neg i}}[log\, p(\theta_i|X,\theta_{\neg i})]\right)}{q(\theta_i)}\,d\theta_i + c$$

$$= -D_{KL}\left(Q_i||\exp\left(\mathbb{E}_{q_{\neg i}}[\log p(\theta_i|X)]\right)\right) \tag{8}$$

From (8) we see that the ELBO with respect to q_i is equivalent to the negative KL divergence between q_i and $\exp\left(\mathbb{E}_{q_{\neg i}}[\log p(\theta_i|X)]\right)$ plus a constant not involving q_i. We then see that the ELBO is maximized with respect to q_i when the KL divergence between q_i and $\exp\left(\mathbb{E}_{q_{\neg i}}[\log p(\theta_i|X)]\right)$ is zero.

$$\underset{q_i}{\operatorname{argmax}}\, ELBO = q_i^* = \exp\left(\mathbb{E}_{q_{\neg i}}[\log p(\theta_i|X,\theta_{\neg i})]\right) \propto \exp\left(\mathbb{E}_{q_{\neg i}}[\log p(\theta_i,\theta_{\neg i},X)]\right) \tag{9}$$

$$\Rightarrow D_{KL}\left(Q_i^*||\exp\left(\mathbb{E}_{q_{\neg i}}[\log p(\theta_i|\theta_{\neg i},X)]\right)\right) = 0.$$

We can use (9) to then derive explicit updates to q_i depending on the parameterization of q and the generative model, $p(X|\theta)$. Before implementing an example of CAVI, we will first discuss some prominent uses of VI.

VI Applications

There exist many applications of Variational Inference across numerous fields of study [1]. Stochastic VI, an alternative approach in which only portions of the data are used for each update step, has enabled the use of VI in the presence of large quantities of data [1]. One well-known application of VI is for a topic modeling framework called Latent Dirichlet Allocation (LDA) [2]. LDA is an unsupervised NLP machine learning algorithm that seeks to model the distribution of document topics and words given topics. LDA applies VI to train a Bayesian network (a graphical model) in an efficient manner. Another prominent application of VI is the Variational Autoencoder (VAE) machine learning architecture [3]. A VAE is a model that consists of two main pieces, an encoder and decoder. A VAE may be applied to learn the joint distribution of some observed data and latent variables. A VAE parametrizes the variational distribution using a neural network (the encoder), and then seeks to reconstruct the data given the variational distribution. VAE's have found many applications within recent machine learning literature such as the generation of synthetic images.

CAVI Example: Gaussian Mixture Model

We generalize the example given by Blei et. al in 1-dimensional space to fit a Gaussian Mixture model in d-dimensional space via the CAVI algorithm [4]. We attribute the derivations, choice of priors, and certain notational choices to [5].

Suppose we observe N d-dimensional datapoints $X = \{x_1, \dots, x_N\} \in \mathbb{R}^d$. We assume each x_n is drawn from one of K mixture components centered at $\mu_1, \dots, \mu_K$ according to a latent one-hot vector ζ_n of dimension K. We collect the means $\mu_1, \dots, \mu_K$ into a matrix M of dimension d by K and the one-hot vectors $\zeta_1, \dots, \zeta_n$ into a matrix Z of dimension K by N. We adopt a Gaussian prior on the mixture means so that $\mu_i \sim N(0, \mathrm{I})$, a uniform prior over the possible values of Z, and an inverse-Wishart prior over the mixture covariance matrices, Σ_i, so that $\Sigma_i \sim \mathcal{W}^{-1}(\Psi, \nu)$. These priors may be adjusted to match the desired modelling assumptions, although this set leads to convenient derivations and updates.

We then define our variational distribution, $q(\mathrm{M}, \Sigma, \mathrm{Z}, \Phi)$. We assume that $\zeta_n \sim Categorical(\phi_n)$ and adopt a Dirichlet prior on ϕ_n so that $\phi_n \sim Dir(\alpha)$. We apply the mean-field approximation so that $q(\mathrm{M}, \Sigma, \mathrm{Z}, \Phi) = q(M)q(\Sigma)q(\mathrm{Z})q(\Phi)$. Following [6], we may then compute the update rule for $q(\mu_k)$, $q(\Sigma_k)$, $q(\zeta_\mathrm{n})$, and $q(\phi_n)$ for each iteration of CAVI.

We begin deriving the update rules by first decomposing the joint distribution of observed and latent variables.

$$p(X, \mathrm{M}, \Sigma, \mathrm{Z}, \Phi) = p(X|\mathrm{M}, \Sigma, \mathrm{Z})p(\mathrm{Z}|\Phi)p(\Phi)p(\mathrm{M}|\Sigma)p(\Sigma)$$

$$p(X|\mathrm{M},\Sigma,\mathrm{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left(\frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}}\exp\left(-\frac{(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)}{2}\right)\right)^{\zeta_{nk}} \tag{10}$$

$$p(Z|\Phi) = \prod_{n=1}^{N}\prod_{k=1}^{K}\phi_k^{\zeta_{nk}} \tag{11}$$

$$p(\Phi) = \frac{1}{B(\alpha)}\prod_{k=1}^{K}\phi_k^{\alpha_0-1} \tag{12}$$

$$p(\mathrm{M}|\Sigma) = \prod_{k=1}^{K}\frac{1}{\sqrt{(2\pi)^k\det(I)}}\exp\left(-\frac{(\mu_k)^T I(\mu_k)}{2}\right) \tag{13}$$

$$p(\Sigma) = \prod_{k-1}^{K}\frac{|\Psi_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 p}{2}}\Gamma_d\left(\frac{\nu_0}{2}\right)}|\Sigma_k|^{-\frac{(\nu_0+d+1)}{2}}\exp\left(-\frac{1}{2}tr\left(\Psi_0\Sigma_k^{-1}\right)\right) \tag{14}$$

Equations (10) and (13) follow from the pdf of a multivariate normal distribution. Equation (11) follows the pdf of a categorical (or multinomial with one draw) pdf. Equation (12) follows the pdf of a Dirichlet distribution. Finally, equation (14) follows from the pdf of an inverse Wishart distribution. From this set of expressions, we may evaluate $q^*(\theta_i)$ for each variational parameter. Due to length constraints, the full derivations are omitted from this paper, to which we refer to the reader to [5-7].

$$Let\ z_{nk} = \mathbb{E}_{q_{\mathrm{M},\Sigma,\Phi}}\left[-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k) + \log\phi_k\right] + C$$

$$Let\ \gamma_{nk} = \frac{z_{nk}}{\sum_{k=1}^{K} z_{nk}}$$

$$Let\ G_k = \sum_{n=1}^{N}\gamma_{nk}$$

$$Let\ \bar{x}_k = \frac{1}{G_k}\sum_{n=1}^{N}\gamma_{nk}x_n$$

$$Let\ S_k = \frac{1}{G_k}\sum_{n=1}^{N}\gamma_{nk}\,(x_n-\bar{x}_k)(x_n-\bar{x}_k)^T$$

Variational Means Update Rule

$$Let\ \mu_k' = \frac{G}{1+G}\bar{x}_k$$

$$q^*(\mu_k|\Sigma_k) = \mathcal{N}(\mu_k|\mu_k', \frac{1}{1+N}\Sigma_k)$$

The above optimal variational distribution follows a normal distribution with an mean consisting of a weighted combination of the prior mean (in our case being 0 for convenience) and the mean of the datapoints assigned to cluster k. The computation for Σ_k is found below.

Variational Covariances Update Rule

$$Let\ \Psi'_k = \Psi_0 + G_k S_k + \frac{G_k}{1+G_k}\bar{x}_k\bar{x}_k^T$$

$$q^*(\Sigma_k) = \mathcal{W}^{-1}(\Sigma_k|\Psi'_k, \nu_o + G_k)$$

Here, the optimal variational distribution for $q^*(\Sigma_k)$ follows an inverse Wishart distribution resulting from a weighted combination of the prior scale parameter and that given the datapoints assigned to cluster k.

Variational Cluster Assignments Update Rule

$$Let\ \alpha'_k = \alpha_0 + G_k$$

$$q^*(\Phi) = Dir(\alpha')$$

Φ now follows a Dirichlet distribution shrunk towards our prior distribution with parameter α_0.

Variational Cluster Probabilities Update Rule

$$Let\ z_{nk} = \mathbb{E}_{q_{M,\Sigma,\Phi}}\left[-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k) + \log\phi_k\right] + C$$

$$Let\ \gamma_{nk} = \frac{z_{nk}}{\sum_{k=1}^{K} z_{nk}}$$

$$q^*(\mathrm{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\gamma_{nk}^{\zeta_{nk}}$$

The optimal variational cluster assignment distribution follows from the weighted likelihood datapoint x_n lies in cluster k given both the cluster distributions and the categorical distribution parametrized by ϕ_k.

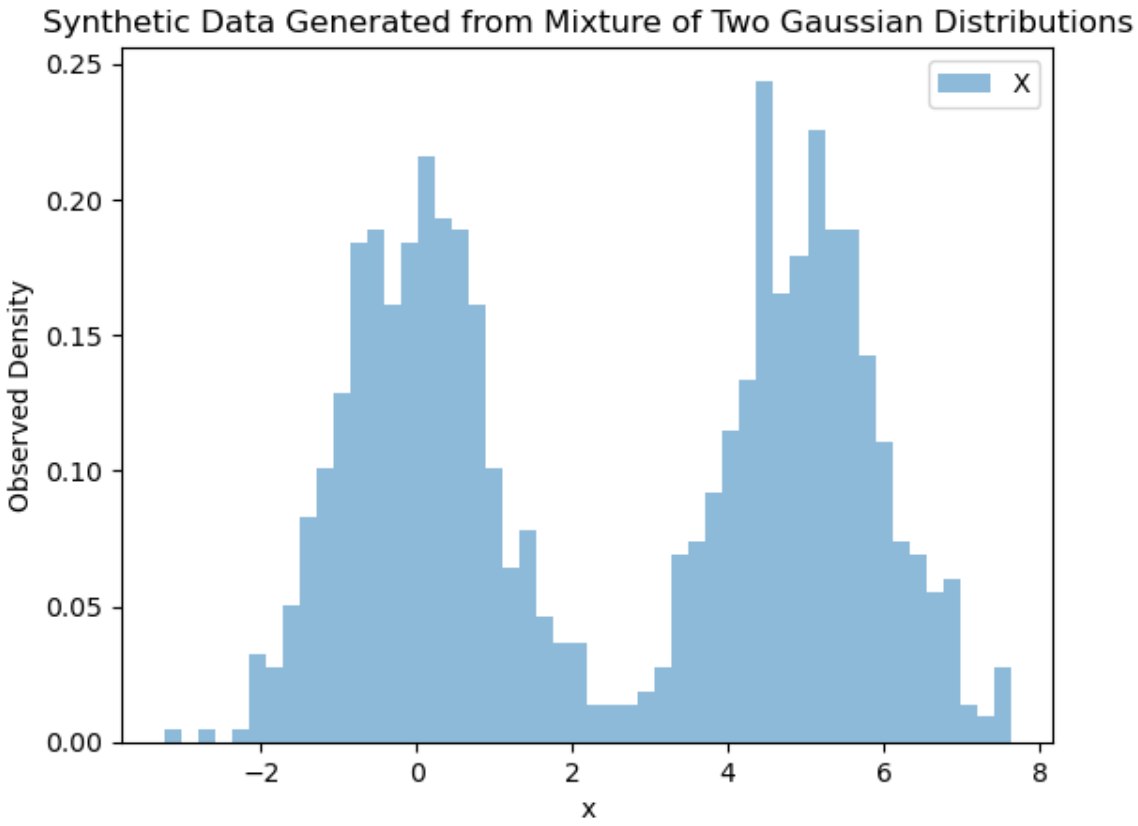
Example: Gaussian Mixture Model via VI Simulation

The above calculations outline the optimal updates for a Gaussian mixture model with an arbitrary number of clusters and dimensionality. While these expressions may not provide immediate intuition towards the VI solution, they demonstrate the feasibility of optimal updates computed by maximizing the ELBO with respect to each variational parameter individually while holding the remaining parameters fixed. Via successive updates (which would also be deterministic assuming the entire dataset is used during each iteration), VI may then converge to some final variational distribution. It is worth noting that the ELBO is unlikely to be convex with respect to the variational parameters, meaning this procedure may converge to a local rather than global optimal

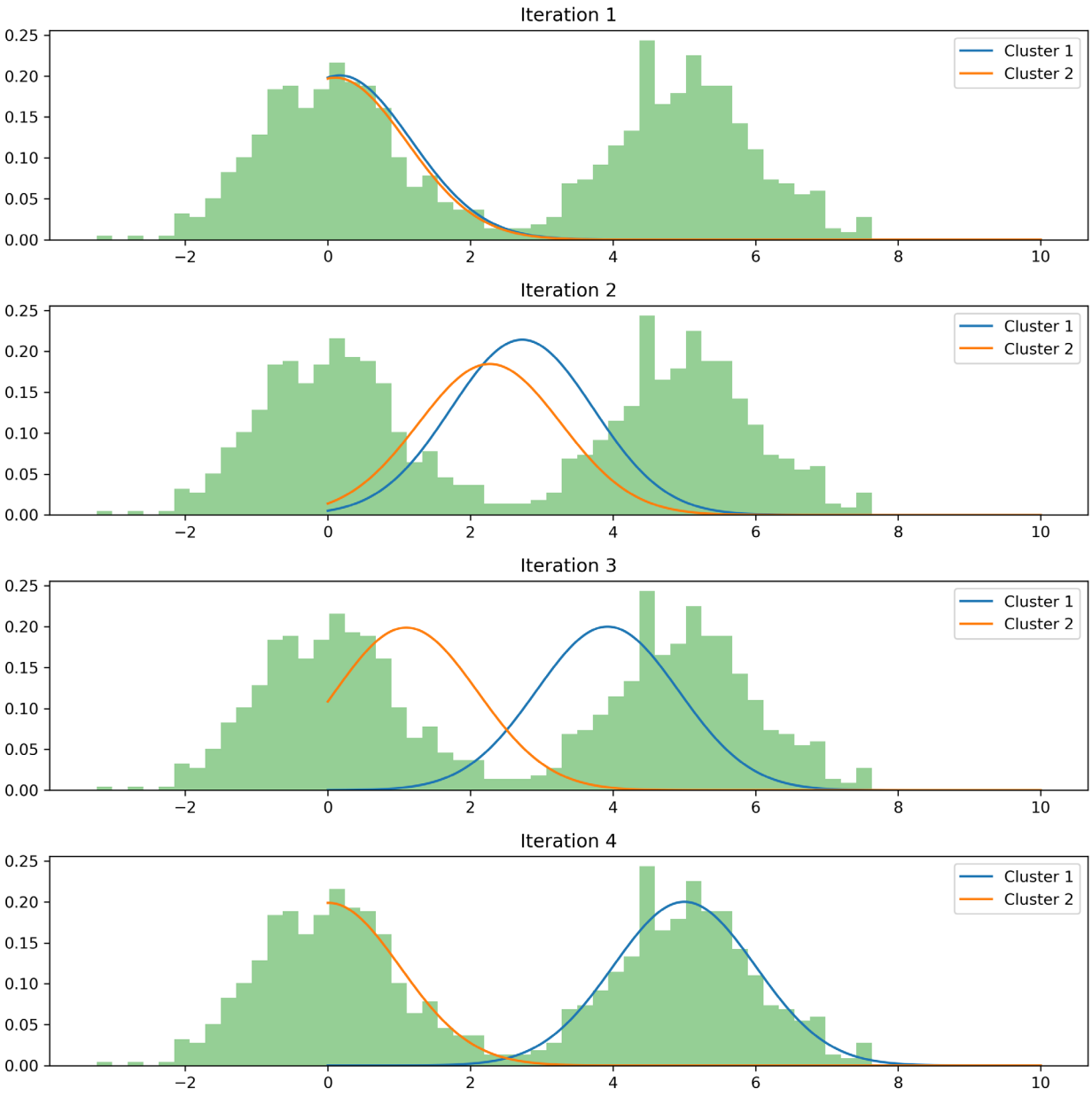
solution. However, this may be remedied by running the VI algorithm several times over randomized initial priors over the variational parameters.

For simplicity and ease of visualization, we demonstrate a working example of fitting a Gaussian Mixture model in one-dimensional space through Variational Inference. We begin by generating sample data according to the following mixture model:

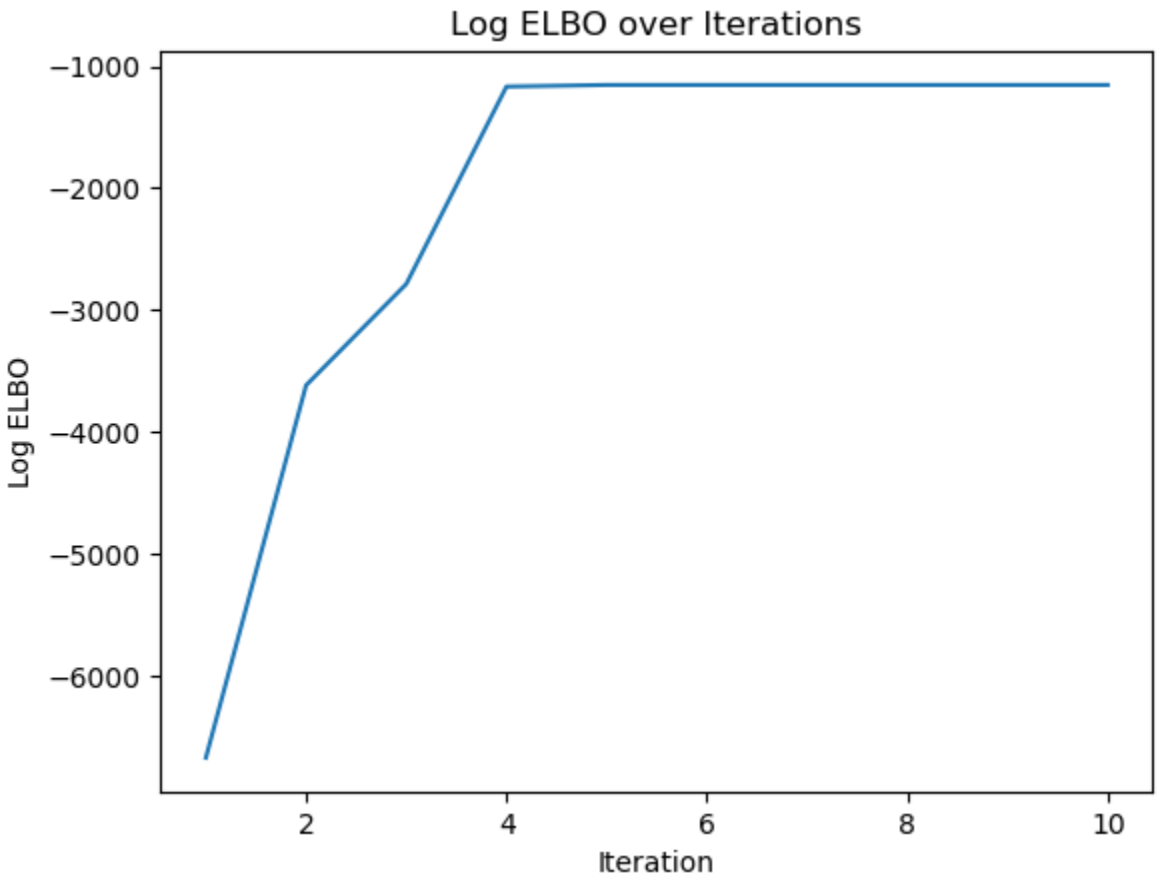
$$x_i \sim \begin{cases} N(\mu_1 = 5, \sigma = 1), & prob. = 1/2 \\ N(\mu_2 = 0, \sigma = 1), & prob. = 1/2 \end{cases}$$



Next, we fit a Gaussian mixture model to the above data with two clusters. In the simulation visualized below, we assume a known fixed cluster variance of 1. The heights of the Gaussian plots indicate our MAP estimates for the cluster assignment probabilities. The Gaussians are centered at the MAP estimates for the cluster centers. The below figures demonstrate gradual convergence over the course of the algorithm.



We may quantify the degree of fit of model to the data by viewing the ELBO over the course of the fitting process. A flatlining ELBO indicates model convergence, meaning we have, at least locally, minimized the KL Divergence between the variational distribution and the target distribution given our modeling assumptions.



As demonstrated above, the model appears to converge after about 4 iterations. This model likely converged exceptionally quickly given its simplicity, which would not necessarily be the case for more complicated models. Further, this model converged to the global optimum in one go, which would also not necessarily hold when implanting CAVI on other models.

References

[1] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in Variational Inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 41, pp. 2008-2026, 2017.

[2] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," 2001. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf.

[3] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *CoRR,* vol. abs/1312.6114, 2013.

[4] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association,* vol. 112, no. 518, pp. 859-877, 2017/04/03 2017, doi: 10.1080/01621459.2017.1285773.

[5] A. Kushwaha. "Variational Inference: Gaussian Mixture model." https://ashkush.medium.com/variational-inference-gaussian-mixture-model-52595074247b (accessed 4/14, 2024).

[6] X. Y. S. Sia. "Coordinate Ascent Mean-field Variational Inference (Univariate Gaussian Example)." https://suzyahyah.github.io/bayesian%20inference/machine%20learning/2019/03/20/CAVI.html (accessed 4/14, 2024).

[7] A. Ganguly and S. W. F. Earp, "An Introduction to Variational Inference," *ArXiv,* vol. abs/2108.13083, 2021.