Upgrading Bayesian Linear Regression with Variational Inference

John Wang
University of Massachusetts
jawang@umass.edu

1 Introduction

1.1 Technical Motivation

In class, we studied MAP estimation given a Gaussian prior and likelihood. The estimator represented the mode of the posterior which remained normal and hence could be calculated analytically. In the infinite dimensional case of Gaussian process regression, all finite subsets of a Gaussian process were multivariate Gaussian. Therefore the posterior, now a distribution over functions, could still be computed as Gaussians are closed under conditioning. However the prior and likelihood need not be Gaussian, leading to intractable posteriors. The issue motivates the need for variational inference (VI) which estimates the posterior with a tractable distribution.

Specifically, VI finds the distribution in a family of tractable distributions that minimizes the KL divergence with the posterior. We require $q^* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(x)|p(x|y))$ where $q \in \mathcal{Q}$ is a tractable distribution, p is the true posterior, x is the latent variable, and y is the observed variable [Mur23]. One choice for a tractable $\mathcal{Q}$ is the normal family. As a result, VI generalizes optimization over function spaces.

1.2 Mathematical Foundations

To start, we familiarize ourselves with the goal of Bayesian inference: computing the posterior distribution. We explain why the posterior is intractable if the prior and likelihood are not conjugate priors. We study the VI objective $q^* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(x)|p(x|y))$ and prove that this minimization problem is equivalent to maximizing the evidence lower bound (ELBO) [BKM17]

$$\mathcal{L}(q) = \mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{p(x,y)}{q(x)} \right) \right].$$

To maximize the ELBO, we derive the coordinate ascent variational inference (CAVI) algorithm. Assuming that our latent variable $X \in \mathbb{R}^p$, we let the components of X be independent under the approximate posterior q so that $q(x) = \prod_{i=1}^{p} q_i(x_i)$. We rewrite the ELBO ($\mathcal{L}(q)$) in terms of the $q_i(x_i)$ and take functional derivatives with respect to $q_i(x_i)$ to get the optimal $q_i(x_i)^* \propto \exp(\mathbb{E}_{q-i}[\log p(x,y)])$ where $q - i(x) = \prod_{j \neq i} q_j(x_j)$ [BN06]. However, $q_i(x_i)^*$ cannot be determined directly as we do not know the other $q_j(x_j)^*$ beforehand. What follows is an iterative algorithm in the spirit of gradient descent where $q_i(x_i)$ is updated with the current value of $\exp(\mathbb{E}_{q-i}[\log p(x,y)])$ until convergence [BKM17]. Additionally, we show that CAVI converges to a local optimum as the ELBO never decreases after an update.

1.3 Objectives

Our first task is to equate minimizing the KL objective to maximizing the ELBO ($\mathcal{L}(q)$) [BKM17]. The second main mathematical result is the derivation of the CAVI update. This is an iterative algorithm similar to gradient descent where $q(x)$ is updated in the direction of the gradient of $\mathcal{L}(q)$ [BKM17].

After the explanation of the CAVI algorithm, we verify it empirically through two experiments. For the first, we apply the algorithm to the setup of Bayesian linear regression. Let the design matrix be $A \in \mathbb{R}^{n \times p}$ and the labels be $Y \in \mathbb{R}^n$. The data model be defined by $Y = A\theta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The prior distribution is given by a multivariate Gaussian $\theta \sim \mathcal{N}(0, \tau^2 I)$. We calculate the true posterior $p(\theta|Y)$ analytically. Then, we use the mean-field Gaussian

to approximate $p(x|y)$ using CAVI [Mur23]. To explain the results, we compare posteriors and plot the ELBO as a function of time.

In the second experiment, we use CAVI to approximate an intractable posterior. In this scenario, we "upgrade" Bayesian linear regression by estimating τ and σ in addition to θ. To approximate τ and σ, we use the gamma family as they have non-negative support. We use Python and PyTorch to conduct the experiments.

2 Formulation

2.1 Introduction to Bayesian Inference

The world of Bayesian statistics relies on a key formula known as Bayes' theorem. If A and B are events, Bayes' theorem states that

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\,\mathbb{P}[A]}{\mathbb{P}[B]}$$

Bayesian statistics takes Bayes' theorem and upgrades events to probability distributions over random variables. If X and Y are random variables on possibly different spaces

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

This result just follows from Bayes' theorem, but it is more complex then it appears. $p(x)$ is the marginal distribution of X and lives on the space of X. $p(y)$ is the marginal distribution of Y and lives on the space of Y. $p(x|y)$ is the conditional distribution of X given Y and lives on the space of X. $p(y|x)$ is the conditional distribution of Y given X and lives on the space of Y.

Bayesians use the upgraded version of Bayes' theorem to predict a *latent variable* X with samples from an *observed variable* Y. We assume the latent variable is related to the observed variable but is hidden. For example, suppose we observe symptoms of flu patients and would like to infer whether they are sick. Let Y be a 3-dimensional random vector depicting a patient's symptoms, namely their temperature, whether they have a sore throat, and whether they have a cough. Thus, Y has observation space $\mathbb{R} \times \{0, 1\}^2$. X is the binary random variable indicating whether the patient has a flu.

The goal is to determine the *posterior* $p(x|y)$ to predict whether flu patients are sick given their symptoms. Applying Bayes' theorem, we must first determine the *prior* $p(x)$. As an aside, this terminology highlights another way of describing Bayes' theorem as a way of improving our prior belief about X using relevant information from Y. We assume a reasonable distribution for $p(x)$ based on past data. The probability that a patient being sick, all else being equal, can be estimated by dividing the number of sick patients by the total patients in the dataset. The *likelihood* $p(y|x)$ can also be estimated using historical data. For example, given a sick patient we can determine the probability they have symptoms with similar Monte Carlo estimates. In most Bayesian inference problems, we assume the prior and likelihood are known.

Additionally, the *joint* $p(x, y) = p(y|x)p(x)$ is computable as we know the prior and likelihood. The last distribution needed to solve for $p(x|y)$ is the *evidence* $p(y)$. The evidence must be computed analytically according to the model we have described for the prior and likelihood for the posterior to be valid. We can solve for $p(y)$ with

$$p(y) = \int p(x, y)\,\mathrm{d}x$$

Integrating the joint over the latent variable X is rarely tractable. To see why consider the case when X is high dimensional. Suppose instead of just predicting whether a patient is sick, we want to know the duration of the sickness and whether the patient is contagious. Integrating over a high dimensional space is rarely tractable! Exceptions include the normal distribution which is closed under conditioning and the binomial-beta conjugate priors.

We can view the above integral as the normalization constant for $p(x|y)$. If we fix $y = y_0$, $p(y = y_0) = \int p(x, y = y_0)\,\mathrm{d}x$ is simply the area under the curve of $p(x, y = y_0)$. Dividing by the area is exactly what is needed for $p(x|y)$ to integrate to 1. As a result, it is possible to numerically solve for $p(y)$ given a particular value for y. However, this can be too expensive if X is high dimensional. Nevertheless, $p(x|y) \propto p(x, y)$ so given a value of $y = y_0$,

$$\arg\max_x p(x|y = y_0) = \arg\max_x p(x, y = y_0)$$

Therefore, we can determine the mode of the posterior with the joint. This is known as the MAP estimator.

In closing, here a well-known example of Bayesian inference we have seen in class. In the one dimensional case, we are given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ and model $y_i = \mu x_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. In this problem, μ is our latent variable and we are trying to solve for $p(\mu|\mathcal{D})$. This model assumes that for all $i \in [n]$, $y_i|x_i, \mu \sim \mathcal{N}(\mu x_i, \sigma^2)$. We also assume that the prior $p(\mu) = \mathcal{N}(\mu_0, \tau_0^2)$. In the case of Bayesian linear regression, the posterior is again normal as we have seen in class.

2.2 Motivation for Variational Inference

Variational inference (VI) proposes a solution to an incomputable posterior $p(x|y)$ caused by an intractable evidence. Instead of solving for the posterior, we find a computable surrogate distribution that approximates the posterior. More specifically, we pick the surrogate distribution within a tractable family of distributions which best approximates the posterior in the KL divergence. This yields the objective

$$\arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(x)|p(x|y)) = \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{x \sim p(x)} \left[\log \left(\frac{q(x)}{p(x|y)} \right) \right]$$

where $q(x)$ is the surrogate in a space of tractable functions $\mathcal{Q}$. For all distributions $q(x)$ and $p(x|y)$, $\mathrm{KL}(q(x)|p(x|y)) \geq 0$ with equality when $q(x) = p(x|y)$. The term variational implies we are optimizing over a function space. We are using the approximation of the true posterior to infer something about our latent variable X. However, our objective is flawed because it is written in terms of $p(x|y)$ which we cannot compute. In fact, our goal was to find $p(x|y)$ in the first place! Knowing the prior and likelihood, we can try to reframe the objective exclusively in terms of the joint $p(x, y)$.

$$\begin{aligned}
\arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(x)|p(x|y)) &= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{q(x)}{p(x|y)} \right) \right] \\
&= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{q(x)p(y)}{p(x, y)} \right) \right] \\
&= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{q(x)}{p(x, y)} \right) \right] + \mathbb{E}_{x \sim q(x)}[\log(p(y))] \\
&= \arg\min_{q \in \mathcal{Q}} -\mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{p(x, y)}{q(x)} \right) \right] + \log(p(y)) \\
&= \arg\max_{q \in \mathcal{Q}} \mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{p(x, y)}{q(x)} \right) \right]
\end{aligned}$$

As a result, we have written the minimizing problem as a maximization of the Evidence Lower Bound (ELBO), $\mathcal{L}(q) = \mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{p(x,y)}{q(x)} \right) \right]$. The ELBO is computable since we know the surrogate and joints distributions are computable. The meaning of the ELBO comes from the fact that

$$\mathrm{KL}(q(x)|p(x|y)) = -\mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{p(x, y)}{q(x)} \right) \right] + \log(p(y))$$

Rearranging we have

$$\begin{aligned}
\mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{p(x, y)}{q(x)} \right) \right] &= -\mathrm{KL}(q(x)|p(x|y)) + \log(p(y)) \\
\mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{p(x, y)}{q(x)} \right) \right] &\leq \log(p(y))
\end{aligned}$$

It is insightful that the ELBO is typically more negative than the log of the evidence. The difference between the ELBO and the evidence is given by the KL divergence.

2.3 Solving for Surrogates

In the previous section, we saw that our VI objective of minimizing the KL divergence between a surrogate distribution and the true posterior is equivalent to maximizing the ELBO.

$$\mathcal{L}(q) = \mathbb{E}_{x \sim q(x)} \left[\log \left(\frac{p(x, y)}{q(x)} \right) \right] = \int \log \left(\frac{p(x, y)}{q(x)} \right) q(x) \, \mathrm{d}x$$

The resulting objective remains challenging to reason about. First of all, we are still trying to solve an optimization problem over a function space. In most cases, this is infeasible and we must restrict the tractable family Q to a

parametric family such as a normal or exponential. Even if we are working with a familiar family of distributions, we may have difficulty computing the ELBO, especially in real life scenarios. As a work around, we can determine an approximation via sampling. The hard work pays off at the end, as empirical results show with minimal sampling we can converge to the optimal surrogate $q(x)$.

To get a better feel for the optimization problem, we look at an example where the ELBO can be found analytically. Let the prior be defined by $X \sim \text{Exponential}(\lambda = 1)$ and the likelihood by $Y|X = x \sim \mathcal{N}(\mu = x, \sigma^2 = 1)$. According to our model, given a value $X = x$ we have the mean for the normal distribution of Y. Then for $x \geq 0$

$$p(x) = \exp(-x)$$

and for $y \in \mathbb{R}$

$$p(y|x) = \frac{1}{2\pi} \exp\left(-\frac{(y-x)^2}{2}\right)$$

The joint for $x \geq 0$, $y \in \mathbb{R}$

$$p(x, y) = \exp(-x) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-x)^2}{2}\right)$$

Since the support of X is non-negative, let $\mathcal{Q}$ be the family Exponential(λ). Plugging our distributions into the objective

$$\arg\max_{q \in \mathcal{Q}} \mathbb{E}_{x \sim q(x)} \left[\log\left(\frac{p(x, y)}{q(x)}\right)\right]$$

we have

$$\begin{aligned}
&\arg\max_{\lambda > 0} \mathbb{E}_{x \sim q(x)} \left[\log\left(\frac{\exp(-x) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-x)^2}{2}\right)}{\lambda \exp(-\lambda x)}\right)\right] \\
&= -\mathbb{E}[x] - \frac{1}{2}\mathbb{E}[(y-x)^2] + \lambda \mathbb{E}[x] - \log \lambda + \text{const} \\
&= -\mathbb{E}[x] - \frac{1}{2}(-2y\mathbb{E}[x] + \mathbb{E}[x^2]) + \lambda \mathbb{E}[x] - \log \lambda + \text{const} \\
&= -\frac{1}{2}\mathbb{E}[x^2] + (y + \lambda - 1)\mathbb{E}[x] - \log \lambda + \text{const} \\
&= -\frac{1}{\lambda^2} + (y-1)\frac{1}{\lambda} - \log \lambda + \text{const}
\end{aligned}$$

Maximizing for λ we have

$$\begin{aligned}
\frac{2}{\lambda^3} - (y-1)\frac{1}{\lambda^2} - \frac{1}{\lambda} &= 0 \\
2 - (y-1)\lambda - \lambda^2 &= 0 \\
\lambda &= \frac{(y-1) \pm \sqrt{(y-1)^2 + 8}}{-2}
\end{aligned}$$

We pick the λ that satisfies the constraint $\lambda > 0$ depending on the value of y.

2.4 Mean-Field Approach

In this section, we generalize our choice for a family of distributions $\mathcal{Q}$. From now on suppose the latent variable X is in $\mathbb{R}^p$ and the observed variable Y is in $\mathbb{R}^d$. With the mean field assumption we assume that our surrogate distribution factors as a product of the component distributions.

$$q(x) = \prod_{i=1}^{p} q_i(x_i)$$

In this way, the components of X are independent under the surrogate distribution. This factorization is the only requirement of $\mathcal{Q}$. The factor distributions need not live within the family of distributions $\mathcal{Q}$. The factor distributions need not live in the same family of distributions. An example of a family that satisfies the mean-field assumption is a multivariate normal that factorizes as the product of univariate normals.

Now, we try to maximize the ELBO over a mean-field family. We will be able to separate the component distributions and solve for them independently. To simplify the computation, we let X be in $\mathbb{R}^3$.

$$
\begin{aligned}
&\mathbb{E}_{x\sim q(x)}\left[\log\left(\frac{p(x,y)}{q(x)}\right)\right] \\
&= \int \log\left(\frac{p(x,y)}{q(x)}\right) q(x)\,\mathrm{d}x \\
&= \int_{x_1}\int_{x_2}\int_{x_3} q_1q_2q_3(\log p - \log q_1 - \log q_2 - \log q_3)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 \\
&= \int_{x_1}\int_{x_2}\int_{x_3} q_1q_2q_3 \log p\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 - \int_{x_1}\int_{x_2}\int_{x_3} q_1q_2q_3(\log q_1 + \log q_2 + \log q_3)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 \\
&= \int_{x_1} q_1\left(\int_{x_2}\int_{x_3} q_2q_3 \log p\,\mathrm{d}x_2\,\mathrm{d}x_3\right)\mathrm{d}x_1 - \int_{x_1} q_1\left(\int_{x_2}\int_{x_3} q_2q_3(\log q_1 + \log q_2 + \log q_3)\,\mathrm{d}x_2\,\mathrm{d}x_3\right)\mathrm{d}x_1
\end{aligned}
$$

Next, we can try to isolate $\log q_1$ in the second term like we did for $\log p$ above.

$$
\begin{aligned}
&- \int_{x_1} q_1\left(\int_{x_2}\int_{x_3} q_2q_3 \log q_1\,\mathrm{d}x_2\,\mathrm{d}x_3\right)\mathrm{d}x_1 - \int_{x_1} q_1\left(\int_{x_2}\int_{x_3} q_2q_3(\log q_2 + \log q_3)\,\mathrm{d}x_2\,\mathrm{d}x_3\right)\mathrm{d}x_1 \\
&= -\int_{x_1} q_1 \log q_1\,\mathrm{d}x_1 \cdot \int_{x_2} q_2q_3\,\mathrm{d}x_2\,\mathrm{d}x_3 - \int_{x_1} q_1\,\mathrm{d}x_1 \cdot \int_{x_2}\int_{x_3} q_2q_3(\log q_2 + \log q_3)\,\mathrm{d}x_2\,\mathrm{d}x_3 \\
&= -\int_{x_1} q_1 \log q_1\,\mathrm{d}x_1 - \int_{x_3} q_2q_3(\log q_2 + \log q_3)\,\mathrm{d}x_2\,\mathrm{d}x_3
\end{aligned}
$$

Combining terms we have

$$
\int_{x_1} q_1\left(\int_{x_2}\int_{x_3} q_2q_3 \log p\,\mathrm{d}x_2\,\mathrm{d}x_3 - \log q_1\right)\mathrm{d}x_1 - \int_{x_3} q_2q_3(\log q_2 + \log q_3)\,\mathrm{d}x_2\,\mathrm{d}x_3
$$

Let $\mathbb{E}_{q-1}[\cdot] = \mathbb{E}_{(x_2,x_3)\sim(q(x_2),q(x_3))}[\cdot]$ so that we have the expectation of (x_2, x_3) under the joint law $(q(x_2), q(x_3))$. Then we have

$$
\int_{x_1} q_1(\mathbb{E}_{q-1}[\log p] - \log q_1)\,\mathrm{d}x_1 - \mathbb{E}_{q-1}[\log q_2 + \log q_3]
$$

Notice that only the first term is in terms of q_1. If we treat q_2 and q_3 as constants, we will be able to optimize for q_1 independently. Since we are optimizing over functions we need to take a functional derivative. We would like to maximize the ELBO with respect to the function q_1. This results in a definition similar to directional derivatives.

$$
\frac{\delta\mathcal{L}(q)}{\delta q_1} = \left.\frac{\mathrm{d}\mathcal{L}(q+\epsilon\phi)}{\mathrm{d}\epsilon}\right|_{\epsilon=0}
$$

We will not apply the definition directly. Instead, we apply the Euler-Lagrange equation. Treat the second term as a constant. Since the integrand in the first term is only in terms of q_1 the functional derivative is equivalent to the derivative of the integrand with respect to q_1. To maximize we set this derivative equal to 0.

$$
\mathbb{E}_{q-1}[\log p] - \log q_1 - 1 = 0
$$

$$
\log q(x_1) \propto \mathbb{E}_{q-1}[\log p(x,y)]
$$

This answer is reasonable because $\mathbb{E}_{q-1}[\log p(x,y)]$ will be a function of x_1. In the last line, we disregard the constant 1 and consider proportionality because $\mathbb{E}_{q-1}[\log p(x,y)]$ must integrate to 1 to be a valid probability distribution. So, we just need to find the normalization constant using information about the factor distribution family. We can generalize this to X in $\mathbb{R}^p$ so that for $i \in [p]$

$$
\log q(x_i) \propto \mathbb{E}_{q-i}[\log p(x,y)]
$$

3 Results

3.1 Applying the Mean Field Result and CAVI

The result of conducting VI on a mean field family is clean yet unsatisfying. The optimal q_i is expressed as an expectation over the rest of the q_j where $j \neq i$. Assuming we do not know the optimal q_j beforehand, we would not be able to determine q_i. However, suppose that we set prior distributions for $q_1, \ldots, q_m$ in the spirit of Bayesian inference.

Then, we perform updates with respect to our priors starting from q_1 all the way to q_p. Since our priors are not optimal, our updated distributions are not either. As a result, we repeat the updates k times to improve our estimates. This is known as the coordinate ascent variation inference algorithm (CAVI) and closely resembles gradient descent.

Sounds reasonable but does it work? We can show that CAVI increases the ELBO after every update. Since it is bounded above by the log-evidence, CAVI converges to a local optimum. To show why CAVI always moves towards the optimum, we rewrite the ELBO as a functional of q_i by fixing all q_j for $j \neq i$

$$\mathcal{L}(q) = \mathbb{E}_{x\sim q(x)}[\log p(x,y)] - \mathbb{E}_{x\sim q(x)}[\log q(x)] = \mathbb{E}_{x_i\sim q_i(x_i)}[\mathbb{E}_{q-i}[\log p(x,y)]] - \mathbb{E}_{x_i\sim q_i(x_i)}[\log q_i(x_i)] + \text{const}$$

Now, let $f(x_i) = \mathbb{E}_{q-i}[\log p(x,y)]$. Then, $\mathcal{L}(q_i) = \mathbb{E}_{x_i\sim q_i(x_i)}[f(x_i)] - \mathbb{E}_{x_i\sim q_i(x_i)}[\log q_i(x_i)] + \text{const}$. Replacing $f(x_i)$ with $\log\exp f(x_i)$ we have

$$\mathcal{L}(q_i) = \mathbb{E}_{x_i\sim q_i(x_i)}[\log\exp f(x_i)] - \mathbb{E}_{x_i\sim q_i(x_i)}[\log q_i(x_i)] + \text{const} = \mathbb{E}_{x_i\sim q_i(x_i)}\left[\log\left(\frac{\exp f(x_i)}{q_i(x_i)}\right)\right] + \text{const}$$

Suppose $z = \int \exp f(x_i)\,\mathrm{d}x_i$ is the normalization constant of $\exp f(x_i)$. After multiplying and dividing by z

$$\begin{aligned}
\mathcal{L}(q_i) &= \mathbb{E}_{x_i\sim q_i(x_i)}\left[\log\left(\frac{\exp f(x_i)/z}{q_i(x_i)}\cdot z\right)\right] + \text{const} \\
&= \mathbb{E}_{x_i\sim q_i(x_i)}\left[\log\left(\frac{\exp f(x_i)/z}{q_i(x_i)}\right)\right] + \text{const} \\
&= -\text{KL}(q_i(x_i)||\exp f(x_i)/z) + \text{const}
\end{aligned}$$

To maximize $\mathcal{L}(q_i)$, we would like to decrease $\text{KL}(q_i(x_i)||\exp f(x_i)/z)$ since it is non-negative. Therefore, we set $\text{KL}(q_i(x_i)||\exp f(x_i)/z) = 0$ which is true iff $q_i(x_i) = \exp f(x_i)/z$. This is exactly the CAVI update. Now that we have seen that CAVI converges, we would like the algorithm to be significantly less expensive than a numerical calculation of the posterior. We hope to show some empirical results in the next section.

3.2 Experiments

3.2.1 CAVI Estimate in Bayesian Linear Regression

In this experiment, we adopt the setup of Bayesian linear regression. The data model is given by $Y = A\theta + \epsilon$ where $\theta \sim \mathcal{N}(0, \tau^2 I_p)$ is in $\mathbb{R}^p$, $A \in \mathbb{R}^{n\times p}$ is the data matrix, and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is in $\mathbb{R}^n$. We assume τ and σ are known. Then we have the prior $p(\theta) = \mathcal{N}(0, \tau^2 I_p)$ and likelihood $p(Y|\theta) = \mathcal{N}(A\theta, \sigma^2 I_n)$.

In Bayesian linear regression the posterior $p(\theta|Y)$ can be computed analytically. Specifically the posterior is normal with

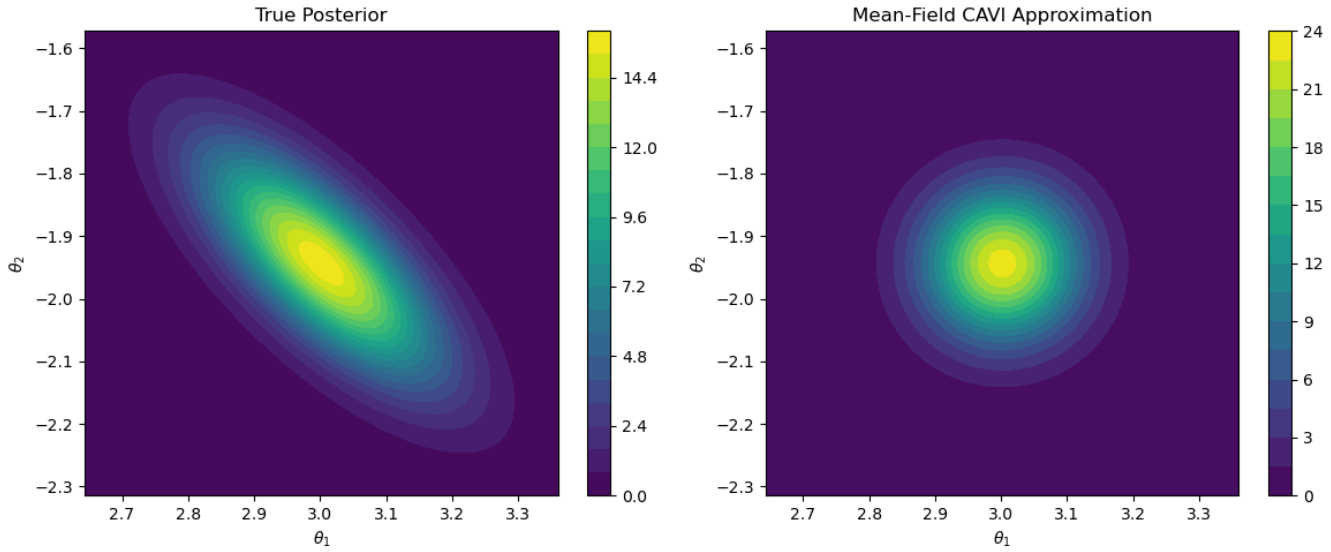
$$\mathbb{E}[\theta|Y] = \left(A^\top A + \frac{\sigma^2}{\tau^2} I_p\right)^{-1}(A^\top Y)$$

$$\text{Var}[\theta|Y] = \frac{1}{\tau^2} I_p + \frac{1}{\sigma^2} A^\top A$$

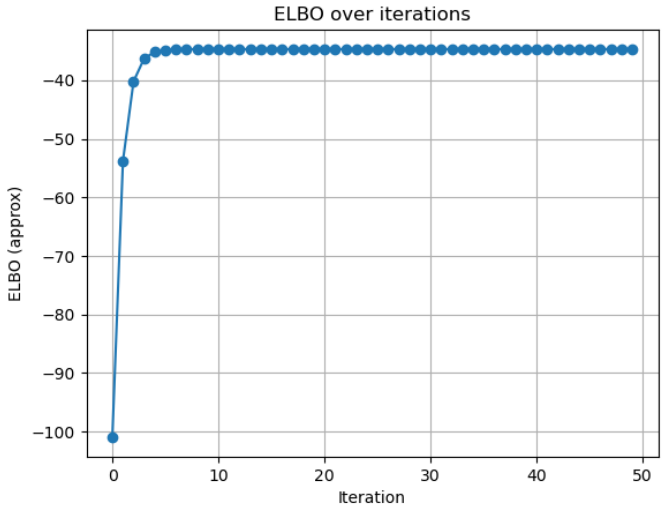
Note that the mean of the posterior is equal to the mode because the normal distribution is symmetrical. As a result, the MAP estimator of the posterior is equivalent to the solution of ridge regression with $\lambda = \frac{\tau^2}{\sigma^2}$.

In this experiment, we generate $n = 50$ samples with $p = 2$ features. For $i \in [n]$, x_i is generated from a multivariate normal distribution with mean $\begin{bmatrix}0\\0\end{bmatrix}$ and covariance $\begin{bmatrix}1.0 & 0.8\\0.8 & 1.0\end{bmatrix}$. Therefore, the two features are positively correlated. We also generate $y_i = x_i^\top \theta_{\text{true}} + \epsilon_i$ where $\theta_{\text{true}} = \begin{bmatrix}3.0\\-2.0\end{bmatrix}$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. We assume that $\sigma = 0.5$ and $\tau = 1.0$.

The goal is to compare the true posterior given the data $p(\theta|Y)$ to an estimate using the CAVI algorithm. Since we do not know the distribution of the posterior beforehand, we let $q(x)$ be a standard multivariate normal distribution $(q_1, q_2 = \mathcal{N}(0,1))$. Note that this choice of $q(x)$ satisfies the mean-field assumption. After running the CAVI algorithm, we can compare the distribution $q(x)$ to the true posterior and graph the convergence of the ELBO.



The true posterior was normal with mean $\begin{bmatrix} 3.00216505 \\ -1.94337853 \end{bmatrix}$ and covariance $\begin{bmatrix} 0.01426359 & -0.01082542 \\ -0.01082542 & 0.01530459 \end{bmatrix}$. The CAVI surrogate distribution was normal with $\begin{bmatrix} 3.00216505 \\ -1.94337853 \end{bmatrix}$ and covariance $\begin{bmatrix} 0.00660643 & 0.0 \\ 0.0 & 0.00708859 \end{bmatrix}$. The mean of the surrogate matches with the true posterior but what about the covariance matrix? The mean-field assumption forces θ_1 and θ_2 to be independent so the correlations must be 0. In general, the mean field approach tries to estimate the true posterior without modeling correlations.



In this graph, one iteration signifies that $q_1(\theta_1)$ and $q_2(\theta_2)$ are updated once. Observe that the ELBO is negative because it must be less than the log-evidence which is negative. The ELBO increases and converges after a few iterations. This is the desired behavior as we want to maximize the ELBO.

3.2.2 Upgrading Bayesian Linear Regression

In the first experiment, we tried to analyze how well the CAVI surrogate estimated a posterior that could be calculated analytically. Now is the time to estimate a intractable posterior. Consider the same setup as the previous experiment. However, suppose we are unsure about the values of τ and σ. We still set the true $\sigma = 0.5$ like the first experiment. Then, τ and σ join θ as latent variables and our surrogate will be given by

$$q(\theta, \tau, \sigma) = q(\theta_1)q(\theta_2)q(\tau_1^{-2})q(\tau_2^{-2})q(\sigma^{-2})$$

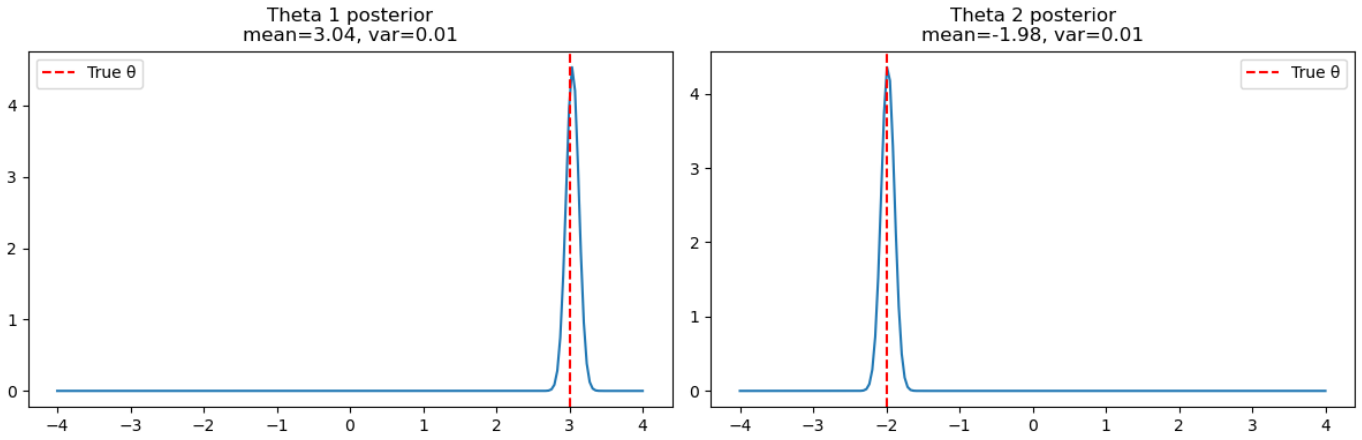
In this case, τ_1 may not necessarily be equal to τ_2. Also, we choose to estimate distributions for τ_1^{-2}, τ_2^{-2}, and σ^{-2} for computation reasons (conjugate priors). We initialize

$$q(\theta_1), q(\theta_2) \sim \mathcal{N}(0, 1)$$

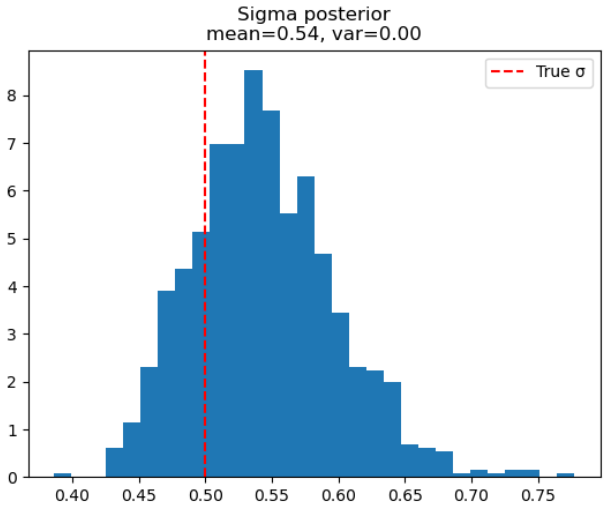
$$q(\tau_1^{-2}), q(\tau_2^{-2}) \sim \text{Gamma}(\alpha = 1, \beta = 1)$$

$$q(\sigma^{-2}) \sim \text{Gamma}(\alpha = 1, \beta = 1)$$

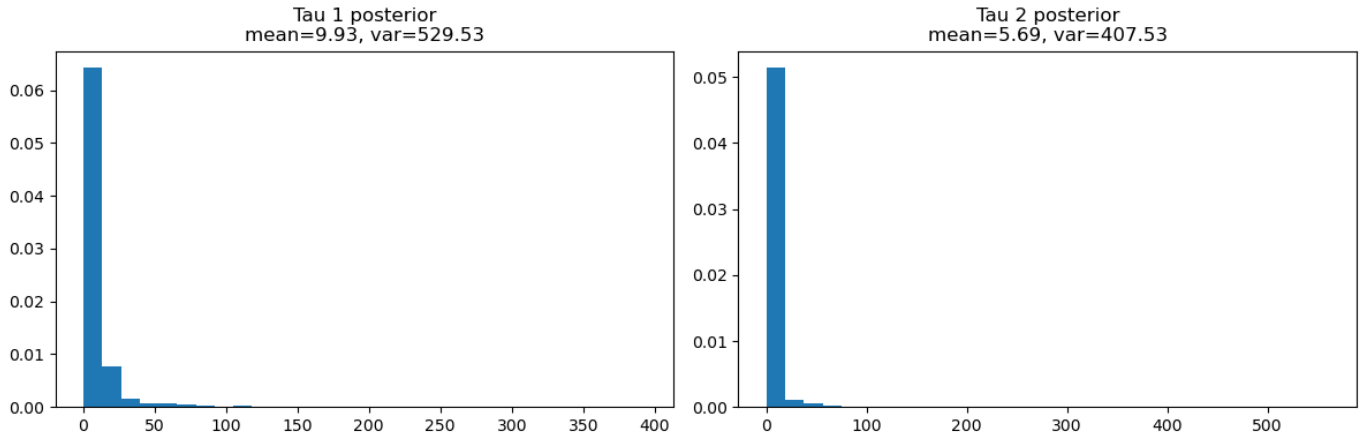
Since we now have a product of normals and gammas our true posterior is intractable. We conduct the analysis by comparing the distributions of θ and σ to the true values. Also, we examine at the distribution of τ and the ELBO curve.



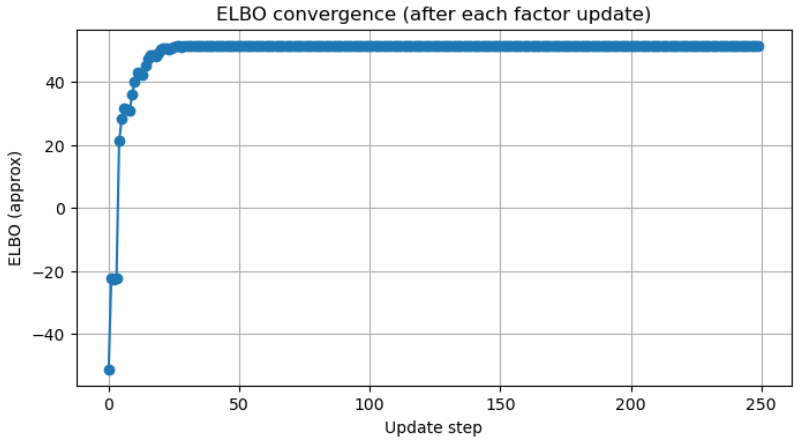
The CAVI surrogates for θ_1 and θ_2 are normal with mean close the true values and small variance.



The CAVI surrogates for σ^{-2} is gamma. After applying a non-linear transformation we obtain a distribution for σ which has mean close to the true value and small variance.



The CAVI surrogate distribution for τ_1^{-2} and τ_2^{-2} is gamma. After applying a non-linear transformation we obtain a distribution for τ_1 and τ_2. τ_1 and τ_2 depict how much our features are allowed to vary. In other words, it is a measure of feature relevance. Since both features are used directly in our data model, we expect τ_1 and τ_2 to be non-zero. The results confirm our intuition.



Lastly, as in the first experiment the ELBO converges to an maximum value. In this graph, we plot the ELBO after every update to a factor distribution.

4 Conclusion

In practice, obtaining a posterior distribution over latent variables is often preferable to relying on a simple MAP estimate. A posterior distribution enables the computation of expectations of arbitrary functions of the latent variables, providing a richer characterization than a single point estimate. Uncertainty can be quantified through posterior variances or credible intervals derived from posterior quantiles. Moreover, predictions can be made by integrating over the posterior distribution, yielding posterior predictive distributions that properly account for parameter uncertainty. These quantities may be computed analytically when closed-form expressions are available, or approximated numerically via sampling from the posterior.

Throughout this report, we examined how variational inference approximates posterior distributions. Under the mean-field assumption, we derived optimal variational distributions for each factor in the surrogate posterior. These distributions cannot be computed directly, as each depends on the expectations of the others, leading naturally to an iterative coordinate ascent variational inference (CAVI) algorithm. The algorithm is implemented by initializing the variational factors and successively updating each factor until convergence. To analyze the behavior of CAVI, we first considered Bayesian linear regression and compared the variational approximation to the true posterior. We then extended the model by treating the prior and likelihood standard deviations, τ and σ, as latent variables and deriving their corresponding variational posteriors. In this hierarchical setting, the true posterior becomes intractable, while

CAVI provides a computationally efficient approximation using normal variational factors θ and gamma factors for τ and σ.

Variational inference under the mean-field assumption has notable limitations. By assuming independence among latent variables, the variational posterior cannot capture correlations present in the true posterior, which may lead to underestimated uncertainty. Additionally, the ELBO objective is generally non-convex, making CAVI sensitive to initialization and prone to local optima. Future methodological work could address these issues by exploring richer variational families, improving convergence strategies, or comparing variational approximations with MCMC samples to assess their accuracy.

Building on the motivations discussed earlier, future work could explore practical uses of posterior distributions in hierarchical and Bayesian linear regression models. This might include leveraging posterior predictive distributions to make informed predictions, studying the influence of hierarchical priors, or applying posterior information to guide real-world decision-making.

References

[BKM17] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[BN06] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[Mur23] Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.