Bayesian Optimal Design - Weekly Report

Rasmus Hag Løvstad, pgq596

May 8, 2023

1 Week 3

1.1 Objective

The objective of this week is to learn about and implement variational inference for Bayesian linear regression. Then we will study convergence rates and accuracy by playing around with different parameters.

1.2 Theory

The point of variational inference is to approximate the posterior distribution $p(\theta|\mathbf{y}, \mathbf{d})$ using optimization techniques. The space of which to optimize in is the parameter space for some distribution $q(\theta)$, which we aim to make as close to the posterior as possible by KL-divergence. For reference, we will recite the definition of our linear model:

$$\theta \sim \mathcal{N}(\mu, \mathbf{\Sigma})$$

$$\epsilon \sim \mathcal{N}(0, \sigma_{\mathbf{y}}^2)$$

$$\mathbf{y}|\mathbf{d}, \theta \sim \theta^T \mathbf{d} + \epsilon$$

The optimization problem we wish to solve is defined like so:

$$q^*(\theta) = \arg \min_{q(\theta) \in \mathcal{L}} \mathrm{KL}(q(\theta)||p(\theta|\mathbf{y}, \mathbf{d}))$$

Computing the KL-divergence is hard though, because it requires computing the evidence $p(\mathbf{y}, \mathbf{d})$, so instead we try to optimize in regards to the expectation lower bound (ELBO):

TODO: show why

$$\mathrm{ELBO}(q) = \mathbb{E}[\log p(\theta, \mathbf{d}, \mathbf{y})] - \mathbb{E}[\log q(\theta)]$$

We are going to pick our distribution q from the mean-field family of the shape:

$$q(\theta) = \prod_{j=1}^{m} q_j(\theta_j; \mu_j, \sigma_j^2)$$

Thus, each weight in our parameter vector θ is going to have its own distribution, and it is the parameters of these that we wish to optimize to

maximize the ELBO.
The algorithm is like so: For every $j \in [m]$ do

$$q_j(\theta_j) \leftarrow \exp[\mathbb{E}_{-j}[\log p(\theta_j|\theta_{-j}, \mathbf{d}, \mathbf{y})]]$$

Then compute ELBO to measure convergence.
The difficult thing is expanding the formula to an expression that is easy to compute.

Comments for review meeting

It makes no sense to use $p(\theta_j|\theta_{-j}, \mathbf{d}, \mathbf{y})$, because if we already have the posterior, why should we approximate it? Instead we probably need to massage this formula in some way or another to reach an expression based on our prior and posterior which we *can* compute. We should probably also take a quick look at how we plan to compute ELBO.

1.3 Design

1.4 Results

1.5 Evaluation