Bayesian Optimal Design - Weekly Report

Rasmus Hag Løvstad, pgq596

May 24, 2023

1 Week 4

1.1 Objective

The objective of this week is to integrate our Bayesian Optimal Design optimizer from week 2 with our linear regression variational inference optimizer from week 3.

1.2 Theory

1.2.1 Finding the gradient of the Mutual Information

This is a working draft and should be sectioned into a more readable format, as well as have made notation consistent Let us first regard our mutual information objective function from week 2:

$$MI(\mathbf{d}) = \int_\Theta \int_\mathbf{Y} p(\theta, \mathbf{y}|\mathbf{d}) \log p(\theta|\mathbf{y}, \mathbf{d}) d\mathbf{y} d\theta - \int_\Theta p(\theta) \log p(\theta) d\theta$$

Since we are optimizing, let us throw away the second term, since it is constant in terms of $\mathbf{d}$:

$$= \int_\Theta \int_\mathbf{Y} p(\theta, \mathbf{y}|\mathbf{d}) \log p(\theta|\mathbf{y}, \mathbf{d}) d\mathbf{y} d\theta$$

To optimize the mutual information, we will need the derivative of it in terms of $\mathbf{d}$:

$$\frac{\partial}{\partial \mathbf{d}} MI(\mathbf{d}) = \frac{\partial}{\partial \mathbf{d}} \int_\Theta \int_\mathbf{Y} p(\theta, \mathbf{y}|\mathbf{d}) \log p(\theta|\mathbf{y}, \mathbf{d}) d\mathbf{y} d\theta$$

$$= \int_\Theta \int_\mathbf{Y} \frac{\partial}{\partial \mathbf{d}} p(\theta, \mathbf{y}|\mathbf{d}) \log p(\theta|\mathbf{y}, \mathbf{d}) d\mathbf{y} d\theta$$

Let us then use the product rule

TODO: maybe change left derivative

$$= \int_\Theta \int_\mathbf{Y} (\frac{\partial}{\partial \mathbf{d}} p(\theta, \mathbf{y}|\mathbf{d}) \log p(\theta|\mathbf{y}, \mathbf{d})) + (p(\theta, \mathbf{y}|\mathbf{d}) \frac{\partial}{\partial \mathbf{d}} \log p(\theta|\mathbf{y}, \mathbf{d})) d\mathbf{y} d\theta$$

Now, let us use the fact that $\frac{\partial}{\partial \mathbf{d}} p(\theta, \mathbf{y}|\mathbf{d}) = p(\theta, \mathbf{y}|\mathbf{d}) \frac{\partial}{\partial \mathbf{d}} \log p(\theta|\mathbf{y}, \mathbf{d})$

TODO: prove this lemma

$$= \int_\Theta \int_\mathbf{Y} p(\theta, \mathbf{y}|\mathbf{d}) \frac{\partial}{\partial \mathbf{d}} \log p(\theta|\mathbf{y}, \mathbf{d}) \log p(\theta|\mathbf{y}, \mathbf{d}) + p(\theta, \mathbf{y}|\mathbf{d}) \frac{\partial}{\partial \mathbf{d}} \log p(\theta|\mathbf{y}, \mathbf{d}) d\mathbf{y} d\theta$$

$$= \int_\Theta \int_\mathbf{Y} p(\theta, \mathbf{y}|\mathbf{d}) \frac{\partial}{\partial \mathbf{d}} \log p(\theta|\mathbf{y}, \mathbf{d})(\log p(\theta|\mathbf{y}, \mathbf{d}) + 1) d\mathbf{y} d\theta$$

$$= \int_\Theta \int_\mathbf{Y} p(\mathbf{y}|\theta, \mathbf{d}) p(\theta) \frac{\partial}{\partial \mathbf{d}} \log p(\theta|\mathbf{y}, \mathbf{d})(\log p(\theta|\mathbf{y}, \mathbf{d}) + 1) d\mathbf{y} d\theta$$

Solving this double integral can be hard. Let us consider it as an expectation of the form

$$\mathbb{E}[f(\theta, \mathbf{y})] = \int_{(\theta, \mathbf{y})} p(\theta, \mathbf{y}) f(\theta, \mathbf{y}) d(\theta, \mathbf{y})$$

with $p(x) = p(\mathbf{y}|\theta, \mathbf{d})p(\theta)$ and $f(x) = \frac{\partial}{\partial \mathbf{d}} \log p(\theta|\mathbf{y}, \mathbf{d})(\log p(\theta|\mathbf{y}, \mathbf{d}) + 1)$. We can then approximate this expectation by sampling by reducing the expectation to:

$$\mathbb{E}[f(x)] \approx \frac{1}{N} \sum_{i=0}^{N} f(\theta_i, \mathbf{y}_i), \quad (\theta_i, \mathbf{y}_i) \sim p(\theta_i, \mathbf{y}_i)$$

which leads to

TODO: Figure out specific notation here

$$\frac{\partial}{\partial \mathbf{d}} MI(\mathbf{d}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \mathbf{d}} \log p(\theta_i|\mathbf{y}_i, \mathbf{d})(\log p(\theta_i|\mathbf{y}_i, \mathbf{d}) + 1)$$

where $(\theta_i, \mathbf{y}_i) \sim p(\mathbf{y}_i|\theta_i, \mathbf{d})p(\theta_i)$. Sampling θ_i is easy from our prior, and we can do reparameterization to sample $\mathbf{y}_{ij} = \theta_i^T \mathbf{d} + z_j$ where $z_j \sim \mathcal{N}(0, \sigma_\mathbf{y}^2)$.

1.2.2 Finding the gradient of the posterior

Notation:
$\vartheta = \mu_\theta$ and A_θ for use in q_ϑ
$\mathbf{y_d} = \mathbf{y}$ calculated from $\mathbf{d}$.
Now, let us consider the posterior $p(\theta_i|\mathbf{y}, \mathbf{d})$. This is the distribution that we try to approximate when performing variational inference. Thus we can expect our variational distribution $q(\theta_i)$ to reasonably approximate it after our inference algorithm has run. We will denote the optimal parameters found $\vartheta^*(\mathbf{d}, \mathbf{y_d}) = \arg\max_\vartheta \text{ELBO}_{\mathbf{d}, \mathbf{y}(\mathbf{d})}(q_\vartheta)$ such that $q_{\vartheta^*}(\theta_i) \approx p(\theta_i|\mathbf{y_d}, \mathbf{d})$.
In our refactored expression for mutual information, we have a term containing $\frac{\partial}{\partial \mathbf{d}} \log p(\theta_i|\mathbf{y_d}, \mathbf{d})$.

$$\frac{\partial}{\partial \mathbf{d}} \log p(\theta_i|\mathbf{y_d}, \mathbf{d}) \approx \frac{\partial}{\partial \mathbf{d}} \log q_{\vartheta^*}(\theta_i)$$

Since ϑ^* is a function of $\mathbf{d}$, and q^* is a function of θ^*, then we can use the chain rule.

$$= \frac{\partial}{\partial \vartheta^*} \log q_{\vartheta^*}(\theta_i) \frac{\partial}{\partial \mathbf{d}} \vartheta^*(\mathbf{y_d}, \mathbf{d})$$

1.2.3 Using The Implicit Function Theorem for finding the indirect gradient

Let $\mathcal{D}$ be $(\mathbf{d}, \mathbf{y})$ encoded in some vector.
If for some $(\mathcal{D}', \vartheta')$, $\frac{\partial}{\partial\vartheta}\text{ELBO}_{\mathcal{D}'}(q_\vartheta)\Big|_{\mathcal{D}=\mathcal{D}',\vartheta=\vartheta'} = 0$ and the Jacobian is invertible, then there exists an open set of datapoints $\mathcal{D} \in \mathcal{X} \times \mathcal{Y}$ such that there exists a function $\vartheta^* \colon \mathcal{X} \times \mathcal{Y} \to \Theta$ such that

TODO: add reference! very similar to litterature

$$\vartheta^*(\mathcal{D}') = \vartheta' \text{ and } \forall \mathcal{D} \in \mathcal{X} \times \mathcal{Y}, \frac{\partial}{\partial\vartheta}\text{ELBO}_{\mathcal{D}}(q_\vartheta)\Big|_{\mathcal{D},\vartheta^*(\mathcal{D})} = 0$$

Another consequence of this is that we can write

$$\frac{\partial\vartheta^*}{\partial\mathbf{d}}\Big|_{\mathbf{d}'} = \left(-\left[\frac{\partial^2\text{ELBO}_{\mathbf{d},\mathbf{y}_i}(q_\vartheta)}{\partial\vartheta\partial\vartheta^T}\right]^{-1} \times \frac{\partial^2\text{ELBO}_{\mathbf{d},\mathbf{y}_i}(q_\vartheta)}{\partial\vartheta\partial\mathbf{d}^T}\right)\Big|_{\mathbf{d}',\vartheta^*(\mathbf{d}',\mathbf{y}_i')}$$

Where $\mathbf{y}_i' = \theta^T\mathbf{d}' + \mathbf{z}$. Now we have the indirect gradient.

1.3 Design

1.4 Results

1.5 Evalution