DERIVATION OF THE COORDINATE ASCENT VARIATIONAL INFERENCE ALGORITHM OPTIMAL UPDATE FOR THE SPARSEPRO MODEL

Gilad Turok
Department of Computer Science
Columbia University
New York, NY 10027
gt2453@columbia.edu

Austin Tao
Department of Applied Mathematics
Columbia University
New York, NY 10027
alt2177@columbia.edu

ABSTRACT

A common problem in modern machine learning and Bayesian statistics is creating probabilistic models that successfully explain how observed data is generated from intermediary, hidden variables. One method for doing so is variational inference, specifically the Coordinate Ascent Variational Inference (CAVI) algorithm. This method has many applications, one of which is the Fine-Mapping Problem in computational genomics, and the SparsePro model is a novel method to conduct this type of fine mapping. In this paper, we derive the optimal update step for the CAVI algorithm applied to the SparsePro model.

Keywords Variational Inference · Coordinate Ascent · SparsePro Model

1 Introduction

One of the key areas of interest in machine learning and modern day statistics is approximating complicated probability distributions, specifically in order to create some type of model that explains how some given observed data is generated by latent variables. Variational inference frames this problem through the lens of optimization by considering an entire family of probability densities over the latent variables, then finding the setting of parameters that produces the conditional density closest to the desired conditional.

Formalizing this, if we are given some observed variables $\mathbf{X}$ and a set of latent variables $\mathbf{z}$ with the joint probability model $p(\mathbf{z}, \mathbf{X})$ known, we seek the posterior

$$p(\mathbf{z} \mid \mathbf{X}) = \frac{p(\mathbf{z}, \mathbf{X})}{p(\mathbf{X})} \tag{1}$$

This posterior captures the latent variables $\mathbf{z}$ conditioned on the observed variables $\mathbf{X}$, and the denominator $p(\mathbf{X})$ is called the evidence. Calculating this evidence involves computing the following integral

$$p(\mathbf{X}) = \int p(\mathbf{z}, \mathbf{X}) d\mathbf{z} \tag{2}$$

Unfortunately, in most cases, this integral is intractable as it either cannot be solved in closed form or requires exponential time to compute [Blei et al., 2017]. Instead, we choose to approximate the posterior using a quantity that depends only on the latent variables $\mathbf{z}$. Recall that in variational inference, we consider a family of probability densities over the latent variables, denoted $\mathcal{D}$. We draw our approximation of the posterior from this family, that is to say, $q(\mathbf{z}) \in \mathcal{D}$, and we denote the optimal approximation as $q^*(\mathbf{z})$ defined by the following

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z}) \in \mathcal{D}} KL(q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{X})) \tag{3}$$

Here, KL is the Kullback-Leibler divergence, where we have

$$\begin{aligned} KL(q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{X})) &= \mathbb{E}_{\mathbf{z}}[\log q(\mathbf{z})] - \mathbb{E}_{\mathbf{z}}[\log p(\mathbf{z} \mid \mathbf{X})] \\ &= \mathbb{E}_{\mathbf{z}}[\log q(\mathbf{z})] - \mathbb{E}_{\mathbf{z}}[\log p(\mathbf{z}, \mathbf{X})] + \log p(\mathbf{X}) \end{aligned}$$

Going forward, we will suppress the notation that the expectation is taken with respect to $\mathbf{z}$.

However, computing the KL divergence requires knowing $\log p(\mathbf{X})$, which is intractable. Because of this, instead of optimizing the KL divergence, we instead optimize the Evidence Lower Bound (ELBO), denoted $\mathcal{L}$. The ELBO is equivalent to the negative KL divergence up to an additive constant. Specifically,

$$\begin{aligned}\mathcal{L}(\mathbf{z}) : &= \mathbb{E}[\log p(\mathbf{z}, \mathbf{X})] - \mathbb{E}[\log q(\mathbf{z})] \quad (4)\\ &= \mathbb{E}[\log p(\mathbf{X} \mid \mathbf{z})] - KL(q(\mathbf{z}) \mid\mid p(\mathbf{z})) \quad (5)\end{aligned}$$

Now, our optimization problem is to maximize the ELBO $\mathcal{L}(\mathbf{z})$ in order to minimize $KL(q(\mathbf{z}) \mid\mid p(\mathbf{z}))$. Our approximation for the posterior $p(\mathbf{z} \mid \mathbf{X})$ is now given by

$$q^*(\mathbf{z}) = \arg\max_{q(\mathbf{z}) \in \mathcal{D}} \mathcal{L}(\mathbf{z}) \quad (6)$$

2 The CAVI Algorithm

Now that we have framed the problem of calculating the posterior as an optimization problem, we will perform this optimization one coordinate at a time for tractability. This is known as the Coordinate Ascent Variational Inference algorithm, and we want to derive the ideal CAVI update for our latent variable $\mathbf{z}_k$ while holding other latent variables $\mathbf{z}_{\setminus k}$ constant until convergence.

Deriving this generalized ideal update for CAVI, we begin with equation 4 and expand from the definition of expectation

$$\begin{aligned}\mathcal{L}(\mathbf{z}) &= \mathbb{E}[\log p(\mathbf{z}, \mathbf{X})] - \mathbb{E}[\log q(\mathbf{z})]\\ &= \int_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{X}) d\mathbf{z} - \left[\mathbb{E}_{\mathbf{z}_k}[\log q(\mathbf{z}_k)] + \sum_{k\prime \neq k} \mathbb{E}_{\mathbf{z}_{k\prime}}[\log q(\mathbf{z}_{k\prime})]\right]\\ &= \int_{\mathbf{z}_k} q(\mathbf{z}_k) \mathbb{E}_{\mathbf{z}_{\setminus k}}[\log p(\mathbf{z}, \mathbf{X})] d\mathbf{z}_k - \int_{\mathbf{z}_k} q(\mathbf{z}_k) \log q(\mathbf{z}_k) d\mathbf{z}_k + C\end{aligned}$$

Now, define $\log \tilde{p}_k := \mathbb{E}_{\mathbf{z}_{\setminus k}}[\log p(\mathbf{z}, \mathbf{X})] + C$. We finish simplifying the ELBO with this new value,

$$\begin{aligned}\mathcal{L}(\mathbf{z}) &= \int_{\mathbf{z}_k} q(\mathbf{z}_k) \log \frac{\tilde{p}_k(\mathbf{z}_k, \mathbf{X})}{q(\mathbf{z}_k)} d\mathbf{z}_k + C\\ &= -KL(q(\mathbf{z}_k) \mid\mid \tilde{p}_k(\mathbf{z}_k, \mathbf{X})) + C\end{aligned}$$

Thus, we confirmed that to maximize the ELBO with respect to the latent variable $\mathbf{z}_k$, we want to minimize the KL divergence. The KL divergence is minimized when the distributions are equal, so we have

$$\log q(\mathbf{z}_k) = \log \tilde{p}_k(\mathbf{z}_k, \mathbf{X}) \propto \mathbb{E}_{\mathbf{z}_{\setminus k}}[\log p(\mathbf{z}, \mathbf{X})] \quad (7)$$

This represents the ideal update step for the CAVI algorithm. We now present the pseudocode for the CAVI algorithm [Blei et al., 2017].

Algorithm 1 CAVI

Input: Data set $\mathbf{X}$, model $p(\mathbf{z}, \mathbf{X})$
Output: Variational Density $q(\mathbf{z}) = \prod_{k=1}^{K} q_k(z_k)$
Initialize: Variational factors $q_k(z_k)$
while ELBO is **not** converged **do**:
    for $k \in \{1, \ldots, K\}$ **do**:
        Set $q_k(z_k) \propto \exp\{\mathbb{E}_{\setminus k}[\log p(z_k, \mathbf{z}_{\setminus k}, \mathbf{X})]\}$
    end
    Compute $\mathcal{L}(z) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{X})] - \mathbb{E}[\log q(\mathbf{z})]$
end
return $q(\mathbf{z})$

3 The SparsePro Model

We will now identify the specific application we are applying the CAVI algorithm to. Here, we focus particularly on the Fine Mapping Problem in computational genomics, which the task of identifying causal variants from Genome Wide Association Studies (GWAS). Given some genetic information – in our case, single-nucleotide polymorphisms (SNPs) – and some phenotype data, we want to be able to identify which genotypes cause a phenotypic trait.

The SparsePro Model [Zhang et al., 2021] is one such approach to tackling this problem that has had demonstrated success in recent years. For this model, the assumption is made that phenotypic traits are the results of sums of K individual causal effects. Additionally, we assume a generative process from genotype data $\mathbf{X}$ to phenotype data $\mathbf{y}$. Our sum of single effects model takes on the following form

$$\mathbf{y} = \mathbf{X}\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}\sum_{k=1}^{K}\mathbf{S}_k\beta_k + \epsilon \tag{8}$$

where we have G SNP measurements for N number of people and $\mathbf{X}_{N\times G}$ is the genotype data matrix. Additionally, we make the assumption that there are no functional priors present, therefore the prior probability for any SNP to be causal is $1/G$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$ be the vector that holds the probability of the g-th SNP being causal (where $g \in G$). Because we assume that there are no functional priors, we immediately know $\pi_g = \frac{1}{G}$.

We also have that $\mathbf{S}_k \sim \mathrm{Cat}(\boldsymbol{\pi})$, meaning $\mathbf{S}_k$ is an indicator vector of length G that identifies the causal SNP under the k-th causal effect. Additionally, we have the causal effect size $\beta_k \sim \mathcal{N}(0, \tau_{\beta_k}^{-1})$ where $\tau_{\beta_k}^{-1}$ is a hyperparameter, as well as Gaussian noise $\epsilon \sim \mathcal{N}(0, \tau_y^{-1}\mathbb{I})$ parameterized by τ_y^{-1}. The phenotype data is $\mathbf{y}_{N\times 1}$.

Connecting this model back to variational inference, we take the setup of SparsePro and frame it so that we can apply CAVI to it. For this, the observed variables are the genotype data matrix $\mathbf{X}$ and phenotype data $\mathbf{y}$, while the latent variables are the sparse indicator vectors $\{\mathbf{S}_1, \ldots, \mathbf{S}_K\}$ and the causal effect sizes $\{\beta_1, \ldots, \beta_K\}$. Our overarching goal is to calculate the intractable true posterior $p(\mathbf{S}, \boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y})$. To do this, we use variational inference to approximate the posterior with $q(\mathbf{S}, \boldsymbol{\beta})$, where we have

$$q(\mathbf{S}, \boldsymbol{\beta}) = \prod_k q(\beta_k \mid \mathbf{S}_k)q(\mathbf{S}_k) \tag{9}$$

4 Deriving the Optimal CAVI Update for SparsePro

Recall that in general, the optimal update step for CAVI is given by

$$\log q(\mathbf{z}_k) \propto \mathbb{E}_{\mathbf{z}_{\setminus k}}[\log p(\mathbf{z}, \mathbf{X})] \tag{10}$$

For the specific case of SparsePro, this update step takes on the following form:

$$\log q^*(\beta_k, \mathbf{S}_k) \propto \mathbb{E}_{\setminus k}[\log p(\mathbf{y}, \mathbf{X}, \mathbf{S}, \boldsymbol{\beta})] \tag{11}$$

In order to now derive the optimal CAVI update for SparsePro, we must first compute the expectation of the log-joint probability function, then identify the variational parameters such that equation 11 is satisfied. It is also important to note that this ideal CAVI update is done with respect to the k-th latent variables β_k and $\mathbf{S}_k$. All terms with a different k are treated as constant and dropped.

4.1 Calculating the Expectation of the Log-Joint

We begin by explicitly writing out the form of the log-joint probability function. Due to the Mean Field Approximation, we can partition our variables and assume that they are independent of each other. Thus, the log-joint is written as follows

$$\log p(\mathbf{z}, \mathbf{X}) = \log p(\mathbf{y}, \mathbf{X}, \mathbf{S}, \boldsymbol{\beta}) = \log\left[p(\mathbf{y} \mid \mathbf{X}, \mathbf{S}, \boldsymbol{\beta}) \prod_{k\prime=1}^{K} p(\mathbf{S}_{k\prime}; \gamma) \prod_{k\prime=1}^{K} p(\boldsymbol{\beta}_{k\prime}; \tau_{\boldsymbol{\beta}_{k\prime}}^{-1})\right] \tag{12}$$

Taking the expectation of the log-joint, we have the following equation by the definition of expectation. Note that going forward, we suppress the notation of hyperparameters unless necessary.

$$\begin{aligned}\mathbb{E}_{\setminus k}[\log p(\mathbf{y}, \mathbf{X}, \mathbf{S}, \boldsymbol{\beta})] = &\int \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{S}, \boldsymbol{\beta}) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) \\ &+ \int \sum_{k\prime} \log p(\mathbf{S}_{k\prime}) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) \\ &+ \int \sum_{k\prime} \log p(\boldsymbol{\beta}_{k\prime}) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})\end{aligned}$$

Notice that the expression that we are evaluating for, $\mathbb{E}_{\setminus k}[\log p(z_k, x_k)]$, is integrated with respect to everything except k, but the function itself only has dependency on k itself. Therefore, we are only interested in terms that are dependent on k. This simplifies our summations, as we only need to worry about the case where $k\prime = k$. Our expression is now

$$\begin{aligned}\mathbb{E}_{\setminus k}[\log p(\mathbf{y}, \mathbf{X}, \mathbf{S}, \boldsymbol{\beta})] = &\int \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{S}, \boldsymbol{\beta}) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) \\ &+ \int \log p(\mathbf{S}_k) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) \\ &+ \int \log p(\boldsymbol{\beta}_k) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})\end{aligned}$$

Now, we will assign each of those 3 integrals a new variable $\mathcal{T}$ such that we have $\mathbb{E}_{\setminus k}[\log p(\mathbf{y}, \mathbf{X}, \mathbf{S}, \boldsymbol{\beta})] = \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3$. More explicitly, we have

$$\begin{aligned}\mathcal{T}_1 &:= \int \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{S}, \boldsymbol{\beta}) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) \\ \mathcal{T}_2 &:= \int \log p(\mathbf{S}_k) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) \\ \mathcal{T}_3 &:= \int \log p(\boldsymbol{\beta}_k) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})\end{aligned}$$

Once we evaluate each of these three terms, we will have successfully calculated the expectation of the log-joint.

4.1.1 Log-Joint Expectation: Computing $\mathcal{T}_1$

To compute the first term $\mathcal{T}_1$, we take advantange of the fact that we know $p(\mathbf{y} \mid \mathbf{X}, \mathbf{S}, \boldsymbol{\beta})$ is sampled from a Gaussian distribution, which we know the PDF of [Prince, 2012]. Define $\boldsymbol{\mu}_y := \mathbf{XS}\boldsymbol{\beta} = \mathbf{X}\sum_{k\prime=1}^{K} \mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}$. We now expand the log of the PDF of the multivariate Gaussian.

$$\begin{aligned}\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{S}, \boldsymbol{\beta}) &= -\frac{1}{2}\left[N \log(2\pi) + \log(\det(\tau_y^{-1}\mathbb{I}_N)) + (\mathbf{y} - \boldsymbol{\mu}_y)^T (\tau_y^{-1}\mathbb{I}_N)^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)\right] \\ &= \frac{N}{2}\log\left(\frac{1}{2\pi}\right) - \frac{1}{2}\log(\det(\tau_y^{-1}\mathbb{I}_N)) - \frac{\tau_y}{2}(\mathbf{y} - \boldsymbol{\mu}_y)^T(\mathbf{y} - \boldsymbol{\mu}_y) \\ &= \frac{N}{2}\log\left(\frac{1}{2\pi}\right) - \frac{\tau_y}{2}(\mathbf{y} - \boldsymbol{\mu}_y)^T(\mathbf{y} - \boldsymbol{\mu}_y)\end{aligned}$$

With respect to k, the first term of the above equation on the RHS drops out, leaving us with

$$\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{S}, \boldsymbol{\beta}) = -\frac{\tau_y}{2}(\mathbf{y} - \boldsymbol{\mu}_y)^T(\mathbf{y} - \boldsymbol{\mu}_y)$$

and we have that

$$\mathcal{T}_1 = \int -\frac{\tau_y}{2}(\mathbf{y} - \boldsymbol{\mu}_y)^T(\mathbf{y} - \boldsymbol{\mu}_y) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})$$

To further simplify this, we use a property of the inner product which states that for two vectors $\mathbf{a}$ and $\mathbf{b}$, the following is true:

$$(\mathbf{a} - \mathbf{b})^T(\mathbf{a} - \mathbf{b}) = \sum_n (a_n - b_n)^2 \tag{13}$$

Using equation 13 to continue simplifying $\mathcal{T}_1$, we create a squared term inside of the integral.

$$\begin{aligned}\mathcal{T}_1 &= -\frac{\tau_y}{2}\int (\mathbf{y}-\boldsymbol{\mu}_y)^T(\mathbf{y}-\boldsymbol{\mu}_y)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k}) \\ &= -\frac{\tau_y}{2}\int \sum_{n\prime=1}^{N}(\mathbf{y}-\boldsymbol{\mu}_y)_{n\prime}^2 q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})\end{aligned}$$

Now expanding the squared term out, applying equation 13 once more, and plugging in the definition of $\boldsymbol{\mu}_y$, we get an equation for $\mathcal{T}_1$ as the sum of three integrals.

$$\begin{aligned}\mathcal{T}_1 = &-\frac{\tau_y}{2}\int \mathbf{y}^T\mathbf{y}q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k}) + \tau_y\int \mathbf{y}^T\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k}) \\ &-\frac{\tau_y}{2}\int \left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)^T\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})\end{aligned}$$

The first integral on the RHS is quite simple to solve upon inspection, as we can pull out the terms with no dependency on $d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})$, leaving just the integral of the PDF $q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})$, which integrates to 1. Our expression is now

$$\begin{aligned}\mathcal{T}_1 = &-\frac{\tau_y}{2}\mathbf{y}^T\mathbf{y} + \tau_y\int \mathbf{y}^T\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k}) \\ &-\frac{\tau_y}{2}\int \left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)^T\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})\end{aligned}$$

Next, we independently compute the second term in the equation for $\mathcal{T}_1$. Similar to before, we can pull out the terms that are not dependent on $d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})$.

$$\tau_y\int \mathbf{y}^T\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k}) = \tau_y\mathbf{y}^T\mathbf{X}\int\left(\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})$$

Now, we take advantage of the fact that we can split the summation over K as

$$\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime} = \mathbf{S}_k\boldsymbol{\beta}_k + \sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime} \tag{14}$$

Continuing to simplify the expression using equation 14,

$$\begin{aligned}&= \tau_y\mathbf{y}^T\mathbf{X}\int\left(\mathbf{S}_k\boldsymbol{\beta}_k + \sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k}) \\ &= \tau_y\mathbf{y}^T\mathbf{X}\mathbf{S}_k\boldsymbol{\beta}_k + \tau_y\mathbf{y}^T\mathbf{X}\int\left(\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})\end{aligned}$$

Now, we notice that in our expression above, the integral and the coefficient in front of it form a term that is completely constant with respect to k. Recall that we are only interested in the log-joint with respect to k. Therefore, we can treat the integral as constant, and we simplify (7) to just be

$$\tau_y\int \mathbf{y}^T\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k}) = \tau_y\mathbf{y}^T\mathbf{X}\mathbf{S}_k\boldsymbol{\beta}_k$$

Our expression for $\mathcal{T}_1$ is now

$$\mathcal{T}_1 = -\frac{\tau_y}{2}\mathbf{y}^T\mathbf{y} + \tau_y\mathbf{y}^T\mathbf{X}\mathbf{S}_k\boldsymbol{\beta}_k - \frac{\tau_y}{2}\int\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)^T\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})$$

Lastly, we proceed to simplify the final integral in the equation for $\mathcal{T}_1$. To do so, we first only consider the integrand of the remaining integral and simplify that before evaluating the integral. We proceed using equation 14:

$$\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)^T\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)=\sum_{n\prime=1}^{N}\left[(\mathbf{X}\mathbf{S}_k\boldsymbol{\beta}_k)_{n\prime}+\left(\mathbf{X}\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)_{n\prime}\right]^2$$

$$=\sum_{n\prime=1}^{N}\left[(\mathbf{X}\mathbf{S}_k\boldsymbol{\beta}_k)_{n\prime}^2\right]+\sum_{n\prime=1}^{N}\left[2(\mathbf{X}\mathbf{S}_k\boldsymbol{\beta}_k)_{n\prime}\left(\mathbf{X}\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)_{n\prime}\right]$$

$$+\sum_{n\prime=1}^{N}\left[\left(\mathbf{X}\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)_{n\prime}\left(\mathbf{X}\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)_{n\prime}\right]$$

Applying equation 13, we turn our summations into vector multiplication.

$$=(\mathbf{X}\mathbf{S}_k\boldsymbol{\beta}_k)^T(\mathbf{X}\mathbf{S}_k\boldsymbol{\beta}_k)+2(\mathbf{X}\mathbf{S}_k\boldsymbol{\beta}_k)^T\left(\mathbf{X}\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)+\left(\mathbf{X}\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)^T\left(\mathbf{X}\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)$$

Now, taking this integrand and placing it back inside the integral, we can split the integral and take advantage of the terms that do not depend on k. This leads us to the following

$$=-\frac{\tau_y}{2}\boldsymbol{\beta}_k^2(\mathbf{X}\mathbf{S}_k)^T(\mathbf{X}\mathbf{S}_k)-\tau_y(\boldsymbol{\beta}_k\mathbf{X}\mathbf{S}_k)^T\mathbf{X}\int\left(\mathbf{X}\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})$$

$$+\int\left(\mathbf{X}\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)^2q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})$$

For notational brevity, we will define $\mathbf{X}_g$ and $\tilde{\boldsymbol{\beta}}$ to be the following:

$$\mathbf{X}_g:=\mathbf{X}\mathbf{S}_k\qquad\qquad\tilde{\boldsymbol{\beta}}:=\mathbb{E}_{q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})}\left[\sum_{\substack{k\prime=1\\k\prime\neq k}}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right]$$

Continuing the simplification, we end with

$$-\frac{\tau_y}{2}\int\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)^T\left(\mathbf{X}\sum_{k\prime=1}^{K}\mathbf{S}_{k\prime}\boldsymbol{\beta}_{k\prime}\right)q(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})d(\mathbf{S}_{\setminus k},\boldsymbol{\beta}_{\setminus k})=-\frac{\tau_y}{2}\boldsymbol{\beta}_k^2(\mathbf{X}_g^T\mathbf{X}_g)-\tau_y\boldsymbol{\beta}_k\mathbf{X}_g^T\mathbf{X}\tilde{\boldsymbol{\beta}}+C$$

where C is a constant.

At this point, we have evaluated all three integrals that sum up to $\mathcal{T}_1$ and our final equation is

$$\mathcal{T}_1=\tau_y\mathbf{y}^T\mathbf{X}_g\boldsymbol{\beta}_k-\frac{\tau_y}{2}\boldsymbol{\beta}_k^2(\mathbf{X}_g^T\mathbf{X}_g)-\tau_y\boldsymbol{\beta}_k\mathbf{X}_g^T\mathbf{X}\tilde{\boldsymbol{\beta}}+C \tag{15}$$

4.1.2 Log-Joint Expectation: Computing $\mathcal{T}_2$ and $\mathcal{T}_3$

We now proceed to finish the calculation of the expectation of the log-joint by computing $\mathcal{T}_2$ and $\mathcal{T}_3$. Thankfully, these are more straightforward than the calculation of $\mathcal{T}_1$. Recall that

$$\mathcal{T}_2 = \int \log p(\mathbf{S}_k; \gamma) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) \qquad \mathcal{T}_3 = \int \log p(\boldsymbol{\beta}_k; \tau^{-1}_{\boldsymbol{\beta}_{k\prime}}) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})$$

Calculating $\mathcal{T}_2$ is done by recalling the PDF of the categorical distribution [Gordon-Rodriguez et al., 2022] and taking the logarithm. However, instead of ending up with two summations, we only end up with one (iterating over G). This is because, again, we are only interested in the terms that have dependency on k.

$$\mathcal{T}_2 = \int \sum_{g\prime=1}^{G} \mathbf{S}_{k_{g\prime}} \log(\boldsymbol{\pi}_{g\prime}) q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})$$

To evaluate the summation, recall what $\mathbf{S}_{k_{g\prime}}$ represents: $\mathbf{S}_k$ represents the indicator vector for the k-th causal SNP, meaning it is a vector of the form

$$\mathbf{S}_k = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where $\mathbf{S}_{k_{g\prime}}$ represents the $g\prime$-th index of the indicator vector $\mathbf{S}_k$. Notice that because $\mathbf{S}_k$ is an indicator vector, there is only one entry that is nonzero. Therefore, we have that

$$\sum_{g\prime=1}^{G} \mathbf{S}_{k_{g\prime}} = 1$$

Also recall that if we do not have functional annotations,

$$\boldsymbol{\pi}_{g\prime} = \frac{1}{G}, \quad \forall g\prime$$

and if we do have functional annotations,

$$\boldsymbol{\pi}_{g\prime} = \text{softmax}(\mathbf{A}_{g\prime}\mathbf{w}) = \frac{\exp(\mathbf{A}_{g\prime}\mathbf{w})}{\sum_{g\prime\prime=1}^{G} \exp(\mathbf{A}_{g\prime\prime}\mathbf{w})}$$

Therefore we can simplify $\mathcal{T}_2$ further by pulling the summation out of the integral

$$\mathcal{T}_2 = \log(\boldsymbol{\pi}_g) \int q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})$$

All that remains in the integral is a probability function, which integrates to 1. We end the simplification of $\mathcal{T}_2$ to be

$$\mathcal{T}_2 = \log(\boldsymbol{\pi}_g) \tag{16}$$

Now to compute $\mathcal{T}_3$, we recall that $\boldsymbol{\beta}_k$ is sampled from a normal distribution to easily simplify

$$\mathcal{T}_3 = C - \frac{\tau_{\boldsymbol{\beta}_k}}{2} \boldsymbol{\beta}_k^2 \int q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k}) d(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})$$

Again, the PDF integrates to 1, and we are left with our simplification for $\mathcal{T}_3$

$$\mathcal{T}_3 = C - \frac{\tau_{\boldsymbol{\beta}_k}}{2} \boldsymbol{\beta}_k^2 \tag{17}$$

Now, we complete the calculation for the expectation of the log-joint probability by combining equations 15, 16, and 17.

$$\mathbb{E}_{\setminus k}[\log p(\mathbf{y}, \mathbf{X}, \mathbf{S}, \boldsymbol{\beta})] = C - \frac{\tau_{\boldsymbol{\beta}_k}}{2} \boldsymbol{\beta}_k^2 - \frac{\tau_y}{2} \boldsymbol{\beta}_k^2 \mathbf{X}_g^T \mathbf{X}_g + \tau_y \boldsymbol{\beta}_k \mathbf{X}_g (\mathbf{y}^T - \mathbf{X}_g^T \tilde{\boldsymbol{\beta}}) + \log(\boldsymbol{\pi}_g) \tag{18}$$

4.2 Defining Variational Parameters

At this point, we have successfully completed the calculation for the expectation of the log-joint probability, which is the RHS of equation 11. Now, we want to identify the distributions that emit the ideal values for the variational parameters $\beta_k, \mathbf{s}_k$. We expand the LHS of equation 11

$$\log q(\mathbf{s}_k, \beta_k) = \log q(\beta_k \mid \mathbf{s}_k) + \log q(\mathbf{s}_k)$$

and recognize our (precision-defined) normal and categorical distributions

$$q(\beta_k \mid \mathbf{s}_k) \sim \mathcal{N}(\mu^*_{kg}, (\tau^*_{kg})^{-1}) \qquad q(\mathbf{s}_k) \sim \text{Cat}(\boldsymbol{\gamma}^*_{kg})$$

have the following parameters:

$$\begin{aligned}
\tau^*_{kg} &:= \tau_y \mathbf{X}_g^T \mathbf{X}_g + \tau_{\beta_k} \\
\mu^*_{kg} &:= \frac{\tau_y}{\tau^*_{kg}} \mathbf{X}_g^T (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\setminus k}) \\
\boldsymbol{\gamma}^*_{kg} &:= \text{softmax}\Big(\log \pi_g - \frac{1}{2} \log \frac{\tau^*_{kg}}{2\pi} + \frac{1}{2} \tau^*_{kg} (\mu^*_{kg})^2 \Big)_g
\end{aligned}$$

To see this, we explicitly expand out our distributions with their parameters. First we expand out $\log q(\beta_k \mid \mathbf{s}_k) \sim \log \mathcal{N}(\mu^*_{kg}, (\tau^*_{kg})^{-1})$ as follows:

$$\begin{aligned}
\log q(\beta_k \mid \mathbf{s}_k) &= \frac{1}{2} \log \frac{\tau^*_{kg}}{2\pi} - \frac{1}{2} (\tau^*_{kg})(\beta_k - \mu^*_{kg})^2 \\
&= \frac{1}{2} \log \frac{\tau^*_{kg}}{2\pi} - \frac{\tau^*_{kg}}{2} \Big(\beta_k^2 - 2\beta_k \mu^*_{kg} + (\mu^*_{kg})^2 \Big) \\
&= \frac{1}{2} \log \frac{\tau^*_{kg}}{2\pi} - \frac{\tau^*_{kg}}{2} \left(\beta_k^2 - 2\beta_k \Big(\frac{\tau_y}{\tau^*_{kg}} \mathbf{X}_g^T (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\setminus k}) \Big) + \Big(\frac{\tau_y}{\tau^*_{kg}} \mathbf{X}_g^T (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\setminus k}) \Big)^2 \right) \\
&= \frac{1}{2} \log \frac{\tau^*_{kg}}{2\pi} - \frac{\tau^*_{kg} \beta_k^2}{2} + \beta_k \tau_y \mathbf{X}_g^T (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\setminus k}) + \frac{1}{\tau^*_{kg}} \Big(\tau_y \mathbf{X}_g^T (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\setminus k}) \Big)^2
\end{aligned}$$

Now we expand out $\log q(\mathbf{s}_k) \sim \log \text{Cat}(\gamma^*_{kg})$. To do so, recognize again that the SNP indicator vector $\mathbf{S}_k$ is zero for all SNPs $g\prime \in \{1 \dots G\}$ except for when index $g\prime = g$, our true causal SNP. Furthermore, at our true causal SNP g, the indicator vector takes a value of $\mathbf{S}_{kg} = 1$.

$$\begin{aligned}
\log q(\mathbf{s}_k) &= \log \prod_{g\prime=1}^{G} (\gamma^*_{kg})^{\mathbb{1}[\mathbf{S}_{kg\prime}=1]} \\
&= \sum_{g\prime=1}^{G} \mathbf{S}_{kg\prime} \log(\gamma^*_{kg\prime}) \\
&= \mathbf{S}_{kg} \log \gamma^*_{kg} \\
&= \log \left(\text{softmax} \left(\log \pi_g - \frac{1}{2} \log \frac{\tau^*_{kg}}{2\pi} + \frac{1}{2} \tau^*_{kg} (\mu^*_{kg})^2 \right)_g \right) \\
&= C + \log \pi_g - \frac{1}{2} \log \frac{\tau^*_{kg}}{2\pi} + \frac{1}{2} \tau^*_{kg} (\mu^*_{kg})^2
\end{aligned}$$

Now we add both distributions $\log q(\beta_k \mid \mathbf{S}_k)$ and $\log q(\mathbf{S}_k)$ and recover our equation for the joint $\log q(\beta_k, \mathbf{S}_k)$ that we derived in equation 18

$$
\begin{aligned}
\log q(\beta_k, \mathbf{S}_k) &= \log q(\beta_k \mid \mathbf{s}_k) + \log q(\mathbf{s}_k) \\
&= \left[\frac{1}{2}\log\frac{\tau^*_{kg}}{2\pi} - \frac{\tau^*_{kg}}{2}\Big(\beta_k^2 - 2\beta_k\mu^*_{kg} + (\mu^*_{kg})^2\Big)\right] + \left[\log\pi_g - \frac{1}{2}\log\frac{\tau^*_{kg}}{2\pi} + \frac{1}{2}\tau^*_{kg}(\mu^*_{kg})^2\right] \\
&= -\frac{\tau^*_{kg}}{2}\beta_k^2 + \log\pi_g + \tau^*_{kg}\beta_k\mu^*_{kg} \\
&= -\frac{\tau^*_{kg}}{2}\beta_k^2 + \log\pi_g + \tau^*_{kg}\beta_k\big(\frac{\tau_y}{\tau^*_{kg}}\mathbf{X}_g^T(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\setminus k})\big) \\
&= -\frac{\tau^*_{kg}}{2}\beta_k^2 + \log\pi_g + \beta_k\tau_y\mathbf{X}_g^T(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\setminus k}) \\
&= -\frac{\beta_k^2}{2}(\tau_y\mathbf{X}_g^T\mathbf{X}_g + \tau_{\beta_k}) + \beta_k\tau_y\mathbf{X}_g^T(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\setminus k}) \\
&= -\frac{\tau_{\beta_k}}{2}\beta_k^2 - \frac{\tau_y}{2}\beta_k^2\mathbf{X}_g^T\mathbf{X}_g + \log\pi_g + \tau_y\beta_k\mathbf{X}_g(\mathbf{y}^T - \mathbf{X}^T\tilde{\boldsymbol{\beta}}_{\setminus k}) + \log\pi_g
\end{aligned}
$$

References

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, apr 2017. doi:10.1080/01621459.2017.1285773. URL `https://doi.org/10.48550/arXiv.1601.00670`.

Wenmin Zhang, Hamed Najafabadi, and Yue Li. Sparsepro: an efficient genome-wide fine-mapping method integrating summary statistics and functional annotations. *bioRxiv*, 2021. doi:10.1101/2021.10.04.463133. URL `https://www.biorxiv.org/content/early/2021/11/02/2021.10.04.463133`.

S.J.D. Prince. *Computer Vision: Models Learning and Inference*. Cambridge University Press, 2012.

Elliott Gordon-Rodriguez, Gabriel Loaiza-Ganem, Andres Potapczynski, and John P. Cunningham. On the normalizing constant of the continuous categorical distribution, 2022. URL `https://arxiv.org/abs/2204.13290`.