# 1 Two Data Augmentations

Suppose we have a joint distribution $p(X, \theta) = p(X \mid \theta)p(\theta)$ specified by a likelihood and a prior. Bayesian statistics frames inferences about the unknown quantity $\theta$ in terms of calculations involving the posterior $p(\theta \mid X)$ given the observations. In many cases, it is helpful to work with an augmented model containing intermediate latent variables $\mu$. In general we have the *hierarchical factorization*

$$p(X, \mu, \theta) = p(X \mid \mu, \theta)p(\mu \mid \theta)p(\theta). \tag{1}$$

In a *sufficient augmentation* (SA), the new variables $\mu$ are sufficient for $\theta$, so the factorization

$$p(X, \mu, \theta) = p(X \mid \mu)p(\mu \mid \theta)p(\theta) \tag{2}$$

holds. In an *ancillary augmentation* (AA), the new variables—denoted $\nu$ for contrast—are independent of $\theta$ a priori, so the joint distribution factorizes as

$$p(X, \nu, \theta) = p(X \mid \nu, \theta)p(\nu)p(\theta). \tag{3}$$

**Example.** In location families, for example, there are natural sufficient and ancillary augmentations. One important example is the normal-normal model. In the sufficient augmentation,

$$\mu \mid \theta \sim \mathcal{N}(\theta, V) \tag{4}$$
$$X \mid \mu, \theta \sim \mathcal{N}(\mu, 1) \tag{5}$$

with a flat prior on $\theta$, the posterior is $\mathcal{N}(X, 1 + V)$. In the ancillary augmentation,

$$\nu \mid \theta \sim \mathcal{N}(0, V) \tag{6}$$
$$X \mid \nu, \theta \sim \mathcal{N}(\nu + \theta, 1) \tag{7}$$

There is a one-to-one relationship $\nu = \mu - \theta$ between the two augmentation schemes, but the performance of approximate posterior inference methods can differ depending on the choice of augmentation.

# 2 Variational Inference

*Mean-field variational inference* finds the factorized distribution over the latent which is closest in KL-divergence to the posterior. In the context of the previous example, our objective is

$$\min_{q(\mu), q(\theta)} D\left(q(\mu)q(\theta) \,\middle\|\, p(\mu, \theta \mid X)\right), \tag{8}$$

and similarly for the ancillary augmentation. This is easily shown to be equivalent to maximizing the evidence lower bound (ELBO)

$$\max_{q = q(\mu)q(\theta)} \underbrace{\mathbb{E}_q\left[\log \frac{p(X, \mu, \theta)}{q(\mu)q(\theta)}\right]}_{\mathcal{L}(q)}, \tag{9}$$

It is also easily shown that maximizing one variational factor $q(\mu)$ with the other $q(\theta)$ held fixed and vice versa is given in closed form by

$$q(\mu) \propto \exp\left\{\mathbb{E}_{q(\theta)}\left[\log p(X, \mu, \theta)\right]\right\} \tag{10}$$
$$q(\theta) \propto \exp\left\{\mathbb{E}_{q(\mu)}\left[\log p(X, \mu, \theta)\right]\right\} \tag{11}$$

Alternating (10) and (11) gives a coordinate ascent algorithm (called *CAVI*) for maximizing (9).

**Example.** (SA) Returning to the sufficient augmentation version of the normal-normal model above,

$$q(\mu) \triangleq \mathcal{N}(\widehat{\mu}, \widehat{\sigma}_{\mu}^2) \tag{12}$$
$$q(\theta) \triangleq \mathcal{N}(\widehat{\theta}_S, \widehat{\sigma}_{\theta_S}^2) \tag{13}$$

Since we are optimizing the variational parameters (denoted by hat$\widehat{\text{s}}$), we include superscripts $\widehat{\mu}^{(t)}$ for the iteration number. Writing out the ELBO

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q\left[\log \frac{p(X, \mu, \theta)}{q(\mu)q(\theta)}\right] \\
&= \mathbb{E}_q\left[\log p(X \mid \mu)\right] - \mathbb{E}_q\left[\log \frac{q(\mu)}{p(\mu \mid \theta)}\right] - \mathbb{E}_q\left[\log \frac{q(\theta)}{p(\theta)}\right] \\
&= \frac{2X\widehat{\mu} - \widehat{\sigma}_{\mu}^2 - \widehat{\mu}^2}{2} + \log \widehat{\sigma}_{\mu} - \frac{\widehat{\sigma}_{\mu}^2 + \widehat{\sigma}_{\theta_S}^2 + (\widehat{\mu} - \widehat{\theta})^2}{2V} + \log \widehat{\sigma}_{\theta_S} + \text{const.}
\end{aligned}$$

The coordinate ascent updates are

$$\widehat{\mu}^{(t+1)} = \frac{VX + \widehat{\theta}^{(t)}}{1+V} \tag{14}$$
$$\widehat{\sigma}_{\mu}^{2(t+1)} = \frac{V}{1+V} \tag{15}$$
$$\widehat{\theta}^{(t+1)} = \widehat{\mu}^{(t+1)} \tag{16}$$
$$\widehat{\sigma}_{\theta_S}^{2(t+1)} = V \tag{17}$$

Thus the variational parameter for the posterior variance of $\theta$ given $X$, $\widehat{\sigma}_{\theta_S}^{2(t+1)} = V$ underestimates the true posterior variance $1 + V$ (this is a common property of variational Bayes). The variational parameter for the posterior mean of $\theta$ given $X$ satisfies

$$\left|\widehat{\theta}^{(t+1)} - X\right| = \left|\frac{VX + \widehat{\theta}^{(t)}}{1+V} - X\right| = \frac{1}{1+V}\left|\widehat{\theta}^{(t)} - X\right|, \tag{18}$$

this parameter converges geometrically with rate $\frac{1}{1+V}$.

**Example.** (AA) For the ancillary augmentation, let

$$\widetilde{q}(\nu) \triangleq \mathcal{N}(\widehat{\nu}, \widehat{\sigma}^2_{\nu}) \tag{19}$$

$$\widetilde{q}(\theta) \triangleq \mathcal{N}(\widehat{\theta}_A, \widehat{\sigma}^2_{\theta_A}). \tag{20}$$

Again writing out the ELBO,

$$\begin{aligned}
\mathcal{L}(\widetilde{q}) &= \mathbb{E}_{\widetilde{q}}\left[\log \frac{p(X,\nu,\theta)}{\widetilde{q}(\nu)\widetilde{q}(\theta)}\right] \\
&= \mathbb{E}_{\widetilde{q}}\left[\log p(X \mid \nu,\theta)\right] - \mathbb{E}_{\widetilde{q}}\left[\log \frac{\widetilde{q}(\nu)}{p(\nu)}\right] - \mathbb{E}_{\widetilde{q}}\left[\log \frac{\widetilde{q}(\theta)}{p(\theta)}\right] \\
&= \frac{2X\widehat{\nu} + 2X\widehat{\theta} - 2\widehat{\nu}\widehat{\theta} - \widehat{\sigma}^2_{\nu} - \widehat{\nu}^2 - \widehat{\sigma}^2_{\theta_A} - \widehat{\theta}^2}{2} + \log \widehat{\sigma}_{\nu} - \frac{\widehat{\sigma}^2_{\nu} + \widehat{\nu}^2}{2V} + \log \widehat{\sigma}_{\theta_A} + \text{const.}
\end{aligned}$$

The coordinate ascent updates are

$$\widehat{\nu}^{(t+1)} = \frac{V(X - \widehat{\theta}^{(t)})}{1+V} \tag{21}$$

$$\widehat{\sigma}^{2(t+1)}_{\nu} = \frac{V}{1+V} \tag{22}$$

$$\widehat{\theta}^{(t+1)} = X - \widehat{\nu}^{(t+1)} \tag{23}$$

$$\widehat{\sigma}^{2(t+1)}_{\theta_A} = 1 \tag{24}$$

The variational parameter for the posterior mean of $\theta$ given $X$ satisfies

$$\left|\widehat{\theta}^{(t+1)} - X\right| = \left|\widehat{\nu}^{(t+1)}\right| = \frac{V}{1+V}\left|X - \widehat{\theta}^{(t)}\right|, \tag{25}$$

this parameter converges geometrically with rate $\frac{V}{1+V}$.

# 3 ASIS-CAVI

Consider the following algorithm for *ancillary sufficient interweaving scheme-coordinate ascent variational inference*, as inspired by Yu and Meng (2011).

1. Update $q(\mu)$ using the CAVI update in the SA model,
2. Update $q(\theta)$ using the CAVI update in the SA model,
3. Reparametrize: choose $\widetilde{q}(\nu)$, $\widetilde{q}(\theta)$ to minimize
$$\min_{\widetilde{q}(\nu),\widetilde{q}(\theta)} D\left(\widetilde{q}(\nu)\widetilde{q}(\theta)\middle\|q(\nu+\theta)q(\theta)\right)$$
4. Update $\widetilde{q}(\nu)$ using the CAVI update in the AA model,
5. Update $\widetilde{q}(\theta)$ using the CAVI update in the AA model,
6. Reparametrize: choose $q(\mu)$, $q(\theta)$ to minimize
$$\min_{q(\mu),q(\theta)} D\left(q(\mu)q(\theta)\middle\|\widetilde{q}(\mu-\theta)\widetilde{q}(\theta)\right)$$
7. Repeat 1 through 6 until convergence.

**Example.** Returning to the normal-normal model, we need to solve the reparametrization steps.

$$D\left(\widetilde{q}(\nu)\widetilde{q}(\theta)\middle\|q(\nu+\theta)q(\theta)\right) = \mathbb{E}_{\widetilde{q}}[\log q(\nu+\theta)q(\theta)] - H(\widetilde{q}) \tag{26}$$

$$= \mathbb{E}_{\widetilde{q}}\left[\log \frac{1}{\sqrt{2\pi\widehat{\sigma}_\mu^2}}\exp\left(-\frac{(\nu+\theta-\widehat{\mu})^2}{2\widehat{\sigma}_\mu^2}\right)\frac{1}{\sqrt{2\pi\widehat{\sigma}_{\theta_S}^2}}\exp\left(-\frac{(\theta-\widehat{\theta}_S)^2}{2\widehat{\sigma}_{\theta_S}^2}\right)\right] - H(\widetilde{q}) \tag{27}$$

$$= \text{const.} + \mathbb{E}_{\widetilde{q}}\left[-\frac{(\nu+\theta-\widehat{\mu})^2}{2\widehat{\sigma}_\mu^2} - \frac{(\theta-\widehat{\theta}_S)^2}{2\widehat{\sigma}_{\theta_S}^2}\right] + \log(2\pi e\widehat{\sigma}_\nu\widehat{\sigma}_{\theta_A}) \tag{28}$$

$$= \text{const.} - \frac{\widehat{\nu}^2 + \widehat{\sigma}_\nu^2 + \widehat{\theta}_A^2 + \widehat{\sigma}_{\theta_A}^2 + \widehat{\mu}^2 - 2\widehat{\nu}\widehat{\mu} - 2\widehat{\theta}_A\widehat{\mu} + 2\widehat{\nu}\widehat{\theta}_A}{2\widehat{\sigma}_\mu^2} \tag{29}$$

$$- \frac{\widehat{\theta}_A^2 + \widehat{\sigma}_{\theta_A}^2 + \widehat{\theta}_S^2 - 2\widehat{\theta}_S\widehat{\theta}_A}{2\widehat{\sigma}_{\theta_S}^2} + \log(2\pi e\widehat{\sigma}_\nu\widehat{\sigma}_{\theta_A}) \tag{30}$$

Setting derivatives equal to zero and finding fixed points,

$$\widehat{\nu} = \widehat{\mu} - \widehat{\theta}_S \tag{31}$$

$$\widehat{\theta}_A = \widehat{\theta}_S \tag{32}$$

$$\sigma_\nu^2 = \widehat{\sigma}_\mu^2 = \frac{V}{V+1} \tag{33}$$

$$\widehat{\sigma}_{\theta_A}^2 = \left(\frac{1}{\widehat{\sigma}_\mu^2} + \frac{1}{\widehat{\sigma}_{\theta_S}^2}\right)^{-1} = \frac{V}{V+2} \tag{34}$$

Similarly deriving step (6),

$$\widehat{\mu} = \widehat{\nu} + \widehat{\theta}_A \tag{35}$$

$$\widehat{\theta}_S = \widehat{\theta}_A \tag{36}$$

# 4 Alternate ASIS-CAVI

Consider the following algorithm for *ancillary sufficient interweaving scheme-coordinate ascent variational inference*, as inspired by Yu and Meng (2011).

1. Update $q_\mu(\mu)$ using the CAVI update in the SA model,
2. Update $q_\theta(\theta)$ using the CAVI update in the SA model,
3. Reparametrize: choose $\widetilde{q}_\nu(\nu)$ to minimize

$$\min_{\widetilde{q}_\nu(\nu)} D\left(q_\mu(\mu)\middle\|\widetilde{q}_\nu(\mu-\theta)\right)$$

4. Update $\widetilde{q}_\theta(\theta)$ using the CAVI update in the AA model,
5. Repeat 1 through 4 until convergence.

**Example.** Returning to the normal-normal model, the only step we have yet to solve is (3)

$$\begin{aligned}
D\left(q_\mu(\mu)\middle\|\widetilde{q}_\nu(\mu-\theta)\right) &= \mathbb{E}_q\left[\log\frac{q_\mu(\mu)}{\widetilde{q}_\nu(\mu-\theta)}\right] &(37)\\
&= \text{const.} - \mathbb{E}_q\left[\log\widetilde{q}_\nu(\mu-\theta)\right] &(38)\\
&= \text{const.} - \mathbb{E}_q\left[\log\frac{1}{\sqrt{2\pi\widehat{\sigma}_\nu^2}}\exp\left\{-\frac{(\mu-\theta-\widehat{\nu})^2}{2\widehat{\sigma}_\nu^2}\right\}\right] &(39)\\
&= \text{const.} + \log\widehat{\sigma}_\nu + \frac{\widehat{\nu}^2 - 2\widehat{\mu}\widehat{\nu} + 2\widehat{\theta}\widehat{\nu} + \widehat{\mu}^2 + \widehat{\sigma}_\mu^2 + \widehat{\theta}^2 + \widehat{\sigma}_{\theta_S}^2 - 2\widehat{\mu}\widehat{\theta}}{2\widehat{\sigma}_\nu^2} &(40)
\end{aligned}$$

this yields

$$\begin{aligned}
\widehat{\nu}^{(t)} &= \widehat{\mu}^{(t)} - \widehat{\theta}^{(t)} &(41)\\
\widehat{\sigma}_\nu^{2(t)} &= \widehat{\sigma}_\mu^{2(t)} + \widehat{\sigma}_{\theta_S}^{2(t)} = \frac{V+2}{V+1}V. &(42)
\end{aligned}$$

So the whole algorithm listed above is

$$\begin{aligned}
\widehat{\mu}^{(t+1)} &= \frac{VX+\widehat{\theta}^{(t)}}{1+V} &(43)\\
\widehat{\sigma}_\mu^{2(t+1)} &= \frac{V}{1+V} &(44)\\
\widehat{\theta}^{(t+1)} &= \widehat{\mu}^{(t+1)} &(45)\\
\widehat{\sigma}_\theta^{2(t+1)} &= V &(46)\\
\widehat{\nu}^{(t+1)} &= \widehat{\mu}^{(t+1)} - \widehat{\theta}^{(t+1)} = 0 &(47)\\
\widehat{\sigma}_\nu^{2(t+1)} &= \frac{V+2}{V+1}V &(48)\\
\widehat{\theta}^{(t+1)} &= X - \widehat{\nu}^{(t+1)} = X &(49)\\
\sigma_\theta^{2(t+1)} &= 1 &(50)
\end{aligned}$$

The algorithm converges in one iteration.