8. Appendix

This appendix contains additional plots and proofs of the results from Section 2.

Lemma 6. The divergence from $q(z)$ to $p(z)$ is

$$KL(q(Z)\|p(Z)) = KL(q(Z|W)\|p(Z)) - I_q[W, Z],$$

where $D_0 = E_{q(W,Z)} \log (q(Z|W)/p(Z))$ is conditional divergence and $I_q$ denotes mutual information under $q$.

Proof. Define the joint distribution $p(w, z) = q(w)p(z)$. Then, the chain-rule of KL-divergence (Cover & Thomas, 2006, Thm. 2.5.3) states that

$$KL(q(Z, W)\|p(Z, W)) = KL(q(W|Z)\|p(W|Z)) + KL(q(Z)\|p(Z)).$$ (27)

The left-hand side simplifies into $D_0$, and the first term on the right-hand side simplifies into $I_q[W, Z]$. $\Box$

Theorem 7. For fixed values of $\beta$ and $p(w|z)$, the distribution $q^*(w)$ that minimizes $D_\beta$ is

$$q^*(w) = \exp(s(w) - A)$$

$$A = \log \int_w \exp s(w)$$

$$s(w) = \log p(w) - KL(q(Z|w)\|p(Z|w))$$

$$- (\beta^{-1} - 1) KL(q(Z|w)\|p(Z)) .$$

Moreover, at $q^*$, the objective value is $D_\beta^* = -\beta A$.

Proof. First, consider derivatives of $D_0$ and $D_1$ with respect to $q(w)$. The first can immediately be seen to be

$$\frac{dD_0}{dq(w)} = KL(q(Z|w)\|p(Z)) .$$

For the second, we can derive

$$\frac{dD_1}{dq(w)} = \frac{d}{dq(w)} \int_{w,z} q(w, z) \log \frac{q(z|w)}{p(w, z)}$$

$$+ \frac{d}{dq(w)} \int_{w,z} q(w) \log q(w)$$

$$= \int_z q(z|w) \log \frac{q(z|w)}{p(w, z)} + \log q(w) + 1$$

$$= KL(q(Z|w)\|p(Z|w)) - \log p(w) + \log q(w) + 1.$$

If we create a Lagrangian for $D_\beta$ with a Lagrange multiplier $\lambda$ to enforce normalization of $q(w)$, we know that at the optimal $q(w)$ its gradient will be zero. Using the above derivatives, we therefore have that

$$0 = (1 - \beta) KL(q(Z|w)\|p(Z)) + \beta KL(q(Z|w)\|p(Z|w))$$

$$- \beta \log p(w) + \beta \log q(w) + \lambda,$$

Which solved for $q(w)$, this gives

$$q(w) \propto \exp( - (1 - \beta^{-1}) KL(q(Z|w)\|p(Z))$$

$$- KL(q(Z|w)\|p(Z|w)) + \log p(w) ) - \beta A.$$ (29)

Now, taking the left-hand side and terms in the bottom line, we can recognize that

$$\int_w q(w) \left( \log \frac{p(w)}{q(w)} - KL(q(Z|w)\|p(Z|w)) \right) = -D_1.$$

Further, if we take the terms from the middle line, we have that

$$-\beta \int_w q(w)(1 - \beta^{-1}) KL(q(Z|w)\|p(Z)) = (\beta - 1) D_0.$$

Thus, we can re-write Eq. 29 as $-\beta A = (1 - \beta) D_0 + \beta D_1$, establishing the value of $D_\beta^*$. $\Box$

Remark 8. In the limit where $\beta \to 0$ the divergence bound becomes

$$\lim_{\beta \to 0} D_\beta^* = \inf_w KL(q(Z|w)\|p(Z)).$$

Proof. Use the representation that $\lim_{\beta \to 0} D_\beta^* = \lim_{\beta \to 0} -\beta A$ is equal to

$$\lim_{\beta \to 0} -\beta \log \int_w \exp \left( \log p(w) - KL(q(Z|w)\|p(Z|w)) \right)$$

$$-(\beta^{-1} - 1) KL(q(Z|w)\|p(Z)) \right)$$

$$= \lim_{\beta \to 0} -\beta \log \int_w \exp \left( -\beta^{-1} KL(q(Z|w)\|p(Z)) \right).$$

The form for $D_\beta^*$ follows from the fact that $\lim_{\beta \to 0} \beta \log \int_w \exp(\beta^{-1} f(w)) = \sup_w f(w)$. $\Box$
Lemma 9. If \( p(w|z) = r(w)q(z|w)/r_z \) and \( r_z \) is a constant, then the solution in Thm. 3 holds with

\[
\begin{aligned}
s(w) &= \log r(w) - \log r_z \\
&\quad \quad + \mathbb{E}_{q(w)}[\beta^{-1}\log p(z) + (1 - \beta^{-1})\log q(z|w)].
\end{aligned}
\]

Proof. First, without using the particular form for \( p(w|z) \), we can write \( s(w) \) as

\[
\begin{aligned}
\log p(w) - \int_z q(z|w) \log \frac{q(z|w)}{p(z|w)} \\
&\quad - (\beta^{-1} - 1) \int_z q(z|w) \log \frac{q(z|w)}{p(z)}
\end{aligned}
\]

Cancelling terms involving \( q(z|w) \) in the numerators, this is

\[
\begin{aligned}
\log p(w) - \int_z q(z|w) \log \frac{p(z)}{p(z|w)} \\
&\quad - \beta^{-1} \int_z q(z|w) \log \frac{q(z|w)}{p(z)}
\end{aligned}
\]

The \( \log p(w) \) can be absorbed into the first term to give, after some cancellation that

\[
s(w) = \int_z q(z|w) \log p(w|z) - \beta^{-1} KL(q(Z|w)\|p(Z)).
\]

Now, using the assumed form for \( p(w|z) \), we can immediately write that \( s(w) \) is

\[
\int_z q(z|w) \log \frac{r(w)q(z|w)}{r_z} - \beta^{-1} \int_z q(z|w) \log \frac{q(z|w)}{p(z)},
\]

equivalent to the form stated. \( \square \)
Figure 5. Examples sampling from a two-dimensional mixture of three gaussians after running inference for $5 \times 10^5$ iterations. The sampled weights $\nu$ are pictured as ellipsoids at one standard deviation. Colored contours show the density $p(z)$. To avoid visual clutter, a smaller number (equally spaced) of samples are shown for smaller $\beta$. 
Figure 6. Examples sampling from a two-dimensional “donut” distribution after running inference for $5 \times 10^5$ iterations. The sampled weights $w$ are pictured as ellipsoids at one standard deviation. Colored contours show the density $p(z)$. To avoid visual clutter, a smaller number (equally spaced) of samples are shown for smaller $\beta$. 

\begin{align*}
\beta &= 0 & \beta &= 0.05 & \beta &= 0.10 & \beta &= 0.15 & \beta &= 0.20 \\
\beta &= 0.25 & \beta &= 0.30 & \beta &= 0.35 & \beta &= 0.40 & \beta &= 0.45 \\
\beta &= 0.50 & \beta &= 0.55 & \beta &= 0.60 & \beta &= 0.65 & \beta &= 0.70 \\
\beta &= 0.75 & \beta &= 0.80 & \beta &= 0.85 & \beta &= 0.90 & \beta &= 0.95
\end{align*}
A Divergence Bound for Hybrids of MCMC and Variational Inference and an Application to Langevin Dynamics and SGVI

Figure 7. Inference for various values of $\beta$ on ionosphere after $10^4$ (top row) $10^5$ (middle row) or $10^6$ (bottom row) iterations. After each iteration, one sample is drawn from $q_w(Z)$, and plots show the first two principal components (computed on samples from Stan). Each plot show samples resulting from the (constant) step-size $\epsilon$ that resulted in the minimum MMD for that $\beta$ and number of iterations. The same sequence of random numbers is for all inference methods. (More results are in the appendix.)

Figure 8. Inference for various values of $\beta$ on a1a after $10^4$ (top row) $10^5$ (middle row) or $10^6$ (bottom row) iterations. In some of these plots, a “tail” is visible, reflecting the path into the high-density region from where $w = 0$ where inference was initialized.
Figure 9. Inference for various values of $\beta$ on *australian* after $10^4$ (top row) $10^5$ (middle row) or $10^6$ (bottom row) iterations.

Figure 10. Inference for various values of $\beta$ on *sonar* after $10^4$ (top row) $10^5$ (middle row) or $10^6$ (bottom row) iterations.