Semi-Implicit Variational Inference: Supplementary Material

Mingzhang Yin and Mingyuan Zhou

Algorithm 1 Semi-Implicit Variational Inference (SIVI)

input : Data $\{x_i\}_{1:N}$, joint likelihood $p(\boldsymbol{x}, \boldsymbol{z})$, explicit variational distribution $q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid \boldsymbol{\psi})$ with reparameterization $\boldsymbol{z} = f(\boldsymbol{\epsilon}, \boldsymbol{\xi}, \boldsymbol{\psi})$, $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$, implicit layer neural network $T_{\boldsymbol{\phi}}(\epsilon)$ and source of randomness $q(\boldsymbol{\epsilon})$

output: Variational parameter $\boldsymbol{\xi}$ for the conditional distribution $q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid \boldsymbol{\psi})$, variational parameter $\boldsymbol{\phi}$ for the mixing distribution $q_{\boldsymbol{\phi}}(\boldsymbol{\psi})$

Initialize $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ randomly

while *not converged* **do**

  Set $\underline{L}_{K_t} = 0$, ρ_t and η_t as step sizes, and $K_t \geq 0$ as a non-decreasing integer; Sample $\boldsymbol{\psi}^{(k)} = T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(k)})$, $\boldsymbol{\epsilon}^{(k)} \sim q(\boldsymbol{\epsilon})$ for $k = 1, \ldots, K_t$; take subsample $\mathbf{x} = \{x_i\}_{i_1:i_M}$

  for $j = 1$ **to** J **do**

    Sample $\boldsymbol{\psi}_j = T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}_j)$, $\boldsymbol{\epsilon}_j \sim q(\boldsymbol{\epsilon})$

    Sample $\boldsymbol{z}_j = f(\tilde{\boldsymbol{\epsilon}}_j, \boldsymbol{\xi}, \boldsymbol{\psi}_j)$, $\tilde{\boldsymbol{\epsilon}}_j \sim p(\boldsymbol{\varepsilon})$

    $\underline{L}_{K_t} = \underline{L}_{K_t} + \frac{1}{J}\big\{ - \log \frac{1}{K_t+1}\big[\sum_{k=1}^{K_t} q_{\boldsymbol{\xi}}(\boldsymbol{z}_j \mid \boldsymbol{\psi}^{(k)}) + q_{\boldsymbol{\xi}}(\boldsymbol{z}_j \mid \boldsymbol{\psi}_j)\big] + \frac{N}{M} \log p(\mathbf{x} \mid \boldsymbol{z}_j) + \log p(\boldsymbol{z}_j)\big\}$

  end

  $t = t + 1$

  $\boldsymbol{\xi} = \boldsymbol{\xi} + \rho_t \nabla_{\boldsymbol{\xi}} \underline{L}_{K_t}\big(\{\boldsymbol{\psi}^{(k)}\}_{1,K_t}, \{\boldsymbol{\psi}_j\}_{1,J}, \{\boldsymbol{z}_j\}_{1,J}\big)$

  $\boldsymbol{\phi} = \boldsymbol{\phi} + \eta_t \nabla_{\boldsymbol{\phi}} \underline{L}_{K_t}\big(\{\boldsymbol{\psi}^{(k)}\}_{1,K_t}, \{\boldsymbol{\psi}_j\}_{1,J}, \{\boldsymbol{z}_j\}_{1,J}\big)$

end

A. Proofs

Proof of Inequility (3). To prove a functional form of Jensen's Inequality, let $h(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})} q(\boldsymbol{z} | \boldsymbol{\psi})$ and $\langle f, g \rangle_{L^2} = \int f(\boldsymbol{z}) g(\boldsymbol{z}) d\boldsymbol{z}$. From Theorem 1, we have convexity, and according to Theorem 6.2.1. of Kurdila & Zabarankin (2005), we have an equivalent first-order definition for convexity as

$$\begin{aligned}\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{\psi})||p(\boldsymbol{z})) \geq & \mathrm{KL}(h(\boldsymbol{z})||p) + \\ & \langle q(\boldsymbol{z}|\boldsymbol{\psi}) - h(\boldsymbol{z}), \nabla_q \mathrm{KL}(q||p)|_{h(\boldsymbol{z})} \rangle_{L^2}\end{aligned}$$

Taking the expectation with respect to $\boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})$ on both sides, we have

$$\begin{aligned}& \mathbb{E}_{\boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})} \mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{\psi})||p(\boldsymbol{z})) \\ \geq & \ \mathrm{KL}(h(\boldsymbol{z})||p(\boldsymbol{z})) \\ & + \mathbb{E}_{\boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})} [\langle q(\boldsymbol{z}|\boldsymbol{\psi}) - h(\boldsymbol{z}), \nabla_q \mathrm{KL}(q||p)|_{h(\boldsymbol{z})} \rangle_{L^2}] \\ = & \ \mathrm{KL}(h(\boldsymbol{z})||p(\boldsymbol{z})) \\ = & \ \mathrm{KL}(\mathbb{E}_{\boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})} q(\boldsymbol{z}|\boldsymbol{\psi})||p(\boldsymbol{z})).\end{aligned}$$

□

Proof of Proposition 1. We show that directly maximizing the lower bound $\underline{\mathcal{L}}$ of ELBO in (4) may drive $q(\boldsymbol{\psi})$ towards degeneracy. For VI that uses $q(\boldsymbol{z} \mid \boldsymbol{\psi})$ as its variational distribution, if supposing $\boldsymbol{\psi}^*$ is the optimum variational parameter, which means

$$\boldsymbol{\psi}^* = \arg\max_{\boldsymbol{\psi}} -\mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{\psi})} \log \frac{q(\boldsymbol{z}|\boldsymbol{\psi})}{p(\boldsymbol{x}, \boldsymbol{z})},$$

then we have

$$\begin{aligned}\underline{\mathcal{L}} &= -\mathbb{E}_{\boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})} \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{\psi})} \log \frac{q(\boldsymbol{z}|\boldsymbol{\psi})}{p(\boldsymbol{x}, \boldsymbol{z})} \\ &= \int q_{\boldsymbol{\phi}}(\boldsymbol{\psi}) [-\mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{\psi})} \log \frac{q(\boldsymbol{z}|\boldsymbol{\psi})}{p(\boldsymbol{x}, \boldsymbol{z})}] d\boldsymbol{\psi} \\ &\leq \int q_{\boldsymbol{\phi}}(\boldsymbol{\psi}) d\boldsymbol{\psi} [-\mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{\psi}^*)} \log \frac{q(\boldsymbol{z}|\boldsymbol{\psi}^*)}{p(\boldsymbol{x}, \boldsymbol{z})}] \\ &= -\mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{\psi}^*)} \log \frac{q(\boldsymbol{z}|\boldsymbol{\psi}^*)}{p(\boldsymbol{x}, \boldsymbol{z})}.\end{aligned}$$

The equality in the above equation is reached if and only if $q(\boldsymbol{\psi}) = \delta_{\boldsymbol{\psi}^*}(\boldsymbol{\psi})$, which means the mixing distribution degenerates to a point mass density and hence SIVI degenerates to vanilla VI. □

Proof of Proposition 2. $B_0 = 0$ is trivial. Denote $\boldsymbol{\psi}^{(0)} = \psi_v$. For iid samples $\boldsymbol{\psi}^{(k)} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})$, when $K \to \infty$, by the strong law of large numbers, $\tilde{h}_K(\boldsymbol{z}) = \frac{\sum_{k=0}^{K} q(\boldsymbol{z} \mid \boldsymbol{\psi}^{(k)})}{K+1}$ converges almost surely to $\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\psi})} q(\boldsymbol{z} \mid \boldsymbol{\psi}) = h_{\boldsymbol{\phi}}(\boldsymbol{z})$. To prove (6), by the strong law of large numbers, we first rewrite it as the limit of a double sequence $S(K, J)$, where $K, J \in \{1, 2, \ldots, \}$, and check the condition for the interchange of iterated limits (Rudin, 1964; Habil, 2016): i) The double limit exists; ii) Fixing one index of the double sequence, the one side limit exists for the other index .

$$\begin{aligned}& \lim_{K \to \infty} \mathbb{E}_{\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \cdots, \boldsymbol{\psi}^{(K)} \sim q(\boldsymbol{\psi})} \log \frac{\sum_{k=0}^{K} q(\boldsymbol{z} \mid \boldsymbol{\psi}^{(k)})}{K+1} \\ = & \lim_{K \to \infty} \lim_{J \to \infty} \frac{1}{J} \sum_{j=1}^{J} \log \frac{1}{K+1} \sum_{k=0}^{K} q(\boldsymbol{z} \mid \boldsymbol{\psi}_j^{(k)}) \\ \triangleq & \lim_{K \to \infty} \lim_{J \to \infty} S(K, J).\end{aligned}$$

Here $\boldsymbol{\psi}_j^{(k)}$ are iid samples from $q(\boldsymbol{\psi})$. For i) we show double limit $\lim_{K,J \to \infty} S(K, J) = \log h(\boldsymbol{z})$. For $\forall \epsilon > 0$, $\exists N(\epsilon)$, when $K, J > N(\epsilon)$, $|\log \frac{1}{K+1} \sum_{k=0}^{K} q(\boldsymbol{z} \mid \boldsymbol{\psi}_j^{(k)}) - \log h(\boldsymbol{z})| < \epsilon$ thanks to the law of large numbers, then

$$\begin{aligned}& \left| \sum_{j=1}^{J} \log \frac{1}{K+1} \sum_{k=0}^{K} q(\boldsymbol{z} \mid \boldsymbol{\psi}_j^{(k)}) - J \log h(\boldsymbol{z}) \right| \\ \leq & \sum_{j=1}^{J} \left| \log \frac{1}{K+1} \sum_{k=0}^{K} q(\boldsymbol{z} \mid \boldsymbol{\psi}_j^{(k)}) - \log h(\boldsymbol{z}) \right| \leq J\epsilon.\end{aligned}$$

Deviding both sides by J we get $|S(K,J) - \log h(\boldsymbol{z})| \leq \epsilon$ when $K, J > N(\epsilon)$. By definition, we have $\lim_{K,J\to\infty} S(K,J) = \log h(\boldsymbol{z})$.
ii) for each fixed $J \in \mathbb{N}$, $\lim_{K\to\infty} S(K,J) = \log h(\boldsymbol{z})$ exists; for each fixed $K \in \mathbb{N}$, $\lim_{J\to\infty} S(K,J) = \mathbb{E}_{\boldsymbol{\psi}^{(0)},\boldsymbol{\psi}^{(1)},\cdots,\boldsymbol{\psi}^{(K)}\sim q(\boldsymbol{\psi})} \log \frac{\sum_{k=0}^{K} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(k)})}{K+1} \leq \log h(\boldsymbol{z})$ also exists. The limitation can then be interchanged and (6) is proved. Therefore, we have

$$
\begin{aligned}
&\lim_{k\to\infty} \underline{\mathcal{L}}_k = \underline{\mathcal{L}} + \mathbb{E}_{\boldsymbol{\psi}}\mathrm{KL}(q(\boldsymbol{z}\,|\,\boldsymbol{\psi})||h_{\boldsymbol{\phi}}(\boldsymbol{z})) \\
&= \mathbb{E}_{\boldsymbol{\psi}\sim q(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}\,|\,\boldsymbol{\psi})}\left[\log\frac{q(\boldsymbol{z}\,|\,\boldsymbol{\psi})}{h_{\boldsymbol{\phi}}(\boldsymbol{z})} - \log\frac{q(\boldsymbol{z}\,|\,\boldsymbol{\psi})}{p(x,\boldsymbol{z})}\right] \\
&= -\mathbb{E}_{\boldsymbol{\psi}\sim q(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}\,|\,\boldsymbol{\psi})}\log\frac{h_{\boldsymbol{\phi}}(\boldsymbol{z})}{p(x,\boldsymbol{z})} \\
&= -\mathbb{E}_{\boldsymbol{z}\sim h_{\boldsymbol{\phi}}(\boldsymbol{z})}\log\frac{h_{\boldsymbol{\phi}}(\boldsymbol{z})}{p(x,\boldsymbol{z})} = \mathcal{L}.
\end{aligned}
$$

□

Proof of Proposition 3. Assume integer $K > M > 0$. Let $\mathcal{I}$ be the set that consists of all the subsets of $\{1,\cdots,K\}$ with cardinality M. Let I be a discrete uniform random variable and for element $\{i_1,\cdots,i_M\} \in \mathcal{I}$, $P(I = \{i_1,\cdots,i_M\}) = \frac{1}{\binom{K}{M}}$. We have $\mathbb{E}_I \frac{1}{M}\sum_{i\in I} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^i) = \frac{1}{K}\sum_{i=1}^{K} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^i)$. To show $\bar{\mathcal{L}}_K = \bar{\mathcal{L}} - A_K$ is monotonic decreasing, we only need to show A_K is monotonic increasing:

$$
\begin{aligned}
A_K =& \mathbb{E}_{\boldsymbol{\psi}\sim q(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{z}\sim h_{\boldsymbol{\phi}}(\boldsymbol{z})}\mathbb{E}_{\boldsymbol{\psi}^{(1)},\cdots,\boldsymbol{\psi}^{(K)}\sim q(\boldsymbol{\psi})} \\
&\log\frac{\frac{1}{K}\sum_{i=1}^{K} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(i)})}{q(\boldsymbol{z}\,|\,\boldsymbol{\psi})} \\
=& \mathbb{E}_{\boldsymbol{\psi}\sim q(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{z}\sim h_{\boldsymbol{\phi}}(\boldsymbol{z})}\mathbb{E}_{\boldsymbol{\psi}^{(1)},\cdots,\boldsymbol{\psi}^{(K)}\sim q(\boldsymbol{\psi})} \\
&\log\mathbb{E}_I\left[\frac{\frac{1}{M}\sum_{i\in I} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(i)})}{q(\boldsymbol{z}\,|\,\boldsymbol{\psi})}\right] \\
\geq& \mathbb{E}_{\boldsymbol{\psi}\sim q(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{z}\sim h_{\boldsymbol{\phi}}(\boldsymbol{z})}\mathbb{E}_{\boldsymbol{\psi}^{(1)},\cdots,\boldsymbol{\psi}^{(K)}\sim q(\boldsymbol{\psi})} \\
&\mathbb{E}_I\log\frac{\frac{1}{M}\sum_{i\in I} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(i)})}{q(\boldsymbol{z}\,|\,\boldsymbol{\psi})} \\
=& \mathbb{E}_{\boldsymbol{\psi}\sim q(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{z}\sim h_{\boldsymbol{\phi}}(\boldsymbol{z})}\mathbb{E}_{\boldsymbol{\psi}^{(1)},\cdots,\boldsymbol{\psi}^{(M)}\sim q(\boldsymbol{\psi})} \\
&\log\frac{\frac{1}{M}\sum_{i=1}^{M} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(i)})}{q(\boldsymbol{z}\,|\,\boldsymbol{\psi})} \\
=& A_M.
\end{aligned}
$$

We now show $\lim_{K\to\infty}\bar{\mathcal{L}}_K = \mathcal{L}$. Again, by the strong law of large numbers, $\frac{1}{K}\sum_{i=1}^{K} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(i)})$ converges almost surely to $\mathbb{E}_{\boldsymbol{\psi}\sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}) = h_{\boldsymbol{\phi}}(\boldsymbol{z})$ and hence

$$
\begin{aligned}
&\lim_{K\to\infty}\bar{\mathcal{L}}_K = \bar{\mathcal{L}} + \mathbb{E}_{\boldsymbol{\psi}}\mathrm{KL}(h_{\boldsymbol{\phi}}(\boldsymbol{z})||q(\boldsymbol{z}\,|\,\boldsymbol{\psi})) \\
&= -\mathbb{E}_{\boldsymbol{z}\sim h_{\boldsymbol{\phi}}(\boldsymbol{z})}\mathbb{E}_{\boldsymbol{\psi}\sim q(\boldsymbol{\psi})}\left[\log\frac{q(\boldsymbol{z}\,|\,\boldsymbol{\psi})}{p(x,\boldsymbol{z})} + \log\frac{h_{\boldsymbol{\phi}}(\boldsymbol{z})}{q(\boldsymbol{z}\,|\,\boldsymbol{\psi})}\right] \\
&= \mathcal{L}.
\end{aligned}
$$

□

Proof of Equation (11). The gradient of B_K with respect to $\boldsymbol{\phi}$ can be expressed as

$$
\begin{aligned}
&\nabla_{\boldsymbol{\phi}} B_K = \nabla_{\boldsymbol{\phi}}\mathbb{E}_{\boldsymbol{\psi}\sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{\psi}^{(1)},\ldots,\boldsymbol{\psi}^{(K)}\sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})}\Bigg[\\
&\quad \mathrm{KL}\left(q(\boldsymbol{z}\,|\,\boldsymbol{\psi})\middle\|\frac{q(\boldsymbol{z}\,|\,\boldsymbol{\psi})+\sum_{k=1}^{K} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(k)})}{K+1}\right)\Bigg] \\
&= \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}^{(1)},\ldots,\boldsymbol{\epsilon}^{(K)}\sim p(\boldsymbol{\epsilon})}\nabla_{\boldsymbol{\phi}}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}\,|\,T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}))}\Bigg[\\
&\quad \log\frac{q(\boldsymbol{z}\,|\,T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}))}{\frac{q(\boldsymbol{z}\,|\,T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}))+\sum_{k=1}^{K} q(\boldsymbol{z}\,|\,T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(k)}))}{K+1}}\Bigg] \\
&= \mathbb{E}_{\boldsymbol{\epsilon},\ldots,\boldsymbol{\epsilon}^{(K)}}\nabla_{\boldsymbol{\phi}}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}\,|\,T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}))}\log\left[r_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\epsilon},\boldsymbol{\epsilon}^{(1:K)})\right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon},\ldots,\boldsymbol{\epsilon}^{(K)}\sim p(\boldsymbol{\epsilon})}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}\,|\,T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}))}\Bigg\{ \\
&\quad q(\boldsymbol{z}\,|\,T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}))\nabla_{\boldsymbol{\phi}}\log\left[r_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\epsilon},\boldsymbol{\epsilon}^{(1:K)})\right] \\
&\quad + \left[\nabla_{\boldsymbol{\phi}}\log q(\boldsymbol{z}\,|\,T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}))\right]\log\left[r_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\epsilon},\boldsymbol{\epsilon}^{(1:K)})\right]\Bigg\}.
\end{aligned}
$$

□

B. Bayesian Logistic Regression

We consider datesets *waveform* ($n = 5000$, $V = 21$, and $400/4600$ for training/testing), *spam* ($n = 3000$, $V = 2$, and 2000/1000 for training/testing), and *nodal* ($n = 53$, $V = 5$, and 25/28 for training/testing). The training-set-size to feature-dimension ratio n_{train}/V varies in these three datasets, and we expect the posterior uncertainty to be large if this ratio is small.

The contribution of observation i to the likelihood can be expressed as

$$
P(y_i\,|\,\boldsymbol{x}_i,\boldsymbol{\beta}) = \frac{e^{y\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} \quad \propto e^{(y-\frac{1}{2})\boldsymbol{x}_i'\boldsymbol{\beta}}\mathbb{E}_{\omega_i}\left[e^{-\frac{\omega_i(\boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}}\right],
$$

where the expectation is taken respect to a Pólya-Gamma (PG) distribution (Polson et al., 2013) as $\omega_i \sim \mathrm{PG}(1,0)$, and hence we have an augmented likelihood as

$$
P(y_i,\omega_i\,|\,\boldsymbol{x}_i,\boldsymbol{\beta}) \propto e^{(y-\frac{1}{2})\boldsymbol{x}_i'\boldsymbol{\beta}-\frac{1}{2}\omega_i(\boldsymbol{x}_i'\boldsymbol{\beta})^2}.
$$

B.1. Gibbs Sampling via Data Augmentation

Denoting $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)'$, $\boldsymbol{y} = (y_1, \ldots, y_N)'$, $\mathbf{A} = \text{diag}(\alpha_0, \ldots, \alpha_V)'$, and $\boldsymbol{\Omega} = \text{diag}(\omega_1, \ldots, \omega_N)$, we have

$$(\omega_i \,|\, -) \sim \text{PG}(1, \boldsymbol{x}_i'\boldsymbol{\beta}), \;\; (\boldsymbol{\beta} \,|\, -) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = (\mathbf{A} + \mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{X}'(\boldsymbol{y} - 1/2)$. To sample from the Pólya-Gamma distribution, a random sample from which can be generated as a weighted sum of an infinite number of iid gamma random variables, we follow Zhou (2016) to truncate the infinite sum to the summation of M gamma random variables, where the parameters of the Mth gamma random variable are adjusted to match the mean and variance of the finite sum with those of the infinite sum. We set $M = 5$ in this paper.

B.2. Mean-Field Variational Inference with Diagonal Covariance Matrix

We choose a fully factorized Q distribution as

$$Q = \left[\prod\nolimits_i q(\omega_i)\right] \left[\prod\nolimits_v q(\beta_v)\right].$$

To exploit conjugacy, defining

$$q(\omega_i) = \text{PG}\left(1, \lambda_i\right),$$
$$q(\beta_v) = \mathcal{N}(\mu_v, \sigma_v^2),$$

we have closed-form coordinate ascent variational inference update equations as

$$\lambda_i = \sqrt{\mathbb{E}[(\boldsymbol{x}_i'\boldsymbol{\beta})^2]} = \sqrt{\boldsymbol{x}_i'\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}']\boldsymbol{x}_i},$$
$$\sigma_v^2 = \left(\mathbb{E}[\alpha_v] + \sum\nolimits_i \mathbb{E}[\omega_i] x_{iv}^2\right)^{-1}$$
$$\mu_v = \sigma_v^2 \sum\nolimits_i x_{iv} \left\{y_i - 1/2 - \mathbb{E}[\omega_i] \sum\nolimits_{\tilde{v} \neq v} x_{i\tilde{v}} \mathbb{E}[\beta_{\tilde{v}}]\right\},$$

where the expectations with respect to the q distributions can be expressed as $\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}'] = \boldsymbol{\mu}\boldsymbol{\mu}' + \text{diag}(\sigma_0^2, \ldots, \sigma_V^2)$ and $\mathbb{E}[\omega_i] = \tanh(\lambda_i/2)/(2\lambda_i)$.

B.3. Mean-Field Variational Inference with Full Covariance Matrix

We choose a fully factorized Q distribution as

$$Q = \left[\prod\nolimits_i q(\omega_i)\right] q(\boldsymbol{\beta}).$$

To exploit conjugacy, defining

$$q(\omega_i) = \text{PG}\left(1, \lambda_i\right),$$
$$q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

we have closed-form coordinate ascent variational inference update equations as

$$\lambda_i = \sqrt{\mathbb{E}[(\boldsymbol{x}_i'\boldsymbol{\beta})^2]} = \sqrt{\boldsymbol{x}_i'\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}']\boldsymbol{x}_i},$$
$$\boldsymbol{\Sigma} = (\mathbb{E}[\mathbf{A}] + \mathbf{X}'\mathbb{E}[\boldsymbol{\Omega}]\mathbf{X})^{-1}, \; \boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{X}'(\boldsymbol{y} - 1/2),$$

where the expectations with respect to the q distributions can be expressed as $\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}'] = \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}$ and $\mathbb{E}[\omega_i] = \tanh(\lambda_i/2)/(2\lambda_i)$. Note the update equations shown above are identical to those shown in Jaakkola & Jordan (2000).

B.4. SIVI Configuration

For inputs in Algorithm 1, we choose a multi-layer perceptron with layer size $[100, 200, 100]$ as $T_{\boldsymbol{\phi}}$ for $\boldsymbol{\psi} = T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon})$, $\boldsymbol{\epsilon}$ as 50 dimensional isotropic Gaussian random variable and $K = 100$, $J = 50$. For the explicit layer, we choose an MVN as $q_{\boldsymbol{\xi}}(\boldsymbol{z} \,|\, \boldsymbol{\psi}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\psi}, \boldsymbol{\xi})$. In this setting, $\boldsymbol{\psi}$ is the mean variable mixed with implicit distribution $q_{\boldsymbol{\phi}}(\boldsymbol{\psi})$ while $\boldsymbol{\xi}$ is the covariance matrix which can be either diagonal or full. In the experiments, we update the neural network parameter $\boldsymbol{\phi}$ by the Adam optimizer, with learning rate 0.01. We update $\boldsymbol{\xi}$ by gradient ascent, with step size $\eta_t = 0.001 * 0.9^{\text{iteration}/100}$. The implicit layer parameter $\boldsymbol{\phi}$ and explicit layer parameter $\boldsymbol{\xi}$ are updated iteratively.

C. Experimental Settings and Results for SIVAE

We implement SIVI with $M = 3$ stochastic hidden layers, with the dimensions of hidden layers $[\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{\ell}_3]$ as $[150, 150, 150]$ and with the dimensions of injected noises $[\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \boldsymbol{\epsilon}_3]$ as $[150, 100, 50]$. Between two adjacent stochastic layers there is a fully connected deterministic layer with size 500 and *ReLU* activation function. We choose binary pepper and salt noise (Im et al., 2017) for $q_t(\boldsymbol{\epsilon})$. The model is trained for 2000 epochs with mini-batch size 200 and step-size $0.001 * 0.75^{\text{epoch}/100}$. K_t is gradually increased from 1 to 100 during the first 1500 epochs. The explicit and implicit layers are trained iteratively. Warm-up is used during the first 300 epochs as suggested by Sønderby et al. (2016) to gradually impose the prior regularization term $\text{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \,|\, \boldsymbol{x}) || p(\boldsymbol{z}))$. The model is trained end-to-end using the Adam optimizer. After training process, as in Rezende et al. (2014) and Burda et al. (2015), we compute the marginal likelihood for test set by importance sampling with $S = 2000$:

$$\log p(\boldsymbol{x}) \approx \log \frac{1}{S} \sum_{s=1}^{S} \frac{p(\boldsymbol{x} \,|\, \boldsymbol{z}_s) p(\boldsymbol{z}_s)}{\hat{h}(\boldsymbol{z}_s \,|\, \boldsymbol{x})}, \;\; \boldsymbol{z}_s \sim h(\boldsymbol{z}_s | \boldsymbol{x}),$$

where

$$\hat{h}(\boldsymbol{z}_s | \boldsymbol{x}) = \frac{1}{M} \sum_{k=1}^{M} q(\boldsymbol{z}_s \,|\, \boldsymbol{\psi}^{(k)}), \;\; \boldsymbol{\psi}^{(k)} \overset{iid}{\sim} q_{\boldsymbol{\phi}}(\boldsymbol{\psi} | \boldsymbol{x})$$

is used to estimate $h(\boldsymbol{z}_s \,|\, \boldsymbol{x})$; we set $M = 100$. The performance of SIVI and a comparison to reported results with other methods are provided in Table 2.

Table 2. Comparison of the negative log evidence between various algorithms.

Methods	$-\log p(\boldsymbol{x})$
Results below form Burda et al. (2015)	
VAE + IWAE	$= 86.76$
IWAE + IWAE	$= 84.78$
Results below form Salimans et al. (2015)	
DLGM + HVI (1 leapfrog step)	$= 88.08$
DLGM + HVI (4 leapfrog step)	$= 86.40$
DLGM + HVI (8 leapfrog steps)	$= 85.51$
Results below form Rezende & Mohamed (2015)	
DLGM+NICE (Dinh et al., 2014) (k = 80)	≤ 87.2
DLGM+NF (k = 40)	≤ 85.7
DLGM+NF (k = 80)	≤ 85.1
Results below form Gregor et al. (2015)	
DLGM	≈ 86.60
NADE	$= 88.33$
DBM 2hl	≈ 84.62
DBN 2hl	≈ 84.55
EoNADE-5 2hl (128 orderings)	$= 84.68$
DARN 1hl	≈ 84.13
Results below form Maaløe et al. (2016)	
Auxiliary VAE (L=1, IW=1)	≤ 84.59
Results below form Mescheder et al. (2017)	
VAE + IAF (Kingma et al., 2016)	$\approx 84.9 \pm 0.3$
Auxiliary VAE (Maaløe et al., 2016)	$\approx 83.8 \pm 0.3$
AVB + AC	$\approx 83.7 \pm 0.3$
SIVI (3 stochastic layers)	$= 84.07$
SIVI (3 stochastic layers)+ IW($\tilde{K} = 10$)	$= 83.25$

D. Additional Figures

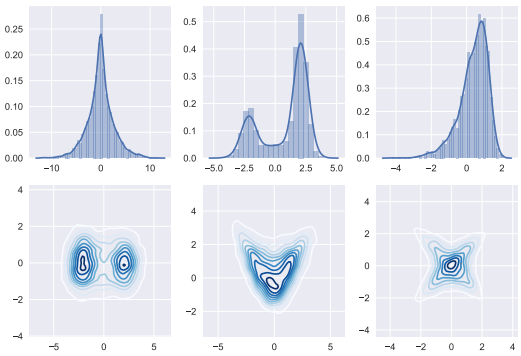

Figure 8. Visualization of the MLP based implicit distributions $\boldsymbol{\psi} \sim q(\boldsymbol{\psi})$, which are mixed with isotropic Gaussian (or Log-Normal) distributions to approximate the target distributions shown in Figure 1.

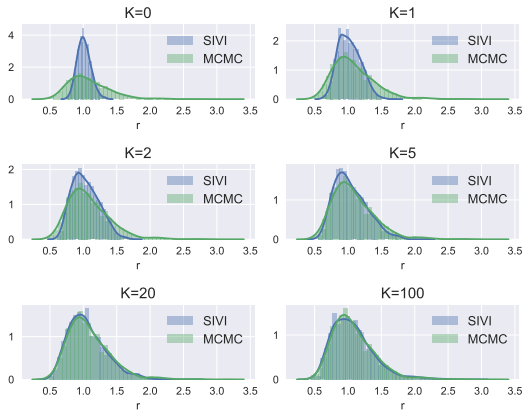

Figure 9. The marginal posterior distribution of the negative binomial dispersion parameter r inferred by SIVI becomes more accurate as K increases

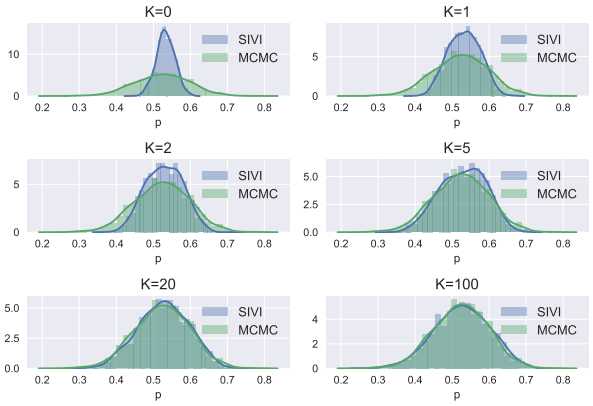

Figure 10. The marginal posterior distribution of the negative binomial probability parameter p inferred by SIVI becomes more accurate as K increases.

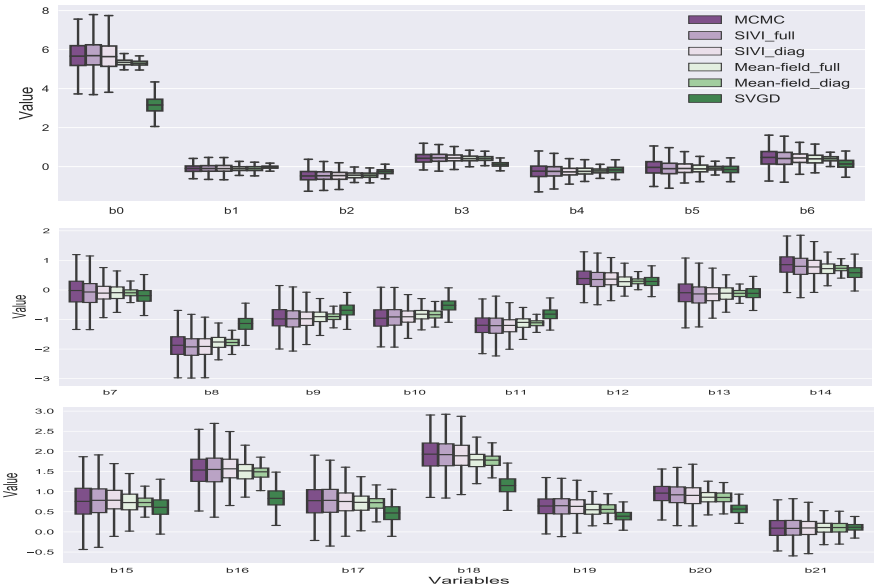

Figure 11. Comparison of all marginal posteriors of β_v inferred by various methods for Bayesian logistic regression on *waveform*.

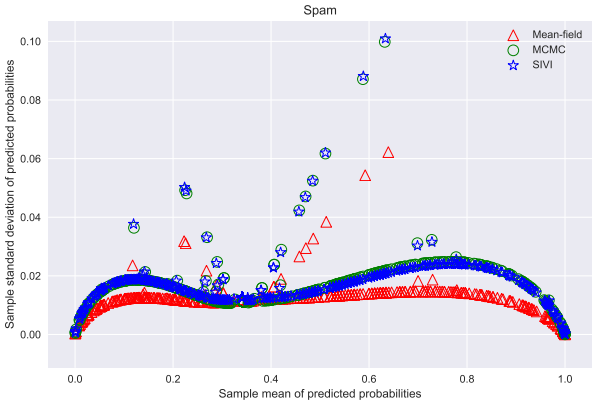

Figure 12. Sample means and standard deviations of predictive probabilities for dataset *spam*.

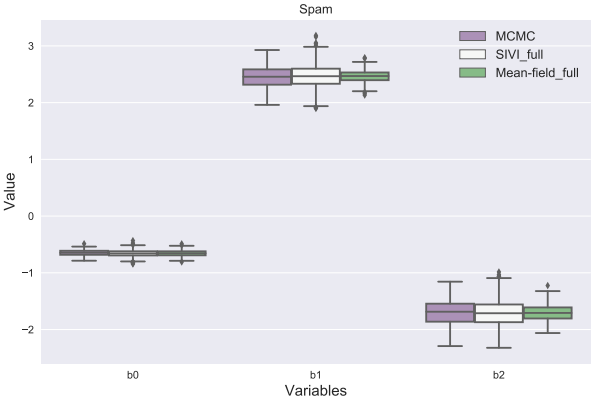

Figure 13. Boxplot of marginal posteriors inferred by MCMC, SIVI, and MFVI for dataset *spam*.

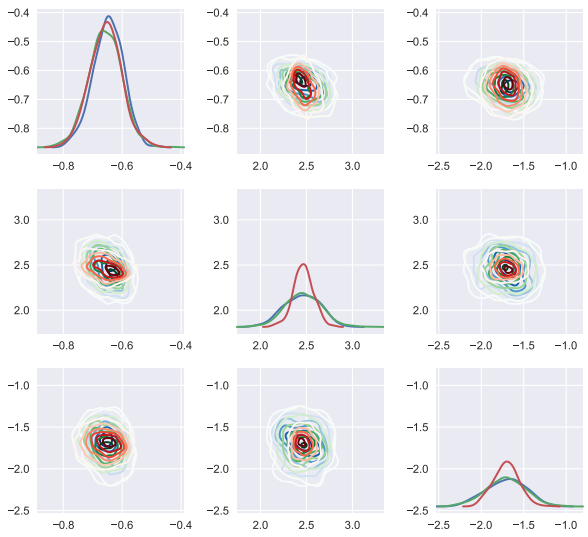

Figure 14. Univariate marginal and pairwise joint posteriors for dataset *spam*. Blue, green, and red are for MCMC, SIVI with a full covariance matrix, and MFVI with a full covariance matrix, respectively.

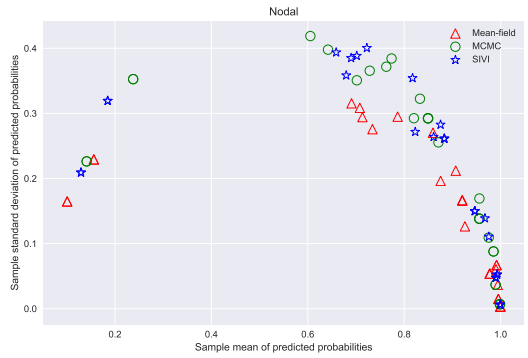

Figure 15. Sample means and standard deviations of predictive probabilities for dataset *nodal*.

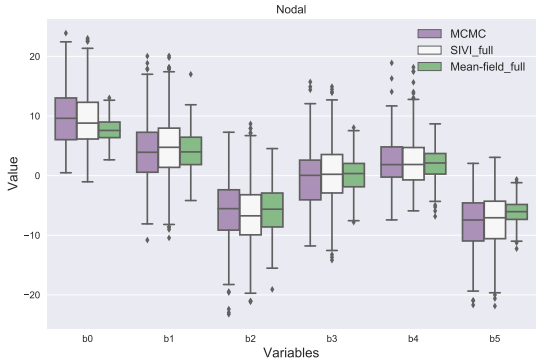

Figure 16. Boxplot of marginal posteriors inferred by MCMC, SIVI, and MFVI for dataset *nodal*.

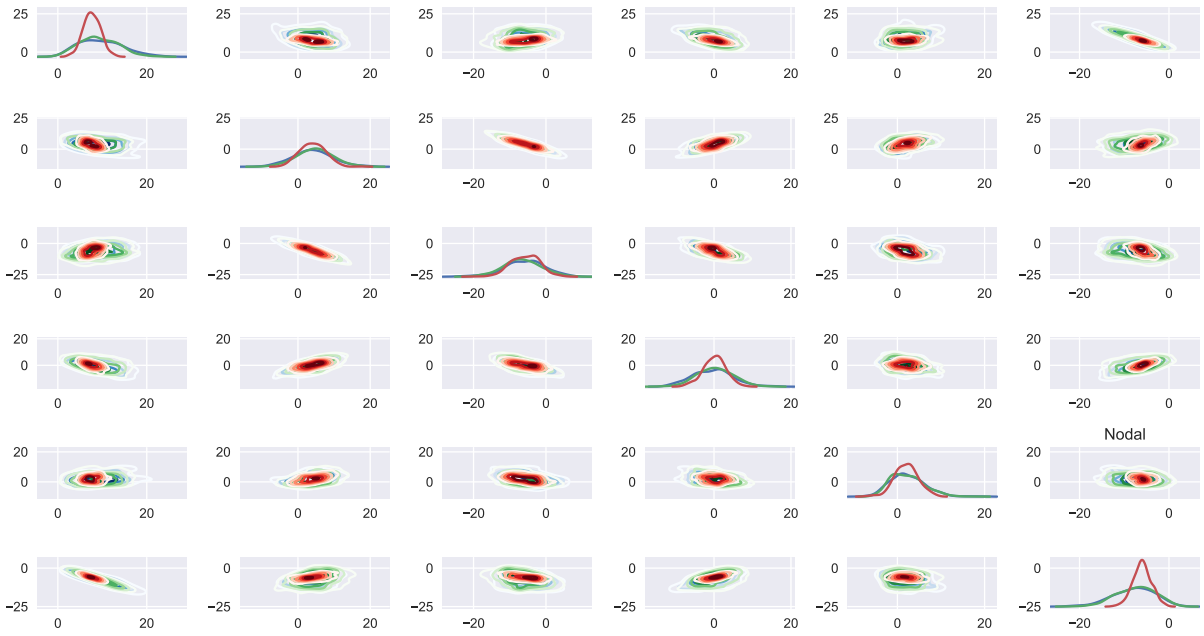

Figure 17. Univariate marginal and pairwise joint posteriors for dataset *nodal*. Blue, green, and red are for MCMC, SIVI with a full covariance matrix, and MFVI with a full covariance matrix, respectively.