Semi-Implicit Variational Inference

Mingzhang Yin[§]
Joint work with Mingyuan Zhou[⋆§]

The University of Texas at Austin
[§]Department of Statistics and Data Sciences
[⋆]IROM Department, McCombs School of Business

International Conference on Machine Learning
Stockholm, Sweden, July 11, 2018

Bayesian Inference

- Bayes' rule:

$$P(\boldsymbol{z} \mid X) = \frac{P(X \mid \boldsymbol{z})P(\boldsymbol{z})}{P(X)} = \frac{P(X \mid \boldsymbol{z})P(\boldsymbol{z})}{\int P(X \mid \boldsymbol{z})P(\boldsymbol{z})d\boldsymbol{z}}$$

$$\text{Posterior of } \boldsymbol{z} \text{ given } X = \frac{\text{Conditional Likelihood} \times \text{Prior}}{\text{Marginal Likelihood}}$$

- Two main ways for approximate Bayesian inference:
  - Draw $\boldsymbol{z} \sim P(\boldsymbol{z}|X)$ using Markov chain Monte Carlo (MCMC) based methods such as Gibbs sampling: iteratively sample $P(z_k \mid X, \boldsymbol{z} \backslash z_k)$
  - Approximate the posterior $P(\boldsymbol{z} \mid X)$ with $Q(\boldsymbol{z})$, which is straightforward to sample from, using an optimization method such as Laplace approximation and variational inference

Variational inference

- Evidence and ELBO:

$$\begin{aligned}\ln P(X) &= \int Q(\boldsymbol{z}) \ln \frac{P(X, \boldsymbol{z})}{Q(\boldsymbol{z})} d\boldsymbol{z} + \int Q(\boldsymbol{z}) \ln \frac{Q(\boldsymbol{z})}{P(\boldsymbol{z} \,|\, X)} d\boldsymbol{z} \\ &= \mathcal{L}(Q) + \mathrm{KL}(Q(\boldsymbol{z}) || P(\boldsymbol{z} \,|\, X)).\end{aligned}$$

- Since $\mathrm{KL}(Q(\boldsymbol{z})||P(\boldsymbol{z} \,|\, X)) \geq 0$, minimizing the Kullback-Leibler (KL) divergence from $P(\boldsymbol{z} \,|\, X)$ to $Q(\boldsymbol{z})$ is the same as maximizing the evidence lower bound:

$$\min_Q \mathrm{KL}(Q(\boldsymbol{z})||P(\boldsymbol{z} \,|\, X)) \Leftrightarrow \max_Q \mathrm{ELBO}$$

$$\begin{aligned}\mathrm{ELBO} = \mathcal{L}(Q) &= \mathbb{E}_Q[\ln P(X, \boldsymbol{z})] - \mathbb{E}_Q[\ln Q(\boldsymbol{z})] \\ &= \mathbb{E}_Q[\ln P(X \,|\, \boldsymbol{z})] - \mathrm{KL}(Q(\boldsymbol{z})||P(\boldsymbol{z}))\end{aligned}$$

- Variational inference converts the problem of posterior inference into an optimization problem

Mean-field variational inference

- Mean-field variational inference (VI) factorizes the Q distribution of $\boldsymbol{z} = (z_1, \ldots, z_K)^T$ as

$$Q(\boldsymbol{z}) = \prod_{i=1}^{K} q_{\phi_i}(z_i)$$

- The factorized assumption allows for closed-form coordinate ascent updates:

$$q^*(z_k) = \frac{\exp\left\{\mathbb{E}_{q(\boldsymbol{z}_{-k})}[\log p(X, z_k, \boldsymbol{z}_{-k})]\right\}}{\int \exp\left\{\mathbb{E}_{q(\boldsymbol{z}_{-k})}[\log p(X, z_k, \boldsymbol{z}_{-k})]\right\} dz_k}, \quad k = 1, \cdots, K$$

  where $\boldsymbol{z}_{-k} = \{z_1, \ldots, z_{k-1}, z_{k+1}, \ldots, z_K\}$.

- However, mean-field VI often clearly underestimates the variance of the posterior, due to the use of KL divergence and two restrictive constraints:
  - $q(z_k)$ are often restricted to the exponential family
  - The dependencies between z_k cannot be captured

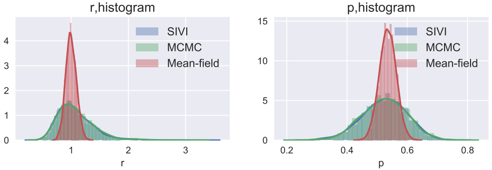
Model:

$$x_i \overset{i.i.d.}{\sim} \text{NB}(r, p), \;\; r \sim \text{Gamma}(a, 1/b), \;\; p \sim \text{Beta}(\alpha, \beta),$$

Mean-filed VI:

$$Q(r, p) = q(r)q(p) = \text{Gamma}(r; \tilde{a}, \tilde{b})\text{Beta}(p; \tilde{\alpha}, \tilde{\beta}),$$

Mean-filed VI underestimates variance (mainly due to the factorized assumption):



"Modern" variational inference

Choose a more flexible $Q_\phi(\boldsymbol{z})$ and infer the variational parameter ϕ via (stochastic) gradient descent (by reparameterization or score method)

$$\nabla_\phi \mathcal{L}(Q_\phi(\boldsymbol{z})) = \nabla_\phi \mathbb{E}_{\boldsymbol{z} \sim Q_\phi(\boldsymbol{z})} \left[\ln \frac{P(X, \boldsymbol{z})}{Q_\phi(\boldsymbol{z})} \right]$$

- There are two major flexibilities we want $Q_\phi(\boldsymbol{z})$ have:
  - We wish $Q_\phi(\boldsymbol{z})$ is not restricted to have an analytic density (but should be easy to sample)
  - We wish $Q_\phi(\boldsymbol{z})$ to incorporate dependencies of latent variables
- We also want to maintain computational tractability for a flexible inference distribution

To achieve the computation and accuracy balance, we use the neural network implicit distribution in a hierarchical model.

Implicit distribution

Implicit distribution consists of a source of randomness $q(\epsilon)$ and a deterministic transform $T_\phi : \mathbb{R}^p \to \mathbb{R}^d$

$$\boldsymbol{z} = T_\phi(\epsilon),\ \epsilon \sim q(\epsilon)$$

- When T_ϕ is invertible and the dimension is low, the density

$$q_\phi(\boldsymbol{z}) = \frac{\partial}{\partial z_1} \cdots \frac{\partial}{\partial z_d} \int_{T_\phi(\epsilon) \le \boldsymbol{z}} q(\epsilon) d\epsilon$$

  can be calculated using change of variables. But in general $\{T_\phi(\epsilon) \le \boldsymbol{z}\}$ cannot be calculated and hence the high dimension integral is intractable, making $q_\phi(\boldsymbol{z})$ become implicit
- Direct inference with implicit distribution can be difficult because of the need to estimate the density ratio $\frac{P(X,\boldsymbol{z})}{Q_\phi(\boldsymbol{z})}$

Hierarchical variational family

Capturing the latent variable dependencies plays the key role to accurately estimate the uncertainty.

- One way is to add a hierarchical structure that assumes z_k to be conditional independent but marginally dependent, using

$$q(\boldsymbol{z} \mid \boldsymbol{\psi}) = \prod_{k=1}^{K} q(z_k \mid \psi_k), \quad \boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi})$$

- Marginalizing $\boldsymbol{\psi}$ out, we can view $\boldsymbol{z}$ as a variable drawn from the distribution family $\mathcal{H}$ which we choose as variational family

$$\mathcal{H} = \left\{ h_\phi(\boldsymbol{z}) : h_\phi(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi})}[q(\boldsymbol{z} \mid \boldsymbol{\psi})] = \int_{\boldsymbol{\psi}} \left[\prod_{k=1}^{K} q(z_k \mid \psi_k) \right] q_\phi(\boldsymbol{\psi}) d\boldsymbol{\psi} \right\}$$

- It is evident that $q(\boldsymbol{z} \mid \boldsymbol{\psi}) \in \mathcal{Q} \subseteq \mathcal{H}$, i.e., $\mathcal{H}$ is an expansion of the original variational distribution family

Semi-implicit variational inference (SIVI)

- We call the hierarchical model semi-implicit because it requires $q(\boldsymbol{z} \mid \psi)$ to be explicit while allows $q_\phi(\psi)$ to be implicit, and $h_\phi(\boldsymbol{z}) = \mathbb{E}_{q_\phi(\psi)} q(\boldsymbol{z} \mid \psi)$ and ELBO is generally not analytic
- KL convexity and Jensen's inequality lead to an ELBO lower bound:

$$
\begin{aligned}
&\underline{\mathcal{L}}(q(\boldsymbol{z} \mid \psi), q_\phi(\psi)) = \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z} \mid \psi)} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \psi)} \\
&= -\mathbb{E}_{\psi \sim q_\phi(\psi)} \mathrm{KL}(q(\boldsymbol{z} \mid \psi) || p(\boldsymbol{z}|\boldsymbol{x})) + \log p(\boldsymbol{x}) \\
&\leq -\mathrm{KL}(\mathbb{E}_{\psi \sim q_\phi(\psi)} q(\boldsymbol{z} \mid \psi) || p(\boldsymbol{z}|\boldsymbol{x})) + \log p(\boldsymbol{x}) \\
&= \mathcal{L} = \mathbb{E}_{\boldsymbol{z} \sim h_\phi(\boldsymbol{z})} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{h_\phi(\boldsymbol{z})}
\end{aligned}
$$

- Using the concavity of the logarithmic function, we have $\log h_\phi(\boldsymbol{z}) \geq \mathbb{E}_{\psi \sim q_\phi(\psi)} \log q(\boldsymbol{z} \mid \psi)$ and hence an ELBO upper bound:

$$
\bar{\mathcal{L}}(q(\boldsymbol{z} \mid \psi), q_\phi(\psi)) = \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{\boldsymbol{z} \sim h_\phi(\boldsymbol{z})} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \psi)} \geq \mathcal{L}
$$

- Note there is a subtle but critical difference between $\underline{\mathcal{L}}$ and $\bar{\mathcal{L}}$

Degeneracy of $\underline{\mathcal{L}}$

Maximizing the surrogate lower bound $\underline{\mathcal{L}}$ may lead to degeneracy that $q_\phi(\psi)$ converges to a point mass density:

Proposition (Degeneracy)

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

$$\underline{\mathcal{L}}(q(\boldsymbol{z}\,|\,\psi), q_\phi(\psi)) \leq -\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}\,|\,\psi^*)} \log \frac{q(\boldsymbol{z}|\psi^*)}{p(\boldsymbol{x},\boldsymbol{z})},$$

where the equality is true if and only if $q_\phi(\psi) = \delta_{\psi^*}(\psi)$.

Asymptotically exact ELBO

- Avoid degeneracy by adding regularization $\underline{\mathcal{L}}_K = \underline{\mathcal{L}} + B_K$

$$B_K = \mathbb{E}_{\boldsymbol{\psi},\boldsymbol{\psi}^{(1)},\ldots,\boldsymbol{\psi}^{(K)}\sim q_\phi(\boldsymbol{\psi})}\mathrm{KL}(q(\boldsymbol{z}\,|\,\boldsymbol{\psi})||\tilde{h}_K(\boldsymbol{z})), \tag{1}$$

where $\tilde{h}_K(\boldsymbol{z}) = \frac{1}{K+1}[q(\boldsymbol{z}\,|\,\boldsymbol{\psi}) + \sum_{k=1}^K q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(k)})]$, $B_K \geq 0$, with $B_K = 0$ if and only if $K = 0$ or $q_\phi(\boldsymbol{\psi})$ degenerates to a point mass density

- The Jensen gap can also be narrowed from upper side by $\bar{\mathcal{L}}_k = \bar{\mathcal{L}} - A_k$

$$A_K = \mathbb{E}_{\boldsymbol{\psi}\sim q_\phi(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{z}\sim h_\phi(\boldsymbol{z})}\mathbb{E}_{\boldsymbol{\psi}^{(1)},\ldots,\boldsymbol{\psi}^{(K)}\sim q_\phi(\boldsymbol{\psi})}\Big[\log\big(\frac{1}{K}\sum_{k=1}^K q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(k)})\big) - \log q(\boldsymbol{z}\,|\,\boldsymbol{\psi})\Big]$$

The regularized lower bound $\underline{\mathcal{L}}_K$ is an asymptotically exact ELBO that satisfies $\underline{\mathcal{L}}_0 = \underline{\mathcal{L}}$ and $\lim_{K\to\infty}\underline{\mathcal{L}}_K = \mathcal{L}$. The regularized upper bound satisfies $\bar{\mathcal{L}}_1 = \bar{\mathcal{L}}$, $\bar{\mathcal{L}}_{K+1} \leq \bar{\mathcal{L}}_K$, and $\lim_{K\to\infty}\bar{\mathcal{L}}_K = \mathcal{L}$.

Algorithm for SIVI

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 $k = 1, \dots, K_t$; 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

  $t = t + 1$

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

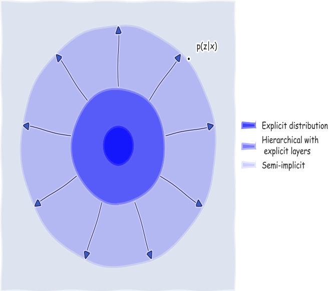
  $\boldsymbol{\phi} = \boldsymbol{\phi} + \eta_t \nabla_{\boldsymbol{\phi}} \underline{L}_{K_t}(\{\boldsymbol{\psi}^{(k)}\}_{1,K_t}, \{\boldsymbol{\psi}_j\}_{1,J}, \{\boldsymbol{z}_j\}_{1,J})$

end

Methods to expand variational distribution family

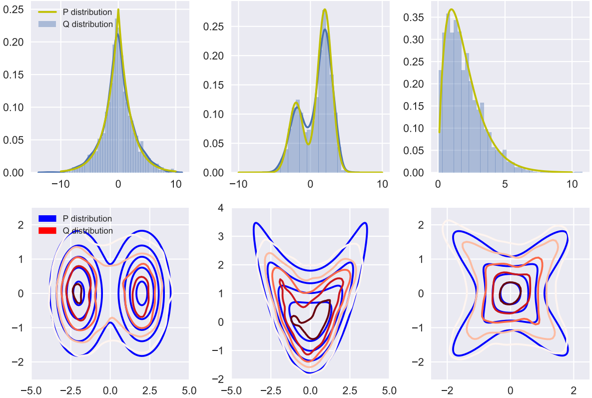
Expand variational family via stochastic and/or deterministic method

- Hierarchical models: eg. Negative Binomial⇔Poisson-Gamma hierarchy; Hierarchical variational model (Ranganath et al., 2016)
- Normalizing Flow: transfer simple distribution with a chain of simple invertible mapping $\boldsymbol{z}_t = f_t \circ \cdots \circ f_0(\boldsymbol{z}_0)$ (Rezende and Mohamed, 2015)
- Modeling the dependencies between univariate marginals with copula (Tran et al., 2015)
- Implicit distribution $\boldsymbol{z} = f(\epsilon)$, where f is not invertible; (Tran et al., 2017)
- Our approach: hierarchy with explicit conditional layer, implicit mixing layers (semi-implicit)



Expressiveness of SIVI

$h(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{\psi}\sim q(\boldsymbol{\psi})} q(\boldsymbol{z} \mid \boldsymbol{\psi})$	$p(\boldsymbol{z})$
$z \sim \mathcal{N}(\psi, 0.1)$, $\psi \sim q(\psi)$	$\text{Laplace}(z; \mu = 0, b = 2)$
	$0.3\mathcal{N}(z; -2, 1) + 0.7\mathcal{N}(z; 2, 1)$
$z \sim \text{Log-Normal}(\psi, 0.1)$, $\psi \sim q(\psi)$	$\text{Gamma}(z; 2, 1)$
$\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{\psi}, \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}\right)$, $\boldsymbol{\psi} \sim q(\boldsymbol{\psi})$	$0.5\mathcal{N}(\boldsymbol{z}; -2, I) + 0.5\mathcal{N}(\boldsymbol{z}; 2, I)$
	$\mathcal{N}(z_1; z_2^2/4, 1)\mathcal{N}(z_2; 0, 4)$
	$0.5\mathcal{N}\left(\boldsymbol{z}; 0, \begin{bmatrix} 2 & 1.8 \\ 1.8 & 2 \end{bmatrix}\right) + 0.5\mathcal{N}\left(\boldsymbol{z}; 0, \begin{bmatrix} 2 & -1.8 \\ -1.8 & 2 \end{bmatrix}\right)$



Model:

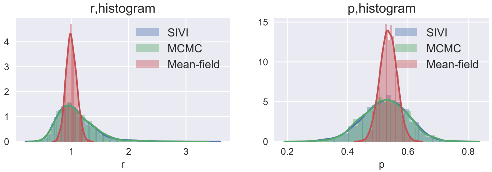
$$x_i \overset{i.i.d.}{\sim} \mathrm{NB}(r,p),\ \ r \sim \mathrm{Gamma}(a, 1/b),\ \ p \sim \mathrm{Beta}(\alpha,\beta),$$

Mean-filed VI:

$$Q(r,p) = q(r)q(p) = \mathrm{Gamma}(r;\tilde{a},\tilde{b})\mathrm{Beta}(p;\tilde{\alpha},\tilde{\beta}),$$

SIVI (both the conditional and mixing q distributions are reparameterizable) :

$$q(r,p\,|\,\psi) = \text{Log-Normal}(r;\mu_r,\sigma_0^2)\text{Logit-Normal}(p;\mu_p,\sigma_0^2),$$
$$\psi = (\mu_r,\mu_p) \sim q(\psi),$$



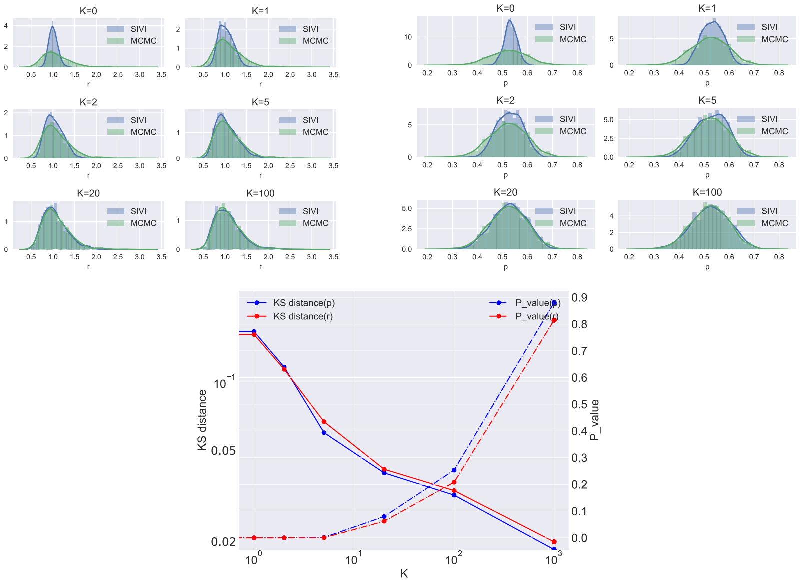


Figure: Kolmogorov-Smirnov (KS) distance and its corresponding p-value between the marginal posteriors of r and p inferred by SIVI and MCMC. SIVI rapidly improves as K increases.

Score function gradient for conjugate model

If $q(\boldsymbol{z} \mid \psi)$ is not reparameterizable, then we introduce a density ratio as

$$r_{\boldsymbol{\xi},\phi}(\boldsymbol{z},\epsilon,\epsilon^{(1:K)}) = \frac{q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_\phi(\epsilon)))}{\frac{1}{K+1}[q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_\phi(\epsilon)) + \sum_{k=1}^K q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_\phi(\epsilon^{(k)}))]}$$

and approximate the gradient of $\underline{\mathcal{L}}_K$ with respect to ϕ as

$$\nabla_\phi \underline{\mathcal{L}}_K \approx \frac{1}{J}\sum_{j=1}^J \Big\{ -\nabla_\phi \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_\phi(\epsilon_j))}[\log \frac{q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_\phi(\epsilon_j))}{p(\boldsymbol{x},\boldsymbol{z})}]$$
$$+ \nabla_\phi \log r_{\boldsymbol{\xi},\phi}(\boldsymbol{z}_j,\epsilon_j,\epsilon^{(1:K)})$$
$$+ [\nabla_\phi \log q_{\boldsymbol{\xi}}(\boldsymbol{z}_j \mid T_\phi(\epsilon_j))] \log r_{\boldsymbol{\xi},\phi}(\boldsymbol{z}_j,\epsilon_j,\epsilon^{(1:K)}) \Big\},$$

- The first summation term is equivalent to the gradient of MFVI's ELBO
- Both the second and third terms correct the restrictions of $q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_\phi(\epsilon_j))$
- $\log r_{\boldsymbol{\xi},\phi}(\boldsymbol{z},\epsilon,\epsilon^{(1:K)})$ in the third term is expected to be small regardless of convergence, effectively mitigating the variance of score function gradient estimation that is usually high in basic black-box VI

Model:

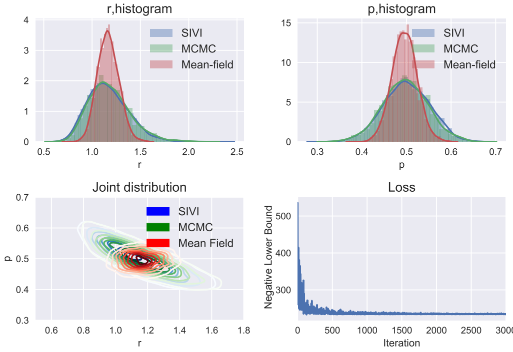
$$p(n_i, l_i \mid r, p) = r^{l_i} p^{n_i} (1-p)^r / Z_i, \ \ r \sim \text{Gamma}(a, 1/b), \ \ p \sim \text{Beta}(\alpha, \beta)$$

Mean-filed VI:

$$Q(r, p) = q(r)q(p) = \text{Gamma}(r; \tilde{a}, \tilde{b})\text{Beta}(p; \tilde{\alpha}, \tilde{\beta}),$$

SIVI (non-reparameterizable conditional q distribution but conjugate model):

$$q(r, p \mid \boldsymbol{\psi}) = \text{Gamma}(r; \psi_1, \psi_2)\text{Beta}(p; \psi_3, \psi_4), \ \ \boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3, \psi_4) \sim q(\boldsymbol{\psi})$$

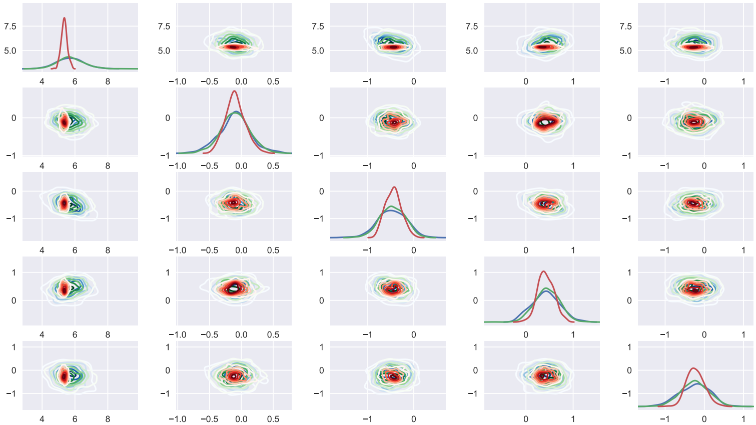


Bayesian logistic regression (pairwise joint distributions)

$$y_i \sim \text{Bernoulli}[(1+e^{-\mathbf{x}_i'\boldsymbol{\beta}})^{-1}], \quad \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}_{V+1})$$

SIVI: $q(\boldsymbol{\beta} \mid \boldsymbol{\psi}) = \mathcal{N}(\boldsymbol{\psi}, \boldsymbol{\Sigma}), \ \boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})$

(Blue: MCMC, Red: VI, Green: SIVI):



Bayesian logistic regression

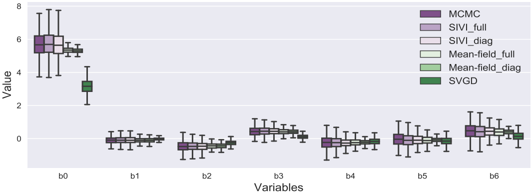


Figure: Comparing univariate marginals

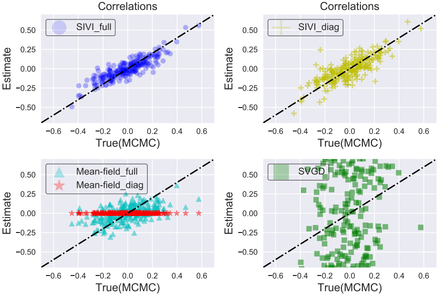


Figure: Comparing posterior covariance matrix

Bayesian logistic regression (predictive uncertainty)

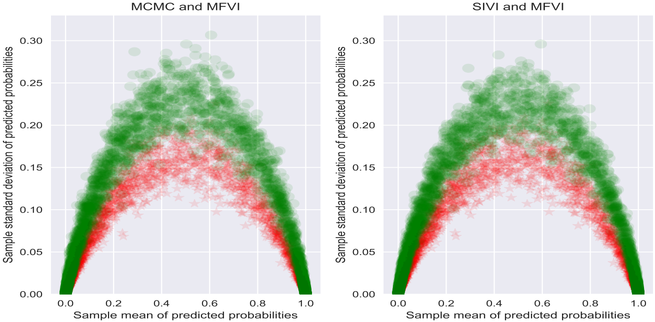


Figure: Comparison of MFVI (red) with a full covariance matrix, MCMC (green on left), and SIVI (green on right) with a full covariance matrix on quantifying predictive uncertainty for Bayesian logistic regression on *waveform*

Semi-implicit variational autoencoder

We construct semi-implicit VAE (SIVAE) by using a hierarchical encoder that injects random noise at M different stochastic layers as

$$\begin{aligned}
&\boldsymbol{\ell}_t = T_t(\boldsymbol{\ell}_{t-1}, \boldsymbol{\epsilon}_t, \boldsymbol{x}; \phi),\ \ \boldsymbol{\epsilon}_t \sim q_t(\boldsymbol{\epsilon}),\ \ t = 1, \ldots, M, \\
&\boldsymbol{\mu}(\boldsymbol{x}, \phi) = f(\boldsymbol{\ell}_M, \boldsymbol{x}; \phi), \quad \boldsymbol{\Sigma}(\boldsymbol{x}, \phi) = g(\boldsymbol{\ell}_M, \boldsymbol{x}; \phi), \\
&q_\phi(\boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}, \phi), \boldsymbol{\Sigma}(\boldsymbol{x}, \phi)),
\end{aligned}$$

where $\boldsymbol{\ell}_0 = \emptyset$ and T_t, f, and g are all deterministic neural networks. Note given data $\boldsymbol{x}_i$, $\boldsymbol{\mu}(\boldsymbol{x}_i, \phi)$, $\boldsymbol{\Sigma}(\boldsymbol{x}_i, \phi)$ are now random variables rather than following vanilla VAE to assume deterministic values. This clearly moves the encoder variational distribution beyond a simple Gaussian form.

Semi-implicit variational autoencoder

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($\tilde{K} = 10$)	$= 83.25$

Summary

- Uncertainty estimation is difficult but important in Variational Inference
- One key to get an accurate uncertainty estimation is to construct a flexible variational distribution that can capture the dependencies between latent variables
- Balancing the expressiveness and tractability, semi-implicit variational inference (SIVI) can approach the accuracy of MCMC in quantifying posterior uncertainty, but often pays a lower computational cost and can generate independent posterior samples fast via the inferred stochastic variational inference network.

Thank you!

Welcome to our poster at Hall B # 177

[1] Reproducible code is at https://github.com/mingzhang-yin/SIVI

Related inference methods

- VAE: Changing the empirical data distribution leads to degenerated $\mathcal{L}$

$$\mathcal{L}_{VAE} = \mathbb{E}_{\boldsymbol{x}\sim D(\boldsymbol{x})}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|f(\boldsymbol{x}))} \log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|f(\boldsymbol{x}))}$$

$$\underline{\mathcal{L}}_{SIVI} = \mathbb{E}_{\epsilon\sim q(\epsilon)}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|f(\epsilon))} \log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|f(\epsilon))}$$

- Data augmentation: iteratively sample from $p(\boldsymbol{z}|\psi)$ and $p(\psi|\boldsymbol{z})$ with

$$p(\boldsymbol{z}) = \int p(\boldsymbol{z},\psi)d\psi$$

- Auxiliary Deep Generative Models(Maaløe et al., 2016): optimize on a less tighter bound

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{h_\phi(\boldsymbol{z}|\boldsymbol{x})} \log \frac{p(\boldsymbol{x},\boldsymbol{z})}{h_\phi(\boldsymbol{z}|\boldsymbol{x})} \geq \mathbb{E}_{q_\phi(\boldsymbol{z},a|\boldsymbol{x})} \log \frac{p(\boldsymbol{x},\boldsymbol{z},a)}{q_\phi(\boldsymbol{z},a|\boldsymbol{x})}$$