Semi-Implicit Variational Inference

Mingzhang Yin and Mingyuan Zhou

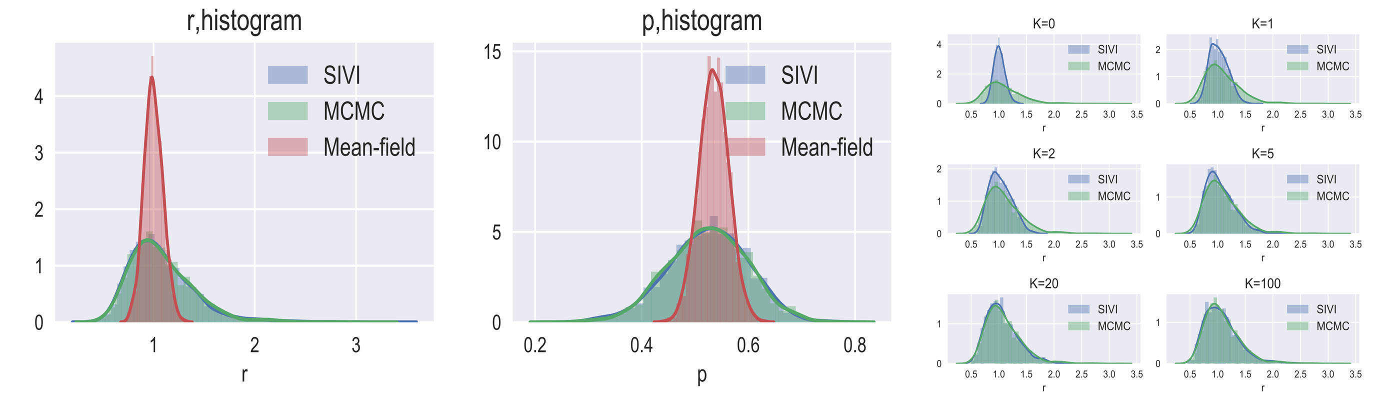
Variational inference

Find $Q(\boldsymbol{z}) \in \mathcal{Q}$ to maximize evidence lower bound (ELBO)

$$\text{ELBO} = \mathcal{L}(Q) = \mathbf{E}_Q[\ln P(\boldsymbol{x}, \boldsymbol{z})] - \mathbf{E}_Q[\ln Q(\boldsymbol{z})]$$
$$= \ln p(\boldsymbol{x}) - \text{KL}(Q(\boldsymbol{z}) || P(\boldsymbol{z}|\boldsymbol{x}))$$

- Optimizing $Q(\boldsymbol{z})$ is considered as approximating posterior inference;
- Mean-field VI makes a fully factorized assumption as $Q(\boldsymbol{z}) = \prod_{i=1}^{K} q_{\phi_i}(z_i)$;
- Capturing latent dependencies is crucial for correct uncertainty estimation.

Negative binomial $x_i \overset{iid}{\sim} \text{NB}(r, p),\ r \sim \text{Gamma}(a, 1/b),\ p \sim \text{Beta}(\alpha, \beta)$



Semi-implicit model

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

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

For an implicit distribution, it is often easy to generate random samples from it but intractable to calculate its probability density function, making variational inference difficult

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

Semi-implicit model is a two-stage model

$$\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{\psi}), \quad \boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi})$$

- The first layer distribution $q(\boldsymbol{z}|\boldsymbol{\psi})$ is explicit, while the mixing distribution $q_\phi(\boldsymbol{\psi})$ is allowed to be implicit;
- The marginal distribution $h_\phi(\boldsymbol{z})$ is used as variational distribution

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

- The components of $\boldsymbol{z}$ are conditionally independent but marginally dependent;
- It is evident that $q(\boldsymbol{z}|\boldsymbol{\psi}) \in \mathcal{Q} \subseteq \mathcal{H}$, i.e., $\mathcal{H}$ forms an expansion;
- Semi-implicit distribution $h_\phi(\boldsymbol{z})$ achieves a balance between expressiveness and tractability.

Lower and upper bound of ELBO

Optimize ELBO $= \mathbf{E}_{h_\phi(\boldsymbol{z})}[\ln p(\boldsymbol{x}, \boldsymbol{z}) - \ln h_\phi(\boldsymbol{z})]$ for SIVI is generally intractable if $h_\phi(\boldsymbol{z}) = \mathbb{E}_{q_\phi(\boldsymbol{\psi})} q(\boldsymbol{z} \mid \boldsymbol{\psi})$ is not analytic

- KL convexity and Jensen's inequality lead to an ELBO lower bound:

$$\underline{\mathcal{L}}(q(\boldsymbol{z} \mid \boldsymbol{\psi}), q_\phi(\boldsymbol{\psi})) = \mathbb{E}_{\boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi})} \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z} \mid \boldsymbol{\psi})} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{\psi})}$$
$$= -\mathbb{E}_{\boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi})} \text{KL}(q(\boldsymbol{z} \mid \boldsymbol{\psi}) || p(\boldsymbol{z}|\boldsymbol{x})) + \log p(\boldsymbol{x})$$
$$\le -\text{KL}(\mathbb{E}_{\boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi})} q(\boldsymbol{z} \mid \boldsymbol{\psi}) || p(\boldsymbol{z}|\boldsymbol{x})) + \log p(\boldsymbol{x}) = \mathcal{L} = \mathbb{E}_{\boldsymbol{z} \sim h_\phi(\boldsymbol{z})} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{h_\phi(\boldsymbol{z})}$$

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

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

- Note there is a subtle but critical difference between $\underline{\mathcal{L}}$ and $\bar{\mathcal{L}}$
- Direct optimizing on $\underline{\mathcal{L}}$ will result in degeneracy; namely, $q_\phi(\boldsymbol{\psi}) \to \delta_{\boldsymbol{\psi}^*}$

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

with $\boldsymbol{\psi}^* = \text{argmax}_{\boldsymbol{\psi}} -KL(q(\boldsymbol{z} \mid \boldsymbol{\psi}) || p(x, \boldsymbol{z}))$

Asymptotically exact surrogate ELBOs

Add regularization as $\underline{\mathcal{L}}_K = \underline{\mathcal{L}} + B_K$

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

The regularized surrogate ELBO can be expressed as

$$\underline{\mathcal{L}}_K = \mathbb{E}_{\boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi})} \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z} \mid \boldsymbol{\psi})} \mathbb{E}_{\boldsymbol{\psi}^{(1)}, \dots, \boldsymbol{\psi}^{(K)} \sim q_\phi(\boldsymbol{\psi})} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{\frac{1}{K+1}[q(\boldsymbol{z} \mid \boldsymbol{\psi}) + \sum_{k=1}^{K} q(\boldsymbol{z} \mid \boldsymbol{\psi}^{(k)})]}$$

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

$$\bar{\mathcal{L}}_K = \mathbb{E}_{\boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi})} \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z} \mid \boldsymbol{\psi})} \mathbb{E}_{\boldsymbol{\psi}^{(1)}, \dots, \boldsymbol{\psi}^{(K)} \sim q_\phi(\boldsymbol{\psi})} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{\frac{1}{K} \sum_{k=1}^{K} q(\boldsymbol{z} \mid \boldsymbol{\psi}^{(k)})}$$

Property: Surrogate ELBOs

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 $\bar{\mathcal{L}}_1 = \bar{\mathcal{L}}$, $\bar{\mathcal{L}}_{K+1} \le \bar{\mathcal{L}}_K$, and $\lim_{K \to \infty} \bar{\mathcal{L}}_K = \mathcal{L}$.

For non-reparameterizable but conjugate model, the gradient can be expressed as

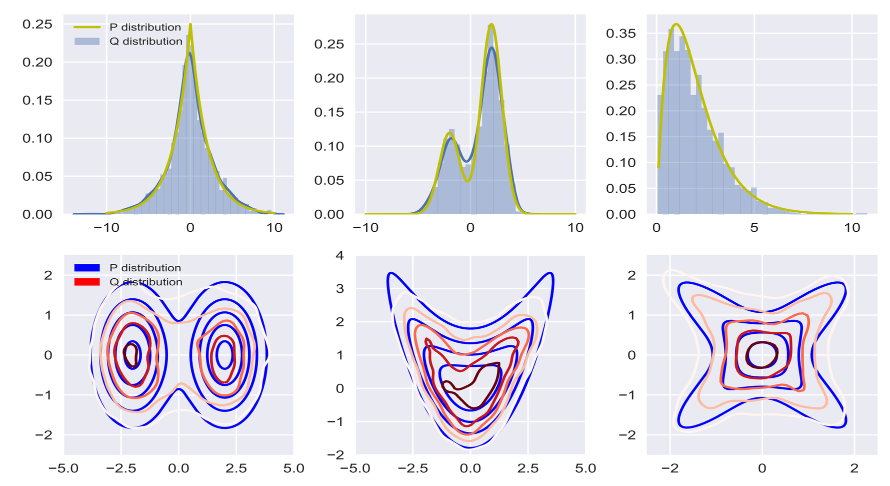
$$\nabla_\phi \underline{\mathcal{L}}_K \approx \tfrac{1}{J} \textstyle\sum_{j=1}^{J} \big\{ -\nabla_\phi \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_\phi(\boldsymbol{\epsilon}_j))} [\log \frac{q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_\phi(\boldsymbol{\epsilon}_j))}{p(\boldsymbol{x},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

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

Experiments

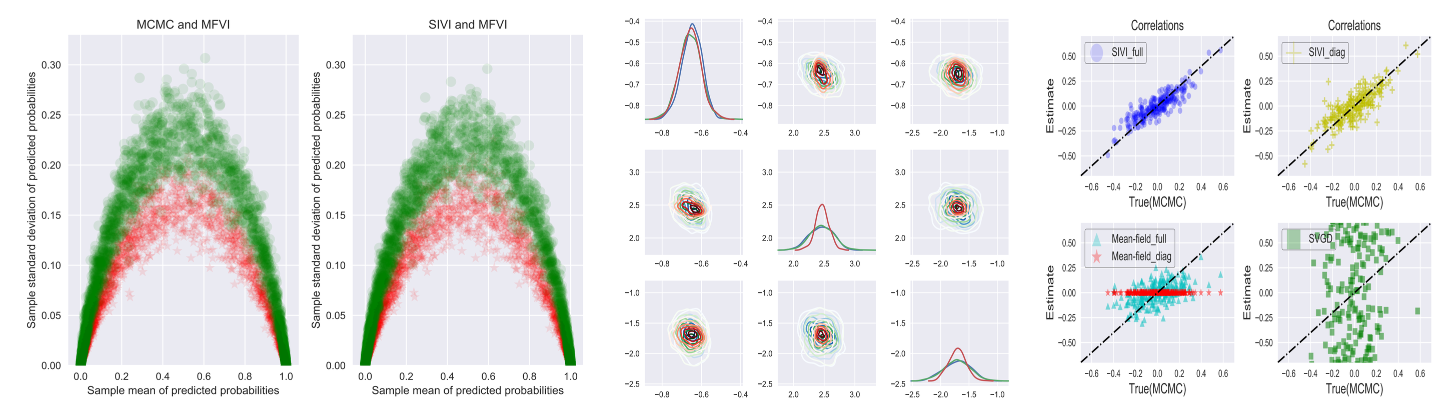
- **Toy examples (capturing skewness, kurtosis, and multimodality)**

$$h(\boldsymbol{z}) = \mathbf{E}_{\boldsymbol{\psi} \sim q(\boldsymbol{\psi})} q(\boldsymbol{z} \mid \boldsymbol{\psi}), \quad q(\boldsymbol{z} \mid \boldsymbol{\psi}) = \text{(log)Normal}(\boldsymbol{\psi}, 0.1)$$



- **Bayesian Logistic regression**

$$y_i \sim \text{Bernoulli}[(1 + e^{-\boldsymbol{x}_i' \boldsymbol{\beta}})^{-1}],\ \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1} \mathbf{I}_{V+1})$$
$$q(\boldsymbol{\beta} \mid \boldsymbol{\psi}) = \mathcal{N}(\boldsymbol{\psi}, \boldsymbol{\Sigma}),\ \boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi})$$



- **Variational autoencoder(VAE)**
  Inject random noise at M different stochastic layers. Let $h(\boldsymbol{z}|\boldsymbol{x}) = \int q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\epsilon}) q(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}$

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

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

Full version at `https://arxiv.org/abs/1805.11183`