Contents

1 Unbiased Implicit Variational Inference 2

1.1 Quality of approximation 3

1.2 Gradient variance 4

1.3 Scratch notes 7

2 Semi-implicit variational inference 8

3 Hierarchical variational inference 8

4 Theoretical guarantees for implicit VI 9

5 Other references 10

1 Unbiased Implicit Variational Inference

Based on Titsias and Ruiz [7].

- Authors introduce unbiased implicit variational inference (UIVI) that defines a flexible variational family. Like semi-implicit variational inference (SIVI), UIVI uses an implicit variational distribution $q_\theta(z) = \int q_\theta(z|\varepsilon)q(\varepsilon)d\varepsilon$ where $q_\theta(z|\varepsilon)$ is a reparameterizable distribution whose parameters can be outputs of some neural network g, i.e., $q_\theta(z|\varepsilon) = h(u; g(\varepsilon;\theta))$ with $u \sim q(u)$. Under two assumptions on the conditional $q_\theta(z|\varepsilon)$, the ELBO can be approximated via Monte Carlo sampling. In particular, the entropy component of the ELBO can be rewritten as an expectation w.r.t. the reverse conditional $q_\theta(\varepsilon|z)$. Efficient approximation of this expectation w.r.t. the reverse conditional is done by reusing samples from approximating the main expectation to initialize a MCMC sampler.
- In SIVI, the variational distribution $q_\theta(z)$ is defined as

$$q_\theta(z) = \int q_\theta(z|\varepsilon)q(\varepsilon)d\varepsilon$$

  where $\varepsilon \sim q(\varepsilon)$.
- UIVI:
  - Like SIVI, UIVI uses an implicit variational distribution $q_\theta(z)$ whose density cannot be evaluated but from which samples can be drawn. Unlike SIVI, UIVI directly maximizes the ELBO rather than a lower bound.
  - The dependence of $q_\theta(z|\varepsilon)$ on ε can be arbitrarily complex. Titsias and Ruiz [7] take the parameters of a reparameterizable distribution (Assumption 1) as the output of a neural network with parameters θ that takes ε as input, i.e.,

$$z = h(u; g_\theta(\varepsilon)) = h_\theta(u;\varepsilon)$$

    where $u \sim q(u)$ and g_θ is some neural network. It is also assumed that $\nabla_z \log q_\theta(z|\varepsilon)$ can be evaluated (Assumption 2).
  - The gradient of the ELBO is given by

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \mathbb{E}_{q_\theta(z)}\left[\log p(x,z) - \log q_\theta(z)\right] \\
&= \nabla_\theta \int \left(\log p(x,z) - \log q_\theta(z)\right) q_\theta(z) dz \\
&= \int \nabla_\theta \left(\left(\log p(x,z) - \log q_\theta(z)\right) q_\theta(z)\right) dz \\
&= \int \nabla_\theta \left(\left(\log p(x,z) - \log q_\theta(z)\right) \int q_\theta(z|\varepsilon)q(\varepsilon)d\varepsilon\right) dz \\
&= \int\int \nabla_\theta \left(\left(\log p(x,z) - \log q_\theta(z)\right)\big|_{z=h_\theta(u;\varepsilon)}\right) q(u)q(\varepsilon)d\varepsilon du \\
&= \mathbb{E}_{q(\varepsilon)q(u)}\left[\nabla_z \log p(x,z)\big|_{z=h_\theta(u;\varepsilon)} \nabla_\theta h_\theta(u;\varepsilon)\right] - \mathbb{E}_{q(\varepsilon)q(u)}\left[\nabla_z \log q_\theta(z)\big|_{z=h_\theta(u;\varepsilon)} \nabla_\theta h_\theta(u;\varepsilon)\right] .
\end{aligned}$$

    (Note that is $\mathbb{E}_{q_\theta(z)}\left[\nabla_\theta \log q_\theta(z)\right] = 0$ is applied as below; see Slide 24) (Gradient can be pushed into expectation using DCT.)

$$\begin{aligned}
\nabla_\theta \mathbb{E}_{q_\theta(z)}\left[\log q_\theta(z)\right] &= \nabla_\theta \mathbb{E}_{q(\varepsilon)}\left[\log q_\theta(f_\theta(\varepsilon))\right] \\
&= \mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta \log q_\theta(z)\big|_{z=f_\theta(\varepsilon)}\right] + \mathbb{E}_{q(\varepsilon)}\left[\nabla_z \log q_\theta(z)\big|_{z=f_\theta(\varepsilon)} \nabla_\theta f_\theta(\varepsilon)\right] \\
&= \mathbb{E}_{q_\theta(z)}\left[\nabla_\theta \log q_\theta(z)\right] + \mathbb{E}_{q(\varepsilon)}\left[\nabla_z \log q_\theta(z)\big|_{z=f_\theta(\varepsilon)} \nabla_\theta f_\theta(\varepsilon)\right] \\
&= \mathbb{E}_{q(\varepsilon)}\left[\nabla_z \log q_\theta(z)\big|_{z=f_\theta(\varepsilon)} \nabla_\theta f_\theta(\varepsilon)\right]
\end{aligned}$$

As $\nabla_z \log q_\theta(z)$ cannot be evaluated, this gradient is rewritten as an expectation using the log-deritative identity: $\nabla_x \log f(x) = \frac{1}{f(x)} \nabla_x f(x)$:

$$\begin{aligned}
\nabla_z \log q_\theta(z) &= \frac{1}{q_\theta(z)} \nabla_z q_\theta(z) \\
&= \frac{1}{q_\theta(z)} \nabla_z \int q_\theta(z|\varepsilon) q(\varepsilon) d\varepsilon \\
&= \frac{1}{q_\theta(z)} \int \nabla_z q_\theta(z|\varepsilon) q(\varepsilon) d\varepsilon \\
&= \frac{1}{q_\theta(z)} \int q_\theta(z|\varepsilon) q(\varepsilon) \nabla_z \log q_\theta(z|\varepsilon) d\varepsilon \\
&= \int q_\theta(\varepsilon|z) \nabla_z \log q_\theta(z|\varepsilon) d\varepsilon \\
&= \mathbb{E}_{q_\theta(\varepsilon|z)} \left[\nabla_z \log q_\theta(z|\varepsilon) \right] .
\end{aligned}$$

$\nabla_z \log q_\theta(z|\varepsilon)$ can be evaluated by assumption.

- UIVI estimates the gradient of the ELBO by drawing S samples from $q(\varepsilon)$ and $q(u)$ (in practice, $S = 1$):

$$\nabla_\theta \mathcal{L}(\theta) \approx \frac{1}{S} \sum_{s=1}^{S} \Big(\nabla_z \log p(x, z) \big|_{z = h_\theta(u_s, \varepsilon_s)} \nabla_\theta h_\theta(u_s; \varepsilon_s) - \mathbb{E}_{q_\theta(\varepsilon|z)} \left[\nabla_z \log q_\theta(z|\varepsilon) \right] \big|_{z = h_\theta(u_s; \varepsilon_s)} \nabla_\theta h_\theta(u_s; \varepsilon_s) \Big) .$$

  To estimate the inner expectation, samples are drawn from the reverse conditional $q_\theta(\varepsilon|z) \propto q_\theta(z|\varepsilon) q(\varepsilon)$ using MCMC. Exploiting the fact that (z_s, ε_s) comes from the joint $q_\theta(z, \varepsilon)$, UIVI initializes the MCMC at ε_s so no burn-in is required. A number of iterations are run to break the dependency between ε_s and the ε'_s that is used to estimate the inner expectation.

1.1 Quality of approximation

TODO: analyze the (best-case) approximation of UIVI. Questions:

1. Approach? Probabilistic bound on KL as function of ELBO optimization iteration?
2. How to deal with implicit mixing component? Do surrogate families simpler than neural networks help? What assumptions would be needed?
3. Posterior contraction in terms of limiting data?

- Can we say something about ELBO maximizer $\hat{\theta}$, e.g.,
  - KL upper bound

$$\begin{aligned}
\mathrm{KL}(q_{\hat{\theta}}(z) \| p(z|x)) &= -\mathbb{E}_{q_{\hat{\theta}}(z)} \left[\log \frac{p(z|x)}{q_{\hat{\theta}}(z)} \right] \\
&= \mathbb{E}_{q_{\hat{\theta}}(z)} \left[\log \frac{q_{\hat{\theta}}(z)}{p(z|x)} \right] \\
&= \mathbb{E}_{q_{\hat{\theta}}(z)} \left[\log \frac{\mathbb{E}_{q(\varepsilon)} \left[q_{\hat{\theta}}(z|\varepsilon) \right]}{p(z|x)} \right]
\end{aligned}$$

  - ELBO lower bound

$$\begin{aligned}
\mathcal{L}(\hat{\theta}) &= \mathbb{E}_{q_{\hat{\theta}}(z)} \left[\log p(x, z) - \log q_{\hat{\theta}}(z) \right] \\
&= \mathbb{E}_{q_{\hat{\theta}}(z)} \left[\log p(x, z) - \log \mathbb{E}_{q(\varepsilon)} \left[q_{\hat{\theta}}(z|\varepsilon) \right] \right]
\end{aligned}$$

  - Simple case:
    $X \sim N(Z, \sigma^2)$, prior $Z \sim N(\mu_0, \sigma_0^2)$, posterior $Z|X_{1:n} \sim N\left(\frac{\mu_0 {\sigma_0}^{-2} + n\bar{X}\sigma^{-2}}{{\sigma_0}^{-2} + n\sigma^{-2}}, \sigma_1^2 = \frac{1}{{\sigma_0}^{-2} + n\sigma^{-2}}\right)$.
    Gaussian $q_\theta(z|\varepsilon)$:

    $$\begin{aligned} \varepsilon &\sim N(0,1) \\ u &\sim N(0,1) \\ z &= h_\theta(u;\varepsilon) = \mu_\theta(\varepsilon) + \sigma_1 u \\ \mu_\theta(\varepsilon) &= \theta + \varepsilon \\ z|\varepsilon &\sim N(\mu_\theta(\varepsilon), \sigma_1^2) = N(\theta + \varepsilon, \sigma_1^2) \\ z|\varepsilon, u &= \theta + \varepsilon + \sigma_1 u \\ z &\sim N(\theta, \sigma_1^2 + 1) \\ z &\sim N\left(\mathbb{E}\left[\mu_\theta(\varepsilon)\right], \sigma_1^2 + \mathrm{Var}\left(\mu_\theta(\varepsilon)\right)\right) \end{aligned}$$

    This says that for this normal-normal model, the true posterior is not in our variational family, and no function $\mu_\theta(\varepsilon)$ is able to change that unless $\mu_\theta(\varepsilon)$ is constant. TODO: problem is that σ in h_θ is misspecified. Learning both fixes issue?

    $$\begin{aligned} z &= \mu_\theta(\varepsilon_1) + \sigma_\theta(\varepsilon_2) u \\ &\sim N\left(\mathbb{E}\left[\mu_\theta(\varepsilon_1)\right], \mathrm{Var}\left(\mu_\theta(\varepsilon_1)\right) + \mathrm{Var}\left(\sigma_\theta(\varepsilon_2)\right)\right) \end{aligned}$$

    if learning independently.
  - Convolution: approximation identities
    Kruijer (2020): convolution error
  - Differential entropy not invariant under change of variables.

Approaches:

- Question mainly boils down to how expressive is the implicit distributional family?
- KL between true posterior and variational distribution:
  - Analytic approach: normal-normal example below shows simple case where true posterior is in variational family and where it is not.
  - More complicated attempt: come up with analytic $q_\theta(z)$ for more complex mixing (e.g., normalizing flow) but likely not generalizable as in general is intractable. Intention: for any well-behaved target and base, there exists a diffeomorphism that can turn the base into the target.
  - Plummer et al. [5] provides probabilistic bounds on KL between true posterior and variational distribution given by a particular implicit model (non-linear latent variable model with a Gaussian process prior), and maybe posterior contraction to true density? Unclear how generalizable results are based on current understanding.
- Posterior contraction/measure of approximation of variational distribution and limiting posterior?
- Dimensionality? Is this just a problem of convergence/complexity?

1.2 Gradient variance

TODOare there scenarios where UIVI breaks down that other VI methods may not?

- Approximating posterior using Gaussian mixtures in high-dimensions.
- Reparameterization gradients and variances + other relevant references:
  - The Generalized Reparameterization Gradient (Ruiz 2016)

  - Reducing Reparameterization Gradient Variance (Miller 2017)
  - Variance reduction properties of the reparameterization trick (Xu 2018): applies CLT over multiple samples so each partial derivative is approximately normal; Taylor expansions?
  - Monte Carlo Gradient Estimation in Machine Learning (Mohamed 2020)
  - EM Algorithm & High Dimensional Data (slides)
  - BBVI, structured SIVI

- Rao-Blackwellization:

$$\begin{aligned}
L(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}') &= \widehat{\nabla}_\theta \mathcal{L}(\theta; \mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}') \\
L(\boldsymbol{\varepsilon}) &= \mathbb{E}_{\mathbf{u}, \boldsymbol{\varepsilon}'}[L(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}')] \\
\mathrm{Var}(L(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}')) &= \mathrm{Var}(L(\boldsymbol{\varepsilon})) + \mathbb{E}[(L(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}') - L(\boldsymbol{\varepsilon}))^2] \\
&\geq \mathrm{Var}(L(\boldsymbol{\varepsilon}))
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\left(L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right) &= \mathbb{E}_{\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_{1:m}}\left[\left(L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}) - \mathbb{E}_{\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_{1:m}}\left[L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right]\right)^2\right] \\
&= \mathbb{E}_{\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_{1:m}}\left[\left(L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}) - \mathbb{E}_{\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_{1:m}}\left[L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right]\right)^2\right] \\
&= \mathrm{Var}\left(L_\theta(\boldsymbol{\varepsilon}'_{1:m})\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\left(L_\theta(\boldsymbol{\varepsilon}'_{1:m})\right) &= \mathbb{E}_{\boldsymbol{\varepsilon}'_{1:m}}\left[\left(L_\theta(\boldsymbol{\varepsilon}'_{1:m}) - \mathbb{E}_{\boldsymbol{\varepsilon}'_{1:m}}\left[L_\theta(\boldsymbol{\varepsilon}'_{1:m})\right]\right)^2\right] \\
&= \mathbb{E}_{\boldsymbol{\varepsilon}'_{1:m}}\left[\left(L_\theta(\boldsymbol{\varepsilon}'_{1:m}) - \mathbb{E}_{\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}}\left[L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right]\right)^2\right] \\
&= \mathbb{E}_{\boldsymbol{\varepsilon}'_{1:m}}\left[\left(L_\theta(\boldsymbol{\varepsilon}'_{1:m}) - L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}) + L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}) - \mathbb{E}_{\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}}\left[L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right]\right)^2\right] \\
&= \mathbb{E}_{\boldsymbol{\varepsilon}'_{1:m}}\left[\left(L_\theta(\boldsymbol{\varepsilon}'_{1:m}) - L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right)^2\right] \\
&\quad - 2\mathbb{E}_{\boldsymbol{\varepsilon}'_{1:m}}\left[\left(L_\theta(\boldsymbol{\varepsilon}'_{1:m}) - L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right)\left(L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}) - \mathbb{E}_{\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}}\left[L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right]\right)\right] \\
&\quad + \mathrm{Var}\left(L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right)
\end{aligned}$$

  TODOCan we show lower bound depends on dimension d of $\mathbf{z}$ somehow assuming that $q_\theta(\mathbf{z}) = p(\mathbf{z}|\mathbf{x})$ (or very close, e.g., in convolution family), meaning that in high dimensions, variance is still high even at the optimum? Take $\boldsymbol{\varepsilon}$ to be of dimension d and take mapping function as inverse CDF (this is just the posterior itself? i.e., $\mathrm{Var}_{q_\theta}(\mathbf{z}) = \mathrm{Var}_{p_{\mathbf{z}|\mathbf{x}}}(\mathbf{z})$ with $\sigma = 0$). Assume that $q_\theta(\mathbf{z}) \approx p(\mathbf{z}|\mathbf{x})$ so that variances are approximately equal?

- $$\widehat{\nabla}_\theta \mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\nabla_{\mathbf{z}} \log p(\mathbf{x}, \mathbf{z})\big|_{\mathbf{z}=h_\theta(\mathbf{u}_i;\boldsymbol{\varepsilon}_i)} - \frac{1}{m}\sum_{j=1}^{m} \nabla_{\mathbf{z}} \log q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}'_j)\big|_{\mathbf{z}=h_\theta(\mathbf{u}_i;\boldsymbol{\varepsilon}_i)}\right)\nabla_\theta h_\theta(\mathbf{u}_i; \boldsymbol{\varepsilon}_i)$$

  TODOUnder above assumptions?

$$\begin{aligned}
L(\boldsymbol{\varepsilon}) &= \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x}, h_\theta(\mathbf{u}; \boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})\right] - \mathbb{E}_{\mathbf{u}, \boldsymbol{\varepsilon}'}\left[\nabla_{\mathbf{z}} \log q_\theta(h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})|\boldsymbol{\varepsilon}')\nabla_\theta h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})\right] \\
&= \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x}, h_\theta(\mathbf{u}; \boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})\right] - \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}}\mathbb{E}_{\boldsymbol{\varepsilon}'}\left[\log q_\theta(h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})|\boldsymbol{\varepsilon}')\right]\nabla_\theta h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})\right] \\
&\geq \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x}, h_\theta(\mathbf{u}; \boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})\right] - \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log q_\theta(h_\theta(\mathbf{u}; \boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})\right] \\
&= 0 \\
\mathrm{Var}(L(\boldsymbol{\varepsilon})) &= \mathrm{Var}\left(\mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x}, h_\theta(\mathbf{u}; \boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})\right]\right) + \mathrm{Var}\left(\mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}}\mathbb{E}_{\boldsymbol{\varepsilon}'}\left[\log q_\theta(h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})|\boldsymbol{\varepsilon}')\right]\nabla_\theta h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})\right]\right) \\
&\quad - \mathrm{Cov}\left(\mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x}, h_\theta(\mathbf{u}; \boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})\right]\left(\mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}}\mathbb{E}_{\boldsymbol{\varepsilon}'}\left[\log q_\theta(h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})|\boldsymbol{\varepsilon}')\right]\nabla_\theta h_\theta(\mathbf{u}; \boldsymbol{\varepsilon})\right]\right)^\top\right)
\end{aligned}$$

TODOinequality of $L(\boldsymbol{\varepsilon})$ not useful?

$$\begin{aligned}
L(\boldsymbol{\varepsilon}) &= \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x}, h_\theta(\mathbf{u};\boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] - \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}}\mathbb{E}_{\boldsymbol{\varepsilon}'}\left[\log q_\theta(h_\theta(\mathbf{u};\boldsymbol{\varepsilon})|\boldsymbol{\varepsilon}')\right]\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] \\
&= \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x}, h_\theta(\mathbf{u};\boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] - \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log q_\theta(h_\theta(\mathbf{u};\boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] \\
&= \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log p(h_\theta(\mathbf{u};\boldsymbol{\varepsilon})|\mathbf{x})\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] - \mathbb{E}_{\mathbf{u}}\left[\nabla_{\mathbf{z}} \log q_\theta(h_\theta(\mathbf{u};\boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right]
\end{aligned}$$

If have θ such that $q_\theta(\mathbf{z}) = p(\mathbf{z}|\mathbf{x})$,

$$\begin{aligned}
L(\boldsymbol{\varepsilon}'_{1:m}) &= \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x}, h_\theta(\mathbf{u};\boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] - \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_{\mathbf{z}} \log q_\theta(h_\theta(\mathbf{u};\boldsymbol{\varepsilon})|\boldsymbol{\varepsilon}'_j)\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] \\
&= \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x}, h_\theta(\mathbf{u};\boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] - \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_\theta \log q_\theta(h_\theta(\mathbf{u};\boldsymbol{\varepsilon})|\boldsymbol{\varepsilon}'_j)\right] \\
&= \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x}, h_\theta(\mathbf{u};\boldsymbol{\varepsilon}))\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] - \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_\theta \log q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}'_j)\big|_{\mathbf{z}=h_\theta(\mathbf{u};\boldsymbol{\varepsilon})}\right]
\end{aligned}$$

Given θ and $\boldsymbol{\varepsilon}'_j$, second term is expected score but unless $q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}') = p(\mathbf{z}|\mathbf{x})$ (p must be Gaussian), expectation is non-zero?

$$\begin{aligned}
\mathrm{Var}(L(\boldsymbol{\varepsilon}'_{1:m})_i) &= \frac{1}{m}\mathrm{Var}\left(\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_{\theta_i} \log q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}')\big|_{\mathbf{z}=h_\theta(\mathbf{u};\boldsymbol{\varepsilon})}\right]\right) \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\varepsilon}'}\left[\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_{\theta_i} \log q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}')\big|_{\mathbf{z}=h_\theta(\mathbf{u};\boldsymbol{\varepsilon})}\right]^2\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\varepsilon}'}\left[\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\left(\Sigma_\theta(\boldsymbol{\varepsilon}')^{-1}(h_\theta(\mathbf{u};\boldsymbol{\varepsilon}) - \mu_\theta(\boldsymbol{\varepsilon}'))\right)^\top \nabla_{\theta_i} h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right]^2\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\varepsilon}'}\left[\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\left(\Sigma_\theta(\boldsymbol{\varepsilon}')^{-1}((\mu_\theta(\boldsymbol{\varepsilon}) + \Sigma_\theta(\boldsymbol{\varepsilon})\mathbf{u}) - \mu_\theta(\boldsymbol{\varepsilon}'))\right)^\top \nabla_{\theta_i} h(\mathbf{u};\boldsymbol{\varepsilon})\right]^2\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\varepsilon}'}\left[\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\left(\Sigma_\theta(\boldsymbol{\varepsilon}')^{-1}(\mu_\theta(\boldsymbol{\varepsilon}) - \mu_\theta(\boldsymbol{\varepsilon}')) + \Sigma_\theta(\boldsymbol{\varepsilon}')^{-1}\Sigma_\theta(\boldsymbol{\varepsilon})\mathbf{u}\right)^\top \nabla_{\theta_i} h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right]^2\right]
\end{aligned}$$

If Σ independent of $\boldsymbol{\varepsilon}$, then

$$\mathrm{Var}(L(\boldsymbol{\varepsilon}'_{1:m})_i) = \frac{1}{m}\mathbb{E}_{\boldsymbol{\varepsilon}'}\left[\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\left(\Sigma^{-1}(\mu_\theta(\boldsymbol{\varepsilon}) - \mu_\theta(\boldsymbol{\varepsilon}')) + \mathbf{u}\right)^\top \nabla_{\theta_i} h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right]^2\right]$$

Can simplify further if Σ diagonal but is there a point?

Notes:

1. $\boldsymbol{\varepsilon}'$ sampled from $q(\boldsymbol{\varepsilon}|\mathbf{z})$ via MCMC, but variance w.r.t. $q(\boldsymbol{\varepsilon})$ considered here?
2. Dimension of $\mathbf{z}$ shows up as score
3. $\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'}\left[\nabla_{\theta_i} \log q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}')\big|_{\mathbf{z}=h_\theta(\mathbf{u};\boldsymbol{\varepsilon})}\right] = 0$
4. Variance is variance of expected score of potentially high-dimensional distribution where parameters are changing according to $\boldsymbol{\varepsilon}$ but map to parameters of distribution is fixed due to fixed θ

- TODO: start with finite mixture, i.e.,

$$\begin{aligned}
q_\theta(\mathbf{z}) &= \sum_{i=1}^K q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}_k)q(\boldsymbol{\varepsilon}_k) \\
\mathbf{z} &= \mu(\boldsymbol{\varepsilon}_k) + \sigma(\boldsymbol{\varepsilon}_k)\mathbf{u} \\
\mathbf{u} &\sim \mathcal{N}(0,1) \\
\boldsymbol{\varepsilon} &\sim q(\boldsymbol{\varepsilon})
\end{aligned}$$

Show variance depends on K, and somehow K is known to be exponential in d?

1.3 Scratch notes

$\hat{\theta} = \frac{\mu_0 \sigma_0^{-2} + n\bar{X}\sigma^{-2}}{\sigma_0^{-2} + n\sigma^{-2}}$:

$$\begin{aligned} \mathrm{KL}(q_\theta(z) \| p(z|x)) &= -\int q_\theta(z) \log \frac{p(z|x)}{q_\theta(z)} dz \\ &= -\int q_\theta(z) \log p(z|x) + \int q_\theta(z) \log q_\theta(z) dz \end{aligned}$$

$$\begin{aligned} u &\sim N(0,1) \\ z &= h_\theta(u;\varepsilon) = \mu_\theta(\varepsilon) + \sigma_1 u \\ \mu_\theta(\varepsilon) &= \theta + \varepsilon \\ u &= h_\theta^{-1}(z;\varepsilon) = \sigma_1^{-1}(z - \mu_\theta(\varepsilon)) \\ \nabla_z h_\theta^{-1}(z;\varepsilon) &= \sigma_1^{-1} \\ q_\theta(z|\varepsilon) &= q_u(h_\theta^{-1}(z;\varepsilon))\sigma_1^{-1} \\ q_\theta(z) &= \int q_\theta(z|\varepsilon) q(\varepsilon) d\varepsilon \\ &= \int \sigma_1^{-1} q_u\left(\sigma_1^{-1}\left(z - \mu_\theta(\varepsilon)\right)\right) q(\varepsilon) d\varepsilon \\ &= \int \sigma_1^{-1} q_u\left(\sigma_1^{-1}\left(z - \theta - \varepsilon\right)\right) q(\varepsilon) d\varepsilon \\ &= \int \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2}\left(\sigma_1^{-2}(z-\theta-\varepsilon)^2\right)\right) q(\varepsilon) d\varepsilon \\ &= \int \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2}\left((z-\theta)^2 - 2(z-\theta)\varepsilon + \varepsilon^2\right)\right) q(\varepsilon) d\varepsilon \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2}(z-\theta)^2\right) \int \exp\left(-\frac{1}{2\sigma_1^2}\left(-2(z-\theta)\varepsilon + \varepsilon^2\right)\right) q(\varepsilon) d\varepsilon \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2}(z-\theta)^2\right) \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_1^2}\left(-2(z-\theta)\varepsilon + \varepsilon^2\right) - \frac{1}{2}\varepsilon^2\right) d\varepsilon \end{aligned}$$

Posterior exact when ?

- If h_θ monotonic, invertible:

$$\begin{aligned} z &= h_\theta(u;\boldsymbol{\varepsilon}) \\ q_\theta(z|\boldsymbol{\varepsilon}) &= q_u\left(h_\theta^{-1}(z;\boldsymbol{\varepsilon})\right) \left|\nabla_z h_\theta^{-1}(z;\boldsymbol{\varepsilon})\right| \\ q_\theta(z) &= \int q_u\left(h_\theta^{-1}(z;\boldsymbol{\varepsilon})\right) \left|\nabla_z h_\theta^{-1}(z;\boldsymbol{\varepsilon})\right| q(\boldsymbol{\varepsilon}) d\boldsymbol{\varepsilon} \end{aligned}$$

TODO: normalizing flow literature? Restrict h_θ to be independent of $\boldsymbol{\varepsilon}$ (e.g., linear flows)?

2 Semi-implicit variational inference

Based on Yin and Zhou [8].

SIVI is addresses the issues of classical VI attributed to the requirement of a conditionally conjugate variational family by relaxing this requirement to allow for implicit distributional families from which samples can be drawn. This implicit family consists of hierarchical distributions with a mixing parameter. While the distribution conditioned on the mixing parameter is required to be analytical and reparameterizable, the mixing distribution can be arbitrarily complex. The use of such a variational family also addresses the problems of conventional mean-field families as dependencies between the latent variables can be introduced through the mixing distribution.

The objective in SIVI is a surrogate ELBO that is only exact asymptotically and otherwise a lower bound of the ELBO [3]. Like in black box VI, the gradients are rewritten as expectations and estimated via Monte Carlo samples.

Molchanov et al. [3] extends SIVI to doubly SIVI for variational inference and variational learning in which both the variational posterior and the prior are semi-implicit distributions. They also show that the SIVI objective is a lower bound of the ELBO.

Molchanova et al. [4] and Moens et al. [2] comment that SIVI and UIVI struggle in high-dimensional regimes. MCMC methods also have high variance [2].

Moens et al. [2] introduce compositional implicit variational inference (CI-VI), which rewrites the SIVI ELBO as a compositional nested form $\mathbb{E}_\nu\left[f_\nu\left(\mathbb{E}_\omega\left[g_\omega(\theta)\right]\right)\right]$. The gradient involves estimating the nested expectations, for which a simple Monte-Carlo estimator would be biased. CI-VI uses an extrapolation-smoothing scheme for which the bias converges to zero with iterations. In practice, the gradient involves matrix-vector products that are expensive but can be approximated via sketching techniques. Under certain assumptions, convergence of the CI-VI algorithm is proved in terms of the number of oracle calls needed to convergence (TODO).

3 Hierarchical variational inference

Based on Ranganath et al. [6].

Predating SIVI and UIVI, HVM first(?) addressed the restricted variational family issue of classical VI by using a hierarchical variational distribution which is enabled by BBVI. HVM considers a mean-field variational likelihood and a variational prior that is differentiable (e.g., a mixture or a normalizing flow). HVM also optimizes a lower bound of the ELBO that is constructed using a recursive variational distribution that approximates the variational prior.

4 Theoretical guarantees for implicit VI

Based on Plummer et al. [5].

TODO: Considers non-linear latent variable model (NL-LVM)

$$\begin{aligned} z &= \mu(\varepsilon) + u \\ u &\sim N(0, \sigma^2) \\ \varepsilon &\sim U(0,1) \\ \mu &\sim \Pi_\mu \\ \sigma &\sim \Pi_\sigma \end{aligned}$$

where Π_μ and Π_σ are priors. Can write as

$$\begin{aligned} z &= \mu(\varepsilon) + \sigma u \\ u &\sim N(0,1) \end{aligned}$$

This leads to density

$$\begin{aligned} f_{\mu,\sigma}(z) = f(z;\mu,\sigma) &= \int_0^1 \phi_\sigma(y - \mu(\boldsymbol{\varepsilon}))d\boldsymbol{\varepsilon} \\ &= \int \phi_\sigma(y-t)d\nu_\mu(t) \end{aligned}$$

where ϕ_σ is the density of a $\mathrm{N}(0, \sigma^2 \mathbf{I}_d)$ distribution, and $\nu_\mu = \lambda \circ \mu^{-1}$ the image measure where λ is the Lebesgue measure and $\mu : [0,1] \to \mathbb{R}$. The second form is a convolution with a Gaussian kernel and suggests that $f_{\mu,\sigma}$ is flexible depending on the choice of μ. Under certain assumptions on f_0, it is known that $\phi_\sigma * f_0$ can approximate f_0 arbitrarily close as bandwidth $\sigma \to 0$. This should hold for UIVI under particular choices of the reparameterization and mixing distributions.

A Gaussian process latent variable model puts a GP prior for the transfer function μ. (Theorem 3.1) If Π_μ has full sup-norm support on $C[0,1]$ and Π_σ has full support on $[0,\infty)$, then the L_1 support of the induced prior $\Pi = (\Pi_\mu \otimes \Pi_\sigma) \circ f_{\mu,\sigma}^{-1}$ contains all densities which have a first finite moment and are non-zero almost everywhere on their support.

TODO: posterior contraction says expected divergence of posterior density and true density goes to 0 given observations of the response z. The response in our case is the latent variable. Can this work with our observations x? This likely does not apply as the true posterior is the one that changes and the variational distribution is only approximating the true posterior. If the true posterior is always in the family and we can approximate it exactly then posterior contraction follows standard Bayesian results.

Introduces Gaussian process implicit VI (GP-IVI), which uses a finite mixture of uniform mixing distributions. TODO: transfer function not necessarily GP? Has probabilistic bound on error of best approximation to posterior and an α-variational Bayes risk bound.

For simple normal-normal model, KL divergence for true normal model and true posterior converges weakly to a χ_1^2 and not to 0.

5 Other references

VI review:

- Advances in Variational Inference (2019)
- Variational Inference: A Review for Statisticians (2017)
- Black Box Variational Inference (2013): dominated convergence theorem used to push gradient into expectation

Possibly related VI approaches/of interest

- Semi-Implicit Variational Inference (2018)

  Doubly Semi-Implicit Variational Inference (2019)

  Structured Semi-Implicit Variational Inference (2019): mentions that previous methods scale exponentially with dimension of the latent variables. Imposes that the high-dimensional semi-implicit distribution factorizes into a product of low-dimensional conditional semi-implicit distributions and shows that the resulting entropy bound is tighter than that of SIVI's and consequently a tighter ELBO objective.

  Efficient Semi-Implicit Variational Inference (2021)
- Variational Inference using Implicit Distributions (2017): implicit with density ratio estimation?
- Importance Weighted Hierarchical Variational Inference (2019)
- Normalizing Flows for Probabilistic Modeling and Inference (2021)

  Stochastic Normalizing Flows (2020)
- Implicit VI:

  Variational Inference with Implicit Models (2018; slides)

  Implicit Variational Inference: the Parameter and the Predictor Space (2020): optimizing over predictor space rather than parameter space?

Theory/analysis

- Statistical Guarantees for Transformation Based Models with Applications to Implicit Variational Inference (2021)

  Statistical and Computational Properties of Variational Inference (2021; thesis)
- Theoretical Guarantees of Variational Inference and Its Applications (2020; thesis)

  α-Variational Inference with Statistical Guarantees (2018): a particular variational family with theoretical guarantees
- Contributions to the theoretical study of variational inference and robustness (2020; thesis)
- On Statistical Optimality of Variational Bayes (2018): general guarantees for variational estimates as approximations for true data-generating parameter for MF-VI using variational risk bounds?

  Statistical guarantees for variational Bayes (2021; slides)
- Statistical Guarantees and Algorithmic Convergence Issues of Variational Boosting (2020)
- Robust, Accurate Stochastic Optimization for Variational Inference (2020) – iterates as MCMC?
- Convergence Rates of Variational Inference in Sparse Deep Learning (2019)

  On the Convergence of Extended Variational Inference for Non-Gaussian Statistical Models (2020)

References

[1] Hwan Chung, Eric Loken, and Joseph L Schafer. Difficulties in drawing inferences with finite-mixture models: A simple example with a simple solution. *The American Statistician*, 58(2):152–158, 2004.

[2] Vincent Moens, Hang Ren, Alexandre Maraval, Rasul Tutunov, Jun Wang, and Haitham Ammar. Efficient semi-implicit variational inference. *arXiv preprint arXiv:2101.06070*, 2021.

[3] Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov. Doubly semi-implicit variational inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2593–2602. PMLR, 2019.

[4] Iuliia Molchanova, Dmitry Molchanov, Novi Quadrianto, and Dmitry Vetrov. Structured semi-implicit variational inference. 2019.

[5] Sean Plummer, Shuang Zhou, Anirban Bhattacharya, David Dunson, and Debdeep Pati. Statistical guarantees for transformation based models with applications to implicit variational inference. In *International Conference on Artificial Intelligence and Statistics*, pages 2449–2457. PMLR, 2021.

[6] Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333. PMLR, 2016.

[7] Michalis K Titsias and Francisco Ruiz. Unbiased implicit variational inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 167–176. PMLR, 2019.

[8] Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *International Conference on Machine Learning*, pages 5660–5669. PMLR, 2018.