Contents

1 Unbiased Implicit Variational Inference

Based on Titsias and Ruiz [5].

- 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.
- Questions: TODO
  1. Can the gradient be pushed into the expectation? (Section 2.2)
- 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 [5] 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

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

    (TODO: where is $\mathbb{E}_{q_\theta(z)} [\nabla_\theta \log q_\theta(z)] = 0$ applied?) (Gradient can be pushed into expectation using DCT.) 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, $S = 1$):

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

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 Analysis

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

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

2 Semi-implicit variational inference

Based on Yin and Zhou [6].

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 [2]. Like in black box VI, the gradients are rewritten as expectations and estimated via Monte Carlo samples.

Molchanov et al. [2] 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. [3] and Moens et al. [1] comment that SIVI and UIVI struggle in high-dimensional regimes. MCMC methods also have high variance [1].

Moens et al. [1] 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 estimated 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. [4].

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 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)
- Stochastic Normalizing Flows (2020)

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)
- Contributions to the theoretical study of variational inference and robustness (2020; thesis)
- On Statistical Optimality of Variational Bayes (2018)

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

[2] 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.

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

[4] Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International conference on machine learning*, pages 324–333. PMLR, 2016.

[5] 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.

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