VARIATIONAL INFERENCE USING NORMALIZING FLOWS

SILAS BRACK AND JESPER MARTINHOFF KNUDSEN

Abstract. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1. Introduction

In this project we will apply normalizing flows in the realm of variational inference. So to give some context to our endeavour we will start out by motivating variational inference and normalizing flows.[1]

1.1. **Variational inference.**

In variational inference we strive to approximate some distribution by an element in some distributional family.

The reason we might wish to do so is often that we have some implicit distribution which we do not have an analytical expression for but we might be able to sample form it by use of some sampling scheme like MCMC chains.

in our case this we have an implicit distribution which arises from forming the posterior form some non conjugate prior-likelihood pair:

$$p(z|\mathcal{D}) = \frac{p(\mathcal{D}|z)p(z)}{p(\mathcal{D})}$$

Here we have two annoyances, one being the intractability of the integral in $p(\mathcal{D}) = \int p(\mathcal{D}|z)p(z)dz$ and the other being annoyance of the non-parametric functional form of: $p(z|\mathcal{D}) \propto \prod_n p(d_n|z)p(z)$ which becomes excessively cumbersome for inference on large data sets. Thus this incentivises us to find a

Thanks to our supervisor Michael Riis Andersen.

[1] `https://github.com/silasbrack/normalizing-flows`

function which can approximate $p(z|\mathcal{D}) \approx q_\theta(z)$ as this would allow us to sample directly and independently from $q\theta$ and have q_θ be a parametric distribution and thus potentially much less cumbersome. However the effectiveness of this approach all depends on how well $q_\theta(z)$ approximates $p(z|\mathcal{D})$. This necessitates rich distributional families. and this is where the frame work of normalizing flows seems particular promising. That is normalizing flows gives us a general frame work for constructing arbitrarily complex distributional families. The general idea is to start out with simple distribution in the sense that they are easy to sample from, then we apply a series of transformations to end up with a possibly much more complex distribution, but one which we can still sample independently from as we know its relation to a sampleable distribution.

With this we are ready to start developing normalizing flows further and start to look at how we can use them to approximate arbitrary distributions.

Letting $p(z|y) \approx q(z)$, we calculate the posterior predictive as:

$$p(y^*|y) = \int p(y^*|z)p(z|y)dz$$
$$p(y^*|y) \approx \int p(y^*|z)q(z)dz$$

2. Normalising Flows

As previously mentioned a normalizing flow can be defined by a sampleable base distribution along with a series of diffeomorphisms:

$$z_0 \sim q(z_0) \qquad\qquad f_{(n)} = f_n \circ f_{n-1} \circ \cdots \circ f_1$$

and well denote the z_0s image through $f_{(n)}$ as z_n, ie $f_{(n)}(z_0) = z_n$.
We can then recursively apply the change of variable formula:

$$p(z') = p(z)|\det(J(f^{-1}))| = p(z)|\det(J(f \circ z))|^{-1} \quad \text{where} \quad z' = f(z)$$

to derive $q(z_n)$, here the last equality is due to the choice of diffeomorphisms. Through recursive application over all f_i we derive:

$$q(z_n) = q(z_0) \prod_k |\det J(f_k \circ z_{k-1})|^{-1}.$$

This can be viewed as parametric distributional family parameterized by the choice of $q(z_0)$ as well as the choice of $f_{(n)}$ whose parameters we will denote $\theta = \{\theta_n, \theta_{n-1}, \cdots, \theta_1\}$ [3]

3. Invertible Linear-Time Transformations

A simple construction of a normalizing flow could be invertible linear-time transformations. Their general form is given by:

$$f(z) = \mathbf{z} + \mathbf{u}h_\theta(\mathbf{w}^T\mathbf{z} + b)$$

with parameters $\lambda = \{\mathbf{w}, \mathbf{u}, b, \theta\}$ and h smooth nonliterary, NOTE[I don't know why as i would assume that it also has to be invertible for f to be invertible] by application of the theory discussed in [2]

$$\begin{aligned} q(z_n) &= q(z_0) \prod_k |\det J(f_k \circ z_{k-1})|^{-1} \\ &= q(z_0) \prod_k |1 + \det(\mathbf{u}h'(z_{k-1})^T\mathbf{w})|^{-1} \\ &= q(z_0) \prod_k |1 + \mathbf{u}^T h'(z_{k-1})\mathbf{w}|^{-1} \end{aligned}$$

a good pick here could be the SE kernel:

$$\begin{aligned} h(x)_\alpha &= \exp(-\frac{x^2}{\alpha}) \\ h'(x)_\alpha &= -\exp(-\frac{x^2}{\alpha}) \cdot 2 \cdot \frac{x}{\alpha} \end{aligned}$$

4. Optimization

When we then to fit the posterior to the variational family we need an objective function. Ie how do we determine what is a good fit. A common chioce is the KL-divergence:

$$\begin{aligned} KL[q_\theta||p] &= \int p(z) \ln \frac{p(z)}{q_\theta(z)} dz \\ &= \ln p(x) - L[q_\theta] \end{aligned}$$

Thus we get:

$$\ln p(x) = KL[q_\theta||p] + L[q_\theta]$$

Here $\ln p(x)$ is the marginal likelihood and thus does not does not depend on q_θ and is thus a constant w.r.t. θ as such maximizing $L[q_\theta]$ over θ, minimizes $KL[q_\theta||p]$ so we will resort to maximizing $L[q_\theta]$ due to their equivalence.

$$\begin{aligned} L[q_\theta] &= \mathbb{E}_{q_\phi(z_n)}\left[\ln q_\theta(z_n)\right] - \mathbb{E}_{q_\phi(z_n)}\left[\ln p(x, z_n)\right] \\ &= \mathbb{E}_{q_0(z_0)}\left[\ln q_\theta(f_{(n)} \circ z_0)\right] - \mathbb{E}_{q_0(z_0)}\left[\ln p(x, f_{(n)} \circ z_0)\right] \\ &= \mathbb{E}_{q_0(z_0)}\left[\ln q(z_0)\right] - \sum_k \mathbb{E}_{q_0(z_0)}\left[\ln|\det J(f_k \circ f_{(k-1)} \circ z_0)|\right] \\ &- \mathbb{E}_{q_0(z_0)}\left[\ln p(x, f_{(n)} \circ z_0)\right] \end{aligned}$$

As $\mathbb{E}_{q_0(z_0)}\left[\ln q(z_0)\right] \perp \phi$ we can leave it out as it will not effect the optimization, we can furthere apply the reparametrization trick to $z_0 \sim q_0$, and then estimate the gradients using that:

$$\begin{aligned} L[q_\theta] &= -\sum_k \mathbb{E}_{q_0(z_0)}\left[\ln|\det J(f_k \circ f_{(k-1)} \circ z_0)|\right] \\ &\quad - \mathbb{E}_{q_0(z_0)}\left[\ln p(x, f_{(n)} \circ z_0)\right] \\ &= -\sum_k \mathbb{E}_{q(\varepsilon)}\left[\ln|\det J(f_k \circ f_{(k-1)} \circ g(\lambda, \varepsilon))|\right] \\ &\quad - \mathbb{E}_{q(\varepsilon)}\left[\ln p(x, f_{(n)} \circ g(\lambda, \varepsilon))\right] \\ \nabla_\phi L[q_\theta] &= -\sum_k \int q(\varepsilon) \nabla_\phi \ln|\det J(f_k^{\phi_k} \circ f_{(k-1)}^{(\phi_{k-1})} \circ g(\lambda, \varepsilon))| d\varepsilon \\ &\quad - \int q(\varepsilon) \nabla_\phi \ln p(x, f_{(n)}^{\phi} \circ g(\lambda, \varepsilon)) d\varepsilon \\ &\approx -\sum_k \sum_{n \in N} \nabla_\phi \ln|\det J(f_k^{\phi_k} \circ f_{(k-1)}^{(\phi_{k-1})} \circ g(\lambda, \varepsilon_n))| \\ &\quad - \sum_{n \in N} \nabla_\phi \ln p(x, f_{(n)}^{\phi} \circ g(\lambda, \varepsilon_n)), \quad \{\varepsilon_n\}^{|N|} \sim q(\varepsilon) \end{aligned}$$

Where we will use autograd to evaluate $\nabla_\phi \ln|\det J(f_k^{\phi_k} \circ f_{(k-1)}^{(\phi_{k-1})} \circ g(\lambda, \varepsilon_n))|$ and

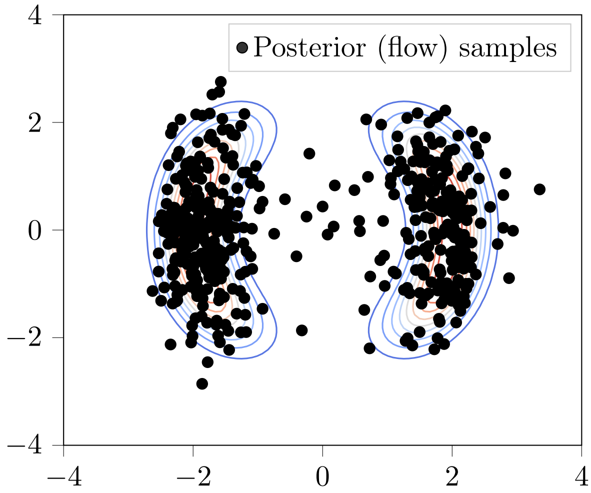

FIGURE 1. Caption

$\nabla_\phi \ln p(x, f^{\phi}_{(n)} \circ g(\lambda, \varepsilon_n))$ at ε_n, this will allow us to recursively approximate the direction of greatest ascent and we can then use something like the Adam algorithm to fit the flow our posterior.

4.1. **Optimation of planar flows.** from the derived formula above we can easily find the

5. Results

TABLE 1. Fefjfiewjfoewjf for seed 0.

	Mean-field	Full-rank	Planar			Radial		
			4	8	16	4	8	16
ELBO								
$\mu - \hat{\mu}$								
$\sigma^2 - \hat{\sigma}^2$								
$\hat{k}$								

6. Conclusion

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

References

[1] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[2] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using real nvp". In: *arXiv preprint arXiv:1605.08803* (2016).

[3] Andrew Gelman et al. *Bayesian Data Analysis*. CRC Press, 2013 (page 3).

[4] George Papamakarios et al. "Normalizing flows for probabilistic modeling and inference". In: *arXiv preprint arXiv:1912.02762* (2019).

[5] Sameera Ramasinghe et al. "Robust normalizing flows using Bernstein-type polynomials". In: *arXiv preprint arXiv:2102.03509* (2021).

[6] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.

[7] Yuling Yao et al. "Yes, but did it work?: Evaluating variational inference". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5581–5590.

Appendix A. Appendix

A.1. **The Reparametrization Trick.** To elaborate a bit on the reparameterization trick used in (4): if we let $q_0 = \mathcal{N}(\mu, \Sigma)$ and thus let $\lambda = \{\mu, \Sigma\}$ we can express q_0 as

$$q_0 = L\varepsilon + \mu, \quad \varepsilon \sim \mathcal{N}(0, \mathbb{I})$$

where