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[1, 2] 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 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

Thanks to our supervisor Michael Riis Andersen.

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

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:

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

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:

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

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. INVERTIBLE LINEAR-TIME TRANSFORMATIONS

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

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

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 section 2.

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

a good pick here could be the SE kernel:

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

4. OPTIMIZATION

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

$$\begin{aligned} D_{\mathrm{KL}}(q_\theta \,||\, p) &= \int q_\theta(z)\ln\frac{q_\theta(z)}{p(z)}dz \\ &= \ln p(x) - L[q_\theta] \end{aligned} \tag{1}$$

Thus we get:

$$\ln p(x) = D_{\mathrm{KL}}(q_\theta \mid\mid p) + L[q_\theta]$$

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 $D_{\mathrm{KL}}(q_\theta \mid\mid p)$ so we will resort to maximizing $L[q_\theta]$ due to their equivalence.

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

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

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

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

Algorithm 1 Variational Inf. with Normalizing Flows

Parameters: $\boldsymbol{\phi}$ variational, $\boldsymbol{\theta}$ generative
while not converged **do**
  $\boldsymbol{x} \leftarrow \{\text{Get mini-batch}\}$
  $\boldsymbol{z}_0 \sim q_0(\bullet \mid \boldsymbol{x})$
  $\boldsymbol{z}_k \leftarrow f_K \circ f_{K-1} \circ \ldots \circ f_1(\boldsymbol{z}_0)$
  $\mathcal{F}(\boldsymbol{x}) \approx \mathcal{F}(\boldsymbol{x}, \boldsymbol{z}_K)$
  $\Delta\boldsymbol{\theta} \propto -\nabla_\theta \mathcal{F}(\boldsymbol{x})$
  $\Delta\boldsymbol{\phi} \propto -\nabla_\phi \mathcal{F}(\boldsymbol{x})$
end while

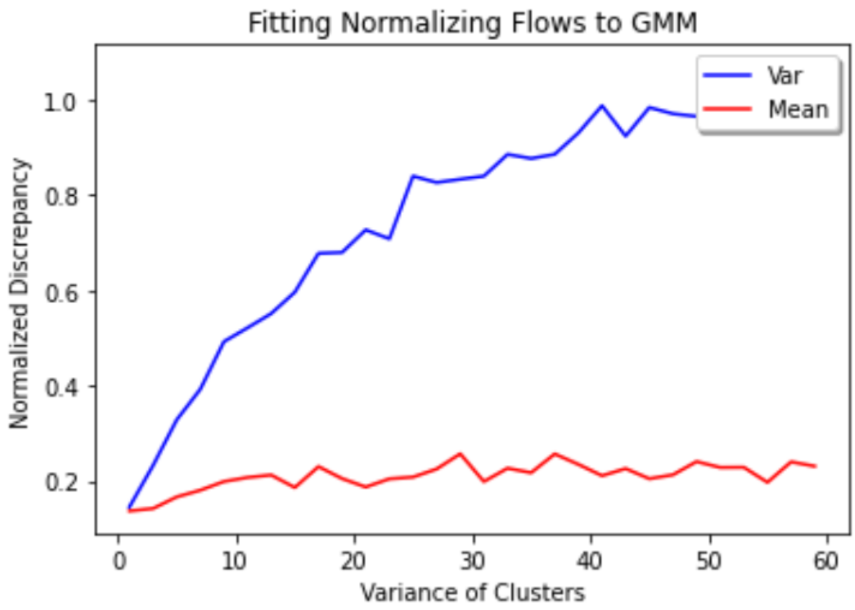

FIGURE 1. Experimentation on fitting planar flows on Gaussian Mixture Model with increasing variance and comparing mean (red) and variance (blue) of the two. Here the x-axis denotes the difference in variance $\sigma^2_{\text{NF}} - \sigma^2_{\text{gmm}}$ and the y-axis denotes the difference in mean $|\mu_{\text{gmm}} - \mu_{\text{NF}}|$

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

4.2. **Training.** Furthermore, for Planar and Radial flows training is limited by a single entry point ... [3, 4]

4.3. **Metrics.** - ELBO / KL divergence - difference of means - difference of variance - k-hat

$$\left\|\Sigma^{-1}(\mu - \hat{\mu})\right\|$$

$$\left\|I - \Sigma^{-1}\hat{\Sigma}\right\|_F$$

5. RESULTS

Pyro[5] was used to perform stochastic variational inference automatically and ArviZ[6] was used to generate simple statistics, summaries and visualizations for the inference results.

5.1. **Multivatiate normal.** Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam

FIGURE 2. Final ELBO of normalizing flow as a function of number of flows for correlated multivariate Normal posterior. The final ELBO was calculated as the mean of the ELBOs of the last 500 iterations from training.

TABLE 1. Final ELBO for different potential energy posteriors from Rezende and Mohamed [1] (table 2).

	Planar			Radial		
	2	8	32	2	8	32
$U_1(z)$	73.50	256.24	279.24	-61.84	236.32	272.28
$U_2(z)$	271.71	501.16	659.48	-94.02	184.09	550.71
$U_3(z)$	418.16	666.00	967.83	211.54	576.92	761.59
$U_4(z)$	473.85	638.39	750.76	229.20	462.36	648.31

FIGURE 3. Average ELBO of the last 1000 iterations for ...

FIGURE 4. 360.0pt

in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

With covariance matrix

5.2. **Energy functions.** 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.

The normalizing flows were trained for 10000 epochs with an ADAM learning rate of 0.005, 16 Markov Chain gradient estimation samples and 256 base distribution samples ($\boldsymbol{z}_0 \sim q_0(\bullet|\boldsymbol{x})$).

Rezende and Mohamed [1] used normalizing flows for variational inference on four target posterior distributions, defined as in table 2.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a

TABLE 2. Test energy functions.

Potential $U(\mathbf{z})$
1: $\frac{1}{2}\left(\frac{\|\mathbf{z}\|-2}{0.4}\right)^2 - \ln\left(e^{-\frac{1}{2}\left[\frac{\mathbf{z}_1-2}{0.6}\right]^2} + e^{-\frac{1}{2}\left[\frac{\mathbf{z}_1+2}{0.6}\right]^2}\right)$
2: $\frac{1}{2}\left[\frac{\mathbf{z}_2-w_1(\mathbf{z})}{0.4}\right]^2$
3: $-\ln\left(e^{-\frac{1}{2}\left[\frac{\mathbf{z}_2-w_1(\mathbf{z})}{0.35}\right]^2} + e^{-\frac{1}{2}\left[\frac{\mathbf{z}_2-w_1(\mathbf{z})+w_2(\mathbf{z})}{0.35}\right]^2}\right)$
4: $-\ln\left(e^{-\frac{1}{2}\left[\frac{\mathbf{z}_2-w_1(\mathbf{z})}{0.4}\right]^2} + e^{-\frac{1}{2}\left[\frac{\mathbf{z}_2-w_1(\mathbf{z})+w_3(\mathbf{z})}{0.35}\right]^2}\right)$
with $w_1(\mathbf{z}) = \sin\left(\frac{2\pi\mathbf{z}_1}{4}\right)$, $w_2(\mathbf{z}) = 3e^{-\frac{1}{2}\left[\frac{(\mathbf{z}_1-1)}{0.6}\right]^2}$, $w_3(\mathbf{z}) = 3\sigma\left(\frac{\mathbf{z}_1-1}{0.3}\right)$ and $\sigma(x) = 1/(1+e^{-x})$.

nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

TABLE 3. Metrics

Number of flows	2			8			32		
ELBO samples	32	128	512	32	128	512	32	128	512
ELBO									
$\mu - \hat{\mu}$									
$\sigma^2 - \hat{\sigma}^2$									
$\hat{k}$									

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

From the log term in eq. (1) it is clear that minimizing the KL-divergence heavily penalizes candidate distributions q for which the probability at z, $q(z)$, is high and $p(z|x)$ is

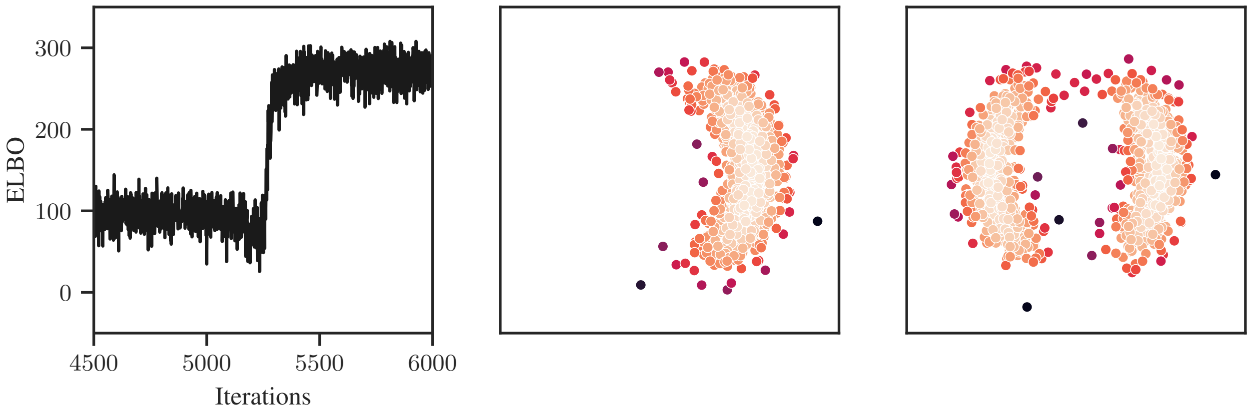

FIGURE 5. Training curve for a normalizing flow with 4 planar flows. When variational inference is performed on this bi-modal target distribution, an 'elbow' forms in the ELBO as the variational approximation changes shape to 'snap' to the new mode[7].

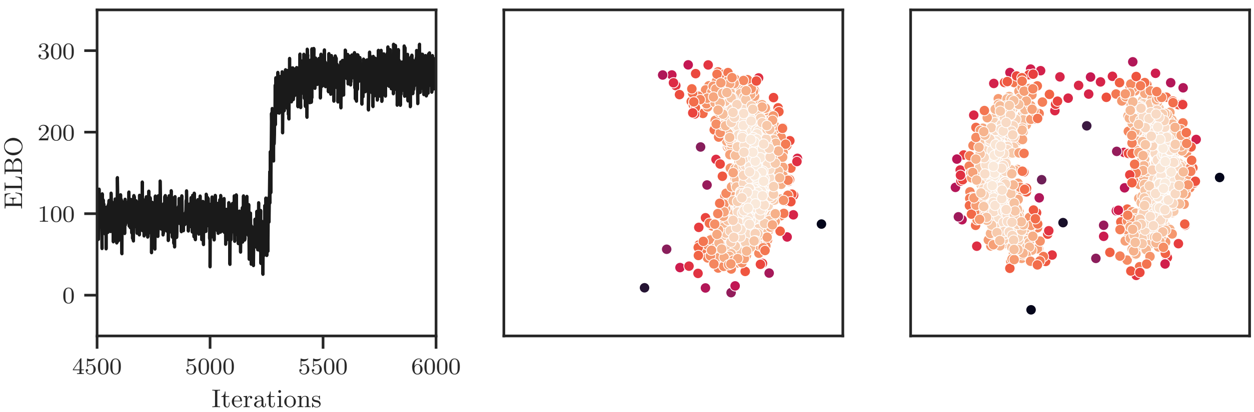

FIGURE 6. Training curve for a normalizing flow with 4 planar flows. When variational inference is performed on this bi-modal target distribution, an 'elbow' forms in the ELBO as the variational approximation changes shape to 'snap' to the new mode[7].

FIGURE 7. Probabilistic graphical model for Poisson regression problem.

low. This is why it takes some time for the second mode to be fitted — the space between the two modes has a low $p(z)$, leading to a penalization on the ELBO for approximations which traverse this space. Indeed, it can be seen in **??** that in iteration 10000, the variational approximation deviates from a uni-modal approximation, and this iteration corresponds to the area in the training curve in fig. 5 where the ELBO drops, before in the second mode is captured and the ELBO rises again in iteration 11000.

FIGURE 8. Posterior predictive distribution $t_*|\mathbf{t}$

FIGURE 9. Probabilistic graphical model for eight schools problem.

TABLE 4. ELBO and $\hat{k}$-statistic[12] for different variational inference algorithms trained on the eight schools test problem. It can be seen that there is a loosely monotonic relationship between the ELBO and the $\hat{k}$-statistic.

		ELBO	$\hat{k}$
Mean-field		-33.406	0.8710
Full-rank		-32.598	0.8226
Planar	4	-33.297	0.8329
	8	-32.728	0.8066
	16	-32.412	0.7700
	32	-31.791	0.6379
Radial	4	-37.756	0.9209
	8	-37.683	0.8979
	16	-35.904	0.8760
	32	-33.292	0.8203

FIGURE 10. Probabilistic graphical model for non-negative matrix factorization.

5.3. **Poisson regression.**

5.4. **Eight schools.**

The eight schools problem[8] is a classic test problem in Bayesian statistics[9, 10].

The results in table 4 were generated for different variational approximations for 10k iterations, an ADAM [11] learning rate of 10^{-2}, and 256 Monte Carlo ELBO gradient estimation samples.

A normalizing flow variational approximation with 32 planar flows which was trained for 1000 iterations achieved a final ELBO of -32.746 and a $\hat{k}$-statistic of 0.7723, which means this approximation performs worse than a fully-converged full-rank family. This suggests that a more complex model whose training is terminated before convergence generally performs worse than a simpler model which has been trained to convergence.

5.5. **Non-negative matrix factorization.**

5.6. **Radon.**

5.7. **Bayesian neural network.**

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.

References

[1] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538 (pages 1, 6).

[2] George Papamakarios et al. "Normalizing flows for probabilistic modeling and inference". In: *arXiv preprint arXiv:1912.02762* (2019) (page 1).

[3] Rianne van den Berg et al. *Sylvester Normalizing Flows for Variational Inference*. 2019. arXiv: `1803.05649 [stat.ML]` (page 5).

[4] Durk P Kingma et al. "Improved variational inference with inverse autoregressive flow". In: *Advances in neural information processing systems* 29 (2016), pp. 4743–4751 (page 5).

[5] Eli Bingham et al. "Pyro: Deep universal probabilistic programming". In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 973–978 (page 5).

[6] Ravin Kumar et al. "ArviZ a unified library for exploratory analysis of Bayesian models in Python". In: *Journal of Open Source Software* 4.33 (2019), p. 1143 (page 5).

[7] 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 (page 8).

[8] Donald B Rubin. "Estimation in parallel randomized experiments". In: *Journal of Educational Statistics* 6.4 (1981), pp. 377–401 (page 9).

[9] Andrew Gelman et al. *Bayesian Data Analysis*. CRC Press, 2013 (page 9).

[10] Bob Carpenter et al. "Stan: A probabilistic programming language". In: *Journal of statistical software* 76.1 (2017), pp. 1–32 (page 9).

[11] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (page 9).

[12] Yuling Yao et al. "Yes, but did it work?: Evaluating variational inference". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5581–5590 (page 9).

Appendix A. Appendix

A.1. **The Reparameterization 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 $z_0 \sim q_0$ as

$$z_0 = \mu + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1)$$

where