Fat-Tailed Variational Inference

April 20, 2021

Given a probabilistic model $p(x, y)$, the goal of automatic variational inference is to (1) construct a variational family of distributions $\{q_\phi\}_{\phi \in \Phi}$ and (2) search for a good variational approximation to the posterior $q_\phi \approx p(x \mid y)$. We focus on (1) and consider variational families specifically constructed to include fat-tailed members. Our contributions:

1. We rectify the failure of ADVI against fat tailed target densities by extending it to utilize Student-T variational family, and extend its applicability by transforming the density through a normalizing flow analogous to (Stefan Webb and Goodman, 2019). Our results on a location-scale representation for a StudentT shows that (1) allowing users to specify a tail-index for variational approximators leads to better fits and (2) even without prior knowledge, estimating the tail coefficient during training also works.

2. We provide alternatives for estimation of the variational approximation's tail index, adapting ideas from classical tail index estimation (Hill (1975); Pickands III et al. (1975)) to the variational inference setting where posterior samples are expensive but joint log density evaluations are cheap

1 Background

Probabilistic machine learning problems are coveniently formalized using probabilistic programming languages (PPLs) which provide high level primitives for probabilistic modelling and inference. In order to enable more flexibile and precise control over variational inference methods, recent work has explored a design space for variational inference APIs ranging from limited but completely automated (ADVI, Kucukelbir et al. (2017)) to manually specified guide programs (WebPPL, Pyro). Our work here focuses on automated variational inference where we seek to automatically construct a reasonable variational family automatically after tracing the execution of the probabilistic program.

Fat-tailed distributions commonly arise in robust machine learning where a standard approach is to replace light-tailed noise distributions fat-tailed ones Tipping and Lawrence (2005). They are also commonly used as weakly informative prior distributions in Bayesian heirarchical models Gelman et al. (2006).

1.1 Related Work

The Box-Cox transform (Box and Cox, 1964) is a commonly used parametric power transform to address fat tails and make data more similar to a normal distribution. This is analogous to composing a power transform before the normalizing flow. Whereas Asar et al. (2017) fit this transform using goodness-of-fit tests against a normal, whereas our optimization signal is back-propagated from an ELBO against the target posterior.

The bulk of related work focuses on fat-tails arising from relaxing priors. Recent work in VAEs has extensively studied the impact of relaxing Gaussian assumptions to heavier-tailed distributions. (Mathieu et al., 2019) consider a StudentT prior distribution $p(z)$ over the latent code z in a VAE with Gaussian encoder $q(z \mid x)$ [1], showing that the anisotropy of a StudentT product distribution (compared to the standard choice of Normal prior) leads to more disentangled representations. Chen et al. (2020) perform a similar modification except in a coupled VAE Cao et al. (2019) and showed improvements in the marginal likelihoods of reconstructed images. In addition, Boenninghoff et al. (2020) consider a mixture of StudentTs for the prior $p(z)$. Finally, (Feynman: maybe remove this? it was rejected from ICLR), Abiri and Ohlsson (2020) considered both StudentT prior and variational approximation family and showed improvements in SSIM (structural similarity) score of between original and reconstructed images. To position our work in context, note that the VAE's encoder $q(z \mid x)$ may be viewed as a variational approximation to the posterior $p(z \mid x)$ defined by the decoder model $p(x \mid z)$ and the prior $p(z)$. Our work differs from (Mathieu et al., 2019; Chen et al., 2020; Boenninghoff et al., 2020) in that we consider heavy-tailed variational approximations $q(z \mid x)$ rather than priors $p(z)$, and although (Abiri and Ohlsson, 2020) also considers a StudentT approximate posterior our work (1) considers a more general variational family comprised of flow transforms of StudentTs and (2) conducts a more thorough investigation across a broader range of models beyond a VAE on FashionMNIST.

Relaxation of priors to heavy-tailed distributions has numerous applications beyond VAEs. In Silnova et al. (2018), the authors perform inference in heavy-tailed probabilistic linear discriminant analysis using Gaussian mean-field variational inference and show improved accuracy in speaker identification. Our work is complementary to these approaches; whereas they consider heavy-tailed priors $p(z)$ we consider heavy-tailed variational families $q(z \mid x)$.

More directly comparable recent work on fat-tailed variational families Ding et al. (2011); Futami et al. (2017) studies the t-exponential family variational approximation (which includes Student-Ts and other heavier-tailed) includes heavy-tailed variational families, but critically do not discuss selection of the parameter t (which is deterministically to the Student-T's DoF v). Other differences include their derivation of expectation propagation update equations whereas we directly backprop a noisy ELBO estimate, and our broader variational family which includes flow transforms.

Most related to our work is Jaini et al. (2020), which shows similar impossibility results with Gaussian tailed base distributions and consider flow transforms of StudentT

[1] `https://github.com/iffsid/disentangling-disentanglement/blob/3396d40f46c34dd928a1241f567a86276b0ff41b/src/main.py#L52`

base distributions. However, our tail adaptive flows (1) do not mandate equal tail index across all dimensions, (2) composes techniques from Kucukelbir et al. (2017) to address target distributions with constrained support, (3) are investigated within the setting of density estimatino ($\propto KL(p, q)$, test set likelihood) as variational inference ($KL(q, p)$, ELBO, since p is a posterior hard to sample), and (4) are automatically generated as part of a probabilistic programming engine.

2 Fat Tailed Variational Inference

2.1 Distribution classes

While there is significant attention on parameterizing the transport map, there is comparatively little interest in considering the distribution p of the latent variable Z. Because the vast majority of transport map parameterizations are Lipschitz continuous, we will find that this choice of p can have a profound impact on the quality of the approximation of the tails of the target distribution. Here, we shall consider the following three main types of tail behaviour.

1. The class $\mathcal{G}$ of subgaussian random vectors X satisfying

$$\limsup_{r\to\infty} r^{-2} \log \mathbb{P}(\|X\| > r) < 0.$$

2. The class $\mathcal{E}$ of subexponential random vectors X satisfying

$$\limsup_{r\to\infty} r^{-1} \log \mathbb{P}(\|X\| > r) < 0.$$

3. The class $\mathcal{P}_\nu$ of random vectors X with tails no heavier than a ν-power law, that is,

$$\limsup_{r\to\infty} \log \mathbb{P}(\|X\| > r)/\log r < -\nu.$$

It is more convenient to characterize these distribution classes according to the concentration function α_X defined for a random vector X by

$$\alpha_X(r) = \sup_{A\,:\,\mathbb{P}(X\in A)\geq 1/2} \mathbb{P}(\mathrm{dist}(X, A) > r),$$

where $\mathrm{dist}(x, A) = \inf_{y\in A} \|x - y\|$. Equivalent characterizations of $\mathcal{G}, \mathcal{E}, \mathcal{P}_\nu$ follow by replacing $\mathbb{P}(\|X\| > r)$ with $\alpha_X(r)$. For example, a random variable $X \in \mathcal{G}$ if and only if $\limsup_{r\to\infty} r^{-2}\alpha_X(r) < 0$. Under this observation, the following lemma is an immediate consequence of (Ledoux, 2001, Proposition 1.3).

Lemma 1. *The classes $\mathcal{G}$, $\mathcal{E}$, and $\mathcal{P}_\nu$ for any $\nu > 0$, are closed under Lipschitz transformations. In other words, if f is Lipschitz, then for any $X \in \mathcal{G}$ (resp. $\mathcal{E}$, $\mathcal{P}_\nu$), $f(X) \in \mathcal{G}$ (resp. $\mathcal{E}$, $\mathcal{P}_\nu$).*

In other words, using only transport map approximators based on Lipschitz transformations, to achieve an accurate approximation of the tails of the target distribution, one must ensure that the reference distribution exhibits the same tail behaviour. Density estimators using a base distribution of incorrect class will inevitably fail at approximating slower tail decay. Note that this is not in violation of universal approximation theory since $\mathcal{G}$ is dense in L^2 (and therefore in $\mathcal{E}$ and $\bigcup_{\nu>0} \mathcal{P}_\nu$).

Two noteworthy choices for p in the literature are the standard Gaussian distribution, as seen in Kingma et al. (2016), and the logistic distribution, as seen in Dinh et al. (2014) (recommended because it "tends to provide a better behaved gradient"). Indeed, in the former case, only subgaussian approximations are available. In the latter case, since the logistic distribution is subexponential and not subgaussian, subexponential approximations are available. However, neither of these choices are effective for distributions which exhibit power-law tails, such as the Student t-distribution.

(Feynman: From 1/29/2021:

New theory? Extend multivariate reuslts of Jaini et al. (2020) to invalidate shared ν, eg product of two studentT with very different DoF. Interaction between ADVI transformed distributions and StudentTs.

)

2.2 Fat tails and tail indices

(Feynman: Reconcile with distribution classes)

Definition 2.1 (Resnick (2007))**.** A random variable X has *fat tails* with tail index $\alpha > 0$ if $\Pr[X > x] \sim x^{-\alpha}$ as $x \to \infty$.

Variance undefined if $\alpha < 3$, generally all moments $> \alpha - 1$ are infinite. α is called the *tail index*.

Can relax power law to regularly/slowly varying [2], which permits deviations without affecting tail exponent.

Pareto: $\Pr[X > x] = \left(\frac{x_m}{x}\right)^\alpha \sim x^{-\alpha}$

Student T [3]: $\Pr[X > x] \sim x \left(\frac{x^2}{\nu}\right)^{-\frac{1}{2}} + \left(\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \sim x^{-\nu}$.

Cauchy: $\Pr[X > x] \sim -\arctan\left(\frac{x-x_0}{\gamma}\right) \sim x^{-1}$, which makes since because standard Cauchy is StudentT($\nu = 1$).

Stable distributions? X is stable if $\forall a, b > 0$, $aX_1 + bX_2 = cX + d$ for some $c > 0$, d where X_1, X_2 independent copies of X (closure under convolution for fixed α). This includes normal, Cauchy, and Levy. Four parameter family with stability parameter α controlling the asymptotic behavior (tail index?). $\alpha = 2$ gives a Guassian, and $\alpha < 2$ gives $p(x) \sim |x|^{-(\alpha+1)}$.

(Feynman: Reconcile definitions with Jaini et al. (2020))

[2] https://journals.aps.org/prresearch/abstract/10.1103/PhysRevResearch.1.033034

[3] https://math.stackexchange.com/questions/3092190/asymptotics-of-hypergeometric-2f-1abcz-for-large-z-to-ir

2.3 Flows in variational inference

Universal approximation theorems about flows are not violated; they are asymptotic so finite realizations will still have support $\mathbb{R}$ and Lipschitz continuity.

IMPORTANT: support $\mathbb{R}$ makes them unsuitable for VI's $KL(q, p)$, need Kucukelbir et al. (2017) transforms

2.4 Failure modes on fat-tailed target densities

One major limitation of $p = N(0, I)$ is:

Theorem 1 (Chapter 2 Wainwright (2019))**.** *Let $(X_i)_1^n$ be a vector of iid $N(0, 1)$ RVs, $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz. Then $f(X) - \mathbb{E}f(X)$ is L-sub-Gaussian.*

In particular, density estimators using a Gaussian base distribution p will inevitably fail at approximating slower tail decay.

When the target density is a fat-tailed Cauchy distribution as in fig. 1, variational inference using flow-transformed Gaussian base distributions Stefan Webb and Goodman (2019) (orange in left) result in tails which decay inappropriately fast as measured using a Kolmogorov-Smirnov goodness-of-fit test (right). In contrast, the learned flow-transformed Student-T base distribution (green left) provides a much better approximation of the tail behavior.

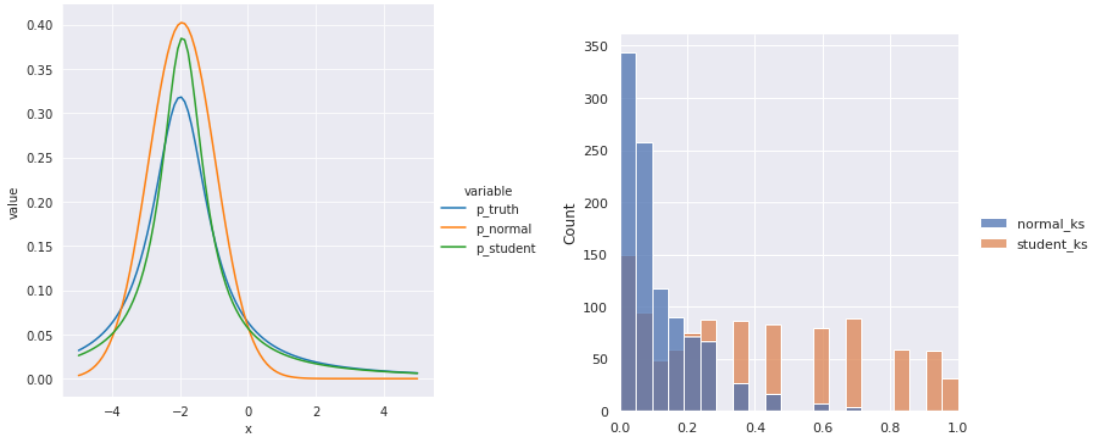


Figure 1: VI against a Cauchy target, PDFs (left) and p-value of Kolmogorov-Smirnov test statistics (right, ≤ 0.05 suggests poor approximation).

2.4.1 Analytical tail index algebra

Restrict to StudentT or Pareto.

Sum, PDFs are convolved, max of coefficients?

Product,

Ratio

Max

(Feynman: Method of steepest descent / saddle point approximations to approximate asymptotics of integrals, so given the PDF (eg from a PP) in theory can work out the power law exponent)

2.5 A flexibile fat-tailed variational family

We consider inverse autoregressive flow Kingma et al. (2016) transforms of heavy-tailed base distributions for variational inference, and investigate the effects of directly learning the tail index versus applying asymptotic approximations.

Lemma 1. *Inverse autoregressive flows Kingma et al. (2016) are Lipschitz continuous (because they are compositions of saturating non-linearities and matrix-vector products)*

To avoid the pitfalls of section 2.4, we see from lemma 1 that the root cause was closure of subgaussian distributions under Lipschitz maps. Accordingly, the two modifications we can consider relaxing are (1) Lipschitz-continuity of the transport map or (2) the distribution class of the base distribution.

2.6 Relaxing to non-Lipschitz transport maps

Classical statistical practice is to apply a power transform in an effort to make the data closer to a normal distribution. As power transforms are not Lipschitz continuous on the real line, this corresponds to a relaxation of assumption (1). We operationalize this idea by precomposing a ??? transform onto the transport map and propose learning the transform parameter through stochastic optimization of ELBO (c.f. estimating the Box-Cox parameter using goodness-of-fit against normal in Asar et al. (2017)).

(Feynman: Need to log transform a Normal, powers of normal are still sub-exponential)

(Feynman: Experiment with this)

2.7 Relaxing the base distribution

Complementary to relaxing Lipschitz continuity, another workaround for lemma 1 is to change the base distribution class. In our work, we restrict attention to fat-tailed distributions with tails no heavier than a ν-power law $\mathcal{P}_\nu$.

Selection of the tail index parameter ν is a well studied problem with classical solutions (Hill, 1975; Pickands III et al., 1975) using order statistics of samples. We consider:

- Defining APIs to allow researchers to succinctly specify fat-tailness of latent variables and their automatic variational inference semantics
- Estimating the tail index from samples (either provided or obtained from MCMC)
- Estimating the tail index of the posterior $p(x \mid y)$ using the log joint density $p(x, y)$, which is easy to evaluate and (given samples) approximately marginalize
- Sample-free estimation by including the tail index as a variational parameter and optimizing ELBO
- (Feynman: TODO: tail-coefficient algebra)

2.7.1 Fine-grained control over tail indices of variational approximations

decorator to instruct mean-field guide with fixed tail index of 2.0.

2.7.2 Estimating tail index without samples

Owing to the difficulty of generating posterior samples, techniques for estimating the tail index from samples (e.g. the Hill estimator) are not directly applicable to inference. As a result, it is of interest to be able to estimate a tail index given only access to an unnormalized log density.

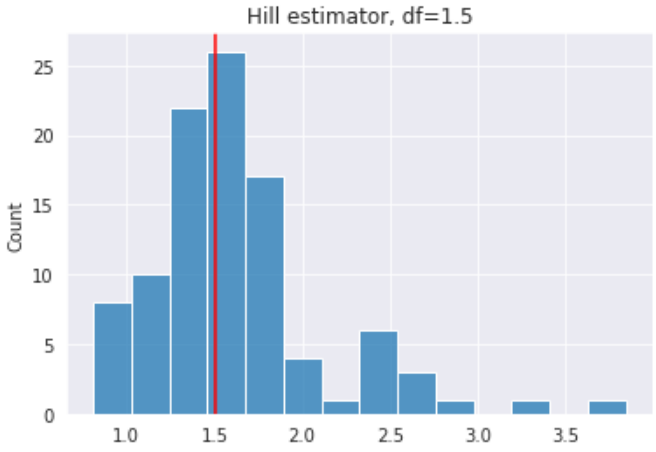


Figure 2: Hill estimator against $\nu = 1.5$ StudentT

Directly optimizing ν in the variational ELBO results in a trade-off between accurate tail behaviour and matching the distribution around the mode.

For x, y large:

$$\begin{aligned}
\Pr[X > x] &\sim x^{-\alpha} \\
p(x) &\sim \alpha x^{-\alpha-1} \\
\log p(x) &\sim \log \alpha - (\alpha+1) \log x \\
\log \frac{p(x)}{p(y)} &\sim (\alpha+1) \log \frac{y}{x} \\
\alpha &\sim \frac{\log p(x) - \log p(y)}{\log y - \log x} - 1
\end{aligned}$$

Note that p can be unnormalized as the partition function is cancelled out. For example. setting $x = 10$ and $y = 20$ yields an estimate of 1.4799 for a $\nu = 1.5$ StudentT.

(Feynman: Consider Richardson extrapolation to accelerate this limit. Rate of convergece in limit must be a power)

2.7.3 Approximate marginalization

(Feynman: Is this feasible for PP?)

Problem: Joint distributions (e.g. location-scale mixtures) require marginalization before we can get $p(x)$.

Our Method: Approximate with discrete mixture

$$p(x) = \int p(x|z)p(z)dz \approx \frac{1}{N}\sum_{i=1}^{N} p(x \mid z_i) \quad \text{where } z_i \overset{\text{iid}}{\sim} p(z)$$

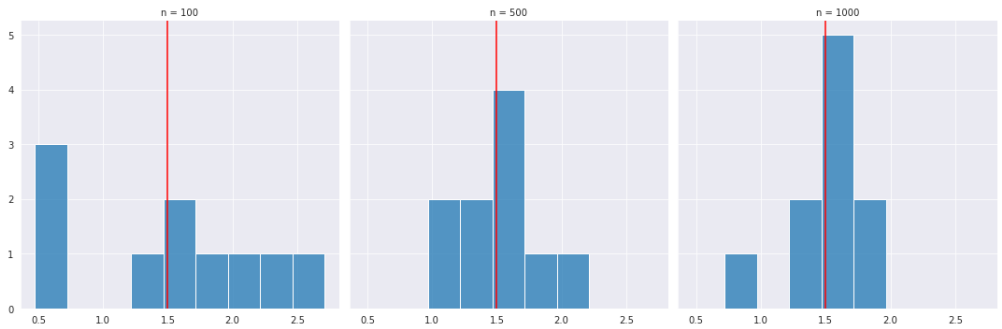


Figure 3: Tail index estimation on location-scale mixture representation for StudentT, where the mixture is discretized to n components.

3 Experiments

These experiments investigate the behavior of neural density estimators with *heavy-tailed base distribution*. Specifically, we consider a masked autoregressive flow Papamakarios et al. (2017) transform of a generalized Student's t distribution as a density estimator $q_\theta(X)$ in a variational inference framework. To fit q_θ to a target distribution π, the ELBO gradient is reparameterized and Monte-Carlo approximated

$$\begin{aligned}
\nabla_\theta \mathbb{E}_{q_\theta} \log \frac{\pi(X)}{q_\theta(X)} &= \nabla_\theta \mathbb{E}_p \log \frac{\pi(X)}{p_\theta(f_\theta^{-1}(X))\left|\det \nabla f_\theta^{-1}(X)\right|} \\
&= \mathbb{E}_p \nabla_\theta \log \frac{\pi(X)}{p_\theta(f_\theta^{-1}(X))\left|\det \nabla f_\theta^{-1}(X)\right|} \\
&\approx \frac{1}{n}\sum_{i}^{n} \nabla_\theta \log \frac{\pi(x_i)}{p_\theta(f_\theta^{-1}(x_i))\left|\det \nabla f_\theta^{-1}(x_i)\right|}
\end{aligned}$$

(Feynman: Careful, $D_{KL}(N(0,1), \text{Cauchy}(0,1) \approx 0.2592 < \infty = D_{KL}(\text{Cauchy}(0,1), \text{N}(0,1)))$

3.1 Applications in PPL inference

The experiments in this section are conducted using the `beanmachine` PPL, where inference is conducted following a Metropolis-within-Gibbs framework.

3.2 Normal-normal location mixture

We first consider a Normal-Normal conjugate inference problem where the posterior is known to be a Normal distribution as well. Figure 4 shows the resulting density approximation, which can be seen to be reasonable for both a Normal base distribution (the "correct" one) and a StudentT base distribution. This suggests that mis-specification (i.e. heavier tails in the base distribution than the target) may not be too problematic.

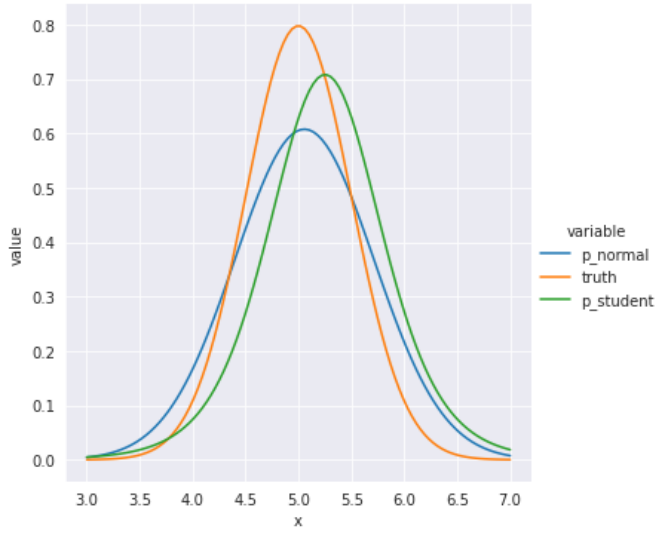


Figure 4: VI against a Normal posterior

3.3 StudentT scale mixture representation

We next consider a heavy-tailed posterior target density by using a scale-mixture representation for Student's t. Specifically, if $v \sim \chi_2(\nu)$ and $y \mid v \sim N(0, v)$ then the marginal distribution of y is StudentT with ν degrees of freedom. In this experiment, we also allow the degrees of freedom for the base StudentT distribution to be optimized as well in `p_student_df_vi` and set the degrees of freedom equal to the true ν in `p_student_vi`. While using the true ν does yield an almost exact fit, optimizing ν is more practical.

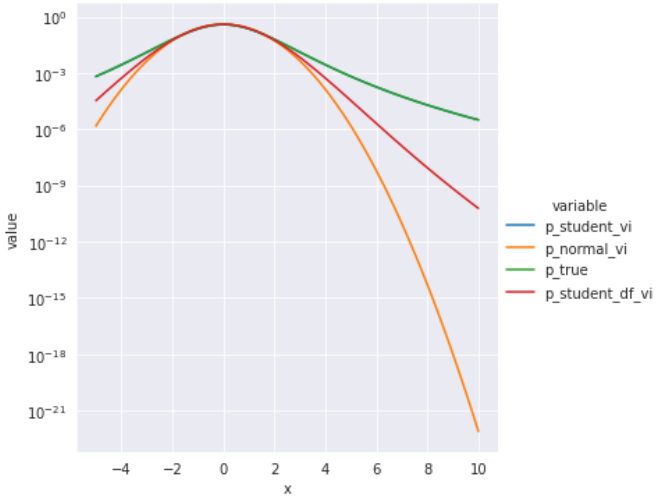


Figure 5: VI against StudentT represented by a Normal scale mixture

3.4 Bayesian Robust Regression

$n = 100$, $d = 10$.

$X_{ij} \overset{\text{iid}}{\sim} N(0, 1)$ for $i \in [n]$, $j \in [d]$.

$y_i \overset{\text{iid}}{\sim} \text{StudentT}(\text{loc} = X\beta, df = 5)$

Improper "flat" prior on β to ensure heavy-tailed posterior.

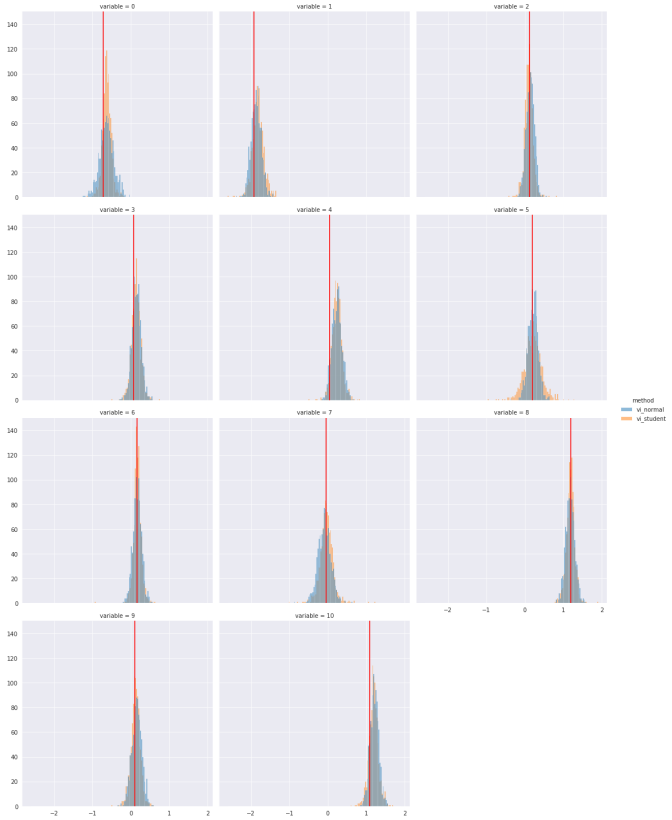
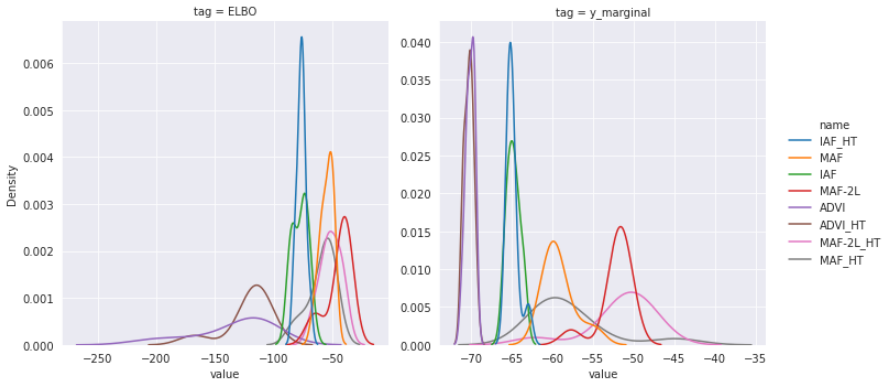


Figure 6: VI of Bayesian robust linear regression

(Feynman: Can work out the exponent with asymptotic approximations; compare?)

3.5 Eight schools



	ELBO	$\log P(y)$
ADVI	-193.86 ± 33.50	-70.11 ± 0.52
ADVI-HT	-121.54 ± 19.59	-70.29 ± 0.54
MAF	-55.06 ± 5.46	-59.14 ± 1.99
MAF-HT	-59.63 ± 6.74	-57.84 ± 4.97
MAF-2L	-45.01 ± 11.02	-52.19 ± 2.06
MAF-2L-HT	-51.67 ± 8.72	-51.53 ± 4.23
IAF	-77.67 ± 6.47	-64.79 ± 0.80
IAF-HT	-76.67 ± 3.71	-64.91 ± 0.76

Figure 7: Final ELBO and (MC estimate of) log marginal $\log P(y)$ after 5000 steps

3.6 Importance weights

When the importance sampling density is more peaked than the target density.

Wang et al. (2018) example 3.1: Let $p = N(0, \sigma_p^2)$, $q = N(0, \sigma_q^2)$, and for $x \sim q$ let $w(x) = \frac{p(x)}{q(x)}$. If $\sigma_p > \sigma_q$, then w has tail index $\frac{\sigma_p^2}{\sigma_p^2 - \sigma_q^2}$. Otherwise, w is not fat-tailed. (Feynman: TODO)

3.7 Some HTVAE example reconstructions and likelihoods

(Feynman: TODO. From Sohn et al. (2015), MNIST, Caltech-UCSD Birds, and Labeled Faces in the Wild.)

4 Single-Site Normalizing Flow Variational Inference

Let $\{q_\phi\}_{\phi \in \Phi}$ be a parameterized (variational) family. Recall the variational ELBO

$$\log p(y) = \log \int dx \frac{q_\phi(x)}{q_\phi(x)} p(x, y) \geq \int dx q_\phi(x) \log \frac{p(x, y)}{q_\phi(x)}$$

The variational family's PDF $q_\phi(x)$ is assumed to be tractable, and the joint PDF $p(x, y)$ may be obtained by running the probabilistic program forwards. Finally, the Monte-Carlo method enables tractable approximation of the integral:

$$\int dx q_\phi(x) \log \frac{p(x,y)}{q_\phi(x)} \approx \log \frac{p(x,y)}{q_\phi(x)} \qquad x \sim q_\phi$$

making the ELBO tractable.

Assume there are d latent variables $x = (x_i)_{i=1}^d$. Single-site variational inference (i.e. CaVI, coordinate-ascent variational inference) sequentially iterates over $i \in [d]$, performing Monte-Carlo

$$\phi \leftarrow \phi + \alpha_t \nabla_\phi \mathbb{E}_{\setminus\{i\}} \mathbb{E}_i q_\phi(x) \log \frac{p(x,y)}{q_\phi(x)}$$

Finish description, vary the number of MC samples between two expectations

4.1 Cached Inverses

The Monte-Carlo ELBO

$$\sum_i \log p(x_i, y) - \log q(x_i) \qquad x_i \sim q$$

requires both sampling and evaluating log densities with respect to q. When q is taken to be a distribution parameterized by a normalizing flow $q(x) = |\frac{df}{dx}| p(f^{-1}(x))$, log density evaluations require evaluating the inverse flow f^{-1} whereas sampling involves pushing forward samples $x \sim p$ along the flow's forward direction.

Modern normalizing flow architectures typically are much more expensive to compute in one direction (c.f. IAF vs MAF), with some flows (NAF, SoS, ResidualFlows) requiring an optimization to approximate the inverse flows. To circumvent this problem, we identify a critical optimization for VI that caching can be used to avoid flow inversion during Monte-Carlo training. That is, by caching pairs $(x_i, y_i = f(x_i))$ sampled during MC ELBO, the entropy term $\log q(y_i) = \log|\frac{df}{dx}(x_i)| + \log p(x_i)$ can be evaluated without explicit inversion.

Run-time comparisons for ELBO with/without caching

4.2 NaMI structured conditional flows

The NaMI algorithm **?** can be used for model inversion, enabling single-site flow distributions which depend on the values of adjacent nodes (as well as amortized variational inference where the variational families $q_{(\phi,y)}(x)$ depend on the conditioned observations y.

5 Meeting Notes

1. We added a knob for heavy tail, how do we make fair comparison? Increase the flexibility of the flow for non-Gaussian.
2. Neural network weights heavy tailed; can we learn VI with heavy-tailed distributions.

6 Discussion

Limitations: only considered symmetric base distributions, could also consider skewness (Gupta, 2003).

References

Abiri, N. and Ohlsson, M. (2020). Variational auto-encoders with student's t-prior. *arXiv preprint arXiv:2004.02581*.

Asar, Ö., Ilk, O., and Dag, O. (2017). Estimating box-cox power transformation parameter via goodness-of-fit tests. *Communications in Statistics-Simulation and Computation*, 46(1):91–105.

Boenninghoff, B., Zeiler, S., Nickel, R. M., and Kolossa, D. (2020). Variational autoencoder with embedded student-t mixture model for authorship attribution. *arXiv preprint arXiv:2005.13930*.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Cao, S., Li, J., Nelson, K. P., and Kon, M. A. (2019). Coupled vae: Improved accuracy and robustness of a variational autoencoder. *arXiv preprint arXiv:1906.00536*.

Chen, K. R., Svoboda, D., and Nelson, K. P. (2020). Use of student's t-distribution for the latent layer in a coupled variational autoencoder. *arXiv preprint arXiv:2011.10879*.

Ding, N., Qi, Y., and Vishwanathan, S. (2011). t-divergence based approximate inference. *Advances in Neural Information Processing Systems*, 24:1494–1502.

Dinh, L., Krueger, D., and Bengio, Y. (2014). Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Futami, F., Sato, I., and Sugiyama, M. (2017). Expectation propagation for t-exponential family using q-algebra. In *Advances in Neural Information Processing Systems*, pages 2245–2254.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Gupta, A. (2003). Multivariate skew t-distribution. *Statistics: A Journal of Theoretical and Applied Statistics*, 37(4):359–363.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174.

Jaini, P., Kobyzev, I., Yu, Y., and Brubaker, M. (2020). Tails of lipschitz triangular flows. In *International Conference on Machine Learning*, pages 4673–4681. PMLR.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

Ledoux, M. (2001). *The concentration of measure phenomenon*. Number 89. American Mathematical Soc.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. (2019). Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347.

Pickands III, J. et al. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, 3(1):119–131.

Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.

Silnova, A., Brummer, N., Garcia-Romero, D., Snyder, D., and Burget, L. (2018). Fast variational bayes for heavy-tailed plda applied to i-vectors and x-vectors. *arXiv preprint arXiv:1803.09153*.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491.

Stefan Webb, J. P. Chen, M. J. and Goodman, N. (2019). Improving automated variational inference with normalizing flows. *6th ICML Workshop on Automated Machine Learning (AutoML)*.

Tipping, M. E. and Lawrence, N. D. (2005). Variational inference for student-t models: Robust bayesian interpolation and generalised component analysis. *Neurocomputing*, 69(1-3):123–141.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, D., Liu, H., and Liu, Q. (2018). Variational inference with tail-adaptive f-divergence. *Advances in Neural Information Processing Systems*, 31:5737–5747.