Fat–Tailed Variational Inference with Anisotropic Tail Adaptive Flows

Feynman Liang
Department of Statistics
University of California, Berkeley
feynman@berkeley.edu

Liam Hodgkinson
Department of Statistics
University of California, Berkeley
liam.hodgkinson@berkeley.edu

Michael Mahoney
Department of Statistics
University of California, Berkeley
mmahoney@stat.berkeley.edu

Abstract

While fat-tailed densities commonly arise as posterior and marginal distributions in robust models and scale mixtures, they present a problematic scenario where the standard Gaussian variational family incurs exponential error in approximation of tail probabilities. We develop a theory for heavy-tails in high dimensions in order to rigorously characterize the limitations of variational inference with normalizing flow transformations of Gaussian base distributions. To mitigate these limitations and enable density modeling and variational inference of fat-tailed targets, we propose anisotropic tail-adaptive flows (ATAF). In particular, our theory sharpens previous work from tail-isotropic elliptical distributions to multivariate fat-tailed distributions with heterogeneous tails and is especially relevant to commonly encountered product distributions arising from grouping/concatenating random variables (e.g. blocked Metropolis-Hastings, Hamiltonian Monte Carlo state). Experimental results confirm ATAF's ability to more accurately approximate targets which exhibit fat-tails and tail-anisotropy, and ablation studies on two Bayesian posterior variational inference tasks suggests ATAF readily composes with other recent advancements to set a new state-of-the-art method for performing VI in the presence of fat tails.

1 Introduction

Flow based methods have proven to be effective techniques to model complex probability densities and compete with the state of the art on density estimation [16, 10, 18], generative modelling [6, 20], and black-box variational inference [21, 2] tasks. These methods start with a random variable X with some simple and tractable distribution μ and apply a learnable transport map f^θ to build another random variable $Y = f^\theta(X)$ with a more expressive pushforward measure $f^\theta_*\mu$. In contrast to the implicit distributions [17] produced by generative adversarial networks (GANs), flow based methods restrict the transport map f^θ to (1) be invertible and (2) have efficiently computable Jacobian determinants. As a result, probability density functions can be tractably computed through direct application of change of variables:

$$p_Y(y) = p_X((f^\theta)^{-1}(y)) \left| \det \left. \frac{d(f^\theta)^{-1}(z)}{dz} \right|_{z=y} \right|$$

Preprint. Under review.

While recent developments have considerably expanded the range of architectures and models used for the transport map f^θ, the base distribution μ has received relatively less investigation. We believe this asymmetric focus is detrimental to the research community because the sensible default choice of Gaussian base distribution $\mu = \mathcal{N}(0, \boldsymbol{I})$ may result in significant limitations to the expressivity of the model (theorem 1). Addressing these shortcomings is an important aim for recent works [18], and our work here represents an additional advancement towards this goal.

In this work, we develop theory for fat-tailed random variables and their transformations under Lipschitz-continuous functions in order to characterize the effect of the base distribution μ in a range of flow-based models (table 1). For the multivariate setting, we sharpen the tail-isotropic theory from [18] by developing a notion of direction-dependent tail parameters and propose *anisotropic tail adaptive flows* (ATAF) in order to implement our additional insights. Our experimental results demonstrate the improvements ATAF provides over prior work, providing a new state-of-the-art for performing variational inference against fat-tailed target densities.

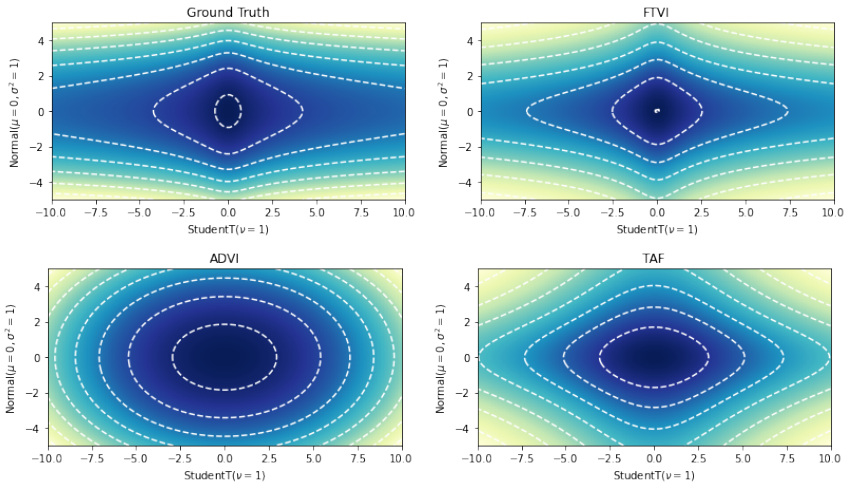


Figure 1: $\mathcal{Q}_{ATAF}$ (definition 4.1) improves over $\mathcal{Q}_{TAF}$ (definition 2.2) for modeling a tail-anisotropic StudentT$(1) \times N(0, 1)$ target. Left to right: target, ATAF, ADVI, TAF.
(Feynman: Play with bounds to see if we can make it look better, explain this figure in isolation (OK to repeat later content) so TLDR readers understand it (two directions, axis-aligned, level sets anistropic, order bad to better to best))

2 Background: flow based models for black-box variational inference

We first review the set up for variational inference and define the setting and notation used in the remainder of the paper. Let $\mathcal{Q} = \{q_\phi\}_{\phi \in \Phi}$ be a parameterized (variational) family. VI with KL-divergence optimizes a Monte-Carlo approximation of the variational ELBO

$$\text{ELBO}(\phi) = \int q_\phi(x) \log \frac{p(x, y)}{q_\phi(x)} dx \approx \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i, y)}{q_\phi(x_i)} \qquad x_i \overset{\text{iid}}{\sim} q_\phi$$

We consider the case where the target density $p(x \mid y)$ is fat-tailed. Such distributions commonly arises during a standard "robustification" approach where light-tailed noise distributions are replaced with fat-tailed ones [29]. They also appear when weakly informative prior distributions are used in Bayesian heirarchical models [12]. Of particular interest is the multivariate case where supp $p(x \mid y) \subset \mathbb{R}^d$, where we introduce a new notion of direction-dependent tail parameter (definition 3.2) which captures anisotropic axis-aligned tails which commonly arise in the process of concatenating or blocking together random variables within a probabilistic model.

In black-box variational inference (BBVI) [26], the variational family $\mathcal{Q}(\Phi)$ is automatically constructed without specific knowledge of the target density.

One of the earliest implementations of BBVI utilizes deterministic transformations of a multivariate normal and is known as automatic differentiation variational inference (ADVI) [22]:

Definition 2.1. ADVI's variational family is given by

$$\mathcal{Q}_{ADVI}(\Phi_{\text{Affine}}) \coloneqq \{(f \circ \Phi_{\text{Affine}})_* \mu\}$$

where $\mu = \text{Normal}(0_d, I_d)$, $\Phi_{\text{Affine}}(x) = Ax + b$ is an affine transform, and f is a pre-determined bijection between constrained supports [22].

Normalizing flow transforms improve variational inference **(author?)** [21, 28] by making the transport map a learnable bijection parameterized by neural networks. However, most major software packages implementing these methods (Pyro's `AutoNormalizingFlow/AutoIAFNormal`, Stan's `method=variational`, PyMC's `NormalizingFlowGroup`) As we will show (**??**), this potentially results in an exponentially bad density approximation in the tails.

To address this issue, **(author?)** [18] utilizes a single degrees-of-freedom $\nu \in \mathbb{R}$ across all dimensions.

Definition 2.2. Tail-adaptive flow (TAF) uses variational family

$$\mathcal{Q}_{TAF}(\nu, \Phi_{\text{NF}}) \coloneqq \{(\Phi_{\text{NF}})_* \mu_\nu\}$$

Importantly, in $\mu_\nu = \prod^d \text{StudentT}(\nu)$ the same ν is used across all dimensions.

2.1 StudentT scale mixture representation

To illustrate ATAF on a toy example, consider a heavy-tailed posterior target density by using a scale-mixture representation for StudentT. 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.

An example of this failure in the exponentially poor tail approximation in fig. 2. When the target density is a fat-tailed distribution, variational inference using flow-transformed Ggaussian base distributions [28] (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 StudentT base distribution (green left) provides a much better approximation of the tail behavior.

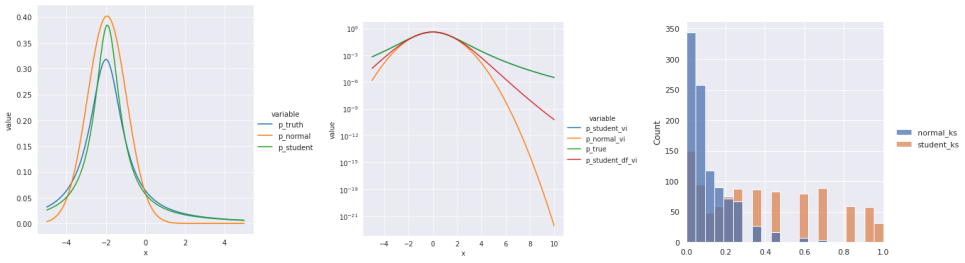


Figure 2: VI against a fat-tailed target, PDFs (left) and p-value of Kolmogorov-Smirnov test statistics (right, ≤ 0.05 suggests poor approximation).

(Feynman: 4-5 lines to explain this improvement of TAF over ADVI, state this is not new)

3 Tail behavior of Lipschitz flows

3.1 Fat-tailed distributions and tail indices

Definition 3.1 (Classification of tails)**.** We classify the tail behavior of a random variable X using its complement CDF $\mathbb{P}(X \geq x)$ and define:

Exponential-type $X \in \mathcal{E}^p_\alpha$ means $\mathbb{P}(|X| \geq x) = \Theta(e^{-\alpha x^p})$

Logarithmic-type $X \in \mathcal{L}^p_\alpha$ means $\mathbb{P}(|X| \geq x) = \Theta(e^{-\alpha (\log x)^p})$

Model	Autoregressive transform	Suff. conditions for Lipschitz
NICE[8]	$z_j + \mu_j \cdot \mathbb{1}_{k \notin [j]}$	μ_j Lipschitz
MAF[25]	$\sigma_j z_j + (1 - \sigma_j)\mu_j$	σ_j bounded
IAF[21]	$z_j \cdot \exp(\lambda_j) + \mu_j$	λ_j bounded, μ_j Lipschitz
Real-NVP[9]	$\exp(\lambda_j \cdot \mathbb{1}_{k \notin [j]}) \cdot z_j + \mu_j \cdot \mathbb{1}_{k \notin [j]}$	λ_j bounded, μ_j Lipschitz
Glow[20]	$\sigma_j \cdot z_j + \mu_j \cdot \mathbb{1}_{k \notin [j]}$	σ_j bounded, μ_j Lipschitz
NAF[16]	$\sigma^{-1}(w^\top \cdot \sigma(\sigma_j z_j + \mu_j))$	Always (logistic mixture CDF)
NSF[10]	$z_j \mathbb{1}_{z_j \notin [-B,B]} + M_j(z_j; z_{<j}) \mathbb{1}_{x_j \in [-B,B]}$	Always (identity outside $[-B, B]$)
Residual Flows[6]	n/a (not autoregressive)	Always (required for invertibility)

Table 1: Conditions for Lipschitz continuity of some popular / recently developed flows. $M(\cdot)$ denotes monotonic rational quadratic spline [10].

We call p the *class index* and α the *tail-parameter* for random variable X.

Notice every $\mathcal{E}_\alpha^p$ and $\mathcal{L}_\beta^q$ are disjoint, and furthermore $\mathcal{E}_\alpha^p \cap \mathcal{L}_\beta^q = \emptyset$ for all α, β, p, q.

For convenience, we define the ascending families $\overline{\mathcal{E}_\alpha^p}$ and $\overline{\mathcal{L}_\alpha^p}$ analogously as before except with $\Theta(\cdot)$ replaced by $\mathcal{O}(\cdot)$. Finally, we define the class of all exponential-type tails with class index at least p as $\overline{\mathcal{E}^p} = \cup_{\alpha \in \mathbb{R}_+} \overline{\mathcal{E}_\alpha^p}$.

Some commonly encountered families include $\alpha^{-1/2}$-sub-Gaussians $\overline{\mathcal{E}_\alpha^2}$ and sub-exponentials $\overline{\mathcal{E}_\alpha^1}$. Particularly relevant to this work are the power-laws $\mathcal{L}_\alpha^1$.

3.2 Failure of Light-Tailed and Elliptical Distribution

Many flow-based models exhibit Lipschitz continuity in their transport map either due to explicit enforcement [14, 6] or as a result choice of architecture and activation function (table 1). Our results on fat-tailed flows are developed within this context, which we encapsulate in the following assumption:

Assumption 1: f^θ is Lipschitz continuous (e.g. sufficient conditions in table 1 are satisfied) and invertible.

Theorem 1 (Lipschitz maps of tail classes)**.** *Under Assumption 1, the distribution classes $\mathcal{E}^p$ and $\overline{\mathcal{L}_\alpha^p}$ (with $p, \alpha \in \mathbb{R}_+$) are closed under every flow transformation in table 1.*

(Feynman: In english; light tailed base cannot get mapped to heavy-tailed target with a Lipschitz transport map) Note this does not violate the universality theorems (e.g. [16]) as they only apply in the infinite hidden unit limit.

Proof. Let X be a random variable from either $\mathcal{E}_\alpha^p$ or $\mathcal{L}_\alpha^p$. Its concentration function [23, Equation 1.6] is given by

$$\alpha_X(r) \coloneqq \sup\{\mu\{x : d(x, A) \geq r\}; A \subset \text{supp } X, \mu(A) \geq 1/2\} = \mathbb{P}(|X - m_X| \geq r)$$

Under Assumption 1, f^θ is Lipschitz (say with Lipschitz constant L) so by [23, Proposition 1.3]

$$\mathbb{P}(|f^\theta(X) - m_{f^\theta(X)}| \geq r) \leq 2\alpha_X(r/L) = \mathcal{O}(\alpha_X(r/L))$$

where $m_{f^\theta(X)}$ is a median of $f^\theta(X)$. Furthermore, by the triangle inequality

$$\begin{aligned}
\mathbb{P}(|f^\theta(X)| \geq r) &= \mathbb{P}(|f^\theta(X) - m_{f^\theta(X)} + m_{f^\theta(X)}| \geq r) \\
&\leq \mathbb{P}(|f^\theta(X) - m_{f^\theta(X)}| \geq r - |m_{f^\theta(X)}|) \\
&= \mathcal{O}(\mathbb{P}(|f^\theta(X) - m_{f^\theta(X)}| \geq r)) \\
&= \mathcal{O}(\alpha_X(r/L)) \qquad (1)
\end{aligned}$$

where the asymptotic equivalence holds because $|m_{f^\theta(X)}|$ is independent of r.

When $X \in \mathcal{E}^p_\alpha$, eq. (1) implies

$$\mathbb{P}(|f^\theta(X)| \geq r) = \mathcal{O}(e^{-\frac{\alpha}{L} r^p})$$
$$\therefore X \in \mathcal{E}^p_{\alpha/L}$$

from which we see that the Lipschitz transform of exponential-type tails continues to possess exponential-type tails with the same class index p, although the tail parameter may have changed which necessitates taking a union over tail parameters. Hence, $\mathcal{E}^p_\infty$ is closed under Lipschitz maps for each $p \in \mathbb{R}_{>0}$.

On the other hand, when $X \in \mathcal{L}^p_\alpha$ then eq. (1) means that

$$\begin{aligned}\mathbb{P}(|f^\theta(X)| \geq r) &= \mathcal{O}(e^{-\alpha(\log(r/L))^p}) \\ &= \mathcal{O}(e^{-\alpha(\log r)^p} e^{-\alpha(-\log L)^p}) \\ &= \mathcal{O}(e^{-\alpha(\log r)^p}) \\ \therefore X &\in \mathcal{L}^p_\alpha\end{aligned}$$

Unlike exponential-type tails, Lipschitz transforms of logarithmic-type tails not only remain logarithmic, but their tails decay no slower than a logarithmic-type tail of the same class index with the *same* tail parameter α. This upper bound suffices to show closure under Lipschitz maps for the ascending family $\overline{\mathcal{L}^p_\alpha}$. □

Corollary 2 (Heavy to light)**.** If in addition f^θ is smooth with no critical points on the interior or boundary of its domain, then $\mathcal{L}^p_\alpha$ is closed (note here the ascending union over tail paramters is absent).

(Feynman: Use English: this means that we cannot map a heavy-tailed base distribution to a light-tailed target)

Proof. Let f^θ be as before with the additional assumptions. Since f^θ is a smooth continuous bijection, it is a diffeomorphism. Furthermore, by assumption f^θ has invertible Jacobian on the closure of its domain hence $\sup_{x \in \text{dom } f^\theta} |(f^\theta)'(x)| \geq M > 0$. By the inverse function theorem, $(f^\theta)^{-1}$ exists and is a diffeomorphism with

$$\frac{d}{dx}(f^\theta)^{-1}(x) = \frac{1}{(f^\theta)'((f^\theta)^{-1}(x))} \leq \frac{1}{M}$$

Therefore, $(f^\theta)^{-1}$ is M^{-1}-Lipschitz and we may apply theorem 1 to conclude the desired result. □

In fact, we can show a stronger result of closure under finite-degree polynomial maps. This result extends previous impossibility theorems [18] from Lipschitz flows to also include polynomial flows such as SoS-flows [19].

Theorem 3 (Closure under polynomial maps)**.** *For all $\alpha, \beta, p, q \in \mathbb{R}_+$, there does not exist a finite-degree polynomial map from $\mathcal{E}^p_\alpha$ to $\mathcal{L}^q_\beta$.*

Proof. Let $X \in \mathcal{E}^p_\alpha$. By considering sufficiently large X such that leading powers dominate, it suffices to consider monomials $Y = X^k$. Notice $\mathbb{P}(Y \geq x) = \mathbb{P}(X \geq x^{1/k}) \asymp e^{-\alpha x^{p/k}} = \sum_l \frac{(-\alpha x^{lp/k})}{l!}$, so $Y \in \mathcal{E}^{p/k}_\alpha$. □

(Feynman: Same comment about english here; polynomial maps of exponential-types remain exponential-type)

Remark 4. There does not exist a inverse polynomial map (e.g. sqrt) from $\mathcal{L}_\alpha$ to $\mathcal{E}$.

As a result, **??** suggests that to avoid the pitfalls of section 3.2 either (1) the base distribution μ should be modified, or (2) the Lipschitz-continuity of Φ should be relaxed. Analogous to [18], we consider (1) by modifying the base distribution μ to define ATAF.

3.3 Tail parameters for multivariate fat-tails

In **(author?)** [18], a multivariate random variable X is defined to be heavy-tailed if $\|X\|_2$ is heavy-tailed and theory around tail parameters is developed for elliptically contoured multivariate distributions $X = \mu + RAU$ for some heavy-tailed random variable R, fixed vectors $\mu \in \mathbb{R}^d$ and $U \in S^{d-1}$, and A a (Cholesky factor) defining the ellipsoid axes.

While elliptically contoured multivariate distributions admit a straightforward generalization from the scalar case, they are severely limited in practical applications. One fundamental limitation of elliptically contoured X is that the tail parameter is the same for every 1-dimension projection. If R has tail index α then for all for $u \in S^{d-1}$

$$\mathbb{P}[\langle X, u\rangle \geq x] = \mathbb{P}[\langle \mu, u\rangle + R\langle AU, u\rangle \geq x] = \mathbb{P}[R \geq \frac{x - \langle \mu, u\rangle}{\langle u, AU\rangle}] \sim C_1 \left(\frac{x - \langle \mu, u\rangle}{\langle u, AU\rangle}\right)^{-\alpha} \sim C_2 x^{-\alpha}$$

Furthermore, the assumption of elliptically contoured distributions are easily violated in many common and useful applications. For example, in probabilistic programming collections of random variables are oftentimes grouped together into a single multivariate random variable (e.g. blocked Gibbs, Hamiltonian Monte Carlo). In fact, Q_{TAF} (used in the experiments in **(author?)** [18]) utilize a StudentT product base distribution which fig. 3 shows is not elliptically contoured.

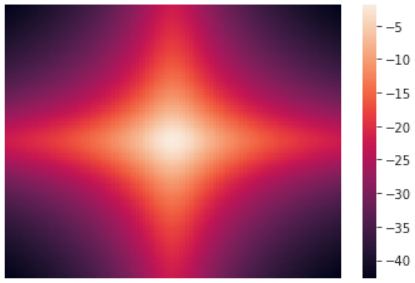


Figure 3: Illustrating multivariate fat-tails using a 2-dimensional StudentT. The multivariate theory from **(author?)** [18] is for elliptically contoured distributions and not applicable to the StudentT product base distributions used in their experiments.

(Feynman: Consider just using words here to delineate from our contributions, this is not a super important fact)

Here, we propose a more fine-grained definition for tail parameters in multivariate distributions. See fig. 4

Definition 3.2. For a d-dimensional random vector X, its *tail parameter function* $\alpha_X : \mathcal{S}^{d-1} \to \mathbb{R}_+$ is defined as $\alpha_X(v) = \lim_{x\to\infty} \frac{\log \mathbb{P}(\langle v, X\rangle \geq x)}{\log x}$ Intuitively, $\alpha_X(v)$ returns the index of the 1-dimensional projection $\langle v, X\rangle$.

Contrast this to **(author?)** [18], where X is characterized by the tail index of $\|X\|$. For $X = (t_1, \ldots, t_d)$ with $t_i \overset{\text{iid}}{\sim} \text{StudentT}(\nu_i)$

$$\mathbb{P}[\|X\|_2 \geq t] = \mathbb{P}[\sup_{v\in\mathcal{S}^{d-1}} \langle X, v\rangle \geq t] \geq \sup_{v\in\mathcal{S}^{d-1}} \mathbb{P}\langle X, v\rangle \geq t]$$

from which we see that the previous theory provides a coarse upper bound over all directions v.

Elliptical distributions are *tail isotropic* i.e. $\alpha(v) \equiv c$ is constant. The base distribution for **(author?)** [18], $\prod_1^d \text{StudentT}(\nu_i)$ with $\nu_i \equiv \nu$, is also tail isotropic because $\langle v, X\rangle$ is a sum of StudentT and by tail index algebra has index $\max_i \nu_i = \nu$.

Proposition 5 (Pushforwards of tail-isotropic distributions)**.** Let μ be tail isotropic with index ν. and suppose f^W is invertible and satisfies Assumption 1. Then $f^W_* \mu$ is tail isotropic with index ν.

Proof. Suppose μ is isotropic, so $\langle X, v\rangle$ has index ν for all $v \in \mathcal{S}^{d-1}$. Then for any affine autoregressive flow, $\langle v, f(X)\rangle$ is an affine combination of random variables with tail index ν hence by the tail index algebra has index ν. □

However, $\alpha(v)$ is non-trivial to work with; it is an asymptotic quantity and is defined for uncountably many $v \in S^{d-1}$. In this work, we propose approximating $\alpha(v)$ using the standard basis vectors:

Definition 3.3. The *standard basis tail parameters* of a fat-tailed $X \in \mathbb{R}^d$ is $\{\alpha(e_i) : i \in [d]\}$ where α is defined in definition 3.2 and e_i is the ith standard basis vector.

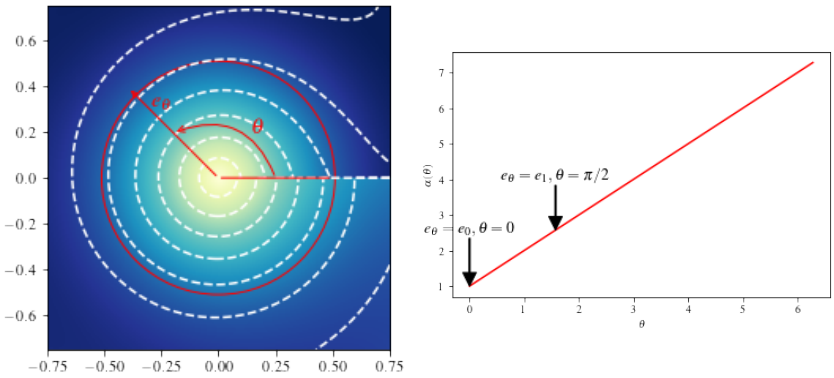


Figure 4: Intuition for the direction-dependent tail-parameter function and the standard basis tail parameters. The distribution shown (left) has PDF $dP(r, \theta) = r^{-\alpha(\theta)} r dr d\theta$ whose tail parameter $\alpha(\theta)$ depends linearly on the direction θ. Whereas prior fat-tailed theory based on $\|X\|_2 = \sup_{\|v\|_2=1} \langle X, v \rangle$ is only sensitive to the largest tail parameter $\max_{\theta \in [0,2\pi]} \alpha(\theta)$, our direction-dependent theory's tail parameter function (right) is sensitive to the tail-anisotropy present. (Feynman: Can we do $\nu = \sin$ so $\nu(0) = \nu(2\pi)$?) (Feynman: Show the TAF theory's tail param, the max of the right, In the text, say this figure is the "worst-case" in the sense that there are infinitely many tail parameters here)

The standard basis vectors provide a natural choice of projections for multivariate product distributions (as commonly encountered during blocking / grouping of random variables). Going back to our previous example, we still have that $\alpha_{\|X\|} = \max_i \alpha(e_i)$ but now the tail indices of the remaining e_j need not be equal. Admittedly, the standard basis is less suitable for correlated multivariate distributions. For example, let $t_i \sim \text{StudentT}(\nu_i)$ and consider the rotated random variable $X = R_{\pi/4}[t_1; t_2]$. Its standard basis tail parameters are

$$\Pr[\langle X, e_1 \rangle > x] = \Pr[\langle X, e_2 \rangle > x] = \Pr[0.5t_1 + 0.5t_2 > x] \leq \max_i \nu_i$$

This example illustrates that the standard basis tail parameters provide multivariate tail parameters which are no worse than previous work.

3.4 Limitations of our theory

- Standard basis parameters not general, only works for axis-aligned tails i.e. independent product distributions
- Still fails on the spiral, but it should do better than TAF
- Theory not fully general, only considers rays from origin rather than path integrals

4 Automatic Fat Tailed Variational Inference

Motivated by our theory for standard basis tail parameters, in order to improve approximation of multivariate fat-tailed random variables with anisotropic tail index, we propose the following family of fat-tailed distributions.

Definition 4.1. Anisotropic Tail-Adaptive Flows (ATAF) uses variational family

$$\mathcal{Q}_{ATAF}(\nu, \Phi_{\text{NF}}) := \{(f \circ \Phi_{\text{NF}})_* \mu_\nu\}$$

$\mu_\nu = \prod_i^d \text{StudentT}(\nu_i)$ is a d-dimensional independent product of StudentT's with degrees-of-freedom $(\nu_i)_{i=1}^d \in \mathbb{R}^d$, Φ_{NF} is a bijection (parameterized here by a normalizing flow) to increase the family's capacity, and f is a bijection between constrained supports [22].

Remark 6. Let $\mu = \prod_1^d \text{StudentT}(\nu_i)$ and suppose f^W is invertible and satisfies Assumption 1. Then $f_*^W \mu$ can be tail anisotropic.

Compared to **(author?)** [18], the degrees of freedom ν is no longer shared across all d dimensions so the variational approximations may exhibit varying degrees of fat-tailedness across different dimensions. As seen in fig. 1, such anisotropy is particularly important in situations such as the "heavy-tailed pancake." This situation is particularly relevant in probabilistic programming, where multiple latent variariables (potentially of different tail index) are blocked together for joint sampling / approximation.

We chose to estimate tail indices parametrically by including ν_i as an additional variable within variational inference optimzation, though we explore non-parametric tail index estimation in **??**.

5 Experiments

These experiments investigate the behavior of neural density estimators with *heavy-tailed base distribution*. Specifically, we consider a masked autoregressive flow [25] 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

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

We found that applying ATAF when the target is light tailed results in minimal error (appendix B); the result agrees with intuition because $\text{StudentT}(\nu) \to N(0,1)$ as $\nu \to \infty$ so it is reasonable to expect ATAF to learn reasonable approximations.

5.1 BLR example

A common Bayesian statistics problem exhibiting anisotropic tails is Bayesian linear regression with unknown location and scale parameters. With a conjugate prior of $\sigma^2 \sim \text{Inv-Gamma}(a_0, b_0)$ and $\beta \mid \sigma^2 \sim N(0, \sigma^2 I)$, the posterior distribution is also $\sigma^2 \sim \text{Inv-Gamma}(a_n, b_n)$ (with parameters $a_n = a_0 + \frac{n}{2}$ and $b_n = b_0 + \frac{1}{2}(y^\top y - \mu_n^\top \Sigma_n \mu_n)$) and conditionally normal for $\beta \sim N(\mu_n, \sigma^2 \Sigma_n)$ (with parameters $\mu_n = \Sigma_n (X^\top X \hat{\beta})$ and $\Sigma_n = (X^\top X + \sigma^{-2} I)^{-1}$). Notice that this posterior is conditionally Gaussian for every fixed σ^2, but the marginal distribution $p(\sigma^2) = \text{Inv-Gamma}(a_n, b_n) \in \mathcal{L}^1_{a_n}$ exhibits fat tails. Hence, it is unsurprising that in fig. 5 we find that (consistent with **??**) normalizing flows with isotropic base distributions (Gaussian and TAF's isotropic StudentT product) fail

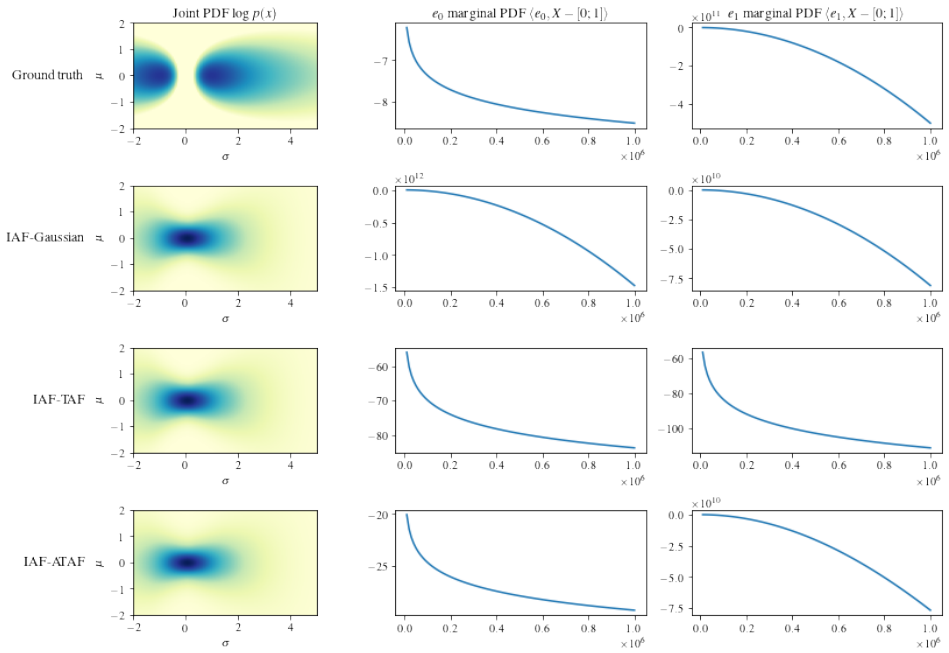
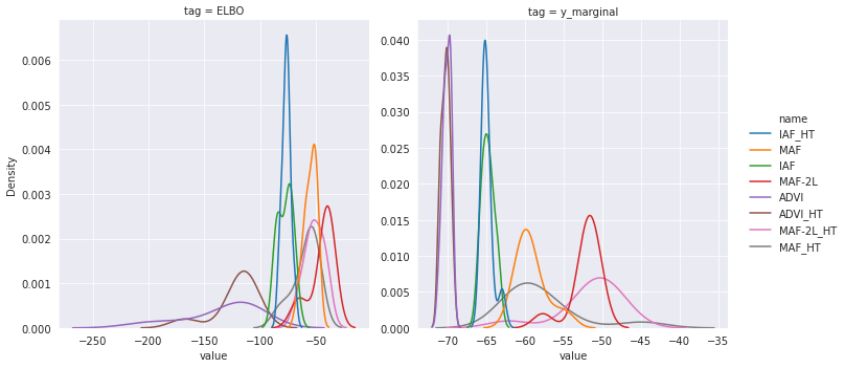


Figure 5: Bayesian linear regression's tail-anisotropic posterior (left column) is light-tailed in $\beta = \langle [\beta, \sigma], e_1 \rangle$ (middle column) and fat-tailed in $\sigma^2 = \langle [\beta, \sigma], e_2 \rangle^2$ (right column). While ATAF (bottom row) is able to model both tail decays accurately, as a consequence of theorem 5 neither tail-isotropic flow (IAF-Gaussian, IAF-TAF) is able to do so.

(Feynman: Make sure to explain this well. Figures on left draw the eye, but the point is the concave/convex on right. Keep left column, explain looking at the bulk may lead to poor conclusions and that examining tails is non-obvious and leads to different conclusions. Consider symmetric log x scale to squish zero probability y-axis, clipping to $\sigma \geq 1$. Mention difficulty of heavy-tails, suggest next steps include winsorization.)

5.2 Eight schools



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

Figure 6: Final ELBO and estimated log marginal $\log P(y)$ after 5000 steps on `eight_schools`

5.3 Bayesian regression analysis on diamonds

Table 2: Final ELBO and (MC estimate of) log marginal $\log P(y)$ after 10000 steps on `diamonds`

	ELBO	$\log P(y)$
ADVI	-374.4	-6912.0
ADVI-HT	-246.3	-6876.6
MAF	-211.1	-6894.8
MAF-HT	-208.5	-7442.0
MAF-2L	-197.9	-6532.4
MAF-2L-HT	-191.9	-6839.9
MAF-3L	-194.1	-7027.3
MAF-3L-HT	-209.8	-7128.1

6 Related Work

Fat-tails in variational inference 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. [24] 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. [5] perform a similar modification except in a coupled VAE [4] and showed improvements in the marginal likelihoods of reconstructed images. In addition, [3] consider a mixture of StudentTs for the prior $p(z)$. 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

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

$p(x \mid z)$ and the prior $p(z)$. Our work differs from [24, 5, 3] in that we consider heavy-tailed variational approximations $q(z \mid x)$ rather than priors $p(z)$, and although [1] 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 [27], 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 [7, 11] 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.

Normalizing flows **(author?)** [2] utilizes a restrictive alternating real-NVP flow (originally designed for generative image modeling) **(author?)** [2, 28] both consider normalizing flows for VI, and while neither addresses fat-tailed target densities **(author?)** [28] docoments improvements over ADVI and NUTS across thirteen different Bayesian linear regression models from **(author?)** [13] and **(author?)** [2] shows normalizing flows can compose nicely with other advances in black-box VI (e.g. stick the landing, importance weighting).

To our knowledge, only **(author?)** [18] explicitly considers flows with tails heavier than Gaussians. However, they primarily consider density estimation rather than variational inference. While they also show similar impossibility results and propose a flow transform of a StudentT base distribution, our proposed model is more flexible by allowing for independent tail parameters for each dimension and we also address the issue of distributions with constrained supports.

7 Discussion

Limitations: only considered symmetric base distributions, could also consider skewness [15].

8 Broader Impacts

Acknowledgments and Disclosure of Funding

Feynman Liang is supported by funding from NPSC and Facebook. Michael Mahoney acknowledges funding from ??.

References

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

[2] A. Agrawal, D. Sheldon, and J. Domke. Advances in black-box vi: Normalizing flows, importance weighting, and optimization. *arXiv preprint arXiv:2006.10343*, 2020.

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

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

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

[6] R. T. Chen, J. Behrmann, D. Duvenaud, and J.-H. Jacobsen. Residual flows for invertible generative modeling. *arXiv preprint arXiv:1906.02735*, 2019.

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

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

[9] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[10] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. Neural spline flows. *arXiv preprint arXiv:1906.04032*, 2019.

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

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

[13] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[14] W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

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

[16] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018.

[17] F. Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

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

[19] P. Jaini, K. A. Selby, and Y. Yu. Sum-of-squares polynomial flow. In *International Conference on Machine Learning*, pages 3009–3018. PMLR, 2019.

[20] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

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

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

[23] M. Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

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

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

[26] R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

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

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

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

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[TODO]**
   (b) Did you describe the limitations of your work? **[TODO]**
   (c) Did you discuss any potential negative societal impacts of your work? **[TODO]**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[TODO]**
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? **[TODO]**
   (b) Did you include complete proofs of all theoretical results? **[TODO]**
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[TODO]**
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[TODO]**
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[TODO]**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **[TODO]**
   (b) Did you mention the license of the assets? **[TODO]**
   (c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]**

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]**

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[TODO]**
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[TODO]**
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[TODO]**

A Appendix

Optionally include extra information (complete proofs, additional experiments and plots) in the appendix. This section will often be part of the supplemental material.

B Normal-normal location mixture

We consider a Normal-Normal conjugate inference problem where the posterior is known to be a Normal distribution as well. Here, we aim to show that ATAF performs no worse than ADVI because StudentT$(\nu) \to N(0, 1)$ as $\nu \to \infty$. Figure 7 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.

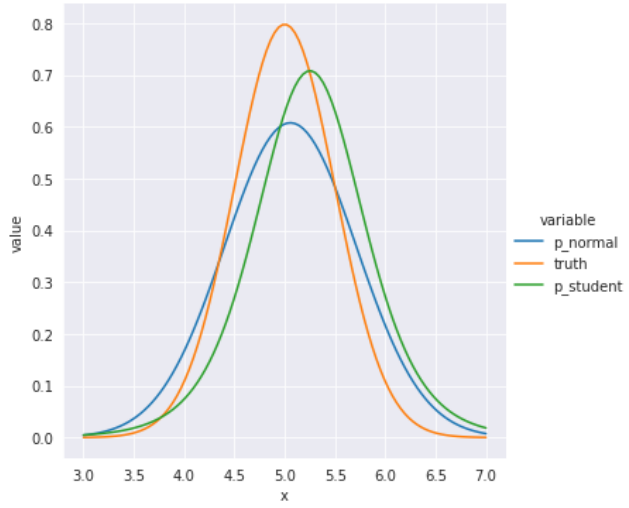


Figure 7: VI against a Normal posterior