Fat–Tailed Variational Inference with Anisotropic Tail Adaptive Flows

Feynman Liang, Liam Hodgkinson, Michael W. Mahoney

UC Berkeley, Meta, ICSI

June 28, 2022

Variational inference

Goal: Given access to a proportional $\bar{\pi} \propto \pi$, approximate $\pi \approx q$

Variational inference

Goal: Given access to a proportional $\bar{\pi} \propto \pi$, approximate $\pi \approx q$

Example: Bayesian inference, $\pi(\theta) = p(\theta \mid x)$ and $\bar{\pi}(\theta) = p(x, \theta)$

Variational inference

Goal: Given access to a proportional $\bar{\pi} \propto \pi$, approximate $\pi \approx q$
Example: Bayesian inference, $\pi(\theta) = p(\theta \mid x)$ and $\bar{\pi}(\theta) = p(x, \theta)$
Variational inference: $\max_{q \in \mathcal{Q}} \text{ELBO}(q, \bar{\pi})$ where

$$
\begin{aligned}
-\text{KL}(q, \pi) \propto \text{ELBO}(q, \bar{\pi}) &= \int q(x) \log \frac{\bar{\pi}(x)}{q(x)} dx \\
&\approx \frac{1}{n} \sum_{i=1}^{n} \log \frac{\bar{\pi}(x_i)}{q(x_i)}, \; x_i \overset{\text{i.i.d.}}{\sim} q
\end{aligned}
$$

More expressive variational family $\mathcal{Q} \Rightarrow$ better approximation quality

Berkeley
UNIVERSITY OF CALIFORNIA

Expressive variational families using flows

Let f_θ be an invertible flow and $p_X(x)$ a probability density (the *base distribution*). Consider variational family $\mathcal{Q} = \{q_\theta : \theta \in \Theta\}$ where

$$q_\theta(y) = p_X(f_\theta^{-1}(y)) \left| \det \left. \frac{df_\theta^{-1}(z)}{dz} \right|_{z=y} \right| . \quad (1)$$

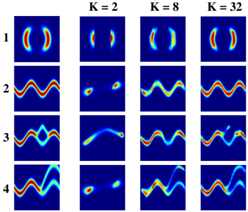


Figure 1: From [15], flows can transform a Gaussian into complex pushforward distributions

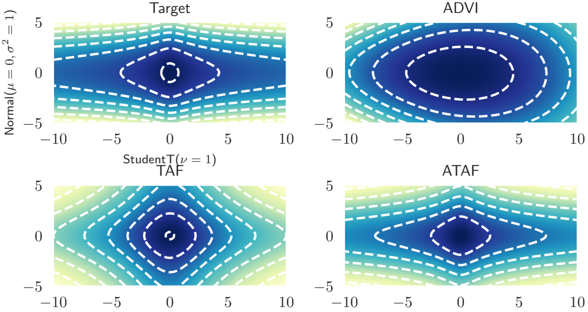
Berkeley
UNIVERSITY OF CALIFORNIA

Model	Autoregressive transform	Lipschitz when
NICE[3]	$z_j + \mu_j \cdot \mathbb{1}_{k \notin [j]}$	μ_j Lipschitz
MAF[14]	$\sigma_j z_j + (1 - \sigma_j)\mu_j$	σ_j bounded
IAF[12]	$z_j \cdot \exp(\lambda_j) + \mu_j$	λ_j bounded, μ_j Lipschitz
Real-NVP[4]	$\exp(\lambda_j \cdot \mathbb{1}_{k \notin [j]}) \cdot z_j + \mu_j \cdot \mathbb{1}_{k \notin [j]}$	λ_j bounded, μ_j Lipschitz
Glow[11]	$\sigma_j \cdot z_j + \mu_j \cdot \mathbb{1}_{k \notin [j]}$	σ_j bounded, μ_j Lipschitz
NAF[8]	$\sigma^{-1}(w^\top \cdot \sigma(\sigma_j z_j + \mu_j))$	Always (logistic mixture CDF)
NSF[5]	$z_j \mathbb{1}_{z_j \notin [-B,B]} + M_j(z_j; z_{<j}) \mathbb{1}_{x_j \in [-B,B]}$	Always (linear outside $[-B, B]$)
FFJORD[7]	n/a (not autoregressive)	Always (required for invertibility)
ResFlow[2]	n/a (not autoregressive)	Always (required for invertibility)

Table 1: Some recently developed invertible flows.

Fat-tailed variational inference

Research aims ([9], this work): What happens when π is fat-tailed? What about when π is multivariate?



Berkeley
UNIVERSITY OF CALIFORNIA

Methods

Automatic Differentiation Variational Inference (ADVI, [13, 17]):
$\mathcal{Q}_{\text{ADVI}} := \{(f_\theta)_* \mu\}$, where $\mu = \text{Normal}(0_d, I_d)$.
Tail Adaptive Flows (TAF, [9]):
$\mathcal{Q}_{\text{TAF}} := \{(f_\theta)_* \mu_\nu\}$, where $\mu_\nu = \prod_{i=1}^{d} \text{StudentT}(\nu)$ with $\nu \in \mathbb{R}_+$.
Anisotropic Tail-Adaptive Flows (ATAF, this work):
$\mathcal{Q}_{\text{ATAF}} := \{(f_\theta)_* \mu_{\boldsymbol{\nu}}\}$, where $\mu_{\boldsymbol{\nu}} = \prod_{i=1}^{d} \text{StudentT}(\nu_i)$ with $\boldsymbol{\nu} \in \mathbb{R}_+^d$.

Sharpening prior univariate theory

Assumption

f_θ is invertible, and both f_θ and f_θ^{-1} are L-Lipschitz continuous (e.g. Table 1).

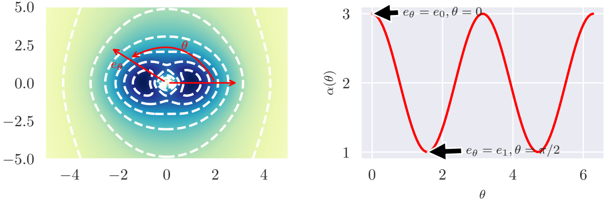
Theorem

- *f_θ cannot make the tails of a fat-tailed distribution fatter (decrease tail parameter α).*
- *If in addition f_θ is smooth with no critical points, then it cannot change the tail parameter of a fat-tailed distribution.*
- *Light-tailed distributions remain light-tailed under polynomial flows [10].*

Multivariate fat tails and tail anisotropy

Definition (Tail parameter function)

For random vector X, define $\alpha_X(v) = -\lim_{x\to\infty} \log \mathbb{P}(\langle v, X\rangle \geq x)/\log x$ when the limit exists, and $\alpha_X(v) = +\infty$ otherwise. X is *tail-isotropic* if $\alpha_X(v) \equiv c < \infty$ is constant.



Necessity of ATAF

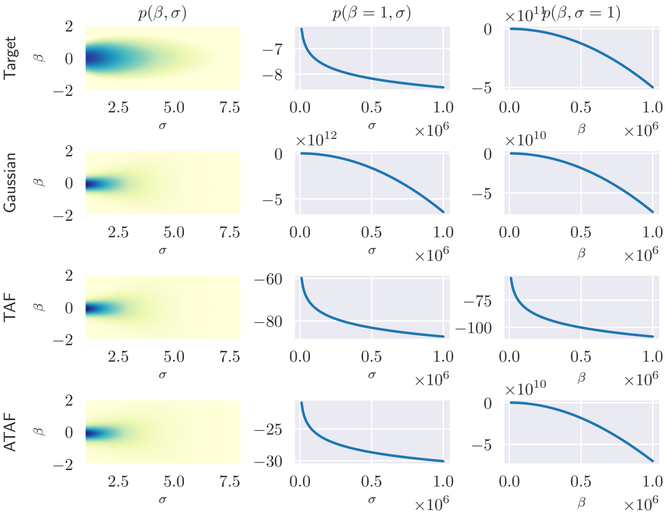
Proposition (Pushforwards of tail-isotropic distributions)

Let μ be tail isotropic with non-integer parameter ν and suppose f_θ satisfies Assumption 1. Then $(f_\theta)_*\mu$ is tail isotropic with parameter ν.

Bayesian linear regression

$$\sigma^2 \sim \text{Inv-Gamma}(a_0, b_0)$$
$$\beta \mid \sigma^2 \sim \mathcal{N}(0, \sigma^2), \qquad y \mid X, \beta, \sigma \sim \mathcal{N}(X\beta, \sigma^2),$$

The posterior is tail-anisotropic:
$p(\sigma^2, \beta = c \mid X, y) \propto \rho(\sigma^2) \in \mathcal{L}^1_{\alpha_n}$ is fat-tailed (power-law)
$p(\sigma^2 = c, \beta \mid X, y) \propto \rho(\beta \mid c) \in \overline{\mathcal{E}^2}$ is light-tailed (sub-Gaussian)



$p(\beta, \sigma)$
$p(\beta = 1, \sigma)$
$p(\beta, \sigma = 1)$
Target
Gaussian
TAF
ATAF
β
σ
$\times 10^{11}$
$\times 10^{12}$
$\times 10^{10}$
$\times 10^{6}$
2.5 5.0 7.5
0.0 0.5 1.0

Eight schools [16]

$$\tau \sim \text{HalfCauchy}(\text{loc} = 0, \text{scale} = 5)$$
$$\mu \sim \mathcal{N}(0, 5), \qquad \theta \sim \mathcal{N}(\mu, \tau), \qquad y \sim \mathcal{N}(\theta, \sigma).$$

	ELBO	$\log p(y)$
ADVI	-72.13 ± 6.89	-53.25 ± 3.44
TAF	-64.64 ± 4.88	-52.51 ± 4.41
ATAF	$-\mathbf{58.63} \pm 4.75$	$-\mathbf{51.01} \pm 3.71$
NUTS	n/a	-47.78 ± 0.093

Berkeley
UNIVERSITY OF CALIFORNIA

Financial [6] and actuarial [1] density modeling

	Fama-French 5 Industry Daily	CMS 2008-2010 DE-SynPUF
ADVI	-5.018 ± 0.056	-1.883 ± 0.012
TAF	-4.703 ± 0.023	-1.659 ± 0.004
ATAF	$\mathbf{-4.699} \pm 0.024$	$\mathbf{-1.603} \pm 0.034$

Table 2: Log-likelihoods (higher is better, $\pm$ standard errors).

Berkeley
UNIVERSITY OF CALIFORNIA

Conclusions

- Flow-based VI can expressively model the bulks of complicated distributions …

Conclusions

- Flow-based VI can expressively model the bulks of complicated distributions …
- But modeling of tails is still limited by choice of base distribution!

Conclusions

- Flow-based VI can expressively model the bulks of complicated distributions …
- But modeling of tails is still limited by choice of base distribution!
- Prior work (TAF, [9]) considers univariate tails, and in this work:
  - Prior univariate theory is refined to include α and closure results are sharpened
  - A multivariate theory is proposed to quantify tail-anisotropy and prove ATAF's necessity
  - Experiments confirm ATAF's improvements on real-world fat-tailed datasets

References I

[1] Centers for Medicare and Medicaid Services. CMS 2008-2010 data entrepreneurs' synthetic public use file (DE-SynPUF), 2010. [Online; accessed 10-March-2020].

[2] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32:9913–9923, 2019.

[3] L Dinh, D Krueger, and Y Bengio. NICE: non-linear independent components estimation. In *3rd International Conference on Learning Representations, Workshop Track Proceedings*, 2015.

[4] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations*, 2017.

[5] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in Neural Information Processing Systems*, 32:7509–7520, 2019.

References II

[6] Eugene F Fama and Kenneth R French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

[7] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

[8] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018.

[9] Priyank Jaini, Ivan Kobyzev, Yaoliang Yu, and Marcus Brubaker. Tails of Lipschitz triangular flows. In *International Conference on Machine Learning*, pages 4673–4681. PMLR, 2020.

[10] Priyank Jaini, Kira A Selby, and Yaoliang Yu. Sum-of-squares polynomial flow. In *International Conference on Machine Learning*, pages 3009–3018. PMLR, 2019.

Berkeley
UNIVERSITY OF CALIFORNIA

References III

[11] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 31:10236–10245, 2018.

[12] Diederik P. Kingma, Tim Salimans, Rafal Józefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational autoencoders with inverse autoregressive flow. *Advances in Neural Information Processing Systems*, 29:4736–4744, 2016.

[13] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

[14] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems*, 30:2338–2347, 2017.

References IV

[15] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

[16] Donald B Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):377–401, 1981.

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