Fat-tailed variational inference

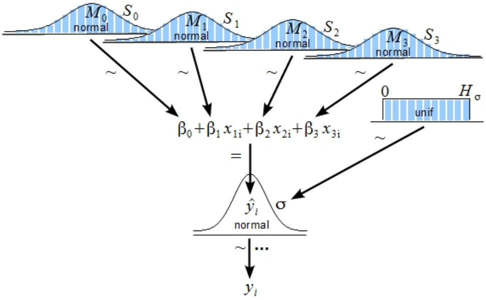
Feynman Liang, Liam Hodgkinson, Michael Mahoney

UC Berkeley

May 6, 2021

Motivating example

$$
\begin{aligned}
\text{Linear regression} \quad & y = X\beta + \epsilon,\ \beta \in \mathbb{R}^d,\ \epsilon \sim N(0,\sigma) \\
\text{Robust regression} \quad & y = X\beta + \epsilon,\ \beta \in \mathbb{R}^d,\ \epsilon \sim StudentT(\sigma) \\
\text{Bayesian robust regression} \quad & y = X\beta + \epsilon,\ \beta \sim P\ ,\ \epsilon \sim StudentT(\sigma)
\end{aligned}
$$



[1]

Goal: approximate (observables of) $\mathbb{P}(\beta \mid X, y)$

[1] https://jkkweb.sitehost.iu.edu/BMLR/

Motivating example

Goal: approximate (observables of) $\mathbb{P}(\beta \mid X, y)$
General solutions:

- (MCMC, see LIC talk) sample $\beta_i \sim \mathbb{P}(\beta \mid X, y)$:

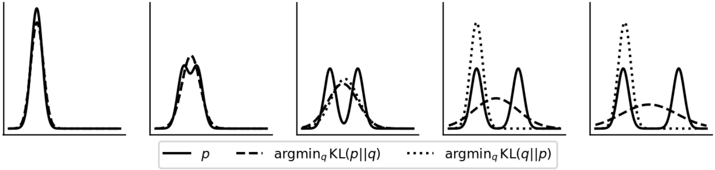
$$n^{-1} \sum_{i}^{n} f(\beta_i) \to \mathbb{E}_{\beta|X,y} f(\beta) \qquad \forall f \in C(\mathbb{R}^d)$$

- (Today's talk) search for variational approximation $q_{\theta^*} \in \mathcal{Q} = \{q_\theta : \theta \in \Theta\}$ "close" to $\mathbb{P}(\beta \mid X, y)$

Variational inference I

Definition

The *forward KL Divergence* $D_{\mathrm{KL}}(P||q_\theta) = \mathbb{E}_P \log \frac{\mathbb{P}(\beta|X,y)}{q_\theta(\beta)}$, and the *reverse KL Divergence* is $D_{\mathrm{KL}}(q_\theta||P)$.



[2]

- Forward KL is mass-covering/mean-seeking, requires sampling/integrating P, density estimation objective
- Reverse KL is zero-forcing/mode-seeking, requires sampling/integrating Q, variational inference objective

Variational inference II

Evaluating $\mathbb{P}(\beta \mid X, y) = \frac{\mathbb{P}(\beta, X, y)}{\int \mathbb{P}(\beta, X, y) d\beta}$ intractable!

Definition

The *evidence lower bound* $ELBO(\theta) := \mathbb{E}_{q_\theta} \log \frac{\mathbb{P}(\beta, X, y)}{q_\theta(\beta)}$.

Can show $D_{\mathrm{KL}}(q_\theta || P) = \text{constant} - ELBO(\theta)$, so VI transforms (intractable) inference problem to a (tractable) optimization of an approximation:

$$\mathbb{P}(\beta \mid X, y) \approx \arg\max_{q_\theta \in \mathcal{Q}} ELBO(\theta)$$

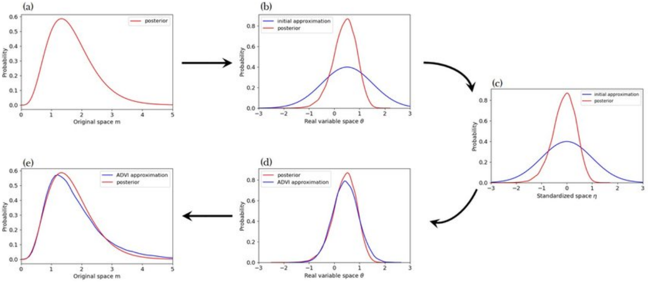
[2] https://www.tuananhle.co.uk/notes/reverse-forward-kl.html

Automatic differentiation variational inference (ADVI) I

Definition ([3])

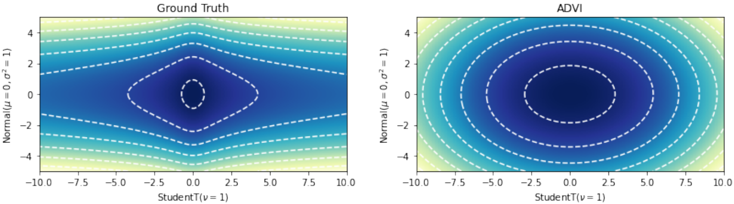
$$\mathcal{Q}_{ADVI} = \{q_\theta(\beta) = f_* N(\beta \mid \theta_0, e^{-\theta_1}) : \theta_0, \theta_1 \in \mathbb{R}^d\}$$

f is a deterministic bijection between supports.



Automatic differentiation variational inference (ADVI) II

Problem: Gaussian approximations are too limited!



[3] Kucukelbir et al. "Automatic differentiation variational inference." JMLR 2017

[4] Zhang, Xin, and Andrew Curtis. "Seismic tomography using variational inference methods." Journal of Geophysical Research: Solid Earth 125.4 (2020)

Normalizing flows I

Normalizing flows: $f = f^W$ is a ~~deterministic~~ learnable bijection represented with neural networks.

Lemma (Change of variable)

If $Y = f(X)$ is an injective pushforward, then

$$p_Y(y) = p_X(f^{-1}(y))|\det Df^{-1}(y)|$$

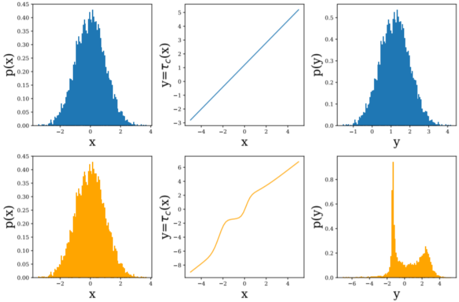
Desiderata:

- Sampling: fast evaluation of f
- Density: fast evaluation of f^{-1} and $\det Df$

Normalizing flows II

Example (Neural autoregressive flows)

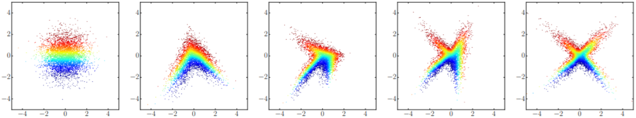
$y_i = \mathrm{DNN}(x_t; W = c(x_{1:t-1}))$ constrained strictly monotonic



Normalizing flows III

Example (Masked autoregressive flows, MAF)

$$y_i = \sigma(x_{1:i-1})x_i + \mu(x_{1:i-1})$$



Beyond sub-Gaussians I

Theorem (Wainwright “High-dimensional statistics” 2019)

Let X be σ-sub-Gaussian and f be L-Lipschitz. Then $f(X) - \mathbb{E}f(X)$ is L-sub-Gaussian.

Observation: Gaussian base distributions are pervasive!

```
# some important defaults
initial_dist_name = "normal"
initial_dist_map = 0.0
```

[5]

```
mg = tfd.MultivariateNormalDiag(
    loc=tf.zeros(ndims), scale_diag=[3.] + [1.] * (ndims - 1))
funnel = tfd.TransformedDistribution(
    mg, bijector=tfb.MaskedAutoregressiveFlow(bijector_fn=bijector_fn))
```

[6]

```
>>> from pyro.nn import AutoRegressiveNN
>>> base_dist = dist.Normal(torch.zeros(10), torch.ones(10))
>>> transform = AffineAutoregressive(AutoRegressiveNN(10, [40]))
>>> pyro.module("my_transform", transform)  # doctest: +SKIP
>>> flow_dist = dist.TransformedDistribution(base_dist, [transform])
```

[7]

Observation: Many f^W used in practice are Lipschitz!

Model	coefficients	$T_j(z_j\,; z_1, \ldots, z_{j-1})$
NICE	$\mu_j(z_{<l})$	$z_j + \mu_j \cdot \mathbf{1}_{j \notin [l]}$
IAF	$\sigma_j(z_{<j}),\ \mu_j(z_{<j})$	$\sigma_j z_j + (1 - \sigma_j)\mu_j$
MAF	$\lambda_j(z_{<j}),\ \mu_j(z_{<j})$	$z_j \cdot \exp(\lambda_j) + \mu_j$
Real-NVP	$\lambda_j(z_{<l}), \mu_j(z_{<l})$	$\exp(\lambda_j \cdot \mathbf{1}_{j \notin [l]}) \cdot z_j + \mu_j \cdot \mathbf{1}_{j \notin [l]}$
Glow	$\sigma_j(z_{<l}), \mu_j(z_{<l})$	$\sigma_j \cdot z_j + \mu_j \cdot \mathbf{1}_{j \notin [l]}$

[8]

Beyond sub-Gaussians II

Definition (Classification of tails)

- Exponential-type: $X \in \mathcal{E}_\alpha^p$ means $\mathbb{P}(X \geq x) = \mathcal{O}(e^{-\alpha x^p})$
- Logarithmic-type: $X \in \mathcal{L}_\alpha^p$ means $\mathbb{P}(X \geq x) = \mathcal{O}(e^{-\alpha (\log x)^p})$

Example

- $\mathcal{E}_\alpha^2$ sub-Gaussians
- $\mathcal{E}_\alpha^1$ sub-Exponentials
- $\mathcal{L}_\alpha^1$ regularly varying (power law)
  - $\mathrm{StudentT}(\nu) \in \mathcal{L}_\nu^1$
  - $\mathrm{Cauchy} \in \mathcal{L}_1^1$

Beyond sub-Gaussians III

Assumption 1: λ_j and σ_j are bounded and μ_j is Lipschitz,

Theorem (LHM, 2021)

Under Assumption 1, the distribution classes $\cup_{\beta\in\mathbb{R}}\mathcal{E}^p_\beta$ and $\mathcal{L}^p_\alpha$ (with $p, \alpha \in \mathbb{R}_+$) are closed under every flow transformation in Table 1.

Theorem (LHM, 2021)

There does not exist a polynomial map between $\mathcal{L}$ and $\mathcal{E}$.

[5] https://github.com/pymc-devs/pymc3/blob/d7172c0a1a76301031d1b3b411d00643c416a0c4/pymc3/variational/opvi.py#L766

[6] https://github.com/tensorflow/probability/blob/22947dc575778318b660303129ee39c2a870e5a9/spinoffs/inference_gym/i L92

[7] https://github.com/pyro-ppl/pyro/blob/d7687ae0f738bd81a792dabbb18a53c0fce73765/pyro/distributions/transforms/affine_autoregressive.py#L46

[8] https://arxiv.org/pdf/1907.04481.pdf

Tail-adaptive flows (TAFs)

Definition

$$\mathcal{Q}_{TAF} = \left\{ \left(f_*^W \left(\prod_{i=1}^{d} \mathsf{StudentT}(\nu) \right) \right) : \nu \in \mathbb{R}_+, W \in \mathbb{R}^{\#\text{ NF params}} \right\}$$

Method	Power	Gas	Hepmass	MiniBoone	BSDS300
MADE	0.40 ± 0.01	8.47 ± 0.02	-15.15 ± 0.02	-12.24 ± 0.47	153.71 ± 0.28
MAF affine (5)	0.14 ± 0.01	9.07 ± 0.02	-17.70 ± 0.02	-11.75 ± 0.44	155.69 ± 0.28
MAF affine (10)	0.24 ± 0.01	10.08 ± 0.02	-17.73 ± 0.02	-12.24 ± 0.45	154.93 ± 0.28
MAF MoG (5)	0.30 ± 0.01	9.59 ± 0.02	-17.39 ± 0.02	-11.68 ± 0.44	156.36 ± 0.28
TAN	0.60 ± 0.01	12.06 ± 0.02	-13.78 ± 0.02	-11.01 ± 0.48	159.80 ± 0.07
NAF DDSF (5)	0.62 ± 0.01	11.91 ± 0.13	-15.09 ± 0.40	-8.86 ± 0.15	157.73 ± 0.04
NAF DDSF (10)	0.60 ± 0.02	11.96 ± 0.33	-15.32 ± 0.23	-9.01 ± 0.01	157.43 ± 0.30
SOS (7)	0.60 ± 0.01	11.99 ± 0.41	-15.15 ± 0.10	-8.90 ± 0.11	157.48 ± 0.41
TAF affine (5)	0.28 ± 0.01	9.87 ± 0.23	-17.41 ± 0.20	-11.71 ± 0.09	156.53 ± 0.52
TAF SOS (7)	0.59 ± 0.01	11.99 ± 0.34	-15.11 ± 0.18	-8.94 ± 0.23	157.52 ± 0.22

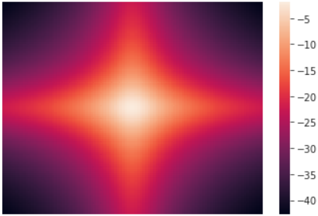
9

[9] Jaini, Priyank, et al. "Tails of Lipschitz Triangular Flows." ICML 2020

Multivariate heavy-tails

Prior work (Jaini, 2020): $X \in \mathbb{R}^d$ is heavy-tailed iff $\|X\|$ is, develop theory for elliptical distributions $\underline{X} \stackrel{d}{=} \underline{\mu} + R\underline{\underline{A}}\,\underline{\underline{U}}^{(d)}$.

Problem 1: TAF's $\prod_1^d \text{StudentT}(\nu)$ is not elliptical



Problem 2: Tail parameter ν is the same in every direction!

Direction-dependent tail parameters

Root cause: $\sup_{v \in \mathcal{S}^{d-1}} \langle v, X \rangle = \|X\|_2$, so scalar tail parameter is an upper bound.

Definition

The *tail parameter function* for a fat-tailed random variable $X \in \mathbb{R}^d$

$$\begin{aligned} \alpha : \mathcal{S}^{d-1} &\to \mathbb{R}_+ \\ v &\mapsto \limsup_{r \to \infty} \frac{\log \mathbb{P}(\| \langle v, X \rangle \| > r)}{\log r} \end{aligned}$$

Example

Elliptical distributions are *tail isotropic* i.e. $\alpha(v) \equiv c$ is constant.

Proposition (LHM, 2021)

Let μ be elliptical or $\prod_1^d \mathrm{StudentT}(\nu)$ and suppose f^W is invertible and satisfies Assumption 1. Then $f_*^W \mu$ is tail isotropic with $\alpha \equiv \nu$.

Standard basis tail parameters

$\alpha(\cdot)$ difficult to work with; need finite-dimensional parameterization which still permits tail anisotropy.

Key observation: multivariate distributions oftentimes obtained from concatenation (blocked Metropolis-Hastings, Hamiltonian Monte-Carlo) $\implies$ tails are axis-aligned.

Definition

The *standard basis tail parameters* are $\{\alpha_i := \alpha(v_i) : i \in [d]\}$

Fat-tailed variational inference (FTVI)

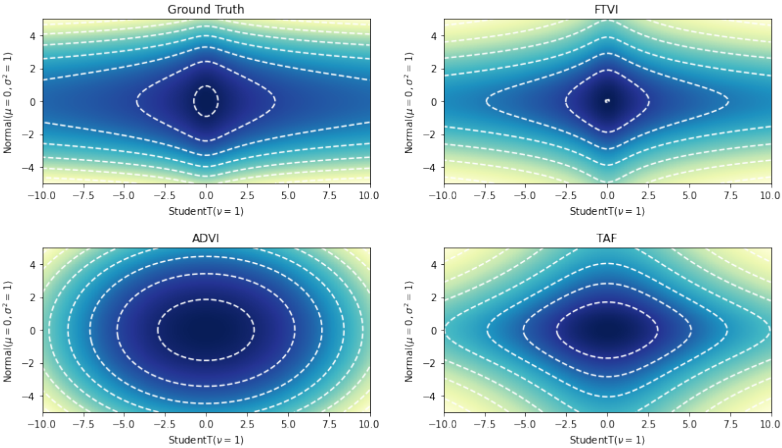
Definition

$$\mathcal{Q}_{FTVI} = \left\{ \left(f_*^W \left(\prod_{i=1}^{d} \mathrm{StudentT}(\nu_i) \right) \right) : \nu \in \mathbb{R}_+^d, W \in \mathbb{R}^{\#\ \text{NF params}} \right\}$$

Remark

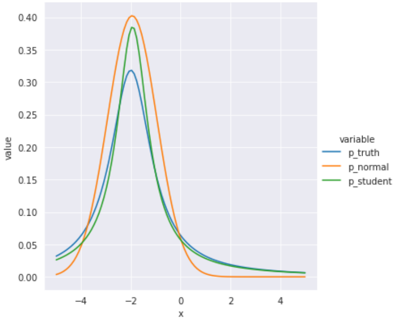
Let $\mu = \prod_1^d \mathit{StudentT}(\nu_i)$ and suppose f^W is invertible and satisfies Assumption 1. Then $f_*^W \mu$ can be tail anisotropic.

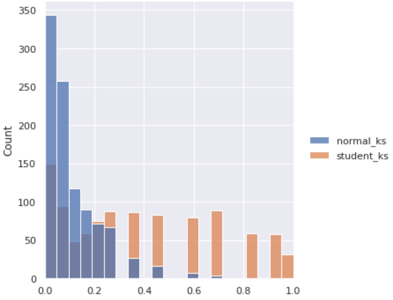
Results: fat-tailed pancake



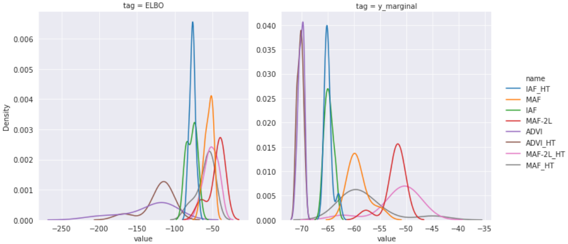
Results: gamma scale mixture

```
scale = InvGamma(1/2, 1/2)
truth = scale.sqrt() * Normal(0, 1)
```





Results: 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

Tail index algebra

- Many of previous proofs rely on a few common lemmas, extract into an easy-to-use algebra
- Enables a priori tail index estimation without samples (quick and dirty upper bounding, initializing ν for VI)
- Handles addition, multiplication, division, concatenation, exp/log, Lipschitz functions
- Conditioning: conditional asymptotically equivalent to joint