Whereas previous chapters considered adaptive methods which learn an approximation's bulk (**??**) and tail (**??**) from samples, in this chapter we develop a systematic approach for analyzing the tails of random variables during the static analysis (before drawing samples) pass of a probabilistic programming language (PPL) compiler. To characterize how the tails change under algebraic operations, we develop an algebra acting on a three-parameter family of tail asymptotics based on the generalized Gamma distribution. Our algebraic operations are closed under addition and multiplication, capable of distinguishing sub-Gaussians with differing scales, and handle ratios sufficiently well to reproduce the tails of most important statistical distributions directly from their definitions. Our experiments confirm that inference algorithms leveraging generalized Gamma algebra metadata attain superior performance across a number of density modeling and variational inference tasks. Parts of this chapter have been submitted for peer review as Feynman Liang et al. "Static Analysis of Tail Behaviour with a Generalized Gamma Algebra". In: *Submitted to AISTATS 2023* (2023).

0.1 Introduction

To facilitate efficient probabilistic modelling and inference, modern probabilistic programming languages (PPLs) draw upon recent developments in functional programming [Tol+16], programming languages [Ber19], and deep variational inference [Bin+19]. Despite their broadening appeal, common pitfalls such as mismatched distribution supports [Lee+19] and non-integrable expectations [WLL18; Veh+15; Yao+18] remain uncomfortably commonplace and challenging to debug. Recent innovations aiming to improve PPLs have automated verification of distribution constraints [Lee+19], tamed noisy gradient estimates [Esl+16] and unruly density ratios [Veh+15; WLL18], and approximated high-dimensional distributions with non-trivial bulks [Pap+21] and non-Gaussian tails [Jai+20].

Continuing this line of work, here we we consider how to statically analyze a probabilistic program in order to automate the inference of tail behavior for any random variables present. At present, correct inference of tail behaviour for target distributions remains an outstanding issue [Yao+18; WLL18], which causes challenges for downstream Monte Carlo tasks. For example, importance sampling estimators can exhibit infinite variance if the tail of the approximating density is lighter than the target. Most prominant black-box variational inference methods are incapable of changing their tail behaviour from an initial proposal distribution [Jai+20; LHM22]. MCMC algorithms may also lose ergodicity when the tail of the target density falls outside of a particular family [RT96]. All of these issues could be avoided if the tail of the target is known before runtime.

To classify tail asymptotics and define calibration, we propose a three-parameter family based on the generalized Gamma distribution (eq. (2)) which interpolates between established asymptotics on sub-Gaussian [Led01] and regularly varying [Mik99] random variables. Algebraic operations on random variables can be lifted to computations on the tail parameters resulting in what we call the *generalized Gamma algebra (GGA)*. Through analyzing opera-

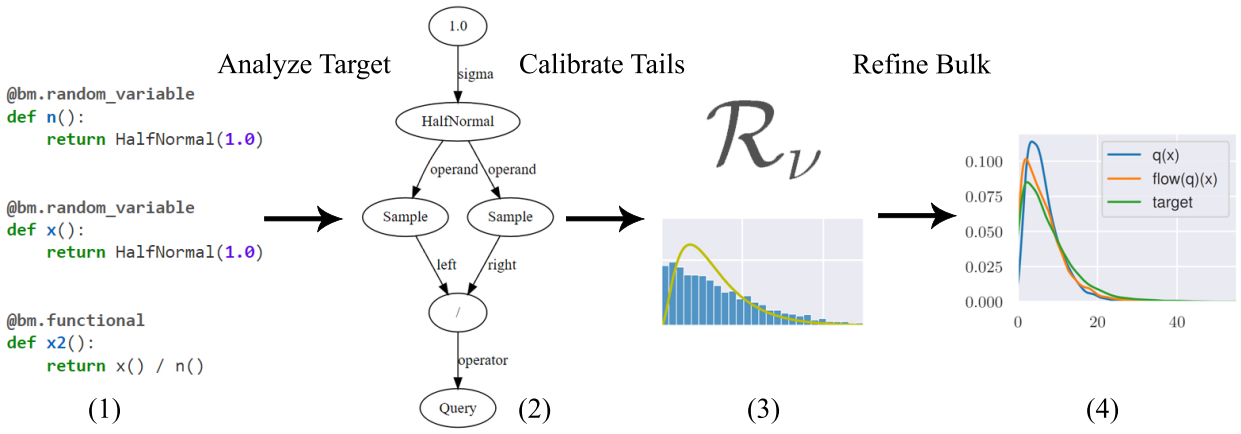

Figure 0.1: Our overall approach for density approximations with calibrated tails. A generative model expressed in a PPL (1) and analyzed using the GGA (2) to compute the tail parameters of the target. A representative distribution with calibrated tails is chosen for the initial approximation (3) and a learnable Lipschitz pushforward (see Lemma 2) is optimized (4) to correct the bulk approximation.

tions like $X + Y$, X^2, and X/Y at the level of densities (e.g. additive convolution $p_X \oplus p_Y$), the tail parameters of a target density can be estimated from the parameters of any input distributions using Table 0.1.

Operationalizing the GGA, we propose *tail inferential* static analysis analogous to traditional *type inference* and provide a reference implementation using the `beanmachine graph` [Teh+20] PPL compiler. GGA tail metadata can be used to diagnose and address tail-related problems in downstream tasks, such as employing Riemannian-manifold methods [GC11] to sample heavy tails or pre-emptively detect unbounded expectations. Here, we consider density estimation and variational inference where we use the GGA-computed tail of the target density to calibrate our density approximation. When composed with a learnable Lipschitz pushforward map (Section 0.4), the resulting combination is a flexible density approximator with provably calibrated tails.

Contributions

- The GGA is introduced, generalizing prior work on classifying tail asymptotics while including both sub-Gaussian / sub-exponentials [Led01] as well as power-law / Pareto-based tail indices [CSN09]. Composing operations outlined in table 0.1, one can compute the GGA tail class for downstream random variables of interest.
- The GGA is implemented in the static analysis phase of a PPL compiler. This unlocks the ability to leverage GGA metadata in order to better tailor the emitted inference algorithm.
- Finally, we propose and evaluate a density estimator which combines GGA tails with normalizing flows in order to simultaneously achieve good bulk approximation as well as

correct tails.

0.2 Related Work

Heavy tails and probabilistic machine learning

For studying heavy tails, methods based on subexponential distributions [GK98] and generalized Pareto distributions (GPD) or equivalently regularly varying distributions [Taj03] have recieved attention historically. Mikosch [Mik99] presents closure theorems for regularly varying which are special cases of Proposition 1 and Lemma 2. Heavy tails can impact probabilistic machine learning methods in a number of ways. The observation that density ratios $\frac{p(x)}{q(x)}$ tend to be heavy tailed has resulted in new methods for smoothing importance sampling [Veh+15], adaptively modifying divergences [WLL18], and diagnosing variational inference through the Pareto $\hat{k}$ diagnostic [Yao+18]. These works are complementary to our paper and our reported results include $\hat{k}$ diagnostics for VI and $\hat{\alpha}$ tail index estimates based on GPD.

Our work considers heavy-tailed targets $p(x)$ which is the same setting as Jaini et al. [Jai+20] and Liang et al. [LHM22]. Whereas those respective works lump the tail parameter in as another variational parameter and may be more generally applicable, the GGA may be applied before samples are drawn and leads to perfectly calibrated tails when applicable.

Probabilistic programming

PPLs can be characterized by the primary use case optimized for, whether that's Gibbs sampling over Bayes nets [Spi+96; Val+17], stochastic control flow [Goo+12; WSG11], deep stochastic variational inference [Tra+18; Bin+19], or Hamiltonian Monte-Carlo [Car+17; Xu+20]. Our implementation target `beanmachine` [Teh+20] is a declarative PPL selected due to availability of a PPL compiler and support for static analysis plugins. Similar to Bingham et al. [Bin+19] and Siddharth et al. [Sid+17], it uses PyTorch [Pas+19] for GPU tensors and automatic differentiation. Synthesizing an approximating distribution during PPL compilation (Section 0.4) is also performed in the Stan language by Kucukelbir et al. [Kuc+17] and normalizing flow extensions in Webb et al. [Web+19]. We compare directly against these related density approximators in Section 0.5.

Static analysis

There is a long history of formal methods and probabilistic programming [Koz79; JP89]. While much of the research [Cla+13] is concerned with defining formal semantics and establishing invariants [WHR18] See [Ber19] for a recent review. Static analysis utilizes the abstract syntax tree (AST) representation of a program in order to compute invariants (e.g. the return type of a function, the number of classes implementing a trait) without executing the underlying program. As dynamic analysis in PPs is less reliable due to non-determinism, static analysis methods for PPs become increasingly important.

Within PPLs, static analysis has traditionally been applied in the context of formalizing semantics [Koz79] and has been used to verify probabilistic programs by ensuring termination,

bounding random values values [SCG13]. [Lee+19] proposes a static analyzer for the Pyro PPL [Bin+19] to verify distribution supports and avoid $-\texttt{Inf}$ log probabilities.

More relevant to our work are applications of static analysis to improve inference. Nori et al. [Nor+14] statically analyzes a probabilistic program and computes pre-images of observations in order to better adapt MCMC proposal distributions. While we also perform static analysis over abstract syntax tree (AST) representations of a probabilistic program, applying GGA yields an upper bound on the tails of all random variables so that calibrated tails can be imposed on distribution estimates.

0.3 The Generalized Gamma Algebra

Here we formulate an algebra of random variables that is closed under most standard elementary operations (addition, multiplication, powers) which forms the foundation for our static analysis.

Definition 1 *A random variable X is said to have a* generalized Gamma tail *if the Lebesgue density of $|X|$ satisfies*

$$p_{|X|}(x) \sim cx^{\nu}e^{-\sigma x^{\rho}}, \qquad \text{as } x \to \infty, \tag{1}$$

for some $c > 0$, $\nu \in \mathbb{R}$, $\sigma > 0$ and $\rho \in \mathbb{R}$. Denote the set of all such random variables by $\mathcal{G}$.

Consider the following equivalence relation on $\mathcal{G}$: $X \equiv Y$ if and only if $0 < p_{|X|}(x)/p_{|Y|}(x) < +\infty$ for all sufficiently large x. The resulting equivalence classes can be represented by their corresponding parameters ν, σ, ρ, and hence, we denote the class of random variables X satisfying eq. (1) by (ν, σ, ρ). In the special case where $\rho = 0$, for a fixed $\nu < -1$, each class $(\nu, \sigma, 0)$ for $\sigma > 0$ is equivalent, and is denoted by $\mathcal{R}_{|\nu|}$, representing *regularly varying* tails. Our algebra operates on these equivalence classes of $\mathcal{G}$, characterizing the change in tail behaviour under various operations.

The form of eq. (1) and the name of the algebra is derived from the generalized Gamma distribution.

Definition 2 *Let $\nu \in \mathbb{R}$, $\sigma > 0$, and $\rho \in \mathbb{R}\backslash\{0\}$ be such that $(\nu + 1)/\rho > 0$. A non-negative random variable X is* generalized Gamma distributed *with parameters ν, σ, ρ if it has Lebesgue density*

$$p_{\nu,\sigma,\rho}(x) = c_{\nu,\sigma,\rho}x^{\nu}e^{-\sigma x^{\rho}}, \qquad x > 0, \tag{2}$$

where $c_{\nu,\sigma,\rho} = \rho\sigma^{(\nu+1)/\rho}/\Gamma((\nu + 1)/\rho)$ is the normalizing constant.

The importance of the generalized Gamma form arises due to a combination of two factors:

(i) The majority of interesting continuous univariate distributions with infinite support satisfy eq. (1), including Gaussians ($\nu = 0$, $\rho = 2$), gamma/exponential/chi-squared ($\nu > -1$, $\rho = 1$), Weibull/Frechet ($\rho = \nu + 1$), and Student T/Cauchy/Pareto ($\mathcal{R}_{\nu}$). However, some notable exceptions include the log-normal distributions.

Ordering	$\max\{(\nu_1,\sigma_1,\rho_1),(\nu_2,\sigma_2,\rho_2)\}$ $\equiv \begin{cases}(\nu_1,\sigma_1,\rho_1) & \text{if } \limsup_{x\to\infty}\frac{x_1^{\nu}e^{-\sigma_1 x^{\rho_1}}}{x_2^{\nu}e^{-\sigma_2 x^{\rho_2}}}<+\infty\\(\nu_2,\sigma_2,\rho_2) & \text{otherwise.}\end{cases}$
Addition	$(\nu_1,\sigma_1,\rho_1)\oplus(\nu_2,\sigma_2,\rho_2)$ $\equiv \begin{cases}\max\{(\nu_1,\sigma_1,\rho_1),(\nu_2,\sigma_2,\rho_2)\} & \text{if } \rho_1\neq\rho_2 \text{ or } \rho_1,\rho_2<1\\(\nu_1+\nu_2+1,\min\{\sigma_1,\sigma_2\},1) & \text{if } \rho_1=\rho_2=1\\(\nu_1+\nu_2+\frac{2-\rho}{2},(\sigma_1^{-\frac{1}{\rho-1}}+\sigma_2^{-\frac{1}{\rho-1}})^{1-\rho},\rho) & \text{if } \rho=\rho_1=\rho_2>1.\end{cases}$
Powers	$(\nu,\sigma,\rho)^{\beta}\equiv(\frac{\nu+1}{\beta}-1,\sigma,\frac{\rho}{\beta})$ for $\beta>0$
Reciprocal*	$(\nu,\sigma,\rho)^{-1}\equiv\begin{cases}(-\nu-2,\sigma,-\rho) & \text{if } (\nu+1)/\rho>0 \text{ and } \rho\neq 0\\\mathcal{R}_2 & \text{otherwise}\end{cases}$
Scalar Multiplication	$c(\nu,\sigma,\rho)\equiv(\nu,\sigma/\lvert c\rvert^{\rho},\rho)$
Multiplication	$(\nu_1,\sigma_1,\rho_1)\otimes(\nu_2,\sigma_2,\rho_2)$ $\equiv\begin{cases}\left(\frac{1}{\mu}\left(\frac{\nu_1}{\lvert\rho_1\rvert}+\frac{\nu_2}{\lvert\rho_2\rvert}+\frac{1}{2}\right),\sigma,-\frac{1}{\mu}\right) & \text{if } \rho_1,\rho_2<0\\\left(\frac{1}{\mu}\left(\frac{\nu_1}{\rho_1}+\frac{\nu_2}{\rho_2}-\frac{1}{2}\right),\sigma,\frac{1}{\mu}\right) & \text{if } \rho_1,\rho_2>0\\\mathcal{R}_{\lvert\nu_1\rvert} & \text{if } \rho_1\leq 0,\rho_2>0\\\mathcal{R}_{\min\{\lvert\nu_1\rvert,\lvert\nu_2\rvert\}} & \text{if } \rho_1=0,\rho_2=0\end{cases}$ where $\mu=\frac{1}{\lvert\rho_1\rvert}+\frac{1}{\lvert\rho_2\rvert}=\frac{\lvert\rho_1\rvert+\lvert\rho_2\rvert}{\lvert\rho_1\rho_2\rvert}$, $\sigma=\mu(\sigma_1\lvert\rho_1\rvert)^{\frac{1}{\mu\lvert\rho_1\rvert}}(\sigma_2\lvert\rho_2\rvert)^{\frac{1}{\mu\lvert\rho_2\rvert}}$.
Product of Densities	$(\nu_1,\sigma_1,\rho_1)\&(\nu_2,\sigma_2,\rho_2)\equiv\begin{cases}(\nu_1+\nu_2,\sigma_1,\rho_1) & \text{if } \rho_1<\rho_2\\(\nu_1+\nu_2,\sigma_1+\sigma_2,\rho) & \text{if } \rho=\rho_1=\rho_2\\(\nu_1+\nu_2,\sigma_2,\rho_2) & \text{otherwise.}\end{cases}$
Functions (L-Lipschitz)	$f(X_1,\ldots,X_n)\equiv L\max\{X_1,\ldots,X_n\}$

Table 0.1: Operations on random variables (e.g.X_1+X_2) are viewed as actions on density functions (e.g. convolution $(\nu_1,\sigma_1,\rho_1)\oplus(\nu_2,\sigma_2,\rho_2)$) and the tail parameters of the result are analyzed and reported.

(ii) The set $\mathcal{G}$ is known to be closed under additive convolution, positive powers, and Lipschitz functions — we will show it is closed under multiplicative convolution as well. This covers the majority of elementary operations on independent random variables, with reciprocals, exponentials and logarithms the only exceptions. However, we will introduce a few "tricks" to handle these cases as well.

The full list of operations in GGA is compiled in table 0.1. All operations in the GGA can be proven to exhibit identical behaviour with their corresponding operations on random variables, with the sole exception of reciprocals (marked by asterisk), where additional assumptions are

required.

Illustrative examples

To further illustrate the GGA through example, in this section we work out explicit GGA computations using distributions from table 0.2 and operations in table 0.1 and recover some common probability identities.

Example 1 (Chi-squared random variables) *Let $X_1, \ldots, X_k$ be k independent standard normal random variables. The variable $Z = \sum_{i=1}^{k} X_i^2$ is* chi-squared distributed *with k degrees of freedom. Using the generalized Gamma algebra, we can accurately determine the tail behaviour of this random variable directly from its construction. Recall that each $X_i \equiv (0, 1/2, 2)$, and by the power operation, $X_i^2 \equiv (-1/2, 1/2, 1)$. Applying the addition operation k times reveals that $Z \equiv (k/2 - 1, 1/2, 1)$ and implies that the density of Z is asymptotically $cx^{k/2-1}e^{-x/2}$ as $x \to \infty$. In fact, the density of Z is exactly $p_Z(x) = c_k x^{k/2-1}e^{-x/2}$ where $c_k = 2^{-k/2}/\Gamma(k/2)$.*

Example 2 (Products of random variables) *To demonstrate the efficacy of the multiplication operation in our algebra, we consider the product of two exponential, Gaussian, and reciprocal Gaussian random variables. In* **??**, *we manually prove the following.*

Lemma 1 *Let $X_1, X_2 \sim Exp(\lambda)$ and $Z_1, Z_2 \sim \mathcal{N}(0, 1)$ be independent. The densities of X_1X_2, Z_1Z_2 and $Z = 1/Z_1 \cdot 1/Z_2$ satisfy as $x \to \infty$,*

$$p_{X_1X_2}(x) \sim \frac{\lambda^{3/2}\sqrt{\pi}}{x^{1/4}} e^{-2\lambda\sqrt{x}}, \quad p_{Z_1Z_2}(x) \sim \frac{1}{\sqrt{2\pi x}} e^{-x}, \quad p_Z(x) \sim \frac{1}{\sqrt{2\pi}|z|^{3/2}} e^{-1/|z|}.$$

With ease, our algebra correctly determines that $X_1X_2 \equiv (-\frac{1}{4}, 2\lambda, \frac{1}{2})$, $Z_1Z_2 \equiv (-\frac{1}{2}, 1, 1)$ and $Z \equiv (-\frac{3}{2}, 1, -1)$.

Example 3 (Reciprocal distributions) *Perhaps the most significant challenge with a tail algebra is correctly identifying the tail behaviour of reciprocal distributions. Here, we test the efficacy of our formulation with known reciprocal distributions.*

- Reciprocal normal: *$X \sim \mathcal{N}(0, 1) \equiv (0, 1/2, 2)$, and $X^{-1} \equiv (-2, 1/2, -2)$.*
- Inverse exponential: *$X \sim Exp(\lambda) \equiv (0, \lambda, 1)$, and $X^{-1} \equiv (-2, \lambda, -1)$.*
- Inverse t-distribution: *$X \equiv \mathcal{R}_\nu$, and $X^{-1} \equiv \mathcal{R}_2$.*
- Inverse Cauchy: *$X \equiv \mathcal{R}_2$, it is known X^{-1} has the same distribution and our theory predicts $X^{-1} \equiv \mathcal{R}_2$.*

Example 4 (Cauchy distribution) *A simple special case of the Student T distribution is the Cauchy distribution, which arises as the ratio of two standard normal random variables. For $X \sim \mathcal{N}(0,1)$, $X \equiv (0, 1/2, 2)$ and $X^{-1} \equiv (-2, 1/2, -2)$. Hence, the multiplication operation correctly predicts that the ratio of two standard normal random variables is in $\mathcal{R}_2$.*

Example 5 (Student T distribution) *Let X be a standard normal random variable, and V a chi-squared random variable with ν degrees of freedom. The random variable $T = X/\sqrt{V/\nu}$ is* t-distributed *with ν degrees of freedom. Since $V \equiv (\nu/2 - 1, 1/2, 1)$, multiplying by the constant $1/\nu$ reveals $V/\nu \equiv (\nu/2 - 1, 1/(2\nu), 1)$. Applying the square root operation, $\sqrt{V/\nu} \equiv (\nu - 1, 1/(2\nu), 2)$. To compute the division operation, we first take the reciprocal to find $(V/\nu)^{-1/2} \equiv (-\nu - 1, 1/(2\nu), -2)$. Finally, since $\rho = -2 < 1$ for this random variable, the multiplication operation with $X \equiv (0, 1/2, 2)$ yields $T \equiv \mathcal{R}_{\nu+1}$, and so the density of T is asymptotically $cx^{-\nu-1}$ as $x \to \infty$. Indeed, the density of T satisfies $p_T(x) = c_\nu(1 + x^2/\nu)^{-(\nu+1)/2}$ where $c_\nu = \Gamma(\frac{\nu+1}{2})/\Gamma(\frac{\nu}{2})(\nu\pi)^{-1/2}$, which exhibits the predicted tail behaviour.*

Example 6 *Log-normal distribution Although the log-normal distribution does not lie in $\mathcal{G}$, the existence of log-normal tails arising from the multiplicative central limit theorem is suggested by our algebra. Let $X_1, X_2, \ldots$ be independent standard normal random variables and let $Z_k = X_1 \cdots X_{2^k}$ for each $k = 1, 2, \ldots$. By the multiplicative central limit theorem, letting $\tau = \exp(\mathbb{E} \log |X_i|) \approx 1.13$, $\left(\frac{X_1 \cdots X_n}{\tau}\right)^{1/\sqrt{n}}$ converges in distribution as $n \to \infty$ to a log-normal random variable Z with density*

$$p_Z(x) = \frac{1}{x\sqrt{2\pi}} \exp(-\tfrac{1}{2}(\log x)^2).$$

Therefore, the same is true for $V_k = (Z_k/\tau)^{2^{-k/2}}$. Using our algebra, we will attempt to reproduce the tail of this density. Letting $\tilde{Z}_k = X_{2^k} \cdots X_{2^{k+1}}$, we see that $Z_{k+1} = Z_k \tilde{Z}_k$, and $Z_k, \tilde{Z}_k$ are iid. Let $Z_k \equiv (\nu_k, \sigma_k, \rho_k)$, by induction using the multiplication operation, we find that $\nu_{k+1} = \frac{1}{\mu}\left(\frac{2\nu_k}{\rho_k} - \frac{1}{2}\right) = \nu_k - \frac{\rho_k}{4}$, $\sigma_{k+1} = \mu\,(\sigma_k\rho_k)^{\frac{2}{\mu\rho_k}} = \frac{2}{\rho_k}\,(\sigma_k\rho_k) = 2\sigma_k$, and $\rho_{k+1} = \frac{1}{\mu} = \frac{\rho_k}{2}$. Since $\rho_0 = 2$, $\sigma_0 = 1/2$, and $\nu_0 = 0$, we find that $\rho_k = 2^{1-k}$ and $\sigma_k = 2^{k-1}$. Furthermore, $\nu_{k+1} = \nu_k - 2^{-k-1}$ and so $\nu_k = -1 + 2^{-k}$. Therefore $Z_k \equiv (-1 + 2^{-k}, 2^{k-1}, 2^{1-k})$, and

$$V_k \equiv (-1 + 2^{-k/2}, 2^{k-1}\tau^{-2^{1-k}}, 2^{1-k/2}),$$

and letting $\epsilon_k = 2^{-k/2}$, the tail behaviour of the density of V_k satisfies

$$\begin{aligned} p_k(x) &\sim c_k x^{-1+\epsilon_k} \exp\left(-\frac{\epsilon_k^{-2}}{2\tau^{-2\epsilon_k^2}} x^{2\epsilon_k}\right) \\ &\sim c_k x^{-1+\epsilon_k} \exp\left(-\frac{1}{2\tau^{-2\epsilon_k^2}}\left(\frac{x^{\epsilon_k} - 1}{\epsilon_k}\right)^2\right) \approx c_k x^{-1} \exp\left(-\frac{1}{2}(\log x)^2\right), \end{aligned}$$

as $x \to \infty$, where the approximation improves as k gets larger. The quality of this approximation is shown in fig. 0.2.

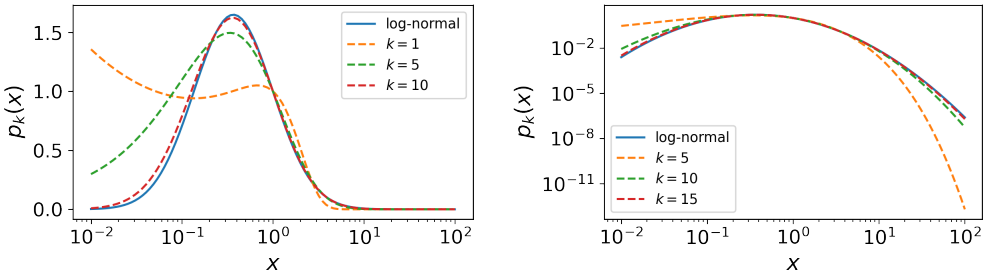

Figure 0.2: Estimation of the log-normal density by tail algebra applied to V_k.

Operations in the Generalized Gamma Algebra

Whereas table 0.1 provides a summary of our theory useful for referencing, in this section we provide additional explanation and references for how operations on random variables affect their GGA tails.

Ordering. A total ordering is imposed on the equivalence classes of $\mathcal{G}$ according to the heaviness of tails. In particular, we say that $(\nu_1, \sigma_1, \rho_1) \leq (\nu_2, \sigma_2, \rho_2)$ if $(x^{\nu_1} e^{-\sigma_1 x^{\rho_1}})/(x^{\nu_2} e^{-\sigma_2 x^{\rho_2}})$ is bounded as $x \to \infty$. As usual, we say $(\nu_1, \sigma_1, \rho_1) < (\nu_2, \sigma_2, \rho_2)$ if $(\nu_1, \sigma_1, \rho_1) \leq (\nu_2, \sigma_2, \rho_2)$ but $(\nu_1, \sigma_1, \rho_1) \not\equiv (\nu_2, \sigma_2, \rho_2)$.

Addition. Tails of this form are closed under addition. Combining subexponentiality for $\rho < 1$ [AA10, Chapter X.1], with [Asm+17, Thm 3.1 & eqn. (8.3)],

Proposition 1 *Denoting the addition of random variables (additive convolution of densities) by* $\oplus$*,*

$$(\nu_1, \sigma_1, \rho_1) \oplus (\nu_2, \sigma_2, \rho_2) \equiv \begin{cases} \max\{(\nu_1, \sigma_1, \rho_1), (\nu_2, \sigma_2, \rho_2)\} & \text{if } \rho_1 \neq \rho_2 \text{ or } \rho_1, \rho_2 < 1 \\ (\nu_1 + \nu_2 + 1, \min\{\sigma_1, \sigma_2\}, 1) & \text{if } \rho_1 = \rho_2 = 1 \\ (\nu_1 + \nu_2 + 1 - \frac{\rho}{2}, (\sigma_1^{-\frac{1}{\rho-1}} + \sigma_2^{-\frac{1}{\rho-1}})^{1-\rho}, \rho) & \text{if } \rho = \rho_1 = \rho_2 > 1. \end{cases} \tag{3}$$

Powers. For all exponents $\beta > 0$, by invoking a change of variables $x \mapsto x^\beta$, it is easy to show that $(\nu, \sigma, \rho)^\beta \equiv \left(\frac{\nu+1}{\beta} - 1, \sigma, \frac{\rho}{\beta}\right)$. We *define* negative powers and reciprocals equivalently to positive powers in the case $\beta < 0$. This equivalence cannot be proven to hold in general since we cannot determine tail asymptotics of the reciprocal without knowledge of its behaviour around zero. Therefore, we implicitly assume that the behaviour around zero mimics the tail behaviour, that is, eq. (1) holds as $x \to 0^+$. However, this can only hold provided

$(\nu+1)/\rho > 0$ and $\rho \neq 0$. In all other cases, including $\mathcal{R}_\nu$, we assume that the density of X approaches a nonzero value near zero, and define the reciprocal to be $\mathcal{R}_2$.

Multiplication. For any $c \in \mathbb{R}\backslash\{0\}$, it can be readily seen from a change of variables $x \mapsto cx$ that $c(\nu, \sigma, \rho) = (\nu, \sigma/|c|^\rho, \rho)$. The class $\mathcal{G}$ is also closed under multiplication (assuming independence of random variables), as we show in the following result — the proof is delayed to Appendix C.

Proposition 2 *Denoting the multiplication of independent random variables (multiplicative convolution) by* $\otimes$*,*

$$(\nu_1, \sigma_1, \rho_1) \otimes (\nu_2, \sigma_2, \rho_2) \equiv \begin{cases} \left(\frac{1}{\mu}\left(\frac{\nu_1}{|\rho_1|} + \frac{\nu_2}{|\rho_2|} + \frac{1}{2}\right), \sigma, -\frac{1}{\mu}\right) & \text{if } \rho_1, \rho_2 < 0 \\ \left(\frac{1}{\mu}\left(\frac{\nu_1}{\rho_1} + \frac{\nu_2}{\rho_2} - \frac{1}{2}\right), \sigma, \frac{1}{\mu}\right) & \text{if } \rho_1, \rho_2 > 0 \\ \mathcal{R}_{|\nu_1|} & \text{if } \rho_1 \leq 0, \rho_2 > 0 \\ \mathcal{R}_{\min\{|\nu_1|, |\nu_2|\}} & \text{if } \rho_1 = 0, \rho_2 = 0 \end{cases}$$

where $\mu = \frac{1}{|\rho_1|} + \frac{1}{|\rho_2|} = \frac{|\rho_1|+|\rho_2|}{|\rho_1\rho_2|}$ *and* $\sigma = \mu(\sigma_1|\rho_1|)^{\frac{1}{\mu|\rho_1|}}(\sigma_2|\rho_2|)^{\frac{1}{\mu|\rho_2|}}$.

Product of Densities. We can also consider a product of densities operation acting on two random variables X, Y, denoted $X\&Y$, by $p_{X\&Y}(x) = cp_X(x)p_Y(x)$, where $c > 0$ is an appropriate normalizing constant and $p_X, p_Y, p_{X\&Y}$ are the densities of X, Y, and $X\&Y$, respectively. In terms of the equivalence classes:

$$(\nu_1, \sigma_1, \rho_1)\&(\nu_2, \sigma_2, \rho_2) \equiv \begin{cases} (\nu_1 + \nu_2, \sigma_1, \rho_1) & \text{if } \rho_1 < \rho_2 \\ (\nu_1 + \nu_2, \sigma_1 + \sigma_2, \rho) & \text{if } \rho = \rho_1 = \rho_2 \\ (\nu_1 + \nu_2, \sigma_2, \rho_2) & \text{otherwise.} \end{cases}$$

Note that this particular operation does not require either p_X or p_Y to be normalized — only the tail behaviour is needed. We may also use this to work out the tail behaviour of a posterior density, provided the tail behaviour of the likelihood in the parameters is known.

Lipschitz Functions. There are many multivariate functions that cannot be readily represented in terms of the operations covered thus far. For these, it is important to specify the tail behaviour of pushforward measures under Lipschitz-continuous functions. Fortunately, this is covered by lemma 2 below, presented in [Led01, Proposition 1.3]. Hölder-continuous functions can also be represented as a composition of a power operation and a Lipschitz-continuous function.

Lemma 2 *For any Lipschitz continuous function* $f : \mathbb{R}^d \to \mathbb{R}$ *satisfying* $\|f(x) - f(y)\| \leq L\|x - y\|$ *for* $x, y \in \mathbb{R}^d$*, there is* $f(X_1, \ldots, X_d) \equiv L \max\{X_1, \ldots, X_d\}$.

Power Law Approximation. Note that as $x \to \infty$, $p_{|X|}(x) \sim cx^{\nu}e^{-\sigma x^{\rho}} = \tilde{c}x^{\nu}e^{-\sigma\rho\frac{x^{\rho}-1}{\rho}} \approx \tilde{c}x^{\nu}e^{-\sigma\rho\log x}$

$$p_{|X|}(x) \sim cx^{\nu}e^{-\sigma x^{\rho}} = \tilde{c}x^{\nu}e^{-\sigma(x^{\rho}-0)} = \tilde{c}x^{\nu}e^{-\sigma\rho\frac{x^{\rho}-1}{\rho}} \approx \tilde{c}x^{\nu}e^{-\sigma\rho\log x} = \tilde{c}x^{\nu-\sigma\rho},$$

where we have used the approximation $\log x = \rho^{-2}(x^{\rho}-1) + \mathcal{O}(\rho^2)$. Consequently, we can represent tails of this form by the Student t distribution with $|\nu - \sigma\rho| - 1$ degrees of freedom. In practice, we find this approximation tends to *overestimate* the heaviness of the tail. Alternatively, the generalized Gamma density (2) satisfies $\mathbb{E}X^r = \sigma^{-r/\rho}\Gamma(\frac{\nu+1+r}{\rho})/\Gamma(\frac{\nu+1}{\rho})$ for $r > 0$. Let $\alpha > 0$ be such that $\mathbb{E}X^{\alpha} = 2$. By Markov's inequality, the tail of X satisfies $\mathbb{P}(X > x) \leq 2x^{-\alpha}$. Therefore, we can represent tails of this form by the Student t distribution with $\alpha + 1$ degrees of freedom (generate $X \sim t_{\alpha}$). In practice, we find this approximation to be more accurate, and is hence used in Section 4.1.

List of univariate distributions

Here we provide an enumeration of common parametric distributions and their corresponding GGA parameterizations.

Table 0.2: List of univariate distributions

Name	Support	Density $p(x)$	Class
Benktander Type II	$(0,\infty)$	$e^{\frac{a}{b}(1-x^b)}x^{b-2}(ax^b - b + 1)$	$(2b-2, \frac{a}{b}, b)$
Beta prime distribution	$(0,\infty)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1+x)^{-\alpha-\beta}$	$\mathcal{R}_{\beta+1}$
Burr distribution	$(0,\infty)$	$ckx^{c-1}(1+x^c)^{-k-1}$	$\mathcal{R}_{ck+1}$
Cauchy distribution	$(-\infty,\infty)$	$(\pi\gamma)^{-1}\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]^{-1}$	$\mathcal{R}_2$
Chi distribution	$(0,\infty)$	$\frac{1}{2^{k/2-1}\Gamma(k/2)}x^{k-1}e^{-x^2/2}$	$(k-1, \frac{1}{2}, 2)$
Chi-squared distribution	$(0,\infty)$	$\frac{1}{2^{k/2}\Gamma(k/2)}x^{\frac{k}{2}-1}e^{-x/2}$	$(\frac{k}{2}-1, \frac{1}{2}, 1)$
Dagum distribution	$(0,\infty)$	$\frac{ap}{x}\left(\frac{x}{b}\right)^{ap}\left(\left(\frac{x}{b}\right)^a + 1\right)^{-p-1}$	$\mathcal{R}_{a+1}$
Davis distribution	$(0,\infty)$	$\propto (x-\mu)^{-1-n}/\left(e^{\frac{b}{x-\mu}} - 1\right)$	$(-1-n, b, -1)$
Exponential distribution	$(0,\infty)$	$\lambda e^{-\lambda x}$	$(0, \lambda, 1)$

F distribution	$(0, \infty)$	$\propto x^{d_1/2-1}(d_1 x + d_2)^{-(d_1+d_2)/2}$	$\mathcal{R}_{d_2/2+1}$
Fisher z-distribution	$(-\infty, \infty)$	$\propto \frac{e^{d_1 x}}{(d_1 e^{2x}+d_2)^{(d_1+d_2)/2}}$	$(0, d_2, 1)$
Frechet distribution	$(0, \infty)$	$\frac{\alpha}{\lambda}\left(\frac{x-m}{\lambda}\right)^{-1-\alpha} e^{-\left(\frac{x-m}{\lambda}\right)^{-\alpha}}$	$(-1-\alpha, \lambda^{\alpha}, -\alpha)$
Gamma distribution	$(0, \infty)$	$\frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$(\alpha - 1, \beta, 1)$
Gamma/Gompertz distribution	$(0, \infty)$	$bse^{bx}\beta^{s}/(\beta - 1 + e^{bx})^{s+1}$	$(0, bs, 1)$
Gen. hyperbolic distribution	$(-\infty, \infty)$	$\propto e^{\beta(x-\mu)} \frac{K_{\lambda-1/2}(\alpha\sqrt{\delta^2+(x-\mu)^2})}{(\delta^2+(x-\mu)^2)^{1/4-\lambda/2}}$	$(\lambda - 1, \alpha - \beta, 1)$
Gen. Normal distribution	$(-\infty, \infty)$	$\frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left(-\left(\frac{\lvert x-\mu\rvert}{\alpha}\right)^{\beta}\right)$	$(0, \alpha^{-\beta}, \beta)$
Geometric stable distribution	$(-\infty, \infty)$	no closed form	$\mathcal{R}_{\alpha+1}$
Gompertz distribution	$(0, \infty)$	$\sigma\eta \exp(\eta + \sigma x - \eta e^{\sigma x})$	$\mathcal{L}$
Gumbel distribution	$(0, \infty)$	$\beta^{-1} e^{-(\beta^{-1}(x-\mu)+e^{-\beta^{-1}(x-\mu)})}$	$(0, \frac{1}{\beta}, 1)$
Gumbel Type II distribution	$(0, \infty)$	$\alpha\beta x^{-\alpha-1} e^{-\beta x^{-\alpha}}$	$(-\alpha - 1, \beta, -\alpha)$
Holtsmark distribution	$(-\infty, \infty)$	no closed form	$\mathcal{R}_{5/2}$
Hyperbolic secant distribution	$(-\infty, \infty)$	$\frac{1}{2}\text{sech}\left(\frac{\pi x}{2}\right)$	$(0, \frac{\pi}{2}, 1)$
Inv. chi-squared distribution	$(0, \infty)$	$\frac{2^{-k/2}}{\Gamma(k/2)} x^{-k/2-1} e^{-1/(2x)}$	$(-\frac{k}{2} - 1, \frac{1}{2}, -1)$
Inv. gamma distribution	$(0, \infty)$	$\frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	$(-\alpha - 1, \beta, -1)$
Levy distribution	$(0, \infty)$	$\sqrt{\frac{c}{2\pi}}(x-\mu)^{-3/2} e^{-\frac{c}{2(x-\mu)}}$	$(-\frac{3}{2}, \frac{c}{2}, -1)$
Laplace distribution	$(-\infty, \infty)$	$\frac{1}{2\lambda} \exp\left(-\frac{\lvert x-\mu\rvert}{\lambda}\right)$	$(0, \frac{1}{\lambda}, 1)$
Logistic distribution	$(-\infty, \infty)$	$\frac{e^{-(x-\mu)/\lambda}}{\lambda(1+e^{-(x-\mu)/\lambda})^2}$	$(0, \frac{1}{\lambda}, 1)$
Log-Cauchy distribution	$(0, \infty)$	$\frac{\sigma}{x\pi}((\log x - \mu)^2 + \sigma^2)^{-1}$	$\mathcal{R}_1$
Log-Laplace distribution	$(0, \infty)$	$\frac{1}{2\lambda x} \exp\left(-\frac{\lvert\log x-\mu\rvert}{\lambda}\right)$	$\mathcal{R}_{1/\lambda+1}$
Log-logistic distribution	$(0, \infty)$	$\frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\left(1+\left(\frac{x}{\alpha}\right)^{\beta}\right)^{-2}$	$\mathcal{R}_{\beta+1}$

Log-t distribution	$(0,\infty)$	$\propto x^{-1}(1+\frac{1}{\nu}(\log x-\mu)^2)^{-\frac{\nu+1}{2}}$	$\mathcal{R}_1$
Lomax distribution	$(0,\infty)$	$\frac{\alpha}{\lambda}\left(1+\frac{x}{\lambda}\right)^{-\alpha-1}$	$\mathcal{R}_{\alpha+1}$
Maxwell-Boltzmann distribution	$(0,\infty)$	$\sqrt{\frac{2}{\pi}}\frac{x^2e^{-x^2/(2\sigma^2)}}{\sigma^3}$	$(2,\frac{1}{2\sigma^2},2)$
Normal distribution	$(-\infty,\infty)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$	$(0,\frac{1}{2\sigma^2},2)$
Pareto distribution	(x_0,∞)	$\alpha x_0^\alpha x^{-\alpha-1}$	$\mathcal{R}_{\alpha+1}$
Rayleigh distribution	$(0,\infty)$	$\frac{x}{\sigma^2}e^{-x^2/(2\sigma^2)}$	$(1,\frac{1}{2\sigma^2},2)$
Rice distribution	$(0,\infty)$	$\frac{x}{\sigma^2}\exp\left(-\frac{(x^2+\nu^2)}{2\sigma^2}\right)I_0\left(\frac{x\nu}{\sigma^2}\right)$	$(\frac{1}{2},\frac{1}{2\sigma^2},2)$
Skew normal distribution	$(-\infty,\infty)$	no closed form	$(0,\frac{1}{2\sigma^2},2)$
Slash distribution	$(-\infty,\infty)$	$\frac{1-e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}x^2}$	$(-2,\frac{1}{2},2)$
Stable distribution	$(-\infty,\infty)$	no closed form	$\mathcal{R}_{\alpha+1}$
Student's t-distribution	$(-\infty,\infty)$	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	$\mathcal{R}_{\nu+1}$
Tracy-Widom distribution	$(-\infty,\infty)$	no closed form	$(-\frac{3\beta}{4}-1,\frac{2\beta}{3},\frac{3}{2})$
Voigt distribution	$(-\infty,\infty)$	no closed form	$\mathcal{R}_2$
Weibull distribution	$(0,\infty)$	$\frac{\rho}{\lambda}\left(\frac{x}{\lambda}\right)^{\rho-1}e^{-(x/\lambda)^\rho}$	$(\rho-1,\lambda^{-\rho},\rho)$

The following densities are not supported by our algebra: Benini distribution; Benktander Type I distribution; Johnson's S_U-distribution; and the log-normal distribution. All of these densities exhibit log-normal tails.

Proofs of new results

Proof [Proof of Lemma 1] The proof relies on the following integral definition [Wat95, pg. 183] and asymptotic relation as $z\to\infty$ [Wat95, pg. 202] of the modified Bessel function $K_\nu(z)$ for $z>0$ and $\nu\geq 0$,

$$K_\nu(z)=\frac{1}{2}\left(\frac{z}{2}\right)^\nu\int_0^\infty u^{-\nu-1}\exp\left(-u-\frac{z^2}{4u}\right)\mathrm{d}u\sim\sqrt{\frac{\pi}{2z}}e^{-z}. \tag{4}$$

We also make use of the known density for the product of two independent continuous random variables: if X and Y have densities p_X and p_Y respectively, then $Z = XY$ has density

$$p_Z(z) = \int_{\mathbb{R}} p_X(x) p_Y(z/x) |x|^{-1} \mathrm{d}x.$$

Exponentials. Recalling that the density of $X \sim \mathrm{Exp}(\lambda)$ is $p_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, for $Z = XY$ where $X \sim \mathrm{Exp}(\lambda_1)$ and $Y \sim \mathrm{Exp}(\lambda_2)$ are independent,

$$p_Z(z) = \int_0^\infty x^{-1} \lambda_1 e^{-\lambda_1 x} \lambda_2 e^{-\lambda_2 z/x} \mathrm{d}x = \lambda_1 \lambda_2 \int_0^\infty x^{-1} e^{-\lambda_1 x - \lambda_2 z/x} \mathrm{d}x.$$

Since $2K_0(2\sqrt{z}) = \int_0^\infty u^{-1} \exp(-u - \frac{z}{u}) \mathrm{d}u$, let $u = \lambda_1 v$, so that $\mathrm{d}u = \lambda_1 \mathrm{d}v$,

$$2K_0(2\sqrt{\lambda_1 \lambda_2 z}) = \int_0^\infty u^{-1} \exp\left(-\lambda_1 v - \lambda_2 \frac{z}{v}\right) \mathrm{d}v.$$

Therefore, letting $\lambda = \sqrt{\lambda_1 \lambda_2}$,

$$p_Z(z) = 2\lambda^2 K_0(2\lambda\sqrt{z}) \sim \sqrt{\pi} \lambda^{3/2} z^{-1/4} e^{-2\lambda z^{1/2}}.$$

Normals. Recalling that the density of $X \sim \mathcal{N}(0,1)$ is $p_X(x) = (2\pi)^{-1/2} \exp(-\frac{1}{2}x^2)$, for $Z = XY$ where $X, Y \sim \mathcal{N}(0,1)$ are independent,

$$\begin{aligned} p_Z(z) &= \frac{1}{2\pi} \int_{\mathbb{R}} |x|^{-1} e^{-\frac{1}{2}x^2} e^{-\frac{1}{2}z^2/x^2} \mathrm{d}x \\ &= \frac{1}{\pi} \int_0^\infty x^{-1} e^{-\frac{1}{2}x^2 - \frac{1}{2}z^2/x^2} \mathrm{d}x \\ &= \frac{1}{\pi} \int_0^\infty x^{-1} e^{-\frac{1}{2}x^2 - \frac{1}{2}z^2/x^2} \mathrm{d}x. \end{aligned}$$

Let $u = \frac{1}{2}x^2$ so that $\mathrm{d}u = x\mathrm{d}x$ and

$$K_\nu(z) = z^\nu \int_0^\infty x^{-2\nu-1} \exp\left(-\frac{1}{2}x^2 - \frac{z^2}{2x^2}\right) \mathrm{d}x.$$

In particular, for any $z \in \mathbb{R}$,

$$K_0(|z|) = \int_0^\infty x^{-1} \exp\left(-\frac{1}{2}x^2 - \frac{z^2}{2x^2}\right) \mathrm{d}x, \tag{5}$$

and so

$$p_Z(z) = \frac{1}{\pi} K_0(|z|) \sim \frac{1}{\sqrt{2\pi|z|}} e^{-|z|}.$$

Reciprocal Normals. Finally, by a change of variables, we note that the density of X^{-1} where $X \sim \mathcal{N}(0,1)$ is $p_{X^{-1}}(x) = (2\pi)^{-1/2}x^{-2}\exp(-\frac{1}{2x^2})$. Therefore, the density of $Z = 1/(XY)$ where $X, Y \sim \mathcal{N}(0,1)$ are independent is given by

$$\begin{aligned} p_Z(z) &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}x^2} e^{-\frac{1}{2x^2}} \frac{x^2}{\sqrt{2\pi}z^2} e^{-\frac{x^2}{2z^2}} \frac{1}{|x|} \mathrm{d}x \\ &= \frac{1}{2\pi z^2} \int_{\mathbb{R}} e^{-\frac{1}{2x^2}-\frac{x^2}{2z^2}} \frac{1}{|x|} \mathrm{d}x \\ &= \frac{1}{\pi z^2} \int_0^\infty e^{-\frac{1}{2x^2}-\frac{x^2}{2z^2}} \frac{1}{x} \mathrm{d}x \\ &= \frac{1}{\pi z^2} K_0(|z|^{-1}) \sim \sqrt{\frac{1}{2\pi}} |z|^{-3/2} e^{-|z|^{-1}}, \end{aligned}$$

where we have once again used (5). ■

Recall that the Mellin transform of a function f on $(0,\infty)$ is given by

$$\mathcal{M}_s[f] = \int_0^\infty x^{s-1} f(x) \mathrm{d}x.$$

Letting p_{XY} denote the density of the product of independent random variables X, Y with respective densities p_X and p_Y, $\mathcal{M}_s[p_{XY}] = \mathcal{M}_s[p_X]\mathcal{M}_s[p_Y]$. There is

$$\mathcal{M}_s[cx^\nu e^{-\sigma x^\rho}] = \frac{c\sigma^{-\nu/\rho}}{\rho} \sigma^{-s/\rho} \Gamma\left(\frac{\nu}{\rho} + \frac{s}{\rho}\right).$$

To facilitate the proof of Proposition 2, we define the Fox H-function

$$H_{p,q}^{m,n}\left[z \left| \begin{smallmatrix} (a_1,A_1),\ldots,(a_p,A_p) \\ (b_1,B_1),\ldots,(b_q,B_q) \end{smallmatrix} \right.\right]$$

as the inverse Mellin transform of

$$\Theta(s) = z^{-s} \frac{\prod_{j=1}^m \Gamma(b_j + B_j s) \cdots \prod_{j=1}^n \Gamma(1 - a_j - A_j s)}{\prod_{j=m+1}^q \Gamma(1 - b_j - B_j s) \prod_{j=n+1}^p \Gamma(a_j + A_j s)}.$$

An important property of the Fox H-function is its asymptotic behaviour as $z \to \infty$. From [MSH09, Theorem 1.3],

$$H_{p,q}^{q,0}\left[z \left| \begin{smallmatrix} (a_1,A_1),\ldots,(a_p,A_p) \\ (b_1,B_1),\ldots,(b_q,B_q) \end{smallmatrix} \right.\right] \sim c x^{(\delta+\frac{1}{2})/\mu} \exp(-\mu\beta^{-1/\mu} x^{1/\mu}), \qquad \text{as } x \to \infty,$$

for some constant $c > 0$, where $\beta = \prod_{j=1}^p (A_j)^{-A_j} \prod_{j=1}^q B_j^{B_j}$, $\mu = \sum_{j=1}^q B_j - \sum_{j=1}^p A_j$, and $\delta = \sum_{j=1}^q b_j - \sum_{j=1}^p a_j + \frac{p-q}{2}$.

Proof [Proof of Proposition 2] The $\rho_1 \leq 0, \rho_2 > 0$ and $\rho_1 = \rho_2 = 0$ cases follow from Breiman's lemma [BDM16, Lemma B.5.1]. Our argument proceeds similar to [Asm+17]. Assume that $\rho_1, \rho_2 > 0$ and let $0 < \epsilon < 1$ be such that $0 < a_- < a_+ < 1$, where

$$a_+ = \frac{(1+\epsilon)\rho_2}{\rho_1 + \rho_2}, \qquad a_- = 1 - \frac{(1+\epsilon)\rho_1}{\rho_1 + \rho_2}.$$

Then for $\rho = \frac{\rho_1\rho_2}{\rho_1+\rho_2}$, if $X \equiv (\nu_1, \sigma_1, \rho_1)$ and $Y \equiv (\nu_2, \sigma_2, \rho_2)$, then

$$\begin{aligned}
\mathbb{P}(XY > x, X \notin [x^{a_-}, x^{a_+}]) &\leq \mathbb{P}(X > x^{a_+}) + \mathbb{P}(Y > x^{1-a_-}) \\
&\sim c_1 x^{\nu_1 a_+} e^{-\sigma_1 x^{\rho_1 a_+}} + c_2 x^{\nu_2(1-a_-)} e^{-\sigma_2 x^{\rho_2(1-a_-)}} \\
&\leq \left(c_1 x^{\nu_1 a_+} + c_2 x^{\nu_2(1-a_-)}\right) e^{-\min\{\sigma_1,\sigma_2\} x^{(1+\epsilon)\rho}} = o(x^\nu e^{-\sigma x^\rho}),
\end{aligned}$$

for any $\nu, \sigma > 0$. Hence, it will suffice to show the claimed tail asymptotics for the generalized Gamma distribution. In this case, since $a_- > 0$ and $a_+ < 1$, the tail of the distribution for the product of X, Y depends only on the tail of the distributions for X and Y.

Therefore, assume without loss of generality that $p_X(x) = c_X x^{\nu_1} e^{-\sigma_1 x^{\rho_1}}$ and $p_Y(x) = c_Y x^{\nu_2} e^{-\sigma_2 x^{\rho_2}}$. Then

$$\mathcal{M}_s[p_{XY}] = c_X c_Y \frac{\sigma_1^{-\nu_1/\rho_1}}{\rho_1} \frac{\sigma_2^{-\nu_2/\rho_2}}{\rho_2} \left(\sigma_1^{1/\rho_1}\sigma_2^{1/\rho_2}\right)^{-s} \Gamma\left(\frac{\nu_1}{\rho_1} + \frac{s}{\rho_1}\right) \Gamma\left(\frac{\nu_2}{\rho_2} + \frac{s}{\rho_2}\right).$$

Consequently,

$$p_{XY}(z) = c_X c_Y \frac{\sigma_1^{-\nu_1/\rho_1}}{\rho_1} \frac{\sigma_2^{-\nu_2/\rho_2}}{\rho_2} H_{p,q}^{m,n}\left[\sigma_1^{1/\rho_1}\sigma_2^{1/\rho_2} z \,\middle|\, \overline{(\tfrac{\nu_1}{\rho_1}, \tfrac{1}{\rho_1}), (\tfrac{\nu_2}{\rho_2}, \tfrac{1}{\rho_2})}\right]$$

Computing the corresponding β, δ, μ for the asymptotic expansion, we find that

$$\mu = \frac{1}{\rho_1} + \frac{1}{\rho_2}, \qquad \delta = \frac{\nu_1}{\rho_1} + \frac{\nu_2}{\rho_2} - 1, \qquad \beta = \rho_1^{-1/\rho_1}\rho_2^{-1/\rho_2}.$$

Consequently, for some $c > 0$,

$$p_{XY}(z) \sim c z^{\frac{1}{\mu}(\frac{1}{2}+\delta)} \exp\left(-\mu\beta^{-\frac{1}{\mu}}(\sigma_1^{1/\rho_1}\sigma_2^{1/\rho_2})^{\frac{1}{\mu}} z^{\frac{1}{\mu}}\right),$$

which completes the $\rho_1, \rho_2 > 0$ case. The final case follows by composing the multiplication and reciprocal operations. Note that

$$\begin{aligned}
(\nu_1, \sigma_1, -\rho_1)^{-1} \otimes (\nu_2, \sigma_2, -\rho_2)^{-1} &\equiv (-\nu_1 - 2, \sigma_1, \rho_1) \otimes (-\nu_2 - 2, \sigma_2, \rho_2) \\
&\equiv \left(\frac{1}{\mu}\left(\frac{-\nu_1 - 2}{\rho_1} + \frac{-\nu_2 - 2}{\rho_2} - \frac{1}{2}\right), \sigma, \frac{1}{\mu}\right) \\
&\equiv \left(\frac{1}{\mu}\left(\frac{-\nu_1}{\rho_1} + \frac{-\nu_2}{\rho_2} - 2\mu - \frac{1}{2}\right), \sigma, \frac{1}{\mu}\right) \\
&\equiv \left(\frac{1}{\mu}\left(\frac{-\nu_1}{\rho_1} + \frac{-\nu_2}{\rho_2} - \frac{1}{2}\right) - 2, \sigma, \frac{1}{\mu}\right),
\end{aligned}$$

and therefore

$$(\nu_1, \sigma_1, -\rho_1) \otimes (\nu_2, \sigma_2, -\rho_2) \equiv \left(\frac{1}{\mu}\left(\frac{\nu_1}{\rho_1} + \frac{\nu_2}{\rho_2} + \frac{1}{2}\right), \sigma, -\frac{1}{\mu}\right).$$

■

0.4 Implementation

Compile-time static analysis

To illustrate an implementation of GGA for static analysis, we sketch the operation of the PPL compiler at a high-level and defer to the supplementary code for details. A probabilistic program is first inspected using Python's built-in `ast` module and transformed to static single assignment (SSA) form [RWZ88]. Next, standard compiler optimizations (e.g. dead code elimination, constant propagation) are applied and an execution of the optimized program is traced [WSG11; Bin+19] and accumulated in a directed acyclic graph representation. A breadth-first type checking pass, as seen in Algorithm 1, completes in linear time, and GGA results may be applied to implement `computeGGA()` using the following steps:

- If a node has no parents, then it is an atomic distribution and its tail parameters are known (Table 0.2)
- Otherwise, the node is an operation taking its potentially stochastic inputs (parents) to its output. Consult Table 0.1 for the output GGA tails.

Algorithm 1 Pseudocode for a GGA tails static analysis pass

Require: Abstract syntax tree for a PPL program
  frontier ← [rv : Parents(rv) = ∅]
  GGAs ← {}
  while frontier ≠ ∅ **do**
    next ← frontier.popLeft()
    GGAs[next] ← computeGGA(next.op, next.parent)
    frontier ← frontier + next.children()
  end while
  return GGA parameter estimates for all random variables

Representative distributions

For each (ν, σ, ρ) we make a carefully defined choice of p on $\mathbb{R}$ such that if $X \sim p$, then $X \equiv (\nu, \sigma, \rho)$. This way, any random variable $f(X)$, where f is 1-Lipschitz, will exhibit the correct tail, and so approximations of this form may be used for variational inference or density estimation. Let $X \equiv (\nu, \sigma, \rho)$ and $0 < \epsilon \ll 1$ denote a small parameter such that tails e^{-x^ϵ} are deemed to be "very heavy" (we chose $\epsilon = 0.1$).

($\rho \leq 0$) If $\rho \leq -1$, then $p_X(x) \sim cx^{-|\nu|}$. A prominent distribution on $\mathbb{R}$ with power law tails is the *Student t distribution*, in this case, with $|\nu| - 1$ degrees of freedom if $\nu < -1$ (generate $X \sim t_{|\nu|-1}$).

($\rho > \epsilon$) For moderately sized $\rho > 0$, we consider a symmetrized variant of the generalized Gamma density (Equation (2)).

($\rho \leq \epsilon$) If $X \equiv (\nu, \sigma, \rho)$ where ρ is small, then X will exhibit much heavier tails, and the generalized Gamma distribution in Case 1 will become challenging to sample from. In these cases, we expect that the tail of X should be well represented by a power law. The generalized Gamma density (Equation (2)) satisfies $\mathbb{E}X^r = \sigma^{-r/\rho}\Gamma(\frac{\nu+1+r}{\rho})/\Gamma(\frac{\nu+1}{\rho})$ for $r > 0$. Let $\alpha > 0$ be such that $\mathbb{E}X^\alpha = 2$. By Markov's inequality, the tail of X satisfies $\mathbb{P}(X > x) \leq 2x^{-\alpha}$. Therefore, we can represent tails of this form by the Student t distribution with $\alpha + 1$ degrees of freedom (generate $X \sim t_\alpha$).

Bulk correction by Lipschitz mapping

While a representative distribution will exhibit the desired tails, the target distribution's bulk may be very different from a generalized Gamma and result in poor distributional approximation. To address this, we propose splicing together the tails from a generalized Gamma with a flexible density approximation for the bulk. While many combinations are possible, in this work we rely on Lemma 2 and post-compose neural spline flows [Dur+19] (which are identity functions outside of a bounded interval) after properly initialized generalized Gamma distributions. Optimizing the parameters of the flow results in good bulk approximation while simultaneously preserving the tail correctness guarantees attained by the GGA.

Example 7 *Let $A \in \mathbb{R}^{k\times k}$, $x, y \in \mathbb{R}^k$, with $x_i, y_i, A_{ij} \overset{iid}{\sim} \mathcal{N}(-1, 1)$. The distribution of $x^\top Ay = \sum_{i,j} x_i A_{ij} y_j$ is convolution of normal-powers [GG08] and has no convenient closed form expression. Using GGA's closure theorems (table 0.1), one can compute its tail parameters to be $(\frac{k}{2} - 1, \frac{3}{2}, \frac{2}{3})$.*

The GGA representative is a gamma distribution with the correct tails, but there is non-negligible error in the bulk where x is small. To address this, a learnable bijector can be optimized as in Figure 0.3 bottom left to correct the bulk approximation. Guaranteed by Lemma 2 and visualized in Figure 0.3 bottom right, the tails of the overall composition remain calibrated.

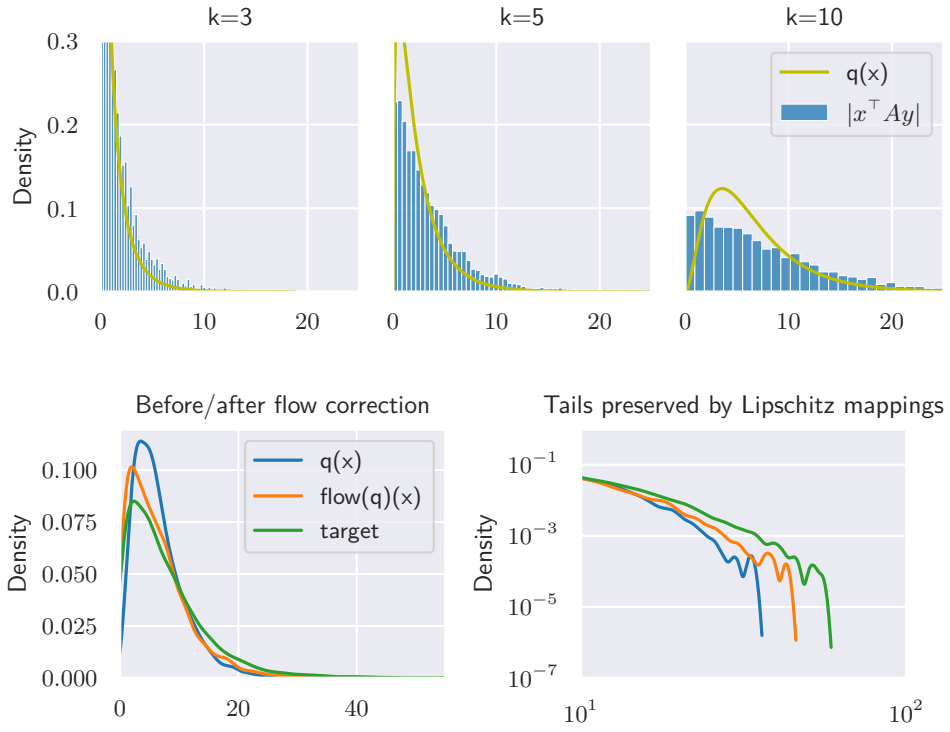

Figure 0.3: (Top) 5000 samples of $|x^\top Ay|$ vs the calibrated GGA density $q(x)$. While calibrated tails are provably guarantees, the target distribution's bulk differs from the assumed generalized Gamma representative distribution (section 0.4) for all k. To fix the bulk approximation, a normalizing flow is composed with the GGA representative to form flow$(q)(x)$. The bulk approximation is improved (bottom) while the tails continue to exhibit the same behavior (bottom right).

0.5 Experiments

In this section we demonstrate that GGA-based density estimation yields improvements across a variety of metrics. We consider the parametric family defined in Section 0.4 and compare against pushfowards of Normal distributions. To understand the individual effect of using a GGA base distribution over standard normals versus more expressive pushforward maps [Dur+19], we also report ablation results where normalizing flows are replaced by affine transforms as originally proposed in [Kuc+17]. All experiments are repeated for 50 trials, trained to convergence using the Adam optimizer with manually tuned learning rate, and conducted on i7-8700K CPU and GTX 1080 GPU hardware.

All target distributions in this section are expressed as generative PPL programs: Cauchy using a reciprocal normal, Chi2 using a sum of squared normals, IG (Inverse Gamma) using a reciprocal exponential, Normal using a sum of normals, and StudentT using a normal and

Cauchy ratio. Doing so tasks the static analyzer to infer the target's tails and makes the analysis non-trivial. See supplementary for full details.

Our results in the following tables share a consistent narrative where a GGA base distribution rarely hurts and can significantly help with heavy tailed targets. Except for when targets are truly light tailed ($\alpha = \infty$ in Chi2 and Normal), GGA-based approximations are the only ones to reproduce appropriate GPD tail index $\hat{\alpha}$ in density estimation and achieve a passing Pareto $\hat{k}$ diagnostic [Yao+18] below 0.2 in variational inference. When viewed through traditional evaluation metrics such as negative cross-entropy $H(p, q) = E_p \log p(X)$, ELBO $E_q \log \frac{q(X)}{p(X)}$, and importance-weighted autoencoder bound [BGS15] $E_q \log \sum_i^{1000} \frac{p(X)}{q(X)}$, GGA-based approximations remain favorable on almost all heavy-tailed targets and have negligible difference for light tailed targets. Less surprising is the result that adding a flow improved approximation metrics, as we expect the additional representation flexibility to be beneficial.

Density Estimation We minimize a Monte-Carlo estimate of the cross entropy $H(p, q) = -E_p[\log q(X)] \approx -\frac{1}{N} \sum_{i=1}^{N} \log q(x_i), \quad x_i \sim p$. The results are shown in Table 0.3 along with power-law tail index estimates [CSN09] $\hat{\alpha}$. Overall, we see that GGA performs better (lower NLL, $\hat{\alpha}$ closer to target) when the target has heavier tails (lower $\hat{\alpha}$ target/theory) and that the difference is smaller but still non-negligible for distributions such as Chi1 which possess tails heavier than Gaussian.

Table 0.3: Density estimation metrics attained (mean, standard deviation in parenthesis) on targets of varying tail index (smaller α = heavier tails). Higher negative cross entropy $-H(p, q) = E_p \log q(X)$ implies a better overall approximation (row maxes bolded) while close agreement between the target Pareto tail index α [CSN09] and its estimate $\hat{\alpha}$ in $q(x)$ suggest calibrated tails (closest in row bolded).

Target	Method Metric	Normal Affine	Normal Flow	GGA Affine	GGA Flow
Cauchy	$\hat{\alpha}$	7.7 (2.5)	7.1 (6.6)	2.1 (0.064)	**2 (0.067)**
($\alpha = 2$)	-H(p,q)	-1.4e7 (6.2e7)	-5.3e+10 (2.6e+11)	**-3.9e3 (56)**	-3.9e3 (55)
Chi2	$\hat{\alpha}$	**6.8 (2.4)**	6.4 (0.88)	5.5 (1.2)	5.2 (1.6)
($\alpha = \infty$)	-H(p,q)	**-2.8e3 (38)**	-2.9e3 (55)	-2.8e3 (26)	-2.8e3 (44)
IG	$\hat{\alpha}$	7.3 (1.7)	27 (39)	**1.9 (0.092)**	1.9 (0.092)
($\alpha = 2$)	-H(p,q)	-1.4e8 (6.2e8)	-4.3e9 (2.1e+10)	-4e3 (54)	**-3.9e3 (47)**
Normal	$\hat{\alpha}$	**8.4 (3.5)**	8.8 (4.6)	8.8 (2.8)	8.2 (4)
($\alpha = \infty$)	-H(p,q)	**-1.4e3 (19)**	-1.4e3 (19)	-1.4e3 (21)	-1.4e3 (24)
StudentT	$\hat{\alpha}$	7.7 (2.3)	13 (11)	**3.1 (0.16)**	3.3 (0.45)
($\alpha = 3$)	-H(p,q)	-3e3 (4.7e2)	**-2.7e3 (6.4e2)**	-3.6e3 (28)	-3.4e3 (42)

Variational Inference The optimization objective is the ELBO

$$E_q \log \frac{p(X)}{q(X)} \approx \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(x_i)}{q(x_i)}, \quad x_i \sim q$$

Here, the density p must also be evaluated so for simplicity experiments in table 0.4 use closed-form marginalized densities for targets. The overall trends also show that GGA yields consistent improvements as measured by both ELBO and importance-weighted estimates of marginal likelihood and that the difference was greater when the tails of $p(z)$ were heavier. The $\hat{k}$ diagnostics [Yao+18] corroborate our findings that variational inference succeeds ($\hat{k} < -1.2$) when a GGA with appropriately matched tails is used and fails ($\hat{k} > 1$) when Gaussian tails are erroneously imposed.

Table 0.4: Variational inference metrics (mean, standard deviation in parenthesis) on targets of varying tail index (smaller α = heavier tails). Both the IWAE bound $E_q \log \sum_i^K \frac{p(X_i)}{q(X_i)}$ and the ELBO ($K = 1$) measure (a lower bound) on the marginal likelihood where larger is better (row maxes bolded). In Yao et al. [Yao+18], a Pareto $\hat{k}$ diagnostic > 0.2 is interpreted as potentially problematic so only values below are bolded.

Target	Method Metric	Normal Affine	Normal Flow	GGA Affine	GGA Flow
Cauchy ($\alpha = 2$)	$\hat{k}$	0.46 (0.13)	0.35 (0.43)	**0.011 (0.0063)**	**0.034 (0.01)**
	ELBO	-0.19 (0.011)	-0.1 (0.028)	**1.4 (0.00027)**	1.4 (0.0015)
	IWAE	6.8 (0.031)	6.9 (0.15)	**8.3 (0.00028)**	8.3 (0.0015)
Chi2 ($\alpha = \infty$)	$\hat{k}$	0.26 (0.094)	0.23 (0.12)	**0.075 (0.07)**	**0.14 (0.1)**
	ELBO	-0.024 (0.0072)	-0.046 (0.034)	**-0.002 (0.003)**	-0.031 (0.031)
	IWAE	**6.9 (0.0066)**	6.9 (0.0098)	6.9 (0.0016)	6.9 (0.0067)
IG ($\alpha = 2$)	$\hat{k}$	13 (3.4)	0.63 (0.55)	11 (3.2)	5.7 (5.7)
	ELBO	-0.63 (6.5)	-1.5 (0.1)	**0.44 (4.2)**	-0.14 (0.9)
	IWAE	**2e3 (3.9e3)**	11 (23)	9.5e2 (1.6e3)	1.6e2 (1.6e2)
Normal ($\alpha = \infty$)	$\hat{k}$	**0.0055 (0.0082)**	**0.022 (0.017)**	**0.007 (0.007)**	**0.017 (0.014)**
	ELBO	**-0.000 (0.001)**	-0.00038 (0.0013)	-0.0002 (0.0006)	-0.00071 (0.001)
	IWAE	**6.9 (0.0005)**	6.9 (0.0013)	6.9 (0.00055)	6.9 (0.00094)
StudentT ($\alpha = 3$)	$\hat{k}$	0.53 (0.17)	0.21 (0.26)	**0.002 (0.003)**	**0.12 (0.064)**
	ELBO	-0.072 (0.0099)	-0.017 (0.0025)	**1.4 (0.00012)**	1.4 (0.0052)
	IWAE	6.9 (0.058)	6.9 (0.01)	**8.3 (0.00012)**	8.3 (0.0052)

The targets in Table 0.3 and Table 0.4 are analyzed using the GGA. Note that Inverse Gamma ("IG") corresponds to the inverse exponential. We selected closed form targets so that the Pareto tail index α is known analytically and the quality of theoretical predictions as well as empirical results can be evaluated against. All experiments are repeated for 100

trials and $1,000$ samples from the model (as well as the approximation in VI) were used to compute each gradient estimate. Losses were trained until convergence, which all occurred in under $10,000$ iterations at a 0.05 learning rate and the Adam [KB14] optimizer.

SGD for least-squares linear regression

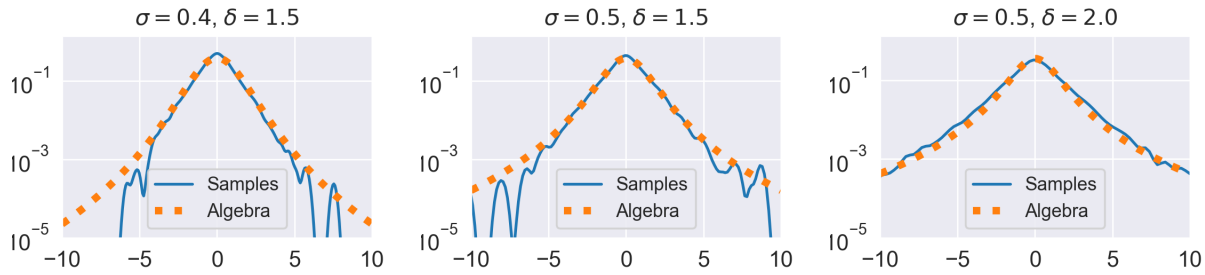

Figure 0.4: Density of iterates of SGD vs. predicted tail behaviour

For inputs X and labels Y from a dataset $\mathcal{D}$, the least squares estimator for linear regression satisfies $\beta = \min_\beta \frac{1}{2}\mathbb{E}_{X,Y\sim\mathcal{D}}(Y - X\beta)^2$. To solve for this estimator, one can apply stochastic gradient descent (SGD) sampling over independent $X_k, Y_k \sim \mathcal{D}$ to obtain the sequence of iterations

$$\beta_{k+1} = (I - \delta X_k X_k^\top)\beta_k + \delta Y_k X_k$$

for a step size $\delta > 0$. For large δ, the iterates β_k typically exhibit heavy-tailed fluctuations; in this regard, this sequence of iterates has been used as a simple model for more general stochastic optimization dynamics [GSZ21; HM21]. In particular, generalization performance has been tied to the heaviness of the tails in the iterates [SSG19]. Here we use our algebra to predict the tail behaviour in a simple one-dimensional setting where $X_k \sim \mathcal{N}(0, \sigma^2)$ and $Y_k \sim \mathcal{N}(0, 1)$. From classical theory [BDM16], it is known that X_k converges in distribution to a power law with tail exponent $\alpha > 0$ satisfying $\mathbb{E}|1 - \delta X_k^2|^\alpha = 1$. In fig. 0.4, we plot the density of the representative obtained using our algebra after 10^4 iterations against a kernel density estimate of the first 10^6 iterates when $\sigma \in \{0.4, 0.5\}$ and $\delta \in \{1.5, 2.0\}$. In all cases, the density obtained from the algebra provides a surprisingly close fit.

Normal target

Consider the toy example of a Normal target. This case is trivial for Gaussian based methods and is oftentimes the initialization. This lack of approximation gap in ADVI is seen in Figure 0.5, where we also see that GGA achieves similar approximation quality. This is unsurprising as the GGA approximation in Table 0.2 is also a Normal distribution.

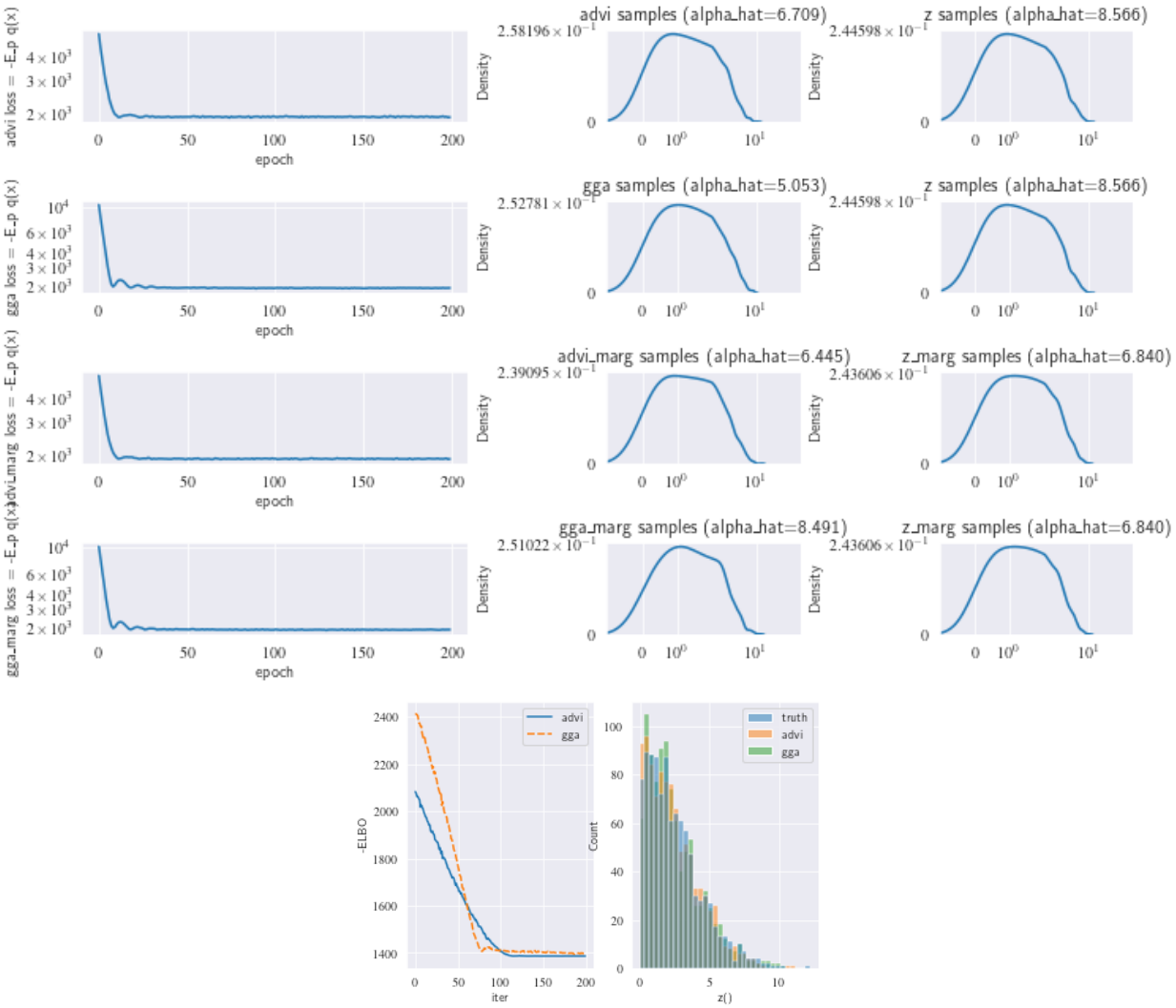

Figure 0.5: Density estimation and VI against a known normal target

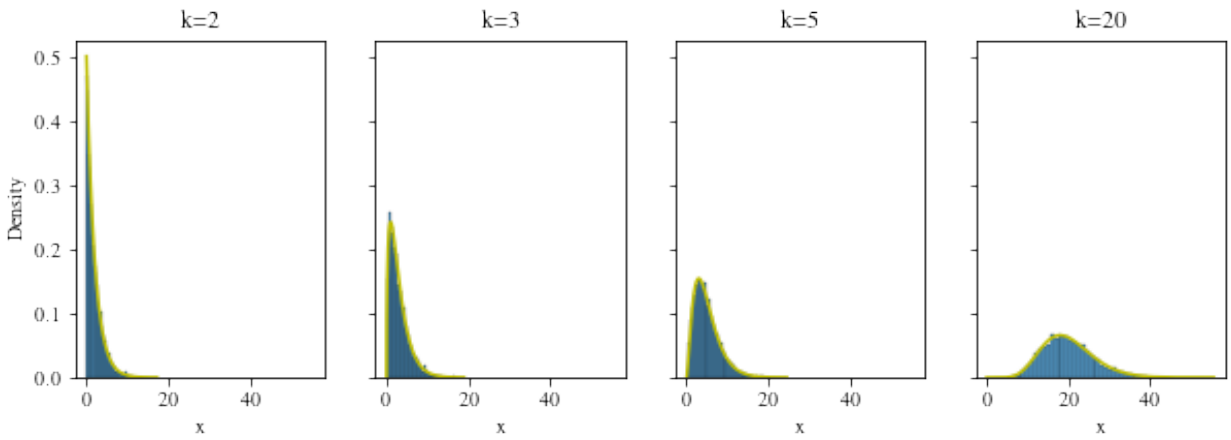

Figure 0.6: 5000 samples of JL matrix trace (blue) vs GGA prediction (yellow)

Chi-square

Now let $X_{ij} \sim N(0,1)$ and consider $trX^\top X$. Such quantities arise in the analysis of random projections. It is important here to recognize that the power operation $X \mapsto X^2$ is not equivalent to the multiplication operation $X \mapsto X \otimes X$, as multiplication assumes independence.

0.6 Conclusion

In this work, we have proposed a novel systematic approach for conducting tail inferential static analysis by implementing a three-parameter generalized Gamma algebra into a PPL compiler. Initial results are promising, showing that improved inference with simpler approximation families is possible when combined with tail metadata. While already useful, the generalized Gamma algebra and its implementation currently has some notable limitations:

- Since the algebra assumes independence, handling of dependencies between defined random variables must be conducted externally. This will inevitably require interoperability with a symbolic package to decompose complex expressions into operations on independent random variables.
- The GGA is formulated for univariate distributions only. Suitably defining multivariate tails is an open problem with interesting alternatives [Jai+20; LHM22] all of which could extend GGA to higher dimensions.
- Conditioning is arguably the most important feature of a PPL and what distinguishes it from a glorified simulator. Exact marginalization in general is NP-hard [KF09], so treatment of conditional distributions using symbolic manipulations is a significant open

problem, with some basic developments [SR17; CJ19]. Since only the tails are required in our setup, it may be possible to construct a dual algebra for operations under conditioning; this is left for future work.

- Compile-time static analysis only applicable to fixed model structure. While out of scope for our current work, open-universe models [MR10] and PPLs to support them [Bin+19] are an important research direction.
- The most significant omission to the algebra itself is classification of log-normal tails; while addition may be treated using [GT16] for example, multiplicative convolution with log-normal tails remains elusive.
- At present, reciprocals are approximated by assuming behaviour near zero. Reciprocals may be better treated by covering near-zero asymptotics separately.

The GGA provides a necessary first step into the static analysis of tails in a probabilistic program. As the above limitations are improved in future work and GGA becomes more broadly applicable, we are excited to see how improved tail modelling will improve downstream PPL applications as well as other researchers will utilize GGA metadata to develop novel PPL applications.

Bibliography

[AA10] Soren Asmussen and Hansjorg Albrecher. *Ruin probabilities*. Vol. 14. World scientific, 2010.

[Asm+17] Søren Asmussen et al. "Tail asymptotics of light-tailed Weibull-like sums". In: *Probability and Mathematical Statistics* 37.2 (2017), pp. 235–256.

[BDM16] Dariusz Buraczewski, Ewa Damek, and Thomas Mikosch. "Stochastic models with power-law tails". In: *Springer Ser. Oper. Res. Financ. Eng., Springer, Cham* 10 (2016), pp. 978–3.

[Ber19] Ryan Bernstein. "Static analysis for probabilistic programs". In: *arXiv preprint arXiv:1909.05076* (2019).

[BGS15] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. "Importance weighted autoencoders". In: *arXiv preprint arXiv:1509.00519* (2015).

[Bin+19] Eli Bingham et al. "Pyro: Deep universal probabilistic programming". In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 973–978.

[Car+17] Bob Carpenter et al. "Stan: A probabilistic programming language". In: *Journal of statistical software* 76.1 (2017).

[CJ19] Kenta Cho and Bart Jacobs. "Disintegration and Bayesian inversion via string diagrams". In: *Mathematical Structures in Computer Science* 29.7 (2019), pp. 938–971.

[Cla+13] Guillaume Claret et al. "Bayesian inference using data flow analysis". In: *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*. 2013, pp. 92–102.

[CSN09] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data". In: *SIAM review* 51.4 (2009), pp. 661–703.

[Dur+19] Conor Durkan et al. "Neural spline flows". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 7509–7520.

[Esl+16] SM Eslami et al. "Attend, infer, repeat: Fast scene understanding with generative models". In: *Advances in Neural Information Processing Systems* 29 (2016).

[GC11] Mark Girolami and Ben Calderhead. "Riemann manifold langevin and hamiltonian monte carlo methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.2 (2011), pp. 123–214.

[GG08] Rameshwar D Gupta and Ramesh C Gupta. "Analyzing skewed data by power normal model". In: *Test* 17.1 (2008), pp. 197–210.

[GK98] Charles M Goldie and Claudia Klüppelberg. "Subexponential distributions". In: *A practical guide to heavy tails: statistical techniques and applications* (1998), pp. 435–459.

[Goo+12] Noah Goodman et al. "Church: a language for generative models". In: *arXiv preprint arXiv:1206.3255* (2012).

[GSZ21] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. "The heavy-tail phenomenon in SGD". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 3964–3975.

[GT16] Archil Gulisashvili and Peter Tankov. "Tail behavior of sums and differences of log-normal random variables". In: *Bernoulli* 22.1 (2016), pp. 444–493.

[HM21] Liam Hodgkinson and Michael Mahoney. "Multiplicative noise and heavy tails in stochastic optimization". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4262–4274.

[Jai+20] Priyank Jaini et al. "Tails of Lipschitz Triangular Flows". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4673–4681.

[JP89] Claire Jones and Gordon D Plotkin. "A probabilistic powerdomain of evaluations". In: *Proceedings. Fourth Annual Symposium on Logic in Computer Science*. IEEE Computer Society. 1989, pp. 186–187.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[Koz79] Dexter Kozen. "Semantics of probabilistic programs". In: *20th Annual Symposium on Foundations of Computer Science (sfcs 1979)*. IEEE. 1979, pp. 101–114.

[Kuc+17] Alp Kucukelbir et al. "Automatic differentiation variational inference". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 430–474.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*. Mathematical surveys and monographs 89. American Mathematical Soc., 2001.

[Lee+19] Wonyeol Lee et al. "Towards verified stochastic variational inference for probabilistic programs". In: *Proceedings of the ACM on Programming Languages* 4.POPL (2019), pp. 1–33.

[LHM22] Feynman Liang, Liam Hodgkinson, and Michael Mahoney. "Fat–Tailed Variational Inference with Anisotropic Tail Adaptive Flows". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. 2022, p. 132.

[LHM23] Feynman Liang, Liam Hodgkinson, and Michael Mahoney. "Static Analysis of Tail Behaviour with a Generalized Gamma Algebra". In: *Submitted to AISTATS 2023* (2023).

[Mik99] T Mikosch. *Regular Variation Subexponentiality and Their Applications in Probability Theory*. 1999. URL: `https://www.eurandom.tue.nl/reports/1999/013-report.pdf`.

[MR10] Brian Milch and Stuart Russell. "Extending Bayesian networks to the open-universe case". In: *Heuristics, Probability and Causality: A Tribute to Judea Pearl. College Publications* (2010).

[MSH09] Arakaparampil M Mathai, Ram Kishore Saxena, and Hans J Haubold. *The H-function: theory and applications*. Springer Science & Business Media, 2009.

[Nor+14] Aditya Nori et al. "R2: An efficient MCMC sampler for probabilistic programs". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 28. 2014.

[Pap+21] George Papamakarios et al. "Normalizing flows for probabilistic modeling and inference". In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64.

[Pas+19] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[RT96] Gareth O Roberts and Richard L Tweedie. "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli* (1996), pp. 341–363.

[RWZ88] Barry K Rosen, Mark N Wegman, and F Kenneth Zadeck. "Global value numbers and redundant computations". In: *Proceedings of the 15th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. 1988, pp. 12–27.

[SCG13] Sriram Sankaranarayanan, Aleksandar Chakarov, and Sumit Gulwani. "Static analysis for probabilistic programs: inferring whole program properties from finitely many paths". In: *Proceedings of the 34th ACM SIGPLAN conference on Programming language design and implementation*. 2013, pp. 447–458.

[Sid+17] N. Siddharth et al. "Learning Disentangled Representations with Semi-Supervised Deep Generative Models". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 5927–5937. URL: `http://papers.nips.cc/paper/7174-learning-disentangled-representations-with-semi-supervised-deep-generative-models.pdf`.

[Spi+96] David Spiegelhalter et al. "BUGS 0.5: Bayesian inference using Gibbs sampling manual (version ii)". In: *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK* (1996), pp. 1–59.

[SR17] Chung-chieh Shan and Norman Ramsey. "Exact Bayesian inference by symbolic disintegration". In: *Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages*. 2017, pp. 130–144.

[SSG19] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. "A tail-index analysis of stochastic gradient noise in deep neural networks". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5827–5837.

[Taj03] Nader Tajvidi. "Confidence intervals and accuracy estimation for heavy-tailed generalized Pareto distributions". In: *Extremes* 6.2 (2003), pp. 111–123.

[Teh+20] Nazanin Tehrani et al. "Bean machine: A declarative probabilistic programming language for efficient programmable inference". In: *International Conference on Probabilistic Graphical Models*. PMLR. 2020.

[Tol+16] David Tolpin et al. "Design and implementation of probabilistic programming language anglican". In: *Proceedings of the 28th Symposium on the Implementation and Application of Functional programming Languages*. 2016, pp. 1–12.

[Tra+18] Dustin Tran et al. "Simple, distributed, and accelerated probabilistic programming". In: *Advances in Neural Information Processing Systems* 31 (2018).

[Val+17] Perry de Valpine et al. "Programming with models: writing statistical algorithms for general model structures with NIMBLE". In: *Journal of Computational and Graphical Statistics* 26.2 (2017), pp. 403–413.

[Veh+15] Aki Vehtari et al. "Pareto smoothed importance sampling". In: *arXiv preprint arXiv:1507.02646* (2015).

[Wat95] George Neville Watson. *A treatise on the theory of Bessel functions*. Cambridge university press, 1995.

[Web+19] Stefan Webb et al. "Improving automated variational inference with normalizing flows". In: *ICML Workshop on Automated Machine Learning*. 2019.

[WHR18] Di Wang, Jan Hoffmann, and Thomas Reps. "PMAF: an algebraic framework for static analysis of probabilistic programs". In: *ACM SIGPLAN Notices* 53.4 (2018), pp. 513–528.

[WLL18] Dilin Wang, Hao Liu, and Qiang Liu. "Variational inference with tail-adaptive f-divergence". In: *Advances in Neural Information Processing Systems* 31 (2018), pp. 5737–5747.

[WSG11] David Wingate, Andreas Stuhlmüller, and Noah Goodman. "Lightweight implementations of probabilistic programming languages via transformational compilation". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 2011, pp. 770–778.

[Xu+20] Kai Xu et al. “AdvancedHMC. jl: A robust, modular and efficient implementation of advanced HMC algorithms”. In: *Symposium on Advances in Approximate Bayesian Inference*. PMLR. 2020, pp. 1–10.

[Yao+18] Yuling Yao et al. “Yes, but did it work?: Evaluating variational inference”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5581–5590.