Notes on the ARC White-box Challenge Problem: Part 2

Zachary Martinot

October 2026

In my previous set of notes (Martinot, "Notes on the Problem (Phase 1)") I argued that structure in $\vec{F}(\vec{x})$ - that is, the Fourier harmonic content - is essentially isotropic when viewed as a black-box because the matrices W_ℓ are drawn independently from an isotropic distribution. However, if we consider each layer function individually, the columns of the weight matrices are a set of special directions. This is shown explicitly in section 5. I think this structure will shed light on the failure of the cumulant propagation scheme to scale with depth, and points the way towards doing better.

The cumulant propagation of Wu et al. (2026) is implicitly modeling the intermediate distribution of each layer 's output by a truncated Edgeworth/Gram-Charlier series but the choice of the standard component basis is misaligned with the important structure. When the series so defined is truncated to very the low order necessitated by finite available compute, it is effectively looking at only a finite set of marginal distributions in directions that are relatively insensitive to the important non-Gaussianity. I'm thinking that capturing this anisotropic non-Gaussian structure more efficiently is how you improve the efficiency of a closure propagation and hence the final estimate.

Contents

1 Preliminaries **2**
1.1 ARC Challenge 2
1.2 Alternate Definition 3

2 Change of Variables **3**

3 Distribution Coordinate Transformation By Variational Approximation **4**

4 Eigenspace Coordinates **6**

5 Anisotropy of Non-Gaussianity **8**

6 Mixtures of Gaussians **10**

7 Ridge Function Approximations **11**

8 Fourier Decomposition 13

9 Fourier Propagator 14
9.1 Quadratic Approximation 14

10 Particle Approximations and Renormalization Flows 15

A Integral Representations of $(z)_+$ 17

B Gaussian Transport 17

C Gaussian Closure Integrals 18

1 Preliminaries

We are generally concerned with the problem of accurately estimating the expectation value of a function $\langle \vec{F}(\vec{x}) \rangle$ given a probability distribution $P(\vec{x})$ over the inputs $\vec{x} \in \mathbb{R}^N$. Explicitly, this means our problem is to evaluate an integral of the form

$$\langle \vec{F}(\vec{x}) \rangle = \int_{\mathbb{R}^N} d\vec{x} P(\vec{x}) \vec{F}(\vec{x}). \tag{1}$$

In particular, we are concerned with functions defined by the L-layered composition

$$\vec{F}(\vec{x}) = \left(\vec{F}_{L-1} \circ \vec{F}_{L-2} \circ \ldots \circ \vec{F}_0 \right)(\vec{x}), \tag{2}$$

or equivalently by the recursion

$$\vec{x}_{\ell+1} = \vec{F}_\ell(\vec{x}_\ell), \text{ for } \ell = 0, 1, \ldots, L-1, \text{ with } \vec{x}_0 = \vec{x}, \tag{3}$$

$$\vec{F}(\vec{x}) = \vec{x}_L. \tag{4}$$

1.1 ARC Challenge

For the ARC challenge, the layers are given by

$$\vec{F}_\ell(\vec{x}_\ell) = (W_\ell \vec{x}_\ell)_+ , \tag{5}$$

where $(\vec{z})_+ = \max(0, \vec{z})$ is the first-degree truncated power (a.k.a. ReLU) applied element-wise, and each of the $N \times N$ matrices W_ℓ is a realization of a random matrix with all elements i.i.d. as zero-mean Gaussian with variance $\frac{2}{N}$. For Phase 1 we had $N = 256$ and $L = 32$, while in Phase 2 this was changed to $N = 1024$ and $L = 16$.

The input distribution is standard normal Gaussian $P(\vec{x}) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|\vec{x}|^2}$.

1.2 Alternate Definition

The variables

$$\vec{z}_\ell = W_\ell \vec{x}_\ell \tag{6}$$
$$\tag{7}$$

are vaguely Gaussian by the CLT, so it is useful to consider an alternate parameterization for the sequence of transformations as

$$\vec{z}_{\ell+1} = \vec{G}_\ell(\vec{z}_\ell), \tag{8}$$
$$\vec{G}_\ell(\vec{z}) = W_{\ell+1}(\vec{z})_+, \tag{9}$$

and then $\vec{F}(\vec{x})$ is

$$\vec{F}(\vec{x}) = \big(\vec{z}_{L-1}(\vec{x})\big)_+. \tag{10}$$

2 Change of Variables

As we all learned in introductory calculus, the way to compute a seemingly intractable integral is to find a good change of variables. This approach applies just as well to the present problem. Each step of the recursion $\vec{z}_\ell = \vec{G}_{\ell-1}(\vec{z}_{\ell-1})$ can be seen as just such a change of variables from $\vec{z}_{\ell-1}$ to $\vec{z}_\ell$.

If we knew the distribution function P_{L-1} of $\vec{z}_{L-1}$, then our expectation integral could be evaluated as

$$\langle \vec{x}_L \rangle = \int d\vec{z} (\vec{z})_+ P_{L-1}(\vec{z}). \tag{11}$$

and accurate evaluation of integrals over $(\vec{z})_+$ is a more tractable proposition than general $\vec{F}(\vec{x})$ - provided we can figure out what P_{L-1} is and it is sufficiently simple.

In principle, we should be able to compute P_{L-1} by recursively computing $P_{\ell+1}$ as the pushforward of P_ℓ under $\vec{G}_\ell$, and thus treat our composite function as a sequence of coordinate transformations.

If $\vec{f}(\vec{z}_0)$ is invertible and $P_0(\vec{z}_0)$ is the distribution of $\vec{z}_0$, then the distribution of $\vec{z} = \vec{f}(\vec{z}_0)$ is given exactly by the well-known change of variables formula

$$P(\vec{z}) = P_0\Big(\vec{f}^{-1}(\vec{z})\Big) \left| \det\left(\frac{\partial \vec{f}^{-1}(\vec{z})}{\partial \vec{z}^\intercal} \right) \right|. \tag{12}$$

This can be applied recursively to a chain of invertible functions $\vec{f}_\ell$ to evaluate the density P_{L-1} as

$$\log P_{L-1}(\vec{z}_{L-1}) = \log P_0(\vec{z}_0) - \sum_{\ell=0}^{L-2} \log \left| \det\left(\frac{\partial \vec{f}_\ell(\vec{z}_\ell)}{\partial \vec{z}_\ell^\intercal} \right) \right|. \tag{13}$$

Unfortunately, in our present case the layer functions $\vec{G}_\ell(\vec{z}) = W_\ell(\vec{z})_+$ are not invertible, and this density formula is not applicable - $(z)_+$ is surjective, mapping all $z \leq 0$ to the point at zero. Instead, the pushforward is specified by the more general but less immediately practical marginal integral definition. If $\vec{x} \sim P_0(\vec{x})$ and $\vec{y} = \vec{f}(\vec{x})$ then the distribution of $\vec{y}$ is given by (e.g. Gillespie (1983))

$$P_1(\vec{y}) = \int d\vec{x} \delta\Big(\vec{y} - \vec{f}(\vec{x})\Big) P_0(\vec{x}). \tag{14}$$

The pushforward of $(\vec{z}_0)_+$ is a mixture of 2^N components which are all singular but for one - each component has a Dirac-delta factor corresponding to a coordinate face of the positive orthant.[1]Consequently, the true distributions $P_\ell(\vec{z}_\ell)$ are singular on the full space $\vec{z}_\ell \in \mathbb{R}^N$ for $\ell > 0$.

Despite this technicality, it seems the large dimension of mixing means we can still approximate these distributions by smooth functions over $\vec{z} \in \mathbb{R}^N$.

3 Distribution Coordinate Transformation By Variational Approximation

The general idea is to find some sequence of approximations $Q_\ell(\vec{z}_\ell)$ to the true distributions $P_\ell(\vec{z}_\ell)$ by determining $Q_{\ell+1}$ as an approximation to the pushforward of Q_ℓ under the map $\vec{G}_\ell(\vec{z})$. The general framework for this is to define a variational problem with some appropriate objective function $\mathcal{L}(\vec{\theta}_\ell)$, and some choice for the approximation function $Q_\ell(\vec{z}) = Q_\ell(\vec{z}, \vec{\theta}_\ell)$ with parameters $\vec{\theta}_\ell$.

For each ℓ, the ideal approximation problem we would solve is to minimize the KL divergence against the true distribution P_ℓ

$$\mathcal{L}(\vec{\theta}_\ell) = \int d\vec{z}_\ell P_\ell(\vec{z}_\ell) \log\left(\frac{P_\ell(\vec{z}_\ell)}{Q_\ell(\vec{z}_\ell, \vec{\theta}_\ell)}\right), \tag{15}$$

and since

$$P_\ell(\vec{z}_\ell) d\vec{z}_\ell = P_{\ell-1}(\vec{z}_{\ell-1}) d\vec{z}_{\ell-1}, \tag{16}$$

this is equivalent to

$$\mathcal{L}(\vec{\theta}_\ell) = \int d\vec{z}_{\ell-1} P_{\ell-1}(\vec{z}_{\ell-1}) \log\left(\frac{P_{\ell-1}(\vec{z}_{\ell-1})}{Q_\ell\Big(\vec{G}_{\ell-1}(\vec{z}_{\ell-1}), \vec{\theta}_\ell\Big)}\right). \tag{17}$$

So given the approximation $Q_\ell(\vec{z}, \vec{\theta}_\ell)$ at layer ℓ we determine the next layer's approximation

[1]If this is unclear, consider the rectified standard normal Gaussian in two dimensions.

$Q_{\ell+1}$ by minimizing

$$\mathcal{L}(\vec{\theta}_{\ell+1}) = \int d\vec{z} Q_\ell(\vec{z}) \log\left(\frac{Q_\ell(\vec{z})}{Q_{\ell+1}\left(\vec{G}_\ell(\vec{z}), \vec{\theta}_{\ell+1}\right)}\right) \tag{18}$$

$$= \mathcal{L}_0 - \int d\vec{z} Q_\ell(\vec{z}) \log Q_{\ell+1}\left(\vec{G}_\ell(\vec{z}), \vec{\theta}_{\ell+1}\right) \tag{19}$$

over the parameters $\vec{\theta}_{\ell+1}$. Assuming each such problem is soluble, this defines a closure relation for the sequence of parameters $\vec{\theta}_\ell$ such that $\vec{\theta}_{\ell+1} = f(\vec{\theta}_\ell)$, and thus produces a sequence of approximations Q_ℓ.

For the simplest applicable example, we can approximate the distributions as a Gaussian for each ℓ.[2] The approximating functions are

$$Q_\ell(\vec{z}) = \mathcal{N}(\vec{z}, \vec{\eta}_\ell, S_\ell) \tag{20}$$

$$= \frac{1}{\sqrt{\det(2\pi S_\ell)}} \exp\left(-\frac{1}{2}(\vec{z} - \vec{\eta}_\ell)^\intercal S_\ell^{-1}(\vec{z} - \vec{\eta}_\ell)\right) \tag{21}$$

the objective function is

$$\mathcal{L}(\vec{\eta}, S) = \frac{1}{2}\log(\det(S)) + \frac{1}{2}\int d\vec{z}\mathcal{N}(\vec{z}, \vec{\eta}_\ell, S_\ell)\left(\vec{G}_\ell(\vec{z}) - \vec{\eta}\right)^\intercal S^{-1}\left(\vec{G}_\ell(\vec{z}) - \vec{\eta}\right) + \text{const.} \tag{22}$$

and the solution $(\vec{\eta}_{\ell+1}, S_{\ell+1})$ is given by

$$(\vec{\eta}_{\ell+1}, S_{\ell+1}) = \arg\min_{\vec{\eta}, S} \mathcal{L}(\vec{\eta}, S) \tag{23}$$

This function has a unique stationary point found by solving

$$\frac{\partial \mathcal{L}}{\partial \vec{\eta}} = 0, \qquad \frac{\partial \mathcal{L}}{\partial S} = 0, \tag{24}$$

where with a little algebra we find the solution is simply matching moments:

$$\vec{\eta}_{\ell+1} = W_{\ell+1}\int d\vec{z}\mathcal{N}(\vec{z}, \vec{\eta}_\ell, S_\ell)(\vec{z})_+ \tag{25}$$

$$S_{\ell+1} = W_{\ell+1}\left(\int d\vec{z}\mathcal{N}(\vec{z}, \vec{\eta}_\ell, S_\ell)\left((\vec{z})_+ - \vec{\eta}_{\ell+1}\right)\left((\vec{z})_+ - \vec{\eta}_{\ell+1}\right)^\intercal\right) W_{\ell+1}^\intercal. \tag{26}$$

Formulas for these integrals are given in Appendix C (reprinted from Martinot).

This Gaussian case is nice and tidy, and gives several accurate digits in the mean estimate of $\vec{F}(\vec{x})$. But it is an approximation with a fixed error at a fixed cost. Our goal is an approximation that can be scaled to trade between accuracy and cost, while maximizing accuracy per unit of computational effort.

The question is how to generalize from here. We want to find a class of approximations $Q_\ell(\vec{z}, \vec{\theta}_\ell)$ that is sufficiently general so that the approximations are extensible to arbitrary

[2] This corresponds to "covariance propagation" of Wu et al. (2026).

accuracy, while at the same time fitting into a variational problem that can produce a computationally efficient transformation algorithm. There are a plethora of ways to go about approximating a relatively complex function in terms of simpler elements. But it is not obvious how to do so in a way that would make the transformation approximation problem computationally efficient.

The KL divergence may be a theoretically optimal objective in terms of entropy maximization, but in practice we are also concerned with the computational cost (not to mention, outright feasibility!). There are other measures besides the KL divergence that may be sufficiently similar while producing more efficient algorithms. There are a number of "f-divergences" of the form

$$\mathcal{L} = \int d\vec{z} P(\vec{z}) f\left(\frac{P(\vec{z})}{Q(\vec{z})}\right) \tag{27}$$

with existing knowledge bases to be considered.

It seems generic that the approximations Q_ℓ be expected to increasingly diverge from P_ℓ with increasing ℓ. At each ℓ we incur some error in Q_ℓ, and this compounds into a larger relative error for $Q_{\ell+1}$, etc. In order to design generally applicable and efficient algorithms it would be useful to understand how and under what conditions this error growth can be controlled.

4 Eigenspace Coordinates

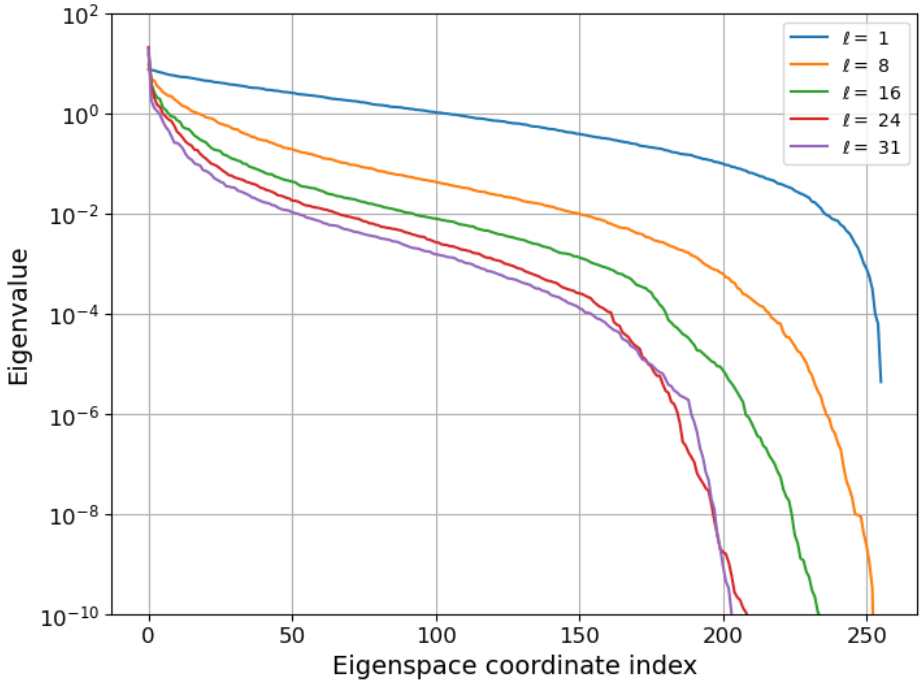


Figure 1: The eigenspectra of covariance matrices S_ℓ.

As ℓ increases the spectrum of the covariance S_ℓ becomes steeper (Figure 1), so that its effective numerical rank at a fixed error tolerance decreases steadily (though not strictly monotonically). Denote the eigendecomposition of the covariance

$$S_\ell = \mathcal{U}_\ell \Lambda_\ell \mathcal{U}_\ell^\intercal. \tag{28}$$

For some tolerance ϵ, say we keep the N_ℓ eigenvalues $\Lambda_{\ell n} > \epsilon$ so that S_ℓ is rank N_ℓ. The N_ℓ eigenvectors span a subspace of $\mathbb{R}^N$ on which the distribution $P_\ell(\vec{z}_\ell)$ has significant variance. On the complement, P_ℓ is effectively singular. It may be worth building the approximation Q_ℓ on that lower dimensional subspace explicitly.

With $V_\ell = \mathcal{U}_\ell \Lambda_\ell^{1/2}$, we define eigenspace ("whitened") coordinates

$$\vec{r} = V_\ell^+(\vec{z} - \vec{\eta}_\ell)$$

where

$$V_\ell^+ = (V_\ell^\intercal V_\ell)^{-1} V_\ell^\intercal = \Lambda_\ell^{-1/2} \mathcal{U}_\ell^\intercal$$

is the pseudo-inverse. With the approximation $\vec{z}_\ell = V_\ell \vec{r}_\ell + \vec{\eta}_\ell$ we can then define a new sequence of transformations

$$\vec{g}_\ell(\vec{r}) = V_{\ell+1}^+ \Big(W_{\ell+1}(V_\ell \vec{r} + \vec{\eta}_\ell)_+ - \vec{\eta}_{\ell+1} \Big), \tag{29}$$

$$\vec{r}_{\ell+1} = \vec{g}_\ell(\vec{r}_\ell), \tag{30}$$

and then approximate the sequence of distributions $Q_\ell(\vec{r})$ such that

$$P_\ell(\vec{z}) \approx Q_\ell\big(V_\ell^+(\vec{z} - \vec{\eta}_\ell)\big)\delta\big((I - \mathcal{U}_\ell \mathcal{U}_\ell^\intercal)\vec{z}\big).$$

Under this approximation, we estimate expectation values as

$$\int d\vec{z} f(\vec{z}) P_\ell(\vec{z}) \approx \int d\vec{r} f(V_\ell \vec{r} + \vec{\eta}_\ell) Q_\ell(\vec{r}). \tag{31}$$

In these whitened coordinates, the Gaussian approximation is simply $\propto e^{-\frac{1}{2}|\vec{r}|^2}$ for all ℓ, while in the ambient space the singular Gaussian is

$$\propto \exp\left(-\frac{1}{2}(\vec{z} - \vec{\eta}_\ell)^\intercal V_\ell (V_\ell^\intercal V_\ell)^{-2} V_\ell^\intercal (\vec{z} - \vec{\eta}_\ell) \right) \tag{32}$$

$$= \exp\left(-\frac{1}{2}(\vec{z} - \vec{\eta}_\ell)^\intercal \mathcal{U}_\ell \Lambda_\ell^{-1} \mathcal{U}_\ell^\intercal (\vec{z} - \vec{\eta}_\ell) \right). \tag{33}$$

The potential value of this eigenspace transformation is not just the dimension reduction $N_\ell < N$, which may only be moderate, but in the efficiency with which the important non-Gaussian structure may be captured and propagated accurately.

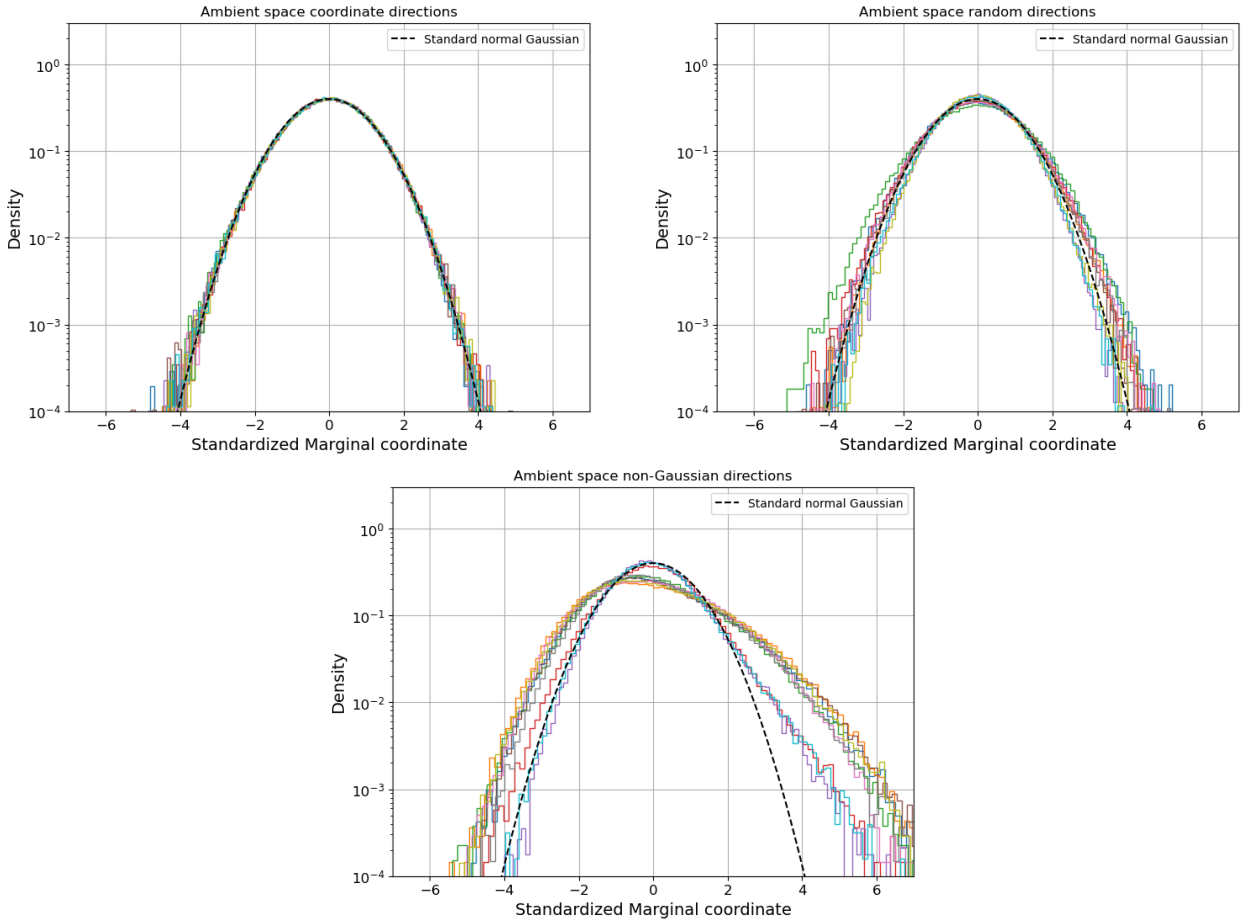


Figure 2: Marginals of the distribution of the standardized variable $(\vec{z} - \vec{\eta})/\vec{\sigma}$ (where $\vec{\sigma} = \sqrt{\text{diag}(S)}$) at $\ell = 2$ for a realization of $\vec{F}(\vec{x})$ with $N = 256$. Each panel shows the marginals along 10 different directions $\hat{u}$. Top-left: standard basis directions $\hat{e}_\alpha$, i.e. components of $\vec{z}_\ell$. Top-right: random directions drawn uniformly from $\mathbb{S}^{N-1}$. Bottom: directions $\hat{u}$ are normalized columns of W_ℓ.

5 Anisotropy of Non-Gaussianity

The deviations of $P_\ell(\vec{z}_\ell)$ from Gaussianity are tied to columns of the matrices W_ℓ. We can see this by considering the 1D marginal distributions along directions $\hat{u}$,

$$\bar{P}_{\hat{u}}(z) = \int d\vec{z}\delta(z - \hat{u} \cdot \vec{z})P(\vec{z}). \tag{34}$$

Marginalization tends to smear out structure from the full distribution, but this effect depends on the direction $\hat{u}$. The complete set of marginals $\bar{P}_{\hat{u}}(z)$ over all direction $\hat{u}$ is the Radon transform, which is invertible, so all of the structure in $P(\vec{z})$ is encoded somewhere in the marginals.

Figure 2 shows marginals over the standardized coordinates $(\vec{z} - \vec{\eta})/\sqrt{\text{diag}(S)}$. The marginal distribution for each component z_α of $\vec{z}$ corresponds to taking the standard basis directions $\hat{u} = \hat{e}_\alpha$. In the top-left panel ofFigure 2 , we find that these distributions are

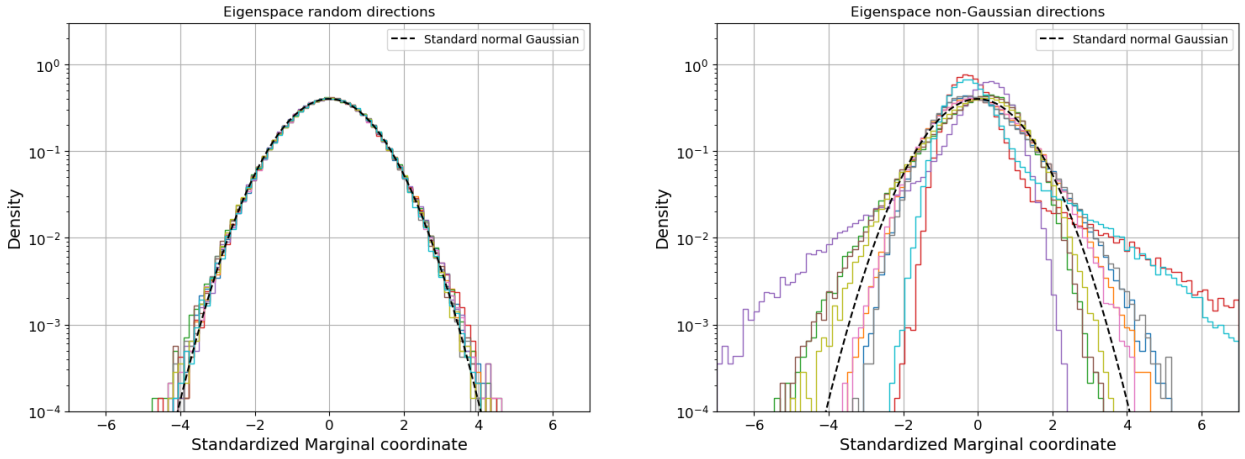


Figure 3: Marginals of the distribution in eigenspace coordinates $\vec{r}_\ell$ at $\ell = 2$ for a realization of $\vec{F}(\vec{x})$ with $N = 256$. Each panel shows the marginals along 10 different directions $\hat{u}\cdot\vec{r}_\ell$. The random directions $\hat{u}$ are uniformly sampled from $\mathbb{S}^{N_\ell - 1}$, while the "non-Gaussian" directions $\hat{u}$ are from the QR decomposition Equation 38.

essentially Gaussian. The top-right panel then shows the marginals for random directions $\hat{u}$, and here we find some more noticeable deviations from Gaussianity, but still close to Gaussian. In contrast, the bottom panel of Figure 2 shows the marginals for the directions corresponding to columns of W_ℓ,

$$\hat{u} = \frac{W_\ell \hat{e}_\alpha}{\|W_\ell \hat{e}_\alpha\|}, \tag{35}$$

and here we find a much higher degree of non-Gaussianity with obvious skew. Hence, these are labeled the "non-Gaussian directions".

The standard basis projections are linear combinations

$$z_\alpha = \sum_\beta W_{\alpha\beta} x_\beta \tag{36}$$

of components of the highly non-Gaussian $\vec{x}$. Since the matrix W is generated as Gaussian noise in a large dimension N, the magnitudes $|W_{\alpha\beta}|$ of the weights are roughly equal, while the signs are roughly evenly distributed between -1 and 1 . The distribution of z_α can then be understood intuitively as N convolutions[3] of roughly equally sized, asymmetric, and sharply truncated kernels. The equal distribution of signs flips half of those asymmetric kernels, and so after N convolutions the result is very nearly Gaussian.

On the other hand, if $\vec{W}_\alpha^\intercal$ is a column of W and $\hat{u} = \frac{\vec{W}_\alpha^\intercal}{\|\vec{W}_\alpha^\intercal\|}$, then

$$\hat{u}\cdot\vec{z} = \frac{1}{\|\vec{W}_\alpha^\intercal\|} \sum_\beta \left(\vec{W}_\alpha^\intercal \cdot \vec{W}_\beta^\intercal\right) x_\beta. \tag{37}$$

[3]The components of $\vec{x}$ are not independent so the distribution of z_α is technically given by the convolution-like integrals over the joint distribution:
https://en.wikipedia.org/wiki/Convolution_of_probability_distributions

Here $\vec{W}_\alpha^\intercal \cdot \vec{W}_\beta^\intercal \approx \delta_{\alpha\beta}$ since the N directions of independent standard normal vectors are nearly orthogonal for large N. Thus there is now a dominant term in the sum $\alpha = \beta$, and we can think of the other $N-1$ smaller terms producing relatively narrower convolution kernels applied to the asymmetric, sharply truncated distribution of x_α. This series of convolutions smooths out the hard edge of that asymmetric distribution, but because each convolution kernel is relatively narrow it remains asymmetric and the end result is a highly skewed distribution.

The projections of the columns of W_ℓ into the non-signular subspace coordinates $\vec{r}_\ell$ form the matrix $V_\ell^+ W_\ell$ which is $N_\ell \times N$. Since the (normalized) columns of W_ℓ are uniformly distributed on the sphere $\mathbb{S}^{N-1}$ the columns $V_\ell^+ W_\ell \hat{e}_n$ of this matrix are generally an over-complete basis that spans $\vec{r}_\ell \in \mathbb{R}^{N_\ell}$. We can find an orthonormal basis for $\mathbb{R}^{N_\ell}$ that is close to the set of vectors $V_\ell^+ W_\ell \hat{e}_n$ by the QR decomposition

$$\mathcal{Q}\mathcal{R} = V_\ell^+ W_\ell, \tag{38}$$

where $\mathcal{Q}$ is an orthogonal $N_\ell \times N_\ell$ matrix. Formally the QR decomposition corresponds to the Gramm-Schmidt orthogonalization process. The n-th column of $\mathcal{Q}$ is a unit vector $\hat{u}_n$ and the set of these N_ℓ directions are a complete orthonormal basis for $\vec{r}_\ell \in \mathbb{R}^{N_\ell}$. What makes this particular basis special is that it corresponds closely with the directions along which the non-Gaussianity in the marginal distributions of $Q_\ell(\vec{r}_\ell)$ are most significant. In contrast, the marginal of the distribution along any random direction in $\mathbb{R}^{N_\ell}$ has the effect of averaging out the non-Gaussian structure such that the resulting marginal is much more nearly Gaussian, as shown in Figure 3. So the set of directions $\{\hat{u}_n\}_{n=1}^{N_\ell}$ is the orthonormal basis in which we can actually detect the significant deviations from Gaussianity and thus capture it in our model.

6 Mixtures of Gaussians

Any function can be approximated as an expansion in Gaussian basis functions and if the coefficients are restricted to be positive such an expansion is a natural model for arbitrary probability density functions. Since a single Gaussian is already a good approximation, it seems natural to ask if we can improve the model by adding additional Gaussian components so that our approximations are like

$$Q(\vec{z}) = \frac{1}{Z} \sum_k e^{a_k} e^{-\frac{1}{2}(\vec{z}-\vec{m}_k)^\intercal C_k^{-1} (\vec{z}-\vec{m}_k)}. \tag{39}$$

If there was a good approximation to $P_{L-1}(\vec{z})$ of this form, then the final expectation integral of $(\vec{z})_+$ would simply be the weighted average of the analytic expectations for each component $\sum_k \omega_k \vec{\mathcal{E}}(\vec{m}_k, C_k)$.

Unfortunately, naively trying to extend the model this way makes the transformation approximation problem incredibly ugly. The variational problem for the parameters $(\omega_k, \vec{m}_k, C_k)$ of the next approximation $Q_{\ell+1}(\vec{z})$ with a KL divergence objective becomes

$$\mathcal{L} = -\int d\vec{z} Q_\ell(\vec{z}) \log\left(\sum_k e^{a_k} e^{-\frac{1}{2}(\vec{z}-\vec{m}_k)^\intercal C_k^{-1}(\vec{z}-\vec{m}_k)}\right) + \log\left(\sum_k e^{a_k} \sqrt{\det(2\pi C_k)}\right), \tag{40}$$

and the problem is obviously not efficiently soluble with some off-the-shelf non-linear solver.

From here we can try to simplify things again. Since a single Gaussian is already known to be a good approximation, we might make the covariance parameter the same for all components $C_k = S$. Further, we can make the model an explicit correction to the single Gaussian approximation $\mathcal{N}(\vec{z}, \vec{\eta}, S)$ as

$$Q_\ell(\vec{z}) = \frac{1}{Z_\ell} e^{-\frac{1}{2}(\vec{z}-\vec{\eta}_\ell)^\intercal S_\ell^{-1}(\vec{z}-\vec{\eta}_\ell)} \left(1 + \sum_{k>0} e^{a_k^\ell + \vec{b}_k^\ell \cdot S_\ell^{-1}\vec{z}}\right). \tag{41}$$

Then under the previous Gaussian mixture approximation $Q_{\ell-1}$

$$Q_{\ell-1}(\vec{z}) = \sum_{k=0}^{K} \omega_k^{\ell-1} \mathcal{N}\left(\vec{z}, \vec{\eta}_{\ell-1} + \vec{b}_k^{\ell-1}, S_{\ell-1}\right), \tag{42}$$

$$\omega_k^{\ell-1} = \frac{e^{a_k^{\ell-1} + \vec{b}_k^{\ell-1}\cdot S_{\ell-1}^{-1}\vec{\eta}_{\ell-1} + \frac{1}{2}\vec{b}_k^{\ell-1}\cdot S_{\ell-1}^{-1}\vec{b}_k^{\ell-1}}}{1 + \sum_{k>0} e^{a_k^{\ell-1} + \vec{b}_k^{\ell-1}\cdot S_{\ell-1}^{-1}\vec{\eta}_{\ell-1} + \frac{1}{2}\vec{b}_k^{\ell-1}\cdot S_{\ell-1}^{-1}\vec{b}_k^{\ell-1}}} \text{ with } a_0 = 0, \vec{b}_0 = 0. \tag{43}$$

the parameters $\vec{\eta}_\ell$, S_ℓ can be set by the analytic evaluation of the mean and covariance of $W_\ell(\vec{z}_{\ell-1})_+$. It then remains to determine the parameters a_k^ℓ, $\vec{b}_k^\ell$ of the corrections

With $\vec{\eta}, S$ held fixed, the part of the KL divergence varying with a_k, $\vec{b}_k$ is

$$\mathcal{L}_2 = \log\left(1 + \sum_{k>0} e^{a_k + \vec{b}_k\cdot S^{-1}\vec{\eta} + \frac{1}{2}\vec{b}_k\cdot S^{-1}\vec{b}_k}\right) - \int d\vec{z} Q_{\ell-1}(\vec{z}) \log\left(1 + \sum_{k>0} \exp\left(a_k + \vec{b}_k \cdot S^{-1}\vec{G}_{\ell-1}(\vec{z})\right)\right). \tag{44}$$

This is perhaps getting somewhat less ugly - it is a difference of convex functions of the parameters Yuille and Rangarajan (2001) and its plausible that the $\vec{b}_k$ could be warm-started with good local estimates. But approximating the integral remains a formidable obstacle, so this is still far from practical.

Avenues for further progress down this road could involve alternative objective functions such as score matching Hyvärinen and Dayan (2005)

$$\mathcal{L} = \int d\vec{z} P(\vec{z}) \frac{1}{2} \left\| \frac{\partial \log Q(\vec{z})}{\partial \vec{z}^\intercal} - \frac{\partial \log P(\vec{z})}{\partial \vec{z}^\intercal} \right\|^2 \tag{45}$$

$$= \mathcal{L}_0 + \int d\vec{z} P(\vec{z}) \left[\frac{1}{2} \left\| \frac{\partial \log Q(\vec{z})}{\partial \vec{z}^\intercal} \right\|^2 + \mathrm{Tr}\left[\frac{\partial \log Q(\vec{z})}{\partial \vec{z} \partial \vec{z}^\intercal} \right] \right]. \tag{46}$$

This removes the normalization term from the objective, but the main problem of approximating the integral quickly and accurately enough to be used in an iterative non-linear solver remains. There may be ways of exploiting the directional structure $\vec{b}_k$, but that is vague wishful thinking at this point.

7 Ridge Function Approximations

Another approach to building on the Gaussian approximation could be based on the non-Gaussian directions of section 5. Assuming we have an initial Gaussian approximation

$\mathcal{N}(\vec{z}, \vec{\eta}, S)$, we model the residual in the eigenspace coordinates $\vec{r}$ (section 4) so the distribution is

$$Q(\vec{r}) = -\log(Z) - \frac{1}{2}|\vec{r}|^2 - U(\vec{r}). \tag{47}$$

Then we can define the variational approximation by the score-matching objective Hyvärinen and Dayan (2005); Lyu (2012)

$$\mathcal{L} = \int d\vec{r} Q_0(\vec{r}) \left[\tfrac{1}{2}\left| \nabla \log Q(\vec{g}(\vec{r})) \right|^2 + \nabla^2 \log Q(\vec{g}(\vec{r}))\right], \tag{48}$$

$$= \int d\vec{r} Q_0(\vec{r}) \left[\tfrac{1}{2}\left|\vec{g}(\vec{r}) + \nabla U(\vec{g}(\vec{r}))\right|^2 - \nabla^2 U(\vec{g}(\vec{r}))\right] + \text{ const.} \tag{49}$$

As shown in section 5 we can find the orthogonal basis directions $\{\hat{u}_n\}_{n=1}^{N_\ell}$ along which the deviations from Gaussianity are concentrated, and so we model the residual as an expansion in ridge functions Candès (2003); Ismailov (2021)

$$U(\vec{r}) = \sum_{n=1}^{N_\ell} \psi_n(\hat{u}_n \cdot \vec{r}). \tag{50}$$

Each $\psi_n(r)$ is a 1D function, and since the $\hat{u}_n$ are an orthonormal basis this makes the approximation $Q(\vec{r})$ separable in the coordinates $r_n = \hat{u}_n \cdot \vec{r}$,

$$Q(\vec{r}) = \frac{1}{Z} e^{-\frac{1}{2}|\vec{r}|^2} e^{-\sum_n \psi_n(\hat{u}_n \cdot \vec{r})} \tag{51}$$

$$= \prod_n \frac{1}{Z_n} e^{-\frac{1}{2}r_n^2} e^{-\psi_n(r_n)}, \tag{52}$$

$$= \prod_n Q_n(r_n) \tag{53}$$

$$\text{where } Z_n = \int_{-\infty}^{\infty} dr e^{-\frac{1}{2}r^2} e^{-\psi_n(r)}. \tag{54}$$

We can further model the ψ_n as a linear combination of a few fixed basis functions e.g. as a cubic spline in the B-spline basis

$$\psi_n(r) = \sum_{k=1}^{K} \rho_{nk} \phi\left(\tfrac{1}{\Delta}(r - \bar{r}_k)\right) \tag{55}$$

where ϕ is the B-spline basis kernel, and the interval containing the equally Δ-spaced knots $\bar{r}_k$ spans the effective support of the distribution (c.f. Figure 3).

With this parameterization, the score-matching objective function becomes a quadratic function of the parameters ρ_{nk}, and further make it separable over n,

$$\mathcal{L} = \int d\vec{r} Q_0(\vec{r}) \sum_{n=1}^{N_\ell} \left[\frac{1}{2}\left(\frac{\partial \log Q_n}{\partial r_n}\left(\hat{u}_n \cdot \vec{g}(\vec{r})\right)\right)^2 + \frac{\partial^2 \log Q_n}{\partial r_n^2}\left(\hat{u}_n \cdot \vec{g}(\vec{r})\right)\right] \tag{56}$$

$$= \sum_{n=1}^{N_\ell} \mathcal{L}_n. \tag{57}$$

Additional regularization may be added, for example a smoothing second derivative penalty. For each n, the coefficients ρ_{nk} are then the solution to a small $K \times K$ linear system $A_n \vec{\rho}_n = \vec{b}_n$.

The expensive part is computing the matrices and vectors $A_n, \vec{b}_n$, e.g. from integrals over pairs of basis function derivatives

$$A_{n,kk'} \propto \int d\vec{r} Q_0(\vec{r}) \phi'\left(\frac{1}{\Delta}\left(\hat{u}_n \cdot \vec{g}(\vec{r}) - \bar{r}_k\right)\right) \phi'\left(\frac{1}{\Delta}\left(\hat{u}_n \cdot \vec{g}(\vec{r}) - \bar{r}_{k'}\right)\right) + \text{regularization}. \tag{58}$$

It remains to be seen if these integrals could be approximated efficiently enough at sufficient accuracy to make this model useful.

8 Fourier Decomposition

Any smooth distribution $Q(\vec{r})$ we might want to approximate can be decomposed in Fourier modes as

$$Q(\vec{r}) = \int d\vec{u} e^{i2\pi \vec{u} \cdot \vec{r}} \tilde{Q}(\vec{u}). \tag{59}$$

The Fourier transform of the distribution (a.k.a. the characteristic function) is defined by

$$\tilde{Q}(\vec{u}) = \int d\vec{r} e^{-i2\pi \vec{u} \cdot \vec{r}} Q(\vec{r}) \tag{60}$$

$$= \mathbb{E}_{\vec{r} \sim Q}\left[e^{-i2\pi \vec{u} \cdot \vec{r}}\right]. \tag{61}$$

Fourier modes are ridge functions (the original) with a preferred spatial direction, and we can construct a Fourier model informed by the section 5 non-Gaussian directions. Similarly to the previous spline model, assume the distribution can be approximated as separable 1D factors along the axes $\hat{u}_n$,

$$Q(\vec{r}) = \prod_n Q_n(\hat{u}_n \cdot \vec{r}). \tag{62}$$

Then the Fourier transform is similarly separable:

$$\tilde{Q}(\vec{u}) = \int d\vec{r} e^{-i2\pi \vec{u} \cdot \vec{r}} \prod_n Q_n(\hat{u}_n \cdot \vec{r}) \tag{63}$$

$$= \prod_n \int dr e^{-i2\pi \hat{u}_n \cdot \vec{u} r} Q_n(r) \tag{64}$$

$$= \prod_n \tilde{Q}_n(\hat{u}_n \cdot \vec{u}). \tag{65}$$

The approximation for the distribution could then be something like

$$Q(\vec{r}) = \square(\vec{r}) \prod_n \left(\sum_k \tilde{Q}_{nk} e^{i2\pi \vec{u}_{nk} \cdot \vec{r}}\right), \tag{66}$$

where $\vec{u}_{nk} = u_k \hat{u}_n$ for some radial coordinates u_k, and $\square(\vec{r})$ is a box function aligned with the axes $\hat{u}_n$. A model like this may allow for efficient expectation integrals based on a series of 1D quadrature rules, and make use of tools like the FFT/NUFFT.

9 Fourier Propagator

It is straightforward to write down a formal expression for the transformation of the Fourier modes induced by a change of variables. Under a transformation $\vec{r}_1 = \vec{g}(\vec{r}_0)$, the distribution transforms as

$$Q_1(\vec{r}_1) = \int d\vec{r}\delta(\vec{r}_1 - \vec{g}(\vec{r}))Q_0(\vec{r}). \tag{67}$$

Taking the Fourier transform over $\vec{r}_1$ of both sides gives us

$$\tilde{Q}_1(\vec{u}) = \int d\vec{r}e^{-i2\pi\vec{u}\cdot\vec{g}(\vec{r})}Q_0(\vec{r}), \tag{68}$$

$$= \mathbb{E}_{\vec{r}\sim Q}\left[e^{-i2\pi\vec{u}\cdot\vec{g}(\vec{r})}\right]. \tag{69}$$

Then inserting the Fourier integral representation of Q_0 we have

$$\tilde{Q}_1(\vec{u}) = \int d\vec{v}\tilde{Q}_0(\vec{v}) \int d\vec{r}e^{i2\pi\vec{v}\cdot\vec{r}}e^{-i2\pi\vec{u}\cdot\vec{g}(\vec{r})}. \tag{70}$$

This defines a propagator

$$K([\vec{g}], \vec{u}, \vec{v}) = \int d\vec{r}e^{i2\pi\vec{v}\cdot\vec{r}}e^{-i2\pi\vec{u}\cdot\vec{g}(\vec{r})} \tag{71}$$

where the argument $[\vec{g}]$ denotes that K is a functional of $\vec{g}(\vec{r})$, and the transformation can be written as a linear transformation of the modes

$$\tilde{Q}_1(\vec{u}) = \int d\vec{v}K([\vec{g}], \vec{u}, \vec{v})\tilde{Q}_0(\vec{v}). \tag{72}$$

9.1 Quadratic Approximation

Suppose $\vec{g}(\vec{r})$ is a quadratic function with components $g_n(\vec{r}) = c_n + \vec{J}_n \cdot \vec{r} + \frac{1}{2}\vec{r}^\intercal H_n \vec{r}$ (where H_n is a matrix for each n). Then the propagator is a Gaussian integral which can be simplified to

$$K_0([\vec{g}], \vec{u}, \vec{v}) = \frac{e^{-i2\pi\vec{u}\cdot\vec{c}}}{\sqrt{\det\left(i\vec{u}\cdot\vec{H}\right)}} \exp\left(i\pi(\vec{v} - J^T\vec{u})^\intercal(\vec{u}\cdot\vec{H})^{-1}(\vec{v} - J^T\vec{u})\right) \tag{73}$$

where $\vec{u}\cdot\vec{H} = \sum_n u_n H_n$ and J are matrices. More generally if write the error with respect to this quadratic explicitly as $\vec{\delta}(\vec{r})$ so $g_n(\vec{r}) = c_n + \vec{J}_n \cdot \vec{r} + \frac{1}{2}\vec{r}^\intercal H_n \vec{r} + \delta_n(\vec{r})$, then

$$e^{-i2\pi\vec{u}\cdot\vec{g}(\vec{r})} = e^{-i2\pi(\vec{u}\cdot\vec{c} + \vec{u}\cdot J\vec{r} + \frac{1}{2}\vec{r}^\intercal(\vec{u}\cdot\vec{H})\vec{r})}\left(1 - i2\pi\vec{u}\cdot\vec{\delta}(\vec{r}) + \mathcal{O}((\vec{u}\cdot\vec{\delta})^2)\right) \tag{74}$$

and the propagator can be expressed as a series of perturbative corrections to the Gaussian.

Consider constructing a quadratric approximation to $\vec{g}(\vec{r})$. A Taylor series expansion of $\vec{G}(\vec{z}) = W_1(\vec{z})_+$ (and thus $\vec{g}(\vec{r})$) at a point is not helpful, but assuming we know the mean

$\vec{\eta}_0$ and covariance S_0 we can construct a quadratic function which is similar in a statistical sense. That is, the best quadratic approximation to $\vec{G}(\vec{z})$ averaged over the distribution Q_0. This is obtained by optimizing

$$\mathbb{E}\left[\left\|\vec{G}(\vec{z}) - \vec{\eta}_1 - A(\vec{z} - \vec{\eta}_0) - \tfrac{1}{2}\sum_{\alpha\beta}\vec{B}_{\alpha\beta}\big((z_\alpha - \eta_{0\alpha})(z_\beta - \eta_{0\beta}) - S_{0\alpha\beta}\big)\right\|^2\right] \tag{75}$$

which produces

$$A_{\alpha\beta} = \sum_\nu W_{1\alpha\nu}\mathbb{E}\left[\frac{\partial(z_\nu)_+}{\partial z_\beta}\right] \tag{76}$$

$$= \sum_\nu W_{1\alpha\nu}\Phi_\nu\delta_{\nu\beta} \tag{77}$$

$$B_{n\alpha\beta} = \sum_\nu W_{1n\nu}\mathbb{E}\left[\frac{\partial(z_\nu)_+}{\partial z_\alpha \partial z_\beta}\right] \tag{78}$$

$$= \sum_\nu W_{1n\nu}\phi_\nu\delta_{\nu\alpha}\delta_{\nu\beta} \tag{79}$$

and hence the quadratic approximation

$$G_\alpha(\vec{z}) \approx \eta_{1\alpha} + \sum_\beta W_{1\alpha\beta}\left[\Phi_\beta(z_\beta - \eta_{0\beta}) + \tfrac{1}{2}\phi_\beta\left((z_\beta - \eta_{0\beta})^2 - S_{0\beta\beta}\right)\right] \tag{80}$$

where (as defined in Appendix C)

$$\Phi_\beta = \Phi\left(\frac{\eta_{0\beta}}{\sqrt{S_{0\beta\beta}}}\right) \tag{81}$$

$$\phi_\beta = \tfrac{1}{\sqrt{S_{0\beta\beta}}}\phi\left(\frac{\eta_{0\beta}}{\sqrt{S_{0\beta\beta}}}\right). \tag{82}$$

10 Particle Approximations and Renormalization Flows

A "particle approximation"

$$Q(\vec{x}) \approx \sum_j \omega_j \delta(\vec{x} - \vec{x}_j) \tag{83}$$

approximates the distribution $P(\vec{x})$ in the sense that it approximates integrals over that distribution as

$$\int d\vec{x} P(\vec{x}) f(\vec{x}) \approx \sum_j \omega_j f(\vec{x}_j). \tag{84}$$

So a "particle approximation" is the same thing as a "quadrature rule" and standard Monte Carlo sampling is constructing such an approximation with M pseudo-random samples $\vec{x}_j \sim P$ and uniform weights $\omega_j = \frac{1}{M}$.

Consider that when we push these samples through the layers of a composite function $\vec{F}(\vec{x})$ we are actually computing the pushforward of the particle approximation at each layer ℓ. For a map $\vec{y} = \vec{f}(\vec{x})$ each component transforms according to the law given in section 2 as

$$\int d\vec{x}\delta(\vec{y} - \vec{f}(\vec{x}))\delta(\vec{x} - \vec{x}_j) = \delta(\vec{y} - \vec{f}(\vec{x}_j)) \tag{85}$$

and so for each ℓ we are computing the next particle approximation

$$Q_{\ell+1}(\vec{x}_{\ell+1}) = \int d\vec{x}\delta\Big(\vec{x}_{\ell+1} - \vec{F}_\ell(\vec{x})\Big) \sum_j \frac{1}{M}\delta(\vec{x} - \vec{x}_{\ell,j}) \tag{86}$$

$$= \sum_j \frac{1}{M}\delta\Big(\vec{x}_{\ell+1} - \vec{F}_\ell(\vec{x}_{\ell,j})\Big) \tag{87}$$

$$= \sum_j \frac{1}{M}\delta(\vec{x}_{\ell+1} - \vec{x}_{\ell+1,j})\,. \tag{88}$$

But said another way, this is how we are constructing a new quadrature rule for the weight function $P_{\ell+1}(\vec{x}_{\ell+1})$. This construction makes no use of the details of the function $\vec{F}_\ell$ - it is the same for any function. Similarly, drawing i.d.d. samples from P uses no information about the specific structure of P in forming the initial approximation Q. Since $P_{\ell+1}$ is determined by the combination of P_ℓ and $\vec{F}_\ell$, each new quadrature rule is similarly independent of any information about the structure of $P_{\ell+1}$.

In contrast, more efficient quadrature rules are determined based on some assumptions about the structure of the function for which the integral is to be approximated. Given the initial Gaussian distribution for P_0, we can generally construct a quadrature rule/particle approximation for it that is more efficient than naive Monte Carlo sampling for the relatively simple individual layer functions $\vec{F}_\ell$ (see Martinot). We can also see from MC simulation and the Gaussian closure that the incremental push-forward for each $\vec{z}_\ell$ stays at least "vaguely Gaussian" - approximated by a smooth, unimodal distribution that is mostly contained inside some elliptic shell. Thus it seems there should be particle approximations that are similarly improved over naive MC samples for each P_ℓ.

It then seems reasonable to ask if there is a different way to approximate the push-forward from a particle approximation

$$\{(\omega_{\ell,j}, \vec{x}_{\ell,j})\}_{j=1}^{M_\ell} \mapsto \{(\omega_{\ell+1,j}, \vec{x}_{\ell+1,j})\}_{j=1}^{M_{\ell+1}}$$

that is informed by the specific function $\vec{F}_\ell$, such that the new approximation $(\omega_{\ell+1,j}, \vec{x}_{\ell+1,j})$ is as good an estimator for $\int d\vec{x}P_{\ell+1}(\vec{x})\vec{F}_{\ell+1}(\vec{x})$ as the initial approximation $(\omega_{\ell,j}, \vec{x}_{\ell,j})$ was for $\int d\vec{x}P_\ell(\vec{x})\vec{F}_\ell(\vec{x})$.

Food for thought in this area:

- Particle Flows: Daum and Huang (2012), Daum and Huang (2015)
- Generalized Quadrature: Keshavarzzadeh et al. (2018); Trefethen (2022)

A Integral Representations of $(z)_+$

The function $(z)_+$ has several integral representations. The Laplace transform is $1/s^2$, so the inverse transform (Bromwich integral) gives us

$$(z)_+ = \frac{1}{2\pi}\int_{-\infty}^{\infty} du \frac{e^{\gamma z + iuz}}{(\gamma + iu)^2}, \text{ for } \gamma > 0. \tag{89}$$

Then in addition:

$$\begin{aligned}
(z)_+ &= \frac{1}{2}z + \frac{1}{2}|z| && (90)\\
&= \frac{1}{2}z + \frac{1}{\pi}\int_0^{\infty} du \frac{1-\cos(uz)}{u^2} \text{ (Dirichlet integral)} && (91)\\
&= \frac{1}{2}z + \frac{z^2}{\pi}\int_0^{\infty} du \frac{1}{z^2+u^2} \text{ (Stieltjes transform)} && (92)\\
&= \frac{1}{2}z + \frac{1}{4\sqrt{\pi}}\int_0^{\infty} du \frac{1-e^{-uz^2}}{u^{3/2}} \text{ (Schwinger trick)} && (93)
\end{aligned}$$

These could be useful for treating integrals of $(\vec{V}\cdot\vec{r}+\vec{\eta})_+$, i.e. the mean under an approximation $Q(\vec{r})$. In particular, in terms of the Fourier transform $\tilde{Q}(\vec{u})$ the cosine form gives

$$\int d\vec{r} Q(\vec{r})(\vec{V}\cdot\vec{r}+\eta)_+ = \frac{1}{2}\eta + \frac{1}{\pi}\int_0^{\infty} du \frac{1}{u^2}\left(1 - \mathrm{Re}\left[e^{iu\eta}\tilde{Q}(u\vec{V})\right]\right). \tag{94}$$

This can approximated accurately by a trapezoid rule. Also see De Hoog et al. (1982) for inverse Laplace quadrature.

B Gaussian Transport

If the distributions of $\vec{z}_0$ and $\vec{z}_1 = \vec{G}(\vec{z}_0)$ are both close to Gaussian, there is a sense in which the map $\vec{G}$ is close to some affine map, since the affine transformation of a Gaussian is another Gaussian. If $\vec{\eta}_0, S_0$ and $\vec{\eta}_1, S_1$ are the mean and covariance of $\vec{z}_0, \vec{z}_1$ respectively, then we can ask: what affine map $\vec{h}(\vec{z}) = A\vec{z} + \vec{b}$ transports $\mathcal{N}(\vec{\eta}_0, S_0)$ to $\mathcal{N}(\vec{\eta}_1, S_1)$? This could be an affine surrogate that is highly correlated with $\vec{G}$.

The parameters A and $\vec{b}$ satisfy

$$\begin{aligned}
\vec{b} &= \vec{\eta}_1 - A\vec{\eta}_0 && (95)\\
S_1 &= AS_0A^\intercal. && (96)
\end{aligned}$$

This determines $\vec{b}$ given A, but the condition for A is degenerate, since for any factorization $S_0 = V_0V_0^\intercal$ the transformation $V_0 \to V_0O$ with any orthogonal matrix O leaves it unchanged.

We can break this degeneracy by choosing A such that $\vec{h}(\vec{z})$ is as close as possible to another affine approximation $J(\vec{z}-\vec{\eta}_0)+\vec{\eta}_1$ where

$$J = \mathbb{E}_{\vec{z}\sim\mathcal{N}(\vec{\eta}_0, S_0)}\left[\frac{\partial \vec{G}(\vec{z})}{\partial \vec{z}^\intercal}\right], \tag{97}$$

i.e. the best affine approximation to $\vec{G}$ averaged over $\mathcal{N}(\vec{\eta}_0, S_0)$. Now A is determined by the constrained optimization

$$\min_A \mathbb{E}_{\vec{x}\sim\mathcal{N}(\vec{\eta}_0,S_0)} \|A(\vec{x}-\vec{\eta}_0) - J(\vec{x}-\vec{\eta}_0)\|^2 \text{ s.t. } S_1 = AS_0A^\intercal \tag{98}$$

According to a language model, the solution is as follows. The expectation reduces the problem to

$$\min_A \mathrm{Tr}[(A-J)S_0(A-J)^\intercal] \text{ s.t. } S_1 = AS_0A^\intercal. \tag{99}$$

Let the eigendecompositions of the (possibly low-rank) covariance matrices be

$$S_0 = \mathcal{U}_0\Lambda_0\mathcal{U}_0^\intercal \tag{100}$$
$$S_1 = \mathcal{U}_1\Lambda_1\mathcal{U}_1^\intercal. \tag{101}$$

Then with

$$P_0 = \mathcal{U}_0\mathcal{U}_0^\intercal, \text{ projector on to range of } S_0 \tag{102}$$
$$M = \Lambda_1^{1/2}\mathcal{U}_1^\intercal J\mathcal{U}_0\Lambda_0^{1/2} \tag{103}$$
$$M = R\Sigma W^\intercal, \text{ (SVD)} \tag{104}$$
$$Q_* = RW^\intercal \tag{105}$$

the solution is

$$A = \mathcal{U}_1\Lambda_1^{1/2}Q_*\Lambda_0^{-1/2}\mathcal{U}_0^\intercal + J(I-P_0). \tag{106}$$

In Monte Carlo testing, the resulting affine approximation $A(\vec{z}-\vec{\eta}_0)+\vec{\eta}_1$ has higher correlation with the exact $\vec{G}(\vec{z}) = W(\vec{z})_+$ than does $J(\vec{z}-\vec{\eta}_0)+\vec{\eta}_1$.

Further, the composition of these Gaussian transport maps over layers ℓ remains well correlated with the true trajectories $\vec{z}_\ell$ when using the Gaussian closure approximation data for $\vec{\eta}_\ell, S_\ell$.

C Gaussian Closure Integrals

Let $\sigma_\alpha = \sqrt{S_{\alpha\alpha}}$ and define $r_\alpha = \frac{\eta_\alpha}{\sigma_\alpha}$. For the first moment of the next layer we have

$$\langle (z_\alpha)_+ \rangle = \frac{1}{(2\pi)^{N/2}\sqrt{\det C}} \int d\vec{z} e^{-\frac{1}{2}(\vec{z}-\vec{\eta})^\intercal C^{-1}(\vec{z}-\vec{\eta})}(z_\alpha)_+ \tag{107}$$
$$= \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \int_0^\infty dz \exp\left(-\frac{(z-\eta_\alpha)^2}{2\sigma_\alpha^2}\right) z \tag{108}$$
$$= \eta_\alpha\Phi(r_\alpha) + \sigma_\alpha\phi(r_\alpha), \tag{109}$$

after substituting $u = (z-\eta_\alpha)/\sigma_\alpha$, where

$$\phi(r) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}r^2}, \tag{110}$$
$$\Phi(r) = \int_{-\infty}^r dx\phi(x) = \frac{1}{2}\left(1 + \mathrm{erf}(r/\sqrt{2})\right). \tag{111}$$

Denote the correlation coefficient by $\rho = \frac{S_{\alpha\beta}}{\sqrt{S_{\alpha\alpha}S_{\beta\beta}}}$. The raw second moment is then

$$\langle (z_\alpha)_+ (z_\beta)_+ \rangle = \frac{1}{2\pi\sigma_\alpha\sigma_\beta\sqrt{1-\rho^2}} \int_0^\infty dz_\alpha \int_0^\infty dz_\beta z_\alpha z_\beta \tag{112}$$

$$\exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(z_\alpha - \eta_\alpha)^2}{\sigma_\alpha^2} - 2\rho \frac{(z-\eta_\alpha)}{\sigma_\alpha} \frac{(z_\beta - \eta_\beta)}{\sigma_\beta} + \frac{(z_\beta - \eta_\beta)^2}{\sigma_\beta^2} \right] \right)$$

$$= \int_{-r_\alpha}^\infty du \int_{-r_\beta}^\infty dv f_\alpha(u) f_\beta(v) \phi_2(u, v, \rho) \tag{113}$$

where we define

$$f_\alpha(u) = \eta_\alpha + \sigma_\alpha u \tag{114}$$

$$\phi_2(u, v, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{u^2 - 2\rho uv + v^2}{2(1-\rho^2)} \right), \tag{115}$$

and denote the covariance by

$$I_{\alpha\beta}(\rho) = \langle (z_\alpha)_+ (z_\beta)_+ \rangle - \langle (z_\alpha)_+ \rangle \langle (z_\beta)_+ \rangle . \tag{116}$$

We simplify through Price's theorem - for any function $g(u, v)$ of the bivariate normally distributed variables (u, v) with correlation ρ, the expectation $\langle g(u, v) \rangle$ satisfies

$$\frac{\partial}{\partial\rho} \langle g(u, v) \rangle = \left\langle \frac{\partial g(u, v)}{\partial u \partial v} \right\rangle. \tag{117}$$

Hence, we have

$$\frac{\partial I_{\alpha\beta}(\rho)}{\partial\rho} = \left\langle \frac{\partial f_\alpha(u)}{\partial u} \frac{\partial f_\beta(v)}{\partial v} \right\rangle \tag{118}$$

$$= \sigma_\alpha\sigma_\beta \Phi_2(r_\alpha, r_\beta, \rho), \tag{119}$$

where Φ_2 is the bivariate cumulative distribution function

$$\Phi_2(r_\alpha, r_\beta, \rho) = \int_{-\infty}^{r_\alpha} du \int_{-\infty}^{r_\beta} dv \phi_2(u, v, \rho). \tag{120}$$

Now from the integral form of Plackett's identity

$$\partial_\rho \Phi_2(u, v, \rho) = \phi_2(u, v, \rho) \longleftrightarrow \Phi_2(u, v, \rho) = \Phi(u)\Phi(v) + \int_0^\rho dt \phi_2(u, v, t) \tag{121}$$

we can write

$$\frac{\partial I_{\alpha\beta}(\rho)}{\partial\rho} = \sigma_\alpha\sigma_\beta \Phi(r_\alpha)\Phi(r_\beta) + \sigma_\alpha\sigma_\beta \int_0^\rho dt \phi_2(r_\alpha, r_\beta, t). \tag{122}$$

Noting that $I_{\alpha\beta}(0) = 0$, we integrate up to ρ to obtain

$$I_{\alpha\beta}(\rho) = \rho\sigma_\alpha\sigma_\beta\Phi(r_\alpha)\Phi(r_\beta) + \sigma_\alpha\sigma_\beta \int_0^\rho dt \int_0^t ds\phi_2(r_\alpha, r_\beta, s), \tag{123}$$

and by Cauchy's formula for repeated integration and $\rho\sigma_\alpha\sigma_\beta = S_{\alpha\beta}$ we have

$$I_{\alpha\beta}(\rho) = \Phi(r_\alpha)S_{\alpha\beta}\Phi(r_\beta) + \sigma_\alpha\sigma_\beta \int_0^\rho dt(\rho - t)\phi_2(r_\alpha, r_\beta, t). \tag{124}$$

Finally, through a change of variables $t = \rho(1 - v^2)$ we obtain

$$I_{\alpha\beta}(\rho) = \Phi(r_\alpha)S_{\alpha\beta}\Phi(r_\beta) + \sigma_\alpha\sigma_\beta\rho^2 \int_0^1 dv2v^3\phi_2(r_\alpha, r_\beta, \rho(1 - v^2)). \tag{125}$$

This form of the integral term avoids singularities and can be evaluated cheaply and accurately by a shifted $(0, 3)$ Gauss-Jacobi rule,

$$\rho^2 \int_0^1 dv2v^3\phi_2(r_\alpha, r_\beta, \rho(1 - v^2)) \approx \frac{\rho^2}{2\pi} \sum_j w_j \frac{1}{\sqrt{1 - t_j^2}} \exp\left(-\frac{r_\alpha^2 + r_\beta^2 - 2r_\alpha r_\beta t_j}{2(1 - t_j^2)}\right), \tag{126}$$

$$t_j = \rho(1 - v_j^2), \tag{127}$$

where v_j, w_j are the appropriate nodes and weights.

References

Emmanuel J Candès. Ridgelets: estimating with ridge functions. *The Annals of Statistics*, 31(5):1561–1599, 2003.

Fred Daum and Jim Huang. Particle flow and monge-kantorovich transport. In *2012 15th International Conference on Information Fusion*, pages 135–142. IEEE, 2012.

Fred Daum and Jim Huang. renormalization group flow in k-space for nonlinear filters, bayesian decisions and transport. In *2015 18th International Conference on Information Fusion (Fusion)*, pages 1617–1624. IEEE, 2015.

Frank R De Hoog, John H Knight, and AN Stokes. An improved method for numerical inversion of laplace transforms. *SIAM Journal on scientific and Statistical Computing*, 3 (3):357–366, 1982.

Daniel T Gillespie. A theorem for physicists in the theory of random variables. *American Journal of Physics*, 51(6):520–533, 1983.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4):695, 2005.

Vugar E Ismailov. *Ridge functions and applications in neural networks*, volume 263. American Mathematical Society, Providence, Rhode Island, 2021.

Vahid Keshavarzzadeh, Robert M Kirby, and Akil Narayan. Numerical integration in multiple dimensions with designed quadrature. *SIAM Journal on Scientific Computing*, 40(4): A2033–A2061, 2018.

Siwei Lyu. Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*, 2012.

Zachary Martinot. Notes on the problem (phase 1). URL https://discourse.aicrowd.com/t/notes-on-the-problem-phase-1/18188?u=zacharymartinot.

Lloyd N Trefethen. Exactness of quadrature formulas. *Siam Review*, 64(1):132–150, 2022.

Wilson Wu, Victor Lecomte, Michael Winer, George Robinson, Jacob Hilton, and Paul Christiano. Estimating the expected output of wide random mlps more efficiently than sampling, 2026. URL https://arxiv.org/abs/2605.05179.

Alan L Yuille and Anand Rangarajan. The concave-convex procedure (cccp). *Advances in neural information processing systems*, 14, 2001.