Notes on the ARC White-box Challenge Problem (Phase 1)

Zachary Martinot

August 2026

Contents

1 The Problem

The problem is to accurately estimate 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, the input distribution for the challenge is the standard normal distribution

$$P(\vec{x}) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|\vec{x}|^2}, \tag{2}$$

so our integral is

$$\langle \vec{F}(\vec{x}) \rangle = \frac{1}{(2\pi)^{N/2}} \int_{\mathbb{R}^N} d\vec{x} e^{-\frac{1}{2}|\vec{x}|^2} \vec{F}(\vec{x}). \tag{3}$$

Additionally, the specific functions of interest are defined by the L-layered composition

$$\vec{F}(\vec{x}) = \left(\vec{F}_{L-1} \circ \vec{F}_{L-2} \circ \ldots \circ \vec{F}_0\right)(\vec{x}), \tag{4}$$

or equivalently by the recursion

$$\vec{x}_{\ell+1} = \vec{F}_\ell(\vec{x}_\ell), \text{ for } \ell = 0, 1, \ldots, L-1, \text{ with } \vec{x}_0 = \vec{x}, \tag{5}$$

$$\vec{F}(\vec{x}) = \vec{x}_L. \tag{6}$$

Each layer in the composition is

$$\vec{F}_\ell(\vec{x}_\ell) = \left(W_\ell \vec{x}_\ell\right)_+, \tag{7}$$

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, the dimension is $N = 256$ and there are $L = 32$ composed layers.

The impetus of the challenge is to do better than naive Monte Carlo "sampling". Rather than a special separate thing, we can simply view naive Monte Carlo sampling as a family of quadrature[1] rules, defined by nodes $\vec{q}_k$ and weights w_k as

$$\int_{\mathbb{R}^N} d\vec{x} P(\vec{x}) \vec{F}(\vec{x}) \approx \sum_{k=1}^{K} w_k \vec{F}(\vec{q}_k). \tag{8}$$

Given K, the naive Monte Carlo nodes are chosen by sampling $\vec{q}_k$ i.i.d. from the distribution $P(\vec{x})$, and the weights are taken to be $w_k = \frac{1}{K}$ uniformly. Averaged over the family, the error for these rules scales as $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$.

When compared to the classical methods of quadrature in one dimension, this error scaling seems quite poor. For "nice" integrands $f(x)$, simple Riemann sums may already converge like $\mathcal{O}\left(\frac{1}{K}\right)$ and this is easily improved to $\mathcal{O}\left(\frac{1}{K^n}\right)$ with simple composite Newton-Cotes (i.e. trapezoid, Simpsons) rules. More sophisticated methods like Gaussian or Clenshaw-Curtis quadrature rules can push this even further to the point that integrals of "smooth" but otherwise arbitrary functions can often be evaluated quite cheaply. In 1D, integral approximation is in many respects considered a "solved" problem.

The generally outstanding problem is that the naive transfer of these methods to dimensions $N >> 1$ suffers from the "curse of dimensionality" - tensor products of 1D quadratures produce node counts that blow up exponentially with N. On the other hand, the $\frac{1}{\sqrt{K}}$ scaling of naive Monte Carlo sampling is independent of dimension and so in high dimensions this simple method seems attractive again.

The question, then, is whether it is possible to once again do better in the high-dimensional regime.

The naive Monte Carlo quadrature is essentially the "method of complete ignorance". You can use it to integrate any function without knowing much of anything about it, but the

[1] or, cubature if you like - I think the distinction is archaic and will use the terms interchangeably

price of that ignorance is to spend time/energy beating down the worst possible convergence rate. In order to do better, you need to *know something* about the kind of function begin integrated and exploit that information to skip doing redundant computational work.

There is good reason to generically expect that, *yes, it is possible*. The key idea that leads to efficient quadrature rules is to exploit some kind of *structure* in the function to be integrated. Functions with no structure are called "noise", and the only thing we might want to do with them is integrate them away (at the frustrating slow/expensive rate of $\frac{1}{\sqrt{K}}$) to uncover something that is actually interesting - something with structure. The classical form of structure is assuming a function is "smooth" in the sense that it is similar to a finite degree polynomial (algebraic or trigonometric). But this not the only kind of structure that might be exploitable. Most functions of "practical" interest will posses some kind of structure or regularity - otherwise, we[2] wouldn't care about them in the first place.

2 Homogeneity

The function $\vec{F}(\vec{x})$ is positive-homogeneous. Observe that for each layer $\vec{F}_\ell(\vec{x}_\ell)$, you can pull a positive constant a through $(\vec{z})_+$, so

$$\vec{F}_\ell(a\vec{x}_\ell) = (aW_\ell\vec{x}_\ell)_+ = a(W_\ell\vec{x}_\ell)_+ = a\vec{F}_\ell(\vec{x}_\ell),$$

and since this is the case for each ℓ, we have

$$\vec{F}(a\vec{x}) = a\vec{F}(\vec{x}).$$

Then, since

$$\vec{F}(\vec{x}) = |\vec{x}|\vec{F}(\hat{x})$$

we can integrate $|\vec{x}|$ out of the problem. Let $\vec{x} = r\hat{s}$ be the radial and angular parts of the input. Then since $d\vec{x} = drr^{N-1}d\hat{s}$ we have

$$\langle\vec{F}(r\hat{s})\rangle = \frac{1}{(2\pi)^{N/2}}\int d\vec{x}e^{-\frac{1}{2}|\vec{x}|^2}r\vec{F}(\hat{s}) \tag{9}$$

$$= \frac{\Omega_{N-1}}{(2\pi)^{N/2}}\int_0^\infty drr^{N-1}e^{-\frac{1}{2}r^2}r\int_{\mathbb{S}^{N-1}}\frac{d\hat{s}}{\Omega_{N-1}}\vec{F}(\hat{s}) \tag{10}$$

$$=\sqrt{2}\frac{\Gamma(\frac{N+1}{2})}{\Gamma(\frac{N}{2})}\int_{\mathbb{S}^{N-1}}\frac{d\hat{s}}{\Omega_{N-1}}\vec{F}(\hat{s}). \tag{11}$$

Here $\Omega_{N-1} = \frac{2\pi^{N/2}}{\Gamma(N/2)}$ denotes the measure of the unit sphere in N-dimensions $\mathbb{S}^{N-1}$. So our problem is equivalently described as evaluating the expected value for the uniform distribution on the sphere.

Here we also find the simple prescription for sampling points uniformly on $\mathbb{S}^{N-1}$ - sample $\vec{x}$ from the N-dimensional standard normal Gaussian distribution, then the directions $\hat{s} = \frac{\vec{x}}{|\vec{x}|}$ are uniformly distributed on $\mathbb{S}^{N-1}$.

[2] some mathematicians excluded, perhaps

Technically, naive Monte Carlo over the uniform distribution on $\mathbb{S}^{N-1}$ instead of the standard normal on $\mathbb{R}^N$ is then a free variance reduction, but its a negligible one for large N. As $N \to \infty$ the mass of $e^{-\frac{1}{2}|\vec{x}|^2}$ concentrates in a spherical shell of width $\sim \frac{1}{\sqrt{N}}$, and so the variance reduction in our $N = 256$ case is roughly a negligible factor of $1 - \frac{1}{2N} \sim 1$.

3 Quadrature on the Sphere

In the spirit of keeping things simple if its possible to do so, a natural place to start is to ask if we could simply find a quadrature rule that can approximate the integral $\int d\hat{s}\vec{F}(\hat{s})$ more accurately than naive Monte Carlo.

Spoiler: no, you can't.

For general quadrature on the sphere $\mathbb{S}^2$ in three dimensions, the standard thing to do is construct a rule that integrates all spherical harmonics $Y_{\ell m}(\hat{s})$ for $\ell < L$. Exact rules are constructed based on a tensor product grid of two angular coordinates and applying Gaussian or Clenshaw-Curtis -type rules in the polar angle. Since its based on a tensor product grid, naively scaling it to larger N is infeasible.

An alternative construction is to take K unstructured points and try to distribute them as uniformly as possible over $\mathbb{S}^2$. This is done by starting with some initial quasi-uniform set of points $\hat{s}_k$, treating the points as particles that repel each other with some potential like $\frac{1}{\|\hat{s}_i - \hat{s}_j\|^\alpha}$, and then finding the configuration that minimizes the energy

$$E = \sum_{i \neq j} \frac{1}{\|\hat{s}_i - \hat{s}_j\|^\alpha}. \tag{12}$$

The optimized point set produces an equal-weights rule which is a bit like a spherical analogue of a Riemann sum. It will integrate all the harmonics up to L with good accuracy - provided K is proportional to the total number of number harmonics with $\ell < L$.

In three dimensions this is $\sim L^2$, but for general N the total number of harmonics goes as $\binom{L+N-1}{N} \sim e^N$. If you have a fixed budget, i.e. a maximum number of points K that can be used, about the best you can do with them is to put them into a minimum-energy configuration, which will be effectively uniform over $\mathbb{S}^{N-1}$. But practically, for large N this is unlikely to buy much improvement.

As N gets large, the number of nearly-orthogonal directions (say within some tolerance ϵ) grows exponentially.[3] So if you randomly draw K points from the uniform distribution on $\mathbb{S}^{N-1}$ and K is only $\sim N$ or even $\sim N^2$, then those K directions will already by nearly orthogonal. There is no significant sense of clustering for a minimum-energy configuration to improve upon, and so a quadrature rule defined by these quasi-optimal point sets may be little different from a randomly selected point set of the same size.

The question is how power is distributed over angular scales λ. Consider the expansion

[3] The quick illustration of this is to consider the angular distance between the directions $(1, 0, \ldots, 0)$ and $(1, 1, \ldots, 1)/\sqrt{N}$. In three dimensions these directions have significant overlap, but for general N the angular distance is $\arccos(1/\sqrt{N}) \to \pi/2$ as $N \to \infty$. Also c.f. the Johnson–Lindenstrauss lemma.

of $\vec{F}(\hat{s})$ in hyperspherical harmonics as

$$\vec{F}(\hat{s}) = \sum_{\lambda=0}^{\infty} \sum_{\mu} \vec{F}_{\lambda\mu} Y_{\lambda\mu}(\hat{s}) \tag{13}$$

(where μ is a multi-index) and define the angular power spectrum C_λ so that

$$\int \frac{d\hat{s}}{\Omega_{N-1}} |\vec{F}(\hat{s})|^2 = \sum_{\lambda=0}^{\infty} C_\lambda. \tag{14}$$

If power is relatively concentrated to the first few λ modes, then there is some significant improvement to be found in employing a quadrature rule with exact precision up to $\lambda = \lambda_{max}$. But for a fixed set of quadrature nodes, the fluctuations in the modes $\lambda > \lambda_{max}$ start to look random along each fixed direction. So if the bulk of power is not limited to modes $\lambda \leq \lambda_{max}$, then the rule will end up being little better than random sampling.

As an illustration, consider the components of the first layer function

$$F_{0\alpha}(\hat{s}) = \left(\vec{W}_{0\alpha} \cdot \hat{s}\right)_+, \tag{15}$$

$$= |\vec{W}_{0\alpha}| \left(\hat{W}_{0\alpha} \cdot \hat{s}\right)_+, \tag{16}$$

For each α the function $(\hat{W}_{0\alpha} \cdot \hat{s})_+$ is a zonal function on the sphere, and we can evaluate its harmonic spectrum in the coordinate system aligned with $\hat{W}_{0\alpha}$. Defining $\cos(\theta) = \hat{W}_{0\alpha} \cdot \hat{s}$ the kernel is

$$(\cos(\theta))_+ = \begin{cases} \cos(\theta), & 0 < \theta < \pi/2 \\ 0, & \pi/2 < \theta < \pi. \end{cases} \tag{17}$$

The spectral coefficients of $(\cos(\theta))_+$ are given by

$$f_\lambda = \frac{\Omega_{N-2}}{\Omega_{N-1}} \int_0^1 dx (1-x^2)^{(N-2)/2} x \mathcal{C}_\lambda^{(N-2)/2}(x) \tag{18}$$

where $\mathcal{C}_\lambda^{(N-2)/2}(x)$ is the normalized Gegenbauer function, and then the angular power spectrum for this kernel is $C_\lambda = |f_\lambda|^2$. In Figure 1 we see that this spectrum does have some significant build up of power in the low harmonics $\lambda \leq 3$, and then a slowly decreasing tail for $\lambda > 3$ due to the kink at $\theta = \pi/2$.

A simple rule that exactly integrates all harmonics $\lambda \leq 3$ (a.k.a. a spherical 3-design) is given simply by taking all the axis unit vector $\hat{e}_n$ and their antithetic complements $-\hat{e}_n$ as nodes, and equal weights $\frac{1}{2N}$, so that

$$\int \frac{d\hat{s}}{\Omega_{N-1}} f(\hat{s}) \approx \frac{1}{2N} \sum_{n=1}^{N} \left(f(\hat{e}_n) + f(-\hat{e}_n)\right). \tag{19}$$

Using this rule with $K = 2N = 512$ points to estimate the mean of the first layer function

$$\langle \vec{F}_0(\hat{s}) \rangle \approx \frac{1}{2N} \sum_{n=1}^{N} \left((W_0 \hat{e}_n)_+ + (-W_0 \hat{e}_n)_+\right)$$

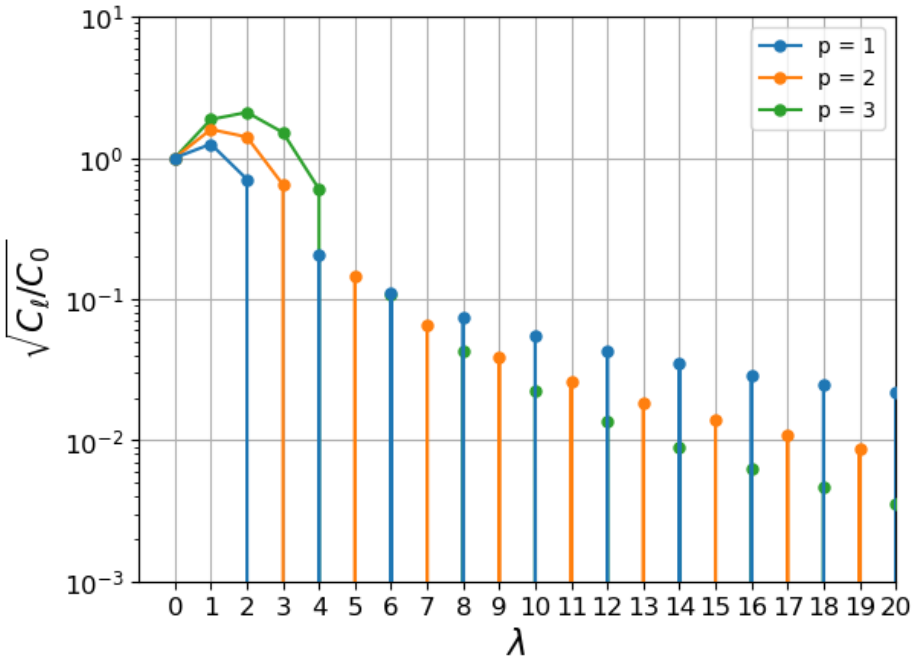

Figure 1: The relative amplitude between angular scales λ of $(\hat{W} \cdot \hat{s})^p_+$ for $N = 256$.

one finds that the estimate is notably better on average than naive Monte Carlo with $K = 512$ randomly generated points. The improvement is a factor of roughly ~ 8 - referring back to Figure 1, we can see that this is consistent with the dynamic range between the $\lambda = 0, 1, 2$ modes and the first non-zero tail mode $\lambda = 4$.

In contrast, using this rule to estimate the mean of the full composite function $\langle \vec{F}(\hat{s}) \rangle$ finds no such improvement over using 512 randomly drawn points. The composite function $\vec{F}(\hat{s})$ does not have the same segmentation in λ as an individual layer. Composition will tend to convolve power between modes so by layer 32 the low-λ modes are no longer relatively large. If there even is still any significant knee to the spectrum, it is at some $\lambda \sim 32$ that would require an exponentially larger number of nodes K to reach.

I tested this (crudely) by using the `symetrichyperspheresampling` code[4] to generate quasi-uniform point sets with $K = 4096$ and $K = 4 \cdot 4096$, with a parity symmetry constraint. This code produces the set of vectors by minimizing an energy function of the pair-wise angular distances, and can do so under symmetry constraints. Along with the 3-design rule with $K = 512$, I use these three node sets to estimate $\langle \vec{F}(\hat{s}) \rangle$ (for a fixed draw of the weights W_ℓ) under a set of random rotations applied to the node vectors. The scaling of the mean squared error relative to the true mean (computed with 10^8 naive MC samples) between the different node sets is precisely $\frac{1}{K}$. Since the low-λ modes aren't a large fraction of the power, the fact that they are integrated more accurately by these specialized point sets doesn't matter much. For *most* modes, they are effectively no more accurate than naive Monte Carlo. Perhaps there is technically a small factor of absolute variance reduction, but nothing large enough to be worth trying to measure.

[4] https://gitlab.com/mammasmias/symetrichyperspheresampling

Perhaps there is some other basis besides harmonics on which a different kind of quadrature rule could be based? I think the answer must be "no". The variation in the family of functions $\vec{F}$ is defined by the weight matrices W_ℓ, and because they are generated from an isotropic Gaussian distribution, there is spherical symmetry which is matched to the spherically symmetric angular power spectrum.

Spherical quadrature rules are not going to put much of a dent in the problem on their own, but they may yet find a place as one part of an algorithm that does. While composite functions may not be simple enough for a low-order quadrature rule to be useful, individual layer functions can be.

3.1 Radial/Angular Quadrature for Gaussian Weight

Consider the Gaussian weighted integral

$$I[f] = \frac{1}{(2\pi)^{N/2}\sqrt{\det(C)}} \int d\vec{x} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^\intercal C^{-1}(\vec{x}-\vec{\mu})} f(\vec{x}). \tag{20}$$

With the change of variables $\vec{r} = Q^{-1}(\vec{x}-\vec{\mu})$ (with $Q = U\Lambda^{1/2}$ given $C = U\Lambda U^\intercal$) it becomes

$$I[f] = \frac{1}{(2\pi)^{N/2}} \int d\vec{r} e^{-\frac{1}{2}|\vec{r}|^2} f(Q\vec{x}+\vec{\mu}) \tag{21}$$

$$= \frac{\Omega_{N-1}}{(2\pi)^{N/2}} \int_0^\infty dr r^{N-1} e^{-\frac{1}{2}r^2} \int \frac{d\hat{s}}{\Omega_{N-1}} f(rQ\hat{s}+\vec{\mu}). \tag{22}$$

Then substituting $t = \frac{1}{2}r^2$ obtains

$$I[f] = \frac{1}{\Gamma(N/2)} \int_0^\infty dt t^{\frac{N-2}{2}} e^{-t} \int \frac{d\hat{s}}{\Omega_{N-1}} f(\sqrt{2t}Q\hat{s}+\vec{\mu}). \tag{23}$$

The radial part of this integral can then be approximated with a with generalized Gauss-Laguerre quadrature rule for the power-law weight exponent $\frac{N-2}{2}$, which can be found in standard scientific libraries. This can be combined with a spherical quadrature rule over $\hat{s}$.

This could find use in in approximating moments of a nearly-Gaussian model of a layer. Schematically, at a layer ℓ the density approximation could have the form

$$P_\ell(\vec{x}_\ell) \propto e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^\intercal C^{-1}(\vec{x}-\vec{\mu})}(1+R_\ell(\vec{x}_\ell)) \tag{24}$$

for some smooth perturbation R_ℓ. The expectation value of a function $F(\vec{x}_\ell)$ is then $I[(1+R_\ell)F]$ which could be approximated by this method, or something similar.

3.2 Zonal Kernel Quadrature

In a similar vein, consider an integral of the form

$$I[f] = \int \frac{d\hat{s}}{\Omega_{N-1}} g(\hat{a}\cdot\hat{s}) f(\hat{s}), \tag{25}$$

where $\hat{a}$ is a fixed unit vector. We can take $\hat{\omega}$ to be in $\mathbb{S}^{N-2}$ orthogonal to $\hat{a}$ such that $\hat{s} = \cos(\theta)\hat{a} + \sin(\theta)\hat{\omega}$. Then $d\hat{s} = d\hat{\omega}d\theta \sin(\theta)^{N-2}$, and the integral can be written as

$$I[f] = \frac{\Omega_{N-2}}{\Omega_{N-1}} \int_0^\pi d\theta \sin(\theta)^{N-2} g(\cos(\theta)) \int \frac{d\hat{\omega}}{\Omega_{N-2}} f(\cos(\theta)\hat{a} + \sin(\theta)\hat{\omega}). \tag{26}$$

With $x = \cos(\theta)$ this becomes

$$I[f] = \frac{\Omega_{N-2}}{\Omega_{N-1}} \int_{-1}^1 dx (1-x^2)^{\frac{N-2}{2}} g(x) \int \frac{d\hat{\omega}}{\Omega_{N-2}} f(x\hat{a} + \sqrt{1-x^2}\hat{\omega}). \tag{27}$$

The integral in x is approximated by a Gauss-Jacobi rule (again available in common libraries), and this is combined with a spherical rule for the $\hat{\omega}$ integral over $\mathbb{S}^{N-2}$. The point is a spherical rule to integrate the product $g(\hat{a} \cdot \hat{s}) f(\hat{s})$ over $\hat{s}$ may be infeasible, but a rule that only needs to integrate f alone may become feasible i.e. a spherical 3- or 5-design.

4 Splitting Radial and Angular Dynamics

We can do a bit more with the homogeneity of the layer functions. For each layer ℓ, define $e^{a_\ell} = |\vec{x}_\ell|$ and the unit vector $\hat{s}_\ell$ so that

$$\vec{x}_\ell = e^{a_\ell} \hat{s}_\ell.$$

Then since each $\vec{F}_\ell(\vec{x}_\ell)$ is homogeneous we have

$$e^{a_{\ell+1}} \hat{s}_{\ell+1} = e^{a_\ell} \vec{F}_\ell(\hat{s}_\ell). \tag{28}$$

Thus we find we can separate the angular dynamics as

$$\hat{s}_{\ell+1} = \frac{\vec{F}_\ell(\hat{s}_\ell)}{|\vec{F}_\ell(\hat{s}_\ell)|} \tag{29}$$

$$= \frac{(W_\ell \hat{s}_\ell)_+}{\sqrt{\sum_\beta (\vec{W}_{\ell\beta} \cdot \hat{s}_\ell)_+^2}}, \tag{30}$$

independent of the amplitude a_ℓ. On the other hand, the amplitude evolves as

$$a_{\ell+1} = a_\ell + \log(|\vec{F}_\ell(\hat{s}_\ell)|). \tag{31}$$

Our actual function is defined by a sequence of functions $\vec{F}_\ell$ for $\ell = 0$ to $\ell = L-1$. Suppose we make a periodic extension by repeating this sequence infinitely for $\ell \geq L$, so that $\vec{F}_\ell = \vec{F}_{\ell+L}$ is periodic as a function of ℓ with period L. It turns out that under this extension the sequence $\hat{s}_\ell$ converges to a periodic attractor $\hat{u}_\ell$, i.e. a periodic sequence $\hat{u}_\ell = \hat{u}_{\ell+L}$. The attractor sequence $\hat{u}_\ell$ is computed simply by evaluating the sequence Equation 29 for some several periods beyond $\ell = L-1$, until convergence of the final chunk of length L. For *any* initial condition $\hat{s}_0 \in \mathbb{S}^{N-1}$, the periodically extended sequence converges to $\hat{u}_\ell$, and the

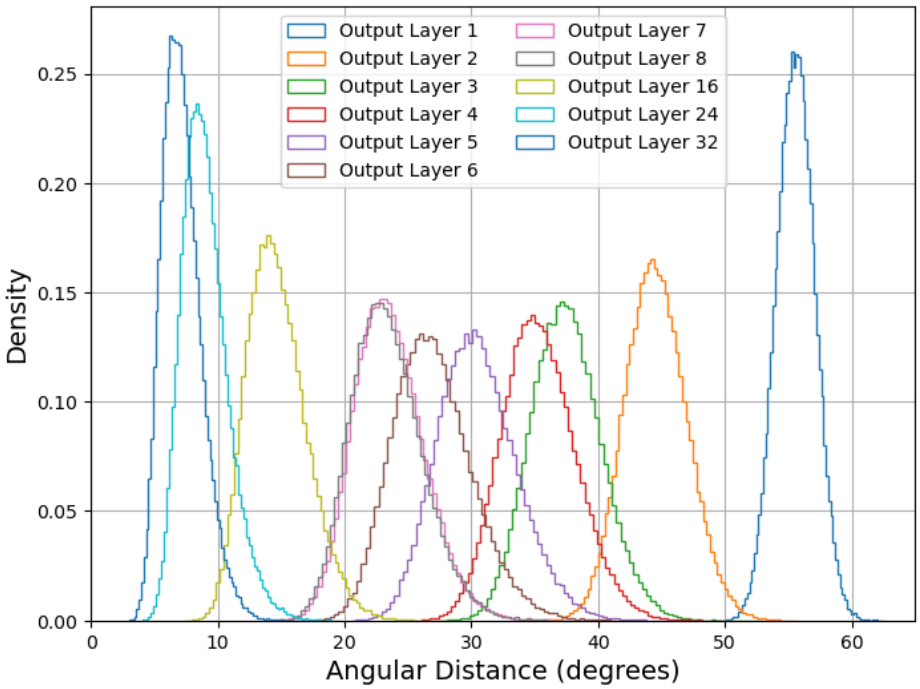

Figure 2: Distribution of angular distances of trajectories $\hat{s}_\ell$ from the mean direction $\propto \langle \hat{s}_\ell \rangle$. Note this does not say that the distribution of directions is isotropic, it just says that it is contracting on average as $\ell \to L$.

direction $\hat{u}_0$ in the periodic attractor sequence is highly correlated with the direction $\hat{\mu}_L$ of the true mean $\vec{\mu}_L = \langle \vec{x}_L \rangle$ at $\ell = L$.

The salient point is that Equation 29 is a contracting map. As $\ell \to L$, the distribution of directions $\hat{s}_\ell$ becomes increasingly concentrated, as shown in Figure 2. This seems like the kind of generic structural property that could be exploitable, though so far I'm not sure how.

The initial condition is uniform on $\mathbb{S}^{N-1}$, but for $\ell > 0$ the dynamics of $\hat{s}_\ell$ are actually constrained to the positive orthant of the sphere $\mathbb{S}^{N-1}_+$ i.e. the part of the sphere with all coordinates ≥ 0. For large N, this space is exponentially smaller than the full sphere by a factor 2^N. It also challenges intuition you might try to carry from the 2-sphere. In $N = 3$ dimensions the smallest angular distance from the center of the positive orthant $(1,1,1)/\sqrt{3}$ to each of the three faces (e.g. $(1,1,0)/\sqrt{2}$) is $\arcsin(1/\sqrt{3}) \approx 35°$, but in $N = 256$ dimensions it is $\arcsin(1/\sqrt{256}) \approx 3.5°$. So a uniform distribution on $\mathbb{S}^{N-1}_+$ is a rather anisotropic distribution on $\mathbb{S}^{N-1}$, and something like a von Mises-Fisher distribution is not going to be a good approximation of concentration on the spherical orthant. The natural analogue is the Spherical Dirichlet distribution - unfortunately this distribution is computationally more difficult to work with.

Maybe there is something useful in the structure here - but maybe its just making things more complicated and confusing than they need to be.

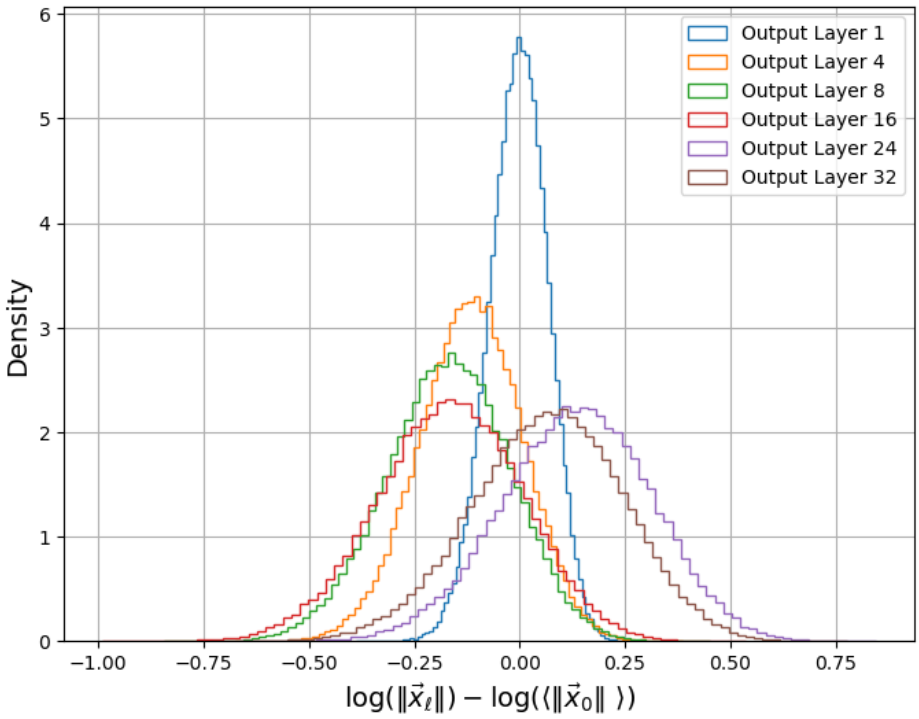

Figure 3: Distribution of the log-norms of layer outputs $\vec{x}_\ell$ with the expected magnitudeEquation 11 of the initial Gaussian distribution subtracted off.

5 Gaussian Closure

The outputs of each layer $\vec{x}_\ell = \vec{F}_{\ell-1}(\vec{x}_{\ell-1})$ are not at all Gaussian. However, the matrices W_ℓ are generated by independent Gaussian noise, and for each α the variables $z_{\ell\alpha} = \vec{W}_{\ell\alpha} \cdot \vec{x}_\ell$ will be approximately Gaussian for large N by the central limit theorem. If $\vec{\mu}_\ell$, C_ℓ are the mean and covariance of $\vec{x}_\ell$, the variable $\vec{z}_\ell$ has mean $\vec{\eta}_\ell = W_\ell \vec{\mu}_\ell$ and covariance $S_\ell = W_\ell C_\ell W_\ell^\intercal$, and by approximating $\vec{z}_\ell$ as Gaussian we obtain a closure relation that approximately propagates the mean and covariance through the nonlinear layer functions.

For the first moment of the next layer we have

$$\langle (z_\alpha)_+ \rangle = \eta_\alpha \Phi(r_\alpha) + \sigma_\alpha \phi(r_\alpha), \tag{32}$$

where $\sigma_\alpha = \sqrt{S_{\alpha\alpha}}$ and $r_\alpha = \frac{\eta_\alpha}{\sigma_\alpha}$. The covariance of the next layer is then, in terms of $\rho_{\alpha\beta} = \frac{S_{\alpha\beta}}{\sqrt{S_{\alpha\alpha} S_{\beta\beta}}}$,

$$\langle (z_\alpha)_+ (z_\beta)_+ \rangle - \langle (z_\alpha)_+ \rangle \langle (z_\beta)_+ \rangle = \Phi(r_\alpha) S_{\alpha\beta} \Phi(r_\beta) + 2 S_{\alpha\beta} \rho_{\alpha\beta} \int_0^1 dv v^3 \phi_2(r_\alpha, r_\beta, \rho_{\alpha\beta}(1 - v^2)). \tag{33}$$

Here ϕ and Φ are the standard normal density and cumulative distribution functions, while ϕ_2 is the bivariate normal density. The integral term can be evaluated accurately and cheaply by

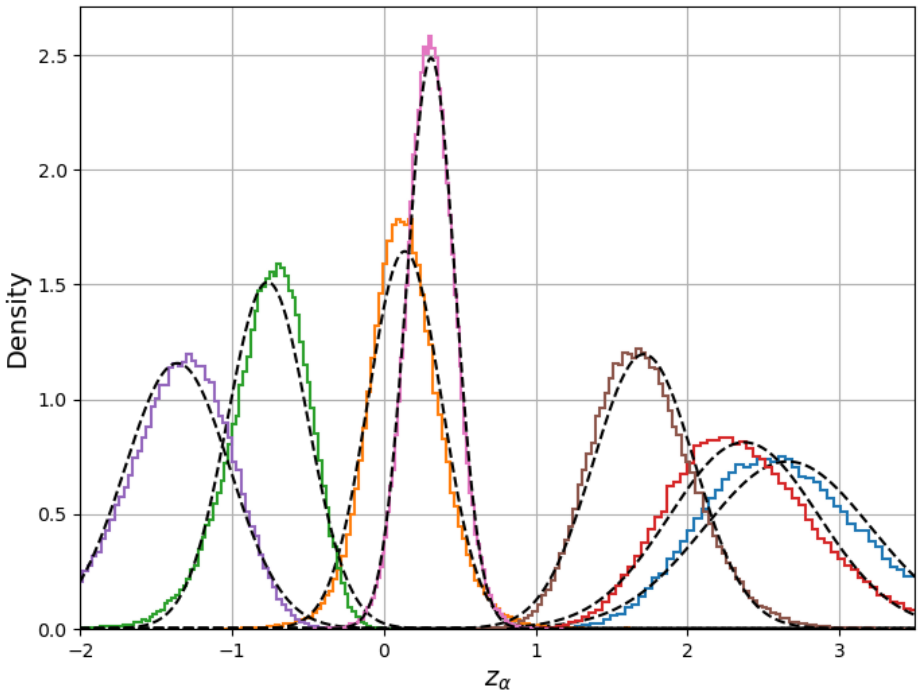

Figure 4: Comparison of the approximate Gaussian (black dashed) to simulated true (histograms) marginal distributions for a few components of the vector $\vec{z}_\ell$ at the final layer $\ell = 31$.

a Gauss-Jacobi quadrature rule with just a couple of nodes (see Equation 54). The complete derivation is described in Appendix A along with definitions and details.

This Gaussian approximation is remarkably good compared to what would initially seem to be a very complex object produced by 32 layers of functional composition. Of course, since $\vec{z}_\ell = W_\ell(\vec{z}_{\ell-1})_+$ is non-linear $\vec{z}_\ell$ cannot be truly Gaussian and each step projects out some residual information. The error induced by this projection builds up with each layer so that each successive estimate of the mean drifts further from the truth. This is illustrated

Fundamentally, the recursive form of this closure

$$(\vec{\eta}_{\ell+1}, S_{\ell+1}) = f(\vec{\eta}_\ell, S_\ell) \tag{34}$$

is precisely the form of a solution that matches the core structure of the functions $\vec{F}(\vec{x})$ - it is compositional in nature. While the appropriate extension of this approximation to improve the accuracy may not be immediately obvious, it is clear that finding such an extension is the way forward.

6 Discussion

The Gaussian closure is biased, but relative to its cost and simplicity it is a remarkably good approximation. Such a thing should be built upon, not discarded.

Since its close, perhaps there is a relatively simple way to refine it with additional MC sampling? Essentially, use the Gaussian approximation as a control variate. If there is something sensible in this area, I imagine it would be something akin a Kalman filter-esque prediction/correction recursion.

In any case, I don't think any control variate that tries to cut directly through 32 layers of composition can work.

The core structural feature of $\vec{F}(\vec{x})$ is that it is compositional. An efficient integrator for $\vec{F}(\vec{x})$ should then follow that structure, which means the integrator will also be compositional. Essentially this means the solution must be some kind of propagating closure approximation.

That doesn't necessarily mean some limited "sampling" couldn't be a part of it. As discussed in the introduction, if we view "sampling" simply as a quadrature rule, then such a rule may be used as part of the propagation in a closure relation. Another way of thinking about it is that any direct use of sampled trajectories, i.e. pushing samples directly through the dynamics $\vec{F}_\ell$, can only provide information relatively locally to a given layer.

That also doesn't mean that it needs to look like the (Wu-Leconte-Winer-Robinson et al 2026) formulation. I suspect the Edgeworth/Gram-Charlier -type model using polynomial Hermite expansions is not the basis for this problem, and I think developing an alterative is the most promising direction.

A 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

$$\langle (z_\alpha)_+ \rangle = \frac{1}{(2\pi)^{N/2}\sqrt{\det C}} \int d\vec{z} e^{-\frac{1}{2}(\vec{z}-\vec{\eta})^\intercal C^{-1}(\vec{z}-\vec{\eta})} (z_\alpha)_+ \tag{35}$$

$$= \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \int_0^\infty dz \exp\left(-\frac{(z-\eta_\alpha)^2}{2\sigma_\alpha^2}\right) z \tag{36}$$

$$= \eta_\alpha \Phi(r_\alpha) + \sigma_\alpha \phi(r_\alpha), \tag{37}$$

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

$$\phi(r) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}r^2}, \tag{38}$$

$$\Phi(r) = \int_{-\infty}^{r} dx \phi(x) = \frac{1}{2}\left(1 + \mathrm{erf}(r/\sqrt{2})\right). \tag{39}$$

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

$$\langle (z_\alpha)_+ (z_\beta)_+ \rangle = \frac{1}{2\pi\sigma_\alpha\sigma_\beta\sqrt{1-\rho^2}} \int_0^\infty dz_\alpha \int_0^\infty dz_\beta z_\alpha z_\beta \tag{40}$$

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

$$= \int_{-r_\alpha}^\infty du \int_{-r_\beta}^\infty dv f_\alpha(u) f_\beta(v) \phi_2(u,v,\rho) \tag{41}$$

where we define

$$f_\alpha(u) = \eta_\alpha + \sigma_\alpha u \tag{42}$$

$$\phi_2(u, v, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{u^2 - 2\rho uv + v^2}{2(1-\rho^2)}\right), \tag{43}$$

and denote the covariance by

$$I_{\alpha\beta}(\rho) = \langle (z_\alpha)_+ (z_\beta)_+ \rangle - \langle (z_\alpha)_+ \rangle \langle (z_\beta)_+ \rangle . \tag{44}$$

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

$$\frac{\partial}{\partial \rho} \langle g(u, v) \rangle = \left\langle \frac{\partial g(u, v)}{\partial u \partial v} \right\rangle. \tag{45}$$

Hence, we have

$$\frac{\partial I_{\alpha\beta}(\rho)}{\partial \rho} = \left\langle \frac{\partial f_\alpha(u)}{\partial u} \frac{\partial f_\beta(v)}{\partial v} \right\rangle \tag{46}$$

$$= \sigma_\alpha \sigma_\beta \Phi_2(r_\alpha, r_\beta, \rho), \tag{47}$$

where Φ_2 is the bivariate cumulative distribution function

$$\Phi_2(r_\alpha, r_\beta, \rho) = \int_{-\infty}^{r_\alpha} du \int_{-\infty}^{r_\beta} dv \phi_2(u, v, \rho). \tag{48}$$

Now from the integral form of Plackett's identity

$$\partial_\rho \Phi_2(u, v, \rho) = \phi_2(u, v, \rho) \longleftrightarrow \Phi_2(u, v, \rho) = \Phi(u)\Phi(v) + \int_0^\rho dt \phi_2(u, v, t) \tag{49}$$

we can write

$$\frac{\partial I_{\alpha\beta}(\rho)}{\partial \rho} = \sigma_\alpha \sigma_\beta \Phi(r_\alpha)\Phi(r_\beta) + \sigma_\alpha \sigma_\beta \int_0^\rho dt \phi_2(r_\alpha, r_\beta, t). \tag{50}$$

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

$$I_{\alpha\beta}(\rho) = \rho \sigma_\alpha \sigma_\beta \Phi(r_\alpha)\Phi(r_\beta) + \sigma_\alpha \sigma_\beta \int_0^\rho dt \int_0^t ds \phi_2(r_\alpha, r_\beta, s), \tag{51}$$

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

$$I_{\alpha\beta}(\rho) = \Phi(r_\alpha) S_{\alpha\beta} \Phi(r_\beta) + \sigma_\alpha \sigma_\beta \int_0^\rho dt (\rho - t) \phi_2(r_\alpha, r_\beta, t). \tag{52}$$

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

$$I_{\alpha\beta}(\rho) = \Phi(r_\alpha) S_{\alpha\beta} \Phi(r_\beta) + \sigma_\alpha \sigma_\beta \rho^2 \int_0^1 dv 2v^3 \phi_2(r_\alpha, r_\beta, \rho(1 - v^2)). \tag{53}$$

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

$$\rho^2 \int_0^1 dv 2v^3 \phi_2(r_\alpha, r_\beta, \rho(1-v^2)) \approx \frac{\rho^2}{2\pi} \sum_j w_j \frac{1}{\sqrt{1-t_j^2}} \exp\left(-\frac{r_\alpha^2 + r_\beta^2 - 2r_\alpha r_\beta t_j}{2(1-t_j^2)}\right), \tag{54}$$

$$t_j = \rho(1-v_j^2), \tag{55}$$

where v_j, w_j are the appropriate nodes and weights.

The first term of $I_{\alpha\beta}$ corresponds to the covariance produced by statistical linearization of $(\vec{z}_\ell)_+$. Under the assumption that $\vec{z}_\ell$ is Gaussian, the statistically optimal linear approximation to $(\vec{z}_\ell)_+$ is given by

$$J_\ell \vec{z}_\ell + \vec{b}_\ell$$

where by Stein's lemma

$$J_\ell = \left\langle \left((\vec{z}_\ell)_+ - \langle (\vec{z}_\ell)_+ \rangle\right) (\vec{z}_\ell - \vec{\eta}_\ell)^\dagger \right\rangle \tag{56}$$

$$= \left\langle \frac{\partial (\vec{z}_\ell)_+}{\partial \vec{z}_\ell^\dagger} \right\rangle \tag{57}$$

$$= \mathrm{diag}(\Phi(r_\alpha)), \tag{58}$$

and such a linear map transforms the covariance as $S_\ell \to J_\ell S_\ell J_\ell^\intercal$.