A MEAN FIELD THEORY OF BATCH NORMALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a mean field theory for batch normalization in fully-connected feed-forward neural networks. In so doing, we provide a precise characterization of signal propagation and gradient backpropagation in wide batch-normalized networks at initialization. We find that gradient signals grow exponentially in depth and that these exploding gradients cannot be eliminated by tuning the initial weight variances or by adjusting the nonlinear activation function. Indeed, batch normalization itself is the cause of gradient explosion. As a result, vanilla batch-normalized networks without skip connections are not trainable at large depths for common initialization schemes, a prediction that we verify with a variety of empirical simulations. While gradient explosion cannot be eliminated, it can be reduced by tuning the network close to the linear regime, which improves the trainability of deep batch-normalized networks without residual connections. Finally, we investigate the learning dynamics of batch-normalized networks and observe that after a single step of optimization the networks achieve a relatively stable equilibrium in which gradients have dramatically smaller dynamical range.

1 INTRODUCTION

Deep neural networks have been enormously successful across a broad range of disciplines. These successes are often driven by architectural innovations. For example, the combination of convolutions (LeCun et al., 1990), residual connections (He et al., 2015), and batch normalization Ioffe & Szegedy (2015) has allowed for the training of very deep networks and these components have become essential parts of models in vision (Zoph et al.), language Chen & Wu (2017), and reinforcement learning Silver et al. (2017). However, a fundamental problem that has accompanied this rapid progress is a lack of theoretical clarity. An important consequence of this gap between theory and experiment is that two important issues become conflated. In particular, it is generally unclear whether novel neural network components improve generalization or whether they merely increase the number of hyperparameter configurations where good generalization can be achieved. Resolving this confusion has the promise of allowing researchers to more effectively and deliberately design neural networks.

Recently, progress has been made (Poole et al., 2016; Schoenholz et al., 2016; Daniely et al., 2016; Pennington et al., 2017) in this direction by considering neural networks at initialization, before any training has occurred. In this case, the parameters of the network are random variables which induces a distribution of the activations of the network as well as the gradients. Studying these distributions is equivalent to understanding the prior over functions that these random neural networks compute. Picking hyperparameters that correspond to well-conditioned priors ensures that the neural network will be trainable and this fact has been extensively verified experimentally. However, to fulfill its promise of making neural network design less of a black box, these techniques must be applied to neural network architectures that are used in practice. Over the past year, this gap has closed significantly and theory for networks with skip connections (Yang & Schoenholz, 2017; 2018), convolutional networks (Xiao et al., 2018), and gated recurrent networks (Chen et al., 2018) have been developed.

Before state-of-the-art models can be studied in this framework, a slowly-decreasing number of architectural innovations must be studied. One particularly important component that has thus-far remained illusive is batch normalization. In this paper, we develop a theory of random, fully-connected networks with batch normalization. A significant complication in the case of batch normalization (compared to e.g. layer normalization or weight normalization) is that the statistics of the network

depend non-locally on the entire batch. Thus, our first main result is to recast the theory for random fully-connected networks so that it can be applied to batches of data. We then extend the theory to include batch normalization explicitly and validate this theory against Monte-Carlo simulations. We show that as in previous cases we can leverage our theory to predict valid hyperparameter configurations.

In the process of our investigation, we identify a number of previously unknown properties of batch normalization that make training unstable. In particular, we show that for any choice of nonlinearity, gradients of fully-connected networks with batch normalization explode exponentially in the depth of the network. This imposes strong limits on the maximum trainable depth of batch normalized networks that can be ameliorated by pushing activation functions to be more linear at initialization. It might seem that such gradient explosion ought to lead to learning dynamics that are unfavorable. However, we show that networks with batch normalization causes the scale of the gradients to naturally equilibrate after a single step of gradient descent (provided the gradients are not so large as to cause numerical instabilities).

Finally, we note that there is a related vein of research that has emerged that leverages the prior over functions induced by random networks to perform exact Bayesian inference (Lee et al., 2017). One of the natural consequences of this work is that the prior for networks with batch normalization can be computed exactly in the wide network limit. As such, it is now possible to perform Bayesian inference in the case of wide neural networks with batch normalization.

2 Related Work

Batch normalization has rapidly become an essential part of the deep learning toolkit. Since then, a number of similar modifications have been proposed including layer normalization Ba et al. (2016) and weight normalization Salimans & Kingma (2016). Comparisons of performance between these different schemes have been challenging and inconclusive Gitman & Ginsburg (2017). Since the original introduction of batchnorm in Ioffe & Szegedy (2015), which proposed that batchnorm prevents "internal covariate shift" as the explanation for their effectiveness. Since then, several papers have approached batchnorm from a theoretical angle, especially following Ali Rahimi's famous call to action at NIPS 2018. Balduzzi et al. (2017) found that batchnorm in resnets allow deep gradient signal propagation in contrast to the case without batchnorm. Santurkar et al. (2018) found that batchnorm does not help covariate shift but helps by smoothing loss landscape. Bjorck et al. (2018) reached the opposite conclusion as our paper for residual networks with batchnorm, that batchnorm works in this setting because it induces beneficial gradient dynamics and thus allows a much bigger learning rate. Luo et al. (2018) explores similar ideas that batchnorm allows large learning rates and likewise uses random matrix theory to support their claims. Kohler et al. (2018) identified situations in which batchnorm can provably induce acceleration in training. Of the above that mathematically analyze batchnorm, all but Santurkar et al. (2018) make simplifying assumptions on the form of batchnorm and typically do not have gradients flowing through the batch variance. Even Santurkar et al. (2018) only analyzes a deep linear network which gets added a batchnorm layer at a single moment in training. Our analysis here works for arbitrarily deep batchnorm networks with any activation function used in practice[1]. It is an initialization time analysis, but we use such insight to predict training and test time behavior.

3 Theory

We begin with a brief recapitulation of mean field theory in the fully-connected setting. In addition to recounting earlier results, we rephrase the formalism developed previously to compute statistics of neural networks over a batch of data. Later, we will extend the theory to include batch normalization. A fully-connected network of depth L whose layers have width N_l is defined by an activation function[2] ϕ along with weights, $W^l \in \mathbb{R}^{N_{l-1} \times N_l}$, and biases, $b^l \in \mathbb{R}^{N_l}$. Given a batch of B inputs[3]

[1] upper bounded by an exponential function, for example.

[2] The activation function may be layer dependent, but for ease of exposition we assume that it is not.

[3] Throughout the text, we assume that all elements of the batch are unique.

$\{x_i : x_i \in \mathbb{R}^{N_0}\}_{i=1,\cdots,B}$, the pre-activations of the network are defined by the recurrence relation,

$$\boldsymbol{z}_i^1 = W^1 \boldsymbol{x}_i + b^1 \qquad \text{and} \qquad \boldsymbol{z}_i^l = W^l \phi(\boldsymbol{z}_i^{l-1}) + b^l \quad \forall\, l > 1. \tag{1}$$

At initialization, we choose the weights and biases to be i.i.d. as $W^l_{\alpha\beta} \sim \mathcal{N}(0, \sigma_w^2/N_{l-1})$ and $b^l_\alpha \sim \mathcal{N}(0, \sigma_b^2)$. We will be concerned with understanding the statistics of the pre-activations and the gradients induced by the randomness in the weights and biases. For ease of exposition we will typically take the network to have constant width $N_l = N$.

In the mean field approximation, we iteratively replace the pre-activations in eq. (2) by Gaussian random variables with matching first and second moments. In the infinite width limit this approximation becomes exact Lee et al. (2017). Since the weights are i.i.d. with zero mean it follows that the mean of each pre-activation is zero and the covariance between distinct neurons are zero. The pre-activation statistics are therefore given by $(z^l_{\alpha_1 1}, \cdots, z^l_{\alpha_B B}) \xrightarrow{N_{l-1}\to\infty} \mathcal{N}(0, \Sigma^l \delta_{\alpha_1 \cdots \alpha_B})$ where Σ^l are $B \times B$ covariance matrices. The covariance matrices are defined by the recurrence relation,

$$\Sigma^l = \sigma_w^2 V_\phi(\Sigma^{l-1}) + \sigma_b^2 \mathbf{1}\mathbf{1}^T \tag{2}$$

where $V_\phi(\Sigma) = \mathbb{E}[\phi(h)\phi(h)^T : h \sim \mathcal{N}(0, \Sigma)]$ computes the matrix of uncentered second moments of $\phi(z)$ for $z \sim \mathcal{N}(0, \Sigma)$. At first eq. (2) may seem challenging since the expectation involves a Gaussian integral in $\mathbb{R}^B$. However, each term in the expectation of V_ϕ involves at most a pair of pre-activations and so the expectation may be reduced to the evaluation of $\mathcal{O}(B^2)$ two-dimensional integrals. For many choices of activation function these integrals may be done analytically and so eq. (2) defines a computationally efficient method for computing the statistics of neural networks after random initialization. This theme of dimensionality reduction will play a prominent role in the forthcoming discussion on batch normalization.

Eq. (2) defines a dynamical system over the space of covariance matrices. Studying the statistics of random feed-forward networks therefore amounts to investigating this dynamical system and is an enormous simplification compared with studying the pre-activations of the network directly. As is common in the dynamical systems literature, a significant amount of insight can be gained by investigating the behavior of eq. (2) in the vicinity of its fixed points. For most common activation functions, eq. (2) has a fixed point at Σ^*. Moreover, when the inputs are non-degenerate, this fixed point generally has a simple structure with $\Sigma^* = q^*[(1 - c^*)I + c^* \mathbf{1}\mathbf{1}^T]$ owing to permutation symmetry among elements of the batch. We refer to fixed points with such symmetry as Batch Symmetry Breaking 1 (BSB1) fixed points. As we will discuss later, in the context of batch normalization other fixed points with fewer symmetries may become preferred. In the fully-connected setting fixed points may efficiently be computed by solving the fixed point equation induced by eq. (2) in the special case $B = 2$. The structure of this fixed point implies that in asymptotically deep feed-forward neural networks all inputs yield pre-activations of identical norm with identical angle between them. Neural networks that are deep enough so that their pre-activation statistics lie in this regime have been shown to be untrainable.

Notation As we often talk about matrices and also linear operator over matrices, we write $\mathcal{T}\{\Sigma\}$ for an operator $\mathcal{T}$ applied to a matrix Σ, and matrix multiplication is still written as juxtaposition. Composition of matrix operators are denoted with $\mathcal{T}_1 \circ \mathcal{T}_2$.

To understand the behavior of eq. (2) near its fixed point we can consider the Taylor series in the deviation from the fixed point, $\delta\Sigma^l = \Sigma^l - \Sigma^*$. To lowest order we generically find,

$$\delta\Sigma^l = \frac{dV_\phi}{d\Sigma}\bigg|_{\Sigma=\Sigma^*} \{\delta\Sigma^{l-1}\} \tag{3}$$

where $J = \frac{dV_\phi}{d\Sigma}\big|_{\Sigma=\Sigma^*}$ is the $B^2 \times B^2$ Jacobian of V_ϕ. In most prior work where ϕ was a pointwise non-linearity one could consider the special case of $B = 2$ which naturally gave rise to linearized dynamics in $q^l = \mathbb{E}[(z_i^l)^2]$ and $c^l = \mathbb{E}[z_i^l z_j^l]/q^l$. However, in the case of batch normalization we will see that one must consider the evolution of eq. (3) as a whole. This is qualitatively reminiscent of the case of convolutional networks studied in Xiao et al. (2018) where the evolution of the entire pixel $\times$ pixel covariance matrix had to be evaluated. The dynamics induced by eq. (3) will be controlled by the eigenvalues of J. Suppose J has eigenvalues λ_i - ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{B^2}$ - with associated eigen"vectors" e_i (note that the e_i will themselves be $B \times B$ matrices). It follows

that if $\delta\Sigma^0 = \sum_i c_i e_i$ for some choice of constants c_i then $\delta\Sigma^l = \sum_i c_i \lambda_i^l e_i$. Thus, if $\lambda_i < 1$ for all i, $\delta\Sigma^l$ will approach zero exponentially and the fixed-point will be stable. The number of layers over which Σ will approach Σ^* will be given by $-1/\log(\lambda_1)$. By contrast if $\lambda_i > 1$ for any i then the fixed point will be unstable. In this case, there is typically a different, stable, fixed point that must be identified. It follows that if the eigenvalues of J can be computed then the dynamics will follow immediately.

At face value, J is a complicated object since it simultaneously has large dimension and possesses an intricate block structure. However, the permutation symmetry of Σ^* induces strong symmetries in J that significantly simplify the analysis [B.4]. In particular J_{ijkl} is a four-index object, however $J_{ijkl} = J_{\pi(i)\pi(j)\pi(k)\pi(l)}$ for all permutations π on B and $J_{ijkl} = J_{jilk}$. We call linear operators possessing such symmetries ultrasymmetric and show that all ultrasymmetric matrices admit an eigen-decomposition that contains three distinct eigenspaces with associated eigenvalues [B.37].

Theorem 1. Let $\mathcal{T}$ be an ultrasymmetric matrix operator. Then it has the following eigenspaces,

1. Two 1-dimensional eigenspaces whose eigenvectors have identical structure to Σ^*,
$$e_i^1 = (\lambda_i^1 - \alpha^1)I + (\beta^1 + \alpha^1 - \lambda_i^1)\mathbf{1}\mathbf{1}^T \tag{4}$$
with eigenvalue λ_i^1.

2. Two $(B-1)$-dimensional eigenspaces whose eigenvectors are permutations of the matrix,
$$e_i^2 = \begin{pmatrix} \lambda_i^2 - \alpha^2 & 0 & -\beta^2 & -\beta^2 & \cdots \\ 0 & -(\lambda_i^2 - \alpha^2) & \beta^2 & \beta^2 & \cdots \\ -\beta^2 & \beta^2 & 0 & 0 & \cdots \\ -\beta^2 & \beta^2 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{5}$$
with eigenvalues λ_i^2.

3. An eigenspace of dimension $B(B-3)/2$ whose eigenvectors are of the form $e^3 = G\Sigma G$ such that e^3 is symmetric and $\text{Diag}(e^3) = 0$. The eigenvalue of all such eigenvectors is λ^3.

The eigenvalues as well as α and β are not arbitrary but depend on the specific choice of ultrasymmetric matrix. In the case of fully-connected networks, the number of distinct eigenspaces reduces to two whose eigenvalues are identical to those found via the simplified analysis presented in Schoenholz et al. (2016) [B.62].

Similar arguments allow us to develop a theory for the statistics of gradients. The backpropogation algorithm gives an efficient method of propagating gradients from the end of the network to the earlier layers as,

$$\frac{\partial L}{\partial W^l} = \sum_i \boldsymbol{\delta}_i^l \phi(\boldsymbol{z}_i^{l-1})^T \qquad \delta_{\alpha i}^l = \phi'(z_{\alpha i}^l) \sum_\beta W_{\beta\alpha}^{l+1} \delta_{\beta i}^{l+1}. \tag{6}$$

Here $\boldsymbol{\delta}_i^l = \frac{\partial L}{\partial \boldsymbol{z}_i^l}$ are N_l-dimensional vectors that describe the error signal from neurons in the l'th layer due to the i'th element of the batch. The preceding discussion gave a precise characterization of the statistics of the $\boldsymbol{z}_i^l$ that we can leverage to understand the statistics of $\boldsymbol{\delta}_i^l$. It is easy to see that $\mathbb{E}[\delta_{\alpha i}^l] = 0$ and $\mathbb{E}[\delta_{\alpha i}^l \delta_{\beta j}^l] = \tilde{\Sigma}_{ij}^l \delta_{\alpha\beta}$ where $\tilde{\Sigma}^l$ is a covariance matrix and we may once again drop the neuron index. We can construct a recurrence relation to compute $\tilde{\Sigma}^l$,

$$\tilde{\Sigma}^l = \sigma_w^2 V_{\phi'}(\Sigma^l) \odot \tilde{\Sigma}^{l+1}. \tag{7}$$

Typically, we will be interested in understanding the dynamics of $\tilde{\Sigma}^l$ when Σ^l has converged exponentially towards its fixed point. Thus, we study the approximation,

$$\tilde{\Sigma}^l \approx \sigma_w^2 V_{\phi'}(\Sigma^*) \odot \tilde{\Sigma}^{l+1}. \tag{8}$$

Since these dynamics are linear, explosion and vanishing of gradients will be controlled by the eigenvalues of $V_{\phi'}(\Sigma^*)$.

3.1 BATCH NORMALIZATION

We now extend the mean field formalism to include batch normalization. Here, the definition for the neural network is modified to the coupled equations,

$$\boldsymbol{z}_i^l = W^l \phi(\tilde{\boldsymbol{z}}_i^{l-1}) + b^l \qquad \tilde{z}_{\alpha i}^l = \gamma_\alpha \frac{z_{\alpha i}^l - \mu_\alpha}{\sigma_\alpha} + \beta_\alpha \tag{9}$$

where $\mu_\alpha = \frac{1}{N_l}\sum_i z_{\alpha i}$ and $\sigma_\alpha^2 = \sqrt{\frac{1}{N_l}\sum_i (z_{\alpha i} - \mu_\alpha)^2 + \epsilon}$ are the per-neuron batch statistics. In practice $\epsilon \approx 10^{-5}$ or so to prevent division by zero, but in this paper, unless stated otherwise (in the last few sections), ϵ is assumed to be 0. Unlike in the case of vanilla fully-connected networks, here the pre-activations are invariant to σ_w^2 and σ_b^2. Without a loss of generality, we therefore set $\sigma_w^2 = 1$ and $\sigma_b^2 = 0$ for the remainder of the text. In principal, batch normalization additionally yields a pair of hyperparameters γ and β which are set to be constants. However, these may be incorporated into the nonlinearity and so without a loss of generality we set $\gamma = 1$ and $\beta = 0$.

The arguments from the previous section can proceed identically and we conclude that as the width of the network grows, the pre-activations will be jointly Gaussian with identically distributed neurons. Thus, we arrive at an analogous expression to eq. (2),

$$\Sigma^l = V_{\tilde{\phi}}(\Sigma^{l-1}) \qquad \text{where} \qquad \tilde{\phi}(h) = \phi\left(\frac{\sqrt{B}Gh}{||Gh||}\right). \tag{10}$$

Here we have introduced the projection operator $G = I - \frac{1}{B}\mathbf{1}\mathbf{1}^T$ which is defined such that $Gx = x - \mu\mathbf{1}$ with $\mu = \sum_i x_i/N$. Unlike ϕ, $\tilde{\phi}$ is does not act component-wise on h. It is therefore not obvious whether $V_{\tilde{\phi}}$ can be evaluated without performing a B dimensional Gaussian integral.

We present a pair of results that simplify eq. (10) to a small number of integrals - independent of B - over V_ϕ by finding integral transforms to relate the two functions. From previous work Poole et al. (2016), V_ϕ can be expressed in terms of a two-dimensional Gaussian integrals independent of B. When ϕ is degree-α positive homogeneous (e.g. rectified linear activations) we can relate V_ϕ and $V_{\tilde{\phi}}$ by the Laplace transform [B.1.1].

Theorem 2. Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is degree-α positive homogeneous. For any positive semi-definite matrix Σ define the projection $\Sigma^G = G\Sigma G$. Then

$$V_{\tilde{\phi}}(\Sigma) = \frac{B^\alpha}{\Gamma(\alpha)} \int_0^\infty ds s^{\alpha-1} \frac{V_\phi(\Sigma^G(I + 2s\Sigma^G)^{-1})}{\sqrt{\det(I + 2s\Sigma^G)}}. \tag{11}$$

Using this parameterization when V_ϕ has a closed form solution $V_{\tilde{\phi}}$ involves only a single integral. We further show that for any ϕ, $V_{\tilde{\phi}}$ can be related to V_ϕ by Fourier transform at the expense of an additional integral to perform the change of variables $r = ||Gh||$ [B.1.2].

Theorem 3. For general $\phi : \mathbb{R} \to \mathbb{R}$ with finite Gaussian moments,

$$V_{\tilde{\phi}}(\Sigma) = \int_0^\infty d(r^2) \int_{-\infty}^\infty \frac{d\lambda}{2\pi} e^{i\lambda r^2} \frac{V_\phi(G(\Sigma^{-1} + 2i\lambda G/B)Gr^{-2})}{\sqrt{\det(I + 2i\lambda G\Sigma/B)}} \tag{12}$$

Together these theorems provide analytic recurrence relations for random neural networks with batch normalization over a wide range of activation functions. By analogy to the fully-connected case we would like to study the dynamical system over covariance matrices induced by these equations.

We begin by investigating the fixed point structure of eq. (10). As in the case of feed-forward networks, permutation symmetry implies that there exist fixed points of the form $\Sigma^* = q^*[(1 - c^*)I + c^*\mathbf{1}\mathbf{1}^T$. A low-dimensional integral expression for q^* and c^* can be obtained by transforming to hyperspherical coordinates [B.3.1].

Theorem 4. For $B \geq 4$ the fixed point $\Sigma^* = q^*[(1 - c^*)I + c^*\mathbf{1}\mathbf{1}^T]$ satisfies,

$$q^* = \frac{\Gamma\left(\frac{B-1}{2}\right)}{\Gamma\left(\frac{B-2}{2}\right)\sqrt{\pi}} \int_0^\pi d\theta_1 \sin^{B-3}\theta_1 \phi(\sqrt{B}\zeta_1(\theta_1))^2 \tag{13}$$

$$q^* c^* = \frac{B-3}{2\pi} \int_0^\pi d\theta_1 \int_0^\pi d\theta_2 \sin^{B-3}\theta_1 \sin^{B-4}\theta_2 \phi(\sqrt{B}\zeta_1(\theta_1))\phi(\sqrt{B}\zeta_2(\theta_1, \theta_2)) \tag{14}$$

where

$$\zeta_1(\theta) = -\sqrt{\frac{B-1}{B}}\cos\theta, \quad \zeta_2(\theta_1,\theta_2) = \frac{1}{\sqrt{B-1}}\left[\frac{1}{\sqrt{B}}\cos\theta_1 - \sqrt{B-2}\sin\theta_1\cos\theta_2\right]. \tag{15}$$

While these equations allow for the efficient computation of fixed points for arbitrary activation functions, significant simplification occurs when the activation functions are α-homogeneous [B.3.2]. In particular, for rectified linear activations we arrive at the following result.

Theorem 5. When $\phi = \text{ReLU}$, there is a unique fixed point of the form $\Sigma^* = u^* I + v^* \mathbf{1}\mathbf{1}^T$ with,

$$q^* = \frac{B-1}{2\sqrt{\pi}} \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{B-1}{2}\right)}{\Gamma\left(\frac{B+1}{2}\right)} \qquad c^* = \mathcal{J}\left(\frac{-1}{B-1}\right) \tag{16}$$

where $\mathcal{J}(c) = \frac{1}{\pi}(\sqrt{1-c^2} + (\pi - \arccos(c))c)$ is the arccosine kernel Cho & Saul (2009a).

Together, these results describe the fixed points for most commonly used activation functions.

In the presence of batch normalization, when the activation function grows quickly, a winner-take-all phenomenon can occur where a subset of samples in the batch have much bigger activations than others. This causes the covariance matrix to form blocks of differing magnitude, breaking the BSB1 symmetry. One observes this, for example, as the degree α of α-relu increases past a point $\alpha_{\text{transition}}(B)$ depending on the batch size B. We examine this in more detail and give concrete examples in the appendix. However, by far most of the nonlinearities used in practice, like ReLU, leaky ReLU, tanh, sigmoid, etc, all lead to BSB1 fixed points. Thus from here on, we assume that any nonlinearity ϕ mentioned induces Σ^l to converge to BSB1 fixed points.

3.1.1 Linearized Dynamics

With the fixed point structure for batch normalized networks having been described, we now investigate the linearized dynamics of eq. (10) in the vicinity of these fixed points. As in the vanilla setting, we leverage the properties of ultrasymmetric matrices; however, as a consequence of mean subtraction with batch normalization here there are only three unique eigenspaces with $\lambda_1^1 = \lambda_2^1$, and $\lambda_1^2 = \lambda_2^2$ and in this case we label them $\mathbb{G}$, $\mathbb{L}$, and $\mathbb{M}$ respectively. These eigenspaces have an intuitive interpretation and in particular $\mathbb{G}$ captures the size of the batch; $\mathbb{L}$ captures the fluctuation between norms of the elements of the batch; $\mathbb{M}$ captures the correlation subject to zero mean constraint.

To determine the eigenvalues of $\frac{dV_{\tilde{\phi}}}{d\Sigma}\big|_{\Sigma=\Sigma^*}$ it is helpful to consider the action of batch normalization in more detail. In particular, we notice that $\tilde{\phi}$ can be decomposed into the composition of three separate operations, $\tilde{\phi} = \phi \circ r \circ G$. As discussed above, Gh subtracts the mean from h and we introduce the new function $r(h) = \sqrt{B}h/||h||$ which normalizes h by its standard deviation. Applying the chain rule, we can rewrite the Jacobian as,

$$\frac{dV_{\tilde{\phi}}}{d\Sigma} = \frac{dV_{\phi\circ r}}{d\Sigma} \circ G^{\otimes 2} \tag{17}$$

where $\circ$ denotes composition and $G^{\otimes 2}$ is the natural extension of G to act on matrices as $G^{\otimes 2}\{\Sigma\} = G\Sigma G = \Sigma^G$. It ends up being advantageous to study $G^{\otimes 2} \circ \frac{dV_{\phi\circ r}}{d\Sigma}$ and to note that the nonzero eigenvalues of this object are identical to the nonzero eigenvalues of the Jacobian [B.42].

As in the previous section there are two distinct ways to make progress on the spectrum of eq. (17). For arbitrary nonlinearity one can transform to hyperspherical coordinates which leads to tractable integral equations for the eigenvalues. The resulting equations for the eigenvalues can be evaluated, but are complicated and the specific form is relatively unenlightening [B.47] . In the case of positive-homogeneous activation functions we arrive at a relatively compact representation for the different eigenvalues [B.68]. Here, we summarize the results for rectified linear networks.

Theorem 6. Let $\phi = \text{ReLU}$ and $B > 3$. The eigenvalues for the different eigenspaces outlined above are

$$\lambda_{\mathbb{G}} = 0 \tag{18}$$

$$\lambda_{\mathbb{M}} = \frac{B}{2\sqrt{\pi}v^*} \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{B+1}{2}\right)}{\Gamma\left(\frac{B+3}{2}\right)} \mathcal{J}'\left(\frac{-1}{B-1}\right) \tag{19}$$

$$\lambda_{\mathbb{L}} = \frac{1}{2\sqrt{\pi}v^*} \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{B+1}{2}\right)}{\Gamma\left(\frac{B+3}{2}\right)} \left\{(B-2)\left[1 - \mathcal{J}\left(\frac{-1}{B-1}\right)\right] + \frac{B}{B-1}\mathcal{J}'\left(\frac{-1}{B-1}\right)\right\}. \tag{20}$$

Together these eigenvalues along with the fixed point outlined in Theorem 5 completely characterize the statistics of pre-activations in deep networks with batch normalization.

3.1.2 Gradient Backpropagation

With a mean field theory of the pre-activations of feed-forward networks with batch normalization having been developed, we turn our attention to the backpropagation of gradients. In contrast to the case of networks without batch normalization, we will see that exploding gradients at initialization are a severe problem here. To this end, one of the main results from this section will be to show that fully-connected networks with batch normalization feature exploding gradients for *any* choice of nonlinearity such that $\Sigma^l \to$ a BSB1 fixed point. Below, by "rate of gradient explosion" we mean the rate at which the gradient norm squared grows with depth.

As a starting point we seek an analog of eq. (8) in the case of batch normalization. However, because the activation functions no longer act point-wise on the pre-activations, the backpropagation equation becomes,

$$\delta^l_{\alpha i} = \sum_{\beta j} \frac{\partial \tilde{\phi}(z^l_{\alpha j})}{\partial z^l_{\alpha i}} W^{l+1}_{\beta\alpha} \delta^{l+1}_{\beta j} \tag{21}$$

where we observe the additional sum over the batch. Computing the resulting covariance matrix $\tilde{\Sigma}^l$, we arrive at the recurrence relation,

$$\tilde{\Sigma}^l = \sigma_w^2 \mathbb{E}\left[\left(\frac{\partial \tilde{\phi}(h)}{\partial h}\right)^T \tilde{\Sigma}^{l+1} \frac{\partial \tilde{\phi}(h)}{\partial h} : h \sim \mathcal{N}(0, \Sigma^l)\right] =: \sigma_w^2 V_{\tilde{\phi}'}(\Sigma^l)^\dagger\{\tilde{\Sigma}^{l+1}\} \tag{22}$$

where and we have defined the linear operator $\tilde{\Sigma} \mapsto V_F(\Sigma)^\dagger\{\tilde{\Sigma}\} = \mathbb{E}[F_h^T \tilde{\Sigma} F_h : h \sim \mathcal{N}(0, \Sigma)]$ for any vector-indexed linear operator F_h. As in the case of vanilla feed-forward networks, here we will be concerned with the behavior of gradients when Σ^l is close to its fixed point. We therefore study the asymptotic approximation to eq. (22) given by $\tilde{\Sigma}^l = V_{\tilde{\phi}'}(\Sigma^*)^\dagger\{\tilde{\Sigma}^{l+1}\}$. In this case the dynamics of $\tilde{\Sigma}$ are linear and are therefore naturally determined by the eigenvalues of $V_{\tilde{\phi}'}(\Sigma^*)$.

As in the forward case, batch normalization is the composition of three operations $\tilde{\phi} = \phi \circ r \circ G$. Applying the chain rule, eq. (22) can be rewritten as,

$$V_{\tilde{\phi}'}(\Sigma)^\dagger = G^{\otimes 2} \circ \mathbb{E}\left[\left(\frac{\partial(\phi \circ r)(z)}{\partial z}\bigg|_{z=Gh}^T\right)^{\otimes 2} : h \sim \mathcal{N}(0, \Sigma)\right] = G^{\otimes 2} \circ F(\Sigma) \tag{23}$$

with $F(\Sigma)$ appropriately defined. Note that $(V_{\tilde{\phi}'}(\Sigma)^\dagger)^n = (G^{\otimes 2} \circ F(\Sigma))^n = (G^{\otimes 2} \circ F(\Sigma) \circ G^{\otimes 2})^{n-1} \circ G^{\otimes 2} \circ F(\Sigma)$, so that it suffices to study the eigendecomposition of $G^{\otimes 2} \circ F(\Sigma^*) \circ G^{\otimes 2}$. Due to the symmetry of Σ^*, this operator is ultrasymmetric, so that its eigenspaces are $\mathbb{G}, \mathbb{L}, \mathbb{M}$ and we can compute its eigenvalues as in Section 3.1.1. However, this computation is not so enlightening as to the dependence of these eigenvalues on the nonlinearity. We instead use the Laplace and Fourier methods to derive more explicit representations of the eigenvalues. Here we highlight our results on the max eigenvalue, $\lambda_{\max} = \lambda_{\mathbb{G}}$, which determines the asymptotic dynamics of $\tilde{\Sigma}^l$.

Theorem 7. For any well-behaved nonlinearity ϕ such that Σ^l converges to a BSB1 fixed point with depth $l \to \infty$, the gradient explodes asymptotically at the rate of

$$\frac{(B(B-2))^{2-B/2}}{q^*(1-c^*)(2\pi)} \int_{Z>0} \left((B-1-z_1^2)\phi'(z_1)^2 + (1+z_1 z_2)\phi'(z_1)\phi'(z_2)\right) Z^{(B-5)/2} \, \mathrm{d}z_1 \, \mathrm{d}z_2 \tag{24}$$

where $Z = Z(z_1, z_2) = B(B-2) - B(z_1^2 + z_2^2) + (z_1 - z_2)^2$.

Theorem 8. In a ReLU-batchnorm network, gradients explode asymptotically at the rate

$$\frac{(B-3+2\alpha)\left((2\alpha-1)\mathcal{J}'\left(\frac{-1}{B-1}\right) + \alpha^2(B-1)\mathcal{J}(1)\right)}{(2\alpha-1)(B-3)(B-1)(\mathcal{J}(1) - \mathcal{J}\left(\frac{-1}{B-1}\right))} - \frac{\alpha^2}{B-3} \tag{25}$$

which decreases to $\frac{\pi}{\pi-1}$ as $B \to \infty$. In contrast, for a linear batchnorm network, the gradients explode asymptotically at the rate $\frac{B-2}{B-3}$, which goes to 1 as $B \to \infty$.

Section 3.1.2 shows theory and simulation for ReLU gradient dynamics.

By noticing that the integral in Theorem 7 diagonalizes over the Gegenbauer basis, we obtain the following lower bound on the gradient explosion rate:

Theorem 9 (Batchnorm causes gradient explosion). Suppose $\phi(z)$ has the Gegenbauer expansion $\phi(z) = \sum_{k=0}^{\infty} a_k \sqrt{\frac{B-3+2k}{(B-3)\binom{B-4+k}{k}}} C_k^{(B-3)/2}\left(\frac{z}{\sqrt{B-1}}\right)$, normalized so that

$$\frac{(B-1)^{(B-3)/2}\Gamma\left(\frac{B}{2}-\frac{1}{2}\right)}{\Gamma\left(\frac{B}{2}-1\right)\sqrt{\pi}} \int_{-\sqrt{B-1}}^{\sqrt{B-1}} \mathrm{d}z \phi(z)^2((B-1)-z^2)^{(B-4)/2} = \sum_{k=0}^{\infty} a_k^2. \tag{26}$$

Then

$$\lambda_{\max} = \frac{\sum_{k=1}^{\infty} k \frac{B-3+k}{B-3} c_k a_k^2}{\sum_{k=1}^{\infty} c_k a_k^2} \tag{27}$$

where $c_k = 1 - (-1)^k {}_2F_1(-k, B-3+k, \frac{B}{2}-1; \frac{B-2}{2(B-1)}) > 0$ for all $k > 0$. Consequently, for any non-constant ϕ (i.e. there is a $j > 0$ such that $a_j \neq 0$), $\lambda_{\max} > 1$; ϕ minimizes $\lambda_{\max}$ iff it is linear (i.e. $a_i = 0 \forall i \geq 2$), in which case gradients explode at the rate of $\frac{B-2}{B-3}$.

This contrasts starkly with the case of nonnormalized fully-connected networks, which can use the weight and bias variances to control its mean field network dynamics Poole et al. (2016); Schoenholz et al. (2016). As a corollary, we disprove the conjecture of the original batchnorm paper Ioffe & Szegedy (2015) that "Batch Normalization may lead the layer Jacobians to have singular values close to 1" in the initialization setting, and in fact prove the exact opposite, that *batchnorm forces the layer Jacobian singular values away from 1*.

ϵ as a hyperparameter In practice, ϵ is usually treated as small constant and is not regarded as a hyperparameter to be tuned. Nevertheless, we can investigate its effect on gradient explosion. A straightforward generalization of the analysis presented above to to the case of $\epsilon > 0$ suggests somewhat larger ϵ values than typically used can ameliorate (but not eliminate) gradient explosion problems. See Fig. 4(c,d).

4 Experiments

Having developed a theory for neural networks with batch normalization at initialization, we now explore the relationship between the properties of these random networks and their learning dynamics. We will see that the trainability of networks with batch normalization is controlled by gradient explosion. We quantify the depth scale over which gradients explode by $\xi = 1/\log \lambda_{\mathbb{G}}$ where, as above, $\lambda_{\mathbb{G}}$ is the largest eigenvalue of the jacobian. Across many different experiments we will see strong agreement between ξ and the maximum trainable depth.

We first investigate the relationship between trainability and initialization for rectified linear networks as a function of batch size. The results of these experiments are shown in fig. 2 where in

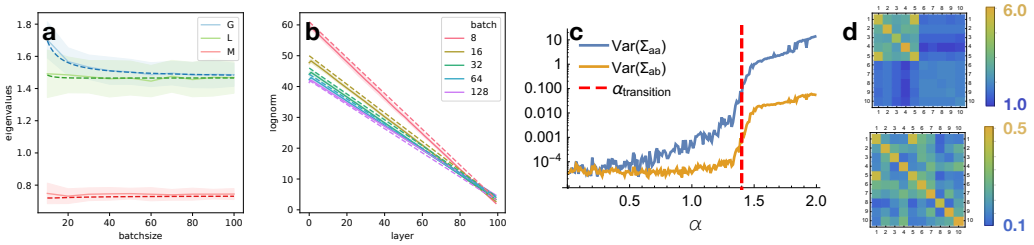

Figure 1: **Numerical confirmation of theoretical predictions.** (a,b) Comparison between theoretical prediction (dashed lines) and Monte Carlo simulations (solid lines) for the eigenvalues of backwards jacobian as a function of batch size and the magnitude of gradients as a function of depth respectively for rectified linear networks. In each case Monte Carlo simulations are averaged over 200 sample networks of width 1000 and shaded regions denote 1 standard deviation. (c,d) Demonstration of the existence of a BSB1 to BSB2 symmetry breaking transition as a function of α for α-homogeneous activation functions. In (c) we plot the empirical variance of the eigenvalues of the covariance matrix which clearly shows a jump at the transition. In (d) we plot representative covariance matrices for the two phases (BSB1 bottom, BSB2 top).

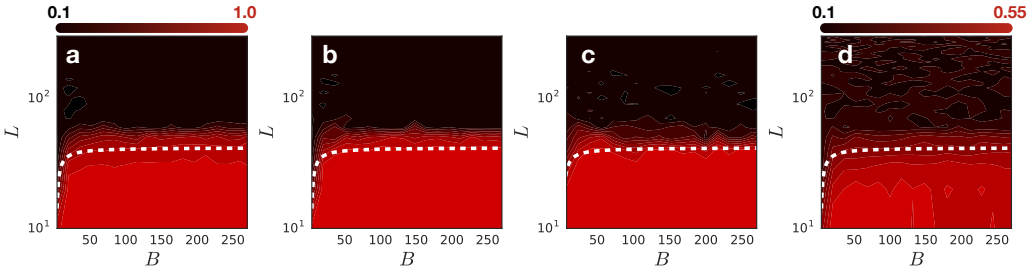

Figure 2: **Batch normalization strongly limits the maximum trainable depth.** Colors show test accuracy for rectified linear networks with batch normalization and $\gamma = 1$, $\beta = 0$, $\epsilon = 10^{-3}$, $N = 384$, and $\eta = 10^{-5}B$. (a) trained on MNIST for 10 epochs (b) trained with fixed batch size 1000 and batch statistics computed over sub batches of size B. (c) trained using RMSProp. (d) Trained on CIFAR10 for 50 epochs.

each case we plot the test accuracy after training as a function of the depth and the batch size and overlay 16ξ in white dashed lines. In fig. 2 (a) we consider networks trained using SGD on MNIST where we observe that networks deeper than about 50 layers are untrainable regardless of batch size. In (b) we compare standard batch normalization with a modified version in which the batch size is held fixed but batch statistics are computed over subsets of size B. This removes subtle gradient fluctuation effects noted in Smith & Le (2018). In (c) we do the same experiment with RMSProp and in (d) we train the networks on CIFAR10. In all cases we observe a nearly identical trainable region.

It is counter intuitive that training can occur at intermediate depths where there is significant gradient explosion. To gain insight into the behavior of the network during learning we record the magnitudes of the weights, the gradients with respect to the pre-activations, and the gradients with respect to the weights for the first 10 steps of training for networks of different depths. The result of this experiment is shown in fig. 3. Here we see that before learning, as expected, the norm of the weights is constant and independent of layer while the gradients feature exponential explosion. However, we observe that two related phenomena occur after a single step of learning: the weights grow exponentially in the depth and the magnitude of the gradients are stable up to some threshold after which they vanish exponentially in the depth. Thus, it seems that although the gradients of batch normalized networks at initialization are ill-conditioned, the gradients appear to quickly reach a stable dynamical equilibrium. Pathologically, in very high depth settings, the relative gradient vanishing can in fact be so severe as to cause lower layers to mostly stay constant during training.

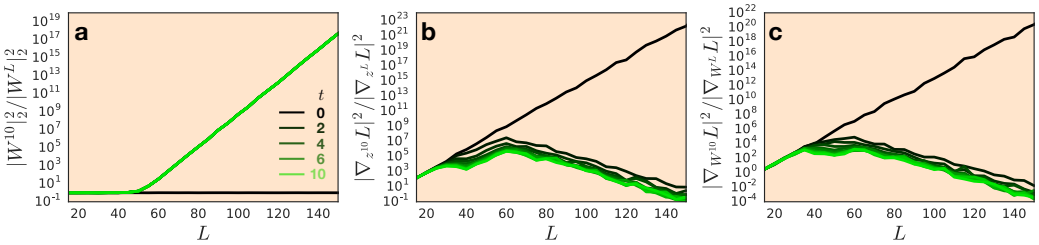

Figure 3: **Gradients in networks with batch normalization quickly achieve dynamical equilibrium.** Plots of the relative magnitudes of (a) the weights (b) the gradients of the loss with respect to the pre-activations and (c) the gradients of the loss with respect to the weights for rectified linear networks of varying depths during the first 10 steps of training. Colors show step number from 0 (black) to 10 (green).

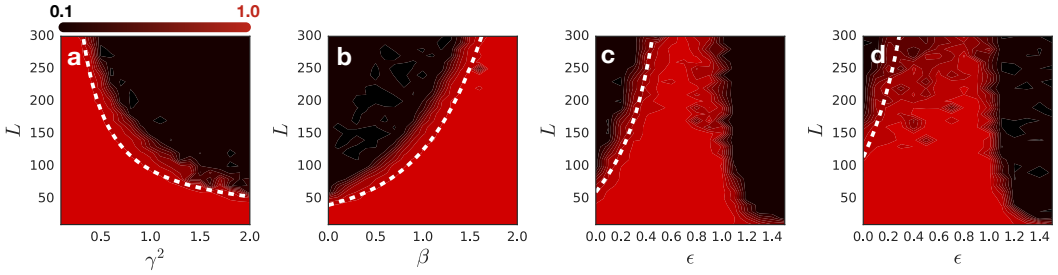

Figure 4: **Three techniques for counteracting gradient explosion.** Test accuracy on MNIST as a function of different hyperparameters along with theoretical predictions (white dashed line) for the maximum trainable depth. (a) $tanh$ network changing the overall scale of the pre-activations, here $\gamma \to 0$ corresponds to the linear regime. (b) Rectified linear network changing the mean of the pre-activations, here $\beta \to \infty$ corresponds to the linear regime. (c,d) $tanh$ and rectified linear networks respectively as a function of ϵ, here we observe a well defined phase transition near $\epsilon \sim 1$. Note that in the case of rectified linear activations we use $\beta = 2$ so that the function is locally linear about 0. We also find initializing β and/or setting $\epsilon > 0$ having positive effect on VGG19 with batchnorm. See Figs. 5 and 6

As discussed in the theoretical exposition above, batch normalization necessarily features exploding gradients for any nonlinearity that converges to a BSB1 fixed point. We performed a number of experiments exploring different ways of ameliorating this gradient explosion. These experiments are shown in fig. 4 with theoretical predictions for the maximum trainable depth overlaid; in all cases we see exceptional agreement. In fig. 4 (a,b) we explore two different ways of tuning the degree to which activation functions in a network are nonlinear. In fig. 4 (a) we tune $\gamma \in [0, 2]$ for networks with $\tanh$-activations and note that in the $\gamma \to 0$ limit the function is linear. In fig. 4 (b) we tune $\beta \in [0, 2]$ for networks with rectified linear activations and we note, similarly, that in the $\beta \to \infty$ limit the function is linear. As expected, we see the maximum trainable depth increase significantly with decreasing γ and increasing β. In fig. 4 (c,d) we vary ϵ for $tanh$ and rectified linear networks respectively. In both cases, we observe a critical point at large ϵ where gradients do not explode and very deep networks are trainable.

5 Conclusion

In this work we have presented a theory for neural networks with batch normalization at initialization. In the process of doing so, we have uncovered a number of counter intuitive aspects of batch normalization and - in particular - the fact that at initialization it causes gradients to explode necessarily. We have introduced several methods to reduce the degree of gradient explosion enabling the training of significantly deeper networks in the presence of batch normalization. Finally, this work paves the way for future work on more advanced, state-of-the-art, network topologies.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *arXiv:1607.06450 [cs, stat]*, July 2016. URL http://arxiv.org/abs/1607.06450. 00329 arXiv: 1607.06450.

David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The Shattered Gradients Problem: If resnets are the answer, then what is the question? In *PMLR*, pp. 342–350, July 2017. URL http://proceedings.mlr.press/v70/balduzzi17b.html.

Johan Bjorck, Carla Gomes, and Bart Selman. Understanding Batch Normalization. June 2018. URL https://arxiv.org/abs/1806.02375.

Minmin Chen, Jeffrey Pennington, and Samuel Schoenholz. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 873–882, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/chen18i.html.

Q. Chen and R. Wu. CNN Is All You Need. *ArXiv e-prints*, December 2017.

Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pp. 342–350, 2009a.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pp. 342–350, 2009b. URL http://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. *arXiv:1602.05897 [cs, stat]*, February 2016. URL http://arxiv.org/abs/1602.05897. arXiv: 1602.05897.

Igor Gitman and Boris Ginsburg. Comparison of Batch Normalization and Weight Normalization Algorithms for the Large-scale Image Classification. *arXiv:1709.08145 [cs]*, September 2017. URL http://arxiv.org/abs/1709.08145. 00002 arXiv: 1709.08145.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv: 1512.03385.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*, February 2015. URL http://arxiv.org/abs/1502.03167. arXiv: 1502.03167.

Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Ming Zhou, Klaus Neymeyr, and Thomas Hofmann. Towards a Theoretical Understanding of Batch Normalization. *arXiv:1805.10694 [cs, stat]*, May 2018. URL http://arxiv.org/abs/1805.10694. 00000 arXiv: 1805.10694.

Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pp. 396–404, 1990.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Ping Luo, Xinjiang Wang, Wenqi Shao, and Zhanglin Peng. Understanding Regularization in Batch Normalization. *arXiv:1809.00846 [cs, stat]*, September 2018. URL http://arxiv.org/abs/1809.00846. arXiv: 1809.00846.

Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in neural information processing systems*, pp. 4785–4795, 2017.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *arXiv:1606.05340 [cond-mat, stat]*, June 2016. URL http://arxiv.org/abs/1606.05340. arXiv: 1606.05340.

Tim Salimans and Diederik P. Kingma. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. February 2016. URL https://arxiv.org/abs/1602.07868. 00149.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How Does Batch Normalization Help Optimization? (No, It Is Not About Internal Covariate Shift). *arXiv:1805.11604 [cs, stat]*, May 2018. URL http://arxiv.org/abs/1805.11604. 00000 arXiv: 1805.11604.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep Information Propagation. *arXiv:1611.01232 [cs, stat]*, November 2016. URL http://arxiv.org/abs/1611.01232. arXiv: 1611.01232.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. 2018.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5393–5402, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/xiao18a.html.

Greg Yang and Samuel S. Schoenholz. Meanfield Residual Network: On the Edge of Chaos. In *Advances in neural information processing systems*, 2017.

Greg Yang and Samuel S Schoenholz. Deep mean field theory: Layerwise variance and width variation as methods to control gradient explosion. 2018.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition.

A VGG19 WITH BATCHNORM ON CIFAR100

We find acceleration effects, especially in initial training, due to setting $\epsilon > 0$ and/or initializing $\beta > 0$. See Figs. 5 and 6.

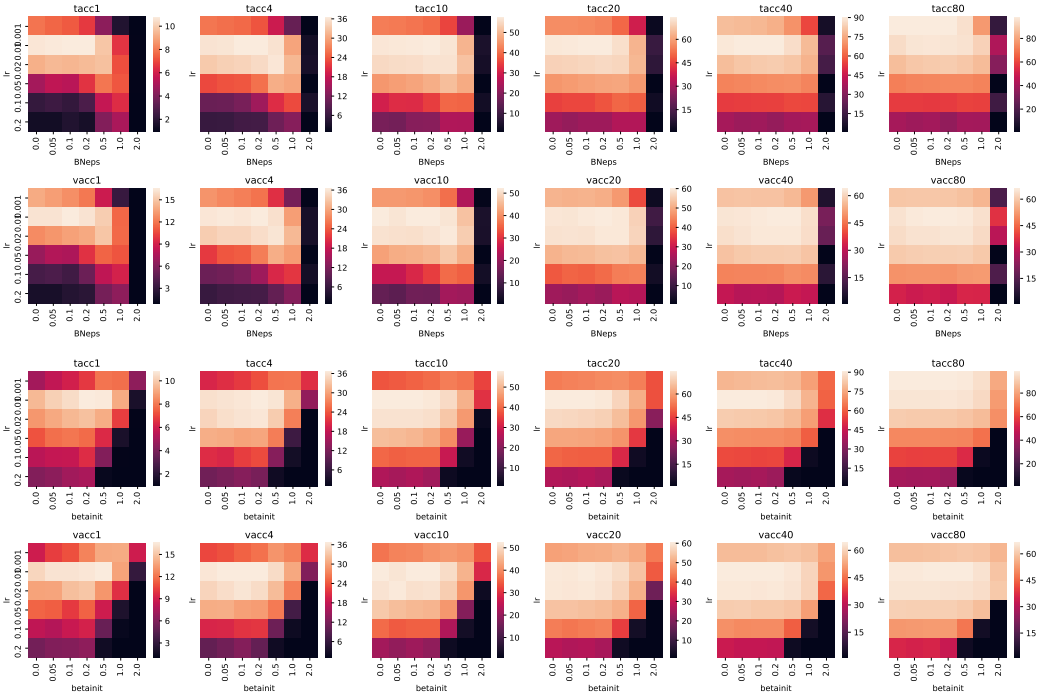

Figure 5: We sweep over different values of learning rate, β initialization, and ϵ, in training VGG19 with batchnorm on CIFAR100 with data augmentation. We use 8 random seeds for each combination, and assign to each combination the median training/validation accuracy over all runs. We then aggregate these scores here. In the first row we look at training accuracy with different learning rate vs β initialization at different epochs of training, presenting the max over ϵ. In the second row we do the same for validation accuracy. In the third row, we look at the matrix of training accuracy for learning rate vs ϵ, taking max over β. In the fourth row, we do the same for validation accuracy.

In what follows, we adopt a slightly different notation from the main text in order to express the mean field theory of batchnorm more faithfullly.

B FORWARD DYNAMICS

Definition B.1. Let $\mathcal{S}_B$ be the space of PSD matrices of size $B \times B$. Given a measurable function $\Phi : \mathbb{R}^B \to \mathbb{R}^B$, define the integral transform $\mathrm{V}\Phi : \mathcal{S}_B \to \mathcal{S}_B$ by $\mathrm{V}\Phi(\Sigma) = \mathrm{E}[\Phi(h)^{\otimes 2} : h \sim \mathcal{N}(0, \Sigma)]$.[4] When $\phi : \mathbb{R} \to \mathbb{R}$ and B is clear from context, we also write $\mathrm{V}\phi$ for V applied to the function acting coordinatewise by ϕ.

Definition B.2. For any $\phi : \mathbb{R} \to \mathbb{R}$, let $\mathfrak{B}_\phi : \mathbb{R}^B \to \mathbb{R}^B$ be batchnorm (applied to a batch of neuronal activations) followed by coordinatewise applications of ϕ, $\mathfrak{B}_\phi(x)_j = \phi\left((x_j - \operatorname{Avg} x) / \sqrt{\frac{1}{B}\sum_{i=1}^B (x_i - \operatorname{Avg} x)^2}\right)$ (here $\operatorname{Avg} x = \frac{1}{B}\sum_{i=1}^B x_i$). When $\phi = \mathrm{id}$ we will also write $\mathfrak{B} = \mathfrak{B}_{\mathrm{id}}$. We write ρ for ReLU, so that $\mathfrak{B}_\rho$ is batchnorm followed by ReLU.

[4] This definition of V absorbs the previous definitions of V and W in Yang & Schoenholz (2017) for the scalar case

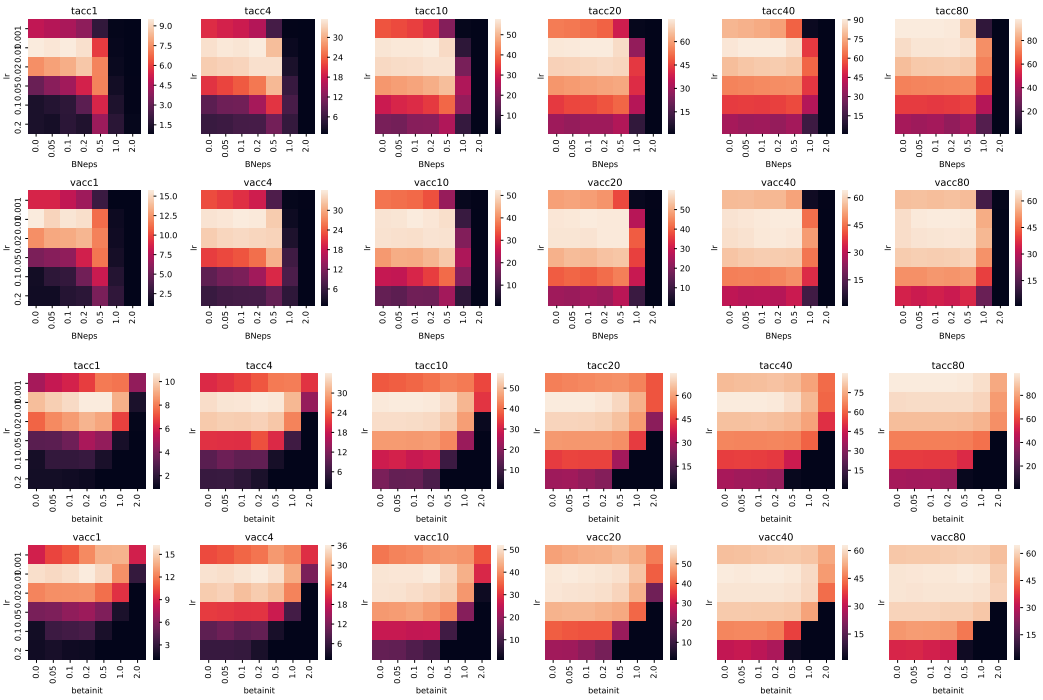

Figure 6: In the same setting as Fig. 5, except we don't take the max over the unseen hyperparameter but rather set it to 0 (the default value).

Definition B.3. Define the matrix $G^B = I - \frac{1}{B}\mathbb{1}\mathbb{1}^T$. Let $\mathcal{S}_B^G$ be the space of PSD matrices of size $B \times B$ with zero mean across rows and columns, $\mathcal{S}_B^G := \{\Sigma \in \mathcal{S}_B : G^B \Sigma G^B = \Sigma\} = \{\Sigma \in \mathcal{S}_B : \Sigma \mathbb{1} = 0\}$.

When B is clear from context, we will suppress the subscript/superscript B. In short, for $h \in \mathbb{R}^B$, Gh zeros the sample mean of h. G is a projection matrix to the subspace of vectors $h \in \mathbb{R}^B$ of zero coordinate sum. With the above definitions, we then have $\mathfrak{B}_\phi(h) = \phi(\sqrt{B}Gh/\|Gh\|)$.

In this section we will be interested in studying the dynamics on PSD matrices of the form

$$\Sigma^{(l)} = \mathrm{V}\mathfrak{B}_\phi(\Sigma^{(l-1)}) = \mathrm{E}[\phi(\sqrt{B}Gh/\|Gh\|)^{\otimes 2} : h \sim \mathcal{N}(0, \Sigma^{(l-1)})] \tag{28}$$

where $\Sigma^{(l)} \in \mathcal{S}_B$ and $\phi : \mathbb{R} \to \mathbb{R}$.

B.1 Simplification

On the face of it, the iteration map $\mathrm{V}\mathfrak{B}_\phi$ requires one to do an B dimensional integral. However, one can reduce this down to a constant number of dimension (independent of B) if the operator $\mathrm{V}\phi$ has a closed form. There are two ways to do this: 1) The Laplace method, which reduces the integral down to 1 dimension but requires the assumption that ϕ is positive homogeneous, and 2) The Fourier method, which reduces the integral down to 2 dimensions but allows ϕ to be any function.

B.1.1 Laplace Method

The key insight in the Laplace Method is to apply Schiwinger parametrization to deal with normalization.

Lemma B.4 (Schwinger parametrization)**.** *For $z > 0$ and $c > 0$,*

$$c^{-z} = \Gamma(z)^{-1} \int_0^\infty x^{z-1} e^{-cx} \, \mathrm{d}x$$

The following is the key lemma in the Laplace method.

Lemma B.5 (The Laplace Method Master Equation)**.** *For $A, A' \in \mathbb{N}$, let $f : \mathbb{R}^A \to \mathbb{R}^{A'}$ and let $k \geq 0$. Suppose $\|f(y)\| \leq h(\|y\|)$ for some nondecreasing function $h : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ such that $\mathrm{E}[h(r\|y\|) : y \in \mathcal{N}(0, I_A)]$ exists for every $r \geq 0$. Define $\wp(\Sigma) := \mathrm{E}[\|y\|^{-2k} f(y) : y \sim \mathcal{N}(0, \Sigma)]$. Then on $\{\Sigma \in \mathcal{S}_A : \operatorname{rank} \Sigma > 2k\}$, $\wp(\Sigma)$ is well-defined and continuous, and furthermore satisfies*

$$\wp(\Sigma) = \Gamma(k)^{-1} \int_0^\infty \mathrm{d}s\, s^{k-1} \det(I + 2s\Sigma)^{-1/2}\, \mathrm{E}[f(y) : y \sim \mathcal{N}(0, \Sigma(I + 2s\Sigma)^{-1})] \tag{29}$$

Proof. $\wp(\Sigma)$ is well-defined for full rank Σ because the $\|y\|^{-2k}$ singularity at $y = 0$ is Lebesgue-integrable in a neighborhood of 0 in dimension $A > 2k$.

We prove Eq. (29) in the case when Σ is full rank and then apply a continuity argument.

Proof of Eq. (29) for full rank Σ. First, we will show that we can exchange the order of integration

$$\int_0^\infty \mathrm{d}s \int_{\mathbb{R}^A} \mathrm{d}y\, s^{k-1} f(y) e^{-\frac{1}{2} y^T (\Sigma^{-1} + 2sI) y} = \int_{\mathbb{R}^A} \mathrm{d}y \int_0^\infty \mathrm{d}s\, s^{k-1} f(y) e^{-\frac{1}{2} y^T (\Sigma^{-1} + 2sI) y}$$

by Fubini-Tonelli's theorem. Observe that

$$\begin{aligned}
&(2\pi)^{-A/2} \det \Sigma^{-1/2} \int_0^\infty \mathrm{d}s\, s^{k-1} \int_{\mathbb{R}^A} \mathrm{d}y\, \|f(y)\| e^{-\frac{1}{2} y^T (\Sigma^{-1} + 2sI) y} \\
&= \int_0^\infty \mathrm{d}s\, s^{k-1} \det(\Sigma(\Sigma^{-1} + 2sI))^{-1/2}\, \mathrm{E}[\|f(y)\| : y \sim \mathcal{N}(0, (\Sigma^{-1} + 2sI)^{-1})] \\
&= \int_0^\infty \mathrm{d}s\, s^{k-1} \det(I + 2s\Sigma)^{-1/2}\, \mathrm{E}[\|f(y)\| : y \sim \mathcal{N}(0, \Sigma(I + 2s\Sigma)^{-1})]
\end{aligned}$$

For $\lambda = \|\Sigma\|_2$,

$$\|f(\sqrt{\Sigma(I + 2s\Sigma)^{-1}} y)\| \leq h(\|\sqrt{\Sigma(I + 2s\Sigma)^{-1}} y\| \leq h\left(\left\|\sqrt{\frac{\lambda}{1 + 2s\lambda}} y\right\|\right) \leq h(\sqrt{\lambda}\|y\|)$$

Because $\mathrm{E}[h(\sqrt{\lambda}\|y\|) : y \in \mathcal{N}(0, I)]$ exists by assumption,

$$\begin{aligned}
&\mathrm{E}[\|f(y)\| : y \sim \mathcal{N}(0, \Sigma(I + 2s\Sigma)^{-1})] \\
&= \mathrm{E}[\|f(\sqrt{\Sigma(I + 2s\Sigma)^{-1}} y)\| : y \sim \mathcal{N}(0, I)] \\
&\to \mathrm{E}[\|f(0)\| : y \sim \mathcal{N}(0, I)] = \|f(0)\|
\end{aligned}$$

as $s \to \infty$, by dominated convergence with dominating function $h(\sqrt{\lambda}\|y\|) e^{-\frac{1}{2}\|y\|^2} (2\pi)^{-A/2}$. By the same reasoning, the function $s \mapsto \mathrm{E}[\|f(y)\| : y \sim \mathcal{N}(0, \Sigma(I + 2s\Sigma)^{-1})]$ is continuous. In particular this implies that $\sup_{0 \leq s \leq \infty} \mathrm{E}[\|f(y)\| : y \sim \mathcal{N}(0, \Sigma(I + 2s\Sigma)^{-1})] < \infty$. Combined with the fact that $\det(I + 2s\Sigma)^{-1/2} = \Theta(s^{-A/2})$ as $s \to \infty$,

$$\begin{aligned}
&\int_0^\infty \mathrm{d}s\, s^{k-1} \det(I + 2s\Sigma)^{-1/2}\, \mathrm{E}[\|f(y)\| : y \sim \mathcal{N}(0, \Sigma(I + 2s\Sigma)^{-1})] \\
&\leq \int_0^\infty \mathrm{d}s\, \Theta(s^{k-1-A/2})
\end{aligned}$$

which is bounded by our assumption that $A/2 > k$. This shows that we can apply Fubini-Tonelli's theorem to allow exchanging order of integration.

Thus,

$$
\begin{aligned}
&\mathrm{E}[\|y\|^{-2k} f(y) : y \sim \mathcal{N}(0,\Sigma)] \\
&= \mathrm{E}[f(y) \int_0^\infty \mathrm{d}s\, \Gamma(k)^{-1} s^{k-1} e^{-\|y\|^2 s} : y \sim \mathcal{N}(0,\Sigma)] \\
&= (2\pi)^{-A/2} \det \Sigma^{-1/2} \int \mathrm{d}y\, e^{-\frac{1}{2} y^T \Sigma^{-1} y} f(y) \int_0^\infty \mathrm{d}s\, \Gamma(k)^{-1} s^{k-1} e^{-\|y\|^2 s} \\
&= (2\pi)^{-A/2} \Gamma(k)^{-1} \det \Sigma^{-1/2} \int_0^\infty \mathrm{d}s\, s^{k-1} \int \mathrm{d}y\, f(y) e^{-\frac{1}{2} y^T (\Sigma^{-1} + 2sI) y} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Fubini-Tonelli)} \\
&= \Gamma(k)^{-1} \int_0^\infty \mathrm{d}s\, s^{k-1} \det(\Sigma(\Sigma^{-1} + 2sI))^{-1/2}\, \mathrm{E}[f(y) : y \sim \mathcal{N}(0, (\Sigma^{-1} + 2sI)^{-1})] \\
&= \Gamma(k)^{-1} \int_0^\infty \mathrm{d}s\, s^{k-1} \det(I + 2s\Sigma)^{-1/2}\, \mathrm{E}[f(y) : y \sim \mathcal{N}(0, \Sigma(I + 2s\Sigma)^{-1})]
\end{aligned}
$$

Domain and continuity of $\wp(\Sigma)$. The LHS of Eq. (29), $\wp(\Sigma)$, is defined and continuous on $\operatorname{rank}\Sigma/2 > k$. Indeed, if $\Sigma = M I_C M^T$, where M is a full rank $A \times C$ matrix with $\operatorname{rank}\Sigma = C \leq A$, then

$$
\begin{aligned}
&\mathrm{E}[\|y\|^{-2k} \|f(y)\| : y \sim \mathcal{N}(0,\Sigma)] \\
&= \mathrm{E}[\|Mz\|^{-2k} \|f(Mz)\| : z \sim \mathcal{N}(0, I_C)].
\end{aligned}
$$

This is integrable in a neighborhood of 0 iff $C > 2k$, while it's always integrable outside a ball around 0 because $\|f\|$ by itself already is. So $\wp(\Sigma)$ is defined whenever $\operatorname{rank}\Sigma > 2k$. Its continuity can be established by dominated convergence.

Proof of Eq. (29) for $\operatorname{rank}\Sigma > 2k$. Observe that $\det(I + 2s\Sigma)^{-1/2}$ is continuous in Σ and, by an application of dominated convergence as in the above, $\mathrm{E}[f(y) : y \sim \mathcal{N}(0, \Sigma(I + 2s\Sigma)^{-1})]$ is continuous in Σ. So the RHS of Eq. (29) is continuous in Σ whenever the integral exists. By the reasoning above, $\mathrm{E}[f(y) : y \sim \mathcal{N}(0, \Sigma(I + 2s\Sigma)^{-1})]$ is bounded in s and $\det(I + 2s\Sigma)^{-1/2} = \Theta(s^{-\operatorname{rank}\Sigma/2})$, so that the integral exists iff $\operatorname{rank}\Sigma/2 > k$.

To summarize, we have proved that both sides of Eq. (29) are defined and continous for $\operatorname{rank}\Sigma > 2k$. Because the full rank matrices are dense in this set, by continuity Eq. (29) holds for all $\operatorname{rank}\Sigma > 2k$. □

If ϕ is degree-α positive homogeneous, i.e. $\phi(ru) = r^\alpha \phi(u)$ for any $u \in \mathbb{R}, r \in \mathbb{R}^+$, we can then compute

$$
\begin{aligned}
\mathrm{V}\mathfrak{B}_\phi(\Sigma) &= \mathrm{E}[\phi(\sqrt{B} Gh/\|Gh\|)^{\otimes 2} : h \sim \mathcal{N}(0,\Sigma)] = B^\alpha\, \mathrm{E}[\phi(Gh)^{\otimes 2}/\|Gh\|^{2\alpha} : h \sim \mathcal{N}(0,\Sigma)] \\
&= B^\alpha (2\pi)^{-B/2} (\det\Sigma)^{-1/2} \int \frac{\phi(Gh)^{\otimes 2}}{\|Gh\|^{2\alpha}} e^{-h^T \Sigma^{-1} h/2}\, \mathrm{d}h \\
&= B^\alpha (2\pi)^{-B/2} (\det\Sigma)^{-1/2} \int \phi(Gh)^{\otimes 2} e^{-h^T \Sigma^{-1} h/2} \int_0^\infty \Gamma(\alpha)^{-1} s^{\alpha-1} e^{-s\|Gh\|^2}\, \mathrm{d}s\, \mathrm{d}h \qquad (30) \\
&= B^\alpha (2\pi)^{-B/2} (\det\Sigma)^{-1/2} \Gamma(\alpha)^{-1} \int_0^\infty s^{\alpha-1} \int \phi(Gh)^{\otimes 2} e^{-\frac{1}{2}(h^T \Sigma^{-1} h + 2s h^T G h)}\, \mathrm{d}h\, \mathrm{d}s \\
&= B^\alpha (\det\Sigma)^{-1/2} \Gamma(\alpha)^{-1} \int_0^\infty \mathrm{d}s\, s^{\alpha-1} (\det(\Sigma^{-1} + 2sG)^{-1})^{1/2}\, \mathrm{E}[\phi(y)^{\otimes 2} : y \sim \mathcal{N}(0, G(\Sigma^{-1} + 2sG)^{-1} G)] \\
&= B^\alpha \Gamma(\alpha)^{-1} \int_0^\infty \mathrm{d}s\, s^{\alpha-1} \det(I + 2s\Sigma G)^{-1/2} \mathrm{V}\phi(G(I + 2s\Sigma G)^{-1} \Sigma G)
\end{aligned}
$$

where in Eq. (30) we applied Schwinger's Parametrization.

If $\phi = \rho$, then

$$
\mathrm{V}\phi(\Sigma)_{ij} = \begin{cases} \frac{1}{2}\Sigma_{ii} & \text{if } i = j \\ \frac{1}{2}\mathfrak{J}_1(\Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}})\sqrt{\Sigma_{ii}\Sigma_{jj}} & \text{otherwise} \end{cases}
$$

and, more succinctly, $\mathrm{V}\phi(\Sigma) = D^{1/2}\mathrm{V}\phi(D^{-1/2}\Sigma D^{-1/2})D^{1/2}$ where $D = \mathrm{Diag}(\Sigma)$. Here $\mathfrak{J}_1(c) := \frac{1}{\pi}(\sqrt{1-c^2} + (\pi - \arccos(c))c)$ Cho & Saul (2009b).

Matrix simplification. We can simplify the expression $G(I + 2s\Sigma G)^{-1}\Sigma G$, leveraging the fact that G is a projection matrix.

Definition B.6. Let $\mathbb{e}$ be an $B \times (B-1)$ matrix whose columns form an orthonormal basis of $\operatorname{im} G := \{Gv : v \in \mathbb{R}^B\} = \{w \in \mathbb{R}^B : \sum_i w_i = 0\}$. Then the $B \times B$ matrix $\tilde{\mathbb{e}} = (\mathbb{e} | B^{-1/2}\mathbb{1})$ is an orthogonal matrix. For much of this paper $\mathbb{e}$ can be any such basis, but at certain sections we will consider specific realizations of $\mathbb{e}$ for explicit computation.

From easy computations it can be seen that $G = \tilde{\mathbb{e}} \begin{pmatrix} I_{B-1} & 0 \\ 0 & 0 \end{pmatrix} \tilde{\mathbb{e}}^T$. Suppose $\Sigma = \tilde{\mathbb{e}} \begin{pmatrix} \Sigma_\square & v \\ v^T & a \end{pmatrix} \tilde{\mathbb{e}}^T$ where $\Sigma_\square$ is $(B-1) \times (B-1)$, v is a column vector and a is a scalar. Then $\Sigma_\square = \mathbb{e}^T \Sigma \mathbb{e}$ and $\Sigma G = \tilde{\mathbb{e}} \begin{pmatrix} \Sigma_\square & 0 \\ v^T & 0 \end{pmatrix} \tilde{\mathbb{e}}^T$ is block lower triangular, and

$$
\begin{aligned}
(I + 2s\Sigma G)^{-1} &= \tilde{\mathbb{e}} \begin{pmatrix} (I + 2s\Sigma_\square)^{-1} & 0 \\ * & 1 \end{pmatrix} \tilde{\mathbb{e}}^T \\
(I + 2s\Sigma G)^{-1}\Sigma G &= \tilde{\mathbb{e}} \begin{pmatrix} (I + 2s\Sigma_\square)^{-1}\Sigma_\square & 0 \\ * & 0 \end{pmatrix} \tilde{\mathbb{e}}^T \\
G(I + 2s\Sigma G)^{-1}\Sigma G &= \tilde{\mathbb{e}} \begin{pmatrix} (I + 2s\Sigma_\square)^{-1}\Sigma_\square & 0 \\ 0 & 0 \end{pmatrix} \tilde{\mathbb{e}}^T \\
&= \mathbb{e}(I + 2s\Sigma_\square)^{-1}\Sigma_\square \mathbb{e}^T \\
&= G\Sigma G(I + 2sG\Sigma G)^{-1} \\
&=: \Sigma^G (I + 2s\Sigma^G)^{-1}
\end{aligned}
$$

where

Definition B.7. For any matrix Σ, write $\Sigma^G := G\Sigma G$.

Similarly, $\det(I_B + 2s\Sigma G) = \det(I_{B-1} + 2s\Sigma_\square) = \det(I_B + 2s\Sigma^G)$. So, altogether, we have

Theorem B.8. *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is degree-α positive homogeneous. Then for any $B \times (B-1)$ matrix $\mathbb{e}$ whose columns form an orthonormal basis of $\operatorname{im} G := \{Gv : v \in \mathbb{R}^B\} = \{w \in \mathbb{R}^B : \sum_i w_i = 0\}$, with $\Sigma_\square = \mathbb{e}^T\Sigma\mathbb{e}$,*

$$
\mathrm{V}\mathfrak{B}_\phi(\Sigma) = B^\alpha \Gamma(\alpha)^{-1} \int_0^\infty \mathrm{d}s\, s^{\alpha-1} \det(I + 2s\Sigma_\square)^{-1/2} \mathrm{V}\phi(\mathbb{e}(I + 2s\Sigma_\square)^{-1}\Sigma_\square)^{-1}\mathbb{e}^T) \tag{31}
$$

$$
= B^\alpha \Gamma(\alpha)^{-1} \int_0^\infty \mathrm{d}s\, s^{\alpha-1} \det(I + 2s\Sigma^G)^{-1/2} \mathrm{V}\phi((I + 2s\Sigma^G)^{-1}\Sigma^G)^{-1}) \tag{32}
$$

B.1.2 Fourier Method

The Laplace method crucially used the fact that we can pull out the norm factor $\|Gh\|$ out of ϕ, so that we can apply Schwinger parametrization. For general ϕ this is not possible, but we can apply

some wishful thinking and proceed as follows

$$
\begin{aligned}
\mathrm{V}\mathfrak{B}_\phi(\Sigma) &= \mathrm{E}[\phi(\sqrt{B}Gh/\|Gh\|)^{\otimes 2} : h \sim \mathcal{N}(0,\Sigma)] \\
&= \mathrm{E}[\phi(Gh/r)^{\otimes 2} \int_0^\infty \mathrm{d}(r^2)\delta(r^2 - \|Gh\|^2/B) : h \sim \mathcal{N}(0,\Sigma)] \\
&= \mathrm{E}[\phi(Gh/r)^{\otimes 2} \int_0^\infty \mathrm{d}(r^2) \int_{-\infty}^\infty \frac{\mathrm{d}\lambda}{2\pi} e^{i\lambda(r^2 - \|Gh\|^2/B)} : h \sim \mathcal{N}(0,\Sigma)] \\
\text{“}=\text{”} &\int_{\mathbb{R}^B} \frac{\mathrm{d}h}{\sqrt{\det(2\pi\Sigma)}} e^{\frac{-1}{2}h^T\Sigma^{-1}h} \phi(Gh/r)^{\otimes 2} \int_0^\infty \mathrm{d}(r^2) \int_{-\infty}^\infty \frac{\mathrm{d}\lambda}{2\pi} e^{i\lambda(r^2 - \|Gh\|^2/B)} \\
\text{“}=\text{”} &\int_0^\infty \mathrm{d}(r^2) \int_{-\infty}^\infty \frac{\mathrm{d}\lambda}{2\pi} \int_{\mathbb{R}^B} \frac{\mathrm{d}h}{\sqrt{\det(2\pi\Sigma)}} e^{\frac{-1}{2}h^T\Sigma^{-1}h} \phi(Gh/r)^{\otimes 2} e^{i\lambda(r^2 - \|Gh\|^2/B)} \qquad (33) \\
&= \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int_0^\infty \mathrm{d}(r^2) \int_{-\infty}^\infty \frac{\mathrm{d}\lambda}{2\pi} e^{i\lambda r^2} \int_{\mathbb{R}^B} \mathrm{d}h \phi(Gh/r)^{\otimes 2} e^{\frac{-1}{2}(h^T\Sigma^{-1}h + 2i\lambda\|Gh\|^2/B)} \qquad (34) \\
&= \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int_0^\infty \mathrm{d}(r^2) \int_{-\infty}^\infty \frac{\mathrm{d}\lambda}{2\pi} e^{i\lambda r^2} \int_{\mathbb{R}^B} \mathrm{d}h \phi(Gh/r)^{\otimes 2} e^{\frac{-1}{2}(h^T(\Sigma^{-1} + 2i\lambda G/B)h)} \\
\text{“}=\text{”} &\frac{1}{\sqrt{\det(2\pi\Sigma)}} \int_0^\infty \mathrm{d}(r^2) \int_{-\infty}^\infty \frac{\mathrm{d}\lambda}{2\pi} e^{i\lambda r^2} \\
&\sqrt{\det(2\pi(\Sigma^{-1} + 2i\lambda G/B)^{-1})}\, \mathrm{E}[\phi(Gh/r)^{\otimes 2} : h \sim \mathcal{N}(0, \Sigma^{-1} + 2i\lambda G/B)] \\
&= \int_0^\infty \mathrm{d}(r^2) \int_{-\infty}^\infty \frac{\mathrm{d}\lambda}{2\pi} e^{i\lambda r^2} \det(I + 2i\lambda G\Sigma/B)^{-1/2}\, \mathrm{E}[\phi(h)^{\otimes 2} : h \sim \mathcal{N}(0, G(\Sigma^{-1} + 2i\lambda G/B)Gr^{-2}))] \qquad (35) \\
&= \int_0^\infty \mathrm{d}(r^2) \int_{-\infty}^\infty \frac{\mathrm{d}\lambda}{2\pi} e^{i\lambda r^2} \det(I + 2i\lambda G\Sigma/B)^{-1/2} \mathrm{V}\phi(G(\Sigma^{-1} + 2i\lambda G/B)Gr^{-2})
\end{aligned}
$$

Here all steps are legal except for possibly Eq. (33), Eq. (34), and Eq. (35). In Eq. (33), we “Fourier expanded” the delta function — this could possibly be justified as a principal value integral, but this cannot work in combination with Eq. (34) where we have switched the order of integration, integrating over h first. Fubini's theorem would not apply here because $e^{i\lambda(r^2 - \|Gh\|^2/B)}$ has norm 1 and is not integrable. Finally, in Eq. (35), we need to extend the definition of Gaussian to complex covariance matrices, via complex Gaussian integration.

Thus the above derivation is not correct mathematically. However, its end result can be justified rigorously by carefully expressing the delta function as the limit of mollifiers.

Theorem B.9. *Let Σ be a positive-definite $D \times D$ matrix. Let $\phi : \mathbb{R} \to \mathbb{R}$. Suppose for any $\mathsf{b} > \mathsf{a} > 0$,*

1. *$\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)|\phi(z_a/r)\phi(z_b/r)|$ exists and is finite for any $z_a, z_b \in \mathbb{R}$ (i.e. $\phi(z_a/r)\phi(z_b/r)$ is locally integrable in r).*

2. *there exists an $\epsilon > 0$ such that any $\gamma \in [-\epsilon, \epsilon]$ and for each $a, b \in [D]$,*

$$\int_{\|z\|^2 \in [\mathsf{a},\mathsf{b}]} \mathrm{d}z\, e^{-\frac{1}{2}z^T\Sigma^{-1}z} |\phi(z_a/\sqrt{\|z\|^2 + \gamma})\phi(z_b/\sqrt{\|z\|^2 + \gamma})|$$

   exists and is uniformly bounded by some number possibly depending on a and b.

3. *for each $a, b \in [D]$, $\int_{\mathbb{R}^D} \mathrm{d}z \int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2) e^{-\frac{1}{2}z^T\Sigma^{-1}z} |\phi(z_a/r)\phi(z_b/r)|$ exists and is finite.*

4. *$\int_{-\infty}^\infty \mathrm{d}\lambda \left|\det(I + 2i\lambda\Sigma)^{-1/2} \mathrm{V}\phi((\Sigma^{-1} + 2i\lambda I)^{-1}r^{-2})\right|$ exists and is finite for each $r > 0$.*

If $e^{-\frac{1}{2}z^T\Sigma^{-1}z}\phi(z/\|z\|)^{\otimes 2}$ is integrable over $z \in \mathbb{R}^D$, then

$$(2\pi)^{-D/2}\det\Sigma^{-1/2}\int \mathrm{d}z\, e^{-\frac{1}{2}z^T\Sigma^{-1}z}\phi(z/\|z\|)^{\otimes 2}$$
$$= \frac{1}{2\pi}\lim_{\mathsf{a}\searrow 0,\mathsf{b}\nearrow\infty}\int_{\mathsf{a}}^{\mathsf{b}}\mathrm{d}s\int_{-\infty}^{\infty}\mathrm{d}\lambda\, e^{i\lambda s}\det(I+2i\lambda\Sigma)^{-1/2}\mathrm{V}\phi((\Sigma^{-1}+2i\lambda I)^{-1}s^{-1}).$$

Similarly,

$$(2\pi)^{-D/2}\det\Sigma^{-1/2}\int \mathrm{d}z\, e^{-\frac{1}{2}z^T\Sigma^{-1}z}\phi(\sqrt{D}z/\|z\|)^{\otimes 2}$$
$$= \frac{1}{2\pi}\lim_{\mathsf{a}\searrow 0,\mathsf{b}\nearrow\infty}\int_{\mathsf{a}}^{\mathsf{b}}\mathrm{d}s\int_{-\infty}^{\infty}\mathrm{d}\lambda\, e^{i\lambda s}\det(I+2i\lambda\Sigma)^{-1/2}\mathrm{V}\phi((\Sigma^{-1}+2i\lambda I/D)^{-1}s^{-1}).$$

If G is the mean-centering projection matrix and $\Sigma^G = G\Sigma G$, then by the same reasoning as in Thm B.8,

$$\mathrm{V}\mathfrak{B}_\phi(\Sigma) = (2\pi)^{-D/2}\det\Sigma^{-1/2}\int \mathrm{d}z\, e^{-\frac{1}{2}z^T\Sigma^{-1}z}\phi(\sqrt{D}Gz/\|Gz\|)^{\otimes 2}$$
$$= \frac{1}{2\pi}\lim_{\mathsf{a}\searrow 0,\mathsf{b}\nearrow\infty}\int_{\mathsf{a}}^{\mathsf{b}}\mathrm{d}s\int_{-\infty}^{\infty}\mathrm{d}\lambda\, e^{i\lambda s}\det(I+2i\lambda\Sigma^G/D)^{-1/2}\mathrm{V}\phi(\Sigma^G(I+2i\lambda\Sigma^G/D)^{-1}s^{-1}). \tag{36}$$

Note that assumption (1) is satisfied if ϕ is continuous; assumption (4) is satisfied if $D \geq 3$ and for all Π, $\|\mathrm{V}\phi(\Pi)\| \leq \|\Pi\|^\alpha$ for some $\alpha \geq 0$; this latter condition, as well as assumptions (2) and (3), will be satisfied if ϕ is polynomially bounded. Thus the common coordinatewise nonlinearities ReLU, identity, tanh, etc all satisfy these assumptions.

Warning: in general, we cannot swap the order of integration as $\int_{-\infty}^{\infty}\mathrm{d}\lambda\int_{-\infty}^{\infty}\mathrm{d}s$. For example, if $\phi = \mathrm{id}$, then

$$(2\pi)^{-D/2}\det\Sigma^{-1/2}\int \mathrm{d}z\, e^{-\frac{1}{2}z^T\Sigma^{-1}z}\phi(z/\|z\|)^{\otimes 2}$$
$$= \frac{1}{2\pi}\lim_{\mathsf{a}\searrow 0,\mathsf{b}\nearrow\infty}\int_{\mathsf{a}}^{\mathsf{b}}\mathrm{d}s\int_{-\infty}^{\infty}\mathrm{d}\lambda\, e^{i\lambda s}\det(I+2i\lambda\Sigma)^{-1/2}(\Sigma^{-1}+2i\lambda I)^{-1}s^{-1}$$
$$\neq \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}\lambda\int_{0}^{\infty}\mathrm{d}s\, e^{i\lambda s}\det(I+2i\lambda\Sigma)^{-1/2}(\Sigma^{-1}+2i\lambda I)^{-1}s^{-1}$$

because the s-integral in the latter diverges (in a neighborhood of 0).

Proof. We will only prove the first equation; the second follows similarly.

By dominated convergence,

$$\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\phi(z/\|z\|)^{\otimes 2}$$
$$= \lim_{\mathsf{a}\searrow 0,\mathsf{b}\nearrow\infty}\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\phi(z/\|z\|)^{\otimes 2}\mathbb{I}(\|z\|^2 \in [\mathsf{a},\mathsf{b}]).$$

Let $\eta : \mathbb{R} \to \mathbb{R}$ be a nonnegative bump function (i.e. compactly supported and smooth) with support $[-1,1]$ and integral 1 such that $\eta'(0) = \eta''(0) = 0$. Then its Fourier transform $\hat{\eta}(t)$ decays like $O(t^{-2})$. Furthermore, $\eta_\epsilon(x) := \epsilon^{-1}\eta(x/\epsilon)$ is a mollifier, i.e. for all $f \in L^1(\mathbb{R})$, $f * \eta_\epsilon \to f$ in L^1 and pointwise almost everywhere.

Now, we will show that

$$\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\phi(z_a/r)\phi(z_b/r)\mathbb{I}(\|z\|^2 \in [\mathsf{a},\mathsf{b}]) = \lim_{\epsilon\searrow 0}\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}}\mathrm{d}(r^2)\phi(z_a/r)\phi(z_b/r)\eta_\epsilon(\|z\|^2 - r^2)$$

by dominated convergence. Pointwise convergence is immediate, because

$$\begin{aligned}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\phi(z_a/r)\phi(z_b/r)\eta_\epsilon(r^2-\|z\|^2) &= \int_{-\infty}^{\infty} \mathrm{d}(r^2)\phi(z_a/r)\phi(z_b/r)\mathbb{I}(r^2\in[\mathsf{a},\mathsf{b}])\eta_\epsilon(\|z\|^2-r^2)\\ &= [(r^2\mapsto \phi(z_a/r)\phi(z_b/r)\mathbb{I}(r^2\in[\mathsf{a},\mathsf{b}]))*\eta_\epsilon](\|z\|^2)\\ &\to \phi(z_a/\|z\|)\phi(z_b/\|z\|)\mathbb{I}(\|z\|^2\in[\mathsf{a},\mathsf{b}])\end{aligned}$$

as $\epsilon\to 0$ (where we used assumption (1) that $\phi(z_a/r)\phi(z_b/r)\mathbb{I}(r^2\in[\mathsf{a},\mathsf{b}])$ is L^1).

Finally, we construct a dominating integrable function. Observe

$$\begin{aligned}&\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)|\phi(z_a/r)\phi(z_b/r)|\eta_\epsilon(\|z\|^2-r^2)\\ =&\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)|\phi(z_a/r)\phi(z_b/r)|\eta_\epsilon(\|z\|^2-r^2)\\ =&\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(\|z\|^2+\gamma)|\phi(z_a/\sqrt{\|z\|^2+\gamma})\phi(z_b/\sqrt{\|z\|^2+\gamma})|\eta_\epsilon(-\gamma)\\ =&\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}-\|z\|^2}^{\mathsf{b}-\|z\|^2} \mathrm{d}\gamma\, |\phi(z_a/\sqrt{\|z\|^2+\gamma})\phi(z_b/\sqrt{\|z\|^2+\gamma})|\eta_\epsilon(-\gamma)\\ =&\int_{\|z\|^2\in[\mathsf{a}-\epsilon,\mathsf{b}+\epsilon]} \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}-\|z\|^2}^{\mathsf{b}-\|z\|^2} \mathrm{d}\gamma\, |\phi(z_a/\sqrt{\|z\|^2+\gamma})\phi(z_b/\sqrt{\|z\|^2+\gamma})|\eta_\epsilon(-\gamma)\quad \text{since } \operatorname{supp}\eta_\epsilon=[-\epsilon,\epsilon]\\ \le&\int_{\|z\|^2\in[\mathsf{a}-\epsilon,\mathsf{b}+\epsilon]} \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{-\epsilon}^{\epsilon} \mathrm{d}\gamma\, |\phi(z_a/\sqrt{\|z\|^2+\gamma})\phi(z_b/\sqrt{\|z\|^2+\gamma})|\eta_\epsilon(-\gamma)\\ =&\int_{-\epsilon}^{\epsilon} \mathrm{d}\gamma\, \eta_\epsilon(-\gamma)\int_{\|z\|^2\in[\mathsf{a}-\epsilon,\mathsf{b}+\epsilon]} \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}|\phi(z_a/\sqrt{\|z\|^2+\gamma})\phi(z_b/\sqrt{\|z\|^2+\gamma})|\end{aligned}$$

For small enough ϵ then, this is integrable by assumption (2), and yields a dominating integrable function for our application of dominated convergence.

In summary, we have just proven that

$$\begin{aligned}&\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\phi(z/\|z\|)^{\otimes 2}\\ =&\lim_{\mathsf{a}\searrow 0,\mathsf{b}\nearrow\infty}\lim_{\epsilon\searrow 0}\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\phi(z_a/r)\phi(z_b/r)\eta_\epsilon(\|z\|^2-r^2)\end{aligned}$$

Now,

$$\begin{aligned}&\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\phi(z_a/r)\phi(z_b/r)\eta_\epsilon(\|z\|^2-r^2)\\ =&\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\phi(z_a/r)\phi(z_b/r)\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\lambda\, \hat{\eta}_\epsilon(\lambda)e^{-i\lambda(\|z\|^2-r^2)}\\ =&\frac{1}{2\pi}\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\phi(z_a/r)\phi(z_b/r)\int_{-\infty}^{\infty} \mathrm{d}\lambda\, \hat{\eta}(\epsilon\lambda)e^{-i\lambda(\|z\|^2-r^2)}\end{aligned}$$

Note that the absolute value of the integral is bounded above by

$$\begin{aligned}&\frac{1}{2\pi}\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)|\phi(z_a/r)\phi(z_b/r)|\int_{-\infty}^{\infty} \mathrm{d}\lambda\, |\hat{\eta}(\epsilon\lambda)e^{-i\lambda(\|z\|^2-r^2)}|\\ =&\frac{1}{2\pi}\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)|\phi(z_a/r)\phi(z_b/r)|\int_{-\infty}^{\infty} \mathrm{d}\lambda\, |\hat{\eta}(\epsilon\lambda)|\\ \le& C\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)|\phi(z_a/r)\phi(z_b/r)|\end{aligned}$$

for some C, by our construction of η that $\hat{\eta}(t) = O(t^{-2})$ for large $|t|$. By assumption (3), this integral exists. Therefore we can apply the Fubini-Tonelli theorem and swap order of integration.

$$
\begin{aligned}
&\frac{1}{2\pi}\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\phi(z_a/r)\phi(z_b/r)\int_{-\infty}^{\infty}\mathrm{d}\lambda\,\hat{\eta}(\epsilon\lambda)e^{-i\lambda(\|z\|^2-r^2)}\\
&=\frac{1}{2\pi}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\int_{-\infty}^{\infty}\mathrm{d}\lambda\,\hat{\eta}(\epsilon\lambda)e^{-i\lambda(\|z\|^2-r^2)}\int \mathrm{d}z\phi(z_a/r)\phi(z_b/r)e^{-\frac{1}{2}z^T\Sigma^{-1}z}\\
&=\frac{1}{2\pi}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\int_{-\infty}^{\infty}\mathrm{d}\lambda\,\hat{\eta}(\epsilon\lambda)e^{i\lambda r^2}\int \mathrm{d}z\phi(z_a/r)\phi(z_b/r)e^{-\frac{1}{2}z^T(\Sigma^{-1}+2i\lambda I)z}\\
&=\frac{1}{2\pi}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)r^D\int_{-\infty}^{\infty}\mathrm{d}\lambda\,\hat{\eta}(\epsilon\lambda)e^{i\lambda r^2}\int \mathrm{d}z\phi(z_a)\phi(z_b)e^{-\frac{1}{2}r^2z^T(\Sigma^{-1}+2i\lambda I)z}\\
&=(2\pi)^{D/2-1}\det\Sigma^{1/2}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\int_{-\infty}^{\infty}\mathrm{d}\lambda\,\hat{\eta}(\epsilon\lambda)e^{i\lambda r^2}\det(I+2i\lambda\Sigma)^{-1/2}\mathrm{V}\phi((\Sigma^{-1}+2i\lambda I)^{-1}r^{-2})
\end{aligned}
$$

By assumption (4),

$$
\begin{aligned}
&\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\int_{-\infty}^{\infty}\mathrm{d}\lambda\,\left|\hat{\eta}(\epsilon\lambda)e^{i\lambda r^2}\det(I+2i\lambda\Sigma)^{-1/2}\mathrm{V}\phi((\Sigma^{-1}+2i\lambda I)^{-1}r^{-2})\right|\\
&\le C\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\int_{-\infty}^{\infty}\mathrm{d}\lambda\,\left|\det(I+2i\lambda\Sigma)^{-1/2}\mathrm{V}\phi((\Sigma^{-1}+2i\lambda I)^{-1}r^{-2})\right|
\end{aligned}
$$

is integrable (where we used the fact that $\hat{\eta}$ is bounded — which is true for all $\eta \in C_0^\infty(\mathbb{R})$), so by dominated convergence (applied to $\epsilon \searrow 0$),

$$
\begin{aligned}
&\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\phi(z/\|z\|)^{\otimes 2}\\
&=\lim_{\mathsf{a}\searrow 0,\mathsf{b}\nearrow\infty}\lim_{\epsilon\searrow 0}\int \mathrm{d}z e^{-\frac{1}{2}z^T\Sigma^{-1}z}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\phi(z_a/r)\phi(z_b/r)\eta_\epsilon(\|z\|^2-r^2)\\
&=(2\pi)^{D/2-1}\det\Sigma^{1/2}\lim_{\mathsf{a}\searrow 0,\mathsf{b}\nearrow\infty}\int_{\mathsf{a}}^{\mathsf{b}} \mathrm{d}(r^2)\int_{-\infty}^{\infty}\mathrm{d}\lambda\,e^{i\lambda r^2}\det(I+2i\lambda\Sigma)^{-1/2}\mathrm{V}\phi((\Sigma^{-1}+2i\lambda I)^{-1}r^{-2})
\end{aligned}
$$

where we used the fact that $\hat{\eta}(0) = \int_{\mathbb{R}}\eta(x)\,\mathrm{d}x = 1$ by construction. This gives us the desired result after putting back in some constants and changing $r^2 \mapsto s$. □

B.2 Global Convergence

Basic questions regarding the dynamics Eq. (28) are 1) does it converge? 2) What are the limit points? 3) How fast does it converge? Here we answer these questions definitively when $\phi = \mathrm{id}$.

The $\phi = \mathrm{id}$ case. First we prove a lemma

Lemma B.10. *Consider the dynamics $\Sigma^{(l)} = \mathrm{E}[(h/\|h\|)^{\otimes 2} : h \sim \mathcal{N}(0, \Sigma^{(l-1)})]$ on $\Sigma^{(l)} \in \mathcal{S}_A$. Suppose $\Sigma^{(0)}$ is full rank. Then*

1. $\lim_{l\to\infty}\Sigma^{(l)} = \frac{1}{A}I$.
2. *This convergence is exponential in the sense that, for any full rank $\Sigma^{(0)}$, there is a constant $K < 1$ such that $\lambda_1(\Sigma^{(l)}) - \lambda_A(\Sigma^{(l)}) < K(\lambda_1(\Sigma^{(l-1)}) - \lambda_A(\Sigma^{(l-1)}))$ for all $l \ge 2$. Here λ_1 (resp. λ_A) denotes that largest (resp. smallest) eigenvalue.*
3. *Asymptotically, $\lambda_1(\Sigma^{(l)}) - \lambda_A(\Sigma^{(l)}) = O((1-\frac{2}{A+2})^l)$.*

Proof. Let $\lambda_1^{(l)} \ge \lambda_2^{(l)} \ge \cdots \ge \lambda_A^{(l)}$ be the eigenvalues of $\Sigma^{(l)}$. It's easy to see that $\sum_i \lambda_i^{(l)} = 1$ for all $l \ge 1$. So WLOG we assume $l \ge 1$ and this equality holds.

We will show the "exponential convergence" statement; that $\lim_{l\to\infty}\Sigma^{(l)} = \frac{1}{A}I$ then follows from the trace condition above.

Proof of Item 2. Using the overbar/underbar notation for brevity, we will now compute the eigenvalues $\overline{\lambda}_1 \geq \cdots \geq \overline{\lambda}_A$ of $\overline{\Sigma}$.

First, notice that $\overline{\Sigma}$ and Σ can be simultaneously diagonalized, and by induction all of $\{\Sigma^{(l)}\}_{l\geq 0}$ can be simultaneous diagonalized. Thus we will WLOG assume that Σ is diagonal, $\Sigma = \mathrm{Diag}(\lambda_1, \ldots, \lambda_A)$, so that $\overline{\Sigma} = \mathrm{Diag}(\gamma_1, \ldots, \gamma_A)$ for some $\{\gamma_i\}_i$. These $\{\gamma_i\}_i$ form the eigenvalues of $\overline{\Sigma}$ but a priori we don't know whether they fall in decreasing order; in fact we will soon see that they do, and $\gamma_i = \overline{\lambda}_i$.

We have

$$
\begin{aligned}
\gamma_i &= \left(\frac{\pi}{2}\right)^{-A/2} \int_{0^A}^{\infty^A} \frac{\lambda_i x_i^2}{\sum_j \lambda_j x_j^2} e^{-\|x\|^2/2} \,\mathrm{d}x \\
&= \left(\frac{\pi}{2}\right)^{-A/2} \int_{0^A}^{\infty^A} \lambda_i x_i^2 e^{-\|x\|^2/2} \,\mathrm{d}x \int_0^\infty e^{-s\sum_j \lambda_j x_j^2} \,\mathrm{d}s \\
&= \left(\frac{\pi}{2}\right)^{-A/2} \int_{0^{A+1}}^{\infty^{A+1}} \lambda_i x_i^2 e^{-\frac{1}{2}\sum_j (1+2s\lambda_j) x_j^2} \,\mathrm{d}x \,\mathrm{d}s \\
&= \left(\frac{\pi}{2}\right)^{-A/2} \int_{0^2}^{\infty^2} \lambda_i x_i^2 e^{-\frac{1}{2}(1+2s\lambda_i)x_i^2} \cdot \left(\frac{\pi}{2}\right)^{\frac{A-1}{2}} \prod_{j\neq i} (1+2s\lambda_j)^{-1/2} \,\mathrm{d}x_i \,\mathrm{d}s \\
&= \left(\frac{\pi}{2}\right)^{-1/2} \int_{0^2}^{\infty^2} \lambda_i x_i^2 e^{-\frac{1}{2}(1+2s\lambda_i)x_i^2} \,\mathrm{d}x_i \prod_{j\neq i} (1+2s\lambda_j)^{-1/2} \,\mathrm{d}s \\
&= \int_0^\infty \lambda_i (1+2s\lambda_i)^{-3/2} \prod_{j\neq i} (1+2s\lambda_j)^{-1/2} \,\mathrm{d}s \\
&= \int_0^\infty \prod_{j=1}^A (1+2s\lambda_j)^{-1/2} \lambda_i (1+2s\lambda_i)^{-1} \,\mathrm{d}s
\end{aligned} \tag{37}
$$

Therefore,

$$
\begin{aligned}
\gamma_i - \gamma_k &= \int_0^\infty \prod_{j=1}^A (1+2s\lambda_j)^{-1/2} (\lambda_i (1+2s\lambda_i)^{-1} - \lambda_k (1+2s\lambda_k)^{-1}) \,\mathrm{d}s \\
&= (\lambda_i - \lambda_k) \int_0^\infty \prod_{j=1}^A (1+2s\lambda_j)^{-1/2} (1+2s\lambda_i)^{-1} (1+2s\lambda_k)^{-1} \,\mathrm{d}s
\end{aligned}
$$

Since the RHS integral is always positive, $\lambda_i \geq \lambda_k \implies \gamma_i \geq \gamma_k$ and thus $\gamma_i = \overline{\lambda}_i$ for each i.

Define $T(\lambda) := \int_0^\infty \prod_{j=1}^A (1+2s\lambda_1)^{-1/2} (1+2s\lambda_1)^{-1} (1+2s\lambda_A)^{-1} \,\mathrm{d}s$, so that $\overline{\lambda}_1 - \overline{\lambda}_A = (\lambda_1 - \lambda_A) T(\lambda)$.

Note first that $\prod_{j=1}^A (1+2s\lambda_j)^{-1/2} (1+2s\lambda_i)^{-1} (1+2s\lambda_k)^{-1}$ is (strictly) log-convex and hence (strictly) convex in λ. Furthermore, $T(\lambda)$ is (strictly) convex because it is an integral of (strictly) convex functions. Thus T is maximized over any convex region by its extremal points, and only by its extremal points because of strict convexity.

The convex region we are interested in is given by

$$
\mathcal{A} := \{(\lambda_i)_i : \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_A \geq 0 \ \ \& \ \sum_i \lambda_i = 1\}.
$$

The extremal points of $\mathcal{A}$ are $\{\omega^k := (\overbrace{1/k, 1/k, \cdots, 1/k}^{k \text{ times}}, 0, 0, \cdots, 0)\}_{k=1}^{A}$. For $k = 1, \ldots, A-1$, we have

$$\begin{aligned} T(\omega^k) &= \int_0^\infty (1 + 2s/k)^{-k/2-1} \, \mathrm{d}s \\ &= \frac{2/k}{-k/2}(1 + 2s/k)^{-k/2}\Big|_0^\infty \\ &= 1. \end{aligned}$$

But for $k = A$,

$$\begin{aligned} T(\omega^A) &= \int_0^\infty (1 + 2s/A)^{-A/2-2} \, \mathrm{d}s \\ &= \frac{2/A}{-A/2-1}(1 + 2s/A)^{-k/2}\Big|_0^\infty \\ &= \frac{A}{A+2} \end{aligned}$$

This shows that $T(\lambda) \leq 1$, with equality iff $\lambda = \omega^k$ for $k = 1, \ldots, A-1$. In fact, because every point $\lambda \in \mathcal{A}$ is a convex combination of ω^k for $k = 1, \ldots, A$, $\lambda = \sum_{k=1}^{A} a_k \omega^k$, by convexity of T, we must have

$$\begin{aligned} T(\lambda) &\leq \sum_{k=1}^{A} a_k T(\omega^k) \\ &= 1 - a_A \frac{2}{A+2} \\ &= 1 - \lambda_A \frac{2}{A+2} \end{aligned} \tag{38}$$

where the last line follows because ω^A is the only point with last coordinate nonzero so that $a_A = \lambda_A$.

We now show that the gap $\lambda_1^{(l)} - \lambda_A^{(l)} \to 0$ as $l \to \infty$. There are two cases: if $\lambda_A^{(l)}$ is bounded away from 0 infinitely often, then $T(\lambda) < 1 - \epsilon$ infinitely often for a fixed $\epsilon > 0$ so that the gap indeed vanishes with l. Now suppose otherwise, that $\lambda_A^{(l)}$ converges to 0; we will show this leads to a contradiction. Notice that

$$\begin{aligned} \overline{\lambda}_A &= \int_0^\infty \prod_{j=1}^{A} (1 + 2s\lambda_j)^{-1/2} \lambda_A (1 + 2s\lambda_A)^{-1} \, \mathrm{d}s \\ \overline{\lambda}_A / \lambda_A &\geq \int_0^\infty \prod_{j=1}^{A} (1 + 2s(1-\lambda_A)/(A-1))^{-(A-1)/2} (1 + 2s\lambda_A)^{-3/2} \, \mathrm{d}s \end{aligned}$$

where the first lines is Eq. (37) and the 2nd line follows from the convexity of $\prod_{j=1}^{A-1}(1+2s\lambda_j)^{-1/2}$ as a function of $(\lambda_1, \ldots, \lambda_{A-1})$. By a simple application of dominated convergence, as $\lambda_A \to 0$, this integral converges to a particular simple form,

$$\begin{aligned} \overline{\lambda}_A / \lambda_A &\geq \int_0^\infty \prod_{j=1}^{A} (1 + 2s/(A-1))^{-(A-1)/2} \, \mathrm{d}s \\ &= 1 + \frac{2}{A-3} \end{aligned}$$

Thus for large enough l, $\lambda_A^{(l+1)}/\lambda_A^{(l)}$ is at least $1 + \epsilon$ for some $\epsilon > 0$, but this contradicts the convergence of λ_A to 0.

Altogether, this proves that $\lambda_1^{(l)} - \lambda_A^{(l)} \to 0$ and therefore $\lambda^{(l)} \to \omega^A$ as $l \to \infty$. Consequently, $\lambda_A^{(l)}$ is bounded from below, say by K' (where $K' > 0$ because by Eq. (37), $\lambda_A^{(l)}$ is never 0), for all l,

and by Eq. (38), we prove Item 2 by taking $K = 1 - K'\frac{2}{A+2}$. In addition, asymptotically, the gap decreases exponentially as $T(\omega^A)^l = \left(\frac{A}{A+2}\right)^l$, proving Item 3. □

Theorem B.11. *Consider the dynamics $\Sigma^{(l)} = \mathrm{E}[(h/\|h\|)^{\otimes 2} : h \sim \mathcal{N}(0, \Sigma^{(l-1)})]$ on $\Sigma^{(l)} \in \mathcal{S}_A$. Suppose $\Sigma^{(0)} = M^T D M$ where M is orthogonal and D is a diagonal matrix. If D has rank C and $D_{ii} \neq 0, \forall 1 \leq i \leq C$, then*

1. *$\lim_{l\to\infty} \Sigma^{(l)} = \frac{1}{C} M^T D' M$ where D' is the diagonal matrix with $D'_{ii} = \mathbb{I}(D_{ii} \neq 0)$.*
2. *This convergence is exponential in the sense that, for any $\Sigma^{(0)}$ of rank C, there is a constant $K < 1$ such that $\lambda_1(\Sigma^{(l)}) - \lambda_C(\Sigma^{(l)}) < K(\lambda_1(\Sigma^{(l-1)}) - \lambda_C(\Sigma^{(l-1)}))$ for all $l \geq 2$. Here λ_i denotes the ith largest eigenvalue.*
3. *Asymptotically, $\lambda_1(\Sigma^{(l)}) - \lambda_C(\Sigma^{(l)}) = O((1 - \frac{2}{C+2})^l)$.*

Proof. Note that $\Sigma^{(l)}$ can always be simultaneously diagonalized with $\Sigma^{(0)}$, so that $\Sigma^{(l)} = M^T D^{(l)} M$ for some diagonal $D^{(l)}$ which has $D^{(l)}_{ii} = 0, \forall 0 \leq i \leq C$. Then we have $(D^{(l)})' = \mathrm{E}[(h/\|h\|)^{\otimes 2} : h \sim \mathcal{N}(0, (D^{(l-1)})')]$, where $(D^{(l)})'$ means the diagonal matrix obtained from $D^{(l)}$ by deleting the dimensions with zero eigenvalues. The proof then finishes by Lemma B.10. □

From this it easily follows the following characterization of the convergence behavior.

Corollary B.12. *Consider the dynamics of Eq. (28) for $\phi = \mathrm{id}$: $\Sigma^{(l)} = B\,\mathrm{E}[(Gh/\|Gh\|)^{\otimes 2} : h \sim \mathcal{N}(0, \Sigma^{(l-1)})]$ on $\Sigma^{(l)} \in \mathcal{S}_B$. Suppose $G\Sigma G$ has rank $C < B$ and factors as $\hat{\mathbb{e}} D \hat{\mathbb{e}}^T$ where $D \in \mathbb{R}^{C\times C}$ is a diagonal matrix with no zero diagonal entries and $\hat{\mathbb{e}}$ is an $B \times C$ matrix whose columns form an orthonormal basis of a subspace of $\operatorname{im} G$. Then*

1. *$\lim_{l\to\infty} \Sigma^{(l)} = \frac{B}{C} \hat{\mathbb{e}} I_C \hat{\mathbb{e}}^T$.*
2. *This convergence is exponential in the sense that, for any $G\Sigma^{(0)}G$ of rank C, there is a constant $K < 1$ such that $\lambda_1(\Sigma^{(l)}) - \lambda_C(\Sigma^{(l)}) < K(\lambda_1(\Sigma^{(l-1)}) - \lambda_C(\Sigma^{(l-1)}))$ for all $l \geq 2$. Here λ_i denotes the ith largest eigenvalue.*
3. *Asymptotically, $\lambda_1(\Sigma^{(l)}) - \lambda_C(\Sigma^{(l)}) = O((1 - \frac{2}{C+2})^l)$.*

General ϕ case. We don't (currently) have a proof of any characterization of the basin of attraction for Eq. (28) for general ϕ. Thus we are forced to resort to finding its fixed points manually and characterize their local convergence properties.

B.3 Limit points

Batchnorm $\mathfrak{B}_\phi$ is permutation-equivariant, in the sense that $\mathfrak{B}_\phi(\pi h) = \pi \mathfrak{B}_\phi(h)$ for any permtuation matrix π. Along with the case of $\phi = \mathrm{id}$ studied above, this suggests that we look into fixed points Σ^* that are invariant under permutation. These are matrices of the form

Definition B.13. We say a matrix $\Sigma \in \mathcal{S}_B$ is *BSB1* (short for "1-Step Batch Symmetry Breaking" or "1-Block Symmetry Breaking") if Σ has one common entry on the diagonal and one common entry on the off-diagonal, i.e.

$$\begin{pmatrix} a & b & b & \cdots \\ b & a & b & \cdots \\ b & b & a & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

We will denote such a matrix as $\mathrm{BSB1}(a, b)$. Note that $\mathrm{BSB1}(a, b)$ can be written as $(a-b)I + b\mathbb{1}\mathbb{1}^T$. Its spectrum is given below.

Lemma B.14. *A $B \times B$ matrix of the form $\upsilon I + \nu \mathbb{1}\mathbb{1}^T$ has two eigenvalues υ and $B\nu + \upsilon$, each with eigenspaces $\{x : \sum_i x_i = 0\}$ and $\{x : x_1 = \cdots = x_B\}$. Equivalently, if it has a on the diagonal and b on the off-diagonal, then the eigenvalues are $a - b$ and $(B-1)b + a$.*

The following simple lemma will also be very useful

Lemma B.15. *Let* $\Sigma := \mathrm{BSB1}_B(a, b)$. *Then* $G\Sigma G = (a - b)G$.

Proof. G and $\mathrm{BSB1}_B(a, b)$ can be simultaneously diagonalized by Lemma B.14. Note that G zeros out the eigenspace $\mathbb{R}\mathbb{1}$, and is identity on its orthogonal complement. The result then follows from easy computations. □

What are the BSB1 fixed points of *Eq.* (28)? We will exhibit two method in this section to compute fixed points of BSB1 form. The spherical coordinates method works for any nonlinearity ϕ and reduces the computation of such fixed points to 2-dimensional integrals. On the other hand, the Laplace method sheds more light for positive-homogenous nonlinearities (like ReLU), where we can give closed form solutions for the fixed points.

B.3.1 Spherical Coordinates Method

The main result (Thm B.18) of this section is an expression of the BSB1 fixed point diagonal and off-diagonal entries in terms of 1- and 2-dimensional integrals. This allows one to numerically compute such fixed points.

By a simple symmetry argument, we have the following

Lemma B.16. *Suppose* X *is a random vector in* $\mathbb{R}^B$*, symmetric in the sense that for any permutation matrix* π *and any subset* U *of* $\mathbb{R}^B$ *measurable with respect to the distribution of* X*,* $P(X \in U) = P(X \in \pi(U))$*. Let* $\Phi : \mathbb{R}^B \to \mathbb{R}^B$ *be a symmetric function in the sense that* $\Phi(\pi x) = \pi\Phi(x)$ *for any permutation matrix* π*. Then* $\mathrm{E}\,\Phi(X)\Phi(X)^T = \upsilon I + \nu\mathbb{1}\mathbb{1}^T$ *for some* υ *and* ν*.*

Then

Lemma B.17. *Suppose* Y *is a random Gaussian vector in* $\mathbb{R}^B$ *with zero mean and covariance matrix* $\underline{\Sigma} := \underline{\upsilon} I + \underline{\nu}\mathbb{1}\mathbb{1}^T$*. Then* $\mathrm{E}\,\mathfrak{B}_\phi(Y)^{\otimes 2} = \mathrm{E}\,\mathfrak{B}_\phi(Y)\mathfrak{B}_\phi(Y)^T = \upsilon I + \nu\mathbb{1}\mathbb{1}^T$ *where*

$$\upsilon + \nu = (2\pi\underline{\upsilon})^{\frac{1-B}{2}} \int_{\mathbb{R}^{B-1}} \phi(\sqrt{B} z_1(x_{B-1})/\|x\|)^2 e^{-\|x\|^2/2\underline{\upsilon}}\,\mathrm{d}x$$
$$\nu = (2\pi\underline{\upsilon})^{\frac{1-B}{2}} \int_{\mathbb{R}^{B-1}} \phi(\sqrt{B} z_1(x_{B-1})/\|x\|)\phi(\sqrt{B} z_2(x_{B-1}, x_{B-2})/\|x\|) e^{-\|x\|^2/2\underline{\upsilon}}\,\mathrm{d}x$$

where $z_1(x_{B-1}) = -\sqrt{\frac{B-1}{B}} x_{B-1}$ *and* $z_2(x_{B-1}, x_{B-2}) = \frac{1}{\sqrt{B(B-1)}} x_{B-1} - \sqrt{\frac{B-2}{B-1}} x_{B-2}$.

Proof. By Lemma B.16 and the permutation symmetry of $\mathfrak{B}_\phi$ and Y, the expected outer product has the form $\upsilon I + \nu\mathbb{1}\mathbb{1}^T$ for some υ and ν, satisfying $\upsilon + \nu = \mathrm{E}\,\mathfrak{B}_\phi(Y)_i^2$, $\nu = \mathrm{E}\,\mathfrak{B}_\phi(Y)_i\mathfrak{B}_\phi(Y)_j$ for any $i \neq j$. So it suffices to show that these expectations, for some i and j, are equal to the integrals in the statement of the theorem.

By Lemma B.14, G has eigenvalues 1 and 0 respectively with eigenspaces $V_0 := \{y : \sum_i y_i = 0\}$ and $V_1 := \{y : y_1 = \cdots = y_B\}$, and Σ has eigenvalues $\underline{\upsilon}$ and $\underline{\upsilon} + B\underline{\nu}$ with the same eigenspaces. Since the linear image of a Gaussian is a Gaussian, $Z := GY$ is normally distributed with covariance $G\Sigma G$; in other words, Z is isotropic on V_0, with variance $\underline{\upsilon}$.

Define $\mathbb{e}_i := (i(i+1))^{-1/2}(\overbrace{1, 1, \ldots, 1}^{i}, -i, 0, \ldots, 0)^T$, for each $i = 1, \ldots, B-1$. Then $\{\mathbb{e}_i\}_i$ forms an orthonormal basis of V_0. Let $x \in \mathbb{R}^{B-1}$ be the random coordinates of V_0 in this basis, so that $Z = \mathbb{e}X$, where $\mathbb{e}$ is the matrix with $\{\mathbb{e}_i\}$ as columns. A short computation shows that $Z_B = -\sqrt{\frac{B-1}{B}} X_{B-1} = z_1(X_{B-1})$ and $Z_{B-1} = \frac{1}{\sqrt{B(B-1)}} X_{B-1} - \sqrt{\frac{B-2}{B-1}} X_{B-2} = z_2(X_{B-1}, X_{B-2})$.

Thus

$$\begin{aligned}
\mathrm{E}\,\mathfrak{B}_\phi(Y)_B^2 &= \mathrm{E}\,\phi(\sqrt{B}Z_B/\|Z\|)^2 \\
&= \mathrm{E}\,\phi(\sqrt{B}z_1(X_{B-1})/\|X\|)^2 \\
\mathrm{E}\,\mathfrak{B}_\phi(Y)_B\mathfrak{B}_\phi(Y)_{B-1} &= \mathrm{E}\,\phi(\sqrt{B}Z_B/\|Z\|)\phi(\sqrt{B}Z_{B-1}/\|Z\|) \\
&= \mathrm{E}\,\phi(\sqrt{B}z_1(X_{B-1})/\|X\|)\phi(\sqrt{B}z_2(X_{B-1},X_{B-2})/\|X\|)
\end{aligned}$$

which correspond to the integrals in the theorem statement, after noting that X is isotropic on $\mathbb{R}^B$ with variance $\underline{\upsilon}$ because it is related to Z by an orthogonal linear transformation.

□

Note that, by rescaling, the integrals in Lemma B.17 actually do not depend on $\underline{\upsilon}$ and $\underline{\nu}$. Thus they yield the fixed point BSB1 Σ^*. We can simplify the integrals further into 1- and 2-dimensions, which give the following

Theorem B.18. *The BSB1 fixed point $\Sigma^* = \upsilon^* I + \nu^* \mathbb{1}\mathbb{1}^T$ to Eq. (28) satisfies*

$$\begin{aligned}
\upsilon^* + \nu^* &= \frac{\Gamma(\frac{B-1}{2})}{\Gamma(\frac{B-2}{2})\sqrt{\pi}} \int_0^\pi \mathrm{d}\theta_1 \sin^{B-3}\theta_1 \phi(\sqrt{B}\zeta_1(\theta_1))^2 \\
\nu^* &= \begin{cases} \frac{B-3}{2\pi}\int_0^\pi \mathrm{d}\theta_1 \int_0^\pi \mathrm{d}\theta_2 \sin^{B-3}\theta_1 \sin^{B-4}\theta_2 \phi(\sqrt{B}\zeta_1(\theta_1))\phi(\sqrt{B}\zeta_2(\theta_1,\theta_2)), & \text{if } B \geq 4 \\ \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\theta \phi(-\sqrt{2}\sin\theta)\phi(\sin\theta/\sqrt{2} - \cos\theta\sqrt{3}/\sqrt{2}), & \text{if } B = 3 \end{cases}
\end{aligned}$$

where

$$\begin{aligned}
\zeta_1(\theta) &= z_1(\cos\theta) = -\sqrt{\frac{B-1}{B}}\cos\theta \\
\zeta_2(\theta_1,\theta_2) &= z_2(\cos\theta_1, \sin\theta_1\cos\theta_2) = \frac{1}{\sqrt{B(B-1)}}\cos\theta_1 - \sqrt{\frac{B-2}{B-1}}\sin\theta_1\cos\theta_2
\end{aligned}$$

with z_1, and z_2 as defined in Lemma B.17,

Proof. We change the integrals in Lemma B.17 to spherical coordinates $(r, \theta_1, \ldots, \theta_{B-2}) \in [0,\infty) \times [0,\pi]^{B-3} \times [0, 2\pi]$, defined by

$$\begin{aligned}
x_{B-1} &= r\cos\theta_1 \\
x_{B-2} &= r\sin\theta_1\cos\theta_2 \\
x_{B-3} &= r\sin\theta_1\sin\theta_2\cos\theta_3 \\
&\vdots \\
x_2 &= r\sin\theta_1\cdots\sin\theta_{B-3}\cos\theta_{B-2} \\
x_1 &= r\sin\theta_1\cdots\sin\theta_{B-3}\sin\theta_{B-2}.
\end{aligned}$$

Then

$$\mathrm{d}x_1\ldots\mathrm{d}x_{B-1} = r^{B-2}\sin^{B-3}\theta_1\sin^{B-4}\theta_2\cdots\sin\theta_{B-3}\,\mathrm{d}r\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_{B-2}.$$

and $z_1(x_{B-1})/\|x\| = z_1(\cos\theta_1)$ and $z_2(x_{B-1}, x_{B-2})\|x\|^{-1} = z_2(\cos\theta_1, \sin\theta_1\cos\theta_2)$. Thus,

$$\begin{aligned}
\upsilon + \nu &= (2\pi\underline{\upsilon})^{\frac{1-B}{2}} \int_0^\infty r^{B-2}e^{-r^2/2\underline{\upsilon}}\,\mathrm{d}r \int_{\mathcal{A}} \phi(\sqrt{B}z_1)^2 \sin^{B-3}\theta_1\cdots\sin\theta_{B-3}\,\mathrm{d}r\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_{B-2} \\
\nu &= (2\pi\underline{\upsilon})^{\frac{1-B}{2}} \int_0^\infty r^{B-2}e^{-r^2/2\underline{\upsilon}}\,\mathrm{d}r \int_{\mathcal{A}} \phi(\sqrt{B}z_1)\phi(\sqrt{B}z_2) \sin^{B-3}\theta_1\cdots\sin\theta_{B-3}\,\mathrm{d}r\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_{B-2}
\end{aligned}$$

where $\mathcal{A} = [0,\pi]^{B-3} \times [0, 2\pi]$. By Lemma B.21 and Lemma B.22 (below), for $B \geq 4$, we can integrate out $r, \theta_3, \ldots, \theta_{B-2}$:

$$\upsilon^* + \nu^* = (2\pi\underline{\upsilon})^{\frac{1-B}{2}} \times 2^{\frac{B-3}{2}} \underline{\upsilon}^{\frac{B-1}{2}} \Gamma\left(\frac{B-1}{2}\right) \times \pi^{\frac{B-4}{2}+1} \Gamma\left(\frac{B-2}{2}\right)^{-1} (2\pi) \int_0^\pi \mathrm{d}\theta_1 \sin^{B-3}\theta_1 \phi(\sqrt{B}z_1)^2$$

$$= \frac{\Gamma(\frac{B-1}{2})}{\Gamma(\frac{B-2}{2})\sqrt{\pi}} \int_0^\pi \mathrm{d}\theta_1 \sin^{B-3}\theta_1 \phi(\sqrt{B}z_1)^2$$

$$\nu^* = (2\pi\underline{\upsilon})^{\frac{1-B}{2}} \times 2^{\frac{B-3}{2}} \underline{\upsilon}^{\frac{B-1}{2}} \Gamma\left(\frac{B-1}{2}\right) \times \pi^{\frac{B-5}{2}+1} \Gamma\left(\frac{B-3}{2}\right)^{-1} (2\pi) \times$$

$$\int_0^\pi \mathrm{d}\theta_1 \int_0^\pi \mathrm{d}\theta_2 \sin^{B-3}\theta_1 \sin^{B-4}\theta_2 \phi(\sqrt{B}z_1)\phi(\sqrt{B}z_2)$$

$$= \frac{B-3}{2\pi} \int_0^\pi \mathrm{d}\theta_1 \int_0^\pi \mathrm{d}\theta_2 \sin^{B-3}\theta_1 \sin^{B-4}\theta_2 \phi(\sqrt{B}z_1)\phi(\sqrt{B}z_2)$$

For $B = 3$ we can similarly obtain the result for ν^*. □

Lemma B.19. *For $j, k \geq 0$ and $0 \leq s \leq t \leq \pi/2$,*

$$\int_s^t \sin^j\theta \cos^k\theta \,\mathrm{d}\theta = \frac{1}{2}\,\mathrm{Beta}\left(\cos^2 t, \cos^2 s; \frac{k+1}{2}, \frac{j+1}{2}\right).$$

By antisymmetry of cos *with respect to $\theta \mapsto \pi - \theta$, if $\pi/2 \leq s \leq t \leq \pi$,*

$$\int_s^t \sin^j\theta \cos^k\theta \,\mathrm{d}\theta = \frac{1}{2}(-1)^k\,\mathrm{Beta}\left(\cos^2 s, \cos^2 t; \frac{k+1}{2}, \frac{j+1}{2}\right).$$

Proof. Set $x := \cos^2\theta \implies \mathrm{d}x = -2\cos\theta\sin\theta\,\mathrm{d}\theta$. So the integral in question is

$$\frac{-1}{2}\int_{\cos^2 s}^{\cos^2 t} (1-x)^{\frac{j-1}{2}} x^{\frac{k-1}{2}}\,\mathrm{d}x$$

$$= \frac{1}{2}(\mathrm{Beta}(\cos^2 s; \frac{k+1}{2}, \frac{j+1}{2}) - \mathrm{Beta}(\cos^2 t; \frac{k+1}{2}, \frac{j+1}{2}))$$

□

As consequences,

Lemma B.20. *For $j, k \geq 0$, $\int_0^\pi \sin^j\theta\cos^k\theta\,\mathrm{d}\theta = \frac{1+(-1)^k}{2}\mathrm{Beta}(\frac{j+1}{2}, \frac{k+1}{2}) = \frac{1+(-1)^k}{2}\frac{\Gamma(\frac{1+j}{2})\Gamma(\frac{1+k}{2})}{\Gamma(\frac{2+j+k}{2})}$.*

Lemma B.21.

$$\int_0^\pi \sin\theta_1\,\mathrm{d}\theta_1 \int_0^\pi \sin^2\theta_2\,\mathrm{d}\theta_2 \cdots \int_0^\pi \sin^k\theta_k\,\mathrm{d}\theta_k = \pi^{k/2}/\Gamma(\frac{k+2}{2}).$$

Proof. By Lemma B.20, $\int_0^\pi \sin^r\theta_r\,\mathrm{d}\theta_r = \mathrm{Beta}(\frac{r+1}{2}, \frac{1}{2}) = \frac{\Gamma(\frac{r+1}{2})\sqrt{\pi}}{\Gamma(\frac{r+2}{2})}$. Thus this product of integrals is equal to

$$\prod_{r=1}^k \sqrt{\pi}\frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r+2}{2})} = \pi^{k/2}\Gamma(1)/\Gamma(\frac{k+2}{2}) = \pi^{k/2}/\Gamma(\frac{k+2}{2}).$$

□

Lemma B.22. *For $B > 1, \upsilon > 0$, $\int_0^\infty r^B e^{-r^2/2\upsilon}\,\mathrm{d}r = 2^{\frac{B-1}{2}}\upsilon^{\frac{B+1}{2}}\Gamma(\frac{B+1}{2})$*

Proof. Apply change of coordinates $z = r^2/2\upsilon$. □

Case of ReLU. One can obtain an analytical solution of the fixed point for ReLU by playing with the integrals above. We present such a derivation in the appendix, but the Laplace method yields a much quicker and cleaner way to do so (Thm B.25).

B.3.2 LAPLACE METHOD

Definition B.23. For any $\alpha \geq 0$, define $K_{\alpha,B} := \mathsf{c}_\alpha \mathrm{P}\left(\frac{B-1}{2}, \alpha\right)^{-1} \left(\frac{B-1}{2}\right)^\alpha$ where $\mathrm{P}(a,b) := \Gamma(a+b)/\Gamma(a)$ is the Pochhammer symbol.

Note that

Proposition B.24. $\lim_{B\to\infty} K_{\alpha,B} = \mathsf{c}_\alpha$.

Theorem B.25. *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is degree α positive-homogeneous. For any BSB1 $\Sigma \in \mathcal{S}_B$, $\mathrm{V}\mathfrak{B}_\phi(\Sigma)$ is BSB1. The diagonal entries are $K_{\alpha,B}\mathfrak{J}_\phi(1)$ and the off-diagonal entries are $K_{\alpha,B}\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)$. Here c_α is as defined in Defn H.2 and $\mathfrak{J}_\phi$ is as defined in Defn I.4. Thus a BSB1 fixed point of Eq. (28) exists and is unique.*

Proof. Let $\mathbb{e}$ be an $B \times (B-1)$ matrix whose columns form an orthonormal basis of $\operatorname{im} G := \{Gv : v \in \mathbb{R}^B\} = \{w \in \mathbb{R}^B : \sum_i w_i = 0\}$. Let $\mathrm{BSB1}(a,b)$ denote a matrix with a on the diagonal and b on the off-diagnals. Note that $\mathrm{V}\phi$ is positive homogeneous of degree α, so for any v,

$$
\begin{aligned}
&\mathrm{V}\phi(\mathbb{e} v I_{B-1}(I_{B-1} + 2svI_{B-1})^{-1}\mathbb{e}^T) \\
&= \mathrm{V}\phi(\frac{v}{1+2sv}\mathbb{e}\mathbb{e}^T) \\
&= \left(\frac{v}{1+2sv}\right)^\alpha \mathrm{V}\phi(\mathbb{e}\mathbb{e}^T) && \text{by } \textit{Proposition I.8} \\
&= \left(\frac{v}{1+2sv}\right)^\alpha \mathrm{V}\phi(G) \\
&= \left(\frac{v}{1+2sv}\frac{B-1}{B}\right)^\alpha \mathrm{V}\phi\left(\mathrm{BSB1}\left(1, \frac{-1}{B-1}\right)\right) \\
&= \left(\frac{v}{1+2sv}\frac{B-1}{B}\right)^\alpha \mathsf{c}_\alpha\mathfrak{J}_\phi\left(\mathrm{BSB1}\left(1, \frac{-1}{B-1}\right)\right) && \text{by Corollary I.3}
\end{aligned}
$$

So by Eq. (31),

$$
\begin{aligned}
\mathrm{V}\mathfrak{B}_\phi(\mathbb{e} v I \mathbb{e}^T) &= B^\alpha \Gamma(\alpha)^{-1} \int_0^\infty \mathrm{d}s\, s^{\alpha-1}(1+2sv)^{-(B-1)/2} \left(\frac{v}{1+2sv}\frac{B-1}{B}\right)^\alpha \mathsf{c}_\alpha\mathfrak{J}_\phi\left(\mathrm{BSB1}\left(1, \frac{-1}{B-1}\right)\right) \\
&= \Gamma(\alpha)^{-1}(B-1)^\alpha \mathsf{c}_\alpha\mathfrak{J}_\phi\left(\mathrm{BSB1}\left(1, \frac{-1}{B-1}\right)\right) \int_0^\infty \mathrm{d}s\, s^{\alpha-1}(1+2sv)^{-(B-1)/2-\alpha} v^\alpha \\
&= \Gamma(\alpha)^{-1}(B-1)^\alpha \mathsf{c}_\alpha\mathfrak{J}_\phi\left(\mathrm{BSB1}\left(1, \frac{-1}{B-1}\right)\right) \mathrm{Beta}(\alpha, (B-1)/2) 2^{-\alpha} \\
&= \mathsf{c}_\alpha \frac{\Gamma((B-1)/2)}{\Gamma(\alpha + (B-1)/2)} \left(\frac{B-1}{2}\right)^\alpha \mathfrak{J}_\phi\left(\mathrm{BSB1}\left(1, \frac{-1}{B-1}\right)\right)
\end{aligned}
$$

□

Corollary B.26. *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is degree α positive-homogeneous. If Σ^* is the BSB1 fixed point of Eq. (28) as given by Thm B.25, then $G\Sigma^* G = v^* I_{B-1}$ where $v^* = K_{\alpha,B}\big(\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi(-1/(B-1))\big)$*

By setting $\alpha = 1$ and $\phi = \rho$, we get

Corollary B.27. *For any BSB1 $\Sigma \in \mathcal{S}_B$, $\mathrm{V}\mathfrak{B}_\rho(\Sigma)$ is BSB1 with diagonal entries $\frac{1}{2}$ and off-diagonal entries $\frac{1}{2}\mathfrak{J}_1\left(\frac{-1}{B-1}\right)$, so that $G^{\otimes 2}\{\mathrm{V}\mathfrak{B}_\rho(\Sigma)\} = G(\mathrm{V}\mathfrak{B}_\rho(\Sigma))G = \left(\frac{1}{2} - \frac{1}{2}\mathfrak{J}_1\left(\frac{-1}{B-1}\right)\right) G$.*

By setting $\alpha = 1$ and $\phi = \mathrm{id} = x \mapsto \rho(x) - \rho(-x)$, we get

Corollary B.28. *For any BSB1 $\Sigma \in \mathcal{S}_B$, $\mathrm{V}\mathfrak{B}_{\mathrm{id}}(\Sigma)$ is BSB1 with diagonal entries 1 and off-diagonal entries $\frac{-1}{B-1}$, so that $G^{\otimes 2}\{\mathrm{V}\mathfrak{B}_{\mathrm{id}}(\Sigma)\} = B/(B-1)G$.*

Remark B.29. One might hope to tweak the Laplace method for computing the fixed point to work for the Fourier method, but because there is no nice relation between $\mathrm{V}\phi(c\Sigma)$ and $\mathrm{V}\phi(\Sigma)$ in general, we cannot simplify Eq. (36) as we can Eq. (31) and Eq. (32).

B.4 LOCAL CONVERGENCE

In this section we consider linearization of the dynamics given in Eq. (28). Thus we must consider linear operators on the space of PSD linear operators $\mathcal{S}_B$. To avoid confusion, we use the following notation: If $\mathcal{T} : \mathcal{S}_B \to \mathcal{S}_B$ (for example the Jacobian of $\mathrm{V}\mathfrak{B}_\phi$) and $\Sigma \in \mathcal{S}_B$, then write $\mathcal{T}\{\Sigma\}$ for the image of Σ under $\mathcal{T}$.

A priori, the Jacobian of $\mathrm{V}\mathfrak{B}_\phi$ at its BSB1 fixed point may seem like a very daunting object. But because of the BSB1 symmetries, the corresponding Jacobian will also have many symmetries that significantly simplify its analysis. We formalize such symmetries as below.

Definition B.30. Let $\mathcal{T} : \mathbb{R}^{B\times B} \to \mathbb{R}^{B\times B}$ be a linear operator. Let $\delta_i \in \mathbb{R}^B$ be the vector with 0 everywhere except 1 in coordinate i; then $\delta_i\delta_j^T$ is the matrix with 0 everywhere except 1 in position (i, j). Write $[kl|ij] := \mathcal{T}(\delta_i\delta_j^T)_{kl}$. Suppose $\mathcal{T}$ has the property that for all $i, j, k, l \in [B] = \{1, \dots, B\}$

- $[kl|ij] = [\pi(k)\pi(l)|\pi(i)\pi(j)]$ for all permutation π on $[B]$, and
- $[ij|kl] = [ji|lk]$.

Then we say $\mathcal{T}$ is *ultrasymmetric*.

Remark B.31. In what follows, we will often "normalize" the representation "$[ij|kl]$" to the unique "$[i'j'|k'l']$" that is in the same equivalence class according to Defn B.30 and such that $i', j', k', l' \in [4]$ and $i' \le j', k' \le l'$ unless $i' = l', j' = k'$, in which case the normalization is $[12|21]$. Explicitly, we have the following equivalence classes and their normalized representations

class	repr.
$i = j = k = l$	$[11\|11]$
$i = j = k$ or $i = j = l$	$[11\|12]$
$i = j$ and $k = l$	$[11\|22]$
$i = j$	$[11\|23]$
$i = k = l$ or $j = k = l$	$[12\|11]$
$i = k$ and $j = l$	$[12\|12]$
$i = k$ or $i = l$	$[12\|13]$
$i = l$ and $j = k$	$[12\|21]$
$k = l$	$[12\|33]$
all different	$[12\|34]$

It's straightforward to verify that

Proposition B.32. *The Jacobian* $\frac{\mathrm{d}\mathrm{V}\mathfrak{B}_\phi}{\mathrm{d}\Sigma}\big|_\Sigma$ *is ultrasymmetric for any BSB1* Σ.

The Jacobian of $\mathrm{V}\mathfrak{B}_\phi$ will be a linear operator from $\mathcal{H}_B$ to $\mathcal{H}_B$ where

Definition B.33. Denote by $\mathcal{H}_B$ the space of symmetric matrices of dimension B. Also write $\mathcal{H}_B^G$ for the space of symmetric matrices Σ of dimension B such that $G\Sigma G = \Sigma$ (which is equivalent to saying rows of Σ sum up to 0).

As in the case of $\mathcal{S}$, we omit subscript B when it's clear from context.

Definition B.34. Let $\mathbb{L}_B := \{GDG : D \text{ diagonal}, \operatorname{tr} D = 0\} \subseteq \mathcal{H}_B^G$ and $\mathbb{M}_B := \{\Sigma \in \mathcal{H}_B^G : \mathrm{Diag}\Sigma = 0\}$. Note that $\dim \mathbb{L}_B = B - 1, \dim \mathbb{M}_B = \frac{B(B-3)}{2}$ and $\mathbb{R}G^B \oplus \mathbb{L}_B \oplus \mathbb{M}_B = \mathcal{H}_B^G$ is an orthogonal decomposition w.r.t Frobenius inner product.

Definition B.35. For any nonzero $a, b \in \mathbb{R}$, set

$$\mathrm{L}_B(a, b) := \begin{pmatrix} a & 0 & -b & -b & \cdots \\ 0 & -a & b & b & \cdots \\ -b & b & 0 & 0 & \cdots \\ -b & b & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \in \mathcal{H}_B.$$

In general, we say a matrix M is $\mathrm{L}_B(a,b)$-shaped if $M = P\mathrm{L}_B(a,b)P^T$ for some permutation matrix P.

It's not hard to see that $\mathrm{L}(B-2,1)$-shaped matrices span $\mathbb{L}_B$.

Lemma B.36. *Let $L = \mathrm{L}_B(a,b)$. Then $G^{\otimes 2}\{L\} = GLG = \frac{a+2b}{B}\mathrm{L}_B(B-2,1)$.*

Proof. GLG can be written as the sum of outer products

$$\begin{aligned}
GLG &= aG_1 \otimes G_1 - aG_2 \otimes G_2 + \sum_{i=3}^{B} -bG_1 \otimes G_i + bG_2 \otimes G_i - bG_i \otimes G_1 + bG_i \otimes G_2 \\
&= a\mathrm{L}_B\left(\left(\frac{B-1}{B}\right)^2 - \frac{1}{B^2}, \frac{B}{B^2}\right) + b\sum_{i=3}^{B}(-\delta_1+\delta_2)\otimes G_i + G_i \otimes (-\delta_1+\delta_2) \\
&= \frac{a}{B}\mathrm{L}_B(B-2,1) + b((-\delta_1+\delta_2)\otimes v + v \otimes (-\delta_1+\delta_2)) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{with } v = \left(-\frac{B-2}{B}, -\frac{B-2}{B}, \frac{2}{B}, \frac{2}{B}, \dots\right) \\
&= \frac{a}{B}\mathrm{L}_B(B-2,1) + \frac{2b}{B}\mathrm{L}_B(B-2,1) \\
&= \frac{a+2b}{B}\mathrm{L}_B(B-2,1)
\end{aligned}$$

□

Lemma B.37. *Let $\mathcal{T} : \mathbb{R}^{B\times B} \to \mathbb{R}^{B\times B}$ be an ultrasymmetric linear operator. Then $\mathcal{T}$ has the following eigendecomposition.*

1. *Two 1-dimensional eigenspace $\mathbb{R}\cdot\mathrm{BSB1}(\lambda^{\mathcal{T}}_{\mathrm{BSB1},i} - \alpha_{22}, \alpha_{21})$ with eigenvalue $\lambda^{\mathcal{T}}_{\mathrm{BSB1},i}$ for $i = 1,2$, where*

$$\begin{aligned}
\alpha_{11} &= [11|11] + [11|22](B-1) \\
\alpha_{12} &= 2(B-1)[11|12] + (B-2)(B-1)[11|23] \\
\alpha_{21} &= 2[12|11] + (B-2)[12|33] \\
\alpha_{22} &= [12|12] + 4(B-2)[12|13] + [12|21] + (B-2)(B-3)[12|34]
\end{aligned}$$

   and $\lambda^{\mathcal{T}}_{\mathrm{BSB1},1}$ and $\lambda^{\mathcal{T}}_{\mathrm{BSB1},2}$ are the roots to the quadratic

$$x^2 - (\alpha_{11}+\alpha_{22})x + \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}.$$

2. *Two $(B-1)$-dimensional eigenspaces $\mathfrak{S}_B \cdot \mathrm{L}(\lambda^{\mathcal{T}}_{\mathbf{L},i} - \beta_{22}, \beta_{21})$ with eigenvalue $\lambda^{\mathcal{T}}_{\mathbf{L},i}$ for $i = 1,2$. Here $\mathfrak{S}_B \cdot W$ denotes the linear span of the orbit of matrix W under simultaneous permutation of its column and rows (by the same permutation), and*

$$\begin{aligned}
\beta_{11} &= [11|11] - [11|22] \\
\beta_{12} &= 2(B-2)([11|23] - [11|12]) \\
\beta_{21} &= -[12|11] + [12|33] \\
\beta_{22} &= [12|21] + [12|12] + 2(B-4)[12|13] - 2(B-3)[12|34].
\end{aligned}$$

   and $\lambda^{\mathcal{T}}_{\mathbf{L},1}$ and $\lambda^{\mathcal{T}}_{\mathbf{L},2}$ are the roots to the quadratic

$$x^2 - (\beta_{11}+\beta_{22})x + \beta_{11}\beta_{22} - \beta_{12}\beta_{21}.$$

3. *Eigenspace $\mathbb{M}$ (dimension $B(B-3)/2$) with eigenvalue $\lambda^{\mathcal{T}}_{\mathbb{M}}$.*

The proof is by careful, but ultimately straightforward, computation.

Proof. We will use the bracket notation of Defn B.30 to denote entries of $\mathcal{T}$, and implicitly simplify it according to Remark B.31.

Item 1. Let $U \in \mathbb{R}^{B\times B}$ be the BSB1 matrix . By ultrasymmetry of $\mathcal{T}$ and BSB1 symmetry of A, $\mathcal{T}\{U\}$ is also BSB1. So we proceed to calculate the diagonal and off-diagonal entries of $\mathcal{T}\{G\}$.

We have

$$\begin{aligned}
\mathcal{T}\{\mathrm{BSB1}(a,b)\}_{11} &= [11|11]a + 2(B-1)[11|12]b + [11|22](B-1)a + [11|23](B-2)(B-1)b \\
&= ([11|11] + [11|22](B-1))a + (2(B-1)[11|12] + (B-2)(B-1)[11|23])b \\
\mathcal{T}\{\mathrm{BSB1}(a,b)\}_{12} &= [12|12]b + 2[12|11]a + (B-2)[12|33]a + 2(B-2)[12|13]b \\
&\quad + [12|21]b + 2(B-2)[12|23]b + (B-2)(B-3)[12|34]b \\
&= (2[12|11] + (B-2)[12|33])a \\
&\quad + ([12|12] + 2(B-2)[12|13] + [12|21] + 2(B-2)[12|13] + (B-2)(B-3)[12|34])b \\
&= (2[12|11] + (B-2)[12|33])a \\
&\quad + ([12|12] + 4(B-2)[12|13] + [12|21] + (B-2)(B-3)[12|34])b
\end{aligned}$$

Thus $\mathrm{BSB1}(\omega_1,\gamma_1)$ and $\mathrm{BSB1}(\omega_2,\gamma_2)$ are the eigenmatrices of $\mathcal{T}$, where (ω_1,γ_1) and (ω_2,γ_2) are the eigenvectors of the matrix

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}$$

with

$$\begin{aligned}
\alpha_{11} &= [11|11] + [11|22](B-1) \\
\alpha_{12} &= 2(B-1)[11|12] + (B-2)(B-1)[11|23] \\
\alpha_{21} &= 2[12|11] + (B-2)[12|33] \\
\alpha_{22} &= [12|12] + 4(B-2)[12|13] + [12|21] + (B-2)(B-3)[12|34]
\end{aligned}$$

The eigenvalues are the two roots $\lambda^{\mathcal{T}}_{\mathrm{BSB1},1}, \lambda^{\mathcal{T}}_{\mathrm{BSB1},2}$ to the quadratic

$$x^2 - (\alpha_{11} + \alpha_{22})x + \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}$$

and the corresponding eigenvectors are

$$\begin{aligned}
(\omega_1,\gamma_1) &= (\lambda_1 - \alpha_{22}, \alpha_{21}) \\
(\omega_2,\gamma_2) &= (\lambda_2 - \alpha_{22}, \alpha_{21}).
\end{aligned}$$

Item 2. We will study the image of $\mathrm{L}_B(a,b)$ (Defn B.35) under $\mathcal{T}$. We have

$$\begin{aligned}
\mathcal{T}\{\mathrm{L}(a,b)\}_{11} &= -\mathcal{T}\{\mathrm{L}(a,b)\}_{22} \\
&= [11|11]a + 2(B-2)[11|12](-b) + [11|22](-a) + 2(B-2)[11|23]b \\
&= ([11|11] - [11|22])a + 2(B-2)([11|23] - [11|12])b \\
\mathcal{T}\{\mathrm{L}(a,b)\}_{12} &= \mathcal{T}\{\mathrm{L}(a,b)\}_{21} \\
&= [12|12]0 + [12|21]0 + [12|11](a-a) + [12|13](b-b) + [12|31](b-b) + [12|33]0 + [12|34]0 \\
&= 0 \\
\mathcal{T}\{\mathrm{L}(a,b)\}_{33} &= \mathcal{T}\{\mathrm{L}(a,b)\}_{ii}, \forall i \geq 3 \\
&= [11|11]0 + [11|12](2b-2b) + [11|22](a-a) + [11|23](2(B-3)b - 2(B-3)b) \\
&= 0 \\
\mathcal{T}\{\mathrm{L}(a,b)\}_{34} &= \mathcal{T}\{\mathrm{L}(a,b)\}_{ij}, \forall i \neq j \ \ \& \ \ i,j \geq 3 \\
&= [12|12]0 + [12|11]0 + [12|13](2b-2b) + [12|21]0 \\
&\quad + [12|31](2b-2b) + [12|33](a-a) + [12|34](2(B-4)b - 2(B-4)b) \\
&= 0 \\
\mathcal{T}\{\mathrm{L}(a,b)\}_{13} &= \mathcal{T}\{\mathrm{L}(a,b)\}_{1j}, \forall j \geq 3 \\
&= \mathcal{T}\{\mathrm{L}(a,b)\}_{j1}, \forall j \geq 3 \\
&= [12|12](-b) + [12|13](B-3-1)(-b) + [12|11]a + [12|21](-b) \\
&\quad + [12|31](B-3-1)(-b) + [12|33](-a) + [12|34]2(B-3)b \\
&= \big([12|11] - [12|33]\big)a + \big(2(B-3)[12|34] - 2(B-4)[12|13] - [12|12] - [12|21]\big)b
\end{aligned}$$

Thus $\mathrm{L}(a,b)$ transforms under $\mathcal{T}$ by the matrix

$$\mathrm{L}(a,b) \mapsto \mathrm{L}(\begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix})$$

with

$$\begin{aligned}
\beta_{11} &= [11|11] - [11|22] \\
\beta_{12} &= 2(B-2)([11|23] - [11|12]) \\
\beta_{21} &= -[12|11] + [12|33] \\
\beta_{22} &= [12|21] + [12|12] + 2(B-4)[12|13] - 2(B-3)[12|34].
\end{aligned}$$

So if $\lambda^{\mathcal{T}}_{\mathbf{L},1}$ and $\lambda^{\mathcal{T}}_{\mathbf{L},2}$ are the roots of the equation

$$x^2 - (\beta_{11} + \beta_{22})x + \beta_{11}\beta_{22} - \beta_{12}\beta_{21}$$

then

$$\begin{aligned}
\mathcal{T}\{\mathrm{L}(\lambda^{\mathcal{T}}_{\mathbf{L},1} - \beta_{22}, \beta_{21})\} &= \lambda^{\mathcal{T}}_{\mathbf{L},1}\mathrm{L}(\lambda^{\mathcal{T}}_{\mathbf{L},1} - \beta_{22}, \beta_{21}) \\
\mathcal{T}\{\mathrm{L}(\lambda^{\mathcal{T}}_{\mathbf{L},2} - \beta_{22}, \beta_{21})\} &= \lambda^{\mathcal{T}}_{\mathbf{L},2}\mathrm{L}(\lambda^{\mathcal{T}}_{\mathbf{L},2} - \beta_{22}, \beta_{21})
\end{aligned}$$

Similarly, any image of these eigenvectors under simultaneous permutation of rows and columns remains eigenvectors with the same eigenvalue. This derives Item 2.

Item 3. Let $M \in \mathbb{M}$. We first show that $\mathcal{T}\{M\}$ has zero diagonal. We have

$$\begin{aligned}
\mathcal{T}\{M\}_{11} &= [11|12](\sum_{i=2}^{B} M_{1i} + M_{i1}) + [11|23](\sum_{i,j=1}^{B} M_{ij} - \left(\sum_{i=2}^{B} M_{1i} + M_{i1}\right)) \\
&= 0 + 0 = 0
\end{aligned}$$

which follows from $M\mathbb{1} = 0$ by definition of $\mathbb{M}$. Similarly $\mathcal{T}\{M\}_{ii} = 0$ for all i.

Now we show that M is an eigenmatrix.

$$\begin{aligned}
\mathcal{T}\{M\}_{12} &= [12|12]M_{12} + [12|11]0 + [12|33]0 + [12|13]\left(\sum_{i=3}^{B} M_{1i} + \sum_{i=1}^{B} M_{i2}\right) \\
&\quad + [12|21]M_{21} + [12|31]\left(\sum_{i=3}^{B} M_{i1} + \sum_{i=1}^{B} M_{2i}\right) + [12|34]\sum_{i\geq 3, j\geq 3} M_{ij} \\
&= [12|12]M_{12} - 2M_{12}[12|13] + M_{21}[12|21] + [12|31](-2M_{12}) + [12|34](\sum_{i=3}^{B} -M_{i1} - M_{1i}) \\
&= M_{12}([12|12] - 2[12|13] + [12|21] - 2[12|31] + 2[12|34]) \\
&= M_{12}([12|12] + [12|21] - 4[12|13] + 2[12|34]) \\
&= \lambda^{\mathcal{T}}_{\mathbb{M}} M_{12}
\end{aligned}$$

Similarly $\mathcal{T}\{M\}_{ij} = \lambda^{\mathcal{T}}_{\mathbb{M}} M_{ij}$ for all $i \neq j$.

□

Lemma B.38. *Let $\mathcal{T} : \mathbb{R}^{B\times B} \to \mathbb{R}^{B\times B}$ be a linear operator. We write $G^{\otimes 2} \circ \mathcal{T} \restriction \mathcal{H}^G_B$ for the operator $\Sigma \mapsto \mathcal{T}\{\Sigma\} \mapsto G(\mathcal{T}\{\Sigma\})G$, restricted to $\Sigma \in \mathcal{H}^G_B$. Then $G^{\otimes 2} \circ \mathcal{T} \restriction \mathcal{H}^G_B$ has the following eigendecomposition.*

1. *Eigenspace $\mathbb{R}G$ with eigenvalue $\lambda^{G,\mathcal{T}}_{\mathbb{R}G} := B^{-1}((B-1)(\alpha_{11} - \alpha_{21}) - (\alpha_{12} - \alpha_{22}))$, where as in Lemma B.37,*

$$\begin{aligned}
\alpha_{11} &= [11|11] + [11|22](B-1) \\
\alpha_{12} &= 2(B-1)[11|12] + (B-2)(B-1)[11|23] \\
\alpha_{21} &= 2[12|11] + (B-2)[12|33] \\
\alpha_{22} &= [12|12] + 4(B-2)[12|13] + [12|21] + (B-2)(B-3)[12|34]
\end{aligned}$$

2. *Eigenspace* $\mathbb{L}$ *with eigenvalue* $\lambda_{\mathbb{L}}^{G,\mathcal{T}} := B^{-1}((B-2)\beta_{11} + \beta_{12} + 2(B-2)\beta_{21} + 2\beta_{22})$, *where as in Lemma B.37,*

$$\begin{aligned}
\beta_{11} &= [11|11] - [11|22] \\
\beta_{12} &= 2(B-2)([11|23] - [11|12]) \\
\beta_{21} &= -[12|11] + [12|33] \\
\beta_{22} &= [12|21] + [12|12] + 2(B-4)[12|13] - 2(B-3)[12|34].
\end{aligned}$$

3. *Eigenspace* $\mathbb{M}$ *with eigenvalue* $\lambda_{\mathbb{M}}^{G,\mathcal{T}} := \lambda_{\mathbb{M}}^{\mathcal{T}} = [12|12] + [12|21] - 4[12|13] + 2[12|34]$.

Proof. Item 1 As in the proof of Lemma B.37, we find

$$\mathcal{T}\{\mathrm{BSB1}(a,b)\} = \mathrm{BSB1}\left(\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}\right)$$

where

$$\begin{aligned}
\alpha_{11} &= [11|11] + [11|22](B-1) \\
\alpha_{12} &= 2(B-1)[11|12] + (B-2)(B-1)[11|23] \\
\alpha_{21} &= 2[12|11] + (B-2)[12|33] \\
\alpha_{22} &= [12|12] + 4(B-2)[12|13] + [12|21] + (B-2)(B-3)[12|34].
\end{aligned}$$

For $a = B-1, b = -1$ so that $\mathrm{BSB1}(B-1,-1) = BG$, we get

$$\begin{aligned}
\mathcal{T}\{\mathrm{BSB1}(B-1,-1)\} &= \mathrm{BSB1}\left((B-1)\alpha_{11} - \alpha_{12}, (B-1)\alpha_{21} - \alpha_{22}\right) \\
G^{\otimes 2} \circ \mathcal{T}\{\mathrm{BSB1}(B-1,-1)\} &= G\,\mathrm{BSB1}\left((B-1)\alpha_{11} - \alpha_{12}, (B-1)\alpha_{21} - \alpha_{22}\right)\, G \\
&= \left((B-1)(\alpha_{11} - \alpha_{21}) - (\alpha_{12} - \alpha_{22})\right) G \\
&= B^{-1}\left((B-1)(\alpha_{11} - \alpha_{21}) - (\alpha_{12} - \alpha_{22})\right) \mathrm{BSB1}(B-1,-1) \\
&= \lambda_{\mathbb{R}G}^{G,\mathcal{T}} \mathrm{BSB1}(B-1,-1)
\end{aligned}$$

by Lemma B.15.

Item 2. It suffices to show that $\mathrm{L}(B-2,1)$ is an eigenmatrix with the eigenvalue $\lambda_{\mathbb{L}}^{G,\mathcal{T}}$.

As in the proof of Lemma B.37, we find

$$\mathcal{T}\{\mathrm{L}(a,b)\} = \mathrm{L}\left(\begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}\right)$$

where

$$\begin{aligned}
\beta_{11} &= [11|11] - [11|22] \\
\beta_{12} &= 2(B-2)([11|23] - [11|12]) \\
\beta_{21} &= -[12|11] + [12|33] \\
\beta_{22} &= [12|21] + [12|12] + 2(B-4)[12|13] - 2(B-3)[12|34].
\end{aligned}$$

So with $a = B-2, b = 1$, we have

$$\begin{aligned}
\mathcal{T}\{\mathrm{L}(B-2,1)\} &= \mathrm{L}((B-2)\beta_{11} + \beta_{12}, (B-2)\beta_{21} + \beta_{22}) \\
G^{\otimes 2} \circ \mathcal{T}\{\mathrm{L}(B-2,1)\} &= G\,\mathrm{L}((B-2)\beta_{11} + \beta_{12}, (B-2)\beta_{21} + \beta_{22})\, G \\
&= B^{-1}((B-2)\beta_{11} + \beta_{12} + 2((B-2)\beta_{21} + \beta_{22}))\mathrm{L}(B-2,1) \\
&= \lambda_{\mathbb{L}}^{G,\mathcal{T}} \mathrm{L}(B-2,1)
\end{aligned}$$

by Lemma B.36.

Item 3. The proof is exactly the same as in that of Lemma B.37. □

Noting that the eigenspaces of Lemma B.38 are orthogonal, we have

Proposition B.39. *$G^{\otimes 2} \circ \mathcal{H}^G$ for any ultrasymmetric operator $\mathcal{T}$ is self-adjoint.*

Now we prepare several helper reults in order to make progress understanding batchnorm.

Definition B.40. Define $\mathfrak{N}(x) = \sqrt{B}x/\|x\|$, i.e. division by sample standard deviation.

Batchnorm $\mathfrak{B}_\phi$ can be decomposed as the composition of three steps, $\phi \circ \mathfrak{N} \circ G$, where G is mean-centering, $\mathfrak{N}$ is division by standard deviation, and ϕ is coordinate-wise application of nonlinearity. We have, as operators $\mathcal{H} \to \mathcal{H}$,

$$\begin{aligned}
\frac{\mathrm{dV}\mathfrak{B}_\phi(\Sigma)}{\mathrm{d}\Sigma} &= \frac{\mathrm{d}\,\mathrm{E}[\mathfrak{B}_\phi(z)^{\otimes 2} : z \in \mathcal{N}(0, \Sigma)]}{\mathrm{d}\Sigma} \\
&= \frac{\mathrm{d}\,\mathrm{E}[(\phi \circ \mathfrak{N})(x)^{\otimes 2} : x \in \mathcal{N}(0, G\Sigma G)]}{\mathrm{d}\Sigma} \\
&= \frac{\mathrm{dV}(\phi \circ \mathfrak{N})(\Sigma^G)}{\mathrm{d}\Sigma^G} \circ \frac{\mathrm{d}\Sigma^G}{\mathrm{d}\Sigma} \\
&= \frac{\mathrm{dV}(\phi \circ \mathfrak{N})(\Sigma^G)}{\mathrm{d}\Sigma^G} \circ G^{\otimes 2}
\end{aligned}$$

Definition B.41. Write $\mathcal{U} := G^{\otimes 2} \circ \frac{\mathrm{dV}(\phi \circ \mathfrak{N})(\Sigma^G)}{\mathrm{d}\Sigma^G}\Big|_{\Sigma^G = G\Sigma^* G} : \mathcal{H}^G \to \mathcal{H}^G$.

It turns out to be advantageous to study $\mathcal{U}$ first and relate its eigendecomposition back to that of $\frac{\mathrm{dV}\mathfrak{B}_\phi(\Sigma)}{\mathrm{d}\Sigma}\Big|_{\Sigma = \Sigma^*}$, where Σ^* is the BSB1 fixed point, by applying Lemma B.42.

Lemma B.42. *Let X and Y be two vector spaces. Suppose linear operators $A : X \to Y, B : Y \to X$. Then*

1. $\operatorname{rank} AB = \operatorname{rank} BA$
2. *If $v \in Y$ is an eigenvector of AB with nonzero eigenvalue, then $X \ni Bv \neq 0$ and Bv is an eigenvector of BA of the same eigenvalue*
3. *Suppose AB has $k = \operatorname{rank} AB$ linearly independent eigenvectors $\{v_i\}_{i=1}^k$ of nonzero eigenvalues $\{\lambda_i\}_{i=1}^k$. Then BA has k linearly independent eigenvectors $\{Bv_i\}_{i=1}^k$ with the same eigenvalues $\{\lambda_i\}_{i=1}^k$ which are all eigenvectors of BA with nonzero eigenvalues, up to linear combinations within eigenvectors with the same eigenvalue.*

With $A = G^{\otimes 2}$ and $B = \frac{\mathrm{dV}(\phi \circ \mathfrak{N})(\Sigma^G)}{\mathrm{d}\Sigma^G}$, Lemma B.37 implies that AB and BA can both be diagonalized, and this lemma implies that all nonzero eigenvalues of $\frac{\mathrm{dV}\mathfrak{B}_\phi(\Sigma)}{\mathrm{d}\Sigma}$ can be recovered from those of $\mathcal{U}$.

Proof. (Item 1) Observe $\operatorname{rank} AB = \operatorname{rank} ABAB \leq \operatorname{rank} BA$. By symmetry the two sides are in fact equal.

(Item 2) Bv cannot be zero or otherwise $ABv = A0 = 0$, contradicting the fact that v is an eigenvector with nonzero eigenvalue. Suppose λ is the eigenvalue associated to v. Then $BA(Bv) = B(ABv) = B(\lambda v) = \lambda Bv$, so Bv is an eigenvector of BA with the same eigenvalue.

(Item 3) Item 2 shows that $\{Bv_i\}_{i=1}^k$ are eigenvectors BA with the same eigenvalues $\{\lambda_i\}_{i=1}^k$. The eigenspaces with different eigenvalues are linearly independent, so it suffices to show that if $\{Bv_{i_j}\}_j$ are eigenvectors of the same eigenvalue λ_s, then they are linearly independent. But $\sum_j a_j Bv_{i_j} = 0 \implies \sum_j a_j v_{i_j} = 0$ because B is injective on eigenvectors by Item 2, so that $a_j = 0$ identically. Hence $\{Bv_j\}_j$ is linearly independent.

Since $\operatorname{rank} BA = k$, these are all of the eigenvectors with nonzero eigenvalues of BA up to linearly combinations.

□

Lemma B.43. *Let $f : \mathbb{R}^B \to \mathbb{R}^A$ be measurable, and $\Sigma \in \mathcal{S}_B$ be invertible. Then for any $\Lambda \in \mathbb{R}^{B\times B}$,*

$$\frac{\mathrm{d}}{\mathrm{d}\Sigma}\,\mathrm{E}[f(z) : z \sim \mathcal{N}(0,\Sigma)]\{\Lambda\} = \frac{1}{2}\,\mathrm{E}[f(z)\left\langle \Sigma^{-1}zz^T\Sigma^{-1} - \Sigma^{-1}, \Lambda\right\rangle : z \sim \mathcal{N}(0,\Sigma)]$$

If f is in addition twice-differentiable, then

$$\frac{\mathrm{d}}{\mathrm{d}\Sigma}\,\mathrm{E}[f(z) : z \sim \mathcal{N}(0,\Sigma)] = \frac{1}{2}\,\mathrm{E}\left[\frac{\mathrm{d}^2 f(z)}{\mathrm{d}^2 z} : z \sim \mathcal{N}(0,\Sigma)\right]$$

whenever both sides exist.

Proof. Let $\Sigma_t, t \in (-\epsilon,\epsilon)$ be a smooth path in $\mathcal{S}_B$, with $\Sigma_0 = \Sigma$. Write $\frac{d}{dt}\Sigma_t = \dot{\Sigma}_t$. Then

$$\begin{aligned}
&\frac{d}{dt}\,\mathrm{E}[f(z) : z \sim \mathcal{N}(0,\Sigma_t)] \\
&= \frac{d}{dt}(2\pi)^{-B/2}\det\Sigma_t^{-1/2}\int \mathrm{d}z\; e^{-\frac{1}{2}z^T\Sigma_t^{-1}z} f(z)\Big|_{t=0} \\
&= (2\pi)^{-B/2}\frac{-1}{2}\det\Sigma_0^{-1/2}\,\mathrm{tr}\left(\Sigma_0^{-1}\dot{\Sigma}_0\right)\int \mathrm{d}z\; e^{-\frac{1}{2}z^T\Sigma_0^{-1}z/2} f(z) \\
&\qquad + (2\pi)^{-B/2}\det\Sigma_0^{-1/2}\int \mathrm{d}z\; e^{-\frac{1}{2}z^T\Sigma_0^{-1}z}\frac{-1}{2}z^T(-\Sigma_0^{-1}\dot{\Sigma}_0\Sigma_0^{-1})zf(z) \\
&= \frac{-1}{2}(2\pi)^{-B/2}\det\Sigma_0^{-1/2}\int \mathrm{d}z\; e^{-\frac{1}{2}z^T\Sigma_0^{-1}z} f(z)(\mathrm{tr}\,\Sigma_0^{-1}\dot{\Sigma}_0 - z^T\Sigma_0^{-1}\dot{\Sigma}_0\Sigma_0^{-1}z) \\
&= \frac{1}{2}(2\pi)^{-B/2}\det\Sigma^{-1/2}\int \mathrm{d}z\; e^{-\frac{1}{2}z^T\Sigma^{-1}z} f(z)\left\langle \Sigma^{-1}zz^T\Sigma^{-1} - \Sigma^{-1}, \dot{\Sigma}_0\right\rangle \\
&= \frac{-1}{2}\,\mathrm{E}[f(z)\left\langle \Sigma^{-1} - \Sigma^{-1}zz^T\Sigma^{-1}, \dot{\Sigma}_0\right\rangle : z \sim \mathcal{N}(0,\Sigma)]
\end{aligned}$$

Note that

$$\begin{aligned}
v^T\frac{\mathrm{d}}{\mathrm{d}z}e^{\frac{-1}{2}z^T\Sigma^{-1}z} &= -v^T\Sigma^{-1}ze^{\frac{-1}{2}z^T\Sigma^{-1}z} \\
w^T\left(\frac{\mathrm{d}^2}{\mathrm{d}z^2}e^{\frac{-1}{2}z^T\Sigma^{-1}z}\right)v &= \left(w^T\Sigma^{-1}zz^T\Sigma^{-1}v - w^T\Sigma^{-1}v\right)e^{\frac{-1}{2}z^T\Sigma^{-1}z} \\
&= \left\langle \Sigma^{-1}zz^T\Sigma^{-1} - \Sigma^{-1}, vw^T\right\rangle e^{\frac{-1}{2}z^T\Sigma^{-1}z}
\end{aligned}$$

so that as a cotensor,

$$\frac{\mathrm{d}^2}{\mathrm{d}z^2}e^{\frac{-1}{2}z^T\Sigma^{-1}z}\{\Lambda\} = \left\langle \Sigma^{-1}zz^T\Sigma^{-1} - \Sigma^{-1}, \Lambda\right\rangle e^{\frac{-1}{2}z^T\Sigma^{-1}z}$$

for any $\Lambda \in \mathbb{R}^{B\times B}$.

Therefore,

$$\begin{aligned}
\frac{\mathrm{d}\,\mathrm{E}[f(z) : z \sim \mathcal{N}(0,\Sigma)]}{\mathrm{d}\Sigma}\{\Lambda\} &= \frac{1}{2}(2\pi)^{-B/2}\det\Sigma^{-1/2}\int \mathrm{d}z\; f(z)\frac{\mathrm{d}^2 e^{-\frac{1}{2}z^T\Sigma^{-1}z}}{\mathrm{d}^2 z}\{\Lambda\} \\
&= \frac{1}{2}(2\pi)^{-B/2}\det\Sigma^{-1/2}\int \mathrm{d}z\; e^{-\frac{1}{2}z^T\Sigma^{-1}z}\frac{\mathrm{d}^2 f(z)}{\mathrm{d}^2 z}\{\Lambda\} \\
&\qquad\qquad \text{(by integration by parts)} \\
&= \frac{1}{2}\,\mathrm{E}\left[\frac{\mathrm{d}^2 f(z)}{\mathrm{d}^2 z} : z \sim \mathcal{N}(0,\Sigma)\right]\{\Lambda\}
\end{aligned}$$

□

B.4.1 SPHERICAL COORDINATES

Note that for any $\Lambda \in \mathcal{H}_B$ with $\|\Lambda\|_2 < \upsilon^*$,

$$\begin{aligned}\mathrm{E}[\phi \circ \mathfrak{N}(z)^{\otimes 2} : z \in \mathcal{N}(0, \upsilon^* G + \Lambda)] &= \mathrm{E}[\phi \circ \mathfrak{N}(Gz)^{\otimes 2} : z \in \mathcal{N}(0, \upsilon^* I + \Lambda)] \\ &= \mathrm{E}[\mathfrak{B}_\phi(z)^{\otimes 2} : z \in \mathcal{N}(0, \upsilon^* I + \Lambda)]\end{aligned}$$

so that we have, for any $\Lambda \in \mathcal{H}_B$,

$$\begin{aligned}\mathcal{U}\{\Lambda\} &= \left.\frac{\mathrm{dV}\mathfrak{B}_\phi(\Sigma)}{\mathrm{d}\Sigma}\right|_{\Sigma = \upsilon^* I}\{\Lambda\} \\ &= \frac{1}{2}\,\mathrm{E}[\mathfrak{B}_\phi(z)^{\otimes 2}\langle \upsilon^{*-1} zz^T - \upsilon^{*-1} I, \Lambda\rangle : z \sim \mathcal{N}(0, \upsilon^* I)] \\ &\qquad\qquad \text{(by Lemma B.43)} \\ &= \frac{1}{2}\upsilon^{*-2}\,\mathrm{E}[\mathfrak{B}_\phi(z)^{\otimes 2}\langle zz^T, \Lambda\rangle : z \sim \mathcal{N}(0, \upsilon^* I)] - \frac{1}{2}\upsilon^{*-1}\Sigma^*\langle I, \Lambda\rangle \\ &\qquad\qquad \text{(by Thm B.25)} \\ &= (2\upsilon^*)^{-1}\left(\mathrm{E}[\mathfrak{B}_\phi(z)^{\otimes 2}\langle zz^T, \Lambda\rangle : z \sim \mathcal{N}(0, I)] - \Sigma^*\langle I, \Lambda\rangle\right) \\ &\qquad\qquad (\mathfrak{B}_\phi \text{ is scale-invariant}) \\ &= (2\upsilon^*)^{-1}\left(\mathrm{E}[\mathfrak{B}_\phi(z)^{\otimes 2}\langle Gzz^T G, \Lambda\rangle : z \sim \mathcal{N}(0, I)] - \Sigma^*\langle I, \Lambda\rangle\right) \\ &\qquad\qquad (\Lambda = G\Lambda G)\end{aligned}$$

Let's extend to all matrices by this formula:

Definition B.44. Define

$$\begin{aligned}\tilde{\mathcal{U}} : \mathbb{R}^{B\times B} &\to \mathcal{S}_B, \\ \Lambda &\mapsto (2\upsilon^*)^{-1}\left(\mathrm{E}[\mathfrak{B}_\phi(z)^{\otimes 2}\langle Gzz^T G, \Lambda\rangle : z \sim \mathcal{N}(0, I)] - \Sigma^*\langle I, \Lambda\rangle\right)\end{aligned}$$

So $\tilde{\mathcal{U}} \restriction \mathcal{H}_B = \mathcal{U} \restriction \mathcal{H}_B$. Ultimately we will apply Lemma B.37 to $G^{\otimes 2} \circ \tilde{\mathcal{U}} \restriction \mathcal{H}_B = G^{\otimes 2} = \mathcal{U}$

Definition B.45. Write $\tau_{ij} = \frac{1}{2}\left(\delta_i\delta_j^T + \delta_j\delta_i^T\right)$.

Then

$$\begin{aligned}\tilde{\mathcal{U}}\{\tau_{ij}\}_{kl} &= (2\upsilon^*)^{-1}\left(\mathrm{E}[\mathfrak{B}_\phi(z)_k\mathfrak{B}_\phi(z)_l(Gz)_i(Gz)_j : z \sim \mathcal{N}(0, I)] - \Sigma^*_{kl}\mathbb{I}(i=j)\right) \\ &= (2\upsilon^*)^{-1}\left(\mathrm{E}[\phi(\mathfrak{N}(y))_k\phi(\mathfrak{N}(y))_l y_i y_j : y \sim \mathcal{N}(0, G)] - \Sigma^*_{kl}\mathbb{I}(i=j)\right) \\ &= (2\upsilon^*)^{-1}\left(\mathrm{E}[\phi(\mathfrak{N}(y))_k\phi(\mathfrak{N}(y))_l y_i y_j : y \sim \mathcal{N}(0, \mathbb{e}\mathbb{e}^T)] - \Sigma^*_{kl}\mathbb{I}(i=j)\right) \\ &\qquad\qquad \text{(See Defn B.6 for defn of } \mathbb{e}) \\ &= (2\upsilon^*)^{-1}\left(\mathrm{E}[\phi(\mathfrak{N}(\mathbb{e}x))_k\phi(\mathfrak{N}(\mathbb{e}x))_l(\mathbb{e}x)_i(\mathbb{e}x)_j : x \sim \mathcal{N}(0, I_{B-1})] - \Sigma^*_{kl}\mathbb{I}(i=j)\right)\end{aligned}$$

Here we will realize $\mathbb{e}$ as the matrix whose columns are $\mathbb{e}_{B-m} := (m(m+1))^{-1/2}(0, \ldots, 0, -m, \overbrace{1, 1, \ldots, 1}^{m})^T$, for each $m = 1, \ldots, B-1$.

$$\mathbb{e}^T = \begin{pmatrix} \frac{1}{\sqrt{B(B-1)}} & & & & \\ & \frac{1}{\sqrt{(B-1)(B-2)}} & & & \\ & & \ddots & & \\ & & & 6^{-1/2} & \\ & & & & 2^{-1/2} \end{pmatrix} \begin{pmatrix} -(B-1) & 1 & 1 & \cdots & 1 & 1 & 1 \\ 0 & -(B-2) & 1 & \cdots & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -2 & 1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \tag{39}$$

Definition B.46. Define $W_{ij|kl} := \mathrm{E}[\phi(\mathfrak{N}(\mathbb{e}x))_k\phi(\mathfrak{N}(\mathbb{e}x))_l(\mathbb{e}x)_i(\mathbb{e}x)_j : x \sim \mathcal{N}(0, I_{B-1})]$.

Then $\tilde{\mathcal{U}}\{\tau_{ij}\}_{kl} = (2v^*)^{-1}(W_{ij|kl} - \Sigma^*_{kl}\mathbb{I}(i = j))$. If we can evaluate $W_{ij|kl}$ then we can use Lemma B.37 to compute the eigenvalues of $G^{\otimes 2} \circ \mathcal{U}$. It's easy to see that $W_{ij|kl}$ is ultrasymmetric. Thus WLOG we can take i, j, k, l from $\{1, 2, 3, 4\}$.

Lemma B.47. *Let $f : \mathbb{R}^4 \to \mathbb{R}^A$ for some $A \in \mathbb{N}$. Suppose $B \geq 6$. Then for any $k < B - 1$*

$$\begin{aligned}&\mathrm{E}[r^{-k} f(v_1, \ldots, v_4) : x \sim \mathcal{N}(0, I_{B-1})] \\ &= \frac{\Gamma((B-1-k)/2)}{\Gamma((B-5)/2)} 2^{-k/2}\pi^{-2} \int_0^\pi \mathrm{d}\theta_1 \cdots \int_0^\pi \mathrm{d}\theta_4 \; f(v_1^\theta, \ldots, v_4^\theta) \sin^{B-3}\theta_1 \cdots \sin^{B-6}\theta_4\end{aligned}$$

where $v_i = x_i/\|x\|$ and $r = \|x\|$, and

$$\begin{aligned}v_1^\theta &= \cos\theta_1 \\ v_2^\theta &= \sin\theta_1 \cos\theta_2 \\ v_3^\theta &= \sin\theta_1 \sin\theta_2 \cos\theta_3 \\ v_4^\theta &= \sin\theta_1 \sin\theta_2 \sin\theta_3 \cos\theta_4.\end{aligned}$$

Lemma B.48. *Let $f : \mathbb{R}^2 \to \mathbb{R}^A$ for some $A \in \mathbb{N}$. Suppose $B \geq 4$. Then for any $k < B - 1$*

$$\begin{aligned}&\mathrm{E}[r^{-k} f(v_1, v_2) : x \sim \mathcal{N}(0, I_{B-1})] \\ &= \frac{\Gamma((B-1-k)/2)}{\Gamma((B-3)/2)} 2^{-k/2}\pi^{-1} \int_0^\pi \mathrm{d}\theta_1 \int_0^\pi \mathrm{d}\theta_2 \; f(v_1^\theta, v_2^\theta) \sin^{B-3}\theta_1 \sin^{B-4}\theta_2\end{aligned}$$

where $v_i = x_i/\|x\|$ and $r = \|x\|$, and

$$\begin{aligned}v_1^\theta &= \cos\theta_1 \\ v_2^\theta &= \sin\theta_1 \cos\theta_2\end{aligned}$$

Lemma B.49. *Let $f : \mathbb{R} \to \mathbb{R}^A$ for some $A \in \mathbb{N}$. Suppose $B \geq 3$. Then for any $k < B - 1$*

$$\begin{aligned}&\mathrm{E}[r^{-k} f(v) : x \sim \mathcal{N}(0, I_{B-1})] \\ &= \frac{\Gamma((B-1-k)/2)}{\Gamma((B-2)/2)} 2^{-k/2}\pi^{-1/2} \int_0^\pi \mathrm{d}\theta \; f(\cos\theta) \sin^{B-3}\theta\end{aligned}$$

where $v_i = x_i/\|x\|$ and $r = \|x\|$.

Lemma B.47 implies Lemma B.48 and Lemma B.49 by integrating out the appropriate θ_is, so we will just prove Lemma B.47.

Proof of Lemma B.47.

$$\mathrm{E}[r^{-k} f(v_1, \ldots, v_4) : x \sim \mathcal{N}(0, I_{B-1})] = (2\pi)^{-(B-1)/2} \int \mathrm{d}x \; r^{-k} f(v_1, \ldots, v_4) e^{-r^2/2}$$

Now we change into spherical coordinates $(r, \theta_1, \ldots, \theta_{B-2}) \in [0, \infty) \times [0, \pi]^{B-3} \times [0, 2\pi]$, defined by

$$\begin{aligned}x_1 &= r\cos\theta_1 \\ x_2 &= r\sin\theta_1\cos\theta_2 \\ x_3 &= r\sin\theta_1\sin\theta_2\cos\theta_3 \\ &\vdots \\ x_{B-2} &= r\sin\theta_1 \cdots \sin\theta_{B-3}\cos\theta_{B-2} \\ x_{B-1} &= r\sin\theta_1 \cdots \sin\theta_{B-3}\sin\theta_{B-2}.\end{aligned}$$

Then

$$v = \begin{pmatrix} \cos\theta_1 \\ \sin\theta_1\cos\theta_2 \\ \sin\theta_1\sin\theta_2\cos\theta_3 \\ \vdots \\ \sin\theta_1 \cdots \sin\theta_{B-3}\cos\theta_{B-2} \\ \sin\theta_1 \cdots \sin\theta_{B-3}\sin\theta_{B-2} \end{pmatrix}$$

and

$$\mathrm{d}x_1 \dots \mathrm{d}x_{B-1} = r^{B-2} \sin^{B-3}\theta_1 \sin^{B-4}\theta_2 \cdots \sin\theta_{B-3}\,\mathrm{d}r\,\mathrm{d}\theta_1 \cdots \mathrm{d}\theta_{B-2}.$$

So

$$\begin{aligned}
&\mathrm{E}[r^{-k} f(v_1,\dots,v_4) : x \sim \mathcal{N}(0, I_{B-1})] \\
&= (2\pi)^{-(B-1)/2} \int \mathrm{d}x\, r^{-k} f(v_1,\dots,v_4) e^{-r^2/2} \\
&= (2\pi)^{-(B-1)/2} \int \left(r^{B-2}\sin^{B-3}\theta_1 \cdots \sin\theta_{B-3}\,\mathrm{d}r\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_{B-2}\right) r^{-k} f(v_1^\theta,\dots,v_4^\theta) e^{-r^2/2} \\
&= (2\pi)^{-(B-1)/2} \int f(v_1^\theta,\dots,v_4^\theta)\sin^{B-3}\theta_1\cdots\sin\theta_{B-3}\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_{B-2} \int r^{B-2-k} e^{-r^2/2}\,\mathrm{d}r \\
&= (2\pi)^{-(B-1)/2} \int f(v_1^\theta,\dots,v_4^\theta)\sin^{B-3}\theta_1\cdots\sin\theta_{B-3}\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_{B-2} \int_0^\infty r^{B-2-k} e^{-r^2/2}\,\mathrm{d}r \\
&= (2\pi)^{-(B-1)/2} \int f(v_1^\theta,\dots,v_4^\theta)\sin^{B-3}\theta_1\cdots\sin\theta_{B-3}\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_{B-2} 2^{(B-3-k)/2}\Gamma((B-1-k)/2) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(change of coordinate } r \leftarrow r^2 \text{ and definition of } \Gamma \text{ function)} \\
&= \Gamma((B-1-k)/2) 2^{-(2+k)/2}\pi^{-(B-1)/2} \int f(v_1^\theta,\dots,v_4^\theta)\sin^{B-3}\theta_1\cdots\sin\theta_{B-3}\,\mathrm{d}\theta_1\cdots\mathrm{d}\theta_{B-2}
\end{aligned}$$

Because $v_1^\theta,\dots,v_4^\theta$ only depends on $\theta_1,\dots,\theta_4$, we can integrate out $\theta_5,\dots,\theta_{B-2}$. By applying Lemma B.21 to $\theta_5,\dots\theta_{B-3}$ and noting that $\int_0^{2\pi}\mathrm{d}\theta_{B-2} = 2\pi$, we get

$$\begin{aligned}
&\mathrm{E}[r^{-k} f(v_1^\theta,\dots,v_4^\theta) : x \sim \mathcal{N}(0, I_{B-1})] \\
&= \Gamma((B-1-k)/2) 2^{-(2+k)/2}\pi^{-(B-1)/2} \int f(v_1^\theta,\dots,v_4^\theta)\sin^{B-3}\theta_1\cdots\sin^{B-6}\theta_4 \pi^{(B-7)/2}\Gamma((B-5)/2)^{-1}(2\pi) \\
&= \frac{\Gamma((B-1-k)/2)}{\Gamma((B-5)/2)} 2^{-k/2}\pi^{-2} \int f(v_1^\theta,\dots,v_4^\theta)\sin^{B-3}\theta_1\cdots\sin^{B-6}\theta_4
\end{aligned}$$

□

Assuming $\{i,j,k,l\} \in \{1,2,3,4\}$,

$$\begin{aligned}
(\mathbb{e}v)_1 &= -\sqrt{\frac{B-1}{B}}\cos\theta_1 \\
(\mathbb{e}v)_2 &= \cos\theta_1 - \sqrt{\frac{B-2}{B-1}}\sin\theta_1\cos\theta_2 \\
(\mathbb{e}v)_3 &= \cos\theta_1 + \sin\theta_1\cos\theta_2 - \sqrt{\frac{B-3}{B-2}}\sin\theta_1\sin\theta_2\cos\theta_3 \\
(\mathbb{e}v)_4 &= \cos\theta_1 + \sin\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2\cos\theta_3 - \sqrt{\frac{B-4}{B-3}}\sin\theta_1\sin\theta_2\sin\theta_3\cos\theta_4
\end{aligned}$$

so in particular they only depend on $\theta_1,\dots,\theta_4$. By Lemma B.47, and the fact that $x \mapsto \mathbb{e}x$ is an isometry,

$$\begin{aligned}
W_{ij|kl} &= \mathrm{E}[r^2\phi(\sqrt{B}\mathbb{e}v)_k\phi(\sqrt{B}\mathbb{e}v)_l(\mathbb{e}v)_i(\mathbb{e}v)_j : x \sim \mathcal{N}(0, I_{B-1})] \\
&= (B-5)(B-3)(B-1)(2\pi)^{-2}\int_0^\pi \mathrm{d}\theta_1 \cdots \int_0^\pi \mathrm{d}\theta_4\, \phi(\sqrt{B}\mathbb{e}v)_k\phi(\sqrt{B}\mathbb{e}v)_l(\mathbb{e}v)_i(\mathbb{e}v)_j \sin^{B-3}\theta_1\cdots\sin^{B-6}\theta_4
\end{aligned}$$

If WLOG we further assume that $k,l \in \{1,2\}$ (by ultrasymmetry), then there is no dependence on θ_3 and θ_4 inside ϕ. So we can expand $(\mathbb{e}v)_i$ and $(\mathbb{e}v)_j$ in trigonometric expressions like above and integrate out θ_3 and θ_4 via Lemma B.20.

B.4.2 LAPLACE METHOD

Differentiating Eq. (31) In what follows, let ϕ be a positive-homogeneous function of degree α. We begin by studying $\frac{\mathrm{dV}(\phi\circ\mathfrak{N})(\Sigma^G)}{\mathrm{d}\Sigma^G}$. We will differentiate Eq. (31) directly at $G^{\otimes 2}\{\Sigma^*\} = G\Sigma^* G$ where Σ^* is the BSB1 fixed point given by Thm B.25. To that end, consider a smooth path $\Sigma(t) \in \mathcal{S}_B^G, t \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$, with $\Sigma(0) = G\Sigma^* G$. Set $\Sigma_\square(t) = \mathbb{e}^T\Sigma(t)\mathbb{e} \in \mathcal{S}_{B-1}$, so that $\Sigma(t) = \mathbb{e}\Sigma_\square(t)\mathbb{e}^T$ and $\Sigma_\square(0) = \upsilon^* I_{B-1}$ where $\upsilon^* = K_{\alpha,B}(\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi(\frac{-1}{B-1}))$ as in Thm B.25. If we write $\dot\Sigma_\square$ and $\dot\Sigma$ for the time derivatives, we have

$$
\begin{aligned}
& B^{-\alpha}\Gamma(\alpha)\frac{d}{dt}\mathrm{V}(\phi\circ\mathfrak{N})(\mathbb{e}\Sigma_\square(t)\mathbb{e}^T)\Big|_{t=0} \\
&= B^{-\alpha}\Gamma(\alpha)\frac{d}{dt}\mathrm{V}\mathfrak{B}_\phi(\mathbb{e}\Sigma_\square(t)\mathbb{e}^T)\Big|_{t=0} \\
&= \int_0^\infty \mathrm{d}s\, s^{\alpha-1}\Bigg[\frac{d}{dt}\det(I+2s\Sigma_\square(t))^{-1/2}\Big|_{t=0} \times \mathrm{V}\phi\left(\mathbb{e}\Sigma_\square(0)(I+2s\Sigma_\square(0))^{-1}\mathbb{e}^T\right) \\
&\qquad + \det(I+2s\Sigma_\square(0))^{-1/2}\frac{d}{dt}\mathrm{V}\phi(\mathbb{e}\Sigma_\square(t)(I+2s\Sigma_\square(t))^{-1}\mathbb{e}^T)\Big|_{t=0}\Bigg] \\
&= \int_0^\infty \mathrm{d}s\, s^{\alpha-1}\Bigg[-\frac{s}{1+2s\upsilon^*}\det(I+2s\Sigma_\square(0))^{-1/2}\operatorname{tr}\left(\dot\Sigma_\square(0)\right) \times \left(\frac{\upsilon^*}{1+2s\upsilon^*}\right)^\alpha \mathrm{V}\phi(G) \\
&\qquad + \det(I+2s\Sigma_\square(0))^{-1/2}\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\Big|_{\Sigma=\mathbb{e}\Sigma_\square(0)(I+2s\Sigma_\square(0))^{-1}\mathbb{e}^T} \\
&\qquad\qquad \left\{\mathbb{e}(I+2s\Sigma_\square(0))^{-1}\dot\Sigma_\square(0)(I+2s\Sigma_\square(0))^{-1}\mathbb{e}^T\right\}\Bigg]
\end{aligned}
$$

(apply Lemma B.51 and Lemma B.52)

$$
\begin{aligned}
&= \int_0^\infty \mathrm{d}s\, s^{\alpha-1}\Bigg[-\frac{s}{1+2s\upsilon^*}(1+2s\upsilon^*)^{-(B-1)/2}\operatorname{tr}\left(\dot\Sigma_\square(0)\right) \times \left(\frac{\upsilon^*}{1+2s\upsilon^*}\right)^\alpha \mathrm{V}\phi(G) \\
&\qquad + (1+2s\upsilon^*)^{-(B-1)/2}\left(\frac{\upsilon^*}{1+2s\upsilon^*}\right)^{\alpha-1}\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\Big|_{\Sigma=G}\left\{(1+2s\upsilon^*)^{-2}\mathbb{e}\dot\Sigma_\square(0)\mathbb{e}^T\right\}\Bigg]
\end{aligned}
$$

(using fact that $\mathrm{dV}\phi/\mathrm{d}\Sigma$ is degree $(\alpha-1)$-positive homogeneous)

$$
\begin{aligned}
&= -\left(\int_0^\infty (1+2s\upsilon^*)^{-\frac{B-1}{2}-1-\alpha}\upsilon^{*\alpha}s^\alpha\right)\operatorname{tr}\left(\dot\Sigma_\square(0)\right)\mathrm{V}\phi(G) \\
&\quad + \left(\int_0^\infty (1+2s\upsilon^*)^{-\frac{B-1}{2}-1-\alpha}\upsilon^{*\alpha-1}s^{\alpha-1}\right)\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\Big|_{\Sigma=G}\left\{\mathbb{e}\dot\Sigma_\square(0)\mathbb{e}^T\right\} \\
&= -\upsilon^{*-1}2^{-1-\alpha}\operatorname{Beta}\left(\frac{B-1}{2},\alpha+1\right)\operatorname{tr}\left(\dot\Sigma_\square(0)\right)\mathrm{V}\phi(G) \\
&\quad + \upsilon^{*-1}2^{-\alpha}\operatorname{Beta}\left(\frac{B+1}{2},\alpha\right)\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\Big|_{\Sigma=G}\left\{\dot\Sigma(0)\right\}
\end{aligned}
$$

(apply Lemma B.53)

if $\frac{B-1}{2} + 2 > \alpha$ (precondition for Lemma B.53).

With some trivial simplifications, we obtain the following

Lemma B.50. *Let ϕ be positive-homogeneous of degree α. Consider a smooth path $\Sigma(t) \in \mathcal{S}_B^G$ with $\Sigma(0) = \upsilon^* G$. If $\frac{B-1}{2} + 2 > \alpha$, then*

$$\begin{aligned}
&\frac{d}{dt}\mathrm{V}(\phi \circ \mathfrak{N})(\Sigma(t))\Big|_{t=0} \\
&= \frac{\mathrm{dV}(\phi \circ \mathfrak{N})(\Sigma)}{\mathrm{d}\Sigma}\Big|_{\Sigma=\upsilon^* G}\{\dot{\Sigma}(0)\} \\
&= -\alpha B^\alpha \upsilon^{*-1} 2^{-1-\alpha} \mathrm{P}\left(\frac{B-1}{2}, \alpha+1\right)^{-1} \mathrm{tr}\left(\dot{\Sigma}_\square(0)\right) \mathrm{V}\phi(G) \\
&\quad + B^\alpha \upsilon^{*-1} 2^{-\alpha} \mathrm{P}\left(\frac{B+1}{2}, \alpha\right)^{-1} \frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\Big|_{\Sigma=G}\left\{\dot{\Sigma}(0)\right\}
\end{aligned} \tag{40}$$

where $\mathrm{P}(a,b) = \Gamma(a+b)/\Gamma(a)$ is the Pochhammer symbol.

Lemma B.51. *For any $\Sigma_\square(0)$, $\frac{d}{dt}\det(I + 2s\Sigma_\square)^{-1/2} = -s\det(I + 2s\Sigma_\square)^{-1/2}\,\mathrm{tr}((I + 2s\Sigma_\square)^{-1} d\Sigma_\square/dt)$.*

For $\Sigma_\square(0) = \upsilon I$, this is also equal to $-\frac{s}{1+2s\upsilon}\det(I + 2s\Sigma_\square)^{-1/2}\,\mathrm{tr}(\frac{d}{dt}\Sigma_\square)$.

Proof. Straightforward computation. □

Lemma B.52. *For any $\Sigma_\square(0)$, $\frac{d}{dt}\Sigma_\square(I + 2s\Sigma_\square)^{-1} = (I + 2s\Sigma_\square(0))\frac{d\Sigma_\square}{dt}(I + 2s\Sigma_\square(0))^{-1}$.*

Proof. Straightforward computation. □

Lemma B.53. *For $a > b + 1$, $\int_0^\infty (1 + 2s\upsilon)^{-a} s^b \,\mathrm{d}s = (2\upsilon)^{-1-b}\,\mathrm{Beta}(b+1, a-1-b)$.*

Proof. Apply change of variables $x = \frac{2\upsilon s}{1+2\upsilon s}$. □

Definition B.54. Let $\mathcal{T} : \mathcal{H}_B \to \mathcal{H}_B$ be such that for any $\Sigma \in \mathcal{H}_B$ and any $i \neq j \in [B]$,

$$\begin{aligned}
\mathcal{T}\{\Sigma\}_{ii} &= u\Sigma_{ii} \\
\mathcal{T}\{\Sigma\}_{ij} &= v\Sigma_{ii} + v\Sigma_{jj} + w\Sigma_{ij}
\end{aligned}$$

Then we say that $\mathcal{T}$ is *diagonal-off-diagonal semidirect*, or *DOS* for short. We write more specifically $\mathcal{T} = \mathrm{DOS}_B(u, v, w)$.

Lemma B.55. *Let $\mathcal{T} := \mathrm{DOS}_B(u, v, w)$. Then $\mathcal{T}\{\mathrm{L}_B(a, b)\} = \mathrm{L}_B(ua, wb - va)$.*

Proof. Let $L := \mathrm{L}_B(a, b)$. It suffices to verify the following, each of which is a direct computation.

1. $\mathcal{T}\{L\}_{3:B,3:B} = 0$.
2. $\mathcal{T}\{L\}_{1,2} = \mathcal{T}\{L\}_{2,1} = 0$
3. $\mathcal{T}\{L\}_{1,1} = -\mathcal{T}\{L\}_{2,2} = ua$
4. $\mathcal{T}\{L\}_{1,i} = -\mathcal{T}\{L\}_{2,i} = \mathcal{T}\{L\}_{i,1} = -\mathcal{T}\{L\}_{i,2} = va - wb$.

□

Lemma B.56. *Let $\mathcal{T} := \mathrm{DOS}_B(u, v, w)$. Then $\mathrm{L}_B(w - u, v)$ and $\mathrm{L}_B(0, 1)$ are its eigenvectors:*

$$\begin{aligned}
\mathcal{T}\{\mathrm{L}_B(w-u, v)\} &= u\mathrm{L}_B(w-u, v) \\
\mathcal{T}\{\mathrm{L}_B(0, 1)\} &= w\mathrm{L}_B(0, 1)
\end{aligned}$$

Proof. The map $(a, b) \mapsto (ua, wb - va)$ has eigenvalues u and w with corresponding eigenvectors $(w - u, v)$ and $(0, 1)$. By Lemma B.55, this implies the desired results. □

Lemma B.57. *Let $\mathcal{T} := \mathrm{DOS}_B(u, v, w)$. Then for any $L \in \mathbb{L}_B$, $G^{\otimes 2} \circ \mathcal{T}\{L\} = \frac{(u-2v)(B-2)+2w}{B} L$.*

Proof. By Lemma B.36 and Lemma B.55, $G^{\otimes 2} \circ \mathcal{T}\{\mathrm{L}(B-2,1)\} = G^{\otimes 2}\{\mathrm{L}(u(B-1), w - v(B-1))\} = \frac{(u-2v)(B-2)+2w}{B}\mathrm{L}_B(B-2,1)$. By permutation symmetry, we also have the general result for any $\mathrm{L}_B(B-2,1)$-shaped matrix L. Since they span $\mathbb{L}_B$, this gives the conclusion we want. □

Lemma B.58. *Let* $\mathcal{T} := \mathrm{DOS}_B(u,v,w)$. *Then* $\mathcal{T}\{\mathrm{BSB1}_B(a,b)\} = \mathrm{BSB1}_B(ua, wb+2va)$.

Proof. Direct computation. □

Lemma B.59. *Let* $\mathcal{T} := \mathrm{DOS}_B(u,v,w)$. *Then* $\mathrm{BSB1}_B(w-u, wv)$ *and* $\mathrm{BSB1}_B(0,1)$ *are eigenvectors of* $\mathcal{T}$.

$$\mathcal{T}\{\mathrm{BSB1}_B(u-w, 2v)\} = u\mathrm{BSB1}_B(u-w, 2v)$$
$$\mathcal{T}\{\mathrm{BSB1}_B(0,1)\} = w\mathrm{BSB1}_B(0,1)$$

Proof. The linear map $(a,b) \mapsto (ua, wb+2va)$ has eigenvalues u and w with corresponding eigenvectors $(u-w, 2v)$ and $(0,1)$. The result then immediately follows from Lemma B.58. □

Lemma B.60. *Let* $\mathcal{T} := \mathrm{DOS}_B(u,v,w)$. *Then* $G^{\otimes 2} \circ \mathcal{T}\{G\} = \frac{(B-1)(u-2v)+w}{B}G$.

Proof. Direct computation with Lemma B.58 and Lemma B.15 □

Definition B.61. Define $\mathbf{M}_B := \{\Sigma \in \mathcal{S}_B : \mathrm{Diag}\Sigma = 0\}$.

For any $a, b \in \mathbb{R}$, set

$$\mathcal{L}_B(a,b) := \begin{pmatrix} a & -b & -b & \cdots \\ -b & 0 & 0 & \cdots \\ -b & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \in \mathcal{H}_B.$$

Define $\mathbf{L}_B(a,b) := \mathrm{span}(P_{1i}^T \mathcal{L}_B(a,b) P_{1i} : i \in [B])$ where P_{1i} is the permutation matrix that swap the first entry with the ith entry, i.e. $\mathbf{L}_B(a,b)$ is the span of the orbit of $\mathcal{L}_B(a,b)$ under permuting rows and columns simultaneously.

Note that

$$\mathrm{L}_B(a,b) = \mathcal{L}_B(a,b) - P_{12}^T \mathcal{L}_B(a,b) P_{12}^T$$
$$\mathrm{BSB1}_B(a,-2b) = \sum_{i=1}^{B} P_{1i}^T \mathcal{L}_B(a,b) P_{1i}$$

So $\mathrm{L}_B(a,b), \mathrm{BSB1}_B(a,-2b) \in \mathbf{L}_B(a,b)$.

Theorem B.62. *Let* $\mathcal{T} := \mathrm{DOS}_B(u,v,w)$. *Then* $\mathcal{T} \restriction \mathcal{H}_B$ *has the following eigendecomposition:*

- $\mathbf{M}_B$ *has eigenvalue* w *(*$\dim \mathbf{M}_B = B(B-1)/2$*)*
- $\mathbf{L}_B(w-u, v)$ *has eigenvalue of* u *(*$\dim \mathbf{L}_B(w-u,v) = B$*)*

Proof. The case of $\mathbf{M}_B$ is obvious. We will show that $\mathbf{L}_B(w-u,v)$ is an eigenspace with eigenvalue u. Then by dimensionality consideration they are all of the eigenspaces of $\mathcal{T}$.

Let $L := \mathcal{L}_B(w-u, v)$. Then it's not hard to see $\mathcal{T}\{L\} = \mathcal{L}_B(a,b)$ for some $a, b \in \mathbb{R}$. It follows that a has to be $u(w-u)$ and b has to be $-(v(w-u) - wv) = uv$, which yields what we want. □

Theorem B.63. *Let* $\mathcal{T} := \mathrm{DOS}_B(u,v,w)$. *Then* $G^{\otimes 2} \circ \mathcal{T} \restriction \mathcal{H}_B^G : \mathcal{H}_B^G \to \mathcal{H}_B^G$ *has the following eigendecomposition:*

- $\mathbb{R}G$ *has eigenvalue* $\frac{(B-1)(u-2v)+w}{B}$ *(*$\dim \mathbb{R}G = 1$*)*

- $\mathbb{M}_B$ *has eigenvalue* w *(*$\dim \mathbb{M}_B = B(B-3)/2$*)*
- $\mathbb{L}_B$ *has eigenvalue* $\frac{(B-2)(u-2v)+2w}{B}$ *(*$\dim \mathbb{L}_B = B-1$*)*

Proof. The case of $\mathbb{M}_B$ is obvious. The case for $\mathbb{R}G$ follows from Lemma B.60. The case for $\mathbb{L}_B$ follows from Lemma B.57. By dimensionality considerations these are all of the eigenspaces. □

This immediately gives the following consequence.

Theorem B.64. *Let* $\phi : \mathbb{R} \to \mathbb{R}$ *be any function with finite first and second Gaussian moments. Then* $G^{\otimes 2} \circ \frac{\mathrm{dV}\phi}{\mathrm{d}\Sigma}\big|_{\Sigma=\mathrm{BSB1}(a,b)} : \mathcal{H}_B^G \to \mathcal{H}_B^G$ *has the following eigendecomposition:*

- $\mathbb{R}G$ *has eigenvalue* $\frac{(B-1)(u-2v)+w}{B}$ *(*$\dim \mathbb{R}G = 1$*)*
- $\mathbb{M}_B$ *has eigenvalue* w *(*$\dim \mathbb{M}_B = B(B-3)/2$*)*
- $\mathbb{L}_B$ *has eigenvalue* $\frac{(B-2)(u-2v)+2w}{B}$ *(*$\dim \mathbb{L}_B = B-1$*)*

where

$$u = \frac{\partial \mathrm{V}\phi(\Sigma)_{11}}{\partial \Sigma_{11}}\bigg|_{\Sigma=\mathrm{BSB1}(a,b)}$$
$$v = \frac{\partial \mathrm{V}\phi(\Sigma)_{12}}{\partial \Sigma_{11}}\bigg|_{\Sigma=\mathrm{BSB1}(a,b)}$$
$$w = \frac{\partial \mathrm{V}\phi(\Sigma)_{12}}{\partial \Sigma_{12}}\bigg|_{\Sigma=\mathrm{BSB1}(a,b)}$$

Theorem B.65. *Let* $\phi : \mathbb{R} \to \mathbb{R}$ *be positive-homogeneous of degree* α*. Then for any* $p \neq 0, c \in \mathbb{R}$*,* $G^{\otimes 2} \circ \frac{\mathrm{dV}\phi}{\mathrm{d}\Sigma}\big|_{\Sigma=\mathrm{BSB1}(p,cp)} : \mathcal{H}_B^G \to \mathcal{H}_B^G$ *has the following eigendecomposition:*

- $\mathbb{M}_B$ *has eigenvalue* $\mathsf{c}_\alpha p^{\alpha-1} \mathfrak{J}'_\phi(c)$
- $\mathbb{L}_B$ *has eigenvalue* $\mathsf{c}_\alpha p^{\alpha-1} \frac{1}{B}\left((B-2)\alpha\left[\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi(c)\right] + (2 + c(B-2))\,\mathfrak{J}'_\phi(c)\right)$
- $\mathbb{R}G$ *has eigenvalue* $\mathsf{c}_\alpha p^{\alpha-1} \frac{1}{B}\left((B-1)\alpha\left[\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi(c)\right] + (1 + c(B-1))\,\mathfrak{J}'_\phi(c)\right)$

Proof. By Proposition I.6, $\frac{\mathrm{dV}\phi}{\mathrm{d}\Sigma}\big|_{\Sigma=\mathrm{BSB1}(p,cp)}$ is $\mathrm{DOS}(u,v,w)$ with

$$\begin{aligned}
u &= \frac{\partial \mathrm{V}\phi(\Sigma)_{11}}{\partial \Sigma_{11}}\bigg|_{\Sigma=\mathrm{BSB1}(p,cp)} \\
&= \mathsf{c}_\alpha \alpha p^{\alpha-1} \mathfrak{J}_\phi(1) \\
v &= \frac{\partial \mathrm{V}\phi(\Sigma)_{12}}{\partial \Sigma_{11}}\bigg|_{\Sigma=\mathrm{BSB1}(p,cp)} \\
&= \frac{1}{2}\mathsf{c}_\alpha \Sigma_{ii}^{\frac{1}{2}(\alpha-2)} \Sigma_{jj}^{\frac{1}{2}\alpha} (\alpha \mathfrak{J}_\phi(c_{ij}) - c_{ij}\mathfrak{J}'_\phi(c_{ij})) \\
&= \frac{1}{2}\mathsf{c}_\alpha p^{\alpha-1}\left(\alpha \mathfrak{J}_\phi(c) - c\mathfrak{J}'_\phi(c)\right) \\
w &= \frac{\partial \mathrm{V}\phi(\Sigma)_{12}}{\partial \Sigma_{12}}\bigg|_{\Sigma=\mathrm{BSB1}(p,cp)} \\
&= \mathsf{c}_\alpha \Sigma_{ii}^{(\alpha-1)/2} \Sigma_{jj}^{(\alpha-1)/2} \mathfrak{J}'_\phi(c_{ij}) \\
&= \mathsf{c}_\alpha p^{\alpha-1} \mathfrak{J}'_\phi(c)
\end{aligned}$$

With Thm B.63, we can do the computation:

- $\mathbb{M}_B$ has eigenvalue $w = \mathsf{c}_\alpha p^{\alpha-1}\mathfrak{J}'_\phi(c)$
- $\mathbb{L}_B$ has eigenvalue

$$\begin{aligned}
&\frac{(u-2v)(B-2)+2w}{B} \\
&= \frac{1}{B}\left((B-2)\left(\mathsf{c}_\alpha \alpha p^{\alpha-1}\mathfrak{J}_\phi(1) - \mathsf{c}_\alpha p^{\alpha-1}\left(\alpha\mathfrak{J}_\phi(c) - c\mathfrak{J}'_\phi(c)\right)\right) + 2\mathsf{c}_\alpha p^{\alpha-1}\mathfrak{J}'_\phi(c)\right) \\
&= \mathsf{c}_\alpha p^{\alpha-1}\frac{1}{B}\left((B-2)\alpha\left[\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi(c)\right] + (2+c(B-2))\mathfrak{J}'_\phi(c)\right)
\end{aligned}$$

- $\mathbb{R}G$ has eigenvalue

$$\begin{aligned}
&\frac{(B-1)(u-2v)+w}{B} \\
&= \frac{1}{B}\left((B-1)\left(\mathsf{c}_\alpha \alpha p^{\alpha-1}\mathfrak{J}_\phi(1) - \mathsf{c}_\alpha p^{\alpha-1}\left(\alpha\mathfrak{J}_\phi(c) - c\mathfrak{J}'_\phi(c)\right)\right) + \mathsf{c}_\alpha p^{\alpha-1}\mathfrak{J}'_\phi(c)\right) \\
&= \mathsf{c}_\alpha p^{\alpha-1}\frac{1}{B}\left((B-1)\alpha\left[\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi(c)\right] + (1+c(B-1))\mathfrak{J}'_\phi(c)\right)
\end{aligned}$$

□

We record the following consequence which will be used frequently in the sequel.

Theorem B.66. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be positive-homogeneous of degree α. Then $G^{\otimes 2} \circ \frac{\mathrm{dV}\phi}{\mathrm{d}\Sigma}\big|_{\Sigma=G} : \mathcal{H}_B^G \to \mathcal{H}_B^G$ has the following eigendecomposition:*

- *$\mathbb{M}_B$ has eigenvalue $\lambda_{\mathbb{M}}^{G,\phi}(B,\alpha) := \mathsf{c}_\alpha\left(\frac{B-1}{B}\right)^{\alpha-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)$*
- *$\mathbb{L}_B$ has eigenvalue $\lambda_{\mathbb{L}}^{G,\phi}(B,\alpha) := \mathsf{c}_\alpha\left(\frac{B-1}{B}\right)^{\alpha-1}\frac{1}{B}\left((B-2)\alpha\left[\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)\right] + \frac{B}{B-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)\right)$*
- *$\mathbb{R}G$ has eigenvalue $\lambda_{\mathbb{R}G}^{G,\phi}(B,\alpha) := \mathsf{c}_\alpha\left(\frac{B-1}{B}\right)^{\alpha}\alpha\left[\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)\right]$*

Proof. Plug in $p = \frac{B-1}{B}$ and $c = -1/(B-1)$ for Thm B.65. □

Theorem B.67. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a degree α positive-homogeneous function. Then for $p \neq 0, c \in \mathbb{R}$, $\frac{\mathrm{dV}\phi}{\mathrm{d}\Sigma}\big|_{\Sigma=\mathrm{BSB1}(p,cp)} : \mathcal{H}_B \to \mathcal{H}_B$ has the following eigendecomposition:*

- *$\mathbf{M}_B$ has eigenvalue $\mathsf{c}_\alpha p^{\alpha-1}\mathfrak{J}'_\phi(c)$*
- *$\mathbf{L}_B(a,b)$ has eigenvalue $\mathsf{c}_\alpha \alpha p^{\alpha-1}\mathfrak{J}_\phi(1)$ where $a = 2\left(\mathfrak{J}'_\phi(c) - \alpha\mathfrak{J}_\phi(1)\right)$ and $b = \alpha\mathfrak{J}_\phi(c) - c\mathfrak{J}'_\phi(c)$*

Proof. Use Thm B.62 with the computations from the proof of Thm B.65 as well as the following computation

$$\begin{aligned}
w - u &= \mathsf{c}_\alpha p^{\alpha-1}\mathfrak{J}'_\phi(c) - \mathsf{c}_\alpha \alpha p^{\alpha-1}\mathfrak{J}_\phi(1) \\
&= \mathsf{c}_\alpha p^{\alpha-1}\left(\mathfrak{J}'_\phi(c) - \alpha\mathfrak{J}_\phi(1)\right) \\
v &= \frac{1}{2}\mathsf{c}_\alpha p^{\alpha-1}\left(\alpha\mathfrak{J}_\phi(c) - c\mathfrak{J}'_\phi(c)\right)
\end{aligned}$$

□

Theorem B.68. *Let ϕ be positive-homogeneous with degree α. Assume $\frac{B-1}{2} > \alpha$. The operator $\mathcal{U} = G^{\otimes 2} \circ \frac{\mathrm{dV}(\phi\circ\mathfrak{N})(\Sigma^G)}{\mathrm{d}\Sigma^G}\big|_{\Sigma^G = v^* G} : \mathcal{H}_B^G \to \mathcal{H}_B^G$ has 3 distinct eigenvalues. They are as follows:*

1. $\lambda_{\mathbb{R}G}^{G,\mathfrak{B}_\phi}(B,\alpha) := 0$ *with dimension 1 eigenspace* $\mathbb{R}G$.
2. $$\begin{aligned}\lambda_{\mathbb{M}}^{G,\mathfrak{B}_\phi}(B,\alpha) &:= B^\alpha \upsilon^{*-1} 2^{-\alpha} \mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1} \lambda_{\mathbb{M}}^{G,\phi}(B,\alpha) \\ &= \mathsf{c}_\alpha B(B-1)^{\alpha-1}\upsilon^{*-1}2^{-\alpha}\mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)\end{aligned}$$
   with dimension $\frac{B(B-3)}{2}$ *eigenspace* $\mathbb{M}$.
3. $$\begin{aligned}\lambda_{\mathbb{L}}^{G,\mathfrak{B}_\phi}(B,\alpha) &:= B^\alpha \upsilon^{*-1} 2^{-\alpha} \mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1} \lambda_{\mathbb{L}}^{G,\phi}(B,\alpha) \\ &= \mathsf{c}_\alpha (B-1)^{\alpha-1}\upsilon^{*-1}2^{-\alpha}\mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1} \\ &\qquad \left((B-2)\alpha\left[\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)\right]+\frac{B}{B-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)\right)\end{aligned}$$
   with dimension $B-1$ *eigenspace* $\mathbb{L}$.

Proof. **Item 1. The case of λ_G.** Since $(\phi\circ\mathfrak{N})(\upsilon^* G) = (\phi\circ\mathfrak{N})(\upsilon^* G + \omega G)$ for any ω, the operator $\frac{\mathrm{dV}(\phi\circ\mathfrak{N})(\Sigma^G)}{\mathrm{d}\Sigma^G}\Big|_{\Sigma^G=\upsilon^* G}$ sends G to 0.

Item 2. The case of λ_2. Assume $\dot\Sigma(0)\in\mathbb{M}$, Thm B.66 gives (with $\odot$ denoting Hadamard product, i.e. entrywise multiplication)

$$\begin{aligned}\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\bigg|_{\Sigma=G}\{\dot\Sigma(0)\} &= \lambda_{\mathbb{M}}^{G,\phi}(B,\alpha)\dot\Sigma(0) \\ &= \mathsf{c}_\alpha\left(\frac{B-1}{B}\right)^{\alpha-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)\dot\Sigma(0)\end{aligned}$$

Since $\mathrm{tr}(\dot\Sigma(0)) = 0$, Eq. (40) gives

$$\begin{aligned}& G^{\otimes 2}\circ\frac{\mathrm{dV}(\phi\circ\mathfrak{N})(\Sigma^G)}{\mathrm{d}\Sigma^G}\bigg|_{\Sigma^G=\upsilon^* G}\{\dot\Sigma(0)\} \\ &= G^{\otimes 2}\circ\frac{d}{dt}\mathrm{V}(\phi\circ\mathfrak{N})(\Sigma(t))\bigg|_{t=0}\{\dot\Sigma(0)\} \\ &= B^\alpha\upsilon^{*-1}2^{-\alpha}\mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1} G^{\otimes 2}\circ\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\bigg|_{\Sigma=G}\left\{\dot\Sigma(0)\right\} \\ &= B^\alpha\upsilon^{*-1}2^{-\alpha}\mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1}\mathsf{c}_\alpha\left(\frac{B-1}{B}\right)^{\alpha-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)G^{\otimes 2}\{\dot\Sigma(0)\} \\ &= \mathsf{c}_\alpha B(B-1)^{\alpha-1}\upsilon^{*-1}2^{-\alpha}\mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)\dot\Sigma(0)\end{aligned}$$

So the eigenvalue for $\mathbb{M}$ is $\lambda_{\mathbb{L}}^{G,\mathfrak{B}_\phi} = \mathsf{c}_\alpha B(B-1)^{\alpha-1}\upsilon^{*-1}2^{-\alpha}\mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)$.

Item 3. The case of $\lambda_{\mathbb{L}}$. Let $\dot\Sigma(0) = \mathrm{L}(B-2,1)$. Thm B.66 gives

$$G^{\otimes 2}\circ\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\bigg|_{\Sigma=G}\{\dot\Sigma(0)\} = \lambda_{\mathbb{L}}^{G,\phi}(B,\alpha)\dot\Sigma(0)$$

where

$$\lambda_{\mathbb{L}}^{G,\phi}(B,\alpha) := \frac{1}{B}\mathsf{c}_\alpha\left(\frac{B-1}{B}\right)^{\alpha-1}\left((B-2)\alpha\left[\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)\right]+\frac{B}{B-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)\right)$$

Since $\mathrm{tr}(\dot{\Sigma}(0)) = 0$, Eq. (40) gives

$$\begin{aligned}
& G^{\otimes 2}\circ\frac{\mathrm{dV}(\phi\circ\mathfrak{N})(\Sigma^G)}{\mathrm{d}\Sigma^G}\bigg|_{\Sigma^G=\upsilon^* G}\{\dot{\Sigma}(0)\} \\
&= G^{\otimes 2}\circ\frac{d}{dt}\mathrm{V}(\phi\circ\mathfrak{N})(\Sigma(t))\bigg|_{t=0} \\
&= B^\alpha\upsilon^{*-1}2^{-\alpha}\mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1}G^{\otimes 2}\circ\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\bigg|_{\Sigma=G}\left\{\dot{\Sigma}(0)\right\} \\
&= B^\alpha\upsilon^{*-1}2^{-\alpha}\mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1}\tilde{\lambda}\dot{\Sigma}(0) \\
&= (B-1)^{\alpha-1}\mathsf{c}_\alpha\upsilon^{*-1}2^{-\alpha}\mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1} \\
&\qquad\times\left((B-2)\alpha\left[\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)\right]+\frac{B}{B-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)\right) \\
&\qquad\times\dot{\Sigma}(0)
\end{aligned}$$

So $\lambda_{\mathbb{L}}^{G,\mathfrak{B}_\phi} = (B-1)^{\alpha-1}\mathsf{c}_\alpha\upsilon^{*-1}2^{-\alpha}\mathrm{P}\left(\frac{B+1}{2},\alpha\right)^{-1}\left((B-2)\alpha\left[\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)\right]+\frac{B}{B-1}\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)\right)$ □

Proposition B.69. *With ϕ and α as above, as $B\to\infty$,*

$$\lambda_{\mathbb{L}} \sim \alpha + B^{-1}(2\alpha^3-4\alpha^2-\alpha+\frac{\mathfrak{J}'_\phi(0)}{\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi(0)}) + O(B^{-2})$$

$$\lambda_{\mathbb{M}} \sim \frac{\mathfrak{J}'_\phi(0)}{\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi(0)} + B^{-1}\left(\left(2\alpha^2-4\alpha+1\right)\frac{\mathfrak{J}'_\phi(0)}{\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi(0)} - \frac{\mathfrak{J}''_\phi(0)}{\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi(0)} - \left(\frac{\mathfrak{J}'_\phi(0)}{\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi(0)}\right)^2\right) + O\left(B^{-2}\right)$$

We can compute, by Proposition I.7,

$$\begin{aligned}
\frac{\mathfrak{J}'_\phi(0)}{\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi(0)} &= \frac{(a+b)^2\frac{1}{\sqrt{\pi}}\frac{\Gamma(\frac{\alpha}{2}+1)^2}{\Gamma(\alpha+\frac{1}{2})}}{(a^2+b^2)-(a-b)^2\frac{1}{2\sqrt{\pi}}\frac{\Gamma(\frac{\alpha}{2}+\frac{1}{2})^2}{\Gamma(\alpha+\frac{1}{2})}} \\
&\le \frac{2\Gamma\left(\frac{\alpha}{2}+1\right)^2}{\sqrt{\pi}\Gamma\left(\alpha+\frac{1}{2}\right)}
\end{aligned}$$

where the last part is obtained by optimizing over a and b. On $\alpha\in(-1/2,\infty)$, this is greater than α iff $\alpha<1$. Thus for $\alpha\ge 1$, the maximum eigenvalue of $\mathcal{U}$ is always achieved by eigenspace $\mathbb{L}$ (for large enough B)

Corollary B.70. $\lambda_{\mathbb{L}} > \lambda_{\mathbb{M}} > \lambda_G$.

C Backward Dynamics

We adopt the matrix convention that, for a multivariate function $f:\mathbb{R}^n\to\mathbb{R}^m$, the Jacobian is

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix},$$

so that $f(x+\Delta x) = \frac{\mathrm{d}f}{\mathrm{d}x}\Delta x + o(\Delta x)$, where x and Δx are treated as column vectors. In what follows we further abbreviate $\mathfrak{B}'_\phi(x) = \frac{\mathrm{d}\mathfrak{B}_\phi(z)}{\mathrm{d}z}\big|_{z=x}$, $\mathfrak{B}'_\phi(x)\Delta x = \frac{\mathrm{d}\mathfrak{B}_\phi(z)}{\mathrm{d}z}\big|_{z=x}\Delta x$.

Let's extend the definition of the V operator:

Definition C.1. Suppose $T_x : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator parametrized by $x \in \mathbb{R}^k$. Then for a PSD matrix $\Sigma \in \mathcal{S}_k$, define $\mathrm{V}T(\Sigma) := \mathrm{E}[T_x \otimes T_x : x \in \mathcal{N}(0,\Sigma)] : \mathbb{R}^{n\times n} \to \mathbb{R}^{m\times m}$, which acts on $n \times n$ matrices Π by

$$\mathrm{V}T(\Sigma)\{\Pi\} = \mathrm{E}[T_x \Pi T_x^T : x \in \mathcal{N}(0,\Sigma)] \in \mathbb{R}^{m\times m}.$$

Under this definition, $\mathrm{V}\mathfrak{B}'_\phi(\Sigma)$ is a linear operator $\mathbb{R}^{B\times B} \to \mathbb{R}^{B\times B}$. Recall the notion of adjoint:

Definition C.2. If V is a (real) vector space, then $V^\dagger$, its *dual space*, is defined as the space of linear functionals $f : V \to \mathbb{R}$ on V. If $T : V \to W$ is a linear operator, then $T^\dagger$, its *adjoint*, is a linear operator $W^\dagger \to V^\dagger$, defined by $T^\dagger(f) = v \mapsto f(T(v))$.

If a linear operator is represented by a matrix, with function application represented by matrix-vector multiplication (matrix on the left, vector on the right), then the adjoint is represented by matrix transpose.

The backward equation. In this section we are interested in the *backward dynamics*, given by the following equation

$$\Pi^{(l)} = \mathrm{V}\mathfrak{B}'_\phi(\Sigma^{(l)})^\dagger\{\Pi^{(l+1)}\}$$

where $\Sigma^{(l)}$ is given by the forward dynamics. Particularly, we are interested in the specific case when we have exponential convergence of $\Sigma^{(l)}$ to a BSB1 fixed point. Thus we will study the asymptotic approximation of the above, namely

$$\Pi^{(l)} = \mathrm{V}\mathfrak{B}'_\phi(\Sigma^*)^\dagger\{\Pi^{(l+1)}\} \tag{41}$$

where Σ^* is the BSB1 fixed point. This is a linear system, so its dynamics is entirely determined by the eigendecomposition of the operator $\mathrm{V}\mathfrak{B}'_\phi(\Sigma^*)^\dagger$. It turns out to be much more convenient to study $\mathrm{V}\mathfrak{B}'_\phi(\Sigma^*)$, which has the same eigenvalues as its adjoint.

Let us start then.

C.1 Eigendecomposition of $\mathrm{V}\mathfrak{B}'_\phi(\Sigma^*)$.

By chain rule, $\mathfrak{B}'_\phi(x) = \frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\big|_{z=Gx}\frac{\mathrm{d}Gx}{\mathrm{d}x} = \frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\big|_{z=Gx}G$, $\mathfrak{B}'_\phi(x)\Delta x = \frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\big|_{z=Gx}G\Delta x$ This

$$\begin{aligned}
\mathrm{V}\mathfrak{B}'_\phi(\Sigma) &= \mathrm{E}\left[\mathfrak{B}'_\phi(x)^{\otimes 2} : x \sim \mathcal{N}(0,\Sigma)\right] \\
&= \mathrm{E}\left[\left(\frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\bigg|_{z=Gx} G\right)^{\otimes 2} : x \sim \mathcal{N}(0,\Sigma)\right] \\
&= \mathrm{E}\left[\frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\bigg|_{z=Gx}^{\otimes 2} \circ G^{\otimes 2} : x \sim \mathcal{N}(0,\Sigma)\right] \\
&= \mathrm{E}\left[\frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\bigg|_{z=Gx}^{\otimes 2} : x \sim \mathcal{N}(0,\Sigma)\right] \circ G^{\otimes 2} \\
&= \mathrm{V}\left[x \mapsto \frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\bigg|_{z=Gx}\right](\Sigma) \circ G^{\otimes 2}
\end{aligned}$$

If we let $F(\Sigma) := \mathrm{V}\left[x \mapsto \frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\Big|_{z=Gx}\right](\Sigma)$, then $\mathrm{V}\mathfrak{B}'_\phi(\Sigma) = F(\Sigma)\circ G^{\otimes 2}$. By Lemma B.42, to obtain the eigendecomposition of $\mathrm{V}\mathfrak{B}'_\phi(\Sigma)$ it suffices to do the same for $G^{\otimes 2}\circ F(\Sigma)$.

Proposition C.3.

$$\left.\frac{\mathrm{d}\mathfrak{N}(z)}{\mathrm{d}z}\right|_{z=y} = \sqrt{B} r^{-1}(I - vv^T),$$

where $r = \|y\|, v = y/\|y\| = \mathfrak{N}(y)/\sqrt{B}$.

Proof. We have $\left.\frac{\mathrm{d}\mathfrak{N}(z)}{\mathrm{d}z}\right|_{z=y} = \sqrt{B} \left.\frac{\mathrm{d}z/\|z\|}{\mathrm{d}z}\right|_{z=y}$, and

$$\begin{aligned}
\frac{\partial(y_i/\|y\|)}{\partial y_j} &= \frac{\delta_{ij}\|y\| - (\partial\|y\|/\partial y_j)y_i}{\|y\|^2} \\
&= \frac{\delta_{ij}}{\|y\|} - \frac{y_i}{\|y\|^2}\frac{\partial\sqrt{y\cdot y}}{\partial y_j} \\
&= \frac{\delta_{ij}}{\|y\|} - \frac{y_i}{\|y\|^2}\frac{y_j}{\|y\|} \\
&= \frac{\delta_{ij}}{\|y\|} - \frac{y_i y_j}{\|y\|^3}
\end{aligned}$$

so that

$$\left.\frac{\mathrm{d}z/\|z\|}{\mathrm{d}z}\right|_{z=y} = r^{-1}(I - vv^T).$$

□

By chain rule, this easily gives

Proposition C.4.

$$\left.\frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\right|_{z=y} = \sqrt{B} D r^{-1}(I - vv^T),$$

where $D = \mathrm{Diag}(\phi'(\mathfrak{N}(y))), r = \|y\|, v = y/\|y\| = \mathfrak{N}(y)/\sqrt{B}$.

With $v = y/\|y\|$, $r = \|y\|$, and $\tilde{D} = \mathrm{Diag}(\phi'(\mathfrak{N}(y)))$, we then have

$$\begin{aligned}
F(\Sigma) &= \mathrm{E}\left[\tilde{D}^{\otimes 2}\circ\left(\sqrt{B}\left.\frac{\mathrm{d}z/\|z\|}{\mathrm{d}z}\right|_{z=Gx}\right)^{\otimes 2} : x\sim\mathcal{N}(0,\Sigma)\right] \\
&= \mathrm{E}\left[\tilde{D}^{\otimes 2}\circ\left(\sqrt{B}\left.\frac{\mathrm{d}z/\|z\|}{\mathrm{d}z}\right|_{z=y}\right)^{\otimes 2} : y\sim\mathcal{N}(0,G\Sigma G)\right] \\
&= B\,\mathrm{E}\left[r^{-2}\left(\tilde{D}^{\otimes 2} + (\tilde{D}vv^T)^{\otimes 2} - \tilde{D}\otimes(\tilde{D}vv^T) - (\tilde{D}vv^T)\otimes\tilde{D}\right) : y\sim\mathcal{N}(0,G\Sigma G)\right]
\end{aligned} \tag{42}$$

$$F(\Sigma)\{\Lambda\} = B\,\mathrm{E}\left[r^{-2}\left(\tilde{D}\Lambda\tilde{D} + \tilde{D}vv^T\Lambda vv^T\tilde{D} - \tilde{D}\Lambda vv^T\tilde{D} - \tilde{D}vv^T\Lambda\tilde{D}\right) : y\sim\mathcal{N}(0,G\Sigma G)\right]$$

C.1.1 SPHERICAL INTEGRAL

$$
\begin{aligned}
F(\Sigma^*)\{\Lambda\} &= B\,\mathrm{E}\left[r^{-2}\left(\tilde{D}\Lambda\tilde{D} + \tilde{D}vv^T\Lambda vv^T\tilde{D} - \tilde{D}\Lambda vv^T\tilde{D} - \tilde{D}vv^T\Lambda\tilde{D}\right) : y \sim \mathcal{N}(0, \upsilon^* G)\right] \\
&= \upsilon^{*-1} B\,\mathrm{E}\left[r^{-2}\left(\tilde{D}\Lambda\tilde{D} + \tilde{D}vv^T\Lambda vv^T\tilde{D} - \tilde{D}\Lambda vv^T\tilde{D} - \tilde{D}vv^T\Lambda\tilde{D}\right) : y \sim \mathcal{N}(0, G)\right] \\
F(\Sigma^*)\{\tau_{ij}\}_{kl} &= \upsilon^{*-1} B\,\mathrm{E}[r^{-2}\Big(\phi'(\sqrt{B}v_k)\phi'(\sqrt{B}v_l)\frac{\mathbb{I}(k=i\ \ \&\ \ l=j) + \mathbb{I}(k=j\ \ \&\ \ l=i)}{2} \\
&\qquad + v_i v_j v_k \phi'(\sqrt{B}v_k) v_l \phi'(\sqrt{B}v_l) \\
&\qquad - \frac{1}{2}\big(\mathbb{I}(i=k)\phi'(\sqrt{B}v_k)v_j v_l \phi'(\sqrt{B}v_l) + \mathbb{I}(j=k)\phi'(\sqrt{B}v_k)v_i v_l \phi'(\sqrt{B}v_l) \\
&\qquad\quad + \mathbb{I}(i=l)\phi'(\sqrt{B}v_l)v_j v_k \phi'(\sqrt{B}v_k) + \mathbb{I}(j=l)\phi'(\sqrt{B}v_l)v_i v_k \phi'(\sqrt{B}v_k)\big)\Big) \\
&\quad : rv \sim \mathcal{N}(0, G)] \\
&= \upsilon^{*-1} B \frac{\Gamma((B-3)/2)}{\Gamma((B-5)/2)} 2^{-2/2}\pi^{-2} \int_0^\pi \mathrm{d}\theta_1 \cdots \int_0^\pi \mathrm{d}\theta_4\, f(v_1^\theta, \ldots, v_4^\theta) \sin^{B-3}\theta_1 \cdots \sin^{B-6}\theta_4 \\
&= \upsilon^{*-1} B(B-5)(2\pi)^{-2} \int_0^\pi \mathrm{d}\theta_1 \cdots \int_0^\pi \mathrm{d}\theta_4\, f(v_1^\theta, \ldots, v_4^\theta) \sin^{B-3}\theta_1 \cdots \sin^{B-6}\theta_4
\end{aligned}
$$

by Lemma B.47, where we assume, WLOG by ultrasymmetry, $k, l \in \{1, 2\}; i, j \in \{1, \ldots, 4\}$, and

$$
\begin{aligned}
f(v_1, \ldots, v_4) := \frac{1}{2}\phi'(\sqrt{B}\mathbb{e}v)_k \phi'(\sqrt{B}\mathbb{e}v)_l \big(& \\
& \mathbb{I}(k=i\ \ \&\ \ l=j) + \mathbb{I}(k=j\ \ \&\ \ l=i) + 2(\mathbb{e}v)_i(\mathbb{e}v)_j(\mathbb{e}v)_k(\mathbb{e}v)_l \\
& - (\mathbb{I}(i=k)(\mathbb{e}v)_j(\mathbb{e}v)_l + \mathbb{I}(j=k)(\mathbb{e}v)_i(\mathbb{e}v)_l + \mathbb{I}(i=l)(\mathbb{e}v)_j(\mathbb{e}v)_k + \mathbb{I}(j=l)(\mathbb{e}v)_i(\mathbb{e}v)_k) \\
& \big) \\
v_1^\theta &= \cos\theta_1 \\
v_2^\theta &= \sin\theta_1 \cos\theta_2 \\
v_3^\theta &= \sin\theta_1 \sin\theta_2 \cos\theta_3 \\
v_4^\theta &= \sin\theta_1 \sin\theta_2 \sin\theta_3 \cos\theta_4.
\end{aligned}
$$

and $\mathbb{e}$ as in Eq. (39).

We can integrate out θ_3 and θ_4 symbolically by Lemma B.20 since their dependence only appear outside of ϕ'. This reduces each entry of $F(\Sigma^*)\{\tau_{ij}\}_{kl}$ to 2-dimensional integrals to evaluate numerically.

C.1.2 LAPLACE METHOD

In this section suppose that ϕ is degree α positive-homogeneous. Set $D = \mathrm{Diag}(\phi'(y))$ (and recall $v = y/\|y\|, r = \|y\|$). Then D is degree $\alpha - 1$ positive-homogeneous in x (because ϕ' is). Consequently we can rewrite Eq. (42) as follows,

$$
\begin{aligned}
F(\Sigma^*) &= B^\alpha \mathop{\mathrm{E}}_{y \sim \mathcal{N}(0, G\Sigma^* G)}\left[r^{-2}\left(r^{-2(\alpha-1)}D^{\otimes 2} + r^{-2(\alpha+1)}(Dvv^T)^{\otimes 2} - r^{-2\alpha}D \otimes (Dvv^T) - r^{-2\alpha}(Dvv^T) \otimes D\right)\right] \\
&= B^\alpha \mathop{\mathrm{E}}_{y \sim \mathcal{N}(0, \upsilon^* G)}\left[r^{-2\alpha}D^{\otimes 2} + r^{-2(\alpha+2)}(Dyy^T)^{\otimes 2} - r^{-2(\alpha+1)}\left(D \otimes (Dyy^T) + (Dyy^T) \otimes D\right)\right] \\
&= B^\alpha(\mathsf{A} + \mathsf{B} - \mathsf{C}) \qquad (43)
\end{aligned}
$$

where Σ^* is the BSB1 fixed point of Eq. (28), $G\Sigma^* G = \upsilon^* G$, and

$$
\begin{aligned}
\mathsf{A} &= \mathrm{E}\left[r^{-2\alpha}D^{\otimes 2} : y \sim \mathcal{N}(0, \upsilon^* G)\right] \\
\mathsf{B} &= \mathrm{E}\left[r^{-2(\alpha+2)}(Dyy^T)^{\otimes 2} : y \sim \mathcal{N}(0, \upsilon^* G)\right] \\
\mathsf{C} &= \mathrm{E}\left[r^{-2(\alpha+1)}\left(D \otimes (Dyy^T) + (Dyy^T) \otimes D\right) : y \sim \mathcal{N}(0, \upsilon^* G)\right]
\end{aligned}
$$

Each term in the sum above is ultrasymmetric, so has the same eigenspaces $\mathbb{R}G, \mathbb{M}, \mathbb{L}$ (this will also become apparent in the computations below without resorting to Lemma B.37). We can thus compute the eigenvalues for each of them in order.

In all computation below, we apply Lemma B.5 first to relate the quantity in question to $\mathrm{V}\phi$.

Computing A. For two matrices Σ, Λ, write $\Sigma \odot \Lambda$ for entrywise multiplication. We have by Lemma B.5

$$\begin{aligned}
\mathsf{A}\{\Lambda\} &= \mathrm{E}\left[r^{-2\alpha} D\Lambda D : y \sim \mathcal{N}(0, \upsilon^* G)\right] \\
&= \mathrm{E}\left[r^{-2\alpha} \Lambda \odot \phi'(y)\phi'(y)^T : y \sim \mathcal{N}(0, \upsilon^* G)\right] \\
&= \Gamma(\alpha)^{-1} \int_0^\infty \mathrm{d}s\, s^{\alpha-1} \det(I + 2s\upsilon^* G)^{-1/2} \Lambda \odot \mathrm{V}\phi'\left(\frac{\upsilon^*}{1+2s\upsilon^*} G\right) \\
&= \Gamma(\alpha)^{-1} \int_0^\infty \mathrm{d}s\, s^{\alpha-1} (1+2s\upsilon^*)^{-(B-1)/2} \Lambda \odot \left(\frac{\upsilon^*}{1+2s\upsilon^*}\right)^{\alpha-1} \mathrm{V}\phi'(G) \\
&= \Gamma(\alpha)^{-1} \Lambda \odot \mathrm{V}\phi'(G) \int_0^\infty \mathrm{d}s\, (\upsilon^* s)^{\alpha-1} (1+2s\upsilon^*)^{-(B-1)/2-\alpha+1} \\
&= \Gamma(\alpha)^{-1} \Lambda \odot \mathrm{V}\phi'(G)\, \mathrm{Beta}\left(\frac{B-3}{2}, \alpha\right) \upsilon^{*-1} 2^{-\alpha} \\
&= \Lambda \odot \mathsf{c}_{\alpha-1} \left(\frac{B-1}{B}\right)^{\alpha-1} \mathrm{BSB1}\left(\mathfrak{J}_{\phi'}(1), \mathfrak{J}_{\phi'}\left(\frac{-1}{B-1}\right)\right) \mathrm{P}\left(\frac{B-3}{2}, \alpha\right)^{-1} \upsilon^{*-1} 2^{-\alpha}
\end{aligned}$$

Then Thm B.63 gives the eigendecomposition for $B^\alpha G^{\otimes 2} \circ \mathsf{A} \restriction \mathcal{H}_B^G$

Theorem C.5. *Let $R_{B,\alpha} := (2\alpha-1)\mathsf{c}_{\alpha-1} B(B-1)^{\alpha-1} \mathrm{P}\left(\frac{B-3}{2}, \alpha\right)^{-1} 2^{-\alpha}$. Then $B^\alpha G^{\otimes 2} \circ \mathsf{A} \restriction \mathcal{H}_B^G$ has the following eigendecomposition.*

$$\begin{aligned}
\lambda_{\mathbb{L}}^{\mathsf{A}} &= R_{B,\alpha} \upsilon^{*-1} \left(\frac{B-2}{B} \mathfrak{J}_\phi'(1) + \frac{2}{B} \mathfrak{J}_\phi'\left(\frac{-1}{B-1}\right)\right) \\
&= \frac{(2\alpha+B-3)\left(\alpha^2(B-2)\mathfrak{J}_\phi(1) + 2(2\alpha-1)\mathfrak{J}_\phi'\left(\frac{1}{1-B}\right)\right)}{(2\alpha-1)(B-3)(B-1)\left(\mathfrak{J}(1) - \mathfrak{J}\left(\frac{1}{1-B}\right)\right)} \\
\lambda_{\mathbb{M}}^{\mathsf{A}} &= R_{B,\alpha} \upsilon^{*-1} \mathfrak{J}_\phi'\left(\frac{-1}{B-1}\right) \\
&= \frac{B(2\alpha+B-3)\mathfrak{J}_\phi'\left(\frac{1}{1-B}\right)}{(B-3)(B-1)\left(\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi\left(\frac{1}{1-B}\right)\right)} \\
\lambda_{G}^{\mathsf{A}} &= R_{B,\alpha} \upsilon^{*-1} \left(\frac{B-1}{B} \mathfrak{J}_\phi'(1) + \frac{1}{B} \mathfrak{J}_\phi'\left(\frac{-1}{B-1}\right)\right) \\
&= \frac{(2\alpha+B-3)\left(\alpha'^2(B-1)\mathfrak{J}_\phi(1) + (2\alpha-1)\mathfrak{J}_\phi\left(\frac{1}{1-B}\right)\right)}{(2\alpha-1)(B-3)(B-1)\left(\mathfrak{J}(1) - \mathfrak{J}\left(\frac{1}{1-B}\right)\right)}
\end{aligned}$$

Proof. We have $B^\alpha \mathsf{A} = (2\alpha-1)^{-1} R_{B,\alpha} \upsilon^{*-1} \mathrm{DOS}\left(\mathfrak{J}_{\phi'}(1), 0, \mathfrak{J}_{\phi'}\left(\frac{-1}{B-1}\right)\right)$. This allows us to apply Thm B.63. We make a further simplification $\mathfrak{J}_{\phi'}(c) = (2\alpha-1)\mathfrak{J}_\phi'(c)$ via Proposition I.6. □

Computing B.

$$\begin{aligned}
\mathsf{B}\{\Lambda\} &= \mathrm{E}\left[r^{-2(\alpha+2)} D y y^T \Lambda y y^T D : y \sim \mathcal{N}(0, \upsilon^* G)\right] \\
&= \mathrm{E}\left[r^{-2(\alpha+2)} \psi(y) y^T \Lambda y \psi(y)^T : y \sim \mathcal{N}(0, \upsilon^* G)\right]
\end{aligned}$$

where $\psi(y) := y\phi'(y)$, which, in this case of ϕ being degree α positive-homogeneous, is also equal to $\alpha\phi(y)$.

By Lemma B.5, $\mathsf{B}\{\Lambda\}$ equals

$$\Gamma(\alpha+2)^{-1}\int_0^\infty \mathrm{d}s\, s^{\alpha+1}\det(I+2sv^*G)^{-1/2}\,\mathrm{E}[\psi(y)y^T\Lambda y\psi(y)^T : y\sim\mathcal{N}(0,\frac{v^*}{1+2sv^*}G)]$$

Definition C.6. Let $\psi:\mathbb{R}\to\mathbb{R}$ be measurable and $\Sigma\in\mathcal{S}_B$. Define $\mathrm{V}^{(4)}\psi(\Sigma):\mathcal{H}_B\to\mathcal{H}_B$ by

$$\mathrm{V}^{(4)}\psi(\Sigma)\{\Lambda\} = \mathrm{E}[\psi(y)y^T\Lambda y\psi(y)^T : y\sim\mathcal{N}(0,\Sigma)]$$

For two matrices Σ,Λ, write $\langle\Sigma,\Lambda\rangle := \mathrm{tr}(\Sigma^T\Lambda)$.

Proposition C.7. $\mathrm{V}^{(4)}\phi(\Sigma)\{\Lambda\} = \mathrm{V}\phi(\Sigma)\langle\Sigma,\Lambda\rangle + 2\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\{\Sigma\Lambda\Sigma\}$.

Proof. We have

$$\begin{aligned}
\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\{\Lambda\} &= (2\pi)^{-B/2}\int \mathrm{d}z\,\phi(z)^{\otimes 2}\left(\frac{\mathrm{d}}{\mathrm{d}\Sigma}\det(\Sigma)^{-1/2}e^{-\frac{1}{2}z^T\Sigma^{-1}z}\right)\{\Lambda\}\\
&= (2\pi)^{-B/2}\int \mathrm{d}z\,\phi(z)^{\otimes 2}\left\langle\left[\frac{-1}{2}\det\Sigma^{-1/2}\Sigma^{-1}+\det\Sigma^{-1/2}\frac{1}{2}\Sigma^{-1}zz^T\Sigma^{-1}\right]e^{-\frac{1}{2}z^T\Sigma^{-1}z},\Lambda\right\rangle\\
&= \frac{1}{2}(2\pi)^{-B/2}\det\Sigma^{-1/2}\int \mathrm{d}z\,\phi(z)^{\otimes 2}e^{-\frac{1}{2}z^T\Sigma^{-1}z}\left[\langle zz^T,\Sigma^{-1}\Lambda\Sigma^{-1}\rangle - \langle\Sigma^{-1},\Lambda\rangle\right]\\
&= \frac{1}{2}\mathrm{V}^{(4)}\phi(\Sigma)\{\Sigma^{-1}\Lambda\Sigma^{-1}\} - \frac{1}{2}\mathrm{V}\phi(\Sigma)\langle\Sigma^{-1},\Lambda\rangle
\end{aligned}$$

Making the substitution $\Lambda\to\Sigma\Lambda\Sigma$, we get the desired result. □

If φ is degree α positive-homogeneous, then $\mathrm{V}^{(4)}\varphi$ is degree $\alpha+1$ positive-homogeneous. Thus,

$$\begin{aligned}
\mathsf{B}\{\Lambda\} &= \Gamma(\alpha+2)^{-1}\int_0^\infty \mathrm{d}s\, s^{\alpha+1}\det(I+2sv^*G)^{-1/2}\mathrm{V}^{(4)}\psi\left(\frac{v^*}{1+2sv^*}G\right)\{\Lambda\}\\
&= \Gamma(\alpha+2)^{-1}\int_0^\infty \mathrm{d}s\, s^{\alpha+1}(1+2sv^*)^{-(B-1)/2}\left(\frac{v^*}{1+2sv^*}\right)^{\alpha+1}\mathrm{V}^{(4)}\psi(G)\{\Lambda\}\\
&= \Gamma(\alpha+2)^{-1}\mathrm{V}^{(4)}\psi(G)\{\Lambda\}\int_0^\infty \mathrm{d}s\,(v^*s)^{\alpha+1}(1+2sv^*)^{-(B-1)/2-\alpha-1}\\
&= \Gamma(\alpha+2)^{-1}\mathrm{V}^{(4)}\psi(G)\{\Lambda\}\,\mathrm{Beta}(\frac{B-3}{2},\alpha+2)2^{-\alpha-2}v^{*-1}\\
&= \mathrm{P}\left(\frac{B-3}{2},\alpha+2\right)^{-1}2^{-\alpha-2}v^{*-1}\mathrm{V}^{(4)}\psi(G)\{\Lambda\}
\end{aligned}$$

By Proposition C.7, $\mathrm{V}^{(4)}\psi(G)\{\Lambda\} = \mathrm{V}\psi(G)\,\mathrm{tr}\,\Lambda + 2\frac{\mathrm{dV}\psi(\Sigma)}{\mathrm{d}\Sigma}\{\Lambda\}$ if $\Lambda\in\mathcal{H}_B^G$. Thus

Theorem C.8. *$B^\alpha G^{\otimes 2}\circ\mathsf{B}\restriction\mathcal{H}_B^G$ has the following eigendecomposition (note that here v^* is still with respect to ϕ, not $\psi=\alpha\phi$)*

1. *Eigenspace $\mathbb{R}G$ with eigenvalue*

$$\lambda_G^{\mathsf{B}} := \frac{\alpha^2}{B-3}$$

2. *Eigenspace $\mathbb{M}$ with eigenvalue*

$$\begin{aligned}
\lambda_{\mathbb{M}}^{\mathsf{B}} &:= B^\alpha\mathrm{P}\left(\frac{B-3}{2},\alpha+2\right)^{-1}2^{-\alpha-2}v^{*-1}2\lambda_{\mathbb{M}}^{G,\psi}(B,\alpha)\\
&= \frac{2\alpha^2B\mathfrak{J}_\phi'\left(\frac{-1}{B-1}\right)}{(B-3)(B-1)(B-1+2\alpha)(\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right))}
\end{aligned}$$

3. *Eigenspace* $\mathbb{L}$ *with eigenvalue*

$$\lambda_{\mathbb{L}}^{\mathsf{B}} := B^{\alpha}\mathrm{P}\left(\frac{B-3}{2}, \alpha+2\right)^{-1} 2^{-\alpha-2}\upsilon^{*-1}2\lambda_{\mathbb{L}}^{G,\psi}(B,\alpha)$$

$$= \frac{2\alpha^3(B-2)}{(B-3)(B-1)(B-1+2\alpha)} + \frac{2\alpha^2 B\mathfrak{J}_{\phi}'\left(\frac{-1}{B-1}\right)}{(B-3)(B-1)^2(B-1+2\alpha)(\mathfrak{J}_{\phi}(1)-\mathfrak{J}_{\phi}\left(\frac{-1}{B-1}\right))}$$

Proof. The only thing to justify is the value of λ_G^{B}. We have

$$\lambda_G^{\mathsf{B}} = B^{\alpha}\mathrm{P}\left(\frac{B-3}{2}, \alpha+2\right)^{-1} 2^{-\alpha-2}\upsilon^{*-1}\left((B-1)\mathsf{c}_{\alpha}\left(\frac{B-1}{B}\right)^{\alpha}\left(\mathfrak{J}_{\psi}(1)-\mathfrak{J}_{\psi}\left(\frac{-1}{B-1}\right)\right)+2\lambda_{\mathbb{M}}^{G,\psi}(B,\alpha)\right)$$

$$= B^{\alpha}\mathrm{P}\left(\frac{B-3}{2}, \alpha+2\right)^{-1} 2^{-\alpha-2}\upsilon^{*-1}(B-1+2\alpha)\mathsf{c}_{\alpha}\left(\frac{B-1}{B}\right)^{\alpha}\left(\mathfrak{J}_{\psi}(1)-\mathfrak{J}_{\psi}\left(\frac{-1}{B-1}\right)\right)$$

$$= \mathsf{c}_{\alpha}(B-1)^{\alpha}(B-1+2\alpha)\mathrm{P}\left(\frac{B-3}{2}, \alpha+2\right)^{-1} 2^{-\alpha-2}\upsilon^{*-1}\left(\mathfrak{J}_{\psi}(1)-\mathfrak{J}_{\psi}\left(\frac{-1}{B-1}\right)\right)$$

$$= \mathsf{c}_{\alpha}(B-1)^{\alpha}\mathrm{P}\left(\frac{B-3}{2}, \alpha+1\right)^{-1} 2^{-\alpha-1}\upsilon^{*-1}\left(\mathfrak{J}_{\psi}(1)-\mathfrak{J}_{\psi}\left(\frac{-1}{B-1}\right)\right)$$

$$= \mathsf{c}_{\alpha}(B-1)^{\alpha}\mathrm{P}\left(\frac{B-3}{2}, \alpha+1\right)^{-1} 2^{-\alpha-1}K_{\alpha,B}^{-1}\alpha^2$$

(since $\mathfrak{J}_{\psi} = \alpha^2\mathfrak{J}_{\phi}$)

$$= \frac{\alpha^2\mathsf{c}_{\alpha}(B-1)^{\alpha}\mathrm{P}\left(\frac{B-3}{2}, \alpha+1\right)^{-1} 2^{-\alpha-1}}{\mathsf{c}_{\alpha}\mathrm{P}\left(\frac{B-1}{2}, \alpha\right)^{-1}\left(\frac{B-1}{2}\right)^{\alpha}}$$

(Defn B.23)

$$= \frac{\alpha^2 2^{-1}}{(B-3)/2}$$

$$= \frac{\alpha^2}{B-3}$$

□

Computing C. By Lemma B.5,

$$\mathsf{C}\{\Lambda\} = \mathrm{E}\left[r^{-2(\alpha+1)}\left(D\Lambda yy^T D + Dyy^T\Lambda D\right) : y \sim \mathcal{N}(0, \upsilon^* G)\right]$$

$$= \Gamma(\alpha+1)^{-1}\int \mathrm{d}s\, s^{\alpha}\det(I+2s\upsilon^* G)^{-1/2}\,\mathrm{E}[D\Lambda yy^T D + Dyy^T\Lambda D : y \sim \mathcal{N}(0, \frac{\upsilon^*}{1+2s\upsilon^*}G)]$$

$$= \Gamma(\alpha+1)^{-1}\int \mathrm{d}s\, s^{\alpha}(1+2s\upsilon^*)^{-(B-1)/2}\left(\frac{\upsilon^*}{1+2s\upsilon^*}\right)^{\alpha}\mathrm{E}[D\Lambda yy^T D + Dyy^T\Lambda D : y \sim \mathcal{N}(0, G)]$$

$$= \Gamma(\alpha+1)^{-1}\int \mathrm{d}s\, (\upsilon^* s)^{\alpha}(1+2s\upsilon^*)^{-(B-1)/2-\alpha}\,\mathrm{E}[D\Lambda yy^T D + Dyy^T\Lambda D : y \sim \mathcal{N}(0, G)]$$

$$= \Gamma(\alpha+1)^{-1}\,\mathrm{Beta}\left(\frac{B-3}{2}, \alpha+1\right)2^{-\alpha-1}\upsilon^{*-1}\,\mathrm{E}[D\Lambda yy^T D + Dyy^T\Lambda D : y \sim \mathcal{N}(0, G)]$$

$$= \mathrm{P}\left(\frac{B-3}{2}, \alpha+1\right)^{-1} 2^{-\alpha-1}\upsilon^{*-1}\,\mathrm{E}[D\Lambda yy^T D + Dyy^T\Lambda D : y \sim \mathcal{N}(0, G)]$$

Lemma C.9. *Suppose* ϕ *is degree* α *positive-homogeneous. Then for* $\Lambda \in \mathcal{H}_B$,

$$\mathrm{E}[D \otimes (Dyy^T) + Dyy^T \otimes D : y \sim \mathcal{N}(0, \Sigma)] = \alpha\left.\frac{\mathrm{d}\mathrm{V}\phi(\Pi\Sigma\Pi)}{\mathrm{d}\Pi}\right|_{\Pi=I}$$

$$\mathrm{E}[D\Lambda yy^T D + Dyy^T\Lambda D : y \sim \mathcal{N}(0, \Sigma)] = \alpha\frac{\mathrm{d}\mathrm{V}\phi(\Sigma)}{\mathrm{d}\Sigma}\{\Sigma\Lambda + \Lambda\Sigma\}$$

where $D = \mathrm{Diag}(\phi'(y))$.

Proof. Let $\Pi_t, t \in (-\epsilon, \epsilon)$ be a smooth path in $\mathcal{H}_B$ with $\Pi_0 = I$. Write $D_t = \mathrm{Diag}(\phi'(\Pi_t y))$, so that $D_0 = D$. Then, using $\dot{\Pi}_t$ to denote t derivative,

$$\begin{aligned}
\frac{d}{dt}\mathrm{V}\phi(\Pi_t \Sigma \Pi_t) &= \frac{d}{dt}\,\mathrm{E}[\phi(\Pi_t y)\phi(\Pi_t y)^T : y \sim \mathcal{N}(0,\Sigma)] \\
&= \mathrm{E}[D_t \dot{\Pi}_t y \phi(\Pi_t y)^T + \phi(\Pi_t y) y^T \dot{\Pi}_t D_t : y \sim \mathcal{N}(0,\Sigma)] \\
\frac{d}{dt}\mathrm{V}\phi(\Pi_t \Sigma \Pi_t)\bigg|_{t=0} &= \mathrm{E}[D \dot{\Pi}_0 y \phi(y)^T + \phi(y) y^T \dot{\Pi}_0 D : y \sim \mathcal{N}(0,\Sigma)]
\end{aligned}$$

Because for $x \in \mathbb{R}$, $\alpha\phi(x) = x\phi'(x)$, for $y \in \mathbb{R}^B$ we can write $\phi(y) = \alpha^{-1}\mathrm{Diag}(\phi'(y))y$. Then

$$\begin{aligned}
\alpha\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\{\Sigma\dot{\Pi}_0 + \dot{\Pi}_0\Sigma\} &= \alpha\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\left\{\frac{d\Pi_t\Sigma\Pi_t}{dt}\right\} \\
&= \alpha\,\frac{d}{dt}\mathrm{V}\phi(\Pi_t\Sigma\Pi_t)\bigg|_{t=0} \\
&= \mathrm{E}[D\dot{\Pi}_0 y y^T D + D y y^T \dot{\Pi}_0 D : y \sim \mathcal{N}(0,\Sigma)]
\end{aligned}$$

□

Therefore, for $\Lambda \in \mathcal{H}_B^G$,

$$\begin{aligned}
\mathsf{C}\{\Lambda\} &= \mathrm{P}\left(\frac{B-3}{2}, \alpha+1\right)^{-1} 2^{-\alpha-1}\upsilon^{*-1}\alpha\,\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\bigg|_{\Sigma=G}\{G\Lambda + \Lambda G\} \\
&= 2\alpha\mathrm{P}\left(\frac{B-3}{2}, \alpha+1\right)^{-1} 2^{-\alpha-1}\upsilon^{*-1}\,\frac{\mathrm{dV}\phi(\Sigma)}{\mathrm{d}\Sigma}\bigg|_{\Sigma=G}\{\Lambda\}
\end{aligned}$$

So Thm B.66 gives

Theorem C.10. $B^\alpha G^{\otimes 2} \circ \mathsf{C}$ *has the following eigendecomposition*

1. *eigenspace* $\mathbb{R}G$ *with eigenvalue*

$$\begin{aligned}
\lambda_G^{\mathsf{C}} &:= 2B^\alpha\alpha\mathrm{P}\left(\frac{B-3}{2}, \alpha+1\right)^{-1} 2^{-\alpha-1}\upsilon^{*-1}\lambda_{\mathbb{R}G}^{G,\phi}(B,\alpha) \\
&= \frac{2\alpha^2}{B-3}
\end{aligned}$$

2. *eigenspace* $\mathbb{L}$ *with eigenvalue*

$$\begin{aligned}
\lambda_{\mathbb{L}}^{\mathsf{C}} &:= 2B^\alpha\alpha\mathrm{P}\left(\frac{B-3}{2}, \alpha+1\right)^{-1} 2^{-\alpha-1}\upsilon^{*-1}\lambda_{\mathbb{L}}^{G,\phi}(B,\alpha) \\
&= \frac{2\alpha^2(B-2)}{(B-3)(B-1)} + \frac{2\alpha B\mathfrak{J}_\phi'\left(\frac{-1}{B-1}\right)}{(B-3)(B-1)^2(\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi\left(\frac{-1}{B-1}\right))}
\end{aligned}$$

3. *eigenspace* $\mathbb{M}$ *with eigenvalue*

$$\begin{aligned}
\lambda_{\mathbb{M}}^{\mathsf{C}} &:= 2B^\alpha\alpha\mathrm{P}\left(\frac{B-3}{2}, \alpha+1\right)^{-1} 2^{-\alpha-1}\upsilon^{*-1}\lambda_{\mathbb{M}}^{G,\phi}(B,\alpha) \\
&= \frac{2\alpha B\mathfrak{J}_\phi'\left(\frac{-1}{B-1}\right)}{(B-3)(B-1)(\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)}
\end{aligned}$$

Altogether, by Eq. (43) and Thms C.5, C.8 and C.10, this implies

Theorem C.11. *$G^{\otimes 2} \circ F(\Sigma^*) : \mathcal{H}_B^G \to \mathcal{H}_B^G$ has eigenspaces $\mathbb{R}G, \mathbb{M}, \mathbb{L}$ respectively with the following eigenvalues*

$$\begin{aligned}
\lambda_G^{G,\mathfrak{B}'_\phi} &= \frac{(B-3+2\alpha)\left((2\alpha-1)\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)+\alpha^2(B-1)\mathfrak{J}_\phi(1)\right)}{(2\alpha-1)(B-3)(B-1)(\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right))} - \frac{\alpha^2}{B-3} \\
&= \frac{\alpha^2(B-3+2\alpha)\left(\mathfrak{J}_\phi(1)+\frac{1}{B-1}\mathfrak{J}_{\alpha^{-1}\phi'}\left(\frac{-1}{B-1}\right)\right)}{(2\alpha-1)(B-3)(\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right))} - \frac{\alpha^2}{B-3} \\
&= \frac{\alpha^2\left((B-2)\mathfrak{J}_\phi(1)+(B-3+2\alpha)\frac{1}{B-1}\mathfrak{J}_{\alpha^{-1}\phi'}\left(\frac{-1}{B-1}\right)+(2\alpha-1)(\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)\right)}{(2\alpha-1)(B-3)(\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right))}
\end{aligned}$$

$$\lambda_{\mathbb{M}}^{G,\mathfrak{B}'_\phi} = \frac{B(B^2+2(\alpha-2)B+2(\alpha-3)\alpha+3)}{(B-3)(B-1)(B-1+2\alpha)} \frac{\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)}{\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)}$$

$$\begin{aligned}
\lambda_{\mathbb{L}}^{G,\mathfrak{B}'_\phi} &= -\frac{2\alpha^2(B-2)(B-1+\alpha)}{(B-3)(B-1)(B-1+2\alpha)} \\
&\quad + 2\frac{\alpha^2(3B-4)+\alpha\left(3B^2-11B+8\right)+(B-3)(B-1)^2}{(B-3)(B-1)^2(B-1+2\alpha)} \frac{\mathfrak{J}'_\phi\left(\frac{-1}{B-1}\right)}{\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)} \\
&\quad + \frac{\alpha^2(B-2)(B-3+2\alpha)}{(2\alpha-1)(B-3)(B-1)} \frac{\mathfrak{J}_\phi(1)}{\mathfrak{J}_\phi(1)-\mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)}
\end{aligned}$$

Proposition C.12. $\lambda_G^{G,\mathfrak{B}'_\phi} \geq \lambda_{\mathbb{L}}^{G,\mathfrak{B}'_\phi} \geq \lambda_{\mathbb{M}}^{G,\mathfrak{B}'_\phi}$

D Cross Batch: Forward Iteration

Definition D.1. For linear operators $\mathcal{T}_i : \mathcal{X}_i \to \mathcal{Y}_i, i = 1, \ldots, k$, we write $\bigoplus_i \mathcal{T}_i : \bigoplus_i \mathcal{X}_i \to \bigoplus_i \mathcal{Y}_i$ for the operator $(x_1, \ldots, x_k) \mapsto (\mathcal{T}_1(x_1), \ldots, \mathcal{T}_k(x_k))$. We also write $\mathcal{T}_1^{\oplus n} := \bigoplus_{j=1}^n \mathcal{T}_1$ for the direct sum of n copies of $\mathcal{T}_1$.

For $k \geq 2$, now consider the extended ("k-batch") dynamics on $\widetilde{\Sigma} \in \mathcal{S}_{kB}$ defined by

$$\begin{aligned}
\widetilde{\Sigma}^{(l)} &= \mathrm{V}\mathfrak{B}_\phi^{\oplus k}(\widetilde{\Sigma}^{(l-1)}) \\
&= \mathrm{E}[(\mathfrak{B}_\phi(z_{1:B}), \mathfrak{B}_\phi(z_{B+1:2B}), \ldots, \mathfrak{B}_\phi(z_{(k-1)B+1:kB}))^{\otimes 2} : z \sim \mathcal{N}(0, \widetilde{\Sigma}^{(l-1)})]
\end{aligned} \tag{44}$$

If we restrict the dynamics to just the upper left $B \times B$ submatrix (as well as any of the diagonal $B \times B$ blocks) of $\widetilde{\Sigma}^{(l)}$, then we recover Eq. (28).

D.1 Limit Points

Definition D.2. We say a matrix $\widetilde{\Sigma} \in \mathcal{S}_{kB}$ is *CBSB1* (short for "1-Step Cross-Batch Symmetry Breaking") if $\widetilde{\Sigma}$ in block form ($k \times k$ blocks, each of size $B \times B$) has one common BSB1 block on the diagonal and one common constant block on the off-diagonal, i.e.

$$\begin{pmatrix}
\mathrm{BSB1}(a,b) & c\mathbb{1}\mathbb{1}^T & c\mathbb{1}\mathbb{1}^T & \cdots \\
c\mathbb{1}\mathbb{1}^T & \mathrm{BSB1}(a,b) & c\mathbb{1}\mathbb{1}^T & \cdots \\
c\mathbb{1}\mathbb{1}^T & c\mathbb{1}\mathbb{1}^T & \mathrm{BSB1}(a,b) & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix}$$

We will study the fixed point $\widetilde{\Sigma}^*$ to Eq. (44) of CBSB1 form.

D.1.1 SPHERICAL COORDINATES

Theorem D.3. *Let $\widetilde{\Sigma} \in \mathcal{S}_{kB}$. If $\widetilde{\Sigma}$ is CBSB1 then $\mathrm{V}\mathfrak{B}_\phi^{\oplus k}(\widetilde{\Sigma})$ is CBSB1 and equals $\widetilde{\Sigma}^* = \begin{pmatrix} \Sigma^* & c^* \\ c^* & \Sigma^* \end{pmatrix}$ where Σ^* is the BSB1 fixed point of Thm B.18 and $c^* \geq 0$ with $\sqrt{c^*} = \frac{\Gamma((B-1)/2)}{\Gamma((B-2)/2)}\pi^{-1/2}\int_0^\pi \mathrm{d}\theta\, \phi(-\sqrt{B-1}\cos\theta)\sin^{B-3}\theta$.*

Proof. We will prove for $k = 2$. The general k cases follow from the same reasoning.

Let $\widetilde{\Sigma} = \begin{pmatrix} \Sigma & c\mathbb{1}\mathbb{1}^T \\ c\mathbb{1}\mathbb{1}^T & \Sigma \end{pmatrix}$ where $\Sigma = \mathrm{BSB1}(a, b)$. As remarked below Eq. (44), restricting to any diagonal blocks just recovers the dynamics of Eq. (28), which gives the claim about the diagonal blocks being Σ^* through Thm B.18.

We now look at the off-diagonal blocks.

$$\begin{aligned}\mathrm{V}\mathfrak{B}_\phi^{\oplus 2}(\widetilde{\Sigma})_{1:B,B+1:2B} &= \mathrm{E}[\mathfrak{B}_\phi(z_{1:B}) \otimes \mathfrak{B}_\phi(z_{B+1:2B}) : z \sim \mathcal{N}(0, \widetilde{\Sigma})] \\ &= \mathrm{E}[\phi \circ \mathfrak{N}(y_{1:B}) \otimes \phi \circ \mathfrak{N}(y_{B+1:2B}) : y \sim \mathcal{N}(0, \widetilde{\Sigma}^{G^{\oplus 2}})]\end{aligned}$$

where $\widetilde{\Sigma}^{G^{\oplus 2}} := G^{\oplus 2}\widetilde{\Sigma}G^{\oplus 2} = \begin{pmatrix} G & 0 \\ 0 & G \end{pmatrix}\widetilde{\Sigma}\begin{pmatrix} G & 0 \\ 0 & G \end{pmatrix} = \begin{pmatrix} G\Sigma G & cG\mathbb{1}\mathbb{1}^T G \\ cG\mathbb{1}\mathbb{1}^T G & G\Sigma G \end{pmatrix} = \begin{pmatrix} (a-b)G & 0 \\ 0 & (a-b)G \end{pmatrix}$ (the last step follows from Lemma B.15). Thus $y_{1:B}$ is independent from $y_{B+1:2B}$, and

$$\mathrm{V}\mathfrak{B}_\phi^{\oplus 2}(\widetilde{\Sigma})_{1:B,B+1:2B} = \mathrm{E}[\phi \circ \mathfrak{N}(x) : x \sim \mathcal{N}(0, (a-b)G)]^{\otimes 2}$$

By symmetry,

$$\begin{aligned}\mathrm{E}[\phi \circ \mathfrak{N}(x) : x \sim \mathcal{N}(0, (a-b)G)] &= \sqrt{c^*}\mathbb{1} \\ \mathrm{E}[\phi \circ \mathfrak{N}(x) : x \sim \mathcal{N}(0, (a-b)G)]^{\otimes 2} &= c^*\mathbb{1}\mathbb{1}^T\end{aligned}$$

where $\sqrt{c^*} := \mathrm{E}[\phi(\mathfrak{N}(x)_1) : x \sim \mathcal{N}(0, (a-b)G)]$. We can compute

$$\begin{aligned}\sqrt{c^*} &= \mathrm{E}[\phi(\mathfrak{N}(x)_1) : x \sim \mathcal{N}(0, G)] \\ &\qquad\qquad \text{because } \mathfrak{N} \text{ is scale-invariant} \\ &= \mathrm{E}[\phi(\mathfrak{N}(\mathbb{e}x)_1) : x \sim \mathcal{N}(0, I_{B-1})] \\ &\qquad\qquad \text{where } \mathbb{e} \text{ is as in Eq. (39)} \\ &= \mathrm{E}[\phi((\mathbb{e}\mathfrak{N}(x))_1) : x \sim \mathcal{N}(0, I_{B-1})] \\ &\qquad\qquad \text{because } \mathbb{e} \text{ is an isometry} \\ &= \mathrm{E}[\phi(\sqrt{B}(\mathbb{e}v)_1) : x \sim \mathcal{N}(0, I_{B-1})] \\ &\qquad\qquad \text{with } v = x/\|x\| \\ &= \frac{\Gamma((B-1)/2)}{\Gamma((B-2)/2)}\pi^{-1/2}\int_0^\pi \mathrm{d}\theta\, \phi(-\sqrt{B-1}\cos\theta)\sin^{B-3}\theta\end{aligned}$$

□

Corollary D.4. *With the notation as in Thm D.3, if ϕ is positive-homogeneous of degree α and $\phi(x) = a\rho_\alpha(x) - b\rho_\alpha(-x)$, then $\sqrt{c^*} = (a-b)\frac{1}{2\sqrt{\pi}}(B-1)^{\alpha/2}\frac{\Gamma((B-1)/2)\Gamma((\alpha+1)/2)}{\Gamma((\alpha+B-1)/2)}$.*

Proof. We compute

$$\int_0^{\pi/2} \mathrm{d}\theta\, (\cos\theta)^\alpha \sin^{B-3}\theta = \frac{1}{2}\operatorname{Beta}\left(0, 1; \frac{\alpha+1}{2}, \frac{B-2}{2}\right) \quad \text{by Lemma B.19}$$
$$= \frac{1}{2}\operatorname{Beta}\left(\frac{\alpha+1}{2}, \frac{B-2}{2}\right)$$
$$\int_{\pi/2}^{\pi} \mathrm{d}\theta\, (-\cos\theta)^\alpha \sin^{B-3}\theta = \int_0^{\pi/2} \mathrm{d}\theta\, (\cos\theta)^\alpha \sin^{B-3}\theta$$
$$= \frac{1}{2}\operatorname{Beta}\left(\frac{\alpha+1}{2}, \frac{B-2}{2}\right)$$

So for a positive homogeneous function $\phi(x) = a\rho_\alpha(x) - b\rho_\alpha(-x)$,

$$\sqrt{c^*} = \frac{\Gamma((B-1)/2)}{\Gamma((B-2)/2)}\pi^{-1/2}\left(a\int_{\pi/2}^{\pi} \mathrm{d}\theta\, (-\sqrt{B-1}\cos\theta)^\alpha \sin^{B-3}\theta - b\int_0^{\pi/2} \mathrm{d}\theta\, (\sqrt{B-1}\cos\theta)^\alpha \sin^{B-3}\theta\right)$$
$$= (a-b)\frac{\Gamma((B-1)/2)}{\Gamma((B-2)/2)}\pi^{-1/2}(B-1)^{\alpha/2}\frac{1}{2}\operatorname{Beta}\left(\frac{\alpha+1}{2}, \frac{B-2}{2}\right)$$
$$= (a-b)\frac{1}{2\sqrt{\pi}}(B-1)^{\alpha/2}\frac{\Gamma((B-1)/2)\Gamma((\alpha+1)/2)}{\Gamma((\alpha+B-1)/2)}$$

Expanding the beta function and combining with Thm D.3 gives the desired result. □

D.1.2 LAPLACE METHOD

While we don't need to use the Laplace method to compute c^* for positive homogeneous functions (since it is already given by Corollary D.4), going through the computation is instructive for the machinery for computing the eigenvalues in a later section.

Lemma D.5 (The Laplace Method Cross Batch Master Equation). *For $A, B, C \in \mathbb{N}$, let $f : \mathbb{R}^{A+B} \to \mathbb{R}^C$ and let $a, b \geq 0$. Suppose for any $y \in \mathbb{R}^A, z \in \mathbb{R}^B$, $\|f(y,z)\| \leq h(\sqrt{\|y\|^2 + \|z\|^2})$ for some nondecreasing function $h : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ such that $\mathrm{E}[h(r\|z\|) : z \in \mathcal{N}(0, I_{A+B})]$ exists for every $r \geq 0$. Define $\wp(\Sigma) := \mathrm{E}[\|y\|^{-2a}\|z\|^{-2b}f(y,z) : (y,z) \sim \mathcal{N}(0,\Sigma)]$. Then on $\{\Sigma \in \mathcal{S}_{A+B} : \operatorname{rank}\Sigma > 2(a+b)\}$, $\wp(\Sigma)$ is well-defined and continuous, and furthermore satisfies*

$$\wp(\Sigma) = \Gamma(a)^{-1}\Gamma(b)^{-1}\int_0^\infty \mathrm{d}s \int_0^\infty \mathrm{d}t\; s^{a-1}t^{b-1}\det(I_{A+B} + 2\Omega)^{-1/2} \mathop{\mathrm{E}}_{(y,z)\sim\mathcal{N}(0,\Pi)} f(y,z) \tag{45}$$

where $D = \begin{pmatrix}\sqrt{s}I_A & 0\\ 0 & \sqrt{t}I_B\end{pmatrix}$, $\Omega = D\Sigma D$, *and* $\Pi = D^{-1}\Omega(I+2\Omega)^{-1}D^{-1}$.

Proof. If Σ is full rank, we can show Fubini-Tonelli theorem is valid in the following computation by the same arguments of the proof of Lemma B.5.

$$\mathrm{E}[\|y\|^{-2a}\|z\|^{-2b}f(y,z) : (y,z) \sim \mathcal{N}(0,\Sigma)]$$
$$= \mathrm{E}[f(y,z)\int_0^\infty \mathrm{d}s\, \Gamma(a)^{-1}s^{a-1}e^{-\|y\|^2 s}\int_0^\infty \mathrm{d}t\, \Gamma(b)^{-1}t^{b-1}e^{-\|y\|^2 t} : (y,z) \sim \mathcal{N}(0,\Sigma)]$$
$$= (2\pi)^{-\frac{A+B}{2}}\Gamma(a)^{-1}\Gamma(b)^{-1}\det\Sigma^{-1/2}\int_0^\infty \mathrm{d}s \int_0^\infty \mathrm{d}t\; s^{a-1}t^{b-1}\int_{\mathbb{R}^{A+B}} \mathrm{d}y\; \mathrm{d}z\; f(y,z)e^{-\frac{1}{2}(y,z)(\Sigma^{-1}+2D^2)(y,z)^T} \quad \text{(Fubini-Tonelli)}$$
$$= \Gamma(a)^{-1}\Gamma(b)^{-1}\int_0^\infty \mathrm{d}s \int_0^\infty \mathrm{d}t\; s^{a-1}t^{b-1}\det(\Sigma(\Sigma^{-1}+2D^2))^{-1/2}\,\mathrm{E}[f(y,z) : (y,z) \sim \mathcal{N}(0,(\Sigma^{-1}+2D^2)^{-1})]$$

We recover the equation in question with the following simplifications.

$$\begin{aligned}(\Sigma^{-1}+2D^2)^{-1} &= \Sigma(I_{A+B}+2D^2\Sigma)^{-1}\\ &= \Sigma(D^{-1}+2D\Sigma)^{-1}D^{-1}\\ &= \Sigma D(I_{A+B}+2D\Sigma D)^{-1}D^{-1}\\ &= D^{-1}\Omega(I_{A+B}+2\Omega)^{-1}D^{-1}\\ \det(\Sigma(\Sigma^{-1}+2D^2)) &= \det(I_{A+B}+2\Sigma D^2)\\ &= \det(I_{A+B}+2D\Sigma D)\\ &= \det(I_{A+B}+2\Omega)\end{aligned}$$

The case of general Σ with $\operatorname{rank}\Sigma > 2(a+b)$ is given by the same continuity arguments as in Lemma B.5. □

Let $\widetilde{\Sigma} \in \mathcal{S}_{2B}, \widetilde{\Sigma} = \begin{pmatrix}\Sigma & \Xi\\ \Xi^T & \Sigma'\end{pmatrix}$ where $\Sigma, \Sigma' \in \mathcal{S}_B$ and $\Xi \in \mathbb{R}^{B\times B}$. Consider the off-diagonal block of $\mathrm{V}\mathfrak{B}_\phi^{\oplus 2}(\widetilde{\Sigma})$.

$$\begin{aligned}&\mathrm{E}[\mathfrak{B}_\phi(z)\otimes\mathfrak{B}_\phi(z') : (z,z')\sim\mathcal{N}(0,\widetilde{\Sigma})]\\ &= \mathrm{E}[\phi(\mathfrak{N}(y))\otimes\phi(\mathfrak{N}(y')) : (y,y')\sim\mathcal{N}(0,\widetilde{\Sigma}^{G^{\oplus 2}})]\\ &= B^\alpha\,\mathrm{E}[\|y\|^{-\alpha}\|y'\|^{-\alpha}\phi(y)\otimes\phi(y') : (y,y')\sim\mathcal{N}(0,\widetilde{\Sigma}^{G^{\oplus 2}})]\\ &= B^\alpha\Gamma(\alpha/2)^{-2}\int_0^\infty \mathrm{d}s\int_0^\infty \mathrm{d}t\,(st)^{\alpha/2-1}\det(I_{2B}+2\Omega)^{-1/2}\mathop{\mathrm{E}}_{(y,y')\sim\mathcal{N}(0,\Pi)}\phi(y)\otimes\phi(y')\\ &\qquad\qquad\text{(by Lemma D.5)}\\ &= B^\alpha\Gamma(\alpha/2)^{-2}\int_0^\infty \mathrm{d}s\int_0^\infty \mathrm{d}t\,(st)^{\alpha/2-1}\det(I_{2B}+2\Omega)^{-1/2}\mathrm{V}\phi(\Pi)_{1:B,B+1:2B}\end{aligned}\tag{46}$$

where $\Omega = \begin{pmatrix}s\Sigma^G & \sqrt{st}\Xi^G\\ \sqrt{st}(\Xi^G)^T & t\Sigma'^G\end{pmatrix}$ and $\Pi = D^{-1}\Omega(I+2\Omega)^{-1}D^{-1}$ with $D = \sqrt{s}I_B \oplus \sqrt{t}I_B = \begin{pmatrix}\sqrt{s}I_B & 0\\ 0 & \sqrt{t}I_B\end{pmatrix}$.

Theorem D.6 (Rephrasing and adding to Corollary D.4)**.** *Let $\widetilde{\Sigma} \in \mathcal{S}_{kB}$ and ϕ be positive homogeneous of degree α. If $\widetilde{\Sigma}$ is CBSB1 then $\mathrm{V}\mathfrak{B}_\phi^{\oplus k}(\widetilde{\Sigma})$ is CBSB1 and equals $\widetilde{\Sigma}^* = \begin{pmatrix}\Sigma^* & c^*\mathbb{1}\mathbb{1}^T\\ c^*\mathbb{1}\mathbb{1}^T & \Sigma^*\end{pmatrix}$ where Σ^* is the BSB1 fixed point of Thm B.18 and*

$$\begin{aligned}c^* &= \mathsf{c}_\alpha\left(\frac{B-1}{2}\right)^\alpha \mathrm{P}\left(\frac{B-1}{2},\frac{\alpha}{2}\right)^{-2}\mathfrak{J}_\phi(0)\\ &= (a-b)^2\frac{1}{4\pi}(B-1)^\alpha\left(\frac{\Gamma((B-1)/2)\Gamma((\alpha+1)/2)}{\Gamma((\alpha+B-1)/2)}\right)^2\end{aligned}$$

Proof. Like in the proof of Corollary D.4, we only need to compute the cross-batch block, which is given above by Eq. (46). Recall that $\widetilde{\Sigma} = \begin{pmatrix}\Sigma & c\mathbb{1}\mathbb{1}^T\\ c\mathbb{1}\mathbb{1}^T & \Sigma\end{pmatrix}$ where $\Sigma = \mathrm{BSB1}(a,b)$. Set $\upsilon := a-b$. Note that because the off-diagonal block $\Xi = c\mathbb{1}\mathbb{1}^T$ by assumption, $\Xi^G = 0$, so that

$$\begin{aligned}\Omega &= s\Sigma^G \oplus t\Sigma^G\\ &= s\upsilon G \oplus t\upsilon G\\ \Omega(I+2\Omega)^{-1} &= \frac{s\upsilon}{1+2s\upsilon}G \oplus \frac{t\upsilon}{1+2t\upsilon}G\\ \Pi &= \frac{\upsilon}{1+2s\upsilon}G \oplus \frac{\upsilon}{1+2t\upsilon}G\end{aligned}$$

Therefore,

$$\det(I+2\Omega) = (1+2s\upsilon)^{B-1}(1+2t\upsilon)^{B-1}$$
$$\mathrm{V}\phi(\Pi) = \upsilon^\alpha \left(\frac{B-1}{B}\right)^\alpha \Delta \mathrm{V}\phi \begin{pmatrix} \mathrm{BSB1}(1, \frac{-1}{B-1}) & 0 \\ 0 & \mathrm{BSB1}(1, \frac{-1}{B-1}) \end{pmatrix} \Delta$$

where $\Delta = \begin{pmatrix} (1+2s\upsilon)^{-\alpha/2} I_B & 0 \\ 0 & (1+2t\upsilon)^{-\alpha/2} I_B \end{pmatrix}$. In particular, the off-diagonal block is constant with entry $\mathsf{c}_\alpha \upsilon^\alpha \left(\frac{B-1}{B}\right)^\alpha (1+2s\upsilon)^{-\alpha/2}(1+2t\upsilon)^{-\alpha/2}\mathfrak{J}_\phi(0)$.

Finally, by Eq. (46),

$$\begin{aligned}
&\mathrm{E}[\mathfrak{B}_\phi(z)\otimes\mathfrak{B}_\phi(z') : (z,z') \sim \mathcal{N}(0,\widetilde{\Sigma})] \\
&= B^\alpha \Gamma(\alpha/2)^{-2} \int_0^\infty \mathrm{d}s \int_0^\infty \mathrm{d}t\ (st)^{\alpha/2-1} \det(I_{2B}+2\Omega)^{-1/2} \mathrm{V}\phi(\Pi)_{1:B,B+1:2B} \\
&= B^\alpha \Gamma(\alpha/2)^{-2} \int_0^\infty \mathrm{d}s \int_0^\infty \mathrm{d}t\ (st)^{\alpha/2-1} [(1+2s\upsilon)(1+2t\upsilon)]^{-\frac{B-1}{2}} \\
\times &\quad \mathsf{c}_\alpha \upsilon^\alpha \left(\frac{B-1}{B}\right)^\alpha [(1+2s\upsilon)(1+2t\upsilon)]^{-\alpha/2} \mathfrak{J}_\phi(0) \mathbb{1}\mathbb{1}^T \\
&= \mathsf{c}_\alpha (B-1)^\alpha \Gamma(\alpha/2)^{-2} \mathfrak{J}_\phi(0)\mathbb{1}\mathbb{1}^T \int_0^\infty \mathrm{d}\sigma \int_0^\infty \mathrm{d}\tau\ (\sigma\tau)^{\alpha/2-1}(1+2\sigma)^{-\frac{B-1+\alpha}{2}}(1+2\tau)^{-\frac{B-1+\alpha}{2}} \\
&\hspace{20em} \text{(with } \sigma = \upsilon s, \tau = \upsilon t) \\
&= \mathsf{c}_\alpha (B-1)^\alpha \Gamma(\alpha/2)^{-2} \mathfrak{J}_\phi(0)\mathbb{1}\mathbb{1}^T \left(2^{-\alpha/2}\,\mathrm{Beta}\left(\frac{B-1}{2}, \frac{\alpha}{2}\right)\right)^2 \\
&= \mathsf{c}_\alpha \left(\frac{B-1}{2}\right)^\alpha \mathrm{P}\left(\frac{B-1}{2}, \frac{\alpha}{2}\right)^{-2} \mathfrak{J}_\phi(0)\mathbb{1}\mathbb{1}^T
\end{aligned}$$

Simplification with Defn H.2 and Proposition I.5 gives the result. □

E Local Convergence

We now look at the linearized dynamics around CBSB1 fixed points of Eq. (44).

Theorem E.1. *eigenvalue of deviation with zero diagonal blocks is*

$$\begin{aligned}
&\frac{B}{2}\left(\frac{B-1}{2}\right)^{\alpha-1} \mathrm{P}\left(\frac{B}{2}, \frac{\alpha}{2}\right)^{-2} \upsilon^{*-1} \mathsf{c}_\alpha \mathfrak{J}'_\phi(0) \\
&= \frac{B}{B-1} \frac{\mathrm{P}\left(\frac{B-1}{2}, \alpha\right)}{\mathrm{P}\left(\frac{B}{2}, \frac{\alpha}{2}\right)^2} \frac{\mathfrak{J}'_\phi(0)}{\mathfrak{J}_\phi(1) - \mathfrak{J}_\phi\left(\frac{-1}{B-1}\right)}
\end{aligned}$$

Proof. Let $\widetilde{\Sigma}^* = \begin{pmatrix} \Sigma^* & c^*\mathbb{1}\mathbb{1}^T \\ c^*\mathbb{1}\mathbb{1}^T & \Sigma^* \end{pmatrix}$ be the CBSB1 fixed point in Thm D.6. Take a smooth path $\widetilde{\Sigma}_\tau = \begin{pmatrix} \Sigma_\tau & \Xi_\tau \\ \Xi_\tau^T & \Sigma'_\tau \end{pmatrix} \in \mathcal{S}_{2B}, \tau \in (-\epsilon, \epsilon)$ such that $\widetilde{\Sigma}_0 = \widetilde{\Sigma}^*, \dot{\widetilde{\Sigma}}_0 := \frac{d}{d\tau}\widetilde{\Sigma} = \begin{pmatrix} 0 & \Upsilon \\ \Upsilon^T & 0 \end{pmatrix}$ for some $\Upsilon \in \mathbb{R}^{B\times B}$. Then with $\Omega_\tau = \begin{pmatrix} s\Sigma_\tau^G & \sqrt{st}\Xi_\tau^G \\ \sqrt{st}(\Xi_\tau^G)^T & t{\Sigma'}_\tau^G \end{pmatrix}$ and $\Pi_\tau = D^{-1}\Omega_\tau(I+2\Omega_\tau)^{-1}D^{-1}$ with

$$D = \sqrt{s} I_B \oplus \sqrt{t} I_B = \begin{pmatrix} \sqrt{s} I_B & 0 \\ 0 & \sqrt{t} I_B \end{pmatrix},$$

$$\begin{aligned}
&\frac{d}{d\tau} \mathrm{E}[\mathfrak{B}_\phi(z) \otimes \mathfrak{B}_\phi(z') : (z, z') \sim \mathcal{N}(0, \widetilde{\Sigma})]\Big|_{\tau=0} \\
&= \frac{d}{d\tau} B^\alpha \Gamma(\alpha/2)^{-2} \int_0^\infty \mathrm{d}s \int_0^\infty \mathrm{d}t \, (st)^{\alpha/2-1} \det(I_{2B} + 2\Omega)^{-1/2} \mathrm{V}\phi(\Pi)_{1:B,B+1:2B}\Big|_{\tau=0} \\
&= B^\alpha \Gamma(\alpha/2)^{-2} \int_0^\infty \mathrm{d}s \int_0^\infty \mathrm{d}t \, (st)^{\alpha/2-1} \Bigg(\frac{d}{d\tau} \det(I_{2B} + 2\Omega)^{-1/2}\Big|_{\tau=0} \mathrm{V}\phi(\Pi)_{1:B,B+1:2B} \\
&\qquad\qquad + \det(I_{2B} + 2\Omega)^{-1/2} \frac{d}{d\tau} \mathrm{V}\phi(\Pi)_{1:B,B+1:2B}\Big|_{\tau=0} \Bigg) \\
&= B^\alpha \Gamma(\alpha/2)^{-2} \int_0^\infty \mathrm{d}s \int_0^\infty \mathrm{d}t \, (st)^{\alpha/2-1} \Bigg(- \det(I + 2\Omega_0)^{-1/2} \operatorname{tr}\left((I + 2\Omega_0)^{-1} \dot{\Omega}_0 \right) \mathrm{V}\phi(\Pi) \\
&\qquad\qquad + \det(I + 2\Omega_0)^{-1/2} (D^{-\alpha})^{\otimes 2} \circ \frac{\mathrm{dV}\phi}{\mathrm{d}\Sigma}\Big|_{\Sigma = \Omega_0 (I + 2\Omega_0)^{-1}} \left\{ \frac{d}{d\tau} \Omega (I + 2\Omega)^{-1}\Big|_{\tau=0} \right\} \Bigg)_{1:B,B+1:2B}
\end{aligned} \tag{47}$$

(by chain rule and Lemma B.51)

We compute

$$\begin{aligned}
\Omega_0 &= \upsilon^* (sG \oplus tG) \\
\dot{\Omega}_0 &= \sqrt{st} \begin{pmatrix} 0 & \Upsilon^G \\ (\Upsilon^G)^T & 0 \end{pmatrix} \\
(I + 2\Omega_0)^{-1} &= ((1 + 2s\upsilon^*)^{-1} G + \frac{1}{B} \mathbb{1}\mathbb{1}^T) \oplus ((1 + 2t\upsilon^*)^{-1} G + \frac{1}{B} \mathbb{1}\mathbb{1}^T) \\
\det(I + 2\Omega_0) &= (1 + 2s\upsilon^*)^{B-1} (1 + 2t\upsilon^*)^{B-1} \\
\Omega_0 (I + 2\Omega_0)^{-1} &= \frac{s\upsilon^*}{1 + 2s\upsilon^*} G \oplus \frac{t\upsilon^*}{1 + 2t\upsilon^*} G \\
\frac{\mathrm{dV}\phi}{\mathrm{d}\Sigma}\Big|_{\Sigma = \Omega_0 (I + 2\Omega_0)^{-1}} &= \left(\left(\frac{s\upsilon^*}{1 + 2s\upsilon^*} \right)^{(\alpha-1)/2} I_B \oplus \left(\frac{t\upsilon^*}{1 + 2t\upsilon^*} \right)^{(\alpha-1)/2} I_B \right)^{\otimes 2} \circ \left(\frac{B-1}{B} \right)^{\alpha-1} \frac{\mathrm{dV}\phi}{\mathrm{d}\Sigma}\Big|_{\Sigma = \mathrm{BSB1}(1, \frac{-1}{B-1})^{\oplus 2}}
\end{aligned}$$

(because $\mathrm{dV}\phi / \mathrm{d}\Sigma$ is degree $\alpha - 1$ positive homogeneous)

$$\frac{d}{d\tau} \Omega (I + 2\Omega)^{-1}\Big|_{\tau=0} = (I + 2\Omega_0)^{-1} \dot{\Omega}_0 (I + 2\Omega_0)^{-1}$$

(by Lemma B.52)

$$\begin{aligned}
&= \frac{\sqrt{st}}{(1 + 2s\upsilon^*)(1 + 2t\upsilon^*)} \begin{pmatrix} 0 & \Upsilon^G \\ (\Upsilon^G)^T & 0 \end{pmatrix} \\
\frac{\mathrm{dV}\phi}{\mathrm{d}\Sigma}\Big|_{\Sigma = \Omega_0 (I + 2\Omega_0)^{-1}} &\left\{ \frac{d}{d\tau} \Omega (I + 2\Omega)^{-1}\Big|_{\tau=0} \right\} \\
&= \left(\frac{B-1}{B} \right)^{\alpha-1} \frac{\sqrt{st}}{(1 + 2s\upsilon^*)(1 + 2t\upsilon^*)} \mathsf{c}_\alpha \mathfrak{J}'_\phi(0) \left(\frac{s\upsilon^*}{1 + 2s\upsilon^*} \right)^{(\alpha-1)/2} \left(\frac{t\upsilon^*}{1 + 2t\upsilon^*} \right)^{(\alpha-1)/2} \\
&\qquad \begin{pmatrix} 0 & \Upsilon^G \\ (\Upsilon^G)^T & 0 \end{pmatrix} \\
&= \left(\frac{B-1}{B} \right)^{\alpha-1} \frac{(st)^{\alpha/2} \upsilon^{*\alpha-1} \mathsf{c}_\alpha \mathfrak{J}'_\phi(0)}{(1 + 2s\upsilon^*)^{(\alpha+1)/2} (1 + 2t\upsilon^*)^{(\alpha+1)/2}} \begin{pmatrix} 0 & \Upsilon^G \\ (\Upsilon^G)^T & 0 \end{pmatrix}
\end{aligned}$$

Thus the product $(I + 2\Omega_0)^{-1}\dot{\Omega}_0$ has zero diagonal blocks so that its trace is 0. Therefore, only the second term in the sum in Eq. (47) above survives, and we have

$$\begin{aligned}
&\frac{d}{d\tau}\,\mathrm{E}[\mathfrak{B}_\phi(z)\otimes\mathfrak{B}_\phi(z') : (z,z')\sim\mathcal{N}(0,\widetilde{\Sigma})]\Big|_{\tau=0} \\
&= B^\alpha\Gamma(\alpha/2)^{-2}\int_0^\infty \mathrm{d}s\int_0^\infty \mathrm{d}t\,(st)^{\alpha/2-1}(1+2s\upsilon^*)^{-\frac{B-1}{2}}(1+2t\upsilon^*)^{-\frac{B-1}{2}}\times \\
&\qquad (st)^{-\alpha/2}\left(\frac{B-1}{B}\right)^{\alpha-1}\frac{(st)^{\alpha/2}\upsilon^{*\alpha-1}\mathsf{c}_\alpha\mathfrak{J}'_\phi(0)}{(1+2s\upsilon^*)^{(\alpha+1)/2}(1+2t\upsilon^*)^{(\alpha+1)/2}}\Upsilon^G \\
&= B(B-1)^{\alpha-1}\Gamma(\alpha/2)^{-2}\int_0^\infty \mathrm{d}s\int_0^\infty \mathrm{d}t\,(st)^{\alpha/2-1}\frac{\upsilon^{*\alpha-1}\mathsf{c}_\alpha\mathfrak{J}'_\phi(0)}{(1+2s\upsilon^*)^{(B+\alpha)/2}(1+2t\upsilon^*)^{(B+\alpha)/2}}\Upsilon^G \\
&= B(B-1)^{\alpha-1}\Gamma(\alpha/2)^{-2}\upsilon^{*-1}\mathsf{c}_\alpha\mathfrak{J}'_\phi(0)\left(\int_0^\infty \mathrm{d}(\upsilon^* s)\,\frac{(\upsilon^* s)^{\alpha/2-1}}{(1+2s\upsilon^*)^{(B+\alpha)/2}}\right)^2\Upsilon^G \\
&= B(B-1)^{\alpha-1}\Gamma(\alpha/2)^{-2}\upsilon^{*-1}\mathsf{c}_\alpha\mathfrak{J}'_\phi(0)\left(2^{-\alpha/2}\,\mathrm{Beta}\left(\frac{B}{2},\frac{\alpha}{2}\right)\right)^2\Upsilon^G \\
&= \frac{B}{2}\left(\frac{B-1}{2}\right)^{\alpha-1}\mathrm{P}\left(\frac{B}{2},\frac{\alpha}{2}\right)^{-2}\upsilon^{*-1}\mathsf{c}_\alpha\mathfrak{J}'_\phi(0)\Upsilon^G
\end{aligned}$$

□

F Cross Batch: Backward Dynamics

For $k \geq 2$, we study the linear equation

$$\tilde{\Pi}^{(l)} = \mathrm{V}\left[(\mathfrak{B}_\phi^{\oplus k})'\right](\widetilde{\Sigma}^*)^\dagger\{\tilde{\Pi}^{(l+1)}\} \tag{48}$$

Its dynamics is essentially given by the eigendecomposition of $\mathrm{V}\left[(\mathfrak{B}_\phi^{\oplus k})'\right](\widetilde{\Sigma}^*)$.

Consider the case $k = 2$ (for illustration purposes).

$$\begin{aligned}
\mathrm{V}\left[(\mathfrak{B}_\phi^{\oplus 2})'\right] &= \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{N}(0,\widetilde{\Sigma}^*)}[(\mathfrak{B}'_\phi(x)\oplus\mathfrak{B}'_\phi(y))^{\otimes 2}] \\
&= \\
\mathrm{V}\left[(\mathfrak{B}_\phi^{\oplus 2})'\right]\left\{\begin{pmatrix}\Sigma & \Xi\\ \Xi^T & \Lambda\end{pmatrix}\right\} &= \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{N}(0,\widetilde{\Sigma}^*)}\left[\begin{pmatrix}\mathfrak{B}'_\phi(x)\Sigma\mathfrak{B}'_\phi(x)^T & \mathfrak{B}'_\phi(x)\Xi\mathfrak{B}'_\phi(y)^T\\ \mathfrak{B}'_\phi(y)\Xi^T\mathfrak{B}'_\phi(x)^T & \mathfrak{B}'_\phi(y)\Lambda\mathfrak{B}'_\phi(y)^T\end{pmatrix}\right]
\end{aligned}$$

From this one sees that $\mathrm{V}\left[(\mathfrak{B}_\phi^{\oplus 2})'\right]$ acts independently on each block, and consequently so does its adjoint. The diagonal blocks of Eq. (48) evolves according to Eq. (41) which we studied in Appendix C. In this section we will study the evolution of the off-diagonal blocks, which is given by

$$\begin{aligned}
\Xi^{(l)} &= \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{N}(0,\widetilde{\Sigma}^*)}[\mathfrak{B}'_\phi(x)\otimes\mathfrak{B}'_\phi(y)]^\dagger\{\Xi^{(l+1)}\} \quad = \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{N}(0,\widetilde{\Sigma}^*)}[\mathfrak{B}'_\phi(y)^T\otimes\mathfrak{B}'_\phi(x)^T]\{\Xi^{(l+1)}\} \\
&= \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{N}(0,\widetilde{\Sigma}^*)}[\mathfrak{B}'_\phi(y)^T\Xi^{(l+1)}\mathfrak{B}'_\phi(x)^T]
\end{aligned} \tag{49}$$

Define $\mathcal{U} := \mathrm{E}_{(x,y)\sim\mathcal{N}(0,\widetilde{\Sigma}^*)}[\mathfrak{B}'_\phi(x)\otimes\mathfrak{B}'_\phi(y)]$. Then

$$\begin{aligned}
\mathcal{U} &= \mathop{\mathrm{E}}_{(\xi,\eta)\sim\mathcal{N}(0,G^{\oplus 2}\widetilde{\Sigma}^*G^{\oplus 2})}\left[\frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\Big|_{z=\xi}\otimes\frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\Big|_{z=\eta}\right]\circ G^{\otimes 2} \\
&= \mathop{\mathrm{E}}_{(\xi,\eta)\sim\mathcal{N}(0,\upsilon^*\left(\begin{smallmatrix}G & 0\\ 0 & G\end{smallmatrix}\right))}\left[\frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\Big|_{z=\xi}\otimes\frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\Big|_{z=\eta}\right]\circ G^{\otimes 2}
\end{aligned}$$

since $\widetilde{\Sigma}^*$ is CBSB1 with diagonal blocks Σ^* (Corollary D.4). Then ξ is independent from η, so this is just

$$\begin{aligned}\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{N}(0,\widetilde{\Sigma}^*)}[\mathfrak{B}'_\phi(x)\otimes\mathfrak{B}'_\phi(y)] &= \mathop{\mathrm{E}}_{\xi\sim\mathcal{N}(0,\upsilon^*G)}\left[\frac{\mathrm{d}\phi\circ\mathfrak{N}(z)}{\mathrm{d}z}\bigg|_{z=\xi}\right]^{\otimes 2}\circ G^{\otimes 2}\\ &= \mathop{\mathrm{E}}_{\xi\sim\mathcal{N}(0,\upsilon^*G)}\left[\sqrt{B}Dr^{-1}(I-vv^T)\right]^{\otimes 2}\circ G^{\otimes 2}\end{aligned}$$

(by Proposition C.4)

where $D = \mathrm{Diag}(\phi'(\sqrt{B}v)), r = \|\xi\|, v = \xi/\|\xi\|$. But notice that $\mathcal{R} := \mathrm{E}_{\xi\sim\mathcal{N}(0,\upsilon^*G)}\left[Dr^{-1}(I-vv^T)\right]$ is actually BSB1, with diagonal

$$\mathrm{E}[r^{-1}(1-v_i^2)\phi'(\sqrt{B}v_i) : rv\sim\mathcal{N}(0,\upsilon^*G)] = \mathrm{E}[r^{-1}(1-v_1^2)\phi'(\sqrt{B}v_1) : rv\sim\mathcal{N}(0,\upsilon^*G)], \forall i\in[B]$$

and off-diagonal

$$\mathrm{E}[r^{-1}(-v_iv_j)\phi'(\sqrt{B}v_i) : rv\sim\mathcal{N}(0,\upsilon^*G)] = \mathrm{E}[r^{-1}(-v_1v_2)\phi'(\sqrt{B}v_1) : rv\sim\mathcal{N}(0,\upsilon^*G)], \forall i\neq j\in[B].$$

Thus by Lemma B.14, $\mathcal{R}$ has two eigenspaces, $\{x : Gx = x\}$ and $\mathbb{R}\mathbb{1}$, with respective eigenvalues $\mathcal{R}_{11}-\mathcal{R}_{12}$ and $\mathcal{R}_{11}+(B-1)\mathcal{R}_{12}$. However, because of the composition with $G^{\otimes 2}$, only the the former eigenspace survives with nonzero eigenvalue in $\mathcal{U}$.

Theorem F.1. *$\mathcal{U}$ has two eigenspaces $\{\Xi : \Xi^G = \Xi\}$ and its orthogonal complement, with respective eigenvalues $\frac{1}{2\pi}B\upsilon^{*-1}\left(\int_0^\pi \mathrm{d}\theta_1\ \sin^{B-1}\theta_1\phi'(-\sqrt{B-1}\cos\theta_1)\right)^2$ and 0.*

Proof. By the above reasoning, the sole eigenspace with nonzero eigenvalue is $\{x : Gx = x\}^{\otimes 2} = \{\Xi : \Xi^G = \Xi\}$. It has eigenvalue

$$\begin{aligned}&B\left(\mathrm{E}[r^{-1}(1-v_1^2+v_1v_2)\phi'(\sqrt{B}v_1) : rv\sim\mathcal{N}(0,\upsilon^*G)]\right)^2\\ &= B\upsilon^{*-1}\left(\mathrm{E}[r^{-1}(1-v_1^2+v_1v_2)\phi'(\sqrt{B}v_1) : rv\sim\mathcal{N}(0,G)]\right)^2\\ &= B\upsilon^{*-1}\left(\mathrm{E}[r^{-1}(1-(\mathbb{e}w)_1^2+(\mathbb{e}w)_1(\mathbb{e}w)_2)\phi'(\sqrt{B}(\mathbb{e}w)_1) : rw\sim\mathcal{N}(0,I_{B-1})]\right)^2\\ &= B\upsilon^{*-1}\left(\frac{\Gamma((B-2)/2)}{\Gamma((B-3)/2)}2^{-1/2}\pi^{-1}\int_0^\pi \mathrm{d}\theta_1\int_0^\pi \mathrm{d}\theta_2\ (1-\zeta(\theta_1)^2+\zeta(\theta_1)\omega(\theta_1,\theta_2))\phi'(\sqrt{B}\zeta(\theta_1))\sin^{B-3}\theta_1\sin^{B-4}\theta_2\right)^2\end{aligned}$$

where we applied Lemma B.48 with $\zeta(\theta) = -\sqrt{\frac{B-1}{B}}\cos\theta$ and $\omega(\theta_1,\theta_2) = \frac{1}{\sqrt{B(B-1)}}\cos\theta_1 - \sqrt{\frac{B-2}{B-1}}\sin\theta_1\cos\theta_2$.

We can further simplify

$$\begin{aligned}&\int_0^\pi \mathrm{d}\theta_1\int_0^\pi \mathrm{d}\theta_2\ (1-\zeta(\theta_1)^2+\zeta(\theta_1)\omega(\theta_1,\theta_2))\phi'(\sqrt{B}\zeta(\theta_1))\sin^{B-3}\theta_1\sin^{B-4}\theta_2\\ &= \int_0^\pi \mathrm{d}\theta_1\ \sin^{B-3}\theta_1\phi'(\sqrt{B}\zeta(\theta_1))\Bigg(\left(1-\zeta(\theta_1)^2+\zeta(\theta_1)\frac{1}{\sqrt{B(B-1)}}\cos\theta_1\right)\int_0^\pi \mathrm{d}\theta_2\ \sin^{B-4}\theta_2\\ &\qquad -\sqrt{\frac{B-2}{B-1}}\sin\theta_1\zeta(\theta_1)\int_0^\pi \mathrm{d}\theta_2\ \sin^{B-4}\theta_2\cos\theta_2\Bigg)\\ &= \int_0^\pi \mathrm{d}\theta_1\ \sin^{B-3}\theta_1\phi'(\sqrt{B}\zeta(\theta_1))\left(\left(1-\zeta(\theta_1)^2+\zeta(\theta_1)\frac{1}{\sqrt{B(B-1)}}\cos\theta_1\right)\mathrm{Beta}\left(\frac{B-3}{2},\frac{1}{2}\right)\right)\end{aligned}$$

(by Lemma B.20)

$$\begin{aligned}&= \int_0^\pi \mathrm{d}\theta_1\ \sin^{B-3}\theta_1\phi'(\sqrt{B}\zeta(\theta_1))\left(\sin^2\theta_1\,\mathrm{Beta}\left(\frac{B-3}{2},\frac{1}{2}\right)\right)\\ &= \mathrm{Beta}\left(\frac{B-3}{2},\frac{1}{2}\right)\int_0^\pi \mathrm{d}\theta_1\ \sin^{B-1}\theta_1\phi'(-\sqrt{B-1}\cos\theta_1)\end{aligned}$$

so the cross-batch backward off-diagonal eigenvalue is

$$
\begin{aligned}
& B\upsilon^{*-1}\left(\frac{1}{\sqrt{2\pi}}\int_0^\pi \mathrm{d}\theta_1\ \sin^{B-1}\theta_1\phi'(-\sqrt{B-1}\cos\theta_1)\right)^2 \\
= & \frac{1}{2\pi}B\upsilon^{*-1}\left(\int_0^\pi \mathrm{d}\theta_1\ \sin^{B-1}\theta_1\phi'(-\sqrt{B-1}\cos\theta_1)\right)^2
\end{aligned}
$$

□

Theorem F.2. *If ϕ is positive-homogeneous of degree α with $\phi(c) = a\rho_\alpha(c) - b\rho_\alpha(-c)$, then the cross-batch backward off-diagonal eigenvalue is*

$$
\frac{1}{8\pi}B(B-1)^{\alpha-1}\upsilon^{*-1}\alpha^2(a+b)^2\,\mathrm{Beta}\left(\frac{\alpha}{2},\frac{B}{2}\right)^2.
$$

Proof. As in the proof of Corollary D.4,

$$
\int_0^{\pi/2}\mathrm{d}\theta\ (\cos\theta)^{\alpha-1}\sin^{B-1}\theta = \int_{\pi/2}^{\pi}\mathrm{d}\theta\ (-\cos\theta)^{\alpha-1}\sin^{B-1}\theta = \frac{1}{2}\,\mathrm{Beta}\left(\frac{\alpha}{2},\frac{B}{2}\right)
$$

So

$$
\begin{aligned}
& \frac{1}{2\pi}B\upsilon^{*-1}\left(\int_0^\pi \mathrm{d}\theta_1\ \sin^{B-1}\theta_1\phi'(-\sqrt{B-1}\cos\theta_1)\right)^2 \\
= & \frac{1}{2\pi}B\upsilon^{*-1}(B-1)^{\alpha-1}\alpha^2\left(a\int_{\pi/2}^{\pi}\mathrm{d}\theta\ (-\cos\theta)^{\alpha-1}\sin^{B-1}\theta + b\int_0^{\pi/2}\mathrm{d}\theta\ (\cos\theta)^{\alpha-1}\sin^{B-1}\theta\right)^2 \\
= & \frac{1}{2\pi}B\upsilon^{*-1}(B-1)^{\alpha-1}\alpha^2(a+b)^2 2^{-2}\,\mathrm{Beta}\left(\frac{\alpha}{2},\frac{B}{2}\right)^2 \\
= & \frac{1}{8\pi}B(B-1)^{\alpha-1}\upsilon^{*-1}\alpha^2(a+b)^2\,\mathrm{Beta}\left(\frac{\alpha}{2},\frac{B}{2}\right)^2
\end{aligned}
$$

□

G WEIGHT GRADIENTS

$$
x_{i\bullet} = \Phi(h_{i\bullet}), \qquad h_{ik} = \sum_{j=1}^{N} w_{ij}\underline{x}_{jk} + b_i.
$$

Suppose we have a loss function E which induces, on a minibatch of inputs $x^{(0)}_{\bullet\bullet}$, a minibatch of gradient vectors at the layer l, $\partial E/\partial x^{(l)}_{\bullet\bullet}$. Then $\frac{\partial E}{\partial h^{(l)}_{i\bullet}} = \frac{\partial E}{\partial x^{(l)}_{i\bullet}}\frac{\mathrm{d}\Phi}{\mathrm{d}h}(h^{(l)}_{i\bullet})$ and $\frac{\partial E}{\partial x^{(l-1)}_{i\bullet}} = \sum_j \frac{\partial E}{\partial h^{(l)}_{j\bullet}} w_{ji}$. The latter is again a sum of a large number of random variables, so converges to a Gaussian (with zero mean) as width $N \to \infty$. Write g_{jb} for $\partial E/\partial x^{(l)}_{jb}$. Then (with gradient independence assumption)

$$
\begin{aligned}
\mathrm{E}[x_{ia}g_{jb}x_{i'c}g_{j'd}] &= \mathrm{E}[x_{ia}x_{i'c}]\,\mathrm{E}[g_{jb}g_{j'd}] \\
&= \Sigma^*_{ac}\mathbb{I}(i=i')\Pi_{bd}\mathbb{I}(j=j')
\end{aligned}
$$

So

$$
\begin{aligned}
\mathrm{E}\left[\left(\frac{\partial E}{\partial w_{ij}}\right)^2\right] &= \mathrm{E}\left[\left(\sum_a x_{ia}g_{ja}\right)^2\right] \\
&= \mathrm{E}\left[\sum_{a,b} x_{ia}g_{ja}x_{ib}g_{jb}\right] \\
&= \sum_{a,b}\Sigma^*_{ab}\Pi_{ab} \\
&= \langle\Sigma^*,\Pi\rangle
\end{aligned}
$$

Since Σ^* is BSB1 with $G\Sigma^* G = \upsilon^* G$ and $\Pi \in \mathcal{S}_B^G$, we get

$$\begin{aligned}
\mathrm{E}\left[\left(\frac{\partial E}{\partial w_{ij}}\right)^2\right] &= \langle \Sigma^*, \Pi \rangle \\
&= \langle \Sigma^*, G^{\otimes 2}\{\Pi\} \rangle \\
&= \langle G^{\otimes 2}\{\Sigma^*\}, \Pi \rangle \\
&= \langle \upsilon^* G, \Pi \rangle \\
&= \upsilon^* \operatorname{tr} \Pi.
\end{aligned}$$

H α-ReLU

Recall that α-ReLUs (Yang & Schoenholz (2017)) are, roughly speaking, the αth power of ReLU.

Definition H.1. The α-ReLU function $\rho_\alpha : \mathbb{R} \to \mathbb{R}$ sends $x \mapsto x^\alpha$ when $x > 0$ and $x \mapsto 0$ otherwise.

This is a continuous function for $\alpha > 0$ but discontinuous at 0 for all other α.

We briefly review what is currently known about the V and W transforms of ρ_α Cho & Saul (2009b); Yang & Schoenholz (2017).

Definition H.2. For any $\alpha > -\frac{1}{2}$, define $\mathsf{c}_\alpha = \frac{1}{\sqrt{\pi}} 2^{\alpha-1}\Gamma(\alpha + \frac{1}{2})$.

When considering only 1-dimensional Gaussians, $\mathrm{V}\rho_\alpha$ is very simple.

Proposition H.3. *If $\alpha > -\frac{1}{2}$, then for any $q \in \mathcal{S}_1 = \mathbb{R}^{\geq 0}$, $\mathrm{V}\rho_\alpha(q) = \mathsf{c}_\alpha q^\alpha$*

To express results of $\mathrm{V}\rho_\alpha$ on $\mathcal{S}_B$ for higher B, we first need the following

Definition H.4. Define

$$J_\alpha(\theta) := \frac{1}{2\pi \mathsf{c}_\alpha} (\sin\theta)^{2\alpha+1} \Gamma(\alpha+1) \int_0^{\pi/2} \frac{\mathrm{d}\eta \cos^\alpha \eta}{(1 - \cos\theta\cos\eta)^{1+\alpha}}$$

and $\mathfrak{J}_\alpha(c) = J_\alpha(\arccos c)$ for $\alpha > -1/2$.

Then

Proposition H.5. *For any $\Sigma \in \mathcal{S}_B$, let D be the diagonal matrix with the same diagonal as Σ. Then*

$$\mathrm{V}\rho_\alpha(\Sigma) = \mathsf{c}_\alpha D^{\alpha/2} \mathfrak{J}_\alpha(D^{-1/2}\Sigma D^{-1/2}) D^{\alpha/2}$$

where $\mathfrak{J}_\alpha$ is applied entrywise.

For example, J_α and $\mathfrak{J}_\alpha$ for the first few integral α are

$$\begin{aligned}
J_0(\theta) &= \frac{\pi - \theta}{\pi} \\
J_1(\theta) &= \frac{\sin\theta + (\pi - \theta)\cos\theta}{\pi} \\
J_2(\theta) &= \frac{3\sin\theta\cos\theta + (\pi - \theta)(1 + 2\cos^2\theta)}{3\pi} \\
\mathfrak{J}_0(c) &= \frac{\pi - \arccos c}{\pi} \\
\mathfrak{J}_1(c) &= \frac{\sqrt{1 - c^2} + (\pi - \arccos c)c}{\pi} \\
\mathfrak{J}_2(c) &= \frac{3c\sqrt{1 - c^2} + (\pi - \arccos c)(1 + 2c^2)}{3\pi}
\end{aligned}$$

One can observe very easily that Daniely et al. (2016); Yang & Schoenholz (2017)

Proposition H.6. *For each $\alpha > -1/2$, $\mathfrak{J}_\alpha(c)$ is an increasing and convex function on $c \in [0,1]$, and is continous on $c \in [0,1]$ and smooth on $c \in (0,1)$. $\mathfrak{J}_\alpha(1) = 1$, $\mathfrak{J}_\alpha(0) = \frac{1}{2\sqrt{\pi}} \frac{\Gamma(\frac{\alpha}{2}+\frac{1}{2})^2}{\Gamma(\alpha+\frac{1}{2})}$, and $\mathfrak{J}_\alpha(-1) = 0$.*

Yang & Schoenholz (2017) also showed the following fixed point structure

Theorem H.7. *For $\alpha \in [1/2, 1)$, $\mathfrak{J}_\alpha(c) = c$ has two solutions: an unstable solution at 1 ("unstable" meaning $\mathfrak{J}'_\alpha(1) > 1$) and a stable solution in $c^* \in (0,1)$ ("stable" meaning $\mathfrak{J}'_\alpha(c^*) < 1$).*

The α-ReLUs satisfy very interesting relations amongst themselves. For example,

Lemma H.8. *Suppose $\alpha > 1$. Then*

$$J_\alpha(\theta) = \cos\theta J_{\alpha-1}(\theta) + (\alpha-1)^2(2\alpha-1)^{-1}(2\alpha-3)^{-1}\sin^2\theta J_{\alpha-2}(\theta)$$
$$\mathfrak{J}_\alpha(c) = c\mathfrak{J}_{\alpha-1}(c) + (\alpha-1)^2(2\alpha-1)^{-1}(2\alpha-3)^{-1}(1-c^2)\mathfrak{J}_{\alpha-2}(c)$$

In additional, surprisingly, one can use differentiation to go from α to $\alpha+1$ *and* from α to $\alpha-1$!

Proposition H.9. *Yang & Schoenholz (2017) Suppose $\alpha > 1/2$. Then*

$$J'_\alpha(\theta) = -\alpha^2(2\alpha-1)^{-1}J_{\alpha-1}(\theta)\sin\theta$$
$$\mathfrak{J}'_\alpha(c) = \alpha^2(2\alpha-1)^{-1}\mathfrak{J}_{\alpha-1}(c)$$

so that

$$\mathfrak{J}_{\alpha-n}(c) = \left[\prod_{\beta=\alpha-n+1}^{\alpha} \beta^{-2}(2\beta-1)\right](\partial/\partial c)^n \mathfrak{J}_\alpha(c)$$

We have the following from Cho & Saul (2009b)

Proposition H.10. *Cho & Saul (2009b) For all $\alpha \geq 0$ and integer $n \geq 1$*

$$\begin{aligned}\mathfrak{J}_{n+\alpha}(c) &= \frac{\mathsf{c}_\alpha}{\mathsf{c}_{n+\alpha}}(1-c^2)^{n+\alpha+\frac{1}{2}}(\partial/\partial c)^n(\mathfrak{J}_\alpha(c)/(1-c^2)^{\alpha+\frac{1}{2}}) \\ &= \left[\prod_{\beta=\alpha}^{\alpha+n-1}(2\beta+1)\right]^{-1}(1-c^2)^{n+\alpha+\frac{1}{2}}(\partial/\partial c)^n(\mathfrak{J}_\alpha(c)/(1-c^2)^{\alpha+\frac{1}{2}})\end{aligned}$$

This implies in particular that we can obtain $\mathfrak{J}'_\alpha$ from $\mathfrak{J}_\alpha$ and $\mathfrak{J}_{\alpha+1}$.

Proposition H.11. *For all $\alpha \geq 0$,*

$$\mathfrak{J}'_\alpha(c) = (2\alpha+1)(1-c^2)^{-1}(\mathfrak{J}_{1+\alpha}(c) - c\mathfrak{J}_\alpha(c)))$$

Proof.

$$\begin{aligned}\mathfrak{J}_{1+\alpha}(c) &= \frac{\mathsf{c}_\alpha}{\mathsf{c}_{1+\alpha}}(1-c^2)^{1+\alpha+\frac{1}{2}}(\partial/\partial c)(\mathfrak{J}_\alpha(c)/(1-c^2)^{\alpha+\frac{1}{2}}) \\ &= (2\alpha+1)^{-1}(1-c^2)^{\alpha+3/2}(\mathfrak{J}'_\alpha(c)/(1-c^2)^{\alpha+1/2} + 2c(\alpha+1/2)\mathfrak{J}_\alpha(c)/(1-c^2)^{\alpha+3/2}) \\ &= (2\alpha+1)^{-1}\mathfrak{J}'_\alpha(c)(1-c^2) + c\mathfrak{J}_\alpha(c) \\ \mathfrak{J}'_\alpha(c) &= (2\alpha+1)(1-c^2)^{-1}(\mathfrak{J}_{1+\alpha}(c) - c\mathfrak{J}_\alpha(c))\end{aligned}$$

□

Note that we can also obtain this via Lemma H.8 and Proposition H.9.

I POSITIVE-HOMOGENEOUS FUNCTIONS IN 1 DIMENSION

Suppose for some $\alpha \in \mathbb{R}$, $\phi : \mathbb{R} \to \mathbb{R}$ is degree α positive-homogeneous, i.e. $\phi(rx) = r^\alpha \phi(x)$ for any $x \in R$, $r > 0$. The following simple lemma says that we can always express ϕ as linear combination of powers of α-ReLUs.

Proposition I.1. *Any degree α positive-homogeneous function $\phi : \mathbb{R} \to \mathbb{R}$ with $\phi(0) = 0$ can be written as $x \mapsto a\rho_\alpha(x) - b\rho_\alpha(-x)$.*

Proof. Take $a = \phi(1)$ and $b = \phi(-1)$. Then positive-homogeneity determines the value of ϕ on $\mathbb{R} \setminus \{0\}$ and it coincides with $x \mapsto a\rho_\alpha(x) - b\rho_\alpha(-x)$. □

As a result we can express the V and W transforms of any positive-homogeneous function in terms of those of α-ReLUs.

Proposition I.2. *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is degree α positive-homogeneous. By Proposition I.1, ϕ restricted to $\mathbb{R} \setminus \{0\}$ can be written as $x \mapsto a\rho_\alpha(x) - b\rho_\alpha(-x)$ for some a and b. Then for any PSD 2×2 matrix M,*

$$\begin{aligned}
\mathrm{V}\phi(M)_{11} &= (a^2+b^2)\mathrm{V}\rho_\alpha(M)_{11} = \mathsf{c}_\alpha(a^2+b^2)M_{11}^\alpha \\
\mathrm{V}\phi(M)_{22} &= (a^2+b^2)\mathrm{V}\rho_\alpha(M)_{22} = \mathsf{c}_\alpha(a^2+b^2)M_{22}^\alpha \\
\mathrm{V}\phi(M)_{12} = \mathrm{V}\phi(M)_{21} &= (a^2+b^2)\mathrm{V}\rho_\alpha(M)_{12} - 2ab\mathrm{V}\rho_\alpha(M')_{12} \\
&= \mathsf{c}_\alpha(M_{11}M_{22})^{\alpha/2}((a^2+b^2)\mathfrak{J}_\alpha(M_{12}/\sqrt{M_{11}M_{22}}) - 2ab\mathfrak{J}_\alpha(-M_{12}/\sqrt{M_{11}M_{22}}))
\end{aligned}$$

where $M' := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} M \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

Proof. We directly compute, using the expansion of ϕ into ρ_αs:

$$\begin{aligned}
\mathrm{V}\phi(M)_{11} &= \mathrm{E}[\phi(x)^2 : x \sim \mathcal{N}(0, M_{11})] \\
&= \mathrm{E}[a^2\rho_\alpha(x)^2 + b^2\rho_\alpha(-x)^2 + 2ab\rho_\alpha(x)\rho_\alpha(-x)] \\
&= a^2\,\mathrm{E}[\rho_\alpha(x)^2] + b^2\,\mathrm{E}[\rho_\alpha(-x)^2] \\
&= (a^2+b^2)\,\mathrm{E}[\rho_\alpha(x)^2 : x \sim \mathcal{N}(0, M_{11})] \qquad (50) \\
&= \mathsf{c}_\alpha(a^2+b^2)M_{11}^\alpha
\end{aligned}$$

where in Eq. (50) we used negation symmetry of centered Gaussians. The case of $\mathrm{V}\phi(M)_{22}$ is similar.

$$\begin{aligned}
\mathrm{V}\phi(M)_{12} &= \mathrm{E}[\phi(x)\phi(y) : (x,y) \sim \mathcal{N}(0, M)] \\
&= \mathrm{E}[a^2\rho_\alpha(x)\rho_\alpha(y) + b^2\rho_\alpha(-x)\rho_\alpha(-y) - ab\rho_\alpha(x)\rho_\alpha(-y) - ab\rho_\alpha(-x)\rho_\alpha(y)] \\
&= (a^2+b^2)\mathrm{V}\rho_\alpha(M)_{12} - 2ab\mathrm{V}\rho_\alpha(M')_{12} \\
&= \mathsf{c}_\alpha(M_{11}M_{22})^{\alpha/2}((a^2+b^2)\mathfrak{J}_\alpha(M_{12}/\sqrt{M_{11}M_{22}}) - 2ab\mathfrak{J}_\alpha(-M_{12}/\sqrt{M_{11}M_{22}}))
\end{aligned}$$

where in the last equation we have applied Proposition H.5. □

This then easily generalizes to PSD matrices of arbitrary dimension:

Corollary I.3. *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is degree α positive-homogeneous. By Proposition I.1, ϕ restricted to $\mathbb{R} \setminus \{0\}$ can be written as $x \mapsto a\rho_\alpha(x) - b\rho_\alpha(-x)$ for some a and b. Let $\Sigma \in \mathcal{S}_B$. Then*

$$\mathrm{V}\phi(\Sigma) = \mathsf{c}_\alpha D^{\alpha/2}\mathfrak{J}_\phi(D^{-1/2}\Sigma D^{-1/2})D^{\alpha/2}$$

where $\mathfrak{J}_\phi$ is defined below and is applied entrywise. Explicitly, this means that for all i,

$$\begin{aligned}
\mathrm{V}\phi(\Sigma)_{ii} &= \mathsf{c}_\alpha \Sigma_{ii}^\alpha \mathfrak{J}_\phi(1) \\
\mathrm{V}\phi(\Sigma)_{ij} &= \mathsf{c}_\alpha \mathfrak{J}_\phi(\Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}})\Sigma_{ii}^{\alpha/2}\Sigma_{jj}^{\alpha/2}
\end{aligned}$$

Definition I.4. Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is degree α positive-homogeneous. By Proposition I.1, ϕ restricted to $\mathbb{R} \setminus \{0\}$ can be written as $x \mapsto a\rho_\alpha(x) - b\rho_\alpha(-x)$ for some a and b. Define $\mathfrak{J}_\phi(c) := (a^2+b^2)\mathfrak{J}_\alpha(c) - 2ab\mathfrak{J}_\alpha(-c)$.

Let us immediately make the following easy but important observations.

Proposition I.5. *$\mathfrak{J}_\phi(1) = a^2 + b^2$ and $\mathfrak{J}_\phi(0) = (a^2 + b^2 - 2ab)\mathfrak{J}_\alpha(0) = (a-b)^2 \frac{1}{2\sqrt{\pi}} \frac{\Gamma(\frac{\alpha}{2}+\frac{1}{2})^2}{\Gamma(\alpha+\frac{1}{2})}$,*

Proof. Use the fact that $\mathfrak{J}_\alpha(-1) = 0$ and $\mathfrak{J}_\alpha(0) = \frac{1}{2\sqrt{\pi}} \frac{\Gamma(\frac{\alpha}{2}+\frac{1}{2})^2}{\Gamma(\alpha+\frac{1}{2})}$ by Proposition H.6. □

As a sanity check, we can easily compute that $\mathfrak{J}_{\mathrm{id}}(c) = 2\mathfrak{J}_1(c) - 2\mathfrak{J}_1(-c) = 2c$ because $\mathrm{id}(x) = \rho(x) - \rho(-x)$. By Corollary I.3 and $\mathsf{c}_1 = \frac{1}{2}$, this recovers the obvious fact that $\mathrm{V}(\mathrm{id})(\Sigma) = \Sigma$.

We record the partial derivatives of $\mathrm{V}\phi$.

Proposition I.6. *Let ϕ be positive homogeneous of degree α. Then for all i with $\Sigma_{ii} \neq 0$,*

$$\frac{\partial \mathrm{V}\phi(\Sigma)_{ii}}{\partial \Sigma_{ii}} = \mathsf{c}_\alpha \alpha \Sigma_{ii}^{\alpha-1} \mathfrak{J}_\phi(1)$$

For all $i \neq j$ with $\Sigma_{ii}, \Sigma_{jj} \neq 0$,

$$\begin{aligned}
\frac{\partial \mathrm{V}\phi(\Sigma)_{ii}}{\partial \Sigma_{ij}} &= 0 \\
\frac{\partial \mathrm{V}\phi(\Sigma)_{ij}}{\partial \Sigma_{ij}} &= \mathsf{c}_\alpha \Sigma_{ii}^{(\alpha-1)/2} \Sigma_{jj}^{(\alpha-1)/2} \mathfrak{J}'_\phi(c_{ij}) \\
\frac{\partial \mathrm{V}\phi(\Sigma)_{ij}}{\partial \Sigma_{ii}} &= \frac{1}{2} \mathsf{c}_\alpha \Sigma_{ii}^{\frac{1}{2}(\alpha-2)} \Sigma_{jj}^{\frac{1}{2}\alpha} (\alpha \mathfrak{J}_\phi(c_{ij}) - c_{ij} \mathfrak{J}'_\phi(c_{ij}))
\end{aligned}$$

where $c_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$ and $\mathfrak{J}'_\phi$ denotes its derivative.

Proposition I.7. *If $\phi(c) = a\rho_\alpha(c) - b\rho_\alpha(-c)$ with $\alpha > 1/2$ on $\mathbb{R} \setminus \{0\}$, then*

$$\begin{aligned}
\mathfrak{J}'_\phi(c) &= (2\alpha - 1)^{-1} \mathfrak{J}_{\phi'}(c) \\
&= (2\alpha+1)(1-c^2)^{-1}((a^2+b^2)\mathfrak{J}_{\alpha+1}(c) + 2ab\mathfrak{J}_{\alpha+1}(-c)) - c\mathfrak{J}_\phi(c) \\
\mathfrak{J}'_\phi(0) &= (a+b)^2 \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{\alpha}{2}+1)^2}{\Gamma(\alpha+\frac{1}{2})}
\end{aligned}$$

Proof. We have $\phi'(c) = a\alpha\rho_{\alpha-1}(c) + b\alpha\rho_{\alpha-1}(-c)$. On the other hand, $\mathfrak{J}'_\phi = (a^2+b^2)\mathfrak{J}'_\alpha(c) + 2ab\mathfrak{J}'_\alpha(-c)$ which by Proposition H.9 is $\alpha^2(2\alpha-1)^{-1}((a^2+b^2)\mathfrak{J}_{\alpha-1}(c) + 2ab\mathfrak{J}_{\alpha-1}(-c)) = (2\alpha-1)^{-1}\mathfrak{J}_{\phi'}(c)$. This proves the first equation.

With Proposition H.11,

$$\begin{aligned}
\mathfrak{J}'_\phi(c) &= (a^2+b^2)\mathfrak{J}'_\alpha(c) + 2ab\mathfrak{J}'_\alpha(-c) \\
&= (a^2+b^2)(2\alpha+1)(1-c^2)^{-1}(\mathfrak{J}_{\alpha+1}(c) - c\mathfrak{J}_\alpha(c)) + 2ab(2\alpha+1)(1-c^2)^{-1}(\mathfrak{J}_{\alpha+1}(-c) + c\mathfrak{J}_\alpha(-c)) \\
&= (2\alpha+1)(1-c^2)^{-1}((a^2+b^2)\mathfrak{J}_{\alpha+1}(c) + 2ab\mathfrak{J}_{\alpha+1}(-c)) - c\mathfrak{J}_\phi(c)
\end{aligned}$$

This gives the second equation.

Expanding $\mathfrak{J}_{\alpha+1}(0)$ With Proposition H.6, we get

$$\begin{aligned}
\mathfrak{J}'_\phi(0) &= (2\alpha+1)((a^2+b^2)\mathfrak{J}_{\alpha+1}(0) + 2ab\mathfrak{J}_{\alpha+1}(0)) \\
&= (2\alpha+1)(a+b)^2 \frac{1}{2\sqrt{\pi}} \frac{\Gamma(\frac{\alpha}{2}+1)^2}{\Gamma(\alpha+\frac{3}{2})}
\end{aligned}$$

Unpacking the definition of $\Gamma(\alpha + \frac{3}{2})$ then yields the third equation. □

In general, we can factor diagonal matrices out of $\mathrm{V}\phi$.

Proposition I.8. *For any $\Sigma \in \mathcal{S}_B$, D any diagonal matrix, and ϕ positive-homogeneous with degree α,*

$$\mathrm{V}\phi(D\Sigma D) = D^\alpha \mathrm{V}\phi(\Sigma) D^\alpha$$

Proof. For any i,

$$\begin{aligned}
\mathrm{V}\phi(D\Sigma D)_{ii} &= \mathrm{E}[\phi(x)^2 : x \sim \mathcal{N}(0, D_{ii}^2 \Sigma_{ii})] \\
&= \mathrm{E}[\phi(D_{ii}x)^2 : x \sim \mathcal{N}(0, \Sigma_{ii})] \\
&= \mathrm{E}[D_{ii}^{2\alpha}\phi(x)^2 : x \sim \mathcal{N}(0, \Sigma_{ii})] \\
&= D_{ii}^{2\alpha}\mathrm{V}\phi(\Sigma)_{ii}
\end{aligned}$$

For any $i \neq j$,

$$\begin{aligned}
\mathrm{V}\phi(D\Sigma D)_{ij} &= \mathrm{E}[\phi(x)\phi(y) : (x, y) \sim \mathcal{N}(0, \begin{pmatrix} D_{ii}^2\Sigma_{ii} & D_{ii}D_{jj}\Sigma_{ij} \\ D_{ii}D_{jj}\Sigma_{ji} & D_{jj}^2\Sigma_{jj} \end{pmatrix}) \\
&= \mathrm{E}[\phi(D_{ii}x)\phi(D_{jj}y) : (x, y) \sim \mathcal{N}(0, \begin{pmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{pmatrix})] \\
&= D_{ii}^{\alpha}D_{jj}^{\alpha}\,\mathrm{E}[\phi(x)\phi(y) : (x, y) \sim \mathcal{N}(0, \begin{pmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{pmatrix})] \\
&= D_{ii}^{\alpha}D_{jj}^{\alpha}\mathrm{V}\phi(\Sigma)_{ij}
\end{aligned}$$

□