Supplementary material

June 11, 2021

Contents

List of symbols

N	Number of samples	1.1
P	Number of features	1.1
$\mathbf{x}$	Input vector $(P \times 1)$	1.1
$\mathbf{X}$	Matrix of inputs $(N \times P)$	1.1
$\boldsymbol{\epsilon}$	Vector of noise $(N \times 1)$	1.1
$\mathbf{y}$	Vector of outputs $(N \times 1)$	1.1
$\mathbf{\Sigma}$	Covariance of $\mathbf{x}$ $(P \times P)$	1.1
$\mathbf{S}^2, \mathbf{U}$	Eigenvalues, eigenvectors of $\mathbf{\Sigma}$	1.1
σ_x^2	Total variance of the data $(\frac{1}{P}\mathrm{Tr}\mathbf{\Sigma})$	1.1
σ^2	Variance of noise	1.1
SNR	Signal to noise ratio	1.1
f_s	Fraction signal power	1.1
f_n	Fraction noise power	1.1
α	Measurement density N/P	1.1
β	Inverse measurement density P/N	1.1
$\mathbf{w}$	True weights	1.1
$\hat{\mathbf{w}}$	Estimated weights	1.1
λ	Regularization coefficient	1.1
$\mathbf{I}_P, \mathbf{I}_N$	Identity matrix $(P \times P,\ N \times N)$	1.1
$\mathcal{V}$	Unexplained variance	1.2
$\mathcal{F}$	Fraction unexplained variance	1.2
$\mathbf{v}$	Vector of mode-wise signal power	1.2
$\tilde{\lambda}$	Effective regularization	1.2
$\rho(\rho_f)$	(Fractional) participation ratio	1.2
z	Generating function parameter $(-1/\lambda)$	1.2
t_i, t_N, T_i, T_N	Generating functions	1.2
ζ	Related to ρ_f	1.2
$M(z)$	Moment generating function	1.3
$R(z)$	$\mathcal{R}$-transform	1.3
γ	Aspect ratio in 2-scale model (S_1/S_2)	5.0.2
H	Hessian of the training objective	6
$\mathbf{B}$	Inverse Hessian	6
$\tilde{\mathbf{B}}$	Related to inverse Hessian	6
D	Number of distinct data scales	6
ϵ_d	Ratio of successive scales	6
γ_x	Data strength measure $(\frac{1}{P}\mathrm{Tr}\mathbf{\Sigma}^2)$	6
$\rho(x)$	Spectral density of $\frac{1}{N}\mathbf{X}\mathbf{X}^T$	6
$G(x)$	Stieltjes transform of $\rho(x)$	6

1 Proof of the generalization error formula

1.1 Setting and statement

We study a generative model of data consisting of N independent identically distributed (iid) random Gaussian input vectors $\mathbf{x}^\mu \in \mathbb{R}^P$ for $\mu = 1, \ldots, N$, each drawn from a zero mean Gaussian with covariance matrix $\mathbf{\Sigma}$ (i.e. $\mathbf{x}^\mu \sim \mathcal{N}(0, \mathbf{\Sigma})$), N corresponding iid scalar noise realizations $\epsilon^\mu \sim \mathcal{N}\left(0, \sigma^2\right)$, and N outputs y^μ given by

$$y^\mu = \mathbf{x}^\mu \cdot \mathbf{w} + \epsilon^\mu,$$

where $\mathbf{w} \in \mathbb{R}^P$ is an unknown ground truth regression vector. The outputs y^μ decompose into a signal component $\mathbf{x}^\mu \cdot \mathbf{w}$ and noise component ϵ^μ and signal to noise ratio (SNR) given by the relative power

$$SNR := \frac{\mathrm{Var}[\mathbf{x} \cdot \mathbf{w}]}{\mathrm{Var}[\epsilon]} = \frac{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}{\sigma^2}, \tag{1}$$

plays a critical role in estimation performance. For convenience below we also define the fractional signal power $f_s := \frac{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w} + \sigma^2}$ and fractional noise power $f_n := \frac{\sigma^2}{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w} + \sigma^2}$. Note $f_s + f_n = 1$ and $SNR = \frac{f_s}{f_n}$.

We construct an estimate $\hat{\mathbf{w}}$ of $\mathbf{w}$ from the data $\{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^N$ using ridge regression:

$$\hat{\mathbf{w}} = \arg\min_w \frac{1}{N} \sum_{\mu=1}^N \left(y^\mu - \mathbf{x}^\mu \cdot w\right)^2 + \lambda ||w||^2. \tag{2}$$

The solution to this optimization problem is given by

$$\hat{\mathbf{w}} = \left(\mathbf{X}^T \mathbf{X} + \lambda N \mathbf{I}_P\right)^{-1} \mathbf{X}^T \mathbf{y}, \tag{3}$$

where $\mathbf{X}$ is an $N \times P$ matrix whose μ^{th} row is $\mathbf{x}^{\mu T}$, $\mathbf{y} \in \mathbb{R}^N$ with components y^μ and $\mathbf{I}_P$ is the $P \times P$ identity matrix.

Finally, define the following: let ($\beta := P/N$) $\alpha := N/P$ denote the (inverse) measurement density, let the eigendecomposition of the data covariance be $\mathbf{\Sigma} = \mathbf{U}\mathbf{S}^2\mathbf{U}^T$, and let the total variance of the data be $\mathbf{s}_x^2 := \frac{1}{P}\mathrm{Tr}\mathbf{\Sigma}$.

1.2 Proof using diagrammatic expansion

Here we prove the high-dimensional error formula, which states that the fraction unexplained variance, defined as

$$\mathcal{F} := \frac{\mathbb{E}_{\mathbf{x}^\mu, \epsilon^\mu, \mathbf{x}, \epsilon}\left[\left(y - \mathbf{x} \cdot \hat{\mathbf{w}}\right)^2\right]}{\mathbb{E}_{\mathbf{x}, \epsilon}\left[y^2\right]}, \tag{4}$$

converges to the following expression in the high-dimensional limit

$$\mathcal{F} = f_n + \frac{1}{\rho_f} \sum_{i=1}^P \left\{ f_s \hat{\mathbf{v}}_i^2 \left(\frac{\tilde{\lambda}}{S_i^2 + \tilde{\lambda}}\right)^2 + f_n \frac{1}{\alpha} \frac{1}{P} \left(\frac{S_i^2}{S_i^2 + \tilde{\lambda}}\right)^2 \right\}. \tag{5}$$

Here $f_n, f_s = \frac{\sigma^2}{\sigma^2+\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}, \frac{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}{\sigma^2+\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$ are the fractional noise and signal power, $\boldsymbol{\Sigma}$ has eigendecomposition $\mathbf{U}\mathbf{S}^2\mathbf{U}^T$, $\hat{\mathbf{v}}$ is the unit vector in the same direction as $\mathbf{v} := \mathbf{S}\mathbf{U}^T\mathbf{w}$, and

$$\lambda = \tilde{\lambda} + \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{\tilde{\lambda}S_j^2}{\tilde{\lambda}+S_j^2} \tag{6}$$

$$\rho_f = \left(\frac{d\tilde{\lambda}}{d\lambda}\right)^{-1}. \tag{7}$$

We will start by computing the raw unexplained variance $\mathcal{V} = \mathbb{E}_{\mathbf{x}^\mu,\epsilon^\mu,\mathbf{x},\epsilon}\left[(y-\mathbf{x}\cdot\hat{\mathbf{w}})^2\right]$, and then divide by $\mathrm{Var}\left[y\right] = \mathbb{E}_{\mathbf{x},\epsilon}\left[y^2\right]$ to obtain the fraction unexplained variance $\mathcal{F}$.

Expectations over $\epsilon^\mu, \mathbf{x}, \epsilon$ The three expectations in (4) over $\epsilon^\mu, \mathbf{x}, \epsilon$, the training noise, and test input and noise, are relatively straightforward. First performing the test example expectation over $\mathbf{x}, \epsilon$, we obtain

$$\begin{aligned}\mathbb{E}_{\mathbf{x}^\mu,\epsilon^\mu,\mathbf{x},\epsilon}\left[(y-\mathbf{x}\cdot\hat{\mathbf{w}})^2\right] &= \mathbb{E}_{\mathbf{x}^\mu,\epsilon^\mu}\mathbb{E}_{\mathbf{x},\epsilon}\left((\mathbf{x}\cdot\mathbf{w}+\epsilon)-\mathbf{x}\cdot\hat{\mathbf{w}}\right)^2\\ &= \mathbb{E}_{\mathbf{x}^\mu,\epsilon^\mu}\mathbb{E}_{\mathbf{x},\epsilon}\left(\mathbf{x}\cdot(\mathbf{w}-\hat{\mathbf{w}})+\epsilon\right)^2\\ &= \mathbb{E}_{\mathbf{x}^\mu,\epsilon^\mu}\left[(\mathbf{w}-\hat{\mathbf{w}})^T\Sigma(\mathbf{w}-\hat{\mathbf{w}})+\sigma^2\right].\end{aligned}$$

Now all of the random dependence on $\mathbf{x}^\mu, \epsilon^\mu$ is in the $(\mathbf{w}-\hat{\mathbf{w}})$'s. Using the definition of $\hat{\mathbf{w}}$, we have

$$\begin{aligned}\mathbf{w}-\hat{\mathbf{w}} &= \mathbf{w} - \left(\mathbf{X}^T\mathbf{X}+\lambda N\mathbf{I}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\mathbf{w}+\boldsymbol{\epsilon}\right)\\ &= \left(\mathbf{I}-\left(\mathbf{X}^T\mathbf{X}+\lambda N\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{X}\right)\mathbf{w} - \left(\mathbf{X}^T\mathbf{X}+\lambda N\mathbf{I}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\\ &= \lambda N\left(\mathbf{X}^T\mathbf{X}+\lambda N\mathbf{I}\right)^{-1}\mathbf{w} - \left(\mathbf{X}^T\mathbf{X}+\lambda N\mathbf{I}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\\ &= \mathcal{R}\left(\mathbf{X}\right)\left(\lambda\mathbf{w}-\frac{1}{N}\mathbf{X}^T\boldsymbol{\epsilon}\right),\end{aligned}$$

where $\mathcal{R}\left(\mathbf{X}\right) = \left(\frac{1}{N}\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}$ is the regularized inverse covariance, and on line 2 we've used the Woodbury matrix identity. This notation makes explicit that $\mathcal{R}\left(\mathbf{X}\right)$ only depends on the training inputs $\mathbf{x}^\mu$ and not the training noise

ϵ^μ. Using this notation, we now perform the expectation over ϵ^μ:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}^\mu,\epsilon^\mu,\mathbf{x},\epsilon}\left[(y-\mathbf{x}\cdot\hat{\mathbf{w}})^2\right] &= \mathbb{E}_{\mathbf{x}^\mu,\epsilon^\mu}\left[(\mathbf{w}-\hat{\mathbf{w}})^T\,\boldsymbol{\Sigma}\,(\mathbf{w}-\hat{\mathbf{w}})+\sigma^2\right] \\
&= \mathbb{E}_{\mathbf{x}^\mu}\mathbb{E}_{\epsilon^\mu}\left[\left(\lambda\mathbf{w}-\frac{1}{N}\mathbf{X}^T\boldsymbol{\epsilon}\right)^T\mathcal{R}\Sigma\mathcal{R}\left(\lambda\mathbf{w}-\frac{1}{N}\mathbf{X}^T\boldsymbol{\epsilon}\right)+\sigma^2\right] \\
&= \mathbb{E}_{\mathbf{x}^\mu}\left[\lambda^2\mathbf{w}^T\mathcal{R}\boldsymbol{\Sigma}\mathcal{R}\mathbf{w}+\frac{1}{N^2}\mathbb{E}_{\epsilon^\mu}\left[\boldsymbol{\epsilon}^T\mathbf{X}\mathcal{R}\boldsymbol{\Sigma}\mathcal{R}\mathbf{X}^T\boldsymbol{\epsilon}\right]+\sigma^2\right] \\
&= \mathbb{E}_{\mathbf{x}^\mu}\left[\lambda^2\mathbf{w}^T\mathcal{R}\boldsymbol{\Sigma}\mathcal{R}\mathbf{w}+\frac{\sigma^2}{N^2}\mathrm{Tr}\left[\mathbf{X}\mathcal{R}\boldsymbol{\Sigma}\mathcal{R}\mathbf{X}^T\right]+\sigma^2\right] \\
&= \sigma^2+\mathbf{w}^T\mathbb{E}_{\mathbf{x}^\mu}\left[\lambda^2\mathcal{R}\boldsymbol{\Sigma}\mathcal{R}\right]\mathbf{w}+\frac{\sigma^2}{N}\mathrm{Tr}\left[\frac{1}{N}\mathbb{E}_{\mathbf{x}^\mu}\left[\mathbf{X}\mathcal{R}\boldsymbol{\Sigma}\mathcal{R}\mathbf{X}^T\right]\right].
\end{aligned}$$

Eigendecomposition and change of variables The formula immediately above depends on the two matrices

$$\mathbb{E}_{\mathbf{x}^\mu}\left[\lambda^2\mathcal{R}\boldsymbol{\Sigma}\mathcal{R}\right] \tag{8}$$

$$\frac{1}{N}\mathbb{E}_{\mathbf{x}^\mu}\left[\mathbf{X}\mathcal{R}\boldsymbol{\Sigma}\mathcal{R}\mathbf{X}^T\right]. \tag{9}$$

Let the eigendecomposition of the data covariance be $\boldsymbol{\Sigma}=\mathbf{U}\mathbf{S}^2\mathbf{U}^T$. We can produce training inputs $\mathbf{X}$ with this covariance from standard normal variables $\mathbf{Z}$ through $\mathbf{X}=\mathbf{Z}\mathbf{S}\mathbf{U}^T$ since then the covariance is

$$\mathbb{E}\left[\frac{1}{N}\mathbf{X}^T\mathbf{X}\right]=\mathbb{E}\left[\frac{1}{N}\mathbf{U}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\mathbf{U}^T\right]=\mathbf{U}\mathbf{S}^2\mathbf{U}^T=\boldsymbol{\Sigma}.$$

Now substitute $\boldsymbol{\Sigma}=\mathbf{U}\mathbf{S}^2\mathbf{U}^T$ and $\mathbf{X}=\mathbf{Z}\mathbf{S}\mathbf{U}^T$ into the matrices (8) (9) we obtain

$$\mathbb{E}_{\mathbf{x}^\mu}\left[\lambda^2\mathcal{R}\boldsymbol{\Sigma}\mathcal{R}\right]=\mathbf{U}\mathbb{E}_{\mathbf{Z}}\left[\left(\mathbf{I}-z\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^{-1}\mathbf{S}^2\left(\mathbf{I}-z\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^{-1}\right]\mathbf{U}^T$$

$$\frac{1}{N}\mathbb{E}_{\mathbf{x}^\mu}\left[\mathbf{X}\mathcal{R}\boldsymbol{\Sigma}\mathcal{R}\mathbf{X}^T\right]=z^2\frac{1}{N}\mathbb{E}_{\mathbf{Z}}\left[\mathbf{Z}\mathbf{S}\left(\mathbf{I}-z\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^{-1}\mathbf{S}^2\left(\mathbf{I}-z\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^{-1}\mathbf{S}\mathbf{Z}^T\right],$$

where we've written the matrices in terms of $z=-\frac{1}{\lambda}$. Defining

$$T_P=\mathbb{E}_{\mathbf{Z}}\left[\left(\mathbf{I}-z\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^{-1}\mathbf{S}^2\left(\mathbf{I}-z\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^{-1}\right] \tag{10}$$

$$T_N=\frac{1}{N}\mathbb{E}_{\mathbf{Z}}\left[z\mathbf{Z}\mathbf{S}\left(\mathbf{I}-z\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^{-1}\mathbf{S}^2\left(\mathbf{I}-z\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^{-1}\mathbf{S}\mathbf{Z}^T\right], \tag{11}$$

the unexplained variance can be written

$$\mathcal{V}=\sigma^2+\mathbf{w}^T\mathbf{U}T_P\mathbf{U}^T\mathbf{w}+z\frac{\sigma^2}{N}\mathrm{Tr}\left[T_N\right]. \tag{12}$$

Recall that $z = -\frac{1}{\lambda}$. We expand all inverses in powers of z giving

$$T_P = \sum_n z^n \mathbb{E}_{\mathbf{Z}} \sum_{p+q=n} \left(\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^p \mathbf{S}^2 \left(\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^q =: \sum_n z^n A_n$$

$$T_N = \sum_n z^{n+1} \mathbb{E}_{\mathbf{Z}} \sum_{p+q=n} \frac{1}{\sqrt{N}}\mathbf{Z}\mathbf{S} \left(\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^p \mathbf{S}^2 \left(\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^q \frac{1}{\sqrt{N}}\mathbf{S}\mathbf{Z}^T =: \sum_n z^{n+1} B_{n+1}.$$

T_P, T_N can be interpreted as generating functions for the matrix sequences A_n, B_n. We will interpret these sequences as counting weighted paths through a certain graph, and then produce a recurrence relation satisfied by A_n, B_n, from which generating function equations will follow.

T_P, T_N can be interpreted as pathsums Generically, for an $h \times w$ matrix M, one can build a bipartite graph with a group of h nodes and on the left and a second group of w nodes on the right, and interpret M_{ij} as an edge weight between the i^{th} node on the left and the j^{th} node on the right. For a matrix product $(ML)_{ij} = \sum_k M_{ik}L_{kj}$, one can interpret the ij^{th} component as a sum over all paths which start at node i on the left, go through an intermediate node k, and end at node j on the right. To compute the total weight for the path, we must chain together the individual "edge weights" contained in M, L, that is $(i \to k \to j) \Rightarrow M_{ik}L_{kj}$

We will now apply this interpretation to the matrix sequences A_n, B_n. Both sequences are defined in terms of repeated products of $\mathbf{Z}S$ and $(\mathbf{Z}S)^T$. We can interpret these two matrices as describing the same edge weights in a bipartite graph with one group of N nodes (for the number of samples) and one group of P nodes (for the number of features). For example, the product

$$\left(\frac{1}{N}S\mathbf{Z}^T\mathbf{Z}S\right)_{ij} = \sum_\gamma \left(\frac{1}{\sqrt{N}}\mathbf{Z}S\right)^T_{i\gamma} \left(\frac{1}{\sqrt{N}}\mathbf{Z}S\right)_{\gamma j},$$

represents a two step path $i \to \gamma \to j$ with individual edge weights given by the matrix $\frac{1}{\sqrt{N}}\mathbf{Z}S$. From this it follows more generally that a power $\left[\left(\frac{1}{N}S\mathbf{Z}^T\mathbf{Z}S\right)^p\right]_{ij}$ will be a pathsum $i \cdots \xrightarrow{2p \text{ steps}} \cdots j$ where each path has a weight equal to the product of the individual traversed edge weights found in $\frac{1}{\sqrt{N}}\mathbf{Z}S$. We will write the set of $2n$ step paths from $i \to j$ as $\mathcal{P}_{2n}(i \to j)$. Note that because every $(\mathbf{Z}S)^T$ is followed by a $(\mathbf{Z}S)$, such paths always alternate between P-type and N-type nodes at each step.

Slightly abusing notation, the full expressions for A_n, B_n are interpreted as

$$\begin{aligned} A_n &= \mathbb{E}_{\mathbf{Z}} \sum_{p+q=n} \left(\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^p \mathbf{S}^2 \left(\frac{1}{N}\mathbf{S}\mathbf{Z}^T\mathbf{Z}\mathbf{S}\right)^q \\ &= \mathbb{E}_{\mathbf{Z}} \sum_{p+q=n} \sum_k \mathcal{P}_{2p}(i \to k) \cdot S_k^2 \cdot \mathcal{P}_{2q}(k \to j), \end{aligned}$$

and

$$\begin{aligned} B_{n+1} &= \mathbb{E}_{\mathbf{Z}} \sum_{p+q=n} \frac{1}{\sqrt{N}} \mathbf{ZS} \left(\frac{1}{N} \mathbf{SZ}^T \mathbf{ZS} \right)^p \mathbf{S}^2 \left(\frac{1}{N} \mathbf{SZ}^T \mathbf{ZS} \right)^q \frac{1}{\sqrt{N}} \mathbf{SZ}^T \\ &= \mathbb{E}_{\mathbf{Z}} \sum_{p+q=n} \sum_k \mathcal{P}_{2p+1} \left(\gamma \to k \right) \cdot S_k^2 \cdot \mathcal{P}_{2q+1} \left(k \to \eta \right). \end{aligned}$$

To distinguish the two halves of the bipartite graph, we'll use Latin indices like i, j, k whenever an index ranges over a feature (ie. one of the P nodes), and Greek indices whenever an index ranges over a sample (ie. one of the N nodes). So a generic path will be of the form $i \to \gamma \to j \to \eta \cdots$, which alternates between the two groups of the bipartite $N - P$ graph.

Pathsum dominated by "tree paths" The weight connecting nodes $\gamma \leftrightarrow i$ is $\left(\frac{1}{\sqrt{N}} \mathbf{ZS} \right)_{\gamma i} = \frac{1}{\sqrt{N}} \mathbf{Z}_{\gamma i} S_i$, which is a standard normal variable $\mathbf{Z}_{\gamma i}$ multiplied by a deterministic scalar $\frac{S_i^2}{\sqrt{N}}$. If the edge $\gamma \leftrightarrow i$ is only traversed once, then $\mathbf{Z}_{\gamma i}$ will appear as a first power in the path weight, and after taking expectations $\mathbb{E}_{\mathbf{Z}}$ the result will be 0. More generally, if any edge $\gamma \leftrightarrow i$ is traversed an odd number of times, the edge weight will include an odd moment of the corresponding standard normal variable $\mathbf{Z}_{\gamma i}$, making the full path weight 0 after taking expectations. So **only paths which traverse their edges an even number of times contribute** to A_n, B_n.

Any path starting and ending at different nodes must traverse some edge an odd number of times, so its path weight will be zero by the previous argument. So **only paths starting and ending at the same node contribute**. This is equivalent to saying A_n, B_n are diagonal, since $(A_n)_{ij}$ consists of paths $i \cdots \to \cdots j$ and similarly for B_n.

For example take the set of 4-step paths starting and ending at i. In order to contribute to the pathsum it must traverse all its edges an even number of times. This leaves three possibilities (Figure 1).

To determine how many paths there are of each type, we simply count the number of distinct P, N-type nodes one can choose (noting that i is the fixed starting/ending point). There are PN of the first type, N^2 of the second type, and N of the third type. Thus in the limit $N, P \to \infty$ with $\alpha = N/P$ fixed, the PN, N^2 paths of the first two types will dominate the N of the third type (the paths will also differ in their path weights, since, for example in type 3, the edge $i \leftrightarrow \gamma$ appears four times giving us a factor of $\mathbb{E}\left[\mathbf{Z}_{\gamma i}^4\right] = 3$, but these are order 1 factors which will not change asymptotic scaling with N, P).

A straightforward extension of this argument shows that for any number of steps, as $N, P \to \infty$ the pathsum is dominated by paths visiting the maximum number of distinct N, P-type nodes. These dominant paths cannot have any cycles (as in Figure 2), since one can always cut a cycle to obtain two independent branches, increasing the N, P scaling of the path type (Figure 2). This implies that **paths whose (paired) edges form a tree dominate the pathsum**. We've already demonstrated a special case of this fact in Figure 1: in terms of

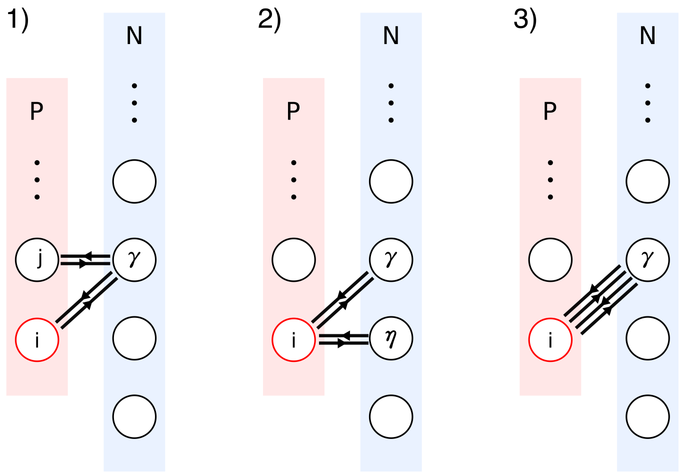

Figure 1: Three possible 4-step paths starting from node i. If edges going in opposite directions between the same two nodes are paired, the first two paths give rise to trees, while the third path will have two edge pairs connecting $i \leftrightarrow \gamma$, and so is not a tree.

edge pairs, path 1) forms a (linear) tree, path 2) forms a tree with root degree 2, and path 3) is not a tree, and so is insignificant in the limit.

Thus whenever we encounter $\mathcal{P}_{2n}\left(i \rightarrow i\right)$, we will restrict attention to the subset $\mathcal{T}ree_n\left(i \rightarrow i\right)$, defined as the set of trees with n edges (a path of $2n$ steps gives a tree with half as many edges n since the steps are paired).

Generating function for simple trees We start by computing the simpler generating functions

$$t_i = \sum_n z^n \cdot \mathcal{P}_{2n}\left(i \rightarrow i\right) = \sum_n z^n \cdot \mathcal{T}ree_n\left(i \rightarrow i\right) \tag{13}$$

$$t_\gamma = \sum_n z^n \cdot \mathcal{P}_{2n}\left(\gamma \rightarrow \gamma\right) = \sum_n z^n \cdot \mathcal{T}ree_n\left(\gamma \rightarrow \gamma\right), \tag{14}$$

which will allow us to illustrate the general approach.

Briefly, we will make use of the notion of a *combinatorial class*, which is a set of *objects* (here, trees), each of which has a *size* (here, number of edges), and a *weight* (here, the full weight computed by multiplying the step weights given in $\frac{1}{\sqrt{N}}\mathbf{Z}S$). Combinatorial classes have an associated *generating function* obtained by collecting all objects by size, and adding up the weights in each size group: $f = \sum_n z^n \left(\sum_{\substack{obj \in class \\ |obj|=n}} w_{obj}\right)$. Using the internal structure of the objects in a combinatorial class, one can often produce relatively simply equations for the generating function. The general framework for this is described in detail in 7.

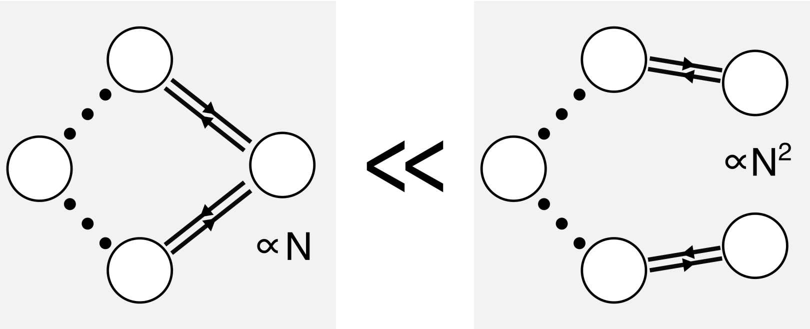

Figure 2: Paths with cycles (in terms of edge pairs) as in the path on the left, will have lower scaling with N, P than those without cycles.

To obtain an expression for $t_i = \sum_n z^n \cdot \mathcal{T}ree_n\,(i \to i)$, we first observe that this is exactly the generating function of the combinatorial class of trees where size n is the number of edges and the weight is computed using the edge weight product described above. Figure 3 shows an example of a 5-edge tree in this class with each distinct node marked by a different letter.

Using Figure 3 as an aid, any tree based at i can be expressed as an ordered sequence of subtrees attached to the root at i, which we'll write as $t_i = \mathrm{SEQ}\,(subtree)$. Each of these subtrees consists of a proper subtree and a root edge connecting the subtree to the root: $t_i = \mathrm{SEQ}\left(\substack{root\\edge} \times \substack{proper\\subtree}\right)$. Each subtree is based at some (ie. any) N-type node γ, so we write $t_i = \mathrm{SEQ}\left(\bigcup_\gamma \substack{root\\edge\\i\leftrightarrow\gamma} \times \substack{proper\\subtree\\at\ \gamma}\right)$. Finally, the proper subtree based at γ is itself a general tree based at node γ, which from (14) is just t_γ, so we have $t_i = \mathrm{SEQ}\left(\bigcup_\gamma \substack{root\\edge\\i\leftrightarrow\gamma} \times t_\gamma\right)$. Identical reasoning gives an analogous specification $t_\gamma = \mathrm{SEQ}\left(\bigcup_i \substack{root\\edge\\\gamma\leftrightarrow i} \times t_i\right)$.

To determine the combinatorial class associated to $\substack{root\\edge\\i\leftrightarrow\gamma}$, observe that a root edge is just a tree of size 1. To determine the weight of a particular root edge $i \leftrightarrow \gamma$, observe that it is traversed exactly twice, each time acquiring an edge factor of $\frac{1}{\sqrt{N}}\left(\mathbf{Z}S\right)_{\gamma i}$ giving $\frac{1}{N}\left(\mathbf{Z}S\right)^2_{\gamma i}$, which after taking expectations becomes $\frac{S_i^2}{N}$ ($\mathbf{Z}_{\gamma i}$ is standard normal and hence $\mathbb{E}\left[\mathbf{Z}^2_{\gamma i}\right] = 1$). Since it has size 1 and weight $\frac{S_i^2}{N}$, its combinatorial class is just $\frac{S_i^2}{N}\mathcal{Z}$. Our specifications are then

$$t_i = \mathrm{SEQ}\left(\bigcup_\gamma \frac{S_i^2}{N}\mathcal{Z} \times t_\gamma\right)$$

$$t_\gamma = \mathrm{SEQ}\left(\bigcup_i \frac{S_i^2}{N}\mathcal{Z} \times t_i\right).$$

Using the rules developed in sec (7), these immediately translate to generating

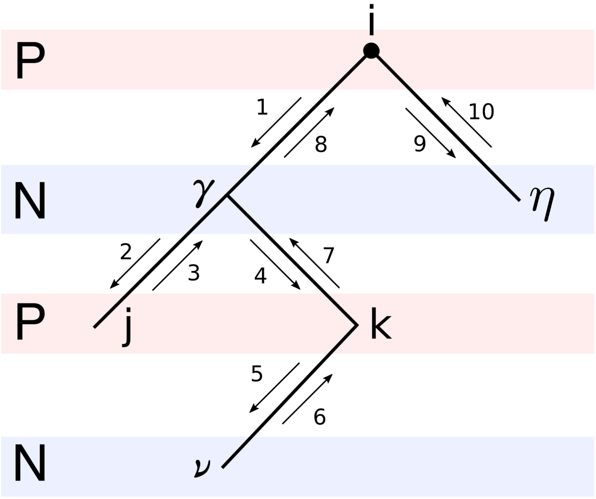

Figure 3:

function equations

$$t_i = \frac{1}{1 - \sum_\gamma \frac{S_i^2}{N} z t_\gamma}$$

$$t_\gamma = \frac{1}{1 - \sum_i \frac{S_i^2}{N} z t_i}.$$

The t_γ equation doesn't depend on γ, so we'll write t_N for the single value of t_γ for all γ. Rearranging then gives

$$t_i = 1 + z S_i^2 t_N t_i$$

$$t_N = 1 + \frac{1}{N} \sum_i (t_i - 1).$$

Full generating functions We will now obtain expressions for T_P, T_N, shown above to be

$$(T_P)_{ii} = \sum_n z^n \cdot \sum_{p+q=n} \sum_k \mathcal{P}_{2p}(i \to k) \cdot S_k^2 \cdot \mathcal{P}_{2q}(k \to i)$$

$$(T_N)_{\gamma\gamma} = \sum_n z^n \cdot \sum_{p+q=n} \sum_k \mathcal{P}_{2p+1}(\gamma \to k) \cdot S_k^2 \cdot \mathcal{P}_{2q+1}(k \to \gamma).$$

$(T_P)_{ii}$ can be intepreted as the generating function of the class of "pointed trees", consisting of trees based at i together with a distinguished node k (the "pointed-to" node) at some point along the path, which multiplies by an extra

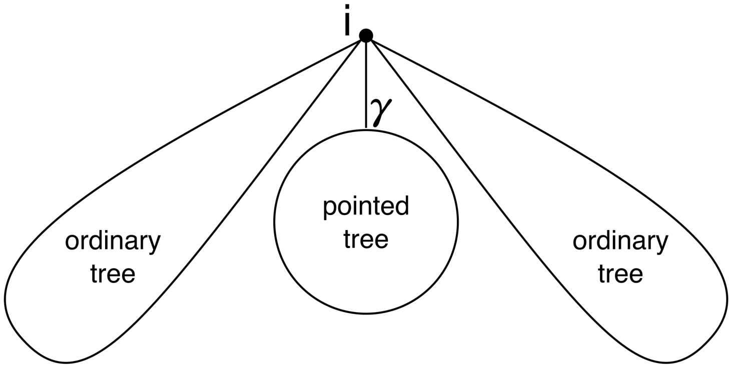

Figure 4: A pointed tree based at i - ie. a tree with a distinguished node visited sometime during the path traversal - can be decomposed into a maximal left tree traversed before the point, a maximal right tree traversed after the point, and finally, when the middle segment is longer than a single node, a root edge connecting the root i to a smaller pointed tree based at some (any) node γ.

weight of S_k^2. Similarly, $(T_N)_{\gamma\gamma}$ is the class of pointed subtrees based at N-type node γ.

We will now describe the structure of these classes in such a way that we can obtain generating function equations. Take a pointed tree based at i. Start at i and follow the path outward. Before and after reaching the pointed-to node k, it will return to the original node i some number of times, eg.

$$i_1 \to \cdots \to i_\ell \to \cdots \to k \to \cdots \to i_{\ell+1} \to \cdots \to i_L$$

We immediately have $(T_P)_{ii} = \binom{\text{left segment}}{i_1 \to \cdots \to i_\ell} \times \binom{\text{mid segment}}{i_\ell \to \cdots k \cdots \to i_\ell} \times \binom{\text{right segment}}{i_{\ell+1} \to \cdots \to i_L}$. The left and right segments $i \to \cdots \to i_\ell$ and $i_{\ell+1} \to \cdots \to i_\ell$ become arbitrary unpointed tree paths based at i, that is, members of t_i from above, so we have $(T_P)_{ii} = t_i \times \binom{\text{mid segment}}{i_k \to \cdots j \cdots \to i_k} \times t_i$.

The middle segment $i_\ell \to \cdots \to k \to \cdots \to i_{\ell+1}$ is either a single node or multiple nodes, $\binom{\text{mid segment}}{i_\ell \to \cdots k \cdots \to i_\ell} = \binom{single}{node} \cup \binom{mult}{nodes}$. If it is a single node, the pointed-to node k is actually i, and the segment is just $\cdots \to i_\ell / i_{\ell+1} \to \cdots$ (where the ℓ's would have been numbered accordingly) and so its class is just $S_i^2 \mathbf{1}$ - ie. a single object of size 0 edges, with the extra pointed weight factor S_i^2. Otherwise it is multiple steps, and it can be decomposed into a proper pointed subtree and a root edge connecting it to the root (Figure 4), ie. $\binom{mult}{nodes} = \bigcup_\gamma \begin{matrix} root \\ edge \\ i \leftrightarrow \gamma \end{matrix} \times \begin{pmatrix} pointed \\ subtree \\ at\ \gamma \end{pmatrix}$ (there is only need for one root edge/subtree rather than a sequence since, by construction, all others were absorbed into the left and right segments). The root edge has class $\frac{S_i^2}{N} \mathcal{Z}$, and the pointed subtree is just an arbitrary pointed subtree based at an N-type node, ie. $\bigcup_\gamma (T_N)_{\gamma\gamma}$. Putting these together, we have

$$(T_P)_{ii} = t_i \times \left(S_i^2 \mathbf{1} \cup \left(\frac{S_i^2}{N} \mathcal{Z} \times \bigcup_\gamma (T_N)_{\gamma\gamma} \right) \right) \times t_i.$$

Similarly, every object of $(T_N)_{\gamma\gamma}$ is a pointed tree path starting at N-space node γ. It can be decomposed the same way as T_P, except that the middle segment must be multiple steps, since γ is an N-type node, but the pointed-to node is P-type. So we just remove the single node case $S_i^2\mathbf{1}$ and obtain the specification

$$(T_N)_{\gamma\gamma} = t_\gamma \times \left(\bigcup_i \frac{S_i^2}{N} \mathcal{Z} \times (T_P)_{ii} \right) \times t_\gamma.$$

Collecting these two specifications and translating to generating function equations, we have

$$\begin{aligned} (T_P)_{ii} &= t_i^2 S_i^2 \left(1 + z \frac{1}{N} \sum_\gamma (T_N)_{\gamma\gamma} \right) \\ (T_N)_{\gamma\gamma} &= z t_\gamma^2 \sum_i \frac{S_i^2}{N} (T_P)_{ii}. \end{aligned}$$

Noting again that the $(T_N)_{\gamma\gamma}$ equation doesn't depend on γ (we showed above t_γ is independent of γ and just write t_N), we'll use write T_N for the single value of $(T_N)_{\gamma\gamma}$ for all γ, which simplifies these equations to

$$\begin{aligned} (T_P)_{ii} &= t_i^2 S_i^2 (1 + z T_N) \\ T_N &= z t_N^2 \sum_i \frac{S_i^2}{N} (T_P)_{ii}. \end{aligned}$$

Substitution into the error expression Collecting these results, we can write the unexplained variance $\mathcal{V}$ in terms of T_P, T_N as follows,

$$\mathcal{V} = \sigma^2 + \mathbf{w}^T U T_P U^T \mathbf{w} + z \frac{\sigma^2}{N} \mathrm{Tr}\left[T_N\right],$$

where T_N, T_P are the solutions of the following system (to make the notation less cumbersome, we'll write T_i for $(T_P)_{ii}$

$$t_i = 1 + z S_i^2 t_N t_i \tag{15}$$

$$t_N = 1 + \frac{1}{N} \sum_i (t_i - 1) \tag{16}$$

$$T_i = (1 + z T_N) t_i^2 S_i^2 \tag{17}$$

$$T_N = z t_N^2 \frac{1}{N} \sum_i S_i^2 T_i. \tag{18}$$

Substituting (17) into (18) gives $T_N = zt_N^2 \left(1 + zT_N\right) \frac{1}{N} \sum_i S_i^4 t_i^2$. Eq. (15) gives $t_N^2 S_i^4 t_i^2 = \frac{1}{z^2} \left(t_i - 1\right)^2$, so substituting gives

$$\begin{aligned} T_N &= \frac{1}{z} \left(1 + zT_N\right) \frac{1}{N} \sum_i \left(t_i - 1\right)^2 \\ \Rightarrow T_N &= \frac{1}{z} \frac{\zeta}{1 - \zeta}, \end{aligned}$$

where $\zeta = \frac{1}{N} \sum_i \left(t_i - 1\right)^2$.

Next, substituting this back into Eq. (17), we obtain

$$T_i = S_i^2 t_i^2 \left(1 + z \frac{1}{z} \frac{\zeta}{1 - \zeta}\right) = \frac{S_i^2 t_i^2}{1 - \zeta}.$$

Using these formulas, we can express the cost entirely in terms of $\tilde{\lambda} := \frac{1}{\frac{1}{N} \mathrm{Tr}\left[\mathbb{E}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T + \lambda \mathbf{I}_N\right)^{-1}\right]\right]}$. By comparing to how t_N was defined, we see $\tilde{\lambda} = \lambda / t_N$. Substituting into $t_N = 1 + \frac{1}{N} \sum_i \left(\frac{1}{1 - zS_i^2 t_N} - 1\right)$ we get

$$\begin{aligned} \lambda &= \tilde{\lambda} - \tilde{\lambda} \frac{1}{N} \sum_i \frac{S_i^2}{\tilde{\lambda} + S_i^2} \\ \zeta &= \frac{1}{N} \sum_i \left(\frac{S_i^2}{\tilde{\lambda} + S_i^2}\right)^2. \end{aligned}$$

Differentiating with respect to λ and solving finally gives

$$\frac{d\tilde{\lambda}}{d\lambda} = \frac{1}{1 - \zeta}. \tag{19}$$

So

$$\begin{aligned} T_N &= \frac{1}{z} \frac{\zeta}{1 - \zeta} = \frac{1}{z} \tilde{\lambda}' \frac{1}{N} \sum_i \left(\frac{S_i^2}{\tilde{\lambda} + S_i^2}\right)^2 \\ T_i &= \frac{S_i^2 t_i^2}{1 - \zeta} = \tilde{\lambda}' S_i^2 \left(\frac{\tilde{\lambda}}{\tilde{\lambda} + S_i^2}\right)^2. \end{aligned}$$

So the unexplained variance is

$$\mathcal{V} = \sigma^2 + \tilde{\lambda}' \left[\sum_i \left(\frac{\tilde{\lambda}}{S_i^2 + \tilde{\lambda}}\right)^2 S_i^2 \left|\mathbf{U}^T \mathbf{w}\right|_i^2 + \sigma^2 \frac{1}{N} \sum_i \left(\frac{S_i^2}{S_i^2 + \tilde{\lambda}}\right)^2\right].$$

with $\tilde{\lambda}$ equal to the solution of

$$\lambda = \tilde{\lambda} - \frac{1}{\alpha} \frac{1}{P} \sum_{j=1}^{P} \frac{\tilde{\lambda} S_j^2}{\tilde{\lambda} + S_j^2}. \tag{20}$$

Finally, we note that using the definition $\tilde{\lambda} := \frac{1}{\frac{1}{N}\mathrm{Tr}\left[\mathbb{E}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T+\lambda\mathbf{I}_N\right)^{-1}\right]\right]}$, we have

$$\frac{d\tilde{\lambda}}{d\lambda} = N\frac{\mathrm{Tr}\left[\mathbb{E}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T+\lambda\mathbf{I}_N\right)^{-2}\right]\right]}{\mathrm{Tr}\left[\mathbb{E}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T+\lambda\mathbf{I}_N\right)^{-1}\right]\right]^2} = N\frac{\mathbb{E}\left[\mathrm{Tr}\left[\tilde{\mathbf{B}}^2\right]\right]}{\mathbb{E}\left[\mathrm{Tr}\left[\tilde{\mathbf{B}}\right]\right]^2} = \frac{N}{\rho} = \frac{1}{\rho_f}, \tag{21}$$

where ρ (ρ_f) is the ordinary (fractional) participation ratio, which proves the relation $\rho_f = \left(\frac{d\tilde{\lambda}}{d\lambda}\right)^{-1}$.

Now define $\mathbf{v} := \mathbf{S}\mathbf{U}^T\mathbf{w}$, and set $\hat{\mathbf{v}}$ to the unit vector in the same direction as $\mathbf{v}$. Note that $\mathbf{v}^T\mathbf{v} = \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} = \sigma^2 SNR$. We now rewrite $\mathcal{V}$ in terms of $\hat{\mathbf{v}}$ and α and substitute $\rho_f = \left(\frac{d\tilde{\lambda}}{d\lambda}\right)^{-1}$ to obtain

$$\mathcal{V} = \sigma^2 + \frac{1}{\rho_f}\sum_{i=1}^{P}\left\{\left(\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}\right)\hat{\mathbf{v}}_i^2\left(\frac{\tilde{\lambda}}{S_i^2+\tilde{\lambda}}\right)^2 + \sigma^2\frac{1}{\alpha}\frac{1}{P}\left(\frac{S_i^2}{S_i^2+\tilde{\lambda}}\right)^2\right\}. \tag{22}$$

Finally, divide (22) by $\mathrm{Var}\left[y\right] = \sigma^2 + \mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$ to obtain the fraction unexplained variance:

$$\mathcal{F} = f_n + \frac{1}{\rho_f}\sum_{i=1}^{P}\left\{f_s\hat{\mathbf{v}}_i^2\left(\frac{\tilde{\lambda}}{S_i^2+\tilde{\lambda}}\right)^2 + f_n\frac{1}{\alpha}\frac{1}{P}\left(\frac{S_i^2}{S_i^2+\tilde{\lambda}}\right)^2\right\}, \tag{23}$$

which is the error formula given above.

1.3 Derivation of $\tilde{\lambda}$ equation using free probability

We will now use the framework of free probability, where we work in the algebra of matrices with expectation functional $\phi\left(A\right) = \frac{1}{N}\mathrm{Tr}\left[\mathbb{E}\left[A\right]\right]$.

We start by observing

$$\begin{aligned}
\tilde{\lambda}^{-1} &= \frac{1}{N}\mathrm{Tr}\left[\tilde{\mathbf{B}}\right] \\
&= \frac{1}{N}\mathrm{Tr}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T+\lambda\mathbf{I}_N\right)^{-1}\right] \\
&= -z\sum_{n\geq 0}z^n\frac{1}{N}\mathrm{Tr}\left[\mathbb{E}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T\right)^n\right]\right] \\
&= -z\left[1+\sum_{n\geq 1}m_n z^n\right] \\
&= -zM\left(z\right),
\end{aligned}$$

where $z = \frac{1}{\lambda}$, and we define the moments $m_n := \frac{1}{N}\mathrm{Tr}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T\right)^n\right]$ and the corresponding moment generating function $M\left(z\right)$ ([1] p198). The R-transform defined as $R(z) := \sum_{n\geq 0} \kappa_{n+1} z^n$, where κ are the free cumulants [1], is related to the moment generating function through

$$\frac{1}{M(z)} + zR(zM(z)) = 1, \tag{24}$$

thus knowledge of the free cumulants of the variable $x := \frac{1}{N}\mathbf{X}\mathbf{X}^T$ is sufficient to determine the moment generating function M.

Let the eigendecomposition of the true covariance $\mathbf{\Sigma}$ be $\mathbf{U}\mathbf{S}^2\mathbf{U}^T$. Thus $\mathbf{X}$ can be generated as $\mathbf{X} = \mathbf{Z}\mathbf{S}\mathbf{U}^T$ where $\mathbf{Z}$ is a matrix of iid standard normal variables, which implies $x = \frac{1}{N}\mathbf{Z}\mathbf{S}^2\mathbf{Z}^T$. Assume for now (and without loss of generality) that $\mathbf{\Sigma}$ is a D-level covariance - that is, the scales in $\mathbf{S}$ appear in D blocks of P_d components with value S_d^2. We can write x as

$$x = \frac{1}{N}\sum_{d=1}^{D} S_d^2 \mathbf{Z}_d \mathbf{Z}_d^T, \tag{25}$$

where $\mathbf{Z}_d$ is a $N \times P_d$ standard normal matrix containing the P_d columns of $\mathbf{Z}$ corresponding to S_d^2.

In the limit of large N, P the terms of the sum (25) are freely independent, so the cumulants of x can be computed by computing the cumulants of the terms separately and adding the results.

To compute the cumulants of the $\mathbf{Z}_d\mathbf{Z}_d^T$, we start by observing that $\mathbf{Z}_d\mathbf{Z}_d^T$ is itself a sum of P_d independent outer products of the columns of $\mathbf{Z}_d$. Each of these can be written zz^T, whose moments are

$$\mu_n = \frac{1}{N}\mathrm{Tr}\left[\mathbb{E}\left[\left(zz^T\right)^n\right]\right] = \frac{1}{N}\mathbb{E}\left[\left(z^T z\right)^n\right] = N^{n-1}, \tag{26}$$

and so the moments of $\frac{1}{N}S_d^2 zz^T$ are $m_n = \frac{N^{n-1}}{N^n}S_d^{2n} = \frac{1}{N}\left(S_d^2\right)^n$. Thus $\frac{1}{N}S_d^2 zz^T$ is a free Bernoulli variable with rate 1 and jump size S_d^2, and its cumulants are $\kappa_n = \frac{1}{N}\left(S_d^2\right)^n + O\left(\frac{1}{N^2}\right)$ ([1] p204). Adding P_d of these, we obtain

$$\kappa_n\left(\frac{1}{N}S_d^2\mathbf{Z}_d\mathbf{Z}_d^T\right) = \frac{1}{\alpha}f_d\left(S_d^2\right)^n + O\left(\frac{1}{N}\right). \tag{27}$$

Now using the free independence of blocks, $\kappa_n(x) = \frac{1}{\alpha}\sum_d f_d\left(S_d^2\right)^n$. Computing the R transform gives

$$R(z) = \sum_{n\geq 0}\kappa_{n+1}z^n = \frac{1}{\alpha}\sum_d \frac{f_d S_d}{1 - zS_d^2}. \tag{28}$$

Finally, plugging this expression into (24), we obtain

$$1 + \frac{1}{\alpha}\sum_d \frac{zf_d S_d M}{1 - zMS_d^2} = M. \tag{29}$$

Now summing over individual scales rather than summing over constant-S_d blocks, we can write this as

$$M = 1 + \frac{1}{\alpha}\frac{1}{P}\sum_{i=1}^{P}\frac{zS_iM}{1 - zS_i^2M}. \tag{30}$$

Finally, substituting $\tilde{\lambda} = -\frac{1}{zM}$ and using the fact that $z = -\frac{1}{\lambda}$, we obtain the following equation for $\tilde{\lambda}$

$$\lambda = \tilde{\lambda}\left(1 - \frac{1}{\alpha}\frac{1}{P}\sum_{i=1}^{P}\frac{S_i}{\tilde{\lambda} + S_i^2}\right), \tag{31}$$

which is exactly the equation for $\tilde{\lambda}$ derived by diagrammatic expansion above.

2 Scalar case

Here we derive an approximate formula for the case of a single input feature, ie. where $P = 1$, as a helpful comparison to the high dimensional case. The weight vector $\mathbf{w}$ is replaced by the scalar weight w, the input matrix $\mathbf{X}$ is replaced by a column vector $\mathbf{x}$, and the covariance $\mathbf{\Sigma}$ is now a single scalar S^2. With these changes, the training objective function is

$$\frac{1}{N}\sum_{i=1}^{N}(y^\mu - x^\mu w)^2 + \lambda w^2, \tag{32}$$

whose minimum occurs at

$$\begin{aligned}\hat{w} &= \left(\frac{1}{N}\mathbf{x}^T\mathbf{x} + \lambda\right)^{-1}\frac{1}{N}\mathbf{x}^T\mathbf{y}\\ &= \frac{\frac{1}{N}\mathbf{x}^T\mathbf{x}}{\frac{1}{N}\mathbf{x}^T\mathbf{x} + \lambda}w + \frac{\frac{1}{N}\mathbf{x}^T\boldsymbol{\epsilon}}{\frac{1}{N}\mathbf{x}^T\mathbf{x} + \lambda}.\end{aligned}$$

Starting with the the unexplained variance gives

$$\begin{aligned}\mathcal{V} &= \mathbb{E}_{x^\mu,\epsilon^\mu,x,\epsilon}\left[(y - \hat{w}x)^2\right]\\ &= \mathbb{E}_{x^\mu,\epsilon^\mu,x,\epsilon}\left[(wx + \epsilon - \hat{w}x)^2\right]\\ &= \sigma^2 + S^2\mathbb{E}_{x^\mu,\epsilon^\mu}\left[(w - \hat{w})^2\right].\end{aligned}$$

The term in the expectation can be written $\frac{\lambda}{\frac{1}{N}\mathbf{x}^T\mathbf{x}+\lambda}w - \frac{\frac{1}{N}\mathbf{x}^T\boldsymbol{\epsilon}}{\frac{1}{N}\mathbf{x}^T\mathbf{x}+\lambda}$. These two terms are uncorrelated, so we have

$$\mathcal{V} = \sigma^2 + S^2w^2\mathbb{E}\left(\frac{\lambda}{\frac{1}{N}\mathbf{x}^T\mathbf{x} + \lambda}\right)^2 + S^2\mathbb{E}\left(\frac{\frac{1}{N}\mathbf{x}^T\boldsymbol{\epsilon}}{\frac{1}{N}\mathbf{x}^T\mathbf{x} + \lambda}\right)^2. \tag{33}$$

If N is large, the quantity $\frac{1}{N}\mathbf{x}^T\mathbf{x}$ approaches $\mathbb{E}\left[\frac{1}{N}\mathbf{x}^T\mathbf{x}\right] = S^2$, while $\frac{1}{N}\mathbf{x}^T\boldsymbol{\epsilon} = \sigma S\left(\frac{1}{N}\mathbf{z}_1^T\mathbf{z}_2\right)$, where $\mathbf{z}_1, \mathbf{z}_2$ are independent standard normal vectors, is $\sigma S \cdot O\left(\frac{1}{\sqrt{N}}\right)$. Substituting these expressions, and dropping the O for clarity, we obtain

$$\mathcal{V} \approx \sigma^2 + S^2 w^2 \left(\frac{\lambda}{S^2+\lambda}\right)^2 + \sigma^2 \frac{1}{N}\left(\frac{S^2}{S^2+\lambda}\right)^2. \tag{34}$$

Finally, dividing (34) by $\mathrm{Var}\left[y\right] = \sigma^2 + S^2$, and noting that $f_s, f_n = \frac{S^2w^2}{\sigma^2+S^2w^2}, \frac{\sigma^2}{\sigma^2+S^2w^2}$, we obtain the fractional variance explained:

$$\mathcal{F}_{scalar} \approx f_n + f_s\left(\frac{\lambda}{S^2+\lambda}\right)^2 + f_n\frac{1}{N}\left(\frac{S^2}{S^2+\lambda}\right)^2. \tag{35}$$

3 Analysis of $\tilde{\lambda}$

Recall from (20) the corrected regularization $\tilde{\lambda}$ satisfies

$$\lambda = \tilde{\lambda}\left(1 - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda}+S_j^2}\right). \tag{36}$$

3.1 Bounds

First we show

$$\begin{cases} \lambda \le \tilde{\lambda} \le \frac{1}{1-\frac{1}{\alpha}}\lambda & \alpha > 1 \\ \lambda \le \tilde{\lambda} & \alpha < 1. \end{cases} \tag{37}$$

Using $0 \le \sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda}+S_j^2} \le P$ we obtain

$$\lambda = \tilde{\lambda}\left(1 - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda}+S_j^2}\right) \le \tilde{\lambda}$$
$$\lambda = \tilde{\lambda}\left(1 - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda}+S_j^2}\right) \ge \left(1 - \frac{1}{\alpha}\right)\tilde{\lambda}.$$

Thus if $\alpha > 1$ we have

$$\lambda \le \tilde{\lambda} \le \frac{1}{1-\frac{1}{\alpha}}\lambda. \tag{38}$$

Otherwise, the second inequality is vacuous and we only have

$$\lambda \le \tilde{\lambda}, \tag{39}$$

which completes the proof of the bounds in (37).

We now show $\lambda \leq \tilde{\lambda} \leq \lambda + \frac{\sigma_x^2}{\alpha}$ where $\sigma_x^2 = \frac{1}{P}\mathrm{Tr}\mathbf{\Sigma}$, and for large λ, $\tilde{\lambda}$ approaches this upper bound with error $O\left(1/\lambda\right)$. The bound $\lambda \leq \tilde{\lambda}$ has already been proven above. Next, for any $\lambda, \tilde{\lambda}$, we can use $\frac{S_j^2/\tilde{\lambda}}{1+S_j^2/\tilde{\lambda}} \leq S_j^2/\tilde{\lambda}$ to obtain

$$\lambda = \tilde{\lambda}\left(1 - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2/\tilde{\lambda}}{1+S_j^2/\tilde{\lambda}}\right) \geq \tilde{\lambda} - \frac{\sigma_x^2}{\alpha}, \tag{40}$$

proving the bound $\tilde{\lambda} \leq \lambda + \frac{\sigma_x^2}{\alpha}$. Finally, if λ is large relative to S_i^2, then so is $\tilde{\lambda} \geq \lambda$. Then

$$\begin{aligned}
\lambda &= \tilde{\lambda}\left(1 - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda}+S_j^2}\right) \\
&= \tilde{\lambda}\left(1 - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda}} - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}O\left(\left(\frac{S_j^2}{\tilde{\lambda}}\right)^2\right)\right) \\
&= \tilde{\lambda}\left(1 - \frac{1}{\tilde{\lambda}}\frac{\sigma_x^2}{\alpha} - O\left(\frac{1}{\tilde{\lambda}^2}\right)\right) \\
&= \tilde{\lambda} - \frac{\sigma_x^2}{\alpha} - O\left(\frac{1}{\tilde{\lambda}}\right).
\end{aligned}$$

Since $\tilde{\lambda} > \lambda$, the error $O\left(\frac{1}{\tilde{\lambda}}\right)$ is upper bounded by $O\left(\frac{1}{\lambda}\right)$, proving the claim.

3.2 Derivatives

We show next that $\tilde{\lambda}$ is an increasing concave function of λ. The bounds given in the previous section show that $\tilde{\lambda}$ is always positive. Differentiating equation (36) with respect to λ gives

$$\begin{aligned}
1 &= \tilde{\lambda}'\left(1 - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda}+S_j^2}\right) + \tilde{\lambda}'\tilde{\lambda}\left(\frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\left(\tilde{\lambda}+S_j^2\right)^2}\right) \\
&= \tilde{\lambda}'\left[\frac{\lambda}{\tilde{\lambda}} + \tilde{\lambda}\left(\frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\left(\tilde{\lambda}+S_j^2\right)^2}\right)\right].
\end{aligned}$$

which shows that $\tilde{\lambda}' \geq 0$, so $\tilde{\lambda}$ is increasing. For the second derivative, recall $\zeta = \frac{1}{\alpha}\frac{1}{P}\sum_{i=1}^{P}(t_i - 1)^2$ and note

$$\begin{aligned}
\frac{d\zeta}{d\lambda} &= \frac{1}{\alpha}\frac{2}{P}\sum_{i=1}^{P}(t_i - 1)\frac{dt_i}{d\lambda} \\
&= \frac{1}{\alpha}\frac{2}{P}\sum_{i=1}^{P}(t_i - 1)\frac{d}{d\lambda}\left(1 - \frac{S_i^2}{\tilde{\lambda} + S_i^2}\right) \\
&= \tilde{\lambda}'\frac{1}{\alpha}\frac{2}{P}\sum_{i=1}^{P}(t_i - 1)\frac{S_i^2}{\left(\tilde{\lambda} + S_i^2\right)^2} \leq 0.
\end{aligned}$$

Next, from above we have $0 \leq \frac{d\tilde{\lambda}}{d\lambda} = \frac{1}{1-\zeta}$, so $0 \leq \zeta \leq 1$. Differentiating this equation gives

$$\begin{aligned}
\frac{d^2\tilde{\lambda}}{d\lambda^2} &= \frac{\zeta'}{(1-\zeta)^2} \\
&= \left(\frac{d\tilde{\lambda}}{d\lambda}\right)^2 \zeta' \leq 0,
\end{aligned}$$

so $\tilde{\lambda}$ is concave.

3.3 Intercept

From the bound $\lambda \leq \tilde{\lambda} \leq \frac{1}{1-\frac{1}{\alpha}}\lambda$ proved in 3.1 to hold when $\alpha > 1$, we have that the y-intercept of $\tilde{\lambda}$ is 0. For $\alpha < 1$, $\tilde{\lambda}(\lambda = 0)$ satisfies

$$0 = \tilde{\lambda}\left(1 - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda} + S_j^2}\right), \tag{41}$$

so that either $\tilde{\lambda} = 0$ or the term in parentheses is 0. Assume $\tilde{\lambda} = 0$. Differentiating the $\tilde{\lambda}$ equation and setting $\lambda, \tilde{\lambda} \to 0$ gives

$$\tilde{\lambda}'(0) = \frac{1}{1 - \frac{1}{\alpha}} < 0, \tag{42}$$

which cannot be since we proved in 3.2 that $\tilde{\lambda}' \geq 0$. So we must have that $\tilde{\lambda} \neq 0$ and the term in the parentheses is 0:

$$1 = \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda}(0) + S_j^2}. \tag{43}$$

Thus, when $\alpha < 1$, $\tilde{\lambda}$'s y-intercept is nonzero and satisfies (43).

3.4 Isotropic data

In the case of a single scale $\mathbf{\Sigma} = \mathbf{S}^2 = S\mathbf{I}_P$, (36) becomes

$$\lambda = \tilde{\lambda}\left(1 - \frac{1}{\alpha}\frac{S^2}{\tilde{\lambda} + S^2}\right). \tag{44}$$

Solving for $\tilde{\lambda}$ gives

$$\tilde{\lambda} = \frac{\lambda - S^2\left(1 - \frac{1}{\alpha}\right) + \sqrt{\left(\lambda - S^2\left(1 - \frac{1}{\alpha}\right)\right)^2 + 4\lambda S^2}}{2}. \tag{45}$$

When $\lambda = 0$, this simplifies to

$$\tilde{\lambda} = \begin{cases} 0 & \alpha > 1 \\ S^2\left(\frac{1}{\alpha} - 1\right) & \alpha < 1, \end{cases} \tag{46}$$

and for large λ, we obtain

$$\tilde{\lambda} = \lambda + \frac{S^2}{\alpha} + O\left(\frac{1}{\lambda}\right) = \lambda + \frac{\sigma_x^2}{\alpha} + O\left(\frac{1}{\lambda}\right), \tag{47}$$

consistent with the result given for general $\mathbf{\Sigma}$ given above.

3.5 Well separated scales

Let $\mathbf{\Sigma}$ be a D-level covariance with (descending) eigenvalues S_d^2 with multiplicities P_d with $\sum_d P_d = P$. Assume the scales are well separated, so that $\epsilon_k^2 = S_{k+1}^2/S_k^2 \ll 1$. In terms of the distinct scales S_d^2, the $\tilde{\lambda}$ equation is

$$\begin{aligned} \lambda &= \tilde{\lambda}\left(1 - \frac{1}{\alpha}\frac{1}{P}\sum_{d=1}^{D} P_d \frac{S_d^2}{\tilde{\lambda} + S_d^2}\right) && (48) \\ &= \tilde{\lambda}\left(1 - \frac{1}{\alpha}\sum_{d=1}^{D} \frac{f_d S_d^2}{\tilde{\lambda} + S_d^2}\right), && (49) \end{aligned}$$

where $f_d := \frac{P_d}{P}$.

We will obtain expressions for $\tilde{\lambda}(\lambda)$ when λ is between the k^{th} and $k+1^{st}$ scales S_k^2 and S_{k+1}^2, in the limit of small ϵ_k. To this end, set $\lambda = x\sqrt{S_k^2 S_{k+1}^2} = x S_k S_{k+1}$ where for small ϵ_k, we keep $x = O(1)$. We first show that we obtain a self consistent solution with $\tilde{\lambda} = y S_k S_{k+1}$ with $y = O(1)$. Making these substitutions, (49) becomes

$$\begin{aligned}
x &= y - \frac{1}{\alpha}\sum_{d=1}^{D}\frac{yf_dS_d^2}{yS_kS_{k+1}+S_d^2}\\
&= y - \frac{1}{\alpha}\sum_{d\le k}\frac{yf_dS_d^2}{yS_kS_{k+1}+S_d^2} - \frac{1}{\alpha}\sum_{d\ge k+1}\frac{yf_dS_d^2}{yS_kS_{k+1}+S_d^2}\\
&= y - \frac{1}{\alpha}\sum_{d\le k}\frac{yf_d}{y\frac{S_kS_{k+1}}{S_d^2}+1} - \frac{1}{\alpha}\sum_{d\ge k+1}\frac{yf_d\frac{S_d^2}{S_kS_{k+1}}}{y+\frac{S_d^2}{S_kS_{k+1}}}\\
&= y - \frac{1}{\alpha}\sum_{d\le k}\frac{yf_d}{y\epsilon_k\left(\prod_{n=d}^{k-1}\epsilon_n^2\right)+1} - \frac{1}{\alpha}\sum_{d\ge k+1}\frac{yf_d\epsilon_k\prod_{n=k+1}^{d-1}\epsilon_n^2}{y+\epsilon_k\prod_{n=k+1}^{d-1}\epsilon_n^2}.
\end{aligned}$$

Keeping terms of order $0,1$ in ϵ_n leaves

$$x = \left(1-\frac{1}{\alpha}\sum_{d\le k}f_d\right)y + \frac{1}{\alpha}f_k\epsilon_ky^2 - \frac{1}{\alpha}f_{k+1}\epsilon_k. \tag{50}$$

Finally, multiplying by S_kS_{k+1} and writing in terms of $\lambda, \tilde{\lambda}$ we obtain

$$\lambda = \left(1-\frac{1}{\alpha}\sum_{d\le k}f_d\right)\tilde{\lambda} + \frac{f_k}{\alpha S_k^2}\tilde{\lambda}^2 - \frac{1}{\alpha}f_{k+1}S_{k+1}^2, \tag{51}$$

which is a self-consistent first order approximation for $\tilde{\lambda}(\lambda)$ in the limit of widely separated scales as long as $\left(1-\frac{1}{\alpha}\sum_{d\le k}f_d\right)>0$.

3.6 Over- and under-sampled limits

Next we obtain approximate expressions for $\tilde{\lambda}, \frac{d\tilde{\lambda}}{d\lambda}, \frac{d}{d\tilde{\lambda}}\frac{d\tilde{\lambda}}{d\lambda}$ in the limit of large or small α. For convenience define $\beta := \frac{1}{\alpha}$.

3.6.1 Oversampled: $\alpha \to \infty$, $\beta \to 0$

For the section below studying the optimal regularization value (sec 5), it will be most useful to derive expressions in the joint limit that $\beta \to 0$ and $\lambda \to \lambda_0$ where λ_0 is an arbitrary value (including possibly 0). To this end, expand λ in powers of β as $\lambda = \lambda_0 + \beta\lambda_1 + O\left(\beta^2\right)$. Now assuming $\tilde{\lambda}$ has the expansion $\tilde{\lambda} = \tilde{\lambda}_0 + \beta\tilde{\lambda}_1 + O\left(\beta^2\right)$, plugging these expansions into the $\tilde{\lambda}$ equation (36) and equating powers up to β^1, we obtain

$$\begin{aligned}
\tilde{\lambda}_0 &= \lambda_0\\
\tilde{\lambda}_1 &= \lambda_1 + \lambda_0\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda}_0+S_j^2},
\end{aligned}$$

which implies $\tilde{\lambda} = \lambda + \beta \left(\frac{1}{P} \sum_{j=1}^{P} \frac{\lambda S_j^2}{\lambda + S_j^2} \right) + O\left(\beta^2\right)$. We have from (19) that

$$\begin{aligned} \frac{d\tilde{\lambda}}{d\lambda} &= \frac{1}{1 - \beta \frac{1}{P} \sum_{j=1}^{P} \left(\frac{S_i^2}{S_i^2 + \tilde{\lambda}} \right)^2} \\ &= 1 + \beta \frac{1}{P} \sum_{j=1}^{P} \left(\frac{S_i^2}{S_i^2 + \lambda} \right)^2 + O\left(\beta^2\right). \end{aligned}$$

Finally, differentiating with respect to $\tilde{\lambda}$ gives

$$\begin{aligned} \frac{d}{d\tilde{\lambda}} \frac{d\tilde{\lambda}}{d\lambda} &= -\frac{2\beta \frac{1}{P} \sum_{j=1}^{P} \frac{S_i^4}{\left(S_i^2 + \tilde{\lambda}\right)^3}}{\left(1 - \beta \frac{1}{P} \sum_{j=1}^{P} \left(\frac{S_i^2}{S_i^2 + \tilde{\lambda}} \right)^2 \right)^2} \\ &= \left(-2 \frac{1}{P} \sum_{j=1}^{P} \frac{S_i^4}{\left(S_i^2 + \lambda\right)^3} \right) \beta + O\left(\beta^2\right). \end{aligned}$$

These are all the formulas we'll need. Summarizing,

$$\tilde{\lambda} = \lambda + \beta \left(\frac{1}{P} \sum_{j=1}^{P} \frac{\lambda S_j^2}{\lambda + S_j^2} \right) + O\left(\beta^2\right) \tag{52}$$

$$\frac{d\tilde{\lambda}}{d\lambda} = 1 + \beta \left(\frac{1}{P} \sum_{j=1}^{P} \left(\frac{S_i^2}{S_i^2 + \lambda} \right)^2 \right) + O\left(\beta^2\right) \tag{53}$$

$$\frac{d}{d\tilde{\lambda}} \frac{d\tilde{\lambda}}{d\lambda} = \beta \left(-2 \frac{1}{P} \sum_{j=1}^{P} \frac{S_i^4}{\left(S_i^2 + \lambda\right)^3} \right) + O\left(\beta^2\right). \tag{54}$$

3.6.2 Undersampled: $\alpha \to 0$, $\beta \to \infty$

Recall the equation for $\tilde{\lambda}$ is

$$\lambda = \tilde{\lambda} \left(1 - \beta \frac{1}{P} \sum_{j=1}^{P} \frac{S_j^2}{\tilde{\lambda} + S_j^2} \right). \tag{55}$$

As β grows, the term in parentheses will become negative unless $\tilde{\lambda}$ grows as well. Thus for any fixed λ, we must have $\tilde{\lambda} \to \infty$, so we expect that $\tilde{\lambda}$ will have (at least) a β^1 term in its expansion.

As before, we assume λ has an expansion in powers of β. For section 5 below, we'll be interested in the case that $\lambda = O(\beta)$ so we expand as follows: $\lambda = \beta\lambda_1 + O(1)$. Assuming $\tilde{\lambda} = \beta\tilde{\lambda}_1 + O(1)$, inserting these expressions into

(55), we obtain $\tilde{\lambda}_1 = \lambda_1 + \sigma_x^2$, so $\tilde{\lambda} = \lambda + \beta\sigma_x^2 + O(1)$. Again using (19), we find

$$\begin{aligned}\frac{d\tilde{\lambda}}{d\lambda} &= \frac{1}{1 - \beta\frac{1}{P}\sum_{j=1}^{P}\left(\frac{S_i^2}{S_i^2+\tilde{\lambda}}\right)^2} \\ &= 1 + \beta\frac{1}{P}\sum_{j=1}^{P}\left(\frac{S_i^2}{S_i^2+\tilde{\lambda}}\right)^2 + O\left(\beta^{-2}\right) \\ &= 1 + O\left(\beta^{-1}\right).\end{aligned}$$

Finally, differentiating with respect to $\tilde{\lambda}$ gives

$$\begin{aligned}\frac{d}{d\tilde{\lambda}}\frac{d\tilde{\lambda}}{d\lambda} &= -\frac{2\beta\frac{1}{P}\sum_{j=1}^{P}\frac{S_i^4}{\left(S_i^2+\tilde{\lambda}\right)^3}}{\left(1 - \beta\frac{1}{P}\sum_{j=1}^{P}\left(\frac{S_i^2}{S_i^2+\tilde{\lambda}}\right)^2\right)^2} \\ &= -2\beta\frac{1}{P}\sum_{j=1}^{P}\frac{S_i^4}{\left(S_i^2+\tilde{\lambda}\right)^3} + O\left(\beta^{-3}\right) \\ &= -2\beta\frac{1}{\tilde{\lambda}^3}\gamma_x + O\left(\beta^{-3}\right),\end{aligned}$$

where $\gamma_x := \frac{1}{P}\mathrm{Tr}[\mathbf{\Sigma}^2]$, and $\tilde{\lambda} = O(\beta)$, so the first term is $O\left(\beta^{-2}\right)$.

Summarizing,

$$\tilde{\lambda} = \lambda + \beta\sigma_x^2 + O(1) \tag{56}$$

$$\frac{d\tilde{\lambda}}{d\lambda} = 1 + O\left(\beta^{-1}\right) \tag{57}$$

$$\frac{d}{d\tilde{\lambda}}\frac{d\tilde{\lambda}}{d\lambda} = -2\beta\frac{1}{\tilde{\lambda}^3}\gamma_x + O\left(\beta^{-3}\right). \tag{58}$$

3.7 Relationship to α

Taking the t_i, t_N equations and writing them in terms of $\beta := \frac{1}{\alpha}$, we obtain

$$\begin{aligned}t_i &= 1 + zS_i^2 t_N t_i \\ t_N &= 1 + \beta\sum_i \beta_i\left(t_i - 1\right).\end{aligned}$$

Differentiating with respect to β gives

$$\begin{aligned}t_i' &= t_N'\frac{1}{t_N}\left(t_i - 1\right)t_i \\ t_N' &= \frac{1}{\beta}\frac{1}{N}\sum_i\left(t_i - 1\right) + \frac{1}{N}\sum_i t_i'.\end{aligned}$$

Solving these, we obtain

$$\frac{dt_N}{d\beta} = \frac{1}{\beta} t_N \frac{t_N - 1}{1 - \zeta}. \tag{59}$$

Using the fact that $t_N = \frac{\lambda}{\tilde{\lambda}}$, we have $\frac{dt_N}{d\beta} = -\frac{\lambda}{\tilde{\lambda}^2}\frac{d\tilde{\lambda}}{d\beta}$, so

$$\frac{d\tilde{\lambda}}{d\beta} = \frac{1}{\beta}\frac{\tilde{\lambda} - \lambda}{1 - \zeta} = \frac{1}{\beta}\left(\tilde{\lambda} - \lambda\right)\frac{d\tilde{\lambda}}{d\lambda} \tag{60}$$

$$\implies \frac{d\tilde{\lambda}}{d\alpha} = -\frac{1}{\alpha}\left(\tilde{\lambda} - \lambda\right)\frac{d\tilde{\lambda}}{d\lambda} \leq 0, \tag{61}$$

which shows that $\tilde{\lambda}$ is a decreasing function of α - consistent with the intuition that as the sampling ratio goes up, the correction, and consequently the increment between $\tilde{\lambda}$ and fixed λ, should go down.

4 Analysis of ρ_f

ρ_f is the fractional participation ratio defined as

$$\rho_f := \frac{\left(\frac{1}{N}\mathrm{Tr}\tilde{\mathbf{B}}\right)^2}{\frac{1}{N}\mathrm{Tr}\tilde{\mathbf{B}}^2} = \frac{\left(\frac{1}{N}\mathrm{Tr}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_N\right)^{-1}\right]\right)^2}{\frac{1}{N}\mathrm{Tr}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_N\right)^{-2}\right]}. \tag{62}$$

Define γ_i to be the eigenvalues of $\tilde{\mathbf{B}}$ (as argued below in 6, these only differ from the eigenvalues of $\mathbf{B} = \left(\frac{1}{N}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_P\right)^{-1}$ in their number of eigenvalues equal to λ). We can then write ρ_f as

$$\rho_f = \frac{\left(\frac{1}{N}\sum_{i=1}^N \gamma_i\right)^2}{\left(\frac{1}{N}\sum_{i=1}^N \gamma_i^2\right)} = \left(\frac{\mathbb{E}_{\gamma_i}\left[\gamma^2\right]}{\mathbb{E}_{\gamma_i}\left[\gamma\right]^2}\right)^{-1} = \left(\left(\frac{\sigma_\gamma}{\mu_\gamma}\right)^2 + 1\right)^{-1}, \tag{63}$$

where $\mu_\gamma, \sigma_\gamma$ are the mean and standard deviation of the eigenvalues γ. From this it's clear that $\rho_f \leq 1$. Since $\gamma_i > 0$, we also have from the first equality that

$$\rho_f = \frac{1}{N}\left(\frac{\sum_{i=1}^N \gamma_i}{\sqrt{\sum_{i=1}^N \gamma_i^2}}\right)^2 = \frac{1}{N}\left(\frac{\|\gamma\|_1}{\|\gamma\|_2}\right)^2 \geq \frac{1}{N}, \tag{64}$$

where $\|\gamma\|_p$ is the p-norm of the vector of eigenvalues γ_i. So $\frac{1}{N} \leq \rho_f \leq 1$.

In section 3.7 of the main text we compare empirical and theoretical values of ρ_f. Empirical values were obtained by evaluating (62) for large gaussian matrices. To obtain analytical values, we use (21) and (19), giving

$$\begin{aligned}\rho_f &= \left(\frac{d\tilde{\lambda}}{d\lambda}\right)^{-1} = 1 - \zeta \\ &= 1 - \frac{1}{\alpha}\frac{1}{P}\sum_{i=1}^{P}(t_i - 1)^2 \\ &= 1 - \frac{1}{\alpha}\frac{1}{P}\sum_{i=1}^{P}\left(\frac{S_i^2}{\tilde{\lambda} + S_i^2}\right)^2 ,\end{aligned}$$

so that ρ_f can be obtained by solving (20) for $\tilde{\lambda}$ and substituting.

Sketch of ρ_f behavior in multiple descent Multiple descent is most pronounced when λ is small, so assume $\lambda \to 0$. We have from the previous paragraph that

$$\rho_f = 1 - \frac{1}{\alpha}\frac{1}{P}\sum_{i=1}^{P}\left(\frac{S_i^2}{\tilde{\lambda} + S_i^2}\right)^2 . \tag{65}$$

We've also shown in 3.3 that when $\alpha < 1$, at $\lambda = 0$, we have

$$1 = \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda} + S_j^2}. \tag{66}$$

Dividing, we find

$$\rho_f = 1 - \frac{\frac{1}{P}\sum_{i=1}^{P}\left(\frac{S_i^2}{\tilde{\lambda}+S_i^2}\right)^2}{\frac{1}{P}\sum_{j=1}^{P}\left(\frac{S_j^2}{\tilde{\lambda}+S_j^2}\right)}. \tag{67}$$

Because $0 \leq \frac{S_j^2}{\tilde{\lambda}+S_j^2} \leq 1$, the numerator is always less than the denominator, showing $\rho_f \geq 0$. The fractional participation ratio ρ_f will become small, and error will grow large, whenever the numerator is approximately equal to the denominator.

If the scales S_i^2 are widely seperated, ie. $\epsilon_i^2 = S_{i+1}^2/S_i^2 \ll 1$, this will happen whenever $\tilde{\lambda} \approx \epsilon_k S_k^2$ since then

$$\frac{S_j^2}{\tilde{\lambda} + S_j^2} = \frac{S_j^2}{\epsilon_k S_k^2 + S_j^2} \approx \begin{cases} 1 & j \leq k \\ 0 & j > k \end{cases}, \tag{68}$$

and so $\rho_f \approx 0$. From (66), this value of $\tilde{\lambda}$ will happen approximately when

$$\begin{aligned}1 &= \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{\tilde{\lambda}+S_j^2} = \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{k} 1 \\ &= \frac{1}{\alpha}\frac{\sum_{j=1}^{k} P_j}{P} = \frac{1}{\alpha}\sum_{j=1}^{k} f_j,\end{aligned}$$

or $\alpha = \sum_{j=1}^{k} f_j$, that is, whenever the number of parameters is equal to the number of featurers at the top k scales, which corresponds exactly with the critical α values leading to phase transitions discussed in 6.

5 Optimal λ

We now derive the following leading order (in α) formulas for the optimal regularization parameter in the over- and under-sampled regimes

$$\lambda^* \to \begin{cases} \frac{1}{\alpha}\sigma^2 \frac{\frac{1}{P}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right]}{\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w}} & \alpha \gg 1 \\ \frac{1}{\alpha}\sigma^2\left(1+SNR\right)\frac{\frac{1}{P}\mathrm{Tr}\left[\boldsymbol{\Sigma}^2\right]}{\mathbf{w}^T\boldsymbol{\Sigma}^2\mathbf{w}} - \frac{\sigma_x^2}{\alpha} & \alpha \ll 1. \end{cases} \tag{69}$$

Recall from (5) the fraction unexplained variance is

$$\mathcal{F} = f_n + \tilde{\lambda}'\sum_{i=1}^{P}\left\{ f_s\hat{\mathbf{v}}_i^2\left(\frac{\tilde{\lambda}}{S_i^2+\tilde{\lambda}}\right)^2 + f_n\beta\frac{1}{P}\left(\frac{S_i^2}{S_i^2+\tilde{\lambda}}\right)^2\right\}, \tag{70}$$

where $\beta := \frac{1}{\alpha}$. As argued previously, both terms in the sum have a well defined limit as $P \to \infty$. We will consider these terms now in the limit that α is either very large or small. To determine the optimal λ we will first solve for the optimal $\tilde{\lambda}$. To this end, we differentiate (70) with respect to $\tilde{\lambda}$ and set the result to 0 when $\tilde{\lambda}$ is equal to its optimal value $\tilde{\lambda}^*$, giving

$$0 = \left(\frac{d\mathcal{F}}{d\tilde{\lambda}}\right)^* = \left(\frac{d\left(\tilde{\lambda}'\right)}{d\tilde{\lambda}}\right)^*\sum_{i=1}^{P}\left\{ f_s\hat{\mathbf{v}}_i^2\left(\frac{\tilde{\lambda}^*}{S_i^2+\tilde{\lambda}^*}\right)^2 + f_n\beta\frac{1}{P}\left(\frac{S_i^2}{S_i^2+\tilde{\lambda}^*}\right)^2\right\} \tag{71}$$

$$+\,2\left(\tilde{\lambda}'\right)^*\sum_{i=1}^{P}\left\{ f_s\hat{\mathbf{v}}_i^2\frac{\tilde{\lambda}^* S_i^2}{\left(S_i^2+\tilde{\lambda}^*\right)^3} - f_n\beta\frac{1}{P}\frac{S_i^4}{\left(S_i^2+\tilde{\lambda}^*\right)^3}\right\}. \tag{72}$$

We now handle each of the two limits in turn.

Oversampled: $\alpha \to \infty$, $\beta \to 0$ In the oversampled limit, $\alpha \to \infty$ or equivalently, $\beta \to 0$. We will show there is a self-consistent solution to the optimality criterion (72) with $\tilde{\lambda}^* = O\left(\alpha^{-1}\right) = O\left(\beta\right)$.

Above we derived the following formulas (eq. (52)) for the oversampled regime.

$$\tilde{\lambda} = \lambda + \beta \left(\frac{1}{P} \sum_{j=1}^{P} \frac{\lambda S_j^2}{\lambda + S_j^2} \right) + O\left(\beta^2\right) \tag{73}$$

$$\frac{d\tilde{\lambda}}{d\lambda} = 1 + \beta \left(\frac{1}{P} \sum_{j=1}^{P} \left(\frac{S_i^2}{S_i^2 + \lambda} \right)^2 \right) + O\left(\beta^2\right) \tag{74}$$

$$\frac{d}{d\tilde{\lambda}} \frac{d\tilde{\lambda}}{d\lambda} = \beta \left(-2 \frac{1}{P} \sum_{j=1}^{P} \frac{S_i^4}{\left(S_i^2 + \lambda\right)^3} \right) + O\left(\beta^2\right), \tag{75}$$

which implies that $\frac{d\left(\tilde{\lambda}'\right)}{d\tilde{\lambda}} = \frac{d}{d\tilde{\lambda}} \frac{d\tilde{\lambda}}{d\lambda} = O\left(\beta\right)$ for all λ. Based on the assumption that $\tilde{\lambda}^* = O\left(\beta\right)$, the first sum in (72) is $O\left(\beta^2\right)$; the second sum, on the other hand, will have terms of order $O\left(\beta\right)$. Using the expressions immediately above, and keeping terms of order $O\left(\beta\right)$, we obtain

$$\tilde{\lambda}^* = \beta \frac{f_n \frac{1}{P} \sum_{i=1}^{P} \frac{1}{S_i^2}}{f_s \sum_{i=1}^{P} \hat{\mathbf{v}}_i^2 \frac{1}{S_i^4}} = \beta \sigma^2 \frac{\frac{1}{P} \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right]}{\mathbf{w}^T \boldsymbol{\Sigma}^{-1} \mathbf{w}}. \tag{76}$$

Finally, $\tilde{\lambda} = \lambda + \beta \left(\frac{1}{P} \sum_{j=1}^{P} \frac{\lambda S_j^2}{\lambda + S_j^2} \right) + O\left(\beta^2\right)$ implies

$$\lambda^* = \beta \sigma^2 \frac{\frac{1}{P} \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right]}{\mathbf{w}^T \boldsymbol{\Sigma}^{-1} \mathbf{w}} + O\left(\beta^2\right). \tag{77}$$

Undersampled: $\alpha \to 0$, $\beta \to \infty$ Here we show there is a self-consistent solution with $\tilde{\lambda}^* = O\left(\beta\right)$. In this limit, the expressions derived above (eq. (56)) are

$$\tilde{\lambda} = \lambda + \beta \sigma_x^2 + O\left(1\right) \tag{78}$$

$$\frac{d\tilde{\lambda}}{d\lambda} = 1 + O\left(\beta^{-1}\right) \tag{79}$$

$$\frac{d}{d\tilde{\lambda}} \frac{d\tilde{\lambda}}{d\lambda} = -2\beta \frac{1}{\tilde{\lambda}^3} \gamma_x + O\left(\beta^{-3}\right). \tag{80}$$

By inspection both sums of (72) will vanish at order $O\left(\beta^{-2}\right)$, so keeping terms up to this order gives

$$\tilde{\lambda}^* = \beta \gamma_x \frac{1}{f_s \sum_{i=1}^{P} \hat{\mathbf{v}}_i^2 S_i^2} = \beta \gamma_x \frac{\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} + \sigma^2}{\mathbf{w}^T \boldsymbol{\Sigma}^2 \mathbf{w}} = \beta \sigma^2 \left(SNR + 1\right) \frac{\frac{1}{P} \mathrm{Tr}\left[\boldsymbol{\Sigma}^2\right]}{\mathbf{w}^T \boldsymbol{\Sigma}^2 \mathbf{w}}, \tag{81}$$

giving the optimal value of $\tilde{\lambda}$. Using (78), we determine the optimal $\tilde{\lambda}$ to be

$$\lambda^* = \beta\sigma^2 \left(SNR + 1\right) \frac{\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^2\right]}{\mathbf{w}^T\mathbf{\Sigma}^2\mathbf{w}} - \beta\sigma_x^2 + O\left(1\right). \tag{82}$$

5.0.1 Effect of alignment on optimal regularization

From the formulas

$$\lambda^* \to \begin{cases} \frac{1}{\alpha}\sigma^2 \frac{\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^{-1}\right]}{\mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w}} & \alpha \gg 1 \\ \frac{1}{\alpha}\sigma^2\left(1 + SNR\right)\frac{\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^2\right]}{\mathbf{w}^T\mathbf{\Sigma}^2\mathbf{w}} - \frac{\sigma_x^2}{\alpha} & \alpha \ll 1, \end{cases} \tag{83}$$

the optimal λ depends on the weights $\mathbf{w}$ through $\mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w}$ in the oversampled regime and $\mathbf{w}^T\mathbf{\Sigma}^2\mathbf{w}$ in the undersampled regime. The eigenvalues of the two matrices $\mathbf{\Sigma}^{-1}, \mathbf{\Sigma}^2$ are in exactly complementary order, so changes which align $\mathbf{w}$ to large eigenvalues of $\mathbf{\Sigma}^{-1}$ will align $\mathbf{w}$ to small eigenvalues of $\mathbf{\Sigma}^2$ and vice versa. Thus changes to $\mathbf{w}$ will tend to have opposite effects on the optimal regularization in the under- and over-sampled limits.

5.0.2 Random w

For the random-weights model with signal to noise ratio SNR the weights are sampled from a multivariate gaussian with covariance $\mathbf{C} = \sigma^2\frac{SNR}{P}\mathbf{\Sigma}^{-1}$. Thus the weights $\mathbf{w}$ are generated by $\sqrt{\sigma^2\frac{SNR}{P}}\mathbf{\Sigma}^{-1/2}\mathbf{z}$ where $\mathbf{z}$ is a standard normal vector. This implies

$$\mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w} = \sigma^2 SNR\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^{-2}\right] \tag{84}$$

$$\mathbf{w}^T\mathbf{\Sigma}^{2}\mathbf{w} = \sigma^2 SNR\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}\right] = \sigma^2 SNR\sigma_x^2, \tag{85}$$

so that the formulas for optimal λ simplify to

$$\lambda^* \to \begin{cases} \frac{1}{\alpha}\frac{1}{SNR}\frac{\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^{-1}\right]}{\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^{-2}\right]} & \alpha \gg 1 \\ \frac{1}{\alpha}\left(1 + \frac{1}{SNR}\right)\frac{\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^2\right]}{\sigma_x^2} - \frac{\sigma_x^2}{\alpha} & \alpha \ll 1. \end{cases} \tag{86}$$

2-scale model Let $\mathbf{\Sigma}$ be a 2-scale covariance with eigenvalues $S_1^2 = \gamma$ and $S_2^2 = \gamma^{-1}$ so the aspect ratio is $S_1/S_2 = \gamma$. For simplicity take the multiplicities to be $P_1 = P_2 = \frac{1}{2}P$. Then $\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^n\right] = \frac{1}{2}\left(\gamma^n + \gamma^{-n}\right)$. In particular, when γ is large, we can write

$$\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^n\right] \approx \begin{cases} \frac{1}{2}\gamma^{-n} = \frac{1}{2}S_{min}^{n/2} & n < 0 \\ 1 & n = 0 \\ \frac{1}{2}\gamma^{n} = \frac{1}{2}S_{max}^{n/2} & n > 0. \end{cases} \tag{87}$$

Analogous formulas hold for very small γ. Thus the formulas for optimal λ in (86) simplify to

$$\lambda^* \to \begin{cases} \frac{1}{\alpha}\frac{1}{SNR}S_{min}^2 & \alpha \gg 1 \\ \frac{1}{\alpha}\left(\frac{1}{2}+\frac{1}{SNR}\right)S_{max}^2 & \alpha \ll 1. \end{cases} \tag{88}$$

6 The spectrum of the Hessian and multiple descent

The training cost for estimated weights $\hat{\mathbf{w}}$ is defined to be $\frac{1}{N}\sum_{i=1}^N \left(y^\mu - \mathbf{x}^\mu \cdot \hat{\mathbf{w}}\right)^2 + \lambda \left|\hat{\mathbf{w}}\right|^2$, which in matrix notation is

$$\frac{1}{N}\left\|\mathbf{y}-\mathbf{X}\mathbf{w}\right\|^2 + \lambda \hat{\mathbf{w}}^T\hat{\mathbf{w}}. \tag{89}$$

The Hessian of this cost function is therefore $\mathbf{H} = \frac{1}{N}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_P$, and the inverse Hessian is $\mathbf{B} := \left(\frac{1}{N}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_P\right)^{-1}$. In the main text we point out that the many of the phenomena of interest can be understood in terms of the spectrum of $\mathbf{B}$, or equivalently in terms of the spectrum of the related matrix $\tilde{\mathbf{B}} := \left(\frac{1}{N}\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_P\right)^{-1}$, which is identical except for the number of eigenvalues equal to λ, corresponding to the zero eigenvalues of $\mathbf{X}^T\mathbf{X}, \mathbf{X}\mathbf{X}^T$. To see this, note that for any nonzero eigenvalue, eigenvector of $\mathbf{X}^T\mathbf{X}$ denoted by μ, v, we have $\mathbf{X}\mathbf{X}^T(\mathbf{X}v) = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}v\right) = \mu(\mathbf{X}v)$ so that $\mathbf{X}v$ is an eigenvector of $\mathbf{X}\mathbf{X}^T$ with the same nonzero eigenvalue μ. Thus, the nonzero spectra are in one-to-one correspondence. The remaining 0 eigenvalues give rise to eigenvalues of λ in $\mathbf{B}, \tilde{\mathbf{B}}$.

The Stieltjes transform of the spectral density $\rho(x)$ of the matrix $\frac{1}{N}\mathbf{X}\mathbf{X}^T$, is defined as $G(x) := \int \frac{\rho(t)}{x-t}dt$ ([1] p198). Thus

$$\begin{aligned} \frac{1}{N}\mathrm{Tr}\tilde{\mathbf{B}} &= \frac{1}{N}\mathrm{Tr}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_N\right)^{-1}\right] \\ &= \int \frac{\rho(t)}{t+\lambda}dt = -G(-\lambda). \end{aligned}$$

Recalling that $\tilde{\lambda}$ was originally defined as $\tilde{\lambda} := \frac{1}{\frac{1}{N}\mathrm{Tr}\tilde{\mathbf{B}}}$, we have $\tilde{\lambda}(\lambda) = -\frac{1}{G(-\lambda)}$. Substituting this into (20), we obtain the following equation for the Stieltjes transform:

$$\lambda = \frac{1}{G} - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{S_j^2}{S_j^2 G - 1}. \tag{90}$$

Although it is difficult to obtain the spectrum from (90) for general $\alpha, \mathbf{S}, \lambda$, it is possible to obtain an exact equation for the boundaries of the support of the spectrum by 1) clearing denominators to obtain a polynomial equation in G, 2) computing the discriminant of this polynomial, which is itself a polynomial in λ, and 3) setting this polynomial to 0. Boundary points of the spectrum must be λ-roots of this polynomial.

6.1 Widely separated scales

We can also use (90) to obtain approximate expressions for the spectral density when the scales are very different from one another: Let $\mathbf{\Sigma}$ be a D-scale covariance with eigenvalues S_d^2 for $d = 1, \ldots, D$ with multiplicities P_d such that $\sum_d P_d = P$. Define $f_d := \frac{P_d}{P}$. Assume the scales are arranged in descending order and are very different from one another, so that $\epsilon_d^2 = S_{d+1}^2 / S_d \ll 1$. In the limit of small ϵ_d^2, the spectral density ρ consists of D disjoint components ρ_d, roughly centered on the D distinct scales S_d^2, satisfying

$$\rho_d\left(x\right) = \frac{\sqrt{\left(x_+ - x\right)\left(x - x_-\right)}}{2\pi S_d^2 \lambda} \tag{91}$$

$$x_\pm = S_d^2 \left(1 - \frac{1}{\alpha}\sum_{d'<d} f_{d'}\right)\left(1 \pm \sqrt{\frac{f_d}{\alpha - \sum_{d'<d} f_{d'}}}\right)^2 . \tag{92}$$

To show this, we will first derive a version of (90) which is valid to first order in the ϵ_d^2. The density formulas above will then come from further specializing these equations to 0^{th} order.

The approximation for small ϵ_d^2 is based on the intuition that for widely separated scales, G should have D more or less distinct regimes when λ is around each of the D scales. When λ is around the k^{th} scale, we have $\lambda = \lambda_k S_k^2$ where $\lambda_k = O\left(1\right)$. We will show there is a self-consistent solution in this regime with $G = G_k \frac{1}{S_k^2}$ where $G_k = O\left(1\right)$. Plugging these two expressions into (90) and keeping terms up to order 1 in ϵ_k^2 gives

$$\begin{aligned}
\lambda_k &= \frac{1}{G_k} - \frac{1}{\alpha}\sum_{d=1}^{D} f_d \frac{S_d^2}{S_d^2 G_k - S_k^2} \\
&= \frac{1}{G_k} - \frac{1}{\alpha}\sum_{d<k} f_d \frac{1}{G_k - \frac{S_k^2}{S_d^2}} - \frac{1}{\alpha} f_k \frac{1}{G_k - 1} - \frac{1}{\alpha}\sum_{d>k} f_d \frac{S_d^2/S_k^2}{\frac{S_d^2}{S_k^2} G_k - 1} \\
&= \frac{1}{G_k} - \frac{1}{\alpha}\sum_{d<k} f_d \frac{1}{G_k - \prod_{n=d}^{k-1}\epsilon_n^2} - \frac{1}{\alpha} f_k \frac{1}{G_k - 1} - \frac{1}{\alpha}\sum_{d>k} f_d \frac{\prod_{n=k}^{d-1}\epsilon_n^2}{G_k \prod_{n=k}^{d-1}\epsilon_n^2 - 1} \\
&= \frac{1}{G_k} - \frac{1}{\alpha}\frac{1}{G_k}\sum_{d<k} f_d - \frac{1}{\alpha}\frac{1}{G_k^2} f_{k-1}\epsilon_{k-1}^2 - \frac{1}{\alpha} f_k \frac{1}{G_k - 1} + \frac{1}{\alpha} f_{k+1}\epsilon_k^2,
\end{aligned}$$

giving us the first order equation

$$\lambda_k = \frac{\left(1 - \frac{1}{\alpha}\sum_{d<k} f_d\right)}{G_k} - \frac{\frac{1}{\alpha} f_k}{G_k - 1} - \frac{\frac{1}{\alpha} f_{k-1}}{G_k^2}\epsilon_{k-1}^2 + \frac{1}{\alpha} f_{k+1}\epsilon_k^2, \tag{93}$$

which we will keep for reference. Specializing now to 0^{th} order in ϵ_k^2 and now writing in terms of the original λ, G, we obtain

$$\frac{\lambda}{S_k^2} = \frac{\left(1 - \frac{1}{\alpha}\sum_{d<k} f_d\right)}{S_k^2 G} - \frac{\frac{1}{\alpha} f_k}{S_k^2 G - 1}, \tag{94}$$

where we stipulate that this equation is approximately valid when $\lambda = O\left(S_k^2\right)$. Using the formulas derived in 6.1.1, we obtain the density

$$\rho_d\left(x\right) = \frac{\sqrt{\left(x_+ - x\right)\left(x - x_-\right)}}{2\pi S_d^2 \lambda}$$
$$x_\pm = S_d^2 \left(1 - \frac{1}{\alpha}\sum_{d'<d} f_{d'}\right)\left(1 \pm \sqrt{\frac{f_d}{\alpha - \sum_{d'<d} f_{d'}}}\right)^2,$$

completing the proof of the formulas above.

Some remarks about this density are in order. First, one can see from the expression for $x_\pm$ that the d^{th} component of the density only appears for values of α satisfying $\alpha > \sum_{d'<d} f_{d'}$. In other words, there is a critical measurement density at which the d^{th} scale in the data first becomes detectable. At this α the support of the d^{th} component is the single point $x = S_d^2 f_d$. As $\alpha \to \infty$ all scales become visible and the components become perfectly concentrated around the values S_d^2. Furthermore, the d^{th} scale has another critical α when its support reaches 0 at the single point $\alpha = \sum_{d' \leq d} f_{d'} = \sum_{d'<d+1} f_{d'}$, which is exactly the α where the $d + 1^{st}$ scale first appears. Thus as α is increased, the spectrum undergoes a sequence of phase transitions where the d^{th} component acquires an extended tail and the $d + 1^{st}$ component appears around $S_{d+1}^2 f_{d+1}$.

Comparing these formulas to (96), we can also see that, whenever it exists, the d^{th} component has total mass

$$\int \rho_d\left(x\right)dx = \begin{cases} 1 - \frac{1}{\alpha}\sum_{d'<d} f_{d'} & \sum_{d'<d} f_{d'} < \alpha < \sum_{d' \leq d} f_{d'} \\ \frac{1}{\alpha} f_d & \sum_{d' \leq d} f_{d'} < \alpha. \end{cases} \tag{95}$$

These masses sum to 1 when $\alpha < 1$, and we also have that when the d^{th} scale first appears it starts out with 0 mass, which gradually increases to a limiting value of $\frac{1}{\alpha} f_d$ for large α. In this limit the D scales have relative mass f_d, corresponding to their multiplicities, and the nonzero scales have total mass $\sum_d \frac{1}{\alpha} f_d = \frac{1}{\alpha}$, which is consistent with the fact that for $\alpha > 1$, $\mathbf{X}\mathbf{X}^T$ has $P < N$ nonzero eigenvalues out of N total, so the fraction of nonzero eigenvalues is $\frac{P}{N} = \frac{1}{\alpha}$.

6.1.1 Formulas for Marchenko-Pastur-type distributions

First we have the following directly from the pdf of the Marchenko-Pastur distribution: if $r_\pm = \sigma^2 \left(1 \pm \sqrt{r}\right)^2$, then

$$\int_{r_-}^{r_+} \frac{1}{2\pi\sigma^2} \frac{\sqrt{\left(r_+ - x\right)\left(x - r_-\right)}}{rx} dx = \begin{cases} 1 & r < 1 \\ 1/r & r > 1. \end{cases} \tag{96}$$

Second, for a density with Stieltjes transform satisfying

$$\frac{\lambda}{S_k^2} = \frac{a}{S_k^2 G} - \frac{b}{S_k^2 G - 1}, \tag{97}$$

we can solve for G in terms of λ to obtain

$$G = \frac{(a-b)\,S_k^2 + \lambda \pm \sqrt{((b-a)\,S_k^2 - \lambda)^{\,2} - 4a\lambda S_k^2}}{2\lambda S_k^2}. \tag{98}$$

From the Stieltjes inversion formula, $\rho(x) = -\frac{1}{\pi}\lim_{\epsilon\to 0^+} \operatorname{Im} G(x+i\epsilon)$ ([1] p31), we see that the spectrum is nonzero only when the Stieltjes transform has a nonzero imaginary part at $x+i\epsilon$ for $\epsilon \to 0^+$. This can only happen when the term in the square root is negative, in which case the imaginary part of G comes exclusively from the square root, and we have

$$\rho\,(x) = \frac{\sqrt{4axS_k^2 - ((b-a)\,S_k^2 - x)^{\,2}}}{2\pi S_k^2 x}. \tag{99}$$

The radicand is a quadratic polynomial with leading coefficient -1 and roots $x_\pm = S_k^2\left(a + b \pm 2\sqrt{ab}\right)$, so this can be written

$$\begin{aligned} \rho\,(x) &= \frac{\sqrt{(x_+ - x)\,(x - x_-)}}{2\pi S_k^2 x} \\ x_\pm &= S_k^2\left(\sqrt{a} \pm \sqrt{b}\right)^2. \end{aligned}$$

7 Rules for manipulating generating functions

This section will serve as a reference for the portions of the diagrammatic proof in 1.2 utilizing generating functions. Here we will introduce a framework for quickly and compactly producing generating function equations for various combinatorial problems, developing what is needed from scratch so as to make the exposition self-contained. We will follow closely along the lines of the exposition in [2].

7.1 Combinatorial classes

A combinatorial class $\mathcal{A}$ is a set of objects a, each of which has a size $|a|$ (in our case, a nonnegative integer) and a weight w_a (any real number). The counting sequence $\{\mathcal{A}_n\}$ is the number of objects in $\mathcal{A}$ having size n. The generating function of a class $\mathcal{A}$ is defined as

$$A\,(z) = \sum_{a\in\mathcal{A}} w_a z^{|a|} = \sum_n z^n \left(\sum_{\substack{a\in\mathcal{A}\\|a|=n}} w_a\right),$$

which is just a weighted generalization of the usual generating function for the sequence $\mathcal{A}_n$ defined as $\sum_n z^n \mathcal{A}_n = \sum_n z^n \left(\sum_{\substack{a\in A\\|a|=n}} 1\right)$. Throughout, we'll write the combinatorial class in script letters, as in $\mathcal{A}$, and its generating function in straight font, as in A.

Given a set of objects, the usual approach to obtaining a generating function involves applying a recursion relation satisfied by the counting sequence to the generating power series. Here we'll use a more direct approach that, by building up a set of basic composition rules, will allow us to avoid error-prone algebraic manipulation of power series later on.

Often, we can describe a combinatorial class as a combination of simpler classes - a kind of combinatorial "recipe". The strategy will be to derive a few rules which allow us to translate combinatorial recipes into generating function relations very quickly and easily.

7.2 Rules

First define two basic classes: **1** will be the class with one object of size 0 with weight 1; and $\mathcal{Z}$ will be the class with one object of size 1 with weight 1. These symbols are chosen since the generating function of **1** is 1 and the generating function of $\mathcal{Z}$ is z.

1 Scalar multiplication *Rule:*

$$\mathcal{A} = \rho\mathcal{B} \implies A = \rho B$$

where the combinatorial equation $\mathcal{A} = \rho\mathcal{B}$ means $\mathcal{A}$'s objects are identical to those of $\mathcal{B}$ but with weights that have an additional factor ρ.

Proof: This is straightforward: $A = \sum_{a\in\mathcal{A}} w_a z^{|a|} = \sum_{b\in\mathcal{A}} \rho w_b z^{|b|} = \rho B$.

2 Disjoint union *Rule:*

$$\mathcal{A} = \mathcal{B} \cup \mathcal{C} \implies A = B + C$$

where $\mathcal{B}$ and $\mathcal{C}$ are assumed to be disjoint (or if not, their shared objects are considered distinct as members of $\mathcal{A}$). This rule easily extends to a union of any finite number of combinatorial classes.

Proof: $A = \sum_{a\in\mathcal{A}} w_a z^{|a|} = \sum_{b\in\mathcal{B}} w_b z^{|b|} + \sum_{c\in\mathcal{C}} w_c z^{|c|} = B + C$.

3 Product *Rule:*

$$\mathcal{A} = \mathcal{B} \times \mathcal{C} \implies A = B \cdot C$$

where × is the usual cartesian product and for an object $(b, c) \in \mathcal{B} \times \mathcal{C}$ the size is just $|b| + |c|$ (ie as if concatenating the two objects) and the weight is $w_b w_c$. This rule easily extends to a product of any finite number of classes.

Proof:

$$A = \sum_{a\in\mathcal{A}} w_a z^{|a|} = \sum_{b\in\mathcal{B},c\in\mathcal{C}} w_b w_c z^{|b|+|c|} = \sum_{b\in\mathcal{B}} w_b z^{|b|} \sum_{c\in\mathcal{C}} w_c z^{|c|} = B \cdot C$$

4 Sequence *Rule:*

$$\mathcal{A} = \mathrm{SEQ}\left(\mathcal{B}\right) \implies A = \frac{1}{1-B}$$

where the notation $\mathrm{SEQ}\left(\mathcal{B}\right)$ means a sequence of any length of objects from $\mathcal{B}$ (including length 0), that is $\mathrm{SEQ}\left(\mathcal{B}\right) = \mathbf{1} \cup \mathcal{B} \cup (\mathcal{B} \times \mathcal{B}) \cup (\mathcal{B} \times \mathcal{B} \times \mathcal{B}) \cup \cdots$.

Proof: using the rules we've alread derived, $A = \sum_n B^n = \frac{1}{1-B}$

Summary

$$\begin{aligned}
\mathbf{1} &\implies 1 \\
\mathcal{Z} &\implies z \\
\mathcal{A} = \rho\mathcal{B} &\implies A = \rho B \\
\mathcal{A} = \mathcal{B} \cup \mathcal{C} &\implies A = B + C \\
\mathcal{A} = \mathcal{B} \times \mathcal{C} &\implies A = B \cdot C \\
\mathcal{A} = \mathrm{SEQ}\left(\mathcal{B}\right) &\implies A = \frac{1}{1-B}
\end{aligned}$$

These are all the rules we'll need.

References

[1] Alexandru Nica and Roland Speicher. *Lectures on the Combinatorics of Free Probability.* London Mathematical Society Lecture Note Series. Cambridge University Press, 2006.

[2] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics.* Cambridge University Press, 2009.