General-Link Gaussian Multi-Index Models: Gradient and Weight Spectra, Hermite Dynamics, and Block-Wishart Hessians

Janis Heran Aiad

July 2026

Keywords: Gaussian multi-index model; Hermite expansion; population gradient flow; gradient spectrum; weight spectrum; block-Wishart Hessian; logarithmic potential; BBP transition.

2020 Mathematics Subject Classification: 60B20, 62R07, 68T07.

Abstract

We study square-loss regression with Gaussian covariates, a fixed-rank teacher $g(\Theta^\top x)$, and a fixed-rank student $\sigma(W^\top x)$, allowing multivariate and non-polynomial links. We prove that the population risk and its gradient flow depend on the ambient parameters only through the Gram matrices $Q = W^\top W$ and $M = W^\top \Theta$. This yields an exact autonomous system of dimension independent of the ambient dimension, as well as exact finite-matrix formulas for the singular spectra of the population gradient and the weights. We then give a Gaussian-Sobolev approximation theorem: finite Hermite projections converge to the population vector field and to its trajectories on every regular finite-time interval. On the Stiefel branch this recovers the Hermite information exponent and its saddle escape scales.

At finite sample size, we derive an exact block weighted-covariance formula for the empirical Hessian. Existing effective spectral theory applies directly to each diagonal $d \times d$ block under its stated hypotheses, giving a scalar bulk law and a finite outlier determinant. For the complete $pd \times pd$ Hessian, we prove the global limiting law for bounded matrix-valued weights by a noncommutative moment expansion and identify its operator-valued Dyson equation. Applying the block-Wishart variational theory of Montanari and Saeed, we identify the limiting left and right bulk edges and the logarithmic potential outside the support. The anisotropic local law required for full-Hessian outliers remains open. Polynomial links add a second issue because they fall outside the available compact-support outlier theorem. Numerical experiments separately test the exact gradient and weight spectra, Hermite approximation, full-Hessian bulk law, variational edges and logarithmic potential, and non-quadratic outlier diagnostics.

1 Introduction

Quadratic phase retrieval has a useful finite-dimensional structure: with Gaussian inputs and fixed ranks p, k, both the population dynamics and the empirical Hessian can be expressed through finitely many overlaps. It is tempting to regard this as a peculiarity of the square link $u \mapsto u^2$. The main point of this paper is that the mechanism is more general. Gaussianity, not quadraticity, is the source of the closure.

Let $x \sim N(0, I_d)$, let $\Theta \in \mathbb{R}^{d\times k}$ have orthonormal columns, and let $W \in \mathbb{R}^{d\times p}$ be the student matrix. Write

$$h = W^\top x, \qquad y = \Theta^\top x.$$

The student and teacher are

$$f_W(x) = \sigma(h), \qquad f_\star(x) = g(y),$$

for smooth links $\sigma : \mathbb{R}^p \to \mathbb{R}$ and $g : \mathbb{R}^k \to \mathbb{R}$. The population loss is

$$R(W) = \frac{1}{2}\mathbb{E}(\sigma(W^\top x) - g(\Theta^\top x))^2.$$

The empirical Hessian is computed on an independent test sample $x_1, \dots, x_n$, with $n/d \to \alpha \in (0, \infty)$.

Contributions. First, we prove an exact population closure on the finite summaries

$$Q = W^\top W, \qquad M = W^\top \Theta.$$

Second, we prove consistency of finite Hermite approximations, including finite-time convergence of the induced trajectories, and derive exact finite-dimensional formulas for the gradient and weight singular spectra. Third, we derive the general-link empirical Hessian exactly, identify the diagonal-block regime covered by published effective spectral theorems, and prove a global matrix-valued compound-Wishart law for the complete Hessian under bounded weights. We then import the variational characterization of the two bulk edges and the exterior logarithmic potential from [7]. Finally, we give numerical checks that keep exact algebra, proved asymptotic consequences, and the remaining local spectral predictions logically separate.

This separation is essential. Bietti, Bruna, and Pillaud-Vivien [4] analyze a two-timescale Grassmannian flow in which the link is learned rapidly; our exact closure also covers a planted, fixed student link, but does not import their global convergence conclusion. Braun, Loureiro, Quang Minh, and Imaizumi [3] treat anisotropic quadratic phase retrieval; their infinite weighted hierarchy is complementary to the finite isotropic closure proved here.

2 Population Closure

We first isolate assumptions for the exact statement. They are weaker than the boundedness hypotheses used later for spectral outliers.

Assumption 2.1 (Population regularity)**.** *The functions σ and g are locally absolutely continuous, the map $W \mapsto R(W)$ is differentiable, and*

$$\mathbb{E}\big[(\sigma(W^\top x) - g(\Theta^\top x))^2\big] + \mathbb{E}\big[\,\big\|x(\sigma(W^\top x) - g(\Theta^\top x))\nabla\sigma(W^\top x)^\top\big\|_{\mathrm{F}}\,\big] < \infty$$

on the parameter region under consideration. These conditions hold, in particular, for C^1 links and first derivatives of polynomial growth.

Let

$$\Delta(h, y) = \sigma(h) - g(y), \qquad G = \begin{pmatrix} Q & M \\ M^\top & I_k \end{pmatrix}.$$

For $U = (h, y) \sim N(0, G)$ define

$$\Psi(h, y) = \Delta(h, y)\nabla\sigma(h) \in \mathbb{R}^p, \qquad \Xi(G) = \mathbb{E}_G[U\Psi(U)^\top] \in \mathbb{R}^{(p+k)\times p}.$$

Let

$$\Gamma(G) = G^\dagger \Xi(G) = \begin{pmatrix} \Gamma_h(G) \\ \Gamma_y(G) \end{pmatrix}.$$

Theorem 2.2 (Exact general-link population closure)**.** *Under Assumption 2.1, the risk is a function of G alone:*

$$R(W) = \mathcal{R}(G) := \frac{1}{2}\mathbb{E}_G[\Delta(h, y)^2].$$

Moreover,

$$\nabla_W R(W) = W\Gamma_h(G) + \Theta\Gamma_y(G).$$

Consequently the gradient flow $\dot{W} = -\nabla_W R(W)$ *induces the closed finite system*

$$\dot{Q} = -\Gamma_h^\top Q - Q\Gamma_h - \Gamma_y^\top M^\top - M\Gamma_y, \tag{2.1}$$

$$\dot{M} = -\Gamma_h^\top M - \Gamma_y^\top. \tag{2.2}$$

Proof. The vector $(W^\top x, \Theta^\top x)$ is centered Gaussian with covariance G, proving the risk identity. Differentiation under the expectation, justified by Assumption 2.1, gives

$$\nabla_W R(W) = \mathbb{E}\left[x\,\Delta(W^\top x, \Theta^\top x)\nabla\sigma(W^\top x)^\top\right].$$

Since (x, U) is jointly Gaussian, including when G is singular,

$$\mathbb{E}[x \mid W^\top x, \Theta^\top x] = [W, \Theta]G^\dagger U.$$

Taking conditional expectation gives $\nabla_W R = [W, \Theta]G^\dagger \mathbb{E}[U\Psi^\top]$, which is the displayed gradient formula. Equations (2.1)–(2.2) follow from $\dot{Q} = \dot{W}^\top W + W^\top \dot{W}$ and $\dot{M} = \dot{W}^\top \Theta$. $\square$

Proposition 2.3 (Exact gradient and weight spectra)**.** *Under the assumptions of Theorem 2.2, the nonzero squared singular values of the population gradient are the eigenvalues of the* $p \times p$ *matrix*

$$\mathcal{G}_\nabla(Q, M) := \Gamma(G)^\top G\Gamma(G) = \Gamma_h^\top Q\Gamma_h + \Gamma_h^\top M\Gamma_y + \Gamma_y^\top M^\top \Gamma_h + \Gamma_y^\top \Gamma_y. \tag{2.3}$$

The nonzero squared singular values of W *are the eigenvalues* $q_1(t), \ldots, q_p(t)$ *of* $Q(t)$*. If* $q_i(t)$ *is simple with normalized eigenvector* $v_i(t)$*, then*

$$\dot{q}_i = -2q_i v_i^\top \Gamma_h v_i - 2v_i^\top M\Gamma_y v_i. \tag{2.4}$$

Thus both spectra are functions of the finite Gram state (Q, M)*, without an ambient-dimensional closure assumption.*

Proof. Writing $B = [W, \Theta]$, Theorem 2.2 gives

$$\nabla_W R = B\Gamma, \qquad B^\top B = G,$$

and therefore

$$(\nabla_W R)^\top \nabla_W R = \Gamma^\top G\Gamma.$$

This proves (2.3). The identity $W^\top W = Q$ gives the weight spectrum. At a simple eigenvalue, the standard perturbation formula and (2.1) yield $\dot{q}_i = v_i^\top \dot{Q} v_i$; using $Qv_i = q_i v_i$ and equality of a scalar quadratic form with that of its transpose gives (2.4). $\square$

Remark 2.4 (Closure does not imply global convergence)**.** *Theorem 2.2 is an exact reduction, not a claim that every planted-link flow reaches the teacher. The reduced system can retain saddles and nonglobal attractors; this distinction is already visible in the planted setting of [4].*

Quadratic specialization. For

$$\sigma(h) = p^{-1}\|h\|^2, \qquad g(y) = y^\top \Lambda y,$$

put

$$A_Q = \frac{2}{p}\left[\left(\frac{\operatorname{Tr} Q}{p} - \operatorname{Tr}\Lambda\right) I_p + \frac{2}{p}Q\right].$$

Wick's formula gives

$$\dot{M} = -A_Q M + \frac{4}{p} M\Lambda, \qquad \dot{Q} = -A_Q Q - QA_Q + \frac{8}{p} M\Lambda M^\top. \tag{2.5}$$

Indeed,

$$\mathbb{E}[xh^\top \|h\|^2] = W(\mathrm{Tr}(Q)I_p + 2Q), \qquad \mathbb{E}[xh^\top y^\top \Lambda y] = W \,\mathrm{Tr}\,\Lambda + 2\Theta\Lambda M^\top,$$

and substitution into the population gradient yields $\nabla_W R = WA_Q - (4/p)\Theta\Lambda M^\top$, from which (2.5) follows. Thus quadratic multi-index phase retrieval is the degree-two specialization of the Gaussian closure, not a separate mechanism.

3 Hermite DMFT Form

Let $\widehat{\mathrm{He}}_\alpha$ be the orthonormal multivariate probabilists' Hermite basis. Formally write

$$\sigma(h) = \sum_{\alpha \in \mathbb{N}^p} a_\alpha \widehat{\mathrm{He}}_\alpha(h), \qquad g(y) = \sum_{\beta \in \mathbb{N}^k} b_\beta \widehat{\mathrm{He}}_\beta(y).$$

For finite expansions, every entry of $\Xi(G)$ is a finite Wick contraction. Away from the normalized branch $Q = I_p$, the coefficients remain those of the functions under standard Gaussian measure while the contractions are evaluated at covariance G; no independence of h and y is assumed.

Assumption 3.1 (Uniform Hermite approximability)**.** *Let $\mathcal{K}$ be a compact subset of the Gram cone such that $\lambda_{\min}(G) \geq c_{\mathcal{K}} > 0$. For $U_G \sim N(0, G)$ assume that, as $L \to \infty$,*

$$\sup_{G \in \mathcal{K}} \mathbb{E}\big[|\sigma_L(h) - \sigma(h)|^2 + \|\nabla\sigma_L(h) - \nabla\sigma(h)\|^2\,\big] \longrightarrow 0,$$

$$\sup_{G \in \mathcal{K}} \mathbb{E}|g_L(y) - g(y)|^2 \longrightarrow 0,$$

and that $\Delta_L \nabla\sigma_L$ is uniformly bounded in L^2 over $G \in \mathcal{K}$. Here $\sigma_L = P_L\sigma$ and $g_L = P_L g$ are standard Gaussian Hermite projections. Assume also that the limiting and projected vector fields are uniformly locally Lipschitz on $\mathcal{K}$.

Proposition 3.2 (Hermite truncation consistency)**.** *Let P_L be projection onto Hermite degree at most L and set $\sigma_L = P_L\sigma$, $g_L = P_L g$. Under Assumption 3.1, the vector fields F_L in (2.1)–(2.2) converge uniformly to F on $\mathcal{K}$. If $G_L(0) \to G(0)$ and both trajectories remain in $\mathcal{K}$ for $0 \leq t \leq T$, then*

$$\sup_{0 \leq t \leq T} \|G_L(t) - G(t)\|_{\mathrm{F}} \longrightarrow 0.$$

Proof. Write $\Psi_L = \Delta_L \nabla\sigma_L$. Adding and subtracting $\Delta_L\nabla\sigma$, then applying Cauchy–Schwarz and Assumption 3.1, gives

$$\sup_{G \in \mathcal{K}} \mathbb{E}_G \|\Psi_L - \Psi\| \longrightarrow 0.$$

The same argument with the factor U_G is valid because centered Gaussians with covariance in $\mathcal{K}$ have uniformly bounded moments. Hence $\Xi_L(G) \to \Xi(G)$ uniformly. The map $G \mapsto G^{-1}$ is bounded and Lipschitz on $\mathcal{K}$, so $\Gamma_L \to \Gamma$ and therefore $F_L \to F$ uniformly there.

Let $L_{\mathcal{K}}$ be a common Lipschitz constant and $\varepsilon_L = \sup_{G \in \mathcal{K}} \|F_L(G) - F(G)\|_{\mathrm{F}}$. As long as both solutions remain in $\mathcal{K}$,

$$\|G_L(t) - G(t)\|_{\mathrm{F}} \leq \|G_L(0) - G(0)\|_{\mathrm{F}} + \int_0^t \big(L_{\mathcal{K}} \|G_L(s) - G(s)\|_{\mathrm{F}} + \varepsilon_L\big)\,\mathrm{d}s.$$

Gronwall's inequality proves the trajectory statement. □

In the diagonal Stiefel branch $Q = I_p$, $p = k$, and $M = \mathrm{diag}(m_1, \dots, m_k)$, this representation gives

$$\mathbb{E}[\sigma(h)g(y)] = \sum_{\alpha} c_\alpha^2 \prod_{i=1}^{k} m_i^{\alpha_i}$$

for a shared planted link with coefficients c_α. In the scalar planted case, spherical gradient flow satisfies

$$\dot{m} = (1 - m^2) \sum_{\ell \geq 1} \ell c_\ell^2 m^{\ell-1}. \tag{3.1}$$

If $s = \min\{\ell \geq 1 : c_\ell \neq 0\}$ and $m(0) \asymp d^{-1/2}$, integration of (3.1) until a fixed small overlap yields

$$T_s(d) \asymp 1 \ (s = 1), \qquad T_2(d) \asymp \log d, \qquad T_s(d) \asymp d^{(s-2)/2} \ (s > 2).$$

Thus the first nonzero Hermite degree in an unseen direction is the information exponent. This is the population mechanism behind the Gaussian multi-index theory of Bietti, Bruna, and Pillaud-Vivien [4]; anisotropic data lead to the weighted hierarchy and Volterra/Stieltjes analysis of Braun, Loureiro, Quang Minh, and Imaizumi [3].

4 Empirical Hessian and Effective Spectrum

Let $X \in \mathbb{R}^{n \times d}$ have rows $x_\ell^\top$. For the square loss, the sample risk is

$$R_n(W) = \frac{1}{2n} \sum_{\ell=1}^{n} \left(\sigma(W^\top x_\ell) - g(\Theta^\top x_\ell)\right)^2.$$

Proposition 4.1 (Exact finite-sample Hessian)**.** *If $\sigma \in C^2$, the Hessian with respect to the columns of W has blocks*

$$H_{ab}(W) = \frac{1}{n} X^\top D_{ab}(W) X,$$

where

$$D_{ab,\ell}(W) = A_{ab}(W^\top x_\ell, \Theta^\top x_\ell)$$

and

$$A(h, y) = \nabla\sigma(h)\nabla\sigma(h)^\top + \Delta(h, y)\nabla_h^2\sigma(h). \tag{4.1}$$

Equivalently, under column-wise vectorization,

$$\nabla^2 R_n(W) = \frac{1}{n} \sum_{\ell=1}^{n} A(h_\ell, y_\ell) \otimes x_\ell x_\ell^\top.$$

Proof. For one observation, the gradient with respect to w_a is $x\Delta\, \partial_a \sigma(h)$. Differentiating with respect to w_b gives $xx^\top[\partial_a\sigma(h)\partial_b\sigma(h) + \Delta\, \partial_{ab}^2\sigma(h)]$. Averaging proves both formulas. □

Proposition 4.2 (Signal–bulk decomposition)**.** *Let $q = p + k$, $B = [W, \Theta]$, and suppose $G = B^\top B$ is positive definite. Set $L = BG^{-1/2}$ and complete L to an orthogonal matrix $[L, L_\perp]$. There are independent standard Gaussian vectors $z_\ell \in \mathbb{R}^q$ and $\xi_\ell \in \mathbb{R}^{d-q}$ such that*

$$(h_\ell, y_\ell) = G^{1/2} z_\ell, \qquad x_\ell = L z_\ell + L_\perp \xi_\ell.$$

Writing $A_\ell = A(G^{1/2} z_\ell)$, the Hessian in the induced orthogonal basis is

$$\begin{pmatrix} T_n & C_n \\ C_n^\top & B_n \end{pmatrix}, \qquad \begin{cases} T_n = n^{-1} \sum_\ell A_\ell \otimes z_\ell z_\ell^\top, \\ C_n = n^{-1} \sum_\ell A_\ell \otimes z_\ell \xi_\ell^\top, \\ B_n = n^{-1} \sum_\ell A_\ell \otimes \xi_\ell \xi_\ell^\top. \end{cases} \tag{4.2}$$

In particular, A_ℓ is independent of ξ_ℓ, and

$$\operatorname{rank}\left[\nabla^2 R_n(W) - (I_p \otimes [L, L_\perp]) \begin{pmatrix} 0 & 0 \\ 0 & B_n \end{pmatrix} (I_p \otimes [L, L_\perp])^\top\right] \le 2pq.$$

Consequently the empirical distribution functions of these two matrices differ by at most $2q/d$.

Proof. Gaussian orthogonal invariance gives independent $z_\ell = L^\top x_\ell$ and $\xi_\ell = L_\perp^\top x_\ell$. Since $B^\top L = G^{1/2}$, the weight A_ℓ depends only on z_ℓ. Conjugating the Kronecker formula in Proposition 4.1 gives (4.2). Its difference from $\operatorname{diag}(0, B_n)$ is supported on the first pq rows or columns, and therefore has rank at most $2pq$. The final claim is the standard rank inequality for empirical spectral distribution functions. □

Fix $a \in [p]$ and write $a_a(h, y) = A_{aa}(h, y)$. The diagonal block $H_{aa} = n^{-1} X^\top D_{aa} X$ is a scalar-weighted self-coupled empirical matrix of the exact form studied by Ben Arous, Gheissari, Huang, and Jagannath [6]. In our aspect-ratio convention its scalar Stieltjes transform is characterized by

$$-\frac{1}{s_a(z)} = z - \mathbb{E}\left[\frac{a_a(h, y)}{1 + \alpha^{-1} s_a(z) a_a(h, y)}\right]. \tag{4.3}$$

The cited outlier equation is a determinant on the finite span of all projections on which a_a depends. Importantly, Corollary 1.16 of [6] states this conclusion for the (a, a) blocks; it does not identify the spectrum of the complete block matrix in Proposition 4.1.

Theorem 4.3 (Diagonal-block effective spectrum; consequence of [6])**.** *Let p, k be fixed, let $n/d \to \alpha \in (0, \infty)$, and suppose $G_d \to G$ with G positive definite. For a fixed a, assume that the scalar weight $a_a(h, y)$ satisfies Assumptions 1–3 of [6]; for outliers this includes uniform compact support and the stated non-concentration condition near zero. Then the empirical spectral distribution of H_{aa} converges to the deterministic law characterized by (4.3). Away from its support, the outlier eigenvalues and the projections of their eigenvectors onto the summary-statistic space obey the finite determinant and effective eigenvector equations of Theorems 1.11–1.12 in [6].*

Proof. The coefficient $a_a(W^\top x, \Theta^\top x)$ is a measurable function of a fixed number of Gaussian projections, so Assumption 1 of [6] holds. Under the remaining assumptions in the statement, the result is a direct application of Theorems 1.10–1.12 of that reference, with $d/n = \alpha^{-1}$. □

The exact signal–bulk reduction also permits a global theorem for the complete Hessian. We state it for a slightly more general triangular array. Put $D = d - q$, $c_D = D/n$, and let

$$B_D = \frac{1}{n} \sum_{\ell=1}^{n} A_\ell \otimes \xi_\ell \xi_\ell^\top, \qquad \xi_\ell \overset{\text{iid}}{\sim} N(0, I_D),$$

where the symmetric $p \times p$ weights are independent of the ξ_ℓ. For every matrix-valued polynomial P in the entries of a symmetric matrix, assume

$$\frac{1}{n} \sum_{\ell=1}^{n} P(A_\ell) \longrightarrow \mathbb{E} P(A) \quad \text{almost surely}, \qquad \sup_{D, \ell} \|A_\ell\| \le K. \tag{4.4}$$

This includes iid bounded weights and bounded continuous Gaussian-projection weights whose limiting Gram matrix is nondegenerate. Related operator-valued Wishart limits and quantitative Cauchy-transform methods are developed in [2]; the elementary moment argument below records the precise indefinite matrix-weight recursion needed here.

Lemma 4.4 (Matrix-valued compound-Wishart moments)**.** *Suppose p is fixed, $c_D \to c \in (0,\infty)$, and (4.4) holds. For every fixed $m \geq 0$,*

$$(\mathrm{id}_p \otimes D^{-1}\operatorname{Tr})(B_D^m) \longrightarrow M_m \qquad \textit{almost surely,}$$

where $M_0 = I_p$ and

$$M_m = \sum_{r=0}^{m-1} c^r \sum_{\substack{j_0+\cdots+j_r=m-r-1 \\ j_i \geq 0}} M_{j_0}\, \mathbb{E}[AM_{j_1}A \cdots M_{j_r}A]\,. \tag{4.5}$$

Proof. Expanding the partial trace, with $\ell_{m+1} = \ell_1$, gives

$$\frac{1}{Dn^m} \sum_{\ell_1,\ldots,\ell_m} A_{\ell_1} \cdots A_{\ell_m} \prod_{a=1}^{m} \langle \xi_{\ell_a}, \xi_{\ell_{a+1}} \rangle. \tag{4.6}$$

Apply Wick's formula to the Gaussian coordinates and group the sample indices according to their kernel partition. If that partition has b blocks, a term has at most $m+1-b$ free coordinate indices. After the normalization in (4.6), it is of order at most $D^{m+1-b}n^b/(Dn^m) = O(1)$. Equality in the coordinate count holds exactly for the planar contractions, equivalently for noncrossing partitions of the cyclic word; every crossing or twisted real-Gaussian contraction loses at least one power of D.

To evaluate the planar terms, expose the block containing the first point. If it contains $r+1$ points, the $r+1$ complementary intervals have lengths $j_0, \ldots, j_r$ with $j_0 + \cdots + j_r = m-r-1$. The block contributes c^r, the first interval contributes M_{j_0}, and independence of its sample label from the labels internal to the remaining intervals gives

$$M_{j_0}\mathbb{E}[AM_{j_1}A \cdots M_{j_r}A].$$

Summing over r and the interval lengths yields (4.5). Assumption (4.4) replaces each bounded sample average by its limit. Apply the same Wick expansion to the fourth power of a centered entry of the partial trace. Every surviving contraction connects at least two trace copies and loses at least two coordinate loops, uniformly for $\|A_\ell\| \leq K$; hence its fourth moment is $O(D^{-2})$. Markov's inequality and Borel–Cantelli give entrywise almost-sure concentration around the conditional planar sum. Since p is fixed, this proves the asserted matrix convergence for every fixed m. ∎

The first three moments, useful for checking the ordering in the noncommutative recursion, are

$$\begin{aligned}
M_1 &= \mathbb{E}A, \\
M_2 &= (\mathbb{E}A)^2 + c\mathbb{E}[A^2], \\
M_3 &= (\mathbb{E}A)^3 + c\{\mathbb{E}[A^2]\mathbb{E}A + \mathbb{E}A\mathbb{E}[A^2] + \mathbb{E}[A(\mathbb{E}A)A]\} + c^2\mathbb{E}[A^3].
\end{aligned}$$

Theorem 4.5 (Global law of the complete bounded-weight Hessian)**.** *Under the assumptions of Lemma 4.4, the empirical spectral distribution of B_D converges almost surely to a compactly supported deterministic probability measure μ. Its Stieltjes transform is $p^{-1}\operatorname{Tr} S(z)$, where $S : \mathbb{C}_+ \to \mathbb{C}^{p\times p}$ is the Herglotz solution normalized by $zS(z) \to -I_p$ at infinity of*

$$-\,S(z)^{-1} = zI_p - \mathbb{E}\left[A\,(I_p + cS(z)A)^{-1}\right]. \tag{4.7}$$

The complete Hessian in Proposition 4.2 has the same limiting measure.

Proof. Since $-KI_p \le A_\ell \le KI_p$,

$$-KI_p \otimes \frac{1}{n}\sum_\ell \xi_\ell \xi_\ell^\top \le B_D \le KI_p \otimes \frac{1}{n}\sum_\ell \xi_\ell \xi_\ell^\top .$$

The extreme-eigenvalue bound for a Gaussian sample covariance matrix therefore confines the spectrum almost surely to a fixed compact interval. The lemma and the method of moments yield a unique compactly supported limit with scalar moments $p^{-1}\operatorname{Tr} M_m$.

For $|w|$ small, let $M(w)=\sum_{m\ge 0} M_m w^m$. Summing (4.5) gives

$$M(w) = I_p + wM(w)\mathbb{E}\Big[A(I_p - cwM(w)A)^{-1}\Big] .$$

Set $S(z) = -z^{-1}M(z^{-1})$. Left multiplication by $M(w)^{-1}$ gives (4.7). The matrix Cauchy transform of the compactly supported moment limit supplies its Herglotz continuation from large z to $\mathbb{C}_+$ and selects the normalized solution. Taking the normalized trace identifies the scalar Stieltjes transform. Finally, the rank inequality in Proposition 4.2 transfers the limiting measure from B_D to the complete Hessian. $\square$

When the limit has a density, it is recovered from

$$\rho(E) = \frac{1}{\pi}\lim_{\eta\downarrow 0} \Im \frac{1}{p}\operatorname{Tr} S(E+i\eta).$$

Equation (4.7) reduces to (4.3) when $p=1$. The theorem is global: it does not by itself provide support convergence, an anisotropic local law, or outlier control. The next result supplies support convergence and exact bulk edges, but not the anisotropic resolvent estimates needed for finite-rank outliers.

The block-Wishart theory of Montanari and Saeed [7] strengthens the preceding global statement for the independent orthogonal bulk. To state the result in our aspect-ratio convention, let ν be a compactly supported law on Sym_p and define, for $U \succ 0$,

$$\mathbb{K}_{\alpha,\nu}(U) = \int A(I_p + UA)^{-1}\,\nu(\mathrm{d}A) - \alpha^{-1}U^{-1}. \tag{4.8}$$

Its derivative on Sym_p is the self-adjoint operator

$$D\mathbb{K}_{\alpha,\nu}(U)[\Delta] = \alpha^{-1}U^{-1}\Delta U^{-1} - \int A(I_p+UA)^{-1}\Delta(I_p+AU)^{-1}A\,\nu(\mathrm{d}A). \tag{4.9}$$

Set

$$\mathcal{P}_-(\alpha,\nu) = \Big\{U\succ 0:\ U^{-1}+A\succ 0 \text{ for all } A\in\operatorname{supp}\nu,\quad D\mathbb{K}_{\alpha,\nu}(U)\succ 0\Big\}. \tag{4.10}$$

Theorem 4.6 (Block-Wishart edges and logarithmic potential)**.** *Let $A_1, A_2, \ldots$ be iid with compactly supported law ν on Sym_p, independent of the iid Gaussian vectors ξ_ℓ, and let $n/D \to \alpha \in (0,\infty)$. For $B_D = n^{-1}\sum_{\ell=1}^n A_\ell\otimes \xi_\ell\xi_\ell^\top$, write $\zeta_- = \inf\operatorname{supp}\mu$ and $\zeta_+ = \sup\operatorname{supp}\mu$ for the edges of its limiting law. Then*

$$\zeta_- = \sup_{U\in\mathcal{P}_-(\alpha,\nu)} \lambda_{\min}(\mathbb{K}_{\alpha,\nu}(U)). \tag{4.11}$$

If ν^- is the pushforward of ν under $A\mapsto -A$, then

$$\zeta_+(\alpha,\nu) = -\zeta_-(\alpha,\nu^-). \tag{4.12}$$

Moreover, for every $u < \zeta_-$,

$$\Phi_{\alpha,\nu}(u) := \int \log|\lambda - u|\,\mu(\mathrm{d}\lambda)$$
$$= \frac{1}{p}\inf_{U\in\mathcal{P}_-(\alpha,\nu)}\left\{-\alpha u\operatorname{Tr}U + \alpha\int \log\det(I_p + AU)\,\nu(\mathrm{d}A) - \log\det U - p(\log\alpha + 1)\right\}. \tag{4.13}$$

The analogous formula to the right follows by the same sign transformation. Almost surely the support of the empirical bulk spectrum converges to $\operatorname{supp}\mu$*, and*

$$\frac{1}{pD}\log|\det(B_D - uI_{pD})| \longrightarrow \Phi_{\alpha,\nu}(u). \tag{4.14}$$

Proof. Up to the fixed permutation exchanging parameter and output indices, B_D is the block-Wishart matrix of [7] with their ambient dimension equal to D, block size equal to p, and aspect ratio $n/D = \alpha$. Gaussian rows satisfy their concentration assumption. Their matrix Stieltjes transform is $U(z) = \alpha^{-1}S(z)$ in the normalization of (4.7); substituting this relation in (4.8) gives $\mathbb{K}_{\alpha,\nu}(U(z)) = zI_p$. Proposition 1 and Theorems 1–2 of [7] now give support convergence, (4.11), and (4.13). Applying the left-edge statement to $-B_D$ proves (4.12). Support convergence makes $\log|\lambda - u|$ bounded and continuous on the empirical spectra for all large D, which yields (4.14). ☐

Corollary 4.7 (Application to a fixed general-link state)**.** *Fix a positive-definite Gram state G and suppose the Hessian weight $A(h,y)$ in Proposition 4.1 is bounded and continuous. Let*

$$\nu_G = \operatorname{Law}\left(A((G^{1/2}Z)_{1:p}, (G^{1/2}Z)_{p+1:p+k})\right), \qquad Z \sim N(0, I_{p+k}).$$

Then the orthogonal bulk in Proposition 4.2 obeys Theorem 4.6 with $\nu = \nu_G$. In particular, its stability edge and exterior log-determinant are functions of (Q, M) alone through (4.11) and (4.13).

Proof. Gaussian regression makes the finite signal coordinates Z_ℓ independent of the orthogonal coordinates ξ_ℓ. The weights are iid functions of Z_ℓ, hence have law ν_G and are independent of ξ_ℓ. Boundedness gives compact support, so the theorem applies. ☐

Remark 4.8 (What the logarithmic-potential formula proves)**.** *If $\zeta_- > 0$, setting $u = 0$ in (4.13) gives the bulk log-determinant entering a Kac–Rice calculation for minima. A finite-rank signal perturbation has the same normalized limit provided it creates no eigenvalue approaching zero exponentially fast. When 0 lies inside the bulk support, however, Theorem 4.6 does not by itself justify the singular logarithmic integral required for general-index critical points. Nor does it address the stationarity-conditioned, same-sample Hessian.*

Remark 4.9 (Unbounded links)**.** *The scalar-block bulk theorem of [6] can tolerate weaker tails, but its outlier theorem assumes compactly supported effective weights. Quadratic and higher Hermite links make $A(h,y)$ unbounded. Even after an operator-valued theorem is established for bounded matrix weights, removing a truncation $A^{(B)}$ requires estimates uniform in B at the edge and at isolated roots. The polynomial-link plots below test both extensions but prove neither one.*

5 Symmetry-Reduced BBP Diagnostics

For each scalar diagonal block, [6] gives a rigorous determinant condition on the finite summary-statistic space. For the complete matrix-weighted Hessian, we use the following Schur-channel prediction. Let ζ be a normalized teacher projection and define

$$K_\zeta(z) = \mathbb{E}\left[\zeta^2 A(h,y)\left(I_p + \alpha^{-1}S(z)A(h,y)\right)^{-1}\right].$$

A simple left branch is a solution of

$$\lambda_{\text{out}} = \lambda_{\min} K_\zeta(\lambda_{\text{out}}), \qquad \lambda_{\text{out}} < x_-,$$

and a simple right branch uses $\lambda_{\max}$ with $\lambda_{\text{out}} > x_+$. More generally one solves $\det(zI_p - K_\zeta(z)) = 0$. At a simple normalized eigenbranch $k(z) = v(z)^\top K_\zeta(z) v(z)$, the scalar residue diagnostic is

$$\Omega_{\text{out}} = \frac{1}{1 - \partial_z k(\lambda_{\text{out}})}.$$

The BBP contact is predicted when the root reaches the corresponding regular bulk edge. Establishing these equations from the complete Hessian requires the matrix-weighted local law identified after (4.7); the cubic experiment additionally requires removal of truncation. Accordingly, K_ζ and Ω_{out} are reported below as diagnostics, not as consequences of [6].

6 Numerical Evidence

The experiments are generated by

`experiments/general_link/hermite_population_general_link.py`.

They test identities separately from asymptotic random-matrix claims. All Gaussian quadratures and Monte Carlo samples use fixed seeds recorded in the script. The population recovery test compares the generic closure with the closed quadratic equations and gives maximum relative errors $7.5 \cdot 10^{-15}$ for Q and $6.7 \cdot 10^{-15}$ for M. A direct finite-difference check of Proposition 4.1 gives relative Frobenius error $2.2 \cdot 10^{-10}$. At the largest Monte Carlo sample, the relative error of the two gradient singular values is 1.22%. The terminal relative error of the two squared weight singular values is 1.02%.

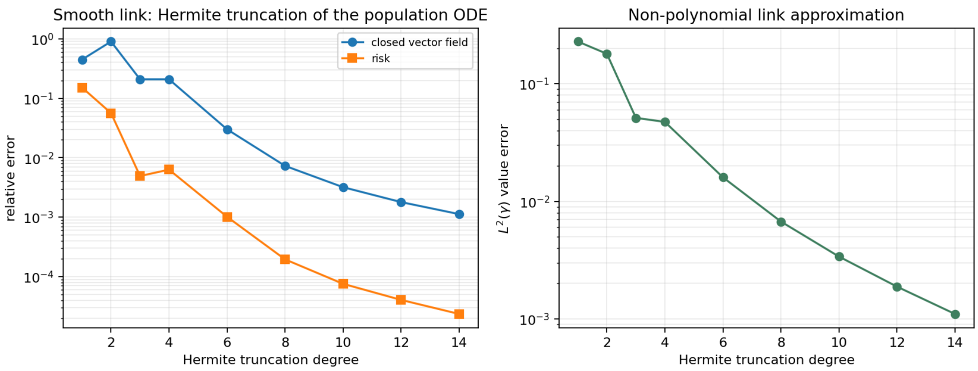

Figure 1: A smooth non-polynomial tanh/sinusoidal link is projected onto finite Hermite degrees. The closed population vector field converges to the direct Gaussian-quadrature vector field; at degree 14 the relative vector-field error is $1.1 \cdot 10^{-3}$.

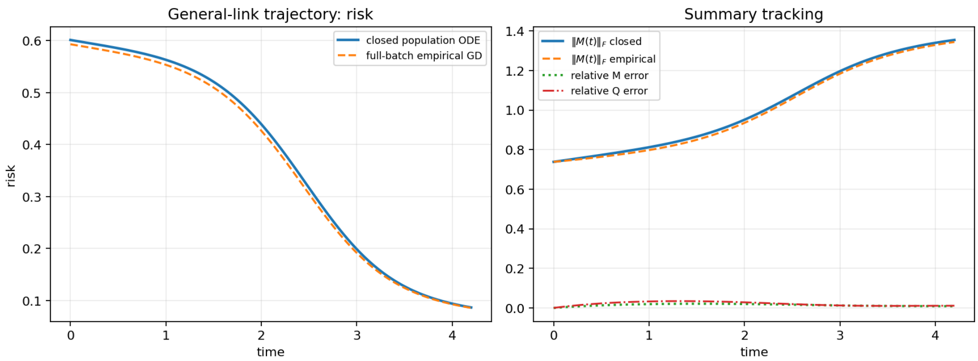

Figure 2: Closed population ODE versus finite full-batch gradient descent for a mixed non-quadratic Hermite link in ambient dimension $d = 56$ using 18,000 independent Gaussian samples. The terminal relative errors are 1.1% for Q and 0.9% for M.

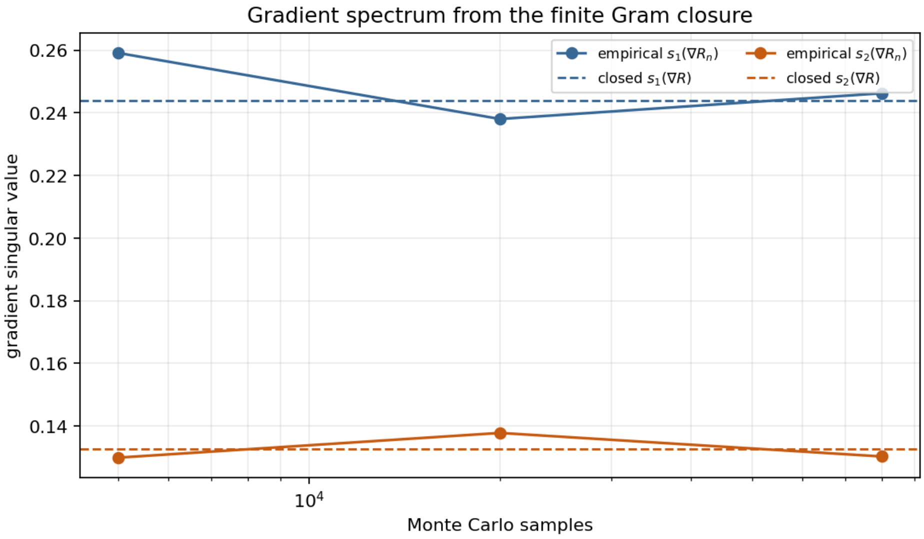

Figure 3: The two singular values of the population gradient obtained from $\mathcal{G}_\nabla(Q, M)$ in (2.3), compared with direct ambient-dimensional Monte Carlo gradients. This tests the spectral statement separately from entrywise gradient agreement.

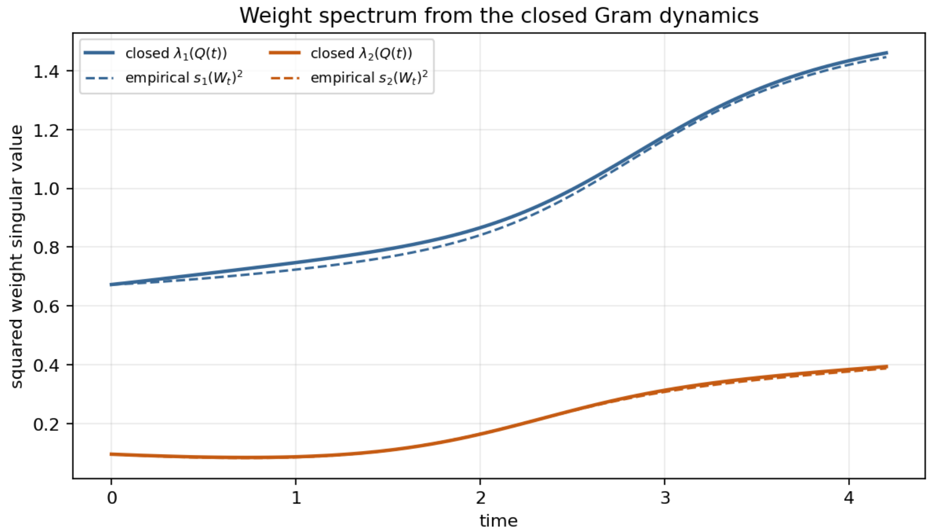

Figure 4: Squared singular values of the student weights. Solid curves are the eigenvalues of the closed Gram trajectory $Q(t)$; dashed curves are the eigenvalues of $W_t^\top W_t$ along finite full-batch gradient descent on the same experiment as Figure 2.

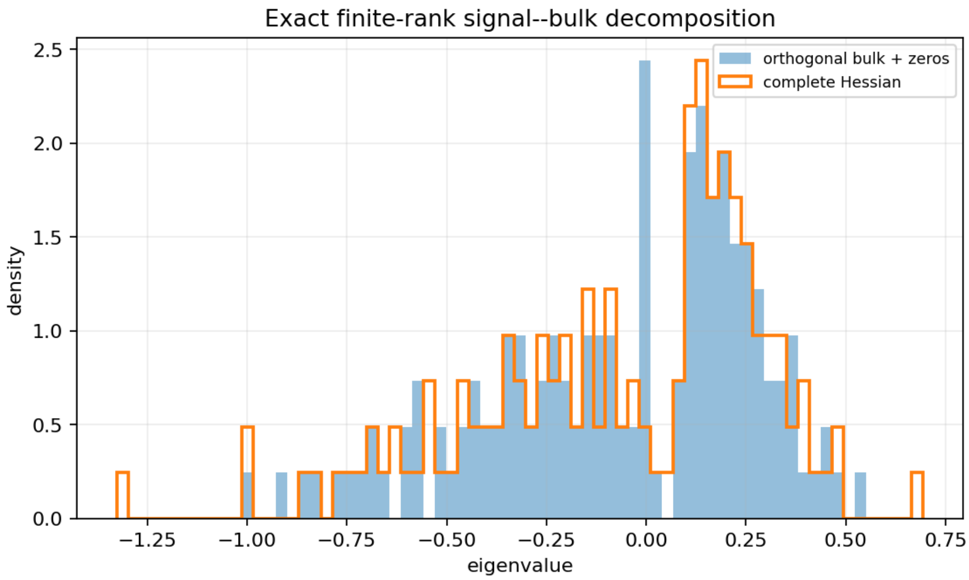

Figure 5: Exact signal–bulk decomposition from Proposition 4.2. The rotated Hessian is reconstructed with relative Frobenius error $6.7 \cdot 10^{-16}$. The signal perturbation has numerical rank 16, equal to the bound $2p(p+k)$, and the spectral CDF distance 0.0417 is below the rank bound $2(p+k)/d = 0.1111$.

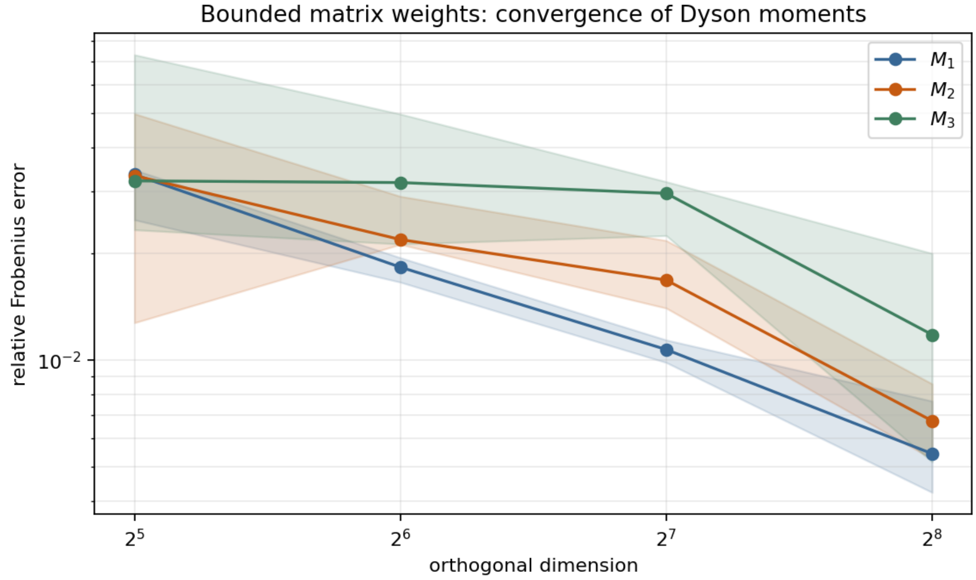

Figure 6: Moment check for the bounded-weight global law with $p = 2$ and $n/D = 5$. Curves show median relative Frobenius errors over five independent repetitions and bands show interquartile ranges. At $D = 128$, the errors for M_1, M_2, M_3 are 0.0107, 0.0168, and 0.0296; at $D = 256$ they fall to 0.0054, 0.0067, and 0.0118, respectively.

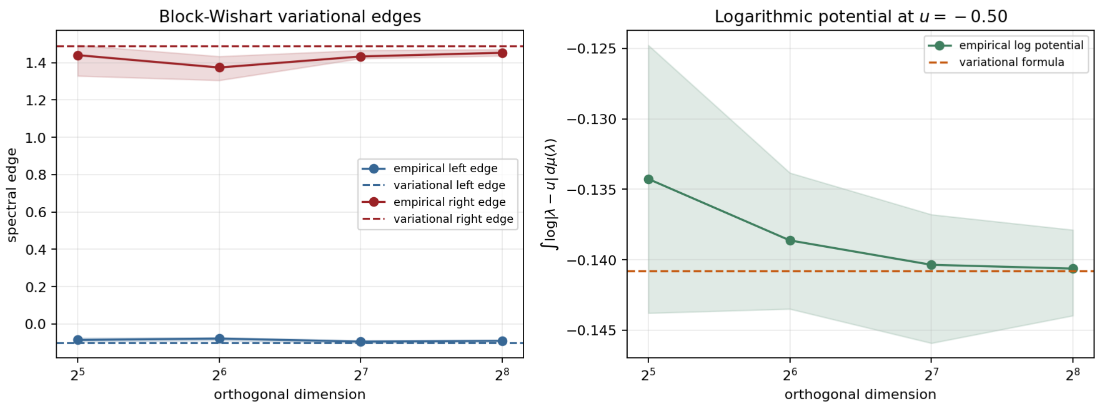

Figure 7: Direct test of Theorem 4.6 for a bounded indefinite 2×2 block-weight law with $n/D = 5$. Medians and interquartile bands use six repetitions. At $D = 256$, the absolute errors of the left edge, right edge, and exterior logarithmic potential are respectively $9.7 \cdot 10^{-3}$, $3.6 \cdot 10^{-2}$, and $1.5 \cdot 10^{-4}$.

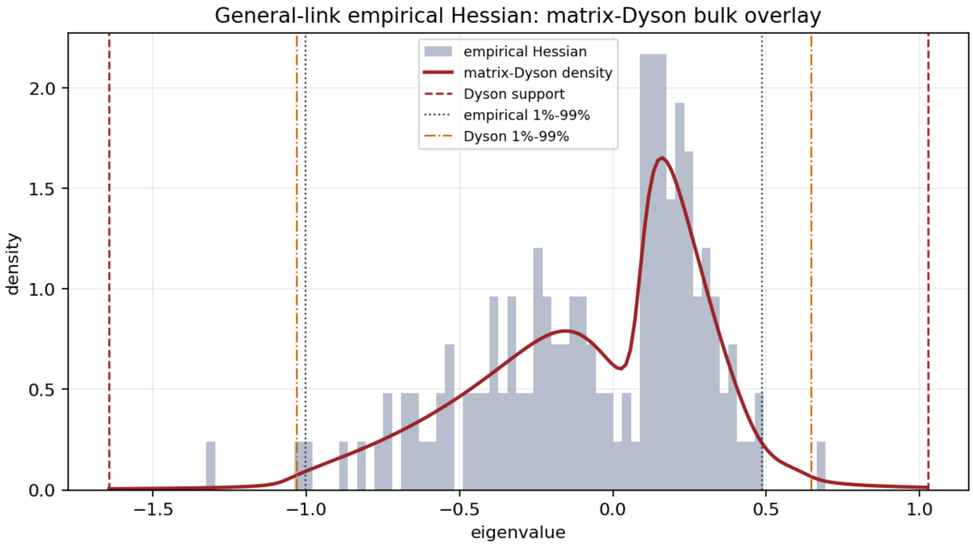

Figure 8: Full-Hessian matrix-Dyson calculation for the general-link empirical Hessian on an independent sample with $d = 72$, $n = 720$, and $p = 2$. The numerical density has mass 0.981 on the displayed grid; its 1% and 99% quantiles differ from the empirical quantiles by 0.028 and 0.163, respectively. These polynomial weights are unbounded, so this panel tests the truncation extension beyond Theorem 4.5.

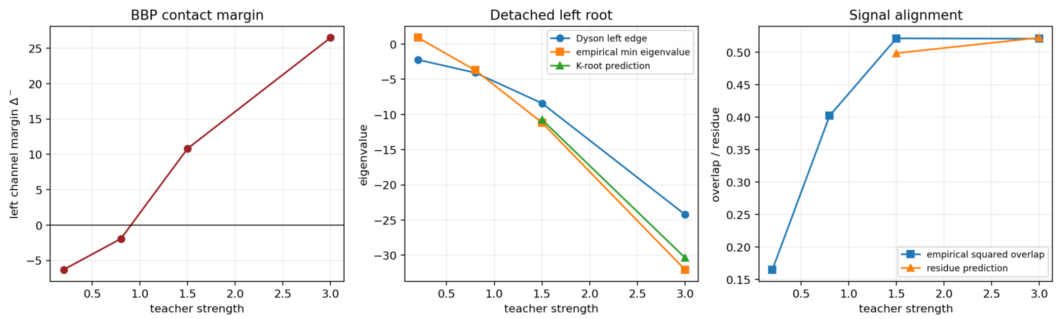

Figure 9: Numerical BBP root diagnostic for a genuinely non-quadratic cubic-Hermite single-index link. As teacher strength increases, the left channel margin crosses zero; only then does the equation $\lambda = \lambda_{\min} K_\zeta(\lambda)$ have an outlier root. At the two detached points the root errors are 0.42 and 1.70, while predicted and empirical squared overlaps are both near 0.5. Since the cubic weights are unbounded, this is evidence for the truncation limit described above.

7 Scope and Open Problems

The exact results of this paper are Theorem 2.2, Proposition 2.3, Proposition 3.2, Propositions 4.1 and 4.2, and the bounded-weight global law in Theorem 4.5. The independent-bulk edges and exterior logarithmic potential follow rigorously from Theorem 4.6. Theorem 4.3 is a carefully scoped consequence of [6] for each diagonal Hessian block. The full-Hessian channel diagnostic is not claimed as a theorem here.

Four extensions remain genuinely open. First, even for bounded links, the complete $pd \times pd$ Hessian requires an anisotropic local law for the finite-rank Schur complement and its derivative. The independent bulk law, support edges, and exterior logarithmic potential are now covered by Theorems 4.5 and 4.6; the missing local theorem must control resolvent bilinear forms in the signal directions. Proposition 4.2 reduces this task to an independent orthogonal bulk plus finite rank. General matrix-Dyson stability theory [1] supplies an analytic template, but an ensemble-specific covariance linearization and local law are still needed. Second, for unbounded polynomial or subexponential links one then needs estimates uniform in the truncation level at regular edges and isolated roots. Third, along a data-dependent training trajectory, a test-sample block theorem is available conditionally, but the same-sample Hessian requires leave-one-out or resolvent comparison estimates. Fourth, replacing isotropic covariates by a power-law covariance destroys finite Gram closure; the correct state is then the weighted infinite hierarchy analyzed in the quadratic case by [3]. Extending that hierarchy to general links requires a separate well-posedness and Volterra analysis.

These are not needed for the finite-dimensional population theorem. They are precisely the additional steps required to turn the polynomial-link spectral diagnostics into a theorem uniform over energy or training time.

Code and data availability. The experiment generator is `experiments/general_link/hermite_population_general_link.py`. Its fixed-seed numerical summaries and all figures used in this article are in `refs/report/3/figures/general_link_hermite_population/`. The detailed theorem target for the full matrix-weighted Hessian is recorded in `refs/report/3/general_link_full_hessian_theorem_target.md`.

References

[1] O. Ajanki, L. Erdős, and T. Krüger, *Stability of the matrix Dyson equation and random matrices with correlations*, Probability Theory and Related Fields 173 (2019), 293–373. https://doi.org/10.1007/s00440-018-0835-z.

[2] M. Banna and G. Cébron, *Operator-valued matrices with free or exchangeable entries*, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 59 (2023), 503–537. https://doi.org/10.1214/22-AIHP1255.

[3] G. Braun, B. Loureiro, H. Quang Minh, and M. Imaizumi, *Fast Escape, Slow Convergence: Learning Dynamics of Phase Retrieval under Power-Law Data*, arXiv:2511.18661, 2025. https://arxiv.org/abs/2511.18661.

[4] A. Bietti, J. Bruna, and L. Pillaud-Vivien, *On Learning Gaussian Multi-index Models with Gradient Flow*, arXiv:2310.19793, 2023. https://arxiv.org/abs/2310.19793.

[5] J. Bruna, L. Pillaud-Vivien, and A. Zweig, *On Single Index Models beyond Gaussian Data*, arXiv:2307.15804, 2023. https://arxiv.org/abs/2307.15804.

[6] G. Ben Arous, R. Gheissari, J. Huang, and A. Jagannath, *Local geometry of high-dimensional mixture models: effective spectral theory and dynamical transitions*, arXiv:2502.15655, version 3, 2026. https://arxiv.org/abs/2502.15655.

[7] A. Montanari and B. Saeed, *Variational Formulas for the Spectrum of Block Wishart Matrices*, arXiv:2606.27774, 2026. https://arxiv.org/abs/2606.27774.