EFFECTIVE REGULARITY: WHEN POINT ESTIMATES GENERALIZE LIKE THE BAYESIAN POSTERIOR IN SINGULAR MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Singular learning theory explains how overparameterized models can generalize well despite having many parameters. Its main guarantees, however, concern Bayesian prediction, whereas neural networks are typically trained to a single solution using gradient descent. In singular models, these two approaches need not have the same generalization behavior. We identify a class of model–data regimes, which we call *effectively regular*, in which they do. Effective regularity arises when the likelihood depends on the parameters only through a lower-dimensional function class that is locally well behaved around the true function. We prove that Bayesian prediction and maximum likelihood then have the same leading generalization error, $d_{\text{eff}}/(2n)$, where d_{eff} is the dimension of the network's function class near the truth rather than its number of parameters. We establish effective regularity for several shallow and deep linear networks and for a shallow nonlinear network. Finally, we examine empirically when gradient descent reaches the corresponding maximum-likelihood solution and therefore achieves the same rate.

1 INTRODUCTION

Deep networks are built with symmetry. Permuting hidden units, rescaling weights across a homogeneous nonlinearity, or rotating an internal representation all leave the function a network computes unchanged, so many distinct parameter settings realize the same map. The data compounds this: real targets are conjectured to have low-dimensional structure, the *manifold hypothesis* (Fefferman et al., 2016; Bengio et al., 2013), and a network only needs a low-rank slice of its weights to fit them. Both effects make the map from parameters to functions non-injective and collapse the Fisher information: the model is *singular*, not regular, in the sense of classical statistics. This degeneracy is not a technicality. It shapes what is learned and how, affecting both gradient descent dynamics and Bayesian learning.

Singular learning theory (SLT) gives the sharpest account of how this degeneracy helps, in the Bayesian setting (Watanabe, 2009). For a singular model the Bayesian generalization error scales as λ/n, where λ is a birational invariant of the model that can lie far below the regular value $d/2$ set by the parameter count.

This guarantee, however, is about the Bayesian posterior. In a singular model, point estimation and Bayesian learning are in general very different: the classical equivalence of maximum likelihood and Bayes, both attaining the rate $d/2$, holds only for *regular* models, and in singular regimes the maximum-likelihood rate can be very different from the Bayesian one. What practitioners run is different again. Gradient descent computes neither the posterior nor an exact maximum-likelihood estimator: it minimizes a non-convex loss, and whether it recovers the MLE depends on the landscape it descends. When the loss is free of bad local minima, gradient descent reaches a global minimizer and coincides with maximum likelihood; when bad local minima are present, its implicit bias can settle on parameters quite different from the MLE. Bayesian posterior, maximum likelihood, and gradient descent are thus three notions of learning that, in a singular model, need not agree.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

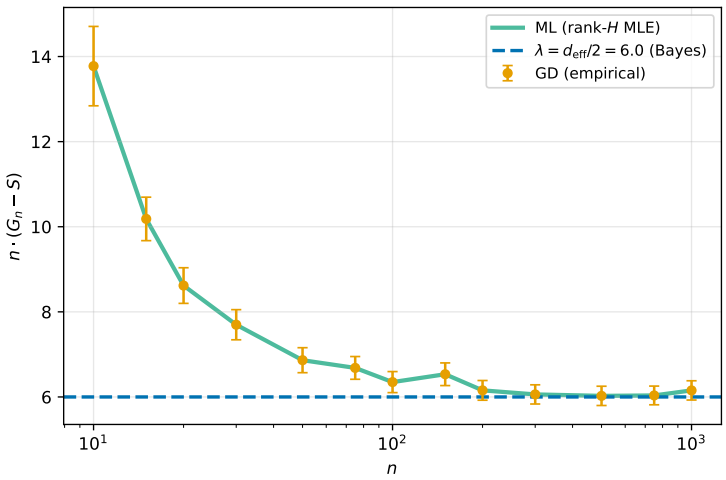


Figure 1: **The three notions of learning coincide.** Reduced-rank network with a bottleneck matched to the rank of the truth ($N = 4$, $H = 2$, $M = 4$, $r = 2$; the matched low-rank case; see the running example of Section 3 for details). We plot the rescaled excess generalization loss $n\,(G_n - S)$ against sample size n. *Green*: maximum likelihood on the realized function class. *Orange*: gradient descent on the weights (A, B) from a small random initialization. *Dashed*: the Bayesian rate $d_{\text{eff}}/2 = 6$ from the factorization principle (Theorem 1). All three curves approach the same constant.

In this work we identify a single structural feature of the model-and-data pair under which Bayesian learning and maximum likelihood coincide asymptotically, and under which gradient descent, when it reaches a global minimum of the loss, coincides with them too.

Contributions.

1. We identify a feature of the model together with the data under which Bayesian learning and maximum likelihood estimation behave the same asymptotically. Both scale like a regular model of an *effective dimension* d_{eff} that depends on both the data and the model, not on the parameter count. We call such models *effectively regular*.
2. We give many configurations with this property, across the reduced-rank network, deep linear networks, and a shallow nonlinear network, and compute their effective dimension analytically in each case.
3. We check empirically whether, in these models, gradient descent finds the maximum-likelihood solution.

2 Preliminaries

We recall the Bayesian-learning setup and the two singular-learning-theory invariants that govern its asymptotics, and then the point estimates practitioners actually compute. The treatment follows Watanabe (2009).

Data. We observe training data $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, drawn i.i.d. from a joint distribution: inputs $x_i \in \mathbb{R}^M$ from a marginal $q(x)$, and outputs $y_i \in \mathbb{R}^N$ from a conditional $q(y \mid x)$, the **true distribution**, unknown to the learner.

Model and prior. A **model** is a parameterized family of conditional densities $\{p(y \mid x, w) : w \in W \subseteq \mathbb{R}^d\}$, and a **prior** $\varphi(w)$ encodes beliefs about w before data.

Learning. Bayesian learning forms the **posterior**

$$p(w \mid D_n) = \frac{1}{Z_n} \prod_{i=1}^{n} p(y_i \mid x_i, w)\, \varphi(w), \qquad Z_n = \int_W \prod_{i=1}^{n} p(y_i \mid x_i, w)\, \varphi(w)\, dw, \tag{1}$$

with Z_n the **marginal likelihood**, and predicts with the **predictive distribution**

$$p^*(y \mid x) = \int_W p(y \mid x, w)\, p(w \mid D_n)\, dw. \tag{2}$$

Losses. Predictive quality is measured against q. The **generalization loss** and **training loss** are

$$G_n = \mathbb{E}_{(x,y)\sim q}\big[-\log p^*(y \mid x)\big], \qquad T_n = \frac{1}{n} \sum_{i=1}^{n} \big[-\log p^*(y_i \mid x_i)\big]. \tag{3}$$

The **entropy** $S = -\int q(x) \int q(y \mid x) \log q(y \mid x)\, dy\, dx$ is the irreducible loss; the excess generalization loss is a Kullback–Leibler (KL) divergence,

$$G_n - S = \mathbb{E}_{x\sim q}\big[\mathrm{KL}\left(q(y \mid x) \,\|\, p^*(y \mid x)\right)\big] \geq 0, \tag{4}$$

vanishing iff $p^* = q$. The rate at which $G_n - S \to 0$ is the central object.

Realizable case. We work in the **realizable** case: the truth lies in the family, $q(y \mid x) = p(y \mid x, w_0)$ for some w_0. The **true set** $W_0 = \{w \in W : p(y \mid x, w) = q(y \mid x) \text{ for all } x, y\}$ may contain more than one point. This non-identifiability is the defining feature of **singular** models. Equivalently, with the KL function $K(w) = \int q(x) \int q(y \mid x) \log \frac{q(y|x)}{p(y|x,w)}\, dy\, dx \geq 0$, we have $W_0 = \{K = 0\}$, and the geometry of this zero set controls learning.

Regular vs. singular. The Fisher information $I(w) = \mathbb{E}_q[\nabla_w \log p\, \nabla_w \log p^\top]$ is, in the realizable case, the Hessian of K at w_0. A model is **regular** when $I(w_0)$ is positive definite at an isolated w_0; then the posterior is asymptotically Gaussian and

$$\mathbb{E}[G_n] = S + \frac{d}{2n} + o(1/n), \qquad \mathbb{E}[G_n - T_n] = \frac{d}{n} + o(1/n), \tag{5}$$

with d the parameter count. A model is **singular** when it is not regular. This is the usual situation for neural networks, where parameter symmetries make W_0 positive-dimensional, and in this case d no longer governs the asymptotics.

The two invariants. SLT replaces d by two birational invariants of the singularity:

$$\mathbb{E}[G_n] = S + \frac{\lambda}{n} + o(1/n), \qquad \mathbb{E}[G_n - T_n] = \frac{2\nu}{n} + o(1/n). \tag{6}$$

Here λ is the **real log-canonical threshold** (RLCT), a rational number with $\lambda \leq d/2$, controlling posterior concentration and the generalization rate; ν is the **singular fluctuation**, controlling the gap between generalization and training loss. The singular fluctuation is read off the posterior **functional variance**

$$V_n = \sum_{i=1}^{n} \mathrm{Var}_{w\sim p(\cdot|D_n)}\big[\log p(y_i \mid x_i, w)\big], \qquad \lim_{n\to\infty} \mathbb{E}[V_n] = 2\nu. \tag{7}$$

For regular models $\nu = \lambda = d/2$. For singular models the two usually differ. For analytic computations of λ one usually relies on resolution of singularities, which we do not need in this paper.

Point estimates: maximum likelihood and gradient descent. The standard point estimate is the **maximum-likelihood estimator (MLE)** $\hat{w}_n = \arg\max_{w\in W} \sum_{i=1}^{n} \log p(y_i \mid x_i, w)$. For Gaussian-output models the negative log-likelihood is a squared-error loss, so **gradient descent** minimizes the empirical risk whose global minimizers are the MLE.

For **regular** models the two agree to leading order: $n\,\mathbb{E}[\mathrm{MLgen} - S] \to d/2 = \lambda$, the same constant as Bayes. For **singular** models this can fail. The ML learning coefficient may exceed the Bayesian value λ, so the same singularity that helps the Bayesian predictive can hurt the point estimate (Fukumizu, 1999).

3 A RUNNING EXAMPLE: THE SHALLOW NETWORK

To make the abstract statements of the next three sections concrete, we carry one model through all of them: the two-layer linear network, also known as reduced-rank regression (RRR).

Model. An input $X \sim \mathcal{N}(0, I_M)$ is mapped to $Y \in \mathbb{R}^N$ through a bottleneck of width H, so M is the input dimension, N the output dimension, and H the hidden dimension:

$$p(Y \mid X, A, B) = \mathcal{N}\Big(\tfrac{1}{c}\, BAX,\ I_N\Big), \qquad A \in \mathbb{R}^{H\times M},\ B \in \mathbb{R}^{N\times H}, \tag{8}$$

with i.i.d. standard-normal priors on the entries of A, B and a coupling constant $c > 0$. The truth is $q(Y \mid X) = \mathcal{N}(W_0 X, I_N)$ with $\operatorname{rank}(W_0) = r \le H$. The likelihood depends on (A, B) only through the **composite weight** $W = BA/c \in \mathbb{R}^{N\times M}$, of rank at most H, so p depends on the parameters only through $\Phi(B, A) = BA/c$. The model is singular through the **gauge symmetry** $(A, B) \mapsto (G^{-1}A, BG)$, $G \in GL(H)$, which preserves W and hence the full likelihood. Because $K(w)$ is constant along each gauge orbit, it has degenerate directions there and its Hessian loses rank.

The running example. We use the **matched low-rank regime** ($r = H < \min(M, N)$: a narrow bottleneck matched to a truth of exactly that rank) as the running example through the next three sections. There the network realizes maps of rank at most H, generically exactly H, so the function class is the smooth rank-H matrix manifold $\mathcal{R}_H = \{W : \operatorname{rank} W = H\}$, of dimension $d_H = H(N + M - H)$, far below the parameter count $H(N + M)$. Figure 1 verifies the generalization side of the prediction in this example, and Appendix F verifies the other invariant, measuring the singular fluctuation directly from the posterior functional variance.

4 THE FACTORIZATION PRINCIPLE

Effective regularity has a simple explanation. The symmetry directions of a network are the directions in which the parameters change without changing the function the network computes. If we quotient these directions out, what is left is a model over the space of functions. When that smaller model is an ordinary regular one, we expect the original model to behave like a regular model of the smaller dimension. The rest of this section makes the statement precise and points to the one assumption that can break it.

Setup. Let $\Phi\colon \Theta \to \mathcal{M}$ be a smooth map from the parameter space into a lower-dimensional space $\mathcal{M}$ of dimension $d_{\text{eff}} = \dim \mathcal{M}$, and let $\{\bar{p}(\cdot \mid m) : m \in \mathcal{M}\}$ be a family of output densities, the **target model** (the model in function space). The prior φ on Θ induces a **pushforward prior** on $\mathcal{M}$, $\bar{\varphi}(m) = \int_{\Phi^{-1}(m)} \varphi(w)\, dw_{\text{fiber}}$, integrating over the fiber $\Phi^{-1}(m)$.

Theorem 1 (Factorization principle)**.** *Suppose*

- *(i)* (factorization) *the likelihood factors through Φ, that is $p(y \mid w) = \bar{p}\big(y \mid \Phi(w)\big)$;*
- *(ii)* (regular target) *the target model is regular at $m_0 = \Phi(w_0)$: its Fisher information is positive definite at m_0, and m_0 is the unique global minimum of the population loss on $\mathcal{M}$; and*
- *(iii)* (benign symmetry) *the pushforward prior $\bar{\varphi}$ is continuous and strictly positive at m_0.*

Then $\nu = \lambda = d_{\text{eff}}/2$, with $d_{\text{eff}} = \dim \mathcal{M}$. When $\mathcal{M}$ is a smooth manifold rather than an open subset of $\mathbb{R}^{d_{\text{eff}}}$, all conditions are read in a local chart at m_0, and are independent of the chart.

The proof is in Appendix A.

Definition 1 (Effectively regular model)**.** A model is **effectively regular** if its likelihood factors through a smooth map $\Phi\colon \Theta \to \mathcal{M}$ onto a lower-dimensional space on which the hypotheses of the factorization principle (Theorem 1) hold at the truth. Its **effective dimension** is the dimension of that target space,

$$d_{\text{eff}} := \dim \mathcal{M},$$

which does not depend on the choice of Φ: the invariant λ fixes $d_{\text{eff}} = 2\lambda$.

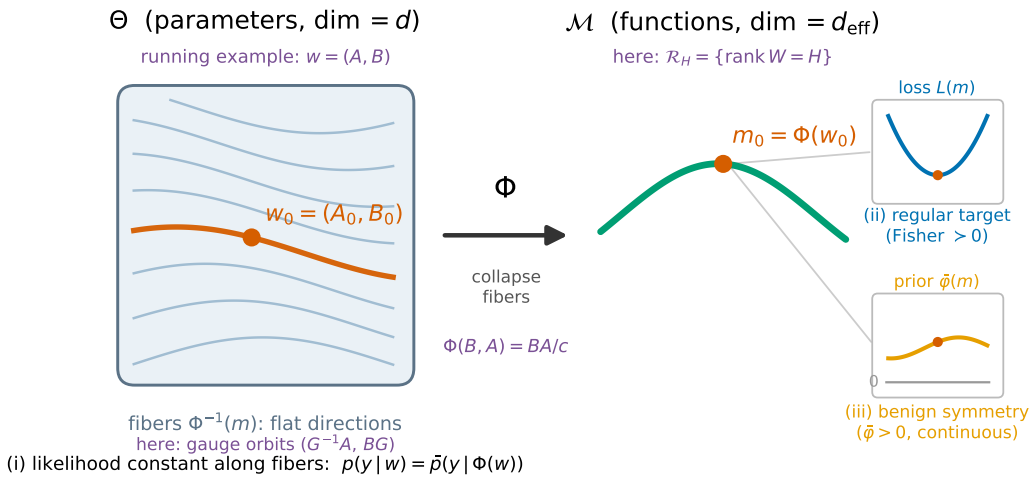


Figure 2: Illustration of the factorization principle.

Remark 1. Theorem 1 gives the regular scaling $\nu = \lambda = d_{\text{eff}}/2$, but with the effective dimension $d_{\text{eff}} = \dim \mathcal{M}$, the dimension of the function class, in place of the parameter count $\dim \Theta$.

Remark 2 (Running example). For the running example we check the three conditions on the **factorization map** $\Phi(B, A) = BA/c$, whose target space is the rank-H manifold $\mathcal{M} = \mathcal{R}_H$ of dimension $d_H = H(N + M - H)$. Condition (i) is clear: the likelihood depends on (B, A) only through $W = BA/c = \Phi(B, A)$. Condition (ii) holds intuitively because, viewed in the composite parameter W, every change of W changes the model and moves it off the zero-loss truth, so the Fisher information on $\mathcal{R}_H$ is positive definite at m_0. Condition (iii), and the rigorous verification of condition (ii), are deferred to Section 7.

5 Maximum Likelihood in Effectively Regular Regimes

The factorization principle (Theorem 1) shows that the Bayesian generalization loss converges at rate $\lambda/n = d_{\text{eff}}/(2n)$ in the effectively regular regimes. We now show that the maximum likelihood estimator achieves the same rate, and in particular that the correct notion of "ML estimator" is determined by the factorization.

Theorem 2 (ML generalization in effectively regular regimes)**.** *Suppose the model factorizes through $\Phi\colon \Theta \to \mathcal{M}$ with regular target model at $m_0 = \Phi(w_0)$, in the sense of Theorem 1. Let $\hat{m}_n$ be the maximum likelihood estimator in $\mathcal{M}$, and define $\hat{w}_n$ to be any preimage $\hat{w}_n \in \Phi^{-1}(\hat{m}_n)$. Then*

$$n \cdot \mathbb{E}\big[\mathrm{ML}_{\text{gen}}(\hat{w}_n) - S\big] \;\to\; \frac{d_{\text{eff}}}{2}, \qquad d_{\text{eff}} = \dim(\mathcal{M}).$$

Proof. **Step 1: the MLE factorizes.** Since $p(y_i \mid w) = \bar{p}(y_i \mid \Phi(w))$, the log-likelihood depends on w only through $m = \Phi(w)$:

$$\sup_{w \in \Theta} \sum_{i=1}^{n} \log p(y_i \mid w) = \sup_{m \in \Phi(\Theta)} \sum_{i=1}^{n} \log \bar{p}(y_i \mid m).$$

Any maximizer $\hat{w}_n$ projects to $\hat{m}_n = \Phi(\hat{w}_n)$, which maximizes the target likelihood. The generalization loss depends only on $\hat{m}_n$, because the KL divergence satisfies $K(w) = \bar{K}(\Phi(w))$.

Step 2: regular ML theory. Since the target model is regular at m_0 with Fisher information $I(m_0)$ positive definite, classical ML theory (van der Vaart, 1998) gives $n \cdot \mathbb{E}[\mathrm{ML}_{\text{gen}}(\hat{m}_n) - S] \to d_{\text{eff}}/2$. □

Remark 3 (Running example). For the shallow network the target maximum likelihood estimator looks for solutions over the rank-H manifold $\mathcal{R}_H = \{W : \operatorname{rank} W = H\}$: it is the rank-$H$ constrained least-squares fit of W.

6 THE CONNECTION TO GRADIENT DESCENT

Theorem 2 concerns the maximum likelihood estimator, defined by a global optimization that, in a singular model, might range over a non-convex parameter space. A practitioner does not solve that optimization in closed form; they run gradient descent on a loss.

The loss. Take the loss to be the empirical negative log-likelihood

$$\ell_n(w) \;=\; -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid w), \tag{9}$$

minimized by full-batch gradient descent over $w \in \Theta$ from a small random initialization. For any model with Gaussian output noise, $Y = f(x, w) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_N)$, this is, up to an additive constant, the squared-error loss $\frac{1}{2\sigma^2 n}\sum_i \|y_i - f(x_i, w)\|^2$ trained over the parameters; the running example (8) is the case $f(x, w) = BAx/c$.

The global minimizers coincide. Gradient descent runs on the original parameters w, not on m. But the loss factors through Φ, $\ell_n(w) = \bar{\ell}_n(\Phi(w))$ with $\bar{\ell}_n(m) = -\frac{1}{n}\sum_i \log \bar{p}(y_i \mid m)$, so (as in the proof of Theorem 2) the two models share a global minimizer: every global minimizer $\hat{w}_n$ of ℓ_n projects to the target-space maximum likelihood estimator $\hat{m}_n = \Phi(\hat{w}_n)$. Hence, if it reaches a global minimum, gradient descent returns exactly the estimator of Theorem 2 and inherits the rate $d_{\text{eff}}/2 = \lambda$.

Whether gradient descent reaches it. Whether the search attains that global minimum is a landscape-dependent optimization question outside the scope of this paper, which we check empirically. In some of the effectively regular configurations gradient descent from a small random initialization always reached the global minimizer; in others, whose landscapes carry local minima, it reached the minimizer only on a fraction of the runs. Figure 1 shows this for the running example. Thorough investigations are in Appendix E.

7 EFFECTIVE REGULARITY IN CONCRETE NETWORKS

We now apply Theorem 1 to three model families: the shallow linear network, its deep linear generalization, and a genuinely nonlinear network. The recipe is the same each time. Condition (i) is free: we let Φ send the weights to the function the network actually computes, the composite weight BA/c for the linear networks and a symmetric matrix for the quadratic network of Section 7.3, so the likelihood sees the weights only through that function. What is left is to check that, viewed in the space of functions, the truth is a well-behaved point: the loss should form a genuine bowl around it (condition (ii)), and the prior, once carried over to function space, should stay smooth and spread out there rather than piling up on the truth (condition (iii)).

Two routes to effective regularity. Across all of our networks this happens in one of two ways, and each secures both conditions at once.

- **Matched rank.** The width is tuned to the truth: the network has just enough capacity to express it and no more, so the truth sits on a smooth part of the function class and the map from weights to functions covers its neighborhood cleanly. That clean geometry is what shapes the loss into a bowl and, at the same time, carries the prior over smoothly, so (ii) and (iii) are two faces of a single fact.
- **Saturated.** The width is far larger than needed. The function class now fills out the whole flat space of functions, where even a simple, low-rank truth is an ordinary interior point with a bowl around it (that is (ii)); and the very wide layer smears the network's random function into a smooth, spread-out density instead of concentrating it on the truth (that

is (iii)). Here the two conditions hold for different reasons, but both trace back to one source, an oversized hidden layer.

Every effectively regular configuration below is one of these two, and in each the point estimate inherits the Bayesian rate, $\nu = \lambda = d_{\text{eff}}/2$ with d_{eff} the dimension of the function class.

7.1 THE SHALLOW NETWORK

We take the shallow network of Section 3. The likelihood factors through $\Phi(B, A) = BA/c$.

The pair (H, r) divides the shallow network into five regimes, which we label A–E and which are tabulated in full in Appendix B.1; three of the five are effectively regular, and it is those three that follow. The remaining two, B and E, are not, and we return to B in Section 8.

Proposition 1 (Shallow network)**.** *The shallow network is effectively regular, with the stated target space $\mathcal{M}$ and effective dimension d_{eff}, in each of the following cases:*

(A) full rank: $H \geq \min(N, M)$ *and* $r = \min(N, M)$*, with* $\mathcal{M} = \mathbb{R}^{N \times M}$ *and* $d_{\text{eff}} = NM$*;*

(D) matched low rank: $H < \min(N, M)$ *and* $r = H$*, with* $\mathcal{M} = \mathcal{R}_H$ *and* $d_{\text{eff}} = d_H = H(N + M - H)$*;*

(C) saturated: $r < \min(N, M)$ *and* $H \geq N + M - r$*, with* $\mathcal{M} = \mathbb{R}^{N \times M}$ *and* $d_{\text{eff}} = NM$.

Cases (A) and (D) are the matched-rank route: the bottleneck carries exactly the rank of the truth, Φ is a submersion onto the realized function class, and the truth sits on a smooth part of it (the whole matrix space when H is wide, the rank-H manifold $\mathcal{R}_H$ when it is narrow). Case (C) is the saturated route: the truth is genuinely low rank and Φ is no longer a submersion, but the very wide hidden layer smears the induced distribution on functions into a smooth density, and the model stays regular on the full matrix space. The proofs are in Appendix B.

7.2 THE SAME PICTURE IN DEEP LINEAR NETWORKS

The shallow network has one hidden layer. Replacing it by a stack of linear layers changes nothing essential: the effective dimension is set by the narrowest layer, and the three effectively regular regimes survive.

An L-layer linear network writes the composite weight as a product of factor matrices,

$$W = \tfrac{1}{c}\, B_L B_{L-1} \cdots B_1, \qquad B_\ell \in \mathbb{R}^{H_\ell \times H_{\ell-1}},$$

with $H_0 = M$, $H_L = N$, and an i.i.d. Gaussian prior on every entry of every factor. The likelihood sees the factors only through their product W, so the factorization principle applies once we read off the target space. The one new quantity is the **bottleneck width**

$$H_* = \min_{1 \leq \ell \leq L-1} H_\ell.$$

A product threaded through a layer of width H_* has rank at most H_*, and every matrix of rank at most H_* factors this way, so the realized function class is the same set of bounded-rank maps as before, with the single hidden width H replaced by H_*. The regime conditions and the targets carry over verbatim under this replacement.

Proposition 2 (Three regimes, deep)**.** *Let H_* be the bottleneck width. In each of the following cases the deep linear network is effectively regular, with the stated effective dimension:*

(A) *if* $H_* \geq \min(N, M)$ *and* $r = \min(N, M)$*, then* $d_{\text{eff}} = NM$ *and* $\nu = \lambda = NM/2$*;*

(D) *if* $H_* < \min(N, M)$ *and* $r = H_*$*, then* $d_{\text{eff}} = d_{H_*} = H_*(N + M - H_*)$ *and* $\nu = \lambda = d_{H_*}/2$*;*

(C) *if* $r < \min(N, M)$ *and* $H_* \geq N + M - r$*, then* $d_{\text{eff}} = NM$ *and* $\nu = \lambda = NM/2$.

The proofs are in Appendix C.

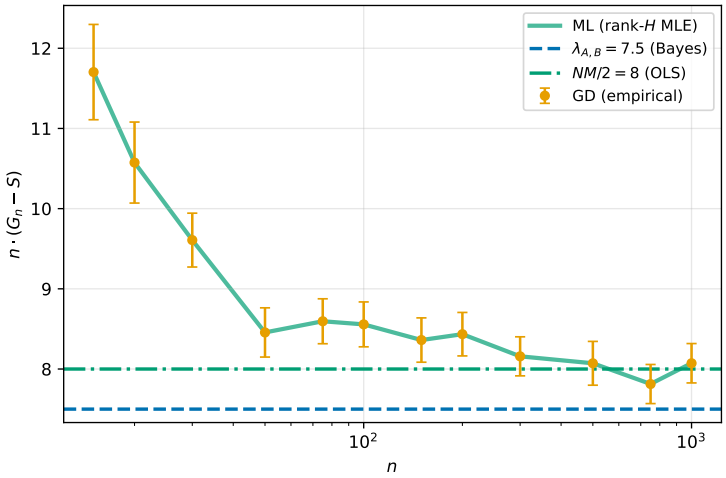


Figure 3: **The point estimate diverges from the Bayesian rate.** Shallow linear network in the singular regime B ($N = 4$, $H = 4$, $M = 4$, $r = 2$: $H \geq \min(N, M)$, $r < \min(N, M)$, $H < N+M-r$); colors as in Figure 1. The pushforward prior diverges at the truth, so condition (iii) fails and the Bayesian rate drops to $\lambda_{A,B} = 7.5$; maximum likelihood and gradient descent instead plateau at the regular OLS rate $NM/2 = 8$, strictly above it. (This configuration sits two units below the saturation threshold $N + M - r = 6$, so $H = 4 \leq N + M - r - 2 = 4$ and the gap is strict.)

7.3 Beyond linearity: a nonlinear network

Everything so far has been linear in the input. One might worry that effective regularity is a feature of linear networks. It is not. We make this concrete with a genuinely nonlinear model, a one-hidden-layer network with a quadratic activation, and show that the same two routes to effective regularity reappear.

Fix $H, M \geq 1$. With $X \sim \mathcal{N}(0, I_M)$ and scalar noise, the quadratic network

$$Y = \sum_{h=1}^{H} a_h \, (b_h^\top X)^2 + \varepsilon, \qquad w = (a, B) \in \mathbb{R}^H \times \mathbb{R}^{H \times M},$$

is nonlinear in X. Yet each hidden unit contributes a quadratic form $a_h(b_h^\top X)^2 = X^\top(a_h b_h b_h^\top)X$, so the network output depends on the weights only through the symmetric matrix

$$G = \sum_{h=1}^{H} a_h \, b_h b_h^\top \; \in \; \mathrm{Sym}^M.$$

The factorization map is $\Phi(a, B) = G$ and the target model is $Y = X^\top G X + \varepsilon$ on a space of symmetric matrices.

Proposition 3 (Low-rank analog)**.** *Let $r = H \leq M$ and let the truth be generic of rank exactly r (all $a_h \neq 0$, the rows of B linearly independent). Then the quadratic network is effectively regular with effective dimension $d_{\text{eff}} = \dfrac{H(2M - H + 1)}{2}$.*

Proposition 4 (Saturated analog)**.** *Let $D = \binom{M+1}{2} = \dim \mathrm{Sym}^M$ and $H \geq 3D$. Then for* every *truth $G^* \in \mathrm{Sym}^M$, of any rank including zero, the quadratic network is effectively regular with effective dimension $d_{\text{eff}} = \binom{M+1}{2}$.*

The proofs are in Appendix D.

8 When the factorization principle fails

Effective regularity is what makes a point estimate inherit the Bayesian rate, and it is not automatic. The whole argument runs through Theorem 1. When it fails, the two rates might part.

The cleanest instance is the linear network with $H \geq \min(N, M)$, $r < \min(N, M)$, and $H < N + M - r$: a low-rank truth with a hidden layer too narrow to saturate. The likelihood still factors through $\Phi(B, A) = BA/c$ onto the full target space $\mathcal{M} = \mathbb{R}^{N \times M}$, and the target model there is regular: condition (ii) holds, exactly as in the full-rank and saturated cases. What fails is condition (iii): the *pushforward prior diverges* at the truth W_0 (Appendix B.5).

This very divergence is what lowers the Bayesian RLCT below the regular value: the closed form of Aoyagi & Watanabe (2005) gives $\lambda \leq NM/2$ throughout Regime B, with equality exactly at $H = N+M-r-1$, one unit short of saturation, and strict inequality whenever $H \leq N+M-r-2$. Away from that boundary, Bayesian learning profits from the singularity. The point estimate does not. Since $H \geq \min(N, M)$ the rank constraint is vacuous, the loss minimizer is the ordinary least squares estimator, and gradient descent inherits its *regular* rate $NM/2$.

9 Related work

Quasi-regularity. The closest notion in singular learning theory is Watanabe's *quasi-regular* model (Watanabe, 2009), which shares with the effectively regular case the identity $\nu = \lambda$. It is nonetheless a different condition: the multiplicity of the RLCT is generally not 1, so the model remains genuinely singular, and maximum likelihood need not track the Bayesian posterior. Effective regularity differs from it: Step 3 of the proof of Theorem 1 gives multiplicity exactly 1.

Learning coefficients of linear networks. For the shallow network, Aoyagi & Watanabe (2005) give λ in closed form at every (N, H, M, r), and Aoyagi (2025) extends the computation to deep linear networks. That line of work computes λ; what effective regularity adds is the second invariant, $\nu = \lambda$, and with it the point estimate.

Point estimates versus Bayes. The generalization of point estimators in linear networks has a longer history. Fukumizu (1999) derives the generalization error of the rank-constrained maximum-likelihood estimator in unidentifiable cases, the estimator we recover in the matched-low-rank regime, and shows it can exceed the Bayesian rate: the point estimate is *penalized* by the rank deficiency rather than aided by it. Nakajima & Watanabe (2005; 2007) analyze the variational and empirical-Bayes solutions of the same networks, a third family of estimators that likewise departs from full Bayes in the singular regimes. Our contribution is complementary: we isolate the condition, effective regularity, under which no such gap opens.

Measuring λ empirically. A parallel line estimates the local learning coefficient of trained networks by sampling (Lau et al., 2023). The effectively regular families here are a natural test bed for such estimators, since λ is known in closed form and equals $d_{\text{eff}}/2$.

10 Discussion

We identified a structural feature of the model-and-data pair, effective regularity, under which Bayesian learning and maximum-likelihood estimation behave the same asymptotically, both governed by an effective dimension that depends on the data and the model rather than on the parameter count, and we checked empirically whether gradient descent finds the maximum-likelihood estimator there.

Several questions are left open. Our account of gradient descent is empirical: it would be interesting to understand analytically how the estimator gradient descent reaches differs from the maximum-likelihood estimator, and whether the gap vanishes or merely stays small. Beyond the linear and quadratic networks treated here, one would like to identify the effectively regular regimes in deep nonlinear models. And more broadly, we found two mechanisms that drive a model into effective regularity, matched rank and saturation; whether these exhaust the possibilities, or whether other structural features can produce the same collapse onto a regular model on function space, remains open.

AI USE STATEMENT

In this work, we used generative AI tools (ChatGPT and Claude) throughout the research process. Specifically, we used them for tasks whose disclosure ICLR requires: developing the theoretical models and conceptual framework, formulating and refining the mathematical claims, supplying critical ingredients for proofs, assisting in the writing of the proofs, proposing and refining hypotheses, designing the numerical experiments and providing feedback on the methodology, implementing the methods, generating the synthetic data sets used in every experiment, preparing and reformatting the resulting data, and interpreting the results. We additionally used them for tasks whose disclosure is recommended: drafting and editing parts of the manuscript for readability, creating and modifying figures, writing and editing code, searching for and summarizing related literature, and formatting references.

We have reviewed all AI-assisted work. Every theorem statement and proof was checked line by line by the authors, and the derivations were independently confirmed against the numerical experiments reported in the paper; AI-generated code was read, tested, and its output cross-checked against analytical predictions wherever a closed form was available; the literature identified with AI assistance was verified by reading the cited sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is theoretical and does not involve human subjects, personal data, or deployed systems. We do not anticipate direct negative societal impacts, though future applications of the methods studied here may require separate ethical evaluation.

REPRODUCIBILITY STATEMENT

All implementation details are given in Appendices E and F. The anonymous supplementary archive contains the code and data reproducing every experiment, figure and table in the paper: the experiment scripts that generate the data, the figure scripts that draw from it, the resulting data files, and the shared model classes they import. The README maps each figure and table to the script and data file it comes from. All data is synthetic and generated inside the scripts, every run is seeded, and everything is CPU-only, so the reported numbers reproduce exactly.

REFERENCES

Miki Aoyagi. Singular learning coefficients and efficiency in learning theory. *arXiv preprint arXiv:2501.12747*, 2025.

Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18(7):924–933, 2005.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Anthony Carbery and James Wright. Distributional and l^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8(3):233–248, 2001.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Kenji Fukumizu. Generalization error of linear neural networks in unidentifiable cases. In *Algorithmic Learning Theory (ALT 1999)*, volume 1720 of *Lecture Notes in Computer Science*, pp. 51–62. Springer, 1999.

Jesper R. Ipsen and Mario Kieburg. Weak commutation relations and eigenvalue statistics for products of rectangular random matrices. *Physical Review E*, 89(3):032106, 2014.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Edmund Lau, Zach Furman, George Wang, Daniel Murfet, and Susan Wei. The local learning coefficient: A singularity-aware complexity measure. *arXiv preprint arXiv:2308.12108*, 2023.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, 2nd edition, 2013.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, 2005.

Shinichi Nakajima and Sumio Watanabe. Generalization error of linear neural networks in an empirical Bayes approach. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pp. 804–810, 2005.

Shinichi Nakajima and Sumio Watanabe. Variational Bayes solution of linear neural networks and its generalization performance. *Neural Computation*, 19(4):1112–1153, 2007.

Fedor Nazarov, Mikhail Sodin, and Alexander Volberg. The geometric Kannan–Lovász–Simonovits lemma, dimension-free estimates for the distribution of the values of polynomials, and the distribution of the zeros of random analytic functions. *St. Petersburg Mathematical Journal*, 14(2): 351–366, 2003.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

APPENDIX

Standing assumptions. The results of SLT require certain regularity conditions on the model and the true distribution, collected in Watanabe's "fundamental conditions" (Watanabe, 2009). In brief: the parameter space W is compact, the log-likelihood ratio $\log(q(x)/p(x \mid w))$ is real analytic in w, and the true distribution has finite moments. The models in this paper have non-compact parameter spaces and Gaussian priors, so the compactness condition is not literally satisfied. However, the Gaussian prior decays fast enough that the contribution from $\|w\| \to \infty$ is negligible, and the asymptotic formulas are widely believed to extend to this setting, though we are not aware of a published proof (Watanabe treats Gaussian priors this way in numerical work (Watanabe, 2010)). All proofs below are stated under these conditions.

A PROOF OF THE FACTORIZATION PRINCIPLE (THEOREM 1)

Write $f_i(w) = \log p(y_i \mid w)$ for the log-likelihood of observation i, and Π_n for the posterior on Θ.

Step 1: variance identity. Since $p(y_i \mid w) = \bar{p}(y_i \mid \Phi(w))$, the log-likelihood depends on w only through $m = \Phi(w)$: with $\bar{f}_i(m) = \log \bar{p}(y_i \mid m)$ we have $f_i = \bar{f}_i \circ \Phi$. Apply the law of total variance under Π_n with m as the conditioning variable:

$$\mathrm{Var}_{\Pi_n}(f_i) = \mathbb{E}_{\bar{\Pi}_n}\big[\mathrm{Var}_{\Pi_n}(f_i \mid m)\big] + \mathrm{Var}_{\bar{\Pi}_n}\big[\mathbb{E}_{\Pi_n}(f_i \mid m)\big], \tag{10}$$

where $\bar{\Pi}_n$ is the pushforward of Π_n to $\mathcal{M}$. Because f_i is constant on each fiber $\Phi^{-1}(m)$, the conditional variance vanishes and the conditional expectation is $\bar{f}_i(m)$. Hence, summing over i,

$$V_n^{\Theta} := \sum_{i=1}^{n} \mathrm{Var}_{\Pi_n}\big(f_i(w)\big) = \sum_{i=1}^{n} \mathrm{Var}_{\bar{\Pi}_n}\big(\bar{f}_i(m)\big) =: V_n^{\mathcal{M}}. \tag{11}$$

The functional variance of the singular model equals that of the target model exactly.

Step 2: regularity of the pushforward posterior. The pushforward posterior $\bar{\Pi}_n$ is the posterior of the target model $\bar{p}$ under the prior $\bar{\varphi}$. By condition (ii) the target model is regular at m_0, and by condition (iii) $\bar{\varphi}$ is continuous and strictly positive at m_0. The regular case of Watanabe's asymptotic theory therefore applies to the target model (Watanabe, 2009): its posterior is asymptotically Gaussian with covariance $n^{-1}I(m_0)^{-1}$, its RLCT is $\lambda_{\mathcal{M}} = d_{\mathrm{eff}}/2$, and its singular fluctuation takes the regular value $\nu_{\mathcal{M}} = d_{\mathrm{eff}}/2$, that is $\mathbb{E}[V_n^{\mathcal{M}}] \to d_{\mathrm{eff}}$.

Step 3: the RLCT transfers. The RLCT is read off the small-temperature partition function $Z_n^0 = \int_\Theta e^{-nK(w)}\varphi(w)\,dw$, whose asymptotic expansion $Z_n^0 = n^{-\lambda_\Theta}(\log n)^{k_\Theta - 1}(c + o(1))$ defines λ_Θ and its multiplicity k_Θ. Because $K = \bar{K} \circ \Phi$, the integrand is constant on the fibers of Φ, so the very change of variables that defines the pushforward prior gives the *exact* identity

$$\int_\Theta e^{-nK(w)}\varphi(w)\,dw \;=\; \int_{\mathcal{M}} e^{-n\bar{K}(m)}\bar{\varphi}(m)\,dm, \tag{12}$$

valid for every n and not merely asymptotically. The two partition functions are the same function of n, hence have the same expansion: $\lambda_\Theta = \lambda_{\mathcal{M}}$ and $k_\Theta = k_{\mathcal{M}}$. By condition (ii) the target model is regular at m_0 and by condition (iii) $\bar{\varphi}$ is continuous and strictly positive there, so Laplace's method applies to the right-hand side of (12) and gives $\lambda_{\mathcal{M}} = d_{\mathrm{eff}}/2$ with $k_{\mathcal{M}} = 1$. Hence $\lambda_\Theta = d_{\mathrm{eff}}/2$, with multiplicity one.

Step 4: combine. From Steps 1 and 2,

$$\nu_\Theta = \tfrac{1}{2}\lim_{n\to\infty} \mathbb{E}\big[V_n^{\Theta}\big] = \tfrac{1}{2}\lim_{n\to\infty} \mathbb{E}\big[V_n^{\mathcal{M}}\big] = \frac{d_{\mathrm{eff}}}{2}, \tag{13}$$

and with Step 3, $\nu = \lambda = d_{\mathrm{eff}}/2$.

The manifold case. When $\mathcal{M}$ is a manifold, Steps 1–4 are read in a local chart at m_0; the Fisher information is the $d_{\text{eff}} \times d_{\text{eff}}$ matrix expressed in that chart. The conclusion is independent of the chart. Two charts about m_0 differ by a diffeomorphism with Jacobian J that is invertible at m_0 and smooth and nonvanishing nearby, so: the Fisher information transforms as $I \mapsto J^\top I J$, preserving positive-definiteness in condition (ii); the pushforward density transforms by the factor $|\det J|$, preserving continuity and strict positivity in condition (iii); and $d_{\text{eff}} = \dim \mathcal{M}$ is intrinsic. Condition (i) is coordinate-free.

B PROOFS FOR THE SHALLOW NETWORK (RRR)

Throughout, $\Phi(B, A) = BA/c$; the target space is $\mathbb{R}^{N\times M}$ in the full-rank (Regime A) and saturated (Regime C) cases and the rank-H manifold $\mathcal{R}_H$ in the matched low-rank case (Regime D), and we verify conditions (ii) target-model regularity and (iii) positivity of $\bar{\varphi}$.

B.1 THE FIVE REGIMES OF THE SHALLOW NETWORK

The pair (H, r) splits the shallow network into five regimes, labelled A–E throughout the paper. Three are effectively regular and are proved so below; the other two are not, for two different reasons, and are the subject of Section 8. The regimes are exhaustive and mutually exclusive: $H \geq \min(N, M)$ splits into A ($r = \min(N, M)$), C and B (both $r < \min(N, M)$, separated by $H \geq N + M - r$ or not), and $H < \min(N, M)$ splits into D ($r = H$) and E ($r < H$).

	Condition on (H, r)	Target $\mathcal{M}$	d_{eff}	Eff. reg.
A	$H \geq \min(N, M)$, $r = \min(N, M)$	$\mathbb{R}^{N\times M}$	NM	yes
C	$r < \min(N, M)$, $H \geq N + M - r$	$\mathbb{R}^{N\times M}$	NM	yes
D	$H < \min(N, M)$, $r = H$	$\mathcal{R}_H$	$H(N + M - H)$	yes
B	$H \geq \min(N, M)$, $r < \min(N, M)$, $H < N + M - r$	$\mathbb{R}^{N\times M}$	—	no: (iii) fails
E	$r < H < \min(N, M)$	$\{\text{rank} \leq H\}$	—	no: no smooth $\mathcal{M}$

The two failures are of different kinds. In Regime B the target space is the full matrix space and the target model there is perfectly regular, so condition (ii) holds; what fails is condition (iii), the pushforward prior diverging at W_0 (Appendix B.5). In Regime E the truth has rank r strictly below the bottleneck H, so W_0 lies in the singular locus of $\text{im}\,\Phi = \{\text{rank} \leq H\}$: that set is not a manifold at W_0, and there is no smooth target on which condition (ii) could be read in the first place. In both, the Bayesian λ falls below the effectively regular value that the corresponding smooth target would give, while the point estimate does not follow it down. In Regime E the drop is always strict, $\lambda < d_H/2$. In Regime B one has $\lambda \leq NM/2$, strictly whenever $H \leq N + M - r - 2$ but with equality on the boundary $H = N + M - r - 1$: there the Bayesian and least-squares constants happen to agree even though condition (iii) fails, so Theorem 1 does not apply and the agreement is a coincidence of the leading constant rather than effective regularity. Regime B is the cleaner of the two and is the one we analyse and plot; Regime E we do not treat.

B.2 REGIME A: Φ IS A SUBMERSION ONTO $\mathbb{R}^{N\times M}$

Assume $H \geq \min(N, M)$ and $r = \min(N, M)$. Target space $\mathcal{M} = \mathbb{R}^{N\times M}$, $d = NM$, target model $Y = WX + \eta$ with $W \in \mathbb{R}^{N\times M}$.

Condition (ii). The Fisher information at W_0 is the $NM \times NM$ matrix $\mathcal{I}_{ij} = \mathbb{E}_X[(\partial_i W \cdot X)^\top (\partial_j W \cdot X)] = \text{tr}(\partial_i W\, \partial_j W^\top)$, using $\mathbb{E}[XX^\top] = I_M$. On a tangent vector $v \in \mathbb{R}^{N\times M}$ the quadratic form is $v^\top \mathcal{I} v = \|\partial_v W\|_F^2 = \|v\|_F^2$, since the parameterization $W \mapsto W$ is the identity. This is strictly positive for $v \neq 0$, so $\mathcal{I} \succ 0$.

Condition (iii). The differential of Φ at (B_0, A_0) is $d\Phi_{(B_0,A_0)}(\dot{B}, \dot{A}) = \frac{1}{c}(\dot{B}A_0 + B_0\dot{A})$. We show it is surjective onto $\mathbb{R}^{N\times M}$.

- *Case $M \le N$:* then $r = M$ and $\operatorname{rank}(A_0) = M$. Let $L = (A_0^\top A_0)^{-1} A_0^\top$ be a left inverse, $LA_0 = I_M$. For any target $\dot W$ set $\dot B = c\dot W L$, $\dot A = 0$; then $d\Phi(\dot B, \dot A) = \dot W L A_0 = \dot W$.
- *Case $N < M$:* then $r = N$ and $\operatorname{rank}(B_0) = N$. Let R be a right inverse, $B_0 R = I_N$. Set $\dot B = 0$, $\dot A = cR\dot W$; then $d\Phi(\dot B, \dot A) = B_0(cR\dot W)/c = \dot W$.

In both cases $d\Phi_{(B_0,A_0)}$ is surjective, so Φ is a submersion at (B_0, A_0). By the submersion theorem there are local coordinates (W, γ) with $\gamma \in \mathbb{R}^{H(N+M)-NM}$ the fiber coordinates and a diffeomorphism Ψ with $\Phi(\Psi(W,\gamma)) = W$, in which $\bar\varphi(W) = \int_\Gamma \varphi(\Psi(W,\gamma))\,|\det J(W,\gamma)|\,d\gamma$. The Gaussian prior φ is smooth and strictly positive, J is smooth and non-degenerate, and the Gaussian decay along the non-compact fiber Γ ensures convergence, so $\bar\varphi$ is finite, smooth, and strictly positive near W_0. Conditions (i)–(iii) hold; Theorem 1 with $d_{\text{eff}} = NM$ gives $\nu = \lambda = NM/2$.

B.3 Regime D: Φ is a submersion onto $\mathcal{R}_H$

Assume $H < \min(N, M)$ and $r = H$. Since $W = BA/c$ has rank at most H, the target space is the rank-H manifold $\mathcal{R}_H = \{W \in \mathbb{R}^{N\times M} : \operatorname{rank} W = H\}$, a smooth submanifold of dimension $d_H = H(N + M - H)$. The locus where BA drops rank is cut out by the vanishing of the $H \times H$ minors, hence is a proper algebraic subset of Θ and has prior probability zero; we restrict Φ to its complement, on which Φ takes values in $\mathcal{R}_H$. Every integral below is an integral against the prior, so none of them is affected.

Local chart. Partition $W = \left(\begin{smallmatrix} C & D \\ E & F \end{smallmatrix}\right)$ with $C \in \mathbb{R}^{H\times H}$. After a change of basis take $W_0 = \left(\begin{smallmatrix} I_H & 0 \\ 0 & 0 \end{smallmatrix}\right)$; near W_0 the block C is invertible and the Schur complement identity gives $\operatorname{rank}(W) = H + \operatorname{rank}(F - EC^{-1}D)$, so $W \in \mathcal{R}_H$ iff $F = EC^{-1}D$. The map $(C, D, E) \mapsto W(C, D, E) = \left(\begin{smallmatrix} C & D \\ E & EC^{-1}D \end{smallmatrix}\right)$ is a smooth chart of dimension $H^2 + H(M-H) + (N-H)H = H(N+M-H) = d_H$.

Condition (ii). In this chart the target model is $Y = W(C, D, E)X + \eta$, with Fisher information $\mathcal{I}_{ij} = \operatorname{tr}(\partial_i W\, \partial_j W^\top)$. The top-left, top-right, and bottom-left blocks depend linearly on (C, D, E), with derivatives v_C, v_D, v_E; for the bottom-right block $F = EC^{-1}D$ the product rule gives $\partial_v F = v_E C^{-1}D + E\,\partial_v(C^{-1})D + EC^{-1}v_D$, every term carrying a factor D or E. At W_0, $D = 0$ and $E = 0$, so $\partial_v F|_{W_0} = 0$ and $\partial_v W|_{W_0} = \left(\begin{smallmatrix} v_C & v_D \\ v_E & 0 \end{smallmatrix}\right)$, which vanishes iff $v = 0$. Hence $\mathcal{I} \succ 0$ at W_0.

Condition (iii). Both factors have maximal rank H at the truth (B_0 full column rank, A_0 full row rank). Any smooth curve in $\mathcal{R}_H$ through W_0 is $W(t) = B(t)A(t)/c$, with tangent $(\dot B A_0 + B_0 \dot A)/c = d\Phi_{(B_0,A_0)}(\dot B, \dot A)$; thus $T_{W_0}\mathcal{R}_H = \operatorname{im} d\Phi_{(B_0,A_0)}$ and Φ is a submersion onto $\mathcal{R}_H$. The same submersion-theorem argument as in Regime A (now in the chart) makes $\bar\varphi$ finite, smooth, strictly positive near W_0. Theorem 1 with $d_{\text{eff}} = d_H$ gives $\nu = \lambda = d_H/2$.

B.4 Regime C: saturation without submersion

Assume $r < \min(N, M)$ and $H \ge N + M - r$. Target space $\mathcal{M} = \mathbb{R}^{N\times M}$, $d = NM$, exactly as in Regime A. Set leftover dimensions $n' = N - r$, $m' = M - r$, $h = H - r$; the regime condition reads $h \ge n' + m'$.

Condition (ii). Target space and target model coincide with Regime A, so the Fisher information at W_0 is the identity on $\mathbb{R}^{N\times M}$ and positive definite.

Why a new argument is needed. With $\operatorname{rank}(W_0) = r < \min(N, M)$, Φ is *not* a submersion. In the canonical factorization $B_0 = \left(\begin{smallmatrix} I_r \\ 0 \end{smallmatrix}\right)$, $A_0 = \left(\begin{smallmatrix} I_r & 0 \end{smallmatrix}\right)$ (rank r, zero-padded), the differential $d\Phi_{(B_0,A_0)}(\dot B, \dot A) = \dot B A_0 + B_0 \dot A$ has zero bottom-right $n' \times m'$ block, so its image has dimension $r(N + M - r) = NM - n'm' < NM$. We verify condition (iii) directly.

Normal form of the factorization map. Partition $B = \left(\begin{smallmatrix} B_T \\ B_B \end{smallmatrix}\right)$, $A = \left(\begin{smallmatrix} A_L & A_R \end{smallmatrix}\right)$ with $B_T \in \mathbb{R}^{r\times H}$, $B_B \in \mathbb{R}^{n'\times H}$, $A_L \in \mathbb{R}^{H\times r}$, $A_R \in \mathbb{R}^{H\times m'}$. Then $W = BA$ has blocks $C = B_T A_L$, $D = B_T A_R$,

$L = B_B A_L$, $G = B_B A_R$. Near W_0, $C \approx I_r$ is invertible, so the Schur complement

$$F = G - LC^{-1}D = B_B\big(I_H - A_L(B_T A_L)^{-1}B_T\big)A_R \tag{14}$$

is well-defined, and $(C, D, L, G) \leftrightarrow (C, D, L, F)$ is a smooth diffeomorphism near W_0; (C, D, L, F) are smooth coordinates, $r^2 + rm' + n'r + n'm' = (r + n')(r + m') = NM$ in total. The matrix $P := I_H - A_L C^{-1} B_T$ is idempotent ($P^2 = P$, using $C = B_T A_L$), with complementary projection $Q = A_L C^{-1} B_T$ of rank r; hence $\operatorname{rank}(P) = H - r = h$, and P factors as $P = ST$ with $S \in \mathbb{R}^{H\times h}$, $T \in \mathbb{R}^{h\times H}$. Therefore

$$F = B_B\, P\, A_R = (B_B S)(T A_R) = UV, \qquad U \in \mathbb{R}^{n'\times h},\ V \in \mathbb{R}^{h\times m'}. \tag{15}$$

The $n'm'$ entries of the Schur complement are generated by a product of an $n' \times h$ and an $h \times m'$ matrix, with h playing the role of a bottleneck rank. The blocks (C, D, L) pass through Φ unchanged; the only nontrivial pushforward is that of $F = UV$ at $F = 0$, with the gauge variables γ integrated out against a smooth Gaussian integrand.

Lemma 1 (Saturation)**.** *Let $U \in \mathbb{R}^{n'\times h}$ have i.i.d. $\mathcal{N}(0, \tau^2)$ entries, let $V \in \mathbb{R}^{h\times m'}$ have i.i.d. columns $\mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$, independent of U, and set $F = UV \in \mathbb{R}^{n'\times m'}$. If $h \geq n' + m'$, the pushforward density of F is bounded, continuous and strictly positive at $F = 0$, with*

$$\rho_F(0) = (2\pi)^{-n'm'/2}\, \mathbb{E}_U\big[\det(U\Sigma U^\top)^{-m'/2}\big] < \infty. \tag{16}$$

Proof. This is the computation of Appendix C.2 at $L = 2$, carrying the column covariance Σ through; we record it here in the form used above.

Conditional on U, the columns of $F = UV$ are i.i.d. $\mathcal{N}(0, U\Sigma U^\top)$, so on the event $\{\operatorname{rank} U = n'\}$ (which has full probability once $h \geq n'$) F has the conditional Gaussian density

$$\rho_F(F \mid U) = (2\pi)^{-n'm'/2} \det(U\Sigma U^\top)^{-m'/2} \exp\Big(-\tfrac{1}{2}\operatorname{tr}\big(F^\top (U\Sigma U^\top)^{-1} F\big)\Big),$$

which is maximized at $F = 0$. Averaging over U, $\rho_F(F) \leq \rho_F(0)$ for every F, with $\rho_F(0)$ as in (16). Since $\det(U\Sigma U^\top) \geq \lambda_{\min}(\Sigma)^{n'} \det(UU^\top)$, finiteness reduces to the negative moment $\mathbb{E}_U[\det(UU^\top)^{-m'/2}]$. The Bartlett decomposition (Muirhead, 2005, Theorem 3.2.15) writes the squared volume as a product of independent chi-squareds, $\det(UU^\top) \overset{d}{=} \tau^{2n'} \prod_{i=0}^{n'-1} \chi^2_{h-i}$, and $\mathbb{E}[(\chi^2_k)^{-m'/2}] = \Gamma((k - m')/2)/\big(2^{m'/2}\Gamma(k/2)\big)$ is finite iff $k > m'$. The narrowest factor is $\chi^2_{h-n'+1}$, so the moment is finite precisely when $h - n' + 1 > m'$, i.e. $h \geq n' + m'$, and then ρ_F is bounded by (16). Continuity at $F = 0$ follows by dominated convergence against that envelope, and strict positivity because the conditional density is positive on a set of positive probability. □

Uniformity over spectators. The mixing matrices S, T in $F = (B_B S)(T A_R)$ depend on the spectator block (B_T, A_L) through P, so (U, V) is not a raw Gaussian; we check Lemma 1's hypothesis fiberwise. Choose $S \in \mathbb{R}^{H\times h}$ an orthonormal basis of $\operatorname{im}(P)$ and $T = S^\top P$, smoothly in (B_T, A_L) on a neighborhood of W_0; since $\operatorname{rank}(P) = h$ throughout, S has orthonormal columns and T has full row rank h. Conditional on the spectators, $U = B_B S$ and $V = T A_R$ are independent Gaussians, $\operatorname{vec}(U) \sim \mathcal{N}(0, \tau_0^2 I_h \otimes I_{n'})$ and $\operatorname{vec}(V) \sim \mathcal{N}(0, \tau_0^2 I_{m'} \otimes TT^\top)$, both non-degenerate ($TT^\top \succ 0$), so Lemma 1 applies at each spectator value with $\tau^2 = \tau_0^2$ and $\Sigma = \tau_0^2 TT^\top$. The nonzero eigenvalues of $PP^\top$ are continuous strictly positive functions of (B_T, A_L) on a compact spectator neighborhood, hence confined to a fixed interval $[\mu_-, \mu_+]$ with $\mu_- > 0$; since the only spectator-dependent quantity in the bound (16) is $\lambda_{\min}(\Sigma) \geq \tau_0^2 \mu_-$, the envelope $(2\pi)^{-n'm'/2}(\tau_0^2\mu_-)^{-n'm'/2}\, \mathbb{E}_U[\det(UU^\top)^{-m'/2}]$ is a single finite constant valid at every spectator value. The bound is therefore uniform, and joint continuity of the pushforward density in (C, D, L, F) follows by dominated convergence.

Conclusion. The blocks (C, D, L) are unchanged and the fiber variables are integrated out against a smooth positive Gaussian, so the induced W-prior in the coordinates (C, D, L, F) is continuous and strictly positive at $(C, D, L, F) = (I_r, 0, 0, 0)$. Condition (iii) holds; with (i) and (ii), Theorem 1 with $d_{\text{eff}} = NM$ gives $\nu = \lambda = NM/2$.

B.5 Regime B: the pushforward prior diverges (the principle fails)

Assume $H \geq \min(N, M)$, $r < \min(N, M)$, and $H < N + M - r$. As in Regime C the target space is $\mathcal{M} = \mathbb{R}^{N \times M}$ with $d_{\text{eff}} = NM$, and condition (ii) holds: the target model is the regular linear-Gaussian model, with identity Fisher information at W_0. Set $n' = N - r$, $m' = M - r$, $h = H - r$; the regime condition is now $h < n' + m'$, the reverse of Regime C.

Same normal form, opposite inequality. The normal form of Appendix B.4 is unchanged. In the coordinates (C, D, L, F) the blocks (C, D, L) pass through Φ unaltered, and the only nontrivial pushforward is that of the Schur complement $F = UV$ with $U \in \mathbb{R}^{n' \times h}$ and $V \in \mathbb{R}^{h \times m'}$ independent non-degenerate Gaussians conditional on the spectators (the construction of U, V from (B_B, A_R) and the projection P is verbatim that of Appendix B.4). The sole difference is the bottleneck width: Lemma 1 required $h \geq n' + m'$ to deliver a bounded, strictly positive density at $F = 0$, and that hypothesis is exactly what fails here.

The density diverges. When $h < n' + m'$ the factors U, V are too thin for the product $F = UV$ to spread regularly over a neighborhood of 0. The identity that proved Lemma 1 proves the converse here. Transposing F if necessary (which swaps n' and m' and replaces (U, V) by $(V^\top, U^\top)$), assume $n' \leq m'$. The regime hypothesis $H \geq \min(N, M)$ then supplies the lower bound

$$h = H - r \;\geq\; \min(N, M) - r \;=\; \min(n', m') \;=\; n', \tag{17}$$

so that, together with the regime condition $h < n' + m'$, the width is pinned in the range $n' \leq h < n' + m'$: the factors are thin enough to spoil the negative moment but never so thin that F loses its density. In particular $\operatorname{rank} U = n'$ almost surely, and the conditional-Gaussian computation of Lemma 1 applies verbatim, giving the density $\rho_F(F) = \mathbb{E}_U[\rho_F(F \mid U)]$ with

$$\rho_F(0) = (2\pi)^{-n'm'/2}\, \mathbb{E}_U\big[\det(U\Sigma U^\top)^{-m'/2}\big].$$

The Bartlett factorization $\det(UU^\top) \overset{d}{=} \tau^{2n'} \prod_{i=0}^{n'-1} \chi^2_{h-i}$ now has narrowest factor $\chi^2_{h-n'+1}$, and the regime condition $h < n' + m'$ forces $h - n' + 1 \leq m'$, where $\mathbb{E}[(\chi^2_k)^{-m'/2}] = \infty$ whenever $k \leq m'$. The chi-squared factors are independent and almost surely positive, so the expectation of the product is the product of the expectations, each lying in $(0, \infty]$; a single infinite factor therefore forces $\rho_F(0) = +\infty$. Moreover, since $\rho_F(\cdot \mid U)$ is continuous and nonnegative, Fatou's lemma gives $\liminf_{F \to 0} \rho_F(F) \geq \mathbb{E}_U\big[\liminf_{F \to 0} \rho_F(F \mid U)\big] = \rho_F(0) = +\infty$, so the pushforward density is not merely infinite at the point but diverges to $+\infty$ as $F \to 0$. It remains integrable, being a probability density.

Condition (iii) therefore fails at W_0. This is precisely the rank-zero singularity: the small-F density behaves like a negative power of $\det(F^\top F)$.

C Proofs for deep linear networks

We prove Proposition 2. Throughout, $\Phi(B_1, \ldots, B_L) = B_L \cdots B_1 / c$, the bottleneck width is $H_* = \min_{1 \leq \ell \leq L-1} H_\ell$, and the target spaces are those of the shallow network with H replaced by H_*.

C.1 Regimes A and D: submersion through the bottleneck

Regime A (Proposition 2, case A).

Proof sketch. Condition (ii) is unchanged from Appendix B.2: the target model $Z = WX + \eta$ on $\mathbb{R}^{N \times M}$ has positive-definite Fisher information at W_0.

For condition (iii), we exhibit a factorization of W_0 at which Φ is a submersion onto $\mathbb{R}^{N \times M}$. Assume $M \leq N$ (the other case follows by transposition). Since every $H_\ell \geq H_* \geq M$ and $\operatorname{rank}(W_0) = M$, we may choose factors (B°_ℓ) so that the partial product $P_{<L} := B^\circ_{L-1} \cdots B^\circ_1 \in \mathbb{R}^{H_{L-1} \times M}$ has full rank M. Freezing all but the top factor, the differential reduces to $d\Phi(\dot{B}_L) = \dot{B}_L\, P_{<L}/c$, which is surjective onto $\mathbb{R}^{N \times M}$ because $P_{<L}$ has rank M. From here the argument is exactly as in Regime A of RRR (Appendix B.2). $\square$

Regime D (Proposition 2, case D).

Proof sketch. The image of Φ lies in $\{\mathrm{rank} \le H_*\}$ and, near the rank-exactly-H_* point W_0, fills the smooth manifold $\mathcal{R}_{H_*}$. As in Appendix B.3, any smooth curve $W(t)$ in $\mathcal{R}_{H_*}$ through W_0 lifts to a curve of factorizations, so its velocity lies in the image of $d\Phi$; combined with $\mathrm{im}\, d\Phi \subseteq T_{W_0}\mathcal{R}_{H_*}$, this gives $\mathrm{im}\, d\Phi = T_{W_0}\mathcal{R}_{H_*}$, so Φ is a submersion onto $\mathcal{R}_{H_*}$. From here the argument is the same as in Regime D of RRR (Appendix B.3). □

C.2 Regime C: saturation through an L-fold product

This subsection proves the saturated case (Proposition 2, case C).

The shallow Regime C proof (Appendix B.4) had two parts. The blocks C, D, L passed through the factorization map Φ as spectators: they appear unchanged in the target coordinates and have nothing to do with the singular behavior. Only the residual Schur block $F \in \mathbb{R}^{n' \times m'}$ (with $n' = N - r$, $m' = M - r$) required work: we had to show that under $F = UV$ with U, V Gaussian, the pushforward density of F is continuous and strictly positive at $F = 0$.

The same reduction applies in the deep case. The spectators C, D, L still pass through unchanged, and the only nontrivial object is the residual Schur block, now generated by an L-fold product

$$F = U_L\, U_{L-1} \cdots U_1,$$

where $h_\ell := H_\ell - r$ are the leftover hidden widths and the U_ℓ have i.i.d. standard Gaussian entries. So the whole deep Regime C question reduces to: when is the pushforward density of this matrix product continuous and strictly positive at $F = 0$?

The shallow mechanism, revisited. First recall what happens at $L = 2$, $F = UV$ with U of shape $n' \times h$ and V of shape $h \times m'$. Conditional on U, the columns of $F = UV$ are i.i.d. $\mathcal{N}(0, UU^\top)$, so F has the Gaussian density

$$\rho_F(F \mid U) = (2\pi)^{-n'm'/2}\, \det(UU^\top)^{-m'/2}\, \exp\!\big(-\tfrac{1}{2}\,\mathrm{tr}\big(F^\top (UU^\top)^{-1} F\big)\big).$$

This is maximized at $F = 0$, so $\rho_F(F) \le \rho_F(0)$ for every F, and

$$\rho_F(0) = (2\pi)^{-n'm'/2}\, \mathbb{E}_U\big[\det(UU^\top)^{-m'/2}\big].$$

The pushforward density at the origin is the average of a **negative moment** of the squared volume $\det(UU^\top)$. This average is finite when $\det(UU^\top)$ does not pile up at zero, that is, when U does not nearly drop rank too often. For a Gaussian $n' \times h$ matrix, the Bartlett decomposition (Muirhead, 2005, Theorem 3.2.15) rewrites the squared volume as a product of independent chi-squareds,

$$\det(UU^\top) \;\overset{d}{=}\; \prod_{i=0}^{n'-1} \chi^2_{h-i},$$

and $\mathbb{E}[(\chi^2_k)^{-m'/2}] = \Gamma((k-m')/2)/(2^{m'/2}\Gamma(k/2))$ is finite iff $k > m'$. The narrowest factor is $\chi^2_{h-(n'-1)}$, so the threshold is $h \ge n' + m'$. This is the shallow saturation condition.

The deep mechanism. At $L \ge 3$ we play the same trick, peeling one layer at a time. Let $G := U_L U_{L-1} \cdots U_2$, so that $F = GU_1$. Conditional on G, the matrix F is Gaussian just as UV was Gaussian conditional on U, and the same calculation gives

$$\rho_F(0) = (2\pi)^{-n'm'/2}\, \mathbb{E}_G\big[\det(GG^\top)^{-m'/2}\big].$$

We now need the negative moment of the squared volume of G, which is itself a product of Gaussian factors. Iterating the Bartlett peel from the innermost factor outward, where each step replaces a deterministic factor by a Gaussian and produces a fresh, independent product of chi-squareds (Ipsen & Kieburg, 2014), gives

$$\det(GG^\top) \;\overset{d}{=}\; \prod_{\ell=1}^{L-1} \prod_{i=0}^{n'-1} \chi^2_{h_\ell - i},$$

all factors mutually independent. (This requires every intermediate width $h_\ell \geq n'$; if any layer falls below n', the product G drops rank almost surely, $\det(GG^\top) = 0$, and the negative moment is infinite.) Taking the expectation factor by factor, every $\mathbb{E}[(\chi^2_{h_\ell - i})^{-m'/2}]$ must be finite, which demands $h_\ell - (n' - 1) > m'$, i.e. $h_\ell \geq n' + m'$, for every layer. Translating back to the original widths, the threshold becomes $H_* \geq N + M - r$. When it holds, the density at the origin has the closed form

$$\rho_F(0) = (2\pi)^{-n'm'/2} \prod_{\ell=1}^{L-1} \prod_{i=0}^{n'-1} \frac{\Gamma\big((h_\ell - i - m')/2\big)}{2^{m'/2}\,\Gamma\big((h_\ell - i)/2\big)}.$$

The physical reading: the random map $F = U_L \cdots U_1$ smears its mass uniformly near $F = 0$ only when every hidden layer is wide enough that the product cannot nearly collapse. Depth does not shift the threshold: it is set by the narrowest leftover width alone.

Conclusion of the proof. The bound $\rho_F(F) \leq \rho_F(0)$ from the conditional calculation above, together with finiteness of $\rho_F(0)$ under the threshold, gives a bounded pushforward density; continuity at $F = 0$ follows by dominated convergence (the integrand is continuous in F and dominated by an integrable function), and strict positivity at $F = 0$ holds because the integrand is positive on a set of positive probability. The remaining ingredients of the factorization argument (the (C, D, L, F) chart giving smooth coordinates on $\mathbb{R}^{N\times M}$, and target-model regularity, condition (ii), reading off as in the shallow case from the spectator structure) carry over unchanged. Theorem 1 then applies with $\mathcal{M} = \mathbb{R}^{N\times M}$, $d = NM$, giving $\nu = \lambda = NM/2$.

D PROOFS FOR THE QUADRATIC NETWORK

This appendix proves Propositions 3 and 4. We first set up the model and the two regime analogs (verifying the three hypotheses of Theorem 1 in each), then prove the lemmas used along the way.

The model. Fix integers $H, M \geq 1$. Let $X \sim \mathcal{N}(0, I_M)$ and $\varepsilon \sim \mathcal{N}(0, 1)$ be independent. The scalar-output quadratic network with H hidden units maps $\mathbb{R}^M \to \mathbb{R}$ via

$$Y = \sum_{h=1}^{H} a_h (b_h^\top X)^2 + \varepsilon,$$

with parameters $w = (a, B) \in \Theta := \mathbb{R}^H \times \mathbb{R}^{H\times M}$, where $a = (a_1, \ldots, a_H)$ and the rows of $B \in \mathbb{R}^{H\times M}$ are $b_h^\top$. The prior $\varphi(w)$ is i.i.d. standard Gaussian on every entry, and $d_{\text{param}} = H(M + 1)$.

Each summand $a_h(b_h^\top X)^2 = X^\top(a_h b_h b_h^\top)X$ is a quadratic form, so $Y = X^\top GX + \varepsilon$ with

$$G \;:=\; \sum_{h=1}^{H} a_h\, b_h b_h^\top \;\in\; \mathrm{Sym}^M.$$

The likelihood depends on w only through G, so the factorization map is

$$\Phi : \Theta \longrightarrow \mathrm{Sym}^M, \qquad \Phi(a, B) \;=\; \sum_{h=1}^{H} a_h\, b_h b_h^\top.$$

D.1 THE REGIME D ANALOG

We now exhibit the analog of Regime D in the linear case. Take $r = H \leq M$ and a generic truth $w_0 = (a^*, B^*)$ with $a_h^* \neq 0$ for every h and rows $b_1^*, \ldots, b_r^*$ linearly independent; then $G^* := \Phi(w_0)$ has rank exactly r.

The image variety. A sum of r rank-≤ 1 symmetric matrices has rank $\leq r$, and any symmetric S of rank $k \leq r$ admits a spectral decomposition $\sum_{i=1}^{k} \mu_i u_i u_i^\top$ which realizes $S = \Phi(a, B)$ on setting $a_i = \mu_i$, $b_i = u_i$ for $i \leq k$ and the remaining $(a_i, b_i) = 0$. Hence

$$\mathrm{im}\,\Phi \;=\; \mathcal{S}_r \;:=\; \{\, S \in \mathrm{Sym}^M : \mathrm{rank}\, S \leq r \,\}.$$

The set $\mathcal{S}_r$ is not a manifold globally: its singular locus is $\mathcal{S}_{r-1}$, the matrices of strictly smaller rank. At any S_0 of rank exactly r it is locally smooth.

Lemma 2 (Smoothness and dimension of $\mathcal{S}_r$)**.** *The set $\mathcal{S}_r \setminus \mathcal{S}_{r-1}$ of symmetric matrices of rank exactly r is a smooth submanifold of Sym^M of dimension*

$$d \;=\; \binom{M+1}{2} - \binom{M-r+1}{2} \;=\; rM - \binom{r}{2}.$$

The differential is surjective. For the factorization principle we need not only image identification but also that Φ is a smooth submersion near w_0, so that the pushforward prior $\bar{\varphi}$ in hypothesis (iii) is a well-defined smooth density on $\mathcal{S}_r$ near G^*. Differentiating $\Phi(a, B) = \sum_h a_h b_h b_h^\top$ at w_0 in the direction $(\dot{a}, \dot{B})$ gives

$$d\Phi_{w_0}(\dot{a}, \dot{B}) \;=\; \sum_{h=1}^{r} \dot{a}_h\, b_h^* (b_h^*)^\top \;+\; \sum_{h=1}^{r} a_h^* \left[\dot{b}_h (b_h^*)^\top + b_h^* \dot{b}_h^\top\right].$$

The image is automatically contained in $T_{G^*}\mathcal{S}_r$ since every smooth curve through w_0 maps into $\mathcal{S}_r$ under Φ. The content of the next lemma is the reverse inclusion: every tangent vector of $\mathcal{S}_r$ at G^* is realized by varying w.

Lemma 3 (Submersion)**.** *Under the generic-truth conditions above, $\mathrm{im}\, d\Phi_{w_0} = T_{G^*}\mathcal{S}_r$, and Φ is a smooth submersion onto an open piece of $\mathcal{S}_r \setminus \mathcal{S}_{r-1}$ on a neighborhood of w_0.*

Verifying the factorization principle. We verify the three hypotheses of Theorem 1 with $\mathcal{M} = \mathcal{S}_r \setminus \mathcal{S}_{r-1}$ near G^* and $d_{\text{eff}} = rM - \binom{r}{2}$.

Hypothesis (i): factorization. By construction $p(Y \mid X, w) = \bar{p}(Y \mid \Phi(w))$ with $\bar{p}(Y \mid G) = \mathcal{N}(Y \mid X^\top G X, 1)$ and $\mathrm{im}\,\Phi = \mathcal{S}_r$.

Hypothesis (ii): target-model regularity. The target model is $Y = X^\top G X + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$ independent of X, and log-likelihood $\log \bar{p}(Y \mid G) = -\frac{1}{2}(Y - X^\top G X)^2 + \text{const}$. For any tangent direction $V \in T_{G^*}\mathcal{S}_r \subset \mathrm{Sym}^M$, the score at the single data point (X, Y) is

$$\partial_V \log \bar{p}(Y \mid G^*) \;=\; (Y - X^\top G^* X)\,(X^\top V X) \;=\; \varepsilon\,(X^\top V X).$$

The Fisher information in direction V is the variance of this score. Independence of ε and X together with $\mathbb{E}[\varepsilon^2] = 1$ reduce it to $\mathbb{E}_X[(X^\top V X)^2]$, which the Isserlis identity for $X \sim \mathcal{N}(0, I_M)$ and symmetric V evaluates as

$$\mathbb{E}_X\left[(X^\top V X)^2\right] \;=\; (\mathrm{tr}\, V)^2 + 2\|V\|_F^2.$$

The Frobenius term is strictly positive whenever $V \neq 0$, so the target Fisher information is positive definite at G^*.

Hypothesis (iii): pushforward prior. By Lemma 3, Φ is a smooth submersion near w_0, so the pushforward-prior argument is identical to the linear Regimes A and D (Appendices B.2 and B.3).

Conclusion. With (i)-(iii) verified, Theorem 1 applies with $d_{\text{eff}} = rM - \binom{r}{2}$ to give $\nu = \lambda = d_{\text{eff}}/2 = H(2M - H + 1)/4$.

D.2 The Regime C analog

We now exhibit the analog of Regime C. Take $H \geq 3D$ where $D := \binom{M+1}{2} = \dim \mathrm{Sym}^M$, and make *no* assumption on $\mathrm{rank}(G^*)$: the truth $G^* \in \mathrm{Sym}^M$ may have any rank, including zero.

The image is all of Sym^M. Once $H \geq M$, the spectral-decomposition argument of the Regime D analog (with r replaced by M) realizes every $S \in \mathrm{Sym}^M$ as $\Phi(a, B)$. Hence $\mathrm{im}\,\Phi = \mathrm{Sym}^M$, a smooth D-dimensional vector space with no rank stratification.

Linearity of Φ in A. A device we will use repeatedly: at fixed B, the map $a \mapsto \Phi(a, B)$ is *linear* in a. Fix once and for all an orthonormal basis of Sym^M for the Frobenius inner product, and write $\mathrm{vec}_{\mathrm{sym}} \colon \mathrm{Sym}^M \xrightarrow{\sim} \mathbb{R}^D$ for the coordinate isomorphism. Set $v(b) := \mathrm{vec}_{\mathrm{sym}}(b\, b^\top) \in \mathbb{R}^D$. Then

$$\mathrm{vec}_{\mathrm{sym}}\big(\Phi(a, B)\big) \;=\; V(B)\, a, \qquad V(B) \;:=\; \big[\, v(b_1) \,\big|\, \cdots \,\big|\, v(b_H) \,\big] \;\in\; \mathbb{R}^{D \times H}. \tag{18}$$

We also write $W(B) := V(B)V(B)^\top \in \mathrm{Sym}^D$ for the associated Gram matrix.

Verifying the factorization principle. We apply Theorem 1 with $\mathcal{M} = \mathrm{Sym}^M$, $m_0 = G^*$, and $d_{\mathrm{eff}} = D$.

Hypothesis (i): factorization. By construction $p(Y \mid X, w) = \bar{p}(Y \mid \Phi(w))$ with $\bar{p}(Y \mid G) = \mathcal{N}(Y \mid X^\top G X, 1)$, now with G ranging over the entire ambient Sym^M.

Hypothesis (ii): target-model regularity. The score and Isserlis computation of the Regime D analog go through verbatim, with the tangent direction now ranging over all of Sym^M instead of $T_{G^*}\mathcal{S}_r$; the Frobenius term in $(\mathrm{tr}\, V)^2 + 2\|V\|_F^2$ is strictly positive for every nonzero $V \in \mathrm{Sym}^M$, so the target Fisher information is positive definite *at every* $G^* \in \mathrm{Sym}^M$.

Hypothesis (iii), lower bound. We need $\bar{\varphi}(G) \geq c_- > 0$ on a neighborhood of G^*. The strategy is the same as in the positivity half of Lemma 1: exhibit a single preimage in $\Phi^{-1}(G^*)$ at which Φ is a submersion onto Sym^M, and integrate the Gaussian prior over the fiber coordinate. The submersion is supplied by the following lemma.

Lemma 4 (Submersion at a well-chosen preimage)**.** *Assume $H \geq D$. For every $G^* \in \mathrm{Sym}^M$ there exists $(a_0, B_0) \in \Phi^{-1}(G^*)$ at which Φ is a submersion.*

Hypothesis (iii), upper bound. For boundedness we use linearity (18). Under the Gaussian prior $a \sim \mathcal{N}(0, I_H)$, the image of a centered Gaussian under a linear map is centered Gaussian, so $G \mid B$ is centered Gaussian on Sym^M with covariance $W(B)$. Its density is controlled by the following two lemmas.

Lemma 5 (Veronese-Gram invertibility)**.** *If $H \geq D$, then $W(B)$ is invertible almost surely under the Gaussian prior on B.*

Lemma 6 (Veronese-Gram negative moment)**.** *Let $b_1, \ldots, b_H \sim \mathcal{N}(0, I_M)$ i.i.d. and let $W(B)$ be as above. If $H \geq 3D$, then the constant*

$$c_+ \;:=\; \mathbb{E}\big[\det W(B)^{-1/2}\big]$$

is finite.

On the full-measure event $\{W(B) \text{ invertible}\}$ of Lemma 5, the conditional density of G given B on $\mathrm{Sym}^M \cong \mathbb{R}^D$ is the centered Gaussian

$$\bar{\varphi}(G \mid B) \;=\; (2\pi)^{-D/2}\, \det W(B)^{-1/2}\, \exp\Big(-\tfrac{1}{2}\, \mathrm{vec}_{\mathrm{sym}}(G)^\top\, W(B)^{-1}\, \mathrm{vec}_{\mathrm{sym}}(G)\Big),$$

maximized at its mean and with exponential factor at most one, so

$$\bar{\varphi}(G \mid B) \;\leq\; (2\pi)^{-D/2}\, \det W(B)^{-1/2} \qquad \forall\, G \in \mathrm{Sym}^M.$$

Marginalizing over B by the tower property and applying Lemma 6,

$$\bar{\varphi}(G) \;=\; \mathbb{E}_B\big[\bar{\varphi}(G \mid B)\big] \;\leq\; (2\pi)^{-D/2}\, c_+ \qquad \text{for every } G \in \mathrm{Sym}^M.$$

Continuity. On the full-measure invertibility event of Lemma 5, the conditional density $\bar{\varphi}(G \mid B)$ is continuous in G and dominated by the B-integrable envelope $(2\pi)^{-D/2} \det W(B)^{-1/2}$, whose expectation is the finite constant c_+ of Lemma 6. Dominated convergence then makes $\bar{\varphi}(G) = \mathbb{E}_B[\bar{\varphi}(G \mid B)]$ continuous in G. Together with the lower bound this gives a continuous, strictly positive pushforward prior at G^*, which is hypothesis (iii).

Conclusion. With (i)-(iii) verified for $\mathcal{M} = \mathrm{Sym}^M$ and $d_{\mathrm{eff}} = D = \binom{M+1}{2}$, Theorem 1 applies to give

$$\nu \;=\; \lambda \;=\; \tfrac{1}{2}\binom{M+1}{2},$$

at *every* truth $G^* \in \mathrm{Sym}^M$, with no assumption on $\mathrm{rank}(G^*)$.

D.3 PROOFS OF THE LEMMAS

Both proofs of Lemmas 2 and 3 below are standard; their strategy and methods appear in (Lee, 2013).

SMOOTHNESS AND DIMENSION OF $\mathcal{S}_r$

The factorization principle demands that $\mathcal{S}_r$ look locally, near G^*, like an open piece of $\mathbb{R}^d$ for some fixed $d = \dim \mathcal{S}_r$. Sets defined by rank *inequalities* are not automatically smooth: their singular locus consists of matrices where the rank drops further. For $\mathcal{S}_r$ the singular locus is $\mathcal{S}_{r-1}$ (matrices of rank strictly less than r); we avoid it by working at G^*, which has rank exactly r by the generic-truth assumption.

The standard route to a smooth structure is to write $\mathcal{S}_r$ near S_0 as the zero set of a smooth submersion $F\colon U \to \mathrm{Sym}^{M-r}$, and then invoke the implicit function theorem (zero sets of submersions are smooth submanifolds of the right codimension). We construct such an F from the Schur complement.

Set-up: block coordinates aligned with the row space of S_0. Fix $S_0 \in \mathcal{S}_r$ with $\operatorname{rank} S_0 = r$. Its row space $\operatorname{im} S_0$ is an r-dimensional subspace of $\mathbb{R}^M$. Choose an orthonormal basis $(e_1, \dots, e_M)$ of $\mathbb{R}^M$ with $e_1, \dots, e_r$ spanning $\operatorname{im} S_0$, and partition any $S \in \mathrm{Sym}^M$ in block form:

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}^\top & S_{22} \end{pmatrix}, \qquad S_{11} \in \mathrm{Sym}^r, \quad S_{12} \in \mathbb{R}^{r\times(M-r)}, \quad S_{22} \in \mathrm{Sym}^{M-r}.$$

At $S = S_0$, the matrix vanishes on the orthogonal complement of its row space, so $(S_0)_{12} = 0$ and $(S_0)_{22} = 0$. The block $(S_0)_{11}$ is the restriction of S_0 to its row space, an isomorphism, hence invertible. By continuity, S_{11} stays invertible on an open neighborhood $U \subset \mathrm{Sym}^M$ of S_0.

The Schur identity reduces "rank $\le r$" to a matrix equation. For block-symmetric S with invertible S_{11} one has the LDU factorization

$$S = \begin{pmatrix} I & 0 \\ S_{12}^\top S_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} S_{11} & 0 \\ 0 & S_{22} - S_{12}^\top S_{11}^{-1} S_{12} \end{pmatrix} \begin{pmatrix} I & S_{11}^{-1} S_{12} \\ 0 & I \end{pmatrix}.$$

The outer triangular factors have determinant 1, hence are rank-preserving; only the middle block-diagonal matrix contributes to the rank. Therefore

$$\operatorname{rank} S \;=\; \operatorname{rank} S_{11} \;+\; \operatorname{rank}\big(S_{22} - S_{12}^\top S_{11}^{-1} S_{12}\big) \;=\; r + \operatorname{rank} F(S), \tag{19}$$

where

$$F(S) \;:=\; S_{22} - S_{12}^\top S_{11}^{-1} S_{12} \;\in\; \mathrm{Sym}^{M-r}$$

is the *Schur complement* of S_{11} in S. The takeaway from (19): on U,

$$S \in \mathcal{S}_r \quad\Longleftrightarrow\quad F(S) = 0.$$

Geometrically: once coordinates are chosen so that the first r directions capture the image of S_0, the bottom-right $(M-r)\times(M-r)$ block of S is forced to a specific function of the other two blocks if S is to remain rank-$\le r$. That forced equality is exactly $F(S) = 0$.

F is a smooth submersion. The entries of F are smooth rational functions of S_{11}, S_{12}, S_{22} (with S_{11} invertible on U). Its S_{22}-partial is the identity on Sym^{M-r}, so dF_S is surjective for every $S \in U$. By the implicit function theorem, $F^{-1}(0) \cap U$ is a smooth submanifold of Sym^M near S_0 of codimension $\dim \mathrm{Sym}^{M-r} = \binom{M-r+1}{2}$.

Proof of Lemma 2. Smoothness is the conclusion of the implicit function theorem applied to F. The dimension is direct:

$$\binom{M+1}{2} - \binom{M-r+1}{2} \;=\; \tfrac{1}{2}\big[M(M+1) - (M-r)(M-r+1)\big] \;=\; rM - \binom{r}{2}.$$

The tangent-space description is standard for the zero set of a submersion: $T_{S_0}\mathcal{S}_r = \ker dF_{S_0}$. □

THE DIFFERENTIAL SURJECTS: LIFTING ARGUMENT

We turn to the proof of Lemma 3. The containment $\operatorname{im} d\Phi_{w_0} \subseteq T_{G^*}\mathcal{S}_r$ is free: every smooth curve in Θ has image in $\mathcal{S}_r$ under Φ, so its tangent at 0 lands in $T_{G^*}\mathcal{S}_r$. The real work is the reverse inclusion: showing that every tangent vector of $\mathcal{S}_r$ at G^* is realized by varying w. The standard device is a *lifting argument*:

(1) Given $V \in T_{G^*}\mathcal{S}_r$, choose a smooth curve $G\colon (-\epsilon, \epsilon) \to \mathcal{S}_r$ with $G(0) = G^*$ and $\dot{G}(0) = V$. (Such a curve exists by Lemma 2.)
(2) Lift it to a smooth curve $w(t)$ in Θ with $\Phi(w(t)) = G(t)$.
(3) Differentiate to obtain $d\Phi_{w_0}\dot{w}(0) = \dot{G}(0) = V$.

For the lift we re-use the Schur-complement parameterization from the previous subsection and combine it with a smooth signed factorization of the upper-left $r \times r$ block.

Proof of Lemma 3. The inclusion $\operatorname{im} d\Phi_{w_0} \subseteq T_{G^*}\mathcal{S}_r$ was explained above.

For the reverse inclusion, fix $V \in T_{G^*}\mathcal{S}_r$. By Lemma 2, $\mathcal{S}_r$ is a smooth submanifold of Sym^M near G^*, so there is a smooth curve $G\colon (-\epsilon, \epsilon) \to \mathcal{S}_r$ with $G(0) = G^*$ and $\dot{G}(0) = V$. By lower-semicontinuity of rank and the fact that $\operatorname{rank} G^* = r$, each $G(t)$ has rank exactly r for small $|t|$.

Step 1: Schur factorization of $G(t)$. Use the basis of the previous subsection in which the first r vectors span $\operatorname{im} G^*$, and partition each $G(t)$ in $r \oplus (M-r)$ blocks $\left(\begin{smallmatrix} S_{11}(t) & S_{12}(t) \\ S_{12}(t)^\top & S_{22}(t) \end{smallmatrix}\right)$. The block $S_{11}(0)$ is the restriction of G^* to its row space and is therefore invertible, so $S_{11}(t)$ stays invertible on a smaller interval. The Schur identity (19) forces $S_{22}(t) = S_{12}(t)^\top S_{11}(t)^{-1} S_{12}(t)$, hence

$$G(t) \;=\; P(t)\, A(t)\, P(t)^\top, \qquad P(t) \;:=\; \begin{pmatrix} I_r \\ S_{12}(t)^\top S_{11}(t)^{-1} \end{pmatrix} \in \mathbb{R}^{M\times r}, \quad A(t) := S_{11}(t) \in \mathrm{Sym}^r. \tag{20}$$

Both $P(t)$ and $A(t)$ are smooth in t, and $A(t)$ stays invertible.

Step 2: smooth signed factorization of $A(t)$. Let $J = \operatorname{diag}(j_1, \dots, j_r)$ with $j_h \in \{+1, -1\}$ be the inertia of $A(0)$ (the signs of its eigenvalues, counted with multiplicity). By Sylvester's law of inertia there exists $R_0 \in GL(r)$ with $A(0) = R_0^\top J\, R_0$. We claim there is a smooth $R\colon (-\epsilon', \epsilon') \to GL(r)$ with $R(0) = R_0$ and

$$A(t) \;=\; R(t)^\top J\, R(t). \tag{21}$$

Write $R(t) = (I + X(t))R_0$, $X(0) = 0$. Equation (21) becomes

$$J\, X(t) + X(t)^\top J + X(t)^\top J\, X(t) \;=\; R_0^{-\top}\big(A(t) - A(0)\big)R_0^{-1}.$$

The right-hand side is symmetric, smooth in t, and vanishes at $t = 0$. The linearization $X \mapsto JX + X^\top J$ from $\mathbb{R}^{r\times r}$ to Sym^r is surjective: for any $S \in \mathrm{Sym}^r$, the choice $X = \frac{1}{2}JS$ satisfies $JX + X^\top J = \frac{1}{2}(S + S^\top) = S$ (using $J^2 = I$). By the implicit function theorem, the equation admits a smooth solution $X(t)$ with $X(0) = 0$ on a smaller interval, and then $R(t) = (I + X(t))R_0$ is the required smooth signed-factorization curve.

Step 3: the lift. Combining (20) and (21),

$$G(t) \;=\; P(t)\, R(t)^\top J\, R(t)\, P(t)^\top \;=\; Q(t)\, J\, Q(t)^\top, \qquad Q(t) := P(t)\, R(t)^\top \in \mathbb{R}^{M\times r}.$$

Writing $Q(t) = \big(q_1(t) \mid \cdots \mid q_r(t)\big)$ column-wise,

$$G(t) \;=\; \sum_{h=1}^{r} j_h\, q_h(t)\, q_h(t)^\top \;=\; \Phi\big(a(t), B(t)\big),$$

where $a_h(t) = j_h$ (constant, $= \pm 1$) and $b_h(t) = q_h(t)$ (smooth in t). Differentiating $\Phi(a(t), B(t)) = G(t)$ at $t = 0$,

$$d\Phi_{w(0)}\, \dot{w}(0) \;=\; \dot{G}(0) \;=\; V.$$

Step 4: matching the basepoint to w_0. At this point we have a smooth curve $w(t) = (a(t), B(t))$ in Θ with $\Phi(w(t)) = G(t)$, and consequently $d\Phi_{w(0)}\dot{w}(0) = V$. The conclusion we want is stronger - $V \in \operatorname{im} d\Phi_{w_0}$, with the differential taken at w_0 specifically, so we need to turn the curve $w(t)$ through $w(0)$ into a curve through w_0.

Why $w(0) \neq w_0$ in general. By construction $a(0) = J \in \{-1, +1\}^r$, a vector of signs, whereas a^* has arbitrary nonzero entries. Likewise $B(0)$ is the factorization produced in Steps 1-3, while B^* is only assumed to have linearly independent rows. The two points $w(0)$ and w_0 are different factorizations of the *same* matrix G^*: both lie in the fiber $\Phi^{-1}(G^*)$, but at different places.

The gauge action. Φ has an explicit gauge group: transformations of Θ that preserve every value of Φ. Three pieces of this group are enough to transport $w(0)$ to w_0.

(a) *Permutation of units.* The symmetric group S_r relabels the index $h \in \{1, \dots, r\}$ on a and B jointly. Manifestly Φ-invariant.

(b) *Per-unit rescaling.* For any $\lambda = (\lambda_1, \dots, \lambda_r) \in (\mathbb{R} \setminus \{0\})^r$, $(a_h, b_h) \mapsto (a_h/\lambda_h^2,\ \lambda_h b_h)$ preserves each summand $a_h b_h b_h^\top$.

(c) *D^*-congruence rotation.* Set $D^* = \operatorname{diag}(a^*)$. For any $O \in GL(r)$ with $O^\top D^* O = D^*$, the map $B \mapsto OB$ (with a unchanged) preserves $\Phi(a^*, B) = B^\top D^* B$. Such O form the $\binom{r}{2}$-dimensional stabilizer of D^* under congruence (concretely $O(p, q)$, where p, q are the numbers of positive and negative entries of a^*).

Matching $a(0)$ to a^* using (a) and (b). Sylvester's law of inertia applied to $G^* = B^{*\top} D^* B^*$ (with B^* of rank r by genericity) and to the alternative factorization $G^* = Q(0) J Q(0)^\top$ from Step 3 implies $\{j_h\} = \{\operatorname{sgn}(a_h^*)\}$ as multisets (both equal the inertia of G^*). A permutation in (a) lines up the orders so that $\operatorname{sgn}(a_h(0)) = \operatorname{sgn}(a_h^*)$ for each h. Then choosing $\lambda_h^2 = j_h / a_h^* > 0$ in (b) converts $a_h(0) = j_h$ into a_h^*.

Matching B to B^* using (c). After (a) and (b), a equals a^* and we have some matrix B' with $B'^\top D^* B' = G^* = B^{*\top} D^* B^*$. Both B' and B^* are $r \times M$ of rank r, so their row spaces both equal $\operatorname{im} G^*$ (an r-dimensional subspace of $\mathbb{R}^M$). Hence $B^* = OB'$ for some $O \in GL(r)$, and the identity $B^{*\top} D^* B^* = B'^\top D^* B'$ together with the injectivity of $B'^\top$ forces $O^\top D^* O = D^*$. So O is exactly an element of the gauge subgroup (c), and applying it takes B' to B^*.

Conclusion. The composition of the three gauge transformations above is a single gauge map $g\colon \Theta \to \Theta$ with two properties:

(I) $g(w(0)) = w_0$ (by construction, this is precisely how we picked the permutation, the rescaling, and the D^*-orthogonal rotation);

(II) $\Phi \circ g = \Phi$ as maps $\Theta \to \operatorname{Sym}^M$ (each of (a), (b), (c) preserves the value of Φ, so does their composition).

Define a new curve $\tilde{w}(t) := g(w(t))$ in Θ. By (I) we have $\tilde{w}(0) = g(w(0)) = w_0$, so $\tilde{w}$ passes through w_0 at $t = 0$. By (II) applied at $w(t)$ for each t,

$$\Phi(\tilde{w}(t)) \;=\; \Phi\big(g(w(t))\big) \;=\; \Phi(w(t)) \;=\; G(t).$$

That is: the curve $\tilde{w}$ maps under Φ to the *same* target curve $G(t)$ as the original lift, even though its basepoint has moved.

We now differentiate the identity $\Phi(\tilde{w}(t)) = G(t)$ at $t = 0$ in two ways. On the one hand, since $\Phi \circ \tilde{w}$ and G are equal as smooth functions of t,

$$\left.\frac{d}{dt}\right|_{t=0} \Phi(\tilde{w}(t)) \;=\; \left.\frac{d}{dt}\right|_{t=0} G(t) \;=\; \dot{G}(0) \;=\; V. \tag{22}$$

On the other hand, by the chain rule applied to the composite $t \mapsto \tilde{w}(t) \mapsto \Phi(\tilde{w}(t))$, using $\tilde{w}(0) = w_0$,

$$\left.\frac{d}{dt}\right|_{t=0} \Phi(\tilde{w}(t)) \;=\; d\Phi_{\tilde{w}(0)}\,\dot{\tilde{w}}(0) \;=\; d\Phi_{w_0}\,\dot{\tilde{w}}(0). \tag{23}$$

Equating (22) and (23) gives $d\Phi_{w_0}\dot{\tilde{w}}(0) = V$. Hence $V \in \operatorname{im} d\Phi_{w_0}$.

Finally, for the constant-rank claim: at every $w \in \Theta$ with $\operatorname{rank}\Phi(w) = r$, the inclusion $\operatorname{im} d\Phi_w \subseteq T_{\Phi(w)}\mathcal{S}_r$ supplies the upper bound $\operatorname{rank} d\Phi_w \le d$. At w_0 equality holds by what we have just shown. The rank of a smooth map's differential is lower-semicontinuous, so equality persists on a neighborhood W of w_0, on which $\Phi|_W$ is a smooth submersion onto an open piece of $\mathcal{S}_r \setminus \mathcal{S}_{r-1}$ containing G^*. □

REGIME C ANALOG: INVERTIBILITY, SUBMERSION, NEGATIVE MOMENT

This subsection proves Lemmas 5, 4, and 6. Notation is that of the body: $v(b) := \operatorname{vec}_{\mathrm{sym}}(b\,b^\top) \in \mathbb{R}^D$ in a fixed orthonormal Frobenius basis of Sym^M, $V(B) := [\,v(b_1) \mid \cdots \mid v(b_H)\,] \in \mathbb{R}^{D\times H}$, and $W(B) := V(B)V(B)^\top \in \mathrm{Sym}^D$, with $D = \binom{M+1}{2}$.

Veronese-Gram invertibility.

Proof of Lemma 5. The set

$$\mathcal{B}_{\mathrm{bad}} \;:=\; \{B \in \mathbb{R}^{HM} : \operatorname{rank} V(B) < D\} \;=\; \bigcap_{\alpha} \{B : p_\alpha(B) = 0\},$$

where $\{p_\alpha\}$ are the $D \times D$ minors of $V(B)$, viewed as polynomials in the entries of B. If we exhibit a single $B^\star \in \mathbb{R}^{H\times M}$ with $\operatorname{rank} V(B^\star) = D$, then some p_α is non-zero at $B^\star$, hence not identically zero as a polynomial. The zero set of a non-zero real polynomial on $\mathbb{R}^{HM}$ has Lebesgue measure zero, so $\mathcal{B}_{\mathrm{bad}}$ has Lebesgue measure zero, and the Gaussian prior on B, being absolutely continuous with respect to Lebesgue measure, assigns it probability zero.

Construction of $B^\star$. Index the $D = \binom{M+1}{2}$ symmetric pairs by $\mathcal{P} := \{(i,j) : 1 \le i \le j \le M\}$ and fix a bijection $h \mapsto (i_h, j_h)$ from $\{1,\dots,D\}$ to $\mathcal{P}$. Define

$$b_h^\star \;:=\; e_{i_h} + e_{j_h} \quad (h = 1,\dots,D), \qquad b_h^\star \;:=\; 0 \quad (h = D+1,\dots,H),$$

and let $B^\star$ be the matrix with rows $(b_h^\star)^\top$ (the choice of the trailing $H - D$ rows is immaterial).

Linear independence of $\{v(b_h^\star)\}_{h=1}^D$. Use the explicit orthonormal Frobenius basis $\{E_{ij}\}_{(i,j)\in\mathcal{P}}$ of Sym^M with

$$E_{ii} \;=\; e_i e_i^\top, \qquad E_{ij} \;=\; \tfrac{1}{\sqrt{2}}\big(e_i e_j^\top + e_j e_i^\top\big) \quad (i < j).$$

A direct expansion gives

$$b_h^\star (b_h^\star)^\top \;=\; \begin{cases} 4\,E_{i_h i_h}, & i_h = j_h, \\ E_{i_h i_h} + E_{j_h j_h} + \sqrt{2}\,E_{i_h, j_h}, & i_h < j_h. \end{cases}$$

The off-diagonal coordinate E_{ij} (with $i < j$) appears in exactly one of the vectors $v(b_h^\star)$, namely the one with $(i_h, j_h) = (i, j)$. A linear combination $\sum_h c_h v(b_h^\star) = 0$ therefore forces $c_h = 0$ for every off-diagonal h; the remaining diagonal-index vectors are pairwise proportional to distinct E_{ii} and so trivially independent. Hence $\{v(b_h^\star)\}_{h=1}^D$ is a basis of $\mathbb{R}^D$ and $\operatorname{rank} V(B^\star) = D$, as required. □

Submersion at a well-chosen preimage.

Proof of Lemma 4. By Lemma 5, the set

$$\Omega \;:=\; \{B \in \mathbb{R}^{H\times M} : \operatorname{rank} V(B) = D\}$$

has full Lebesgue measure once $H \ge D$. Fix any $B_0 \in \Omega$ (chosen independently of G^*). Since $V(B_0) \in \mathbb{R}^{D\times H}$ has full row rank D, the linear system

$$V(B_0)\,a_0 \;=\; \operatorname{vec}_{\mathrm{sym}}(G^*) \tag{24}$$

admits a solution $a_0 \in \mathbb{R}^H$ for arbitrary right-hand side; in particular for the given G^*. By the linearity (18), $\Phi(a_0, B_0) = G^*$.

The partial map $a \mapsto \Phi(a, B_0)$ is the composition of the linear map $a \mapsto V(B_0)\,a$ with the linear isomorphism $\operatorname{vec}_{\mathrm{sym}}^{-1} \colon \mathbb{R}^D \xrightarrow{\sim} \mathrm{Sym}^M$; the differential of a linear map equals the map itself, so

$$d_a\Phi\big|_{(a_0,B_0)}(\dot a) \;=\; \operatorname{vec}_{\mathrm{sym}}^{-1}\big(V(B_0)\,\dot a\big) \quad \in \mathrm{Sym}^M.$$

Because $V(B_0)$ has full row rank, $\dot{a} \mapsto V(B_0)\,\dot{a}$ surjects $\mathbb{R}^H$ onto $\mathbb{R}^D$; composing with the isomorphism gives a surjection $\mathbb{R}^H \twoheadrightarrow \mathrm{Sym}^M$. The full differential $d\Phi_{(a_0,B_0)}(\dot{a},\dot{B}) = d_a\Phi(\dot{a}) + d_B\Phi(\dot{B})$ contains the image of $d_a\Phi$, which is all of Sym^M, hence is itself surjective. □

The Veronese-Gram negative moment. The proof of Lemma 6 is the analytic core of the Regime C analog. It combines anti-concentration of polynomials in Gaussian variables (Carbery-Wright), the Gram-Schmidt expression for a determinant as a product of perpendicular heights, and a Cauchy-Binet-plus-AM-GM split into three blocks.

The Carbery-Wright inequality. Throughout, $b, b_1, b_2, \cdots \sim \mathcal{N}(0, I_M)$ i.i.d., and for $U \in \mathrm{Sym}^M$ we write $u = \mathrm{vec}_{\mathrm{sym}}(U) \in \mathbb{R}^D$, so that $\langle v(b), u\rangle = b^\top U b$. The Carbery-Wright anti-concentration inequality (Carbery & Wright, 2001; Nazarov et al., 2003) asserts that for every degree-k real polynomial P in a Gaussian vector $b \sim \mathcal{N}(0, I_M)$ there is an absolute constant C_* with

$$\mathbb{P}\big(|P(b)| \le s\big) \;\le\; \frac{C_*\, s^{1/k}}{\|P\|_{L^2}^{1/k}} \qquad \forall\, s > 0.$$

We apply this at $k = 2$, to $P(b) = b^\top U b$ with $\|U\|_F = 1$. The fourth-moment identity $\mathbb{E}[(X^\top S X)^2] = (\mathrm{tr}\, S)^2 + 2\|S\|_F^2$ used in the body yields $\|P\|_{L^2}^2 \ge 2$, hence

$$\mathbb{P}\big(|b^\top U b| \le s\big) \;\le\; C_1\, s^{1/2}, \qquad C_1 \;:=\; \frac{C_*}{2^{1/4}}, \tag{25}$$

uniformly over all $U \in \mathrm{Sym}^M$ with $\|U\|_F = 1$.

Lemma 7 (Negative moment of a single quadratic form)**.** *For every* $0 < p < \frac{1}{2}$,

$$M_p \;:=\; \sup_{\|U\|_F=1} \mathbb{E}\big[\,|b^\top U b|^{-p}\,\big] \;\le\; 1 + \frac{C_1}{\frac{1}{2p} - 1} \;<\; \infty.$$

Proof. Set $X := |b^\top U b|$. The layer-cake identity gives

$$\mathbb{E}[X^{-p}] \;=\; \int_0^\infty \mathbb{P}\big(X^{-p} > \lambda\big)\, d\lambda \;=\; \int_0^\infty \mathbb{P}\big(X < \lambda^{-1/p}\big)\, d\lambda.$$

Split at $\lambda = 1$. For $\lambda < 1$, use the trivial bound $\mathbb{P} \le 1$, contributing 1. For $\lambda \ge 1$, apply (25) with $s = \lambda^{-1/p}$:

$$\mathbb{P}\big(X < \lambda^{-1/p}\big) \;\le\; C_1\, \lambda^{-1/(2p)}.$$

The tail integral converges since $1/(2p) > 1$, and equals $C_1/(\frac{1}{2p} - 1)$. Adding the two contributions gives the bound. Uniformity in U follows from the uniformity of (25). □

Lemma 8 (Negative moment of one $D \times D$ Veronese block)**.** *Let* $V_D := [\, v(b_1) \mid \cdots \mid v(b_D)\,] \in \mathbb{R}^{D\times D}$. *For every* $0 < p < \frac{1}{2}$,

$$\mathbb{E}\big[\,|\det V_D|^{-p}\,\big] \;\le\; M_p^D.$$

Proof. The volume of the parallelepiped spanned by D vectors in $\mathbb{R}^D$ is the product of D perpendicular heights:

$$|\det V_D| \;=\; \prod_{j=1}^{D} d_j, \qquad d_j \;:=\; \mathrm{dist}\big(v(b_j),\, \mathrm{span}\{v(b_1), \ldots, v(b_{j-1})\}\big).$$

We bound each conditional negative moment of d_j uniformly by M_p, then assemble via the tower property.

Filtration. Let $\mathcal{F}_{j-1} := \sigma(b_1, \ldots, b_{j-1})$. Since $\dim \mathrm{span}\{v(b_1), \ldots, v(b_{j-1})\} \le j - 1 < D$, its orthogonal complement $L_{j-1}^\perp \subseteq \mathbb{R}^D$ has dimension ≥ 1, so one may select a unit vector $u_j \in L_{j-1}^\perp$ in an $\mathcal{F}_{j-1}$-measurable way (e.g. by Gram-Schmidt applied to a fixed basis of $\mathbb{R}^D$). Set $U_j := \mathrm{vec}_{\mathrm{sym}}^{-1}(u_j) \in \mathrm{Sym}^M$; since $\mathrm{vec}_{\mathrm{sym}}$ is a Frobenius isometry, $\|U_j\|_F = \|u_j\| = 1$, and both u_j, U_j are $\mathcal{F}_{j-1}$-measurable.

Geometric lower bound on the height. Let $P_{L_{j-1}}$ be the orthogonal projection onto L_{j-1}; then $d_j = \|v(b_j) - P_{L_{j-1}} v(b_j)\|$. Since $u_j \perp L_{j-1}$ and $\|u_j\| = 1$, Cauchy-Schwarz gives

$$d_j \;\geq\; \big|\langle v(b_j) - P_{L_{j-1}} v(b_j),\; u_j\rangle\big| \;=\; \big|\langle v(b_j), u_j\rangle\big| \;=\; |b_j^\top U_j\, b_j|.$$

Conditional bound via Lemma 7. Fix an outcome ω in the underlying probability space and freeze the past: $b_1(\omega), \ldots, b_{j-1}(\omega)$ are specific vectors in $\mathbb{R}^M$, so $U_j(\omega) \in \mathrm{Sym}^M$ is a single deterministic matrix with $\|U_j(\omega)\|_F = 1$. From the perspective of the regular conditional distribution $\mathbb{P}(\cdot \mid \mathcal{F}_{j-1})(\omega)$:

- U_j is constant: it equals $U_j(\omega)$;
- b_j retains its unconditional law $\mathcal{N}(0, I_M)$, by the i.i.d. independence of b_j from $\mathcal{F}_{j-1}$.

Conditionally, $b_j^\top U_j b_j$ has the same law as $b^\top U b$ with $U = U_j(\omega)$ a fixed unit-Frobenius matrix and $b \sim \mathcal{N}(0, I_M)$. By the uniformity in U of Lemma 7,

$$\mathbb{E}\big[\, |b_j^\top U_j b_j|^{-p} \,\big|\, \mathcal{F}_{j-1}\,\big](\omega) \;\leq\; \sup_{\|U\|_F = 1} \mathbb{E}\big[\, |b^\top U b|^{-p}\,\big] \;=\; M_p.$$

Combined with the geometric lower bound on d_j, this gives $\mathbb{E}[d_j^{-p} \mid \mathcal{F}_{j-1}] \leq M_p$ almost surely.

Tower assembly. Iterating the conditional bound $\mathbb{E}[d_j^{-p} \mid \mathcal{F}_{j-1}] \leq M_p$ for $j = D, D-1, \ldots, 1$ yields, by the tower property, $\mathbb{E}\big[\prod_j d_j^{-p}\big] \leq M_p^D$. Since $|\det V_D| = \prod_j d_j$, this is the claim. □

Proof of Lemma 6. Assume $H \geq 3D$, and partition $\{1, \ldots, H\}$ into three disjoint index sets S_1, S_2, S_3 of size D each (any unused indices are dropped). By the Cauchy-Binet formula,

$$\det W(B) \;=\; \det\big(V(B)V(B)^\top\big) \;=\; \sum_{\substack{S \subseteq \{1,\ldots,H\} \\ |S| = D}} \big(\det V_S\big)^2 \;\geq\; \sum_{k=1}^{3} \big(\det V_{S_k}\big)^2,$$

where $V_S \in \mathbb{R}^{D \times D}$ is the column submatrix of $V(B)$ indexed by S. AM-GM on the three nonnegative terms gives

$$\tfrac{1}{3} \sum_{k=1}^{3} \big(\det V_{S_k}\big)^2 \;\geq\; \Big(\prod_{k=1}^{3} \big(\det V_{S_k}\big)^2\Big)^{1/3},$$

hence (absorbing the factor of 3)

$$\det W(B)^{-1/2} \;\leq\; \prod_{k=1}^{3} \big|\det V_{S_k}\big|^{-1/3}.$$

The three factors on the right are *independent*, since the b_h's indexed by different S_k's are independent. Therefore

$$\mathbb{E}\big[\det W(B)^{-1/2}\big] \;\leq\; \prod_{k=1}^{3} \mathbb{E}\big[\, |\det V_{S_k}|^{-1/3}\,\big].$$

Each block factor has D i.i.d. Gaussian columns, the setup of Lemma 8 with $p = \frac{1}{3} < \frac{1}{2}$, so $\mathbb{E}[|\det V_{S_k}|^{-1/3}] \leq M_{1/3}^D$. Combining,

$$\mathbb{E}\big[\det W(B)^{-1/2}\big] \;\leq\; \big(M_{1/3}^D\big)^3 \;=\; M_{1/3}^{3D} \;<\; \infty.$$

□

E GRADIENT DESCENT IN THE EFFECTIVELY REGULAR NETWORKS

This appendix reports the gradient-descent experiments summarized in Section 6. Section 6 establishes that, because the likelihood factors through Φ, the global minimizers of the training loss are exactly the maximum likelihood estimator $\hat{w}_n$ of Theorem 2; whether plain gradient descent

reaches such a minimizer is an optimization question we settle empirically. We do so in each of the effectively regular examples of the paper: the shallow linear network (Section 7.1), the deep linear network (Section 7.2), and the nonlinear network (Section 7.3). For every model we (i) give a construction of the model-class maximum likelihood estimator, against which the trained weight is compared, (ii) state the exact generalization excess, and (iii) report the run.

E.1 Constructing the maximum likelihood estimator

In each model the likelihood depends on the parameters only through the composite object $m = \Phi(w)$, so the model-class MLE is the minimizer of the empirical squared error over the realized class $\operatorname{im}\Phi$. We compute it as follows.

Linear networks (shallow and deep). The composite weight is $W = B_L \cdots B_1/c$ and $\operatorname{im}\Phi = \{\, W \,:\, \operatorname{rank} W \le H^* \,\}$, with bottleneck width $H^* = H$ for the shallow network and $H^* = \min_\ell H_\ell$ for the deep one. The model-class MLE is the rank-constrained least-squares (reduced-rank regression) solution

$$W_{\mathrm{ML}} \;=\; \arg\min_{\operatorname{rank} W \le H^*} \; \tfrac{1}{n}\sum_{i=1}^{n} \big\| y_i - W x_i \big\|^2. \tag{26}$$

Writing $\mathcal{A} = \sum_i x_i x_i^\top$ and $M_0 = \big(\sum_i y_i x_i^\top\big)\mathcal{A}^{-1/2}$, the substitution $Z = W\mathcal{A}^{1/2}$ turns (26) into the unweighted Eckart–Young problem $\min_{\operatorname{rank} Z \le H^*} \|Z - M_0\|_F^2$, whose solution truncates M_0 to its top H^* singular components:

$$W_{\mathrm{ML}} \;=\; \lfloor M_0 \rfloor_{H^*}\, \mathcal{A}^{-1/2}, \tag{27}$$

where $\lfloor\cdot\rfloor_{H^*}$ is the rank-H^* singular-value truncation. This is a closed form and, crucially, depends only on (N, M, H^*, r), not on the depth L: adding layers changes the parameterization, not the global optimum. When $H^* \ge \min(N, M)$ (regimes A and C) the truncation is vacuous and W_{ML} is ordinary least squares.

Nonlinear network, saturated regime (Regime C analog). For the quadratic network $Y = \sum_h a_h (b_h^\top X)^2 + \varepsilon = X^\top G X + \varepsilon$ with $G = \sum_h a_h b_h b_h^\top \in \mathrm{Sym}^M$, the prediction $X^\top G X = \langle G, XX^\top\rangle$ is *linear* in G. In the saturated regime the realized class is all of Sym^M, so the MLE is an ordinary least-squares fit in the $D = \binom{M+1}{2}$ symmetric coordinates of G, again a closed form.

Nonlinear network, matched-low-rank regime (Regime D analog). Here the realized class is the rank-$\le r$ symmetric matrices, and the rank constraint makes the fit non-convex with no closed form. We take the model-class MLE to be the best of $R = 12$ independent gradient-descent restarts, and measure whether a single small-init run reaches this reference.

E.2 Exact excess and protocol

Generalization is measured by the exact population excess, which for $X \sim \mathcal{N}(0, I_M)$ is available in closed form with no test set. For the linear networks, with fitted composite weight W and truth W_0,

$$G_n - S \;=\; \tfrac{1}{2}\big\| W_0 - W \big\|_F^2, \tag{28}$$

and for the quadratic network, with $\Delta = G_0 - \widehat{G}$, the Isserlis identity gives

$$G_n - S \;=\; \tfrac{1}{2}\Big[(\operatorname{tr}\Delta)^2 + 2\|\Delta\|_F^2 \Big]. \tag{29}$$

Throughout, the noise variance is $\sigma^2 = 1$ and the coupling is $c = 1$ (so $W = B_L \cdots B_1$), the excess is computed exactly from the learned weight (no test-set Monte-Carlo noise) and averaged over K datasets, and we plot $n(G_n - S)$ against n; the effectively regular prediction is $n(G_n - S) \to d_{\mathrm{eff}}/2 = \lambda$. The optimization diagnostic is the distance $\|\widehat{W} - W_{\mathrm{ML}}\|_F$ (or $\|\widehat{G} - G_{\mathrm{ML}}\|_F$) between the trained weight and the model-class MLE of Section E.1: it certifies that gradient descent reached the global optimum, so that the observed rate is the estimator's rate and not an artifact of under-optimization.

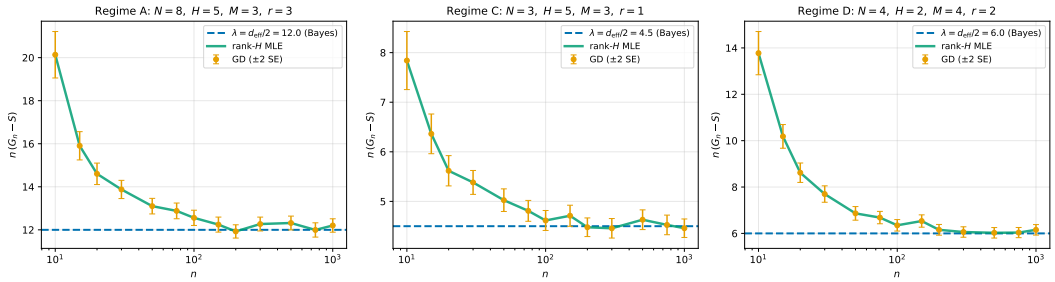


Figure 4: Shallow linear network, effectively regular regimes. $n(G_n - S)$ for gradient descent (orange, ± 2 SE) and the rank-H MLE (green), against the Bayesian rate $\lambda = d_{\text{eff}}/2$ (dashed). In each regime GD coincides with the MLE and approaches λ.

E.3 Shallow linear network

Full-batch gradient descent is run on the (A, B) factors from a small random initialization (scale 0.01, learning rate 0.1), with $K = 500$ datasets per n; regime A uses 2000 steps and regimes C, D use 8000. In every effectively regular regime GD reaches the rank-H MLE and generalizes at λ (Figure 4). At $n = 1000$:

Regime	(N, H, M, r)	$n(G_n - S)$	$\lambda = d_{\text{eff}}/2$	$\|\widehat{W} - W_{\text{ML}}\|_F$
A	$(8, 5, 3, 3)$	12.20	12.0	1.3×10^{-15}
C	$(3, 5, 3, 1)$	4.46	4.5	8.2×10^{-5}
D	$(4, 2, 4, 2)$	6.15	6.0	2.3×10^{-15}

The diagnostic is at machine precision in regimes A and D (in D the $r = H$ fit leaves no weak modes to grow), and $\leq 10^{-4}$ in C, where the noise modes are grown to that level within 8000 steps; the GD and MLE excess curves coincide to three decimals.

E.4 Deep linear network

With two hidden layers, $W = B_3 B_2 B_1$ over widths $[M, H_1, H_2, N]$ and bottleneck $H^* = \min(H_1, H_2)$. Gradient descent uses $K = 250$ datasets, learning rate 0.1, 8000 steps, and initialization scale 0.3 (a 0.01 init stalls through three factors). The MLE (27) with H^* is identical to the shallow MLE with $H \to H^*$, since it does not depend on depth. In every effectively regular regime (including D, where the bottleneck $H^* = 2$ sits in the *second* layer) deep GD reaches the rank-H^* MLE and generalizes at λ (Figure 5). At $n = 1000$:

Regime	$[M, H_1, H_2, N], r$	$n(G_n - S)$	$\lambda = d_{\text{eff}}/2$	$\|\widehat{W} - W_{\text{ML}}\|_F$
A	$[3, 5, 5, 8]$, r=3	12.00	12.0	2.8×10^{-15}
C	$[3, 5, 5, 3]$, r=1	4.50	4.5	1.4×10^{-6}
D	$[4, 3, 2, 4]$, r=2	6.31	6.0	2.2×10^{-15}

The bottleneck width governs the entire picture exactly as the shallow H did.

Why it happens. That GD reaches the global optimum here is expected: for a linear network of any depth, including a rank-constrained bottleneck, every local minimum of the squared loss is global and all other critical points are saddles (Kawaguchi, 2016), so small-init GD is not trapped.

E.5 Nonlinear network, saturated regime (Regime C analog)

For the quadratic network the prediction $X^\top G X$ is linear in G, so in the saturated regime the realized class is all of Sym^M and the MLE is the closed-form ordinary least-squares fit in the $D = \binom{M+1}{2}$ symmetric coordinates (Section E.1). The quartic loss is stiff, so gradient descent uses

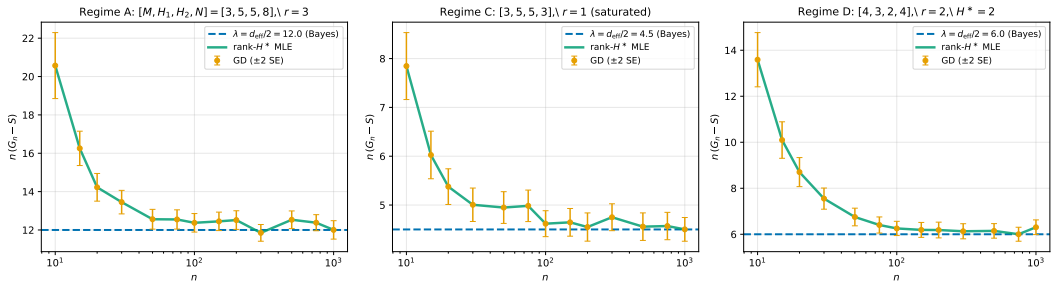


Figure 5: Deep linear network (two hidden layers), effectively regular regimes. $n(G_n - S)$ for gradient descent (orange, ± 2 SE) and the rank-H^* MLE (green), against $\lambda = d_{\text{eff}}/2$ (dashed). The bottleneck $H^* = \min(H_1, H_2)$ replaces the shallow H; GD again coincides with the MLE.

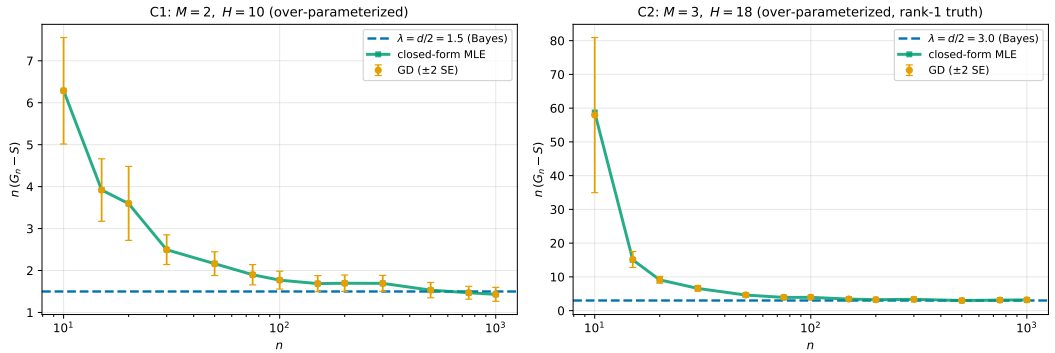


Figure 6: Nonlinear network, saturated regime (Regime C analog), $n(G_n - S)$ vs n. GD (± 2 SE, orange) coincides with the closed-form Sym^M MLE (green) and approaches $\lambda = d/2$ (dashed).

global-norm gradient clipping (unclipped fixed-step GD diverges), with $K = 250$ datasets, learning rate 0.01, and 12000 steps. The landscape is benign: GD reaches the closed-form MLE at machine precision ($\|\widehat{G} - G_{\mathrm{ML}}\|_F \sim 10^{-15}$ for $n \gtrsim 75$) and $n(G_n - S) \to \lambda = d/2$, and the true rank is irrelevant (C2 uses a rank-1 truth) (Figure 6). At $n = 1000$ (columns as in the linear tables, with the optimization diagnostic now $\|\widehat{G} - G_{\mathrm{ML}}\|_F$):

Regime	(M, H), truth	$n(G_n - S)$	$\lambda = d/2$	$\|\widehat{G} - G_{\mathrm{ML}}\|_F$
C1	$(2, 10)$, full rank	1.43	1.5	1.9×10^{-15}
C2	$(3, 18)$, rank 1	3.18	3.0	1.8×10^{-11}

E.6 Nonlinear network, matched-low-rank regime (Regime D analog)

Here the realized class is the rank-$\leq r$ symmetric matrices; the rank constraint makes the fit non-convex with no closed form, so the only reference for the model-class MLE is best-of-restarts GD, which grades the optimizer against itself. We therefore make no "GD reaches the MLE" rate claim from this regime, and the figure reports only the robust quantity: GD conditioned on reaching the best basin found, whose excess descends to $\lambda = d/2$ (Figure 7). Gradient descent uses learning rate 0.004, 20000 steps, clip 5, $K = 250$ datasets. The GD|success excess sits close to λ and slightly below it: in D1 ($M = 3, H = r = 2$, $\lambda = 2.5$) it is 2.49 at $n = 750$ and 2.22 at $n = 1000$; in D2 ($M = 4, H = r = 3$, $\lambda = 4.5$) it is 4.23 at $n = 750$ and 4.21 at $n = 1000$. The shortfall is expected, because conditioning is a selection: we keep the runs whose training loss reached the best basin, and training loss and generalization are positively correlated across runs on a dataset, so the runs we drop are also the ones that would have generalized worst. The conditioned average is

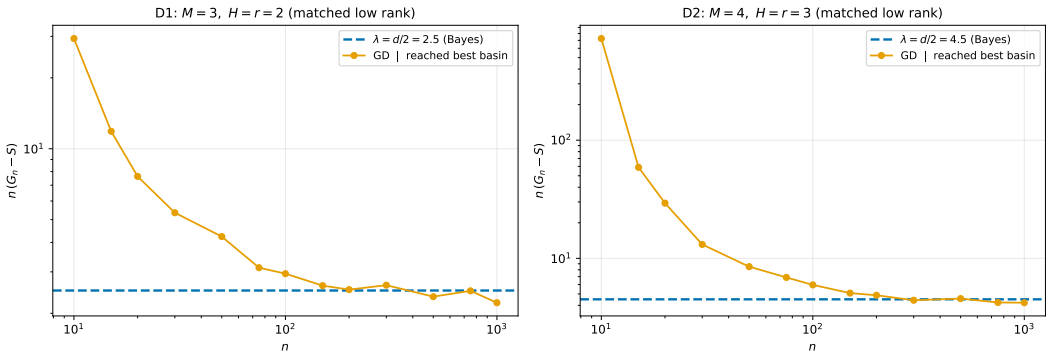


Figure 7: Nonlinear network, matched-low-rank regime (Regime D analog), $n(G_n - S)$ vs n (log-y). GD conditioned on reaching the best basin found (orange) descends to $\lambda = d/2$ (dashed). With no closed-form MLE we plot only this conditioned-on-success curve; the trapping is a fixed-n landscape effect, reported in the text at $n = 500$.

therefore a downward-biased estimate of the excess of the effectively regular estimator, and these values should be read as approaching λ from below rather than as an unbiased measurement of it.

The estimator here is still effectively regular (the rank-$\leq r$ MLE approaches λ), but the landscape is not benign: plain small-init GD does not reliably reach the minimum of its own training loss. This is a fixed-n statement about the landscape and needs no MLE reference, only that some trajectory reaches a strictly lower loss than another. At $n = 500$ in D2, with a reference of best of $R = 16$ restarts each run to 1.5×10^5 steps and a run counted a success when its training loss falls within 10^{-2} of that reference ($K = 100$ datasets), a single small-init trajectory reaches the best basin on 88% of datasets; on the remaining 12% it converges to a strictly higher-loss minimum. This is genuine trapping, not slow convergence: a $10\times$ increase in training steps ($5 \times 10^4 \to 5 \times 10^5$) leaves the reached-best fraction pinned at 88% and the stuck trajectories frozen. We do not plot a success rate against n: with the reference itself best-of-restarts GD, its quality drifts with n and the curve would be ill-defined.

What we observe. Bad local minima appear here and nowhere else among our examples. We do not offer a landscape theory that predicts where they occur; the statement is purely empirical. The over-parameterized Regime C analog is benign in every run, whereas the rank-constrained Regime D analog traps plain small-init gradient descent on a non-negligible fraction of datasets at the finite n we probe (the 88% reached-best figure above). Conditioned on reaching the best basin, gradient descent still matches the effectively regular estimator, so this is a statement about the optimizer, not the estimator: effective regularity is a property of the estimator, and whether an algorithm inherits it is a separate, empirical fact about the landscape.

F THE SINGULAR FLUCTUATION IN THE RUNNING EXAMPLE

Theorem 1 concludes $\nu = \lambda = d_{\text{eff}}/2$, but the experiments of Appendix E measure only the generalization side, which pins λ. This appendix measures the other invariant directly in the running example, the shallow network of Figure 1 with $(N, H, M, r) = (4, 2, 4, 2)$, where the prediction is

$$\nu = \lambda = d_{\text{eff}}/2 = 6, \qquad \text{equivalently} \qquad \mathbb{E}[V_n] \longrightarrow 2\nu = d_{\text{eff}} = 12.$$

Protocol. We sample the posterior by blocked Gibbs on (A, B), which is exact for this model: with a Gaussian prior and Gaussian noise, $A \mid B$ is Gaussian with the Kronecker precision $(B^\top B) \otimes (X^\top X)/\sigma^2 + I/\tau_0^2$, diagonalized by eigendecomposing the two factors separately, and $B \mid A$ is Gaussian with the $H \times H$ precision $A(X^\top X)A^\top/\sigma^2 + I/\tau_0^2$ shared by its rows. There is no Metropolis step and no step size to tune. Only the sufficient statistics $X^\top X$ and $Y^\top X$ enter the conditionals, so the cost per sweep does not grow with n. We take $\sigma^2 = 1$, $c = 1$ and the standard-normal prior of (8) ($\tau_0^2 = 1$), and run 1500 warmup sweeps followed by 3000 kept draws at thinning 2, on $K = 200$ independent datasets per n. The functional variance is evaluated on the

kept draws as $V_n = \sum_{i=1}^{n} \mathrm{Var}_s\left[-\|y_i - W^{(s)}x_i\|^2/2\sigma^2\right]$, the additive constant in $\log p$ dropping out of the variance.

Validation. Because a functional variance is easy to get subtly wrong, we first run the identical code path on $(N, H, M, r) = (1, 1, 1, 1)$, which is Regime A with $d_{\text{eff}} = NM = 1$ and therefore has the known target $\mathbb{E}[V_n] \to 1$. It returns 1.005 ± 0.007 at $n = 1000$.

Result. The validation run is consistent with its exact target at every n, fluctuating on both sides of it within Monte-Carlo error, while the running example rises towards $d_{\text{eff}} = 12$ (monotonically apart from a single dip at $n = 150$, within noise of its neighbours) and reaches it to within 0.3%, about 1.4 standard errors, at $n = 1000$:

	validation, $d_{\text{eff}} = 1$		running example, $d_{\text{eff}} = 12$	
n	$\mathbb{E}[V_n]$	$\mathbb{E}[V_n]/d_{\text{eff}}$	$\mathbb{E}[V_n]$	$\mathbb{E}[V_n]/d_{\text{eff}}$
100	0.993 ± 0.017	0.993	11.802 ± 0.082	0.984
200	1.013 ± 0.013	1.013	11.765 ± 0.055	0.980
500	0.991 ± 0.008	0.991	11.940 ± 0.037	0.995
1000	1.005 ± 0.007	1.005	11.963 ± 0.026	0.997

At $n = 1000$ the running example gives $\hat{\nu} = 5.98 \pm 0.01$ against the predicted $\nu = 6$, the same constant as λ. Together with Figure 1, which shows the generalization side approaching $\lambda = 6$, this confirms both halves of Theorem 1 in the one configuration.