OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Chebyshev Moment Regularization (CMR): Condition-Number Control with Moment Shaping

Jinwoo Baek BAEKJI@OREGONSTATE.EDU
School of Electrical Engineering and Computer Science
Department of Mathematics
Oregon State University
Corvallis, OR 97331, USA

Abstract

Training deep networks often fails due to ill-conditioned weight matrices and poorly shaped spectra. We propose Chebyshev Moment Regularization (CMR), an orthogonally invariant spectral penalty that jointly controls a log-condition proxy and higher-order Chebyshev moments of the normalized Gram spectrum. CMR is implemented in exact and fast matrix-free modes, and its gradients are mixed with the task loss via a decoupled, norm-capped rule. On a deliberately ill-conditioned "VANISH" testbed, CMR reliably revives collapsed training, stabilizes gradient norms, and produces compact spectra, outperforming spectral norm and condition-only baselines. We further prove monotone descent of the condition proxy, bounded moment gradients, and orthogonal invariance, providing theoretical support for CMR as a practical conditioning tool.

1. Introduction

Training deep networks in aggressive regimes (large depth, high learning rates, low precision) often produces ill-conditioned weight matrices with skewed spectra, leading to vanishing or unstable gradients. Common fixes such as careful initialization, architectural tweaks, or spectral norm constraints only indirectly affect conditioning or control a single spectral edge.

We propose Chebyshev Moment Regularization (CMR), a simple drop-in spectral penalty that directly shapes the Gram spectrum of each layer. CMR combines a smooth log-condition proxy with higher-order Chebyshev moments of an affinely normalized spectrum, is orthogonally invariant, and admits both exact and matrix-free implementations. To avoid destabilizing optimization, we compute task and spectral gradients separately and mix them with a time-dependent weight and a global norm cap, so that the spectral update never overwhelms the task update.

On a deliberately ill-conditioned "VANISH" testbed, CMR reliably rescues collapsed runs, stabilizes gradient norms, and yields compact spectra, outperforming spectral norm and condition-only baselines. We further prove monotone descent of the condition proxy under its gradient flow and bounded moment gradients, supporting CMR as a practical tool for conditioning-sensitive training.

2. Method: Chebyshev Moment Regularization

Let $W \in \mathbb{R}^{m\times n}$ be a weight matrix (for conv layers we flatten each kernel to an $m \times n$ matrix) and $G = W^\top W$ with eigenvalues $0 \le \lambda_{\min}(G) \le \lambda_{\max}(G)$.

Spectral proxies. We control the condition number via

$$\rho_{\text{cond}}(W) = \tfrac{1}{2}\Big(\log \lambda_{\max}(G) - \log \big(\lambda_{\min}(G) + \epsilon\big)\Big), \tag{1}$$

where $\epsilon > 0$ is a small floor.
To capture higher–order spectral shape we normalize G to $[-1, 1]$. Let

$$c = \tfrac{1}{2}\big(\lambda_{\max}(G) + \lambda_{\min}(G)\big), \quad d = \max\Big\{\tfrac{1}{2}\big(\lambda_{\max}(G) - \lambda_{\min}(G)\big),\, \epsilon\Big\}, \tag{2}$$

and define $\widehat{G} = (G - cI)/d$. With Chebyshev polynomials T_k (via the standard three–term recurrence), we define for $K \geq 3$ and $\beta > 0$

$$s_k(W) = \frac{1}{n}\mathrm{Tr}\big(T_k(\widehat{G})\big), \qquad w_k = \exp\big(\beta(k-3)\big), \quad k = 3, \ldots, K, \tag{3}$$

and the moment proxy

$$\rho_{\text{moment}}(W) = \sum_{k=3}^{K} w_k\, s_k(W)^2. \tag{4}$$

We start at $k \geq 3$ because $k = 0, 1, 2$ mostly encode mass/mean/variance of the normalized spectrum and are already constrained by the edge normalization and ρ_{cond}.
For a model with weight matrices $\{W^{(\ell)}\}_{\ell=1}^{L}$ we define the total spectral penalty

$$\mathcal{L}_{\text{spec}}(\theta) = \sum_{\ell=1}^{L} \Big(\alpha_1\, \rho_{\text{cond}}(W^{(\ell)}) + \alpha_2\, \rho_{\text{moment}}(W^{(\ell)})\Big). \tag{5}$$

Exact and fast modes. In *exact* mode we form $G = W^\top W$, compute all eigenvalues, and evaluate $s_k(W)$ exactly by applying the Chebyshev recurrence to $\widehat{G}$ and taking traces. In *fast* mode we only use matrix–vector products with W and $W^\top$: $\lambda_{\max}(G)$ is obtained by power iteration, $\lambda_{\min}(G)$ by inverse power iteration on $(G + \epsilon I)^{-1}$, and $s_k(W)$ are estimated with a Hutchinson estimator using S Rademacher probe vectors and the same recurrence applied in a matrix–free way.

2.1. Decoupled and Capped Spectral Gradients

Let $\mathcal{L}_{\text{task}}(\theta)$ be the usual training loss. We use a time–dependent regularization weight

$$\lambda_t = \lambda \cdot \min\Big\{1,\, \frac{t}{T_{\text{w}}}\Big\}, \tag{6}$$

where T_{w} is a warmup horizon (`warmup_steps` in code). At step t the conceptual objective is

$$\mathcal{L}_t(\theta) = \mathcal{L}_{\text{task}}(\theta) + \lambda_t\, \mathcal{L}_{\text{spec}}(\theta). \tag{7}$$

Rather than differentiating $\mathcal{L}_t$ in one pass, CMR decouples the gradients. We first compute

$$g_{\text{task}} = \nabla_\theta \mathcal{L}_{\text{task}}(\theta), \qquad g_{\text{spec,raw}} = \nabla_\theta \mathcal{L}_{\text{spec}}(\theta),$$

and then apply the warmup to the spectral part

$$g_{\text{spec}} = \lambda_t\, g_{\text{spec,raw}}.$$

Let $\rho_{\text{spec}} \in (0, 1]$ be a cap and $\delta > 0$ a small stabilizer. With global ℓ_2 norms $\|\cdot\|$ over all trainable parameters we set

$$\gamma_t = \begin{cases} 0, & \text{if } \|g_{\text{spec}}\| = 0, \\ \min\left\{1,\ \dfrac{\rho_{\text{spec}}\|g_{\text{task}}\|}{\|g_{\text{spec}}\| + \delta}\right\}, & \text{otherwise,} \end{cases} \tag{8}$$

and update with

$$g_t = g_{\text{task}} + \gamma_t\, g_{\text{spec}}. \tag{9}$$

Thus the spectral update is always bounded as $\|\gamma_t g_{\text{spec}}\| \le \rho_{\text{spec}}\|g_{\text{task}}\|$.

3. Theoratical Analysis

We show: (i) the condition proxy decreases under its gradient flow, (ii) moment gradients are bounded under mild spread, and (iii) the penalty is orthogonally invariant. Proofs are in App. B; the rationale for using $k \ge 3$ moments is in App. A.

Assumptions and notation. We take $\epsilon > 0$ so that $\lambda_{\min}(W^\top W + \epsilon I) > 0$. Gradients of spectral terms are well-defined whenever the extremal singular values are simple; otherwise, statements hold almost everywhere and extend via Clarke subgradients of spectral functions [1, 7, 10]. For Lemma 3 we assume a nontrivial spectral spread $\lambda_{\max}(G) - \lambda_{\min}(G) \ge \theta\, \lambda_{\max}(G)$ for some $\theta \in (0, 1]$; see Remark 4 for the $\theta \to 0$ case.

Theorem 1 (Monotone Descent for the Condition Proxy) *Let $W(t)$ evolve under the gradient flow $\dot{W}(t) = -\eta \nabla_W \rho_{cond}(W(t))$ for $\eta > 0$. The condition proxy exhibits a strict descent property:*

$$\frac{d}{dt}\rho_{cond}(W(t)) = -\eta\, \|\nabla_W \rho_{cond}(W(t))\|_F^2 \le 0.$$

Corollary 2 (Control over the Log-Condition Number) *The condition proxy ρ_{cond} and the true log-condition number $\log \kappa(W)$ are related by the identity $\log \kappa(W) = \rho_{cond}(W) + \frac{1}{2}\log(1 + \epsilon/\sigma_{\min}^2(W))$. Thus, the monotone decrease of ρ_{cond} guaranteed by Theorem 1 forces a non-increasing trend in $\log \kappa(W)$, bounded by an additive term that vanishes as $\epsilon \to 0$.*

Discrete-step behavior. While Theorem 1 is stated for gradient flow, in our experiments we observe a *discrete-step monotonic trend* for ρ_{cond} under standard optimizers (Sec. 4), supporting its use as a direct conditioning signal.

Note. The identity holds pointwise; the monotonicity conclusion applies almost everywhere along trajectories where ρ_{cond} is differentiable, and extends in the sense of energy dissipation inequalities using Clarke subgradients [7].

Lemma 3 (Moment Gradients are Bounded and Scale-Friendly) *Assume the spectral spread $\lambda_{\max}(G) - \lambda_{\min}(G) \ge \theta\, \lambda_{\max}(G)$ for some $\theta \in (0, 1]$. Then the Frobenius norm of the moment penalty's gradient satisfies*

$$\|\nabla_W \rho_{moment}\|_F \le \frac{C \cdot K}{\|W\|_2} + \mathcal{O}(\|W\|_2^{-3}).$$

Moreover, without the spread assumption one always has the general bound

$$\|\nabla_W \rho_{moment}\|_F \ \le\ \frac{C'}{\sqrt{n}\,\epsilon}\, K\, \|W\|_2,$$

so the claimed $1/\|W\|_2$ *decay is precisely the regime where the affine normalization is edge–dominated (i.e.,* $d = \Theta(\|W\|_2^2)$*).*

Remark 4 *When the spectrum is nearly degenerate (spread* $\to 0$*),* d *is set by* ϵ *and the bound becomes linear in* $\|W\|_2$*. In practice we keep* ϵ *small and rely on the condition proxy to widen the edges, quickly entering the favorable* $1/\|W\|_2$ *regime.*

Proposition 5 (Orthogonal Invariance) *The* CMR *penalty is invariant under orthogonal transformations. For any orthogonal matrices* Q, R *of appropriate dimensions,*

$$\rho_{cond}(QWR) = \rho_{cond}(W) \quad \textit{and} \quad \rho_{moment}(QWR) = \rho_{moment}(W).$$

4. Stylized VANISH Testbed

To isolate the effect of CMR on gradient stability, we consider a deliberately adversarial "VANISH" setting. We train a fully–connected MLP on MNIST with depth $L = 20$ and hidden width $d = 256$, using $\tanh$ activations and orthogonal initialization with a small spectral scale ("κ–stress"). This configuration is chosen so that standard training either fails completely or converges to a poor local minimum.
We compare three methods:

1. **Vanilla:** Adam with cross–entropy loss, no explicit spectral control.
2. **SN:** spectral normalization on every linear layer, enforcing $\|W^{(\ell)}\|_2 \le 1$.
3. **CMR:** Chebyshev Moment Regularization (cond + moment penalty) in fast mode with Hutchinson probes and the decoupled, capped mixing rule of Sec. 2.

All methods share the same optimizer hyperparameters, batch size, and initialization.

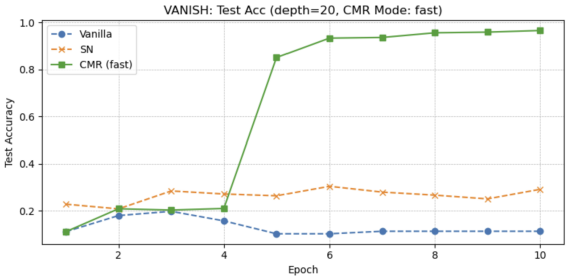

Figure 1: **VANISH testbed.** Test accuracy vs. epoch for a depth–20 MLP on MNIST. Vanilla collapses near chance, SN partially stabilizes but plateaus at low accuracy, while CMR (fast mode) exhibits a sharp recovery and reaches high test accuracy.

Test accuracy and gradients. Figure 1 shows test accuracy over epochs. In this κ–stressed regime, *Vanilla* quickly collapses to near–random performance (≈ 0.11) and never recovers. *SN* keeps training numerically stable but saturates around 0.25–0.30 accuracy, consistent with an overly contractive mapping. In contrast, CMR exhibits a sharp transition: after a few low–accuracy epochs it rapidly "revives" training, reaching ≈ 0.85 test accuracy by epoch 5 and $\gtrsim 0.96$ by epoch 10.

Along this trajectory, we monitor the global gradient norm $\|\nabla_\theta \mathcal{L}_{\text{task}}\|_2$. Vanilla frequently enters a vanishing regime with extremely small gradients; SN stays in a narrow band close to the lower edge, matching slow learning; CMR keeps gradient norms comfortably within a "stable band" where gradients are informative but non–exploding. Together with decreasing condition proxies of hidden layers, this supports the view that CMR turns a numerically pathological configuration into a well–behaved one without changing the architecture.

5. Ablation: Moments and Mixing Strategy

We now dissect which parts of CMR are responsible for the stabilization observed in Sec. 4. We keep the VANISH setup fixed and compare four variants:

- **Vanilla:** no spectral penalty.
- **CondOnly–Capped:** condition proxy only ($\alpha_2 = 0$) with capped mixing.
- **CondMoment–Capped (full CMR):** both condition and moment penalties with capped mixing.
- **CondMoment–Naive:** same spectral penalty as full CMR, but mixed via naive addition $\mathcal{L}_{\text{task}} + \lambda_t \mathcal{L}_{\text{spec}}$ (no decoupling or capping).

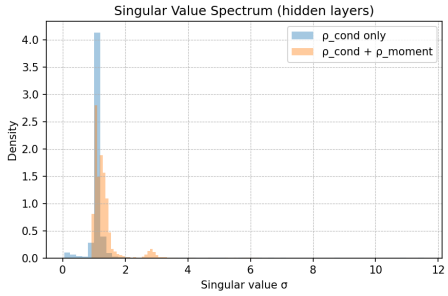

Figure 2: **Effect of higher–order moments.** Hidden–layer singular value spectra under CondOnly–Capped vs. CondMoment–Capped. The moment term removes a heavy tail and produces a compact, better–spread spectrum.

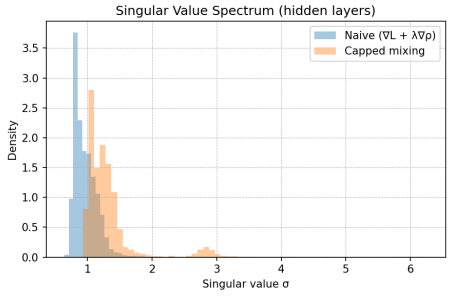

Figure 3: **Mixing strategy.** Singular value spectra for naive mixing $\nabla \mathcal{L}_{\text{task}} + \lambda_t \nabla \mathcal{L}_{\text{spec}}$ vs. decoupled capped mixing (Alg. 1).

5.1. Role of Higher–Order Moments

CondOnly–Capped and CondMoment–Capped both reduce the condition proxy and eventually escape the collapsed regime, but CondMoment–Capped does so earlier and with smoother training curves (cf. Fig. 1). The singular value histograms in Fig. 2 make the difference explicit: CondOnly compresses the condition number but leaves a heavy tail—a few large singular values and many very small ones. Adding the moment proxy substantially reshapes the spectrum: singular values concentrate in a compact interval and the tail is dramatically shortened. Equivalently, the Chebyshev moments $s_k(W)$ remain small for $k \geq 3$ only under full CMR, confirming that the moment term controls spectral *shape* beyond what the log–condition number can see.

5.2. Decoupled Capped Mixing vs. Naive Addition

Naive mixing uses

$$g_{\text{naive}} = \nabla_\theta \mathcal{L}_{\text{task}}(\theta) + \lambda_t \nabla_\theta \mathcal{L}_{\text{spec}}(\theta),$$

which allows the spectral component to dominate early in training. CMR instead uses the decoupled, capped rule of Alg. 1, enforcing $\|\gamma_t g_{\text{spec}}\| \leq \rho_{\text{spec}} \|g_{\text{task}}\|$.
Both CondMoment–Naive and CondMoment–Capped can reach high final accuracy (≈ 0.96–0.97), but their optimization traces differ. Naive mixing exhibits large gradient spikes and over–shrinks the spectrum, producing histograms with mass piled near the smallest singular values (Fig. 3). Capped mixing maintains a moderate spread of singular values and keeps gradients inside the stable band from Sec. 4, avoiding the "regularizer fights the task" regime. Overall, these ablations indicate that both ingredients are necessary for robust stabilization: (i) higher–order Chebyshev moments are needed to control the full spectral shape, and (ii) decoupled capped mixing prevents the spectral penalty from destabilizing or over–shrinking the network during optimization.

6. Related Work

Stability has been pursued via **architecture** (residual, normalization) [4, 6] and **initialization** [3], which help indirectly rather than optimizing spectra. Among **regularizers**, spectral norm constrains only $\|W\|_2$ [8], while orthogonality pushes near-isometries [2]. In contrast, CMR is a loss-level, orthogonally invariant regularizer that jointly controls spectral edges (log-condition) and higher-order shape (Chebyshev) with a capped mixing rule, offering a more direct handle on conditioning with descent and bounded-gradient guarantees.

References

[1] Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, New York, 1997.

[2] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 854–863, 2017.

[3] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[5] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[7] Adrian S. Lewis. Convex analysis on the hermitian matrices. *SIAM Journal on Optimization*, 6(1):164–177, 1996.

[8] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[9] Theodore J. Rivlin. *Chebyshev Polynomials*. Wiley, 2nd edition, 1990.

[10] Ji-Guang Sun. A note on simple non-zero singular values. *Journal of Computational Mathematics*, 6:258–266, 1988.

Appendix A. Why Moments Start at $k = 3$

Let $G = W^\top W$ with eigenvalues $\{\lambda_i\}_{i=1}^n$ and define the affine normalization $\widehat{G} = (G - cI)/d$ with $c = \frac{1}{2}(\lambda_{\max} + \lambda_{\min})$ and $d = \max\{\frac{1}{2}(\lambda_{\max} - \lambda_{\min}), \epsilon\}$ so that $\sigma(\widehat{G}) \subseteq [-1, 1]$. Chebyshev moments are $s_k = \frac{1}{n}\mathrm{Tr}(T_k(\widehat{G}))$ with $T_0(x) = 1$, $T_1(x) = x$, and $T_2(x) = 2x^2 - 1$. They satisfy

$$s_0 \equiv 1, \qquad s_1 = \tfrac{1}{n}\mathrm{Tr}(\widehat{G}) = \tfrac{\overline{\lambda} - c}{d}, \qquad s_2 = \tfrac{1}{n}\mathrm{Tr}(2\widehat{G}^2 - I) = 2 \cdot \tfrac{1}{n}\sum_i \widehat{\lambda}_i^2 - 1,$$

hence $\mathrm{Var}(\widehat{\lambda}) = \left(\frac{s_2+1}{2}\right) - s_1^2$. Thus s_0, s_1, s_2 encode mass/mean/variance of the *normalized* spectrum—quantities already fixed by the edge-based normalization (c, d) and largely governed by the condition proxy. Penalizing $k \leq 2$ would double-count edge/scale control and can interfere with ρ_{cond}. We therefore use $k \geq 3$ to isolate higher-order shape (tails, asymmetry, peaky structure), complementing edge control without redundancy.

Appendix B. Full Theoretical Results and Proofs

This appendix provides detailed derivations and proofs for the theoretical claims made in Section 3. We adopt the notation from the main text.

B.1. Gradient of the Condition Proxy

To prove our main results, we first require the explicit form of the gradient for $\rho_{\text{cond}}(W)$. We state the formula under simplicity of the extremal singular values; otherwise, subgradients exist and the identities hold almost everywhere [1, 7].

Lemma 6 (Gradient of ρ_{cond}) *Let $W = U\Sigma V^\top$ be the singular value decomposition of W. Assume the largest singular value $\sigma_{\max}(W)$ and smallest singular value $\sigma_{\min}(W)$ are simple. Let (u_1, v_1) and (u_r, v_r) be the corresponding pairs of left and right singular vectors. The gradient of the condition proxy is given by:*

$$\nabla_W \rho_{\text{cond}}(W) = \frac{1}{\sigma_{\max}(W)} u_1 v_1^\top - \frac{\sigma_{\min}(W)}{\sigma_{\min}(W)^2 + \epsilon} u_r v_r^\top.$$

Consequently, the squared Frobenius norm of the gradient is:

$$\|\nabla_W \rho_{\text{cond}}(W)\|_F^2 = \frac{1}{\sigma_{\max}(W)^2} + \left(\frac{\sigma_{\min}(W)}{\sigma_{\min}(W)^2 + \epsilon}\right)^2.$$

Proof The condition proxy is defined as $\rho_{\text{cond}}(W) = \log \sigma_{\max}(W) - \frac{1}{2}\log(\sigma_{\min}(W)^2 + \epsilon)$. The gradient of a simple singular value $\sigma_i(W)$ with corresponding vectors (u_i, v_i) is the rank-one matrix $\nabla_W \sigma_i(W) = u_i v_i^\top$. Applying the chain rule yields:

$$\begin{aligned}
\nabla_W \log \sigma_{\max}(W) &= \frac{1}{\sigma_{\max}(W)} \nabla_W \sigma_{\max}(W) = \frac{1}{\sigma_{\max}(W)} u_1 v_1^\top, \\
\nabla_W \left(\tfrac{1}{2}\log(\sigma_{\min}(W)^2 + \epsilon)\right) &= \frac{1}{2}\frac{1}{\sigma_{\min}(W)^2 + \epsilon} \nabla_W (\sigma_{\min}(W)^2) \\
&= \frac{2\sigma_{\min}(W)}{2(\sigma_{\min}(W)^2 + \epsilon)} \nabla_W \sigma_{\min}(W) = \frac{\sigma_{\min}(W)}{\sigma_{\min}(W)^2 + \epsilon} u_r v_r^\top.
\end{aligned}$$

Subtracting the second term from the first gives the gradient formula. For the Frobenius norm, we use the fact that singular vectors form orthonormal sets, meaning $\langle u_1 v_1^\top, u_r v_r^\top \rangle_F = \mathrm{Tr}(v_1 u_1^\top u_r v_r^\top) = (u_1^\top u_r)(v_1^\top v_r) = 0$ for $1 \neq r$. Thus, the squared norm is the sum of the squared norms of the two orthogonal rank-one components:

$$\|\nabla_W \rho_{\text{cond}}(W)\|_F^2 = \left\| \tfrac{1}{\sigma_{\max}} u_1 v_1^\top \right\|_F^2 + \left\| \tfrac{\sigma_{\min}}{\sigma_{\min}^2+\epsilon} u_r v_r^\top \right\|_F^2 = \tfrac{1}{\sigma_{\max}^2} \left\| u_1 v_1^\top \right\|_F^2 + \left(\tfrac{\sigma_{\min}}{\sigma_{\min}^2+\epsilon} \right)^2 \left\| u_r v_r^\top \right\|_F^2 .$$

Since $\left\| u_i v_i^\top \right\|_F^2 = \|u_i\|_2^2 \, \|v_i\|_2^2 = 1$, the result follows. ■

B.2. Proof of Theorem 1 and Corollary 2

Proof [Proof of Theorem 1] We analyze the time derivative of $\rho_{\text{cond}}(W(t))$ along the gradient flow $\dot{W}(t) = -\eta \nabla_W \rho_{\text{cond}}(W(t))$. By the chain rule:

$$\frac{d}{dt} \rho_{\text{cond}}(W(t)) = \left\langle \nabla_W \rho_{\text{cond}}(W(t)), \dot{W}(t) \right\rangle_F .$$

Substituting the definition of the gradient flow:

$$\frac{d}{dt} \rho_{\text{cond}}(W(t)) = \langle \nabla_W \rho_{\text{cond}}(W(t)), -\eta \nabla_W \rho_{\text{cond}}(W(t)) \rangle_F = -\eta \, \|\nabla_W \rho_{\text{cond}}(W(t))\|_F^2 .$$

Since the squared Frobenius norm is always non-negative, we have $\frac{d}{dt} \rho_{\text{cond}}(W(t)) \leq 0$. The descent is strict whenever $\nabla \rho_{\text{cond}} \neq 0$ (the generic case). ■

Proof [Proof of Corollary 2] The log-condition number is $\log \kappa(W) = \log \sigma_{\max}(W) - \log \sigma_{\min}(W)$. The condition proxy is $\rho_{\text{cond}}(W) = \log \sigma_{\max}(W) - \frac{1}{2} \log(\sigma_{\min}(W)^2 + \epsilon)$. We can write:

$$\begin{aligned}
\log \kappa(W) &= \left(\log \sigma_{\max} - \tfrac{1}{2} \log(\sigma_{\min}^2 + \epsilon) \right) + \tfrac{1}{2} \log(\sigma_{\min}^2 + \epsilon) - \log \sigma_{\min} \\
&= \rho_{\text{cond}}(W) + \tfrac{1}{2} \log(\sigma_{\min}^2 + \epsilon) - \tfrac{1}{2} \log(\sigma_{\min}^2) \\
&= \rho_{\text{cond}}(W) + \tfrac{1}{2} \log \left(\frac{\sigma_{\min}^2 + \epsilon}{\sigma_{\min}^2} \right) \\
&= \rho_{\text{cond}}(W) + \tfrac{1}{2} \log \left(1 + \frac{\epsilon}{\sigma_{\min}(W)^2} \right) .
\end{aligned}$$

This is the stated identity. Since $\rho_{\text{cond}}(W)$ decreases monotonically under the flow (Theorem 1), $\log \kappa(W)$ must also follow a non-increasing trend, perturbed only by the additive term which is positive and depends on the ratio $\epsilon / \sigma_{\min}^2$. ■

B.3. Proof of Lemma 3 (Moment Gradient Bounds)

Proof The moment penalty is $\rho_{\text{moment}}(W) = \sum_{k=3}^K w_k s_k(W)^2$. Its gradient is $\nabla_W \rho_{\text{moment}} = \sum_{k=3}^K 2 w_k s_k \nabla_W s_k$. We focus on bounding $\|\nabla_W s_k\|_F$. Recall $s_k = \frac{1}{n} \mathrm{Tr}(T_k(\widehat{G}))$ where $\widehat{G} = (G - cI)/d$ and $G = W^\top W$. Using the chain rule for matrix derivatives: $\nabla_W s_k = 2W \nabla_G s_k$, and $\nabla_G s_k = \frac{\partial s_k}{\partial \widehat{G}} \frac{\partial \widehat{G}}{\partial G}$. The main term is $\frac{\partial \mathrm{Tr}(T_k(\widehat{G}))}{\partial \widehat{G}} = T_k'(\widehat{G})$; by $T_k'(x) = k \, U_{k-1}(x)$ and

$\sup_{x\in[-1,1]} |U_{k-1}(x)| = k$, we have $\|T_k'(\cdot)\|_\infty \le k^2$ [9, Thm. 1.2]. Thus the dominant part of the gradient is

$$\nabla_W s_k \approx \frac{2}{nd}\, W\, T_k'(\widehat{G}).$$

Taking the Frobenius norm and using $\|AB\|_F \le \|A\|_2 \|B\|_F$ gives

$$\|\nabla_W s_k\|_F \lesssim \frac{2}{nd}\, \|W\|_2 \left\|T_k'(\widehat{G})\right\|_F \le \frac{2}{nd}\, \|W\|_2 \sqrt{n}\, \|T_k'(\widehat{G})\|_2 \le \frac{2k^2}{\sqrt{n}d}\, \|W\|_2 .$$

Under the spread assumption, $d = \Theta(\|W\|_2^2)$, yielding $\|\nabla_W s_k\|_F = \mathcal{O}(k^2/\|W\|_2)$. Summing over $k = 3, \dots, K$ and absorbing constants gives the stated bound; terms from $\nabla c, \nabla d$ scale as d^{-2} and become $\mathcal{O}(\|W\|_2^{-3})$ after multiplying by W. If the spread is negligible, $d = \epsilon$ and the alternative bound follows. ■

B.4. Proof of Proposition 5 (Orthogonal Invariance)

Proof Let Q, R be orthogonal matrices.

1. **Condition Proxy ρ_{cond}:** The singular values of a matrix W are defined from the eigenvalues of $W^\top W$. The singular values of QWR are defined from the eigenvalues of $(QWR)^\top (QWR) = R^\top W^\top Q^\top QWR = R^\top (W^\top W) R$. Since $W^\top W$ and $R^\top (W^\top W) R$ are related by an orthogonal similarity transformation, they have the same eigenvalues. Therefore, W and QWR have the same singular values. Since $\rho_{\text{cond}}(W)$ depends only on $\sigma_{\max}(W)$ and $\sigma_{\min}(W)$, it follows that $\rho_{\text{cond}}(QWR) = \rho_{\text{cond}}(W)$.

2. **Moment Proxy ρ_{moment}:** Let $G_W = W^\top W$ and $G_{QWR} = (QWR)^\top (QWR) = R^\top G_W R$. As shown above, G_W and G_{QWR} have the same set of eigenvalues. The affine normalization constants $c = \frac{1}{2}(\lambda_{\max} + \lambda_{\min})$ and $d = \max\{\frac{1}{2}(\lambda_{\max} - \lambda_{\min}), \epsilon\}$ depend only on the extremal eigenvalues, and are thus identical for G_W and G_{QWR}. Let's call them c and d. The normalized Gram matrices are $\widehat{G}_W = (G_W - cI)/d$ and $\widehat{G}_{QWR} = (G_{QWR} - cI)/d = (R^\top G_W R - cR^\top IR)/d = R^\top (G_W - cI) R/d = R^\top \widehat{G}_W R$. By the functional calculus for matrix polynomials [5, Chap. 1], $T_k(R^\top \widehat{G}_W R) = R^\top T_k(\widehat{G}_W) R$. Using the cyclic property of the trace,

$$s_k(QWR) = \frac{1}{n}\mathrm{Tr}(R^\top T_k(\widehat{G}_W) R) = \frac{1}{n}\mathrm{Tr}(RR^\top T_k(\widehat{G}_W)) = \frac{1}{n}\mathrm{Tr}(T_k(\widehat{G}_W)) = s_k(W).$$

   Since the moments s_k are invariant, the penalty $\rho_{\text{moment}}(W) = \sum_{k=3}^K w_k s_k(W)^2$ is also invariant.

This completes the proof that the entire CMR penalty is orthogonally invariant. ■

Appendix C. Additional Diagnostics for the VANISH Testbed

For completeness we report additional diagnostics for the VANISH experiments.

Gradient norms. First, we plot the epoch–wise global gradient norm $\|\nabla_\theta L_{\text{task}}\|_2$ for all methods. Vanilla frequently falls below the stable band, confirming gradient vanishing. SN stays within the band but very close to its lower edge, matching its slow improvement in accuracy. All CMR variants keep gradient norms within a comfortable interior of the band, with CondMoment–Capped showing the most stable trajectory.

Layer–wise statistics. Second, we track per–layer condition proxies and singular values at early and late epochs. Vanilla exhibits very skewed spectra with large condition proxies in deep layers. SN enforces $\sigma_{\max} \approx 1$ but still drives many singular values to near zero, especially in the middle layers. CondOnly–Capped reduces condition proxies but retains a noticeable tail. CondMoment–Capped yields uniformly compact spectra and low condition proxies across all hidden layers, matching the design of the penalty.

Taken together, these diagnostics corroborate the main findings of Secs. 4 and 5: without CMR, the VANISH testbed suffers from either vanishing gradients or excessively skewed spectra, while full CMR with capped mixing produces stable gradients and compact spectra throughout the network.

Appendix D. Main Algorithm

Algorithm 1: CMR-SGD with decoupled, capped spectral gradients

Input: model θ, task loss $\mathcal{L}_{\text{task}}$, weights $(\lambda, \alpha_1, \alpha_2)$, cap ρ_{spec}, warmup T_{w}, stabilizer δ

for $t = 0, 1, 2, \ldots$ **do**

$g_{\text{task}} \leftarrow \nabla_\theta \mathcal{L}_{\text{task}}(\theta)$;

$\lambda_t \leftarrow \lambda \cdot \min\{1,\ t/T_{\text{w}}\}$;

$g_{\text{spec,raw}} \leftarrow \nabla_\theta \sum_\ell \big(\alpha_1 \rho_{\text{cond}}(W^{(\ell)}) + \alpha_2 \rho_{\text{moment}}(W^{(\ell)})\big)$;

$g_{\text{spec}} \leftarrow \lambda_t\, g_{\text{spec,raw}}$;

$\gamma_t \leftarrow \min\{1,\ \rho_{\text{spec}} \|g_{\text{task}}\| / (\|g_{\text{spec}}\| + \delta)\}$;

$g_t \leftarrow g_{\text{task}} + \gamma_t g_{\text{spec}}$;

Update θ with gradient g_t;

end