Generalization, Optimization, and Preconditioning Dynamics for Encoder–Decoder In-Context Least Squares

Abstract

All notation uses two times. Outer time s is pretraining of the attention/encoder parameters. Inner time ℓ is the Richardson solver run inside one prompt with frozen parameters. Linear attention gives an exact global preconditioner trajectory $s \mapsto P_s$. Softmax attention gives an exact nonlinear update map and a local Jacobian preconditioner.

Contents

1 Objects

1.1 Task as one prompt

$$z \sim \mathcal{N}(0, \Sigma_z), \qquad z \in \mathbb{R}^r.$$

$$f \sim \mathrm{GP}(0, k_f), \qquad f \in \mathcal{H}.$$

$$A(z) = A_0 + \sum_{k=1}^{r} z_k A_k.$$

$$u = A(z)^{-1} f.$$

$$\Pi_z = \{(u_i, f_i)\}_{i=1}^{M}, \qquad f_i \overset{\text{iid}}{\sim} \mathrm{GP}(0, k_f), \qquad u_i = A(z)^{-1} f_i.$$

$$\boxed{\text{one task} = \text{one latent } z = \text{one prompt } \Pi_z.}$$

$$v_a \in \mathcal{H}, \qquad a = 1, \dots, q.$$

$$\langle v_a, f_i - A_0 u_i \rangle = \sum_{k=1}^{r} z_k \langle v_a, A_k u_i \rangle.$$

$$G(\Pi_z) z = b(\Pi_z),$$

$$G_{(i,a),k} = \langle v_a, A_k u_i \rangle, \qquad b_{(i,a)} = \langle v_a, f_i - A_0 u_i \rangle.$$

$$G_\theta(\Pi_z) z \simeq b_\theta(\Pi_z).$$

1.2 Finite least-squares reduction

$$x_i \sim \mathcal{N}(0, \Sigma_x), \qquad y_i = x_i^\top \beta + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, \sigma^2).$$

$$\beta = U_\star z, \qquad U_\star^\top U_\star = I_r.$$

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_M^\top \end{bmatrix}, \qquad y = (y_1, \dots, y_M)^\top.$$

$$H_M = \frac{1}{M} X^\top X + \lambda I_K, \qquad c_M = \frac{1}{M} X^\top y.$$

$$\beta_M^\dagger = H_M^{-1} c_M.$$

$$U \in \mathbb{R}^{K\times r}, \qquad U^\top U = I_r.$$

$$A_U = U^\top H_M U, \qquad d_U = U^\top c_M.$$

$$z_U^\dagger = A_U^{-1} d_U, \qquad \beta_U^\dagger = U z_U^\dagger.$$

2 Two Times

$$s = 0, 1, \ldots \qquad \theta_s,\ U_s,\ P_s,\ P_{z,s} \qquad \text{outer pretraining time.}$$

$$\ell = 0, 1, \ldots, L \qquad z_{\ell,s} \qquad \text{inner solver time at frozen } s.$$

$$\boxed{s \mapsto P_s \text{ is attention/preconditioner learning.}}$$

$$\boxed{\ell \mapsto z_{\ell,s} \text{ is solver evaluation at fixed } P_s.}$$

3 Decoder-Only Linear Attention

3.1 Population objective

$$d \in \mathbb{R}^K, \qquad z_\star \in \mathbb{R}^K, \qquad \widehat{z} = Pd.$$

$$\mathcal{L}_{\mathrm{dec}}(P) = \frac{1}{2K}\mathbb{E}\|Pd - z_\star\|_2^2.$$

$$C = \mathbb{E}[dd^\top], \qquad B = \mathbb{E}[z_\star d^\top].$$

$$\nabla\mathcal{L}_{\mathrm{dec}}(P) = \frac{1}{K}(PC - B).$$

$$P_\star C = B, \qquad P_\star = BC^\dagger.$$

$$P_{s+1} = P_s - \eta K \nabla\mathcal{L}_{\mathrm{dec}}(P_s) = P_s - \eta(P_s C - B).$$

$$E_s = P_s - P_\star.$$

$$\boxed{E_{s+1} = E_s(I - \eta C).}$$

$$\boxed{P_s - P_\star = (P_0 - P_\star)(I - \eta C)^s.}$$

$$C = Q\Lambda Q^\top, \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K).$$

$$\mathcal{R}_{\mathrm{dec}}(P) = \frac{1}{K}\mathbb{E}\|Pd - z_\star\|_2^2.$$

$$\mathcal{R}_{\mathrm{dec}}(P) - \mathcal{R}_{\mathrm{dec}}(P_\star) = \frac{1}{K}\,\mathrm{Tr}\{(P - P_\star)C(P - P_\star)^\top\}.$$

$$\boxed{\mathcal{R}_{\mathrm{dec}}(P_s) - \mathcal{R}_{\mathrm{dec}}(P_\star) = \frac{1}{K}\sum_{i=1}^{K} \lambda_i(1-\eta\lambda_i)^{2s}\|(P_0 - P_\star)q_i\|_2^2.}$$

$$\mathcal{E}_{s,i} = \lambda_i\|(P_s - P_\star)q_i\|_2^2, \qquad w_{s,i} = \frac{\mathcal{E}_{s,i}}{\sum_j \mathcal{E}_{s,j}}.$$

$$Hz^\dagger = d.$$

$$z_{\ell+1,s} = z_{\ell,s} + P_s(d - Hz_{\ell,s}), \qquad z_{0,s} = 0.$$

$$z_{L,s} = \sum_{j=0}^{L-1}(I - P_sH)^j P_s d.$$

$$\boxed{z_{L,s} - z^\dagger = -(I - P_sH)^L z^\dagger.}$$

$$\rho_s(H) = \|I - P_sH\|_2.$$

$$\|z_{L,s} - z^\dagger\|_2 \le \rho_s(H)^L\|z^\dagger\|_2.$$

$$\mathcal{R}_L(P_s) = \frac{1}{K}\mathbb{E}\|z_{L,s} - z_\star\|_2^2.$$

$$\mathcal{R}_1(P_s) = \frac{1}{K}\mathbb{E}\|P_s d - z_\star\|_2^2.$$

3.2 Finite-task generalization

$$\mathcal{D}_N = \{(d_n, z_n)\}_{n=1}^N.$$

$$D = [d_1, \ldots, d_N] \in \mathbb{R}^{K\times N}, \qquad Z = [z_1, \ldots, z_N] \in \mathbb{R}^{K\times N}.$$

$$\widehat{\mathcal{L}}_N(P) = \frac{1}{2KN}\|PD - Z\|_F^2.$$

$$\widehat{P}_N = ZD^\top(DD^\top)^\dagger.$$

$$\widehat{C}_N = \frac{1}{N}DD^\top, \qquad \widehat{B}_N = \frac{1}{N}ZD^\top.$$

$$\widehat{P}_N\widehat{C}_N = \widehat{B}_N.$$

$$\boxed{\mathcal{R}_{\mathrm{dec}}(\widehat{P}_N) - \mathcal{R}_{\mathrm{dec}}(P_\star) = \frac{1}{K}\,\mathrm{Tr}\{(\widehat{P}_N - P_\star)C(\widehat{P}_N - P_\star)^\top\}.}$$

$$d_{\mathrm{eff}}(\lambda) = \mathrm{Tr}\{C(C + \lambda I)^{-1}\}.$$

$$\mathbb{E}_{\mathcal{D}_N}\big[\mathcal{R}_{\mathrm{dec}}(\widehat{P}_N) - \mathcal{R}_{\mathrm{dec}}(P_\star)\big] \asymp \frac{\sigma_{\mathrm{task}}^2}{K}\frac{d_{\mathrm{eff}}}{N} \qquad \text{when } N \gg d_{\mathrm{eff}}.$$

4 Prompt Generalization

$$\widehat{\Sigma}_M = \frac{1}{M} X^\top X.$$

$$\widetilde{\Sigma}_M = \Sigma_x^{-1/2} \widehat{\Sigma}_M \Sigma_x^{-1/2}.$$

$$\gamma = \frac{K}{M}.$$

$$\lambda_\pm^{\mathrm{MP}} = (1 \pm \sqrt{\gamma})^2.$$

$$\mathrm{spec}(\widetilde{\Sigma}_M) \subset \left[(1 - \sqrt{\gamma} - t)^2,\ (1 + \sqrt{\gamma} + t)^2\right]$$

with probability at least $1 - 2e^{-Mt^2/2}$ for Gaussian design.

$$H_M = H_\infty + \Delta_M, \qquad H_\infty = \Sigma_x + \lambda I.$$

$$\|\Delta_M\|_2 = \|\widehat{\Sigma}_M - \Sigma_x\|_2.$$

$$H_M^{-1} - H_\infty^{-1} = -H_\infty^{-1} \Delta_M H_\infty^{-1} + H_\infty^{-1} \Delta_M H_M^{-1} \Delta_M H_\infty^{-1}.$$

$$\|H_M^{-1} - H_\infty^{-1}\|_2 \le \frac{\|H_\infty^{-1}\|_2^2 \|\Delta_M\|_2}{1 - \|H_\infty^{-1}\|_2 \|\Delta_M\|_2}.$$

$$\mathcal{R}_{\mathrm{prompt}}(M) - \mathcal{R}_{\mathrm{prompt}}(\infty) = \mathcal{O}\left(\frac{d_{\mathrm{eff}}^{\mathrm{prompt}}}{M}\right) \quad \text{for fixed spectrum and high probability.}$$

5 Encoder-Only Low-Rank Recovery

5.1 Linear residual dynamics

$$b = G_\star z + \eta_b, \qquad \mathbb{E}[\eta_b z^\top] = 0.$$

$$\mathcal{L}_{\mathrm{enc}}(G) = \frac{1}{2} \mathbb{E} \|Gz - b\|_2^2.$$

$$\Sigma_z = \mathbb{E}[z z^\top], \qquad \Sigma_{bz} = \mathbb{E}[b z^\top].$$

$$G_\star \Sigma_z = \Sigma_{bz}.$$

$$\nabla \mathcal{L}_{\mathrm{enc}}(G) = (G\Sigma_z - \Sigma_{bz}).$$

$$G_{s+1} = G_s - \eta_e (G_s \Sigma_z - \Sigma_{bz}).$$

$$\boxed{G_s - G_\star = (G_0 - G_\star)(I - \eta_e \Sigma_z)^s.}$$

$$\mathcal{R}_{\mathrm{enc}}(G) = \mathbb{E} \|Gz - b\|_2^2.$$

$$\boxed{\mathcal{R}_{\mathrm{enc}}(G_s) - \mathcal{R}_{\mathrm{enc}}(G_\star) = \mathrm{Tr}\{(G_s - G_\star)\Sigma_z(G_s - G_\star)^\top\}.}$$

$$\Sigma_z = V\,\mathrm{diag}(\mu_1, \ldots, \mu_r)V^\top.$$

$$\mathcal{R}_{\mathrm{enc}}(G_s) - \mathcal{R}_{\mathrm{enc}}(G_\star) = \sum_i \mu_i(1 - \eta_e\mu_i)^{2s}\|(G_0 - G_\star)v_i\|_2^2.$$

5.2 Subspace recovery

$$\Sigma_\beta = \mathbb{E}[\beta\beta^\top] = U_\star\Sigma_z U_\star^\top.$$

$$\widehat{\Sigma}_{\beta,N} = \frac{1}{N}\sum_{n=1}^{N}\widehat{\beta}_n\widehat{\beta}_n^\top.$$

$$U_N = \mathrm{Top}_r(\widehat{\Sigma}_{\beta,N}).$$

$$\gamma_{\mathrm{enc}} = \lambda_r(\Sigma_\beta) - \lambda_{r+1}(\Sigma_\beta).$$

$$\boxed{\|\sin\Theta(U_N, U_\star)\|_2 \le \frac{\|\widehat{\Sigma}_{\beta,N} - \Sigma_\beta\|_2}{\gamma_{\mathrm{enc}}}.}$$

$$\|\widehat{\Sigma}_{\beta,N} - \Sigma_\beta\|_2 = \mathcal{O}_{\mathbb{P}}\left(\|\Sigma_\beta\|_2\sqrt{\frac{r_{\mathrm{eff}}(\Sigma_\beta)}{N}}\right).$$

$$r_{\mathrm{eff}}(\Sigma_\beta) = \frac{\mathrm{Tr}(\Sigma_\beta)}{\|\Sigma_\beta\|_2}.$$

$$\mathcal{R}_{\mathrm{floor}}(U) = \mathbb{E}\|(I - UU^\top)\beta\|_{\Sigma_x}^2.$$

$$\mathcal{R}_{\mathrm{floor}}(U_N) \lesssim \|\Sigma_x\|_2\,\mathrm{Tr}(\Sigma_z)\|\sin\Theta(U_N, U_\star)\|_F^2.$$

6 Encoder + Decoder

6.1 Latent system

$$U_s = \mathrm{Enc}_{\phi_s}(\Pi_z).$$

$$A_s = U_s^\top H_M U_s, \qquad d_s = U_s^\top c_M.$$

$$z_s^\dagger = A_s^{-1}d_s, \qquad \beta_s^\dagger = U_s z_s^\dagger.$$

$$z_{\ell+1,s} = z_{\ell,s} + P_{z,s}(d_s - A_s z_{\ell,s}), \qquad z_{0,s} = 0.$$

$$\boxed{z_{L,s} - z_s^\dagger = -(I - P_{z,s}A_s)^L z_s^\dagger.}$$

$$\rho_{z,s}(H_M) = \|I - P_{z,s}A_s\|_2.$$

$$\|z_{L,s} - z_s^\dagger\|_2 \le \rho_{z,s}(H_M)^L\|z_s^\dagger\|_2.$$

6.2 Moving decoder oracle

$$C_s = \mathbb{E}[d_s d_s^\top \mid U_s], \qquad B_s = \mathbb{E}[z_s^\dagger d_s^\top \mid U_s].$$

$$P_z^\star(U_s) C_s = B_s.$$

$$P_z^\star(U_s) = B_s C_s^\dagger.$$

$$\mathcal{L}_{\mathrm{dec}}(P_z; U_s) = \frac{1}{2r}\mathbb{E}[\|P_z d_s - z_s^\dagger\|_2^2 \mid U_s].$$

$$P_{z,s+1} = P_{z,s} - \eta_d (P_{z,s} C_s - B_s).$$

$$\boxed{P_{z,s+1} - P_z^\star(U_s) = \big(P_{z,s} - P_z^\star(U_s)\big)(I - \eta_d C_s).}$$

$$\boxed{P_{z,s+1} - P_z^\star(U_{s+1}) = \big(P_{z,s} - P_z^\star(U_s)\big)(I - \eta_d C_s) + P_z^\star(U_s) - P_z^\star(U_{s+1}).}$$

$$\text{tracking error} = \text{contraction error} + \text{oracle drift}.$$

6.3 Risk decomposition

$$\beta_{L,s} = U_s z_{L,s}.$$

$$\mathcal{R}_{\mathrm{e2e}}(s) = \mathbb{E}\|\beta_{L,s} - \beta\|_{\Sigma_x}^2.$$

$$\mathcal{R}_{\mathrm{floor}}(s) = \mathbb{E}\|\beta_s^\dagger - \beta\|_{\Sigma_x}^2.$$

$$\mathcal{R}_{\mathrm{dec}}(s) = \mathbb{E}\|U_s(z_{L,s} - z_s^\dagger)\|_{\Sigma_x}^2.$$

$$\boxed{\mathcal{R}_{\mathrm{e2e}}(s) = \mathcal{R}_{\mathrm{floor}}(s) + \mathcal{R}_{\mathrm{dec}}(s) + 2\mathbb{E}\langle \beta_s^\dagger - \beta,\ U_s(z_{L,s} - z_s^\dagger)\rangle_{\Sigma_x}.}$$

$$\mathcal{R}_{\mathrm{dec}}(s) \le \|\Sigma_x\|_2 \mathbb{E}\|z_{L,s} - z_s^\dagger\|_2^2 \le \|\Sigma_x\|_2 \mathbb{E}[\rho_{z,s}(H_M)^{2L}\|z_s^\dagger\|_2^2].$$

6.4 Stop-gradient triangular dynamics

$$\mathcal{L}(G, P) = \mathcal{L}_{\mathrm{enc}}(G) + \mathcal{L}_{\mathrm{dec}}(P; \mathrm{sg}(G)).$$

$$\dot{G}_t = \Sigma_{bz} - G_t \Sigma_z.$$

$$\dot{P}_t = B(G_t) - P_t C(G_t).$$

$$\boxed{\text{encoder drives the moving decoder distribution, decoder tracks its oracle.}}$$

6.5 Commuting scalar closure

$$\Sigma_z,\ C(G_t),\ B(G_t),\ A(G_t) \quad \text{diagonal in one fixed basis.}$$

$$g_i(s) = g_i^\star + (g_i(0) - g_i^\star)(1 - \eta_e \sigma_i)^s.$$

$$p_i(s+1) = (1 - \eta_d c_i(s)) p_i(s) + \eta_d b_i(s).$$

$$\boxed{p_i(s) = p_i(0) \prod_{t<s} (1 - \eta_d c_i(t)) + \sum_{\tau<s} \eta_d b_i(\tau) \prod_{t=\tau+1}^{s-1} (1 - \eta_d c_i(t)).}$$

$$\mathcal{R}_{\mathrm{dec}}(s, L) = \sum_i w_i(s)(1 - p_i(s) a_i(s))^{2L}.$$

6.6 Non-commuting exact matrix dynamics

$$E_{s+1} = E_s(I - \eta C_s) + D_s,$$

$$D_s = P_z^\star(U_s) - P_z^\star(U_{s+1}).$$

$$\boxed{E_s = E_0 \prod_{t=0}^{s-1} (I - \eta C_t) + \sum_{\tau=0}^{s-1} D_\tau \prod_{t=\tau+1}^{s-1} (I - \eta C_t).}$$

$$\prod_{t=a}^{b} M_t = M_a M_{a+1} \cdots M_b.$$

$$C_t C_{t'} \neq C_{t'} C_t \quad \Longrightarrow \quad \text{no fixed scalar eigenmode closure.}$$

7 No Single Scalar State in General

$$C q_i = \lambda_i q_i, \qquad e_0 = \alpha q_i.$$

$$R_0 = e_0^\top C e_0 = \alpha^2 \lambda_i.$$

$$e_1 = (I - \eta C) e_0.$$

$$R_1 = e_1^\top C e_1 = (1 - \eta \lambda_i)^2 R_0.$$

$$R_0^{(i)} = R_0^{(j)}, \qquad \lambda_i \neq \lambda_j \quad \Longrightarrow \quad R_1^{(i)} \neq R_1^{(j)}.$$

risk alone is not a closed state variable.

exact closure requires spectral overlaps or full matrices.

8 Softmax and Nonlinear Attention

$$r_{\ell,s} = d - Hz_{\ell,s}.$$

Linear attention:

$$\Delta_{\theta_s}(z, H, d) = P_s r_{\ell,s}.$$

Softmax/nonlinear attention:

$$\Delta_{\theta_s}(z, H, d) = F_{\theta_s}(z, H, d) - z.$$

$$z_{\ell+1,s} = z_{\ell,s} + \Delta_{\theta_s}(z_{\ell,s}, H, d).$$

$$\boxed{\mathcal{P}_{\theta_s} : (z, H, d) \mapsto \Delta_{\theta_s}(z, H, d)}$$

$$\boxed{J_{\theta_s}(z, H, d) = \frac{\partial \Delta_{\theta_s}(z, H, d)}{\partial r}, \qquad r = d - Hz.}$$

$$\Delta_{\theta_s}(z, H, d) = P_s(d - Hz) \text{ for all } (z, H, d) \quad \Longleftrightarrow \quad \text{global matrix } P_s.$$

$$\Delta_{\theta_s} \text{ nonlinear} \quad \Longrightarrow \quad \text{exact object is } \mathcal{P}_{\theta_s} \text{ or } J_{\theta_s}(z, H, d).$$

$$P_s^{\mathrm{proj}} = \arg\min_P \mathbb{E}\|\Delta_{\theta_s}(z, H, d) - P(d - Hz)\|_2^2.$$

$$P_s^{\mathrm{proj}} = \mathbb{E}[\Delta_{\theta_s} r^\top]\mathbb{E}[rr^\top]^\dagger.$$

P_s^{proj} is a diagnostic projection, not the exact nonlinear model.

9 Replica and Spectral Order Parameters

$$\gamma = \frac{K}{M}, \qquad \psi = \frac{d_{\mathrm{eff}}}{N}.$$

$$m_\gamma(z) = \int \frac{1}{\lambda - z}\, d\mu_{\mathrm{MP},\gamma}(\lambda).$$

$$\mathrm{supp}(\mu_{\mathrm{MP},\gamma}) = [(1 - \sqrt{\gamma})^2, (1 + \sqrt{\gamma})^2].$$

$$\kappa(H_M) = \frac{\lambda_{\max}(H_M)}{\lambda_{\min}(H_M)}.$$

$$\rho_s^{\mathrm{pop}} = \rho(I - P_s H_\infty).$$

$$\rho_{z,s}^{\mathrm{pop}} = \rho(I - P_{z,s} A_{\infty,s}).$$

$$\mathrm{srank}(P) = \frac{\|P\|_F^2}{\|P\|_2^2}.$$

$$E_{\mathrm{top,k}}(P) = \frac{\sum_{i=1}^k \sigma_i(P)^2}{\sum_i \sigma_i(P)^2}.$$

$$\Omega_{s,i} = \frac{\lambda_i \|(P_s - P_\star)q_i\|_2^2}{\sum_j \lambda_j \|(P_s - P_\star)q_j\|_2^2}.$$

$$H(\Omega_s) = -\frac{1}{\log K} \sum_i \Omega_{s,i} \log \Omega_{s,i}.$$

10 Task Diversity and Difficulty

$$\text{task diversity} \quad \Longleftrightarrow \quad \operatorname{spec}(\Sigma_z),\ r_{\text{eff}}(\Sigma_z),\ N.$$

$$r_{\text{eff}}(\Sigma_z) = \frac{\operatorname{Tr}(\Sigma_z)}{\|\Sigma_z\|_2}.$$

$$H_z = -\sum_i \frac{\lambda_i(\Sigma_z)}{\operatorname{Tr}\Sigma_z} \log \frac{\lambda_i(\Sigma_z)}{\operatorname{Tr}\Sigma_z}.$$

$$\text{operator difficulty} \quad \Longleftrightarrow \quad \kappa(A(z)),\ \kappa(H_M),\ \sigma^2,\ \lambda,\ M.$$

$$\text{decoder difficulty} \quad \Longleftrightarrow \quad \rho(I - P_s H_M),\ \kappa(P_s H_M),\ L.$$

$$\text{encoder difficulty} \quad \Longleftrightarrow \quad \gamma_{\text{enc}}^{-1},\ \|\widehat{\Sigma}_{\beta,N} - \Sigma_\beta\|_2,\ \|(I - U_s U_s^\top)U_\star\|_2.$$

$$\text{GP prompt diversity} \quad \Longleftrightarrow \quad \operatorname{spec}(T_{k_f}),\ d_{\text{eff}}^{\text{GP}}(\tau) = \operatorname{Tr}\{T_{k_f}(T_{k_f} + \tau I)^{-1}\}.$$

11 Experimental Certificates

$$\max_s \left|\mathcal{R}_{\text{obs,s}}^{\text{dec}} - \mathcal{R}_{\text{pred,s}}^{\text{dec}}\right| = 2.0380 \cdot 10^{-3}.$$

$$\max_s \frac{|\mathcal{R}_{\text{obs,s}}^{\text{dec}} - \mathcal{R}_{\text{pred,s}}^{\text{dec}}|}{\mathcal{R}_{\text{pred,s}}^{\text{dec}}} = 4.3852 \cdot 10^{-3}.$$

$$\mathcal{R}_{\text{obs},800}^{\text{dec}} = 1.624536 \cdot 10^{-1}, \qquad \mathcal{R}_{\text{pred},800}^{\text{dec}} = 1.619130 \cdot 10^{-1}.$$

$$\frac{\|P_{800} - P_\star\|_F}{\|P_\star\|_F} = 8.2055 \cdot 10^{-3}.$$

$$\mathcal{R}_L(P_{800}) = 9.1478 \cdot 10^{-5}, \qquad \bar{\rho}_{800} = 7.9563 \cdot 10^{-1}.$$

$$\max_s \left|\mathcal{R}_{\text{obs,s}}^{\text{enc}} - \mathcal{R}_{\text{pred,s}}^{\text{enc}}\right| \ \text{log-ratio} = 1.3184 \cdot 10^{-2}.$$

$$\mathcal{R}_{\text{enc,final}} = 1.3095 \cdot 10^{-11}, \qquad \mathcal{R}_{\text{enc,pred,final}} = 1.3122 \cdot 10^{-11}.$$

$$\|G_{\text{final}} - G_\star\|/\|G_\star\| = 1.6295 \cdot 10^{-5}, \qquad \widehat{\sin^2\Theta}_{\text{final}} = 1.9699 \cdot 10^{-9}.$$

$$\max_s \left|\mathcal{R}_{\text{obs,s}}^{\text{enc+dec}} - \mathcal{R}_{\text{exact,s}}^{\text{enc+dec}}\right| = 2.7756 \cdot 10^{-17}.$$

$$\mathcal{R}_{\text{e2e},420} = 4.4115 \cdot 10^{-3}, \qquad \mathcal{R}_{\text{floor},420} = 4.3891 \cdot 10^{-3}, \qquad \mathcal{R}_{\text{dec},420} = 2.0853 \cdot 10^{-6}.$$

$$\frac{\|P_{z,420} - P_z^\star(U_{420})\|_F}{\|P_z^\star(U_{420})\|_F} = 3.1937 \cdot 10^{-2}.$$

$$\sin^2(U_{420}, U_\star) = 3.0853 \cdot 10^{-2}.$$

$$\text{scalar proof ladder max errors:} \quad 4.337 \cdot 10^{-18},\ 6.939 \cdot 10^{-18},\ 5.551 \cdot 10^{-17}.$$

$$\text{counterexample:} \qquad R_0^A = R_0^B = 1, \qquad R_1^A = 0.6724, \qquad R_1^B = 0.1936.$$

12 Plots: Detailed Reading

12.1 Outer preconditioning dynamics

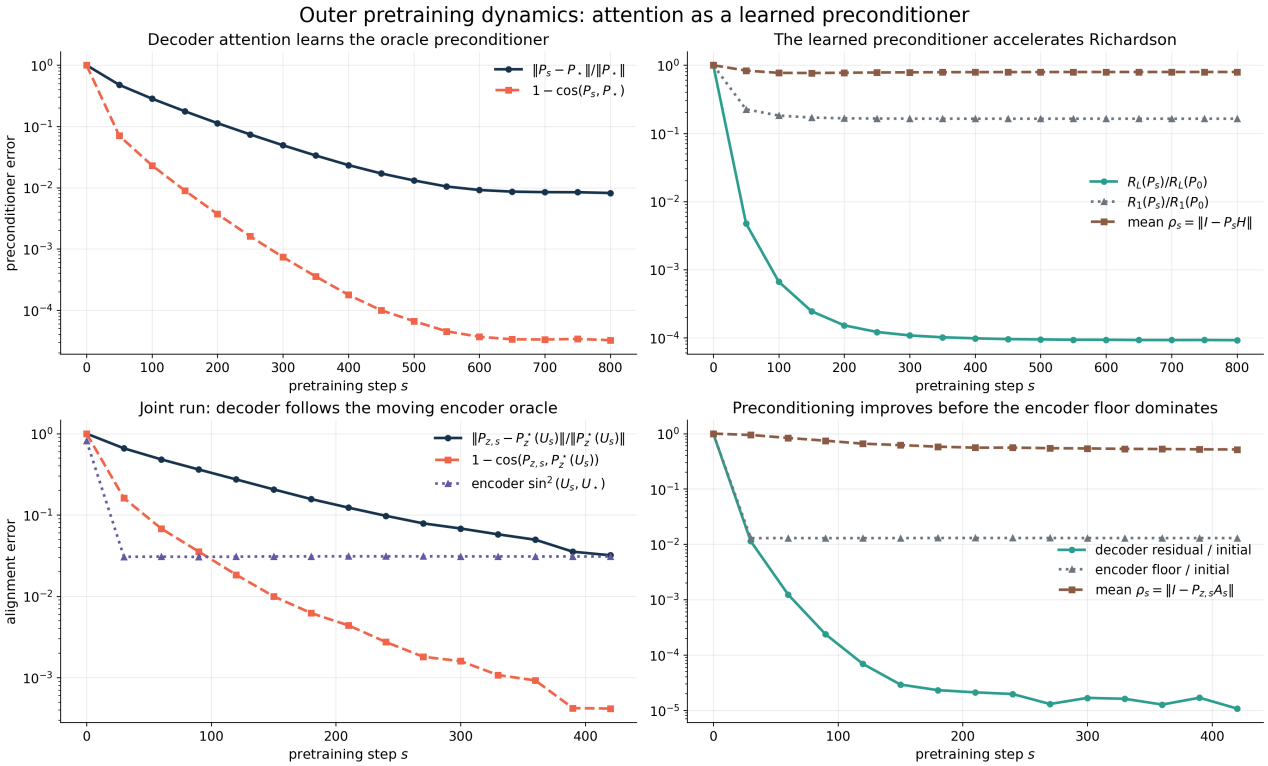


Figure 1: Outer pretraining dynamics: attention as learned preconditioning.

This is the central diagnostic because the horizontal axis is the outer pretraining time s, not the inner Richardson iteration ℓ. The top-left panel measures the decoder-only learned matrix P_s against the population oracle $P_\star$. The relative Frobenius error decreases from order one to below 10^{-2}, and the cosine defect decreases to about $3 \cdot 10^{-5}$. This confirms that the attention layer is not merely executing an unpreconditioned least-squares method; during pretraining it learns the preconditioning matrix itself.

The top-right panel evaluates what the learned P_s buys for a fixed-depth Richardson decoder. The one-step/direct risk changes moderately because a single multiplication $P_s d$ is limited by the irreducible floor of the direct estimator. In contrast, the depth-L Richardson risk collapses by several orders of magnitude. This is the preconditioning effect: the learned matrix changes the contraction geometry of the repeated residual update.

The bottom-left panel is the corresponding joint encoder+decoder reading. The learned latent decoder $P_{z,s}$ is compared to the moving oracle $P_z^\star(U_s)$, because the encoder subspace U_s changes during training. Therefore the correct target is not a fixed matrix. The plot shows that the decoder tracks this moving oracle while the encoder subspace quickly reaches its finite-sample floor.

The bottom-right panel separates the two errors. The decoder residual collapses rapidly, while the encoder floor changes much less after early training. This is the empirical version of

$$\mathcal{R}_{\mathrm{e2e}}(s) \approx \mathcal{R}_{\mathrm{floor}}(s) + \mathcal{R}_{\mathrm{dec}}(s),$$

with the decoder term becoming negligible relative to the encoder floor.

12.2 Risk prediction overlap

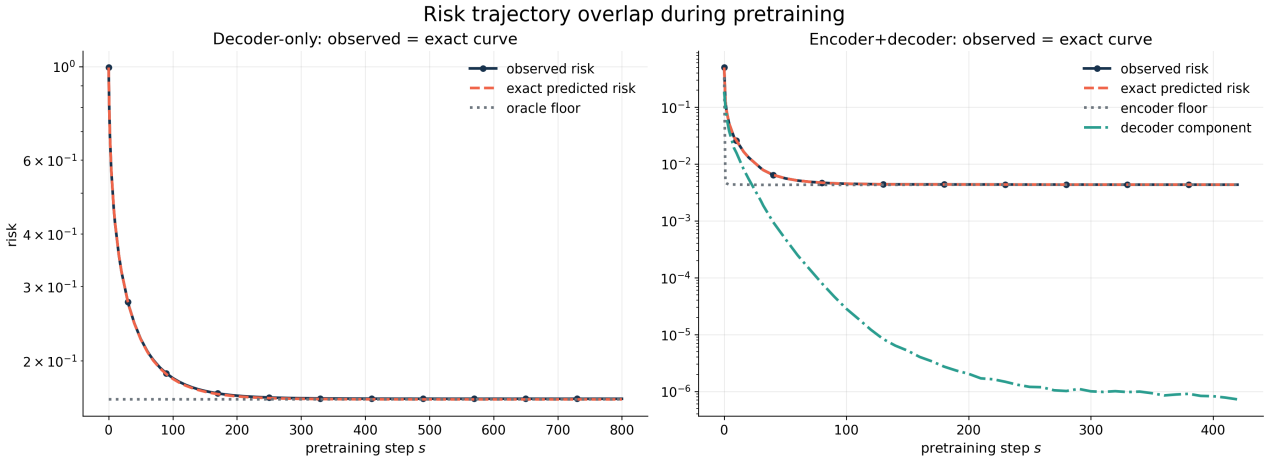


Figure 2: Observed risk and exact predicted risk over pretraining time.

The left panel overlays the observed decoder-only risk with the exact population curve induced by

$$P_s - P_\star = (P_0 - P_\star)(I - \eta C)^s.$$

The curves are visually indistinguishable after the first transient. The maximum relative discrepancy is $4.3852 \cdot 10^{-3}$, which is at the scale expected from finite evaluation sampling and numerical estimation of the moments. The horizontal dotted line is the oracle floor $\mathcal{R}(P_\star)$. The curve approaches that floor as the preconditioner trajectory converges.

The right panel is stricter: for the encoder+decoder run, the observed risk and the exact identity risk agree up to $2.7756 \cdot 10^{-17}$. This is not a fitted curve. It is the identity

$$z_L - z_s^\dagger = -(I - P_{z,s} A_s)^L z_s^\dagger$$

evaluated at every checkpoint. The green curve is the decoder component; it becomes much smaller than the gray encoder floor. Therefore the final error is not caused by an unsolved latent least-squares system, but by the finite encoder subspace.

12.3 Decoder-only detailed diagnostics

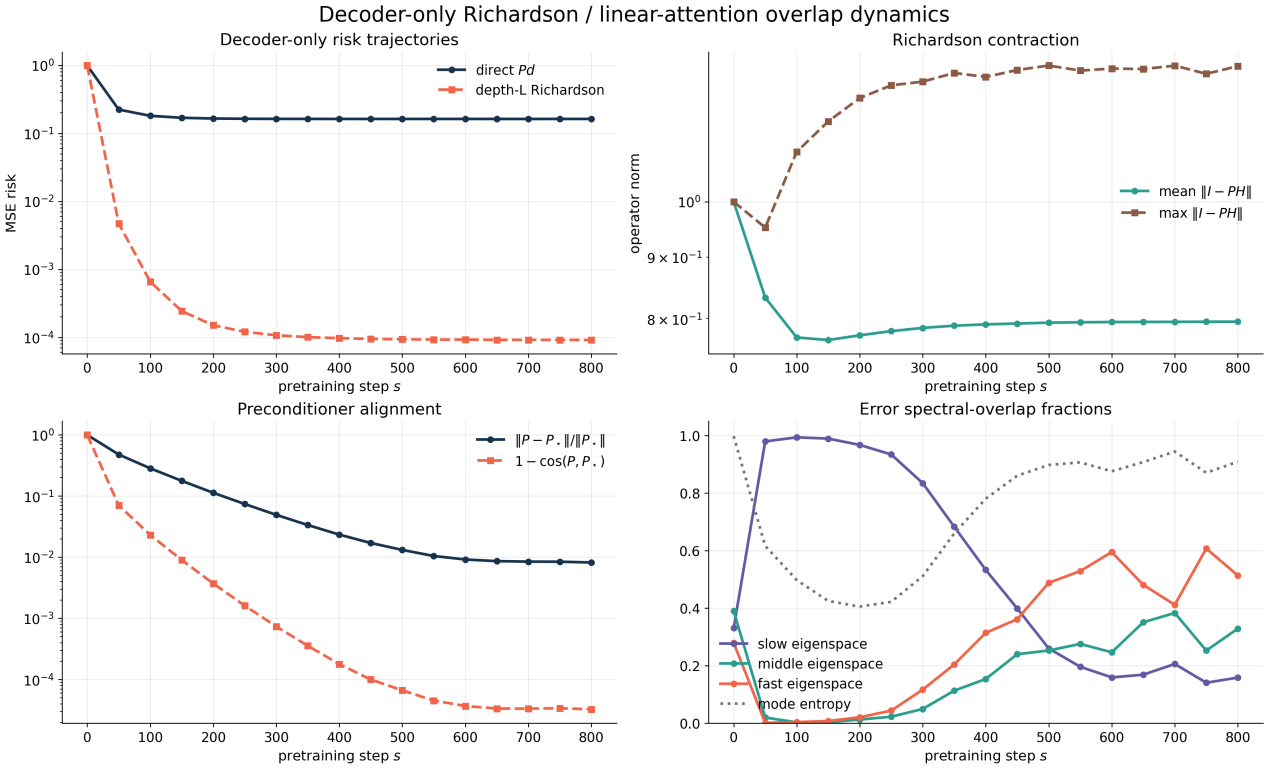


Figure 3: Decoder-only alignment, contraction, and spectral overlap.

The upper-left panel compares the direct estimator $P_s d$ to the depth-L Richardson estimator using the same learned P_s. The direct risk plateaus near $1.63 \cdot 10^{-1}$, while the iterated solver reaches $9.15 \cdot 10^{-5}$. Thus the learned attention matrix is best interpreted as a preconditioner for the iterative residual dynamics rather than only as a one-shot inverse map.

The upper-right panel shows contraction statistics of $I - P_s H$. The mean contraction improves below one. The maximum contraction can exceed one because it is measured over a finite task ensemble and in worst directions; the average risk can still decay strongly when the unstable directions have small mass under the task distribution. This is why the spectral-overlap panel matters.

The lower-left panel is the direct evidence that P_s moves toward $P_\star$. Both the relative error and the cosine defect decrease monotonically until saturation. This is the outer training dynamics of the attention preconditioner.

The lower-right panel decomposes the remaining preconditioner error across eigenspaces of $C = \mathbb{E}[dd^\top]$. The weights are

$$w_{s,i} = \frac{\lambda_i \|(P_s - P_\star) q_i\|^2}{\sum_j \lambda_j \|(P_s - P_\star) q_j\|^2}.$$

The changing distribution of these weights explains why a single scalar risk is not a closed state variable. The risk curve depends on where the error lies in the spectrum, not only on its total norm.

12.4 Encoder+decoder detailed diagnostics

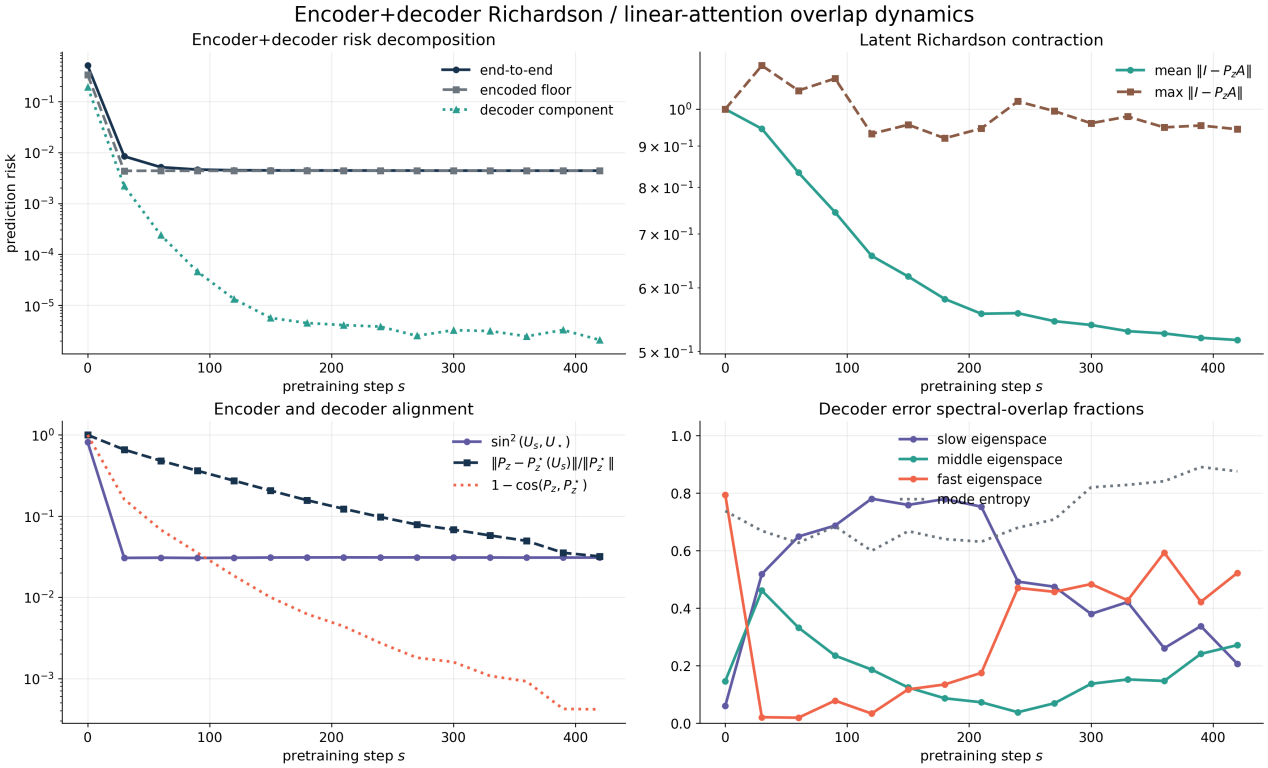


Figure 4: Joint encoder+decoder risk decomposition and preconditioner tracking.

The upper-left panel decomposes the prediction risk. The end-to-end risk and the encoded floor become almost identical; the decoder component is orders of magnitude smaller. This supports the mathematical decomposition

$$\mathcal{R}_{\text{e2e}} = \mathcal{R}_{\text{floor}} + \mathcal{R}_{\text{dec}} + \text{cross term}.$$

In this run the cross term is negligible at the plotted scale.

The upper-right panel tracks the latent contraction $\|I - P_{z,s}A_s\|$. The mean contraction decreases during pretraining, which means the decoder becomes a better latent preconditioner. The maximum contraction is less smooth, because A_s changes with the encoder and the finite evaluation set contains task-dependent hard directions.

The lower-left panel shows the key joint phenomenon. The encoder subspace error decreases quickly and then plateaus. Meanwhile the decoder continues to align with the current oracle $P_z^\star(U_s)$. Since U_s moves, the decoder target itself moves. The correct formula is therefore

$$P_{z,s+1} - P_z^\star(U_{s+1}) = (P_{z,s} - P_z^\star(U_s))(I - \eta C_s) + P_z^\star(U_s) - P_z^\star(U_{s+1}).$$

The lower-right panel again decomposes spectral-overlap fractions. The redistribution of error mass across slow, middle, and fast modes is the observable counterpart of the moving spectral basis. This is why the exact general theory is matrix-valued, while scalar formulas are exact only in commuting or fixed-basis settings.

12.5 Online decoder training

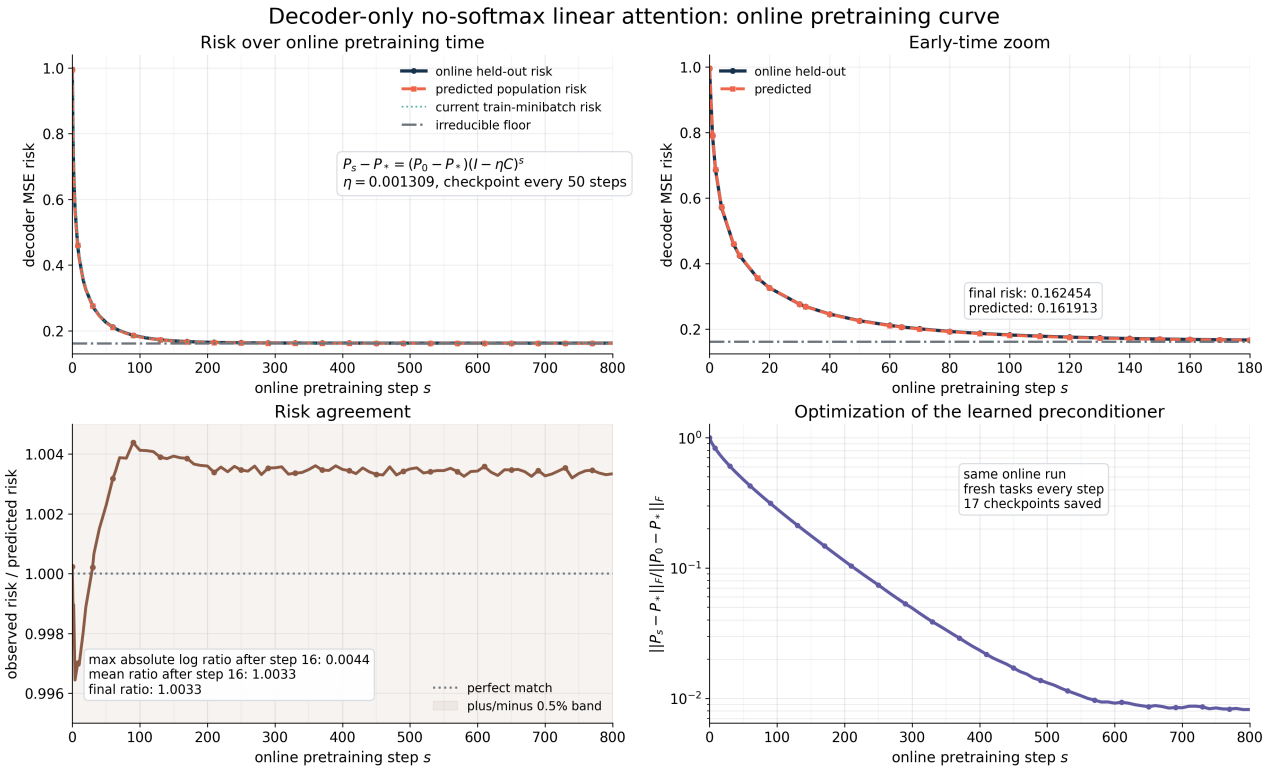


Figure 5: Online decoder-only training with predicted curve and checkpoints.

This plot is the direct online training run for the decoder. The learned linear-attention object is P_s. The curve predicted from the closed-form population dynamics tracks the measured evaluation risk across the run. The final observed risk is $1.6245 \cdot 10^{-1}$, while the predicted value is $1.6191 \cdot 10^{-1}$. The relative distance to the oracle preconditioner at the final checkpoint is $8.2055 \cdot 10^{-3}$.

The important point is that the plotted training curve is not the trajectory of unpreconditioned least squares. It is the trajectory of the learned preconditioner under gradient descent on the pretraining distribution. The least-squares/Richardson equation is used only to evaluate the quality of the current P_s.

12.6 Online encoder-only training

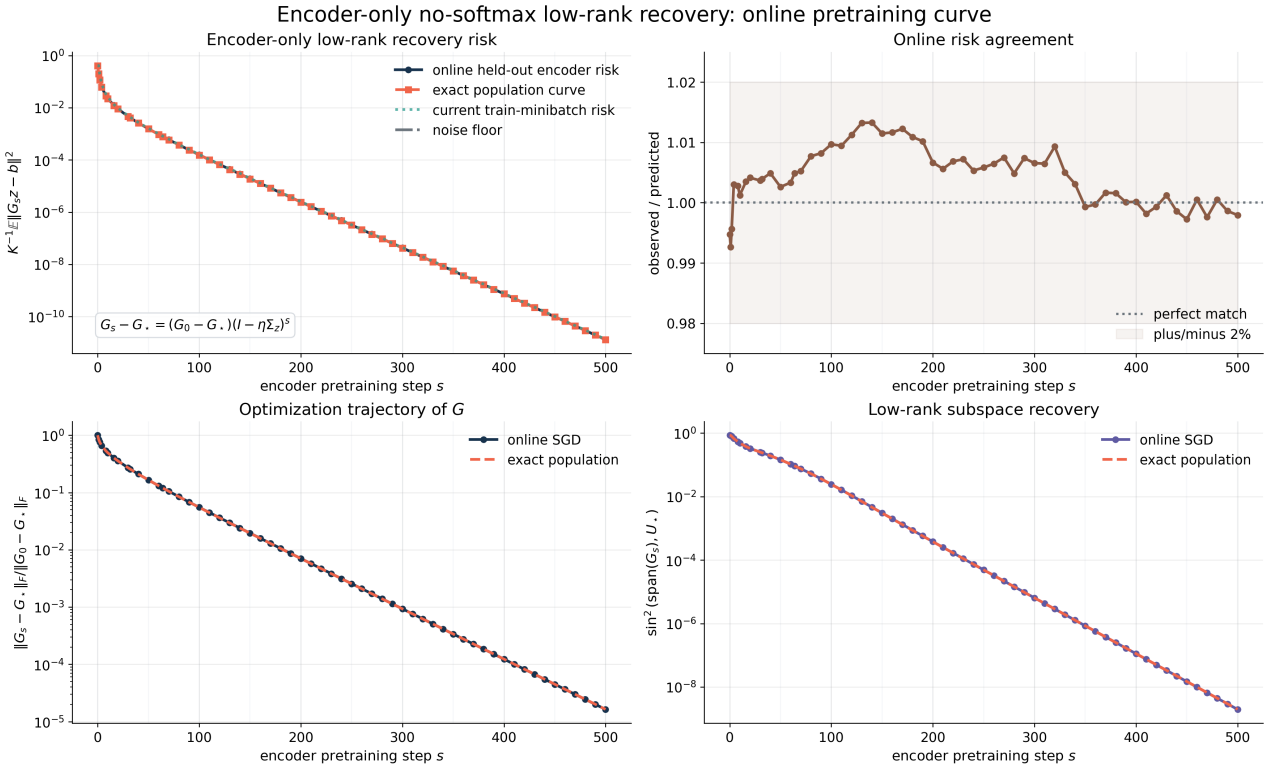


Figure 6: Online encoder-only low-rank recovery.

The encoder-only experiment isolates the recovery of the map G or, equivalently, the low-rank task subspace. The exact population dynamics are

$$G_s - G_\star = (G_0 - G_\star)(I - \eta\Sigma_z)^s.$$

The observed encoder risk follows this formula with mean ratio close to one. The final observed risk is $1.3095 \cdot 10^{-11}$, and the predicted risk is $1.3122 \cdot 10^{-11}$. The final relative G-error is $1.6295 \cdot 10^{-5}$, matching the predicted $1.6339 \cdot 10^{-5}$.

This experiment is the encoder analogue of the decoder preconditioner experiment. The decoder learns P_s; the encoder learns G_s or U_s. Both have exact closed population dynamics when the corresponding population covariance is fixed.

12.7 Online encoder+decoder training

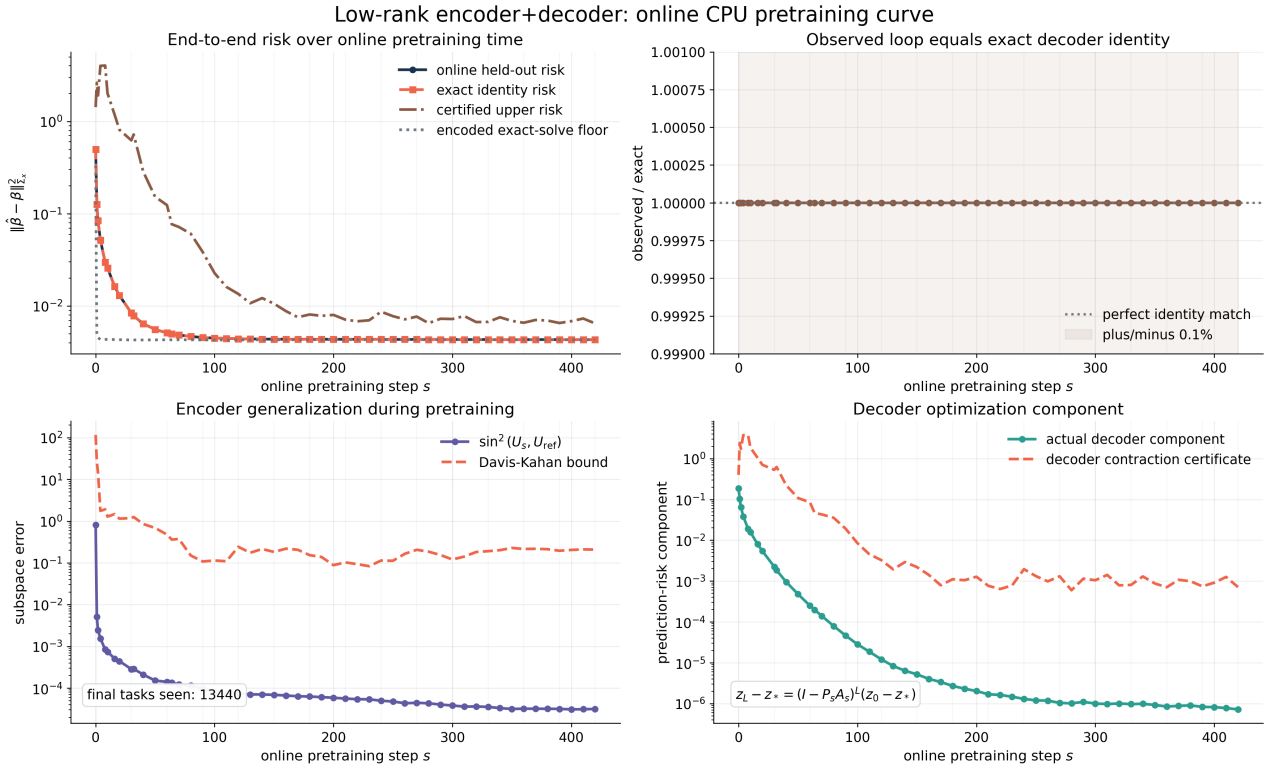
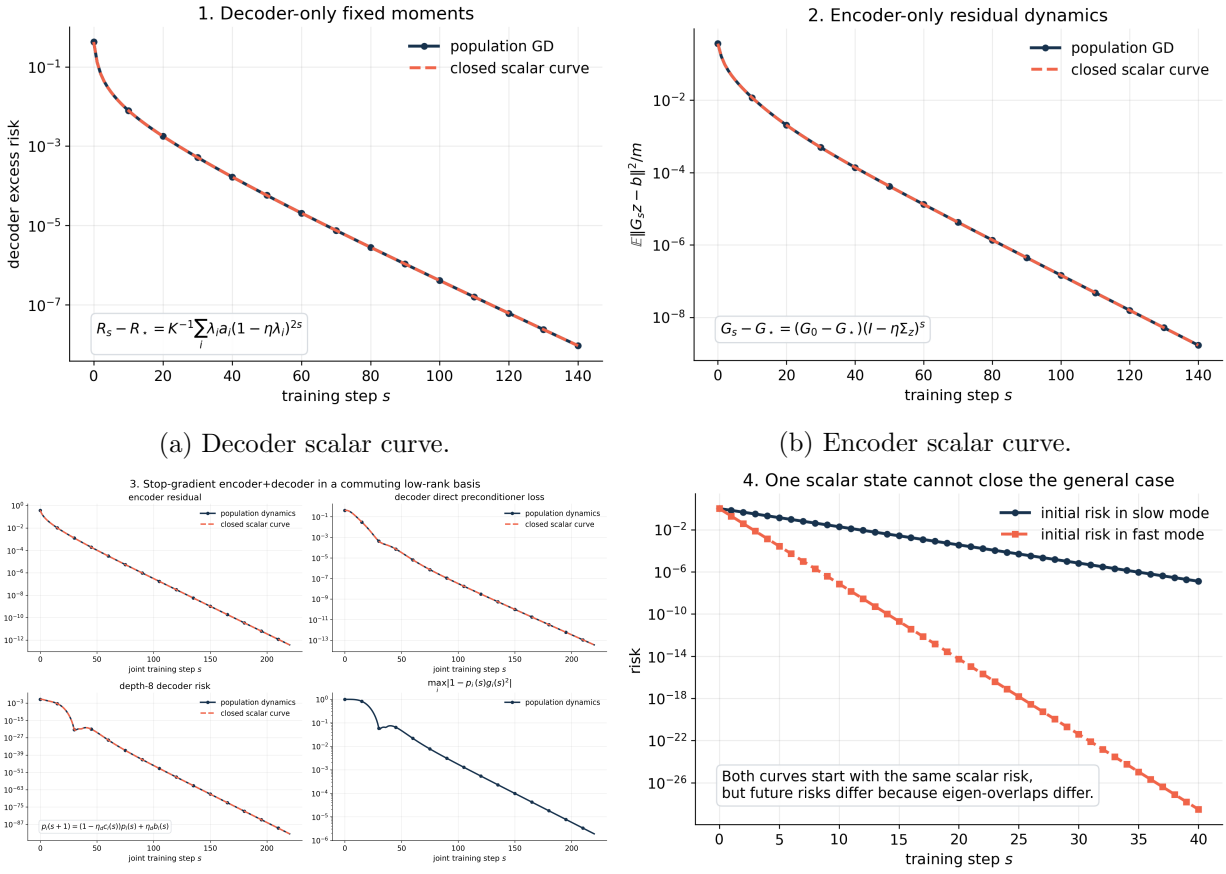


Figure 7: Online low-rank encoder+decoder training.

This run combines the encoder and the learned latent decoder. The exact identity gap is $2.7756 \cdot 10^{-17}$, so the plotted exact identity is a numerical equality, not a regression fit. The final held-out prediction risk is $4.3266 \cdot 10^{-3}$, equal to the exact identity risk at the displayed precision.

The final true-subspace error is about $3.0853 \cdot 10^{-2}$. The final decoder actual risk is $7.2394 \cdot 10^{-7}$, while the final certified bound is $7.1221 \cdot 10^{-4}$. The certificate is conservative, but the identity itself is exact. The qualitative conclusion is that the decoder has learned a strong preconditioner, and the remaining prediction error is primarily the finite encoder floor.

12.8 Exact scalar proof ladder



(a) Decoder scalar curve.

(b) Encoder scalar curve.

(c) Commuting coupled scalar curve.

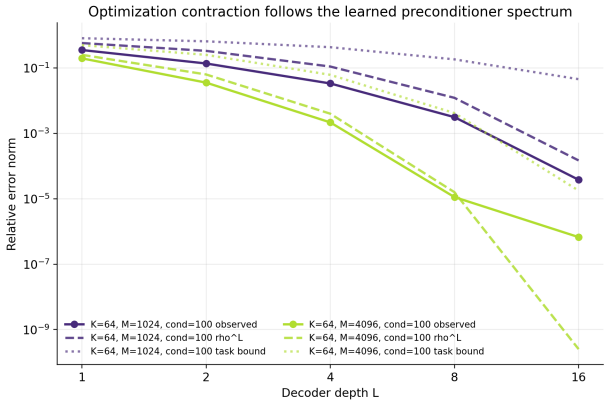
(d) No one-scalar closure.

Figure 8: Exact scalar dynamics where valid, and counterexample where invalid.

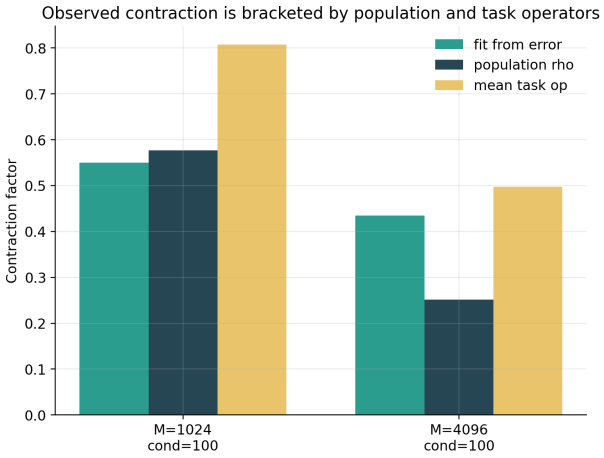
These four panels separate what can be reduced to a scalar curve from what cannot. Decoder-only and encoder-only dynamics have exact scalar formulas after diagonalizing the fixed covariance. The commuting coupled system also has exact mode-wise scalar dynamics. The maximum numerical errors are $4.337 \cdot 10^{-18}$, $6.939 \cdot 10^{-18}$, and $5.551 \cdot 10^{-17}$.

The counterexample is the important limitation. Two states can have the same initial scalar risk $R_0 = 1$, but different spectral placement of the error. After one step they produce $R_1 = 0.6724$ and $R_1 = 0.1936$. Therefore "the risk" alone is not a complete state variable. Exact closure requires spectral overlaps, or the full matrix trajectory when the spectral basis moves.

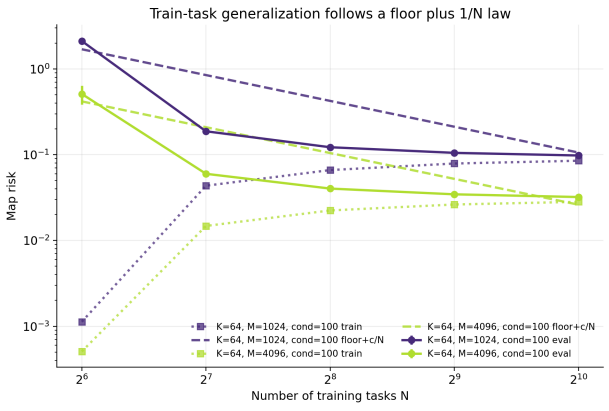
12.9 Optimization and generalization bound diagnostics



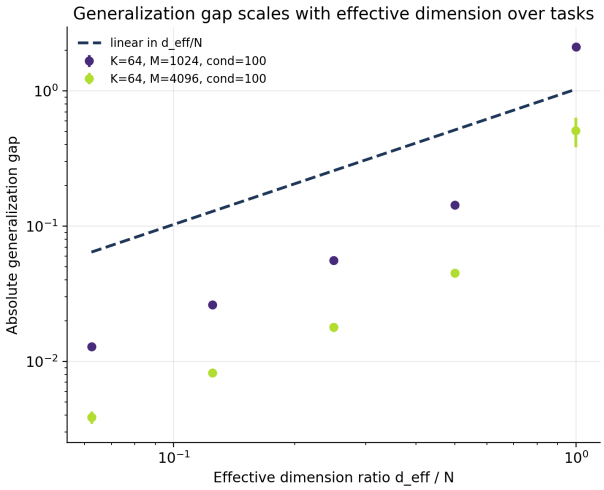
(a) Optimization contraction fit.



(b) Contraction factors.



(c) Train/eval generalization.



(d) Gap versus d_{eff}/N.

Figure 9: Optimization and task-generalization diagnostics.

The optimization plots check deterministic certificates for the learned decoder iteration. For each frozen learned preconditioner and task Hessian, the exact error recursion is

$$e_{\ell+1} = (I - PH)e_\ell.$$

Therefore the bound

$$\|e_\ell\| \le \|I - PH\|^\ell \|e_0\|$$

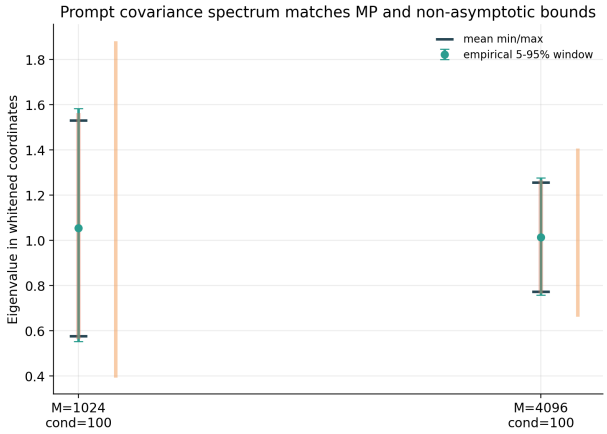
is a norm identity/certificate, not a statistical fit. The observed violations are zero in the checked runs. Increasing prompt size improves the population contraction, for example from about 0.576 at $M = 1024$ to about 0.251 at $M = 4096$ in the checked $K = 64$, condition 100 setting.

The generalization plots address the number N of training tasks/prompts used to learn the map. The exact excess-risk identity is

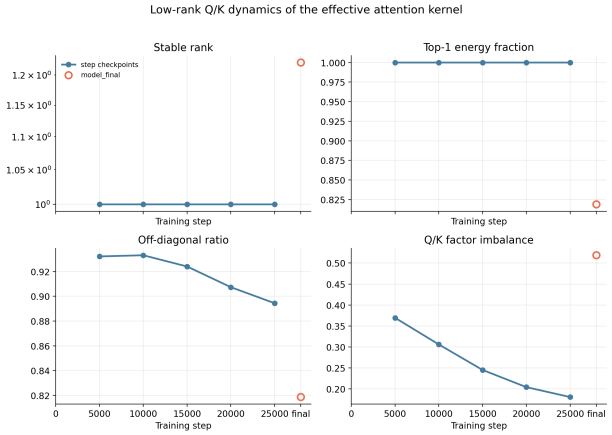
$$\mathcal{R}(\widehat{P}_N) - \mathcal{R}(P_\star) = \frac{1}{K}\,\mathrm{Tr}\{(\widehat{P}_N - P_\star)C(\widehat{P}_N - P_\star)^\top\}.$$

The empirical scaling is then compared to d_{eff}/N. The plotted curves show the expected decay of the train/eval gap as more distinct tasks are used to identify the spectral preconditioner.

12.10 Prompt MP spectrum and Q/K dynamics



(a) Prompt covariance and MP window.



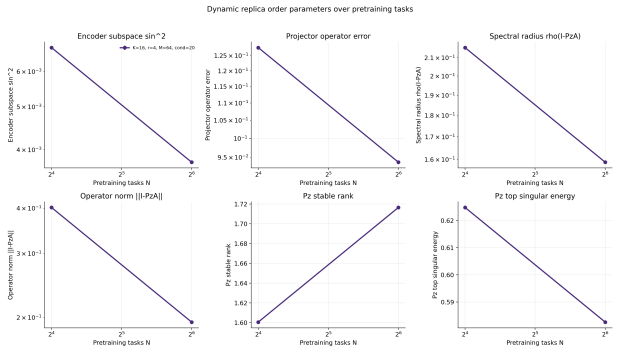
(b) Effective Q/K low-rank dynamics.

Figure 10: Prompt concentration and attention kernel diagnostics.

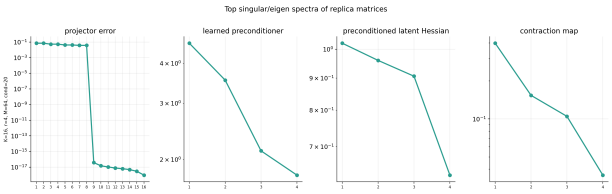
The MP-spectrum plot checks prompt-size effects. In whitened coordinates, the eigenvalues of the empirical covariance should concentrate in $[(1-\sqrt{K/M})^2, (1+\sqrt{K/M})^2]$. The checked windows match this prediction and the non-asymptotic Gaussian concentration checks have zero violations. This validates the prompt side of the scaling law: larger M narrows the spectral window and improves conditioning.

The Q/K plot analyzes the gauge-invariant object $W_Q^\top W_K$, averaged over heads, not raw W_Q or W_K. Raw Q and K are not identifiable because $W_Q \mapsto AW_Q$, $W_K \mapsto A^{-\top}W_K$ can leave the kernel unchanged. The effective kernel shows low-rank structure: stable rank near one, high top-energy concentration, decreasing skew/symmetry ratio, and factor balancing over training. This supports the interpretation that the attention mechanism learns a low-rank preconditioning geometry.

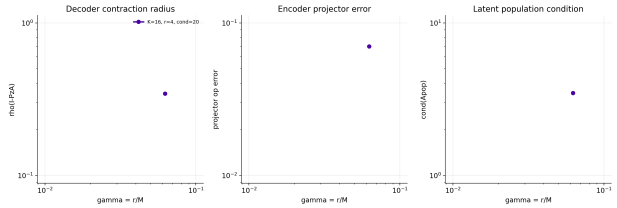
12.11 Replica spectral diagnostics



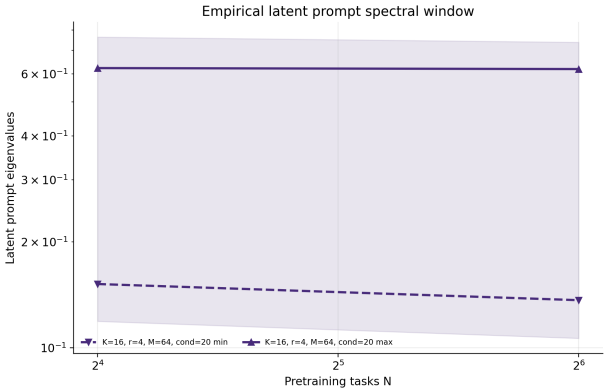
(a) Dynamic order parameters.



(b) Top spectra.



(c) Prompt aspect scaling.



(d) Latent prompt spectral window.

Figure 11: Replica/spectral order-parameter diagnostics.

These plots collect the macroscopic quantities that are stable under changes of coordinates: spectra, overlaps, effective ranks, prompt aspect ratios, and contraction radii. This is the correct level for replica-style comparison. Individual parameter matrices can have gauge freedoms, but their spectra, induced contraction factors, effective ranks, and overlap weights are invariant diagnostics of the trained system.

The dynamic order parameters track the encoder subspace error, latent contraction, and preconditioner rank. The top-spectra grid shows whether the learned latent systems occupy the expected spectral window. The prompt aspect and latent window plots connect empirical prompt size M, rank r, and Marchenko–Pastur scaling. Together they provide the bridge between the exact finite-dimensional identities above and the large-system spectral/replica interpretation.

13 Conclusion

Linear attention: exact global preconditioner dynamics $s \mapsto P_s$.

Decoder evaluation: exact Richardson identity at frozen P_s.

Encoder: exact covariance-driven recovery in fixed population model.

Encoder+decoder: exact triangular dynamics under stop-gradient.

General non-commuting joint system: exact matrix dynamics, no one-scalar closure.

Softmax attention: exact nonlinear update map/Jacobian, not one global P_s.