Optimization and Generalization for Encoder, Decoder, and Encoder–Decoder

MMNN experiments

July 2026

1 Common finite-dimensional problem

Each task is represented by a latent vector $z \in \mathbb{R}^K$. A prompt gives linear weak equations

$$G_m z \simeq b_m, \qquad H_m z_\star = c_m, \qquad H_m = G_m^\top G_m + \lambda I, \qquad c_m = G_m^\top b_m.$$

The encoder estimates the task representation. The decoder solves the induced linear system. The encoder–decoder model composes both:

$$\widehat{z} = \mathrm{Enc}_\Theta(\mathrm{prompt}), \qquad x_{\ell+1} = x_\ell + P_\Theta(c(\widehat{z}) - H(\widehat{z})x_\ell).$$

2 Generalization certificate

For any held-out clipped loss

$$\ell_c = \min(\ell, c), \qquad 0 \le \ell_c \le c,$$

the empirical Bernstein certificate used in the experiments is

$$\boxed{R_c \le \widehat{R}_c + \sqrt{\frac{2\widehat{V}_c \log(2/\delta)}{n}} + \frac{7c\log(2/\delta)}{3(n-1)}.}$$

This is distribution-free after clipping. Raw Gaussian losses are reported separately. For the encoder-only Gaussian model, we also have an exact population risk curve, so the held-out risk can be compared directly to the population prediction.

3 Encoder only

3.1 Model

The encoder-only low-rank recovery experiment uses

$$z \sim \mathcal{N}(0, \Sigma_z), \qquad b = G_\star z + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I).$$

The learned encoder dictionary G_s is trained by

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

3.2 Exact optimization

Population gradient descent gives the exact trajectory

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

Therefore

$$R_{\mathrm{enc}}(s) = \frac{1}{K}\mathbb{E}\|G_s z - b\|_2^2 = \sigma^2 + \frac{1}{K}\operatorname{Tr}\Big[(G_s - G_\star)\Sigma_z(G_s - G_\star)^\top\Big],$$

and, if $0 < \eta < 2/\lambda_{\max}(\Sigma_z)$,

$$R_{\mathrm{enc}}(s) - \sigma^2 \le \|I - \eta\Sigma_z\|_{\mathrm{op}}^{2s}\frac{1}{K}\operatorname{Tr}\Big[(G_0 - G_\star)\Sigma_z(G_0 - G_\star)^\top\Big].$$

3.3 Encoder generalization

The population target is the exact $R_{\mathrm{enc}}(s)$. The independent held-out estimate $\widehat{R}_{\mathrm{enc}}(s)$ matched the exact curve:

$$\widehat{R}_{\mathrm{enc}}(800) = 4.0020466789 \times 10^{-4}, \qquad R_{\mathrm{enc}}(800) = 4.0000000000 \times 10^{-4},$$

$$\frac{\widehat{R}_{\mathrm{enc}}(800)}{R_{\mathrm{enc}}(800)} = 1.0005116697.$$

Table 1: Encoder-only low-rank recovery, final checkpoint.

quantity	value
held-out encoder risk	4.0020×10^{-4}
exact population risk	4.0000×10^{-4}
observed / predicted	1.0005
relative G error	4.5093×10^{-4}
subspace $\sin^2\Theta$	1.7845×10^{-6}
irreducible noise floor	4.0000×10^{-4}

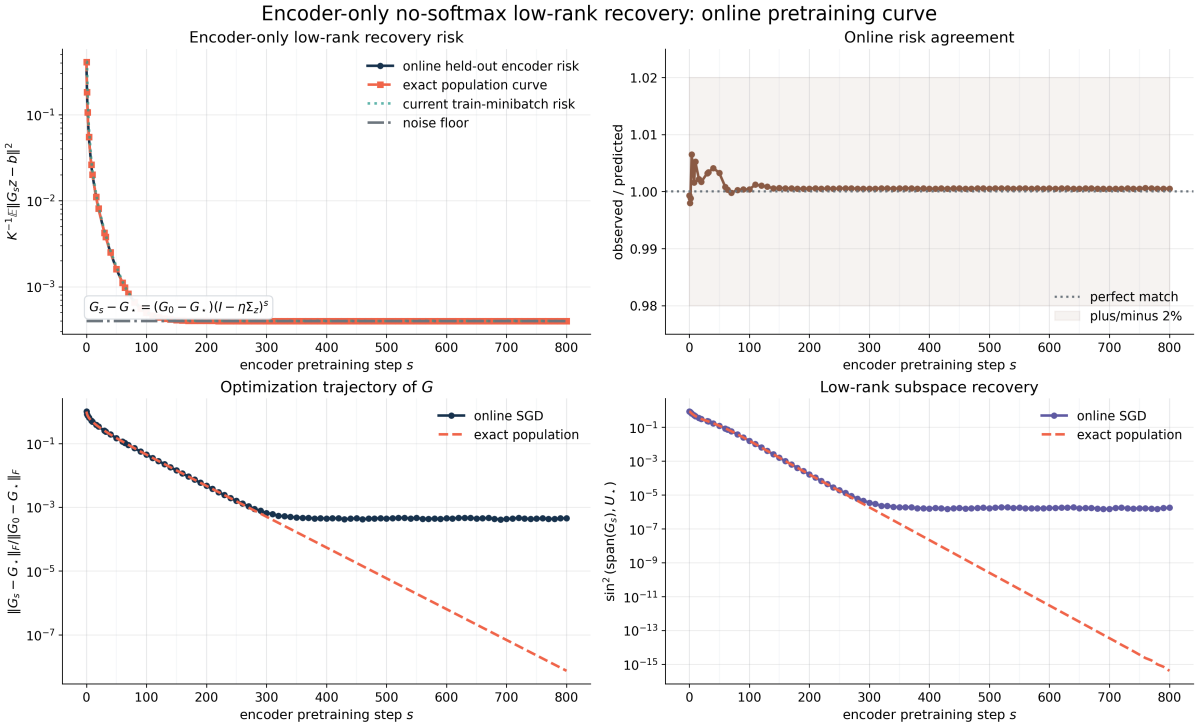

Figure 1: Encoder-only optimization and population-risk curve.

4 Decoder only

4.1 Model

For a fixed task system $Hx_\star = d$, the decoder is a depth-L preconditioned Richardson iteration

$$x_{\ell+1} = x_\ell + P_\Theta(d - Hx_\ell).$$

4.2 Exact optimization

With $e_\ell = x_\ell - x_\star$,

$$\boxed{e_L = (I - P_\Theta H)^L e_0.}$$

For each task,

$$\|e_L\|_2 \le \rho(H, P_\Theta)^L \|e_0\|_2, \qquad \rho(H, P_\Theta) = \|I - P_\Theta H\|_2.$$

Thus the population decoder risk satisfies

$$R_{\mathrm{dec}}(L) = \frac{1}{K}\mathbb{E}\|(I - P_\Theta H)^L x_\star\|_2^2,$$

and the certified upper risk is

$$R_{\mathrm{dec}}(L) \le \frac{1}{K}\mathbb{E}\big[\rho(H, P_\Theta)^{2L}\|x_\star\|_2^2\big].$$

4.3 Decoder generalization

The held-out decoder loss is certified by the clipped empirical-Bernstein bound. At the final decoder-only Flexformer checkpoint,

$$\widehat{R}_{\mathrm{dec}} = 7.674719 \times 10^{-3}, \qquad U_{\mathrm{EB}} = 1.771005 \times 10^{-1},$$

and the pointwise Richardson certificate had no violations:

$$\Pr\left(\|e_L\|_2 > \rho(H, P_\Theta)^L \|e_0\|_2\right) = 0.$$

Table 2: Decoder-only learned-preconditioner checkpoint.

quantity	value
held-out depth risk	7.6747×10^{-3}
clipped EB upper	1.7710×10^{-1}
optimization violation rate	0
error / bound ratio	2.2833×10^{-2}
mean contraction $\mathbb{E}\rho$	1.1031
relative distance to oracle $P_\star$	2.6788×10^{-1}
stable rank of P_Θ	2.9198

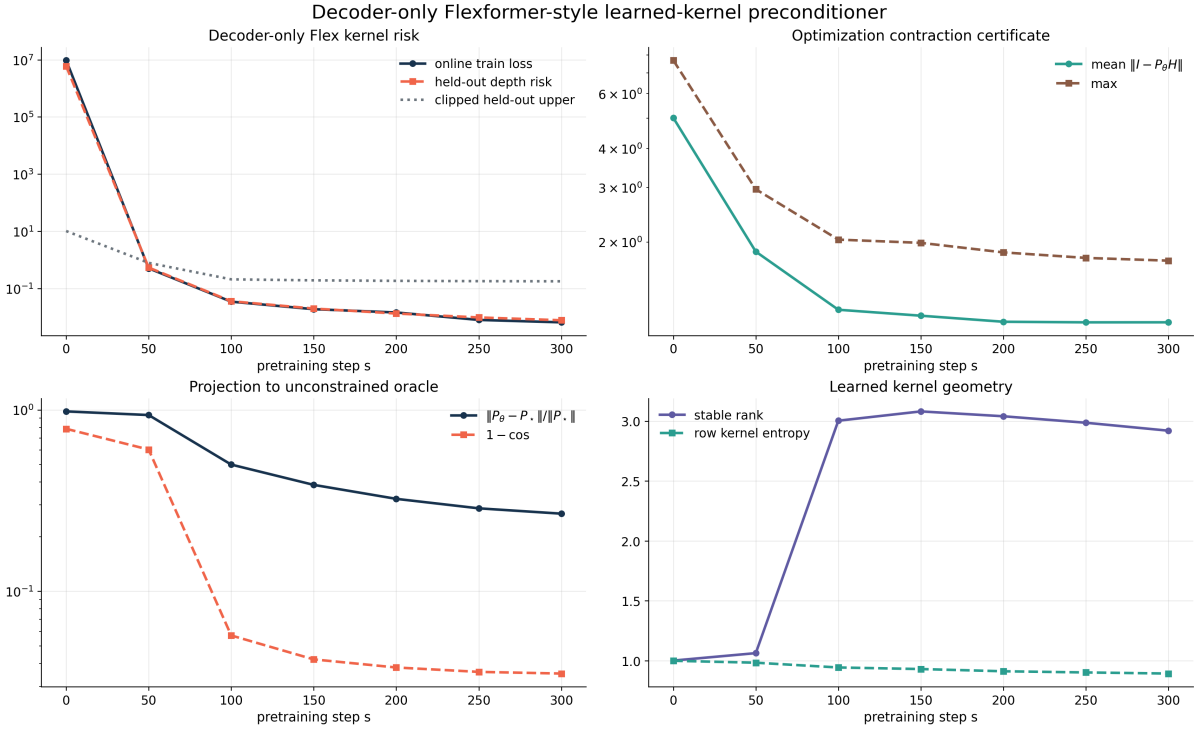

Figure 2: Decoder-only generalization and Richardson optimization certificates.

5 Encoder–decoder

5.1 Abstract decomposition

Let $x^\dagger(\widehat{z})$ be the exact solution of the system induced by the encoded task $\widehat{z}$. Then

$$\widehat{x}_L - x_\star = \underbrace{\widehat{x}_L - x^\dagger(\widehat{z})}_{\text{decoder}} + \underbrace{x^\dagger(\widehat{z}) - x_\star}_{\text{encoder}}.$$

Consequently,

$$\|\widehat{x}_L - x_\star\|_2^2 \le 2\|\widehat{x}_L - x^\dagger(\widehat{z})\|_2^2 + 2\|x^\dagger(\widehat{z}) - x_\star\|_2^2.$$

The decoder part is again controlled exactly by Richardson:

$$\widehat{x}_L - x^\dagger(\widehat{z}) = (I - P_\Theta H(\widehat{z}))^L(\widehat{x}_0 - x^\dagger(\widehat{z})).$$

5.2 Low-rank encoder–decoder

For low-rank linear-regression tasks,

$$\beta = U_\star z, \qquad A_s = U_s^\top H U_s, \qquad d_s = U_s^\top c,$$

$$z_{\ell+1} = z_\ell + P_s(d_s - A_s z_\ell).$$

The exact identity checked in the online run is

$$\boxed{z_L - z_\star = (I - P_s A_s)^L (z_0 - z_\star).}$$

The encoder subspace certificate is Davis–Kahan:

$$\sin^2 \Theta(U_s, U_{\mathrm{ref}}) \le \left(\frac{\|S_s - \Sigma_\beta\|_{\mathrm{op}}}{\Delta} \right)^2.$$

Table 3: Online low-rank encoder–decoder final checkpoint.

quantity	value
held-out prediction risk	4.3266×10^{-3}
exact identity risk	4.3266×10^{-3}
max identity absolute gap	2.7756×10^{-17}
decoder actual risk	7.2394×10^{-7}
decoder certified risk	7.1221×10^{-4}
encoder true subspace $\sin^2 \Theta$	3.0853×10^{-2}
Davis–Kahan $\sin^2 \Theta$ bound	2.0888×10^{-1}

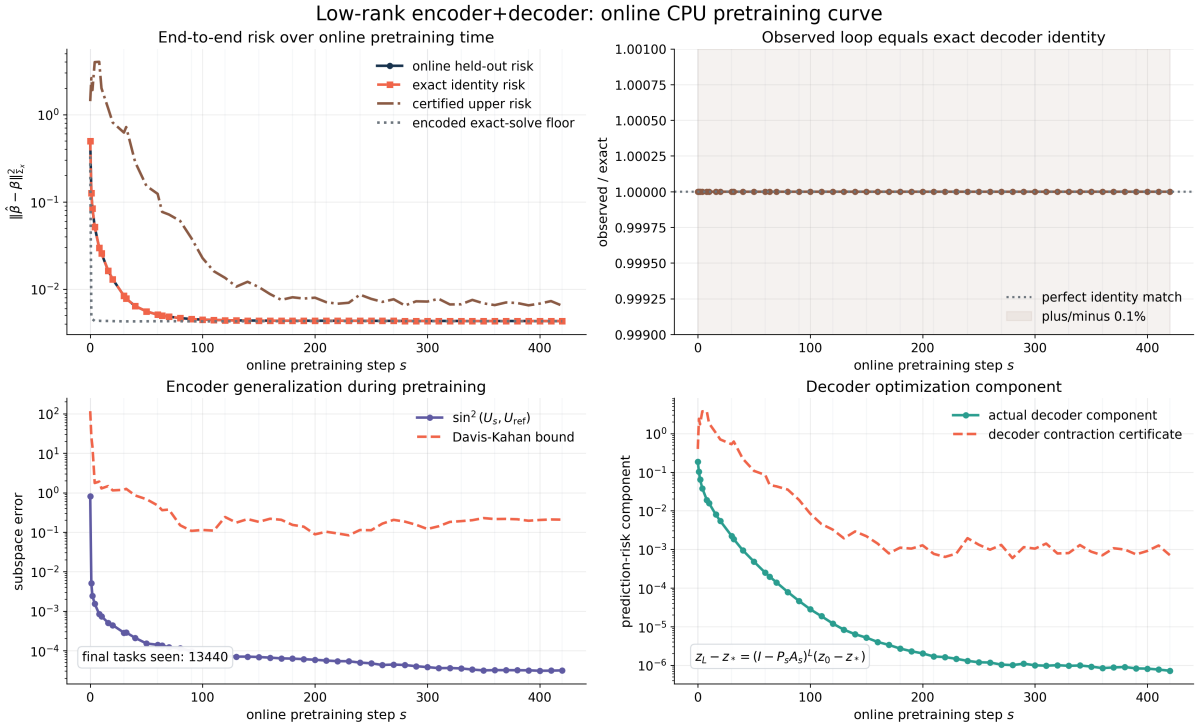

Figure 3: Online low-rank encoder–decoder optimization identity and subspace certificate.

5.3 Flexformer encoder–decoder

The Flexformer encoder–decoder run uses a learned kernel preconditioner in the latent decoder. The same exact decoder identity holds after freezing the checkpoint. The measured end-to-end risk decomposes into an encoder floor and a decoder component:

$$R_{\mathrm{e2e}} \approx R_{\mathrm{floor}}(\mathrm{Enc}) + R_{\mathrm{dec}}.$$

Table 4: Flexformer encoder–decoder final checkpoint.

quantity	value
held-out prediction risk	5.8931×10^{-3}
clipped EB upper	1.7511×10^{-1}
encoded floor risk	2.4169×10^{-3}
decoder component risk	1.7448×10^{-3}
optimization violation rate	0
error / bound ratio	4.5485×10^{-1}
mean latent contraction $\mathbb{E}\rho$	8.2571×10^{-1}
encoder $\sin^2\Theta$ to true subspace	3.6699×10^{-2}

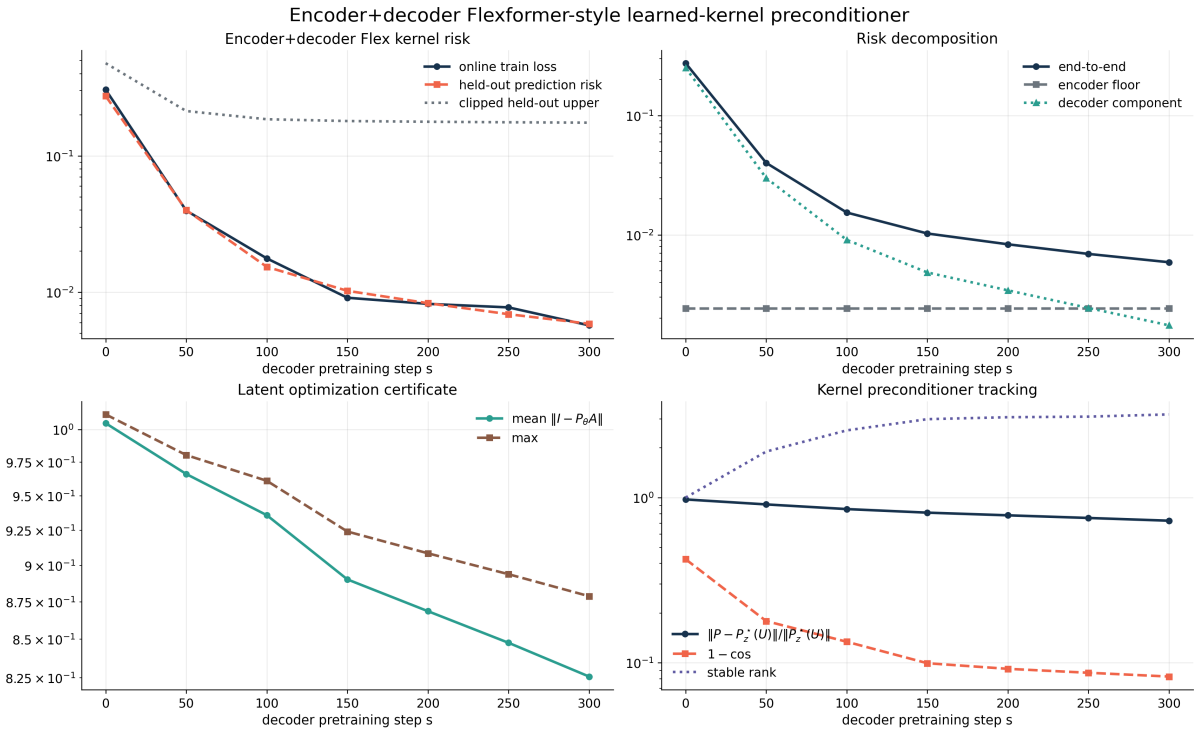

Figure 4: Flexformer encoder–decoder generalization and optimization certificates.

5.4 PDE encoder–decoder recast

For the Galerkin PDE recast,

$$A(z) = A_0 + \sum_{k=1}^{K} z_k A_k, \qquad A(z)u_i = f_i.$$

The prompt encoder is the exact ridge estimator

$$G_{i,r,k} = \langle A_k u_i, v_r \rangle, \qquad b_{i,r} = \langle f_i - A_0 u_i, v_r \rangle,$$

$$\boxed{z_{\mathrm{enc}} = (G^\top G + \lambda_z I)^{-1} G^\top b.}$$

The query decoder is

$$u_{\ell+1} = u_\ell + P_\Theta(f_\star - A(z_{\mathrm{enc}})u_\ell),$$

with exact frozen error identity

$$e_L = (I - P_\Theta A(z_{\mathrm{enc}}))^L e_0.$$

Table 5: PDE encoder–decoder K=4 recast.

quantity	value
encoder z MSE	6.1356×10^{-8}
encoder A relative error	6.0764×10^{-5}
exact encoded-solve floor query MSE	1.7456×10^{-9}
best end-to-end method	free linear
best end-to-end MSE	5.9760×10^{-6}
max optimization violation rate	0
best clipped EB upper	8.4150×10^{-2}

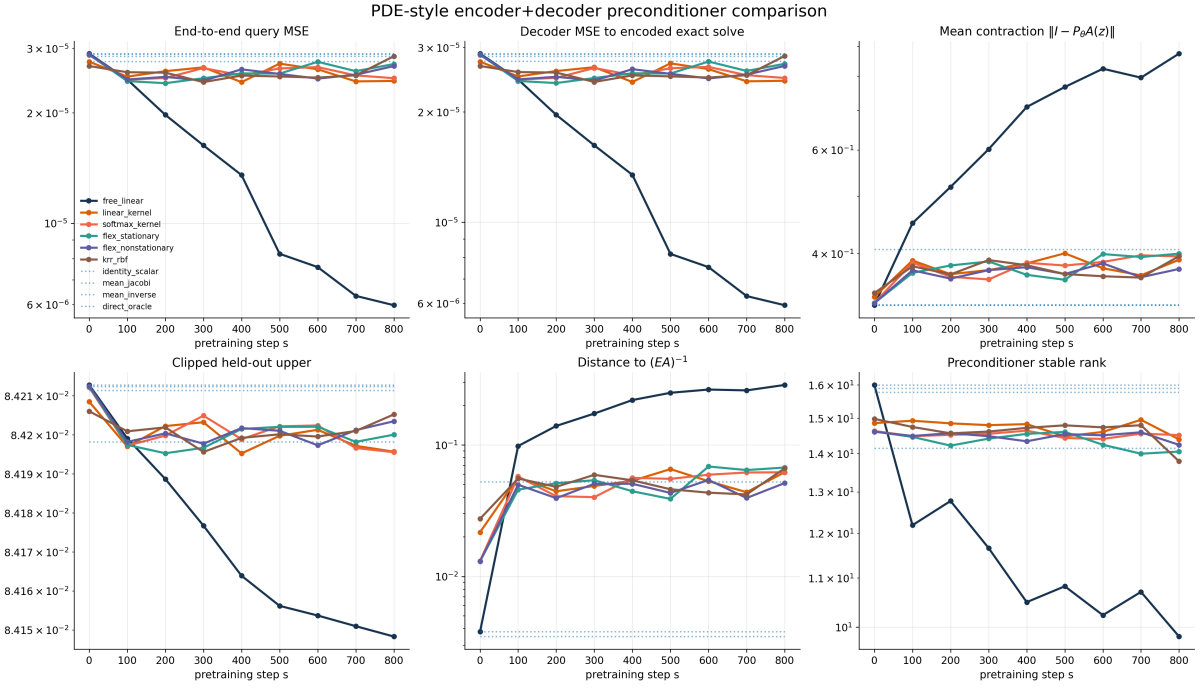

Figure 5: PDE encoder–decoder K=4 generalization and optimization comparison.

6 Summary

The confirmed optimization statements are exact identities:

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

$$e_L = (I - P_\Theta H)^L e_0,$$
$$\widehat{x}_L - x^\dagger(\widehat{z}) = (I - P_\Theta H(\widehat{z}))^L(\widehat{x}_0 - x^\dagger(\widehat{z})).$$

The confirmed generalization statements are:

$$\widehat{R}_{\mathrm{enc}}(800)/R_{\mathrm{enc}}(800) = 1.0005116697,$$

$$U_{\mathrm{EB}}^{\mathrm{dec}} = 1.7710 \times 10^{-1}, \qquad U_{\mathrm{EB}}^{\mathrm{enc+dec}} = 1.7511 \times 10^{-1},$$

$$\max\{\text{Richardson bound violation rates in reported runs}\} = 0.$$