Six Requirements for Inference-Time Self-Iterating Models: Representation, Uncertainty, and Local Credit

Anonymous authors
Paper under double-blind review

Abstract

Inference-time self-iteration lets a model run an iterative learner on its context before it answers. For multi-branch in-context relations—how different two items are along an attribute the model was never trained on, or compressive phase retrieval—a lifted inner loop (per-example global inversion, multipliers, local least squares) provably reaches what test-time training's gradient update, fixed-feature closed-form heads and additional layers cannot. Whether a pretrained model can host such a loop depends on its representation. We prove six requirements. Two concern the representation and hold for every learner: the task must be decodable from the base, which sets an explicit error floor floor(ν) in terms of the base's undecodable fraction ν (R1), and it must be identifiable from the context, which ties the task dimension to the context size and calls for posterior mixtures near the threshold (R2). Two concern uncertainty: inversion and readout must be consistent with the base's residual noise (R3), and in-context noise estimates must be corrected for degrees of freedom (R4). Two concern local credit: the multiplier leak must vanish on exactly decodable bases and grow with the noise (R5), and the knobs that control branch changes must be selected, not trained by backpropagation (R6). R1 and R2 give measurable diagnostics of a base. On word vectors (GloVe), a molecular graph network (GINE) and an image code (PCA), which satisfy them, the matched self-iterating learner beats gradient test-time training, quadratic closed-form heads and Transformers, with a second-seed replication and paired tests; on a base that violates decodability (MiniLM word vectors) every learner, including an oracle probe, stays above the proven floor.

1 Introduction

Test-time training (TTT) turns a layer's hidden state into a learner that is updated on the current context (Sun et al., 2023; 2024); fast-weight programmers and test-time regression place many sequence models in the same design space, with inner updates that are single gradient steps or closed-form regressions (Schlag et al., 2021; Wang et al., 2026). The natural next step is self-iteration: let the inner learner run an iterative solver on the context before answering. Self-iteration is only as good as the representation it iterates on. This paper asks the question that decides whether it can be used: what must the pretrained base satisfy, and what must the self-iterating learner satisfy on that base?

We answer for multi-branch in-context relations. When a context shows pairs of items with how different they are along an unseen attribute, the label is $y = |u^\top(k_a - k_b) + \varepsilon|$, a sign-lost function of a task-specific direction plus the part of the attribute the base does not encode; phase retrieval has the same form.

Why self-iteration. On such relations a lifted inner loop—each context example gets an activity that is inverted globally in one dimension, multipliers coordinate the activities with a shared memory, and the memory is refit by local least squares, a predictive-coding

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.

augmented Lagrangian (Seely & Gould, 2026)—solves tasks that are out of reach for fixed-feature closed-form heads (which have a sample-independent floor, Theorem 1), for correlational gradient updates, and for added layers (Section 3).

Six requirements. Representation (Section 4), provable for every learner: (R1) decodability—if a fraction ν of the attribute is not decodable from the base, every learner has normalized error at least floor(ν), with floor(0.5) = 0.77, and deployment directions outside the keys' span add to ν; (R2) identifiability—fewer context examples than task dimensions leave the task unidentified for every learner, and near the threshold the Bayes predictor is a posterior mixture. Uncertainty and local credit (Section 5), for the self-iterating learner on a base with $\nu > 0$: (R3) inversion and readout consistent with residual noise inside the nonlinearity; (R4) degrees-of-freedom-corrected in-context noise estimates; (R5) a multiplier leak that vanishes when $\nu = 0$ and grows with ν; (R6) derivative-free selection of branch-controlling solver knobs, because backpropagation through the solver omits exactly the branch-change term.

Evidence. Each theorem is confirmed quantitatively on controlled bases (Section 7). The base diagnostics then predict the outcome on real bases (Section 8): GloVe word vectors, a GINE molecular encoder and a PCA image code satisfy R1–R2, and the self-iterating learner that satisfies R3–R6 is the best inner loop on all of them; MiniLM word vectors violate R1, and every learner, including an oracle probe trained on all labels, stays above the proven floor.

2 Setting

Base, TTT layer, inner loop. A pretrained base maps an input x to an embedding; a TTT layer projects it to keys $k = \theta(x) \in \mathbb{R}^d$. Given a context $D = \{(k_i, y_i)\}_{i=1}^n$ ($K \in \mathbb{R}^{n\times d}$), an inner loop computes a memory $w = \mathcal{A}(K, y)$ and answers a query by $\hat{y}_q = R(k_q^\top w)$; θ, solver scalars and readout are trained on training tasks. We compare inner loops with identical keys, data and outer training: TTT gradient updates (official single pass with one step per mini-batch; full-batch descent with a learned step) (Sun et al., 2024); closed-form heads (ridge on keys; ridge on quadratic features vec($kk^\top$) predicting y^2, which contains the exact class $|k^\top w|^2$) (Bertinetto et al., 2019; Wang et al., 2026); a Transformer trained in context (additional layers) (Garg et al., 2022); and the lifted self-iterating loop.

Multi-branch relations on a base. An attribute of an item decomposes as $A(x) = u^\top k(x) + \epsilon(x)$, where ϵ is the part not linearly decodable from the base. With standardized attributes, the base noise ratio of the task is $\nu = \mathrm{Var}(\epsilon)/\mathrm{Var}(A) = 1 - R^2$. The pair relation is

$$y = |m + e|, \qquad m = u^\top(k_a - k_b) \sim \mathcal{N}(0, 2(1-\nu)), \quad e = \epsilon_a - \epsilon_b \sim \mathcal{N}(0, 2\nu), \tag{1}$$

with the residual inside the nonlinearity; compressive phase retrieval $y = |a^\top x|$ on an image code has the same form with $e = a^\top x_\perp$ the energy outside the code. We write $F_\sigma(m) = \mathbb{E}|m + e| = \sigma\sqrt{2/\pi}e^{-m^2/2\sigma^2} + m\,\mathrm{erf}(m/\sqrt{2}\sigma)$, $\sigma^2 = \mathrm{Var}(e)$.

The lifted inner loop. With an activity s_i per example, local loss $\ell_y(s) = (|s| - y)^2$, coupling $s = Kw$, multipliers λ, penalty ρ and leak $v \geq 0$:

$$s_i \leftarrow \arg\min_s \ \ell_{y_i}(s) + \tfrac{\rho}{2}(s - a_i)^2, \quad a_i = k_i^\top w - \lambda_i/\rho, \tag{2}$$

$$\lambda \leftarrow (\lambda + \rho(s - Kw))/(1 + \rho v), \tag{3}$$

$$w \leftarrow (K^\top K + \mu I)^{-1} K^\top (s + \lambda/\rho). \tag{4}$$

The activity step is a one-dimensional global minimization with two closed-form branch candidates; the parameter step uses no chain rule. With $v = 0$, fixed ρ, exact local steps and $\mu \to 0$ the iteration is proximal Douglas–Rachford on the lifted problem (Appendix B), in the family of HIO/DR phase-retrieval methods (Fienup, 1982; Bauschke et al., 2002).

Table 1: Why self-iteration ($\nu = 0$). Left: success rate (query MSE < 0.05) at $n = 4d$, 64 tasks per cell. Right: end-to-end TTT layer ($d = 32$; 64-dim observations with unknown 32-dim signal subspace), success at $n = 4d$ (two seeds: 0.90, 0.90) and overlap of the learned keys with the signal subspace.

Inner loop / head	$d = 16$	32	64	128	256
Lifted self-iteration (1 start)	0.94	0.94	0.95	1.00	1.00
Backprop, Adam (64×3 starts)	1.00	0.83	0.34	0.00	0.00
Backprop, Levenberg–Marquardt (64)	0.94	0.55	0.02	0.00	0.00
Random features / RBF / cubic kernel	0.00	0.00	0.00	0.00	0.00
Additional layer (2-layer MLP)	0.00	0.00	0.00	0.00	0.00

TTT inner loop	success	overlap
Lifted self-iteration	0.90	0.999
official GD	0.00	0.51
full-batch GD (30/150 steps)	0.00	0.49
closed-form ridge	0.00	0.42
Transformer (30k steps)	0.00	–

3 Why self-iteration

On an exactly decodable base ($\nu = 0$) the question is which inner loop can learn the relation at all. We use the multi-branch family $y = g(u^\top x)$, $g = \mathrm{He}_3/\sqrt{6}$, $x \sim \mathcal{N}(0, I_d)$, u uniform on the sphere (three preimage branches, information exponent 3).

Theorem 1 (Fixed-feature floor). For any u-independent features $\varphi : \mathbb{R}^d \to \mathbb{R}^M$, $M \geq d$, and any coefficients $a(u)$, $\mathbb{E}_u \mathbb{E}_x \big(g(u^\top x) - a(u)^\top \varphi(x)\big)^2 \geq \frac{d-1}{d+2}\big(1 - \frac{(M-d)_+}{N_3}\big)$, where $N_3 = \binom{d+2}{3} - d$.

The bound holds with unlimited context and an oracle fit, so it covers ridge, random-feature and kernel heads (Appendix A.4). Gradient credit is correlational: the one-pass online update of official TTT needs $n \gtrsim d^{3/2}$ samples (Ben Arous et al., 2021; Damian et al., 2023), while full-batch reuse helps (Dandi et al., 2024; Lee et al., 2024). The lifted activity step is a nonlinear label transformation with nonzero first Hermite coefficient, so its first parameter step already has a dimension-free overlap with u; multipliers break self-consistent wrong branch assignments, because at a block-stationary state with residual r_0 they tilt the energy gap to a legal competitor as $\Delta - t\rho\kappa$ (Appendix A.5).

Table 1: the lifted loop's sample threshold is $n^\star \approx 3$–$4d$ for d from 32 to 256 (single-start full-batch backpropagation needs 16–$24d$); closed-form heads sit at the floor of Theorem 1 (0.957 measured, 0.966 predicted at $d = 128$, $M = 4096$); inside a trained TTT layer the outer loop learns the signal subspace through the lifted loop and through no other inner loop. For i.i.d. Gaussian single-index problems, message passing is a strong alternative (Rangan, 2011); our comparison is among inner loops a TTT layer can host.

4 Representation: requirements that hold for every learner

A learner is any map from the context and a query key to a prediction, with any number of iterations; query items are fresh draws. NMSE is the mean squared error divided by $\mathrm{Var}(y)$.

Theorem 2 (R1: decodable representation). Under equation 1 with base noise ratio $\nu \in (0, 1)$: (a) every learner has NMSE $\geq \mathrm{floor}(\nu) := \mathbb{E}[\mathrm{Var}(y \mid m)]/\mathrm{Var}(y)$, where $\mathrm{Var}(y \mid m) = m^2 + \sigma^2 - F_\sigma(m)^2 > 0$; floor increases from $\mathrm{floor}(0) = 0$ to $\mathrm{floor}(1) = 1$ (numerically $\mathrm{floor}(0.06) = 0.14$, $\mathrm{floor}(0.2) = 0.40$, $\mathrm{floor}(0.3) = 0.55$, $\mathrm{floor}(0.5) = 0.77$); (b) every learner's context carries information $I(u; D) \leq n\, C(\nu)$ about the task direction, with $C(\nu) \leq \frac{1}{2}\log(1/\nu)$ nats per example; by Fano's inequality, recovering u to within a δ-packing of the unit sphere with error probability at most P_e requires $n \geq \big((1 - P_e)(d-1)\log(1/\delta) - \log 2\big)/C(\nu)$. For Gaussian inputs the sign loss reduces the per-example information from 0.347 (unsigned channel) to 0.107 nats at $\nu = 0.5$.

Part (a) holds even for a learner told u; it is the price of the base, not of the learner. A base with floor(ν) close to one supports no useful inference-time learning of this relation by any method.

Corollary 3 (R1, coverage). If inputs are isotropic Gaussian, keys are an orthogonal projection $k = Px$, and a deployment direction u has a component $u_\perp$ outside range(P), then every learner that sees only keys faces Theorem 2 with the effective noise ratio $\nu_{\text{eff}} = \nu + (1-\nu)\|u_\perp\|^2/\|u\|^2$.

Proposition 4 (R2: identifiability). (a) If $n < d$, for every sign pattern the set of directions consistent with a noiseless context is an affine subspace of dimension at least $d - n$; under a prior with a density, every learner has strictly positive risk. For $u \sim \mathcal{N}(0, d^{-1}I)$ and $k \sim \mathcal{N}(0, I)$, every learner—even one told the signs of all $k_i^\top u$—has NMSE $\geq 1 - n/d$. (b) For $n \leq 2d - 2$, $w \mapsto |Kw|$ is not injective modulo sign for any K, so no learner recovers every direction; for $n \geq 2d - 1$ it is injective for generic K (Balan et al., 2006).

R1–R2 are properties of the base and the deployment protocol. They are measurable: ν is one minus the R^2 of a linear probe from the base, and d is the key dimension.

5 Uncertainty and local credit: requirements for the self-iterating learner

On a base with $\nu > 0$ the lifted loop of Section 3, designed for $\nu = 0$, must be adapted. The following conditions are proved in Appendix A; $k \sim \mathcal{N}(0, I)$ and $\|w^\star\| = 1$.

Theorem 5 (R3: consistent inversion and readout). Let $\sigma > 0$ and $c_\sigma = \mathbb{E}[(F_\sigma(m) - |m|)|m|]$ with $m \sim \mathcal{N}(0, 1)$. (a) Every fixed point of equation 2–equation 4 with $v = 0$ and $k_i^\top w \neq 0$ is a stationary point of $\hat{R}_n(w) = \sum_i (|k_i^\top w| - y_i)^2 + \frac{\rho\mu}{2}\|w\|^2$. (b) The population risk $R(w) = \mathbb{E}(|k^\top w| - y)^2$ has $\nabla R(w^\star) = -2c_\sigma w^\star \neq 0$ and is minimized along $\{\alpha w^\star\}$ at $\alpha^\star = 1 + c_\sigma > 1$. (c) $\mathbb{E}[y \mid k] = F_\sigma(k^\top w^\star)$ and $F_\sigma(0) = \sigma\sqrt{2/\pi} > 0$, so a readout $|k^\top w|$ is never Bayes-optimal. (d) The posterior inversion $T_\tau(y, a) = y \tanh(ya/\tau)$ satisfies $\mathbb{E}[T_{\sigma^2}(y, m) \mid k] = m$, so $w^\star$ solves the population fixed-point equation of $w \leftarrow \mathrm{LS}(K, T_{\sigma^2}(y, Kw))$; for $\tau \to 0$ the identity fails.

The implemented activity step averages the two branch candidates of equation 2 with Gibbs weights at temperature τ; it equals $\frac{\rho a}{2+\rho} + \frac{2y}{2+\rho}\tanh\big(\frac{2\rho y a}{(2+\rho)\tau}\big)$ and tends to T_{σ^2} as $\rho \to 0$ with $\tau/\rho = \sigma^2$.

Theorem 6 (R5: noise-dependent multiplier leak). With hard inversion and leak $v \geq 0$: (a) the fixed points are exactly the self-consistent branch assignments $b \in \{\pm 1\}^n$ with $w = \arg\min_w \sum_i (k_i^\top w - b_i y_i)^2 + \mu_v\|w\|^2$, $\mu_v = \frac{\rho\mu}{2}(1 + \frac{2v}{1+\rho v})$—the multiplier decides which fixed point is reached, not the set of fixed points; (b) at a block-stationary state with residual r_0 and a competitor with gap $\Delta > 0$ and tilt $\kappa > 0$, the gap after t steps is $\Delta - \kappa c_t$, $c_t = \rho\sum_{j=1}^{t}(1 + \rho v)^{-j}$; for $v = 0$ every such state is left, for $v > 0$ exactly those with $\Delta < \kappa/v$; (c) for a fixed correct assignment, prior $w^\star \sim \mathcal{N}(0, d^{-1}I)$ and lifted noise of variance σ^2, the MSE-optimal ridge is $\mu^\star = d\sigma^2$.

So v must vanish at $\nu = 0$ (escape from barriers of any height) and grow with the base noise (regularization); we set the slack variance from the in-context noise estimate, $v = g\hat{r}^2$.

Proposition 7 (R4: unbiased in-context noise estimate). For a linear smoother $H = K(K^\top K + \mu I)^{-1}K^\top$ and targets $t = Kw^\star + \xi$, $\mathrm{Cov}(\xi) = \sigma^2 I$: $\mathbb{E}\,\mathrm{RSS} = \|(I - H)Kw^\star\|^2 + \sigma^2(n - 2\,\mathrm{tr}H + \mathrm{tr}H^2)$; for $\mu \to 0$, $\mathbb{E}[\mathrm{RSS}/n] = \sigma^2(1 - d/n)$ and $\mathbb{E}[\mathrm{RSS}/(n - d)] = \sigma^2$. With $v = g\,\mathrm{RSS}/n$ the loop has positive feedback: smaller v lowers μ_v, raises $\mathrm{df} = \mathrm{tr}H$ and lowers $\mathbb{E}\,\mathrm{RSS}/n$.

Proposition 8 (R2, learner side: posterior mixtures). If the posterior over memories is $\pi\delta_{w_1} + (1 - \pi)\delta_{w_2}$ with $w_1 \neq \pm w_2$ and F is strictly increasing in $|m|$, then for almost every query $\Delta(k) = F(k^\top w_1) - F(k^\top w_2) \neq 0$, and the point predictor $F(k^\top w_j)$ has excess risk $(1-\pi)^2\Delta^2$ or $\pi^2\Delta^2$ over the Bayes mixture.

Theorem 9 (R6: branch-controlling knobs must be selected). Let $w_K(\phi; D)$ be the output of K iterations with hard inversion, ϕ a solver knob (ρ, v or a temperature scale), $L(\phi) =$

Algorithm 1 Self-iterating inner loop satisfying R2–R6 (R2 through the chain mixture) (per context; R chains)

Require: keys K, targets y; $\rho = 1$; outer-trained μ, c, p, s_r, a_0, a_1; leak scale g selected on training data (R6)

df $\leftarrow \operatorname{tr}\big((K^\top K + \mu I)^{-1} K^\top K\big)$
for chain $r = 1, \dots, R$ from a random w and $\lambda = 0$ do
  for $t = 1, \dots, T$ do
    $u \leftarrow Kw$; $\hat{r}^2 \leftarrow \frac{1}{n - \mathrm{df}} \sum_i (y_i - |u_i|)^2$ ▷ R4
    $s \leftarrow$ Gibbs average of the two minimizers of equation 2 at $a = u - \lambda/\rho$, temperature $c\,\hat{r}^2$ ▷ R3
    $\lambda \leftarrow (\lambda + \rho(s - u))/(1 + \rho g \hat{r}^2)$ ▷ R5
    $w \leftarrow (K^\top K + (\mu + p)I)^{-1}\big(K^\top (s + \lambda/\rho) + p\,w\big)$ ▷ proximal least squares
  end for
  $\hat{y}_q^{(r)} \leftarrow a_0 F_{\sqrt{s_r}\hat{r}}(k_q^\top w) + a_1$; $\omega_r \propto \exp(-\frac{n}{2} \hat{r}_r^2 / \min_{r'} \hat{r}_{r'}^2)$ ▷ R3, R2
end for
return $\sum_r \omega_r \hat{y}_q^{(r)}$

$\mathbb{E}_D\, \ell(w_K(\phi; D))$ the outer loss, and suppose the branch pattern partitions task space into regions whose boundary $\Gamma(\phi)$ moves smoothly with normal velocity V, with $\ell \circ w_K$ smooth inside each region. Then

$$L'(\phi) = \underbrace{\mathbb{E}_D\big[\partial_\phi\, \ell(w_K(\phi; D))\big]}_{\text{backpropagation through the solver}} + \underbrace{\int_{\Gamma(\phi)} [\ell]_\Gamma\, V\, p\, dS}_{\text{change of the branch reached}} . \tag{5}$$

With a Gibbs temperature $\tau > 0$ the pathwise derivative is exact but concentrates with magnitude $O(1/\tau)$ in bands of width $O(\tau)$ around Γ. For $v = g\hat{r}^2$ every derivative with respect to g carries a factor $\hat{r}_t^2$ and vanishes on noiseless training tasks.

Backpropagation therefore cannot tune the knobs whose job is to change branches, nor transfer a leak learned on low-noise training tasks to a noisier deployment; in the TTT layer of Section 3, a learned ρ grows from 1 to 219 and success falls from 90% to 0%. The knobs are selected on training data from the deployment regime, and their per-task adaptation comes from in-context statistics (R4).

The learner and its matching rule. Algorithm 1 implements R2–R6. Its members differ only in the leak scale g: the exact member ($\tau, v \to 0$), and noise-calibrated members $g \in \{1, 10\}$. Theorems 5 and 6 predict the matching: $\nu = 0$ calls for the exact member, small ν for a weak leak, larger ν for a strong leak, and a context near $2d$ for mixtures.

6 Mechanism experiments: every prediction against its measurement

Each theorem makes a numerical prediction that can be checked in seconds with true keys and exact arithmetic (float64, 64–512 tasks per point). Table 2 and Figure 1 put every prediction next to its measurement.

What the checks establish. The floor of R1 is not an upper estimate: the Bayes oracle attains it, so any learner on a base with $\nu = 0.5$ is limited to NMSE ≥ 0.78, whatever it iterates. The R2 bound holds with every sign revealed, so it binds every learner that sees less. The fixed-point theorem of R5 is exact, and it explains a second effect of the leak visible in (d): beyond regularizing, it turns cycling nonconvex iterations into convergent ones. The ERM inflation of R3 and the tightness of Theorem 1 are matched to three decimals.

Table 2: Mechanism experiments. Every prediction next to its measurement (true keys, float64).

Claim	Setting	Predicted	Measured
Theorem 1 is tight at $M = d$	$d = 8/16/32/64$, optimal features	.700/.833/.912/.955	.697/.831/.914/.954
R1: the Bayes oracle sits on floor(ν)	$\nu = 0.1/0.3/0.5$	.223/.550/.776	.224/.552/.778
R1: no learner goes below floor(ν)	$\nu \in [0.05, 0.6]$, four learners	ratio ≥ 1	minimum ratio 1.04
R2: NMSE $\geq 1 - n/d$ for $n < d$	$n/d = 0.5/0.75$, signs revealed	$\geq .50/.25$	.75/.45
R3: ERM inflation $\alpha^\star = 1 + c_\sigma$	$\sigma = 0.2/0.5/0.8$, $n = 64d$	1.0017/1.0232/1.0797	1.0025/1.0246/1.0862
R5: fixed point = ridge(μ_v)	$v \in \{0, \ldots, 10\}$, converged tasks	0	$\leq 1.2 \times 10^{-15}$

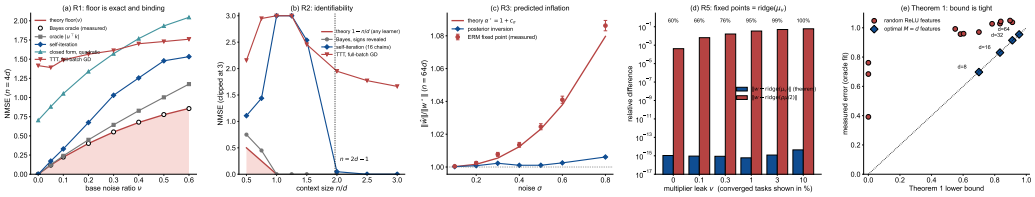


Figure 1: Mechanism experiments. (a) R1: on bases with known ν the Bayes oracle $F_s(u^\top k)$ lies on the theoretical floor (largest deviation 0.004); an oracle $|u^\top k|$ probe, self-iteration, the quadratic closed form and TTT gradient updates all stay above it ($n = 4d$, $d = 16$). (b) R2: noiseless contexts; even the Bayes predictor told every sign cannot beat $1 - n/d$ below $n = d$; self-iteration (16 chains) solves from $n \approx 2d$ (0.048 at $2d$, 0 from $2.5d$), full-batch gradient updates do not. (c) R3: norm of the ERM fixed point started at $w^\star$ (mean ± 2 s.e., 128 tasks, $n = 64d$) follows $\alpha^\star(\sigma)$; posterior inversion stays at 1; the remaining excess shrinks with n. (d) R5: at every converged fixed point of the leaky iteration, w equals ridge regression with μ_v to machine precision, while the unleaked ridge differs by up to 6%; the percentage of tasks that converge rises from 60% ($v = 0$) to 100% ($v = 10$), since the leak removes limit cycles of the nonconvex iteration. (e) Theorem 1: with the optimal $M = d$ features $x_i(\|x\|^2 - d - 2)$ the oracle-fit error equals the bound; random ReLU features stay above it.

7 Verification on controlled bases

We verify the learner conditions on controlled bases with known ν and true keys ($d = 32$, 512 tasks per cell, float64; Figure 2).

For R3, gradient descent on the ERM started at $w^\star$ reaches $\|w\| = 1.0257$ at $\sigma = 0.5$, $n = 64d$ (theory 1.0232); posterior inversion stays at 1.002 and its error falls from 0.365 ($4d$) to 0.074 ($64d$), while exact inversion, a Douglas–Rachford iteration on an inconsistent problem (Bauschke & Moursi, 2017), keeps twice that error. For R4, at $\sigma = 0.5$ the raw estimate is 0.37/0.67/0.84 of the truth at $n = 2d/4d/8d$ (theory 0.50/0.75/0.88, the gap being the extra freedom of branch choices) and the corrected one 0.75/0.89/0.96. For R6, in every configuration the pathwise derivative differs from the true one in sign or by orders of magnitude; with a Gibbs temperature of 0.05 it reaches 10^4 in magnitude with random sign.

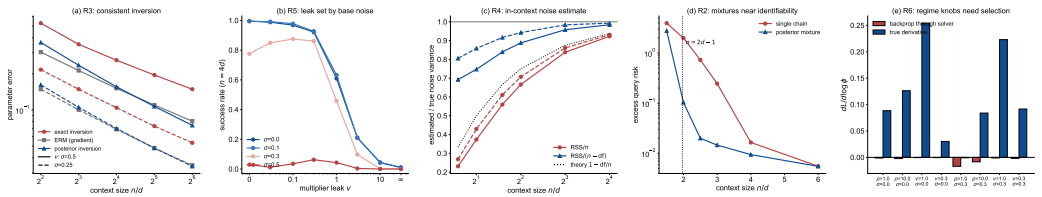


Figure 2: Learner conditions on controlled bases. (a) R3: started at $w^\star$, ERM stationary points move to the norm predicted by Theorem 5(b) (1.026 measured vs $\alpha^\star = 1.023$ at $\sigma = 0.5$, $n = 64d$; 1.0045 vs 1.0032 at $\sigma = 0.25$), exact inversion does not settle, and posterior inversion is consistent ($\|w\| \to 1$, error $\propto n^{-1/2}$). (b) R5: on an exactly decodable base the unleaked multiplier is optimal and success decreases monotonically in the leak (0.996 at $v = 0$, 0.01 without multiplier); at $\sigma = 0.3$ the optimum moves to $v \in [0.03, 0.1]$. (c) R4: the raw noise estimate follows $1 - \mathrm{df}/n$; the corrected one is within 6% of the truth for $n \geq 4d$. (d) R2: near $n = 2d - 1$ a likelihood-weighted mixture of 16 chains lowers excess risk from 1.99 to 0.10 ($n = 2d$) and 0.25 to 0.015 ($n = 3d$); far above the threshold single chains suffice. (e) R6: derivative of the outer loss in $\log \rho$ and $\log v$; backpropagation through the solver returns ≈ 0 where the true derivative (central differences on identical tasks) is $+0.03$ to $+0.25$.

8 Real bases

Bases and tasks. NLP: 13 attributes of the MRC psycholinguistic database (Coltheart, 1981); bases GloVe (Pennington et al., 2014) and MiniLM (mean-pooled sentence encoder); pairs of words with $|\Delta\text{attribute}|$; outer training on 7 attributes, deployment on 5 held-out attributes (word-frequency cluster, age of acquisition). Graph: QM9 (Ramakrishnan et al., 2014) with a 4-layer GINE (Hu et al., 2020) pretrained on 14 properties and frozen; deployment on 5 held-out properties. CV: compressive phase retrieval of CIFAR-10 test images (Krizhevsky, 2009) (grayscale 32×32) on a 64-dimensional PCA code of the training images, $n \in \{2, 3, 4, 6\} \times 64$ Gaussian measurements. All inner loops share keys ($d = 16$ from 64-dim whitened embeddings for NLP/Graph), data, 3000 outer steps and query loss; 1280 (NLP/Graph) or 1024 (CV) held-out sequences per cell, identical across methods.

Diagnostics. Table 3 lists each base's diagnostics. The measured floors agree with Theorem 2: the oracle probe of every admissible base lies just above floor(ν), and for the image code the predicted floor(0.026) = 0.065 matches the directly measured error of the true image projected on the code (0.067). GloVe, GINE and the PCA code satisfy R1 and R2 ($n \geq 2d$ throughout); MiniLM, with $\nu = 0.52$, does not. On GINE the training properties are nearly exact ($\nu \approx 0.01$) while the deployment properties are not ($\nu = 0.22$): the regime shift that R6 says backpropagation cannot bridge, and that training-data selection handles.

Bases that satisfy R1–R2. Table 4 and Figure 4: on every admissible base the self-iterating learner is the best inner loop in every NLP and Graph cell and at every CV context size from $3m$, where it approaches the base floor (0.088 vs 0.065); at $2m$, on the identifiability threshold, it ties the quadratic closed form in mean. The margin over TTT's full-batch gradient update is 0.035–0.041 on NLP ($4d$, $8d$) and 0.04–0.11 on Graph, and over the quadratic closed form 0.07–0.26. The first seed gives the same ordering (Appendix C). The member selected on training data (Table 3) is the one Theorems 5–6 predict from ν: the exact member for $\nu = 0$, a weak leak for the nearly exact image code, a strong leak for $\nu \approx 0.2$–0.3; in each case it was also the best member on held-out data (Appendix C).

A base that violates R1. On MiniLM ($\nu = 0.52$, floor = 0.79) and on the synthetic base with $\sigma = 1$ ($\nu = 0.51$, floor = 0.78) every learner stays above the floor: self-iteration 0.94/1.00, TTT gradient update 0.97/1.01, quadratic closed form 0.99/1.06, Transformer 1.03/1.02 and even the oracle probe trained on all labels 1.02/1.03 (Figure 3b). This is the failure Theorem 2 proves for every learner; it is a property of the base.

Table 3: Base diagnostics predict the outcome. ν: mean over deployment tasks of $1 - R^2$ of a linear probe from the base; floor(ν): Theorem 2; oracle probe: $|\text{probe}(a) - \text{probe}(b)|$ trained on all labels; last columns: held-out NMSE at $n = 4d$ (CV: $4m$). Selection of the member used training data only.

Base	ν	floor(ν)	oracle probe	R1–R2	member	self-iteration	best baseline
synthetic, $\sigma = 0$	0.00	0.00	0.00	yes	exact	0.312	1.008
PCA-64 image code (CV)	0.026	0.065	0.067	yes	$g = 1$, $R = 8$	0.096	0.182
synthetic, $\sigma = 0.5$	0.20	0.41	0.47	yes	$g = 10$	0.909	1.027
GINE (Graph)	0.22	0.39	0.45	yes	$g = 10$	0.780	0.862
GloVe (NLP)	0.29	0.53	0.57	yes	$g = 10$	0.700	0.735
MiniLM (NLP)	0.52	0.79	1.02	R1 violated	–	0.940	0.972
synthetic, $\sigma = 1$	0.51	0.78	1.03	R1 violated	–	1.002	1.011

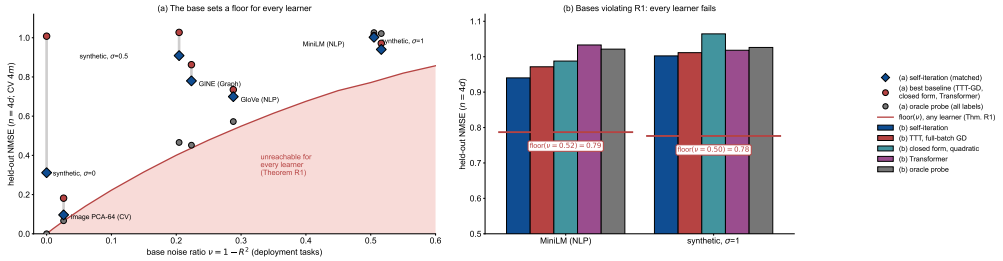


Figure 3: The base decides. (a) Held-out NMSE against the base noise ratio ν. The shaded region lies below floor(ν) and is unreachable for every learner (Theorem 2); the oracle probe sits on the floor. On bases satisfying R1–R2 the matched self-iterating learner (diamonds) is below the best baseline (red). (b) On bases that violate R1, every learner—self-iteration, TTT gradient update, quadratic closed form, Transformer and the oracle probe—stays above the proven floor.

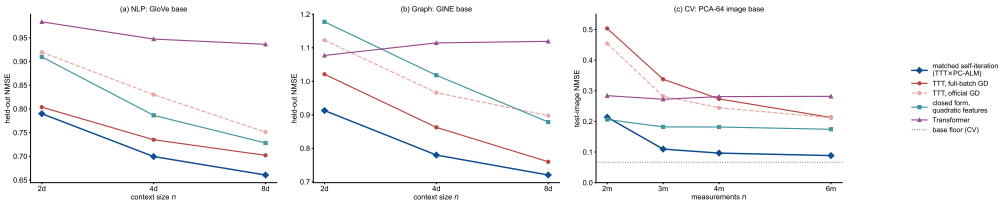


Figure 4: Bases that satisfy R1–R2. Held-out NMSE against context size. NLP and Graph: second seed (new key initialization, training batches and evaluation sequences). CV: test images; dotted line: projection floor of the image code.

Table 4: Bases that satisfy R1–R2: held-out NMSE (NLP/Graph at $n = 2d/4d/8d$, seed 2; CV at $n = 2m/3m/4m/6m$). Paired Wilcoxon tests of self-iteration against each baseline: $p \leq 0.034$ in all NLP/Graph cells (all $p < 10^{-10}$ except full-batch GD at NLP $2d$), $p < 10^{-150}$ for CV at $n \geq 3m$; at CV $2m$ the quadratic head ties in mean (difference $+0.007$, 95% CI $[-0.007, 0.023]$) while self-iteration is better on 70% of images. CV self-iteration entries use $R = 8$ likelihood-weighted chains (R2); a single chain gives .761/.177/.102/.089; all baselines are single models. First-seed NLP and Graph values are in Appendix C.

Inner loop	NLP (GloVe)	Graph (GINE)	CV (PCA-64)
Self-iteration, Algorithm 1	.790/.700/.661	.913/.780/.720	.213/.109/.096/.088
TTT, full-batch GD	.804/.735/.702	1.021/.862/.760	.503/.338/.273/.213
TTT, official GD	.920/.830/.751	1.123/.966/.898	.454/.282/.245/.212
Closed form, quadratic features	.910/.786/.728	1.177/1.018/.878	.206/.182/.182/.174
Closed form, linear ridge	1.051/1.024/1.015	1.055/1.029/1.025	2.07/1.54/1.35/1.22
Transformer (additional layers)	.984/.947/.936	1.077/1.115/1.119	.284/.272/.281/.282
floor(ν) / oracle probe	.53 / .57	.39 / .45	.065 / .067

9 Related work

Test-time training and in-context learning. TTT layers (Sun et al., 2023; 2024), fast weights (Schlag et al., 2021) and test-time regression (Wang et al., 2026) define the inner-loop design space; Transformers can implement gradient-like in-context learners (Garg et al., 2022; von Oswald et al., 2023). We characterize which bases admit an iterating nonlinear inner loop and what that loop must satisfy.

Meta-learning through inner solvers. MAML (Finn et al., 2017) and differentiable closed-form solvers (Bertinetto et al., 2019) backpropagate through inner learners; Theorem 9 identifies the boundary term that pathwise gradients miss for branch-controlling knobs (cf. Mohamed et al., 2020).

Predictive coding, lifting and ALM. Predictive coding (Whittington & Bogacz, 2017) and PC-ALM (Seely & Gould, 2026) introduce activities and multipliers; lifted, MAC and ADMM training decouple layers (Carreira-Perpiñán & Wang, 2014; Taylor et al., 2016; Askari et al., 2018); ALM and proximal methods are classical (Hestenes, 1969; Rockafellar, 1976).

Phase retrieval and single-index learning. HIO/DR/RAAR (Fienup, 1982; Bauschke et al., 2002), Wirtinger flow (Candès et al., 2015) and AMP (Rangan, 2011) solve multi-branch problems; injectivity needs $2d - 1$ real measurements (Balan et al., 2006). Single-index theory separates correlational and data-reuse algorithms (Ben Arous et al., 2021; Damian et al., 2023; Dandi et al., 2024; Lee et al., 2024). EM (Dempster et al., 1977) underlies R3 and degrees-of-freedom corrections (Craven & Wahba, 1979) underlie R4.

10 Scope and conclusion

Our bases are frozen encoders with a trained projection, the relations are pair differences and simulated measurements of real images, and the models are small; Theorem 2 assumes the Gaussian residual model equation 1, Theorem 6(c) a fixed branch assignment, and Theorem 9 smoothly moving branch boundaries.

Inference-time self-iteration earns its compute on multi-branch relations, where gradient, closed-form and additional-layer inner loops cannot follow. Whether a pretrained model can use it is decided by the base: its undecodable fraction sets a floor no learner can pass, its keys must cover the deployment directions, and its task dimension must fit the context. On a base that meets these conditions, a self-iterating learner that is consistent under the base's residual noise, leaks its multipliers accordingly, corrects its noise estimates, mixes near identifiability and selects its branch-controlling knobs on data, is the best inner loop across language, molecular and image bases.

Reproducibility statement

Code, run configurations, per-sequence result files and figure scripts are kept in the authors' research repository; they are not attached to this submission and will be released as an anonymized archive. Every number in the paper is read from a saved result file, and all figures are regenerated from the raw result files. Section 8 and Appendix C give data sources, splits, seeds and hyper-parameters. Member selection used training data only; evaluation sequences are fixed by seed and identical across models, which makes the paired tests valid.

AI use statement

In this work, we used generative AI tools for two purposes: to polish the writing, and to create the figures. All figures were drawn by coding agents (Claude Code and Codex) following figure-making skills, i.e., instruction files that fix the house style, the panel layout and automatic legibility checks. Each figure is produced by a matplotlib script that reads the archived experimental outputs; no image-generation model was used. We have reviewed

all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Armin Askari, Geoffrey Negiar, Rajiv Sambharya, and Laurent El Ghaoui. Lifted neural networks. arXiv preprint arXiv:1805.01532, 2018.

Radu Balan, Pete Casazza, and Dan Edidin. On signal reconstruction without phase. Applied and Computational Harmonic Analysis, 20(3):345–356, 2006.

Heinz H. Bauschke and Walaa M. Moursi. On the Douglas–Rachford algorithm. Mathematical Programming, 164:263–284, 2017.

Heinz H. Bauschke, Patrick L. Combettes, and D. Russell Luke. Phase retrieval, error reduction algorithm, and Fienup variants: A view from convex optimization. Journal of the Optical Society of America A, 19(7):1334–1345, 2002.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. Journal of Machine Learning Research, 22(106):1–51, 2021.

Luca Bertinetto, Philip H. S. Torr, Andrea Vedaldi, and João F. Henriques. Meta-learning with differentiable closed-form solvers. In International Conference on Learning Representations, 2019.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. IEEE Transactions on Information Theory, 61(4):1985–2007, 2015.

Miguel Á. Carreira-Perpiñán and Weiran Wang. Distributed optimization of deeply nested systems. In International Conference on Artificial Intelligence and Statistics, 2014.

Max Coltheart. The MRC psycholinguistic database. Quarterly Journal of Experimental Psychology, 33A:497–505, 1981.

Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. Numerische Mathematik, 31:377–403, 1979.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. In Advances in Neural Information Processing Systems, 2023.

Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. In International Conference on Machine Learning, 2024.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B, 39(1):1–38, 1977.

James R. Fienup. Phase retrieval algorithms: A comparison. Applied Optics, 21(15):2758–2769, 1982.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning, pp. 1126–1135, 2017.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In Advances in Neural Information Processing Systems, 2022.

Magnus R. Hestenes. Multiplier and gradient methods. Journal of Optimization Theory and Applications, 4(5):303–320, 1969.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In International Conference on Learning Representations, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. In Advances in Neural Information Processing Systems, 2024.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte Carlo gradient estimation in machine learning. Journal of Machine Learning Research, 21(132):1–62, 2020.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In Conference on Empirical Methods in Natural Language Processing, 2014.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. Scientific Data, 1: 140022, 2014.

Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In IEEE International Symposium on Information Theory, 2011.

R. Tyrrell Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Mathematics of Operations Research, 1(2):97–116, 1976.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In International Conference on Machine Learning, 2021.

Jeffrey Seely and Julian Gould. Augmented lagrangian predictive coding. arXiv preprint arXiv:2605.31022, 2026.

Yu Sun, Xinhao Li, Karan Dalal, Chloe Hsu, Sanmi Koyejo, Carlos Guestrin, Xiaolong Wang, Tatsunori Hashimoto, and Xinlei Chen. Learning to (learn at test time). arXiv preprint arXiv:2310.13807, 2023.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. arXiv preprint arXiv:2407.04620, 2024.

Gavin Taylor, Ryan Burmeister, Zheng Xu, Bharat Singh, Ankit Patel, and Tom Goldstein. Training neural networks without gradients: A scalable ADMM approach. In International Conference on Machine Learning, 2016.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In International Conference on Machine Learning, 2023.

Ke Alexander Wang, Jiaxin Shi, and Emily B. Fox. Test-time regression: A unifying framework for designing sequence models with associative memory. Journal of Machine Learning Research, 27(169):1–41, 2026.

James C. R. Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. Neural Computation, 29(5):1229–1262, 2017.

A Proofs

A.1 Theorem 2

(a) The query label y_q is conditionally independent of the context given (m_q, u), because query items are fresh. For any learner $\hat{y}_q = h(k_q, D)$, the law of total variance gives $\mathbb{E}(y_q - \hat{y}_q)^2 \geq \mathbb{E}\,\mathrm{Var}(y_q \mid k_q, D) \geq \mathbb{E}\,\mathrm{Var}(y_q \mid k_q, D, u) = \mathbb{E}\,\mathrm{Var}(y_q \mid m_q)$. For $y = |m + e|$ with $e \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{E}[y^2 \mid m] = m^2 + \sigma^2$ and $\mathbb{E}[y \mid m] = F_\sigma(m)$; since y is non-degenerate given m, $\mathrm{Var}(y \mid m) > 0$. Dividing by $\mathrm{Var}(y) = 2 - \frac{4}{\pi}$ (for $m + e \sim \mathcal{N}(0, 2)$) gives the floor, which is 0 at $\nu = 0$ and 1 at $\nu = 1$; the listed values are Monte Carlo evaluations with 4×10^5 samples. (b) By the chain rule $I(u; D) = \sum_i I(u; y_i \mid k_i, y_{<i}) \leq \sum_i I(m_i; y_i \mid k_i)$, since given k_i, $u \to m_i \to y_i$ is Markov and y_i is independent of $y_{<i}$ given m_i. As $y_i = |m_i + e_i|$ is a function of $m_i + e_i$, data processing gives $I(m_i; y_i \mid k_i) \leq I(m_i; m_i + e_i \mid k_i) \leq \frac{1}{2}\log(1 + \mathrm{Var}(m_i)/\mathrm{Var}(e_i)) = \frac{1}{2}\log(1/\nu)$ by the Gaussian channel capacity. Fano's inequality over a δ-packing of size $M \geq \delta^{-(d-1)}$ gives $(1 - P_e)\log M - \log 2 \leq I(u; D)$. The folded-channel values for Gaussian m follow from $h(y) = \frac{1}{2}\log(4\pi e) - \log 2$ and a Monte Carlo evaluation of $h(y \mid m)$.

A.2 Corollary 3

With $x \sim \mathcal{N}(0, I)$ and P an orthogonal projection, $u^\top x = u_\parallel^\top x + u_\perp^\top x$ where $u_\perp^\top x$ is independent of Px. For a learner that sees only Px, $u_\perp^\top x$ is indistinguishable from the base residual, so the undecodable variance is $\nu + (1 - \nu)\|u_\perp\|^2/\|u\|^2$ of the attribute variance and Theorem 2 applies.

A.3 Proposition 4

(a) For a fixed sign pattern b, noiseless consistency means $Kw = b \odot y$, whose solution set is empty or an affine subspace of dimension $d - \mathrm{rank}K \geq d - n$. Under a prior with a density the posterior then has positive variance along this subspace, so $k_q^\top w$ has positive posterior variance for almost every k_q, and every predictor has positive excess risk. Quantitatively, reveal the signs, so that the context determines Ku exactly; a learner with less information cannot do better. With $u \sim \mathcal{N}(0, d^{-1}I)$ the posterior of $m_q = k_q^\top u$ is $\mathcal{N}(\mu_q, v_q)$ with $v_q = \|P_\perp k_q\|^2/d$, $P_\perp$ the projection onto the null space of K, so $\mathbb{E}v_q = (d - n)/d$. The variance of a folded normal $|\mathcal{N}(\mu, v)|$ is minimized at $\mu = 0$, where it equals $(1 - 2/\pi)v$; hence $\mathbb{E}\,\mathrm{Var}(y_q \mid D, \mathrm{signs}) \geq (1 - 2/\pi)(d - n)/d$, while $\mathrm{Var}(y_q) = (1 - 2/\pi)\mathbb{E}m_q^2 = 1 - 2/\pi$. Dividing gives NMSE $\geq 1 - n/d$. (b) is the complement-property characterization of Balan et al. (2006).

A.4 Theorem 1

For $f_u(x) = g(u^\top x)$ and Gaussian x, $\langle f_u, f_v \rangle = (u^\top v)^3$. The covariance operator $\mathbb{E}_u f_u \otimes f_u$ has the nonzero spectrum of the kernel $(u^\top v)^3$ on the sphere, which the Funk–Hecke formula diagonalizes in spherical harmonics; t^3 has Gegenbauer components of degrees 1 and 3 only. The degree-1 eigenvalue is $3/(d(d+2))$ (multiplicity d) and, since the trace is 1, the degree-3 eigenvalue is $(d-1)/((d+2)N_3)$ (multiplicity N_3). The best M-dimensional u-independent approximation leaves at least the eigenvalues outside the top M.

A.5 Multiplier escape and Theorem 6(b)

While the primal stays at z_0 with residual r_0, the multiplier accumulates $\lambda_t = \rho r_0 \sum_{j=1}^t \beta^j$ with $\beta = (1 + \rho v)^{-1}$ ($\lambda_t = t\rho r_0$ for $v = 0$). Block energies are affine in λ at fixed primal states, so the gap to a legal competitor z' is $\Delta - \kappa c_t$ with $\kappa = -r_0^\top (r(z') - r_0)$ in the block normalization. For $v > 0$, $c_t \uparrow \rho\beta/(1 - \beta) = 1/v$.

A.6 Theorem 5

(a) At a fixed point with $v = 0$, $s = Kw$. The activity step's first-order condition at $s_i \neq 0$ is $\ell'(s_i) + \lambda_i = 0$, i.e. $\lambda_i = -2(|k_i^\top w| - y_i)\operatorname{sign}(k_i^\top w)$, and the parameter step gives $\mu w = K^\top \lambda/\rho$. Together, $\sum_i 2(|k_i^\top w| - y_i)\operatorname{sign}(k_i^\top w)k_i + \rho\mu w = 0$. (b) e takes both signs with positive probability, so $F_\sigma(m) > |m|$ strictly. Conditioning on k, $\nabla R(w^\star) = 2\mathbb{E}[(|m| - F_\sigma(m))\operatorname{sign}(m)k]$; writing $k = mw^\star + k_\perp$ with $k_\perp$ centered and independent of m, $\nabla R(w^\star) = -2c_\sigma w^\star$. On the ray, $R(\alpha w^\star) = \alpha^2 - 2\alpha(1 + c_\sigma) + \mathbb{E}y^2$. (c) $\mathbb{E}[y \mid k] = F_\sigma(m)$ and $F_\sigma(0) = \mathbb{E}|e|$. (d) For $s \sim \mathcal{N}(m, \sigma^2)$, $\Pr(s > 0 \mid |s| = y) = (1 + e^{-2ym/\sigma^2})^{-1}$, hence $\mathbb{E}[s \mid |s| = y] = T_{\sigma^2}(y, m)$ and $\mathbb{E}[T_{\sigma^2}(y, m) \mid k] = \mathbb{E}[s \mid k] = m$. As $\tau \to 0$, $\mathbb{E}[T_\tau \mid k] \to \operatorname{sign}(m)F_\sigma(m) \neq m$. Gibbs activity: the branch minimizers are $s_\pm = (\pm 2y + \rho a)/(2 + \rho)$ with values $\frac{\rho}{2+\rho}(y \mp a)^2$, whose difference is $\frac{4\rho y a}{2+\rho}$.

A.7 Theorem 6

(a) For $v > 0$ the fixed-point multiplier is $\lambda = r/v$ with $r = s - Kw$; the parameter step gives $\mu w = K^\top r(1 + \frac{1}{\rho v})$; on the selected branch, $2(s_i - b_i y_i) + \rho r_i + \lambda_i = 0$ gives $r_i = -2(k_i^\top w - b_i y_i)/(2 + \rho + 1/v)$. Hence $\mu w = -2c\sum_i k_i(k_i^\top w - b_i y_i)$ with $c = \frac{1+\rho v}{\rho(1+\rho v+2v)}$, the normal equation of the stated ridge problem with $\mu_v = \mu/(2c)$. Conversely every self-consistent (b, w) defines r, $\lambda = r/v$, $s = Kw + r$ fixed by the three updates; $v = 0$ is Theorem 5(a). (c) With $t = Kw^\star + \xi$, $w^\star \sim \mathcal{N}(0, d^{-1}I)$, $\xi \sim \mathcal{N}(0, \sigma^2 I)$, the posterior mean $(K^\top K + d\sigma^2 I)^{-1}K^\top t$ minimizes the MSE among all estimators.

A.8 Propositions 7 and 8

R4: $\mathbb{E}\|(I - H)t\|^2 = \|(I - H)Kw^\star\|^2 + \sigma^2\operatorname{tr}((I - H)^2)$; for $\mu \to 0$, H is a rank-d projection. $\operatorname{tr}H = \sum_j \eta_j/(\eta_j + \mu)$ decreases in μ, $2h - h^2$ increases on $[0, 1]$ and μ_v increases in v, giving the feedback. R2: for any predictor a, $\mathbb{E}[(y - a)^2 \mid k, D] - \mathbb{E}[(y - \hat{y}_B)^2 \mid k, D] = (a - \hat{y}_B)^2$ with $\hat{y}_B = \pi F_1 + (1 - \pi)F_2$. If $w_1 \neq \pm w_2$, $|k^\top w_1| = |k^\top w_2|$ only on two hyperplanes.

A.9 Theorem 9

Write $L(\phi) = \sum_B \int_{\Omega_B(\phi)} \ell_B(\phi, D)p(D)dD$ with smooth ℓ_B; Reynolds' transport theorem gives the interior term plus boundary integrals, which combine across adjacent regions into the jump term. Backpropagation differentiates ℓ_B at a fixed pattern B and returns only the interior term. With a Gibbs temperature, the branch weight $\sigma((F_- - F_+)/\tau)$ has derivative $O(1/\tau)$ in a band of width $O(\tau)$. For $v = g\hat{r}^2$, $\partial\lambda_{t+1}/\partial g = -\rho\hat{r}_t^2(\lambda_t + \rho r_t)/(1 + \rho g\hat{r}_t^2)^2$.

B PC-ALM as proximal Douglas–Rachford

With $u = \lambda/\rho$, $a = Kw - u$, h the sum of local losses and $C = \operatorname{range}(K)$, fixed ρ, exact proximal steps, $\mu \to 0$ and the multiplier updated before the parameter, eliminating λ gives $S = \operatorname{prox}_{h/\rho}(a)$ and $a^+ = a + P_C(2S - a) - S$.

C Experimental details

Controlled bases. $d = 32$, 512 tasks per cell, float64, 300–400 inner iterations (2000 gradient steps for the ERM), success = overlap > 0.95, excess risk relative to $F_\sigma(k^\top w^\star)$ normalized by the query variance; finite differences with steps 0.05 and 0.2 in $\log\phi$ on 2048 identical tasks.

Real bases. 64-dim PCA-whitened embeddings, $d = 16$ keys; Adam ($3 \cdot 10^{-3}$, batch 64, 3000 steps; CV batch 32, 1000 steps for fixed-code models); $T = 30$ inner iterations (CV $T = 100$, $R = 8$). Base diagnostics: ridge probe on 80% of items, R^2 on the rest. Deployment attributes were chosen from decodability alone.

Member selection on training data. Training-attribute NMSE for exact/g=1/g=10 members: NLP 0.977/0.861/0.847; Graph (median over properties) 0.157/0.146/0.065; synthetic $\sigma = 0$ 0.066/0.381/0.092; training-image NMSE for CV at $2m$ 0.989/0.210/0.582. The selected member was the best on held-out data in every domain.

First seed. NLP ($2d$/$4d$/$8d$): self-iteration .809/.711/.674, full-batch GD .811/.752/.708, official GD .927/.840/.759, quadratic .877/.771/.727, Transformer .986/.953/.922. Graph: .867/.742/.688, 1.010/.871/.783, 1.078/1.009/.987, 1.162/1.029/.902, 1.097/1.146/1.166.