RETHINKING NONPARAMETRIC TEACHING VIA GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Nonparametric teaching studies how example selection can accelerate learning in reproducing kernel Hilbert spaces (RKHS). Existing formulations adopt a first-order perspective, selecting examples with large functional gradients while retaining the learner's native RKHS geometry. Through Mercer spectral analysis, we explain how this fixed geometry links the RKHS cost of prediction errors to the kernel-scaled response of first-order updates. To address the resulting spectral imbalance, we propose Geometry-Aware Teaching (GAT), a second-order framework that integrates curvature-aware batch selection with preconditioned functional updates. For realizable squared loss on finite pools, we justify its teaching score through a common lower bound on prediction-risk decrease and establish convergence guarantees for GAT and first-order teaching baselines. These guarantees distinguish first-order dependence on kernel conditioning from GAT's dependence on batch geometry. With exact batch selection and preconditioning, this spectral independence extends to GAT's prediction trajectory, which remains invariant to positive kernel-eigenvalue rescaling under a fixed teaching protocol. To identify the contribution of example selection, we further establish finite-horizon teaching optimality for a fixed preconditioned learner in a complete pool space. Empirical results confirm the practical benefits of GAT.

1 INTRODUCTION

Nonparametric teaching (Zhang et al., 2023b;a) studies how example selection can accelerate learning in reproducing kernel Hilbert spaces (RKHS) (Schölkopf & Smola, 2001; Hollander et al., 2013; Corder & Foreman, 2014). By formulating learning directly in function space, it isolates the role of data selection in shaping optimization trajectories (Singer, 1974; Zoppoli et al., 2002). In this framework, a learner updates a function via functional gradient descent to approximate a target mapping, while a teacher controls which examples are revealed at each step to influence convergence. Beyond its theoretical elegance, nonparametric teaching has demonstrated practical utility: it accelerates multilayer perceptron training for implicit neural representations that map coordinates to signal values (Zhang et al., 2024), improves training efficiency of graph convolutional networks for implicit graph-to-property mappings (Zhang et al., 2025), and enhances efficiency in training attention-based neural networks (Zhang et al., 2026a).

Despite these successes, existing nonparametric teaching paradigms predominantly rely on a first-order principle: select examples that induce large functional gradient norms under the RKHS inner product (Friedman, 2001; Zhang et al., 2023b). While effective, this gradient-based perspective leaves the learner's update metric fixed by the kernel (Schölkopf & Smola, 2001; Gunasekar et al., 2018). The spectral properties of the associated kernel operator therefore remain relevant to functional gradient dynamics (Aronszajn, 1950; Hofmann et al., 2005; Chagouel et al., 2016; Chada et al., 2024). For squared loss, the full-pool update responds weakly along small-eigenvalue directions. Adaptive teaching can change the batch covariance and couple spectral components, but each first-order batch update still passes through the kernel scaling. Thus gradient amplification alone need not correct spectral imbalance, and first-order convergence guarantees can remain sensitive to kernel conditioning (Müler, 1994; Davies, 2007).

This reveals a key limitation: current teaching strategies select examples to amplify gradients *within* the fixed RKHS geometry but do not reshape the metric under which descent is measured (Zhang

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.

et al., 2023b;a). Yet in Hilbert spaces, convergence speed is inherently geometric. The *curvature operator*, given by the Hessian of the loss functional, describes the local variation of gradients under this metric. In traditional optimization, incorporating curvature information via preconditioning or second-order methods can mitigate ill-conditioning by changing the effective update geometry (Di et al., 1995; Nocedal & Wright, 2006; Yao et al., 2021). This leads to the central question: *Can example selection and update rules jointly reshape the optimization geometry in RKHS, rather than merely operate within it?*

To answer this question, we first study the spectral structure of gradient-based teaching through the Mercer decomposition of the kernel operator. The squared RKHS error assigns inverse-eigenvalue weights to prediction-error components, while the full-pool first-order update scales their corrections by the eigenvalues themselves. Small eigenvalues therefore increase the RKHS cost of a fixed prediction error and weaken its first-order correction. To jointly address these two manifestations of spectral imbalance, we introduce *Geometry-Aware Teaching (GAT)*, a second-order framework that incorporates the Hessian of the loss functional into both teaching and learning. GAT measures batch gradients using the full-pool inverse Hessian and uses the same operator to precondition functional updates. For squared loss, the Hessian shares the positive empirical kernel spectrum, so exact preconditioning removes the kernel-eigenvalue factor from the update while preserving the teacher's influence through batch covariance.

We further analyze GAT under realizable squared loss on a finite pool, where the target belongs to the learner's working RKHS and prediction risk is evaluated over the pool. At the prescribed step size, every candidate batch admits a risk-decrease lower bound proportional to its curvature-aware score, with the same coefficient across batches. Exact score maximization therefore selects the batch with the largest certified decrease. Our convergence analysis compares GAT with first-order baselines using RKHS gradient-norm selection (Greedy) or uniform sampling (Random). First-order bounds retain the kernel condition number, while GAT's bound depends on batch covariance concentration after preconditioning. With exact batch selection and preconditioning, GAT's prediction trajectory is invariant to positive eigenvalue rescaling under the fixed protocol. Within a fixed pool eigenbasis, rescaling positive eigenvalues preserves GAT's trajectory from identical initial predictions on the same teaching task under a fixed protocol. To isolate the contribution of teaching from preconditioning, we further prove that GAT is finite-horizon optimal for a fixed preconditioned learner in a complete pool space.

Our main contributions are summarized as follows:

- We develop a geometric framework for nonparametric teaching that separates kernel scaling from teacher-controlled batch covariance and explains how small eigenvalues increase the RKHS cost of fixed prediction errors while weakening their full-pool correction.
- We propose *Geometry-Aware Teaching (GAT)*, a second-order framework that combines curvature-aware batch selection with preconditioned functional updates.
- For realizable squared loss on finite pools, we justify its teaching score through a common lower bound on prediction-risk decrease. We establish convergence guarantees for GAT, Greedy, and Random, showing how exact preconditioning removes explicit kernel-condition-number dependence from GAT's bound and, with exact batch selection, yields spectral-rescaling invariance under a fixed teaching protocol. We further isolate the contribution of example selection through finite-horizon optimality for a fixed preconditioned learner in a complete pool space.
- We evaluate GAT on representative tasks of function and image teaching, demonstrating faster convergence and improved reconstruction relative to first-order Greedy and Random at matched teaching budgets.

2 Related Works

Nonparametric Teaching (Zhang et al., 2023b;a) provides a principled framework for teaching nonparametric target mappings in reproducing kernel Hilbert spaces (RKHS) through adaptive example selection. Its function-space formulation extends machine teaching (Zhu, 2015; Zhu et al., 2018), particularly iterative machine teaching (Liu et al., 2017; 2018), which typically assumes parametric target models and operates in parameter space. Unlike active learning (Settles, 2009), which emphasizes label efficiency under querying costs, and curriculum learning (Bengio et al., 2009),

which organizes training in an order inspired by increasing difficulty, nonparametric teaching assumes label availability and focuses on selecting examples that directly influence the functional descent dynamics within the RKHS. Applications include accelerated training of multilayer perceptrons for implicit coordinate-to-signal mappings (Zhang et al., 2024; 2026b), graph convolutional networks for implicit graph-to-property mappings (Zhang et al., 2025), and attention-based learners (Zhang et al., 2026a), including Transformer-based models (Vaswani et al., 2017). Nevertheless, existing formulations predominantly select examples by individual functional-gradient norms and retain first-order updates under the native RKHS metric. In contrast, GAT uses the full-pool Hessian to score batches and precondition functional updates, extending nonparametric teaching to a *second-order framework*.

Geometry-Aware Learning. Beyond first-order gradient information, convergence in Hilbert spaces depends on both the ambient metric and the curvature of the loss functional (Aronszajn, 1950; Chagouel et al., 2016; Gunasekar et al., 2018). Traditional optimization exploits this structure through preconditioning and second-order techniques such as Newton-type updates and AdaHessian (Di et al., 1995; Nocedal & Wright, 2006; Yao et al., 2021), which adjust updates according to curvature to accelerate convergence, especially in ill-conditioned settings. In kernel methods, spectral properties of the kernel operator influence functional gradient dynamics, with small-eigenvalue directions responding weakly under full-pool squared-loss updates (Hofmann et al., 2005; Chada et al., 2024). For squared loss, the risk Hessian is the covariance of kernel evaluation sections and shares the positive empirical kernel spectrum; other twice continuously differentiable convex losses introduce scalar curvature weights. GAT extends these insights to nonparametric teaching by using the full-pool Hessian both to assess candidate batches and to precondition functional updates on a finite working RKHS. Our squared-loss analysis connects the teaching score to a common risk-decrease guarantee and isolates the contribution of example selection through comparisons with the learner held fixed.

3 Background

Notation and RKHS Evaluation. We write $\|\cdot\|_{\mathrm{op}}$ for the operator norm and $A \preceq B$ when $B - A$ is positive semidefinite. Let $\mathcal{X} \subseteq \mathbb{R}^p$ be the input space, where p is the input dimension, and let $\mathcal{Y} \subseteq \mathbb{R}$ be the output space. We consider scalar-valued targets; vector-valued extensions are developed in nonparametric teaching (Zhang et al., 2023a). A symmetric positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines an RKHS $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$. Write $K_{\boldsymbol{x}} = K(\boldsymbol{x}, \cdot)$. The space is the completion of finite linear combinations of kernel sections, with

$$f = \sum_i u_i K_{\boldsymbol{x}_i}, \qquad g = \sum_j v_j K_{\boldsymbol{z}_j}, \qquad \langle f, g \rangle_{\mathcal{H}} = \sum_{i,j} u_i v_j K(\boldsymbol{x}_i, \boldsymbol{z}_j), \tag{1}$$

where $u_i, v_j \in \mathbb{R}$ and $\boldsymbol{x}_i, \boldsymbol{z}_j \in \mathcal{X}$ (Schölkopf & Smola, 2001; Zhang et al., 2023b).

Definition 1 (Evaluation functional)**.** *For $\boldsymbol{x} \in \mathcal{X}$, the bounded linear evaluation functional $\mathcal{E}_{\boldsymbol{x}} : \mathcal{H} \to \mathbb{R}$ satisfies*

$$\mathcal{E}_{\boldsymbol{x}}(f) = f(\boldsymbol{x}) = \langle f, K_{\boldsymbol{x}} \rangle_{\mathcal{H}}, \qquad f \in \mathcal{H}. \tag{2}$$

The reproducing property turns a scalar prediction into an inner product in function space (Lax, 2002; Schölkopf & Smola, 2001). In particular, the label supplied by a target $f^* \in \mathcal{H}$ is $y = f^*(\boldsymbol{x})$, and a perturbation $h \in \mathcal{H}$ changes that prediction by $\langle h, K_{\boldsymbol{x}} \rangle_{\mathcal{H}}$.

Definition 2 (Fréchet derivative and functional gradient)**.** *A functional $F : \mathcal{H} \to \mathbb{R}$ is Fréchet differentiable at f if there is a bounded linear functional $DF(f)$ such that*

$$F(f + h) = F(f) + DF(f)[h] + o(\|h\|_{\mathcal{H}}). \tag{3}$$

Its functional gradient $\nabla F(f) \in \mathcal{H}$ is the unique Riesz representer satisfying $DF(f)[h] = \langle \nabla F(f), h \rangle_{\mathcal{H}}$ for all $h \in \mathcal{H}$.

Nonparametric Teaching. A standard teaching objective balances target approximation against the number of examples (Zhang et al., 2023b):

$$\mathcal{D}^* \in \arg\min_{\mathcal{D}} \mathcal{M}\big(\mathcal{A}(\mathcal{D}), f^*\big) + \tau |\mathcal{D}|, \qquad \tau > 0. \tag{4}$$

Here $\mathcal{D}$ is a finite nonempty labeled teaching set, $\mathcal{A}$ is the learner, $\mathcal{M}$ measures function discrepancy, and τ weights teaching cost. A learner commonly seeks to minimize empirical loss,

$$\min_{f \in \mathcal{H}} \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}} \mathcal{L}\big(f(\boldsymbol{x}), y\big), \tag{5}$$

where the scalar point loss $\mathcal{L}(u, y)$ is differentiable and convex in its prediction argument u. In iterative teaching, $\mathcal{A}$ includes the initialization and update rule, while the teacher controls the sequence and repetition of examples or batches.

For a single example $\boldsymbol{x}^t$ at round t, a first-order learner updates its current function f^t by

$$f^{t+1} = f^t - \eta_F \mathcal{G}(\mathcal{L}, f^*; f^t, \boldsymbol{x}^t), \qquad \eta_F > 0, \tag{6}$$

where η_F is the first-order step size. The chain rule and reproducing property give the sample functional gradient

$$\mathcal{G}(\mathcal{L}, f^*; f, \boldsymbol{x}) = \partial_u \mathcal{L}\big(f(\boldsymbol{x}), f^*(\boldsymbol{x})\big) K_{\boldsymbol{x}}. \tag{7}$$

Here $\partial_u \mathcal{L}$ denotes differentiation with respect to the first scalar argument; see Lemmas 11 and 12 for the underlying identities (Gelfand & Fomin, 2000; Coleman, 2012). For a batch, the learner averages these sample gradients. Example selection and the learner's update geometry jointly determine the resulting change in the function.

4 GAT: Geometry-Aware Teaching

In this section, we first analyze spectral bias in first-order nonparametric teaching. We then introduce GAT through curvature-aware selection and preconditioned functional updates. Lastly, for realizable squared loss on finite pools, we establish convergence guarantees for GAT, Greedy, and Random, characterize spectral-rescaling invariance under exact batch selection and preconditioning, and prove finite-horizon optimality for a fixed preconditioned learner in complete pool spaces.

4.1 Spectral Bias under the Fixed RKHS Geometry

First-order Greedy teaching (Zhang et al., 2023b;a) selects an example with the largest functional-gradient norm:

$$\boldsymbol{x}^t \in \arg\max_{\boldsymbol{x} \in \mathcal{X}} \|\mathcal{G}(\mathcal{L}, f^*; f^t, \boldsymbol{x})\|_{\mathcal{H}}^2. \tag{8}$$

For squared loss $\mathcal{L}(u, y) = \frac{1}{2}(u - y)^2$, define the residual function $r_t(\boldsymbol{x}) = f^t(\boldsymbol{x}) - f^*(\boldsymbol{x})$. The score becomes $r_t(\boldsymbol{x})^2 K(\boldsymbol{x}, \boldsymbol{x})$. Thus, examples with the same prediction error can receive different priorities because their kernel sections have different norms. The batch version of Greedy selects the B largest individual scores.

To understand how the same geometry enters the learner's update, consider a rank-d Mercer kernel (Aronszajn, 1950; Steinwart & Scovel, 2012)

$$K_d(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=1}^{d} \lambda_k \phi_k(\boldsymbol{x}) \phi_k(\boldsymbol{x}'), \qquad \lambda_k > 0, \tag{9}$$

where $\phi_1, \ldots, \phi_d$ are orthonormal in $L^2(\mu)$ for a reference probability measure μ. The working RKHS is $\mathcal{H}_d = \operatorname{span}\{\sqrt{\lambda_k}\phi_k\}_{k=1}^d$. We use its kernel K_d throughout the finite-pool analysis; when obtained by restricting a larger RKHS, its sections are the corresponding projected evaluation sections. For $f, f^* \in \mathcal{H}_d$, write

$$f = \sum_{k=1}^{d} c_k \sqrt{\lambda_k}\phi_k, \quad \|f\|_{\mathcal{H}_d}^2 = \|c\|_2^2, \quad a = \Lambda^{1/2}(c - c_*), \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d), \tag{10}$$

where $c = (c_1, \ldots, c_d)^\top$ and $c_* = (c_1^*, \ldots, c_d^*)^\top$ are the RKHS-orthonormal coefficients of f and f^*, respectively. The vector a contains their prediction-error coefficients in the $L^2(\mu)$ basis, giving $\|f - f^*\|_{\mathcal{H}_d}^2 = \sum_k a_k^2/\lambda_k$.

To relate prediction error to RKHS distance, consider the RKHS-normalized basis function $\sqrt{\lambda_k}\phi_k$. A unit change in its RKHS coefficient therefore produces a prediction change with $L^2(\mu)$ norm $\sqrt{\lambda_k}$. To represent a fixed prediction-error coefficient a_k, the required RKHS coefficient error therefore satisfies

$$c_k - c_k^* = \frac{a_k}{\sqrt{\lambda_k}}, \qquad \|a_k \phi_k\|_{\mathcal{H}_d} = \frac{|a_k|}{\sqrt{\lambda_k}}.$$

Thus, for a fixed prediction discrepancy, a smaller eigenvalue requires a larger RKHS coefficient error and yields a larger RKHS norm through the factor $1/\sqrt{\lambda_k}$. Equivalently, this component contributes a_k^2/λ_k to the squared RKHS error. This inverse weighting can also be expressed through the integral operator $T_d h = \int K_d(\cdot, \boldsymbol{x}) h(\boldsymbol{x})\, d\mu(\boldsymbol{x})$:

$$\langle f, g\rangle_{\mathcal{H}_d} = \langle T_d^{-1} f, g\rangle_{L^2(\mu)}, \qquad f, g \in \mathcal{H}_d. \tag{11}$$

The inverse acts on the finite working space $\mathcal{H}_d$. We use φ_k for experimental basis functions; their normalization is specified in Appendix C.3.

Now let μ be uniform on n pool points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Setting $\Phi_{ik} = \phi_k(\boldsymbol{x}_i)$, we have $\Phi^\top \Phi/n = I_d$, where I_d is the d-dimensional identity. For a set of B indices $\mathcal{B}$, $1 \le B \le n$, let $\Phi_{\mathcal{B}}$ collect the corresponding rows and define $Q_{\mathcal{B}} = \Phi_{\mathcal{B}}^\top \Phi_{\mathcal{B}}/B$. Let c_t contain the coefficients of f^t, with $c_k^t = (c_t)_k$, and set $a_t = \Lambda^{1/2}(c_t - c_*)$. The exact first-order batch update with step size $\eta_F > 0$ is

$$a_{t+1} = (I_d - \eta_F \Lambda Q_{\mathcal{B}_t}) a_t. \tag{12}$$

To examine the corresponding gradient response, consider the full-pool squared-loss risk $F(f) = \frac{1}{2}\|f - f^*\|_{L^2(\mu)}^2$. Substituting the kernel decomposition in Equation 9 and using empirical orthonormality gives

$$\nabla F(f^t) = \frac{1}{n}\sum_{i=1}^{n} r_t(\boldsymbol{x}_i) K_d(\cdot, \boldsymbol{x}_i) = \sum_{k=1}^{d} \lambda_k a_{t,k} \phi_k,$$

where $a_{t,k} = (a_t)_k$. The multiplier λ_k comes directly from the kernel decomposition: a prediction-error coefficient $a_{t,k}$ induces a gradient coefficient $\lambda_k a_{t,k}$ in the same $L^2(\mu)$ basis. Consequently, for the full pool, $Q_{\mathcal{B}} = I_d$, and Equation 12 becomes

$$a_{t+1,k} = (1 - \eta_F \lambda_k) a_{t,k}.$$

Under a common step size $0 < \eta_F \le 1/\max_j \lambda_j$, the prediction correction has magnitude $\eta_F \lambda_k |a_{t,k}|$. For a fixed prediction error, a smaller eigenvalue therefore produces a weaker first-order response and slower contraction.

The two effects therefore have explicit spectral origins: the inverse weight $1/\lambda_k$ enlarges the squared RKHS error, while the kernel weight λ_k attenuates the first-order prediction update. This spectral bias motivates GAT, which combines curvature-aware example selection with preconditioned functional updates.

4.2 Reshaping RKHS Geometry with Geometry-Aware Teaching

GAT uses curvature in both decisions: how to assess a candidate batch and how to update the function after that batch is presented. The relevant curvature is the Hessian of the risk, whose operator form follows directly from evaluation in the RKHS.

Definition 3 (Hessian operator in RKHS). *For a twice continuously Fréchet differentiable functional F, its Hessian $\nabla^2 F(f)$ is the bounded self-adjoint operator satisfying, for increments $h \in \mathcal{H}$,*

$$F(f + h) = F(f) + \langle \nabla F(f), h\rangle_{\mathcal{H}} + \tfrac{1}{2}\langle h, \nabla^2 F(f) h\rangle_{\mathcal{H}} + o(\|h\|_{\mathcal{H}}^2). \tag{13}$$

Proposition 4 (Risk Hessian Operator). *Let $F(f) = \mathbb{E}_{\boldsymbol{x}\sim\mu}\mathcal{L}(f(\boldsymbol{x}), f^*(\boldsymbol{x}))$, with $\mathcal{L}(\cdot, y)$ twice continuously differentiable and convex. Assume F is finite at a reference function in the neighborhood under consideration, with locally uniform integrable bounds on $|\partial_u \mathcal{L}|\|K_{\boldsymbol{x}}\|_{\mathcal{H}}$ and $|\partial_{uu}^2 \mathcal{L}|\|K_{\boldsymbol{x}}\|_{\mathcal{H}}^2$, where the derivatives act on the first loss argument. Then*

$$\nabla^2 F(f) = \mathbb{E}_{\boldsymbol{x}\sim\mu}\big[\partial_{uu}^2 \mathcal{L}(f(\boldsymbol{x}), f^*(\boldsymbol{x}))(K_{\boldsymbol{x}} \otimes K_{\boldsymbol{x}})\big], \tag{14}$$

where $(K_{\boldsymbol{x}} \otimes K_{\boldsymbol{x}})h = \langle h, K_{\boldsymbol{x}}\rangle_{\mathcal{H}} K_{\boldsymbol{x}}$. For a finite pool the expectation is a finite sum.

On $\mathcal{H}_d$, write $K_{d,\boldsymbol{x}} = K_d(\boldsymbol{x}, \cdot)$ for the projected evaluation section. For squared loss, the restricted Hessian is the covariance of these sections: it is diagonal with entries λ_k in the RKHS-orthonormal basis associated with μ. Thus the loss curvature inherits exactly the kernel's positive spectrum. More generally, if the scalar loss curvature is bounded between constants $0 < m_\ell \le M_\ell < \infty$, then $m_\ell C_\mu \preceq \nabla^2 F(f) \preceq M_\ell C_\mu$, where $C_\mu = \mathbb{E}_{\boldsymbol{x}\sim\mu}[K_{d,\boldsymbol{x}} \otimes K_{d,\boldsymbol{x}}]$ is the evaluation covariance operator.

These inequalities relate loss curvature to the kernel spectrum even when curvature weighting changes the eigendirections. The proof of Proposition 4 is in Appendix A.1.

For a fixed teaching pool $\mathcal{P} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $y_i = f^*(\boldsymbol{x}_i)$, define $F_{\mathcal{B}}(f) = B^{-1}\sum_{i\in\mathcal{B}} \mathcal{L}(f(\boldsymbol{x}_i), y_i)$ and the full-pool risk $F(f) = n^{-1}\sum_i \mathcal{L}(f(\boldsymbol{x}_i), y_i)$. If $H_t = \nabla^2 F(f^t)$ is positive definite on $\mathcal{H}_d$, GAT selects

$$\mathcal{B}_t^{\mathrm{GAT}} \in \arg\max_{\substack{\mathcal{B}\subseteq\{1,\dots,n\}\\ |\mathcal{B}|=B}} q_t(\mathcal{B}), \quad q_t(\mathcal{B}) = \langle \nabla F_{\mathcal{B}}(f^t), H_t^{-1}\nabla F_{\mathcal{B}}(f^t)\rangle_{\mathcal{H}_d}, \tag{15}$$

and presents this batch to the learner, which uses step size $\eta_A > 0$ to update

$$f^{t+1} = f^t - \eta_A H_t^{-1}\nabla F_{\mathcal{B}_t^{\mathrm{GAT}}}(f^t). \tag{16}$$

The same preconditioner is used for every candidate batch at a given state. The teacher ranks squared gradient norms in that geometry, and the learner moves using its inverse curvature. Because q_t contains interactions between examples, its batch maximization generally differs from top-B singleton selection. The analysis uses exact batch maximization; Appendix G.1 defines exact selection and approximate batch search and describes their use in the experiments.

In the RKHS-orthonormal coordinates above, squared loss has $H = \Lambda$, where we identify an operator with its matrix in the stated orthonormal basis. Equations 15–16 become

$$q_t(\mathcal{B}) = \|Q_{\mathcal{B}} a_t\|_2^2, \qquad a_{t+1} = (I_d - \eta_A Q_{\mathcal{B}_t^{\mathrm{GAT}}})a_t. \tag{17}$$

Compared with Equation 12, preconditioning removes the explicit kernel factor Λ. The selected batch covariance $Q_{\mathcal{B}_t^{\mathrm{GAT}}}$ governs the remaining update and its coupling across spectral components. We next characterize the associated prediction-risk descent guarantee.

4.3 Convergence Comparison and Condition-Number Mitigation

We analyze prediction risk on a fixed pool under the following squared-loss model and teaching protocol.

Assumption 5 (Finite-pool teaching with squared loss)**.** *Let $\mathcal{H}_d$ be a d-dimensional working RKHS. In an RKHS-orthonormal basis, the pool evaluation vectors $\psi_i \in \mathbb{R}^d \setminus \{0\}$ satisfy $f(\boldsymbol{x}_i) = \psi_i^\top c$ and $K_d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \psi_i^\top \psi_j$. Assume $f^0, f^* \in \mathcal{H}_d$, $y_i = \psi_i^\top c_*$, and*

$$F(c) = \frac{1}{2n}\sum_{i=1}^n (\psi_i^\top c - y_i)^2, \qquad H = \frac{1}{n}\sum_{i=1}^n \psi_i\psi_i^\top \succ 0. \tag{18}$$

Write $\alpha = \lambda_{\min}(H)$ and $\beta = \lambda_{\max}(H)$ for the minimum and maximum pool curvatures, $\kappa = \beta/\alpha$ for their condition number, and $\chi = \max_i \|\psi_i\|_2^2 / \min_i \|\psi_i\|_2^2$ for the kernel-diagonal variation.

Here $F(c)$ expresses the pool risk $F(f)$ in RKHS coordinates. Positive definiteness means that the pool identifies every direction in the working space. With coefficient error $e_t = c_t - c_*$, define residuals $r_{t,i} = \psi_i^\top e_t$, risk $E_t = F(c_t) - F(c_*) = \frac{1}{2}e_t^\top H e_t$, and half the squared RKHS distance $V_t = \frac{1}{2}\|e_t\|_2^2$. They obey $\alpha V_t \le E_t \le \beta V_t$. For each size-$B$ subset define

$$H_{\mathcal{B}} = \frac{1}{B}\sum_{i\in\mathcal{B}} \psi_i\psi_i^\top, \quad g_{t,\mathcal{B}} = H_{\mathcal{B}} e_t, \quad Q_{\mathcal{B}} = H^{-1/2} H_{\mathcal{B}} H^{-1/2} \preceq R I_d. \tag{19}$$

The vector $g_{t,\mathcal{B}}$ represents the batch functional gradient. The dimensionless bound $R \ge 1$ limits the largest eigenvalue of every whitened batch covariance, measuring its concentration relative to the pool; $R = n/B$ always works.[1] In any RKHS-orthonormal working basis define $a_t = H^{1/2} e_t$, so $E_t = \frac{1}{2}\|a_t\|_2^2$. Under empirical orthonormality, $H = \Lambda$ and these definitions of a_t and $Q_{\mathcal{B}}$ coincide with the spectral ones above.

Teaching protocol. Each round presents B distinct pool points, with repeats allowed across rounds. GAT exactly maximizes $q_t(\mathcal{B}) = g_{t,\mathcal{B}}^\top H^{-1} g_{t,\mathcal{B}}$ and uses $c_{t+1} = c_t - \eta_A H^{-1} g_{t,\mathcal{B}}$. Greedy selects

[1] Since $H_{\mathcal{B}} = B^{-1}\sum_{i\in\mathcal{B}} \psi_i\psi_i^\top \preceq B^{-1}\sum_{i=1}^n \psi_i\psi_i^\top = (n/B)H$, we have $Q_{\mathcal{B}} \preceq (n/B) I_d$.

the top-B scores $r_{t,i}^2\|\psi_i\|_2^2$; Random draws a fresh uniform size-B subset conditional on the history. Both use $c_{t+1} = c_t - \eta_F g_{t,\mathcal{B}}$. Ties follow a fixed rule.

The teacher knows the target pool values and observes the current predictions, choosing batches for a learner. Teaching complexity counts rounds to a prescribed pool risk. The dimensionless step-size multiplier ν scales the reference steps as $\eta_A = \nu/R$ and $\eta_F = \nu/(\beta R)$; smaller ν gives shorter updates. The analysis below uses a fixed $0 < \nu \leq 1$ and the derived descent coefficient $\omega_\nu = \nu(2-\nu)$.

Lemma 6 (Geometry-Aware Teaching Descent). *Under Assumption 5, let $S_t(\mathcal{B}) = B^{-1}\sum_{i\in\mathcal{B}} r_{t,i}^2$. For any candidate batch, the preconditioned update satisfies*

$$E_t - E_{t+1} = \eta_A S_t(\mathcal{B}) - \tfrac{1}{2}\eta_A^2 q_t(\mathcal{B}), \qquad q_t(\mathcal{B}) \leq R S_t(\mathcal{B}). \tag{20}$$

At $\eta_A = \nu/R$ and $\eta_F = \nu/(\beta R)$, respectively, the preconditioned and first-order updates from this state obey

$$E_t - E_{t+1}^A \geq \frac{\omega_\nu}{2R^2} q_t(\mathcal{B}), \qquad V_t - V_{t+1}^F \geq \frac{\omega_\nu}{2\beta R} S_t(\mathcal{B}). \tag{21}$$

Here superscripts A and F distinguish these two candidate updates.

The first identity separates a batch's useful residual energy from the second-order cost of its update. The uniform covariance bound controls the latter by the former. As a result, maximizing q_t maximizes a risk-decrease lower bound with the same coefficient for every candidate batch. For a first-order learner, the analogous bound controls V_t, explaining why the comparison must account for the conversion from squared RKHS distance to prediction risk.

Lemma 7 (Selection Gains). *Under Assumption 5 and the teaching protocol above, at a state with $E_t > 0$ define*

$$\gamma_t = \frac{q_t(\mathcal{B}_t^{\mathrm{GAT}})}{2E_t}, \qquad \theta_t = \frac{S_t(\mathcal{B}_t^{\mathrm{Greedy}})}{2E_t}. \tag{22}$$

Then $\gamma_t \geq 1$, $\theta_t \geq 1/\chi$, and $\mathbb{E}[S_t(\mathcal{B}_t^{\mathrm{Random}}) \mid c_t] = 2E_t$. Constant kernel diagonal gives $\theta_t \geq 1$. If $E_t = 0$, every subsequent error is zero and the ratios are unnecessary.

The dimensionless gains γ_t and θ_t compare the selected curvature-aware score and residual energy, respectively, with the pool mean squared residual $2E_t$. Greedy weights pointwise residual energy by the kernel diagonal, whose variation is measured by χ. With a constant diagonal, Greedy ranks residuals directly and selects a batch whose mean residual energy is at least the pool average. Random achieves the pool average in conditional expectation.

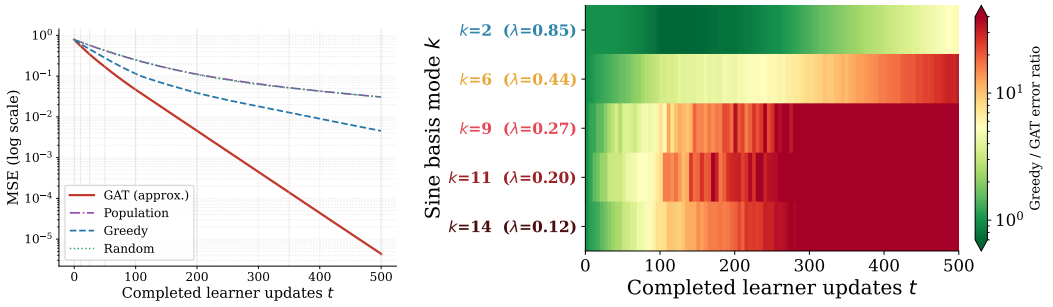
Theorem 8 (Convergence of GAT, Greedy, and Random). *Under Assumption 5 and the teaching protocol above, use $\eta_A = \nu/R$ for GAT and $\eta_F = \nu/(\beta R)$ for Greedy and Random. From a common initialization, for every integer $t \geq 0$,*

$$\begin{aligned} E_t^{\mathrm{GAT}} &\leq (1 - \omega_\nu/R^2)^t E_0, \\ E_t^{\mathrm{Greedy}} &\leq \kappa\big(1 - \omega_\nu/(\chi\kappa R)\big)^t E_0, \\ \mathbb{E}E_t^{\mathrm{Random}} &\leq \kappa\big(1 - \omega_\nu/(\kappa R)\big)^t E_0. \end{aligned} \tag{23}$$

More precisely, the one-round contraction upper bounds are $1 - \omega_\nu\gamma_t/R^2$ for GAT's E_t and $1 - \omega_\nu\theta_t/(\kappa R)$ for Greedy's V_t, with gains evaluated on their respective positive-error states.

Theorem 8 separates kernel conditioning from teaching geometry. First-order bounds retain κ in their factors and in $\alpha V_t \leq E_t \leq \beta V_t$. GAT instead depends on R: when $\kappa > R$, its displayed geometric factor is smaller than both worst-case first-order factors. For constant kernel diagonal, $\theta_t \geq 1$ makes Greedy's decrease certificate in V_t at least as large as Random's conditional-expected certificate from the same state. Appendix B.1 complements these bounds with an exact multi-step risk comparison on a spectrally separated problem family.

For the same exact preconditioner, Random satisfies $\mathbb{E}E_t^{\mathrm{Random}+H^{-1}} \leq (1-\omega_\nu/R)^t E_0$, and residual top-$B$ selection admits the same geometric factor $1 - \omega_\nu/R$ (Appendix A.4). To distinguish the effects of preconditioning and example selection, we next study spectral invariance and teaching optimality with the learner held fixed.



(a) Overall MSE (log scale). (b) Greedy / GAT per-mode error ratio.

Figure 1: Curve teaching with common step-size multiplier $\nu = 0.1$. GAT uses approximate batch search and exact empirical-Hessian preconditioning; Greedy and Random use first-order updates. (a) Pool MSE; Population uses the full pool, and the Random trajectory uses the first prespecified replicate. (b) Greedy/GAT coefficient-error ratios, with values above one indicating lower GAT error.

Corollary 9 (Invariance to Kernel Eigenvalue Rescaling)**.** *Under Assumption 5 and the teaching protocol above, use* $\eta_A = \nu/R$. *Let* $\psi_i = \Lambda^{1/2}\boldsymbol{\phi}_i$, *where* $\boldsymbol{\phi}_i = (\phi_1(\boldsymbol{x}_i), \ldots, \phi_d(\boldsymbol{x}_i))^\top$ *and* $\Phi^\top\Phi/n = I_d$. *Fix* Φ, B, R, ν, *target pool values, initial prediction-error coordinates* a_0, *and an eigenvalue-independent tie-breaking rule. Changing only the positive entries of* Λ *leaves GAT's scores, batches, and prediction trajectory unchanged. Define* $T_{\mathrm{GAT}}(\varepsilon) = \min\{t \in \mathbb{Z}_{\geq 0} : E_t^{\mathrm{GAT}} \leq \varepsilon\}$. *This teaching complexity satisfies*

$$T_{\mathrm{GAT}}(\varepsilon) \leq \left\lceil \frac{R^2}{\omega_\nu} \log \frac{E_0}{\varepsilon} \right\rceil, \qquad 0 < \varepsilon < E_0. \tag{24}$$

Under the fixed-pool comparison in Corollary 9, Equation 17 makes the invariance explicit: exact preconditioning expresses both the score and the update through $Q_{\mathcal{B}}$ and a_t. Batch size and basis concentration enter the teaching bound through R.

Proposition 10 (Finite-Horizon Optimality for a Fixed Learner)**.** *Under Assumption 5 and the teaching protocol above, suppose* $d = n$. *Fix the preconditioned learner and step size* $\eta_A = \nu B/n$. *For every integer budget* $T \geq 0$, *GAT minimizes* E_T *over all legal sequences of size-*B *teaching batches. In particular,*

$$\begin{aligned} E_T^{\mathrm{GAT}} &\leq E_T^{\mathrm{Greedy}+H^{-1}}, \\ E_T^{\mathrm{GAT}} &\leq \mathbb{E}E_T^{\mathrm{Random}+H^{-1}} = (1 - \omega_\nu B/n)^T E_0. \end{aligned} \tag{25}$$

Here both comparison policies use the same learner, and Greedy retains its original pointwise score. At $\nu = 1$, *GAT reaches zero pool risk within* $\lceil n/B \rceil$ *rounds.*

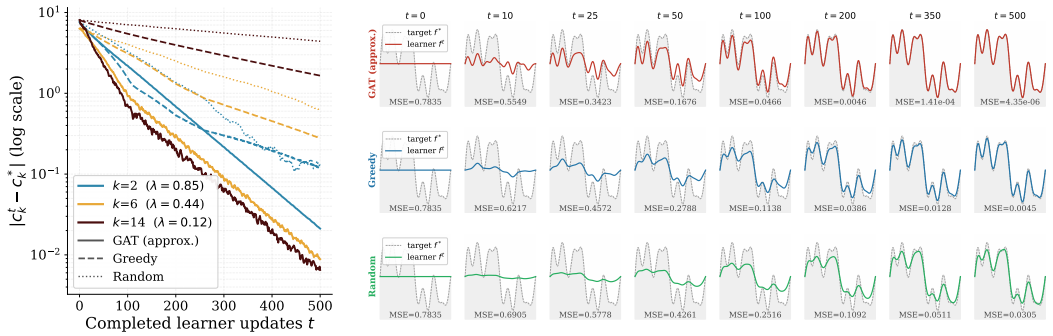
In a complete pool space, exact preconditioning makes the effect of each example independent of the other pool points. GAT allocates batches to maximize the resulting risk reduction. A counting and exchange argument establishes optimality over the full teaching budget (the proof is in Appendix A.6). The comparison includes Greedy and Random teachers under the same learner, isolating the contribution of example selection.

Appendix B.1 compares the original three strategies on a specified problem family, while Appendix B.2 states the additional gradient-accuracy and unit-step requirements for a local quadratic Newton regime.

5 Experiments and Results

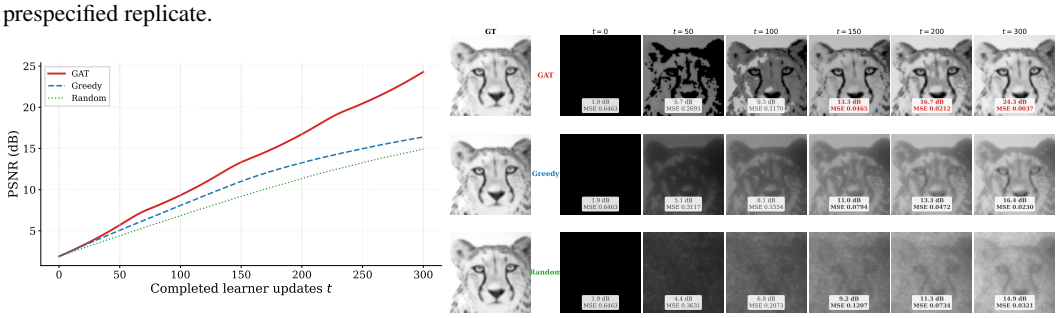
We empirically validate GAT on function and image teaching, comparing it with Greedy, which selects the top-B examples by RKHS gradient norm (Zhang et al., 2023b;a), and Random, which samples uniformly. For curve teaching, Population presents all n examples at each update. The minibatch methods use the same batch size B and dimensionless step-size multiplier ν within each curve or complete-basis image reconstruction task. Appendix C specifies the full protocol.

Curve Teaching. The curve task uses $n = 100$ points, $d = 20$ sine modes, $B = 10$, and $T = 500$, with a realizable target constructed from the mixed-frequency signal $y_{\mathrm{raw}}(x) = \sin(2\pi x) +$



(a) Per-mode coefficient error. (b) Reconstruction snapshots.

Figure 2: Curve learning trajectories at $\nu = 0.1$. (a) Absolute RKHS coefficient errors for modes $k \in \{2, 6, 14\}$. (b) Reconstructions after $t \in \{0, 10, 25, 50, 100, 200, 350, 500\}$ updates. Gray shading and dashed curves show the target; colored curves show the learners. GAT uses approximate batch search, and Random uses the first prespecified replicate.



(a) PSNR convergence. (b) Reconstruction snapshots with PSNR.

Figure 3: Exact GAT versus first-order Greedy and Random on cheetah (64×64, $B = 50$). (a) PSNR trajectories, with the first prespecified Random replicate. (b) Reconstruction after $t \in \{0, 50, 100, 150, 200, 300\}$ completed updates, with ground truth. Random snapshots use the same replicate; aggregate measurements use ten seeds.

$0.6\sin(6\pi x) + 0.4\sin(14\pi x) + 0.3\cos(10\pi x)$ and Hessian condition number $\kappa_H = 21.9345$. GAT selects batches using the curvature-aware score in Equation 15.

At $t = 50$, GAT, Greedy, and Random have MSE 0.16757, 0.27885, and 0.42592. By $t = 500$, these decrease to 4.35×10^{-6}, 4.53×10^{-3}, and 0.03080 ± 0.00082. Figures 1–2 show the reconstruction and coefficient trajectories; Table 3 reports intermediate budgets.

Cheetah Image Teaching. The 64×64 cheetah task uses a complete midpoint cosine basis, with frequency indices j, k and kernel weights $\lambda_{j,k} = (j^2 + k^2 + 1)^{-0.4}$, batch size $B = 50$, and $T = 300$ updates. GAT maximizes the curvature-aware batch score using the full-pool Hessian. At $t = 300$, PSNR reaches 24.30 dB for GAT, 16.38 dB for Greedy, and 14.94 dB for Random (ten-seed mean). Figure 3 shows the reconstructions; Appendix E.2 evaluates cameraman. GAT recovers image structure earlier, matching its faster PSNR improvement.

Selection and update geometry. We cross Random, Greedy, and GAT selection with first-order and preconditioned updates to assess the contribution of example selection. Appendix F reports these controls, while Appendix G evaluates batch-search accuracy and computational cost.

6 Concluding Remarks and Future Work

We develop a geometric framework for nonparametric teaching and introduce GAT, combining curvature-aware batch selection with preconditioned functional updates. Our spectral analysis explains spectral imbalance in first-order teaching. For realizable squared loss on finite pools, we establish risk-decrease and convergence guarantees, with spectral-rescaling invariance under exact selection and preconditioning. In complete pool spaces, GAT is finite-horizon optimal for a fixed preconditioned learner. Experiments show faster convergence than first-order teaching baselines at matched teaching budgets. Future work will study scalable batch optimization, approximate Hessians, broader losses, and neural function learners.

AI USE STATEMENT

We used generative AI tools to improve the clarity and presentation of mathematical proofs and to assist with the interpretation of experimental comparisons. These tools also assisted with manuscript organization and language editing. The authors take full responsibility for all content.

ETHICS STATEMENT

This work studies the approximation of fixed target functions using synthetic signals, pre-existing reconstruction images, and previously released BSDS500 photographs and annotations. Existing contour annotations serve as supervised targets for finite-pool optimization. Applications involving personal or sensitive data would require assessment of privacy, representativeness, and the consequences of adaptive sample selection.

REPRODUCIBILITY STATEMENT

Section 4 specifies the teaching framework and assumptions; Appendix A provides the proofs. Appendix C gives the evaluation protocol, update rules, batch search, and reconstruction settings. Appendices D and E report curve and image reconstruction. Appendix F specifies the controlled comparisons and natural-image tasks, including data splits, parameter selection, and aggregation. Appendix G evaluates batch-search accuracy and computational cost.

REFERENCES

Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 25

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950. 1, 3, 4

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009. 2

Neil K Chada, Quanjun Lang, Fei Lu, and Xiong Wang. A data-adaptive rkhs prior for bayesian learning of kernels in operators. *Journal of Machine Learning Research*, 25(317):1–37, 2024. 1, 3

Isaak Chagouel, Michael Stessin, and Kehe Zhu. Geometric spectral theory for compact operators. *Transactions of the American Mathematical Society*, 368(3):1559–1582, 2016. 1, 3

Rodney Coleman. *Calculus on normed vector spaces*. Springer Science & Business Media, 2012. 4

Gregory W Corder and Dale I Foreman. *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons, 2014. 1

E Brian Davies. *Linear operators and their spectra*, volume 106. Cambridge University Press, 2007. 1

Geng Di, Yu Qingyu, and Qu Changzheng. The eigenvalue problem for hessian operators. *Nonlinear Analysis: Theory, Methods & Applications*, 25(1):27–40, 1995. 2, 3

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001. 1

I. M. Gelfand and S. V. Fomin. *Calculus of variations*. Courier Corporation, 2000. 4

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *ICML*, 2018. 1, 3

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. A tutorial review of rkhs methods in machine learning. *Technical Report*, 2005. 1, 3

Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*. John Wiley & Sons, 2013. 1

Peter D Lax. *Functional analysis*, volume 55. John Wiley & Sons, 2002. 3

Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative machine teaching. In *ICML*, 2017. 2

Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James Rehg, and Le Song. Towards black-box iterative machine teaching. In *ICML*, 2018. 2

Vladimír Müler. On the regular spectrum. *Journal of Operator Theory*, pp. 363–380, 1994. 1

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006. 2, 3

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. the MIT Press, 2001. 1, 3

Burr Settles. Active learning literature survey. 2009. 2

Ivan Singer. *The theory of best approximation and functional analysis*. SIAM, 1974. 1

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35(3):363–417, 2012. 4

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael Mahoney. Adahessian: An adaptive second order optimizer for machine learning. In *AAAI*, 2021. 2, 3

Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, and James Kwok. Nonparametric teaching for multiple learners. In *NeurIPS*, 2023a. 1, 2, 3, 4, 8

Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, and James Kwok. Nonparametric iterative machine teaching. In *ICML*, 2023b. 1, 2, 3, 4, 8, 12

Chen Zhang, Steven Tin Sui Luo, Jason Chun Lok Li, Yik-Chung Wu, and Ngai Wong. Nonparametric teaching of implicit neural representations. In *ICML*, 2024. 1, 3

Chen Zhang, Weixin Bu, Zeyi Ren, Zhengwu Liu, Yik-Chung Wu, and Ngai Wong. Nonparametric teaching for graph property learners. In *ICML*, 2025. 1, 3

Chen Zhang, Jianghui Wang, Bingyang Cheng, Zhongtao Chen, Wendong Xu, Cong Wang, Marco Canini, Francesco Orabona, Yik-Chung Wu, and Ngai Wong. Nonparametric teaching for attention learners. In *ICLR*, 2026a. 1, 3, 21

Chen Zhang, Wei Zuo, Bingyang Cheng, Yikun Wang, Wei-Bin Kou, Yik-Chung Wu, and Ngai Wong. NTK-guided implicit neural teaching. In *CVPR*, 2026b. 3

Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, 2015. 2

Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018. 2

R Zoppoli, M Sanguineti, and Thomas Parisini. Approximating networks and extended ritz method for the solution of functional optimization problems. *Journal of Optimization Theory and Applications*, 112(2):403–440, 2002. 1

Appendix

A PROOFS

Lemma 11 (Chain rule for functionals (Zhang et al., 2023b)). *Let $J : \mathcal{H} \to \mathbb{R}$ be Fréchet differentiable and let $h : \mathbb{R} \to \mathbb{R}$ be differentiable. Then*

$$\nabla(h \circ J)(f) = h'(J(f))\nabla J(f).$$

Lemma 12 (Gradient of the evaluation functional (Zhang et al., 2023b)). *For $\boldsymbol{x} \in \mathcal{X}$, define the evaluation functional $\mathcal{E}_{\boldsymbol{x}}(f) = f(\boldsymbol{x})$. The reproducing property gives $\nabla\mathcal{E}_{\boldsymbol{x}}(f) = K_{\boldsymbol{x}}$, where $K_{\boldsymbol{x}} = K(\boldsymbol{x}, \cdot)$ is the kernel section in $\mathcal{H}$.*

A.1 PROOF OF PROPOSITION 4

Proof. Write $\ell_{\boldsymbol{x}}(u) = \mathcal{L}(u, f^*(\boldsymbol{x}))$. On a convex neighborhood U of a point $\bar{f}$ with $|F(\bar{f})| < \infty$, the assumed domination gives integrable functions M_1, M_2 such that, for $f \in U$,

$$|\ell'_{\boldsymbol{x}}(f(\boldsymbol{x}))|\|K_{\boldsymbol{x}}\|_{\mathcal{H}} \le M_1(\boldsymbol{x}), \qquad |\ell''_{\boldsymbol{x}}(f(\boldsymbol{x}))|\|K_{\boldsymbol{x}}\|^2_{\mathcal{H}} \le M_2(\boldsymbol{x}).$$

All expectations below are over $\boldsymbol{x} \sim \mu$. Since $h(\boldsymbol{x}) = \langle h, K_{\boldsymbol{x}}\rangle_{\mathcal{H}}$, the scalar integral remainder yields

$$\mathbb{E}|\ell_{\boldsymbol{x}}(f(\boldsymbol{x})) - \ell_{\boldsymbol{x}}(\bar{f}(\boldsymbol{x}))| \le \|f - \bar{f}\|_{\mathcal{H}}\mathbb{E}M_1 < \infty,$$
$$\big|F(f+h) - F(f) - \langle G(f), h\rangle_{\mathcal{H}}\big| \le \frac{1}{2}\|h\|^2_{\mathcal{H}}\mathbb{E}M_2,$$

where $G(f) := \mathbb{E}[\ell'_{\boldsymbol{x}}(f(\boldsymbol{x}))K_{\boldsymbol{x}}]$ and the segment from f to $f + h$ lies in U. Hence $DF(f)[h] = \langle G(f), h\rangle_{\mathcal{H}}$. Define

$$A(f) := \mathbb{E}[\ell''_{\boldsymbol{x}}(f(\boldsymbol{x}))(K_{\boldsymbol{x}} \otimes K_{\boldsymbol{x}})].$$

The same integral expansion gives

$$\frac{\|G(f+h) - G(f) - A(f)h\|_{\mathcal{H}}}{\|h\|_{\mathcal{H}}}$$
$$\le \mathbb{E}\Big[\sup_{0\le s\le 1} |\ell''_{\boldsymbol{x}}((f+sh)(\boldsymbol{x})) - \ell''_{\boldsymbol{x}}(f(\boldsymbol{x}))|\|K_{\boldsymbol{x}}\|^2_{\mathcal{H}}\Big] \longrightarrow 0,$$

by continuity of $\ell''_{\boldsymbol{x}}$ and domination by $2M_2$. Likewise,

$$\|A(f+h) - A(f)\|_{\mathrm{op}} \le \mathbb{E}\big[|\ell''_{\boldsymbol{x}}((f+h)(\boldsymbol{x})) - \ell''_{\boldsymbol{x}}(f(\boldsymbol{x}))|\|K_{\boldsymbol{x}}\|^2_{\mathcal{H}}\big] \longrightarrow 0.$$

Consequently, F is twice continuously Fréchet differentiable on U, with

$$D^2F(f)[h, v] = \langle v, A(f)h\rangle_{\mathcal{H}}$$
$$= \mathbb{E}[\ell''_{\boldsymbol{x}}(f(\boldsymbol{x}))\langle h, K_{\boldsymbol{x}}\rangle_{\mathcal{H}}\langle v, K_{\boldsymbol{x}}\rangle_{\mathcal{H}}],$$
$$A(f)^* = A(f), \qquad \|A(f)\|_{\mathrm{op}} \le \mathbb{E}M_2, \qquad \langle h, A(f)h\rangle_{\mathcal{H}} = \mathbb{E}[\ell''_{\boldsymbol{x}}(f(\boldsymbol{x}))h(\boldsymbol{x})^2] \ge 0.$$

The last inequality follows from convexity. This proves Equation 14. Restriction to $\mathcal{H}_d$ replaces $K_{\boldsymbol{x}}$ by $K_{d,\boldsymbol{x}}$. In an RKHS-orthonormal working basis, the finite-pool squared loss therefore gives

$$\nabla^2 F\big|_{\mathcal{H}_d} = \frac{1}{n}\sum_{i=1}^{n} K_{d,\boldsymbol{x}_i} \otimes K_{d,\boldsymbol{x}_i} \quad \longleftrightarrow \quad H = \frac{1}{n}\sum_{i=1}^{n} \psi_i\psi_i^\top.$$

■

A.2 PROOF OF LEMMA 6

Proof. Fix t and a candidate batch $\mathcal{B}$. With $a_t = H^{1/2}e_t$ and $Q_{\mathcal{B}} = H^{-1/2}H_{\mathcal{B}}H^{-1/2}$,

$$E_t = \tfrac{1}{2}\|a_t\|_2^2, \qquad S_t(\mathcal{B}) = e_t^\top H_{\mathcal{B}} e_t = a_t^\top Q_{\mathcal{B}} a_t, \qquad q_t(\mathcal{B}) = a_t^\top Q_{\mathcal{B}}^2 a_t. \tag{26}$$

The preconditioned update and its exact risk decrease are

$$a_{t+1} = H^{1/2}(e_t - \eta_A H^{-1} H_{\mathcal{B}} e_t) = (I_d - \eta_A Q_{\mathcal{B}}) a_t,$$
$$E_t - E_{t+1} = \tfrac{1}{2} a_t^\top [I_d - (I_d - \eta_A Q_{\mathcal{B}})^2] a_t = \eta_A S_t(\mathcal{B}) - \tfrac{1}{2}\eta_A^2 q_t(\mathcal{B}).$$

Since $0 \preceq Q_{\mathcal{B}} \preceq R I_d$,

$$Q_{\mathcal{B}}^2 \preceq R Q_{\mathcal{B}} \quad \Longrightarrow \quad q_t(\mathcal{B}) \le R S_t(\mathcal{B}) \quad \Longrightarrow \quad E_t - E_{t+1} \ge \left(\frac{\eta_A}{R} - \frac{\eta_A^2}{2}\right) q_t(\mathcal{B}).$$

For $\eta_A = \nu/R$, $0 < \nu \le 1$, the coefficient equals $\omega_\nu/(2R^2)$, where $\omega_\nu = \nu(2-\nu)$.

For the first-order update, $e_{t+1} = e_t - \eta_F H_{\mathcal{B}} e_t$ and

$$H_{\mathcal{B}} \preceq R H \preceq \beta R I_d \quad \Longrightarrow \quad H_{\mathcal{B}}^2 \preceq \beta R H_{\mathcal{B}},$$
$$V_t - V_{t+1} = \eta_F S_t(\mathcal{B}) - \tfrac{1}{2}\eta_F^2 e_t^\top H_{\mathcal{B}}^2 e_t$$
$$\ge \left(\eta_F - \tfrac{1}{2}\eta_F^2 \beta R\right) S_t(\mathcal{B}) = \frac{\omega_\nu}{2\beta R} S_t(\mathcal{B}), \qquad \eta_F = \frac{\nu}{\beta R}.$$

Finally, the universal choice $R = n/B$ follows from

$$\frac{n}{B} H - H_{\mathcal{B}} = \frac{1}{B} \sum_{i \notin \mathcal{B}} \psi_i \psi_i^\top \succeq 0 \quad \Longrightarrow \quad Q_{\mathcal{B}} \preceq \frac{n}{B} I_d.$$

■

A.3 Proof of Lemma 7

Proof. Fix c_t and draw a uniform size-B reference subset $\mathcal{U}$. Its inclusion probabilities give

$$\mathbb{P}(i \in \mathcal{U}) = \frac{B}{n}, \qquad \mathbb{E} H_{\mathcal{U}} = \frac{1}{B} \sum_{i=1}^n \frac{B}{n} \psi_i \psi_i^\top = H, \qquad \mathbb{E} Q_{\mathcal{U}} = I_d.$$

Exact batch maximization and Jensen's inequality imply

$$q_t(\mathcal{B}_t^{\mathrm{GAT}}) = \max_{|\mathcal{B}|=B} \|Q_{\mathcal{B}} a_t\|_2^2 \ge \mathbb{E}\|Q_{\mathcal{U}} a_t\|_2^2$$
$$\ge \|(\mathbb{E} Q_{\mathcal{U}}) a_t\|_2^2 = \|a_t\|_2^2 = 2E_t.$$

Only the reference subset $\mathcal{U}$ is averaged; no unbiasedness is imposed on GAT's selected gradient.

For Greedy, define $k_i = \|\psi_i\|_2^2 = K_d(\boldsymbol{x}_i, \boldsymbol{x}_i)$, $k_{\min} = \min_i k_i > 0$, and $k_{\max} = \max_i k_i$, so $\chi = k_{\max}/k_{\min}$. The top-B rule yields

$$B S_t(\mathcal{B}_t^{\mathrm{Greedy}}) = \sum_{i \in \mathcal{B}_t^{\mathrm{Greedy}}} r_{t,i}^2 \ge \frac{1}{k_{\max}} \sum_{i \in \mathcal{B}_t^{\mathrm{Greedy}}} k_i r_{t,i}^2$$
$$= \frac{1}{k_{\max}} \max_{|\mathcal{B}|=B} \sum_{i \in \mathcal{B}} k_i r_{t,i}^2 \ge \frac{B}{n k_{\max}} \sum_{i=1}^n k_i r_{t,i}^2$$
$$\ge \frac{B}{n\chi} \sum_{i=1}^n r_{t,i}^2 = \frac{2B}{\chi} E_t.$$

If k_i is constant, $\chi = 1$ and the same chain gives $\theta_t \ge 1$. For Random,

$$\mathbb{E}[S_t(\mathcal{B}_t^{\mathrm{Random}}) \mid c_t] = \frac{1}{B} \sum_{i=1}^n r_{t,i}^2 \mathbb{P}(i \in \mathcal{B}_t^{\mathrm{Random}} \mid c_t) = \frac{1}{n} \sum_{i=1}^n r_{t,i}^2 = 2E_t.$$

Dividing the deterministic bounds by $2E_t$ proves the gain claims for $E_t > 0$. The remaining case satisfies

$$E_t = 0 \implies e_t = 0 \implies g_{t,\mathcal{B}} = 0 \text{ for every } \mathcal{B} \implies E_s = 0 \quad (s \ge t),$$

using $H \succ 0$, so no gain ratio is needed. ■

A.4 Proof of Theorem 8

Proof. At positive-error states, Lemmas 6–7, evaluated on each method's own trajectory, give

$$\begin{aligned} E_{t+1}^{\mathrm{GAT}} &\le E_t^{\mathrm{GAT}} - \frac{\omega_\nu}{2R^2} q_t(\mathcal{B}_t^{\mathrm{GAT}}) = \left(1 - \frac{\omega_\nu \gamma_t}{R^2}\right) E_t^{\mathrm{GAT}} \\ &\le \left(1 - \frac{\omega_\nu}{R^2}\right) E_t^{\mathrm{GAT}}, \\ V_{t+1}^{\mathrm{Greedy}} &\le V_t^{\mathrm{Greedy}} - \frac{\omega_\nu \theta_t}{\beta R} E_t^{\mathrm{Greedy}} \le \left(1 - \frac{\omega_\nu \theta_t}{\kappa R}\right) V_t^{\mathrm{Greedy}} \\ &\le \left(1 - \frac{\omega_\nu}{\chi \kappa R}\right) V_t^{\mathrm{Greedy}}, \end{aligned}$$

where the second chain uses $E_t \ge \alpha V_t$. Let $\mathcal{F}_t$ denote the history before drawing the round-t batch. Fresh uniform sampling conditional on $\mathcal{F}_t$ gives

$$\begin{aligned} \mathbb{E}[S_t(\mathcal{B}_t^{\mathrm{Random}}) \mid \mathcal{F}_t] &= 2E_t^{\mathrm{Random}}, \\ \mathbb{E}[V_{t+1}^{\mathrm{Random}} \mid \mathcal{F}_t] &\le V_t^{\mathrm{Random}} - \frac{\omega_\nu}{\beta R} E_t^{\mathrm{Random}} \le \left(1 - \frac{\omega_\nu}{\kappa R}\right) V_t^{\mathrm{Random}}. \end{aligned}$$

All factors used in these recursions are nonnegative, since

$$q_t(\mathcal{B}) \le R^2 \|a_t\|_2^2, \qquad S_t(\mathcal{B}) \le R\|a_t\|_2^2 \quad \Longrightarrow \quad \gamma_t \le R^2, \quad \theta_t \le R, \qquad 0 < \omega_\nu \le 1, \quad \kappa \ge 1.$$

Iteration and the tower property therefore imply

$$\begin{aligned} E_t^{\mathrm{GAT}} &\le (1 - \omega_\nu / R^2)^t E_0, \\ E_t^{\mathrm{Greedy}} &\le \beta V_t^{\mathrm{Greedy}} \le \beta \left(1 - \frac{\omega_\nu}{\chi \kappa R}\right)^t V_0 \le \kappa \left(1 - \frac{\omega_\nu}{\chi \kappa R}\right)^t E_0, \\ \mathbb{E} E_t^{\mathrm{Random}} &\le \beta \mathbb{E} V_t^{\mathrm{Random}} \le \beta \left(1 - \frac{\omega_\nu}{\kappa R}\right)^t V_0 \le \kappa \left(1 - \frac{\omega_\nu}{\kappa R}\right)^t E_0, \end{aligned}$$

where $V_0 \le E_0/\alpha$. This proves Equation 23, including the first-order prefactor κ. At $t = 0$ use the common initialization and the convention $z^0 = 1$ also for $z = 0$. ■

For Random with the same exact preconditioner, Lemma 6 instead gives, at $\eta_A = \nu/R$,

$$\begin{aligned} E_t - E_{t+1}(\mathcal{B}) &= \eta_A S_t(\mathcal{B}) - \tfrac{1}{2}\eta_A^2 q_t(\mathcal{B}) \\ &\ge \left(\eta_A - \tfrac{1}{2}\eta_A^2 R\right) S_t(\mathcal{B}) = \frac{\omega_\nu}{2R} S_t(\mathcal{B}), \\ \mathbb{E}[E_{t+1}^{\mathrm{Random}+H^{-1}} \mid \mathcal{F}_t] &\le \left(1 - \frac{\omega_\nu}{R}\right) E_t^{\mathrm{Random}+H^{-1}}, \\ \mathbb{E} E_t^{\mathrm{Random}+H^{-1}} &\le \left(1 - \frac{\omega_\nu}{R}\right)^t E_0. \end{aligned}$$

For the residual teacher, let $\mathcal{B}_t^{\mathrm{res}} \in \arg\max_{|\mathcal{B}|=B} S_t(\mathcal{B})$ and use the same preconditioned learner. Then

$$S_t(\mathcal{B}_t^{\mathrm{res}}) \ge \mathbb{E}_{\mathcal{B}} S_t(\mathcal{B}) = 2E_t, \qquad E_{t+1}^{\mathrm{res}} \le \left(1 - \frac{\omega_\nu}{R}\right) E_t^{\mathrm{res}}, \qquad E_t^{\mathrm{res}} \le \left(1 - \frac{\omega_\nu}{R}\right)^t E_0.$$

A.5 Proof of Corollary 9

Proof. Let $\Phi_{\mathcal{B}}$ denote the rows of Φ indexed by $\mathcal{B}$. Empirical orthonormality gives

$$\begin{aligned} H &= \Lambda^{1/2} \frac{\Phi^\top \Phi}{n} \Lambda^{1/2} = \Lambda, \\ Q_{\mathcal{B}} &= \Lambda^{-1/2} \left(\frac{1}{B} \Lambda^{1/2} \Phi_{\mathcal{B}}^\top \Phi_{\mathcal{B}} \Lambda^{1/2}\right) \Lambda^{-1/2} = \frac{1}{B} \Phi_{\mathcal{B}}^\top \Phi_{\mathcal{B}}, \\ q_t(\mathcal{B}) &= \|Q_{\mathcal{B}} a_t\|_2^2, \qquad a_{t+1} = \left(I_d - \frac{\nu}{R} Q_{\mathcal{B}_t}\right) a_t, \qquad \boldsymbol{r}_t = \Phi a_t. \end{aligned}$$

For two positive diagonal spectra Λ and $\widetilde{\Lambda}$, denote the second trajectory by tildes. Fixed Φ, B, R, ν and identical initial a_0 imply inductively

$$a_t = \widetilde{a}_t \implies q_t(\mathcal{B}) = \widetilde{q}_t(\mathcal{B}) \quad \text{for every } \mathcal{B} \implies \mathcal{B}_t = \widetilde{\mathcal{B}}_t \implies a_{t+1} = \widetilde{a}_{t+1},$$

where the middle implication uses the same eigenvalue-independent tie rule. Thus $\boldsymbol{r}_t = \widetilde{\boldsymbol{r}}_t$ for all t, and fixed target values imply identical predictions. The fixed initialization is in prediction-error coordinates:

$$e_0 = \Lambda^{-1/2} a_0, \qquad \widetilde{e}_0 = \widetilde{\Lambda}^{-1/2} a_0;$$

the RKHS coordinate errors need not coincide.

Put $z = \omega_\nu / R^2 \in (0, 1]$. For $0 < z < 1$, Theorem 8 gives

$$E_t \le (1 - z)^t E_0 \le e^{-zt} E_0, \qquad t \ge \frac{1}{z} \log \frac{E_0}{\varepsilon} \implies E_t \le \varepsilon.$$

For $z = 1$, the same theorem gives $E_1 = 0$. Since $0 < \varepsilon < E_0$, both cases yield

$$T_{\mathrm{GAT}}(\varepsilon) \le \left\lceil \frac{R^2}{\omega_\nu} \log \frac{E_0}{\varepsilon} \right\rceil.$$

The threshold is measured in fixed-pool risk; neither RKHS-distance thresholds nor computational costs enter this statement. ■

A.6 Proof of Proposition 10

Proof. Let $\Psi \in \mathbb{R}^{n \times n}$ have rows $\psi_i^\top$. Since $d = n$ and $H = \Psi^\top \Psi / n \succ 0$, the matrix Ψ is invertible and

$$\Psi H^{-1} \Psi^\top = n \Psi (\Psi^\top \Psi)^{-1} \Psi^\top = n I_n. \tag{27}$$

Define $\boldsymbol{r}_t = (r_{t,1}, \ldots, r_{t,n})^\top$ and $P_{\mathcal{B}} = \mathrm{diag}(\mathbf{1}_{\{i \in \mathcal{B}\}})_{i=1}^n$. Then

$$\begin{aligned}
g_{t,\mathcal{B}} &= B^{-1} \Psi^\top P_{\mathcal{B}} \boldsymbol{r}_t, \\
q_t(\mathcal{B}) &= B^{-2} \boldsymbol{r}_t^\top P_{\mathcal{B}} \Psi H^{-1} \Psi^\top P_{\mathcal{B}} \boldsymbol{r}_t = \frac{n}{B^2} \sum_{i \in \mathcal{B}} r_{t,i}^2, \\
\boldsymbol{r}_{t+1} &= \boldsymbol{r}_t - \frac{\eta_A}{B} \Psi H^{-1} \Psi^\top P_{\mathcal{B}} \boldsymbol{r}_t = (I_n - \nu P_{\mathcal{B}}) \boldsymbol{r}_t, \qquad \eta_A = \frac{\nu B}{n}.
\end{aligned}$$

The batch objective is additive in these coordinates. Put $\rho = (1 - \nu)^2 = 1 - \omega_\nu \in [0, 1)$ for the remaining squared-error fraction after a presentation. The exact decrease is

$$E_t - E_{t+1}(\mathcal{B}) = \frac{1 - \rho}{2n} \sum_{i \in \mathcal{B}} r_{t,i}^2 = \frac{\omega_\nu}{2n} \sum_{i \in \mathcal{B}} r_{t,i}^2. \tag{28}$$

For a horizon T, define $w_i = r_{0,i}^2$ and let m_i count the rounds presenting point i. Every legal sequence satisfies

$$E_T = \frac{1}{2n} \sum_{i=1}^n w_i \rho^{m_i}, \qquad m_i \in \{0, \ldots, T\}, \qquad \sum_{i=1}^n m_i = BT, \tag{29}$$

with $\rho^0 = 1$, including $\rho = 0$. These constraints are also sufficient. For $T \ge 1$, set $M_0 = 0$, $M_i = \sum_{j=1}^i m_j$ and define, for $s = 1, \ldots, T$,

$$\mathcal{B}_s := \{i : \text{some integer } \ell \in [M_{i-1}, M_i) \text{ satisfies } \ell \bmod T = s - 1\}.$$

Because $m_i \le T$, each interval contributes at most one integer to a given residue class. Moreover,

$$\bigcup_{i=1}^n \left([M_{i-1}, M_i) \cap \mathbb{Z} \right) = \{0, \ldots, BT - 1\}, \qquad |\mathcal{B}_s| = B, \qquad \sum_{s=1}^T \mathbf{1}_{\{i \in \mathcal{B}_s\}} = m_i.$$

This constructs a legal batch sequence for every feasible count vector.

Define the feasible set and value function

$$\mathcal{M}_T := \{m \in \{0,\dots,T\}^n : \textstyle\sum_i m_i = BT\}, \qquad J_T(w) := \min_{m\in\mathcal{M}_T} \frac{1}{2n}\sum_i w_i\rho^{m_i}.$$

We prove by induction that repeated top-B selection attains $J_T(w)$ for every nonnegative vector w. The case $T = 0$ is $J_0(w) = (2n)^{-1}\sum_i w_i$. For $T \geq 1$, relabel $w_1 \geq \cdots \geq w_n$ with the prescribed tie rule. An optimal count vector may be chosen with $m_1 \geq \cdots \geq m_n$: exchanging any inverted pair $i < j$, $m_i < m_j$, changes its objective by

$$\frac{1}{2n}\left(w_i\rho^{m_j} + w_j\rho^{m_i} - w_i\rho^{m_i} - w_j\rho^{m_j}\right) = \frac{1}{2n}(w_i - w_j)(\rho^{m_j} - \rho^{m_i}) \leq 0.$$

For such an optimum, the total-count constraint implies

$$m_B = 0 \implies BT = \sum_{i=1}^{n} m_i = \sum_{i=1}^{B-1} m_i \leq (B-1)T < BT,$$

$$m_{B+1} = T \quad (B < n) \implies BT = \sum_{i=1}^{n} m_i \geq \sum_{i=1}^{B+1} m_i = (B+1)T > BT.$$

Therefore $m_B \geq 1$ and, for $B < n$, $m_{B+1} \leq T - 1$; when $B = n$, necessarily $m_i = T$ for every i. Write

$$b_i = \mathbf{1}_{\{i\leq B\}}, \qquad w_i' = w_i\rho^{b_i}, \qquad m_i' = m_i - b_i.$$

The preceding inequalities and the sum constraint give

$$0 \leq m_i' \leq T-1, \qquad \sum_i m_i' = B(T-1), \qquad m' \in \mathcal{M}_{T-1}, \qquad \sum_i w_i'\rho^{m_i'} = \sum_i w_i\rho^{m_i}.$$

Conversely, every $\ell \in \mathcal{M}_{T-1}$ satisfies $\ell + b \in \mathcal{M}_T$. Hence both directions of the optimal-value comparison hold:

$$J_{T-1}(w') \leq \frac{1}{2n}\sum_i w_i'\rho^{m_i'} = J_T(w),$$

$$J_T(w) \leq \min_{\ell\in\mathcal{M}_{T-1}} \frac{1}{2n}\sum_i w_i\rho^{\ell_i+b_i} = J_{T-1}(w').$$

Thus $J_T(w) = J_{T-1}(w')$. The first top-B update produces w', and the induction hypothesis makes its top-B continuation attain $J_{T-1}(w') = J_T(w)$. This proves finite-horizon optimality without repeating a point within a batch.

Every realization of a competing adaptive or randomized teacher is a legal sequence. Holding the learner and step fixed therefore gives

$$E_T^{\mathrm{GAT}} = J_T(w) \leq E_T^{\pi}$$

for every such realization π, including Greedy+H^{-1} and Random+H^{-1}. For the latter, uniform inclusion and Equation 28 yield

$$\mathbb{E}[E_{t+1}^{\mathrm{Random}+H^{-1}} \mid \mathcal{F}_t] = E_t^{\mathrm{Random}+H^{-1}} - \frac{\omega_\nu}{2n}\sum_i r_{t,i}^2\frac{B}{n}$$

$$= \left(1 - \omega_\nu\frac{B}{n}\right)E_t^{\mathrm{Random}+H^{-1}},$$

$$E_T^{\mathrm{GAT}} \leq \mathbb{E}E_T^{\mathrm{Random}+H^{-1}} = \left(1 - \omega_\nu\frac{B}{n}\right)^T E_0.$$

Equality in the Greedy comparison holds when the kernel diagonal is constant and the policies share a tie-breaking rule. At $\nu = 1$, the residuals presented in a round vanish. If N_t is the number of nonzero residuals under GAT, then

$$N_{t+1} = \max\{N_t - B, 0\}, \qquad N_t \leq \max\{n - tB, 0\}, \qquad E_{\lceil n/B\rceil} = 0.$$

The allocation proof uses the complete-pool identity and the common fixed factor ρ. ■

B Additional Discussion

B.1 Multi-Step Comparison on a Spectrally Separated Teaching Pool

Proposition 10 holds the learner fixed to isolate the teacher's contribution. Here we compare the original GAT, Greedy, and Random teacher–learner pairs along their respective trajectories. An explicit realizable family separates the effects of example selection and update geometry.

Take $n = d \geq 2$, and let u_i be the ith standard unit vector in $\mathbb{R}^n$. Choose positive eigenvalues with $\lambda_1 = \min_i \lambda_i = \kappa^{-1}$ and $\max_i \lambda_i = 1$, and set

$$\psi_i = \sqrt{n\lambda_i}\, u_i, \qquad K_d(\boldsymbol{x}_i, \boldsymbol{x}_j) = n\lambda_i \mathbf{1}_{\{i=j\}}, \qquad H = \operatorname{diag}(\lambda_1, \ldots, \lambda_n). \tag{30}$$

Here $\mathbf{1}_{\{i=j\}}$ equals one when $i = j$ and zero otherwise. Choose a common initialization with $r_{0,1} \neq 0$ and $r_{0,i} = 0$ for $i > 1$. This is realizable because the feature matrix with rows $\psi_i^\top$ is invertible. With $R = n/B$ and $\beta = 1$, the step protocol of Theorem 8 gives $\eta_A = \eta_F = \nu B/n$, where $0 < \nu \leq 1$.

For any selected batch $\mathcal{B}_t$, the diagonal geometry gives the exact residual updates

$$\begin{aligned} r^{\mathrm{GAT}}_{t+1,i} &= \big(1 - \nu \mathbf{1}_{\{i \in \mathcal{B}_t\}}\big) r^{\mathrm{GAT}}_{t,i}, \\ r^{s}_{t+1,i} &= \big(1 - \nu\lambda_i \mathbf{1}_{\{i \in \mathcal{B}_t\}}\big) r^{s}_{t,i}, \qquad s \in \{\mathrm{Greedy}, \mathrm{Random}\}. \end{aligned}$$

Thus all residuals except the first remain zero. While the first residual is nonzero, GAT and Greedy both select it: their objectives reduce, respectively, to

$$q_t(\mathcal{B}) = \frac{n}{B^2} r_{t,1}^2 \mathbf{1}_{\{1 \in \mathcal{B}\}}, \qquad r_{t,i}^2 \|\psi_i\|_2^2 = n\lambda_1 r_{t,1}^2 \mathbf{1}_{\{i=1\}}.$$

Random includes it with conditional probability B/n. Since $E_t = r_{t,1}^2/(2n)$ on each trajectory,

$$\begin{aligned} E^{\mathrm{GAT}}_{t+1} &= (1-\nu)^2 E^{\mathrm{GAT}}_t, \\ E^{\mathrm{Greedy}}_{t+1} &= (1-\nu/\kappa)^2 E^{\mathrm{Greedy}}_t, \\ \mathbb{E}[E^{\mathrm{Random}}_{t+1} \mid r^{\mathrm{Random}}_{t,1}] &= \left[1 - \frac{B}{n} + \frac{B}{n}(1-\nu/\kappa)^2\right] E^{\mathrm{Random}}_t. \end{aligned}$$

Iteration yields, for every integer $t \geq 0$,

$$\begin{aligned} E^{\mathrm{GAT}}_t &= (1-\nu)^{2t} E_0, \\ E^{\mathrm{Greedy}}_t &= (1-\nu/\kappa)^{2t} E_0, \\ \mathbb{E}E^{\mathrm{Random}}_t &= \left[1 - \frac{B}{n}\left(\frac{2\nu}{\kappa} - \frac{\nu^2}{\kappa^2}\right)\right]^t E_0. \end{aligned} \tag{31}$$

The factors satisfy

$$(1-\nu)^2 \leq (1-\nu/\kappa)^2 \leq 1 - \frac{B}{n} + \frac{B}{n}(1-\nu/\kappa)^2,$$

so the actual multi-step risks obey

$$E^{\mathrm{GAT}}_t \leq E^{\mathrm{Greedy}}_t \leq \mathbb{E}E^{\mathrm{Random}}_t. \tag{32}$$

Both inequalities are strict for $t \geq 1$ if $\kappa > 1$ and $B < n$. GAT and Greedy present the same informative point, so their difference comes from the update geometry. Greedy and Random share the first-order learner, so their difference comes from the frequency of presenting that point.

For deterministic s, let $T_s(\varepsilon) = \min\{t : E^s_t \leq \varepsilon\}$, and let $T_{\mathrm{Random}}(\varepsilon) = \min\{t : \mathbb{E}E^{\mathrm{Random}}_t \leq \varepsilon\}$. The latter measures an expected-risk threshold. For fixed $0 < \nu < 1$ and $0 < \varepsilon \leq E_0/\mathrm{e}$,

$$\begin{aligned} T_{\mathrm{GAT}}(\varepsilon) &= \Theta_\nu\left(\log \frac{E_0}{\varepsilon}\right), \\ T_{\mathrm{Greedy}}(\varepsilon) &= \Theta_\nu\left(\kappa \log \frac{E_0}{\varepsilon}\right), \\ T_{\mathrm{Random}}(\varepsilon) &= \Theta_\nu\left(\frac{\kappa n}{B} \log \frac{E_0}{\varepsilon}\right). \end{aligned} \tag{33}$$

Here Θ_ν permits constants depending only on ν. These orders follow from Equation 31 and

$$z \le -\log(1-z) \le \frac{z}{1-z} \quad (0 < z < 1),$$
$$\frac{\omega_\nu B}{n\kappa} \le \frac{B}{n}\left(\frac{2\nu}{\kappa} - \frac{\nu^2}{\kappa^2}\right) \le \frac{2\nu B}{n\kappa}, \qquad (1-\nu/\kappa)^2 \ge (1-\nu)^2 > 0.$$

At $\nu = 1$, GAT reaches zero risk in one round. The resulting separation characterizes this family under the stated fixed-step protocol.

B.2 Batch-Gradient Accuracy and a Local Newton Regime

For a smooth pool risk, second-order batch-gradient accuracy and unit Newton steps yield a local quadratic regime under the following conditions.

Proposition 13 (Local Quadratic Convergence in a Newton Regime)**.** *Let F be a twice continuously differentiable pool risk on $\mathcal{H}_d$, with $\nabla F(f^*) = 0$. On a convex neighborhood of f^*, assume*

$$\alpha I_{\mathcal{H}_d} \preceq \nabla^2 F(f) \preceq \beta I_{\mathcal{H}_d}, \qquad \|\nabla^2 F(f) - \nabla^2 F(\widetilde{f})\|_{\mathrm{op}} \le L_H \|f - \widetilde{f}\|_{\mathcal{H}_d},$$

where $\alpha, \beta > 0$ are local curvature bounds, $L_H \ge 0$ bounds Hessian variation, and $I_{\mathcal{H}_d}$ is the identity operator. Suppose the selected batches satisfy, throughout this neighborhood,

$$\|\nabla F_{\mathcal{B}_t}(f^t) - \nabla F(f^t)\|_{\mathcal{H}_d} \le C_g \|f^t - f^*\|^2_{\mathcal{H}_d} \tag{34}$$

where $C_g \ge 0$ controls batch-gradient approximation error. With the exact Hessian and unit step $\eta_A = 1$, sufficiently close initialization gives

$$\|f^{t+1} - f^*\|_{\mathcal{H}_d} \le C_N \|f^t - f^*\|^2_{\mathcal{H}_d}, \qquad C_N = \frac{L_H/2 + C_g}{\alpha}. \tag{35}$$

Moreover, $E_{t+1} \le (2\beta C_N^2/\alpha^2) E_t^2$, where $E_t = F(f^t) - F(f^*)$.

Proof. Set $h_t = f^t - f^*$, $H_t = \nabla^2 F(f^t)$, and $\Delta_t = \nabla F_{\mathcal{B}_t}(f^t) - \nabla F(f^t)$. All three act or lie in $\mathcal{H}_d$ as appropriate. Stationarity and the fundamental theorem of calculus give

$$\nabla F(f^t) = \int_0^1 \nabla^2 F(f^* + s h_t) h_t \, ds,$$
$$H_t h_t - \nabla F(f^t) = \int_0^1 [H_t - \nabla^2 F(f^* + s h_t)] h_t \, ds,$$
$$\|H_t h_t - \nabla F(f^t)\|_{\mathcal{H}_d} \le \int_0^1 L_H (1-s) \|h_t\|^2_{\mathcal{H}_d} \, ds = \frac{L_H}{2} \|h_t\|^2_{\mathcal{H}_d}.$$

For a general positive step η_t, adding and subtracting the full-gradient Newton direction yields

$$h_{t+1} = (1-\eta_t) h_t + \eta_t H_t^{-1} [H_t h_t - \nabla F(f^t)] - \eta_t H_t^{-1} \Delta_t.$$

Since $\|H_t^{-1}\|_{\mathrm{op}} \le \alpha^{-1}$,

$$\|h_{t+1}\|_{\mathcal{H}_d} \le |1-\eta_t| \|h_t\|_{\mathcal{H}_d} + \frac{\eta_t L_H}{2\alpha} \|h_t\|^2_{\mathcal{H}_d} + \frac{\eta_t}{\alpha} \|\Delta_t\|_{\mathcal{H}_d}. \tag{36}$$

Setting $\eta_t = 1$ and applying Equation 34 proves Equation 35. Choose a closed ball of radius $r_* > 0$ inside the neighborhood with $C_N r_* < 1$. For $\|h_0\|_{\mathcal{H}_d} \le r_*$, induction gives

$$\|h_{t+1}\|_{\mathcal{H}_d} \le C_N \|h_t\|^2_{\mathcal{H}_d} \le (C_N r_*) \|h_t\|_{\mathcal{H}_d} \le r_*.$$

Thus the iterates remain in the neighborhood and converge quadratically. The curvature bounds further imply

$$\frac{\alpha}{2} \|h_t\|^2_{\mathcal{H}_d} \le E_t \le \frac{\beta}{2} \|h_t\|^2_{\mathcal{H}_d}, \qquad E_{t+1} \le \frac{\beta}{2} C_N^2 \|h_t\|^4_{\mathcal{H}_d} \le \frac{2\beta C_N^2}{\alpha^2} E_t^2.$$

■

Full-pool gradients satisfy Equation 34 with $C_g = 0$. For selected batches, the condition requires second-order accuracy relative to the full gradient. In the realizable squared-loss setting of Assumption 5,

$$g_{t,\mathcal{B}} - g_t = (H_{\mathcal{B}} - H) e_t, \qquad g_t = H e_t,$$

which is generally linear in e_t. Unit steps remove the first term in Equation 36; second-order gradient accuracy controls the last. These conditions define the local quadratic regime of Proposition 13.

C Experimental Protocol

C.1 Evaluation and teaching budgets

Learners start at $c_0 = 0$, except for the spectral-rescaling initialization in Appendix F.2. Each batch contains B distinct points, with repetition across rounds. Score-based teachers evaluate the current learner state. For scalar targets, pool MSE is $n^{-1}\|\boldsymbol{r}_t\|_2^2$, evaluated after every completed update using unclipped predictions. The two-channel color metric is defined in Appendix F.4. For unit-range images, PSNR is $-10\log_{10}(\mathrm{MSE})$. Random reconstruction summaries average ten runs, with sample standard deviations where indicated. Reconstruction figures use the first fixed replicate; geometry figures aggregate across targets.

Steps and budgets. We use the batch bound $R = n/B$ and dimensionless step-size multiplier $\nu > 0$, giving $\eta_A = \nu B/n$ and $\eta_F = \nu B/(\beta n)$, where β is the largest eigenvalue of H. The reference choice $\nu = 1$ sets both step scales. A minibatch budget of T rounds presents BT examples, counting repeated presentations.

Numerical implementation. Feature construction, empirical Hessian solves, and updates use double precision and one linear-algebra thread per worker, with an unregularized Hessian. Cosine-basis image tasks use orthonormal discrete cosine transforms; the colorization task uses Gaussian features and singular-value coordinates. Ties follow pool index. NumPy random generators use independently drawn 32-bit seeds fixed before evaluation; The Random cells in each paired ablation share the same batch streams. Errors below 10^{-14} indicate numerical saturation.

C.2 Update rules and batch search

Let $\Psi = \Phi\Lambda^{1/2}$, $\widehat{\mathbf{y}}_t = \Psi c_t$, $\boldsymbol{r}_t = \widehat{\mathbf{y}}_t - \mathbf{y}$, and $\mathbf{K} = \Psi\Psi^\top$. For a batch $\mathcal{B}$,

$$\mathrm{proj}_t = \Phi_{\mathcal{B}}^\top \boldsymbol{r}_{t,\mathcal{B}}/B, \qquad g_{t,\mathcal{B}} = \Psi_{\mathcal{B}}^\top \boldsymbol{r}_{t,\mathcal{B}}/B, \qquad H = \Psi^\top\Psi/n. \tag{37}$$

Scoring and updating use the same pool Hessian: solve $Hz = g_{t,\mathcal{B}}$, evaluate $q_t = g_{t,\mathcal{B}}^\top z$, and update $c_{t+1} = c_t - \eta_A z$. Under empirical orthonormality, $\Phi^\top\Phi/n = I$, the Hessian is $H = \Lambda$.

Table 1: Teaching rules and learner updates. The realizable reconstruction comparisons use the listed updates; the factorial ablation crosses Random, Greedy, and GAT selection with both updates.

Teacher	Batch selection	Pair update Δc_t
GAT	exact or approximate maximization of $q_t(\mathcal{B})$	$-\eta_A H^{-1} g_{t,\mathcal{B}}$
Greedy	top-B of $r_{t,i}^2 K_{ii}$	$-\eta_F g_{t,\mathcal{B}}$
Random	uniform size-B subset	$-\eta_F g_{t,\mathcal{B}}$
Population	all n points (curve reference)	$-\eta_F g_t$

Exact score and approximate batch search. The general score, retaining all cross terms, is

$$q_t(\mathcal{B}) = \frac{1}{B^2}\boldsymbol{r}_{t,\mathcal{B}}^\top \Psi_{\mathcal{B}} H^{-1} \Psi_{\mathcal{B}}^\top \boldsymbol{r}_{t,\mathcal{B}}. \tag{38}$$

For $B = 1$, we evaluate all singleton scores; the small-pool control enumerates all batches. For complete pool bases, the identity in Equation 27 makes the score additive, allowing exact maximization by sorting its per-point contributions. The main curve experiment, contour teaching, and colorization use approximate batch search. Colorization sums the scores of its two output channels and replaces r_i^2 by the squared two-channel residual norm in the initialization rules. Initial batches come from four top-B rules—r_i^2, $r_i^2 K_{ii}$, r_i^2/K_{ii}, and $r_i^2 \psi_i^\top H^{-1}\psi_i$—and a forward construction maximizing the score at each addition. From each distinct batch, we evaluate every one-point exchange and accept the largest improvement, stopping when the relative score improvement is at most 10^{-13} or after ten exchanges for curves and four for contour fields and colorization. We return the highest-scoring final batch.

Batch dynamics. For realizable targets in arbitrary empirical coordinates, $a_t = H^{1/2}(c_t - c_*)$ and $Q_{\mathcal{B}} = H^{-1/2} H_{\mathcal{B}} H^{-1/2}$ give

$$a_{t+1}^A = (I - \eta_A Q_{\mathcal{B}_t}) a_t, \qquad a_{t+1}^F = (I - \eta_F H Q_{\mathcal{B}_t}) a_t.$$

The off-diagonal entries of $Q_{\mathcal{B}_t}$ couple the whitened error coordinates as the selected batch changes.

C.3 Reconstruction settings

Table 2: Curve and complete-basis image settings. Here n is pool size, d is basis rank, B is batch size, and T is the number of updates. The empirical Hessian is $H = \Psi^\top \Psi/n$, with extreme eigenvalues α, β and condition number $\kappa_H = \beta/\alpha$.

	Curve	Images (cheetah and cameraman)
Pool and rank	$n = 100, d = 20$	$n = d = 4096$ (64×64)
Basis	$\sqrt{2/n}\sin(k\pi x),\ k = 1{:}20$	complete midpoint cosine tensor
Kernel construction weights	$\lambda_k = 0.85^{k-1}$	$(j^2 + k^2 + 1)^{-0.4},\ j, k = 0{:}63$
α, β	$0.000452249,\ 0.00991983$	$0.0275482,\ 1$
κ_H	21.9345	36.3000
B, T, ν	$10, 500, 0.1$	$50, 300, 0.5$
η_A, η_F	$0.01, 1.008081$	$0.006103516, 0.006103516$
Random replicates	10	10

Curve normalization and target. The pool contains 100 equally spaced points from 0.001 to 0.999, with $\Phi_{ik} = \sqrt{2/n}\sin(k\pi x_i)$. From $y_{\text{raw}}(x) = \sin(2\pi x) + 0.6\sin(6\pi x) + 0.4\sin(14\pi x) + 0.3\cos(10\pi x)$, we construct $c_* = \Lambda^{-1/2}\Phi^\top \mathbf{y}_{\text{raw}}$ and the realizable target $\mathbf{y} = \Psi c_* = \Phi\Phi^\top \mathbf{y}_{\text{raw}}$. The basis satisfies $\|\Phi^\top\Phi/n - I\|_2 = 0.99009$; we compute $H = \Psi^\top\Psi/n$ directly. Its condition number is $\kappa_H = 21.9345$, and the spectral-weight ratio is $\kappa_\Lambda := \max_k \lambda_k / \min_k \lambda_k = 21.9301$.

Cheetah and cameraman normalization and target. Pixel centers are $u_a = (a + 0.5)/64$ and $v_b = (b + 0.5)/64$, with $a, b = 0, \ldots, 63$. Define $\phi_0(u) = 1$ and $\phi_j(u) = \sqrt{2}\cos(\pi j u)$ for $j = 1, \ldots, 63$, and use the complete tensor basis $\phi_j(u_a)\phi_k(v_b)$. Equivalently, $\Phi = \sqrt{n}U$ for the orthogonal two-dimensional cosine synthesis matrix U associated with the type-II discrete cosine transform (DCT-II). Therefore $\Phi^\top\Phi/n = I$, $H = \Lambda$, and every target image is realizable. Frequencies range from zero to 63 in each direction, giving $\lambda_{\min} = (63^2 + 63^2 + 1)^{-0.4} = 0.0275482$.

The cheetah source is a 368×401 RGB image, converted to grayscale before resizing; the cameraman source is the standard 512×512 grayscale test image. Both are scaled to $[0, 1]$ and resized directly to 64×64, without cropping, using bilinear interpolation with antialiasing and reflection at the boundary. The resized values are rounded to single precision before conversion to double precision, fixing the evaluation targets. We use scikit-image 0.25.2 for preprocessing. Mean PSNR averages the per-run values in decibels.

Step selection. The primary curve uses $\nu = 0.1$; cheetah, cameraman, singleton, exhaustive small-pool, and spectral-rescaling controls use $\nu = 0.5$. The primary curve multiplier is shared by all methods and selected from $\{0.05, 0.1, 0.2, 0.5\}$ to display reconstruction and coefficient evolution over 500 updates. The controlled-feature and contour studies in Appendix F select the multiplier on separate training targets; grayscale-to-color teaching uses the common reference multiplier $\nu = 1$. Population uses all n curve points per update at the first-order step in Table 2.

D Curve Teaching

D.1 Teaching budgets and Population comparison

Table 3: Curve pool MSE after completed updates at $\nu = 0.1$. GAT, Greedy, and Random use $B = 10$; Population uses all $n = 100$ points per update. GAT uses approximate batch search. Random is mean $\pm$ sample standard deviation over ten seeds. All methods start at MSE 0.78350 ($t = 0$).

Strategy	$t = 50$	$t = 100$	$t = 150$	$t = 200$	$t = 300$	$t = 500$
GAT (approx.)	0.16757	0.04661	0.01455	0.00456	4.49×10^{-4}	4.35×10^{-6}
Greedy	0.27885	0.11379	0.06321	0.03860	0.01810	0.00453
Random	0.42592	0.24961	0.16110	0.11167	0.06453	0.03080
	$\pm$0.00670	$\pm$0.00340	$\pm$0.00213	$\pm$0.00250	$\pm$0.00205	$\pm$0.00082
Population	0.42294	0.24871	0.15979	0.11123	0.06437	0.03069

GAT attains the lowest pool MSE among the three teacher–learner pairs at every reported budget.

At $t = 500$, Population MSE is 0.03069 versus Random's 0.03080 ± 0.00082. Random has an unbiased pool gradient and an additional gradient-variance term in its expected squared error. Population uses all 100 points per update, while the minibatch methods use ten.

D.2 Spectral trajectories

The spectral panels plot $|c_{t,k} - c_{*,k}|$ in the sine-basis RKHS coordinates. Figure 2a shows $k \in \{2, 6, 14\}$; Figure 1b shows five modes. A Greedy/GAT ratio above one favors GAT. The logarithmic color scale spans 0.7–40 and saturates outside this range, with a five-update display stride. Ratios are masked when the GAT denominator falls below 10^{-8}.

Adaptive changes in the off-diagonal terms of $Q_{\mathcal{B}_t}$ can reverse coordinate updates, producing coefficient fluctuations along a decreasing aggregate-risk trajectory.

E Image Reconstruction

E.1 Cheetah

The cheetah image is attributed to (Zhang et al., 2026a) (CC-BY 4.0).

Table 4: Cheetah reconstruction at fixed teaching budgets ($B = 50$). GAT is exact; Greedy and Random use first-order learners. Random entries average ten runs, with PSNR computed separately for each run.

Strategy	$t = 0$	$t = 50$	$t = 100$	$t = 150$	$t = 200$	$t = 300$
PSNR (dB)						
GAT	1.90	5.70	9.32	13.33	16.74	24.30
Greedy	1.90	5.06	8.09	11.00	13.26	16.38
Random	1.90	4.42	6.87	9.19	11.34	14.94
MSE						
GAT	0.64628	0.26932	0.11700	0.04647	0.02121	0.00372
Greedy	0.64628	0.31167	0.15539	0.07945	0.04716	0.02300
Random	0.64628	0.36181	0.20583	0.12043	0.07337	0.03209

Convergence. At $t = 300$, GAT's MSE is 0.0037184, compared with 0.0229985 for Greedy and 0.0320883 ± 0.0002998 for Random. GAT attains 16.74 dB at $t = 200$, exceeding Greedy's 16.38 dB at $t = 300$ with one-third fewer presentations.

Pixel selection and reconstruction. Figure 4 compares the spatial allocation induced by the curvature-aware batch score, Greedy RKHS-gradient scores, and uniform sampling as reconstruction proceeds.

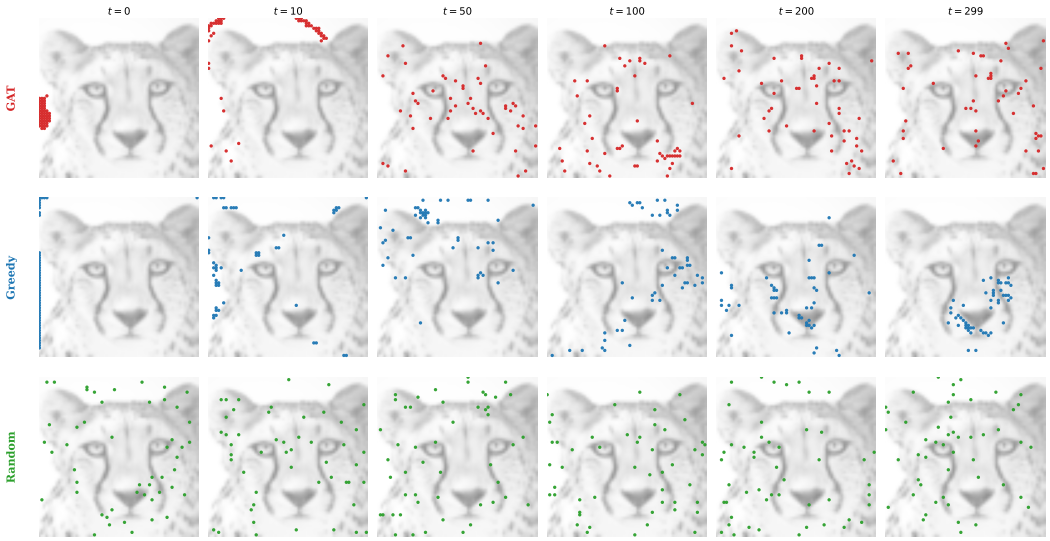


Figure 4: Cheetah batches selected after $t \in \{0, 10, 50, 100, 200, 299\}$ updates. Rows show GAT (red), Greedy (blue), and Random (green, first prespecified replicate); GAT uses the exact complete-pool objective.

For the complete image space, preconditioned updates give $r_{t+1,i} = (1 - \nu)r_{t,i}$ on selected pixels and leave other pixels unchanged. At $\nu = 0.5$, each presentation halves the selected residual.

E.2 CAMERAMAN

Cameraman uses the same complete 64×64 basis, spectrum, batch size, steps, and seed set as cheetah (Table 5 and Figure 5).

Table 5: Cameraman reconstruction at fixed teaching budgets ($B = 50$). GAT is exact; Greedy and Random use first-order learners. Random entries average ten runs, with PSNR computed separately for each run.

Strategy	$t = 0$	$t = 50$	$t = 100$	$t = 150$	$t = 200$	$t = 300$
PSNR (dB)						
GAT	4.79	9.84	14.49	18.83	22.65	30.21
Greedy	4.79	8.44	11.66	14.27	15.91	18.35
Random	4.79	7.11	9.31	11.36	13.24	16.39
MSE						
GAT	0.33207	0.10368	0.03556	0.01310	0.00544	0.00095
Greedy	0.33207	0.14337	0.06818	0.03738	0.02565	0.01463
Random	0.33207	0.19446	0.11711	0.07319	0.04747	0.02294

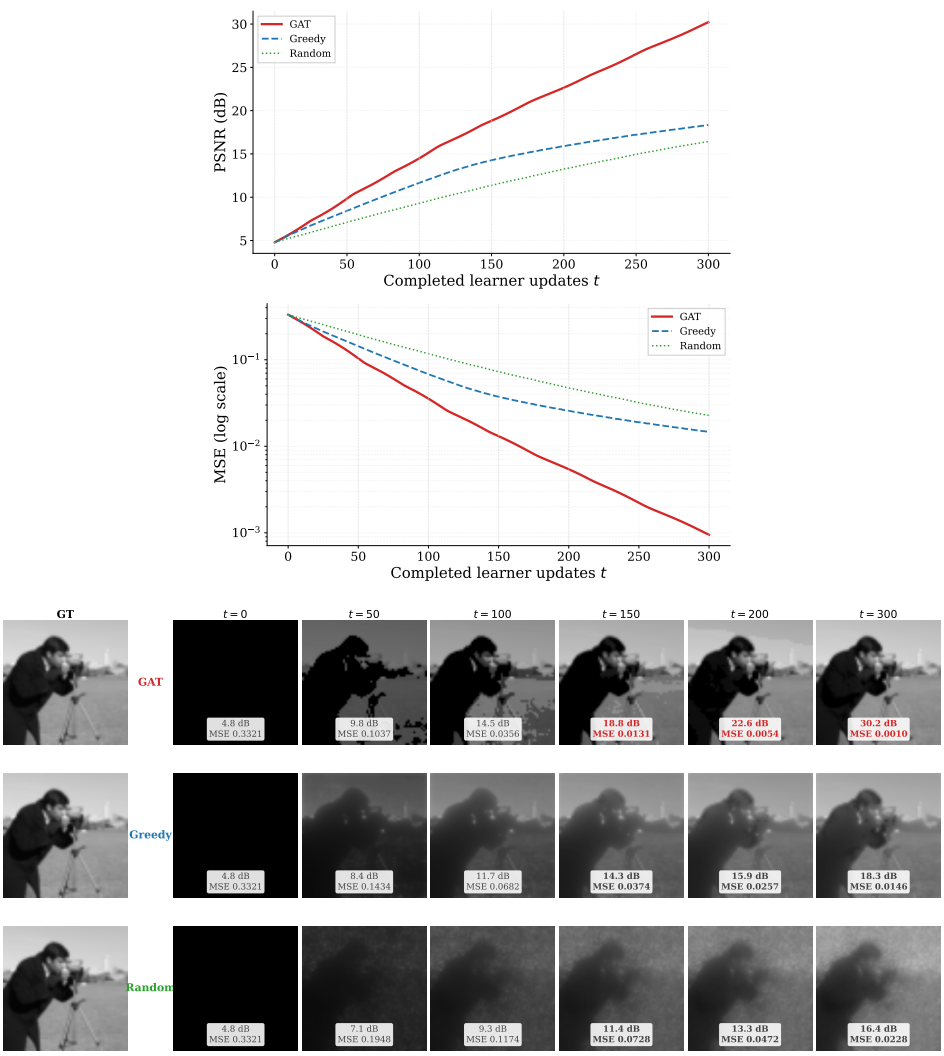


Figure 5: Cameraman PSNR, MSE, and reconstruction. Exact GAT reaches 30.21 dB at $t = 300$, versus 18.35 for Greedy and a ten-seed Random mean of 16.39 dB; all plotted Random trajectories and reconstruction snapshots use the first prespecified replicate.

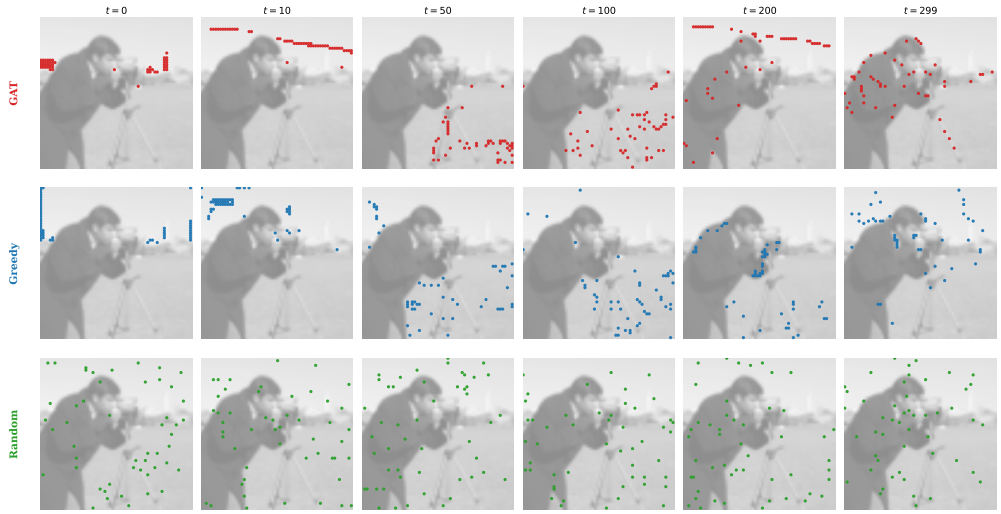


Figure 6: Cameraman selection locations after $t \in \{0, 10, 50, 100, 200, 299\}$ completed updates, with exact GAT, first-order Greedy, and Random (first prespecified replicate).

GAT reaches 18.83 dB at $t = 150$, exceeding Greedy's 18.35 dB at $t = 300$. GAT accelerates reconstruction on both images, with gains determined by their residuals and selection schedules.

F Selection and Learner Geometry

We cross Random, Greedy, and GAT selection with first-order and preconditioned updates on a controlled feature family and grayscale-to-color teaching, and evaluate curvature-aware selection on human contour fields.

F.1 Controlled teacher–learner matching

Design. Let $x_i = i/n$, $n = d = 32$, and let U be the complete real orthonormal discrete Fourier basis. Frequency k has weight $\lambda_k = [1 + (k/4)^2]^{-1}$, shared by its sine and cosine columns. For $C \in \{1, 4, 16, 64\}$, define

$$\Psi_C = \mathrm{diag}(s_C) U \Lambda^{1/2}, \qquad s_C(x_i) = \frac{\exp\{\frac{1}{4}\log(C)\cos[2\pi(x_i - \phi)]\}}{\left(n^{-1}\sum_j \exp\{\frac{1}{2}\log(C)\cos[2\pi(x_j - \phi)]\}\right)^{1/2}}.$$

The parameter C controls spatial contrast in feature amplitude, and ϕ shifts this pattern along the domain. Unit mean-square amplitude preserves the mean kernel diagonal as C varies. The target uses independent standard-normal sine/cosine coefficients at frequencies $1, \ldots, 5$, weighted by $1/k$ and normalized to unit initial MSE; ϕ is uniform on $[0, 1)$. Each target and phase are paired across all settings. We compute the empirical Hessian $H_C = \Psi_C^\top \Psi_C / n$ directly and use $B = 4$, $c_0 = 0$, and 32 updates.

Ten validation targets select the step-size multiplier ν independently for each cell from $\{0.1, 0.25, 0.5, 1\}$ by mean first-pass error, averaged over all four contrast levels. All six cells select $\nu = 1$ with the update steps in Appendix C.1. Evaluation uses 30 further target/phase pairs, averaging ten paired Random streams within each pair. The primary metric is $A_8 = 8^{-1}\sum_{t=0}^{7} \mathrm{MSE}_t/\mathrm{MSE}_0$; uncertainty uses 5,000 paired target-level bootstrap resamples.

Results. With preconditioned updates, GAT reduces A_8 relative to Greedy by 5.9%, 15.7%, and 24.3% at $C = 4, 16, 64$, improving all 30 targets at each contrast level. At $C = 16$, the paired difference is 0.0562 with 95% interval $[0.0467, 0.0671]$. With first-order updates, Greedy outperforms GAT selection at $C > 1$ (Table 6). The constant-diagonal case $C = 1$ provides the reference. At $\nu = 1$, both deterministic selectors reach numerical zero after eight preconditioned updates; A_8 captures their teaching progress during this first pass (Figure 7).

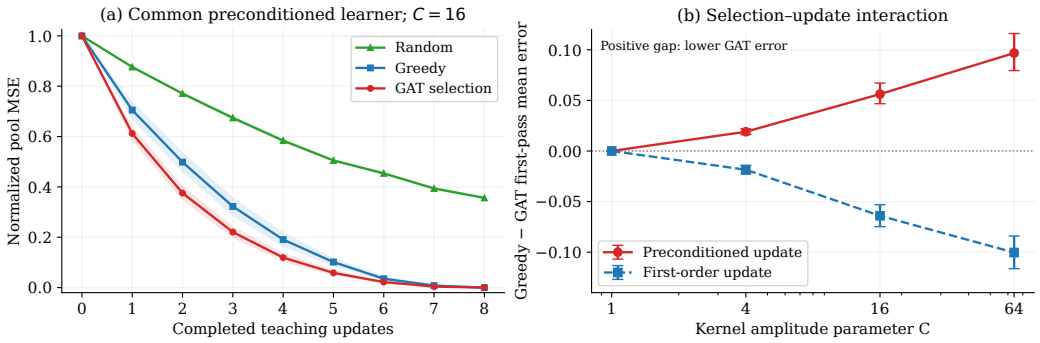


Figure 7: Selection and update geometry across 30 target/phase pairs. (a) Random, Greedy, and GAT selection with the same preconditioned learner at $C = 16$. (b) Paired Greedy-minus-GAT first-pass error under each update rule; positive values favor GAT. Shading and bars show target-level bootstrap 95% intervals after averaging Random streams within targets.

Table 6: First-pass normalized MSE A_8 (mean $\pm$ sample standard deviation over 30 targets). All six cells use validation-selected $\nu = 1$.

	First-order update			Preconditioned update		
C	Random	Greedy	GAT selection	Random	Greedy	GAT selection
1	0.5632 ± 0.0261	0.3132 ± 0.0319	0.3132 ± 0.0319	0.6572 ± 0.0115	0.3015 ± 0.0382	0.3015 ± 0.0382
4	0.6863 ± 0.0408	0.4522 ± 0.0482	0.4707 ± 0.0529	0.6572 ± 0.0115	0.3205 ± 0.0425	0.3015 ± 0.0382
16	0.7733 ± 0.0544	0.5777 ± 0.0828	0.6418 ± 0.0940	0.6572 ± 0.0115	0.3577 ± 0.0519	0.3015 ± 0.0382
64	0.8203 ± 0.0594	0.6489 ± 0.1016	0.7493 ± 0.1196	0.6572 ± 0.0115	0.3983 ± 0.0620	0.3015 ± 0.0382

F.2 Spectral rescaling

The spectral control uses $\Phi = \sqrt{16}\,Q$ from the reduced QR decomposition of a standard-normal 16×6 matrix, target $\Phi(0.3, -0.1, 0.2, 0.4, -0.5, 0.6)^\top$, and initial predictions $\Phi(-0.1, 0.2, 0, 0.1, 0.2, -0.1)^\top$. For $B = 2$, $T = 100$, and $\nu = 0.5$, all 120 batches are enumerated under geometrically spaced spectra with condition numbers $1, 10, 100, 1000$. Selected batches coincide and predictions differ by at most 1.17×10^{-15} (Corollary 9).

F.3 Teaching human contour fields

Task and learner. We teach coordinate-to-contour mappings from BSDS500 human annotations (Arbeláez et al., 2010). This low-rank real-image task evaluates GAT empirically beyond the realizable setting of our convergence analysis. Targets average all annotators' binary boundary masks and use antialiased bilinear resizing with reflection at the boundary. Each full scene retains its aspect ratio, with short side 32 and $n = 1536$ pixels. We fit the resized consensus field directly with a 384-dimensional learner. Twelve official training images determine settings and step sizes;[2] evaluation uses all 200 official test images. The displayed scenes, 100007 and 100039, are the first two test images in filename order.

Let $U \in \mathbb{R}^{n \times 384}$ contain the orthonormal two-dimensional cosine basis functions associated with the type-II discrete cosine transform, ordered by $j^2 + k^2$ with lexicographic ties. Set

$$\lambda_{j,k} = (1 + j^2 + k^2)^{-0.4}, \qquad \Psi = \sqrt{n}\,U\Lambda^{1/2}, \qquad H = \Psi^\top \Psi / n = \Lambda, \qquad P = UU^\top. \quad (39)$$

Here j, k index the two spatial frequencies, and P projects pixel values onto the retained cosine basis. Every policy starts from $c_0 = 0$, presents $B = 8$ coordinates, and uses

$$c_{t+1} = c_t - \frac{\nu B}{n} H^{-1} g_{t,\mathcal{B}_t}, \qquad \boldsymbol{r}_{t+1} = \boldsymbol{r}_t - \nu P_{:,\mathcal{B}_t} \boldsymbol{r}_{t,\mathcal{B}_t}. \quad (40)$$

[2] Training image IDs: 100075, 100080, 100098, 103041, 104022, 105019, 105053, 106020, 106025, 108041, 108073, and 109034.

GAT evaluates the complete curvature-aware score

$$q_t(\mathcal{B}) = \frac{n}{B^2} \boldsymbol{r}_{t,\mathcal{B}}^\top P_{\mathcal{B},\mathcal{B}} \boldsymbol{r}_{t,\mathcal{B}}, \tag{41}$$

including cross terms, using the batch search in Appendix C.2. Greedy selects by $r_i^2 K_{ii}$, where $K_{ii} = n \sum_k U_{ik}^2 \lambda_k$. Residual and singleton-curvature controls use r_i^2 and $r_i^2 P_{ii}$. Random draws a fresh uniform size-eight subset each round.

Evaluation. Step sizes minimize training-scene normalized MSE at $0.5n$ presentations over $\nu \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 3, 4\}$. GAT, Greedy, Residual, and singleton curvature select $\nu = 0.75$; Random selects $\nu = 3$. We evaluate at 768 presentations (96 updates) and run trajectories through $2n$. Each scene uses three shared pixel-priority permutations applied jointly to targets and features. Random averages ten continuation streams per permutation; averaging permutations gives one value per scene. Displayed predictions use the first replicate.

Writing $M_{s,m}(b)$ for scene s's seed-averaged MSE under method m after bn presentations, with initial MSE $M_{s,0}$ and budget fraction b, we report

$$L_m(b) = \frac{1}{200} \sum_s \frac{M_{s,m}(b)}{M_{s,0}}. \tag{42}$$

Relative reductions average $100[1 - M_{s,\mathrm{GAT}}/M_{s,m}]$ over scenes. Figure intervals use 3,000 paired image-bootstrap resamples and are pointwise 95%. Contour-weighted MSE weights squared residuals by consensus target values. All numerical metrics use unclipped predictions, while displayed fields share a common range within each scene.

Table 7: Human-contour teaching at $0.5n$ presentations with a common preconditioner. L_m is mean initial-error-normalized MSE.

Selector	ν	L_m
GAT	0.75	0.2341
Greedy	0.75	0.2648
Random	3	0.3617
Residual	0.75	0.2650
Singleton curvature	0.75	0.2647

Results. At $0.5n$ presentations, mean per-scene MSE reductions are 11.8% relative to Greedy and 35.6% relative to Random, with lower seed-averaged error on all 200 scenes (Table 7). Contour-weighted MSE decreases by 14.4% and 56.2%. At this budget, GAT attains normalized MSE 0.2341, compared with 0.2650 for Residual and 0.2647 for singleton-curvature selection. Figure 8 shows the learning curves and two reconstructions. In mean initial-error-normalized MSE L_m, GAT remains below Greedy from $0.089n$ through $2n$ presentations, while Random becomes slightly lower than GAT after $1.755n$ presentations.

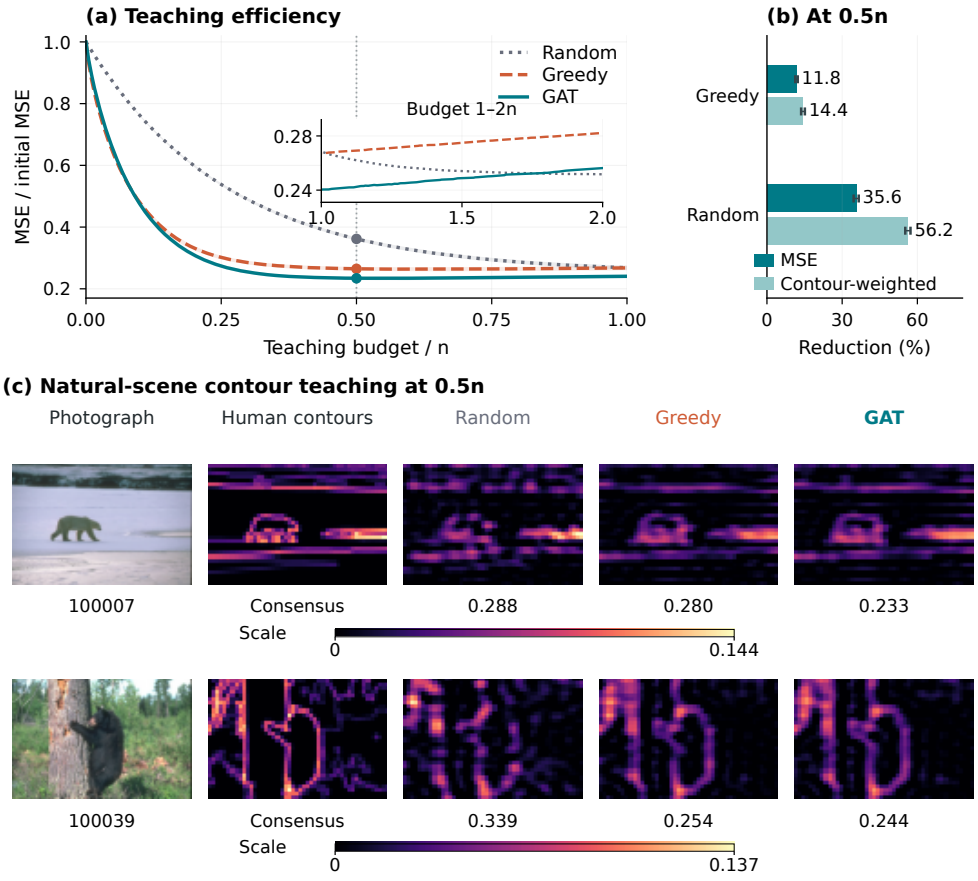


Figure 8: Human-contour teaching with preconditioned updates ($\nu = 0.75$ for GAT and Greedy; $\nu = 3$ for Random). (a) Mean normalized MSE over 200 BSDS500 scenes; the inset extends to $2n$. (b) Mean per-scene reductions at $0.5n$. Shading and bars show pointwise 95% image-bootstrap intervals. (c) Two scenes at $0.5n$, with normalized MSE labels. Predictions use the first replicate and a shared display range per scene.

F.4 Grayscale-to-color teaching

Task and features. We teach color from lightness and spatial coordinates on all 200 BSDS500 test photographs. This low-rank real-image task evaluates GAT empirically beyond the realizable setting of our convergence analysis. Each image is resized with antialiasing to 16×24 pixels or its transpose, giving $n = 384$ teaching points. CIELAB lightness $L \in [0, 100]$ and normalized coordinates (x, y) form the inputs; the two chroma channels $(a, b)/128$ are the targets. Predictions start at zero chroma.

For $z = (x/0.2, y/0.2, (L/100)/0.3)$, the features are 96 Gaussian functions $\exp(-\|z - z_j\|^2/2)$ and a constant, scaled by $1/\sqrt{97}$. Centers follow farthest-point sampling, starting at the input nearest the mean. We select the center count $\{48, 96\}$, spatial bandwidth $\{0.2, 0.5\}$, and lightness bandwidth $\{0.1, 0.3\}$ by mean normalized chroma MSE at $0.5n$ presentations on four training images (100075, 100080, 100098, 103041), averaged equally over Random, Greedy, and GAT with a common preconditioned learner and $\nu = 1$. Training Random results average three fixed streams. The selected features have dimension $d = 97$ on every test image.

Selection and learning. Write the feature matrix's thin singular value decomposition as $\Psi = U\Sigma V^\top$, with $H = \Psi^\top\Psi/n$, $K = \Psi\Psi^\top$, and $P = UU^\top$. For two-channel residuals $\mathsf{R}_t \in \mathbb{R}^{n\times 2}$ and $B = 4$, the GAT score sums the two scalar-output scores:

$$q_t(\mathcal{B}) = \frac{n}{B^2}\,\mathrm{tr}\big(\mathsf{R}_{t,\mathcal{B}}^\top P_{\mathcal{B},\mathcal{B}}\mathsf{R}_{t,\mathcal{B}}\big)\,. \tag{43}$$

Greedy, Residual, and singleton curvature rank $\|\mathsf{R}_{t,i}\|_2^2$ multiplied by K_{ii}, 1, and P_{ii}, respectively. Random samples a uniform size-four subset each round; GAT uses the batch search in Appendix C.2. All cells use $\nu = 1$, giving $\eta_A = B/n$ and $\eta_F = B/(\beta n)$, where $\beta = \|\Psi\|_2^2/n$. The preconditioned and first-order updates are

$$\mathsf{R}_{t+1} = \mathsf{R}_t - P_{:,\mathcal{B}_t}\mathsf{R}_{t,\mathcal{B}_t}, \qquad \mathsf{R}_{t+1} = \mathsf{R}_t - \frac{K_{:,\mathcal{B}_t}}{\|\Psi\|_2^2}\mathsf{R}_{t,\mathcal{B}_t}, \tag{44}$$

respectively. We average three coordinate-priority permutations per photograph and ten paired Random streams per permutation. Chroma MSE is $\|\mathsf{R}_t\|_{\mathrm{F}}^2/(2n)$, normalized by its initial value. The primary budget is $0.5n$ presentations; trajectories extend to n.

RGB PSNR is evaluated at $0.5n$ on the resized 384-pixel pool, combining predicted chroma with observed lightness and converting to sRGB in $[0, 1]$. RGB MSE averages over pixels and channels against the resized image. We compute $-10\log_{10}(\mathrm{MSE}_{\mathrm{RGB}})$ per run, then average over Random streams, coordinate-priority permutations, and photographs.

Table 8: Grayscale-to-color teaching on 200 photographs. Mean initial-error-normalized chroma MSE with $B = 4$ and $\nu = 1$.

Selector	First order $0.5n$	Preconditioned $0.5n$	Preconditioned n presentations
Random	0.5874	0.4270	0.1901
Greedy	0.4783	0.2304	0.1058
GAT	0.5143	0.1898	0.0633
Residual	—	0.2188	0.0953
Singleton curvature	—	0.2427	0.1045

Results. At $0.5n$ presentations, GAT reduces chroma MSE relative to preconditioned Greedy, Residual, and Random by 17.1%, 13.1%, and 55.6% per photograph on average, improving 200, 199, and 198 photographs, respectively (Table 8). With Greedy selection, preconditioning reduces error from 0.4783 to 0.2304; GAT selection further reduces it to 0.1898. With preconditioned updates, mean RGB PSNR at $0.5n$ is 29.60, 28.76, and 25.90 dB for GAT, Greedy, and Random, respectively. Figure 9 shows the normalized chroma-MSE trajectories and RGB reconstructions.

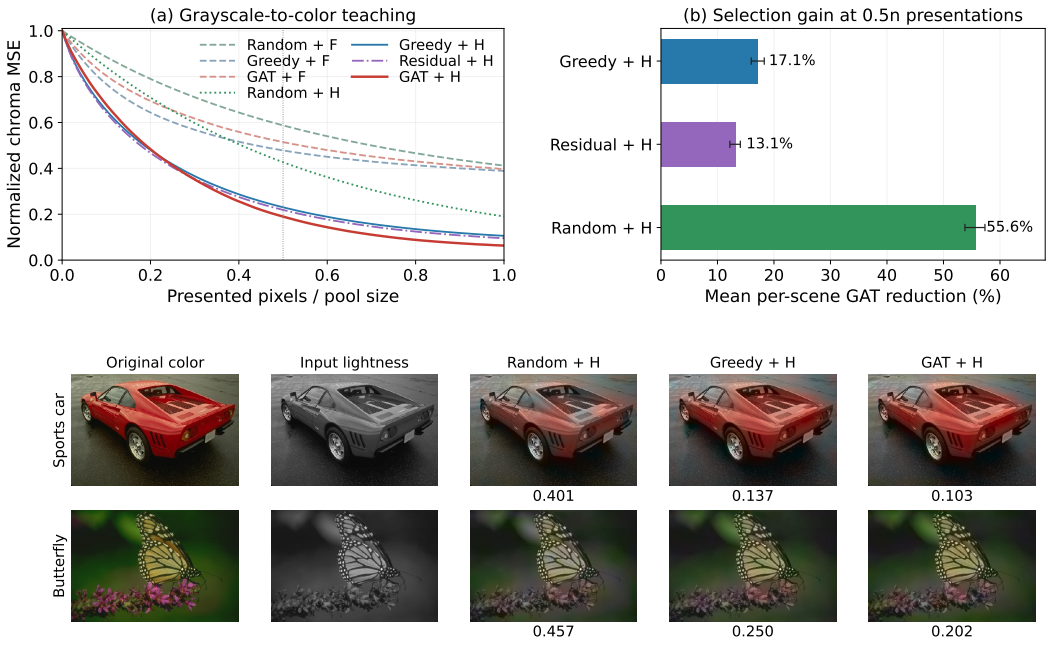


Figure 9: Grayscale-to-color teaching on 200 BSDS500 photographs. F/H denote first-order/preconditioned updates. (a) Mean normalized chroma MSE. (b) Mean per-scene reductions with pointwise 95% paired image-bootstrap intervals (20,000 resamples). Sports-car and butterfly reconstructions use the first replicate at $0.5n$ presentations; values below show normalized MSE on the 384-point pool. Displayed chroma is bilinearly upsampled and combined with original lightness.

G Batch Search and Computation

G.1 Exact and approximate batch selection

Exact selection attains the global maximum of $q_t(\mathcal{B})$ in Equation 15 over all size-B subsets. Approximate batch search seeks a high-scoring batch through a restricted candidate search or local improvement, as specified in Appendix C.2. Selection accuracy concerns this maximization; Hessian accuracy concerns the curvature operator used to evaluate the score and update the learner. Our approximate batch search uses the exact empirical Hessian.

For squared loss, the score expands as

$$q_t(\mathcal{B}) = \frac{1}{B^2} \sum_{i,j \in \mathcal{B}} r_{t,i} r_{t,j} \psi_i^\top H^{-1} \psi_j.$$

The cross terms couple the selected examples, making the general problem a joint batch optimization. In a complete pool space ($d = n$), Equation 27 gives $\psi_i^\top H^{-1} \psi_j = n \mathbf{1}_{\{i=j\}}$, so $q_t(\mathcal{B}) = nB^{-2} \sum_{i \in \mathcal{B}} r_{t,i}^2$. Selecting the B largest squared residuals therefore solves the maximization exactly by sorting. Exhaustive enumeration provides an exact reference for the small low-rank pools below.

Exhaustive controls. The singleton curve control evaluates all scores ($B = 1, T = 500, \nu = 0.5$). A small-pool control uses $n = 12, d = 4, B = 3, T = 150$, and $\nu = 0.5$, with $\Psi = ZA\Lambda^{1/2}$: Z is standard normal, A is upper bidiagonal with unit diagonal and superdiagonal $(0.2, 0.3, 0.2)$, $\Lambda = \mathrm{diag}(1, 0.4, 0.15, 0.04)$, and $\mathbf{y} = ZA(0.3, -0.6, 0.7, 1)^\top$. Local search matches exhaustive enumeration of all 220 batches at every state of the exact-GAT trajectory.

Natural-image batch search. We assess batch-search accuracy on all 200 BSDS500 test scenes, area-averaging each complete contour field to a 4×6 grid (or its transpose), giving $n = 24$ coordinates, and retaining six cosine modes. With $B = 4$, exact selection enumerates all $\binom{24}{4} = 10{,}626$ batches. Exact selection and the four-swap search in Appendix C.2 use the same preconditioned learner, zero

initialization, and $\nu = 0.75$. At each state visited by either policy, we compare both selected batches with the exact optimum. On these reduced contour pools, the search attains 99.86% of the optimal score on average and selects a globally optimal batch at 97.19% of the 9,600 state visits across two policies, two coordinate-priority permutations, and 12 updates per scene. Table 9 reports search accuracy and the resulting teaching trajectories.

Table 9: Exhaustive versus approximate selection on 24-coordinate contour pools with six retained modes. MSE is normalized by its initial value and averaged over all 200 scenes. Score comparisons evaluate both selectors at the same states from both trajectories.

	Exact	Approximate
Mean score / optimum (%)	100	99.86
Minimum score / optimum (%)	100	80.07
Globally optimal batch (%)	100	97.19
Normalized MSE, $0.5n$ presentations	0.3138	0.3136
Normalized MSE, n presentations	0.1964	0.1968
Normalized MSE, $2n$ presentations	0.1791	0.1793

G.2 Image-teaching computation

Table 10 reports computation at fixed presentation budgets with a common preconditioned learner. Measurements use one AMD Ryzen Threadripper PRO 5975WX core and one BLAS thread. We rotate method order over three repeats and average case-wise median times. Reconstruction measurements use three paired streams; colorization uses 20 uniformly spaced test photographs. Reconstruction selectors use cached scores and stable partial sorting; colorization uses compiled selection kernels. The optimized implementations reproduce the reference batches and errors. GAT and Greedy have similar runtimes on both reconstruction images. GAT takes 2.00 ms more per photograph than Greedy at the primary colorization budget.

Table 10: Image-teaching time in milliseconds for GAT, Greedy, and Random selection with a common preconditioned learner. BT counts example presentations. Method times include feature and curvature construction, policy initialization, selection, and updates, starting from prepared image inputs. Common setup is averaged across methods.

Task	BT	Common setup	GAT	Greedy	Random
Cheetah	15,000	0.13	6.48	6.54	3.44
Cameraman	15,000	0.12	7.08	7.16	3.53
Colorization	192	5.12	7.66	5.66	6.05