# ProjKAN: Model Compression via KAN Projections to Bridge the Hypothesis and Capacity Gaps

**Mazlum Ferhat Arslan**[1,2,*] **Weihong Guo**[1] **Shuo Li**[2,3]
[1]Dept. of Mathematics, Applied Mathematics and Statistics
[2]Dept. of Biomedical Engineering
[3]Dept. of Computer and Data Sciences
Case Western Reserve University
mferhata@gmail.com, {wx49,shuo.li11}@case.edu
[*]Corresponding author

## Abstract

Multilayer perceptron (MLP)-based model compression often overlooks two distinct sources of error: the hypothesis gap and the capacity gap. The hypothesis gap measures the additional error from choosing a student class at a fixed parameter budget, while the capacity gap is the best teacher-approximation error attainable at that budget across the candidate classes. We show that Kolmogorov–Arnold Networks (KANs) can reduce the hypothesis gap for smooth low-dimensional target functions. Furthermore, with fixed layer inputs, basis, and compressed architecture, KAN layer coefficient fitting is a convex problem. Building on this formulation, we derive projection-based algorithms for width and depth reduction, as well as a function transfer method that maps a trained MLP teacher into a compact KAN student. These are unified in the Gap-aware Projections via KANs (ProjKAN) framework, which integrates the error decomposition, theoretical guarantees, and compression algorithms into a single methodology. Empirically, in regimes where pruning and standard knowledge distillation degrade sharply, ProjKAN achieves improved accuracy–size tradeoffs, reduced train–test gaps, and greater robustness in data-scarce settings.

## 1 Introduction

MLP hypothesis class-based model compression has two problems: hypothesis gap and capacity gap Fig. 1a. Most existing methods restrict a trained MLP via pruning (LeCun et al., 1989), quantization (Han et al., 2015a), or distillation (Hinton et al., 2015), so the hypothesis class is fixed by the teacher. Therefore any mismatch between the target structure and this class is paid as a hypothesis gap (Cho & Hariharan, 2019). Moreover, even within a well-chosen class, a small student can be intrinsically too limited, which creates a capacity gap (Mirzadeh et al., 2020). The gap refinement of the standard error decomposition (Vapnik, 2013; Shalev-Shwartz & Ben-David, 2014; Bottou & Bousquet, 2007) serves to separate the two failure modes, allowing to address each directly.

Kolmogorov Arnold networks (KANs) (Liu et al., 2024) are a promising student hypothesis class in structured smooth settings, yet using them for compression is difficult for two reasons. First, KAN-native compression is not a drop-in version of MLP pruning, because KAN layers are built from univariate edge functions rather than multivariate weight matrices. Existing KAN pruning is framed as simplification for interpretability, so it does not directly yield a general within-KAN reduction operator. In Fig. 1b, this is the challenge of finding an appropriate vector $v_2$. Second, large KAN models are hard to obtain because KAN training is much slower than MLP training (Liu et al., 2024; Li, 2024). Hence one often starts from a trained MLP, however MLP to KAN transfer is non-trivial

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

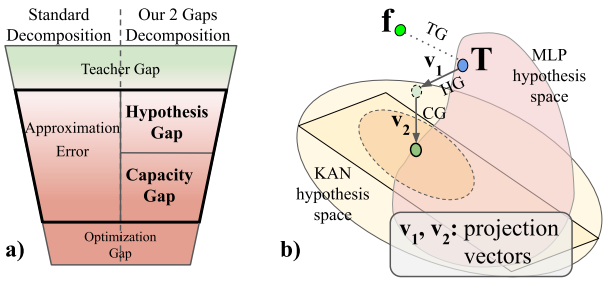



Figure 1: **a) Two-gap view and b) our solution as its consequence.** a) Refinining the standard error decomposition splits the approximation error term into a hypothesis gap ( additional error from the chosen class at a fixed budget) and a capacity gap ( best remaining teacher-approximation error across the candidate classes at that budget). Thus *compression before optimization gap* becomes two-staged. b) Function-space view of our solution. Our two distinct steps: *i.* Find $v_1$ which projects $T$ into KAN hypothesis space, *ii.* find $v_2$ mapping the initial KAN model into a smaller constrained KAN. $f$ denotes the unknown target function; $T$, the teacher.

because multivariate layers must be rewritten as sums of univariate functions; in Fig. 1b, this is the challenge of finding a suitable vector $v_1$.

We propose ProjKAN, a KAN-based framework guided by a two-gap decomposition, justified by theory, and instantiated as three algorithms (Fig. 2). First, 2GD separates loss due to hypothesis-class choice from the best remaining error across candidate classes at a fixed budget, while keeping teacher mismatch and optimization effects conceptually separate. Therefore, it clarifies whether improvement requires changing the hypothesis class or compressing more effectively within it. Second, we give a KAN approximation bound for analytic ridge sums and an explicit axis-aligned analytic subclass with a better error–size rate than shallow ReLU MLPs. Third, for fixed coefficient subspaces, KAN compression reduces to orthogonal projection in an $L_2(\nu)$ coefficient space, which yields the best approximation under the stated constraint. Consequently, ProjKAN results in Algorithm 1 (depth merging), Algorithm 2 (width pruning), and Algorithm 3 (MLP to KAN weight transfer) as a coherent pipeline.

Projection-based pruning predates our work. Our novelty is the KAN-specific coefficient-linear formulation of depth and width compression, its coefficient-space certificates, and the heterogeneous MLP-to-KAN transfer algorithm. Our contributions include: *i)* showing that restricting students to the teacher family can be suboptimal, and giving an analytic ridge-sum KAN bound with an explicit shallow-ReLU separation on an axis-aligned subclass *ii)* casting KAN layer coefficient fitting with fixed inputs, basis, and architecture as a convex problem, with projection certificates for the stated objectives *iii)* proposing projection-based MLP to KAN transfer with auxiliary latents and empirical data-efficiency gains.

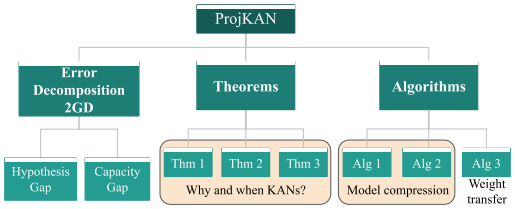



Figure 2: Gap-aware Projections via KAN (ProjKAN) addresses model compression via *i)* refined error decomposition (2GD), *ii)* identifying limits of nested students ( Thm 1), an analytic approximation regime (Thm 2), and fixed-subspace projections (Thm 3) and *iii)* depth merging (Alg. 1), width pruning (Alg. 2) and MLP to KAN weight transfer (Alg.3) algorithms.

## 2 Related Work

**Pruning compresses within a fixed hypothesis family.** Most pruning-based compression methods define the student as a subnetwork of a trained teacher (unstructured weight sparsification (LeCun et al., 1989; Han et al., 2015b), structured channel/filter removal (Li et al., 2017; Wen et al., 2016)), and then recover accuracy by fine-tuning. This framing is effective when the best budget model is well-approximated by deleting parameters from the teacher, but it hard-codes a hypothesis choice: the student must live inside a nested restriction of the teacher's parameterization (Frankle & Carbin, 2018). As a result, pruning primarily targets the budget-error tradeoff *within* one family, and can miss cases where a different parameterization achieves lower error at the same budget. ProjKAN is motivated by this gap: it treats hypothesis

choice as part of compression, separating the loss due to committing to a suboptimal hypothesis space from the irreducible loss due to the budget itself. In settings with smooth, low-effective-dimensional structure, ProjKAN uses KANs (Liu et al., 2024) as an alternative student hypothesis space rather than as a pruned subset of the teacher, aiming to reduce the loss that comes from staying inside the teacher's geometry.

**Heterogeneous transfer usually imitates outputs, not the function form.** When teacher and student differ architecturally, the dominant approach is knowledge distillation (KD): train the student to match the teacher's logits, softened probabilities (Hinton et al., 2015), or intermediate features (Romero et al., 2015). Distillation is broadly applicable, but the supervision is indirect: matching outputs does not guarantee that the student preserves how the teacher partitions the input space, and the transfer can deform decision boundaries when capacity is tight. Other cross-architecture approaches align representations (e.g., matching attention maps (Zagoruyko & Komodakis, 2017) or feature statistics) or use function-preserving network morphisms (Wei et al., 2016; Chen et al., 2015), but these typically rely on strong architectural compatibility or still require iterative retraining to settle into a good approximation. ProjKAN instead focuses on *heterogeneous weight transfer* into a different parameterization that is designed to be efficient in structured smooth regimes: it aims to transfer functional structure into KAN edge functions so the student starts closer to a functionally consistent representation, reducing reliance on post-hoc imitation alone.

**Certifiable compression is limited; KAN layers make it natural.** A smaller set of works seek formal guarantees for compression, typically by recasting parts of pruning or layer reconstruction as convex programs (Carreira-Perpinán & Idelbayev, 2018; Aghasi et al., 2017), deriving stability bounds, or certifying worst-case deviations under assumptions. These guarantees are often surrogate-driven, or loose, because the end-to-end mapping of standard MLP models is nonlinear in parameters and exact global optimization under combinatorial constraints is generally intractable. NMR-KAN uses function-space regularization (Arslan et al., 2025). ProjKAN differs by exploiting a property specific to KAN parameterizations (Liu et al., 2024): under an $L^2(\nu)$ coefficient representation and fixed structural choices, closed linear coefficient constraints admit an optimal orthogonal projection. This yields an optimal compressed layer under the stated objective and an explicit, computable error certificate, directly tying budget enforcement to a tractable best-approximation primitive rather than a heuristic or surrogate objective.

## 3 Methodology

This section presents a gap-driven approach to compression goals (Fig.3). This approach motivates transferring to KAN hypothesis space for low-dimensional target functions along with principled KAN-native pruning algorithms.

**Notation and Setup** Let $\mathcal{X}$ be the input space and let $\mathcal{P}$ be a distribution over $\mathcal{X}$. Let $f : \mathcal{X} \to \mathbb{R}^m$ denote the unknown target function, and let $g : \mathcal{X} \to \mathbb{R}^m$ denote a fixed teacher model. Population discrepancy between two functions $u, v : \mathcal{X} \to \mathbb{R}^m$ is $\|u - v\| := \mathbb{E}_{x\sim\mathcal{P}}\big[\|u(x) - v(x)\|\big]$. **Hypothesis families.** Let $\mathcal{I}$ be an index set of *representations* (e.g., architectural primitives,

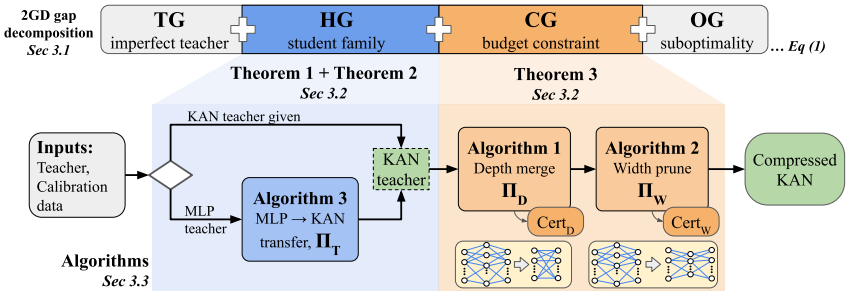



Figure 3: **Methodology flow as conceptual lens−justification−procedure.** Sec. 3.1 separates hypothesis gap from capacity gap; Sec. 3.2 identifies an analytic approximation regime and fixed-subspace KAN projections; Sec. 3.3 turns these results into compression procedures and projection-motivated MLP→KAN transfer.

nonlinearities, parameterizations). For each $i \in \mathcal{I}$ and parameter budget $p \in \mathbb{N}$, let $\mathcal{H}_i^p$ be a hypothesis class with (approximately) $p$ parameters. For a family of functions $\mathcal{H} := \{\mathcal{H}_i^p\}_{i\in\mathcal{I},\, p\in\mathbb{N}}$ with fixed $(i,p)$, $\hat{h} \in \mathcal{H}_i^p$ denotes the compressed model by a given compression procedure. **Teacher surrogate.** Since $f$ is unknown, all terms below that involve approximation within $\mathcal{H}$ are evaluated with respect to $g$, while $\|f-g\|$ is treated as a constant term.

### 3.1 2GD: Two Gaps Between Teacher and Optimization

Two-gap decomposition (2GD) identifies otherwise entangled notions to show that the overall training error has a fundamentally different semantic parts, in line with other works in statistical learning theory (Vapnik, 2013; Bottou & Bousquet, 2007; Shalev-Shwartz & Ben-David, 2014). It separates hypothesis-family mismatch (HG), best budget-limited approximation across candidate families (CG), and residual optimization/compression error within that family (OG) as

$$\begin{aligned}\|f-\hat{h}\| \;\le\; \underbrace{\|f-g\|}_{\mathrm{TG}} \;+\; \underbrace{\Big(\inf_{h\in\mathcal{H}_i^p}\|g-h\| - \inf_{j\in\mathcal{I}}\inf_{h\in\mathcal{H}_j^p}\|g-h\|\Big)}_{\mathrm{HG}} \\ +\; \underbrace{\inf_{j\in\mathcal{I}}\inf_{h\in\mathcal{H}_j^p}\|g-h\|}_{\mathrm{CG}} \;+\; \underbrace{\Big(\|g-\hat{h}\| - \inf_{h\in\mathcal{H}_i^p}\|g-h\|\Big)}_{\mathrm{OG}},\end{aligned} \tag{1}$$

where $i$ is a fixed student family at parameter budget $p$, and $\hat{h} \in \mathcal{H}_i^p$.

The quantity $\|f-\hat{h}\|$ admits the bound (1), separating *i)* teacher mismatch (TG), *ii)* hypothesis-space choice at fixed budget (HG), *iii)* best achievable error at budget $p$ (CG), and *iv)* suboptimality of the training/compression procedure within $\mathcal{H}_i^p$ (OG).

*Remark* 1. Assuming $\mathcal{H}_i^{p_1} \subset \mathcal{H}_i^{p_2}$ whenever $p_1 < p_2$, the function $p \mapsto \min_{h\in\mathcal{H}_i^p}\|g-h\|$ is nonincreasing. Under this nesting, many practical heuristics (e.g., identifying dead/constant neurons, structured pruning rules that remove provably redundant units) primarily target the subproblem of reducing $p$ while controlling $\min_{h\in\mathcal{H}_i^p}\|g-h\|$, i.e., the chosen-family approximation error $\mathrm{HG}+\mathrm{CG}$ is controlled without comparing representation families.

### 3.2 Hypothesis Choice and Convex Compression for KANs

While 2GD exposes HG and CG as different concepts, it does not suggest a better hypothesis family in general. In the following Sec. 3.2.1 investigates HG and gives a concrete preferred region for KANs over MLPs, while Sec. 3.2.2 studies budget-constrained approximation within the KAN family.

#### 3.2.1 A concrete win-region for depth-2 KANs over depth-2 ReLU-MLPs

Theorem 1 formalizes a structural limitation of *compressing within the teacher*: the best student at a fixed budget need not be in the same hypothesis family. The proof is in Appendix C.

**Theorem 1** (Optimal students need not be nested)**.** *Let $(\mathcal{X},\mathbb{P}_X)$ be a distributional input space and let $\mathcal{R}(f) := \|f-f^\star\|_{L^2(\mathbb{P}_X)}$ denote the root mean squared error with respect to a target $f^\star \in L^2(\mathbb{P}_X)$. Fix a parameter budget $p \in \mathbb{N}$. Let $\mathcal{H}_T \subset L^2(\mathbb{P}_X)$ be a "teacher" hypothesis class and let $\mathcal{H}_S \subseteq \mathcal{H}_T$ be a strict subset (the "student" class). Define the best achievable errors at budget $p$ by*

$$\begin{aligned}\mathcal{R}_T(p) &:= \inf\{\mathcal{R}(f) : f\in\mathcal{H}_T,\ \mathrm{params}(f)\le p\},\\ \mathcal{R}_S(p) &:= \inf\{\mathcal{R}(f) : f\in\mathcal{H}_S,\ \mathrm{params}(f)\le p\}.\end{aligned}$$

*Then there exist choices of $(\mathcal{X},\mathbb{P}_X)$, a target $f^\star$, a teacher class $\mathcal{H}_T$, a strict subset $\mathcal{H}_S \subsetneq \mathcal{H}_T$, and a budget $p$ such that*

$$\mathcal{R}_S(p) \;>\; \inf\{\mathcal{R}(f) : f\in\mathcal{H},\ \mathrm{params}(f)\le p\}$$

*for some alternative class $\mathcal{H}$ that is not a subset of $\mathcal{H}_T$. Equivalently: at a fixed parameter budget $p$, restricting students to lie in a subset of the teacher hypothesis space can be strictly suboptimal.*

Theorem 1 states that restricting the student to be in the teacher's hypothesis space can exclude a minimizer that would be achievable by switching representations. This limitation persists even if we could solve the within-family approximation problem exactly (so the optimization gap vanishes).

We next state an approximation guarantee for the structured regime $f(x) = \sum_{k=1}^{r} g_k(w_k^\top x), \quad \|w_k\|_2 = 1$. For each $k$, define $I_k = w_k^\top D$ for $D = [0,1]^d$. We assume that, after affine rescaling of $I_k$ to $[-1,1]$, each $g_k$ extends analytically to a Bernstein ellipse of parameter at least $\rho > 1$, and is bounded there by $M_k$.

**Theorem 2** (Ordinary KAN rate for analytic ridge-sum targets)**.** *Under the assumptions above, for $n \geq 2$ there exists a KAN of shape $[d, r, 1]$ whose edge functions are represented with $n$ Chebyshev coefficients per edge and whose allocated coefficient count is $O(rdn)$, such that*

$$\|f - \widehat{f}_{\mathrm{KAN}}\|_{L^\infty(D)} \leq C_f \rho^{-n}, \qquad C_f := C_\rho \sum_{k=1}^{r} M_k. \tag{2}$$

*Consequently, uniform error at most $\varepsilon$ is achieved with $O\Big(rd \log \frac{C_f}{\varepsilon}\Big)$ KAN edge coefficients.*

Theorem 2 gives the parameter rate to achieve arbitrarily small error to smooth ridge-sum target functions $f$ (with the same role as $f$ in (1)).

**Comparison to shallow ReLU networks.** The KAN upper bound holds over the stated analytic ridge-sum class. The ReLU comparison is an explicit separation on a subclass: for an axis-aligned analytic target $f(x) = g(x_1)$ with $g'' \geq c_0 > 0$ on a compact interval, such as $g(t) = e^t$, the best width-$m$ one-hidden-layer ReLU $L^2(P_X)$ error under the uniform product measure is $\Omega_g(m^{-2})$. The KAN construction attains $O_g(\rho_g^{-n})$, with $\rho_g > 1$. At fixed $d$ and matched parameter budget $P$, this gives $\Omega_g(P^{-2})$ ReLU error versus $O_g(e^{-\gamma_g P})$ KAN error. Appendix C.2.1 gives the proof using classical approximation results (Barrow & Smith, 1978; Siegel & Xu, 2022). The broader comparative claim for arbitrary ridge directions with pairwise interactions remains a conjecture.

### 3.2.2 Locally convex, budget-constrained compression within KAN hypothesis space

This section shows that KAN layers admit *certified coefficient-space projections*. With the layer inputs, basis, and compressed architecture fixed, constraints defining a closed linear coefficient subspace admit a globally optimal orthogonal projection under the stated norm. The following theorem applies Hilbert-space projection theorem to the KAN layer's coefficient vector which is viewed as a linearized vector.

**Theorem 3** (Certified compressibility of KAN layers)**.** *Fix a measure $\nu$ on $\mathbb{R}$ and let $\{\psi_\ell\}_{\ell \geq 1}$ be an orthonormal basis of $L_2(\nu)$. Consider a KAN layer/operator of the form $(\Phi(x))_j = \sum_{i=1}^{m_{\mathrm{in}}} \phi_{ij}(x_i)$ with edge functions $\phi_{ij} \in L_2(\nu)$ admitting expansions $\phi_{ij}(t) = \sum_{\ell \geq 1} a_{ij,\ell} \psi_\ell(t)$. Let $\mathcal{S}$ be any closed linear subspace of the coefficient Hilbert space, and define the induced compressed class*

$$\mathfrak{C}(\mathcal{S}) = \Big\{ \widetilde{\Phi} : \textit{ the coefficient collection of } \widetilde{\Phi} \textit{ lies in } \mathcal{S} \Big\}.$$

*Then the best approximation of $\Phi$ in $\mathfrak{C}(\mathcal{S})$ under the $L_2(\nu)$-induced norm exists and is unique, and satisfies the Pythagorean identity (certificate):*

$$\widetilde{\Phi}^\star = \Pi_{\mathcal{S}} \Phi = \arg\min_{\widetilde{\Phi} \in \mathfrak{C}(\mathcal{S})} \|\Phi - \widetilde{\Phi}\|^2, \qquad \|\Phi - \widetilde{\Phi}\|^2 = \|\Phi - \widetilde{\Phi}^\star\|^2 + \|\widetilde{\Phi}^\star - \widetilde{\Phi}\|^2, \qquad \forall\, \widetilde{\Phi} \in \mathfrak{C}(\mathcal{S}).$$

Theorem 3 has three practical implications: *i) Optimality:* $\widetilde{\Phi}^\star$ minimizes error in the fixed coefficient subspace. *ii) Certificate:* the optimal error $\|\Phi - \widetilde{\Phi}^\star\|^2$ is computable from coefficients. *iii) Suboptimality gap:* any other $\widetilde{\Phi}$ pays exactly $\|\widetilde{\Phi} - \widetilde{\Phi}^\star\|^2$ extra error. The norm here sums edge-function energies. For fixed features, coefficient least squares is convex. Theorem 3 gives the unregularized orthogonal projection; ridge fitting minimizes the corresponding penalized objective.

### 3.3 Algorithms: KAN Compression and Heterogeneous Weight Transfer via Projections

Theoretical insights from Sec. 3.2 identify an analytic approximation regime and fixed-subspace KAN projections with error certificates. Algorithms 1,2 capitalize on these insights (Fig. 4).

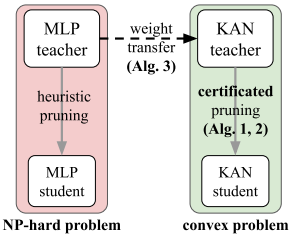



Figure 4: **Different hypothesis gaps cause the change in vertical axis.** Alg. 1 and Alg. 2 addresses layer and width pruning. Algorithm 3 transfers an MLP teacher into a KAN through coefficient fits and auxiliary-state updates.

#### 3.3.1 Certified KAN-Layer Pruning by Projection Merge

Algorithm 1 prunes depth by removing one intermediate KAN layer via *two-layer merging*: given two consecutive layers $\Phi_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^M$ and $\Phi_{\ell+1} : \mathbb{R}^M \to \mathbb{R}^{d_{\ell+2}}$, we approximate their composition by a single KAN layer $\widetilde{\Phi}_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^{d_{\ell+2}}$ with the *same* basis family and basis size $Q$. We choose $\nu$ as the empirical distribution of the layer input activations:

$$\hat{\nu} = \frac{1}{N}\sum_{k=1}^{N} \delta_{z^{(k)}},\ z^{(k)} = \Phi_{\ell-1} \circ \cdots \circ \Phi_1(x^{(k)}),\ \ x^{(k)} \sim \mathcal{D}.$$

Let the target outputs be $y^{(k)} = (\Phi_{\ell+1} \circ \Phi_\ell)(z^{(k)})$ and denote by $\psi(\cdot) \in \mathbb{R}^Q$ the fixed univariate basis feature map used by our KAN layers (including the layer's range map). Define the concatenated feature map

$$\phi(z) = \left[\psi(z_1)^\top, \ldots, \psi(z_{d_\ell})^\top\right]^\top \in \mathbb{R}^{d_\ell Q},$$

and the design matrix $\Phi \in \mathbb{R}^{N \times (d_\ell Q)}$ with rows $\phi(z^{(k)})^\top$. Because a KAN layer is linear in its coefficients given $\phi(\cdot)$, merging reduces to multi-output ridge regression (Hoerl & Kennard, 1970):

$$(\widetilde{W}, \widetilde{\beta}) = \arg\min_{W,\beta} \frac{1}{N}\sum_{k=1}^{N} \left\|\Phi_{k:}W + \beta - y^{(k)}\right\|_2^2 \ + \ \frac{\lambda}{N}\|W\|_F^2.$$

Writing centered matrices $\Phi_c = \Phi - \mathbf{1}\bar{\Phi}$ and $Y_c = Y - \mathbf{1}\bar{Y}$, the solution is given in closed form:

$$\widetilde{W} = (\Phi_c^\top \Phi_c + \lambda I)^\dagger \Phi_c^\top Y_c, \qquad \widetilde{\beta} = \bar{Y} - \bar{\Phi}\widetilde{W}.$$

We reshape $\widetilde{W}$ into KAN coefficients $\widetilde{A} \in \mathbb{R}^{d_{\ell+2} \times d_\ell \times Q}$ so that $\widetilde{\Phi}_\ell(z) = \widetilde{\beta} + \langle \widetilde{A}, \psi(z) \rangle$. We report the empirical reconstruction residual: $\mathrm{Cert}_D = \frac{1}{N}\sum_{k=1}^{N} \left\|y^{(k)} - \widetilde{\Phi}_\ell(z^{(k)})\right\|_2^2$. Here † denotes the Moore–Penrose inverse; for $\lambda > 0$ it is the ordinary inverse. At $\lambda = 0$, fitted outputs are the unique empirical least-squares projection, although coefficients need not be unique.

#### 3.3.2 Certified KAN-Width Pruning by Group-Selection Projection

Algorithm 2 instantiates Theorem 3's projection view for width pruning by choosing $\nu$ as the empirical distribution of scalar layer inputs from a small calibration set $\mathcal{D}_c = \{x^{(k)}\}_{k=1}^N$

$$\hat{\nu}_1 = \frac{1}{Nd_\ell}\sum_{k=1}^{N}\sum_{i=1}^{d_\ell} \delta_{z_i^{(k)}}, \quad \hat{\nu}_2 = \frac{1}{NM}\sum_{k=1}^{N}\sum_{m=1}^{M} \delta_{h_m^{(k)}},$$

where $z^{(k)} = \Phi_{\ell-1} \circ \cdots \circ \Phi_1(x^{(k)})$ and $h^{(k)} = \Phi_\ell(z^{(k)})$. Parameterizing KAN edge functions in a fixed basis enables the computation of the *induced* $L^2(\nu)$ inner product via an empirical Gram matrix in coefficient space:

$$\hat{G}_1 := \mathbb{E}_{t\sim\hat{\nu}_1}\left[\psi(t)\,\psi(t)^\intercal\right], \quad \hat{G}_2 := \mathbb{E}_{t\sim\hat{\nu}_2}\left[\psi(t)\,\psi(t)^\intercal\right],$$

then for any edge function with coefficient vector $a$,

$$\|\varphi\|_{L^2(\hat{\nu}_\ell)}^2 \equiv \mathbb{E}_{t\sim\hat{\nu}_\ell}\left[\varphi(t)^2\right] = a^\intercal \hat{G}_\ell a.$$

In the direct sum of incoming and outgoing edge-coefficient spaces, hidden units occupy disjoint coordinate blocks. Retaining the $r$ largest group energies minimizes the discarded coefficient energy (Yuan & Lin, 2006):

$$E_m = \sum_{i=1}^{d_\ell} (a_{im}^{(\ell)})^\intercal\, \hat{G}_1\, a_{im}^{(\ell)} \ + \ \sum_{j=1}^{d_{\ell+1}} (a_{mj}^{(\ell+1)})^\intercal\, \hat{G}_2\, a_{mj}^{(\ell+1)},$$

The discarded tail energy is the exact projection error for this direct-sum objective: $\text{Cert}_W = \sum_{m \notin J^*} E_m$, where $J^*$ indexes the $r$ largest $E_m$. Incoming edges are included because deleting a unit removes both incident coefficient blocks. With $J^*$ fixed, the successor layer is optionally refit by convex ridge regression. The refit residual measures the successor layer's output fit; $\text{Cert}_W$ measures the coefficient energy discarded by selection. Algorithm 2 adds $\gamma I$ to the empirical Grams, so its tail energy uses the corresponding regularized coefficient norm. Without damping, a singular Gram defines a seminorm; the projection is understood on the quotient by its nullspace.

### 3.3.3 Auxiliary-State Projection for Data-Efficient Weight Transfer

While Algorithms 1–2 use fixed-feature coefficient fits for KAN teachers, current pretrained models are typically MLP backbones. Training a KAN teacher from scratch at scale can require large datasets and expensive gradient-based optimization. This section therefore proposes an innovative *iterative* MLP→KAN transfer scheme that remains in the spirit of projection: it alternates closed-form ridge *projection updates* of KAN coefficients with updates of intermediate (student) states. The key advantage over end-to-end gradient training is that each layer update is a convex regression in KAN coefficient space, enabling stable updates with data-efficiency.

Let $g$ be a fixed MLP teacher of depth $L$ and let $\Phi = \Phi_{L-1} \circ \cdots \circ \Phi_0$ be a KAN student with matched macro-architecture (same depth and widths) and fixed univariate basis $\psi(\cdot)$. Compression by Algorithms 1–2 follows transfer. Given an unlabeled calibration set $\mathcal{D}_c = \{x^{(n)}\}_{n=1}^N$, denote teacher activations by $H_0^T(x) = x$ and $H_{k+1}^T(x)$ for $k = 0, \ldots, L-1$.

**Auxiliary-state objective.** We introduce per-layer auxiliary states $U_k^{(n)} \in \mathbb{R}^{d_k}$ (student hidden representations) , motivated by the penalized splitting objective

$$\min_{\{\Phi_k\},\{U_k^{(n)}\}} \sum_{n=1}^{N} \|U_L^{(n)} - H_L^T(x^{(n)})\|_2^2 + \frac{\rho}{2} \sum_{k=0}^{L-1} \sum_{n=1}^{N} \|U_{k+1}^{(n)} - \Phi_k(U_k^{(n)})\|_2^2 + \sum_{k=0}^{L-1} \lambda_k \|\theta_k\|_2^2, \quad (3)$$

with a weak anchor $\sum_{k=1}^{L-1} \alpha_k \sum_n \|U_k^{(n)} - H_k^T(x^{(n)})\|_2^2$. Unlike strict feature matching, $U$ is *not* forced to equal teacher activations; it is allowed to evolve to representations compatible with KAN layer form. This is to overcome the mismatch between KANs (univariate nonlinearities) and MLPs (multivariate nonlinearities).

**Alternating updates.** Starting from layer-wise ridge fits, we alternate:

1. **$\Phi$-step (projection update).** For each layer $k$, this step solves

$$\Phi_k \leftarrow \arg\min_{\Phi \in \mathcal{C}_k} \frac{1}{N} \sum_{n=1}^{N} \|U_{k+1}^{(n)} - \Phi(U_k^{(n)})\|_2^2 + r_k \|\theta\|_2^2,$$

   where $r_k = 2\lambda_k/(\rho N)$ for the weights in the splitting objective above. This reduces to multi-output ridge regression because a KAN layer is linear in its coefficient parameters given the basis features.

2. **$U$-step (latent feature update).** With $\Phi$ fixed, this step updates $U$ by

$$U_{k+1}^{(n)} \leftarrow \frac{\rho \Phi_k(U_k^{(n)}) + \alpha_{k+1} H_{k+1}^T(x^{(n)})}{\rho + \alpha_{k+1}}, \alpha_{k+1} \geq 0.$$

In practice, a small number of sweeps (10 in reported results) suffices to reduce latent feature mismatch while retaining the closed-form, data-efficient layer updates. Algorithm 3 combines exact ridge fits at fixed auxiliary states with forward anchored updates of the auxiliary states.

## 3.4 Discussion of Methodical Advantages

**Advantage for model compression:** *Data efficient compression.* We decompose the student-teacher loss into explicit gaps (teacher mismatch, hypothesis, capacity, optimization), turning model compression into targeted, auditable objectives. For fixed KAN coefficient subspaces, orthogonal projection in an $L^2(\nu)$ coefficient Hilbert space gives an optimal fit and a certificate for that objective. This enables data-efficient model compression.

**Advantage over MLPs:** *Tractable pruning.* KAN layer coefficient fitting with fixed inputs, basis, and architecture is convex, whereas selecting a pruning structure is generally combinatorial. MLP layer-reconstruction methods can also have convex refits once their structure is fixed. Therefore, the models can be compressed without gradient-based updates which demand high data and computation resources.

**Advantage over pure-KAN approaches:** *Function preservation.* Heterogeneous weight transfer initializes KAN students from MLP teachers to better preserve the teacher function and avoid distortions that can arise from naive distillation. This mitigates the known optimization difficulty of large KANs trained from scratch.

## 4 Results and Discussion

Experiments use implicit neural representations (INR) for encoding signed distance functions (SDFs) and MNIST : INR is a continuous approximation task with interpretable geometry, while MNIST is a controlled classification benchmark. CIFAR10/CIFAR100 are used in the results reported in Appendix B. **Methods.** INR: KD into budget-matched MLP/KAN students (Hinton et al., 2015; Liu et al., 2024). MNIST: KD (rand. init) (Hinton et al., 2015); ridge initialization (Hoerl & Kennard, 1970); pruning+KD MLPs via $\ell_1$ structured pruning (Paszke et al., 2019), Coreset (Mussay et al., 2019), SNIP (Lee et al., 2018), and Fisher pruning (Theis et al., 2018). We denote Alg. 1 as $D$ (depth merging) and Alg. 2 as $W$ (width pruning).

### 4.1 INR at Extreme Compression

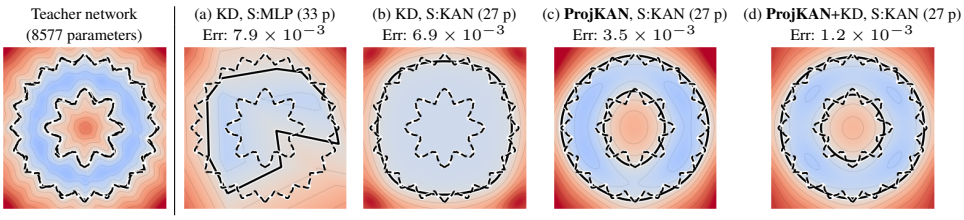



Figure 5: **ProjKAN preserves topology (enclosed hole) while attaining overall high fidelity (MSE) at extreme compression.** Left: Teacher network (8577 parameters, 3 hidden layers each with 64 p) defines the target SDF. (a) Standard distillation into MLP (33 p) fails to capture the shape. (b) Distillation into KAN (27 p) improves the outlines but not the topology. (c) ProjKAN (27 p) recovers the topology. (d) ProjKAN followed by KD achieves the best fidelity ($1.2\times10^{-3}$) with over $300\times$ compression.

**Geometry preservation under fixed budget** Figure 5 compares students at comparable extreme parameter budgets against an Implicit Neural Representation (Park et al., 2019) teacher defining a 2D signed-distance function (SDF) target. Under $\sim 300\times$ compression, standard distillation into an MLP student fails to preserve the target topology, while distillation into a KAN student improves topology but remains noisy. Projection-based KAN pruning substantially improves shape recovery at this scale, and ProjKAN achieves the strongest fidelity, yielding the lowest MSE among the compared methods.

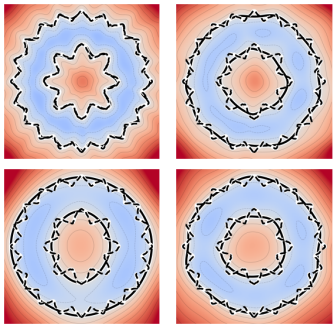


Figure 6: **ProjKAN's steps gradually remove details.** Clockwise: Teacher (8577 params), one-layer merge (4545 p), two-layer merge (417 p), width pruning (27 p).

**Projection-based pruning trajectory** Figure 6 visualizes the progressive reduction of a trained teacher through depth merging followed by width pruning. From left to right, the sequence applies one-layer merge, two-layer merge, and width pruning, reducing the model from $8577p \to 4545p \to 417p \to 27p$. The intermediate reconstructions exhibit gradual simplification rather than abrupt failure, suggesting that projection-based depth/width reductions can preserve salient global structure even when driven to very small budgets.

Table 1: **ProjKAN preserves accuracy under large parameter reduction, while MLP baselines collapse at small $n$.** LeNet-300-100 (MNIST) compression: student accuracies pre/post KD for $n \in \{512, 256, 128, 64\}$, comparing MLP baselines with KAN variants (Ours-D/W/DW) at $11\times$ and $28\times$ parameter reduction. Additional CIFAR results and reconstruction controls are in Appendix B.

| Method | T | Rate | Pre-KD acc | | | | Post-KD acc | | | | Rate | Pre-KD acc | | | | Post-KD acc | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | n=512 | 256 | 128 | **64** | 512 | 256 | 128 | **64** | | 512 | 256 | 128 | **64** | 512 | 256 | 128 | **64** |
| Baseline | MLP | 11.1× | — 21.4 — | | | | 86.3 | 82.5 | 76.4 | 69.1 | 27.7× | — 13.9 — | | | | 82.1 | 77.6 | 70.0 | 61.8 |
| $\ell_1$ | MLP | 11.1× | — 29.6 — | | | | 87.9 | 83.9 | 76.5 | 68.2 | 27.7× | — 18.3 — | | | | **82.6** | 77.9 | 70.6 | 61.4 |
| Coreset | MLP | 11.1× | — 27.2 — | | | | 78.4 | 74.7 | 67.7 | 56.8 | 27.7× | — 17.6 — | | | | 34.2 | 36.6 | 35.8 | 33.1 |
| SNIP | MLP | 11.1× | 21.3 | 25.0 | 22.2 | 23.5 | **88.3** | 84.3 | 78.9 | 69.3 | 27.7× | 17.7 | 13.8 | 15.6 | 16.9 | 82.3 | 76.8 | 71.3 | 62.8 |
| Fisher | MLP | 11.1× | 22.5 | 24.2 | 20.1 | 25.6 | 88.2 | 84.3 | 77.9 | 68.0 | 27.7× | 16.5 | 14.9 | 14.9 | 16.3 | 80.9 | **78.0** | 71.0 | 62.5 |
| **Ours-DW** | KAN | 11.2× | **84.7** | **81.4** | **80.1** | **71.7** | **88.3** | **85.3** | **83.2** | **74.5** | 27.7× | 52.7 | 44.2 | 39.7 | 39.7 | 81.1 | 76.6 | **74.3** | **67.6** |
| **Ours-D** | KAN | 11.4× | 79.5 | 78.4 | 75.3 | 67.5 | 80.7 | 79.6 | 76.1 | 68.5 | – | (network out of depth) | | | | | | | |
| **Ours-W** | KAN | 11.2× | 79.4 | 78.4 | 74.8 | 66.7 | 84.6 | 82.6 | 78.9 | 69.8 | 27.5× | **55.7** | **55.5** | **53.6** | **49.3** | 66.7 | 65.0 | 62.2 | 55.8 |

### 4.2 LeNet-300-100 on MNIST

**Low-data advantage.** At $28\times$ compression, Ours-DW approaches the strongest MLP baseline at $n = 512$ and $n = 256$. Its deficit remains within 1.5 points. At $n = 128$ and $n = 64$, Ours-DW leads by 3.0 and 4.8 points.

**Changing role of KD.** For Ours-DW, KD adds 2.8–3.9 points at $11\times$ compression. At $28\times$, KD adds 27.9–34.6 points. Severe compression therefore creates substantially more accuracy for KD to recover.

**Joint depth–width pruning under severe constraints.** At $28\times$, Ours-W achieves higher pre-KD accuracy across all data sizes. Ours-DW finishes 11.6–14.4 points higher after KD. Depth-only pruning cannot reach this budget because the network exhausts its removable depth. Width reduction enables extreme compression, while joint pruning yields the strongest post-KD KAN results.

**Data efficient function transfer.** Figure 7 reports test accuracy versus labeled samples for MLP to KAN function transfer. KD (random init) degrades most sharply at small $n$ and exhibits the largest seed-to-seed variability. Ridge initialization only ($\lambda = 10$) improves accuracy without a subsequent KD phase, indicating that coefficient-space ridge fitting provides a favorable initialization. Ridge initialization + KD further improves or matches performance across $n$. Our Algorithm 3 achieves the best and the most steady performance improving low-sample accuracy relative to KD (random initialization), while Alg. 3 + KD is slightly stronger for most $n$. As $n$ increases, curves converge and variability decreases, suggesting reduced sensitivity to the transfer procedure in the data-rich regime.

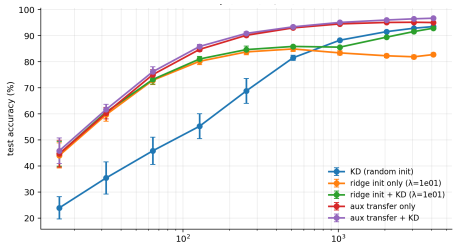



Figure 7: **Algorithm 3 yields large data-efficiency gains.** Test accuracy versus labeled samples $n = 2^k$, $k \in \{4, \ldots, 12\}$. Compared against KD, ridge init only ($\lambda = 10$), ridge init+KD ($\lambda = 10$), Alg.3, and Alg.3+KD. Points are averaged over 5 seeds; error bars denote std.

**Empirical train–test gaps.** Figure 8 reports the in-domain train–test gap as a function of $n$. The KD (random init) baseline exhibits the largest gap magnitude in the smallest-data regime, consistent with overfitting when optimization starts from an uninformed initialization. Ridge-based procedures reduce the gap and its variance; in particular, ridge init + KD narrows the gap while improving test accuracy (Fig. 7), indicating better generalization rather than improved training fit alone. Alg. 3 shows a similar trend: Alg. 3+KD achieves a smaller gap than KD (random init) at small $n$, while maintaining higher test accuracy. For all methods, the gap decreases toward zero as $n$ increases, and differences between transfer procedures diminish once sufficient labeled data is available.

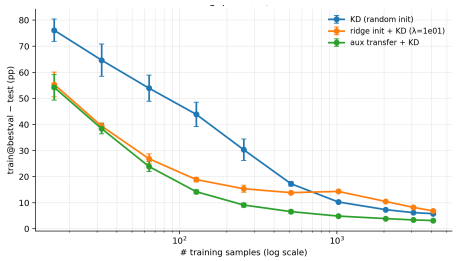



Figure 8: **Smaller empirical train–test gaps at small $n$.** Algorithm 3 yields smaller gaps than KD at small $n$ in this in-domain evaluation.

## 5 Summary and Conclusion

ProjKAN is a framework for model compression identifying and targeting the gaps between teacher networks and the optimization step. Specifically, 2GD decomposition separates models the approximation error as a combined effect of hypothesis and capacity gaps. Within this model, KAN students provide a principled alternative to MLP-based model compression for low-dimensional target functions, and projection-based constructions yield both practical procedures and certificates that quantify the constrained fit at the layer level. Empirically, in the targeted regimes where pruning and standard distillation degrade, ProjKAN delivers stronger accuracy-size tradeoffs, smaller train–test gaps, and more stable behavior at data scarce scenarios. **Limitations.** The approximation guarantees assume analytic ridge structure. The empirical evaluation uses in-domain tests . The certificates quantify layer-level reconstruction and coefficient approximation. **Future work.** A Barron-type extension could separate finite-width and finite edge-basis errors; Appendix C.4 states this proposed next step.

**Broader Impact.** Our principled approach on function classes and approximation regimes aims at clarifying when alternative architectures can offer better efficiency. These may contribute to more energy efficient model design and to the effectiveness of deployed systems. We do not anticipate unique ethical concerns beyond those broadly associated with more efficient ML models.

**Code availability.** Code: `https://github.com/kaptres/kan_compression`.

## Acknowledgments and Disclosure of Funding

The authors declare no competing interests.

## References

Aghasi, A., Abdi, A., Nguyen, N., and Romberg, J. Net-trim: Convex pruning of deep neural networks with performance guarantee. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Arslan, M. F., Guo, W., and Li, S. Neuromanifold-regularized KANs for shape-fair feature representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12790–12799, 2025.

Arslan, M. F., Guo, W., and Li, S. Emergence, retention and mitigation of ill-conditioning due to basis lifting in KANs. In *Advances in Neural Information Processing Systems*, 2026.

Barrow, D. L. and Smith, P. W. Asymptotic properties of best $L_2[0, 1]$ approximation by splines with variable knots. *Quarterly of Applied Mathematics*, 36(3):293–304, 1978.

Bottou, L. and Bousquet, O. The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20, 2007.

Carreira-Perpinán, M. A. and Idelbayev, Y. "learning-compression" algorithms for neural net pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8532–8541, 2018.

Chen, T., Goodfellow, I., and Shlens, J. Net2net: Accelerating learning via knowledge transfer. *arXiv preprint arXiv:1511.05641*, 2015.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794–4802, 2019.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015a.

Han, S., Pool, J., Tran, J., and Dally, W. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, volume 28, 2015b.

He, Y., Zhang, X., and Sun, J. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1389–1397, 2017.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

LeCun, Y., Denker, J., and Solla, S. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Lee, N., Ajanthan, T., and Torr, P. H. SNIP: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340*, 2018.

Li, H., Kadav, A., Durdanovic, I., Samet, H., and Graf, H. P. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, 2017.

Li, Z. Kolmogorov-arnold networks are radial basis function networks. *arXiv preprint arXiv:2405.06721*, 2024.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., and Tegmark, M. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

Luo, J.-H., Wu, J., and Lin, W. ThiNet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Luo, J.-H., Zhang, H., Zhou, H.-Y., Xie, C.-W., Wu, J., and Lin, W. ThiNet: Pruning CNN filters for a thinner net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2525–2538, 2019.

Mirzadeh, S. I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., and Ghasemzadeh, H. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Mussay, B., Osadchy, M., Braverman, V., Zhou, S., and Feldman, D. Data-independent neural pruning via coresets. *arXiv preprint arXiv:1907.04018*, 2019.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Siegel, J. W. and Xu, J. High-order approximation rates for shallow neural networks with cosine and $\mathrm{ReLU}^k$ activation functions. *Applied and Computational Harmonic Analysis*, 58:1–26, 2022.

Theis, L., Korshunova, I., Tejani, A., and Huszár, F. Faster gaze prediction with dense networks and fisher pruning. *arXiv preprint arXiv:1801.05787*, 2018.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 2013.

Wei, T., Wang, C., Rui, Y., and Chen, C. W. Network morphism. In *International Conference on Machine Learning*, pp. 2329–2337, 2016.

Wen, W., Wu, C., Wang, Y., Chen, Y., and Li, H. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

# A More on Implicit Neural Representations with MLPs

Figure 9 indicates the extent to which optimization choices can improve MLP-based SDF compression. By reporting the best result over learning rate and weight decay, the curve approximates the best error that this student family can reach at each parameter budget under our training setup. When the budget is very small, the residual error is mainly a capacity effect: the MLP has too few degrees of freedom to match the teacher SDF, so further tuning cannot remove the mismatch.

Variation in which hyperparameters win across budgets also implies that small MLPs are optimization-sensitive: stable training requires a narrow range of step sizes and regularization, and modest deviations can increase error substantially. As the budget grows, the minimum achievable error decreases and becomes less sensitive to these knobs, consistent with larger models being easier to fit and less dependent on regularization to avoid underfitting. Exhaustive hyperparameter-tuning cannot compensate for tight parameter limits, so improving fidelity at extreme compression requires a different student class or a compression method that preserves function structure.

# B Additional experiments

## B.1 Matched reconstruction baselines

We evaluated both the ICCV and TPAMI ThiNet algorithms (Luo et al., 2017, 2019), plus LASSO channel selection with OLS refitting (He et al., 2017). Tables 2–4 report accuracy before and after KD at the two student budgets used for each dataset. The MNIST settings correspond to the roughly $11\times$ and $28\times$ reductions in Table 1.

Ours-DW wins 89 of 96 ThiNet/LASSO comparisons and 47 of 48 post-KD comparisons.

Table 2: MNIST reconstruction comparisons at two student budgets (A/B). Entries are test accuracy (%), pre-KD above and post-KD below.

| Budget | $n$ | ThiNet-I | ThiNet-T | LASSO | Ours-DW |
|---|---|---|---|---|---|
| A | 64 | 21.5 | 39.2 | 59.6 | 71.7 |
| | | 34.4 | 61.1 | 69.2 | 74.5 |
| A | 128 | 22.8 | 37.6 | 66.7 | 80.1 |
| | | 42.7 | 67.7 | 79.3 | 83.2 |
| A | 256 | 23.2 | 37.0 | 70.8 | 81.4 |
| | | 41.5 | 71.1 | 86.4 | 85.3 |
| A | 512 | 21.5 | 36.6 | 69.1 | 84.7 |
| | | 36.4 | 76.8 | 87.9 | 88.3 |
| B | 64 | 19.5 | 29.7 | 47.3 | 39.7 |
| | | 28.3 | 50.9 | 59.7 | 67.6 |
| B | 128 | 16.6 | 30.4 | 51.3 | 39.7 |
| | | 26.8 | 55.0 | 70.6 | 74.3 |
| B | 256 | 18.9 | 25.9 | 53.7 | 44.2 |
| | | 26.6 | 55.1 | 73.8 | 76.6 |
| B | 512 | 20.1 | 27.0 | 56.7 | 52.7 |
| | | 31.1 | 60.2 | 79.2 | 81.1 |

Table 3: CIFAR-10 reconstruction comparisons at two student budgets (A/B). Entries are test accuracy (%), pre-KD above and post-KD below.

| Budget | $n$ | ThiNet-I | ThiNet-T | LASSO | Ours-DW |
|---|---|---|---|---|---|
| A | 64 | 23.0 | 23.8 | 24.0 | 29.6 |
| | | 26.0 | 25.0 | 25.0 | 30.1 |
| A | 512 | 21.3 | 24.1 | 33.0 | 39.4 |
| | | 30.4 | 30.2 | 36.7 | 40.8 |
| B | 64 | 19.8 | 20.9 | 18.1 | 27.0 |
| | | 24.7 | 24.8 | 18.1 | 27.9 |
| B | 512 | 19.1 | 20.9 | 33.0 | 32.8 |
| | | 33.4 | 31.0 | 33.6 | 34.9 |

Table 4: CIFAR-100 reconstruction comparisons at two student budgets (A/B). Entries are test accuracy (%), pre-KD above and post-KD below.

| Budget | $n$ | ThiNet-I | ThiNet-T | LASSO | Ours-DW |
|---|---|---|---|---|---|
| A | 100 | 5.0 | 5.2 | 6.8 | 7.8 |
| | | 6.8 | 7.0 | 7.4 | 8.3 |
| A | 400 | 5.3 | 5.0 | 10.4 | 11.7 |
| | | 8.6 | 9.1 | 11.4 | 13.0 |
| B | 100 | 3.2 | 3.1 | 5.2 | 5.5 |
| | | 4.8 | 4.1 | 5.6 | 6.1 |
| B | 400 | 3.1 | 3.0 | 8.0 | 7.3 |
| | | 5.4 | 5.5 | 9.2 | 9.3 |

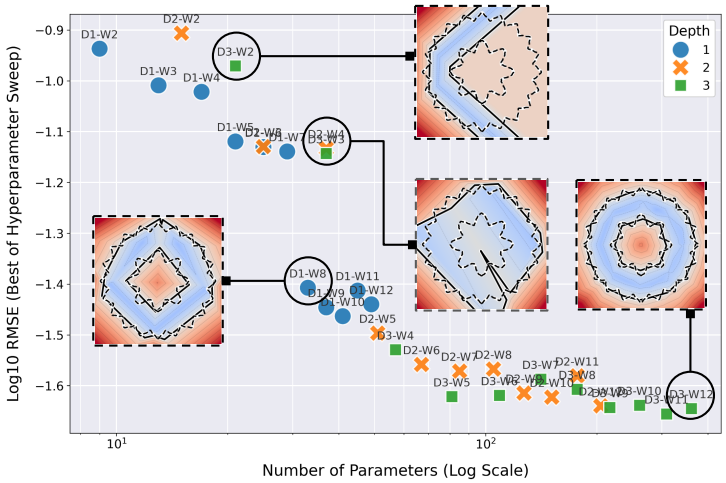



Figure 9: Best approximating MLP student models for SDF compression over hyperparameter search versus number of parameters. The grid-search is for learning rate and weight decay in the range $[10^{-5}, 10^{-1}]$ for both.

### B.2 CIFAR-100 ResNet56 stage compression

We trained a native KANResNet56 on CIFAR-100 (89.84% top-5 accuracy, 856k parameters) and compressed its second stage with 100 images. Random, SNIP, and Fisher were applied to a conventional ResNet56 (91.92% top-5 accuracy, 861k parameters). Table 5 reports top-5 accuracy before and after KD, averaged over three seeds. Ours-Conv is strongest among these methods before and after KD at each reported budget. Its post-KD advantage over the strongest baseline increases from 0.30 points at 70% compression to 8.06 at 80% and 32.92 at 90%.

Table 5: CIFAR-100 ResNet56 second-stage compression. Top-5 accuracy (%), before/after KD on separate lines, averaged over three seeds. Percentages refer to the compressed stage.

| Stage compression | Ours-Conv | Random | SNIP | Fisher |
|---|---|---|---|---|
| 70% | 74.57 | 12.51 | 13.10 | 12.32 |
| | 81.24 | 80.82 | 80.94 | 80.38 |
| 80% | 64.67 | 8.38 | 8.59 | 8.54 |
| | 75.90 | 55.78 | 67.84 | 67.77 |
| 90% | 45.35 | 7.23 | 7.38 | 7.68 |
| | 59.24 | 17.54 | 26.32 | 23.92 |

### B.3 Certificate diagnostic

For retained channel sets $S_b$, we calculate the normalized certificate as

$$\mathcal{C}(S) = \frac{\sum_b \sum_{m \notin S_b} E_{b,m}}{\sum_{b,m} E_{b,m}},$$

where the energy of convolutional channel $m$ in block $b$ is

$$E_{b,m} = \sum_{i,k} (a^{(b,1)}_{m,i,k})^\top G_{b,1} a^{(b,1)}_{m,i,k} + \sum_{o,k} (a^{(b,2)}_{o,m,k})^\top G_{b,2} a^{(b,2)}_{o,m,k}.$$

The Gram matrices are computed from the calibration samples. This is the discarded energy in the direct-sum coefficient objective.

Because discarded energy and accuracy loss both increase mechanically with compression, we compare channel masks at the same compression rate. Using 100 class-balanced calibration images,

we compressed the teacher network's second stage by 70%, 80%, and 90%, and evaluated on the full test set. All four selectors use the same parameter budget and the same coefficient refit. Table 6 reports the certificate and pre-KD accuracy, averaged over three seeds per method.

Table 6: Coefficient certificate and pre-KD top-5 accuracy (%) at fixed stage budgets. Certificate above, accuracy below; averages over three seeds. Lower certificate values are preferred.

| Stage compression | Ours-Conv | KAN-$\ell_1$ | Random-0 | Random-1 |
|---|---|---|---|---|
| 70% | 0.650 | 0.680 | 0.702 | 0.703 |
| | 74.57 | 72.94 | 73.07 | 71.52 |
| 80% | 0.764 | 0.789 | 0.802 | 0.803 |
| | 64.67 | 59.38 | 56.48 | 55.51 |
| 90% | 0.875 | 0.892 | 0.904 | 0.903 |
| | 45.35 | 41.17 | 32.67 | 34.99 |

Across the three budgets, the mean within-budget rank agreement is Spearman $\rho = 0.933$ and Kendall $\tau = 0.889$. The certificate provides an empirical within-budget diagnostic for choosing a channel mask.

### B.4 ImageNet ResNet18 transfer

As a proof-of-concept scaling experiment for Algorithm 3, we tested heterogeneous transfer of a ResNet18 pretrained on ImageNet. Using a convolutional extension, we replaced its second stage with ConvKAN blocks. The validation set is split into 1,000 calibration images and 49,000 test images. The transferred network reaches 87.36% top-5 accuracy, compared with 89.44% for the original ResNet18.

# C Proofs for Section 3

## C.1 Theorem 1

The formal statement appears in Theorem 1.

*Proof.* We prove the claim by exhibiting two concrete counterexamples, each showing that the budget-optimal student can lie outside a natural subset restriction.

**Counterexample 1 (quadratic target; ReLU–MLP vs. quadratic class).** Let $\mathcal{X} = [-1, 1]$ and $\mathbb{P}_X$ be the uniform distribution. Take the target $f^\star(x) = x^2$. Let the teacher class $\mathcal{H}_T$ be a depth-2 ReLU MLP family (width varying) and let the student class $\mathcal{H}_S \subseteq \mathcal{H}_T$ be any strict subset that, at a given budget $p$, excludes all functions representable by a degree-2 polynomial parameterization with $\leq p$ parameters (for example, a subset restriction that enforces a small width or a sparsity pattern that cannot realize $x^2$ exactly at that budget). Now consider the alternative class $\mathcal{H}$ consisting of all degree-2 polynomials $q(x) = ax^2 + bx + c$, with $\mathrm{params}(q) = 3$. For any budget $p \geq 3$, we can choose $q = f^\star$ exactly, hence

$$\inf\{\mathcal{R}(f) : f \in \mathcal{H},\ \mathrm{params}(f) \leq p\} = 0.$$

At a fixed parameter budget, these univariate ReLU networks have a bounded number of affine pieces. The best piecewise-affine approximation to $x^2$ with that many pieces has strictly positive $L^2$ error, so $\mathcal{R}_S(p) > 0$. Therefore, for this problem and budget $p$, the best achievable student within the subset $\mathcal{H}_S \subseteq \mathcal{H}_T$ is strictly worse than a student taken from $\mathcal{H}$, which is not a subset of $\mathcal{H}_T$. This witnesses strict suboptimality of the subset restriction.

**Counterexample 2 (ReLU target; polynomial KAN vs. ReLU).** Let $\mathcal{X} = [-1, 1]$ with $\mathbb{P}_X$ uniform and take the target $f^\star(x) = \mathrm{ReLU}(x) = \max\{x, 0\}$. Let the teacher class $\mathcal{H}_T$ be the class of one-layer polynomial KANs whose edge functions are polynomials of degree at most $n$ (with degree $n$ varying with the budget), and let $\mathcal{H}_S \subseteq \mathcal{H}_T$ be any strict subset restriction (e.g. a subspace/compression constraint on coefficients) at the same parameter budget $p$. Because every $f \in \mathcal{H}_S$ is a polynomial on $[-1, 1]$, each such $f$ is continuous and differentiable everywhere on $(-1, 1)$, hence in particular its left and right derivatives at 0 agree. However, $f^\star(x) = \mathrm{ReLU}(x)$ has a kink at 0: the left derivative is 0 and the right derivative is 1. Consequently, no polynomial can equal $f^\star$ almost everywhere on $[-1, 1]$, and therefore the approximation error cannot be driven to 0 at any fixed finite degree. Formally, for any fixed degree bound $n$ (hence any fixed parameter budget $p$ tied to that $n$),

$$\begin{aligned}&\inf\{\mathcal{R}(f) : f \in \mathcal{H}_T,\ \mathrm{params}(f) \leq p\}\\ &=\ \inf\{\|f - \mathrm{ReLU}\|_{L^2(\mathbb{P}_X)} : \rho(f) \leq n\}\ >\ 0,\end{aligned}$$

where $\rho(f)$ stands for the degree of the polynomial . The fixed-degree polynomial space is finite-dimensional and closed in $L^2(\mathbb{P}_X)$, so its distance to the non-polynomial target is positive. The same lower bound holds for the subset $\mathcal{H}_S \subseteq \mathcal{H}_T$. In contrast, if we take the alternative class $\mathcal{H}$ to be depth-2 ReLU networks, then $f^\star \in \mathcal{H}$ is represented exactly by a single ReLU unit (up to an irrelevant constant/bias parameterization), so for some constant-sized budget $p_0$,

$$\inf\{\mathcal{R}(f) : f \in \mathcal{H},\ \mathrm{params}(f) \leq p_0\} = 0.$$

Thus, at budget $p = p_0$, the optimal student lies outside the teacher space $\mathcal{H}_T$ (hence outside any subset $\mathcal{H}_S \subseteq \mathcal{H}_T$).

In both counterexamples, we exhibit a budget $p$ and an alternative hypothesis class $\mathcal{H}$ (not contained in the teacher class $\mathcal{H}_T$) achieving strictly smaller optimal risk than any student restricted to a subset $\mathcal{H}_S \subseteq \mathcal{H}_T$. This proves that, at a fixed parameter budget, optimal students do not necessarily lie in a subset of the teacher's hypothesis space. □

## C.2 Theorem 2

*Proof of Theorem 2.* For each $k$, choose the input-hidden edge functions as $\phi_{ki}(t) = w_{k,i}t$. Then the $k$-th hidden node computes $w_k^\top x$. Let $p_{k,n}$ be the degree-$(n-1)$ Chebyshev approximant to $g_k$ on $I_k$. By the standard Bernstein-ellipse approximation estimate,

$$\|g_k - p_{k,n}\|_{L^\infty(I_k)} \leq C_\rho M_k \rho^{-n}.$$

Set the $k$-th hidden-output edge to $p_{k,n}$. The resulting ordinary KAN satisfies

$$\|f - \widehat{f}_{\mathrm{KAN}}\|_{L^\infty(D)} \le \sum_{k=1}^{r} \|g_k - p_{k,n}\|_{L^\infty(I_k)} \le C_\rho \sum_{k=1}^{r} M_k \rho^{-n} = C_f \rho^{-n}.$$

The architecture has $rd + r$ univariate edges, each allocated $n$ coefficients, so the allocated coefficient count is $O(rdn)$. Solving $C_f \rho^{-n} \le \varepsilon$ gives $n \gtrsim \log(C_f/\varepsilon)/\log \rho$, and therefore the required allocated coefficient count is $O(rd \log(C_f/\varepsilon))$. □

### C.2.1 Axis-aligned analytic separation from shallow ReLU networks

Barrow & Smith (1978, Theorem 2) prove that, for every fixed $g \in C^2(I)$, best $L^2(I)$ approximation by linear splines with $m$ free knots has asymptotic error

$$\inf_{s \in \mathcal{S}_m} \|g - s\|_{L^2(I)} \sim A_g m^{-2},$$

where $A_g > 0$ whenever $g'' \not\equiv 0$. Independently, Siegel & Xu (2022, Theorem 7) give a fixed-target lower bound for unrestricted shallow ReLU networks.

Consider a fixed non-polynomial analytic profile with $g'' \ge c_0 > 0$ on $I$, such as $g(t) = e^t$, and $f(x) = g(x_1)$ on $I^d$ with the uniform product measure $P_X$. For each fixed $x_{2:d}$, the restriction of an unrestricted width-$m$ one-hidden-layer ReLU network to $x_1$ is a linear spline with at most $m$ knots. Fubini and the cited asymptotic therefore give

$$\inf_{h \in \mathcal{M}_m} \|f - h\|_{L^2(P_X)} = \Omega_g(m^{-2}).$$

For the same fixed target, the Chebyshev construction above gives a depth-2 KAN approximation $O_g(\rho_g^{-n})$, with $\rho_g > 1$. At fixed dimension, the allocated parameter counts grow linearly with $m$ and $n$, respectively. Thus at matched parameter budget $P$,

$$\inf_{h \in \mathcal{M}_P} \|f - h\|_{L^2(P_X)} = \Omega_g(P^{-2}), \qquad \inf_{h \in \mathcal{K}_P} \|f - h\|_{L^2(P_X)} = O_g(e^{-\gamma_g P}).$$

This establishes the stated axis-aligned analytic subclass separation. The comparison for arbitrary ridge directions with pairwise interactions remains conjectural.

### C.3 Theorem 3

**Proof sketch.** We view the layer as a collection of edge functions $\{\phi_{ij}\}_{i \le m_{\mathrm{in}},\, j \le m_{\mathrm{out}}}$ with $\phi_{ij} \in L_2(\nu)$ and expansions $\phi_{ij}(t) = \sum_{\ell \ge 1} a_{ij,\ell}\, \psi_\ell(t)$. Because $\{\psi_\ell\}$ is an orthonormal basis of $L_2(\nu)$, Parseval's identity gives

$$\|\phi_{ij}\|_{L_2(\nu)}^2 = \sum_{\ell \ge 1} |a_{ij,\ell}|^2,$$

$$\|\phi_{ij} - \widetilde{\phi}_{ij}\|_{L_2(\nu)}^2 = \sum_{\ell \ge 1} |a_{ij,\ell} - \widetilde{a}_{ij,\ell}|^2.$$

We then stack all coefficients $\{a_{ij,\ell}\}$ into a single coefficient object $a$ living in the Hilbert space obtained by taking the square-summable direct sum over $(i, j)$ of $\ell_2$ sequences. Under the layer norm induced by $L_2(\nu)$ (the sum of the $L_2(\nu)$ energies across edges), the identification $\Phi \leftrightarrow a$ is an isometry:

$$\begin{aligned} \|\Phi - \widetilde{\Phi}\|^2 &= \sum_{i,j} \|\phi_{ij} - \widetilde{\phi}_{ij}\|_{L_2(\nu)}^2 \\ &= \sum_{i,j} \sum_{\ell \ge 1} |a_{ij,\ell} - \widetilde{a}_{ij,\ell}|^2 = \|a - \widetilde{a}\|^2. \end{aligned}$$

The compressed class $\mathfrak{C}(\mathcal{S})$ corresponds exactly to restricting the coefficient object to lie in a closed linear subspace $\mathcal{S}$ of this Hilbert space. Therefore the approximation problem

$$\min_{\widetilde{\Phi} \in \mathfrak{C}(\mathcal{S})} \|\Phi - \widetilde{\Phi}\|^2$$

is *identical to the Hilbert-space least-squares problem* $\min_{\widetilde{a}\in\mathcal{S}} \|a-\widetilde{a}\|^2$. By the projection theorem in Hilbert spaces, there exists a unique minimizer $\widetilde{a}^\star \in \mathcal{S}$, namely the orthogonal projection $\widetilde{a}^\star = \Pi_{\mathcal{S}} a$, characterized by the orthogonality condition

$$a - \widetilde{a}^\star \;\perp\; \mathcal{S}.$$

Mapping back through the coefficient expansion defines the unique best compressed layer $\widetilde{\Phi}^\star \in \mathfrak{C}(\mathcal{S})$ with coefficients $\widetilde{a}^\star$, and we have $\widetilde{\Phi}^\star = \Pi_{\mathcal{S}}\Phi$ in the sense of this identification.

Finally, for any other $\widetilde{\Phi} \in \mathfrak{C}(\mathcal{S})$ with coefficients $\widetilde{a} \in \mathcal{S}$, we write

$$a - \widetilde{a} = (a - \widetilde{a}^\star) + (\widetilde{a}^\star - \widetilde{a}).$$

Here $a - \widetilde{a}^\star$ is orthogonal to $\mathcal{S}$, while $\widetilde{a}^\star - \widetilde{a} \in \mathcal{S}$, so the two terms are orthogonal. Taking squared norms yields the Pythagorean identity

$$\|a - \widetilde{a}\|^2 = \|a - \widetilde{a}^\star\|^2 + \|\widetilde{a}^\star - \widetilde{a}\|^2.$$

Using again the isometry between coefficient distance and layer distance, this becomes

$$\|\Phi - \widetilde{\Phi}\|^2 = \|\Phi - \widetilde{\Phi}^\star\|^2 + \|\widetilde{\Phi}^\star - \widetilde{\Phi}\|^2, \qquad \forall\, \widetilde{\Phi} \in \mathfrak{C}(\mathcal{S}),$$

which is the desired certificate. □

### C.4 Proposed Barron-type extension

We currently analyze finite ridge sums $f(x) = \sum_{k=1}^{r} g_k(w_k^\top x)$ with analytic ridge profiles. A concrete extension is to replace the finite ridge count $r$ by a variation norm over a specified class $\mathcal{A}$ of normalized ridge atoms. For

$$\|f\|_{\mathcal{B}} := \inf_{\mu:\, f=\int \phi_\theta \, d\mu(\theta)} \|\mu\|_{\mathrm{TV}},$$

standard finite-width arguments motivate estimates of the form

$$\inf_{f_m \in \mathrm{span}_m(\mathcal{A})} \|f - f_m\|_{L^2(P_X)} \le C\|f\|_{\mathcal{B}} m^{-1/2},$$

where $\mathrm{span}_m(\mathcal{A})$ denotes sums of at most $m$ ridge atoms. Combining this width term with our univariate basis-approximation term would separate the error due to using finitely many ridge atoms from the error due to finite edge-basis resolution. Establishing the constants and the required domain or tail assumptions under $P_X$ is the concrete next theoretical step.

# D Algorithms

Basis lifting can make KAN training ill-conditioned (Arslan et al., 2026); Algorithms 1–3 use ridge regression for their coefficient updates.

**Require:** Consecutive KAN layers $\Phi_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^M$ and $\Phi_{\ell+1} : \mathbb{R}^M \to \mathbb{R}^{d_{\ell+2}}$, basis features $\psi(t) \in \mathbb{R}^Q$ (fixed basis, incl. range map), calibration set $\mathcal{D} = \{x^{(k)}\}_{k=1}^N$, ridge $\lambda \geq 0$.

**Ensure:** Merged KAN layer $\widetilde{\Phi}_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^{d_{\ell+2}}$ replacing $\Phi_{\ell+1} \circ \Phi_\ell$, and empirical residual $\mathrm{Cert}_{\text{depth}}$.

1: **Forward to layer inputs.** For each $x^{(k)} \in \mathcal{D}$, compute $z^{(k)} = \Phi_{\ell-1} \circ \cdots \circ \Phi_1(x^{(k)}) \in \mathbb{R}^{d_\ell}$.

2: **Targets.** Set $y^{(k)} \leftarrow (\Phi_{\ell+1} \circ \Phi_\ell)(z^{(k)}) \in \mathbb{R}^{d_{\ell+2}}$.

3: **Design matrix.** For each $k$, form $\phi(z^{(k)}) = [\psi(z_1^{(k)})^\top, \ldots, \psi(z_{d_\ell}^{(k)})^\top]^\top \in \mathbb{R}^{d_\ell Q}$. Let $\Phi \in \mathbb{R}^{N \times (d_\ell Q)}$ have rows $\phi(z^{(k)})^\top$, and let $Y \in \mathbb{R}^{N \times d_{\ell+2}}$ have rows $y^{(k)\top}$.

4: **Center.** $\Phi_c \leftarrow \Phi - \mathbf{1}\bar{\Phi}, \quad Y_c \leftarrow Y - \mathbf{1}\bar{Y}$, where $\bar{\Phi} = \frac{1}{N}\mathbf{1}^\top \Phi$ and $\bar{Y} = \frac{1}{N}\mathbf{1}^\top Y$.

5: **Ridge projection solve.**

$$W \leftarrow (\Phi_c^\top \Phi_c + \lambda I)^\dagger \Phi_c^\top Y_c,$$
$$\beta \leftarrow \bar{Y} - \bar{\Phi} W.$$

6: **Instantiate merged layer.** Reshape $W^\top$ into coefficients $\widetilde{A} \in \mathbb{R}^{d_{\ell+2} \times d_\ell \times Q}$ and set $\widetilde{\Phi}_\ell(z) = \beta + \langle \widetilde{A}, \psi(z) \rangle$.

7: **Empirical reconstruction residual.**

$$\mathrm{Cert}_{\text{depth}} \leftarrow \frac{1}{N} \sum_{k=1}^N \left\| y^{(k)} - \widetilde{\Phi}_\ell(z^{(k)}) \right\|_2^2.$$

8: **Return:** $\widetilde{\Phi}_\ell,\ \mathrm{Cert}_{\text{depth}}$.

**Algorithm 1:** KAN Layer Pruning via Two-Layer Projection Merge

**Require:** Adjacent KAN layers $\Phi_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^M$ and $\Phi_{\ell+1} : \mathbb{R}^M \to \mathbb{R}^{d_{\ell+1}}$, basis features $\psi(t) \in \mathbb{R}^Q$ (fixed basis), calibration set $\mathcal{D} = \{x^{(k)}\}_{k=1}^N$, target width $r \leq M$, ridge $\gamma \geq 0$.

**Ensure:** Pruned layers $(\Phi'_\ell, \Phi'_{\ell+1})$ of width $r$ and certificate Cert.

1: **Forward to get scalars.** For each $x^{(k)} \in \mathcal{D}$, compute $z^{(k)} = \Phi_{\ell-1} \circ \cdots \circ \Phi_1(x^{(k)})$ and $h^{(k)} = \Phi_\ell(z^{(k)})$.

2: **Empirical scalar measures.** $S_1 = \{z_i^{(k)} :\ k \in [N],\ i \in [d_\ell]\}, \quad S_2 = \{h_m^{(k)} :\ k \in [N],\ m \in [M]\}$.

3: **Empirical Gram matrices.**

$$\hat{G}_1 \leftarrow \frac{1}{|S_1|} \sum_{t \in S_1} \psi(t)\psi(t)^\top + \gamma I,$$
$$\hat{G}_2 \leftarrow \frac{1}{|S_2|} \sum_{t \in S_2} \psi(t)\psi(t)^\top + \gamma I.$$

4: **Per-unit coefficient energies.** For each hidden unit $m \in [M]$, set

$$E_m \leftarrow \sum_{i=1}^{d_\ell} (a_{im}^{(\ell)})^\top \hat{G}_1\, a_{im}^{(\ell)} + \sum_{j=1}^{d_{\ell+1}} (a_{mj}^{(\ell+1)})^\top \hat{G}_2\, a_{mj}^{(\ell+1)}.$$

5: **Select units.** $J^* \leftarrow \text{Top-}r\{E_m\}_{m=1}^M$ (indices of the $r$ largest energies).

6: **Prune consistently across layers.** $\Phi'_\ell \leftarrow$ keep only outputs $m \in J^*$ (delete other hidden units);; $\Phi'_{\ell+1} \leftarrow$ keep only inputs $m \in J^*$ (delete corresponding rows).

7: **Coefficient-space certificate (tail energy).** $\mathrm{Cert} \leftarrow \sum_{m \notin J^*} E_m$.

8: **Post-prune refit.** With $J^*$ fixed, refit $\Phi'_{\ell+1}$ by ridge least-squares on $\mathcal{D}$ . The tail certificate above refers to the selection step.

9: **Return:** $(\Phi'_\ell, \Phi'_{\ell+1}),\ \mathrm{Cert}$.

**Algorithm 2:** Certified KAN-Width Pruning via Group-Selection Projection

**Require:** Teacher $g$ (depth $L$), KAN layers $\{\Phi_k : \mathbb{R}^{d_k} \to \mathbb{R}^{d_{k+1}}\}_{k=0}^{L-1}$, calibration set $\mathcal{D}_c = \{x^{(n)}\}_{n=1}^{N}$, penalty $\rho > 0$, anchor weights $\{\alpha_k\}_{k=1}^{L}$, sweeps $T$.

**Require:** Per-layer effective ridge parameters $\{r_k\}_{k=0}^{L-1}$ (used in $\Phi$-step) and initialization ridge parameters $\{r_k^{(0)}\}_{k=0}^{L-1}$.

**Ensure:** Transferred KAN $\Phi = \Phi_{L-1} \circ \cdots \circ \Phi_0$.

**Teacher states.** Forward-pass $g$ on $\mathcal{D}_c$ to obtain stacked activations $\{H_k^T\}_{k=0}^{L}$, where $H_k^T \in \mathbb{R}^{N \times d_k}$.

**Auxiliary states.** Initialize $U_0 \leftarrow H_0^T$ and $U_k \leftarrow H_k^T$ for $k = 1, \ldots, L$.

**Initialization (ridge fit on teacher states).**

**for** $k = 0$ to $L - 1$ **do**

Fit $\Phi_k$ by multi-output ridge regression:

$$\Phi_k \leftarrow \arg\min_{\Phi \in \mathcal{C}_k} \frac{1}{N} \left\| \Phi(U_k) - U_{k+1} \right\|_F^2 \; + \; r_k^{(0)} \, \|\theta\|_2^2.$$

**end for**

**for** $t = 1$ to $T$ **do**

$\Phi$**-step (projection update).**

**for** $k = 0$ to $L - 1$ **do**

Refit $\Phi_k$ by ridge on current auxiliary pairs:

$$\Phi_k \leftarrow \arg\min_{\Phi \in \mathcal{C}_k} \frac{1}{N} \left\| \Phi(U_k) - U_{k+1} \right\|_F^2 \; + \; r_k \, \|\theta\|_2^2.$$

**end for**

$U$**-step (latent feature update).**

Set $U_0 \leftarrow H_0^T$.

**for** $k = 0$ to $L - 1$ **do**

$\widehat{U}_{k+1} \leftarrow \Phi_k(U_k)$

$$U_{k+1} \leftarrow \frac{\rho \, \widehat{U}_{k+1} + \alpha_{k+1} \, H_{k+1}^T}{\rho + \alpha_{k+1}}$$

**end for**

**end for**

**Return:** $\{\Phi_k\}_{k=0}^{L-1}$.

**Algorithm 3:** MLP→KAN Transfer via Alternating Ridge Fit and Anchor Projection

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state the target-function scope. Sections 3.2–3.3 distinguish layer-level projection guarantees from empirical transfer results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 states the approximation assumptions and in-domain scope. Appendix B.3 describes the certificate diagnostic.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Section 3 states the assumptions for Theorems 1–3, and Appendix C gives the proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Algorithms 1–3 specify the compression procedures. Sections 3–4 and Appendix B describe the evaluations; experiment configurations and reproduction instructions accompany the code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and reproduction instructions: `https://github.com/kaptres/kan_compression`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 3–4 and Appendix B describe the architectures, calibration sizes, compression settings and comparisons. Experiment configurations accompany the code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The transfer figures report standard deviations over five seeds. The ResNet and certificate tables report averages over three seeds.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiments report accuracy and parameter counts.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: After conclusions.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: –

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: –

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The implementation, experiment configurations and reproduction commands are documented in the repository.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: –

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: –

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: –

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.