Be Greedy, Stay Linear: Universally Robust Feature Engineering

Yunze Leng[1] YLENG@U.NUS.EDU
Rohan Ghosh RGHOSH92@GMAIL.COM
Mehul Motani[1] MOTANI@NUS.EDU.SG

[1] *Department of Electrical and Computer Engineering, National University of Singapore*

Abstract

Automated feature engineering (FE) often fails a critical test: strictly improving heavily optimized, state-of-the-art (SOTA) predictors. We show that existing FE methods, while boosting untuned baselines, very often make densely tuned GBDT and Neural Networks (NNs) perform worse. We attribute this to a common design paradigm: modern FE methods rely on complex proxy models under the assumption that sophisticated downstream learners require equally complex selection mechanisms. We show that this is counterproductive from two theoretical standpoints. From a feature selection perspective, we demonstrate that complex selection proxies can significantly inflate a feature's apparent benefit, making the process highly susceptible to spurious correlations. From a performance perspective, we show that complex selection proxies yield significantly higher upper bounds on Universality Regret: a metric capturing the expected performance gap between downstream models trained on selected versus optimal feature sets. Together, these two observations highlight the importance of restricting the selection model complexity. To that end, we introduce LinFE, a greedy symbolic feature construction method that enforces a strictly linear selector, constructing the feature set iteratively using the Gram-Schmidt optimized form of the classical Orthogonal Least Squares (OLS) approach. LinFE essentially represents a modern revival of OLS, fine-tuned with (a) evolutionary elements (for faster search), (b) optimized categorical variable representation using their error residuals, and (c) a patience-based cross-validation stopping criterion that is ubiquitous in neural network training. Via a robust evaluation framework on 20 OpenML benchmarks, we demonstrate that while competing FE methods fail to outperform optimized raw baselines in most cases, LinFE consistently elevates the performance of tuned LightGBM, CatBoost, Random Forests, and NNs. Our results demonstrate that for tabular data regression, restricting selector capacity is not a limitation, but a vital step to ensuring the learning of universally robust features.

1. Introduction

Automated feature engineering (FE) methodologies suffer from a critical evaluation blind spot: they are rarely benchmarked against densely hyperparameter optimized predictors using methods such as Optuna [1, 6, 8, 15]). Evaluating on untuned baselines can artificially inflate perceived feature value, as performance gains can result from compensating for unoptimized learning dynamics or inadequate capacity. Furthermore, genuinely effective features should exhibit *universal robustness*, yielding improvements across a diverse array of downstream model architectures, rather than being focused towards a single type.

While current state-of-the-art (SOTA) approaches (e.g., OpenFE [15]) utilize complex selection models under the assumption that complex downstream learners require equally complex selectors, our theoretical analysis proves otherwise. We demonstrate that, counter-intuitively, employing strictly

© Y. Leng[1], R. Ghosh & M. Motani[1].

simpler selection models yields features with superior generalization and cross-architecture robustness. Beyond statistical generalization, linear selectors offer a profound computational advantage. Non-linear selectors rely on noisy gradient heuristics to estimate a candidate feature's utility. In contrast, linear selectors allow for the *exact* computation of marginal loss reduction over the training set using Orthogonal Least Squares (OLS [4]), where a candidate's gain is precisely the normalized projection of its Gram-Schmidt orthogonalized form onto the current residual.

Although OLS is primarily a feature selection approach, it has been combined with suitable feature generation (e.g. symbolic features in [13]) to yield feature generation approaches. Building on this, we introduce LINFE, a framework that further modernizes OLS based feature generation via three key ideas:

1. **Evolutionary Symbolic Search:** Rather than relying on fixed or predefined basis expansions, we employ a genetic programming approach to dynamically synthesize and refine non-linear features, effectively navigating a large search space.

2. **Residual-Based Categorical Encoding:** To prevent the significant dimensional increase and input redundancy caused by one-hot encoding categorical variables, we efficiently replace each categorical element with their average initial-stage error residual.

3. **Generalization-Aware Stopping:** Moving beyond heuristic training-error thresholds, we implement a validation-set based stopping criterion with patience. This strictly regulates feature complexity, ensuring that only features contributing to robust generalization, rather than those exploiting training noise, are preserved.

Finally, we empirically answer whether FE can still improve heavily optimized models. Across 20 OpenML regression benchmarks, we find that surprisingly, established baselines frequently fail to outperform raw features once downstream learners are densely tuned. In stark contrast, LINFE consistently improves predictive performance across diverse, optimized families (Neural Networks, CatBoost, LightGBM, and Random Forests). By achieving SOTA aggregate performance, LINFE demonstrates that a simple, exact linear selector engineers universally robust features capable of elevating even the most heavily optimized predictors.

Related Work. We present a more rigorous discussion of related work in Appendix A.

2. Theoretical Motivation for Linear Selectors

2.1. Feature Gain Accuracy in a General Functional OLS Setting

Evaluating a candidate feature $x \in \mathbb{R}$ for a target y often relies on the reduction in training mean squared error (MSE) achieved by introducing x into a selection model, termed *feature gain*. While measuring this by explicitly retraining complex, non-linear models is computationally prohibitive, standard Ordinary Least Squares (OLS) yields a closed-form expression for this loss reduction [4]. Given an i.i.d. sample $S = \{(x_i, y_i)\}_{i=1}^N$, let $\langle f(x), y\rangle_N = \frac{1}{N}\sum_{i=1}^N f(x_i)y_i$ and $\|f(x)\|_N = \sqrt{\frac{1}{N}\sum_{i=1}^N f(x_i)^2}$ denote the empirical inner product and norm for a transformation $f(x)$, with population counterparts $\mathbb{E}[f(x)y]$ and $\sqrt{\mathbb{E}[f(x)^2]}$. At the first step of feature addition in linear OLS, the exact MSE decrease simplifies to the ratio $\frac{\langle x,y\rangle_N}{\|x\|_N}$.

To generalize this to non-linear selectors, we consider a general **functional OLS** setting where the model is parameterized as $ah(x) + b$ for constants a, b and a functional hypothesis $h \in \mathcal{H}$. Here,

the optimal loss reduction from the first feature is $\sup_{h\in\mathcal{H}} \frac{\langle h(x),y\rangle_N}{\|h(x)\|_N}$, as the model actively finds the feature $h(x)$ that maximizes the feature gain. We bound the deviation between this empirical estimate and its true population expectation under two mild structural assumptions.

Definition 1 (Scale-Invariant Hypothesis Space) *We assume the underlying hypothesis space $\mathcal{H}$ is scale-invariant (closed under scaling). That is, for any function $h \in \mathcal{H}$ and any scalar $\alpha \in \mathbb{R}$, the scaled function $\alpha h \in \mathcal{H}$.*

Assumption 2 (Uniform Boundedness of the Normalized Subclass) *There exists a constant $M > 0$ such that every function in the normalized subclass $\mathcal{H}_1 \subset \mathcal{H}$, defined as $\mathcal{H}_1 = \left\{h \in \mathcal{H} : \mathbb{E}[h(x)^2] = 1\right\}$, is uniformly bounded: $\sup_x |h(x)| \leq M \quad \forall h \in \mathcal{H}_1$.*

Theorem 3 (Uniform Convergence of the Feature Gain Estimate) *Let $\delta \in (0,1)$, and let $\mathcal{R}_N(\mathcal{H}_1)$ denote the empirical Rademacher complexity of the normalized subclass. Assume the sample size N is sufficiently large such that $4M\mathcal{R}_N(\mathcal{H}_1) + M^2\sqrt{\frac{2\ln(4/\delta)}{N}} \leq 0.5$. Then, with probability at least $1-\delta$ over the random draw of the sample S, the following bound holds uniformly:*

$$\left|\sup_{h\in\mathcal{H}} \frac{\langle h(x),y\rangle_N}{\|h(x)\|_N} - \sup_{h\in\mathcal{H}} \frac{\mathbb{E}[h(x)y]}{\sqrt{\mathbb{E}[h(x)^2]}}\right| \leq \sqrt{2}\left(2(2M+1)\mathcal{R}_N(\mathcal{H}_1) + M(M+1)\sqrt{\frac{2\ln(4/\delta)}{N}}\right). \tag{1}$$

Remark 4 (The Linear Advantage and Rademacher Collapse) *Theorem 3 explicitly isolates the statistical penalty incurred by the capacity of the feature selection model. When evaluating raw features directly via a linear selector (as in OLS), the hypothesis space effectively reduces to a single fixed feature evaluation. Consequently, the empirical Rademacher complexity collapses to zero ($\mathcal{R}_N(\mathcal{H}_1) = 0$). Provided N is sufficiently large such that $M^2\sqrt{\frac{2\ln(4/\delta)}{N}} \leq 0.5$, the uniform bound simplifies to a pure concentration term: $\left|\frac{\langle x,y\rangle_N}{\|x\|_N} - \frac{\mathbb{E}[xy]}{\sqrt{\mathbb{E}[x^2]}}\right| \leq 2M(M+1)\sqrt{\frac{\ln(4/\delta)}{N}}$. This establishes a stark structural advantage: by avoiding the capacity overhead associated with searching over complex non-linear transformations, linear feature selectors guarantee that empirical estimates of feature importance converge to their true population counterparts significantly faster.*

2.2. Performance Analysis via Universality Regret

Our earlier analysis primarily touched upon how the accuracy of a feature's true value w.r.t the current model decreases significantly as the complexity of the model increases. While this informs us that the feature sets output by complex selectors may have a significantly inflated degree of value, it does not directly indicate how this will affect final generalization performance when pairing these features with downstream classifiers. To address this we present an analysis where we study the *Universality Regret* of a FE output, which is formally defined as follows.

Definition 5 (True and Empirical Functional Gain) *Given a baseline feature set B, a candidate feature $g \in \mathcal{G}$, and a hypothesis space $\mathcal{H}$, the true functional gain $\Delta(g,\mathcal{H})$ is the reduction in true risk achieved by adding g:$\Delta(g,\mathcal{H}) = \min_{h\in\mathcal{H}} \mathbb{E}_P[\ell(Y,h(B))] - \min_{h\in\mathcal{H}} \mathbb{E}_P[\ell(Y,h(B\cup\{g\}))]$. The empirical functional gain $\widehat{\Delta}(g,\mathcal{H},S_n)$ is its corresponding empirical estimate evaluated over a finite sample S_n of size n:$\widehat{\Delta}(g,\mathcal{H},S_n) = \min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(Y_i,h(B_i)) - \min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(Y_i,h(B_i\cup\{g_i\}))$.*

Definition 6 (Universality Regret) *Let $\mathcal{H}_F$ be a model class used for feature selection, and let $\mathcal{H}_D$ be a downstream model class drawn from a distribution $\mathcal{D}$. Let $g_D^* = \arg\max_{g\in\mathcal{G}} \Delta(g, \mathcal{H}_D)$ be the globally optimal feature for $\mathcal{H}_D$, and let $\hat{g}_F = \arg\max_{g\in\mathcal{G}} \widehat{\Delta}(g, \mathcal{H}_F, S_n)$ be the empirically chosen feature via the proxy. The Universality Regret $\mathcal{R}_{univ}(\mathcal{H}_F)$ is the expected performance suboptimality downstream caused by selecting features via the proxy space:$\mathcal{R}_{univ}(\mathcal{H}_F) = \mathbb{E}_{\mathcal{H}_D\sim\mathcal{D}}\left[\Delta(g_D^*, \mathcal{H}_D) - \Delta(\hat{g}_F, \mathcal{H}_D)\right]$.*

Definition 7 (Functional Hausdorff-Lipschitz Regularity) *We assume the existence of a data, loss and generator class $\mathcal{G}$ dependent regularity constant $L_H > 0$ such that the maximum functional gain mismatch over $\mathcal{G}$ is continuously bounded by the distribution-dependent geometric separation of the hypothesis spaces. Formally: $\text{dist}_{\mathcal{G}}(\mathcal{H}_1, \mathcal{H}_2) = \sup_{g\in\mathcal{G}} |\Delta(g, \mathcal{H}_1) - \Delta(g, \mathcal{H}_2)| \leq L_H \cdot d_{H,P}(\mathcal{H}_1, \mathcal{H}_2)$, where $d_{H,P}(\mathcal{H}_1, \mathcal{H}_2)$ denotes the standard Hausdorff distance between the hypothesis spaces: $d_{H,P}(\mathcal{H}_1, \mathcal{H}_2) = \max\left\{\sup_{f\in\mathcal{H}_1} \inf_{h\in\mathcal{H}_2} \|f-h\|_{L^1(P)},\ \sup_{h\in\mathcal{H}_2} \inf_{f\in\mathcal{H}_1} \|f-h\|_{L^1(P)}\right\}$, where $\|f-h\|_{L^1(P)} = \mathbb{E}_P[|f(X) - h(X)|]$.*

Lemma 8 (Feasibility and Existence of the Hausdorff-Lipschitz Bound) *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be compact hypothesis spaces under the $L^1(P)$ norm over a domain $\mathcal{X}$. Assuming the loss function ℓ is L-Lipschitz continuous, the following holds: $\sup_{g\in\mathcal{G}} |\Delta(g, \mathcal{H}_1) - \Delta(g, \mathcal{H}_2)| \leq 2L \cdot d_{H,P}(\mathcal{H}_1, \mathcal{H}_2)$ thereby establishing that the Functional Hausdorff-Lipschitz Regularity condition is strictly feasible, with the true data-dependent constant satisfying $L_H \leq 2L$.*

Theorem 9 (Universality Regret and Selector Complexity) *Let $\mathcal{G}$ be a symbolic feature space and $\mathcal{D}$ a distribution over downstream model classes $\mathcal{H}_D$. Let $\mathcal{H}_F$ be the feature selection model's hypothesis space. Under the Functional Hausdorff-Lipschitz Regularity assumption, for a C-bounded loss ℓ, with probability at least $1-\delta$ over the choice of S_n, the Universality Regret is bounded by:*

$$\mathcal{R}_{univ}(\mathcal{H}_F) \leq 4\mathfrak{R}_n(\ell \circ \mathcal{H}_F \circ \mathcal{G}) + 2L_H \cdot \mathbb{E}_{\mathcal{H}_D\sim\mathcal{D}}\left[d_{H,P}(\mathcal{H}_F, \mathcal{H}_D)\right] + 2C\sqrt{\frac{\log(1/\delta)}{2n}}. \quad (2)$$

Remark 10 (The Complexity-Diversity Tradeoff) *Theorem 9 formalizes a fundamental tension when selecting features via $\mathcal{H}_F$. While the complexity term $\mathfrak{R}_n(\ell\circ\mathcal{H}_F\circ\mathcal{G})$ grows aggressively as $\mathcal{H}_F$ scales towards deep neural networks, the **expected inductive mismatch** $\mathbb{E}_{\mathcal{H}_D\sim\mathcal{D}}\left[d_{H,P}(\mathcal{H}_F, \mathcal{H}_D)\right]$ is more nuanced. When $\mathcal{D}$ consists of fundamentally distinct architectural families (e.g., mixing tree ensembles and neural networks), the expected inductive mismatch should have less variation across choice of H_P. It is well known that NNs struggle to efficiently approximate the jagged, axis-aligned step functions characteristic of tree ensembles (see, e.g., [5]). Because these cross-family functional disparities are fundamentally unresolvable when $\mathcal{D}$ is diverse, the expected inductive mismatch should see less variation across choice of the selector $\mathcal{H}_F$. This highlights overall that choosing low complexity (e.g., linear) $\mathcal{H}_F$ can yield potentially much lower Universality Regret when the evaluation is done across a diverse set of downstream classes.*

3. LINFE: Methods

We outline the core algorithmic structure of LINFE below. For comprehensive theoretical derivations, detailed pseudo-code (Algorithms 1–3) and standard OLS perspectives please see Appendix B.

LINFE Algorithm Summary:

1. **Categorical Encoding:** To unify heterogeneous features, each discrete categorical level is mapped to a static scalar computed as the mean of the initial residual $\mathbf{r}_0$ over the specific training support where that level is active (Appendix B.4).
2. **Evolutionary Symbolic Search:** At iteration t, genetic programming generates a population of tree-structured candidate expressions $\mathcal{G}_t = \{g_i\}$ over a fixed, domain-agnostic mathematical grammar (Appendix B.3).
3. **Incremental Orthogonalization:** Following OLS, for each candidate feature vector $\mathbf{z}_g = g(\mathbf{X}) \in \mathbb{R}^n$, we isolate its novel predictive component $\tilde{\mathbf{z}}_g = (\mathbf{I} - \mathbf{P}_t)\mathbf{z}_g$. To avoid explicit matrix inversions, this projection is efficiently tracked via an incrementally updated orthonormal basis $\mathcal{Q}_t$, reducing the orthogonalization complexity to $\mathcal{O}(n \cdot m_t)$ per candidate (Appendix B.1).
4. **Residual-Gain Scoring:** Candidate features are ranked by their exact finite-sample reduction in mean squared error relative to the current target residual $\mathbf{r}_t$ (from classical OLS):

$$\widehat{\Delta}_t(g) = \frac{1}{n} \frac{(\mathbf{r}_t^\top \tilde{\mathbf{z}}_g)^2}{\tilde{\mathbf{z}}_g^\top \tilde{\mathbf{z}}_g + \varepsilon},$$

   where $\varepsilon \to 0^+$ handles redundant or linearly dependent feature trajectories (i.e. $\tilde{\mathbf{z}}_g^\top \tilde{\mathbf{z}}_g = 0$).
5. **Feature Set Update & Early Stopping:** The highest-gain feature is added to the generated feature set, the target residual $\mathbf{r}_t$ is updated, and the cycle recurses (from Step 2). If validation RMSE (also using a linear model) fails to improve for a pre-specified *patience* threshold of successive feature additions, execution terminates.

4. Experiments

4.1. Densely Optimized Downstream Learners

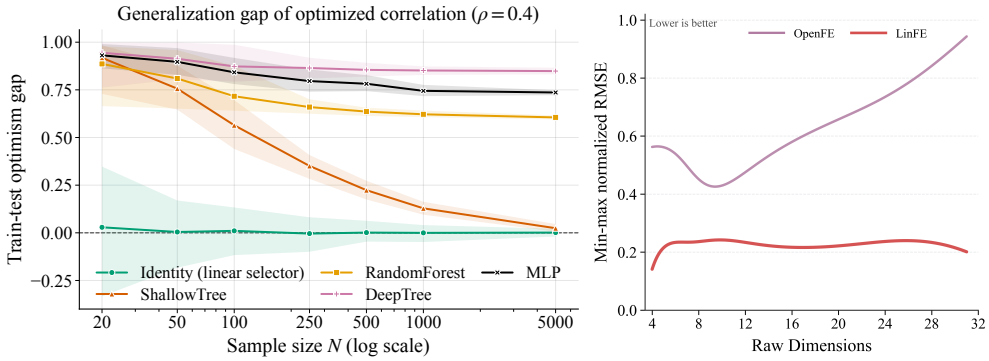
Table 1 presents our central evaluation: testing representations under densely tuned downstream learners via *Optuna*. This rigorous regime inherently challenges automated FE, as optimized GBDTs and NNs should constitute harder baselines to improve via FE. Consequently, we find that competitor methods (AutoFeat, FEAT, MI-obj, OpenFE) frequently degrade performance (<40% win rate). In stark contrast, LINFE outperforms all baselines and is the only method that consistently beats the densely tuned raw data. By strictly using an incremental linear selector, LINFE achieves a dominant 60%–85% win rate across all downstream families, delivering substantial mean improvements of +6.7% for RF and +7.0% for NNs. These gains over thoroughly optimized baselines prove that LINFE does not merely rescue weak models; it can generate structurally unique features that can elevate even densely optimized learners.

4.2. Exploring the Robustness of LINFE

We analyze LINFE's robustness in two ways: (a) evaluating the *optimism gap* for various selectors in Section 2.1's setting, and (b) comparing LINFE and OpenFE's average RMSE across varying dimensionalities. To validate Theorem 3's structural advantage, our first experiment measures the generalization of the empirical feature gain $\sup_{h \in \mathcal{H}} \frac{\langle h(x), y \rangle_N}{\|h(x)\|_N}$. Generating Gaussian covariates $X =$

Table 1: **Performance on densely tuned (Optuna) downstream learners across 20 OpenML datasets.** Raw (−) denotes the original dataset without any FE. ↓ indicates that lower values are better. Rank represents the method's average relative standing across all datasets. Norm. stands for min–max normalized RMSE. All values are dataset-level averages over seeds with 95% confidence intervals. Best results are **bolded**, second-best are underlined.

Method	CatBoost			LightGBM			Random Forest			NNs		
Metric	Rank ↓	RMSE ↓	Norm. ↓	Rank ↓	RMSE ↓	Norm. ↓	Rank ↓	RMSE ↓	Norm. ↓	Rank ↓	RMSE ↓	Norm. ↓
Raw (−)	2.750 ± 0.485	4.462 ± 4.453	0.154 ± 0.092	2.850 ± 0.427	4.612 ± 4.602	0.206 ± 0.104	3.400 ± 0.393	5.023 ± 5.122	0.263 ± 0.126	3.150 ± 0.518	4.909 ± 4.905	0.247 ± 0.113
FEAT	5.700 ± 0.405	8.122 ± 9.972	0.939 ± 0.089	5.450 ± 0.674	8.173 ± 9.898	0.868 ± 0.145	5.500 ± 0.612	8.289 ± 9.964	0.883 ± 0.131	5.200 ± 0.645	7.930 ± 9.269	0.826 ± 0.136
AutoFeat	2.800 ± 0.520	4.467 ± 4.473	**0.141 ± 0.092**	3.200 ± 0.435	4.632 ± 4.607	0.255 ± 0.112	3.500 ± 0.421	5.003 ± 5.072	0.267 ± 0.121	3.150 ± 0.591	4.837 ± 4.812	0.253 ± 0.138
MI-obj	4.350 ± 0.686	5.733 ± 6.223	0.485 ± 0.148	4.150 ± 0.686	5.749 ± 6.161	0.453 ± 0.138	3.800 ± 0.774	5.847 ± 6.303	0.391 ± 0.151	4.050 ± 0.732	5.925 ± 6.414	0.474 ± 0.158
OpenFE	2.950 ± 0.745	5.212 ± 5.676	0.228 ± 0.124	3.000 ± 0.752	5.496 ± 6.115	0.273 ± 0.137	2.900 ± 0.737	5.788 ± 6.502	0.250 ± 0.132	3.550 ± 0.628	5.877 ± 6.489	0.379 ± 0.145
LinFE	**2.450 ± 0.438**	**4.449 ± 4.524**	0.143 ± 0.103	**2.350 ± 0.608**	**4.581 ± 4.630**	**0.198 ± 0.130**	**1.900 ± 0.491**	**4.943 ± 5.119**	**0.127 ± 0.104**	**1.900 ± 0.602**	**4.667 ± 4.690**	**0.139 ± 0.114**



(a) Proxy optimism gap ($\rho = 0.4$)

(b) Dimensional scaling

Figure 1: **Robustness analysis of LINFE. (a)** Train-test optimism gap for normalized correlation based feature selection for various selection models. Validating Theorem 3, high-capacity selectors (e.g., Random Forest, MLP) show a significant inflation of true feature value, which is completely avoided by simple linear proxies (Identity). **(b)** Smoothed min–max normalized RMSE across optimized learners, illustrating LINFE's stability compared to OpenFE as dimensionality increases.

$(X_0, Z_1, \ldots, Z_{50})$ and targets $Y = \rho X_0 + \sqrt{1-\rho^2}\epsilon$, we fit selector classes $\mathcal{H}$ of increasing capacity. By tracking the *train-test optimism gap* $\left|\sup_{h\in\mathcal{H}} \frac{\langle h(x), y\rangle_N}{\|h(x)\|_N} - \sup_{h\in\mathcal{H}} \frac{\mathbb{E}[h(x)y]}{\sqrt{\mathbb{E}[h(x)^2]}}\right|$, we confirm high-capacity models exhibit a substantial mismatch between empirical and true feature gains (Figure 1a), whereas linear selectors concentrate reliably.

Next, we compute min-max normalized RMSE across datasets grouped by input dimensionality. Theorem 9 directly links the selector's Rademacher complexity to the expected population RMSE. Because function classes exhibit $\mathcal{O}(\sqrt{\log d})$ or $\mathcal{O}(\sqrt{d})$ complexity guarantees depending on the weight constraints [7, 12], generalization inherently degrades as dimensionality grows. However, because linear classes maintain minimal capacity relative to other alternatives, LINFE should exhibit significantly greater stability under dimensional scaling than OpenFE's GBDT-based selectors. Figure 1b precisely confirms this behavior, further validating the robustness of our approach.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[2] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4):537–550, 1994.

[3] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, 13:27–66, 2012.

[4] Sheng Chen, Stephen A Billings, and Wan Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of control*, 50(5):1873–1896, 1989.

[5] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

[6] Franziska Horn, Robert Pack, and Michael Rieger. The autofeat python library for automated feature engineering and selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 111–120. Springer, 2019.

[7] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

[8] William La Cava, Tilak Raj Singh, James Taggart, Srinivas Suri, and Jason H Moore. Learning concise representations for regression by evolving networks of trees. *arXiv preprint arXiv:1807.00981*, 2018.

[9] Yuanfei Luo, Mengshuo Wang, Hao Zhou, Quanming Yao, Wei-Wei Tu, Yuqiang Chen, Wenyuan Dai, and Qiang Yang. Autocross: Automatic feature crossing for tabular data in real-world applications. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1936–1945, 2019.

[10] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

[11] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[12] Patrick Rebeschini. Algorithmic foundations of learning, lecture 3: Rademacher complexity. Examples. Lecture slides, Department of Statistics, University of Oxford, 2022.

[13] Stéphane JPS Thunus, Julian D Parker, and Stefan Weinzierl. Arborescent orthogonal least squares regression for narmax-based black-box fitting. *IEEE Access*, 12:155578–155597, 2024.

[14] Binh Tran, Bing Xue, and Mengjie Zhang. Genetic programming for feature construction and selection in classification on high-dimensional data. *Memetic Computing*, 8(1):3–15, 2016.

[15] Tianping Zhang, Zheyu Aqa Zhang, Zhiyuan Fan, Haoyan Luo, Fengyuan Liu, Qian Liu, Wei Cao, and Li Jian. Openfe: Automated feature generation with expert-level performance. In *International Conference on Machine Learning*, pages 41880–41901. PMLR, 2023.

Appendix A. Related Work

Automated FE attempts to navigate the intractable search space of feature transformations. Existing methods generally adopt one of four structural paradigms, each facing distinct challenges when paired with highly expressive downstream learners.

Filter-based Selection. Filter methods evaluate candidates through pure statistical association with the target, aiming to balance relevance against redundancy without directly querying a downstream model's loss [2, 3, 11]. While computationally efficient, mutual information (MI) and statistical filters rely on proxy metrics that often fail to align with the actual error surface of the final predictor, leaving complex conditional redundancies unaddressed.

Global Expansion and Sparse Selection. These methods construct a massive predefined pool of nonlinear transformations and subsequently apply strong regularization (e.g., L_1 penalties) to extract a sparse subset. AutoFeat [6] and AutoCross [9] effectively use an interpretable linear bottleneck to filter this globally expanded space. However, generating the global pool scales poorly with feature dimensionality, and simultaneous sparse selection does not dynamically account for the changing residual as features are added one by one.

Evolutionary Symbolic Search. To avoid static global pools, evolutionary methods dynamically navigate the open-ended space of symbolic expressions. By evolving compact networks of mathematical operators, methods like FEAT [8] and [14] represent the SOTA in finding nonlinear representations with explicit accuracy-complexity trade-offs. Despite their flexibility, these methods typically optimize global fitness metrics over the entire feature set concurrently, making it difficult to guarantee that newly mutated features strictly target the unexplained variance of the existing set.

Greedy Residual Approximation. Taking a step toward MP, greedy methods iteratively ask how much a candidate improves the predictions of the *current* representation. OpenFE [15] represents the forefront of this category. Its FeatureBoost procedure estimates incremental loss reduction by fitting a residual model on top of the current feature set. However, because OpenFE tightly couples the selection logic to a complex GBDT, it is forced to use pruning and approximate attribution (e.g., gradient heuristics) to make the search tractable. LinFE diverges fundamentally from this approach. By replacing the complex tree-based selector with a linear geometric projection, LinFE completely avoids gradient approximations, granting every candidate an exact, closed-form local utility score and mathematically enforcing zero gain for redundant transformations.

Appendix B. Extended Methodology and Theoretical Details

B.1. Linear Residual-Gain Scoring

On the construction training split, let $y \in \mathbb{R}^n$ be the target vector. At stage t, let S_t denote the accepted symbolic atoms and define the full current representation as

$$F_t = B \cup Q \cup S_t.$$

Let $Z_t \in \mathbb{R}^{n \times m_t}$ be the design matrix obtained by evaluating all features in F_t on the n construction training samples. LinFE uses an affine OLS selector $A_t = [\mathbf{1}, Z_t], P_t = A_t A_t^+, r_t = (I - P_t)y$, where A_t^+ is the Moore–Penrose pseudoinverse. For a candidate expression g, write $z_g = (g(x_1), \ldots, g(x_n))^\top$ and remove the component already explained by the affine span of the current representation,

$$\tilde{z}_g = (I - P_t) z_g.$$

Computationally, this orthogonalization is realized via an evolving orthonormal basis $\mathcal{Q}_t$ rather than explicit matrix multiplication, as detailed below.

Computational Note & Time Complexity. While the theoretical projection utilizes an $n \times n$ matrix, it is never explicitly instantiated to avoid prohibitive $\mathcal{O}(n^2)$ memory constraints. Instead, mirroring OMP implementations [10], LinFE incrementally maintains an orthonormal basis $\mathcal{Q}_t$ of the current representation space via modified Gram-Schmidt updates. This mathematically exact approach restricts the orthogonal projection of any newly generated candidate vector to strictly $\mathcal{O}(n \cdot m_t)$ operations, where m_t is the number of currently selected features. Consequently, scoring an entire population of C candidates during evolutionary search requires only $\mathcal{O}(C \cdot n \cdot m_t)$ time. This guarantees theoretical scalability for large tabular datasets and fundamentally bypasses the expensive iterative model refitting required by tree-based selectors.

LinFE scores the candidate by the finite-sample residual gain

$$\widehat{\Delta}_t(g) = \frac{1}{n} \frac{(r_t^\top \tilde{z}_g)^2}{\tilde{z}_g^\top \tilde{z}_g + \varepsilon}, \tag{S.1}$$

where $\varepsilon \to 0^+$ is a microscopic numerical stabilizer. This term exists strictly to safely assign zero gain to redundant features (where $\|\tilde{z}_g\|_2 \approx 0$) during floating-point arithmetic, practically preserving the exact geometric projection without introducing approximation heuristics. Thus Eq. (S.1) selects the residualized atom with the largest normalized alignment with the current residual. The score has no complexity penalty and no downstream-model term. Complexity is controlled by the symbolic grammar and by validation stopping.

B.2. LinFE Algorithm

The detailed algorithm of LinFE is given in Algorithms 1–3.

B.3. Symbolic Candidate Search

The candidate class is an implicit grammar of expression trees over numerical terminals, fixed categorical residual-mean terminals, and scalar constants. In our experiments, LinFE uses the fixed

Algorithm 1: LinFE Main Loop

Require: train/val splits, base terminals B, categories $\mathcal{C}$
Require: safe grammar $\mathcal{G}_{\text{safe}}$, patience p, max stages T
Fit initial affine OLS on B, yielding initial residual r_0
$Q \leftarrow \textsc{CategoricalTerminals}(\mathcal{C}, r_0, \lambda)$ ▷ **Alg. 2**
Let $\mathcal{Q}_0$ be the orthonormal basis of $B \cup Q$, and r_0 its residual
$S_0 \leftarrow \emptyset$, $S_{\text{best}} \leftarrow \emptyset$, $m_{\text{best}} \leftarrow \infty$, $k \leftarrow 0$
for $t = 0, \ldots, T-1$ **do**
  $\mathcal{A}_t \leftarrow \textsc{Search}(\mathcal{G}_{\text{safe}}, B \cup Q, r_t)$ ▷ evolution
  $g_t \leftarrow \textsc{GreedyScore}(\mathcal{A}_t, r_t, \mathcal{Q}_t)$ ▷ **Alg. 3**
  $S_{t+1} \leftarrow S_t \cup \{g_t\}$
  $z \leftarrow g_t(X_{\text{tr}})$
  $\tilde{z} \leftarrow z - \sum_{q \in \mathcal{Q}_t} \langle z, q \rangle q$ ▷ orthogonalize
  $q_{t+1} \leftarrow \tilde{z} / \|\tilde{z}\|_2$ ▷ new basis
  $\mathcal{Q}_{t+1} \leftarrow \mathcal{Q}_t \cup \{q_{t+1}\}$
  $r_{t+1} \leftarrow r_t - \langle r_t, q_{t+1} \rangle q_{t+1}$ ▷ $\mathcal{O}(n)$ update
  $m_t \leftarrow$ val RMSE of affine OLS on $B \cup Q \cup S_{t+1}$
  if $m_t < m_{\text{best}} - \tau$ **then**
    $S_{\text{best}} \leftarrow S_{t+1}$, $m_{\text{best}} \leftarrow m_t$, $k \leftarrow 0$
  else
    $k \leftarrow k + 1$; **if** $k \geq p$ **then break**
return exported representation $B \cup Q \cup S_{\text{best}}$

Algorithm 2: Categorical Terminals

Require: categories $\mathcal{C} = \{C_1, \ldots, C_K\}$, residual r_0, smoothing λ
$Q \leftarrow \emptyset$
for all $C \in \mathcal{C}$ **do**
  Let $\mathcal{L}_C$ be the observed levels of C
  for all $c \in \mathcal{L}_C$ **do**
    $I_c \leftarrow \{i : C_i = c\}$
    $\theta_c \leftarrow \sum_{i \in I_c} r_{0,i} / (\lambda + |I_c|)$
  Define $f_C(C_i) = \theta_{C_i}$
  $Q \leftarrow Q \cup \{f_C(C)\}$
return Q

Algorithm 3: Greedy Feature Scoring

Require: candidate pool $\mathcal{A}_t$, residual r_t, basis set $\mathcal{Q}_t$
for all $g \in \mathcal{A}_t$ **do**
  $z_g \leftarrow g(X_{\text{tr}})$
  $\tilde{z}_g \leftarrow z_g - \sum_{q \in \mathcal{Q}_t} \langle z_g, q \rangle q$
  $\widehat{\Delta}_t(g) \leftarrow \dfrac{(r_t^\top \tilde{z}_g)^2}{n(\tilde{z}_g^\top \tilde{z}_g + \varepsilon)}$
return $\arg\max_{g \in \mathcal{A}_t} \widehat{\Delta}_t(g)$

shallow grammar

$$
\begin{aligned}
& e ::= u \mid a \mid \phi(e) \mid \psi(e, e), \\
& u \in B \cup Q, \qquad a \in [-2, 2], \\
& \phi \in \{\sin, \cos, \tanh, |\cdot|, \log(1 + |\cdot|), \sqrt{|\cdot|}, (\cdot)^2, -(\cdot)\}, \\
& \psi \in \{+, -, \times, \text{safeDiv}, \max, \min\}.
\end{aligned}
\tag{S.2}
$$

Here B denotes processed numerical raw terminals, Q denotes fixed categorical residual-mean terminals, and a denotes an ephemeral scalar constant sampled from $[-2, 2]$. The maximum operator depth is fixed to $d = 4$, yielding shallow expressions that reduce search variance, limit overfitting, and keep the selected features inspectable. In implementation, the grammar is safe and typed: invalid expressions producing NaNs, infinities, or near-constant outputs are discarded, and categorical terminals are only used through admissible linear combinations or interactions, rather than arbitrary nonlinear wrappers. We denote the expressions chosen from the symbolic candidate search via S.

Rather than enumerate the grammar, LinFE uses an evolutionary search procedure: it generates a population of valid expressions, screens candidates with a cheap residual-alignment proxy, exactly evaluates a shortlist using Eq. (S.1), and updates the population by selection, mutation, crossover, and elitism. Thus, the symbolic search is heuristic, while the score assigned to every shortlisted candidate is the exact OLS residual gain relative to the current full representation. The search is repeated at every construction stage against the current residual, so the dictionary is accessed adaptively rather than as a fixed global pool.

In Algorithm 1, B denotes the processed non-categorical raw terminals retained throughout construction, and S denotes the accepted symbolic atoms. Categorical terminals Q are statically generated prior to the search using the initial residual r_0.

Initialization of r_0: We compute r_0 by fitting an affine OLS strictly on the raw continuous features B, yielding $r_0 = (I - P_B)y$. This forces the categorical embeddings to capture only the

target variance unexplained by B, mathematically eliminating redundancy between the discrete and continuous bases from the start. For purely categorical datasets ($B = \emptyset$), the initial OLS reduces to an intercept-only model, naturally simplifying r_0 to the mean-centered target, $y - \bar{y}$.

B.4. Categorical Variables as Symbolic Terminals

Algorithm 2 implements the categorical mapping strategy. Rather than expanding a categorical variable into many independent one-hot terminals, we construct a single depth-one tree over that variable prior to the symbolic search. Let r_0 be the initial residual vector obtained by fitting an affine OLS on the continuous base features B. If C has levels $\mathcal{L}_C$, the fitted terminal is

$$f_C(c) = \theta_c, \qquad \theta_c = \frac{\sum_{i:C_i=c} r_{0,i}}{\lambda + |\{i : C_i = c\}|}. \tag{S.3}$$

To prevent target leakage and over-reliance on rare categorical levels, we incorporate standard additive smoothing (e.g., Laplace smoothing) directly into the terminal definition. $\lambda > 0$ is a regularization parameter that shrinks the representation of low-frequency categories toward zero (the expected residual prior).

Consequently, K categorical variables contribute exactly K static functional variables to form the set Q. Crucially, this mapping is performed exactly once before the greedy construction begins. During the main LinFE loop, the set Q acts as a fixed set of continuous terminals alongside B.

To maintain semantic validity and interpretability, LinFE enforces a *safe grammar* for these categorical terminals. While numerical features may undergo arbitrary transformations, the grammar restricts $f_C(C)$ to linear combinations or pairwise interactions (e.g., cross-features like $x_j f_C(C)$ or $f_{C_i}(C_i) f_{C_j}(C_j)$). Arbitrary nonlinear wrappers around categorical means, such as $\sin(f_C(C))$, are strictly prohibited. This ensures that the constructed features logically represent category-specific slopes or scaling factors.

B.5. Stopping and Downstream Export

LinFE uses a linear model twice during construction. First as the residual-gain selector and second as the early-stopping monitor. After each accepted atom, LinFE refits a linear model using the current feature set ($B \cup Q \cup S$) and evaluates the RMSE on the validation split. If the validation RMSE does not improve for p stages, the algorithm stops and restores the best validation snapshot $B \cup Q \cup S_{best}$. The final generated feature set includes the non-categorical features B, the fixed categorical residual terminals Q, and the selected symbolic expressions S_{best}.

Appendix C. Proofs

Proof [Theorem 3]

Because $\mathcal{H}$ is scale-invariant, for any $h \in \mathcal{H}$ and scalar $c > 0$, the ratio remains invariant since the scalar factors out of both the dot product and the norm. Consequently, evaluating the supremum over the entire space $\mathcal{H}$ is algebraically equivalent to evaluating it over the normalized subclass $\mathcal{H}_1$:

$$\sup_{h \in \mathcal{H}} \frac{\langle h(x), y \rangle_N}{\|h(x)\|_N} = \sup_{h \in \mathcal{H}_1} \frac{\langle h(x), y \rangle_N}{\|h(x)\|_N} \tag{S.4}$$

Furthermore, a standard property of supremums dictates that for any two real-valued functionals F and G, the difference of their supremums is bounded by the supremum of their absolute difference: $|\sup_h F(h) - \sup_h G(h)| \leq \sup_h |F(h) - G(h)|$. Therefore, we can bound the target difference by bounding the uniform deviation over $\mathcal{H}_1$. Recognizing that for all $h \in \mathcal{H}_1$, the population norm is strictly $\sqrt{\mathbb{E}[h(x)^2]} = 1$, the goal reduces to bounding:

$$\sup_{h \in \mathcal{H}_1} \left| \frac{\langle h(x), y \rangle_N}{\|h(x)\|_N} - \mathbb{E}[h(x)y] \right| \tag{S.5}$$

We begin by establishing standard uniform convergence bounds for the empirical dot product and the squared empirical norm.

For the dot product, the summand is $h(x)y$. Since $|h(x)| \leq M$ and $|y| \leq 1$, the product $|h(x)y| \leq M$. By standard Rademacher concentration (via McDiarmid's inequality and symmetrization), with probability at least $1 - \delta/2$:

$$\sup_{h \in \mathcal{H}_1} |\langle h(x), y \rangle_N - \mathbb{E}[h(x)y]| \leq 2\mathcal{R}_N(\mathcal{H}_1 \cdot y) + M\sqrt{\frac{2\ln(4/\delta)}{N}} \tag{S.6}$$

By Talagrand's Contraction Lemma, because multiplication by $y \in [-1, 1]$ is a 1-Lipschitz operation, $\mathcal{R}_N(\mathcal{H}_1 \cdot y) \leq \mathcal{R}_N(\mathcal{H}_1)$. Thus, the right-hand side is bounded by ϵ_A.

For the squared empirical norm, the summand is $h(x)^2$, which is bounded by M^2. Applying Rademacher concentration again, with probability at least $1 - \delta/2$:

$$\sup_{h \in \mathcal{H}_1} \left| \|h(x)\|_N^2 - \mathbb{E}[h(x)^2] \right| \leq 2\mathcal{R}_N(\mathcal{H}_1^2) + M^2\sqrt{\frac{2\ln(4/\delta)}{N}} \tag{S.7}$$

Since $\mathbb{E}[h(x)^2] = 1$ by definition of $\mathcal{H}_1$, and applying the Contraction Lemma to the squared function (which is $2M$-Lipschitz on $[-M, M]$), we have $\mathcal{R}_N(\mathcal{H}_1^2) \leq 2M\mathcal{R}_N(\mathcal{H}_1)$. Thus, the right-hand side is bounded by ϵ_B.

By the union bound, both events hold simultaneously with probability at least $1 - \delta$. We condition the remainder of the proof on this intersection event.

By the explicit assumption of the theorem, N is large enough such that $\epsilon_B \leq 0.5$. Because we established that $\left| \|h(x)\|_N^2 - 1 \right| \leq \epsilon_B$, it strictly follows that:

$$\|h(x)\|_N^2 \geq 1 - \epsilon_B \geq 0.5 \tag{S.8}$$

Taking the square root and inverting yields a deterministic uniform upper bound on the inverse norm:

$$\frac{1}{\|h(x)\|_N} \leq \frac{1}{\sqrt{0.5}} = \sqrt{2} \tag{S.9}$$

We wish to bound $\left|\|h(x)\|_N^{-1} - 1\right|$. We apply the Mean Value Theorem to the function $f(x) = x^{-1/2}$ over the interval $x \in [1-\epsilon_B, 1+\epsilon_B]$. The MVT states:

$$|f(x) - f(1)| \leq \left(\sup_{z \in [1-\epsilon_B, 1+\epsilon_B]} |f'(z)| \right) |x - 1| \tag{S.10}$$

The derivative is $f'(z) = -\frac{1}{2z^{3/2}}$. Its absolute value is maximized at the lower bound of the interval, $z = 1 - \epsilon_B$. Therefore:

$$\left| \frac{1}{\|h(x)\|_N} - 1 \right| \leq \frac{1}{2(1-\epsilon_B)^{3/2}} \left| \|h(x)\|_N^2 - 1 \right| \tag{S.11}$$

Because $\epsilon_B \leq 0.5$, the maximum absolute derivative is upper-bounded by:

$$\frac{1}{2(0.5)^{3/2}} = \frac{1}{2 \cdot (2^{-3/2})} = \sqrt{2} \tag{S.12}$$

Substituting our known bound $\left|\|h(x)\|_N^2 - 1\right| \leq \epsilon_B$ yields:

$$\left| \frac{1}{\|h(x)\|_N} - 1 \right| \leq \sqrt{2}\epsilon_B \tag{S.13}$$

We decompose the target ratio using the standard algebraic identity:

$$\frac{\langle h(x), y \rangle_N}{\|h(x)\|_N} - \mathbb{E}[h(x)y] = \frac{1}{\|h(x)\|_N} \Big(\langle h(x), y \rangle_N - \mathbb{E}[h(x)y] \Big) + \mathbb{E}[h(x)y] \left(\frac{1}{\|h(x)\|_N} - 1 \right) \tag{S.14}$$

Taking the absolute value and applying the triangle inequality:

$$\left| \frac{\langle h(x), y \rangle_N}{\|h(x)\|_N} - \mathbb{E}[h(x)y] \right| \leq \frac{1}{\|h(x)\|_N} \left| \langle h(x), y \rangle_N - \mathbb{E}[h(x)y] \right| + \left| \mathbb{E}[h(x)y] \right| \left| \frac{1}{\|h(x)\|_N} - 1 \right| \tag{S.15}$$

We apply the Cauchy-Schwarz inequality to bound the magnitude of the population expectation. Since $h \in \mathcal{H}_1 \implies \mathbb{E}[h(x)^2] = 1$ and $y \in [-1, 1] \implies \mathbb{E}[y^2] \leq 1$:

$$|\mathbb{E}[h(x)y]| \leq \sqrt{\mathbb{E}[h(x)^2]\mathbb{E}[y^2]} \leq \sqrt{1 \cdot 1} = 1 \tag{S.16}$$

Finally, we substitute all four bounded components into the decomposed inequality:

1. $1/\|h(x)\|_N \leq \sqrt{2}$
2. $|\langle h(x), y \rangle_N - \mathbb{E}[h(x)y]| \leq \epsilon_A$
3. $|\mathbb{E}[h(x)y]| \leq 1$ (from Cauchy-Schwarz)
4. $|1/\|h(x)\|_N - 1| \leq \sqrt{2}\epsilon_B$

This yields:

$$\left| \frac{\langle h(x), y \rangle_N}{\|h(x)\|_N} - \mathbb{E}[h(x)y] \right| \leq \sqrt{2} \cdot \epsilon_A + 1 \cdot \sqrt{2}\epsilon_B = \sqrt{2}(\epsilon_A + \epsilon_B) \tag{S.17}$$

Taking the supremum over all $h \in \mathcal{H}_1$ gives the final desired result. ■

Proof [Lemma 8]

Let $R(h) = \mathbb{E}_P[\ell(Y, h(B))]$ and $R_g(h_g) = \mathbb{E}_P[\ell(Y, h_g(B \cup \{g\}))]$ denote the true risks over the baseline and augmented feature spaces, respectively. By definition, the functional gain is $\Delta(g, \mathcal{H}) = \min_{h \in \mathcal{H}} R(h) - \min_{h_g \in \mathcal{H}_g} R_g(h_g)$.

Let $h_P^* \in \arg\min_{h \in \mathcal{H}_F} R(h)$ and $h_D^* \in \arg\min_{h \in \mathcal{H}_D} R(h)$. By the definition of the Hausdorff distance $d_{H,P}$, for the optimal proxy hypothesis $h_P^* \in \mathcal{H}_F$, there exists a downstream hypothesis $h_D \in \mathcal{H}_D$ such that $\|h_P^* - h_D\|_{L^1(P)} \leq d_{H,P}(\mathcal{H}_F, \mathcal{H}_D)$.

Using the L-Lipschitz continuity of the loss function, we directly bound the risk difference:

$$\begin{aligned} R(h_D^*) - R(h_P^*) &\leq R(h_D) - R(h_P^*) \\ &= \mathbb{E}_P[\ell(Y, h_D(B)) - \ell(Y, h_P^*(B))] \\ &\leq L\mathbb{E}_P[|h_D(B) - h_P^*(B)|] \\ &= L\|h_D - h_P^*\|_{L^1(P)} \leq L \cdot d_{H,P}(\mathcal{H}_F, \mathcal{H}_D) \end{aligned}$$

By symmetry, reversing the roles of $\mathcal{H}_F$ and $\mathcal{H}_D$ yields $|R(h_P^*) - R(h_D^*)| \leq L \cdot d_{H,P}(\mathcal{H}_F, \mathcal{H}_D)$.

Following an identical projection and bounding argument for the augmented optimal hypotheses $h_{P,g}^* \in \mathcal{H}_{P,g}$ and $h_{D,g}^* \in \mathcal{H}_{D,g}$, we obtain $|R_g(h_{P,g}^*) - R_g(h_{D,g}^*)| \leq L \cdot d_{H,P}(\mathcal{H}_{P,g}, \mathcal{H}_{D,g}) \leq L \cdot d_{H,P}(\mathcal{H}_F, \mathcal{H}_D)$.

Expanding the target term via the triangle inequality yields:

$$\begin{aligned} |\Delta(g, \mathcal{H}_F) - \Delta(g, \mathcal{H}_D)| &\leq \big(|R(h_P^*) - R(h_D^*)| + |R_g(h_{P,g}^*) - R_g(h_{D,g}^*)|\big) \\ &\leq (L \cdot d_{H,P}(\mathcal{H}_F, \mathcal{H}_D) + L \cdot d_{H,P}(\mathcal{H}_F, \mathcal{H}_D)) \\ &= 2L \cdot d_{H,P}(\mathcal{H}_F, \mathcal{H}_D) \end{aligned}$$

Taking the supremum over $\mathcal{G}$ confirms that $L_H \leq 4L$ serves as a valid upper bound, completing the proof. ■

Proof [Theorem 9]

1. Decomposition of Regret: For any fixed downstream class $\mathcal{H}_D \in \mathcal{D}$ and a feature $\hat{g}_F$ selected on S_n, we decompose the regret relative to the downstream-optimal feature g_D^* by bridging through the population proxy utility $\Delta(\cdot, \mathcal{H}_F)$:

$$\begin{aligned} \text{Regret}(\mathcal{H}_D) &= \Delta(g_D^*, \mathcal{H}_D) - \Delta(\hat{g}_F, \mathcal{H}_D) \\ &= \underbrace{\Delta(g_D^*, \mathcal{H}_D) - \Delta(g_D^*, \mathcal{H}_F)}_{(A)} + \underbrace{\Delta(g_D^*, \mathcal{H}_F) - \Delta(\hat{g}_F, \mathcal{H}_F)}_{(B)} + \underbrace{\Delta(\hat{g}_F, \mathcal{H}_F) - \Delta(\hat{g}_F, \mathcal{H}_D)}_{(C)} \end{aligned}$$

2. Bounding the Inductive Mismatch (A + C): By applying the definition of the supremum over the feature set $\mathcal{G}$, we bound the transfer loss:

$$\begin{aligned} (A) + (C) &\leq |\Delta(g_D^*, \mathcal{H}_D) - \Delta(g_D^*, \mathcal{H}_F)| + |\Delta(\hat{g}_F, \mathcal{H}_F) - \Delta(\hat{g}_F, \mathcal{H}_D)| \\ &\leq 2 \sup_{g \in \mathcal{G}} |\Delta(g, \mathcal{H}_F) - \Delta(g, \mathcal{H}_D)| \leq 2L_H d_H(\mathcal{H}_F, \mathcal{H}_D) \end{aligned}$$

3. Bounding Selection Overfitting (B): Let $g_P^* = \arg\max_{g \in \mathcal{G}} \Delta(g, \mathcal{H}_F)$. Since $\Delta(g_P^*, \mathcal{H}_F) \geq \Delta(g_D^*, \mathcal{H}_F)$, then $(B) \leq \Delta(g_P^*, \mathcal{H}_F) - \Delta(\hat{g}_F, \mathcal{H}_F)$. Introducing the empirical proxy gain $\widehat{\Delta}(g, \mathcal{H}_F, S_n)$:

$$(B) \leq \left[\Delta(g_P^*, \mathcal{H}_F) - \widehat{\Delta}(g_P^*, \mathcal{H}_F, S_n)\right] + \left[\widehat{\Delta}(g_P^*, \mathcal{H}_F, S_n) - \widehat{\Delta}(\hat{g}_F, \mathcal{H}_F, S_n)\right] + \left[\widehat{\Delta}(\hat{g}_F, \mathcal{H}_F, S_n) - \Delta(\hat{g}_F, \mathcal{H}_F)\right]$$

Because $\hat{g}_F$ maximizes $\widehat{\Delta}$ by construction, the middle term is ≤ 0. Thus:

$$(B) \leq 2 \sup_{g \in \mathcal{G}} |\widehat{\Delta}(g, \mathcal{H}_F, S_n) - \Delta(g, \mathcal{H}_F)|$$

4. Concentration and Complexity: Let $\mathcal{F} = \{(x, y) \mapsto \ell(y, h(B(x), g(x))) : h \in \mathcal{H}_F, g \in \mathcal{G}\}$. Define the uniform deviation $\Phi(S_n) = \sup_{f \in \mathcal{F}} |\mathbb{E}_P[f] - \mathbb{E}_{S_n}[f]|$.

Assuming the loss ℓ is bounded in $[0, C]$, changing a single sample z_i modifies $\Phi(S_n)$ by at most C/n. By McDiarmid's inequality, for any $\delta > 0$, with probability at least $1 - \delta$:

$$\Phi(S_n) \leq \mathbb{E}_{S_n}[\Phi(S_n)] + C\sqrt{\frac{\log(1/\delta)}{2n}}$$

By standard symmetrization, $\mathbb{E}_{S_n}[\Phi(S_n)] \leq 2\mathfrak{R}_n(\mathcal{F})$. Since Term $(B) \leq 2\Phi(S_n)$ (noting that gain estimation error is bounded by risk estimation error), we combine these results to find that with probability $1 - \delta$:

$$(B) \leq 4\mathfrak{R}_n(\mathcal{F}) + 2C\sqrt{\frac{\log(1/\delta)}{2n}}$$

5. Final Assembly: Substituting the high-probability bound for Term (B) back into the regret decomposition and taking the expectation over $\mathcal{H}_D \sim \mathcal{D}$ for the inductive mismatch terms:

$$\mathcal{R}_{univ}(\mathcal{H}_F) \leq 4\mathfrak{R}_n(\ell \circ \mathcal{H}_F \circ \mathcal{G}) + \mathbb{E}_{\mathcal{H}_D \sim \mathcal{D}}\left[L_H d_H(\mathcal{H}_F, \mathcal{H}_D)\right] + 2C\sqrt{\frac{\log(1/\delta)}{2n}}$$

■