CONVEX POLYNOMIAL REGULARIZATION FOR LEARNING REPRESENTATIONS WITH DEGREE-ORDERED POWER GROWTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning effective representations requires inducing structured inductive biases over function spaces, where polynomial degree serves as a natural hierarchy of structural complexity. While linear degree weighting enforces this ordering, standard degree-dependent power penalties inadvertently invert capacity control near zero. This study resolve this by introducing two shape-corrected regularizers that preserve global degree monotonicity including a linear-power penalty (LiP) and a shifted-power penalty (ShiP). These formulations directly constrain the learned coordinate representations within coercive, strictly convex optimization landscapes, guaranteeing unique global solutions and admitting exact coordinate-wise minimization. Across 09 regression benchmarks, regularizing representations via ShiP achieves competitive predictive performance that rivals non-parametric Gaussian process baselines. By bridging degree-ordered complexity constraints with non-linear power penalties, this framework ensures that expressive feature representations remain well-conditioned, interpretable, and computationally tractable without distorting the underlying capacity hierarchy.

1 INTRODUCTION

Polynomial regression makes model structure explicit because each coefficient multiplies a term with a known degree. Once the candidate degree is large, regularization must control both coefficient magnitude and polynomial order. Ridge (Hoerl and Kennard, 1970), lasso (Tibshirani, 1996), and elastic net (Zou and Hastie, 2005) shrink coefficients without distinguishing degree. A simple degree-weighted lasso, $\sum_k k|b_k|$, already imposes the desired ordering at every fixed coefficient magnitude. If that linear degree ordering is the only goal, the weighted lasso is a simpler and sufficient choice.

For the weighted lasso, $p_k(t) = kt$ is linear in t, so its marginal penalty is constant once a coefficient is nonzero. The obvious construction, $|b_k|^k$, has exactly the wrong local behavior: for $0 < t < 1$, $t^{k+1} < t^k$, so increasing the degree *reduces* the penalty at a fixed small coefficient magnitude. Multiplying each power by a finite degree-dependent weight cannot remove this reversal because the ratio between higher and lower powers still vanishes as $t \downarrow 0$. The failure is therefore a property of the pure-power shape, not a tuning accident (Goyal and Bengio, 2022).

There are two properties required to hold simultaneously, *global degree ordering* and *degree-dependent power growth*. We formalize the first as coefficient-level degree monotonicity at every fixed nonzero coefficient magnitude, every higher polynomial degree must incur strictly greater cost. LiP and ShiP bridge the natural endpoints.LiP adds a degree-weighted linear component to the pure powers. ShiP shifts the powers so that their slopes at the origin are positive and increase with degree while the higher-order power behavior remains away from zero.

These nonlinear degree-monotone penalties against the linear degree-weighted comparator and more flexible learners (Tran et al., 2022). Across 09 public univariate datasets and 11 methods, DAMP–ShiP attains the second-lowest nominal average rank (4.89), close to Gaussian-process regression (4.67). The global Friedman test does not separate the methods, and corrected paired tests show dataset-specific wins and losses. The benchmark also identifies a representation boundary on motorcycle-impact data, where a single global polynomial is poorly matched to the response curve, flexible nonparametric methods perform substantially better.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The main contribution retains global degree ordering while safely enabling nonlinear power growth, overcoming the near-zero obstruction. Two convex corrections, LiP and ShiP, ensure this geometry remains uniquely solvable and computationally explicit. Distinct from multivariate inclusion hierarchies, these novel penalties strictly control magnitude-dependent shrinkage geometries across fixed polynomial basis expansion dictionaries. The optimization results show that the corrected geometry remains uniquely solvable and computationally explicit (Bien et al., 2013; Bertsimas and Van Parys, 2020).

2 Monotone Degree Bias

2.1 Impossibility of Finite Power to Enforce Degree Ordering

Let z_i denote a scaled predictor, with the observed training range mapped to $[-1, 1]$, and let $X_{ik} = z_i^k$ for $k = 1, \ldots, n$. With an unpenalized intercept b_0, we consider

$$\min_{b_0,\, \boldsymbol{b}} \ \frac{1}{2} \left\| \boldsymbol{y} - b_0 \mathbf{1} - X\boldsymbol{b} \right\|^2 + \lambda P(\boldsymbol{b}). \tag{1}$$

Because coefficient magnitudes depend on the basis scale, all claims below are made after this preprocessing convention is fixed. In the experiments, response scaling is also estimated from the training fold and undone for evaluation.

Definition 1 (Coefficient-level degree monotonicity)**.** *A family of scalar penalties $\{p_k\}_{k=1}^n$ is degree-monotone if, for every $t > 0$ and $1 \le \ell < k \le n$, $p_k(t) > p_\ell(t)$.*

The definition compares equal coefficient magnitudes and therefore isolates the structural preference encoded by the regularizer. It is stronger than requiring a tuning sequence to increase with degree: the ordering must hold for every coefficient magnitude, including arbitrarily small ones.

Proposition 1. *For degrees $1 \le \ell < k$ and any finite $w_\ell, w_k > 0$, there exists $t > 0$ such that $w_k t^k < w_\ell t^\ell$. Hence no finite static weighting of t^k is uniformly degree-monotone on $(0, \infty)$.*

The proof is immediate from $(w_k/w_\ell)t^{k-\ell} \to 0$ as $t \downarrow 0$; Appendix A gives the full argument. Proposition 1 shows the finite weighting cannot repair pure powers. It identifies the obstruction any construction that retains the pure-power shape and changes only finite weights must fail near zero. It does *not* rule out linear degree weighting $p_k(t) = kt$ is globally degree-monotone.

In Figure 1, pure degree-dependent powers reverse the intended ordering near zero. LiP restores it through a degree-weighted linear component. ShiP shifts the power so that every degree has a positive slope at the origin. Curves are illustrative parameter settings used only to show the structural effect.

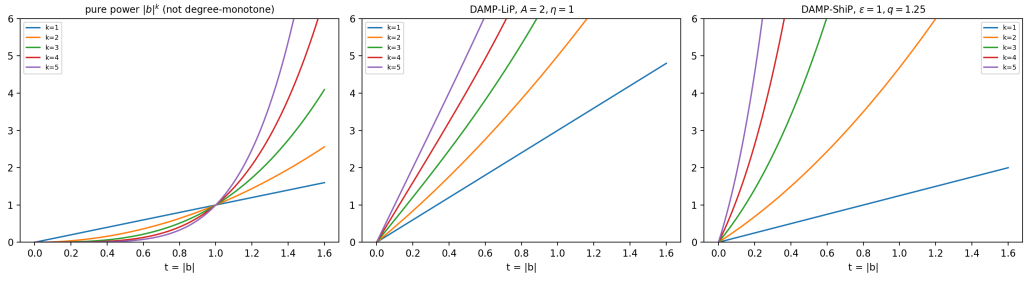

Figure 1: Penalty shape at equal coefficient magnitude.

2.2 Two Convex Corrections to the Near-Zero Geometry

The linear–power construction is:

$$P_{\mathrm{LiP}}(\boldsymbol{b}) = \sum_{k=1}^{n} \left(Ak|b_k| + \eta|b_k|^k \right), \qquad A, \eta > 0. \tag{2}$$

The shifted–power construction is:

$$P_{\mathrm{ShiP}}(\boldsymbol{b}) = \sum_{k=1}^{n} q^k \left[\left(1 + \frac{|b_k|}{\varepsilon} \right)^k - 1 \right], \qquad \varepsilon > 0,\ q \geq 1. \tag{3}$$

LiP keeps the original degree-dependent power growth but adds a nonsmooth term whose slope grows linearly with degree. ShiP changes the local geometry directly:

$$p_k^{\mathrm{ShiP}}(t) = \frac{kq^k}{\varepsilon} t + O(t^2), \tag{4}$$

so every degree has a genuine lasso-like threshold and the leading slope increases with k.

Theorem 1 (Global degree monotonicity)**.** *For the standard degree set $\{1, \ldots, n\}$ with $n \geq 2$:*

(i) LiP is coefficient-level degree-monotone for every $t > 0$ if and only if $A \geq \eta$;

(ii) ShiP is coefficient-level degree-monotone for every $t > 0$ whenever $q \geq 1$ and $\varepsilon > 0$.

The LiP condition is sharp because the degree-1/2 difference equals $t[(A - \eta) + \eta t]$. If $A < \eta$, it is negative for sufficiently small t. Sufficiency follows by treating $t \geq 1$ and $0 < t < 1$ separately. For ShiP, $(1 + t/\varepsilon)^k - 1$ is strictly increasing in k, and q^k is nondecreasing for $q \geq 1$. Theorem 1 therefore resolves the structural question directly: LiP has an exact parameter boundary for global degree ordering, while ShiP achieves the ordering throughout its stated parameter range. Complete proofs appear in Appendix A.

2.3 Degree ordering without sacrificing well-posedness

A degree-ordering penalty is useful only if the resulting estimator remains well posed. The next result shows that both shape corrections retain convexity and determine a unique fit in coefficient space even when the polynomial design is rank deficient.

Theorem 2 (Well-posed penalized regression)**.** *For either P_{LiP} or P_{ShiP}, objective equation 1 is convex and coercive and has a unique global minimizer for every data matrix X and every $\lambda > 0$. The result includes rank-deficient polynomial designs and an unpenalized intercept.*

For $k > 1$, each penalized coordinate is strictly convex. The remaining (b_0, b_1) block is uniquely identified either by the quadratic loss or, in the collinear edge case, by the positive absolute-value penalty on b_1. Collectivity follows from $P_{\mathrm{LiP}} \geq A \sum_k |b_k|$ and $P_{\mathrm{ShiP}} \geq \varepsilon^{-1} \sum_k |b_k|$.Thus, global degree ordering does not require giving up existence or uniqueness of the penalized estimator. Appendix A gives the details.

Let $g_k(\boldsymbol{b}) = \boldsymbol{x}_k^\top (b_0 \mathbf{1} + X\boldsymbol{b} - \boldsymbol{y})$. For LiP, the zero thresholds are

$$\tau_1^{\mathrm{LiP}} = \lambda(A + \eta), \qquad \tau_k^{\mathrm{LiP}} = \lambda A k \quad (k \geq 2), \tag{5}$$

and for ShiP

$$\tau_k^{\mathrm{ShiP}} = \lambda \frac{kq^k}{\varepsilon}. \tag{6}$$

For $k \geq 2$, $b_k^\star = 0$ if and only if $|g_k(\boldsymbol{b}^\star)| \leq \tau_k$. At $k = 1$, the equality case has the ordinary lasso ambiguity: zero implies the weak inequality, while a nonzero coefficient can occur at equality. Notice also that when $A = \eta$, LiP has $\tau_1 = \tau_2$ and strict threshold growth only thereafter; when $A > \eta$, its thresholds increase strictly from degree one onward.

2.4 Exact coordinate minimization

For optimization, separability makes cyclic coordinate descent natural (Friedman et al., 2010; Tseng, 2001), and every coordinate has an exact one-dimensional minimizer. Fix all coordinates except b_k and define the partial residual $\boldsymbol{r}_k = \boldsymbol{y} - b_0 \mathbf{1} - \sum_{j \neq k} \boldsymbol{x}_j b_j$, $a_k = \|\boldsymbol{x}_k\|^2$, and $c_k = \boldsymbol{x}_k^\top \boldsymbol{r}_k$. Both coordinate subproblems have a unique minimizer. If $a_k = 0$, then $\boldsymbol{x}_k = 0$ and $c_k = 0$, so the unique update is $b_k = 0$; below we assume $a_k > 0$.

For LiP, if $k \geq 2$ and $|c_k| \leq \lambda A k$, the update is zero; otherwise $b_k \leftarrow \mathrm{sign}(c_k) s_k$, where the unique $s_k > 0$ solves

$$a_k s + \lambda \eta k s^{k-1} = |c_k| - \lambda A k. \tag{7}$$

Degree one is soft thresholding at $\lambda(A + \eta)$, and degree two has

$$b_2 \leftarrow \frac{\mathrm{sign}(c_2)(|c_2| - 2\lambda A)_+}{a_2 + 2\lambda\eta}. \tag{8}$$

For ShiP, if $|c_k| \leq \lambda k q^k / \varepsilon$ the update is zero; otherwise $b_k \leftarrow \mathrm{sign}(c_k) s_k$ with

$$a_k s + \lambda \frac{k q^k}{\varepsilon} \left(1 + \frac{s}{\varepsilon}\right)^{k-1} = |c_k|. \tag{9}$$

Again degree one is soft thresholding, and degree two is closed form. The left sides of Eq. (7) and Eq. (9) are continuous and strictly increasing, giving unique roots and simple brackets for bisection or safeguarded Newton updates. Coordinate objectives are used to numerical tolerance and monitor coefficient change and KKT residuals. Each coordinate is minimized exactly and the objective is nonincreasing under the cyclic updates, within the standard block-coordinate framework (Tseng, 2001). Pseudocode and numerical checks are in Appendix B.

3 Experiments of degree-monotone regularization

3.1 Dataset and Pre-processing

The benchmark tests how that geometry behaves predictively under the same train/validation/test protocol as standard polynomial regularizers and flexible nonlinear learners. There are 09 public univariate regression datasets: 07 from R/Rdatasets and 02 NIST Statistical Reference Datasets, including the cerftified degree-10 Filip problem and quadratic Pontius problem. Dataset details and provenance are in Table 1:

Table 1: Public univariate regression datasets used in the benchmark.

Dataset	Source	Relationship	N
mcycle	Rdatasets	time → head acceleration	133
cars	Rdatasets	speed → stopping distance	50
ethanol	Rdatasets	equivalence ratio → NOx	88
boston_lstat	Rdatasets	lower-status share → home value	506
faithful	Rdatasets	eruption duration → waiting time	272
gagurine	Rdatasets	age → urinary GAG	314
wage_age	Rdatasets	age → wage	3000
nist_filip	NIST StRD	certified degree-10 polynomial	82
nist_pontius	NIST StRD	certified quadratic calibration	40

For each dataset, we generate 20 deterministic outer splits, holding out 20% for testing. All eleven methods receive the same outer split and the same inner folds. Hyperparameters are selected by five-fold cross-validation using only the remaining 80%, the selected configuration is refit once on the full training portion, and the test set is evaluated once. Predictor and response scaling are estimated within the relevant training fold. The benchmark compares DAMP–LiP and DAMP–ShiP with pure-power regularization, ordinary polynomial least squares, ridge, lasso, elastic net, degree-weighted lasso $\sum_k k|b_k|$, random forests (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), and Gaussian-process regression (Rasmussen and Williams, 2006). Degree and regularization grids, implementation details, and diagnostic checks are in Appendix C.

We report repeated-split RMSE. Because the random holdouts overlap, splitwise errors are dependent; within each dataset we use the Nadeau–Bengio corrected resampled t-test (Nadeau and Bengio, 2003). For $R = 20$ paired RMSE differences with sample variance s^2, the corrected standard error is $\{(1/R + n_{\mathrm{test}}/n_{\mathrm{train}})s^2\}^{1/2}$, using the exact split sizes for that dataset. For each DAMP variant, the ten comparisons are Holm-corrected within dataset.

3.2 EXPERIMENTAL RESULTS

3.2.1 BENCHMARK RESULTS

Figure 2 represents fitted curves from the benchmark datasets. The panels show the representation scope of a global polynomial. Some responses admit a compact polynomial fit, whereas motorcycle impact exhibits local structure that a single global polynomial does not capture well.

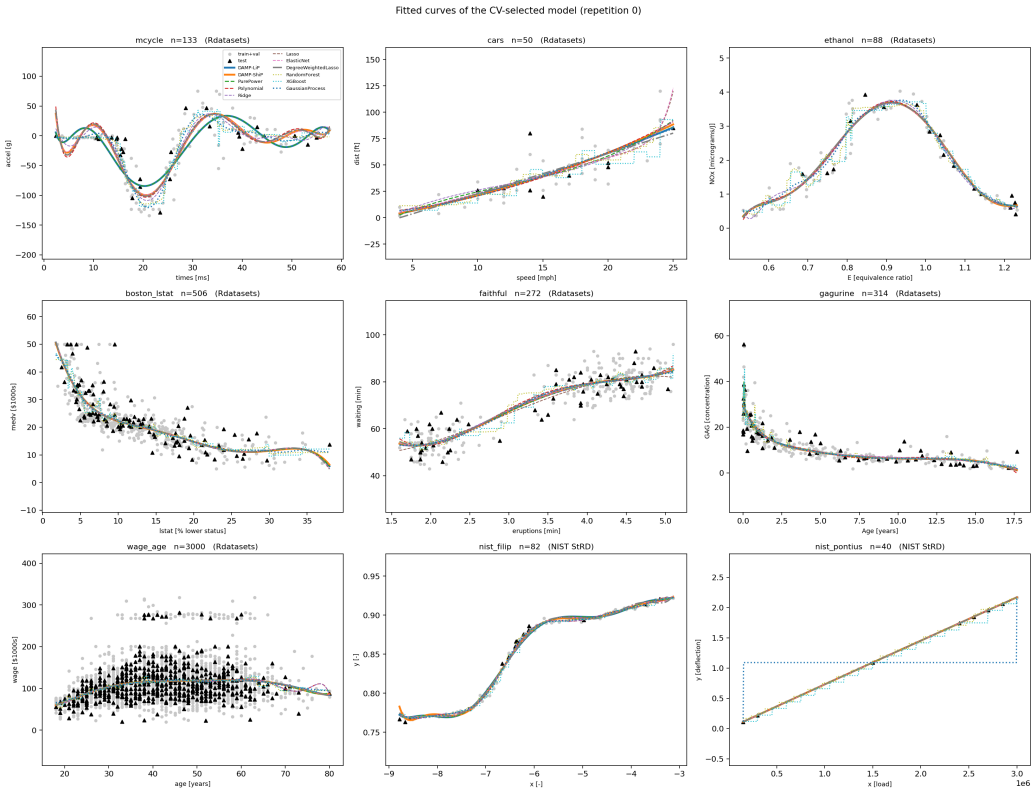

Figure 2: Representative fitted curves from the benchmark.

Table 2 and Figure 3 present mean test RMSE over 20 paired outer splits. All statistical claims rely on the corrected paired tests detailed in Appendix C rather than visual overlap alone.

Table 2: Mean test RMSE over 20 paired outer splits.

Dataset	LiP	ShiP	Pure	Poly	Ridge	Lasso	EN	DWL	RF	XGB	GP
boston_lstat	5.20	5.20	5.20	5.21	5.20	5.21	5.20	5.20	5.36	5.38	**5.16**
cars	16.76	16.97	16.72	16.70	17.20	17.51	17.86	16.72	16.27	**15.93**	16.49
ethanol	0.356	0.356	0.354	0.360	0.451	0.361	0.348	0.366	0.353	0.379	**0.327**
faithful	**5.65**	5.65	5.67	5.67	5.66	5.65	5.67	5.66	5.78	5.84	5.66
gagurine	4.76	4.76	4.75	4.78	4.77	4.75	4.76	4.75	4.47	**4.43**	4.78
mcycle	30.81	28.87	30.81	46.77	39.68	27.76	27.72	27.68	23.26	24.36	**21.76**
nist_filip	0.0051	0.0052	0.0051	0.0045	0.0044	0.0054	0.0058	0.0055	0.0036	0.0046	**0.0031**
nist_pontius	**0.00021**	0.000213	0.000215	0.000216	0.000216	0.000215	0.000215	0.000217	0.049	0.046	0.267
wage_age	39.61	39.60	39.60	39.60	39.60	**39.60**	39.60	39.60	39.71	39.80	39.61

Bold marks the smallest mean in each dataset.

The aggregate corrected paired outcomes reported in the main text are generated directly from the saved statistical-comparison CSV. Against the three external machine-learning methods, the two DAMP variants contribute 54 comparisons: 2 significant wins, 5 significant losses, and 47 nonsignificant results. Against regression and regularization methods, the 108 comparisons contain 0 wins, 3 losses, and 105 nonsignificant results. These denominators exclude LiP–ShiP internal comparisons.

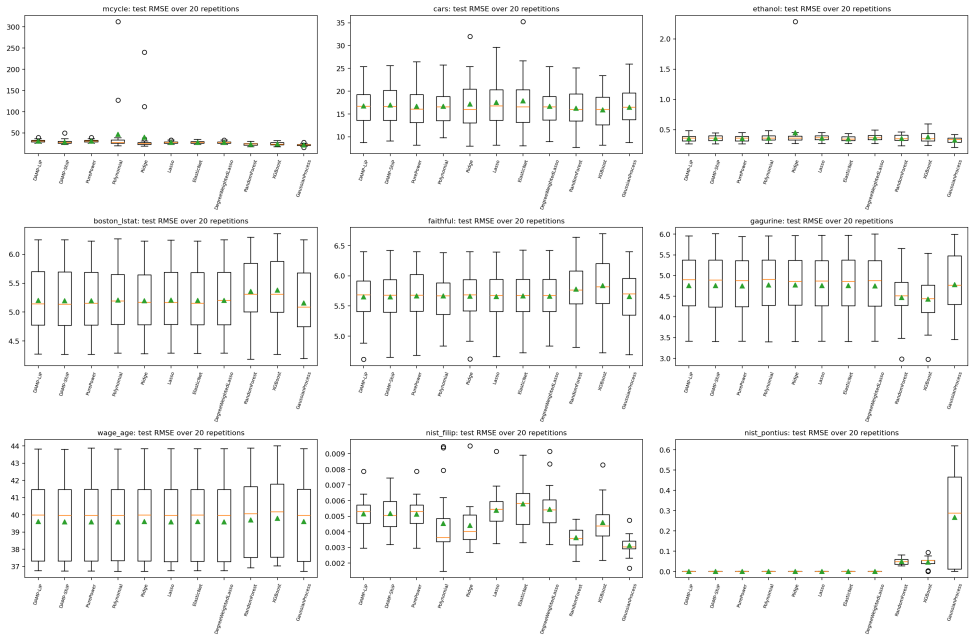

Figure 3: Repeated-split test RMSE distributions for all benchmark methods and datasets.

3.2.2 PREDICTIVE SCOPE

Table 3 shows descriptive average rank by mean RMSE over nine real-data tasks. DAMP–ShiP attains the second-lowest nominal average rank, 4.89, close to Gaussian-process regression at 4.67. ShiP ranks ahead of pure-power regularization (5.33), random forests (5.67), DAMP–LiP and lasso (both 5.89), degree-weighted lasso (6.11), XGBoost and ordinary polynomial regression (both 6.89), and ridge (7.00). The Friedman statistic is $\chi^2 = 5.56$ with $p = 0.851$, so the cross-dataset ranking is descriptive rather than evidence of a global performance ordering.

Table 3: Descriptive average rank by mean RMSE over nine real-data tasks (lower is better).

Method	Avg. rank ↓
Gaussian process	4.67
DAMP–ShiP	**4.89**
Pure power	5.33
Random forest	5.67
DAMP–LiP	**5.89**
Lasso	5.89
Degree-weighted lasso	6.11
Elastic net	6.78
XGBoost	6.89
Polynomial OLS	6.89
Ridge	7.00

Appendix D details the convergence status of the benchmark records, iteration counts, cross-validation candidate eligibility, and tuning-grid boundary selections. These diagnostics document the numerical conditions under which the reported comparisons were obtained.

From Table 3, the corrected task-level comparisons show where the polynomial representation helps and where it becomes the limiting factor. Against random forests, XGBoost, and Gaussian-process regression, LiP and ShiP together produce 2 Holm-significant wins, 5 significant losses, and 47 nonsignificant comparisons across 54 dataset–method pairs. Both significant wins occur against random forest on the NIST Pontius calibration problem.

Table 4 presents Holm-corrected Nadeau–Bengio corrected-resampled-t outcomes aggregated over datasets. Against regression and regularization baselines, 105 of 108 Holm-corrected comparisons are nonsignificant. The remaining three are LiP losses on motorcycle impact against lasso, elastic net, and degree-weighted lasso. Thus, the benchmark does not show a systematic predictive advantage for LiP or ShiP over the simpler weighted lasso. Its role is instead to test whether the stronger penalty geometry can retain competitive predictive accuracy. Degree-weighted lasso supplies global ordering with linear magnitude dependence, pure powers supply nonlinear degree-dependent growth but fail the global ordering requirement. LiP and ShiP are designed to combine those two properties.

Table 4: Holm-corrected Nadeau–Bengio corrected-resampled-t outcomes aggregated over datasets.

Comparison	Total	Wins	Losses	Nonsignificant
LiP vs ML	27	1	4	22
ShiP vs ML	27	1	1	25
LiP vs regression	54	0	3	51
ShiP vs regression	54	0	0	54

"Nonsignificant" means the corrected test does not establish a difference.

The degree-monotone fits also remain explicit. Across all dataset–split fits, LiP and ShiP retain on average 7.11 and 7.58 nonzero polynomial coefficients, respectively. Each retained coefficient corresponds directly to a polynomial term, so the fitted mapping can be written down and inspected term by term. Tree nodes and Gaussian-process training points are reported separately in the supplement rather than treated as a common complexity unit.

4 Discussion

The central distinction is between linear degree weighting and degree-monotone penalties that also retain nonlinear power growth. A weighted lasso such as $\sum_k k|b_k|$ already orders equal-sized coefficients by degree, but it uses the same linear magnitude dependence for every term: for $t > 0$, the marginal cost of kt is the constant k. Pure powers provide degree-dependent curvature but lose the ordering near zero because of the factor $t^{k-\ell}$. LiP and ShiP combine the two capabilities by changing the local geometry, where LiP adds a degree-weighted ℓ_1 slope to the powers, and ShiP builds an increasing near-zero slope, kq^k/ε, into a shifted power. For degrees above one, their marginal penalties then increase with coefficient magnitude rather than remaining constant.

This distinction also affects well-posedness in coefficient space. Theorem 2 gives LiP and ShiP a unique coefficient minimizer for every design matrix, including rank-deficient polynomial designs (Nguyen et al., 2021). Degree-weighted lasso general does not provide uniqueness coefficient under rank deficiency. That difference matters when the fitted polynomial is intended to be inspected term by term rather than being used only for prediction. Nevertheless, if linear degree ordering is the only structural requirement, degree-weighted lasso remains the simpler regularizer.

In the experimental results, ShiP averages second place across the nine tasks, though the Friedman test shows no global statistical difference. Compared to the weighted degree lasso, corrected tests reveal no systematic predictive advantage for nonlinear geometry. The results show that imposing added structural requirements maintains competitive accuracy, but DAMP does not dominate simpler regularizers. Specifically, both DAMP variants outperform random forest on the NIST Pontius dataset, whereas the motorcycle impact task reveals limits of global polynomial bases. Appendix D reports the convergence and tuning-boundary diagnostics associated with these comparisons.

Degree monotonicity also addresses a different structural question from polynomial hierarchy. Hierarchical interaction models constrain which monomials may coexist (Bien et al., 2013; Maruri-Aguilar and Lunagomez, 2020), and DAMP orders the penalty paid by equal-sized coefficients. Hierarchy controls support while degree-monotone penalties manage shrinkage in feature learning. Extending this univariate ordering to multivariate monomial partial orders remains open (Nguyen and Sakama, 2021).

The theory here is deterministic: once the basis and scaling convention are fixed, it establishes penalty geometry, convexity, uniqueness, thresholding, and coordinate computation. Statistical consistency,

degree-selection consistency, and risk guarantees require an explicit data-generating model and are not consequences of these geometric results. Those questions define the next theoretical layer beyond the capability established here.

5 CONCLUSION

This paper introduces an approach to preserve global degree ordering while facilitating safe nonlinear power scaling. LiP and ShiP repair the local geometry while retaining nonlinear growth. LiP has the sharp global condition $A \geq \eta$, and ShiP is globally degree-monotone for $q \geq 1$. Both yield convex, uniquely solvable penalized least-squares problems with degree-dependent thresholding and exact coordinate updates.

The experiment results show that ShiP ranks second on average across 09 public tasks, but neither the Friedman test nor the corrected comparisons against degree-weighted lasso establish a systematic predictive advantage. The empirical result is therefore compatibility rather than dominance. The stronger penalty geometry can be imposed while retaining competitive predictive performance.

In the future, this research will identify specific data regimes and task characteristics where this stronger penalty geometry got a clear empirical advantage over standard baseline methods. Theoretical extensions could also adapt the linear-power and shifted-power shape corrections to manage multivariate interaction hierarchies in more complex, high-dimensional feature spaces.

AI USE STATEMENT

Generative AI tools were used to assist manuscript editing for clarity and structure. They were not used to generate the raw datasets or to alter numerical outputs. The authors take responsibility for the final text, claims, code, and reported results.

ETHICS STATEMENT

The study uses existing public benchmark datasets and involves no new data collection, intervention on human subjects, access to private data, or deployment in a high-stakes decision system. The contribution is methodological. Dataset provenance and processing are documented in the appendix and anonymous supplementary material.

REPRODUCIBILITY STATEMENT

Appendix C gives the complete experimental protocol and hyperparameter grids; Appendix A contains full proofs of the central mathematical claims. The anonymous supplementary archive contains the benchmark source code, the canonical processed dataset files used by the reported run together with the original downloaded files and source URLs, the deterministic seed routine, validation scripts, raw repeated-split results, statistical-comparison tables, and figure-generation scripts. All numerical values in this manuscript are derived from those saved outputs.

REFERENCES

Bertsimas, D., and Van Parys, B. Sparse hierarchical regression with polynomials. *Machine Learning*, 109(5):973–997, 2020. doi:10.1007/s10994-020-05868-6.

Bien, J., Taylor, J., and Tibshirani, R. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013. doi:10.1214/13-AOS1096.

Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi:10.1023/A:1010933404324.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi:10.1145/2939672.2939785.

Friedman, J.H, Hastie, T. and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi:10.18637/jss.v033.i01.

Hoerl, A., and Kennard, R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi:10.1080/00401706.1970.10488634.

Maruri-Aguilar, H., and Lunagomez, S. Lasso for hierarchical polynomial models. *The Canadian Journal of Statistics*, Article e70075, 2026. doi: 10.1002/cjs.70075.

Nadeau, C., and Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003. doi: 10.1023/A:1024068626366.

Rasmussen, C.E., and Williams, C. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. ISBN 978-0-262-18253-9.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi:10.1111/j.2517-6161.1996.tb02080.x.

Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001. doi:10.1023/A:1017501703105.

Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

Tran, Q., Nguyen, H.T., Huynh, T., Nguyen, K., Hoang, S., and Pham, V. Measuring the influence and amplification of users on social network with unsupervised behaviors learning and efficient interaction-based knowledge graph. *Journal of Combinatorial Optimization*, 44(4): 2919-2945, 2022. doi: 10.1007/s10878-021-00815-0

Nguyen, H., Sakama, C., Sato, T., and Inoue, K. An efficient reasoning method on logic programming using partial evaluation in vector spaces. *Journal of Logic and Computation*, 31(5): 1298-1316, 2021. doi: 10.1093/logcom/exab010

Nguyen, H.D., and Sakama, C. Feature learning by least generalization. In *30th International Conference on Inductive Logic Programming (ILP 2021)*, pp. 193-202, 2021. 10.1007/978-3-030-97454-1_14

Goyal, A., and Bengio, Y. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266), 20210068, 2022. 10.1098/rspa.2021.0068

A Full proofs of the structural results

A.1 Impossibility of finite weighting

Proof of Proposition 1. Suppose for contradiction that $w_k t^k \geq w_\ell t^\ell$ for every $t > 0$, where $k > \ell$ and w_k, w_ℓ are finite and positive. Dividing by $w_\ell t^\ell$ gives

$$\frac{w_k}{w_\ell} t^{k-\ell} \geq 1 \qquad \text{for every } t > 0.$$

As $t \downarrow 0$, the left side tends to zero because $k - \ell > 0$, a contradiction. Strict degree monotonicity would imply this weak inequality, so it is impossible as well. □

A.2 Degree monotonicity of LiP

Proof of Theorem 1(i). Write $p_k(t) = Akt + \eta t^k$. Assume first that $A \geq \eta$ and let $1 \leq \ell < k$. If $t \geq 1$, then $t^k \geq t^\ell$ and $Akt > A\ell t$, so $p_k(t) > p_\ell(t)$. If $0 < t < 1$, then

$$p_k(t) - p_\ell(t) = A(k - \ell)t - \eta(t^\ell - t^k).$$

Since $0 < t^\ell - t^k < t^\ell \leq t$,

$$p_k(t) - p_\ell(t) > A(k - \ell)t - \eta t \geq 0,$$

where the last inequality uses $k - \ell \geq 1$ and $A \geq \eta$. Thus the difference is strictly positive.

For necessity on $\{1, \ldots, n\}$ with $n \geq 2$, compare degrees one and two:

$$p_2(t) - p_1(t) = 2At + \eta t^2 - (At + \eta t) = t\{(A - \eta) + \eta t\}.$$

If $A < \eta$, any $0 < t < (\eta - A)/\eta$ makes the bracket negative, contradicting degree monotonicity. Therefore $A \geq \eta$ is necessary and sufficient. □

A.3 Degree monotonicity of ShiP

Proof of Theorem 1(ii). For $t > 0$ set $s = t/\varepsilon > 0$. The scalar penalty is

$$p_k(t) = q^k\{(1 + s)^k - 1\}.$$

If $k > \ell$, then $(1 + s)^k - 1 > (1 + s)^\ell - 1 > 0$. Since $q \geq 1$, $q^k \geq q^\ell$. Hence

$$q^k\{(1 + s)^k - 1\} > q^\ell\{(1 + s)^\ell - 1\},$$

which is the desired strict ordering. □

A.4 Convexity, coercivity, and uniqueness

For LiP, every scalar coordinate $b \mapsto Ak|b| + \eta|b|^k$ is convex, and it is strictly convex for $k > 1$. For ShiP,

$$q^k \left[\left(1 + \frac{|b|}{\varepsilon} \right)^k - 1 \right] = q^k \sum_{j=1}^{k} \binom{k}{j} \varepsilon^{-j} |b|^j,$$

which is convex and, for $k > 1$, contains a positive quadratic term and is therefore strictly convex.

Proof of Theorem **??**. Convexity follows from the convex squared-error loss and the coordinatewise convex penalties. For coercivity,

$$P_{\mathrm{LiP}}(\boldsymbol{b}) \geq A \sum_{k=1}^{n} |b_k|.$$

For ShiP, $(1 + u)^k - 1 \geq u$ for $u \geq 0$ and $q \geq 1$, so

$$P_{\mathrm{ShiP}}(\boldsymbol{b}) \geq \frac{1}{\varepsilon} \sum_{k=1}^{n} |b_k|.$$

Thus either penalty diverges when $\|\boldsymbol{b}\| \to \infty$. If the full pair $(b_0, \boldsymbol{b})$ diverges while $\boldsymbol{b}$ remains bounded, then $|b_0| \to \infty$ and the squared-error term diverges because $N \geq 1$. Hence the intercept objective is coercive and continuous, so a minimizer exists.

For uniqueness, suppose two minimizers exist. Equality at their midpoint in the convexity inequality forces equality in every strictly convex penalty coordinate, so they agree in $b_2, \ldots, b_n$. Conditional on those coordinates, the remaining problem has the form

$$\min_{b_0, b_1} \frac{1}{2} \|\boldsymbol{r} - b_0\mathbf{1} - b_1\boldsymbol{x}_1\|^2 + \rho|b_1|,$$

with $\rho = \lambda(A + \eta) > 0$ for LiP and $\rho = \lambda q/\varepsilon > 0$ for ShiP. If $\mathbf{1}$ and $\boldsymbol{x}_1$ are linearly independent, the quadratic loss is strictly convex in (b_0, b_1) and the minimizer is unique. If $\boldsymbol{x}_1 = c\mathbf{1}$, the loss depends only on $m = b_0 + cb_1$. For fixed m, the positive term $\rho|b_1|$ is uniquely minimized by $b_1 = 0$, after which m is uniquely determined by the one-dimensional least-squares problem. Thus the full minimizer is unique in all cases. □

B Optimality conditions and coordinate updates

Let $\boldsymbol{b}^\star$ be the unique minimizer and $g_k = \boldsymbol{x}_k^\top(b_0\mathbf{1} + X\boldsymbol{b} - \boldsymbol{y})$. For LiP, the scalar subdifferential at zero is $[-(A+\eta), A+\eta]$ for $k=1$ and $[-Ak, Ak]$ for $k \geq 2$. Therefore

$$b_k^\star = 0 \iff |g_k(\boldsymbol{b}^\star)| \leq \lambda A k, \qquad k \geq 2,$$

while at $k=1$ zero implies $|g_1| \leq \lambda(A+\eta)$ and a nonzero coefficient implies equality. For ShiP, the subdifferential at zero is $[-kq^k/\varepsilon, kq^k/\varepsilon]$, giving

$$b_k^\star = 0 \iff |g_k(\boldsymbol{b}^\star)| \leq \lambda k q^k/\varepsilon, \qquad k \geq 2,$$

with the analogous lasso equality ambiguity at $k=1$.

For a coordinate subproblem, define $a_k = \|\boldsymbol{x}_k\|^2$ and $c_k = \boldsymbol{x}_k^\top \boldsymbol{r}_k$. If a nonzero minimizer exists, its sign is $\operatorname{sign}(c_k)$: for any $s > 0$, comparing the two points $t = \pm s$ leaves the quadratic term $\frac{1}{2}a_k s^2$ and the penalty unchanged, while the linear term $-c_k t$ is minimized at $t = \operatorname{sign}(c_k)s$. Hence the magnitude solves the scalar equations shown in Section 2.4.

If $a_k = 0$, then $\boldsymbol{x}_k = 0$ and the unique penalized coordinate update is $b_k = 0$. In the formulas and brackets below, assume $a_k > 0$. For LiP, degree one is

$$b_1 \leftarrow \frac{\operatorname{sign}(c_1)(|c_1| - \lambda(A+\eta))_+}{a_1},$$

and degree two is given by equation 8. For ShiP,

$$b_1 \leftarrow \frac{\operatorname{sign}(c_1)(|c_1| - \lambda q/\varepsilon)_+}{a_1},$$

and

$$b_2 \leftarrow \frac{\operatorname{sign}(c_2)(|c_2| - 2\lambda q^2/\varepsilon)_+}{a_2 + 2\lambda q^2/\varepsilon^2}.$$

For $k \geq 3$, each root equation has a continuous strictly increasing left side. For LiP, a valid bracket is

$$0 \leq s \leq \frac{|c_k| - \lambda A k}{a_k};$$

for ShiP,

$$0 \leq s \leq \frac{|c_k| - \lambda k q^k/\varepsilon}{a_k}.$$

The implementation uses safeguarded Newton updates within these brackets and falls back to bisection as needed.

Cyclic solver. The implementation alternates the exact unpenalized intercept update

$$b_0 \leftarrow \bar{y} - \frac{1}{N}\mathbf{1}^\top X\boldsymbol{b}$$

with the coordinate updates above. It uses warm starts along regularization paths during tuning. Validation scripts compare hundreds of coordinate updates against direct scalar minimization, check KKT residuals, verify intercept identities, confirm objective nonincrease, and test multiple initializations. These are numerical checks rather than proofs of a new convergence theorem.

C Complete experimental protocol

There are 20 outer repetitions per dataset. The bundled canonical two-column processed files are the benchmark inputs for the reported run; the original downloaded files are retained for provenance. A single canonical deterministic seed routine uses

$$\texttt{seed}(d, r) = 10{,}000r + \operatorname{crc32}(d) \bmod 9973.$$

The seed is independent of Python's randomized string hash and is shared by the benchmark, analysis, and figure scripts. The resulting outer split allocates 80% to train/validation and 20% to test. Within

the 80% portion, five-fold cross-validation with shuffled folds is generated once and reused by all methods in that dataset–repetition cell.

The polynomial degree grid is $\{3, 5, 8, 10, 12\}$ for DAMP, pure power, ridge, lasso, elastic net, and degree-weighted lasso; ordinary polynomial least squares searches all degrees $1, \ldots, 12$. Regularization strengths use 16 logarithmically spaced values from 10^{-8} to 10^{2}. LiP searches $(A, \eta) \in \{(1,1), (2,1), (5,1), (10,1)\}$. ShiP searches $\varepsilon \in \{0.1, 1, 10\}$ and $q \in \{1, 1.25, 1.75\}$. Elastic net searches ℓ_1 ratios $\{0.15, 0.5, 0.85\}$. The degree-weighted lasso uses $\sum_k k|b_k|$ with an unpenalized intercept. A candidate is eligible for selection only if every required cross-validation fold returns a finite score and converges under the configured solver; otherwise the candidate is excluded before the mean CV RMSE is computed.

Random forest searches 200, 500, or 1000 trees; maximum depth $\{\text{None}, 5, 10\}$; and minimum leaf size $\{1, 3, 5\}$. XGBoost searches 200 or 500 boosting rounds, maximum depth $\{2, 4, 6\}$, learning rate $\{0.05, 0.1\}$, and subsample fraction 0.9. Gaussian-process regression searches RBF, Matérn-$3/2$, and Matérn-$5/2$ kernels, each with a constant factor and white-noise term. Selection always minimizes mean inner-CV RMSE.

Statistical analysis. The 20 outer holdouts overlap, so splitwise RMSE differences are dependent. For every dataset and DAMP variant, we compare its RMSE with each other method using the Nadeau–Bengio corrected resampled t-test (Nadeau and Bengio, 2003). If $d_1, \ldots, d_R$ are the $R = 20$ paired RMSE differences and s_d^2 is their sample variance, the test uses

$$\mathrm{SE}_{\mathrm{corr}} = \sqrt{\left(\frac{1}{R} + \frac{n_{\mathrm{test}}}{n_{\mathrm{train}}}\right) s_d^2}, \qquad t = \bar{d} / \mathrm{SE}_{\mathrm{corr}},$$

with $R - 1$ degrees of freedom and the exact train/test sizes of that dataset. Ten p-values within each dataset–DAMP-variant family are corrected by Holm's method. The global Friedman test operates on each method's mean RMSE on the nine datasets. Because the global test is nonsignificant, average ranks are descriptive rather than evidence of a globally best method. Nonsignificant pairwise tests are not interpreted as equivalence.

D Numerical and tuning diagnostics

These diagnostics document the numerical conditions under which the reported comparisons were obtained.

Table 5: Final-fit and cross-validation convergence diagnostics. "CV max-iter fits" counts candidate-fold fits that reached the configured iteration cap.

Method	Final conv.	Mean iter.	CV invalid cand.	CV nonconv. fits
DAMP–LiP	100.0%	13107	1396	2821
DAMP–ShiP	100.0%	25133	5300	20402
Pure power	100.0%	10546	570	1424
Ridge	100.0%	–	0	0
Lasso	99.4%	29480	3240	14930
Elastic net	99.4%	31964	8341	39291
Degree-weighted lasso	100.0%	29430	2752	12625
Random forest	100.0%	–	0	0
XGBoost	100.0%	–	0	0
Gaussian process	100.0%	–	0	0
Polynomial OLS	100.0%	–	0	0

All 180 final fits converge for LiP, ShiP, pure-power regularization, and degree-weighted lasso. Lasso and elastic net each converge in 179 of 180 final fits (99.4%). During cross-validation, any candidate with a nonconverged or nonfinite required fold is excluded from selection; every dataset–repetition–method cell retains at least one eligible candidate, so the eligibility rule is never relaxed. The table reports the resulting candidate exclusions and nonconverged fold fits explicitly.

After expanding the search space, the lower regularization boundary is selected in 21.7% of LiP fits, 17.2% of ShiP fits, and 23.9% of pure-power fits; the maximum candidate degree is selected in 21.7%, 23.9%, and 19.4%, respectively. ShiP selects an endpoint of the ε grid in 76.1% of fits, while

Table 6: Hyperparameter boundary-selection rates. Degree "min/max" refers to the smallest/largest candidate degree; other rows refer to the corresponding tuning grid.

Method	Parameter	At min	At max
DAMP–LiP	λ	21.7%	0.0%
DAMP–LiP	degree	15.0%	21.7%
DAMP–LiP	A	67.2%	16.1%
DAMP–ShiP	λ	17.2%	0.0%
DAMP–ShiP	degree	12.2%	23.9%
DAMP–ShiP	ε	42.2%	33.9%
DAMP–ShiP	q	45.6%	30.6%
Pure power	λ	23.9%	0.0%
Pure power	degree	18.9%	19.4%
Ridge	λ	13.9%	0.0%
Ridge	degree	14.4%	23.9%
Lasso	λ	9.4%	0.0%
Lasso	degree	13.9%	8.9%
Elastic net	λ	5.6%	0.0%
Elastic net	degree	11.7%	20.0%
Elastic net	ℓ_1 ratio	46.7%	36.1%
Degree-weighted lasso	λ	8.9%	0.0%
Degree-weighted lasso	degree	17.2%	11.7%

$q = 1$ is itself the theory-constrained edge of the admissible set rather than an unconstrained tuning boundary. These diagnostics define the search-space scope of the empirical results; the mathematical degree-monotonicity results are independent of the tuning grids. The supplementary CSV reports the same diagnostics for every method and dataset.