WHAT DETERMINES THE OPTIMAL TARGET TRANSFORM? RESOLUTION VERSUS RETRANSFORMATION BIAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Regression models are often trained on a transformed target, such as its logarithm, and their predictions are then mapped back. We study which transform gives the smallest error on the original scale and show that the answer is set by two competing effects. The first is a retransformation bias: mapping predictions back shifts them away from the conditional mean by an amount that grows with the noise in the target, so this effect favors transforms close to the identity. The second is resolution: the transform decides where along the target range the model spends its limited precision, so this effect favors stretching the ranges where most targets lie. We derive an expression that contains these two effects and prove that no rule based on the target distribution alone can pick the best transform for every dataset. In controlled experiments with a fixed target distribution, the best transform moves with the predictability of the target as the analysis predicts, and removing the bias moves it toward stronger compression. Trained models keep their error roughly constant in the units in which they are trained, so standardizing the transformed target matters, but their error is far from uniform across the target range. On 41 tabular datasets, a simple statistic computed from the training targets predicts where the choice of transform matters, the uniform quantile transform is worse than a cube-root transform on every dataset with a neural network, and choosing the transform on a validation set works best in practice.

1 INTRODUCTION

A regression target that spans orders of magnitude is rarely modelled on its own scale. It is common practice to apply a logarithmic or Box–Cox transformation (Box & Cox, 1964) or a quantile transformation to the target, to train the model on the transformed values, and to map the predictions back to the original scale. The choice matters: on our benchmark a poor transform can multiply the original-scale error several times. Nuyts & Davis (2025) formalized this as the *target transformation problem*, compared transforms across datasets and model classes, and proposed heuristics for when a transform helps and which to consider. What such an empirical account leaves open is *why* the same transform helps on one dataset and hurts on another.

Two classical results each explain one side of the answer. In quantization, a signal is compressed, quantized with uniform precision and expanded back, and the squared reconstruction error is smallest when the compressor's slope grows as the cube root of the signal density (Bennett, 1948; Panter & Dite, 1951). In econometrics, a model fitted to a transformed response is biased for the conditional mean after back-transformation, by an amount set by the conditional distribution (Duan, 1983; Manning & Mullahy, 2001; Silva & Tenreyro, 2006). We call the two effects *resolution* and *retransformation bias*. They pull in opposite directions: resolution rewards stretching the parts of the target range where most targets lie, while the bias penalizes any curvature and so favors the identity.

In this work, we place both effects in one original-scale risk decomposition and follow its consequences. For an idealized learner that fits the transformed conditional mean up to a learning error, the excess risk is, to leading order, the sum of a bias term that grows with the conditional variance and the curvature of the transform and a resolution term that grows with the learning error and with how strongly the transform compresses the target; each classical result is one of its limits. Because the

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.

balance depends on the conditional law, no rule that sees only the target distribution can be optimal for every joint law with that distribution, which we prove with an exact construction that needs no small-noise expansion. When the learning error scales with the variance of the transformed target, the leading-order optimal Box–Cox power of a power-law target has a closed form in the wide-range limit, three quarters instead of the classical two thirds. The two ingredients are classical; what is new is the joint expression, which makes their balance and its dependence on the conditional law analysable, the impossibility theorem, which has no counterpart in the retransformation literature where a fixed pair of options is compared for a fixed model, and the consequences that follow, the closed form above and the condition for learners whose error varies along the target range.

The analysis rests on an idealized learner, so we measure its assumptions where the conditional law is known. The learning error scales with the transform roughly as the variance of the *standardized* training target, but it is far from uniform along the target range, and trained learners reproduce the idealized optimum only partly. The mechanism itself is visible: with the target distribution fixed, the learned optimum moves with predictability, and removing the retransformation bias moves it toward stronger compression, as the expansion predicts. On 41 tabular datasets, effects grow with a leverage statistic computed from training targets, the variance-normalized objective ranks transforms better than the classical one, and the uniform quantile transform is worse than cube-root companding on all of them, while validation remains the better selector.

The main claim is therefore not that one transform is better than another but that the optimal transform is a property of the joint law and the training protocol, not of the targets. We do not propose a new transform. Cube-root companding (CRC) serves as the reference optimum of the resolution term; on real data it is rarely better than a well-chosen standard transform, and the leverage statistic and the variance-normalized objective are diagnostics, not selectors.

2 Related Work

Choosing a target transform. Nuyts & Davis (2025) define the target transformation problem, compare quantile, Yeo–Johnson and log transforms on ten skewed datasets and four model classes, and give heuristics for when a transform helps; we add a risk decomposition and measurements of the learner's error. Benchmark studies treat target transforms as a binary hyperparameter (Grinsztajn et al., 2022); forecasting work derives the bias of back-transformed Box–Cox forecasts (Pankratz & Dudley, 1987) and selects the power with original-scale criteria (Guerrero, 1993; Proietti & Lütkepohl, 2013), for one power family and a linear model and without a resolution term; validation-based selection of preprocessing can overfit (Schneider et al., 2025).

Retransformation. Smearing corrects the bias of a back-transformed prediction nonparametrically (Duan, 1983); heteroscedastic log errors invalidate a single smearing factor (Manning, 1998); whether to model $\log Y$ or Y with a log link depends on the kurtosis and heteroscedasticity of log-scale residuals (Manning & Mullahy, 2001; Silva & Tenreyro, 2006). TranSUN removes retransformation bias during training for an arbitrary transform and a generalized family of models (Yu et al., 2026). Variance-stabilizing transformations (Anscombe, 1948; Bartlett, 1947), transform-both-sides and weighting (Carroll & Ruppert, 2017) and generalized linear models with a link function (McCullagh, 2019) instead target the conditional variance or the form of the conditional mean. These works correct the bias of a given transform or choose between modelling options; we analyse how bias and resolution together determine which transform to use. With an exact bias correction only the resolution term of equation 1 remains, and the optimal transform then no longer moves toward the identity at low predictability; Section 5 tests this in a controlled setting.

Learned transformations. ACE and AVAS (Breiman & Friedman, 1985; Tibshirani, 1988), warped Gaussian processes (Snelson et al., 2003) and conditional transformation models (Hothorn et al., 2018) learn h by correlation, variance stabilization or likelihood, not by original-scale risk. ACE and AVAS predict by inverting h and inherit its retransformation bias; warped GPs and transformation models integrate a predictive distribution and do not.

Quantization. Bennett's integral and the Panter–Dite law (Bennett, 1948; Panter & Dite, 1951; Gersho, 1979; Gray & Neuhoff, 1998) give the optimal point density of a quantizer; when a function of the source is computed, the density acquires a sensitivity weight (Misra et al., 2011), and under

noise the relevant source is the conditional mean (Wolf & Ziv, 1970). Our resolution term is of this type; the bias term has no quantization analogue.

3 A JOINT EXPANSION

Setup. Y takes values in $\mathcal{Y} = [a, b]$ with density $q > 0$; X are arbitrary covariates, $\mu(x) = \mathbb{E}[Y \mid x]$, $v(x) = \mathrm{Var}(Y \mid x)$. A transform h is a strictly increasing C^3 bijection of $\mathbb{R}$ with $h' \geq c_0 > 0$ and h'', h''' bounded (a C^3 transform on $\mathcal{Y}$ with $h' > 0$ can be extended so); $\kappa_h = h''/h'$. The extension makes h^{-1} defined for every prediction; implementations clip instead, which does not change the leading-order terms below. The excess risk of $\hat{y}$ is $\mathbb{E}[(Y - \hat{y})^2] - \mathbb{E}[v(X)]$.

Definition 1 (Idealized learner)**.** *Trained on $h(Y)$, the learner outputs $\hat{s}(x) = m_h(x) + \sqrt{r_h}\, Z$ with $m_h(x) = \mathbb{E}[h(Y) \mid X = x]$ and Z independent of (X, Y) with a fixed law, $\mathbb{E}[Z] = 0$, $\mathbb{E}[Z^2] = 1$, $\mathbb{E}[Z^6] < \infty$. The prediction is $\hat{y} = h^{-1}(\hat{s})$.*

The learner is Bayes-optimal in transformed space up to a learning error whose *size* r_h may depend on h but whose *shape* does not; this rules out errors whose rare large values do not shrink with r_h. Definition 1 is meant as a model of how transforms are *ranked*, not of the level of the risk: trained learners violate its homogeneity strongly (Section 5), and Proposition 1 and Corollary 2 show what changes when they do. Two idealizations of how r_h depends on h:

Assumption A1 (Fixed error)**.** *$r_h = r$ for every h with $h(a) = -1$, $h(b) = 1$.*

Assumption A2 (Fixed normalized error)**.** *$r_h = \tau^2 \mathrm{Var}[h(Y)]$ for every h.*

Theorem 1 (Small-noise expansion)**.** *Under Definition 1, if $\mathbb{E}[|Y - \mu|^3 \mid x] \leq C v^{3/2}$ for all x, then with $\bar{v} = \sup_x v(x)$*

$$\mathbb{E}[(Y - \hat{y})^2] - \mathbb{E}[v(X)] = \underbrace{\mathbb{E}\Big[\tfrac{1}{4}\kappa_h(\mu)^2\, v^2\Big]}_{\text{retransformation bias}} + \underbrace{r_h\, \mathbb{E}\big[h'(\mu)^{-2}\big]}_{\text{resolution}} + O\big(\bar{v}\, r_h + r_h^{3/2} + \bar{v}^{5/2}\big). \quad (1)$$

Under A2 the leading terms are invariant to $h \mapsto ch + d$ ($c > 0$).

The proof (Appendix A) expands $m_h = h(\mu) + \frac{1}{2}h''(\mu)v + \dots$ and h^{-1}; the cross term vanishes because Z is independent of the regression noise, and the moment condition on Z controls the remainder. The first term is the square of the second-order retransformation bias $h^{-1}(\mathbb{E}[h(Y) \mid x]) - \mu \approx \frac{1}{2}\kappa_h v$ (Duan, 1983); the second is Bennett's integral over the distribution of $\mu(X)$.

Corollary 1 (Limits and optimality condition)**.** *Let p_μ be the density of $\mu(X)$ and $w = h'$. (i) If $v \equiv 0$, then $p_\mu = q$; under A1 the leading term $rJ[h]$, $J[h] = \int q/h'^2$, is minimized uniquely by $h' \propto q^{1/3}$ (Panter–Dite); under A2 it is $\tau^2 K[h]$, $K[h] = \mathrm{Var}[h(Y)]\, J[h]$. (ii) As $r_h \to 0$ the leading term vanishes iff $\kappa_h(\mu(X))\, v(X) = 0$ a.s. If moreover $p_\mu > 0$ on an interval I and $v > 0$ a.s. on $\{\mu(X) \in I\}$, then h is affine on I; in particular, if $p_\mu > 0$ on $[a, b]$ and $v > 0$ a.s., h is affine on $[a, b]$. (iii) Let $\tilde{v}^2(t) = \mathbb{E}[v(X)^2 \mid \mu(X) = t]$. Under A1, a smooth minimizer satisfies, with $\phi = \log w$ and a multiplier λ, $(p_\mu \tilde{v}^2 \phi')' = 2\lambda e^{\phi} - 4r p_\mu e^{-2\phi}$ and $p_\mu \tilde{v}^2 \phi' = 0$ at the boundary.*

Where h must bend is thus decided by the balance of $\tilde{v}^2$ (which flattens it) and r (which pulls it toward the cube-root law); the two terms are comparable when $v \sim \sqrt{r}$. Part (ii) is local: a transform may bend freely where the conditional mean carries no conditional noise, and on disjoint intervals it may be affine with different slopes.

Proposition 1 (Heterogeneous and biased learning error)**.** *If instead $\hat{s}(x) = m_h(x) + b_h(x) + \sqrt{r_h(x)}\, Z$ with bounded b_h, r_h and $\bar{b} = \sup |b_h|$, $\bar{r} = \sup r_h$, then under the conditions of Theorem 1*

$$\mathbb{E}[(Y - \hat{y})^2] - \mathbb{E}[v(X)] = \mathbb{E}\Big[\Big(\tfrac{1}{2}\kappa_h(\mu)v + \frac{b_h(X)}{h'(\mu)}\Big)^2\Big] + \mathbb{E}\Big[\frac{r_h(X)}{h'(\mu)^2}\Big] + O\big((\bar{v} + \bar{b})(\bar{r} + \bar{b}^2) + \bar{r}^{3/2} + \bar{v}^{5/2}\big).$$

A systematic error b_h enters as an additional bias that can cancel or add to the retransformation bias, and an x-dependent error size weights the resolution term by where the learner is inaccurate. Both are what we measure in trained learners in Section 5. The optimality condition changes accordingly:

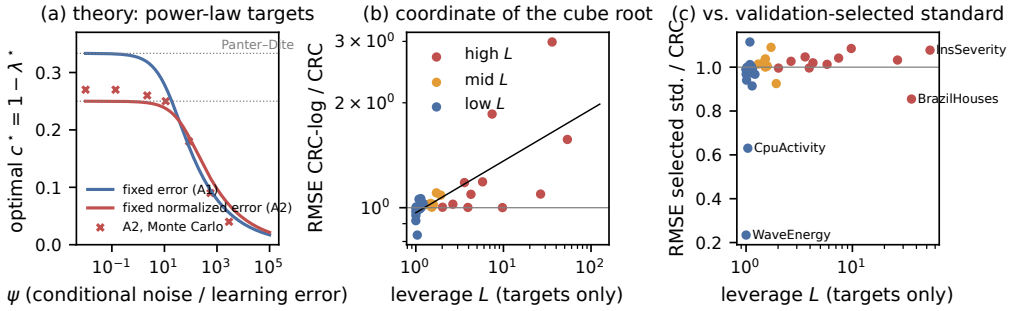

Figure 1: (a) Optimal Box–Cox $c^\star = 1 - \lambda^\star$ for power-law targets (Theorem 2) against the ratio ψ of conditional noise to learning error (left-hand sides of equation 2); crosses: Monte Carlo of the idealized learner under A2 ($U = 10^6$). (b,c) 41 datasets, MLP, mean per-seed RMSE ratio over 20 seeds against the leverage L (median over splits of L computed from training targets): (b) CRC-log over CRC, line: log-log fit; (c) validation-selected standard transform (raw, log1p, Box–Cox) over CRC.

Corollary 2 (Weighted resolution)**.** *If $v \equiv 0$, $b_h \equiv 0$ and $r_h(x) = r\,\omega(\mu(x))$ with a weight $\omega > 0$ that does not depend on h, then under the normalization of A1 the leading term $r \int q\,\omega/h'^2$ is minimized uniquely by $h' \propto (q\,\omega)^{1/3}$; with $r_h(x) = \tau^2 \mathrm{Var}[h(Y)]\,\omega(\mu(x))$ the objective becomes $K_\omega[h] = \mathrm{Var}[h(Y)] \int q\,\omega/h'^2$.*

The transform should thus stretch the target where the learner is inaccurate, as in task-weighted quantization (Misra et al., 2011). For trained learners, however, ω itself depends on h (Section 5), so the corollary gives a fixed-point condition rather than a rule that can be evaluated from the targets alone.

A family and a leverage statistic. Writing $h' \propto q^a (y+\epsilon)^{-b}$ gives one family containing raw targets $(0,0)$, Box–Cox with power λ on $y+\epsilon$ $(0, 1-\lambda)$, the log $(0,1)$, the uniform quantile ("equal-mass") transform $h = 2F - 1$ $(1,0)$, cube-root companding in y (**CRC**, $(1/3, 0)$) and in $\log(y+\epsilon)$ (**CRC-log**, $(1/3, 2/3)$). Under A1 with $v \equiv 0$, the equal-mass transform's excess over CRC is $\int q^{-1} / (\int q^{1/3})^3 \geq 1$, with equality only for uniform q (Appendix A): it is large when parts of the support carry little mass. To summarize how much the resolution term can distinguish transforms we use the *leverage* $L = J[\text{CRC-log}]/J[\text{CRC}] \geq 1$, the first-order cost of placing the cube root in the wrong coordinate, computed from the training targets of each split (Appendix H). Differences between transforms that act through resolution should grow with L; this is a heuristic reading of the resolution term, not a theorem.

A closed form under fixed normalized error. For power-law targets the leading-order optimum can be written down.

Theorem 2 (Leading-order Box–Cox power in the wide-range limit)**.** *Let μ be log-uniform on $[1, U]$, $\ell = \ln U$, and $Y = \mu(1 + \sigma\varepsilon)$ with ε independent of μ, bounded, $\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}[\varepsilon^2] = 1$, and $\sigma|\varepsilon| < 1$, so that Y has compact support, $\mathbb{E}[Y \mid \mu] = \mu$ and $v = \sigma^2\mu^2$. Let $h' \propto y^{-c}$ (Box–Cox with $\lambda = 1 - c$), $0 < c < 1$. Under A1 normalize h on $[1, U]$, the range of μ (normalizing on the support of Y multiplies the resolution term by $1 + O(\sigma)$). Take the leading terms of equation 1 and, for fixed σ, r and τ, their leading order as $\ell \to \infty$ (dropping terms of relative order σ^2 in $\mathrm{Var}h(Y)$). This leading-order objective (not the exact risk) is minimized at the unique $c^\star$ solving*

$$A1: \ \frac{\sigma^4}{r} = \frac{1 - 3c}{2c^3(1-c)^3}, \qquad A2: \ \frac{\sigma^4 \ell}{\tau^2} = \frac{1 - 4c}{c^3(1-c)^4}. \tag{2}$$

Thus $c^\star = 1/3$ ($\lambda = 2/3$) under A1 and $c^\star = 1/4$ ($\lambda = 3/4$) under A2 when $\sigma = 0$, $c^\star$ decreases to 0 (raw targets) as the left-hand side grows, and $c^\star \leq 1/3$: in this model the logarithm ($c = 1$) is never optimal.

The statement concerns the minimizer of the leading-order objective, taken first in the small-noise expansion and then as $\ell \to \infty$; since $\bar{v} = \sigma^2 U^2$ grows with U, the remainder of Theorem 1 is

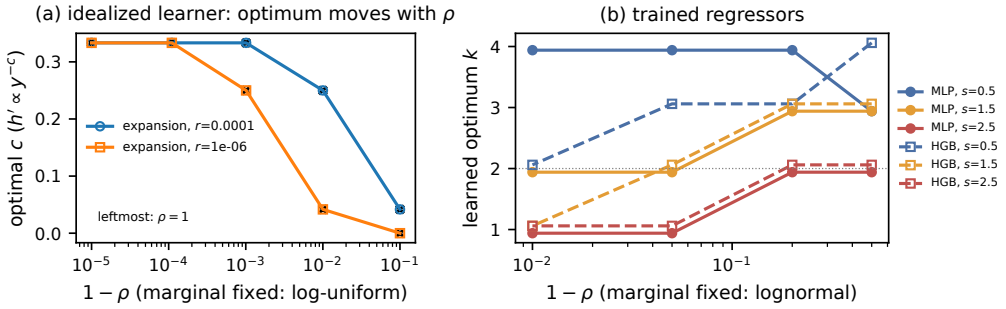

Figure 2: The optimum moves with predictability at a fixed marginal. (a) Idealized learner under A1, log-uniform marginal on $[1, 10^4]$: minimizer of the expansion (lines) and of a Monte Carlo (crosses) over $h' \propto y^{-c}$ ($c = 1/3$ is CRC, $c = 0$ raw). (b) Trained MLP (standardized target) and HGB, lognormal marginal: learned optimum over the sweep $h' \propto q^{1/3}(y + \epsilon)^{(k-2)/3}$ ($k = 2$ is CRC; larger k is closer to raw), 10 seeds.

not uniform in this limit and we do not claim that $c^\star$ minimizes the exact risk. The left-hand sides compare conditional noise with learning error (Figure 1a). A Monte Carlo of the idealized learner with $U = 10^6$ reproduces the exact-moment optimum at all grid points but one (Appendix B).

4 No Rule on the Target Distribution Alone

A rule Φ maps a target density q to a transform $\Phi(q)$; it is *asymptotically optimal* on a joint law P if the excess risk of $\Phi(q)$ is at most $(1 + o(1))$ times the infimum over transforms as the learning error vanishes ($r \to 0$ under A1, $\tau \to 0$ under A2).

Theorem 3 (Impossibility)**.** *Let q be continuously differentiable and positive on $[a, b]$. (i) Under A1, if q is not uniform, and (ii) under A2, for every such q, there are joint laws P_1, P_2 with marginal q such that no rule is asymptotically optimal on both.*

Proof sketch. P_1 is deterministic, $Y = X \sim q$: by Theorem 1 its excess risk is $r_h J[h]$ to first order, so asymptotic optimality forces $\Phi(q)$ to be the Panter–Dite transform (A1) or a minimizer of K (A2); neither is affine—under A2 because an affine h is never a stationary point of K on a compact support (Lemma 1). P_2 draws two independent values from q, reveals the unordered pair as X, and sets Y to one of them at random. Its marginal is q, and as $r_h \to 0$ the prediction tends to the quasi-midpoint $h^{-1}(\frac{1}{2}h(y_1) + \frac{1}{2}h(y_2))$, which equals the midpoint for all pairs only if h is affine (Jensen's functional equation); by continuity a non-affine h misses it on a set of pairs of positive probability, so its excess risk stays bounded away from zero while an affine h has excess $O(r_h)$. This part is exact and uses no expansion (Appendix A).

The theorem makes precise a point implicit in the retransformation literature (Manning & Mullahy, 2001): the transform must depend on the *joint* law. Its quantitative content is the path between the two extremes. Fixing a log-uniform marginal and lowering the correlation ρ of a Gaussian copula moves the optimal power of the idealized learner continuously from the cube-root law toward the identity; the minimizer of the expansion agrees with a Monte Carlo in all 10 settings we tried (Figure 2a, Appendix B).

5 How Well Does the Idealization Describe Trained Learners?

Controlled regression. We fix the target marginal to $\mathrm{Lognormal}(0, s^2)$ and draw $y = \exp\{s(\rho f(X) + \sqrt{1-\rho^2}\, e)\}$, $X \sim N(0, I_{10})$, with a nonlinear score f standardized to $N(0, 1)$ and $e \sim N(0, 1)$: the marginal is the same for every ρ, and $m_h(x) = \mathbb{E}[h(Y) \mid x]$ is computable by quadrature (Appendix C). For $s \in \{0.5, 1.5, 2.5\}$ (1.7 to 8 decades) and $\rho \in \{0.5, 0.8, 0.95, 0.99\}$, 10 seeds, $n = 10{,}000$, we train on the sweep $h' \propto q^{1/3}(y + \epsilon)^{(k-2)/3}$, $k = 0, \ldots, 4$ ($k = 0$ is CRC-log, $k = 2$ CRC; larger k is closer to raw), an MLP on the standardized transformed target

Table 1: Testing Definition 1 (12 cells, 10 seeds). *Spread*: median / max over cells of max/min over the five transforms of $\hat{r}_h$ (abs.) and $\hat{r}_h/\mathrm{Var}[h(Y)]$ (norm.). ρ_S: mean Spearman correlation between true excess risk and that of the idealized learner under A1 or A2 (calibrated to the cell median). *Decile ratio*: max/min of the learning error over deciles of $m_h(x)$. Last columns: idealized over true excess risk with scalar or decile-wise $\hat{r}_h$ (median).

learner	abs. spread	norm. spread	ρ_S A1	ρ_S A2	decile ratio	scalar r	decile $r(x)$
MLP, standardized target	9.6 / 500	2.71 / 14.1	0.24	0.88	11.8 / 123	0.45	0.59
MLP, unstandardized	6.7 / 21	7.52 / 21.6	0.45	0.62	4.3 / 21	0.47	0.66
HGB	12.0 / 1009	2.09 / 6.6	0.37	0.77	23.6 / 265	0.59	0.68

(our default), the same MLP on the unstandardized one ($h \in [-1, 1]$), and a histogram gradient-boosted tree (HGB). We measure the learning error *directly*, $\hat{r}_h = \mathrm{mean}_i(\hat{s}(x_i) - m_h(x_i))^2$ on test points, which needs no subtraction of an irreducible part. The lognormal marginal is unbounded and so does not satisfy the compact support of Theorem 1 literally; the experiment tests trained learners, not the theorem. A rerun with score and noise truncated at ± 3 (compact support) gives the same picture (Appendix P): median spreads 4.6 vs. 2.08 for the MLP and 4.5 vs. 2.18 for HGB.

The error is fixed roughly in the units the learner is trained in. With the standardized target, the learning error varies across transforms by a median factor of 9.6 (up to 500) but after dividing by $\mathrm{Var}[h(Y)]$ by 2.71 (up to 14.1, on wide, highly predictable targets); for HGB, which is essentially invariant to rescaling the target, 12.0 against 2.09 (Table 1). Correspondingly, the idealized learner under A2 ranks the transforms by true excess risk better than under A1 (ρ_S 0.88 vs. 0.24 for the MLP, 0.77 vs. 0.37 for HGB). The ratio of absolute to normalized spread has a median over cells of 1.70 $[1.56, 11.50]$ for the MLP and 2.17 $[1.24, 21.82]$ for HGB (95% bootstrap intervals over cells). Without standardization this reverses for the MLP (0.97 $[0.78, 1.14]$): the absolute spread (6.7) is now *smaller* than the normalized one (7.52). A2 is therefore a property of the training protocol rather than of the learner: an MLP keeps its error roughly fixed in the units of the target it is fitted to. It is a better approximation than A1 under the usual standardization, not an exact law (normalized spreads up to $14.1\times$).

The error is not homogeneous in x. Definition 1 lets the error depend on h only through its size. In fact the learning error differs across deciles of $m_h(x)$ by a median factor of 11.8 (MLP; up to 123) and 23.6 (HGB). An idealized learner with the measured scalar $\hat{r}_h$ reproduces only a median 0.45 of the MLP's true excess risk, 0.59 with decile-wise error sizes and 0.60 when the decile-wise mean error b_h of Proposition 1 is added (HGB: 0.59, 0.68, 0.74). Heterogeneity at the resolution of deciles thus accounts for part of the gap; the rest lies in finer variation of the error with x. The expansion captures the ordering of transforms better than the level of their risk.

Does the closed form describe trained learners? On data matching Theorem 2 (exactly log-uniform μ, bounded multiplicative noise, c swept on a grid including $1/4$; Appendix D), HGB's optimum is $c = 0.33$ (the Panter–Dite value; $c = 1/4$ is worse by 1.035 $[1.012, 1.062]$ at $\sigma = 0$, but indistinguishable at two of four noise levels) and the MLP's is $c = 0.00$ (raw targets). Neither confirms the A2 prediction: HGB's absolute error varies less across c than its normalized error, and the MLP concentrates its accuracy where the raw target is large. Theorem 2 describes the idealized learner; it is not a recommendation.

The learned optimum moves with ρ at a fixed marginal. With the marginal, and hence CRC, J and K, identical across ρ, the learned optimum ranges over $k \in \{1, 2, 3, 4\}$ for both learners (Figure 2b): an empirical instance of Theorem 3. On wide, highly predictable targets both prefer $k = 1$, more compressed than CRC, which neither idealization predicts; this is consistent with, but not proven to be caused by, the heavier tail of the regression function under raw-like transforms (Appendix F).

Removing the bias term. Because $\mu(x)$ and $m_h(x)$ are known, we can subtract the exact conditional retransformation bias $h^{-1}(m_h(x)) - \mu(x)$, leaving only the resolution term (Appendix E). At $\rho = 0.5$ the learned optimum then moves from $k \in \{2, 3, 4\}$ to $k \in \{0\}$, the most compressed members; at $\rho \geq 0.95$ it is unchanged in 10 of 12 cells, and it is at least as compressed as without

Table 2: Median over datasets of the mean per-seed RMSE ratio *method/CRC* (MLP, 20 seeds). +*smearing*: selected standard vs. CRC with global smearing (Duan, 1983) on both sides. *sig.*: datasets on which CRC is significantly better / worse than the selected standard transform (corrected resampled t-test, Holm over datasets). Per-dataset values in Appendix I.

L group	n	U.q.	G.q.	YJ	Log1p	BC CV	Std	Std+G.q.	Std, sm.	CRC-log	Std sig. $+/-$
high ($L \geq 2$)	11	2.873	1.161	1.218	1.035	0.998	1.027	1.020	1.015	1.094	0/0
mid	8	1.191	1.086	1.064	1.053	0.989	1.010	1.010	1.006	1.021	0/0
low ($L < 1.3$)	22	1.109	1.057	0.999	1.001	0.976	0.985	0.985	0.985	0.999	0/2

correction in 22 of 24 cells (the exceptions are HGB on the widest, most predictable targets). The move toward the identity at low predictability is thus produced by the bias term, as the expansion says; global smearing removes only part of it. A *learned* correction (a boosted regressor fitted on validation rows to $y - h^{-1}(\hat{s})$, in the spirit of TranSUN (Yu et al., 2026) but not a reimplementation) moves the MLP optimum the same way ($k \in \{0\}$ at $\rho = 0.5$; at least as compressed in 11 of 12 cells); the oracle bounds what any such correction can achieve.

6 When Does the Choice Matter? Forty-One Datasets

Protocol. We use every OpenML-CTR23 regression task with a nonnegative target (Fischer et al., 2023), the ten datasets of Nuyts & Davis (2025), and six further wide-range datasets, keeping duplicates once: 41 datasets with 167 to 72,000 rows (Appendix G). Each of 20 seeds draws a $60/20/20$ split; all target statistics, including L, are computed from the training split; datasets are grouped by the median L over splits into high ($L \geq 2$, 11), mid (8) and low ($L < 1.3$, 22). Transformed targets are standardized; the model is an MLP (128,64) unless stated. Candidates: raw, log1p, Box–Cox with maximum-likelihood λ, Gaussian and uniform quantile transforms, and family members: 20 transform shapes $\times$ 3 data-relative ϵ (60 candidates), or, in splits with training $L < 1.3$, 10 shapes $\times$ 3 ϵ plus equal-mass (31). *Box–Cox CV* selects among the 15 Box–Cox members on validation; the *selected standard* transform is the best of raw, log1p and Box–Cox on validation; CRC selects its ϵ on validation (CRC_{med} fixes it at the median). We report per-seed RMSE ratios *method/CRC* (> 1 favours CRC) and, because the 20 splits of a dataset overlap, test each dataset with the corrected resampled t-test of Nadeau & Bengio (1999), Holm-corrected over the 41 datasets (the anticonservative Wilcoxon test is reported in Appendix Q). Datasets are the independent units, so claims about groups rest on effect sizes and on tests across datasets, not on counts of significant datasets. Excluding the six datasets we added changes none of the conclusions below (Appendix G).

Effects grow with leverage. CRC-log and CRC differ mainly in resolution, and their learned RMSE ratio rises with L (Spearman 0.76, $p < 10^{-8}$; Figure 1b), from about 1 at low leverage to a group median of 1.094 at high leverage. The ratio of the selected standard transform to CRC also grows with L (0.54, $p < 10^{-3}$; Figure 1c), but it is small: a median of 1.027 at high leverage, 1.015 with smearing on both sides, and 0.998 against Box–Cox CV (Table 2). Per dataset, CRC is significantly better than the selected standard transform on 0 and worse on 2 datasets. Across datasets (one-sample Wilcoxon of the per-dataset log ratios), the high-leverage advantage of CRC is not significant ($p = 0.1$), whereas at low leverage, where Box–Cox CV beats CRC with a median ratio of 0.976, CRC is significantly worse ($p = 8 \times 10^{-4}$). Global smearing uses one residual distribution for all x; under heteroscedasticity it does not remove the retransformation bias (Manning, 1998), so we read the smeared comparison as a robustness check, not as a bias-free one.

Uniform quantile transforms, and boundary handling. The uniform quantile transform is worse than CRC on all 41 datasets (across datasets $p = 9 \times 10^{-13}$; per dataset significantly on 18), with a median ratio of 2.873 at high leverage. Its first-order excess $\int q^{-1} / (\int q^{1/3})^3$ is large when parts of the support carry little mass. Because CRC extends its map past the training support while the baselines clip to it, we reran CRC and the equal-mass family member with the baselines' clipping (Appendix J). Clipping changes CRC's RMSE by a median factor of 0.999; with identical estimator and boundary handling, equal-mass is worse than CRC on all 41 datasets (18 individually significant), and scikit-learn's uniform quantile transform is within about 1% of it (median 0.994). The failure is a property of the transform's shape, not of its implementation.

Table 3: Left: mean Spearman correlation, over all family members (law $\times \epsilon$, 31–60 per dataset), between learned test MSE and J or K computed from training targets only; datasets where K ranks better. Right: selector regret, test RMSE of the selected candidate over the best candidate of that split in hindsight (median over seeds, mean over datasets). argmin J, argmin K and $\mathrm{CRC}_{\mathrm{med}}$ use training targets only; CRC chooses ϵ on validation; validation chooses among all candidates.

L group	n	$\rho_S(J)$	$\rho_S(K)$	$K > J$	argmin J	argmin K	$\mathrm{CRC}_{\mathrm{med}}$	CRC	Log1p	validation
high	11	0.667	0.837	11/11	1.160	1.088	1.152	1.149	1.131	1.066
mid	8	0.717	0.835	6/8	1.052	1.041	1.059	1.046	1.104	1.025
low	22	-0.183	0.337	17/22	1.261	1.068	1.256	1.246	1.100	1.007

Gaussian quantile transform. The Gaussian quantile and Yeo–Johnson transforms, which Nuyts & Davis (2025) found most useful for linear models, are worse than CRC at high leverage (medians 1.161 and 1.218); adding the Gaussian quantile transform to the selected standard set changes the median ratios only to 1.020, 1.010 and 0.985 (Table 2).

Design choices. Neither the reduced family at low leverage nor the training protocol drives these results (Appendices L, M).

Nearly deterministic targets. CRC's largest losses occur on targets that the features determine almost completely: WaveEnergy (raw-target $R^2 = 1.000$; ratio 0.23), CpuActivity (0.978; 0.63) and BrazilHouses (ridge $R^2 = 0.95$; 0.85 despite $L = 35.9$). The bias term cannot explain this, since it vanishes as $v \to 0$; the A2 resolution term K can, because it penalizes CRC's strong compression of the sparse extremes: the member minimizing K on the training targets is close to the identity on these ranges, with regrets 1.139, 1.384 and 1.131 against CRC's 4.624, 1.655 and 1.973, and CRC's excess error sits in the sparse extreme deciles (Appendix N). This is consistent, not verified: on real data we cannot separate conditional noise from learning error.

Other learners. Because the claim is that the optimum depends on the learner, we repeated the comparison with XGBoost (Chen & Guestrin, 2016) and a tabular ResNet (Gorishniy et al., 2021) on all 41 datasets (10 splits; Appendix O) and with HGB and ridge regression. The qualitative findings hold for every learner, although with 10 splits we read effect sizes rather than per-dataset tests: the uniform quantile transform is worse than CRC on 38 (XGBoost) and 41 (ResNet) of 41 datasets, and CRC-log over CRC grows with L (Spearman 0.38 and 0.54). The size of the effects does depend on the learner: for HGB, XGBoost and the ResNet the selected standard transform is within about 1% of CRC in every group (XGBoost medians 1.008, 1.001, 0.992; ResNet 1.005, 1.002, 0.995), whereas the uniform quantile and CRC-log members are penalized most by the neural learners (ResNet high-leverage medians 2.004 and 1.079). Poisson and gamma losses for HGB are within 2% of CRC. With ridge regression on the datasets of Nuyts & Davis (2025) and our protocol, the best transform is between 0.913 and 1.026 of raw-target RMSE, far from their large linear-model gains. To locate the difference we reran their protocol from their released code (5×2 cross-validation, min-max features, ridge with a validation-tuned penalty; Appendix R). Evaluated with their evaluation code, our rerun reproduces their gains (median RSE change of the Gaussian quantile transform over raw targets -13.3%, Yeo–Johnson -5.5%). That code, however, divides the squared error by $\sum_i(\bar{y} - \hat{y}_i)^2$ rather than by $\sum_i(y_i - \bar{y})^2$ as in their Eq. 8, a quantity that is not monotone in the MSE; with Eq. 8 the same predictions give $+6.0\%$ and $+1.8\%$ (RMSE: $+2.9\%$, $+0.8\%$). The disagreement is therefore in the metric, not in the models. A linear model's error also depends on how well h linearizes the conditional mean, a misspecification effect outside Definition 1.

7 Implications for Choosing a Transform

Target-only objectives explain; validation selects. This section is about selection, not about CRC. Computed from the training targets alone, K orders the family members by learned error better than J on 34 of 41 datasets, with a mean gain in rank correlation of 0.35 $[0.21, 0.50]$ (Table 3). As a selector, however, argmin K beats fixed CRC but not validation: its regret exceeds that of validation by 0.042 $[0.021, 0.070]$, while fixed CRC exceeds it by 0.154 $[0.034, 0.362]$ (95% dataset-bootstrap

Table 4: Selector regret against training size on the high-leverage datasets with at least n_{train} training rows (*sets*): median over 20 seeds and mean over datasets, and 90th percentile over seeds and datasets (q90). Candidates here are the standard transforms, the $\alpha = 1/3$ sweep and equal-mass (a subset of those of Table 3); the validation split has $\max(n_{\text{train}}/3, 30)$ rows and the test split is fixed. Bootstrap intervals in Appendix K.

n_{train}	sets	validation	argmin K	CRC	val. q90	K q90
250	10	1.091	1.106	1.100	1.644	1.560
500	8	1.034	1.077	1.070	1.277	1.321
1000	7	1.023	1.040	1.093	1.222	1.129
2000	6	1.011	1.254	1.131	1.246	1.427
5000	6	1.015	1.045	1.107	1.187	1.235

intervals). As Theorem 3 predicts, the best transform depends on the conditional law, which a validation split measures and no target-only statistic sees.

Small data. Validation noise could dominate on small data (Schneider et al., 2025), but on subsampled high-leverage datasets validation has the lower median regret at every size (Table 4), while the 90th percentiles are not ordered and all their differences have intervals containing zero; validation is the better selector typically, not demonstrably the safer one in the tail.

Practical reading. For squared error, avoid the uniform quantile transform on targets with sparse regions. Compute L first: below about 1.3, the identity did as well as any transform. Above it, include the cube-root family among the candidates and choose on validation, with the transformed targets standardized in the same way for every candidate. On nearly deterministic targets, expect transforms close to the identity to win.

8 Discussion

Derived, observed, conjectured. The expansion, the impossibility theorem and the closed form are derived under Definition 1; the behaviour of trained learners and the benchmark results are observed. Two explanations remain conjectures: that near-identity transforms win on nearly deterministic targets because of the A2 resolution term and error heterogeneity, and that learners compress wide, predictable targets strongly because of the tail of the regression function.

Scope. The theory is explanatory: it identifies the forces that move the optimal transform and the direction in which they move it, not the exact optimum of a trained learner. It concerns smooth transforms of continuous targets, while the estimator is piecewise linear on empirical cells; the two agree as cells shrink, without a finite-sample bridge. How learners allocate their error along the target range, the ω of Corollary 2, is the natural next question. The study is tabular, with nonnegative targets and squared error; image and text targets, other losses and learned transforms are left open.

9 Conclusion

We showed that the transform minimizing original-scale error is set by two competing effects, resolution and retransformation bias, and derived an expression that contains both, with the cube-root law of quantization and the retransformation bias of econometrics as its limits. From it we proved that no rule based on the target distribution alone can be optimal for every joint law, and obtained the leading-order optimal Box–Cox power for an idealized learner. Measuring the idealization directly, we found that trained learners keep their error roughly fixed in the units they are trained in but spread it unevenly along the target range, and that with the target distribution fixed the optimum moves with predictability through the bias term. On 41 datasets, a leverage statistic computed from the training targets indicates when the choice can matter, the variance-normalized objective explains how transforms rank, and validation decides which to choose. The optimal target transform is thus a property of the joint law and the training protocol, not of the targets.

AI USE STATEMENT

An LLM-based assistant was used for coding, running and aggregating experiments, formatting LaTeX tables, assisting with mathematical derivations and reviewing proofs for correctness, and editing the manuscript. The authors independently verified all experimental results, mathematical arguments, and proofs, and take full responsibility for the final content.

ETHICS STATEMENT

This work uses public, de-identified tabular datasets and involves no human subjects or personal data beyond what those releases contain.

REPRODUCIBILITY STATEMENT

Proofs are in Appendix A, numerical checks of the theory in Appendix B, datasets and preprocessing in Appendix G, and algorithms and settings in Appendix H. Anonymized code, per-seed results and the script that regenerates every table, figure and in-text number are in the supplementary material.

REFERENCES

Francis J Anscombe. The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254, 1948.

Maurice S Bartlett. The use of transformations. *Biometrics*, 3(1):39–52, 1947.

William Ralph Bennett. Spectra of quantized signals. *The Bell System Technical Journal*, 27(3): 446–472, 1948.

George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243, 1964.

Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

Raymond J Carroll and David Ruppert. *Transformation and weighting in regression*. Chapman and Hall/CRC, 2017.

Tianqi Chen and Carlos Guestrin. 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.

Paulo Cortez and ADJR Morais. A data mining approach to predict forest fires using meteorological data. pp. 512–523, 2007.

Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011.

Naihua Duan. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383):605–610, 1983.

Christophe Dutang, Arthur Charpentier, and Maintainer Christophe Dutang. Package 'casdatasets', 2020.

Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez. A proactive intelligent decision support system for predicting the popularity of online news. In *Portuguese conference on artificial intelligence*, pp. 535–546. Springer, 2015.

Sebastian Felix Fischer, Matthias Feurer, and Bernd Bischl. Openml-ctr23–a curated tabular regression benchmarking suite. In *AutoML Conference 2023 (Workshop)*, 2023.

Allen Gersho. Asymptotically optimal block quantization. *IEEE Transactions on information theory*, 25(4):373–380, 1979.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. volume 34, pp. 18932–18943, 2021.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE transactions on information theory*, 44 (6):2325–2383, 1998.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? volume 35, pp. 507–520, 2022.

Victor M Guerrero. Time-series analysis supported by power transformations. *Journal of forecasting*, 12(1):37–48, 1993.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

Torsten Hothorn, Lisa Möst, and Peter Bühlmann. Most likely transformations. *Scandinavian Journal of Statistics*, 45(1):110–134, 2018.

Willard G Manning. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of health economics*, 17(3):283–295, 1998.

Willard G Manning and John Mullahy. Estimating log models: to transform or not to transform? *Journal of health economics*, 20(4):461–494, 2001.

Peter McCullagh. *Generalized linear models*. Routledge, 2019.

Vinith Misra, Vivek K Goyal, and Lav R Varshney. Distributed scalar quantization for computing: High-resolution analysis and extensions. *IEEE Transactions on Information Theory*, 57(8):5298–5325, 2011.

Andy Moya and Roberto J López-Sastre. Stellar mass and radius estimation using artificial intelligence. *Astronomy & Astrophysics*, 663:A112, 2022.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Advances in neural information processing systems*, 12, 1999.

Loren Nuyts and Jesse Davis. The when and how of target variable transformations. In *International Symposium on Intelligent Data Analysis*, pp. 113–126. Springer, 2025.

Alan Pankratz and Underwood Dudley. Forecasts of power-transformed series. *Journal of Forecasting*, 6(4):239–248, 1987.

Philip F Panter and Ward Dite. Quantization distortion in pulse-count modulation with nonuniform spacing of levels. *Proceedings of the IRE*, 39(1):44–48, 1951.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Tommaso Proietti and Helmut Lütkepohl. Does the box–cox transformation help in forecasting macroeconomic time series? *International Journal of Forecasting*, 29(1):88–99, 2013.

Lennart Schneider, Bernd Bischl, and Matthias Feurer. Overtuning in hyperparameter optimization. In *Proceedings of the Fourth International Conference on Automated Machine Learning*, volume 293 of *Proceedings of Machine Learning Research*, pp. 17/1–43. PMLR, 08–11 Sep 2025.

JMC Santos Silva and Silvana Tenreyro. The log of gravity. *The Review of Economics and statistics*, pp. 641–658, 2006.

Edward Snelson, Zoubin Ghahramani, and Carl Rasmussen. Warped gaussian processes. volume 16, 2003.

Robert Tibshirani. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83(402):394–405, 1988.

J Wolf and Jacob Ziv. Transmission of noisy information to a noisy receiver with minimum distortion. *IEEE Transactions on Information Theory*, 16(4):406–411, 1970.

Jiahao Yu, Haozhuang Liu, Yeqiu Yang, Lu Chen, Jian Wu, Yuning Jiang, and Bo Zheng. Transun: A preemptive paradigm to eradicate retransformation bias intrinsically from regression models in recommender systems. volume 38, pp. 140918–140954, 2026.

A PROOFS

Throughout, $\varepsilon = Y - \mu(X)$, so $\mathbb{E}[\varepsilon \mid x] = 0$ and $\mathbb{E}[\varepsilon^2 \mid x] = v(x)$; $\kappa = \kappa_h(\mu)$, $h' = h'(\mu)$, and $\eta = \sqrt{r_h}Z$. Because h is a C^3 bijection of $\mathbb{R}$ with $h' \geq c_0$ and bounded h'', h''', h^{-1} is C^3 on $\mathbb{R}$ with bounded first three derivatives, so every Taylor remainder below holds globally with a constant that does not depend on the point.

Theorem 1. (i) $m_h(x) = \mathbb{E}[h(\mu+\varepsilon) \mid x] = h(\mu) + \frac{1}{2}h''(\mu)v + \rho_1$ with $|\rho_1| \leq \frac{1}{6}\sup|h'''|\,\mathbb{E}[|\varepsilon|^3 \mid x] \leq C_1 v^{3/2}$. (ii) Let $\Delta = m_h + \eta - h(\mu) = \frac{1}{2}h''v + \eta + \rho_1$. Then $h^{-1}(h(\mu)+\Delta) = \mu + \Delta/h' - \frac{1}{2}(h''/h'^3)\Delta^2 + \rho_2$ with $|\rho_2| \leq C_2|\Delta|^3$, and substituting,

$$\hat{y} - \mu = \tfrac{1}{2}\kappa v + \eta/h' - \tfrac{1}{2}\kappa\,\eta^2/h'^2 + \rho, \qquad |\rho| \leq C\big(v^{3/2} + v|\eta| + |\eta|^3\big),$$

where the η-free part of ρ depends on x only. (iii) $\hat{y}$ is a function of (X, Z) and Z is independent of (X, Y), so $\mathbb{E}[\varepsilon(\hat{y}-\mu) \mid x] = \mathbb{E}[\varepsilon \mid x]\,\mathbb{E}[\hat{y} - \mu \mid x] = 0$ and $\mathbb{E}[(Y-\hat{y})^2 \mid x] = v + \mathbb{E}[(\hat{y}-\mu)^2 \mid x]$. Squaring and taking expectations over Z, with $\mathbb{E}[\eta] = 0$, $\mathbb{E}[\eta^2] = r_h$ and $\mathbb{E}[|\eta|^k] = r_h^{k/2}\mathbb{E}[|Z|^k]$ for $k \leq 6$,

$$\mathbb{E}[(\hat{y}-\mu)^2 \mid x] = \tfrac{1}{4}\kappa^2 v^2 + r_h/h'^2 - \tfrac{1}{2}\kappa^2 v\, r_h/h'^2 - \kappa\,\mathbb{E}[\eta^3]/h'^3 + O\big(v\, r_h + r_h^2 + v^{5/2}\big);$$

the cross terms of ρ with $\frac{1}{2}\kappa v$ are $O(v^{5/2} + v^2 r_h^{1/2}\mathbb{E}[|Z|])$, whose second part is $O(vr_h + v^3)$ by $2ab \leq a^2 + b^2$, and those with η/h' are $O(vr_h + r_h^2)$ because the x-only part of ρ multiplies $\mathbb{E}[\eta] = 0$; $\mathbb{E}[\rho^2] = O(v^3 + v^2 r_h + r_h^3)$ uses $\mathbb{E}[Z^6] < \infty$. The third and fourth terms are $O(vr_h)$ and $O(r_h^{3/2})$. Taking expectations over X gives equation 1; the $r_h^{3/2}$ term vanishes when Z is symmetric. Without a uniform moment condition of this kind the remainder need not be small: Z taking values $\pm\delta/\sqrt{r_h}$ with probability $r_h/(2\delta^2)$ each has $\mathbb{E}[\eta^2] = r_h$ but $\mathbb{E}[|\eta|^3] = \delta r_h$, which is why Definition 1 fixes the law of Z. Under A2, replacing h by $ch + d$ leaves κ_h unchanged and multiplies $\mathrm{Var}[h(Y)]$ by c^2 and h'^{-2} by c^{-2}. □

Corollary 1. (i) If $v \equiv 0$ then $Y = \mu(X)$ and $p_\mu = q$. With $w = h'$ and $\int w = 2$, Hölder's inequality with exponents 3 and 3/2 gives $\int q^{1/3} = \int (q/w^2)^{1/3} w^{2/3} \leq J^{1/3}\,2^{2/3}$, with equality iff $q/w^2 \propto w$, i.e. $w \propto q^{1/3}$ (Panter & Dite, 1951). For $\mathbb{E}[|\hat{y} - y|^p]$ with fixed $\mathbb{E}[|\eta|^p]$ the same argument with exponents $1 + p$ and $(1 + p)/p$ gives $w \propto q^{1/(1+p)}$. Under A2 the leading term is $\tau^2\mathrm{Var}[h(Y)]\int q/h'^2 = \tau^2 K[h]$. (ii) As $r_h \to 0$ the leading term is $\frac{1}{4}\mathbb{E}[\kappa_h(\mu)^2 v^2]$, which is zero iff $\kappa_h(\mu(X))v(X) = 0$ a.s. This constrains h'' only on the set of conditional means that carry conditional noise, and only up to null sets: a transform that bends outside it also has zero leading term (e.g. a mixture of a noisy component whose targets stay in a subinterval and a deterministic component that covers the rest of the support); on two disjoint intervals h may be affine with different slopes; and at an isolated point nothing is implied about h''. If $p_\mu > 0$ on an interval I and $v > 0$ a.s. on $\{\mu(X) \in I\}$, then $\kappa_h = 0$ Lebesgue-a.e. on I, so $h'' = 0$ on I by continuity and h is affine on I. (iii) Write the leading term as $\int p_\mu[\frac{1}{4}\phi'^2\tilde{v}^2 + re^{-2\phi}]$ with $\tilde{v}^2(t) = \mathbb{E}[v(X)^2 \mid \mu(X) = t]$ (written v in the text) and add $\lambda(\int e^\phi - 2)$. The Euler–Lagrange equation of $F = \frac{1}{4}p_\mu v^2\phi'^2 + rp_\mu e^{-2\phi} + \lambda e^\phi$ is $-2rp_\mu e^{-2\phi} + \lambda e^\phi - (\frac{1}{2}p_\mu v^2\phi')' = 0$, with natural boundary condition $\frac{1}{2}p_\mu v^2\phi' = 0$. With $v = 0$, $\lambda e^{3\phi} = 2rp_\mu$, the Panter–Dite law. □

Proposition 1. Repeat the proof of Theorem 1 with $\Delta = \frac{1}{2}h''v + b_h + \sqrt{r_h(x)}Z + \rho_1$. Then $\hat{y} - \mu = \frac{1}{2}\kappa v + b_h/h' + \sqrt{r_h(x)}Z/h' + \rho'$ with $|\rho'| \leq C(v^{3/2} + (v + |b_h|)(|b_h| + \sqrt{r_h}|Z|) + |b_h|^2 + r_h|Z|^2)$. Conditionally on x, Z is independent of ε, so the cross term with ε still vanishes; squaring, the terms linear in Z vanish, and collecting the remainders gives the stated bound. □

Corollary 2. With $v \equiv 0$ and $b_h \equiv 0$, Proposition 1 gives the leading term $r\mathbb{E}[\omega(Y)/h'(Y)^2] = r\int q\,\omega/h'^2$. Hölder with exponents 3 and 3/2 as in Corollary 1(i), applied to $q\omega$ instead of q, gives $\int(q\omega)^{1/3} \leq (\int q\omega/w^2)^{1/3}2^{2/3}$ with equality iff $w \propto (q\omega)^{1/3}$. Under the normalized version the objective is $\tau^2\mathrm{Var}[h(Y)]\int q\omega/h'^2$. □

Equal-mass excess. For $h = 2F - 1$, $h' = 2q$ and $J = \frac{1}{4}\int q^{-1}$; for CRC, $J = \frac{1}{4}(\int q^{1/3})^3$. The ratio is ≥ 1 by (i), with equality iff $2q \propto q^{1/3}$, i.e. q constant. For $Y \sim \mathrm{Uniform}[1, 10^6]$ the equal-mass transform is optimal to first order despite six decades of range: sparse regions, not range, make it poor. □

Lemma 1. *Let $q > 0$ be continuously differentiable on $[a, b]$. Then the identity is not a stationary point of K among strictly increasing C^1 transforms, and hence (by affine invariance) no affine h minimizes K.*

Proof. Let $h_t(y) = y + tg(y)$ with $g \in C^1$, $g(a) = g(b) = 0$; h_t is strictly increasing for small $|t|$. At $t = 0$, $J = 1$ and $V = \mathrm{Var}Y =: s^2$, with mean $\bar{y}$. Since $\int q(y - \bar{y}) = 0$, $\frac{d}{dt}V = 2\int q(y - \bar{y})g$, and $\frac{d}{dt}J = -2\int qg' = 2\int q'g$ by parts. Hence $\frac{d}{dt}K = 2\int g\,[q(y-\bar{y})+s^2q']$, which vanishes for all g iff $q'/q = -(y - \bar{y})/s^2$ on (a, b), i.e. q is a Gaussian density with location $\bar{y}$ and scale s truncated to $[a, b]$. A truncated Gaussian with location m and scale s has mean m only if the truncation is symmetric, $a - m = -(b - m) = -\beta s$, and then its variance is $s^2(1 - 2\beta\varphi(\beta)/(2\Phi(\beta) - 1)) < s^2$. So the location cannot equal the mean while the scale equals the standard deviation. □

Theorem 3. Let Φ be any rule and $h_0 = \Phi(q)$; it is the same transform on every joint law with marginal q.

P_1: $Y = X \sim q$. Here $v \equiv 0$ and by Theorem 1 the excess risk is $r_h J[h] + O(r_h^{3/2})$. (i) Under A1, asymptotic optimality requires $J[h_0] \leq \inf J = J[h_{\mathrm{PD}}]$, i.e. $h_0 = h_{\mathrm{PD}}$ by uniqueness, and h_{PD} is affine iff q is uniform. (ii) Under A2, it requires $K[h_0] = \inf K$ (if the infimum is not attained, no h_0 is asymptotically optimal on P_1 at all), and by Lemma 1 h_0 is not affine.

P_2: a random endpoint of an unordered pair. Draw Y_1, Y_2 i.i.d. from q, let $X = (\min(Y_1, Y_2), \max(Y_1, Y_2))$ and $Y = Y_1$. Then $Y \sim q$, and by exchangeability of (Y_1, Y_2), given $X = (u, w)$, Y is u or w with probability $\frac{1}{2}$ each, so $\mu = \frac{1}{2}(u + w)$ and $m_h = \frac{1}{2}(h(u) + h(w))$. As $r_h \to 0$, $\hat{y} \to M_h(u, w) = h^{-1}(\frac{1}{2}h(u) + \frac{1}{2}h(w))$ by continuity of h^{-1} and bounded convergence, so the excess risk converges to $B(h) = \mathbb{E}[(M_h(U, W) - \frac{1}{2}(U + W))^2]$. For affine h, M_h is the midpoint and the excess risk is $r_h\mathbb{E}[h'^{-2}] \to 0$. If $B(h) = 0$ for a continuous strictly increasing h, then $h(\frac{u+w}{2}) = \frac{1}{2}(h(u) + h(w))$ for almost every, hence by continuity every, pair in $[a, b]^2$ (the pair has a positive density $q(u)q(w)$ there); by Jensen's functional equation for continuous functions, h is affine on $[a, b]$. So on P_2 only affine h_0 are asymptotically optimal. This step uses neither the expansion nor small noise.

Combining, the requirements of P_1 and P_2 are incompatible. □

Remark (a second exact example). If q is strictly decreasing and $Y \perp X$, then $m_h = \mathbb{E}[h(Y)]$ and the limiting excess is $(\mathbb{E}[Y] - h^{-1}(\mathbb{E}[h(Y)]))^2$; h_{PD} is strictly concave, so Jensen gives $h_{\mathrm{PD}}^{-1}(\mathbb{E}[h_{\mathrm{PD}}(Y)]) < \mathbb{E}[Y]$ strictly.

Theorem 2. With μ log-uniform on $[1, U]$, $\mathbb{E}[\mu^m] = (U^m - 1)/(m\ell)$ for $m \neq 0$, so $\mathbb{E}[\mu^m] = U^m/(m\ell)\,(1 + o(1))$ for fixed $m > 0$ as $\ell \to \infty$. Given μ, $v = \sigma^2\mu^2$ exactly and $\mathbb{E}[|Y - \mu|^3 \mid \mu] = \sigma^3\mu^3\mathbb{E}[|\varepsilon|^3] \leq Cv^{3/2}$, so Theorem 1 applies on the compact support. For $h' = k_0 y^{-c}$, $\kappa_h = -c/y$. *Bias:* $\frac{1}{4}\mathbb{E}[\kappa^2 v^2] = \frac{1}{4}c^2\sigma^4\mathbb{E}[\mu^2] = c^2\sigma^4U^2/(8\ell)\,(1 + o(1))$. *Resolution under A1:* with the normalization on the range of μ, $h(1) = -1$, $h(U) = 1$ gives $k_0 = 2(1 - c)/(U^{1-c} - 1)$ and $r\mathbb{E}[h'^{-2}] = r\mathbb{E}[\mu^{2c}]/k_0^2 = rU^2/(8\ell\, c(1 - c)^2)\,(1 + o(1))$. Normalizing instead on the support $[1 - \sigma B, U(1 + \sigma B)]$ of Y ($|\varepsilon| \leq B$) multiplies k_0 by $1 + O(\sigma)$ and the resolution term by $1 + O(\sigma)$, which changes equation 2 under A1 only at relative order σ. *Resolution under A2:* by affine invariance take $k_0 = 1$, $h = y^{1-c}/(1 - c)$; $\mathrm{Var}h(Y) = \mathrm{Var}h(\mu)(1 + O(\sigma^2))$ and $\mathrm{Var}h(\mu) = \mathbb{E}[\mu^{2(1-c)}]/(1 - c)^2 - (\mathbb{E}[\mu^{1-c}])^2/(1 - c)^2 = U^{2(1-c)}/(2\ell(1 - c)^3)\,(1 + O(\ell^{-1}))$, so $\tau^2\mathrm{Var}[h(Y)]\,\mathbb{E}[\mu^{2c}] = \tau^2U^2/(4\ell^2c(1 - c)^3)\,(1 + o(1))$. Dividing by $U^2/(8\ell)$, the leading-order objectives are $r/(c(1 - c)^2) + c^2\sigma^4$ and $2\tau^2/(\ell c(1 - c)^3) + c^2\sigma^4$. Setting the derivatives to zero with $\frac{d}{dc}[c(1 - c)^2] = (1 - c)(1 - 3c)$ and $\frac{d}{dc}[c(1 - c)^3] = (1 - c)^2(1 - 4c)$ gives equation 2. The right-hand sides are strictly decreasing from $+\infty$ to 0 on $(0, 1/3]$ and $(0, 1/4]$ and negative beyond, so the stationary point is unique and is the minimizer, with $c^\star \leq 1/3 < 1$. The order of limits matters: the expansion is taken first (small σ and small learning error at fixed U), then the leading

order in ℓ; the $o(1)$ terms are of relative order ℓ^{-1}, which explains the gap of about 0.02 between the limit formula and the finite-U optimum in Appendix B. □

B NUMERICAL CHECKS OF THE THEORY

Expansion (Theorem 1). Log-uniform marginal on $[1, 10^4]$ built as $Y = U^{\Phi_N(L)}$, $L = \rho Z + \sqrt{1-\rho^2}E$, $X = Z$; power transforms $h' \propto y^{-c}$, $c \in \{0, \frac{1}{24}, \dots, 1\}$; idealized learner under A1 with Gaussian Z; 2×10^5 draws, 60-point Gauss–Hermite quadrature for conditional moments, common random numbers across c.

Table 5: Minimizer of the Monte Carlo excess risk and of the expansion, their ratio at CRC ($c = 1/3$), and CRC and raw targets relative to the best c (Monte Carlo).

r	ρ	$c^\star$ (MC)	$c^\star$ (expansion)	MC/exp. at CRC	CRC/best	Raw/best
1e-6	0.9	0.000	0.000	0.758	750.11	1.00
1e-6	0.99	0.042	0.042	0.969	12.79	1.02
1e-6	0.999	0.250	0.250	1.006	1.08	2.04
1e-6	0.9999	0.333	0.333	1.010	1.00	2.50
1e-6	1	0.333	0.333	1.011	1.00	2.50
1e-4	0.9	0.042	0.042	0.763	9.01	1.03
1e-4	0.99	0.250	0.250	0.989	1.07	1.79
1e-4	0.999	0.333	0.333	1.009	1.00	2.16
1e-4	0.9999	0.333	0.333	1.011	1.00	2.17
1e-4	1	0.333	0.333	1.011	1.00	2.17

Impossibility on the full family. Same marginal, all 60 family members and raw targets, $r \in \{10^{-4}, 10^{-6}, 10^{-8}\}$. At $\rho = 1$ CRC is exactly optimal for $r \geq 10^{-6}$; at $r = 10^{-8}$ the cells of the 2×10^5-point sample used to fit the compressors are wider than the noise and the minimizer moves by one grid step (0.5% from CRC), a discretization effect. At $\rho = 0$ raw targets (or the equivalent member $(0, 1)$) are optimal and CRC is 4.3% worse.

Table 6: Best family member (α, β) in log coordinates (CRC is $(1/3, 2/3)$) and CRC and raw targets relative to it.

r	ρ	best law (α, β)	CRC/best	Raw/best
1e-8	0	0.000, 1.000	1.043	1.000
1e-8	0.5	Raw	1.038	1.000
1e-8	0.8	Raw	1.023	1.000
1e-8	0.95	0.000, 1.000	1.008	1.000
1e-8	0.99	0.000, 1.000	1.002	1.000
1e-8	1	0.000, 0.583	1.005	2.594
1e-6	0	0.000, 1.000	1.043	1.000
1e-6	0.5	Raw	1.038	1.000
1e-6	0.8	Raw	1.023	1.000
1e-6	0.95	Raw	1.008	1.000
1e-6	0.99	Raw	1.002	1.000
1e-6	1	0.333, 0.667	1.000	2.460
1e-4	0	Raw	1.043	1.000
1e-4	0.5	Raw	1.038	1.000
1e-4	0.8	Raw	1.023	1.000
1e-4	0.95	0.667, 1.000	1.007	1.001
1e-4	0.99	0.667, 0.833	1.001	1.006
1e-4	1	0.333, 0.667	1.000	2.138

Closed form (Theorem 2). $U = 10^6$, $\tau = 10^{-3}$, bounded noise $\varepsilon \sim \mathrm{Uniform}[-\sqrt{3}, \sqrt{3}]$ (20-point Gauss–Legendre quadrature), c on a grid of step 0.01; Monte Carlo of the idealized learner under A2 with 6×10^4 draws.

Table 7: Optimal c: wide-range limit equation 2, exact log-uniform moments in the leading-order objective, and Monte Carlo.

σ	ψ_3	$c^\star$ (limit)	$c^\star$ (exact moments)	$c^\star$ (MC)
0.00	0.0	0.250	0.270	0.270
0.01	0.1	0.250	0.270	0.270
0.02	2.2	0.247	0.260	0.260
0.03	11.2	0.237	0.250	0.250
0.05	86.3	0.188	0.180	0.180
0.08	565.9	0.116	0.100	0.090
0.12	2864.8	0.070	0.040	0.040

C CONTROLLED REGRESSION

$X \sim N(0, I_{10})$; $f_0(X) = \sin x_1 + x_2 x_3 + \frac{1}{2}x_4^2 + x_5 + \tanh(x_6 x_7) + \frac{1}{2}x_8$, mapped to $N(0,1)$ through its CDF estimated on 10^6 draws; $y = \exp\{s(\rho f + \sqrt{1-\rho^2}e)\}$; $s \in \{0.5, 1.5, 2.5\}$ (about 1.7, 5 and 8 decades over 10^4 draws), $\rho \in \{0.5, 0.8, 0.95, 0.99\}$; $n = 10{,}000$, $60/20$ training/test rows of a $60/20/20$ split, 10 seeds, ϵ at the median. $m_h(x) = \mathbb{E}[h(Y) \mid x]$ and $\mu(x)$ by 40-point Gauss–Hermite quadrature. Per transform and seed we record the direct learning error $\hat{r}_h$, $\mathrm{Var}[h(Y_{\mathrm{train}})]$, the mean and MSE of $\hat{s} - m_h$ in each decile of m_h on the test points, the true excess risk $\mathrm{mean}(\hat{y} - \mu)^2$, and the excess risk of the idealized learner (Gaussian Z, 20 replicates) with (a) the measured scalar $\hat{r}_h$, (b) decile-wise MSEs, (c) the median $\hat{r}_h$ of the cell for every transform (A1), (d) the median $\hat{r}_h/\mathrm{Var}h$ of the cell times $\mathrm{Var}h$ (A2). An earlier version estimated the learning error as the measured transformed-space test MSE minus $\mathbb{E}[\mathrm{Var}(h(Y) \mid X)]$, which is unstable when the two are close; the direct estimator replaces it.

Table 8: All cells (means over 10 seeds). MLP-u: unstandardized target. $\hat{k}$: learned optimum by test MSE; $k^\star_{A1}$, $k^\star_{A2}$: optimum of the calibrated idealized learner; spreads of $\hat{r}_h$ (abs.) and $\hat{r}_h/\mathrm{Var}h$ (norm.); ρ_S as in Table 1; *decile*: max/min decile error; *mean/RMSE*: largest decile mean error over the overall RMSE; last two: idealized over true excess risk.

learner	s	ρ	$\hat{k}$	$k^\star_{A1}$	$k^\star_{A2}$	abs.	norm.	ρ_S A1	ρ_S A2	decile	mean/RMSE	scalar	decile r
MLP	0.5	0.5	3	2	4	1.7	1.10	0.3	1.0	2.0	0.44	0.93	0.96
MLP	0.5	0.8	4	1	3	1.6	1.11	-0.6	1.0	2.8	0.43	0.78	0.88
MLP	0.5	0.95	4	2	3	1.7	1.06	-0.3	0.9	4.0	0.41	0.64	0.88
MLP	0.5	0.99	4	2	3	1.7	1.03	-0.3	0.9	4.2	0.33	0.45	0.67
MLP	1.5	0.5	3	2	3	49.3	1.27	-0.1	1.0	4.9	0.84	0.79	0.81
MLP	1.5	0.8	3	2	3	15.9	2.89	0.8	0.5	12.8	0.54	0.40	0.51
MLP	1.5	0.95	2	2	2	10.7	3.45	0.5	1.0	23.1	0.41	0.18	0.38
MLP	1.5	0.99	2	2	3	7.5	4.57	0.5	0.9	26.3	0.37	0.13	0.33
MLP	2.5	0.5	2	2	2	499.9	2.54	0.4	0.9	10.8	0.85	0.77	0.77
MLP	2.5	0.8	2	2	3	76.3	3.18	0.6	0.6	43.6	0.62	0.44	0.46
MLP	2.5	0.95	1	1	1	16.7	9.58	0.4	1.0	87.0	0.44	0.27	0.33
MLP	2.5	0.99	1	1	2	8.5	14.12	0.7	0.9	123.4	0.41	0.21	0.32
MLP-u	0.5	0.5	2	1	4	1.4	1.34	0.5	-0.1	2.1	0.37	0.97	1.00
MLP-u	0.5	0.8	2	1	3	1.3	1.35	-0.1	0.8	2.3	0.44	0.84	0.94
MLP-u	0.5	0.95	3	2	3	1.4	1.21	-0.3	0.9	3.0	0.45	0.69	0.94
MLP-u	0.5	0.99	4	2	3	1.5	1.18	-0.3	0.9	3.4	0.38	0.52	0.75
MLP-u	1.5	0.5	2	2	3	8.3	10.70	1.0	0.4	2.5	0.40	0.84	0.87
MLP-u	1.5	0.8	2	2	3	7.3	7.68	0.6	0.3	5.2	0.49	0.41	0.56
MLP-u	1.5	0.95	2	2	3	6.1	7.36	0.8	0.8	7.7	0.42	0.20	0.45
MLP-u	1.5	0.99	2	2	3	5.7	7.27	0.5	0.9	9.2	0.42	0.20	0.42
MLP-u	2.5	0.5	1	2	2	19.8	18.18	0.7	0.2	3.0	0.27	0.93	0.92
MLP-u	2.5	0.8	1	2	2	20.9	14.26	0.6	0.7	6.9	0.41	0.42	0.44
MLP-u	2.5	0.95	1	1	2	11.1	17.40	0.7	0.9	13.3	0.37	0.30	0.39
MLP-u	2.5	0.99	1	1	2	7.3	21.65	0.7	0.7	21.5	0.36	0.33	0.45

Table 9: Table 8 continued: HGB.

learner	s	ρ	$\hat{k}$	$k^\star_{A1}$	$k^\star_{A2}$	abs.	norm.	ρ_S A1	ρ_S A2	decile	mean/RMSE	scalar	decile r
HGB	0.5	0.5	4	1	3	1.8	1.05	-0.7	0.9	4.2	0.97	0.75	0.82
HGB	0.5	0.8	3	1	3	1.6	1.15	-0.8	0.9	5.2	0.92	0.70	0.96
HGB	0.5	0.95	3	2	3	1.4	1.25	0.3	0.9	6.0	1.03	0.75	1.30
HGB	0.5	0.99	2	2	3	1.4	1.27	0.9	0.7	6.7	1.09	0.73	1.42
HGB	1.5	0.5	3	2	3	62.8	1.26	-0.1	0.9	9.1	0.96	0.66	0.74
HGB	1.5	0.8	3	2	3	20.9	2.19	-0.1	1.0	23.5	1.05	0.38	0.63
HGB	1.5	0.95	2	2	3	8.5	4.23	0.9	0.5	39.7	1.21	0.23	0.53
HGB	1.5	0.99	1	2	3	6.1	5.39	0.8	0.3	51.6	1.24	0.27	0.52
HGB	2.5	0.5	2	2	2	1009.1	4.59	1.0	1.0	23.7	0.97	0.74	0.73
HGB	2.5	0.8	2	2	3	130.6	2.00	0.7	0.8	77.1	1.02	0.51	0.50
HGB	2.5	0.95	1	1	2	28.0	5.63	0.7	0.7	218.3	1.10	0.22	0.48
HGB	2.5	0.99	1	1	2	15.5	6.59	0.9	0.7	265.3	1.11	0.19	0.43

D THE CLOSED FORM WITH TRAINED LEARNERS

$X \sim N(0, I_{10})$, $\mu = U^{\Phi_N(f(X))}$ with $U = 10^4$ and f as in Appendix C (so μ is exactly log-uniform), and $Y = \mu(1 + \sigma\varepsilon)$ with $\varepsilon \sim \mathrm{Uniform}[-\sqrt{3}, \sqrt{3}]$, $\sigma \in \{0, 0.02, 0.1, 0.3\}$; $n = 10{,}000$, 6,000 training and 2,000 test rows, 10 seeds. h_c is Box–Cox with $h' \propto y^{-c}$ mapped to $[-1, 1]$ on $[1, U]$, and the transformed target is standardized. We record the excess risk $\mathrm{mean}(\hat{y} - \mu)^2$, the direct learning error, and the predicted optimum of the idealized learner with the median r (A1) or the median τ^2 (A2) over c, using the exact bias of each transform.

Table 10: Learned and predicted optimal c, excess risk at selected c relative to the best, and max/min over c of the absolute and normalized learning error.

learner	σ	learned $c^\star$	A1 pred.	A2 pred.	exc. c=0	c=1/4	c=1/3	c=1	r spread	τ^2 spread
HGB	0	0.333	0.333	0.250	1.089	1.036	1.000	5.6	1.39	2.75
HGB	0.02	0.500	0.333	0.250	1.068	1.016	1.016	6.1	1.39	2.71
HGB	0.1	0.333	0.333	0.250	1.106	1.008	1.000	5.1	1.43	2.80
HGB	0.3	0.333	0.333	0.250	1.130	1.045	1.000	4.9	1.58	2.87
MLP	0	0.000	0.333	0.250	1.000	1.275	1.378	53.2	1.58	2.92
MLP	0.02	0.000	0.333	0.250	1.000	1.218	1.361	53.0	1.56	3.04
MLP	0.1	0.000	0.333	0.250	1.000	1.183	1.267	35.7	1.77	3.46
MLP	0.3	0.000	0.333	0.250	1.000	1.075	1.167	53.3	1.79	3.24

For HGB the learned optimum is the Panter–Dite value $c = 1/3$ at every σ but one, and its absolute error varies less across c than its normalized error, i.e. it behaves like the A1 learner here although it behaved more like A2 in Appendix C. The MLP prefers raw targets at every σ: with raw targets it spends its accuracy where the target, and hence the squared error, is large, a reallocation along the target range that Definition 1 excludes. Neither learner reaches the A2 optimum $c = 1/4$. The optimum is not sharp (Table 11): HGB's per-seed optimum lies mostly between $c = 1/4$ and $c = 1/2$, and $c = 1/4$ and $c = 1/3$ are statistically indistinguishable at two of the four noise levels.

Table 11: Uncertainty of the learned optimum: per-seed optimal c (value:count over 10 seeds) and the excess risk at $c = 1/4$ over that at $c = 1/3$ (geometric mean over seeds, 95% seed-bootstrap interval).

learner	σ	per-seed optimum c (count of 10)	exc. c=1/4 / c=1/3 [95% CI]
HGB	0	1/3:4, 5/12:1, 1/2:4, 2/3:1	1.035 [1.012, 1.062]
HGB	0.02	0:1, 1/6:1, 1/4:2, 1/3:3, 5/12:1, 1/2:2	1.001 [0.978, 1.022]
HGB	0.1	1/6:1, 1/4:4, 1/3:1, 5/12:2, 1/2:2	1.007 [0.986, 1.029]
HGB	0.3	1/3:5, 5/12:3, 1/2:1, 2/3:1	1.048 [1.015, 1.083]
MLP	0	0:5, 1/12:4, 1/6:1	0.944 [0.871, 1.036]
MLP	0.02	0:4, 1/12:3, 1/6:3	0.898 [0.851, 0.945]
MLP	0.1	0:5, 1/12:3, 1/4:1, 1/2:1	0.942 [0.891, 0.987]
MLP	0.3	0:4, 1/12:3, 1/6:3	0.924 [0.905, 0.948]

E CORRECTING THE RETRANSFORMATION BIAS

In the controlled regression of Appendix C (MLP and HGB, standardized targets, 10 seeds), we compare three predictions for every transform: plain retransformation $h^{-1}(\hat{s})$; global smearing with 500 validation residuals; and the oracle correction $h^{-1}(\hat{s}) - [h^{-1}(m_h(x)) - \mu(x)]$, which removes the retransformation bias a perfect transformed-space regressor would have and is available only because the conditional law is known.

Table 12: Learned optimum k (of the sweep $h' \propto q^{1/3}(y+\epsilon)^{(k-2)/3}$; smaller k is more compressed) without correction, with global smearing, and with the oracle bias correction; share of CRC's excess risk that is the bias of a perfect regressor; best uncorrected over best corrected excess.

learner	s	ρ	$\hat{k}$ plain	$\hat{k}$ smeared	$\hat{k}$ bias-corr.	bias share (CRC)	best plain / best corr.
HGB	0.5	0.5	4	2	0	0.18	1.62
HGB	0.5	0.8	3	3	1	0.07	1.30
HGB	0.5	0.95	3	3	3	0.01	1.09
HGB	0.5	0.99	2	2	2	0.00	1.03
HGB	1.5	0.5	3	1	0	0.55	10.64
HGB	1.5	0.8	3	1	1	0.21	3.00
HGB	1.5	0.95	2	1	2	0.03	1.41
HGB	1.5	0.99	1	1	1	0.01	1.18
HGB	2.5	0.5	2	1	0	0.74	326.26
HGB	2.5	0.8	2	1	1	0.38	11.23
HGB	2.5	0.95	1	1	2	0.18	2.66
HGB	2.5	0.99	1	1	4	0.06	1.71
MLP	0.5	0.5	4	2	0	0.36	1.57
MLP	0.5	0.8	3	3	3	0.13	1.17
MLP	0.5	0.95	4	4	4	0.02	1.02
MLP	0.5	0.99	4	4	4	0.00	1.00
MLP	1.5	0.5	3	1	0	0.71	13.89
MLP	1.5	0.8	2	2	2	0.32	2.83
MLP	1.5	0.95	2	2	2	0.08	1.51
MLP	1.5	0.99	2	2	2	0.05	1.21
MLP	2.5	0.5	3	1	0	0.77	468.63
MLP	2.5	0.8	1	1	1	0.44	17.33
MLP	2.5	0.95	1	1	1	0.26	2.31
MLP	2.5	0.99	1	1	1	0.10	1.45

F LEARNABILITY BEYOND VARIANCE: THE TAIL OF THE REGRESSION FUNCTION

Controlled experiment. To isolate the tail of the regression function from its variance, we draw $u = f(X)$ as in Appendix C, standardize $g_a(u) = \sinh(au)/a$ ($a \to 0$: $g_0(u) = u$) to m, and set $t = \sqrt{1-\nu}\, m + \sqrt{\nu}\, e$ with noise share $\nu \in \{5\%, 30\%\}$. The target variance and noise share are fixed; only the kurtosis of the regression function $\sqrt{1-\nu}\, m$ changes. We train the MLP on t ($n = 10{,}000$, 5 seeds) and report the normalized learning error τ^2, the test MSE against $\sqrt{1-\nu}\, m$ divided by $\mathrm{Var}t$.

Table 13: Normalized learning error of the MLP against the kurtosis of the regression function.

a	kurtosis of m	τ^2 (noise 5%)	τ^2 (noise 30%)
0.0	3.0	0.023	0.037
0.5	4.5	0.028	0.047
1.0	28.2	0.066	0.106
1.5	341.9	0.187	0.252
2.0	1437.4	0.491	0.439
2.5	2867.7	0.741	0.579

τ^2 grows 32-fold over the range, roughly as $\mathrm{kurtosis}^{0.35}$ (log-log slope over all seeds). Changing a changes not only the tail but also the local slopes of the regression function, so this is not an isolated manipulation of kurtosis; we read it as evidence that heavier-tailed regression functions are harder to learn at fixed variance. Raw-like transforms of wide-range targets have heavy-tailed

regression functions, which is consistent with the preference of trained learners for transforms more compressed than CRC on wide, predictable targets (Section 5). As a target-only correction we multiplied K by $(\mathrm{kurtosis}[h(Y)]/3)^{0.44}$ (the exponent fitted to the per-a means of this experiment). On an earlier run of the controlled regression it moved the predicted optimum in the right direction (exact hits $5/11$ vs. $3/11$ for K) without reaching the learned optimum on wide targets, and on the four wide-range datasets of group (3) of Appendix G it ranked family members *worse* than K (mean Spearman 0.67 vs. 0.75; J: 0.57). The kurtosis of $h(Y)$ is a poor proxy for that of $\mathbb{E}[h(Y) \mid X]$, and we do not use the correction elsewhere.

G DATASETS

Selection. (1) Every task of OpenML-CTR23 (Fischer et al., 2023) whose target is nonnegative (30 of 35), identified by OpenML data id: BrazilianHouses 44990, FIFA 45012, Socmob 44987, CPS88Wages 44984, video transcoding 44974, auction verification 44958, solar flare 44966, Kings County 44989, diamonds 44979, Miami housing 44983, physiochemical protein 44963, cars 44994, FPS benchmark 44992, California housing 44977, health insurance 44993, concrete 44959, abalone 44956, QSAR fish toxicity 44970, energy efficiency 44960, kin8nm 44980, student performance 44967, CPU activity 44978, white wine 44971, red wine 44972, moneyball 41021, airfoil self-noise 44957, wave energy 44975, naval propulsion 44969; forest fires and superconductivity are also in CTR23 and are used in the versions of (3). (2) The ten datasets of Nuyts & Davis (2025): auto MPG (OpenML 196), bike sharing (42712), liver disorders (8), and from the UCI repository combined cycle power plant, energy efficiency (cooling load; the OpenML copy is binned), real estate valuation and servo (the OpenML copy is binarized); concrete and energy efficiency (heating) coincide with CTR23 and online news with (3). (3) Six wide-range datasets: forest fires (Cortez & Morais, 2007), online news popularity (Fernandes et al., 2015), superconductivity (Hamidieh, 2018), French motor insurance claim severity (freMTPL2 from CASdatasets, OpenML data ids 41214/41215, version 1; claims summed per policy) (Dutang et al., 2020), Ames house prices without the identifier (De Cock, 2011), and stellar radii (AI4MR) without every column computed from the radius (Moya & López-Sastre, 2022).

Preprocessing. For (1) and (2): categorical columns one-hot encoded, numeric missing values imputed with the column median, constant columns dropped, rows with a missing target dropped. For (3): numeric features as given, categoricals one-hot encoded, rows with missing values dropped. Features are standardized inside the model pipeline. *Selection.* Group (3) was chosen by us; without it (35 datasets from CTR23 and Nuyts & Davis (2025) only), CRC-log over CRC still rises with L (Spearman 0.77), the selected standard transform over CRC also rises with L (0.44), the uniform quantile transform is worse than CRC on 35 of 35, and K ranks transforms better than J on 28 of 35. *Size.* A first run capped the five datasets with more than $33{,}334$ rows at that size (one fixed subsample); every result reported uses all rows except the training-size experiment and the diagnosis of Appendix N, which use the capped copies (Appendix I compares the two).

Table 14: Datasets: source (CTR23; ND: Nuyts & Davis (2025); extra: group (3)), rows, features after encoding, leverage L (median over splits, training targets), and raw-target test $R^2_{\mathrm{raw}} = 1 - \overline{\mathrm{MSE}}_{\mathrm{raw}}/\mathrm{Var}(Y)$ of the MLP (negative where it is worse than the mean).

Dataset	Source	N	p	L	R^2_{raw}
InsSeverity	extra	24,944	48	53.70	-0.683
BrazilHouses	CTR23	10,692	49	35.94	0.944
OnlineNews	extra+ND	39,644	59	26.55	-0.762
ForestFires	extra	517	12	9.75	0.109
AI4MR	extra	649	14	7.42	0.565
FIFA	CTR23	19,178	190	5.78	0.578
Socmob	CTR23	1,156	39	4.26	0.719
CPS88Wages	CTR23	28,155	12	3.93	0.295
VideoTransc.	CTR23	68,784	24	3.59	0.985
AuctionVerif.	CTR23	2,043	16	2.65	0.926
SolarFlare	CTR23	1,066	28	2.02	0.011
KingsCounty	CTR23	21,613	132	1.92	0.852
Servo	ND	167	12	1.72	0.424
Superconduct	extra	21,263	81	1.58	0.885
LiverDisorders	ND	345	5	1.54	0.068
Diamonds	CTR23	53,940	26	1.52	0.974
MiamiHousing	CTR23	13,932	15	1.51	0.908
BikeSharing	ND	17,379	20	1.48	0.931
PhysiochemicalProtein	CTR23	45,730	9	1.30	0.561
House	extra	1,121	36	1.21	0.725
Cars	CTR23	804	17	1.15	0.945
FpsBenchmark	CTR23	24,624	122	1.14	0.998
CaliforniaHousing	CTR23	20,640	8	1.13	0.788
HealthInsurance	CTR23	22,272	25	1.13	0.390
Concrete	CTR23+ND	1,030	8	1.12	0.883
QsarFishToxicity	CTR23	908	6	1.11	0.572
Abalone	CTR23	4,177	10	1.11	0.538
RealEstate	ND	414	6	1.11	0.675
StudentPerformancePor	CTR23	649	56	1.09	0.147
EnergyEfficiency	CTR23+ND	768	8	1.09	0.938
Kin8Nm	CTR23	8,192	8	1.09	0.917
EnergyEfficiencyCooling	ND	768	8	1.06	0.912
AutoMpg	ND	398	25	1.06	0.811
CpuActivity	CTR23	8,192	21	1.04	0.978
WhiteWine	CTR23	4,898	11	1.02	0.389
RedWine	CTR23	1,599	11	1.02	0.376
Moneyball	CTR23	1,232	74	1.01	0.881
AirfoilSelfNoise	CTR23	1,503	5	1.00	0.805
WaveEnergy	CTR23	72,000	48	1.00	1.000
Ccpp	ND	9,568	4	1.00	0.946
NavalPropulsionPlant	CTR23	11,934	14	1.00	0.999

H ALGORITHMS AND SETTINGS

Family and estimator. In log coordinates $z = \log(y + \epsilon)$, with p the density of z, the family is $g' \propto p^{\alpha}(y + \epsilon)^{\beta}$; in y it is $h' \propto q^{a}(y + \epsilon)^{-b}$ with $a = \alpha$, $b = 1 - \alpha - \beta$. CRC is $(\alpha, \beta) = (1/3, 2/3)$, CRC-log $(1/3, 0)$, equal-mass $(1, 0)$, log $(0, 0)$, raw $(0, 1)$, and Box–Cox with power λ on $y + \epsilon$ is $(0, \lambda)$. The grid of shapes is $\alpha \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$, $\beta \in \{0, \frac{1}{3}, \frac{2}{3}, 1, \frac{4}{3}\}$ (20 shapes), each with $\epsilon \in \{$first percentile, median, $10\times$ median of positive training targets$\}$ (60 candidates); in splits with training $L < 1.3$, $\alpha \in \{0, \frac{1}{3}\}$ (10 shapes, 30 candidates) plus equal-mass at the median ϵ. The estimator discretizes the empirical measure: unique training values $\tilde{z}_1 < \cdots < \tilde{z}_K$ with masses m_k own Voronoi cells of width Δ_k; cell k receives the increment $d_k = m_k^{\alpha} e^{\beta \tilde{z}_k} \Delta_k^{1-\alpha}$; the compressor is piecewise linear with knots at the cell edges and values $-1 + 2 \sum_{j \le k} d_j / \sum_j d_j$. For CRC, $(\alpha, \beta) = (1/3, 2/3)$, the increments $d_k \propto m_k^{1/3} e^{2\tilde{z}_k/3} \Delta_k^{2/3}$ exactly minimize the cell-wise Bennett sum $J_d = \sum_k m_k e^{2\tilde{z}_k} / w_k^2$ over cell-constant slopes w_k with a fixed budget (Hölder, as

in Corollary 1(i)); other members use the same construction with their own increments and do not minimize J_d. Past the support the default extends the map with a generalized-Pareto boundary slope fitted to the outer 5% (at least 20) of z, capped at 10% of the support width, with $\hat{s}$ clipped to 0.3 beyond the knots on inversion; the *clip* variant instead clips to the training support, like the baselines (Appendix J).

Target-only statistics. In each split, from the training targets only: $L = J_d[\text{CRC-log}]/J_d[\text{CRC}]$ with ϵ at the median (used for the candidate gate and, as the median over splits, for the groups); for every family member, $J = J_d$ evaluated on the training targets with the fitted slopes and $K = \mathrm{Var}[h(Y_{\text{train}})]\, J$. argmin J and argmin K choose among all family candidates of the split, including ϵ, with no validation information.

Baselines. Log1p: $\log(1+y)$, predictions clipped to the training range. Box–Cox: λ by maximum likelihood on $y + 10^{-6}$, predictions clipped to the training range. Gaussian and uniform quantile: `QuantileTransformer` with $\min(1000, n)$ quantiles and a fixed seed (Pedregosa et al., 2011), which clips to the training range. HGB with Poisson or gamma loss on raw targets. Smearing (Duan, 1983): $\hat{y}(x) = \frac{1}{R}\sum_i h^{-1}(\hat{t}(x) + e_i)$ over up to $R = 500$ validation residuals (global, i.e. the same residuals for every x), applied to both sides of a comparison.

Models. MLP: scikit-learn `MLPRegressor` (128,64), ReLU, Adam at 10^{-3}, L_2 10^{-4}, at most 1000 epochs, early stopping on an internal 10% split. HGB: 500 iterations, learning rate 0.05, early stopping. Ridge: `RidgeCV` over 10^{-3}–10^{3}. Features standardized; transformed targets standardized with training statistics unless stated.

Statistics. Per seed, RMSE ratio method/CRC; dataset values are means over 20 seeds; per dataset, the corrected resampled t-test of Nadeau & Bengio (1999) on the 20 per-split log RMSE ratios (variance inflated by $1/J + n_{\text{test}}/n_{\text{train}}$), Holm-corrected over the 41 datasets within each comparison; across datasets, one-sample Wilcoxon tests of the per-dataset mean log ratios. Regret: test RMSE of the selected candidate over the minimum test RMSE among the candidates of that split. Intervals: 2,000 bootstrap resamples of datasets. Repeated seeds describe variability on these datasets, not on new ones.

I FULL RESULTS

Table 15: All datasets (MLP): mean per-seed RMSE ratio *method/CRC* over 20 seeds; * Holm-adjusted $p < 0.05$. Rows ordered as in Table 14.

Dataset	U.q.	G.q.	YJ	Log1p	BC-CV	Sel. std.	Std+G.q.	+sm.	CRC-log
InsSeverity	2.918	1.262	1.088	1.095	0.998	1.078	1.078	1.088	1.570
BrazilHouses	19.759	2.311	1.511	1.455	0.914	0.855	1.005	0.602	2.989
OnlineNews	1.392	1.040	1.047	1.035	1.024	1.032	1.033	1.027	1.094
ForestFires	1.045	1.047	1.032	1.011	1.073	1.085	1.085	1.015	1.001
AI4MR	2.962	1.747	6.040	1.037	1.010	1.041	1.036	1.062	1.856
FIFA	2.873*	1.364	1.218	1.026	0.994	1.013	1.013	1.014	1.186
Socmob	2.964*	2.085	1.660	1.003	1.008	1.020	1.020	1.003	1.093
CPS88Wages	1.024*	1.007*	1.007*	1.015*	0.995	0.996	0.996	1.001	1.002
VideoTransc.	4.722*	1.161	1.345	1.038	0.945	1.047	1.043	1.047	1.181
AuctionVerif.	1.221	1.151	1.294*	1.253	1.007	1.027	1.018	1.018	1.024
SolarFlare	1.119*	1.115*	1.117	0.996	0.996	0.996	0.996	0.996	1.003
KingsCounty	5.055*	1.078	1.011	0.930	0.897	0.925	0.925	0.931	1.087
Servo	1.749	1.593	1.499	1.063	1.070	1.090	1.109	1.028	1.103
Superconduct	1.107	1.095*	1.083*	1.126*	0.989	1.006	1.006	0.998	1.022
LiverDisorders	1.049	1.056	1.053	1.043	0.990	1.002	1.002	0.994	1.007
Diamonds	1.080	1.032	1.042	1.018	0.984	1.038	1.037	1.010	1.004
MiamiHousing	1.866*	1.069	1.076	1.011	0.988	1.001	1.001	1.002	1.007
BikeSharing	1.274*	1.196*	1.062	1.233*	0.988	1.024	1.024	1.022	1.029
PhysiochemicalProtein	1.057*	1.097*	1.066*	1.073*	1.000	1.014	1.014	1.014	1.020
House	1.634	0.992	0.967	0.942	1.031	0.968	0.969	0.964	1.029
Cars	1.368	1.074	1.027	0.992	0.985	0.990	0.990	0.993	1.002
FpsBenchmark	3.480*	1.158	1.016	1.025	0.792	0.914	0.910	0.918	1.062
CaliforniaHousing	1.032	1.145*	1.006	1.010	0.996	0.999	0.999	1.001	1.006
HealthInsurance	1.099*	1.405*	1.029*	1.192*	0.994	0.994	0.994	0.986	0.996
Concrete	1.104	1.040	0.986	1.031	0.974	0.983	0.982	0.982	1.007
QsarFishToxicity	1.078	0.995	0.981	0.994	0.978	0.980	0.983	0.983	1.004
Abalone	1.113*	1.006	0.982	0.982	0.975	0.974	0.974	0.984	1.005
RealEstate	1.012	1.033	0.997	1.027	0.988	0.990	0.992	0.992	0.993
StudentPerformancePor	1.043	1.093	1.004	1.173*	1.008	1.012	1.012	0.993	0.999
EnergyEfficiency	1.190	1.270	1.265	1.282	0.842	1.115	1.090	1.104	1.058
Kin8Nm	1.150*	0.992	0.981	0.992	0.970	0.977	0.977	0.979	1.001
EnergyEfficiencyCooling	1.101	1.161	1.034	1.097	0.970	0.987	0.987	0.987	0.984
AutoMpg	1.038	1.032	0.994	0.996	0.977	0.981	0.965	0.980	0.992
CpuActivity	2.246*	1.454	1.040	0.892	0.622*	0.630*	0.630*	0.611*	0.835*
WhiteWine	1.128*	1.103*	0.999	1.003	0.997	1.000	1.000	0.996	0.999
RedWine	1.100*	1.088*	0.995	0.999	0.994	0.993	0.993	0.987	0.998
Moneyball	1.198*	0.972	0.948	0.959	0.939	0.941	0.942	0.951	1.005
AirfoilSelfNoise	1.070	1.027	1.010	1.013	0.960	0.968	0.956	0.971	0.951
WaveEnergy	4.388*	0.577*	0.289*	0.283*	0.215*	0.234*	0.234*	0.232*	0.973
Ccpp	1.013	1.010	0.999	0.995	0.989	0.991	0.992	0.991	0.996
NavalPropulsionPlant	1.311	6.730*	1.027	1.120	0.844	0.989	0.989	0.991	0.918

Table 16: HGB: median over datasets of mean per-seed RMSE ratio *method/CRC* by leverage group.

L group	n	U. quantile	Raw	Log1p	Sel. std.	Poisson loss	Gamma loss
high	11	1.086	1.046	1.006	1.003	0.998	1.019
mid	8	1.105	1.002	1.037	0.999	1.000	0.996
low	22	1.061	1.001	1.003	0.994	0.997	1.005

Table 17: Ridge regression on the datasets of Nuyts & Davis (2025): mean per-seed RMSE ratio *method/raw* (< 1: the transform helps).

Dataset	L	Log1p	G. quantile	U. quantile	CRC	CRC-log
Ccpp	1.00	0.986	1.031	1.006	0.990	0.986
AutoMpg	1.06	0.913	0.957	1.010	0.998	0.946
EnergyEfficiencyCooling	1.06	0.989	1.034	1.035	0.988	0.986
EnergyEfficiency	1.09	0.976	1.117	0.978	0.990	0.965
RealEstate	1.11	0.954	0.986	1.079	1.029	0.989
Concrete	1.12	1.114	1.029	1.099	1.026	1.033
BikeSharing	1.48	1.138	1.028	1.043	1.006	1.009
LiverDisorders	1.54	1.065	1.061	1.065	1.013	1.018
Servo	1.72	1.013	1.099	1.157	0.955	1.015
OnlineNews	26.55	1.023	1.212	1.590	1.060	1.093

Table 18: Datasets with more than 33,334 rows, MLP: capped subsample versus all rows. p: Wilcoxon (uncorrected) for the selected standard transform vs. CRC; regret as in Table 3.

Dataset	rows	U. quantile	Sel. std.	p	+smearing	CRC-log	regret val.	regret CRC
VideoTransc.	33,334	4.189	1.069	0.044	1.067	1.265	1.016	1.071
VideoTransc.	full	4.722	1.047	0.165	1.047	1.181	1.006	1.112
WaveEnergy	33,334	3.364	0.473	<0.001	0.474	0.984	1.000	2.099
WaveEnergy	full	4.388	0.234	<0.001	0.232	0.973	1.000	4.624
Diamonds	33,334	1.073	1.108	0.956	1.040	0.995	1.018	1.035
Diamonds	full	1.080	1.038	0.729	1.010	1.004	1.004	1.022
PhysiochemicalProtein	33,334	1.055	1.011	0.003	1.008	1.015	1.001	1.008
PhysiochemicalProtein	full	1.057	1.014	<0.001	1.014	1.020	1.000	1.006
OnlineNews	33,334	1.636	0.986	<0.001	0.979	1.038	1.000	1.006
OnlineNews	full	1.392	1.032	0.011	1.027	1.094	1.001	1.005

J BOUNDARY HANDLING

CRC (three ϵ, selected on validation) and the equal-mass member (median ϵ) rerun with the *clip* variant, all 41 datasets, MLP, 20 seeds, same splits.

Table 19: Boundary-handling control (medians over datasets of mean per-seed RMSE ratios). Subscript c: clip variant; EM: equal-mass member; std.: selected standard transform; U.q.: scikit-learn uniform quantile. *sig.*: CRC_c significantly better / worse than std.; *EM_c worse*: datasets with EM_c significantly worse than CRC_c (Holm).

L group	n	CRC: clip / GPD	std. / CRC_c	sig. $+/-$	EM_c / CRC_c	EM_c worse	U.q. / EM_c
high	11	1.000	1.032	0 / 0	3.029	5/11	0.970
mid	8	0.998	1.011	0 / 0	1.226	4/8	0.992
low	22	0.998	0.989	0 / 2	1.108	9/22	0.997

K VALIDATION SELECTION ON SMALL DATA

For the high-leverage datasets and each seed, we subsample n_{train} rows of the training split (without replacement) and $\max(n_{\text{train}}/3, 30)$ rows of the validation split, keep the full test split, and fit the MLP under the standard transforms, the sweep $\alpha = 1/3$ and equal-mass. A dataset enters the row for n_{train} only if its training split has at least n_{train} rows, so the number of datasets falls from 10 at $n = 250$ to 6 at $n = 5000$ (Table 4); the two largest datasets were used at their capped size. argmin K is computed from the training subsample. The candidate set is smaller than in the main suite (standard transforms, the $\alpha = 1/3$ sweep and equal-mass), so regrets are not comparable with Table 3 in level.

Table 20: Training-size experiment with the differences argmin K − validation of the median regret and of q90, each with a 95% dataset-bootstrap interval.

n_{train}	sets	val.	argmin K	CRC	Log1p	val. q90	K q90	CRC q90	Δ median $K-$val.	Δ q90 $K-$val.
250	10	1.091	1.106	1.100	1.137	1.644	1.560	1.560	0.015 [-0.026, 0.055]	-0.08 [-0.67, 0.16]
500	8	1.034	1.077	1.070	1.138	1.277	1.321	1.383	0.042 [-0.042, 0.160]	0.04 [-0.17, 0.72]
1000	7	1.023	1.040	1.093	1.093	1.222	1.129	1.197	0.017 [-0.001, 0.035]	-0.09 [-0.23, 1.49]
2000	6	1.011	1.254	1.131	1.081	1.246	1.427	1.423	0.243 [0.001, 0.710]	0.18 [-0.33, 2.98]
5000	6	1.015	1.045	1.107	1.152	1.187	1.235	1.355	0.030 [0.005, 0.061]	0.05 [-0.09, 1.67]

L FULL FAMILY ON LOW-LEVERAGE DATASETS

To save computation the main suite fitted only $\alpha \in \{0, \frac{1}{3}\}$ (plus equal-mass) in splits with training $L < 1.3$. We fitted the remaining 29 members ($\alpha \in \{\frac{2}{3}, 1\}$) on all 22 low-leverage datasets (MLP, 20 seeds). Columns: family member chosen on validation over the selected standard transform, with the reduced and the full family; best additional member over best reduced member (test RMSE, oracle in hindsight); and regret of validation over all candidates, with each candidate set.

Table 21: Low-leverage datasets with the full family (means over 20 seeds).

Dataset	fam. (gated) / std.	fam. (all 60) / std.	best extra / best gated	val. regret (gated)	val. regret (all)
Abalone	1.002	1.002	1.062	1.006	1.006
AirfoilSelfNoise	0.971	0.960	1.070	1.014	1.005
AutoMpg	0.993	0.993	1.041	1.021	1.021
CaliforniaHousing	0.994	0.994	1.011	1.004	1.004
Cars	0.997	1.003	1.050	1.017	1.023
Ccpp	0.996	0.996	1.008	1.002	1.002
Concrete	1.009	0.997	1.060	1.021	1.018
CpuActivity	0.985	0.985	2.706	1.009	1.009
EnergyEfficiency	0.783	0.787	1.139	1.013	1.017
EnergyEfficiencyCooling	0.928	0.913	1.047	1.030	1.011
FpsBenchmark	0.868	0.868	1.697	1.001	1.001
HealthInsurance	1.000	1.000	1.019	1.001	1.001
House	1.129	1.125	1.252	1.123	1.148
Kin8Nm	0.993	0.993	1.095	1.006	1.006
Moneyball	0.999	0.999	1.232	1.004	1.004
NavalPropulsionPlant	0.824	0.810	1.035	1.017	1.000
QsarFishToxicity	1.008	1.009	1.066	1.015	1.016
RealEstate	0.996	0.994	1.014	1.014	1.012
RedWine	1.007	1.006	1.017	1.015	1.014
StudentPerformancePor	0.992	0.996	1.018	1.016	1.017
WaveEnergy	0.926	0.926	13.959	1.001	1.001
WhiteWine	0.997	0.997	1.017	1.004	1.004

M TRAINING PROTOCOL

On the high- and mid-leverage datasets (10 seeds, the first ten of the main suite) we refit twelve transforms (raw, log1p, Box–Cox, both quantile transforms, equal-mass, CRC and CRC-log with three ϵ) under three protocols: *orig.-scale early stopping* (MLP (128,64) trained epoch by epoch with `partial_fit`, keeping the epoch with the best original-scale validation RMSE, patience 10, at most 200 epochs); *large* (MLP (256,128,64) with the default early stopping); *budget* (MLP (128,64), 200 epochs, no early stopping).

Table 22: Training-protocol robustness: median over datasets of the Spearman correlation between the mean test RMSE of the twelve transforms under the protocol and under the default; medians of mean per-seed RMSE ratios; mean validation regret (among the twelve).

protocol	sets	rank corr. w/ default	Std/CRC	U.q./CRC	CRC-log/CRC	val. regret
default (internal early stop.)	19	1.00	1.012	1.360	1.033	1.067
orig.-scale early stop.	19	0.93	1.004	1.526	1.038	1.021
MLP (256,128,64)	19	0.94	0.998	1.240	1.014	1.032
no early stop., 200 ep.	19	0.91	1.008	2.090	1.056	1.033

N WHERE CRC LOSES ON NEARLY DETERMINISTIC TARGETS

On WaveEnergy and CpuActivity (capped versions, 5 seeds) we split test RMSE by bands of the training-target distribution and replace CRC's cell-wise density by a Gaussian kernel density estimate, or its cell-edge knots by point knots.

Table 23: Test RMSE overall and by target band (training quantiles), mean over 5 seeds.

Dataset	method	RMSE	$< q_{10}$	q_{10}–q_{50}	q_{50}–q_{90}	$> q_{90}$
WaveEnergy	raw	2,970	3,207	3,247	2,863	1,698
WaveEnergy	CRC	5,528	13,652	3,355	2,984	6,191
WaveEnergy	CRC, KDE density	5,348	12,669	3,540	3,169	6,017
WaveEnergy	CRC, point knots	5,417	13,275	3,352	2,974	6,155
WaveEnergy	CRC-log	5,770	14,175	3,533	3,212	6,525
CpuActivity	raw	2.88	5.49	3.05	1.95	1.45
CpuActivity	CRC	4.56	13.02	2.80	1.86	1.46
CpuActivity	CRC, KDE density	4.44	12.47	2.82	1.88	1.51
CpuActivity	CRC, point knots	4.57	13.04	2.81	1.85	1.41
CpuActivity	CRC-log	3.88	10.37	2.80	1.87	1.51

CRC's excess is concentrated in the lowest decile (and, on WaveEnergy, the highest), where the density is low and the cube-root law assigns little resolution; in the central bands it is within a few percent of raw targets. Neither a smoothed density nor different knots removes it, so the loss is not an artefact of the density estimate. Section 6 discusses the explanation that is consistent with the theory and its limits.

O MORE LEARNERS

XGBoost: 500 trees, depth 6, learning rate 0.05, row and column subsampling 0.8, no early stopping. Tabular ResNet (Gorishniy et al., 2021): linear input layer to width 192, three residual blocks (batch norm, linear to 384, ReLU, dropout 0.1, linear to 192, dropout 0.05), batch norm, ReLU and a linear head; AdamW (10^{-3}, weight decay 10^{-5}), batch 256 (1024 when the training split has more than 10,000 rows), at most 200 epochs, early stopping with patience 16 on an internal 10% split. Transforms: raw, log1p, Box–Cox, both quantile transforms, equal-mass, CRC and CRC-log (three ϵ each, chosen on validation); standardized transformed targets; the first 10 splits of the main suite.

Table 24: XGBoost and tabular ResNet: median over datasets of mean per-seed RMSE ratios *method/CRC* by leverage group; *sig.*: CRC significantly better / worse than the selected standard transform (corrected resampled t-test with 10 splits, Holm over datasets; with so few splits the test has little power and we read effect sizes only); validation regret among these candidates.

learner	L group	n	U.q.	G.q.	equal-mass	Std	Std, sm.	CRC-log	Std sig. $+/-$	val. regret
XGBoost	high	11	1.313	1.050	1.473	1.008	1.000	1.010	1/0	1.049
XGBoost	mid	8	1.157	1.049	1.230	1.001	0.999	1.004	0/0	1.027
XGBoost	low	22	1.068	1.063	1.061	0.992	0.991	0.998	0/3	1.020
ResNet	high	11	2.004	1.087	2.595	1.005	1.000	1.079	0/0	1.022
ResNet	mid	8	1.097	1.066	1.202	1.002	0.995	1.005	0/0	1.012
ResNet	low	22	1.110	1.065	1.146	0.995	0.988	1.000	0/0	1.008

P COMPACT SUPPORT AND A LEARNED BIAS CORRECTION

The controlled regression of Appendix C uses a lognormal, hence unbounded, target. We reran it with the score clipped to $[-3, 3]$ and the noise drawn from a normal truncated to $[-3, 3]$, so that Y has compact support; conditional expectations use 40-point Gauss–Legendre quadrature over the truncated noise. In both versions we add a *learned* conditional bias correction: a histogram gradient-boosted regressor (300 iterations, rate 0.05) fitted on the validation rows to $y - h^{-1}(\hat{s})$, whose prediction is added to the test prediction.

Table 25: Compact versus lognormal support (12 cells, 10 seeds each): median spreads of the learning error across the sweep (absolute and normalized), mean Spearman correlation of the A1/A2 idealized learners with the true excess risk, the learned optimum k at $\rho = 0.5$ for the three s without correction, with the oracle and with the learned bias correction, and the number of cells in which the corrected optimum is at least as compressed as the uncorrected one.

support	learner	cells	abs.	norm.	ρ_S A1	ρ_S A2	$\hat{k}_{.5}$ plain	oracle	learned	or. $\leq$ pl.	le. $\leq$ pl.
compact	MLP	12	4.6	2.08	0.30	0.93	4,4,3	0,0,0	1,3,2	12/12	12/12
compact	HGB	12	4.5	2.18	0.41	0.73	4,4,3	0,0,0	0,0,1	12/12	12/12
lognormal	MLP	12	9.6	2.71	0.24	0.88	4,3,3	0,0,0	0,0,0	12/12	11/12
lognormal	HGB	12	12.0	2.09	0.37	0.77	4,3,2	0,0,0	0,0,1	10/12	12/12

Q Dependence between splits

The 20 random $60/20/20$ splits of a dataset share most of their rows, so the per-seed RMSE ratios are positively correlated and a paired Wilcoxon test that treats them as independent is anticonservative. The corrected resampled t-test of Nadeau & Bengio (1999) inflates the variance of the mean log ratio by $1/J + n_{\text{test}}/n_{\text{train}}$ ($J = 20$, $n_{\text{test}}/n_{\text{train}} = 1/3$). This test is used throughout the paper. With it, and Holm over the 41 datasets, CRC is significantly worse than the selected standard transform on 2 datasets (Wilcoxon: 8) and better on 0 (Wilcoxon: 1); the uniform quantile transform is significantly worse than CRC on 18 (Wilcoxon: 37), and CRC-log on 0 (Wilcoxon: 7). Effect sizes are unchanged; the counts of significant datasets shrink, and we base no conclusion on a single dataset's p-value. Treating datasets as the units, one-sample Wilcoxon tests of the per-dataset mean log ratios give, for the selected standard transform over CRC, $p = 0.1$ (high), 0.15 (mid), 8×10^{-4} (low); for the uniform quantile transform $p = 9 \times 10^{-13}$ over all datasets; and for CRC-log $p = 1 \times 10^{-3}$ at high leverage.

R Replicating the linear-model protocol of Nuyts and Davis

Following the released code of Nuyts & Davis (2025) (github.com/ML-KULeuven/target_transformations, commit `a4474e1`, 27 February 2025; the metric is `relative_squared_error` in `src/experiments/utils/evaluation.py`): `RepeatedKFold` (2 splits, 5 repeats, seed 0); a 70/30 train/validation split of each training fold; min-max feature scaling fitted on the training part; scikit-learn `Ridge` with penalty $10^{-4} \cdot 10^{m}$, $m \in [-3, 4]$ (29 values), chosen by validation RMSE of the transformed target; QN and QU with `QuantileTransformer`, YJ with `PowerTransformer`, Ln as $\log(y + 1)$, all with default settings, and CRC (median ϵ) for comparison. The one deliberate difference is that we use our preprocessed features (one-hot categoricals, median imputation) instead of integer category codes. For every fold we compute three metrics of the back-transformed test predictions: the RSE as implemented in the released code, $\sum_i (y_i - \hat{y}_i)^2 / \sum_i (\bar{y} - \hat{y}_i)^2$; the RSE of their Eq. 8, $\sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$, a rescaled MSE; and the RMSE. Table 26 reports, as they do, the median over datasets of the relative change of the fold-averaged metric against raw targets. The implemented ratio rewards predictions whose spread around $\bar{y}$ is large, so a transform can reduce it while increasing the squared error.

Table 26: Median over the ten datasets of the change (%) of the fold-averaged metric relative to raw targets; negative is better.

metric	QN	QU	YJ	Ln	CRC
RSE as implemented	-13.3	+1.5	-5.5	-5.6	+7.2
RSE of their Eq. 8	+6.0	+8.2	+1.8	-1.4	-1.0
RMSE	+2.9	+4.0	+0.8	-0.7	-0.5

Table 27: Fold-averaged RSE as implemented in the released code, for comparison with Table 2 (Ridge) of Nuyts & Davis (2025).

Dataset	Base	QN	QU	YJ	Ln	CRC
AutoMpg	0.189	0.166	0.211	0.172	0.168	0.196
BikeSharing	1.518	1.241	1.854	1.428	1.177	1.825
Ccpp	0.077	0.074	0.077	0.072	0.075	0.076
Concrete	0.687	0.588	0.877	0.666	0.618	0.744
EnergyEfficiency	0.095	0.102	0.088	0.090	0.092	0.093
EnergyEfficiencyCooling	0.131	0.128	0.135	0.127	0.127	0.129
LiverDisorders	3.804	2.008	2.886	2.957	3.189	4.482
OnlineNews	21.223	10.008	3.196	11.578	22.629	49.048
RealEstate	0.830	0.780	1.050	0.812	0.765	0.892
Servo	1.114	0.823	1.001	1.771	1.618	1.192