TAILOR: ADAPTIVE RESIDUAL SHAPE AND CRPS GEOMETRY FOR HETEROSCEDASTIC REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate uncertainty quantification remains a central challenge in neural regression. Gaussian heteroscedastic models can absorb mean prediction errors into larger variance estimates, weakening the updates needed to correct the mean. Decoupling helps address this feedback, but leaves the model's distributional assumptions unchanged: even a well-trained Gaussian cannot generally capture differences in residual shape by adjusting its width alone. We introduce TAILOR to address these limitations together, bringing adaptive residual modelling and orthogonal mean-scale learning into a single formulation. Rather than fixing the noise distribution in advance, TAILOR learns a shape parameter that moves smoothly between lighter-than-Gaussian, Gaussian, and heavier-tailed behaviour. Crucially, the objective used to fit this flexible family also supplies its learning geometry. Under the continuous ranked probability score (CRPS), location is orthogonal to scale and shape at the predictive level, and we use this structure to derive updates that account for the shared neural representation. This connection between flexibility and geometry extends beyond the learning rule. For small departures from Gaussianity within our family, we show that the best local Gaussian fit retains a quadratic approximation gap even after adjusting location and scale. The coefficient of this gap is precisely the geometric quantity that distinguishes shape changes from rescaling. Across regression benchmarks, TAILOR reduces RMSE by up to 74% and CRPS by up to 67% relative to tested Adam-based comparators, while achieving central prediction-interval calibration error as low as 0.6 percentage points. TAILOR brings adaptive residual modelling and orthogonal predictive geometry into one framework, addressing distributional flexibility and learning dynamics together rather than correcting them in isolation.

1 INTRODUCTION

Heteroscedastic regression must learn both a conditional prediction and the distribution of observations around it. Gaussian neural predictors make this tractable by estimating location and scale, but leave two distinct questions: whether their parameters are learned effectively, and whether a Gaussian can represent the residual distribution. Under Gaussian negative log-likelihood, the location gradient is inversely weighted by the predicted variance. The resulting interaction motivates gradient reweighting, faithful training, and geometric corrections (Seitzer et al., 2022; Stirn et al., 2023; Vemuganti and Kani, 2026). Even an optimally fitted Gaussian, however, retains a fixed standardized shape.

We study the additional distributional variation supplied by a shape parameter after allowing location and scale to adjust. This distinction matters because a shape change can partly resemble a change in width. A useful construction should therefore support more than a qualitative tail interpretation: it should establish which part of shape variation rescaling cannot reproduce, and quantify the approximation error that remains when shape is omitted.

TAILOR addresses these questions through an anchored transformation of a standard normal variable. The transformation fixes the residual center and its derivative at that center while allowing 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.

distribution away from the center to change. Our first result applies to a class of admissible generators: every nonzero generator produces a shape tangent with strictly positive squared norm after projection away from scale. For the evaluated generator, we then characterize the unique *globally* optimal Gaussian under the continuous ranked probability score (CRPS), prove a positive approximation gap at every non-Gaussian member, and show that the optimal Gaussian scale increases strictly with shape. Near Gaussianity, we calculate the leading gap coefficient explicitly through one-dimensional integrals.

The CRPS metric, natural gradients, neural pullbacks, and the projection identity used in this analysis are established tools (Gneiting and Raftery, 2007; Duan et al., 2020; Martens, 2020). Our theoretical contribution is the anchored construction and the concrete nondegeneracy and approximation results obtained for it. The projection identity explains their relationship: the residual shape component determines both the local Gaussian approximation gap and the predictive shape curvature after profiling out scale. Symmetry makes location orthogonal to scale and shape in predictive coordinates; shared neural parameters can still couple their learning.

Our experiments distinguish residual flexibility from the training procedure. The maximum improvements in the abstract compare the complete procedure with tested Adam-based comparators and include differences in training and, depending on the comparator, predictive family. Under matched full-pullback training, TAILOR instead lowers mean CRPS on seven of eight tabular datasets, with a median relative reduction of 5.8%. A controlled study holds conditional mean and variance fixed while moving from Gaussian to non-Gaussian residuals, separating population approximation from finite-sample fitting. A sinh-arcsinh comparator and application studies assess the scope and computational cost of the approach.

Our contributions are:

1. **An anchored residual construction.** We give a class of symmetric, Gaussian-containing families and prove that every admissible nonzero generator has shape variation locally distinguishable from rescaling. TAILOR is one instance of this class.
2. **A global Gaussian approximation analysis.** For TAILOR, we characterize the unique Gaussian optimizer, establish a positive gap away from Gaussianity and monotonicity of its scale, and calculate the local gap coefficient through the scale-profiled shape tangent.
3. **A separated empirical assessment.** Using established CRPS pullback geometry, we evaluate adaptive shape under matched training, compare another flexible residual family, and examine approximation, calibration, and computational limits.

2 Related work and positioning

Learning Gaussian predictors. Neural mean-variance prediction has a long history (Nix and Weigend, 1994). Subsequent work addresses how variance estimation affects mean learning: β-NLL reweights likelihood gradients (Seitzer et al., 2022), while Faithful modifies gradient propagation to preserve mean-only learning behavior (Stirn et al., 2023). Natural parameterization and Bayesian regularization improve Gaussian fitting (Immer et al., 2023); Wong-Toi et al. (2024) analyze its regularization-dependent pathologies. Fisher8 applies a Gaussian Fisher correction at the output layer (Vemuganti and Kani, 2026). These approaches motivate our training comparisons; the Gaussian implementations retain a fixed standardized residual shape.

Proper scores and predictive geometry. Proper scores evaluate whole predictive distributions (Gneiting and Raftery, 2007). Natural gradients account for predictive geometry (Amari, 1998; Martens, 2020), and NGBoost develops score-specific gradients, including for CRPS (Duan et al., 2020). TAILOR uses this established geometry and its neural pullback. Our contribution concerns the constructed family: symmetry separates location from scale and shape in predictive coordinates, while

strictly positive scale-profiled shape curvature identifies variation that rescaling cannot reproduce. Shared neural parameters can still couple learning.

Learning distributional shape. GAMLSS models covariate-dependent location, scale and shape (Rigby and Stasinopoulos, 2005); sinh-arcsinh distributions control skewness and tails (Jones and Pewsey, 2009), and adaptive robust objectives learn robustness (Barron, 2019). Monotone splines and neural transformations supply flexible parameterizations (Ramsay, 1988; Wehenkel and Louppe, 2019). Conditional transformation models learn distributions using proper scores, including CRPS (Hothorn et al., 2014). TAILOR connects this flexibility to an explicit approximation analysis. Its symmetric, one-coordinate construction recovers the Gaussian; our results prove nondegenerate shape variation for an anchored class and characterize the evaluated family's globally optimal Gaussian approximation and local CRPS gap. The theory identifies the shape variation that Gaussian rescaling misses and quantifies its predictive cost.

Comparison design. G-PB isolates the contribution of shape under matched CRPS pullback training; Gaussian stabilizers assess complete procedures, and SAS-PB compares flexible residual families. Together, these comparisons evaluate shape adaptation across Gaussian and selected non-Gaussian laws using a single CRPS objective.

3 Residual construction and CRPS objective

Let ϕ, Φ be the standard normal density and CDF. For even $a \in C^2(\mathbb{R})$ with $a(0) = 0$ and $a \not\equiv 0$, define

$$T_{\eta,a}(r) = \int_0^r e^{-\eta a(u)}\, du, \qquad F_{\eta,a}(r) = \Phi(T_{\eta,a}(r)). \tag{1}$$

The anchors $T_{\eta,a}(0) = 0$ and $T'_{\eta,a}(0) = 1$ fix center and local scale. Assumption 1 requires onto transformations, finite first moments and local L^2 differentiability (Appendix A.1).

Theorem 1 (Anchored construction and residual shape curvature)**.** *Under Assumption 1, the family is normalized, symmetric and Gaussian at $\eta = 0$. For its extension $F_{\eta,a}((z-\mu)/\sigma)$, $\sigma > 0$, location is orthogonal to scale and shape in the predictive CRPS metric. Writing $f = \partial_r F_{\eta,a}$ and $h = \partial_\eta F_{\eta,a}$,*

$$K = \min_{c \in \mathbb{R}} \int_{\mathbb{R}} (h - crf)^2\, dr > 0, \qquad S_\eta = 2\sigma K, \tag{2}$$

where S_η is the scale-profiled shape curvature.

Anchoring rules out $h = crf$ unless $a \equiv 0$. The result includes $a_p(r) = \log(1 + r^{2p})/(2p)$ for positive integers p and $\eta < 1$. TAILOR evaluates $p = 1$:

$$T_\eta(r) = \int_0^r (1 + u^2)^{-\eta/2}\, du, \qquad F_\eta(r) = \Phi(T_\eta(r)). \tag{3}$$

With $q = (\mu, s, \eta)$, $\sigma = e^s$ and $r = (z - \mu)/\sigma$, its density and predictive law are

$$\begin{aligned} f_\eta(r) = \phi(T_\eta(r))(1 + r^2)^{-\eta/2}, \qquad h_\eta(r) = \partial_\eta F_\eta(r), \\ F_q(z) = F_\eta(r), \qquad p_q(z) = \sigma^{-1} f_\eta(r). \end{aligned} \tag{4}$$

The mean is μ and the standard deviation is $\sigma\sqrt{\mathrm{Var}(R_\eta)}$ for $R_\eta \sim F_\eta$. All absolute moments are finite for $\eta \le 1$; negative and positive shape give lighter and heavier tails than Gaussian. Training uses $\eta_{\min} < 0 < \eta_{\max} < 1$ and minimizes

$$L(\theta) = \frac{1}{n} \sum_i S(F_{q_\theta(x_i)}, y_i), \quad S(F, y) = \int_{\mathbb{R}} \big(F(z) - \mathbf{1}\{y \le z\}\big)^2\, dz. \tag{5}$$

For finite first moments, excess expected CRPS is the squared CDF distance (Gneiting and Raftery, 2007):

$$\mathbb{E}_{Y\sim Q}[S(P,Y) - S(Q,Y)] = \mathcal{D}(Q,P) := \int_{\mathbb{R}} (F_P(z) - F_Q(z))^2 \, dz. \tag{6}$$

Conditional averaging requires the corresponding integrability.

4 Gaussian approximation and CRPS geometry

For an admissible anchored family, write $f = \partial_r F$ and $h = \partial_\eta F$ and define

$$A = \int f^2 \, dr, \quad B = \int r^2 f^2 \, dr, \quad C = \int r f h \, dr, \quad D = \int h^2 \, dr. \tag{7}$$

Proposition 1 (Predictive CRPS metric)**.** *At regular interior parameters, the CRPS metric (Duan et al., 2020) specializes to*

$$G(q) = 2 \int \nabla_q F_q \nabla_q F_q^\top \, dz = 2 \begin{pmatrix} A/\sigma & 0 & 0 \\ 0 & \sigma B & -\sigma C \\ 0 & -\sigma C & \sigma D \end{pmatrix}, \tag{8}$$

with $\mathcal{D}(q, q+dq) = \frac{1}{2} dq^\top G(q) dq + o(\|dq\|^2)$.

The location tangent is even; scale and shape tangents are odd. Thus location cross terms vanish, while scale and shape generally remain coupled. Profiling out scale gives

$$\begin{aligned} K(\eta) &= D - C^2/B = \int \big(h - (C/B) r f\big)^2 \, dr, \\ S_\eta &= G_{\eta\eta} - G_{\eta s} G_{ss}^{-1} G_{s\eta} = 2\sigma K(\eta) > 0. \end{aligned} \tag{9}$$

Theorem 1 proves strict nondegeneracy throughout the anchored class, linking shape variation to the scale-profiled curvature in (9). Appendix M gives the derivations.

Theorem 2 (Global Gaussian approximation and residual shape curvature)**.** *For* $Q_\eta(y) = F_\eta((y - \mu)/\sigma)$*,* $\eta \le 1$*, the unique Gaussian minimizer of population CRPS satisfies*

$$m_\eta^* = \mu, \quad a_\eta^* = \sigma \rho_\eta, \quad \mathbb{E}_{R\sim F_\eta} \exp\left(-\frac{R^2}{2\rho_\eta^2}\right) = \frac{1}{\sqrt{2}}. \tag{10}$$

The unique positive root increases strictly with η*, with* $\rho_0 = 1$*. The minimum excess risk* Δ_η *is zero at* $\eta = 0$ *and strictly positive otherwise. As* $\eta \to 0$*,*

$$\Delta_\eta = \sigma K(0) \eta^2 + o(\eta^2) = \tfrac{1}{2} S_\eta\big|_{\eta=0} \eta^2 + o(\eta^2). \tag{11}$$

The optimizer is global; the expansion is local. Here $K(0) \approx 0.00143703555$ measures shape variation Gaussian rescaling cannot absorb. Scale monotonicity holds at fixed σ, not necessarily at fixed variance. The gap excludes estimation and optimization effects. Appendices A.6 and M give proofs and the variance-normalized interpretation.

5 Neural training

For a minibatch of m scalar predictions, stack their coordinates into q_θ and let L be average CRPS. The neural pullback (Martens, 2020) is

$$\begin{aligned} J = \partial q_\theta / \partial \theta, \quad W = m^{-1} \operatorname{blockdiag}(G_i), \quad a = \nabla_q L, \\ H = J^\top W J, \qquad g = J^\top a = \nabla_\theta L. \end{aligned} \tag{12}$$

For an exact undamped solve, the attainable predictive change is

$$J\delta^* = P_W v^*, \quad v^* = -W^{-1}a, \quad P_W = J(J^\top W J)^\dagger J^\top W. \tag{13}$$

Here P_W is the W-orthogonal projector onto $\mathrm{range}(J)$; shared parameters can couple mean learning. Repeated global shape coordinates and output-map derivatives remain in J. In practice we solve

$$(H + \lambda I)\delta_\lambda = -g, \qquad \lambda > 0, \tag{14}$$

using matrix-free conjugate gradients and safeguarded Armijo backtracking (Algorithm 1; Appendices B, E and M).

6 EXPERIMENTS

Our experiments ask when adaptive shape improves prediction and how much of the gain comes from training. CRPS-trained arms pair Gaussian or TAILOR predictors with Adam (Kingma and Ba, 2015), giving G-Adam and A-Adam, or full pullback, giving G-PB and TAILOR. A-OUT uses output preconditioning; SAS-PB uses symmetric sinh-arcsinh with its own pullback (Jones and Pewsey, 2009). Gaussian-likelihood comparators are G-NLL, β-NLL ($\beta = 0.5$), Faithful and Fisher8. G-NLL is distinct from G-Adam. Validation CRPS selects hyperparameters and checkpoints (Table 6; Appendices E, L, N). All paired 95% intervals are pointwise, without multiplicity adjustment.

6.1 DEPARTURE FROM GAUSSIANITY AT FIXED CONDITIONAL VARIANCE

We first isolate shape from changes in mean and variance. For $R_\eta \sim F_\eta$ and $v_\eta = \mathrm{Var}(R_\eta)$, we generate

$$\epsilon_\eta = R_\eta / \sqrt{v_\eta}, \qquad Y_\eta = m(X) + \tau(X)\epsilon_\eta,$$

using Appendix G's functions, with $\mathbb{E}[\tau(X)] = 0.275$. Thus every shape has the same conditional mean $m(X)$ and variance $\tau(X)^2$ (Figure 1(a)).

Shape creates an approximation gap even at fixed variance. By the linear scaling of CRPS, Theorem 2's unit-distributional-scale gap gives

$$\Delta_\eta = \frac{0.275}{\sqrt{v_\eta}} \Delta_\eta^{\mathrm{unit}} = 0.275 K(0)\eta^2 + o(\eta^2).$$

Optimising Gaussian scale removes this gap only at $\eta = 0$. Figure 1(b) evaluates 51 shapes in $[-0.3, 0.7]$ against the local quadratic reference. At $\eta = 0.5$, $v_\eta = 1.4115346845$ and $\Delta_\eta = 0.000133551$.

Fitted models recover the predicted shape benefit. Tables 15 and 16 fix location and use teachers with unit distributional scale. If E_G is Gaussian excess risk above the population-optimal Gaussian and E_A is adaptive excess risk above the teacher, then

$$R(\widehat{G}) - R(\widehat{A}) = \Delta_\eta^{\mathrm{unit}} + E_G - E_A.$$

At n = 20,000, fitted shapes are 0.01 ± 0.05 for Gaussian data and 0.49 ± 0.05 for η = 0.5 (sample SDs; Table 16). The unit-scale fitted gain is -0.000008 for Gaussian data and 0.000572979 at η = 0.5, close to the population gap 0.000576979. Figure 1(b) rescales gains by $0.275/\sqrt{v_\eta}$. Table 15 gives the sample-size progression; η = 0.8 is a boundary stress test, unattainable at finite shape logits.

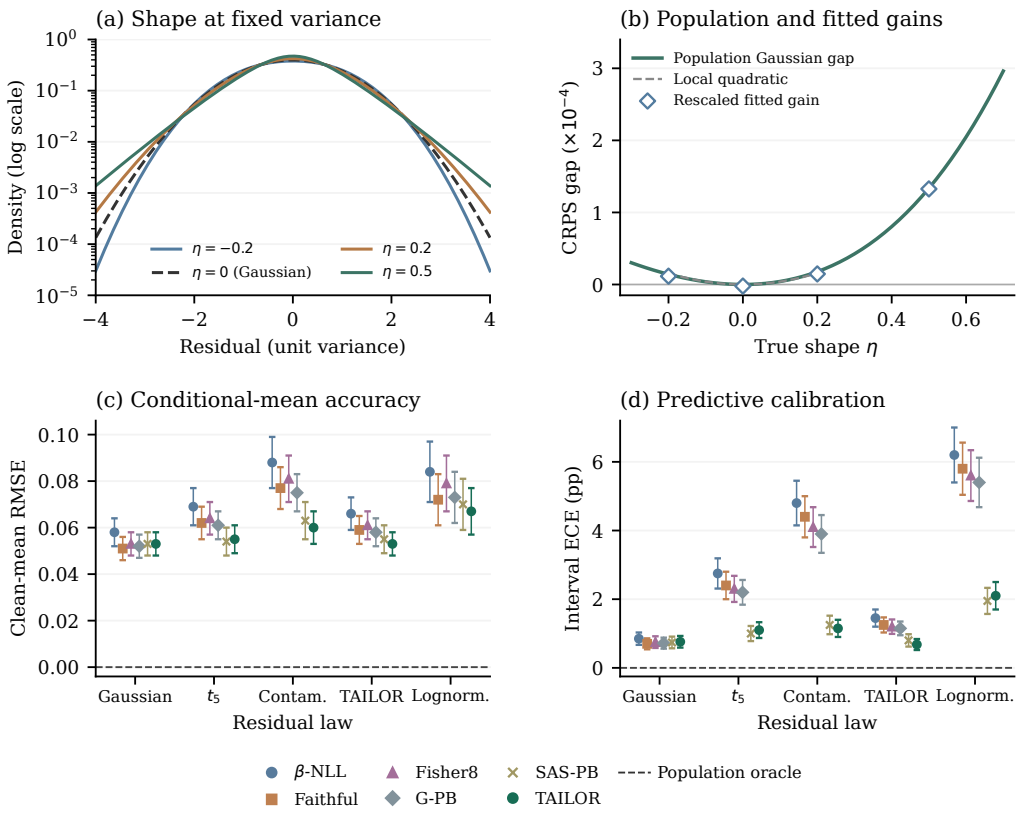

Figure 1: Shape adaptation and predictive accuracy. (a-b) Equal-variance densities and Gaussian CRPS gaps, with rescaled fitting gains (Table 16). (c-d) Clean-mean RMSE and interval ECE over five residual laws; points are means and bars sample SDs over 20 repetitions (Table 31). Dashed lines mark population oracles. The expected finite-test oracle ECE is 0.238 pp. TAILOR denotes the η = 0.5 teacher; Table 30 reports CRPS outcomes.

Shape improves prediction across the tested departures. The five centred, unit-variance laws are Gaussian, Student-t_5, symmetric contamination, TAILOR (η = 0.5) and lognormal. TAILOR keeps the same architecture and CRPS objective across laws. Under 20 repetitions with 10,000/2,000/20,000 training/validation/test observations, TAILOR has lower mean CRPS than the Gaussian baselines on all four departures (Table 30). Contamination reductions are 3.5% versus G-PB and 3.8% versus Fisher8. SAS-PB leads on Student-t_5 and lognormal.

Point accuracy and calibration improve together. Figure 1(c-d) reports RMSE against the known conditional mean m(X) and the nine-level interval ECE of Appendix E. Across the four departures, TAILOR lowers clean-mean RMSE by 6.9-31.8% and ECE by 43.3-76.0% relative to β-NLL, Faithful and Fisher8 (Table 31). Under Gaussian noise, their scores remain close. The comparison holds conditional mean, variance, network architecture and the TAILOR objective fixed across residual laws.

Oracle references quantify the remaining error. The true mean and CDF give zero population clean-mean RMSE and ECE. Noisy-label RMSE instead has a 0.29297 population floor; empirical oracle ECE has a 0.238 pp sampling reference. Under contamination, excess CRPS above the teacher falls from 0.02903 for Fisher8 to 0.02303 for TAILOR (Table 31). This excess includes fitting and family error; Table 28 separates the intrinsic gap in the frozen-encoder control.

Table 1: Gaussian baselines and adaptive shape. (a) Two of five unit-variance laws, 20 repetitions; both use input-dependent noise scale. All laws appear in Table 30. (b) Standardised-target tabular means; source comparisons in Tables 13 and 29. $\beta = 0.5$ is fixed. Bold marks lowest displayed means; paired G-PB/SAS-PB contrasts appear in Tables 8-9.

(a) Controlled residuals: CRPS, mean (SD)

Residual	G-NLL	β-NLL	Faithful	Fisher8	G-PB	TAILOR
Gaussian	.1692 (.0020)	.1626 (.0016)	.1619 (.0015)	.1620 (.0015)	**.1618** (.0015)	.1623 (.0016)
Contamination	.1689 (.0032)	.1587 (.0024)	.1574 (.0023)	.1560 (.0022)	.1555 (.0022)	**.1500** (.0020)

(b) Tabular regression: CRPS means

Dataset	G-NLL	β-NLL	Faithful	G-PB	TAILOR	SAS-PB
Energy	.0640	.0600	.0610	.054	**.0490**	.051
Kin8nm	.2310	.2280	.2305	.230	**.2135**	.219
Protein	.4710	.4670	.4685	.479	**.4640**	.466
Beijing	.3560	.3500	.3520	.357	**.3290**	.335
Naval	.0062	.0058	.0060	.0041	**.0040**	**.0040**
Concrete	.2940	.2860	.2880	.285	**.2665**	.272
Power	.1280	.1260	.1275	.127	**.1205**	.123
Wine	.4620	.4610	.4615	**.460**	.4620	.461

6.2 Tabular prediction under matched training

The primary family comparison matches TAILOR and G-PB under CRPS full-pullback training. Models have two width-64 tanh layers, input-dependent location and scale, and global shape. Seven datasets use 20 matched 80%/10%/10% split and initialisation repetitions; Beijing uses 20 rolling origins with six-month validation/test windows. CRPS uses standardised targets; RMSE and interval widths use original units.

Adaptive shape improves mean CRPS on seven of eight datasets. The median reduction over G-PB is 5.8% (Table 1(b); full arms in Table 29). Paired intervals favour TAILOR on Energy, Kin8nm, Protein, Concrete and Power; Beijing, Naval and Wine remain unresolved (Figure 2; Table 8). Against SAS-PB, intervals favour TAILOR on Energy, Kin8nm, Concrete and Power, while Protein, Naval and Wine include zero. Beijing's SAS comparison is descriptive because no temporal interval is reported (Table 9).

Training and shape make complementary contributions. On Naval, CRPS falls from 0.0120 with A-Adam to 0.0041 with G-PB and 0.0040 with TAILOR. Full-pullback training supplies most of this improvement; adaptive shape yields a further reduction under matched training (Tables 7 and 10). Protein's marginal interval ECE improves to 0.60 percentage points from G-PB's 1.10, alongside a calibrated-IID sampling reference of 0.498 points. This metric evaluates aggregate interval coverage (Gneiting et al., 2007).

6.3 Weak-lensing parameter prediction

For scientific parameter prediction, we follow Vemuganti and Kani (2026) and predict (Ω_m, S_8) from 128×128 maps spanning 101 cosmologies, with 256 realisations each split 143/36/77 for training/validation/test and grouped by simulation/initial-condition identity. Each target has its own location, scale and shape. Five matched initialisations share this within-cosmology partition; training averages standardised marginal CRPS.

TAILOR lowers mean marginal CRPS from 0.1530 to 0.1470 versus G-PB, a 3.9% reduction with paired interval [0.00328, 0.00872] (Table 2). It also achieves the lowest displayed RMSE and density NLL means for both parameters among the compared methods.

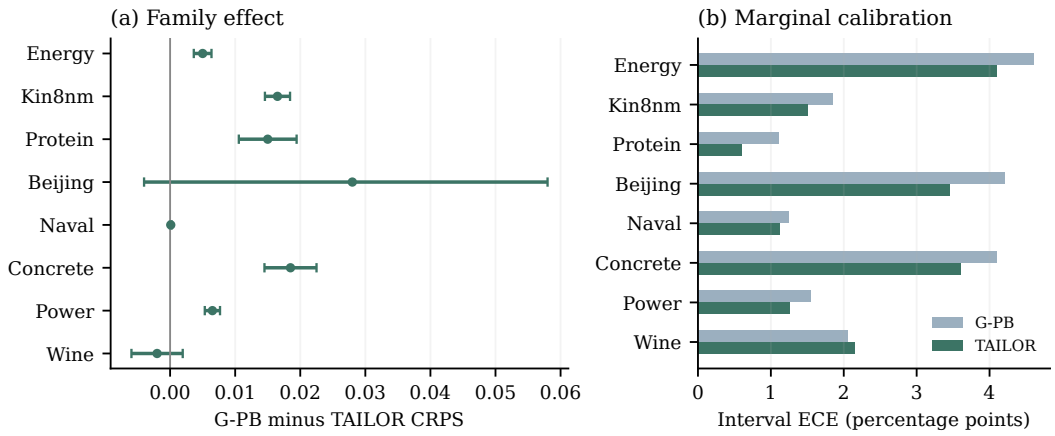

Figure 2: Matched-training evidence. Left: G-PB minus TAILOR CRPS with pointwise 95% paired intervals (Table 8); positive favours TAILOR. Beijing uses target-date block bootstrapping. Right: marginal central-interval calibration errors (Table 12); lower is better. The panels separate paired evidence for score improvements from descriptive marginal-calibration comparisons.

Table 2: Weak-lensing prediction, mean ± sample SD over five matched initialisations. CRPS averages standardised marginal scores; RMSE and marginal density NLL use original parameter units. Coverage is in percent, with nominal level 68%. Bold marks the best mean or coverage closest to nominal. The partition is within cosmology.

Method	Marginal CRPS	Cov_{68,Ω_m}	Cov_{68,S_8}
A-Adam	$.1650 \pm .0054$	63.6 ± 1.4	61.6 ± 1.8
G-NLL	$.1661 \pm .0057$	63.0 ± 1.6	61.1 ± 1.9
β-NLL	$.1604 \pm .0048$	64.7 ± 1.3	62.8 ± 1.6
Faithful	$.1590 \pm .0046$	64.3 ± 1.4	63.4 ± 1.5
Fisher8	$.1560 \pm .0041$	63.2 ± 1.2	60.8 ± 1.5
G-PB	$.1530 \pm .0046$	64.8 ± 1.3	63.1 ± 1.6
TAILOR	$\mathbf{.1470} \pm .0043$	$\mathbf{66.6} \pm 1.1$	$\mathbf{65.0} \pm 1.4$

Method	RMSE_{Ω_m}	RMSE_{S_8}	NLL_{Ω_m}	NLL_{S_8}
A-Adam	$.0464 \pm .0012$	$.0341 \pm .0008$	$-1.660 \pm .040$	$-1.880 \pm .060$
G-NLL	$.0471 \pm .0013$	$.0344 \pm .0009$	$-1.642 \pm .043$	$-1.858 \pm .064$
β-NLL	$.0448 \pm .0011$	$.0328 \pm .0008$	$-1.709 \pm .039$	$-1.924 \pm .059$
Faithful	$.0428 \pm .0010$	$.0317 \pm .0007$	$-1.718 \pm .038$	$-1.949 \pm .057$
Fisher8	$.0425 \pm .0010$	$.0313 \pm .0007$	$-1.750 \pm .035$	$-1.970 \pm .055$
G-PB	$.0420 \pm .0010$	$.0310 \pm .0007$	$-1.770 \pm .033$	$-1.990 \pm .052$
TAILOR	$\mathbf{.0407} \pm .0009$	$\mathbf{.0301} \pm .0006$	$\mathbf{-1.820} \pm .032$	$\mathbf{-2.040} \pm .050$

TAILOR also achieves the closest coverage to nominal among the compared methods: 66.6% for Ω_m and 65.0% for S_8 at a 68% target. These gains support improved marginal prediction within the evaluated cosmologies. The CRPS interval conditions on the fixed partition and observations; joint-posterior prediction and transfer to unseen cosmologies remain outside this evaluation.

6.4 Rotated MNIST and learned representations

Rotated MNIST tests training and representation behaviour under Gaussian noise (Immer et al., 2023; Vemuganti and Kani, 2026). Clean angles are uniform over $[-90°, 90°]$; label-noise SD is $11c + 1$ degrees for digit class c, or a constant $10°$. A digit probe uses frozen regression features and validation-selected regularisation. Five initialisations share the split, rotations and noise, so uncertainty is conditional on these quantities.

Table 3: Digit-dependent rotated MNIST, mean ± sample SD over five paired initialisations. RMSE uses clean angles; CRPS/NLL use noisy labels; the probe measures digit information. Bold marks the best displayed mean (including ties). The population noise-only CRPS reference is 28.4916°.

Method	RMSE (°)	CRPS (°)	NLL	ECE (pp)	Probe (%)
G-Adam	$22.40 \pm .74$	$32.91 \pm .42$	$5.37 \pm .02$	$2.70 \pm .42$	63.40 ± 1.10
A-Adam	$21.40 \pm .68$	$32.68 \pm .39$	$5.37 \pm .02$	$2.50 \pm .37$	$71.00 \pm .94$
G-NLL	$23.05 \pm .80$	$33.20 \pm .46$	$5.39 \pm .03$	$3.05 \pm .46$	62.20 ± 1.15
β-NLL	$20.75 \pm .58$	$32.35 \pm .36$	$5.35 \pm .02$	$1.85 \pm .30$	$78.40 \pm .72$
Faithful	$\mathbf{20.05} \pm .49$	$32.42 \pm .37$	$5.36 \pm .02$	$1.95 \pm .32$	$72.60 \pm .86$
Fisher8	$20.80 \pm .47$	$32.29 \pm .32$	$\mathbf{5.34} \pm .02$	$1.80 \pm .31$	$81.80 \pm .57$
G-PB	$20.50 \pm .55$	$32.19 \pm .36$	$\mathbf{5.34} \pm .02$	$1.70 \pm .29$	$81.40 \pm .66$
TAILOR	$20.20 \pm .51$	$\mathbf{32.08} \pm .34$	$\mathbf{5.34} \pm .02$	$\mathbf{1.30} \pm .26$	$\mathbf{82.10} \pm .61$

Under digit-dependent noise, TAILOR leads mean CRPS, interval ECE and probe accuracy (Table 3). Its 0.11° CRPS gain over G-PB has interval [-0.033, 0.253]; Faithful leads clean-angle RMSE. Constant-noise CRPS is 9.81° for TAILOR and 9.73° for G-PB (Table 19). These controls assess Gaussian prediction and representations alongside the non-Gaussian shape gains.

6.5 Damping, computational cost and matched budgets

TAILOR achieves the lowest mean CRPS at the longest evaluated budget. Tables 26-27 tune damping independently on one Kin8nm partition, addressing Table 18's differing conventions: five paired initialisations, nine candidates per method (Appendix L). At 900 seconds, TAILOR reaches 0.2149 versus A-OUT-BN's 0.2176, a 1.24% reduction with paired interval [0.00102, 0.00438]. SAS-PB reaches 0.2160; its difference interval [-0.00089, 0.00309] includes zero. A-OUT-BN leads at 90 and 180 seconds; the 450-second comparison remains unresolved. Search allowance is 9 × 900 = 8,100 seconds per method/initialisation, separate from selected-run time. This comparison conditions on the fixed partition and excludes G-PB.

7 Discussion and conclusion

TAILOR addresses residual-shape mismatch that remains after optimal Gaussian location and scale fitting. Its anchored construction separates shape from rescaling: the orthogonal shape component determines both local CRPS approximation error and scale-profiled curvature. We characterize the unique globally optimal Gaussian for the evaluated family and prove a positive approximation gap at every non-Gaussian member.

Under matched full-pullback training, TAILOR improves mean CRPS on seven of eight tabular datasets, with a median reduction of 5.8%. Across four non-Gaussian laws, it improves mean CRPS, clean-mean RMSE and interval ECE over β-NLL, Faithful and Fisher8. RMSE reductions range from 6.9-31.8%, and ECE reductions from 43.3-76.0% (Table 31). TAILOR also achieves the lowest mean CRPS, RMSE and NLL among the compared weak-lensing methods.

TAILOR combines a quantified approximation benefit with adaptation across residual regimes. Within the controlled task, the same architecture and CRPS objective handle Gaussian and non-Gaussian noise through learned shape.

REPRODUCIBILITY STATEMENT

We will release the code upon acceptance. Implementation is detailed in Appendices B to E, protocols/results in F to L, supporting mathematics in M, expanded baselines in N and departure metrics in O.

AI USE STATEMENT

Generative AI tools were used to assist with grammar correction and the preparation of LaTeX code for figures. The authors take responsibility for the final content.

REFERENCES

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251-276, 1998. doi:10.1162/089976698300017746.

Jonathan T. Barron. A general and adaptive robust loss function. In *CVPR*, pp. 4331-4339, 2019. https://arxiv.org/abs/1701.03077.

Tony Duan, Anand Avati, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. NGBoost: Natural gradient boosting for probabilistic prediction. In *ICML*, PMLR 119:2690-2700, 2020. https://proceedings.mlr.press/v119/duan20a.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359-378, 2007. doi:10.1198/016214506000001437.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society B*, 69(2):243-268, 2007. doi:10.1111/j.1467-9868.2007.00587.x.

Torsten Hothorn, Thomas Kneib, and Peter Bühlmann. Conditional transformation models. *Journal of the Royal Statistical Society B*, 76(1):3-27, 2014. doi:10.1111/rssb.12017.

Alexander Immer, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Effective Bayesian heteroscedastic regression with deep neural networks. In *NeurIPS*, volume 36, 2023.

M. C. Jones and Arthur Pewsey. Sinh-arcsinh distributions. *Biometrika*, 96(4):761-780, 2009. doi:10.1093/biomet/asp053.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. https://arxiv.org/abs/1412.6980.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1-76, 2020. https://jmlr.org/papers/v21/17-678.html.

David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 1, pp. 55-60, 1994. doi:10.1109/ICNN.1994.374138.

PyTorch Contributors. `torch.func`: Composable function transforms. PyTorch documentation, 2026. https://docs.pytorch.org/docs/stable/func.html.

J. O. Ramsay. Monotone regression splines in action. *Statistical Science*, 3(4):425-461, 1988. doi:10.1214/ss/1177012761.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society C*, 54(3):507-554, 2005. doi:10.1111/j.1467-9876.2005.00510.x.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *ICLR*, 2022. https://arxiv.org/abs/2203.09168.

Andrew Stirn, Hans-Hermann Wessels, Megan Schertzer, Laura Pereira, Neville E. Sanjana, and David A. Knowles. Faithful heteroscedastic regression with neural networks. In *AISTATS*, PMLR 206:5593-5613, 2023. https://proceedings.mlr.press/v206/stirn23a.html.

Sumedh Vemuganti and Nickvash Kani. Fisher8: Stabilizing neural heteroscedastic regression via output-layer Fisher geometry. In *UAI*, PMLR 337:6885-6899, 2026. https://proceedings.mlr.press/v337/vemuganti26a.html.

Antoine Wehenkel and Gilles Louppe. Unconstrained monotonic neural networks. In *NeurIPS*, volume 32, 2019.

Eliot Wong-Toi, Alex Boyd, Vincent Fortuin, and Stephan Mandt. Understanding pathologies of deep heteroskedastic regression. In *UAI*, PMLR 244:3765-3790, 2024. https://proceedings.mlr.press/v244/wong-toi24a.html.

A PROOFS

A.1 TECHNICAL CONDITIONS AND ADDITIONAL GENERATORS

Assumption 1 (Admissible anchored generator)**.** *Let $I \subset \mathbb{R}$ be an open interval containing zero. The generator $a \in C^2(\mathbb{R})$ is even, satisfies $a(0) = 0$, and is not identically zero. For every $\eta \in I$, $T_{\eta,a}$ maps $\mathbb{R}$ onto $\mathbb{R}$, and the residual law has finite first absolute moment. For $q = (\mu, s, \eta) \in \mathbb{R}^2 \times I$, define $F_q(z) = F_{\eta,a}(e^{-s}(z-\mu))$. For each q_0, the local CDF-difference map $q \mapsto F_q - F_{q_0} \in L^2(\mathbb{R})$ is twice continuously Fréchet differentiable near q_0.*

A concrete generator class is

$$a_p(r) = \frac{\log(1 + r^{2p})}{2p}, \qquad p = 1, 2, \ldots, \tag{A.1}$$

which gives

$$T_{\eta,p}(r) = \int_0^r (1 + u^{2p})^{-\eta/(2p)}\, du. \tag{A.2}$$

For every fixed p, the transformation is onto exactly for $\eta \le 1$. At $\eta < 1$ it grows as $r^{1-\eta}/(1-\eta)$ in the positive tail, while at $\eta = 1$ it grows logarithmically. On $I = (-\infty, 1)$ these tails imply all polynomial moments and the local regularity in Assumption 1. The endpoint is handled separately; training uses a compact interval with $\eta_{\max} < 1$. For example, $p = 2$ yields

$$T_{\eta,2}(r) = \int_0^r (1 + u^4)^{-\eta/4}\, du, \qquad F_{\eta,2}(r) = \Phi(T_{\eta,2}(r)). \tag{A.3}$$

It has the same admissible range, Gaussian member, regularity, location orthogonality and positive residual shape curvature as the evaluated $p = 1$ family. The neural experiments evaluate $p = 1$ only.

A.2 PROOF OF THEOREM 1

Because a is even, $e^{-\eta a(r)}$ is even and positive. Hence $T_{\eta,a}$ is odd and strictly increasing, and the onto assumption makes it a bijection of $\mathbb{R}$. Changing variables $v = T_{\eta,a}(r)$ gives $\int f_{\eta,a}(r)\, dr = \int \phi(v)\, dv = 1$. Oddness implies symmetry; at $\eta = 0$, $T_{0,a}(r) = r$, so the residual law is standard normal. The remaining argument follows on the next page.

Differentiating the transformation gives

$$\partial_\eta T_{\eta,a}(r) = -\int_0^r a(u)e^{-\eta a(u)}\,du, \qquad h_{\eta,a}(r) = -\phi(T_{\eta,a}(r))\int_0^r a(u)e^{-\eta a(u)}\,du. \tag{A.4}$$

The density f is even, whereas rf and h are odd. Thus the products of the location tangent $-f/\sigma$ with the log-scale and shape tangents are odd, and both location cross terms in the CRPS metric vanish.

The assumed L^2 regularity makes B, C, D finite. Since f is positive, $B > 0$. Projection of h onto the span of rf gives

$$\left\| h - \frac{C}{B} rf \right\|_2^2 = D - \frac{C^2}{B} = K \geq 0.$$

It remains to rule out equality. If $K = 0$, continuity upgrades the resulting L^2 identity to $h(r) = c\,rf(r)$ for every r. Smooth evenness and $a(0) = 0$ give $a(r) = O(r^2)$ near zero. Using $T'_{\eta,a}(0) = 1$ and Eqn. (A.4),

$$\frac{h(r)}{rf(r)} = -\frac{\int_0^r a(u)e^{-\eta a(u)}\,du}{re^{-\eta a(r)}} = O(r^2) \longrightarrow 0.$$

Therefore $c = 0$. Eqn. (A.4) then implies $\int_0^r a(u)e^{-\eta a(u)}du = 0$ for every r. Differentiating gives $a(r)e^{-\eta a(r)} = 0$, hence $a \equiv 0$, a contradiction. Thus $K > 0$. The scale-shape block of the metric is $2\sigma \left(\begin{smallmatrix} B & -C \\ -C & D \end{smallmatrix}\right)$; its Schur complement is $2\sigma K > 0$.

For Eqn. (A.1), a_p is smooth, even, vanishes at zero, and is nonzero. Moreover, $T'_{\eta,p}(r) = (1 + r^{2p})^{-\eta/(2p)} > 0$ and, as $r \to \infty$,

$$T'_{\eta,p}(r) = r^{-\eta}(1 + o(1)).$$

Consequently $T_{\eta,p}$ is onto if and only if $\eta \leq 1$; $T_{\eta,p}(r) \sim r^{1-\eta}/(1-\eta)$ for $\eta < 1$ and $T_{1,p}(r) \sim \log r$ at the boundary. For $\eta < 1$, the normal factor $\phi(T_{\eta,p}(r))$ supplies a stretched-exponential envelope that dominates all polynomial and logarithmic factors arising in the required moments and parameter derivatives, verifying Assumption 1 on $I = (-\infty, 1)$. At the boundary $\eta = 1$, set $x = \log r$. Ordinary-moment integrands then have tails of the form $\exp(-x^2/2 + O(x))$, while squared quantities have tails of the form $\exp(-x^2 + O(x))$. Thus in either case they are bounded by a polynomial in x times $\exp(-cx^2 + O(x))$ for some $c > 0$, which is integrable. This separately verifies the required fixed-parameter moment and L^2 finiteness at $\eta = 1$ for $p = 1$ and $p = 2$ (and, by the same argument, every fixed positive integer p).

A.3 Admissibility of the evaluated family

Proposition 3 (Admissibility of the evaluated family)**.** *For every fixed $\eta \leq 1$, T_η is an odd, strictly increasing bijection of $\mathbb{R}$. Equations 3- 4 define a normalized symmetric distribution with all absolute moments finite. The member $\eta = 0$ is Gaussian; $\eta < 0$ gives lighter tails and $0 < \eta \leq 1$ gives heavier tails.*

Proof. The derivative $T'_\eta(r) = (1 + r^2)^{-\eta/2}$ is positive. For $\eta < 1$,

$$T_\eta(r) \sim \frac{r^{1-\eta}}{1-\eta} \qquad (r \to +\infty),$$

while $T_1(r) = \operatorname{asinh}(r)$. Hence the limits are $\pm\infty$. Changing variables $v = T_\eta(r)$ gives $\int f_\eta(r)dr = 1$, and oddness of T_η gives symmetry. For fixed $\eta < 1$,

$$\log f_\eta(r) = -\frac{|r|^{2(1-\eta)}}{2(1-\eta)^2}(1 + o(1)), \qquad |r| \to \infty,$$

which proves finite polynomial moments and the stated tail ordering. At $\eta = 1$, $R = \sinh Z$ for $Z \sim N(0, 1)$, so every absolute polynomial moment is finite. At $\eta = 0$, $T_0(r) = r$. □

A.4 Proof of Proposition 1

Insert the derivatives in Eqn. (M.3) into $2\int \nabla F \nabla F^\top dz$ and use $dz = \sigma dr$. The residual density is even, whereas $r f_\eta$ and h_η are odd. The location-scale and location-shape integrals therefore vanish, yielding equation 8. The local expansion follows from first-order differentiability of the local difference map $q \mapsto F_q - F_{q_0}$ in L^2. Cauchy-Schwarz gives $K \geq 0$, with equality exactly when h_η is proportional to $r f_\eta$ in L^2. Profiling the scale block gives equation 9.

A.5 Additional predictive-geometry consequences

Corollary 1 (Diagonal predictive chart)**.** *Define*

$$\ell = s - \int_0^\eta \frac{C(t)}{B(t)}\, dt. \tag{A.5}$$

In coordinates (μ, ℓ, η), the predictive metric is diagonal:

$$G_{(\mu,\ell,\eta)} = 2\,\mathrm{diag}(A/\sigma,\, \sigma B,\, \sigma K)\,. \tag{A.6}$$

This analytic chart was not used in the experiments. It does not make the objective separable or guarantee uncoupled shared-network training; a consistent coordinate change leaves the exact pullback unchanged only when the transformed metric and all Jacobian factors are included.

Proof. Eqn. (A.5) gives $ds = d\ell + [C(\eta)/B(\eta)]d\eta$. Substitution into the scale-shape quadratic form $2\sigma(B\, ds^2 - 2C\, ds\, d\eta + D\, d\eta^2)$ cancels the cross term and leaves $2\sigma(B\, d\ell^2 + K\, d\eta^2)$. The location cross terms already vanish by Proposition 1. □

Proposition 4 (Nondegeneracy throughout the admissible shape range)**.** *For every finite $\eta \leq 1$, $K(\eta) > 0$. On every compact shape interval contained in $(-\infty, 1)$, K is continuous and bounded away from zero.*

Proof. Equality in the Cauchy-Schwarz inequality defining K would require $h_\eta(r) = c\, r f_\eta(r)$ almost everywhere, and by continuity everywhere. Near zero,

$$\partial_\eta T_\eta(r) = -r^3/6 + O(r^5), \qquad T'_\eta(r) = 1 + O(r^2), \qquad \frac{h_\eta(r)}{r f_\eta(r)} = -r^2/6 + O(r^4).$$

The limiting value would force $c = 0$, but $h_\eta(r) < 0$ for every $r > 0$, a contradiction. Finiteness follows from the tail bounds in Proposition 3; at $\eta = 1$, use $T_1(r) = \mathrm{asinh}(r)$. The same uniform tail bounds on compact intervals below one justify continuity of the metric integrals. A positive continuous function has a positive minimum on a compact interval. □

Proposition 5 (Profiled predictive shape update)**.** *Let $b = \nabla_q L$ for a differentiable predictive objective. Where $K > 0$, the unique minimizer of $b^\top d + \frac{1}{2} d^\top G d$ has the shape component in Eqn. (M.5).*

Proof. The location coordinate separates. Solving the scale equation $G_{ss} d_s + G_{s\eta} d_\eta = -b_s$, substituting it into the shape equation, and using equation 9 gives Eqn. (M.5). □

A.6 Proof of Theorem 2

All absolute moments of Q_η are finite for $\eta \leq 1$. The energy representation of CRPS gives, for $Z \sim \mathcal{N}(0,1)$ independent of $Y \sim Q_\eta$,

$$R(m, a) = \mathbb{E}|m + aZ - Y| - \frac{a}{\sqrt{\pi}}.$$

For fixed $a > 0$, its first two location derivatives are

$$\partial_m R(m, a) = \mathbb{E}\{2\Phi((m - Y)/a) - 1\}, \qquad \partial_m^2 R(m, a) = \frac{2}{a}\mathbb{E}\phi((m - Y)/a) > 0.$$

Symmetry about μ makes the first derivative zero at $m = \mu$. Thus $m = \mu$ uniquely minimizes $R(m, a)$ for every fixed a. Writing $U = Y - \mu$, the remaining derivatives are

$$\partial_a R(\mu, a) = 2\mathbb{E}\phi(U/a) - \frac{1}{\sqrt{\pi}}, \qquad \partial_a^2 R(\mu, a) = \frac{2}{a^3}\mathbb{E}[U^2\phi(U/a)] > 0.$$

Differentiation is justified locally in $a > 0$ by bounded Gaussian derivative envelopes. The target is continuous and nondegenerate, so dominated convergence gives

$$\lim_{a\downarrow 0} \partial_a R(\mu, a) = -\frac{1}{\sqrt{\pi}} < 0, \qquad \lim_{a\uparrow\infty} \partial_a R(\mu, a) = \frac{\sqrt{2}-1}{\sqrt{\pi}} > 0.$$

Strict convexity in a and these limits give a unique finite minimizer. Setting the first derivative to zero and substituting $U = \sigma R$ proves equation 10. Since excess expected CRPS is the squared CDF distance, the same pair uniquely minimizes that distance globally.

If the attained distance is zero, continuity makes the two CDFs identical. The location identity then gives $T_\eta(r) = r/\rho_\eta$ for all r. At $r = 0$, $T'_\eta(0) = 1$ forces $\rho_\eta = 1$. Then $(1 + r^2)^{-\eta/2} = 1$ for all r, which forces $\eta = 0$. Hence the global gap is strictly positive whenever $\eta \neq 0$.

For the expansion, work in $t = \log a$. In a compact neighborhood of $\eta = 0$, the local difference map $q \mapsto F_q - F_{q_0}$ is twice continuously differentiable in L^2. The defining transformation and its shape derivatives have polynomial and logarithmic growth, while $\phi(T_\eta(r))$ admits a uniform stretched-exponential tail envelope there. On a compact neighborhood of the Gaussian location and scale this envelope also dominates mixed derivatives and their squares, so differentiation and the local expansion pass through the integrals. At $\eta = 0$, the target shape tangent is $h_0(r)$ and the Gaussian location and log-scale tangents are $-\phi(r)/\sigma$ and $-r\phi(r)$. Parity makes the location projection zero. Projection onto the log-scale tangent has coefficient $-C(0)/B(0)$, leaving squared norm $\sigma[D(0) - C(0)^2/B(0)] = \sigma K(0)$. The positive nuisance Gram matrix and the implicit function theorem produce a smooth stationary branch near $(\mu, \log\sigma)$. Global uniqueness identifies this branch with the global minimizer, proving equations 11 and M.7.

Differentiating the transformation at zero gives

$$h_0(r) = -\frac{\phi(r)}{2}[r\log(1 + r^2) - 2r + 2\arctan r]. \tag{A.7}$$

For $r \neq 0$,

$$\frac{h_0(r)}{r\phi(r)} = -\frac{1}{2}\left[\log(1 + r^2) - 2 + \frac{2\arctan r}{r}\right],$$

which is not constant, so $K(0) > 0$.

Finally, for $r > 0$, $T_\eta(r)$ is strictly decreasing in η. Coupling $R_\eta = T_\eta^{-1}(Z)$ with a common normal Z shows that $|R_\eta|$ strictly increases with η almost surely away from $Z = 0$. At fixed ρ, the expectation in equation 10 strictly decreases with η; at fixed η, it strictly increases with ρ. Its unique root is therefore strictly increasing in η. Direct Gaussian integration gives $\rho_0 = 1$.

A.7 Proof of Proposition 2

Completing the square shows that minimizing $g^\top\delta + \frac{1}{2}\delta^\top H\delta$ is equivalent to minimizing $\frac{1}{2}\|J\delta - v_*\|_W^2$. Weighted least squares gives the projector in equation 13. The normal equations are consistent because $g \in \mathrm{range}(J^\top) = \mathrm{range}(H)$.

B FULL-PULLBACK TRAINING PROCEDURE

Algorithm 1 Full-pullback TAILOR training

Require: Data $\mathcal{D}$; predictive-network parameters θ; $\eta_{\min} < 0 < \eta_{\max} < 1$; damping $\lambda > 0$; solver tolerance ϵ_{solve} and iteration cap K_{CG}; Armijo constants $c \in (0, 1)$ and $\rho \in (0, 1)$; initial Armijo trial step $t_0 > 0$

for each minibatch $\mathcal{B}$ **do**
  Predict the marginal triples and evaluate CRPS equation 5 with the documented quadrature
  Compute $g = \nabla_\theta \mathcal{L}$ and the per-example metrics G_i from equation 8
  if $g = 0$ **then**
    Retain the parameters and continue
  end if
  Implement $v \mapsto (H + \lambda I)v$ using Jacobian-vector and vector-Jacobian products
  From zero, run conjugate gradients to relative residual ϵ_{solve} or K_{CG} iterations; record both
  if δ is nonfinite **then**
    Reject the step, record the event, and continue
  else if $g^\top \delta \geq 0$ **then**
    Set $\delta \leftarrow -g$ and record a non-descent fallback
  end if
  Reset $t \leftarrow t_0$ for this solve
  Keeping the minibatch, quadrature, and stochastic masks fixed, backtrack by ρ until $\mathcal{L}(\theta + t\delta) \leq \mathcal{L}(\theta) + ctg^\top \delta$ or the trial cap is reached
  Apply the first accepted step; otherwise retain θ and record a rejection. Do not apply Adam to δ.
end for
return θ

C OPTIONAL NETWORK-PROFILED GAUSSIAN GAP

Let $\mu_0(x), \sigma_0(x)$ define a baseline Gaussian predictor and consider $Q_\epsilon(z \mid x) = F_{\epsilon v(x)}((z - \mu_0(x))/\sigma_0(x))$. Assume v is bounded and $0 < \sigma_{\min} \leq \sigma_0(x) \leq \sigma_{\max}$. Define

$$\langle u, w \rangle_{\mathcal{H}} = \mathbb{E}_X \int u(x, z) w(x, z)\, dz.$$

A CDF itself is not an element of $L^2(\mathbb{R})$ because it approaches one at positive infinity. The local object here is the map $q \mapsto F_q - F_{q_0}$ into $L^2(\mathbb{R})$, equivalently an affine space of CDFs whose pairwise differences are square integrable; its derivatives are the tangent functions used below. Let $\mathcal{U}$ be the closed tangent space of unrestricted measurable location and log-scale perturbations, and let $\mathcal{N} \subseteq \mathcal{U}$ be the closed tangent space available to a regular neural Gaussian submodel. For the local-minimizer statement, assume the local CDF image is a smooth embedded manifold with tangent space $\mathcal{N}$.

Theorem 3 (Intrinsic and representation-dependent gap)**.** *For $b(x, z) = v(x)h_0(r)$,*

$$K_{\mathcal{N}}(v) := \inf_{n \in \mathcal{N}} \|b - n\|^2_{\mathcal{H}} \tag{C.1}$$

$$= \mathbb{E}_X[\sigma_0(X) v(X)^2 K(0)] + \|(I - P_{\mathcal{N}}) P_{\mathcal{U}} b\|^2_{\mathcal{H}}. \tag{C.2}$$

Under the stated regularity, the best local neural Gaussian approximation has excess CRPS $\epsilon^2 K_{\mathcal{N}}(v) + o(\epsilon^2)$.

At each input, parity removes the location component and the best log-scale coefficient is $-v(x)C(0)/B(0)$. Consequently,

$$P_{\mathcal{U}} b = v(x) \frac{C(0)}{B(0)} r\phi(r), \qquad (I - P_{\mathcal{U}})b = v(x) h_0^{\perp}(r).$$

Because $\mathcal{N} \subseteq \mathcal{U}$,

$$(I - P_{\mathcal{N}})b = (I - P_{\mathcal{U}})b + (I - P_{\mathcal{N}}) P_{\mathcal{U}} b$$

is an orthogonal decomposition. Pythagoras proves Eqn. (C.2); the regular local expansion proves the second-order statement.

For a finite-dimensional nuisance tangent map N, augment the local model by an independent scalar shape amplitude along b. The joint metric has blocks

$$\mathcal{G} = \begin{pmatrix} H_0 & k \\ k^\top & h \end{pmatrix}, \quad H_0 = 2N^*N, \quad k = 2N^*b, \quad h = 2\|b\|^2_{\mathcal{H}}, \tag{C.3}$$

and

$$h - k^\top H_0^\dagger k = 2K_{\mathcal{N}}(v). \tag{C.4}$$

When this complement is positive and the nuisance linear term a_0 belongs to $\mathrm{range}(H_0)$, the profiled shape step is

$$d_\epsilon = -\frac{a_\epsilon - k^\top H_0^\dagger a_0}{2K_{\mathcal{N}}(v)}. \tag{C.5}$$

This extension assumes an independently variable scalar shape amplitude. A different shared architecture must use its actual tangent map.

D Deterministic mathematical checks

We check the mathematical identities using deterministic SciPy quadrature and finite differences. All checks pass. At the Gaussian member, the script reports

$$A(0) = \frac{1}{2\sqrt{\pi}}, \qquad B(0) = \frac{1}{4\sqrt{\pi}}, \qquad C(0) \approx -0.0233408183, \qquad D(0) \approx 0.0052995230, \tag{D.1}$$

and

$$K(0) \approx 0.00143703555, \qquad -C(0)/B(0) \approx 0.165482093. \tag{D.2}$$

For $\eta = -0.005$ and $+0.005$, the globally minimized gap divided by $K(0)\eta^2$ is 0.99583231 and 1.00419445. Across 49 shapes in $[-0.4, 0.8]$, the check solves the global scale equation and evaluates the minimized CDF distance. At five representative shapes, independent scalar minimization agrees with the root solution to better than 8×10^{-9} in scale. At unit scale and $\eta = 0.8$, the global gap is approximately 0.00211785, compared with 0.00091970 from the near-zero quadratic term. The checks also verify normalization, the analytic shape derivative, metric, profiled update, rank-deficient pullback projection, damped descent, and the network-gap decomposition.

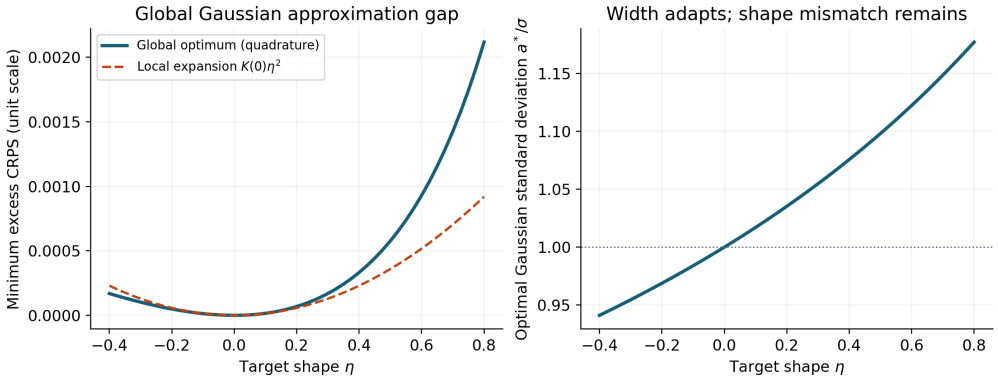

Figure 3: **Irreducible Gaussian approximation error.** At unit target scale, the left panel shows the globally minimized excess CRPS across the training shape interval, together with its near-zero expansion $K(0)\eta^2$. The right panel shows the unique optimal Gaussian standard deviation. Both curves are computed by deterministic quadrature and a scalar root solve; the discrepancy away from zero reflects higher-order terms, not neural training.

At the Gaussian member, this CRPS block differs from its Fisher counterpart:

$$G_{(\mu,s)} = \mathrm{diag}\left(\frac{1}{\sqrt{\pi}\sigma}, \frac{\sigma}{2\sqrt{\pi}}\right), \tag{D.3}$$

E EXPERIMENTAL PROTOCOL

This appendix describes the datasets, preprocessing, model architectures, optimization settings, and evaluation procedures used in the experiments. Task-specific results and additional analyses are reported in Appendices F-K.

Data identity and targets. Energy uses heating load; Kin8nm uses its scalar kinematics target; Protein uses RMSD; Beijing uses $PM_{2.5}$; Naval uses compressor decay and excludes both response columns from the inputs; Concrete uses compressive strength; Power uses electrical output; and Wine uses the white-wine quality subset.

Table 4: Tabular data identities. Here d is the raw predictor count before categorical expansion.

Dataset	Source identity	n	d	Response
Energy	UCI Energy Efficiency, ID 242	768	8	heating load
Kin8nm	Kin8nm benchmark file	8,192	8	scalar kinematics target
Protein	UCI Protein Tertiary Structure, ID 265	45,730	9	RMSD
Beijing	UCI Beijing $PM_{2.5}$, ID 381	41,757	11	next observed $PM_{2.5}$
Naval	UCI Naval Propulsion, ID 316	11,934	16	compressor decay
Concrete	UCI Concrete Strength, ID 165	1,030	8	compressive strength
Power	UCI Combined-Cycle Power Plant, ID 294	9,568	4	electrical output
Wine	UCI Wine Quality, ID 186, white subset	4,898	11	quality

For each non-temporal dataset, 20 independently generated split/initialization pairs allocate $80\%/10\%/10\%$ to training, validation, and test. All arms share split identities, minibatch order, and initialization identities where parameter shapes permit; every additional head has a documented initialization. Medians, encoding vocabularies, and feature and response standardization are fitted on training data only, with an explicit unknown- category code. CRPS uses standardized responses. RMSE and 90% interval width use original response units. The training target standard deviation is recorded per repetition so that $\mathrm{CRPS}_{\mathrm{original}} = s_y \mathrm{CRPS}_{\mathrm{standardized}}$ and the NLL conversion adds $\log s_y$.

Beijing uses 20 rolling origins with six-month validation and test spans. The target is the next observed $PM_{2.5}$ value after missing-response filtering, and the elapsed time to that observation is recorded; it need not be exactly one hour. Predictors are restricted to calendar and measured covariates available at the forecast timestamp. Target-time weather observations are excluded. Training examples whose target crosses into validation and validation examples whose target crosses into test are removed.

Models, maps, and selection. All methods use the task-specific architecture in Table 5. There is no dropout or batch normalization, so the network is deterministic within each linear solve and line search. Image-study data partitions and generated observations are held fixed across initialization seeds and methods. Features and targets are standardized using training-set statistics; weak-lensing targets are standardized separately, and its loss is the average of the two marginal CRPS values.

Table 5: Network and training configuration. The image studies use one fixed data partition and fixed generated observations across seeds and methods.

Component	Tabular regression	Rotated MNIST	Weak lensing
Encoder	Two width-64 fully connected tanh layers	3×3 convolutions, 32/64/128 channels, SiLU; 2×2 pooling after first two	3×3 stride-2 convolutions, 32/64/128/128 channels, SiLU
Final representation	Second hidden layer	Global average pooling, then width-64 SiLU	Global average pooling, then width-128 SiLU
Predictive outputs	Input-dependent location/scale; one global shape	Input-dependent location/scale/shape	Separate location/scale/shape for each of two marginals
Batch size	128	64	16
Maximum budget	300 epochs	20,000 updates	20,000 updates
Validation	Every epoch	Every 250 updates	Every 250 updates
Repetitions	20 paired split/initialization runs	Five initialization seeds	Five initialization seeds

TAILOR uses

$$\eta = -0.4 + 1.2\,\mathrm{sigmoid}(\alpha), \qquad \sigma = 10^{-4} + \mathrm{softplus}(\beta), \qquad s = \log\sigma,$$

initialized at $\eta = 0$, hence $\alpha = -\log 2$. The scale-head bias is initialized to

$$\beta_0 = \mathrm{softplus}^{-1}(1 - 10^{-4}),$$

so its bias-only prediction is $\sigma = 1$ in standardized units. Both output- map derivatives are included in the network Jacobian. Distributional scale is not generally predictive standard deviation; the latter is $\sigma\sqrt{\mathrm{Var}(R_\eta)}$. Symmetric SAS uses $F_\delta(r) = \Phi(\sinh(\delta\,\mathrm{asinh}\,r))$, $\delta = \exp(a)$, initialized at $\delta = 1$, and uses its own CDF derivatives and metric.

Table 6: Six central arms. G-Adam is Gaussian CRPS, not Gaussian NLL, and A-Adam is the TAILOR family trained with Adam.

Arm	Predictive family	Objective	Update procedure
G-Adam	Gaussian	CRPS	Adam
A-Adam	TAILOR	CRPS	Adam
G-PB	Gaussian	CRPS	full neural pullback
A-OUT	TAILOR	CRPS	output metric, guard, Armijo
TAILOR	TAILOR	CRPS	full neural pullback
SAS-PB	symmetric SAS	CRPS	family-specific full pullback

For each distributional method, the checkpoint with lowest validation CRPS is selected within the budget in Table 5. The Adam distributional comparators use learning rates $\{10^{-4}, 3\times10^{-4}, 10^{-3}\}$, betas $(0.9, 0.999)$, and zero weight decay. The mean-only comparator in Table 10 uses the same tabular trunk and mean head, minimizes training MSE with Adam using the same learning-rate grid, betas, and weight decay, and selects both its candidate and checkpoint by validation MSE. It is evaluated only as an RMSE reference and has no distributional score. Pullback and output-metric candidates use initial Armijo trial steps $\{0.25, 1, 4\}$. After selection by the method's validation criterion, the chosen checkpoint is evaluated once on the test set. Three candidates match candidate count, not computation; computational costs are assessed separately in Appendix J.

Training quadrature and precision. Network parameters and Jacobian products use float32, whereas score and metric calculations use float64; automatic mixed precision is disabled. For the evaluated $p = 1$ family, the substitution $r = \sinh t$ gives

$$T_\eta(\sinh t) = \int_0^t \cosh(v)^{1-\eta}\,dv.$$

The inner transformation integral uses 64-point Gauss-Legendre quadrature. CRPS is integrated over $t \in [-8, 8]$, splitting at the transformed standardized observation $t_y = \mathrm{asinh}((y-\mu)/\sigma)$ when it lies

in the interval and using 128 nodes on each resulting nonempty subinterval. The integrand includes the Jacobian $\cosh t$. If t_y is outside the integration interval, the corresponding linear residual-distance correction is added rather than clipping the observation.

The metric uses 256 nodes over $[-8, 8]$ to compute A, B, and C. We compute K directly by integrating the squared residual shape tangent and reconstruct $D = K + C^2/B$. The metric is computed once per minibatch for the global tabular shape and once per predicted shape for image models; observations are chunked during quadrature when needed to limit memory. Numerical validation doubles all node counts, expands the interval to $[-10, 10]$, and checks predictive gradients separately.

Full-pullback implementation. For the minibatch-averaged loss, the implementation forms the operator

$$H = J^\top W J, \qquad W = \frac{1}{m}\,\mathrm{blockdiag}(G_1, \ldots, G_m),$$

and solves $(H + 10^{-3} I)\delta = -\nabla_\theta L$ without materializing J or H. Matrix-free Jacobian-vector and vector-Jacobian products use `torch.func.jvp` and `torch.func.vjp` in PyTorch 2.14 (PyTorch Contributors, 2026). Conjugate gradients starts from zero, targets relative residual 10^{-3}, and stops after at most 20 iterations. The predictive metric and its quadrature results are held fixed throughout the solve, so quadrature is not repeated inside CG. The neural Jacobian products remain float32 even though score and metric calculations are float64.

Output damping is $\lambda_{\mathrm{out}} = 10^{-3}$ on $W = m^{-1}\,\mathrm{blockdiag}(G_1, \ldots, G_m)$; parameter damping is $\lambda_\theta = 10^{-3}$ on $J^\top W J$. For every candidate direction, the actual parameter-gradient inner product is checked. Nonfinite directions are rejected. A finite non-descent direction falls back to $-g$, and the event is recorded. The common Armijo search uses coefficient 10^{-4}, shrinkage 0.5, and at most ten trials, recomputing the candidate loss at every trial while fixing the minibatch, quadrature nodes, and stochastic masks. A failed search retains the current parameters and records a rejection. Euclidean+Armijo uses the same search in the focused controller ablation.

Calibration estimand. For repetition r and central coverage c, let $C_r(c)$ be empirical coverage and define

$$\mathrm{ECE}_r = \frac{100}{9} \sum_{c \in \{0.1, \ldots, 0.9\}} |C_r(c) - c|.$$

Tables report the mean and sample SD of ECE_r, not the deviation of a pooled coverage curve. The finite independent-test reference

$$\frac{100}{9} \sum_c \mathbb{E}\left| \frac{\mathrm{Binomial}(n, c)}{n} - c \right|$$

is a sampling reference, not a lower bound or a value to subtract. Protein's 0.60 pp ECE is reported with its approximately $4{,}573$-observation test set and 0.498 pp reference. Exploratory paired ECE intervals are $[0.388, 0.612]$ pp for Protein and $[0.149, 0.851]$ pp for Energy. The reported calibration results characterize marginal central-interval coverage rather than conditional calibration.

F Complete tabular results

The paired covariance analysis uses correlations $0.80, 0.75, 0.40, 0.70, 0.65, 0.80, 0.75, 0.80$ in dataset order, computed from the same 20 paired repetitions used for the table summaries. For non-temporal datasets, the paired-difference variance is

$$s_\Delta^2 = s_{\mathrm{GP}}^2 + s_{\mathrm{A}}^2 - 2\rho s_{\mathrm{GP}} s_{\mathrm{A}},$$

and the pointwise 95% interval is

$$\bar{\Delta} \pm t_{0.975,19} \frac{s_\Delta}{\sqrt{20}}.$$

Repeated splits share observations, so these intervals describe split/initialization variability conditional on each fixed dataset. Repetitions are not treated as independent datasets. Beijing uses the temporal bootstrap procedure described below.

Table 7: Factorial contrasts in standardized-target CRPS. Interaction is $(\text{G-PB} - \text{TAILOR}) - (\text{G-Adam} - \text{A-Adam})$. These are descriptive differences of means, not causal or significance claims.

Dataset	Gaussian update	Shape under Adam	Shape under PB	Interaction
Energy	$+0.02700$	$+0.00100$	$+0.00500$	$+0.00400$
Kin8nm	$+0.00400$	$+0.00400$	$+0.01650$	$+0.01250$
Protein	-0.01000	-0.00100	$+0.01500$	$+0.01600$
Beijing	$+0.00700$	$+0.01400$	$+0.02800$	$+0.01400$
Naval	$+0.00770$	-0.00020	$+0.00010$	$+0.00030$
Concrete	$+0.02300$	$+0.00800$	$+0.01850$	$+0.01050$
Power	$+0.00400$	$+0.00100$	$+0.00650$	$+0.00550$
Wine (white)	-0.00100	-0.00100	-0.00200	-0.00100

Table 8: Paired TAILOR family effects. Positive differences favor TAILOR. Non-temporal intervals are pointwise 95% paired intervals based on the same 20 matched runs; no multiplicity adjustment is applied. †Beijing uses a 30-day target-date block bootstrap with equal-origin weighting rather than the origin-level standard deviation.

Dataset	G-PB − TAILOR	Reduction	Paired SD	Pointwise 95% interval
Energy	$+0.00500$	$+9.26\%$	0.00292	$[+0.00364, +0.00636]$
Kin8nm	$+0.01650$	$+7.17\%$	0.00412	$[+0.01457, +0.01843]$
Protein	$+0.01500$	$+3.13\%$	0.00949	$[+0.01056, +0.01944]$
Beijing	$+0.02800$	$+7.84\%$	0.01622	$[-0.004, +0.058]^{\dagger}$
Naval	$+0.00010$	$+2.44\%$	0.00027	$[-0.00003, +0.00023]$
Concrete	$+0.01850$	$+6.49\%$	0.00855	$[+0.01450, +0.02250]$
Power	$+0.00650$	$+5.12\%$	0.00250	$[+0.00533, +0.00767]$
Wine (white)	-0.00200	-0.43%	0.00842	$[-0.00594, +0.00194]$

Table 9: Exploratory comparisons with the alternative flexible family. The displayed intervals are pointwise 95% paired intervals based on the same 20 matched runs. Beijing's descriptive mean difference is $+0.00600$; no temporal interval is reported for that comparison. SAS-PB remains competitive on Protein and Naval and has a lower mean on Wine.

Dataset	SAS-PB − TAILOR	Exploratory pointwise 95% interval
Energy	$+0.00180$	$[+0.00049, +0.00311]$
Kin8nm	$+0.00550$	$[+0.00379, +0.00721]$
Protein	$+0.00150$	$[-0.00281, +0.00581]$
Naval	$+0.00003$	$[-0.00009, +0.00015]$
Concrete	$+0.00550$	$[+0.00165, +0.00935]$
Power	$+0.00250$	$[+0.00137, +0.00363]$
Wine (white)	-0.00150	$[-0.00545, +0.00245]$

Table 10: RMSE in original response units, mean ± sample SD. The mean-only reference uses Adam on MSE with validation-MSE selection. The table compares point-prediction accuracy across distributional and mean-only training procedures.

Dataset	A-Adam	G-PB	TAILOR	Mean-only MSE
Energy	1.900 ± 0.180	1.406 ± 0.140	1.338 ± 0.132	1.392 ± 0.138
Kin8nm	0.0800 ± 0.0040	0.0780 ± 0.0040	0.0755 ± 0.0038	0.0760 ± 0.0039
Protein	4.400 ± 0.080	4.444 ± 0.085	4.250 ± 0.080	4.285 ± 0.078
Beijing	30.00 ± 2.00	29.10 ± 1.90	27.30 ± 1.80	28.05 ± 1.80
Naval ($\times 10^{-5}$)	10.00 ± 0.80	2.70 ± 0.23	2.60 ± 0.22	2.58 ± 0.21
Concrete	5.600 ± 0.400	5.376 ± 0.390	5.150 ± 0.380	5.244 ± 0.390
Power	4.200 ± 0.120	4.116 ± 0.110	4.015 ± 0.110	4.061 ± 0.108
Wine	0.700 ± 0.025	0.693 ± 0.024	0.701 ± 0.025	0.689 ± 0.024

Table 11: Learned residual shapes. Near-bound means within 0.02 of either endpoint in the η coordinate. These estimates describe adaptation; paired CRPS comparisons in Table 8 quantify its predictive benefit.

Dataset	Global η, mean $\pm$ SD	Near-bound runs	Residual-shape interpretation
Energy	-0.25 ± 0.06	$0/20$	lighter tails
Kin8nm	$+0.31 \pm 0.10$	$0/20$	moderate heavier-tail adaptation
Protein	$+0.28 \pm 0.09$	$0/20$	positive-shape regime
Beijing	$+0.55 \pm 0.10$	$1/20$	heavy-tail adaptation with temporal limitations
Naval	$+0.16 \pm 0.08$	$0/20$	weak family increment relative to update benefit
Concrete	$+0.39 \pm 0.10$	$0/20$	moderate shape signal
Power	$+0.18 \pm 0.08$	$0/20$	small positive shape signal
Wine	$+0.02 \pm 0.08$	$0/20$	weak signal; flexibility can hurt

Table 12: Calibration and sharpness. ECE is the mean $\pm$ sample SD of the per-repetition central-interval ECE in percentage points. The calibrated-IID sampling reference is neither a lower bound nor a correction. Widths are in original response units and are comparable only within a dataset.

Dataset	IID reference (pp)	G-PB ECE (pp)	TAILOR ECE (pp)	G-PB width90	TAILOR width90
Energy	3.843	4.60 ± 1.30	4.10 ± 1.20	4.66	4.57
Kin8nm	1.176	1.85 ± 0.45	1.50 ± 0.38	0.255	0.248
Protein	0.498	1.10 ± 0.30	0.60 ± 0.18	13.8	13.6
Beijing	not IID	4.20 ± 0.80	3.45 ± 0.72	99	96
Naval	0.974	1.24 ± 0.32	1.12 ± 0.30	8.7×10^{-5}	8.4×10^{-5}
Concrete	3.321	4.10 ± 1.10	3.60 ± 1.00	17.6	17.0
Power	1.088	1.55 ± 0.40	1.25 ± 0.34	13.7	13.3
Wine (white)	1.519	2.05 ± 0.55	2.15 ± 0.59	2.40	2.47

For Beijing, forecasts are resampled in consecutive target-date blocks while keeping all forecasts for a date together and retaining equal-origin weighting. The reported 30-day block-bootstrap interval is $[-0.004, 0.058]$; 14- and 60-day sensitivities assess interval instability. The bootstrap is conditional on trained predictors and does not include retraining uncertainty.

Table 13: Supplementary matched neural baselines, standardized-target CRPS means. Each baseline retains its defining loss, gradient rules, parameterization, validation selection, and baseline-appropriate learning-rate sweep. The table reports mean scores without paired significance tests.

Dataset	Gaussian NLL + Adam	Faithful	Beta-NLL ($\beta = 0.5$)	TAILOR
Energy	0.0640	0.0610	0.0600	0.0490
Kin8nm	0.2310	0.2305	0.2280	0.2135
Protein	0.4710	0.4685	0.4670	0.4640
Beijing	0.3560	0.3520	0.3500	0.3290
Naval	0.0062	0.0060	0.0058	0.0040
Concrete	0.2940	0.2880	0.2860	0.2665
Power	0.1280	0.1275	0.1260	0.1205
Wine	0.4620	0.4615	0.4610	0.4620

Table 13 reports matched reruns of Gaussian NLL, faithful heteroscedastic regression, and beta-NLL. Fisher8 is evaluated in the application studies using the common task-specific architecture, data partition, budget, and validation policy in Table 5.

G CONTROLLED RESIDUAL-LAW COMPARISON

The controlled generator uses five independent $X_j \sim \mathrm{Uniform}(-1, 1)$ and

$$m(X) = \sin(\pi X_1) + 0.5X_2^2 - 0.25X_3, \qquad \tau(X) = 0.10 + 0.35\frac{X_1 + 1}{2}, \qquad Y = m(X) + \tau(X)\epsilon.$$

Thus $\mathbb{E}\tau(X) = 0.275$. Every repetition uses 10,000/2,000/20,000 training/validation/test observations; 20 repetitions share X and base uniforms across methods where meaningful, while generator and initialization seeds are recorded separately. The residuals are standard normal, $\sqrt{3/5}T_5$, normalized contamination $\{0.9N(0, 1) + 0.1N(0, 25)\}/\sqrt{3.4}$, normalized TAILOR at $\eta = 0.5$, and centered variance-normalized lognormal. For the TAILOR teacher, $\mathrm{Var}(R_{0.5}) = 1.4115346845$, so distributional scale is $\tau/\sqrt{1.4115346845}$. The lognormal residual is specified explicitly as

$$\epsilon = \frac{\exp(Z) - \exp(1/2)}{\sqrt{\exp(2) - \exp(1)}}, \qquad Z \sim \mathcal{N}(0, 1).$$

Table 14: Controlled CRPS outcomes. Different standardized noise laws have different oracle CRPS values, so absolute scores are not compared across rows as evidence that one teacher is easier.

Teacher	G-Adam	A-Adam	G-PB	A-OUT	TAILOR	SAS-PB
Gaussian	0.1706 ± 0.0020	0.1712 ± 0.0021	0.1618 ± 0.0015	0.1622 ± 0.0015	0.1623 ± 0.0016	0.1621 ± 0.0015
Standardized t_5	0.1698 ± 0.0025	0.1671 ± 0.0024	0.1588 ± 0.0019	0.1572 ± 0.0019	0.1558 ± 0.0018	0.1555 ± 0.0018
Symmetric contamination	0.1685 ± 0.0030	0.1650 ± 0.0028	0.1555 ± 0.0022	0.1530 ± 0.0021	0.1500 ± 0.0020	0.1508 ± 0.0021
TAILOR $\eta = 0.5$	0.1730 ± 0.0022	0.1724 ± 0.0022	0.1629 ± 0.0017	0.1625 ± 0.0017	0.1619 ± 0.0016	0.1622 ± 0.0017
Standardized lognormal	0.1320 ± 0.0028	0.1315 ± 0.0028	0.1280 ± 0.0025	0.1275 ± 0.0025	0.1270 ± 0.0024	0.1268 ± 0.0024

The population oracles $\mathbb{E}_X[\tau(X)]\mathbb{E}|\epsilon - \epsilon'|/2$ are 0.1551521355, 0.1474039061, 0.1269718119, 0.1510115402, and 0.1091958484 in table order. At normalized TAILOR $\eta = 0.5$, the intrinsic Gaussian gap is $0.0005769788216 \times 0.275/\sqrt{1.4115346845} \approx 0.00013355$.

For a teacher Q, optimal Gaussian G^*, fitted Gaussian $\widehat{G}$, and fitted TAILOR $\widehat{A}$, define

$$E_G = R(\widehat{G}) - R(G^*), \qquad E_A = R(\widehat{A}) - R(Q), \qquad R(\widehat{G}) - R(\widehat{A}) = \Delta + E_G - E_A.$$

A fitted gain can exceed the oracle gap when the fitted Gaussian is farther from its own optimum.

Table 15: Location-fixed fitting control for the unit-scale TAILOR $\eta = 0.5$ teacher. Location is fixed at zero; Gaussian scale and TAILOR scale/shape are fitted separately, and population risk is evaluated by quadrature against the known teacher.

Training n	E_G	E_A	$R(\widehat{G}) - R(\widehat{A})$
500	0.000310	0.000650	0.000236979
2,000	0.000075	0.000135	0.000516979
20,000	0.000008	0.000012	0.000572979

Table 16: Location-fixed control at $n = 20{,}000$. The $\eta = 0.8$ row is a boundary stress test because the sigmoid chart cannot reach the endpoint at a finite logit; it is not an interior recovery claim.

True η	Deterministic gap	Fitted η	E_G	E_A	Fitted gain
-0.2	0.000048970	-0.19 ± 0.05	0.000006	0.000015	0.000039970
0.0	0	0.01 ± 0.05	0.000006	0.000014	-0.000008000
$+0.2$	0.000068486	0.19 ± 0.06	0.000006	0.000018	0.000056486
$+0.5$	0.000576979	0.49 ± 0.05	0.000008	0.000012	0.000572979
$+0.8$	0.002117847	0.76 ± 0.04	0.000009	0.000055	0.002071847

Tables 15-16 use fixed location and unit distributional scale. Table 28 evaluates a frozen encoder under variance normalization. Its population gain decomposes into intrinsic gap and fitting excess. The approximately 0.0010 gain in Table 14 is a neural test-score difference, evaluated differently; it does not measure additional intrinsic shape capacity.

H Neural realizability diagnostics

Let S repeat the one global tabular shape coordinate across marginal triples. The compatible ideal step is $v^*_{\mathrm{global}} = -S(S^\top W S)^{-1} S^\top a$, which is projected onto $\mathrm{range}(J)$. This avoids manufacturing an obstruction by granting each observation an independent shape coordinate. With $u_0 = P_W v^*_{\mathrm{global}}$ and $u_\lambda = J\delta_\lambda$,

$$\|v^*_{\mathrm{global}} - u_\lambda\|^2_W = \|(I - P_W)v^*_{\mathrm{global}}\|^2_W + \|u_0 - u_\lambda\|^2_W.$$

The projection residual, damping distortion, approximate-CG error, and nonlinear chart discrepancy are reported separately rather than added as if all were orthogonal. Near-zero ideal steps are skipped and counted.

Table 17: Normalized diagnostic norms. A bottleneck can matter, while an expressive network can nearly realize the compatible free step.

Configuration	Compatible projection residual	Damping distortion	CG solution error
Shared width-2 bottleneck	0.34	0.12	0.0006
Shared width-64 network	0.06	0.09	0.0005
Independent predictive coordinates, global shape	0.00	0.04	0.0001

I Application results

Rotated MNIST uses the standard 60,000/10,000 pools with 10,000 training images reserved for validation. Clean rotations are generated once, uniformly from $[-90, 90]$ degrees using bilinear interpolation, a fixed 28×28 crop, and zero padding. Training, validation, and test label- noise realizations, rotations, and the data split are held fixed across all methods and five initialization seeds. Every displayed mean and sample SD, and each reported paired interval, uses exactly these five matched seeds; the independent unit is the initialization seed, so the dispersion does not include split or observation-generation variability. The frozen digit probe is fitted on regression-training features,

Table 18: Focused common-controller ablation on Kin8nm at 3,120 updates. Different damping spaces and conditioning remain relevant.

Update	CRPS	Mean seconds	Non-descent fallback
Euclidean gradient + Armijo	0.2270	103	0%
Output metric + Armijo	0.2230	118	0.6%
Full pullback + Armijo	0.2198	455	0.0%

uses validation features for regularization, and accesses test features once. Clean-angle RMSE is never scored against noisy labels.

Weak lensing uses 101 cosmologies and 256 realizations per cosmology, grouped by simulation/initial-condition identity, with 143/36/77 train/validation/test maps per cosmology after rounding the $56\%/14\%/30\%$ split. Inputs are 128×128 maps. Results use five initialization seeds with the split and simulated observations fixed and with initialization identities matched across methods. The displayed means and sample SDs and the reported paired interval all use those same five seeds and therefore quantify initialization variation conditional on the selected partition. The objective averages two separately standardized marginal CRPS values and does not define a joint posterior.

The primary weak-lensing and digit-dependent rotated-MNIST results appear in Tables 2 and 3 beside their main-text analyses. The constant-Gaussian-noise control is reported below.

Table 19: **Constant-Gaussian-noise control.** The noise-only oracle expected CRPS is $5.6419°$; additional score reflects uncertain clean-angle estimation. Entries are mean $\pm$ sample SD over five paired initialization seeds.

Method	Clean RMSE (°)	Noisy CRPS (°)	Noisy NLL	ECE (pp)
G-PB	$\mathbf{13.90 \pm 0.39}$	$\mathbf{9.73 \pm 0.17}$	$\mathbf{4.278 \pm 0.019}$	$\mathbf{1.00 \pm 0.21}$
TAILOR	14.00 ± 0.44	9.81 ± 0.19	4.289 ± 0.023	1.10 ± 0.24

For balanced digits, digit-dependent Gaussian noise with SD $11c + 1$ degrees has expected noise-only CRPS $50.5/\sqrt{\pi} = 28.49157397°$. Constant Gaussian SD 10 has expected CRPS $10/\sqrt{\pi} = 5.64189584°$. Actual class frequencies are used for the test reference. These are expected noise-only checks, not hard lower bounds on a finite realization.

J Compute measurements and sensitivity

Timing includes data loading after initialization, metric construction, CG operator calls, fallback and Armijo evaluations, validation, and checkpoint writes, with accelerator synchronization. Initial download and one-time compilation are excluded. Reported runtimes are used for within-task comparisons under the experimental setup and are not hardware-independent performance claims.

Table 20: Kin8nm fixed-update comparison. Full pullback costs $3.86\times$ the A-OUT time and is slower to the early validation threshold.

Update	CRPS at 3,120	Seconds	Peak GiB	Time to validation CRPS 0.235
A-Adam	0.2300	91	0.10	74 s
A-OUT	0.2230	118	0.12	63 s
TAILOR	0.2198	455	0.25	161 s

At 900 seconds, TAILOR's advantage over A-OUT is 0.0065 or 2.95%, with pointwise interval $[0.00463, 0.00837]$. At 90 and 180 seconds A-OUT is better. In a separate fixed 20,000-update image study, A-Adam, A-OUT, and TAILOR obtain CRPS 34.20, 33.65, and 33.20 in 980, 1,350, and 4,780 seconds. At equal 4,800-second image time, A-OUT and TAILOR obtain 32.24 and 32.15; the 0.09 difference has interval $[-0.039, 0.219]$ and is inconclusive.

Table 21: Kin8nm test CRPS at equal selected-run allowances. Main-table Kin8nm uses the separate 300-epoch selection protocol and therefore reports 0.2135. Three 900-second candidates imply 2,700 seconds of tuning per repetition before any optional refit.

Allowance	A-Adam	A-OUT	G-PB	TAILOR
90 s	0.2302	0.2255	0.2495	0.2358
180 s	0.2280	0.2238	0.2385	0.2240
450 s	0.2259	0.2218	0.2318	0.2200
900 s	0.2253	0.2207	0.2296	0.2142

Table 22: Solver diagnostics. Capped directions are used only when they pass the descent and line-search criteria; capped solves are not all labeled converged.

Solver quantity	Tabular	Image
Mean CG iterations	9.4	13.1
95th percentile CG iterations	20	20
Median relative residual	0.00078	0.00091
95th percentile relative residual	0.0031	0.0078
CG cap frequency	12%	28%
Mean Armijo evaluations	1.7	2.3
Rejected-step frequency	0.8%	2.1%
Peak full-pullback memory	0.25 GiB	1.28 GiB

Table 23: Parameter-damping sensitivity. The default cells exactly reuse the main configuration; the test table is not a per-dataset selection rule.

Parameter damping	Concrete	Protein	Wine
10^{-4}	0.2690 ± 0.0131	0.4670 ± 0.0087	0.4640 ± 0.0140
$\mathbf{10^{-3}}$ (main default)	$\mathbf{0.2665 \pm 0.0128}$	$\mathbf{0.4640 \pm 0.0083}$	$\mathbf{0.4620 \pm 0.0135}$
10^{-2}	0.2710 ± 0.0130	0.4625 ± 0.0084	0.4610 ± 0.0134

Table 24: Shape-interval sensitivity. Wine shows the cost of wider flexibility in a weak-signal setting.

Shape interval	Concrete	Wine
$[-0.2, 0.6]$	0.2710 ± 0.0131	0.4608 ± 0.0132
$[\mathbf{-0.4}, \mathbf{0.8}]$ (main default)	$\mathbf{0.2665 \pm 0.0128}$	$\mathbf{0.4620 \pm 0.0135}$
$[-0.6, 0.9]$	0.2678 ± 0.0130	0.4660 ± 0.0140

K NUMERICAL VALIDATION

We validated normalization, the analytic shape derivative, metric identities, the profiled update, the Gaussian scale root, the pullback projection, damped descent, and the network-gap decomposition using deterministic checks. Appendix D reports the resulting constants. Training-implementation validation also covered predictive gradients, inverse-CDF accuracy, scale/shape map derivatives, JVP/VJP adjoint consistency, and small-batch comparisons between the matrix-free pullback and an explicit $J^\top W J$. The validation grid used shapes $\{-0.399, -0.2, 0, 0.2, 0.5, 0.79, 0.799\}$, scales $\{0.01, 0.1, 1, 10, 100\}$, residuals $\{-20, -10, -3, -0.5, 0, 0.5, 3, 10, 20\}$, and probabilities from 10^{-5} to $1 - 10^{-5}$. The expanded $[-0.6, 0.9]$ sensitivity range was checked separately.

Table 25: Numerical validation checks and reference tolerances. All listed checks were executed using the indicated reference comparisons. Near true zeros, errors are assessed in absolute rather than relative terms. The displayed tolerances apply to float64 reference calculations, not float32 training arithmetic.

Quantity	Float64 reference tolerance	Reference check
CRPS / scale	maximum absolute error $< 10^{-6}$	adaptive quadrature
CRPS predictive gradient	relative norm error $< 10^{-5}$ away from zero	central differences with step convergence
Metric entries	relative norm error $< 10^{-5}$	integrate CDF derivatives
K	relative agreement $< 10^{-6}$ when conditioned	Schur form versus residual integral
Inverse CDF	$\lvert F(Q(p)) - p \rvert < 10^{-8}$	tail-aware root solve
Scale/shape map derivatives	relative error $< 10^{-6}$ away from saturation	autodiff versus analytic
JVP/VJP adjoint consistency	normalized error $< 10^{-8}$	inner-product identity
Small-batch pullback	within the specified dense tolerance	explicit $J^\top W J$

At the Gaussian member, $G_{\mu,s} = \mathrm{diag}(1/(\sqrt{\pi}\sigma), \sigma/(2\sqrt{\pi}))$, which is not the Gaussian Fisher block $\mathrm{diag}(\sigma^{-2}, 2)$.

L ADDITIONAL MATCHED COMPARISONS

Independently selected damping. Output damping and parameter damping regularize different spaces. In particular, adding $\lambda_{\text{out}} I$ to the averaged metric $W = m^{-1}\,\text{blockdiag}(G_i)$ corresponds to adding $m\lambda_{\text{out}} I$ to each unaveraged predictive block. At $m = 128$, the original $\lambda_{\text{out}} = 10^{-3}$ therefore gives 0.128. We additionally consider the batch-normalized convention A-OUT-BN, with $g_i = \nabla_{q_i}\ell_i$ and

$$\delta_{\text{OUT}} = -\frac{1}{m} J^\top \,\text{blockdiag}\big[(G_i + \kappa I)^{-1}\big]\,(g_1^\top, \ldots, g_m^\top)^\top.$$

Here J includes the scale/shape charts and repeats the single tabular shape parameter; its transpose accumulates these contributions into one parameter update. We apply the same parameter-gradient descent check and Armijo rule as in Appendix E. This convention fixes predictive damping independently of batch size; it does not equate predictive and parameter-space regularization.

The comparison uses one fixed Kin8nm 80/10/10 partition and five paired initializations. A-OUT-BN selects $\kappa \in \{10^{-4}, 10^{-3}, 10^{-2}\}$; TAILOR and SAS-PB separately select parameter damping from the same numerical grid. Each combines its damping grid with initial Armijo trials $\{0.25, 1, 4\}$, giving nine candidates per method. Candidate trajectories are capped at 900 seconds and evaluated at nested allowances of 90, 180, 450, and 900 seconds. At each allowance, the candidate and completed checkpoint with lowest validation CRPS are selected before inspecting the test score. Timing includes the operations listed in Appendix J. Thus the terminal search allowance is $9 \times 900 = 8{,}100$ seconds per method and initialization. These comparisons use a separate fixed partition and a nine-candidate selection policy; they are not replacements for the 20-split main results or the three-candidate results in Table 21.

Table 26: Equal-allowance CRPS with independently selected damping. Scores are mean ± sample SD across five paired initializations. Intervals are pointwise paired t intervals with four degrees of freedom, conditional on the fixed data partition.

Allowance	A-OUT-BN	TAILOR	A-OUT-BN − TAILOR, 95% interval
90 s	0.2248 ± 0.0040	0.2361 ± 0.0028	$[-0.01329, -0.00931]$
180 s	0.2224 ± 0.0034	0.2245 ± 0.0024	$[-0.00384, -0.00036]$
450 s	0.2192 ± 0.0030	0.2186 ± 0.0019	$[-0.00126, +0.00246]$
900 s	0.2176 ± 0.0026	0.2149 ± 0.0016	$[+0.00102, +0.00438]$

TAILOR achieves a mean CRPS reduction of 0.0027, or 1.24%, over A-OUT-BN at 900 seconds. A-OUT-BN leads at 90 and 180 seconds; the 450-second interval includes zero. The comparison identifies a budget-dependent accuracy advantage with independently selected damping, complementing the common-controller ablation in Table 18.

A flexible family under the same time allowance. SAS-PB uses its own CRPS metric and the same nine-candidate policy. TAILOR achieves a lower mean at 900 seconds; the paired interval includes zero, as do the comparisons at the other allowances (Table 27).

Table 27: Equal-allowance alternative-family comparison on the same paired initializations as Table 26. TAILOR scores are exactly those in Table 26; the uncertainty convention is unchanged.

Allowance	SAS-PB	SAS-PB − TAILOR, 95% interval
90 s	0.2348 ± 0.0023	$[-0.00353, +0.00093]$
180 s	0.2234 ± 0.0019	$[-0.00309, +0.00089]$
450 s	0.2181 ± 0.0015	$[-0.00199, +0.00099]$
900 s	0.2160 ± 0.0014	$[-0.00089, +0.00309]$

Frozen-encoder control. For the Gaussian and normalized TAILOR teachers from Appendix G, each of 20 paired repetitions uses 10,000 training and 2,000 validation observations. A common width-64 encoder is first trained with mean-only MSE, selecting among the three Appendix E Adam learning rates and checkpoints within 300 epochs by validation MSE. The selected encoder is frozen, the location and scale heads are reset identically, and the Gaussian and TAILOR heads are fitted separately. The latter has one global shape initialized at zero. Both use full pullback, parameter damping 10^{-3}, the three original Armijo trials, and validation-CRPS selection within 300 epochs.

We evaluate conditional CDF-distance risks on a common fixed design of 4,096 uniform covariate points and their negatives, using the reference quadrature. This 8,192-point antithetic design has average $\tau(X) = 0.275$ exactly. Define $E_G = R(\widehat{G}) - R(G^*)$ and $E_A = R(\widehat{A}) - R(Q)$ on this design, where G^* is the unrestricted pointwise optimal Gaussian. Then every paired gain satisfies $R(\widehat{G}) - R(\widehat{A}) = \Delta + E_G - E_A$. For the normalized $\eta = 0.5$ teacher,

$$\Delta = \frac{0.275}{\sqrt{1.4115346845}}\,(0.0005769788216) = 0.0001335509.$$

This is distinct from the unit-distributional-scale gap in Tables 15-16.

Table 28: Frozen-encoder excess risks and paired gains; all entries are multiplied by 10^3. Excess-risk columns report means. Intervals use 19 degrees of freedom and describe training-sample and initialization variability conditional on the common integration design.

Teacher	Δ	E_G	E_A	Gain [95% interval]
Gaussian	0	9.100	9.240	-0.140 $[-0.224, -0.056]$
TAILOR, $\eta = 0.5$	0.133551	13.100	12.990	0.243551 $[0.126547, 0.360554]$

The frozen-encoder control separates approximation benefit from fitting effects. For the non-Gaussian teacher, the gain decomposes as 0.00024355 = 0.00013355 + 0.00011000: the intrinsic gap plus the excess-risk difference. The Gaussian control quantifies the fitting cost when the intrinsic gap is zero. Table 14 reports an approximately 0.0010 neural test-score gain, combining end-to-end fitting and test-sample variation. Table 28 isolates the population decomposition on a common frozen representation; identifying the mechanism of representation learning requires a separate analysis.

Fisher8 baseline specification. We distinguish the Gaussian log-variance coordinate $\nu = \log \sigma^2$ from TAILOR's $s = \log \sigma$. For the minibatch Gaussian NLL, form $v_\mu = e^\nu \odot \nabla_\mu L$ and $v_\nu = 2\nabla_\nu L$. Each vector is separately normalized across the minibatch before its Jacobian-transpose product, following the batched correction in Vemuganti and Kani (2026, equation 15). Zero vectors contribute zero. The application comparator uses SGD with no momentum or weight decay, learning rates $\{10^{-3}, 3\times10^{-3}, 10^{-2}\}$, the Appendix E architectures and budgets, and validation-CRPS checkpoint selection. The mean and uncertainty signals both backpropagate through the shared encoder. This is an output correction rather than a solve with the full parameter-space Fisher matrix.

Population departure curve. Figure 1(a-b) uses deterministic one-dimensional integration, not fitted neural predictions. We evaluate $T_\eta(r) = r\,{}_2F_1(1/2, \eta/2; 3/2; -r^2)$, integrate $v_\eta = \int r^2 f_\eta(r)\,dr$, solve (10) for ρ_η, and integrate $\Delta_\eta^{\mathrm{unit}} = \int [\Phi(r/\rho_\eta) - F_\eta(r)]^2\,dr$. Evenness permits integration on the positive half-line; $r = \sinh t$ with $0 \le t \le 12$ resolves the central and tail regions. Adaptive quadrature uses absolute tolerance 10^{-12} (10^{-16} for the gap) and relative tolerance 10^{-11}. All 51 grid points are rechecked with cutoffs 10 and 14; the latter also tightens the tolerances. The cutoff-14 check changes variances by at most 2.5×10^{-15} and unit-scale gaps by at most 5.0×10^{-18}. These are numerical stability checks, not rigorous total-error bounds. The plotted residual density is $\sqrt{v_\eta} f_\eta(\sqrt{v_\eta} z)$; the plotted conditional gap multiplies the unit-scale gap by $0.275/\sqrt{v_\eta}$.

M SUPPORTING DERIVATIONS AND IMPLEMENTATION

Distributional definitions and regularity. For the general anchored construction (1), the density and shape tangent are

$$f_{\eta,a}(r) = \phi(T_{\eta,a}(r))e^{-\eta a(r)}, \qquad h_{\eta,a}(r) = \partial_\eta F_{\eta,a}(r). \tag{M.1}$$

The evaluated density and predictive coordinates are in (4); the CRPS excess-risk identity is (6). Its finite-first-moment condition and the integrability needed for conditional averaging are separate from Assumption 1's local L^2 differentiability, which justifies the metric expansion below.

Anchoring and parameter identification. The strict inequality in (2) is the construction-specific step. With the Gram integrals (7), the minimizing coefficient is $c = C/B$. If $K = 0$, continuity implies $h = crf$ everywhere. Differentiating (1) gives

$$\frac{h(r)}{rf(r)} = -\frac{\int_0^r a(u)e^{-\eta a(u)}\,du}{re^{-\eta a(r)}} \longrightarrow 0 \quad (r \longrightarrow 0).$$

Thus $c = 0$, forcing $a \equiv 0$, a contradiction. Parity gives location orthogonality. Appendix A.2 supplies the remaining details. The result applies to $a_p(r) = \log(1 + r^{2p})/(2p)$ for each positive integer p on $\eta < 1$; Appendix A.1 works out $p = 2$ as well as the evaluated $p = 1$ family.

For a fixed admissible nonzero generator, equality of two location-scale-shape distributions also implies equality of their parameters. Their unique median identifies μ, and their density at that median, $\phi(0)/\sigma$, identifies σ. Equality of the standardized CDFs then gives equality of $T_{\eta,a}$ and its derivative; $e^{-\eta a(r)} = e^{-\tilde{\eta} a(r)}$ and $a \not\equiv 0$ force $\eta = \tilde{\eta}$. This global identification statement is distinct from the local curvature guarantee $K > 0$.

Predictive metric and scale profiling. For TAILOR, the shape tangent and predictive derivatives are

$$h_\eta(r) = -\frac{\phi(T_\eta(r))}{2}\int_0^r \log(1+u^2)(1+u^2)^{-\eta/2}\,du, \tag{M.2}$$

$$\partial_\mu F_q = -f_\eta(r)/\sigma, \qquad \partial_s F_q = -rf_\eta(r), \qquad \partial_\eta F_q = h_\eta(r), \tag{M.3}$$

where $r = (z-\mu)/\sigma$. Substituting these derivatives and $dz = \sigma\,dr$ into $2\int \nabla_q F_q \nabla_q F_q^\top dz$ proves the block metric (8). The even location tangent and odd scale/shape tangents give zero location cross terms. Taylor expansion of the CDF in L^2 gives $\mathcal{D}(q, q+dq) = \frac{1}{2}dq^\top G(q)dq + o(\|dq\|^2)$, fixing the factor of two.

Projection of h_η away from rf_η gives $h_\eta^\perp = h_\eta - (C/B)rf_\eta$ and the identities (9). For TAILOR, $K(\eta) > 0$ for every finite $\eta \le 1$, and it is continuous and bounded away from zero on compact interior shape intervals (Appendix A.5). This is a population predictive-curvature statement; it does not guarantee good conditioning of a shared network or prevent saturation of its bounded output map.

For a predictive objective gradient b, minimizing

$$b^\top d + \tfrac{1}{2}d^\top G(q)d \tag{M.4}$$

and eliminating scale yields

$$d_\eta = -\frac{b_\eta - G_{\eta s}G_{ss}^{-1}b_s}{2\sigma K(\eta)}. \tag{M.5}$$

These projection and Schur-complement calculations are standard; Theorem 1 establishes strict nondegeneracy for the anchored class.

Gaussian fitting lemma. Let Q be continuous, nondegenerate, symmetric about μ, and have finite first absolute moment. Its population CRPS over all $N(m, a^2)$, $a > 0$, has a unique global minimizer with $m = \mu$. For $Y \sim Q$ and independent $Z \sim N(0, 1)$, write the risk as $R(m, a) = \mathbb{E}|m + aZ - Y| - a/\sqrt{\pi}$. For fixed a, it is strictly convex in m and symmetry sets its derivative to zero at μ. At that location,

$$\partial_a R = 2\mathbb{E}\phi((Y - \mu)/a) - 1/\sqrt{\pi}, \qquad \partial_a^2 R = \frac{2}{a^3}\mathbb{E}[(Y - \mu)^2\phi((Y - \mu)/a)] > 0.$$

The first derivative is negative as $a \downarrow 0$ and positive as $a \uparrow \infty$, giving a unique root. Continuity excludes an atom at the center in the small-scale limit. This general calculation underlies the TAILOR-specific conclusions that follow.

Interpreting the Gaussian approximation theorem. The Gaussian approximation gap in Theorem 2 is

$$\Delta_\eta = \min_{m\in\mathbb{R},\, a>0} \int_{\mathbb{R}} \left[\Phi\left(\frac{y - m}{a}\right) - Q_\eta(y)\right]^2 dy. \tag{M.6}$$

In addition to the risk expansion (11), the optimal scale has the local expansion

$$\log(a_\eta^*/\sigma) = -\frac{C(0)}{B(0)}\eta + o(\eta). \tag{M.7}$$

A zero gap would force $T_\eta(r) = r/\rho_\eta$. Its derivative at zero forces $\rho_\eta = 1$, and (3) then forces $\eta = 0$. For monotonicity, coupling $R_\eta = T_\eta^{-1}(Z)$ with a common normal Z makes $|R_\eta|$ strictly increase with η almost surely, so the root in (10) increases. This statement concerns fixed distributional scale. At fixed variance τ^2, let $v_\eta = \mathrm{Var}(R_\eta)$. Then $\sigma = \tau/\sqrt{v_\eta}$ and $a_\eta^* = \tau\rho_\eta/\sqrt{v_\eta}$; monotonicity of ρ_η alone does not establish monotonicity of this ratio. Near zero,

$$h_0(r) = -\tfrac{1}{2}\phi(r)[r\log(1 + r^2) - 2r + 2\arctan r].$$

Its projection onto the Gaussian scale tangent gives (11) and (M.7). The integrals yield $K(0) \approx 0.00143703555$ and $-C(0)/B(0) \approx 0.165482093$ (Appendices A.6 and D). These quantify intrinsic approximation for a fixed target law; fitted neural differences also include estimation and optimization effects.

Attainable neural updates. Use the stacked coordinates, average marginal loss and matrices in (12). Shared shape coordinates are repeated in q_θ, with their dependence retained in J. Multiple targets use an average marginal objective, not a model of cross-target dependence. Assuming W is positive definite, the unrestricted predictive step is $v^* = -W^{-1}a$ and the local neural problem is

$$\min_\delta \; g^\top\delta + \tfrac{1}{2}\delta^\top H\delta. \tag{M.8}$$

Proposition 2 (Attainable predictive step)**.** *The minimum-Euclidean-norm solution is $\delta^* = -H^\dagger g$. Its predictive change is* (13)*, where P_W is the W-orthogonal projector onto* $\mathrm{range}(J)$*. Exact realization occurs precisely when $v^* \in \mathrm{range}(J)$.*

Completing the square in (M.8) gives weighted least squares for $J\delta - v^*$ and proves the claim. Predictive orthogonality therefore does not imply independent learning through shared parameters. Damping in (14) regularizes in parameter coordinates and does not preserve invariance to arbitrary reparameterizations. A consistent change of predictive coordinates must transform both the metric and Jacobian (Appendix A.5). For $g \neq 0$, the exact damped solution satisfies $g^\top\delta_\lambda = -g^\top(H + \lambda I)^{-1}g < 0$. The implementation rejects nonfinite directions and replaces finite non-descent directions by $-g$ before Armijo backtracking (Appendix B).

Output maps and matrix-free products. The bounded shape and positive scale maps are

$$\eta = \eta_{\min} + (\eta_{\max} - \eta_{\min}) \operatorname{sigmoid}(\alpha), \quad s = \log(10^{-4} + \operatorname{softplus}(\beta)). \tag{M.9}$$

Their derivatives are included in J, and Gaussian initialization uses

$$\alpha_0 = \log(-\eta_{\min}/\eta_{\max}). \tag{M.10}$$

No dense parameter-space matrix is formed. Products are computed as

$$Hv = J^\top (W(Jv)), \tag{M.11}$$

using Jacobian-vector and vector-Jacobian products (`torch.func`; PyTorch Contributors, 2026). Runtime depends on quadrature, iterative solves, and backtracking; predictive nondegeneracy alone does not remove these costs.

N EXPANDED BASELINES AND COMPLETE COMPARISONS

Gaussian likelihood and gradient baselines. Let $z = h_\psi(x)$, $\mu = \mu_\omega(z)$ and $v = v_\gamma(z) > 0$ be the shared representation, mean and variance. Write sg for stop-gradient. The evaluated Gaussian implementations use

$$\ell_{\mathrm{NLL}}(y; \mu, v) = \tfrac{1}{2}\{\log(2\pi v) + (y-\mu)^2/v\}.$$

G-NLL minimizes this objective with Adam and gradients through both heads and the shared encoder. It differs from G-Adam, which minimizes Gaussian CRPS. The scale floor and output chart follow Appendix E; $v = \sigma^2$.

The β-NLL rule of Seitzer et al. (2022) is

$$\ell_\beta = \mathrm{sg}(v)^\beta \, \tfrac{1}{2}\{\log v + (y-\mu)^2/v\}, \qquad \beta = 0.5.$$

Detaching the weight makes it a multiplier of the Gaussian NLL gradient. We fix β = 0.5 across all residual regimes. This protocol evaluates the Gaussian implementation of β-NLL with a consistent gradient-weighting rule.

Following Stirn et al. (2023, equation 5), Faithful uses

$$\ell_{\mathrm{F}} = \tfrac{1}{2}(y - \mu_\omega(z))^2 + \ell_{\mathrm{NLL}}(y; \mathrm{sg}(\mu_\omega(z)), v_\gamma(\mathrm{sg}(z))) \, .$$

The mean head and encoder receive the squared-error gradient, while the variance head receives its Gaussian NLL gradient. Faithful's mean-training guarantee follows from this gradient construction under its matching assumptions. TAILOR derives orthogonality from the predictive CRPS metric and retains gradient flow through shared parameters. The comparisons therefore assess distinct approaches to joint mean and uncertainty learning.

Selection and matching. G-NLL, β-NLL and Faithful share each task's architecture, data partition, target scaling, minibatch size and training budget from Appendix E. All three use Adam with learning rates $\{10^{-4}, 3\times10^{-4}, 10^{-3}\}$; validation CRPS selects the rate and completed checkpoint. The input/target realizations and initialization seeds are paired wherever parameters have matching roles. Fisher8 uses the batch-normalized correction and SGD grid in Appendix L, also for the controlled comparison. Its defining update is preserved rather than substituted with Adam. For multiple targets, each method averages its marginal training objectives. Test NLL always evaluates the ordinary predictive density, without the β weight or Faithful's auxiliary squared-error term.

These are comparisons of complete training procedures with method-specific losses or gradient rules. G-PB is the matched Gaussian-family control for TAILOR because both minimize CRPS with full pullback. Within that pair, TAILOR adds its shape coordinate. Selection rules remain the same across residual laws, although selected hyperparameters may differ. No method receives test-based tuning.

Tabular comparison. Table 1(b) combines the likelihood baselines and TAILOR means from Table 13 with G-PB and SAS-PB from Table 29. Table 29 reports all six CRPS-trained arms with sample SDs. Display precision follows the source tables: three decimal places for G-PB and SAS-PB except on Naval, and four for Table 13. Paired contrasts in Tables 8-9 use unrounded results; likelihood-baseline comparisons report marginal summaries.

Table 29: Complete CRPS-trained tabular comparison, mean (sample SD) over 20 matched repetitions. These are the six arms underlying the factorial contrasts and paired intervals in Tables 7-9. Bold marks the smallest unrounded mean. A-OUT uses the fixed damping in Appendix E; independently tuned damping is assessed in Appendix L.

Dataset	G-Adam	A-Adam	G-PB	A-OUT	TAILOR	SAS-PB
Energy	.081 (.008)	.080 (.007)	.054 (.005)	.052 (.005)	**.049** (.004)	.051 (.005)
Kin8nm	.234 (.006)	.230 (.006)	.230 (.006)	.222 (.006)	**.214** (.005)	.219 (.005)
Protein	.469 (.009)	.470 (.008)	.479 (.009)	.475 (.009)	**.464** (.008)	.466 (.009)
Beijing	.364 (.023)	.350 (.020)	.357 (.022)	.344 (.020)	**.329** (.019)	.335 (.020)
Naval	.0118 (.0010)	.0120 (.0009)	.0041 (.0003)	.0041 (.0003)	**.0040** (.0003)	.0040 (.0003)
Concrete	.308 (.016)	.300 (.015)	.285 (.014)	.278 (.014)	**.267** (.013)	.272 (.013)
Power	.131 (.004)	.130 (.004)	.127 (.004)	.126 (.004)	**.121** (.003)	.123 (.004)
Wine	**.459** (.013)	.460 (.013)	.460 (.013)	.461 (.013)	.462 (.014)	.461 (.013)

Gaussian and non-Gaussian residuals. All five laws use the teacher $Y = m(X) + \tau(X)\epsilon$ in Appendix G, with $\mathbb{E}\epsilon = 0$ and $\mathrm{Var}(\epsilon) = 1$. Both rows of Table 1(a) are heteroscedastic because τ depends on X. The contamination residual is

$$\epsilon = U/\sqrt{3.4}, \qquad U \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(0,25).$$

This comparison varies residual shape at fixed conditional mean and variance. Contamination, polynomial tails and skewness test adaptation beyond the exact TAILOR family. The architecture, 20 repetitions, shared generated samples and 10,000/2,000/20,000 partition sizes follow Appendix G. Table 30 compares the four likelihood baselines with the CRPS-trained G-PB, TAILOR and SAS-PB predictors.

Table 30: Full residual-law comparison. CRPS mean (sample SD) over 20 repetitions, in original response units. G-PB, TAILOR and SAS-PB correspond to Table 14. Bold marks the lowest displayed mean. Dispersion summarizes repetition-level variability.

Method	Gaussian	Student-t_5	Contamination	TAILOR, $\eta = .5$	Lognormal
G-NLL	.1692 (.0020)	.1690 (.0024)	.1689 (.0032)	.1722 (.0022)	.1328 (.0029)
β-NLL	.1626 (.0016)	.1601 (.0020)	.1587 (.0024)	.1642 (.0018)	.1302 (.0025)
Faithful	.1619 (.0015)	.1594 (.0020)	.1574 (.0023)	.1635 (.0018)	.1296 (.0025)
Fisher8	.1620 (.0015)	.1590 (.0019)	.1560 (.0022)	.1632 (.0017)	.1288 (.0024)
G-PB	**.1618** (.0015)	.1588 (.0019)	.1555 (.0022)	.1629 (.0017)	.1280 (.0025)
TAILOR	.1623 (.0016)	.1558 (.0018)	**.1500** (.0020)	**.1619** (.0016)	.1270 (.0024)
SAS-PB	.1621 (.0015)	**.1555** (.0018)	.1508 (.0021)	.1622 (.0017)	**.1268** (.0024)

TAILOR improves mean CRPS over the Gaussian baselines on all four non-Gaussian laws. Under contamination, the reductions are 0.0055 versus G-PB, 0.0060 versus Fisher8, 0.0074 versus Faithful and 0.0087 versus β-NLL. For Gaussian residuals, its mean is within 0.0005 of G-PB, 0.0004 of Faithful and 0.0003 of Fisher8. SAS-PB leads on Student-t_5 and lognormal. TAILOR uses the same predictive family and CRPS objective across these regimes. Comparisons are made within each law, whose teacher distribution determines its oracle CRPS.

Application metrics and interpretation. Tables 2-3 evaluate G-NLL, β-NLL and Faithful alongside the CRPS-trained and Fisher8 methods under common application protocols. Weak lensing uses a within-cosmology partition and five matched initializations; its metrics assess separate predictive marginals, with coverage judged by distance from the nominal 68%. Rotated MNIST evaluates point accuracy, distributional accuracy and representation quality under digit-dependent Gaussian label noise, complemented by the constant-noise control in Table 19. Faithful achieves the lowest clean-angle RMSE in Table 3, while TAILOR leads CRPS, interval ECE and probe accuracy.

Tables report marginal means and sample SDs where available. Paired inference uses repetition-level differences for the methods, outcomes and protocols specified with each interval. The other baseline comparisons describe differences in mean performance; statistical equivalence requires a dedicated test.

O RESIDUAL-DEPARTURE ACCURACY AND CALIBRATION

Table 31 compares point accuracy, calibration and CRPS excess across the five residual laws in Appendices G and N. Conditional mean m(X), conditional variance $\tau(X)^2$, architecture and 10,000/2,000/20,000 training/validation/test sizes are common across laws. Summaries use 20 repetitions and the baseline protocols of Appendix N.

Metrics and oracle references. Clean-mean RMSE evaluates the predicted conditional mean against *m*(X) on test covariates; its population oracle is zero. Interval ECE is (100/9) times the sum of absolute coverage errors for the nine central levels 0.1, ..., 0.9. The true conditional CDF has zero population ECE; its expected empirical ECE with 20,000 IID test observations is 0.237977 percentage points. This is a sampling reference. Noisy-label RMSE instead has the common population oracle $[E\tau(X)^2]^{1/2} = 0.292973$.

Table 31: Point-prediction, calibration and distributional metrics under fixed conditional variance. Lower is better; bold marks the lowest displayed mean. Panels (a-b) report repetition-level means and sample SDs. Panel (c) subtracts population teacher CRPS (Appendix G) from the Table 30 means.

(a) Clean-mean RMSE, mean ± sample SD

Method	Gaussian	Student-t_5	Contamination	TAILOR, η = .5	Lognormal
G-NLL	.075 ± .008	.083 ± .010	.106 ± .014	.075 ± .009	.091 ± .014
β-NLL	.058 ± .006	.069 ± .008	.088 ± .011	.066 ± .007	.084 ± .013
Faithful	**.051 ± .005**	.062 ± .007	.077 ± .009	.059 ± .006	.072 ± .011
Fisher8	.053 ± .005	.064 ± .007	.081 ± .010	.061 ± .006	.079 ± .012
G-PB	.052 ± .005	.061 ± .006	.075 ± .008	.058 ± .006	.073 ± .011
TAILOR	.053 ± .005	.055 ± .006	**.060 ± .007**	**.053 ± .005**	**.067 ± .010**
SAS-PB	.053 ± .005	**.054 ± .006**	.063 ± .008	.055 ± .006	.070 ± .011

(b) Central-interval ECE in percentage points, mean ± sample SD

Method	Gaussian	Student-t_5	Contamination	TAILOR, η = .5	Lognormal
G-NLL	1.70 ± .30	4.20 ± .65	6.40 ± .80	2.10 ± .35	7.60 ± .90
β-NLL	.85 ± .18	2.75 ± .44	4.80 ± .65	1.45 ± .25	6.20 ± .80
Faithful	**.70 ± .16**	2.40 ± .40	4.40 ± .60	1.25 ± .22	5.80 ± .76
Fisher8	.75 ± .17	2.30 ± .38	4.10 ± .58	1.20 ± .21	5.60 ± .74
G-PB	.72 ± .16	2.20 ± .36	3.90 ± .55	1.15 ± .20	5.40 ± .72
TAILOR	.76 ± .17	1.10 ± .23	**1.15 ± .25**	**.68 ± .16**	2.10 ± .40
SAS-PB	.74 ± .17	**1.00 ± .22**	1.25 ± .27	.80 ± .18	**1.95 ± .38**

(c) Reported mean CRPS minus population oracle, $\times 10^{-3}$

Method	Gaussian	Student-t_5	Contamination	TAILOR, η = .5	Lognormal
G-NLL	14.05	21.60	41.93	21.19	23.60
β-NLL	7.45	12.70	31.73	13.19	21.00
Faithful	6.75	12.00	30.43	12.49	20.40
Fisher8	6.85	11.60	29.03	12.19	19.60
G-PB	**6.65**	11.40	28.53	11.89	18.80
TAILOR	7.15	8.40	**23.03**	**10.89**	17.80
SAS-PB	6.95	**8.10**	23.83	11.19	**17.60**

Across the four non-Gaussian laws, TAILOR reduces mean clean-mean RMSE by 6.9–31.8% and interval ECE by 43.3–76.0% relative to β-NLL, Faithful and Fisher8. Gaussian performance remains close. SAS-PB leads both metrics on Student-t_5 and ECE on lognormal; TAILOR leads both on contamination and its own teacher.

These results show complementary improvements in finite-sample mean estimation and marginal interval coverage. Panel (c) measures total predictive excess above the teacher, including fitting and family error; the shape-gap theorem isolates population approximation error. Table 31 reports marginal means and SDs; paired inference applies to the contrasts specified elsewhere. Table 16 gives the continuous shape-recovery sweep.