σTRANSFER: UNCERTAINTY TRANSFER FROM SMALL TO LARGE NETWORKS UNDER μP

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable predictive uncertainty in Laplace approximations depends critically on the prior precision, yet selecting it requires a posterior sweep that is prohibitively expensive for neural networks with billions of parameters. Under the Maximal Update Parametrization (μP), we derive a rescaling of the prior covariance that makes the selected precision stable as model width grows. This leads to σTransfer: we select the precision on a smaller model and zero-shot transfer it to the much larger model, i.e., without searching for the precision on the larger model at all. We show convergence of the prior kernel, posterior covariance, selected precision, and posterior-derived decisions under explicit conditions, and verify σTransfer across regression, image classification, and Transformer readouts. For example, measured precision-sweep speedups reach $\sim 5000\times$ when transferring from width 128 to 4096 on MNIST, at a target-NLL degradation of 0.002; transferring from a public 1B to 7B model gives a median search speedup of $\sim 2.3\times$ (up to $\sim 330\times$), with a mean measured target-NLL increase below 10^{-4} across ten tasks. The same posterior stability also enables transfer of acquisition, OOD-detection, and abstention decisions without constructing a target posterior.

1 INTRODUCTION

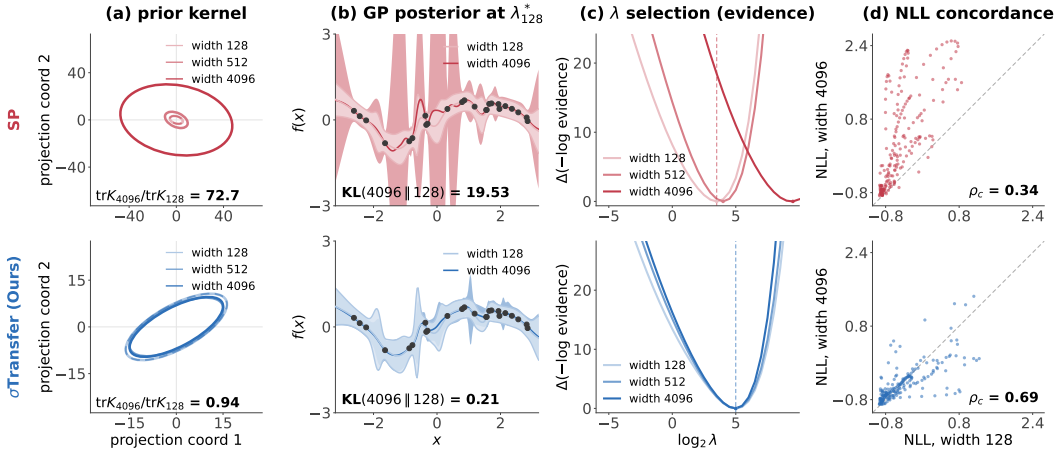

Figure 1: **Why prior precision transfers under σTransfer.** The same regression architecture at widths 128, 512, and 4096, under standard parametrization (SP, red) and σTransfer (blue). **(a)** The prior kernel, in a common two-dimensional projection. **(b)** Predictive posteriors using the precision selected at width 128. **(c)** Plug-in evidence curves used to select it; the dashed line marks the width-128 optimum. **(d)** Predictive NLL per example, width 128 versus wider networks; points on the diagonal indicate agreement. Overall, σTransfer remains stable across width, while SP can drift.

The Laplace approximation for neural networks (Denker & LeCun, 1990; MacKay, 1992) provides predictive uncertainty estimates and supports evidence-based model selection. Its linearised form

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.

places a Gaussian posterior on a trained network, yielding predictive distributions useful for out-of-distribution detection (Daxberger et al., 2021; Kristiadi et al., 2020), active learning (Gal et al., 2017; Kirsch et al., 2019), and selective prediction. The resulting uncertainty estimates depend on the prior precision, a hyperparameter that controls the scale of posterior uncertainty and is commonly selected using validation or marginal likelihood (Kristiadi et al., 2021; Immer et al., 2021a; Antorán et al., 2022). Selecting it typically requires repeated posterior evaluations across candidate precisions, making the sweep increasingly costly as network width grows.

Can we instead search for the prior precision on a small network and reuse it, unchanged, on a much larger one? Under standard parametrization (SP), PyTorch's default and the convention of a single global learning rate, this transfer can be unreliable: the same precision can produce different uncertainty at different widths. The Maximal Update Parametrization (μP; Yang & Hu, 2021; Yang et al., 2021) is an alternative convention that rescales each layer's initialization, forward multiplier, and learning rate with width. This stabilizes training dynamics across width, allowing hyperparameters such as the learning rate to be tuned at a small width and reused at a larger one. Applying the Laplace approximation to a μP network is not enough on its own: the prior precision that is optimal at one width is still not optimal at another.

We introduce $\sigma\mathrm{Transfer}$, which builds on μP to make the optimal prior precision stable across width (Figure 1(c)), so a precision chosen on a small (cheap) network is near-optimal for a large (expensive) one. $\sigma\mathrm{Transfer}$ normalizes the prior geometry: each layer's prior is scaled by a factor fixed by the width and μP's coordinate scalings (Section 4). This adds no new hyperparameter.

Roughly speaking, under $\sigma\mathrm{Transfer}$, networks of different widths approximate the same limiting prior kernel: the function-space covariance that the parameter prior induces on predictions, once the network is linearized around its trained weights. If this kernel is stable across width, we show that the Laplace posterior is stable too. We use this stability in two ways: a small network can select the prior precision for a larger one (*prior-precision transfer*), or make uncertainty-based decisions itself, such as which data to label or when the larger network should abstain, without constructing a Laplace posterior for it (*decision transfer*). Under the conditions of Section 5, our main guarantees can be summarized by the progression

$$\underbrace{K_n \to K_\infty}_{\text{prior-kernel stability}} \implies \underbrace{m_n^\lambda \to m_\infty^\lambda,\ C_n^\lambda \to C_\infty^\lambda}_{\text{posterior mean and covariance stability}} \implies \underbrace{\lambda_n^\star \to \lambda^\star}_{\text{prior-precision transfer}} \quad , \quad \underbrace{\text{same decision}}_{\text{decision transfer}} .$$

Figure 1 illustrates this progression at finite width, from the prior kernel to the posterior, the selected precision, and per-point negative log-likelihood (NLL) concordance. Our contributions are:

1. We introduce $\sigma\mathrm{Transfer}$, a method that stabilizes the prior kernel and therefore the Laplace posterior across width, so that uncertainty computed on a small (cheap) network transfers to a much larger (expensive) one.
2. *Prior-precision transfer*: the optimal prior precision is stable across width (Figure 1(c)), so a precision chosen on the small network is near-optimal for the large one, and we bound the excess NLL from using it. The large network then fits a single posterior instead of a full sweep.
3. *Decision transfer*: decisions made from the small network's posterior are approximately those the large network would make, because the posterior mean and covariance both approach a common limit and so approach each other. We demonstrate this for data acquisition, out-of-distribution detection, and abstention, where the large network needs no Laplace posterior at all.

2 Related work

The Tensor Programs series developed a mathematical framework for neural networks at large width, settling questions such as the convergence of standard architectures to Gaussian processes in the infinite-width limit (Yang, 2020a;b; 2019). A key practical outcome was the Maximal Update Parametrization (μP; Yang & Hu, 2021), which scales initialization and per-layer multipliers so that feature updates neither vanish nor explode as width grows, keeping training stable and expressive regardless of model size. Because μP networks converge to a common limit as width grows, optimization hyperparameters such as the learning rate tuned on a small proxy carry over to a larger target. Yang et al. (2021) demonstrated this with $\mu\mathrm{Transfer}$, transferring these settings from a 40M-parameter proxy to a 6.7B-parameter language model at a search cost of about 7% of one pretraining run, then training the large model once with the optimal learning rate found on the smaller proxy.

Denker & LeCun (1990) first used the Hessian to obtain predictive error bars for neural networks; MacKay (1992) developed this into a full Bayesian framework, placing a Gaussian posterior over trained weights and selecting the prior precision by maximising the marginal likelihood. Laplace posteriors have since achieved strong results in out-of-distribution detection, active learning, and selective prediction (Daxberger et al., 2021). The main computational barrier is the Hessian: storing it costs $O(P^2)$ and inverting it $O(P^3)$ in the number of parameters P. Last-layer restrictions (Daxberger et al., 2021), Kronecker-factored curvature (Ritter et al., 2018), and online evidence maximisation (Immer et al., 2021a) have each reduced this cost. Even so, selecting the prior precision requires sweeping over many candidates and evaluating the posterior at each one. When the target model has millions or billions of parameters, this search alone can dominate the cost of Laplace inference.

3 Preliminaries

Linearized Laplace. The Laplace approximation represents the parameter posterior by a Gaussian centred at the trained parameters $\hat{\theta}_n$ of a network with width n. Under the standard Gaussian prior $\theta \mid \lambda \sim \mathcal{N}(0, \lambda^{-1} I)$, the Laplace approximation gives

$$\theta \mid D, \lambda \;\sim\; \mathcal{N}\Big(\hat{\theta}_n, (H_n + \lambda I)^{-1}\Big), \qquad H_n = J_n(D)^\top \Omega_n J_n(D).$$

Here H_n is the generalized Gauss–Newton curvature, D the training inputs, and X, X' arbitrary finite sets of evaluation inputs. The function $f_n(x;\theta) \in \mathbb{R}^{d_{\text{out}}}$ denotes the neural-network output: a scalar in our regression experiments and the pre-softmax logit vector in classification; $J_n(X) = \partial f_n(X;\hat{\theta}_n)/\partial\theta$ is its Jacobian with respect to the stored parameters, stacked over the inputs in X, and Ω_n is the negative-log-likelihood Hessian with respect to these outputs, evaluated at the trained predictions. Linearized Laplace (Immer et al., 2021b) uses a first-order expansion of f_n around $\hat{\theta}_n$ to propagate the Gaussian parameter posterior to the output space of f_n. The resulting Gaussian output posterior has mean and covariance

$$m_n^\lambda(X) = f_n(X;\hat{\theta}_n), \qquad C_n^\lambda(X, X') = J_n(X)(H_n + \lambda I)^{-1} J_n(X')^\top.$$

With the trained parameters $\hat{\theta}_n$ held fixed, λ leaves the mean unchanged and controls the scale of posterior uncertainty: larger values contract the covariance, whereas smaller values permit greater uncertainty. For classification, predictive class probabilities are defined by averaging softmax over the Gaussian logit posterior; we approximate this expectation using either a probit approximation or Monte Carlo sampling. For Gaussian regression, observation noise is added when forming the predictive distribution. The validation-NLL curve is the map $\lambda \mapsto \mathrm{NLL}_n(\lambda)$ from prior precision to the NLL of the width-n posterior predictive on a fixed validation set.

μP and hyperparameter transfer. The Maximal Update Parametrization (μP), introduced by Yang & Hu (2021), enables training hyperparameters such as the learning rate to transfer across model widths (Yang et al., 2021). μP achieves this through width-dependent scaling of the initialization, forward pass, and parameter updates. Each layer is updated as much as possible without causing logits or activations to diverge as width tends to infinity, so hidden representations keep changing during training instead of staying close to their random initialization as the network widens. The same μP network can be expressed using different parameter units. Writing θ_n for the stored weights of a network with width n, we introduce rescaled parameter coordinates ϕ_n through

$$\phi_n = T_n^{-1}\theta_n, \qquad T_n = \mathrm{blockdiag}\big(t_1(n)I,\; t_2(n)I, \ldots\big).$$

Here T_n applies one positive scale factor to each parameter tensor. These factors depend on width but remain fixed during training. The initialization and update rules are adjusted consistently with these parameter units (Yang & Littwin, 2023, Proposition 2.2.3).

4 Method: σTransfer

We introduce σTransfer, a Laplace method that rescales the prior so that a selected prior precision or posterior-based decisions can transfer from a smaller network to a larger one. This section defines σTransfer and develops the intuition for its stability across width; Section 5 states the conditions and formal guarantees for both, and Appendix K provides the proofs.

The transfer problem. To transfer a prior precision or posterior-based decision across width, the same numerical value of the prior λ must induce similar linearized Laplace output posteriors as the network width grows. If the predictive mean m_n^λ and covariance C_n^λ each approach fixed limits, then sufficiently wide networks of widths n and N approximate the same limiting posterior and hence one another, even when $N \gg n$. We call a quantity evaluated on fixed inputs *width-stable* when it approaches a fixed limit as width grows; Definition 1 makes this precise.

Definition 1 (Width-stability)**.** *Fix a finite input set X. A quantity Q_n indexed by the network width n and evaluated on X is* width-stable *if it converges entrywise, $Q_n \to Q_\infty$ as $n \to \infty$.*

Predictive-mean stability has already been established.[1] We therefore focus on stabilizing the covariance. Section 3 gives the posterior covariance under an isotropic prior, which we write here as

$$C_{n,I}^\lambda(X, X') = J_n(X)(H_n + \lambda I)^{-1} J_n(X')^\top, \qquad H_n = J_n(D)^\top \Omega_n J_n(D). \tag{1}$$

Here n denotes the network width, J_n is the network-output Jacobian at the trained parameters, and Ω_n is the output-space NLL curvature on the training data. This covariance expression does not reveal when width-stability holds, because J_n and H_n change dimension with width.

To identify how to make this covariance width-stable, we apply the Woodbury identity (Woodbury, 1950) to rewrite it in function space. The key step is to do so for a general Gaussian prior with covariance $\lambda^{-1} S_n$, rather than the isotropic prior alone. For the full GGN and $\lambda > 0$,

$$C_n^\lambda(X, X') = \frac{K_n(X, X')}{\lambda} - \frac{K_n(X, D)}{\lambda^2} \Omega_n^{1/2} \big(I + \lambda^{-1} \Omega_n^{1/2} K_n(D, D) \Omega_n^{1/2}\big)^{-1} \Omega_n^{1/2} K_n(D, X'), \tag{2}$$

where $K_n(X, X') = J_n(X) S_n J_n(X')^\top$ is the prior kernel. In this form, the prior geometry S_n enters the posterior covariance only through K_n. The posterior also depends on the prior precision λ, which is fixed, and the likelihood curvature Ω_n, which is width-stable for the likelihoods considered here.[2] However, K_n is not guaranteed to be width-stable: for instance, the isotropic choice $S_n = I$ gives $K_{n,I} = J_n J_n^\top$, which can diverge as width grows. We therefore need to choose S_n so that K_n is width-stable; Eq. (2) then implies that C_n^λ is width-stable.

The σTransfer prior kernel. We now construct S_n so that the prior kernel K_n is width-stable. Building on the width scalings of μP, we introduce width-normalized parameter coordinates $\phi_n = T_n^{-1}\theta_n$. The block-diagonal map T_n rescales each parameter tensor so that its contribution to the prior kernel remains at a comparable scale across widths. Appendix J derives these factors.

The reason for working in these coordinates is that **we prove that the isotropic prior kernel in these width-normalized coordinates is width-stable**. Under the conditions of Theorem 1 (see Section 5 for the precise statement),

$$K_n(X, X') = J_{\phi,n}(X)\, J_{\phi,n}(X')^\top \xrightarrow{n\to\infty} K_\infty(X, X').$$

Here $J_{\phi,n}$ is the network-output Jacobian with respect to the width-normalized coordinates ϕ_n, rather than the stored parameters. Since $\phi_n = T_n^{-1}\theta_n$, the chain rule gives $J_{\phi,n} = J_n T_n$. We can therefore express this same stable kernel in the stored parameter coordinates:

$$K_n(X, X') = J_{\phi,n}(X)\, J_{\phi,n}(X')^\top = J_n(X)\, T_n T_n^\top\, J_n(X')^\top = J_n(X)\, S_n\, J_n(X')^\top, \tag{3}$$

where $S_n = T_n T_n^\top$. Equivalently, λS_n^{-1} is the isotropic precision λI of the width-normalized coordinates expressed in the stored coordinates, which is why we call it a width-normalized prior. The posterior covariance in the stored parameter coordinates is therefore

$$C_n^\lambda(X, X') = J_n(X)\big(H_n + \lambda\, S_n^{-1}\big)^{-1} J_n(X')^\top. \tag{4}$$

Thus σTransfer changes only the prior term from λI to λS_n^{-1}; the trained checkpoint, predictions, and likelihood curvature remain unchanged, and λ remains the only tuned prior hyperparameter. Section 5 states the conditions under which this gives prior-precision and decision transfer across width. In practice, S_n is never constructed as a dense matrix: within a parameter tensor with width factor t, the prior precision is the scalar λ/t^2, which is what one passes to the Laplace implementation. Appendix J gives the block scales and verifies the coordinate identity numerically.

[1] For linearized Laplace, the predictive mean is the trained μP network output, whose width-stability under matched fixed-step training follows from Yang & Hu (2021); Yang & Littwin (2023).

[2] For fixed-noise Gaussian regression, Ω_n is independent of width. For softmax classification, it converges whenever the trained logits converge.

5 THEORETICAL GUARANTEES

Section 4 constructed the σTransfer prior so that the prior kernel K_n is width-stable. We now propagate this stability through the posterior to practical transfer guarantees; formal statements and proofs appear in Appendix K. Together with convergence of the likelihood curvature and trained predictive mean, the results build on each other under the conditions below:

$$\underbrace{K_n \to K_\infty}_{\substack{\text{Kernel stability}\\ \text{Thm. 1}}} \implies \underbrace{C_n^\lambda \to C_\infty^\lambda}_{\substack{\text{Covariance stability}\\ \text{Thm. 2}}} \implies \underbrace{\lambda_n^\star \to \lambda^\star}_{\substack{\text{Precision transfer}\\ \text{Thm. 3}}}, \quad \underbrace{a_{n,\lambda}^\star = a_{\infty,\lambda}^\star}_{\substack{\text{Decision transfer}\\ \text{Thm. 4}}}$$

Theorem 1 (Prior-kernel stability; informal)**.** *Consider the scalar-output, fixed-depth, equal-width ReLU μP MLP of Appendix J. Suppose its checkpoint is obtained after a fixed finite number of width-matched μP training steps and satisfies the regularity conditions stated precisely in Appendix K.1. Then, on every fixed finite input set, the σTransfer prior kernel $K_n(x, x') = J_n(x) S_n J_n(x')^\top$ converges jointly almost surely as $n \to \infty$; the trained outputs converge jointly as well. The same conclusion holds for any fixed output dimension. The full statement, assumptions, and proof are in Appendix K.1.*

The proof builds on Tensor Programs (Yang & Littwin, 2023), which establish large-width limits for empirical averages arising during a fixed number of training steps. The σTransfer normalization expresses the prior kernel as a finite combination of these averages, so their convergence implies convergence of the kernel. Appendix K.1 provides the smoothing argument needed to extend this reasoning to ReLU.

Theorem 2 (Posterior-covariance stability; informal)**.** *For full-GGN Laplace, given prior-kernel convergence and width-stable positive-semidefinite likelihood curvature, the linearized-Laplace output covariance C_n^λ converges uniformly over any interval $[\lambda_-, \lambda_+]$ with $0 < \lambda_- \le \lambda_+ < \infty$ (proof in Appendix K.3).*

Thus the same numerical precision produces increasingly similar posterior uncertainty as the networks become wider.

Theorem 3 (Prior-precision transfer; informal)**.** *If the posterior mean and covariance on a fixed validation set converge uniformly in λ over an interval $\Lambda = [\lambda_-, \lambda_+]$ with $0 < \lambda_- \le \lambda_+ < \infty$, then the validation-NLL curves converge uniformly. If the limiting curve has a unique minimizer $\lambda^\star$, any sequence of minimizers satisfies $\lambda_n^\star \to \lambda^\star$. Even without a unique minimizer, the excess NLL from using the width-n proxy's precision on a target of width N satisfies*

$$\mathrm{NLL}_N(\lambda_n^\star) - \min_{\lambda \in \Lambda} \mathrm{NLL}_N(\lambda) \le 2 \sup_{\lambda \in \Lambda} |\mathrm{NLL}_n(\lambda) - \mathrm{NLL}_N(\lambda)|. \tag{5}$$

Full conditions and proof in Appendix K.5.

Theorem 4 (Posterior-derived decision transfer; informal)**.** *Fix $\lambda > 0$ and a finite action set. Suppose the posterior mean and covariance converge on the fixed finite set containing all inputs used by the action scores. If the scores depend continuously on these moments through scoring rules shared across widths, they converge uniformly. If the limiting best action has a positive gap over all alternatives, two sufficiently wide networks choose the same action. At finite widths, even without this gap, the loss in target score from using the proxy's action is at most twice the largest proxy–target score difference. Full conditions and proof appear in Appendix K.6.*

Scope. The posterior-transfer results apply beyond MLPs: for any width-indexed architecture with $\theta_n = T_n \phi_n$, Theorems 2–4 apply under their stated conditions whenever the width-normalized prior kernel, predictive mean, and likelihood curvature converge. Theorem 1 establishes kernel convergence for the stated finite-step ReLU MLP setting. For last-layer Laplace, the kernel condition reduces to convergence of the normalized final-feature Gram matrix. Extensions to diagonal/KFAC curvature, to finitely many sequential acquisition rounds, and to decisions at the transferred precision appear in Appendices K.4 and K.6, under additional conditions. Precision transfer using fixed-checkpoint evidence additionally requires convergence of the prior-weighted squared parameter norm and a unique limiting maximizer; posterior convergence alone is insufficient (Appendix K.5).

6 RESULTS

We evaluate σTransfer in two deployment regimes. Under Regime A (prior transfer), a small proxy selects the Laplace prior precision λ, and the larger target constructs one posterior at the transferred value. This removes the target-side precision sweep, not the target posterior itself. Under Regime B (decision transfer), the proxy supplies posterior-derived decisions, such as acquisition rankings and active-learning selections, and the target requires no posterior at deployment. Target posteriors in our Regime-B experiments are constructed only as evaluation references. To measure transfer quality we track two estimands: the selection gap $|\Delta\lambda|$, the distance between the proxy and target optimal precisions in bits, and the predictive transfer gap $|\Delta\mathrm{NLL}|$, the absolute target test-NLL difference between proxy-selected and target-selected precisions. All posteriors are post-hoc GGN Laplace around fixed checkpoints; formal definitions of all evaluation metrics are in Appendix B.

The main tables compare four methods. SP with an isotropic prior is standard Laplace practice and serves as the main baseline. We add three further comparisons to isolate the source of any improvement. μP with an isotropic prior uses the μP parametrization but keeps the standard Laplace prior, testing whether μP training alone helps transfer. SP + metric applies our prior geometry under the standard parametrization; this combination is not covered by Theorem 1, but it serves as a mis-specification control and, as we see below, sometimes helps empirically. σTransfer combines μP training with the width-normalized prior. The two μP methods share MAP checkpoints, so any difference between them is due to the prior geometry alone. Appendix C.4 evaluates additional empirical scale-matching controls on the same checkpoints.

6.1 REGRESSION

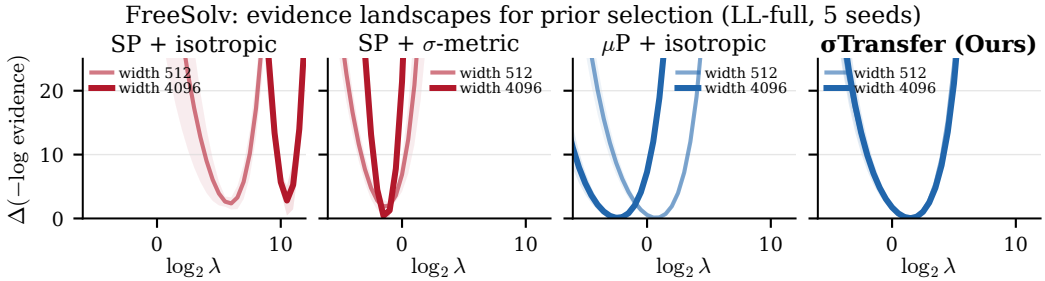

Figure 2: Evidence as a function of prior precision λ on FreeSolv. Light and dark curves show proxy (width 512) and target (width 4096); bands are mean $\pm$ SEM over five seeds. Under SP the landscape shifts with width, so the proxy's optimum is far from the target's. Under σTransfer the two curves nearly overlap. Table 1 reports numerical results for all three datasets.

Table 1: Prior-precision transfer on three regression datasets (proxy width 512, target width 4096, full-covariance last-layer Laplace, five paired seeds, mean $\pm$ s.d.). $|\Delta\mathrm{NLL}|$ scaled by 10^3; lower is better for all columns. See Appendix B for metric definitions.

Method	ESOL			FreeSolv			Lipophilicity		
	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\mathrm{KL}_{4096\Vert512}$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\mathrm{KL}_{4096\Vert512}$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\mathrm{KL}_{4096\Vert512}$
SP + isotropic	$1.5_{\pm0.4}$	$41.8_{\pm30.7}$	$0.208_{\pm0.073}$	$5.0_{\pm2.2}$	$452.8_{\pm429.5}$	$0.079_{\pm0.055}$	$4.0_{\pm2.7}$	$140.2_{\pm119.7}$	$0.173_{\pm0.027}$
SP + metric	$1.5_{\pm0.5}$	$8.40_{\pm6.59}$	$0.208_{\pm0.071}$	$1.2_{\pm1.4}$	$9.65_{\pm10.95}$	$0.079_{\pm0.055}$	$2.1_{\pm1.2}$	$3.39_{\pm1.93}$	$0.173_{\pm0.028}$
μP + isotropic	$3.0_{\pm1.4}$	$0.947_{\pm0.566}$	$\mathbf{0.167}_{\pm0.087}$	$3.1_{\pm0.5}$	$1.564_{\pm0.825}$	$\mathbf{0.008}_{\pm0.004}$	$3.2_{\pm1.0}$	$0.094_{\pm0.024}$	$\mathbf{0.057}_{\pm0.058}$
σTransfer **(ours)**	$\mathbf{0.9}_{\pm0.8}$	$\mathbf{0.548}_{\pm0.412}$	$\mathbf{0.167}_{\pm0.087}$	$\mathbf{0.5}_{\pm0.4}$	$\mathbf{0.461}_{\pm0.430}$	$\mathbf{0.008}_{\pm0.004}$	$\mathbf{0.4}_{\pm0.2}$	$\mathbf{0.033}_{\pm0.019}$	$\mathbf{0.057}_{\pm0.058}$

We test Regime A on ESOL (Delaney, 2004), FreeSolv (Mobley & Guthrie, 2014), and Lipophilicity (Wu et al., 2018) under scaffold shift, transferring an evidence-selected precision from width 512 to 4096 using full-covariance last-layer Laplace and five paired seeds (full protocol in Appendix C.3). On FreeSolv the SP + isotropic evidence landscape shifts markedly across widths, whereas the σTransfer curves nearly overlap (Figure 2). Across all three datasets σTransfer gives the smallest selection gap $|\Delta\lambda|$ (0.4–0.9 bits, against 1.5–5.0 for SP + isotropic) and a predictive transfer gap $|\Delta\mathrm{NLL}|$ orders of magnitude smaller (Table 1). We also report KL divergence to measure agreement between the proxy's and target's predictive distributions (Appendix B).

6.2 IMAGE CLASSIFICATION

Table 2: Prior-precision transfer on MNIST and Fashion-MNIST across three last-layer Laplace approximations (proxy width 128, target width 4096, validation-NLL-selected λ, 10 paired seeds, mean $\pm$ s.d.). $|\Delta\text{NLL}|$ scaled by 10^2; lower is better for all columns. See Appendix B for metric definitions and Appendix D for evidence-selected results.

Method	MNIST						FMNIST					
	LL full		LL diag		LL KFAC		LL full		LL diag		LL KFAC	
	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\text{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\text{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\text{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\text{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\text{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\text{NLL}\rvert$
SP + isotropic	$1.75_{\pm.26}$	$9.8_{\pm4.6}$	$0.45_{\pm.28}$	$0.8_{\pm0.9}$	$1.50_{\pm.24}$	$6.9_{\pm3.2}$	$1.55_{\pm.28}$	$9.0_{\pm4.1}$	$\mathbf{0.25_{\pm.35}}$	$0.9_{\pm1.2}$	$1.85_{\pm.24}$	$13.2_{\pm3.2}$
SP + metric	$3.25_{\pm.26}$	$9.6_{\pm2.1}$	$4.55_{\pm.28}$	$14.8_{\pm1.9}$	$3.50_{\pm.24}$	$10.7_{\pm1.9}$	$3.45_{\pm.28}$	$39.0_{\pm6.3}$	$5.25_{\pm.35}$	$67.2_{\pm7.3}$	$3.15_{\pm.24}$	$33.3_{\pm5.5}$
μP + isotropic	$4.75_{\pm.26}$	$15.7_{\pm3.0}$	$5.30_{\pm.26}$	$20.6_{\pm3.2}$	$4.75_{\pm.26}$	$16.5_{\pm2.7}$	$4.60_{\pm.21}$	$34.9_{\pm2.8}$	$4.55_{\pm.28}$	$47.2_{\pm5.0}$	$4.45_{\pm.28}$	$30.5_{\pm2.6}$
σTransfer **(ours)**	$\mathbf{0.25_{\pm.26}}$	$\mathbf{0.2_{\pm0.3}}$	$\mathbf{0.30_{\pm.26}}$	$\mathbf{0.1_{\pm0.2}}$	$\mathbf{0.25_{\pm.26}}$	$\mathbf{0.2_{\pm0.3}}$	$\mathbf{0.40_{\pm.21}}$	$\mathbf{0.4_{\pm0.4}}$	$0.45_{\pm.28}$	$\mathbf{0.6_{\pm0.6}}$	$\mathbf{0.55_{\pm.28}}$	$\mathbf{0.3_{\pm0.3}}$

Prior transfer (Regime A). Table 2 reports both estimands on MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017) across three last-layer Laplace approximations (full, diagonal, KFAC; see Appendix C.1 for definitions). Among the four methods in Table 2, σTransfer has the smallest predictive transfer gap $|\Delta\text{NLL}|$ in every column, and its worst case (0.6×10^{-2}, FMNIST LL-diagonal) still beats every other method on every approximation and dataset. Its selection gap $|\Delta\lambda|$ is smallest in five of the six columns and never exceeds 0.55 bits, against up to 1.85 for SP + isotropic; the exception is FMNIST LL-diagonal, where SP + isotropic selects marginally closer. The same selection-gap ordering holds under plug-in evidence selection, although its optimum is not always test-NLL-optimal (Appendix D). Table 3 reports the compute: proxy-side selection is $\sim$5000$\times$ faster than the target-side sweep for the full last-layer posterior and $94\times$ for all-parameter diagonal Laplace, and after transfer the target builds one posterior at the transferred precision for a test-NLL degradation of only 0.002. When the selection criterion is OOD detection instead of validation NLL, σTransfer also transfers: at width 4096 the proxy-selected precision achieves 0.950 AUROC on held-out OOD data, matching target-side tuning, whereas SP loses 0.072 and μP with an isotropic prior loses 0.327. The match holds at every width above the proxy (Appendix E).

Decision transfer (Regime B). Figure 3 shows a sequential active-learning experiment on MNIST ($1{,}000$ labelled points, 50 rounds). Selecting data with the proxy produces nearly the same learning curve on the target as using the target's own posterior, with final accuracy only 0.003 below the target reference and a paired area-under-the-learning-curve difference covering zero; under SP the curves diverge, meaning decisions are not transferable. Per-round Spearman correlation between proxy and target EPIG scores (Bickford Smith et al., 2023) is $\bar{\rho} = 0.75$ under σTransfer, against 0.07 for SP + isotropic (Figure 3, right panel), consistent with our theoretical guarantees (Section 5). On a fixed candidate pool the agreement is higher still: $\rho = 0.854 \pm 0.025$ under σTransfer, against 0.669 for SP. The same agreement ordering holds on Fashion-MNIST; we claim decision transfer on both datasets and downstream utility on MNIST. Appendix F gives full learning curves, AULC contrasts, and per-dataset acquisition-agreement tables.

6.3 LANGUAGE MODELS

We train a ladder of decoder-only GPT models at widths 128, 256, 512, 1024, and 2048 (8 layers, 8 heads) on 10^9 FineWeb-Edu tokens (Penedo et al., 2024) under both SP and μP. Training and posterior details are in Appendix G.

Prior transfer (Regime A). Figure 4 (left) shows the evidence-selected prior precision across this ladder (ten downstream tasks). Under σTransfer the optimum converges with width; under SP it drifts about 0.65 bits per width doubling without saturating, so no proxy width is safe. The proxy's choice under σTransfer lands 0.65 ± 0.34 bits from the target's own and degrades target test NLL by only 0.0014 on average, compared to 2.65 ± 0.63 bits and a 0.013 degradation under SP (nearly an order of magnitude larger). Three of ten tasks have absolute target-NLL changes below 10^{-4}; the largest change is 0.0082 (Figure 4, middle). On the public u-μP pairs (Blake et al., 2025) (1B$\rightarrow$7B), seven of ten tasks select the same precision as the target, and eight have absolute target-NLL changes below 10^{-4} (Figure 4, right; Appendices H and I).

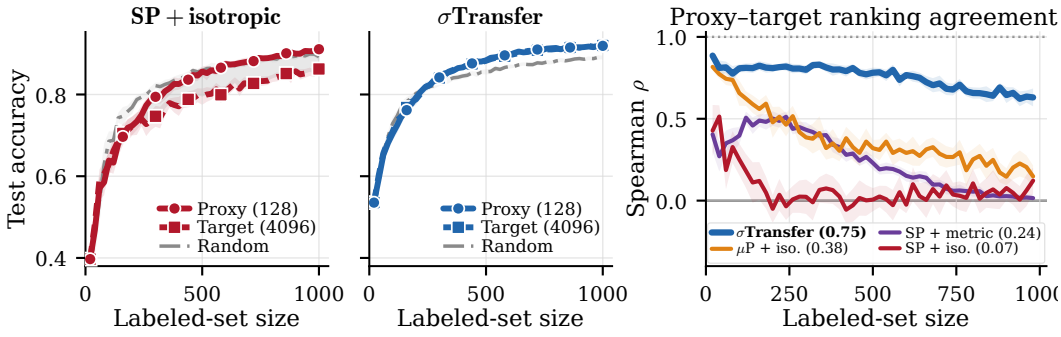

Figure 3: Sequential active learning on MNIST with proxy acquisition (EPIG, proxy width 128, target width 4096, 20 seeds, bands: 95% CI). Left and centre: target test accuracy under proxy, target-reference, and random acquisition. Under σTransfer the proxy and target curves nearly overlap; under SP + isotropic the target's own acquisition falls below random. Right: per-round Spearman ρ between proxy and target rankings. See Appendix F for Fashion-MNIST and AULC contrasts.

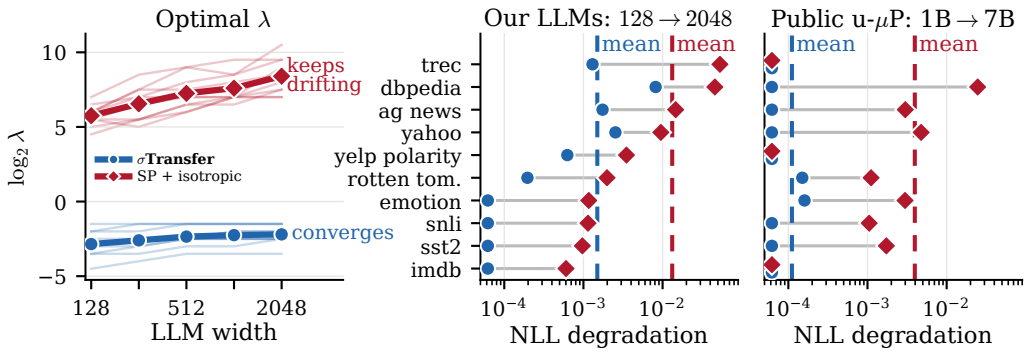

Figure 4: A prior precision chosen on a proxy transfers under σTransfer, not under SP. Left: evidence-based selection on in-distribution training data, per task (thin) and task mean (thick). Middle and right: target NLL at the proxy's λ minus the target's own (lower is better), same ten tasks.

Decision transfer (Regime B). Figure 5 runs sequential EPIG on AG News (Zhang et al., 2015): proxy widths 128 to 1024 acquire labels for a width-2048 target, all at a common, untuned $\lambda = 1$. σTransfer has the highest mean Spearman correlation between proxy and target rankings of which data to label, across all proxy widths and initial label budgets. For $128 \to 2048$ transfer with 5,000 initial labels, target accuracy is lower by 0.0014 with proxy-selected labels than with target-selected labels (Table 3). The bottom row trains the larger model on the datapoints the smaller one selected, and measures the absolute test-NLL gap to training it on the target's own selections, an oracle that requires the very target posterior we are avoiding. σTransfer has the smallest mean absolute test-NLL gap in nine of twelve settings. Appendix G reports a second Regime-B test, where a width-128 proxy gate lets the target answer by MAP alone: it abstains on 86–88% of unseen OOD prompts while serving 96% of in-distribution traffic.

6.4 Computational savings and limitations

Table 3 shows the saving scaling with the cost of the target-side procedure: up to $\sim$5000$\times$ on MNIST, where the target would otherwise sweep a full last-layer Laplace, and a median $\sim$2.3$\times$ on the public 1B$\to$7B pair (Blake et al., 2025), ranging from 1$\times$ to $\sim$330$\times$ across nine timed tasks. It is far smaller where that procedure is already cheap, and σTransfer does not always win: on FMNIST under last-layer diagonal curvature, SP + isotropic selects a marginally closer precision. Appendix C.5 gives the PenDigits and Letter setup.

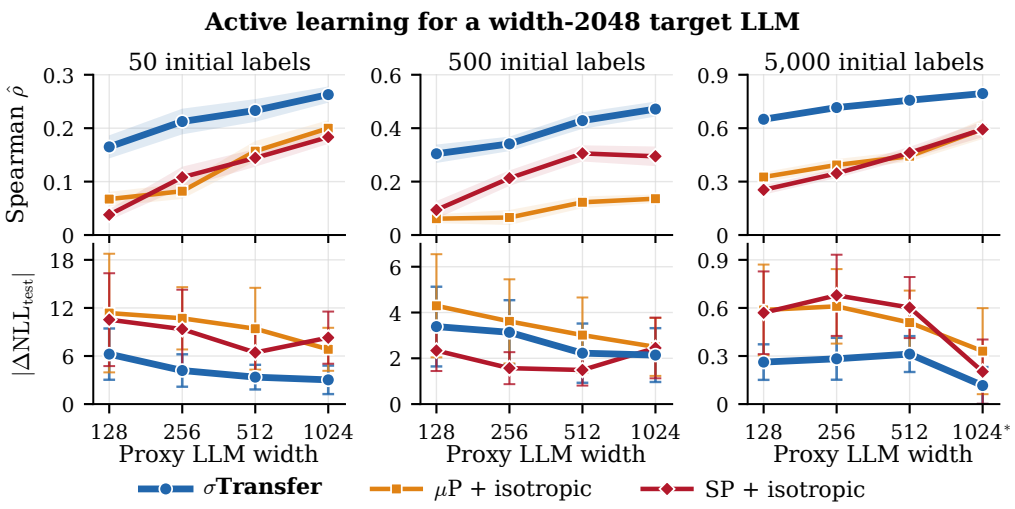

Figure 5: Sequential active learning on AG News, starting with 50, 500 or 5,000 labelled examples and acquiring 250 additional labels using EPIG. Top: Spearman correlation between the proxy's and target's rankings of unlabelled data by informativeness (EPIG), averaged over acquisition steps. Bottom: mean absolute test-NLL gap when the target is trained on proxy-selected rather than target-selected labels (ticks $\times 10^{-3}$; lower is better). All methods use $\lambda = 1$. Means and 95% CIs over ten seeds; the starred setting uses five.

Task	Dataset	Transfer	Approx.	Proxy (s)	Target (s)	Speedup	Degradation
λ sweep	MNIST	128→4096	LL full	1.5	7,560	∼5000×	0.002 NLL
OOD-aware λ	MNIST	128→4096	all + diag	4.1	384	94×	0.001 AUROC
evidence search	LLM (ours)	128→2048	LL full	3.5	25	∼7×	0.0014 NLL
evidence search	LLM (public)	1B→7B	LL full	4	9	∼2.3×	0 NLL
EPIG acquisition	AG News	128→2048	LL full	64.9	187.6	∼3×	0.0014 ACC
evidence search	PenDigits	32→128	all + full	0.68	14.95	∼22×	−0.007 NLL
evidence search	letter	32→128	all + full	0.77	24.74	∼32×	0.078 NLL

Table 3: σTransfer does the expensive selection on the proxy rather than the target. Speedup is target ÷ proxy selection time; deployment still fits one target posterior, so the end-to-end saving is capped by the 61-point grid. Degradation is the resulting change in the stated metric, positive meaning worse. EPIG times average 10 seeds; all times exclude common MAP training and one-off frozen-trunk feature extraction. For the public 1B→7B pair, the row reports AG News, whose speedup is the median across nine timed tasks, calculated from the rounded times in Table 15. LL = last-layer.

7 Conclusion

We introduced σTransfer, a method that uses a small (cheap) network to make uncertainty estimation and uncertainty-based decision-making cheaper for a much larger (expensive) network. It supports two forms of transfer. In *prior-precision transfer*, the precision search is performed on the small network and the selected precision is transferred zero-shot to the much larger network, without any target-side search. In *decision transfer*, the small network's uncertainty estimates are used directly to select data, flag out-of-distribution inputs, or decide when the larger network should abstain. We justify both forms of transfer theoretically through a series of results showing accurate transfer across network widths. Empirically, σTransfer preserves predictive performance while reducing the cost of uncertainty-sensitive tasks by up to several orders of magnitude. We discuss the limitations of our approach and directions for future work in Appendix A.

AI USE STATEMENT

We derived the main theoretical results. For Theorem 1, generative AI tools helped us navigate the Tensor Programs literature (Yang, 2019; 2020a; Yang & Hu, 2021; Yang & Littwin, 2023) and identify its key tools, including the Master Theorem (Yang & Littwin, 2023, Theorem 2.6.10), which we then instantiated for our setting. AI tools also assisted with writing up proofs and some derivation steps, implementing the method and experiments, analysing results, literature search, feedback on the experiments and manuscript, and drafting and editing. We reviewed all AI-assisted work, checked numerical claims against saved outputs and theoretical statements against their proofs, and take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

The appendices provide the parameterization conventions, implementation details, assumptions, and complete proofs used in this work. Experimental protocols report model widths, random seeds, posterior approximations, hyperparameter-selection rules, evaluation metrics, and compute measurements. Code, configuration files, and result manifests will be included in the supplementary material.

REFERENCES

Ethem Alpaydin and Fevzi Alimoglu. Pen-based recognition of handwritten digits. UCI Machine Learning Repository, 1996. URL https://doi.org/10.24432/C5MG6K.

Javier Antorán, David Janz, James U. Allingham, Erik Daxberger, Riccardo Barbano, Eric Nalisnick, and José Miguel Hernández-Lobato. Adapting the linearised Laplace model evidence for modern deep learning. In *International Conference on Machine Learning*, pp. 796–821. PMLR, 2022.

Javier Antorán, Shreyas Padhy, Riccardo Barbano, Eric Nalisnick, David Janz, and José Miguel Hernández-Lobato. Sampling-based inference for large linear models, with application to linearised Laplace. In *International Conference on Learning Representations*, 2023.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented Bayesian active learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 7331–7348. PMLR, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Charlie Blake, Constantin Eichenberg, Josef Dean, Lukas Balles, Luke Y. Prince, Björn Deiseroth, Andres Felipe Cruz-Salinas, Carlo Luschi, Samuel Weinbach, and Douglas Orr. u-μP: The unit-scaled maximal update parametrization. In *International Conference on Learning Representations*, 2025.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 632–642, 2015.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021.

John S Delaney. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3):1000–1005, 2004.

John Denker and Yann LeCun. Transforming neural-net output levels to probability distributions. *Advances in neural information processing systems*, 3, 1990.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning*, pp. 4563–4573. PMLR, 2021a.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *International conference on artificial intelligence and statistics*, pp. 703–711. PMLR, 2021b.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pp. 5436–5446. PMLR, 2020.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Learnable uncertainty under laplace approximations. In *Uncertainty in Artificial Intelligence*, pp. 344–353. PMLR, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Xin Li and Dan Roth. Learning question classifiers. In *Coling 2002: The 19th international conference on computational linguistics*, 2002.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

David L Mobley and J Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28(7):711–720, 2014.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pp. 115–124, 2005.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International conference on learning representations*, 2018.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3687–3697, 2018.

David Slate. Letter recognition. UCI Machine Learning Repository, 1991. URL https://doi.org/10.24432/C5ZP40.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Max A Woodbury. *Inverting modified matrices*. Department of Statistics, Princeton University, 1950.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Greg Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *Advances in neural information processing systems*, 32, 2019.

Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020a.

Greg Yang. Tensor programs iii: Neural matrix laws. *arXiv preprint arXiv:2009.10685*, 2020b.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021.

Greg Yang and Etai Littwin. Tensor programs ivb: Adaptive optimization in the infinite-width limit. *arXiv preprint arXiv:2308.01814*, 2023.

Greg Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34:17084–17097, 2021.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Table 4: Roadmap to the appendix.

Location	Contents
Appendix A	Limitations and directions for future work.
Appendix B	Metrics for prior-precision and decision transfer.
Appendix C	Curvature approximations, the geometric mechanism, regression setup, and additional normalization baselines.
Appendix D	Evidence-selected precision transfer on MNIST and Fashion-MNIST.
Appendix E	Additional results under distribution shift.
Appendix F	Sequential active-learning setup and results.
Appendix G	OOD detection and abstention setup, calibration, and per-dataset results.
Appendix H	Per-dataset precision transfer and the neural g-prior comparison.
Appendix I	External validation on public u-μP/SP model pairs.
Appendix J	Coordinate definition, MLP derivation, block scales, and implementation details.
Appendix K	Formal proofs, structured curvature, and covered decision rules.

A LIMITATIONS AND FUTURE WORK

Theorem 1 establishes kernel convergence for fixed-depth ReLU MLPs under finite-step training and regularity assumptions. Establishing the corresponding convergence conditions for our Transformer settings and determining the proxy width needed for a given transfer accuracy remain open.

Future work could extend $\sigma\mathrm{Transfer}$ to applications in which reliable uncertainty is especially important, including medical prediction and generative models such as diffusion models. Another important direction is to move beyond width-only transfer and derive prior-scaling rules for models whose width and depth increase together. This would extend the broader μP scaling programme to Laplace priors and could enable prior precision and uncertainty-based decisions to transfer across entire model families, rather than only across networks of different widths.

B EVALUATION METRICS

All experiments evaluate transfer fidelity: how well does a quantity chosen or computed at proxy width reproduce the corresponding quantity at target width? The metrics below are grouped by regime. Where a 95% confidence interval is reported, it is an interval for the mean using the Student-t critical value with $n-1$ degrees of freedom, where n is the number of seeds. Tables and figures that report $\pm$ give sample standard deviations unless stated otherwise. For language-model prior transfer, $\pm$ denotes sample standard deviation across tasks. Our models use one pretraining seed per width and parametrization; the public models use one checkpoint per size and parametrization.

Regime A (prior transfer). Each method selects $\lambda_n^\star$ independently at every width n by minimizing validation NLL or maximizing the fixed-checkpoint plug-in Laplace evidence score over a fixed grid $\log_2 \lambda \in [-15, 15]$. Target-side selection is performed only to define the reference optimum for these metrics. In zero-shot deployment, the proxy-selected value is reused and the target constructs its posterior once at that value.

Selection gap. The absolute distance between the proxy and target optima, measured in bits of log-precision:

$$|\Delta\lambda| \;=\; |\log_2 \lambda^\star_{\mathrm{proxy}} - \log_2 \lambda^\star_{\mathrm{target}}|. \tag{6}$$

A gap of zero means both widths select the same grid point. We compute the absolute value per seed or task and report the mean $\pm$ sample standard deviation.

Predictive transfer gap. The absolute difference in target test NLL between using the proxy-selected precision and the target's own optimum:

$$|\Delta\mathrm{NLL}| \;=\; \left|\mathrm{NLL}_{\mathrm{target}}(\lambda^\star_{\mathrm{proxy}}) - \mathrm{NLL}_{\mathrm{target}}(\lambda^\star_{\mathrm{target}})\right|. \tag{7}$$

This measures the predictive discrepancy induced by trusting the proxy's choice. As with the selection gap, absolute values are taken per seed before averaging. In the regression tables the values are scaled by 10^3; in the classification tables by 10^2.

Cross-width predictive KL. A post-hoc diagnostic used in the regression experiments:

$$\mathrm{KL}_{t\|p} \;=\; \mathbb{E}_x\, \mathrm{KL}\big[p^{\lambda^\star_{\mathrm{target}}}_{\mathrm{target}}(\cdot \mid x) \;\big\|\; p^{\lambda^\star_{\mathrm{proxy}}}_{\mathrm{proxy}}(\cdot \mid x)\big], \tag{8}$$

in nats (including observation noise). This measures overall predictive-distribution agreement, not action agreement or a transferred-minus-local performance difference. Because methods using isotropic and width-normalized priors share trained checkpoints within each parametrization family, their KL values are expected to be nearly identical.

OOD AUROC. When the selection criterion is OOD detection rather than validation NLL, we report the area under the receiver operating characteristic curve for distinguishing in-distribution from out-of-distribution test inputs at the selected precision. The OOD-AUROC *regret* is the gap between the transferred and target-selected AUROC values.

Regime B (decision transfer). *Spearman rank correlation (ρ).* The rank correlation between the proxy's and target's acquisition scores (e.g. EPIG values) over a common candidate pool. Higher ρ means the proxy ranks candidates in the same order as the target. Reported per round in sequential active learning and on a single fixed pool for one-shot acquisition.

Top-K overlap. The fraction of the proxy's top-K selected candidates that also appear in the target's top-K. It complements rank correlation by focusing on the items that would actually be chosen.

Area under the learning curve (AULC). In sequential active learning, we integrate the test accuracy over acquisition rounds. The proxy-AULC is compared to the target-reference AULC and to random acquisition to assess whether proxy-guided selection translates into downstream utility.

OOD abstention gate. In the language-model gating experiment (Section 6.3), the proxy's BALD score (Houlsby et al., 2011) is thresholded to decide whether the target should abstain. We report the fraction of OOD prompts correctly abstained on (OOD-accuracy) and the fraction of in-distribution traffic served, calibrated on one OOD source and evaluated on a different, unseen one.

C Experimental details

Our main benchmark is a two-hidden-layer ReLU MLP on flattened MNIST ($d = 784$), with hidden width $n \in \{128, 256, 512, 1024, 2048, 4096\}$. We compare **SP**, standard parametrization with an isotropic stored-coordinate prior; μP **+ isotropic**, μP training with the same stored-coordinate prior; and σTransfer, μP training with the width-normalized prior (ours). The two μP methods share the same trained checkpoints, isolating the effect of the prior geometry. All models reach 100% training accuracy; at width 4096, mean validation accuracy is 0.950 for SP and 0.934 for the μP models, so any SP uncertainty failure below is not an optimization failure.

For MNIST and Fashion-MNIST, we sample $5{,}000$ training and $1{,}000$ validation examples from the official training split, and $2{,}000$ test examples from the official test split. Flattened pixels lie in $[0, 1]$, and the classification head has no bias. We use Adam for SP and MuAdam for μP, with the same base learning rate 10^{-3} across widths and parametrizations, for $2{,}000$ full-batch updates without early stopping or weight decay.

Unless stated otherwise, we use a GGN Laplace approximation and search a fixed 61-point grid, $\log_2 \lambda \in [-15, 15]$. The width-$128$ model is the proxy; target widths use the proxy-selected numerical λ unchanged. Each target constructs one posterior at this transferred value; target-side sweeps are run only to obtain reference optima and predictive-transfer-gap metrics. We report 10 seeds; numerical summaries use sample standard deviations. We select by validation NLL in the primary experiment and retain the fixed-checkpoint plug-in Laplace evidence score as a diagnostic of criterion dependence. The implementation is verified against hand-built references: full-curvature posterior precisions agree to relative Frobenius error $\sim 10^{-8}$.

C.1 Curvature approximations

The GGN Hessian H can be stored and inverted in several ways, offering different accuracy–compute trade-offs. The *full* approximation stores the entire $P \times P$ matrix ($O(P^2)$ memory, $O(P^3)$ inversion), giving the exact linearized posterior. The *diagonal* approximation retains only the diagonal of H ($O(P)$ memory), discarding all off-diagonal correlations between parameters. The *Kronecker-factored* (KFAC) approximation (Martens & Grosse, 2015; Ritter et al., 2018) stores a Kronecker product of two factors for each layer block, capturing within-layer structure at reduced computational cost. In all three cases the prefix *last-layer* (LL) means the approximation is applied only to the parameters of the final linear layer, with earlier layers frozen at their MAP values; this reduces P from the full parameter count to the output-layer dimension alone.

C.2 The geometric mechanism

We isolate the mechanism on two moons, selecting λ on validation data and then measuring kernel and uncertainty agreement on a disjoint held-out pool. Relative kernel drift from width 128 to 4096 falls from 20.4 ± 3.9 under standard practice to 0.64 ± 0.11 under σTransfer; the selected precision moves 2.9 ± 0.1 bits versus -0.5 ± 0.8 bits, and held-out uncertainty-rank agreement rises from 0.22 ± 0.05 to 0.69 ± 0.04. The geometric correction therefore changes not only the scale of the prior kernel but the posterior quantities used downstream.

On MNIST, linear CKA (Kornblith et al., 2019) is about 0.999 for every method, indicating that in this last-layer diagnostic the μP width-dependent head scale stabilizes kernel magnitude rather

than changing its shape. Kernel stability alone does not imply posterior stability: the exact last-layer posteriors of Section 6 test whether it survives conditioning.

C.3 Regression under scaffold shift

The regression benchmark is a three-hidden-layer tanh MLP on binary Morgan fingerprints (Rogers & Hahn, 2010) ($d = 2048$, radius 2, computed with RDKit), with a bias-free readout and base width 128. Each dataset uses the OGB scaffold split (Hu et al., 2020) of its MoleculeNet original (`ogbg-molesol`, `ogbg-molfreesolv`, `ogbg-mollipo`), so training, validation, and test molecules occupy disjoint scaffolds. Inputs are left unstandardized, so a bit absent from the training scaffolds stays $0/1$ when it first appears at validation or test time; targets use training-split standardization.

We train for 20,000 iterations at batch size 100 and learning rate 10^{-3}, validating every 250 iterations. The width-512 model is the proxy and width 4096 the target. We report five seeds, shared across all four arms; as in the classification benchmark the two μP arms share trained checkpoints, so their difference isolates the prior geometry.

C.4 Additional normalization baselines

We construct empirical controls that match the average prior-kernel magnitude across widths. Their scales use trained features; σTransfer derives its scales from the architecture and parametrization.

For proxy width n_0 and training inputs D, set

$$r_n = \frac{1}{|D|} \sum_{x \in D} \|J_{\text{head},n}(x)\|_F^2, \qquad S_n^{\text{meas}} = s_0 \frac{r_{n_0}}{r_n} I,$$

where $J_{\text{head},n}$ is the Jacobian with respect to stored readout weights, including any forward multiplier. The prior covariance is $\lambda^{-1} S_n^{\text{meas}}$. We take $s_0 = 1$ for SP and the proxy's architectural covariance scale for μP. These controls use training features from both the proxy and target, without a target precision search. They match the average prior-kernel trace; this alone does not ensure kernel or posterior convergence. With one bias-free readout block, global and blockwise normalization coincide, giving one "+ empirical" control per parametrization. "SP + metric" uses the width-normalized prior under SP.

Tables 5 and 6 extend Tables 1 and 2, respectively, with the empirical controls. For matching seeds within each parametrization, priors share checkpoints, data splits, and likelihood settings. All use the same 61-point grid $\log_2 \lambda \in [-15, 15]$ with half-bit spacing; selection criteria and seed counts are given in the captions.

Table 5: Additional normalization baselines for regression: width $512 \to 4096$, full-covariance last-layer Laplace, and evidence-selected prior precision. Results are mean $\pm$ sample standard deviation over five seeds for the methods reproduced from Table 1 and ten seeds for the empirical controls. $|\Delta\text{NLL}|$ is multiplied by 10^3; other metrics follow Table 1. Bold marks the smallest reported mean precision and NLL transfer gaps; lower is better.

Method	ESOL			FreeSolv			Lipophilicity		
	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\text{NLL}\rvert$	$\text{KL}_{4096\Vert 512}$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\text{NLL}\rvert$	$\text{KL}_{4096\Vert 512}$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\text{NLL}\rvert$	$\text{KL}_{4096\Vert 512}$
SP + isotropic	$1.5_{\pm 0.4}$	$41.8_{\pm 30.7}$	$0.208_{\pm 0.073}$	$5.0_{\pm 2.2}$	$452.8_{\pm 429.5}$	$0.079_{\pm 0.055}$	$4.0_{\pm 2.7}$	$140.2_{\pm 119.7}$	$0.173_{\pm 0.027}$
SP + metric	$1.5_{\pm 0.5}$	$8.40_{\pm 6.59}$	$0.208_{\pm 0.071}$	$1.2_{\pm 1.4}$	$9.65_{\pm 10.95}$	$0.079_{\pm 0.055}$	$2.1_{\pm 1.2}$	$3.39_{\pm 1.93}$	$0.173_{\pm 0.028}$
μP + isotropic	$3.0_{\pm 1.4}$	$0.947_{\pm 0.566}$	$0.167_{\pm 0.087}$	$3.1_{\pm 0.5}$	$1.564_{\pm 0.825}$	$0.008_{\pm 0.004}$	$3.2_{\pm 1.0}$	$0.094_{\pm 0.024}$	$0.057_{\pm 0.058}$
σTransfer **(ours)**	$\mathbf{0.9}_{\pm 0.8}$	$\mathbf{0.548}_{\pm 0.412}$	$0.167_{\pm 0.087}$	$\mathbf{0.5}_{\pm 0.4}$	$\mathbf{0.461}_{\pm 0.430}$	$0.008_{\pm 0.004}$	$\mathbf{0.4}_{\pm 0.2}$	$\mathbf{0.033}_{\pm 0.019}$	$0.057_{\pm 0.058}$
SP + empirical	$1.3_{\pm 0.7}$	$17.6_{\pm 1.65}$	$0.250_{\pm 0.27}$	$3.0_{\pm 2.2}$	$35.7_{\pm 8.2}$	$0.10_{\pm 0.22}$	$2.8_{\pm 2.8}$	$2.5_{\pm 2.0}$	$0.200_{\pm 0.26}$
μP + empirical	$1.1_{\pm 0.9}$	$0.650_{\pm 0.412}$	$0.200_{\pm 0.26}$	$0.7_{\pm 0.537}$	$0.650_{\pm 0.475}$	$0.05_{\pm 0.2}$	$0.5_{\pm 0.782}$	$0.050_{\pm 0.16}$	$0.10_{\pm 0.22}$

σTransfer has the smallest reported mean NLL transfer gap across all regression and classification settings, tying μP + empirical on MNIST LL-diagonal and FMNIST LL-full. These comparisons concern mean transfer gaps, not statistical significance or absolute predictive quality.

Table 6: Additional normalization baselines for image classification: width $128 \to 4096$, validation-NLL-selected prior precision, and full, diagonal, and KFAC last-layer Laplace. Results are mean $\pm$ sample standard deviation over ten paired seeds. $|\Delta\mathrm{NLL}|$ is multiplied by 10^2; other metrics follow Table 2. Bold marks the smallest reported mean in each column; lower is better.

Method	MNIST						FMNIST					
	LL full		LL diag		LL KFAC		LL full		LL diag		LL KFAC	
	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$
SP + isotropic	$1.75_{\pm 0.26}$	$9.8_{\pm 4.6}$	$0.45_{\pm 0.28}$	$0.8_{\pm 0.9}$	$1.50_{\pm 0.24}$	$6.9_{\pm 3.2}$	$1.55_{\pm 0.28}$	$9.0_{\pm 4.1}$	$\mathbf{0.25}_{\pm 0.35}$	$0.9_{\pm 1.2}$	$1.85_{\pm 0.24}$	$13.2_{\pm 3.2}$
SP + metric	$3.25_{\pm 0.26}$	$9.6_{\pm 2.1}$	$4.55_{\pm 0.28}$	$14.8_{\pm 1.9}$	$3.50_{\pm 0.24}$	$10.7_{\pm 1.9}$	$3.45_{\pm 0.28}$	$39.0_{\pm 6.3}$	$5.25_{\pm 0.35}$	$67.2_{\pm 7.3}$	$3.15_{\pm 0.24}$	$33.3_{\pm 5.5}$
μP + isotropic	$4.75_{\pm 0.26}$	$15.7_{\pm 3.0}$	$5.30_{\pm 0.26}$	$20.6_{\pm 3.2}$	$4.75_{\pm 0.26}$	$16.5_{\pm 2.7}$	$4.60_{\pm 0.21}$	$34.9_{\pm 2.8}$	$4.55_{\pm 0.28}$	$47.2_{\pm 5.0}$	$4.45_{\pm 0.28}$	$30.5_{\pm 2.6}$
σTransfer **(ours)**	$\mathbf{0.25}_{\pm 0.26}$	$\mathbf{0.2}_{\pm 0.3}$	$\mathbf{0.30}_{\pm 0.26}$	$\mathbf{0.1}_{\pm 0.2}$	$\mathbf{0.25}_{\pm 0.26}$	$\mathbf{0.2}_{\pm 0.3}$	$\mathbf{0.40}_{\pm 0.21}$	$\mathbf{0.4}_{\pm 0.4}$	$0.45_{\pm 0.28}$	$\mathbf{0.6}_{\pm 0.6}$	$\mathbf{0.55}_{\pm 0.28}$	$\mathbf{0.3}_{\pm 0.3}$
SP + empirical	$1.35_{\pm 0.34}$	$7.50_{\pm 3.3}$	$\mathbf{0.3}_{\pm 0.35}$	$0.5_{\pm 0.5}$	$0.50_{\pm 0.41}$	$1.7_{\pm 0.4}$	$1.60_{\pm 0.32}$	$10.1_{\pm 4.1}$	$0.45_{\pm 0.28}$	$0.75_{\pm 0.36}$	$2.05_{\pm 0.37}$	$14.8_{\pm 4.0}$
μP + empirical	$0.35_{\pm 0.34}$	$0.65_{\pm 0.42}$	$0.35_{\pm 0.34}$	$\mathbf{0.1}_{\pm 0.3}$	$0.4_{\pm 0.211}$	$0.35_{\pm 0.3}$	$0.6_{\pm 0.3}$	$\mathbf{0.4}_{\pm 0.5}$	$0.60_{\pm 0.32}$	$1.10_{\pm 0.8}$	$0.65_{\pm 0.34}$	$0.6_{\pm 0.4}$

C.5 Additional classification datasets

We evaluate prior-precision transfer on PenDigits (Alpaydin & Alimoglu, 1996) and Letter Recognition (Slate, 1991), using two-hidden-layer ReLU MLPs with proxy width 32 and target width 128. For each dataset, we use 5,000 training and 1,000 validation examples and standardize inputs using training-set statistics. PenDigits retains its provided test split; Letter uses its final 4,000 examples for testing. The μP models are trained with MuSGD for 500 epochs, batch size 128, momentum 0.9, and an initial learning rate of 0.05 with cosine decay.

We use full-GGN Laplace over all parameters, with the common-logit head direction held fixed. At each trained checkpoint, we select the prior precision by maximizing the fixed-checkpoint Laplace log-evidence. The proxy-selected precision is then applied unchanged to the target. Table 3 reports the resulting target test-NLL change relative to target-side evidence selection. Timings include the complete Laplace fit and precision search, excluding network training.

These all-parameter evidence searches extend beyond the convergence guarantee of Corollary 5. Empirically, proxy-side selection gives approximately $22\times$ and $32\times$ speedups on PenDigits and Letter, respectively, with target test-NLL changes of -0.007 and 0.078 relative to target-side evidence selection.

D Evidence-selected transfer on MNIST/FMNIST

Table 7 repeats the main-text comparison with λ chosen by the fixed-checkpoint plug-in Laplace evidence score instead of validation NLL. The selection gaps tell the same story, but the NLL column must be read with care: at width 4096 the target's own plug-in score optimum sits away from the test optimum, so a transferred precision can attain a *lower* test NLL than the target's own selection. This is criterion miscalibration, not a transfer gain, which is why validation NLL is the primary criterion here.

Table 7: Evidence-selected prior transfer (proxy $128 \to$ target 4096, 10 seeds, exact implementations; per-seed absolute values, $|\Delta\mathrm{NLL}|$ scaled by 10^2; notation as in Table 2).

Method	MNIST						FMNIST					
	LL full		LL diag		LL KFAC		LL full		LL diag		LL KFAC	
	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$	$\lvert\Delta\lambda\rvert$	$\lvert\Delta\mathrm{NLL}\rvert$
SP + isotropic	$1.15_{\pm .24}$	$21.8_{\pm 5.2}$	$2.45_{\pm .16}$	$69.2_{\pm 5.4}$	—	—	$1.75_{\pm .26}$	$35.0_{\pm 5.6}$	$3.10_{\pm .32}$	$82.6_{\pm 7.3}$	—	—
SP + metric	$6.15_{\pm .24}$	$133.5_{\pm 2.1}$	$7.45_{\pm .16}$	$144.1_{\pm 3.2}$	—	—	$6.75_{\pm .26}$	$105.2_{\pm 4.1}$	$8.10_{\pm .32}$	$75.7_{\pm 6.2}$	—	—
μP + isotropic	$5.75_{\pm .35}$	$58.3_{\pm 5.0}$	$5.60_{\pm .32}$	$68.2_{\pm 7.1}$	—	—	$5.30_{\pm .26}$	$34.2_{\pm 3.3}$	$4.70_{\pm .26}$	$7.9_{\pm 4.6}$	—	—
σTransfer **(ours)**	$\mathbf{0.75}_{\pm .35}$	$\mathbf{17.4}_{\pm 8.1}$	$\mathbf{0.60}_{\pm .32}$	$\mathbf{17.8}_{\pm 9.3}$	—	—	$\mathbf{0.30}_{\pm .26}$	$\mathbf{6.2}_{\pm 5.4}$	$\mathbf{0.30}_{\pm .26}$	$\mathbf{5.8}_{\pm 5.0}$	—	—

E Additional OOD-transfer results

Figure 6 shows the time-to-quality tradeoff, and Table 8 gives the width-4096 OOD transfer comparison summarized in the main text. Table 9 expands this across all measured target widths. Selection and evaluation use the same disjoint datasets. The signed regret can be slightly negative because each target selects on FMNIST but is evaluated on unseen EMNIST and on

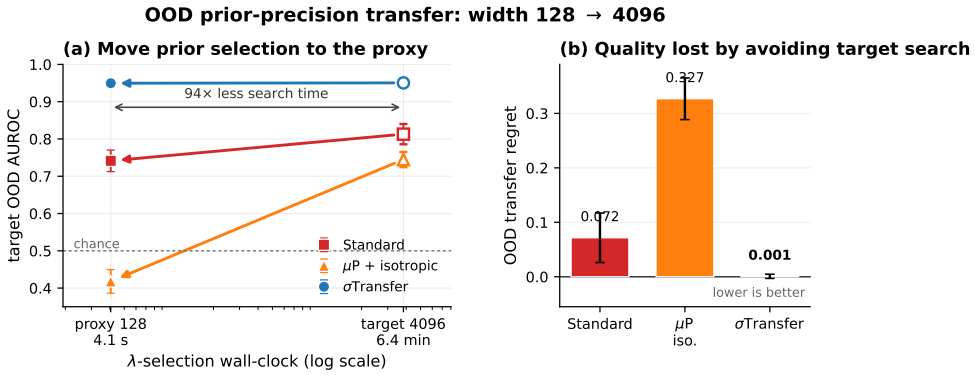

Figure 6: **Time-to-quality tradeoff for OOD precision transfer** ($128 \to 4096$, all-parameter diagonal GGN). Filled markers use the proxy-selected precision; open markers use target-side selection. Arrows point toward the cheaper proxy search. The reported A100 times compare proxy-side and target-side selection procedures (4.1 seconds versus 6.4 minutes, $94\times$). Zero-shot transfer removes the target sweep but retains one target posterior at the transferred precision. The right panel reports paired transfer regret over 10 seeds. Only σTransfer removes the target search without measurable OOD-quality loss.

KMNIST (Clanuwat et al., 2018); in those cases the proxy-selected precision generalizes marginally better to the evaluation distributions.

Method	**AUROC using $\lambda^\star_{128}$**	**AUROC using $\lambda^\star_{4096}$**	**Transfer regret ↓**
SP	0.741 ± 0.029	0.813 ± 0.027	0.072 ± 0.046
μP + isotropic	0.418 ± 0.032	0.745 ± 0.020	0.327 ± 0.038
σTransfer (ours)	$\mathbf{0.950 \pm 0.003}$	$\mathbf{0.950 \pm 0.004}$	$\mathbf{0.001 \pm 0.004}$

Table 8: **OOD-aware precision transfer at width** 4096 (10 seeds; mean ± sample s.d.). The precision is selected on MNIST-validation versus FMNIST and evaluated on disjoint MNIST-test versus EMNIST+KMNIST. Transfer regret is $\mathrm{AUROC}(\lambda^\star_{4096}) - \mathrm{AUROC}(\lambda^\star_{128})$; lower is better.

Method	**Target width n**	**AUROC using $\lambda^\star_{128}$**	**AUROC using $\lambda^\star_n$**	**Regret ↓**
SP	128	0.957 ± 0.005	0.957 ± 0.005	0.0000 ± 0.0000
	512	0.909 ± 0.008	0.909 ± 0.008	0.0000 ± 0.0000
	2048	0.807 ± 0.012	0.802 ± 0.013	-0.0044 ± 0.0193
	4096	0.741 ± 0.029	0.813 ± 0.027	0.0716 ± 0.0455
μP + isotropic	128	0.957 ± 0.004	0.957 ± 0.004	0.0000 ± 0.0000
	512	0.868 ± 0.024	0.868 ± 0.024	0.0000 ± 0.0000
	2048	0.445 ± 0.025	0.743 ± 0.022	0.2977 ± 0.0280
	4096	0.418 ± 0.032	0.745 ± 0.020	0.3269 ± 0.0383
σTransfer	128	0.962 ± 0.003	0.962 ± 0.003	0.0000 ± 0.0000
	512	0.958 ± 0.004	0.958 ± 0.004	0.0003 ± 0.0005
	2048	0.952 ± 0.004	0.951 ± 0.004	-0.0006 ± 0.0012
	4096	$\mathbf{0.950 \pm 0.003}$	$\mathbf{0.950 \pm 0.004}$	$\mathbf{0.0008 \pm 0.0039}$

Table 9: **OOD-aware precision transfer across width** (10 seeds; mean ± sample s.d.). Both AUROC columns evaluate target width n; $\lambda^\star_{128}$ is selected on the proxy and $\lambda^\star_n$ on that target. Regret is $\mathrm{AUROC}(\lambda^\star_n) - \mathrm{AUROC}(\lambda^\star_{128})$ on the held-out evaluation distributions. The main text reports $n = 4096$.

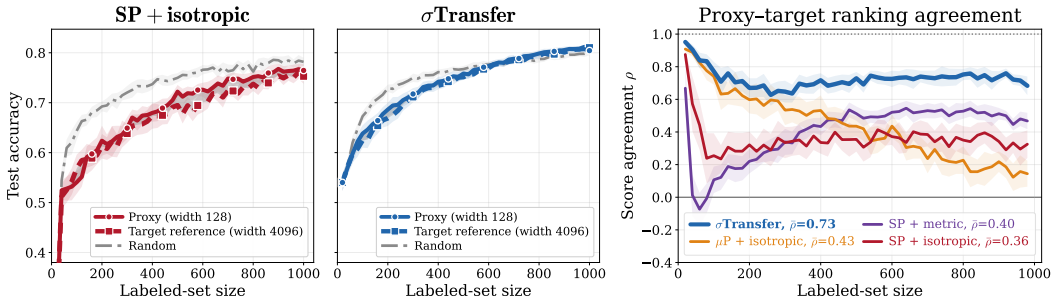

Figure 7: **Sequential active learning on Fashion-MNIST** (20 seeds; panels and encoding as in Figure 3). The agreement ordering replicates with a compressed baseline spread ($\bar{\rho} = 0.73$ for σTransfer against 0.36–0.43). On this task the acquisition curves do not separate from random, so we claim decision transfer on both datasets and downstream utility on MNIST only.

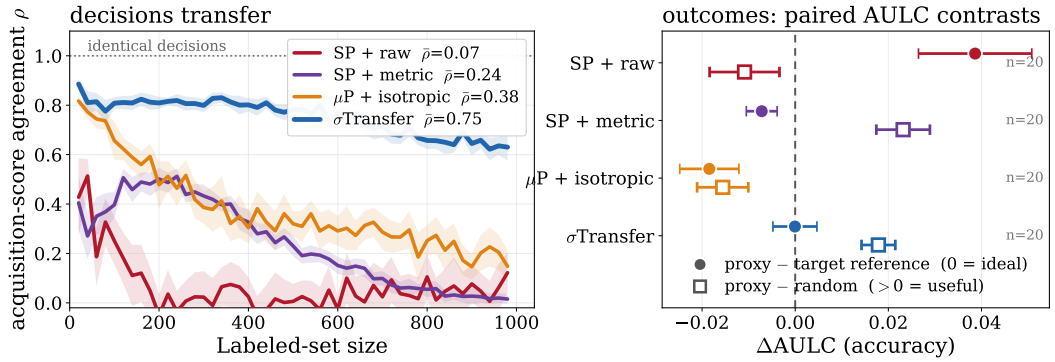

Figure 8: **Sequential active learning: decisions and outcomes** (10-class MNIST, last-layer diagonal posterior at fixed $\log_2 \lambda = 0$, width $128 \to 4096$, 20 seeds). *Left:* per-round Spearman correlation between the width-128 and width-4096 acquisition scores (band: 95% CI of the mean). σTransfer holds $\rho \approx 0.6$–0.8 for the whole trajectory while every baseline decays toward zero, and the trajectory means are $\bar{\rho} = 0.75 \pm 0.01$ (σTransfer), 0.38 ± 0.02 (μP + isotropic), 0.24 ± 0.01 (SP + metric) and 0.07 ± 0.03 (SP + isotropic). *Right:* paired area-under-the-learning-curve contrasts (per seed, then averaged; bars: 95% CI). Filled: proxy − target reference, so 0 means the proxy acquires as well as the target's own posterior; open: proxy − random. Only σTransfer is simultaneously faithful (its fidelity interval is the only one covering zero, -0.0001, $[-0.005, +0.005]$) and useful ($+0.018$, $[+0.014, +0.022]$); SP + isotropic's apparent $+0.039$ fidelity reflects a degenerate target reference that itself falls 0.050 below random. Full learning curves: Figure 10; Fashion-MNIST: Figures 9 and 11.

F Sequential Active Learning Details

For sequential acquisition on MNIST and Fashion-MNIST, we use three-hidden-layer ReLU MLPs of widths 128 and 4096. Each run starts with 20 labelled examples and acquires 20 per round until 1,000 are labelled, giving 50 evaluations including the initial fit. The unlabelled pool initially contains 2,000 examples. Each round scores 500 randomly sampled remaining candidates using EPIG, 200 reference inputs and 256 posterior samples. We use 500 validation and 1,000 test examples. Initial labelled, pool and validation sets are disjoint, class-balanced subsets of the official training split; reference and test inputs are disjoint, class-balanced subsets of the official test split. EPIG uses only the reference inputs.

Acquisition uses last-layer diagonal Laplace, including the head bias, at fixed $\lambda = 1$. We warm-start the weights each round and use full-batch Adam (SP) or MuAdam (μP), both at learning rate 10^{-3}, for at most 500 updates. Validation NLL is checked every 10 updates, with patience 150 updates, and the best checkpoint is restored. The strategies share initial splits across 20 seeds.

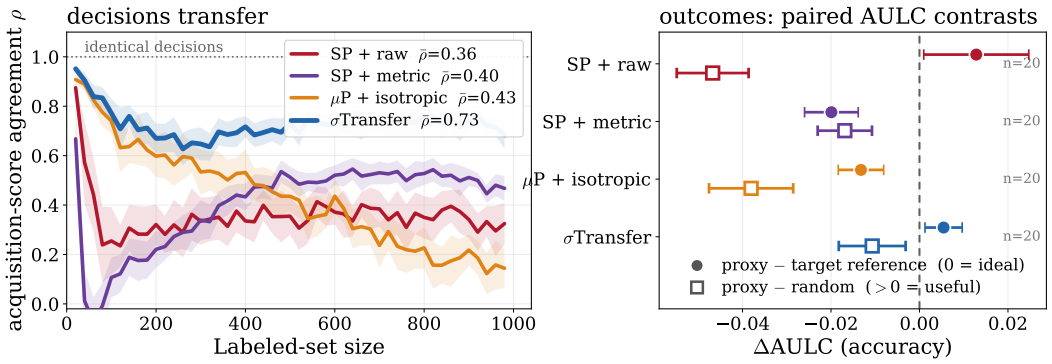

Figure 9: **Sequential active learning on Fashion-MNIST: decisions and outcomes** (20 seeds; same protocol and panels as Figure 8). The agreement ordering replicates ($\bar{\rho} = 0.73 \pm 0.02$ for σTransfer against 0.36–0.43; see also Figure 7), with a compressed baseline spread. The outcome panel is negative for every method: on this task EPIG acquisition does not beat random over the trajectory (a cold-start effect that endpoint accuracy alone conceals). At 20 seeds every method's utility interval excludes zero; σTransfer is the least affected (-0.011, $[-0.018, -0.003]$) against -0.047 $[-0.055, -0.039]$ for SP + isotropic. We therefore claim decision transfer on both datasets and downstream utility on MNIST only.

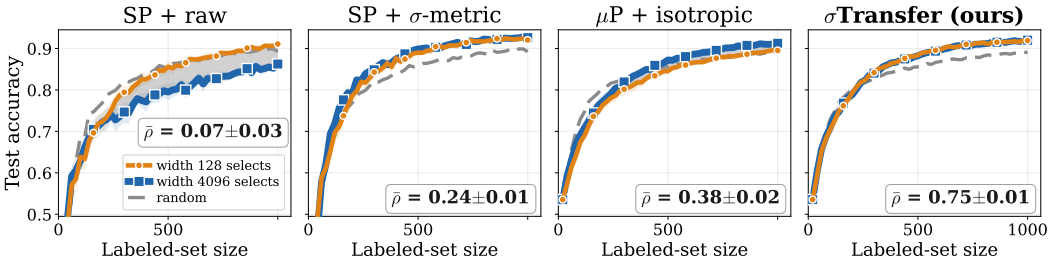

Figure 10: **Sequential active learning on 10-class MNIST** (last-layer diagonal posterior, 20 seeds; bands: 95% CI of the mean). Every curve is averaged over the same complete-trajectory seed set. One panel per method, titled above each panel. Orange selects each acquisition batch with the width-128 posterior and blue with the width-4096 posterior; both are then evaluated on the same width-4096 model, and the grey band between them is the acquisition-transfer gap. The thin orange line is drawn over the thicker blue one, so a method in which the proxy reproduces the target's decisions shows orange tracing blue with no blue visible beneath. Markers (circles for width 128, squares for width 4096) are stamped every seventh acquisition round and carry no information beyond series identity. $\bar{\rho}$ is the mean rank correlation between the two posteriors' acquisition scores over the trajectory, the direct measure of how closely the small model reproduces the large model's decisions. Accuracy on this task is largely insensitive to acquisition quality, but the decisions are not: $\bar{\rho}$ rises from 0.07 (SP + isotropic) through 0.24 (SP + metric) and 0.38 (μP + isotropic) to 0.75 under σTransfer, so neither the parametrization nor the prior geometry alone suffices.

Figure 10 gives the full per-method learning curves for the 50-round sequential active-learning experiment summarised in Section 6.2; Table 11 gives the endpoint statistics at $n = 1000$ labelled points, and Figure 12 and Table 10 the one-shot EPIG acquisition-agreement metrics for both datasets.

G Decision transfer for language models

We use a width-128 model to decide whether a width-2048 model should answer a prompt. The small model computes the uncertainty score, and the large model supplies the prediction.

Models and pretraining. We train decoder-only Transformers at hidden widths $128, 256, 512, 1024, 2048$. All models have 8 layers, 8 attention heads, context length 512,

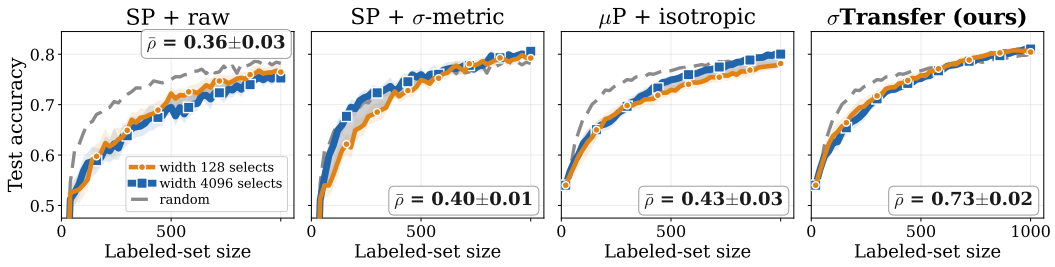

Figure 11: **Sequential active learning on Fashion-MNIST** (last-layer diagonal posterior, 20 seeds; bands: 95% CI of the mean; same protocol and encoding as Figure 10). The decision-agreement ordering matches MNIST, with $\bar{\rho} = 0.73$ under σTransfer, compared with 0.36–0.43 for the three baselines. SP + isotropic has higher agreement on Fashion-MNIST than on MNIST (0.36 versus 0.07).

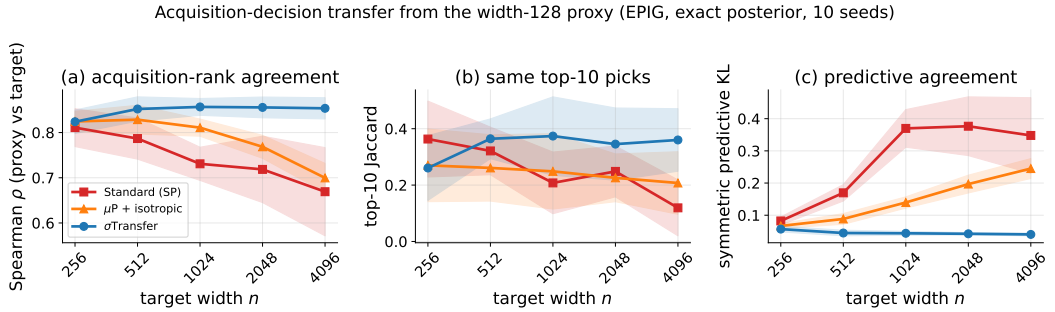

Figure 12: **Direct acquisition transfer** (EPIG, exact last-layer posterior, validation-NLL-selected transferred precision, 10 seeds; bands 95% CI of the mean). **(a)** Proxy–target acquisition-rank agreement. **(b)** Top-10 Jaccard overlap. **(c)** Symmetric predictive KL. The target posterior is used only to measure agreement; the proxy choices do not require it.

and the Pythia padded vocabulary of 50,304 entries (Biderman et al., 2023). They are trained on the same stream of 10^9 FineWeb-Edu tokens (Penedo et al., 2024). We use one pretraining seed for each width and parametrization.

The μP models use `MuReadout`, `set_base_shapes`, and MuAdamW, with base width 128 and base learning rate 0.01. SP uses `nn.Linear` and AdamW at learning rate 0.0003. These rates are fixed across widths. Both use $\beta = (0.9, 0.95)$, weight decay 0.1, 2% warmup, cosine decay to 10% of the initial rate, and gradient-norm clipping at 1, without dropout. Each update accumulates four batches of 32 length-512 sequences.

Classification heads. We freeze the backbone and fit a linear classifier to the final-layer-normalization output h at the last real token. Its logits are $W(h/c) + b$, where c^2 is the mean squared feature norm on the training inputs. We minimize summed cross-entropy plus $(\|W\|_F^2 + \|b\|^2)/2$ using L-BFGS in float64. For each backbone and task, prior comparisons share the fitted head, which is held fixed throughout the precision sweep.

We apply last-layer Laplace with width-normalized or isotropic priors. For raw features of dimension d, their weight covariance scales are $d^{-1}I$ and I, respectively. In the fitted head's coordinates, these become $S_W = (c^2/d)I$ and $S_W = c^2 I$; both use $S_b = I$. These definitions apply to both backbone parametrizations; μP with the width-normalized prior gives σTransfer. We use full-GGN curvature in float64 and evaluate prior precision on a 61-point $\log_2 \lambda$ grid. For precision selection and OOD gating, scores use 200 Monte Carlo samples of the Gaussian logits, with common random numbers.

Active learning on AG News. Figure 5 starts with 50, 500 or 5,000 labelled examples and acquires 250 labels one at a time from a common pool of 512 examples, disjoint from the initial labelled sets. EPIG uses 256 validation reference inputs and 2,048 posterior samples. We refit the heads after each

Table 10: **EPIG acquisition transfer on MNIST and FMNIST** (proxy 128 → target 4096, validation-NLL-selected λ, exact last-layer posterior; MNIST 10 seeds, FMNIST 5 seeds). Agreement between the proxy and target posteriors, with the target evaluated at the transferred precision: Spearman rank correlation of EPIG scores, top-10 acquisition-set Jaccard, and symmetric predictive KL (lower better). **Bold** = best per column.

Method	MNIST			FMNIST		
	Spearman ρ	Jacc_{10}	KL_{sym}	Spearman ρ	Jacc_{10}	KL_{sym}
SP + isotropic	$0.669_{\pm.103}$	$0.119_{\pm.105}$	$0.348_{\pm.124}$	$0.589_{\pm.137}$	$0.118_{\pm.099}$	$0.944_{\pm.193}$
μP + isotropic	$0.700_{\pm.034}$	$0.207_{\pm.117}$	$0.245_{\pm.032}$	$0.629_{\pm.063}$	$0.134_{\pm.124}$	$0.418_{\pm.060}$
σTransfer **(ours)**	$\mathbf{0.854_{\pm.025}}$	$\mathbf{0.360_{\pm.117}}$	$\mathbf{0.040_{\pm.004}}$	$\mathbf{0.750_{\pm.066}}$	$\mathbf{0.198_{\pm.104}}$	$\mathbf{0.167_{\pm.021}}$

		Test acc.	Test NLL	Regret
SP	random	0.901 ± 0.016	0.976 ± 0.141	$+0.009$
SP	proxy EPIG	0.920 ± 0.012	0.774 ± 0.171	-0.011
SP	oracle EPIG	0.910 ± 0.008	0.931 ± 0.111	—
μP + isotropic	random	0.890 ± 0.009	0.920 ± 0.120	$+0.028$
μP + isotropic	proxy EPIG	0.889 ± 0.021	0.909 ± 0.298	$+0.029$
μP + isotropic	oracle EPIG	0.918 ± 0.004	0.694 ± 0.107	—
σTransfer	random	0.888 ± 0.006	0.953 ± 0.152	$+0.037$
σTransfer	proxy EPIG	$\mathbf{0.923 \pm 0.011}$	$\mathbf{0.700 \pm 0.109}$	$\mathbf{+0.003}$
σTransfer	oracle EPIG	0.926 ± 0.010	0.661 ± 0.153	—

Table 11: **Sequential active-learning utility at** $n = 1000$ (10-class MNIST, last-layer diagonal posterior at fixed $\log_2 \lambda = 0$, width $128 \to 4096$; mean $\pm$ sample s.d.). Regret = oracle-EPIG accuracy − method accuracy. Round wall-clock is dominated by retraining the width-4096 model and is the same for every strategy, so this experiment is a decision-transfer result, not a speed result.

acquisition and use full-GGN Laplace at fixed $\lambda = 1$. After the final acquisition, we evaluate the target's MAP predictions on 7,600 test examples. The σTransfer arm uses the width-normalized prior defined above. The feature normalization c is fixed from the full training feature cache throughout acquisition.

OOD data. For each classification task, we take OOD prompts from two other datasets, denoted A and B. We score them using the ID task's classification head and count abstention as the correct action. Sentiment-to-sentiment pairs are excluded. Table 12 gives the source assignments and test-set sizes.

Table 12: OOD source assignments and test-set sizes.

ID tasks (test-set sizes)	Source A (pool size)	Source B (pool size)
AG News (7600), DBpedia (8000)	SST-2 (872)	TREC (500)
IMDB (8000), Rotten Tomatoes (1066), Yelp Polarity (8000), SST-2 (872)	AG News (7600)	TREC (500)
Emotion (2000), SNLI (8000), Yahoo (8000), TREC (500)	AG News (7600)	SST-2 (872)

Dataset sources are Zhang et al. (2015) for AG News, DBpedia, Yahoo Answers, and Yelp Polarity; Socher et al. (2013) for SST-2; Li & Roth (2002) for TREC; Maas et al. (2011) for IMDB; Pang & Lee (2005) for Rotten Tomatoes; Saravia et al. (2018) for Emotion; and Bowman et al. (2015) for SNLI.

Decision rule and calibration. We choose the proxy's prior precision using ID validation data and OOD source A. The selected λ_{ood} maximizes the AUROC of BALD (Houlsby et al., 2011) for distinguishing ID prompts from source A. At this precision, we choose a threshold τ that maximizes Youden's J, defined as ID retention plus OOD recall minus one.

The large model answers a prompt when the proxy's BALD score is at most τ; otherwise, the system abstains. Answered prompts use the large model's MAP predictions. We evaluate on ID test data and OOD source B, which was not used for calibration. We then exchange A and B and average the two results.

Metrics and results. *OOD recall* is the fraction of OOD prompts rejected, and *ID retention* is the fraction of ID prompts answered. *Answered accuracy* is the accuracy among answered prompts, counting answered OOD prompts as errors. *Abstention agreement* is the fraction of the proxy's rejected prompts that are also rejected by the separately calibrated target model.

Table 13 gives the results for each task. The final row reports the mean and population standard deviation across the ten tasks, after averaging the two source orderings within each task.

Table 13: Proxy-based abstention results for width $128 \to 2048$. Each entry reports OOD recall / ID retention / answered accuracy / abstention agreement, averaged over the two source orderings.

Dataset	SP + isotropic	σTransfer
AG News	0.846 / 0.969 / 0.872 / 0.784	0.895 / 0.963 / 0.866 / 0.702
DBpedia	0.981 / 0.999 / 0.899 / 0.985	0.987 / 0.998 / 0.869 / 0.966
Emotion	1.000 / 0.999 / 0.485 / 0.998	0.998 / 0.999 / 0.479 / 0.998
IMDB	0.605 / 0.826 / 0.603 / 0.846	0.725 / 0.837 / 0.643 / 0.873
Rotten Tomatoes	0.875 / 0.977 / 0.477 / 0.955	0.937 / 0.984 / 0.547 / 0.969
SNLI	0.958 / 0.992 / 0.496 / 0.953	0.974 / 0.986 / 0.512 / 0.921
SST-2	0.836 / 0.968 / 0.456 / 0.950	0.732 / 0.939 / 0.431 / 0.912
TREC	0.998 / 1.000 / 0.696 / 1.000	0.999 / 0.996 / 0.655 / 0.998
Yahoo	0.676 / 0.976 / 0.468 / 0.774	0.662 / 0.968 / 0.454 / 0.816
Yelp Polarity	0.775 / 0.927 / 0.713 / 0.729	0.854 / 0.926 / 0.760 / 0.734
Mean $\pm$ SD	$0.855_{\pm.13}$ / $0.963_{\pm.05}$ / $0.616_{\pm.16}$ / $0.898_{\pm.10}$	$0.876_{\pm.12}$ / $0.959_{\pm.05}$ / $0.621_{\pm.16}$ / $0.889_{\pm.10}$

Comparison with target decisions. For evaluation, we also apply the same calibration procedure to the target's own posterior. This gives mean OOD recall of $0.972/0.980$, ID retention of $0.975/0.978$, and answered accuracy of $0.693/0.694$ for SP/σTransfer. Using the proxy instead reduces OOD recall by approximately 10–12 percentage points, ID retention by 1–2 points, and answered accuracy by 7–8 points. Deployment requires only the proxy posterior and target MAP predictions.

Additional comparisons. The target's mean NLL on answered ID prompts differs from ungated prediction by $-0.004/-0.005$ nats for SP/σTransfer. Replacing BALD with one minus the maximum probability of the posterior predictive ($1 - \mathrm{MSP}$; Hendrycks & Gimpel, 2016) gives proxy OOD recall of 0.41–0.42 while retaining approximately half the ID prompts. Applying this alternative to the target gives OOD recall of 0.84–0.85 and answered accuracy of 0.64–0.66.

Selecting λ using ID evidence alone gives OOD recall of 0.85–0.86; selecting τ still uses source A. BALD magnitudes vary across widths. In the experiment reported here, the threshold is always applied to the proxy's scores.

H Prior-precision transfer in language models

Both tables use the fixed-checkpoint plug-in evidence-score criterion on ID training data at the proxy, the shared half-bit grid, and the target's own evidence search only as an evaluation reference. In zero-shot deployment, the target constructs one posterior at the transferred proxy precision. ΔNLL is transferred-minus-own test NLL at the target; sweep arrays use 200 MC samples.

H.1 Comparison with the diagonal neural g-prior

We test whether the diagonal neural g-prior of Antorán et al. (2023) supports prior-precision transfer in our fixed-checkpoint Laplace setting. Its precision is αD, where $D = \mathrm{diag}(H)$ and H is the summed last-layer GGN. We compute D from training data at each width and transfer α from width 128 to 2048. The fitted classification heads remain fixed throughout the precision search.

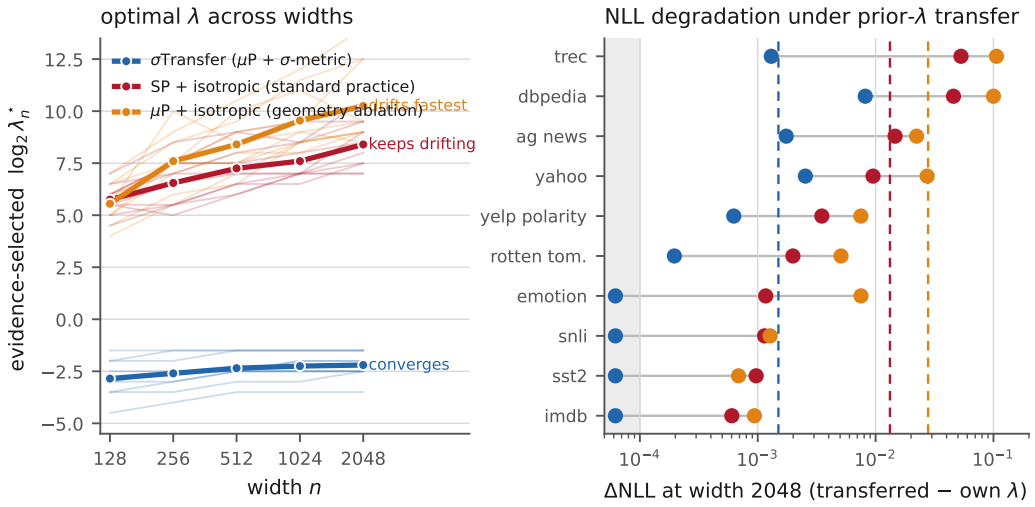

Figure 13: Figure 4 with the geometry ablation added: the same μP checkpoints with an isotropic prior drift faster than SP ($5.5 \to 10.2$ bits mean optimum, 4.7 ± 1.3 bits at $128 \to 2048$) and produce the largest mean target-NLL degradation (0.028, reaching 0.1 on TREC).

Table 14: Our language models, width $128 \to 2048$ (Section 6.3). λ values are $\log_2$.

	σTransfer			SP + isotropic		
Dataset	$\lambda^\star_{128}$	$\lambda^\star_{2048}$	ΔNLL	$\lambda^\star_{128}$	$\lambda^\star_{2048}$	ΔNLL
AG News	−3.5	−2.5	+0.0017	+4.5	+7.0	+0.0146
DBpedia	−4.5	−3.5	+0.0082	+5.0	+7.5	+0.0458
Emotion	−3.0	−2.5	−0.0001	+6.0	+8.5	+0.0012
IMDB	−2.0	−1.5	−0.0000	+6.5	+9.5	+0.0006
Rotten Tomatoes	−2.0	−1.5	+0.0002	+7.0	+10.5	+0.0020
SNLI	−1.5	−1.5	+0.0000	+5.5	+7.5	+0.0011
SST-2	−3.0	−2.5	+0.0000	+6.0	+9.0	+0.0010
TREC	−2.5	−2.0	+0.0013	+5.5	+8.0	+0.0530
Yahoo	−3.5	−2.5	+0.0025	+5.5	+7.0	+0.0095
Yelp Polarity	−3.0	−2.0	+0.0006	+6.0	+9.5	+0.0035
Mean gap (bits) / ΔNLL	0.65 ± 0.34 / +0.0014			2.65 ± 0.63 / +0.0132		

The mean absolute precision gap is 3.95 bits for SP and 3.50 bits for μP, with mean target test-NLL increases of 0.0367 and 0.0229, respectively, when transferring the proxy precision (Table 16). Thus this normalization leaves substantial precision drift in our fixed-checkpoint experiment.

I External validation on public u-μP/SP pairs: details

Checkpoints. We use public unit-scaled μP (u-μP) and SP checkpoints from Blake et al. (2025) at approximately 1B (d=2048, 16 layers) and 7B (d=4096, 32 layers), pretrained on 300B SlimPajama tokens. Width and depth increase together. We evaluate one checkpoint per size and parametrization; SP training hyperparameters were not retuned for each size. Checkpoint revisions and inference code are pinned. The extracted final-normalization representations reproduce the released logits for both 1B models. Last-layer inference follows Appendix G.

Precision transfer. Table 17 compares precision transfer across the four combinations of parametrization and prior. The corrected evidence score adds a quadratic estimate of the gain in log posterior from locally adjusting the trained weights at each prior precision. All selected precisions are interior to the grid. The correction changes the selection gaps and method ordering:

Table 15: Public u-μP/SP pairs, 1B→7B, all ten tasks (primary cells; extends Table 17). Each model's search time is its curvature eigendecomposition plus grid sweep (u-μP family; SP within 10%); Yahoo and DBpedia run on CPU because their 41k/57k-dimensional eigenproblems exceed the GPU solver. The reported ratios compare proxy- and target-side search procedures, not end-to-end Regime-A deployment; the 7B model still constructs one posterior at the transferred precision. Timing results are available for nine tasks; SST-2 is excluded from the timing summary. Speedups are approximate ratios of the displayed, rounded times.

	u-μP + σ-metric		SP + isotropic				
Dataset	gap	ΔNLL	gap	ΔNLL	search 1B	search 7B	speedup
AG News	0.0	+0.0000	0.5	+0.0030	4 s	9 s	∼2.3×
DBpedia	0.0	+0.0000	1.0	+0.0247	40 s	3.7 h	330×
Emotion	0.5	+0.0002	1.0	+0.0030	5 s	24 s	5×
IMDB	0.0	+0.0000	0.5	−0.0016	2 s	4 s	2×
Rotten Tomatoes	0.5	+0.0002	0.5	+0.0011	2 s	3 s	2×
SNLI	0.5	+0.0000	1.5	+0.0011	3 s	6 s	2×
SST-2	0.0	+0.0000	1.0	+0.0017	—	—	—
TREC	0.0	+0.0000	0.0	+0.0000	5 s	23 s	5×
Yahoo	0.0	+0.0000	0.5	+0.0048	19 s	81 min	260×
Yelp Polarity	0.0	+0.0000	0.5	−0.0015	3 s	4 s	1×
Mean	0.15 ± 0.24	+0.0000	0.70 ± 0.42	+0.0036			

Table 16: Diagonal neural g-prior on our language models ($128 \to 2048$, ten tasks, one checkpoint seed). $|\Delta \log_2 \alpha|$ is the mean absolute proxy–target selection gap and ΔNLL is mean transferred-minus-target-local test NLL, both under the fixed-checkpoint plug-in evidence score. Counts give tasks with both proxy and target selections strictly inside the search grid. Validation-NLL gaps are omitted because only five of twenty pairs satisfy this condition.

Parametrization	$\lvert\Delta \log_2 \alpha\rvert$	ΔNLL	Evidence: interior	Val. NLL: interior
SP	3.95	+0.0367	10/10	4/10
μP	3.50	+0.0229	10/10	1/10

σTransfer and SP with an isotropic prior both have a mean gap of 0.56 bits under the corrected criterion.

Using the gate protocol in Appendix G, both public-model proxy gates retain approximately 99.8% of in-distribution inputs and reject 96% of out-of-distribution inputs on average, with similar performance on a common target.

J Technical formulation of the σTransfer prior

We derive the prior scales used in Section 4 for the MLP setting of Theorem 1 and give the implementation details.

Prior coordinates. For the zero-mean prior $\theta_n \sim \mathcal{N}(0, \lambda^{-1} S_n)$ with $\lambda > 0$, write $S_n = \mathrm{blockdiag}(s_{1,n} I, \ldots, s_{G,n} I)$, with one block per parameter tensor and $s_{g,n} > 0$. Taking $T_n = \mathrm{blockdiag}(\sqrt{s_{1,n}} I, \ldots, \sqrt{s_{G,n}} I)$ gives $S_n = T_n T_n^\top$. The prior $\phi_n \sim \mathcal{N}(0, \lambda^{-1} I)$ then induces the stated prior on $\theta_n = T_n \phi_n$. At the trained checkpoint, the Jacobians satisfy $J_{\phi,n} = J_n T_n$, so

$$K_n(x, x') = J_n(x) S_n J_n(x')^\top = J_{\phi,n}(x) J_{\phi,n}(x')^\top.$$

A common width-independent factor in S_n can be absorbed into λ; relative block constants specify the prior.

Table 17: Public 1B→7B prior-precision transfer on eight tasks excluding Yahoo and DBpedia. Precision gaps are in bits (mean ± s.d. across tasks); the 61-point grid has 0.5-bit spacing. ΔNLL is mean target test NLL at the proxy's precision minus that at the target's own precision, both selected by fixed-checkpoint plug-in evidence. The two primary comparisons cover all ten tasks in Table 15. Sweep arrays use 200 MC samples in fp16.

Cell	gap, plug-in evid. (bits)	gap, corrected evid.	ΔNLL
u-μP + σ-metric	0.19 ± 0.26	0.56 ± 0.32	$+4 \times 10^{-5}$
SP + σ-metric	0.06 ± 0.18	0.38 ± 0.23	-1×10^{-4}
u-μP + isotropic	0.88 ± 0.23	1.19 ± 0.65	$+1.1 \times 10^{-3}$
SP + isotropic (standard)	0.69 ± 0.46	0.56 ± 0.42	$+8.6 \times 10^{-4}$

MLP construction. Consider a scalar-output MLP with fixed depth L, fixed input dimension $n_0 = d_{\text{in}}$, and hidden widths $n_1 = \cdots = n_L = n$. Set $h^0(x) = x$ and

$$z_i^\ell(x) = \frac{1}{\sqrt{n_{\ell-1}}} \sum_j W_{ij}^\ell h_j^{\ell-1}(x) + b_i^\ell, \qquad h_i^\ell(x) = \mathrm{ReLU}(z_i^\ell(x)),$$
$$f_n(x) = \frac{1}{n} a^\top h^L(x) + c.$$

All quantities below are evaluated at $\hat{\theta}_n = \theta_{n,T}$, with T fixed independently of width and training satisfying Assumptions 1 and 2. In these coordinates, choose

$$S_{W^1,n} = n\, d_{\text{in}} I, \qquad S_{W^\ell,n} = nI \quad (2 \le \ell \le L),$$
$$S_{b^\ell,n} = nI, \qquad S_{a,n} = nI, \qquad S_{c,n} = 1.$$

Thus T_n has factor $\sqrt{n d_{\text{in}}}$ on the input-weight block, $\sqrt{n}$ on the remaining weight and hidden-bias blocks, and 1 on the output bias. For an internal matrix, $\widetilde{W}^\ell = W^\ell/\sqrt{n}$ is the canonical μP matrix parameter; parameter and optimizer scalings transform together under Yang & Littwin (2023, Proposition 2.2.3 and Definition 2.9.12). The prior map T_n is defined by the choice of S_n above.

Kernel derivation. For $\ell = 1, \ldots, L$, define the sensitivity $r_i^\ell(x) = n\, \partial f_n(x)/\partial z_i^\ell(x)$ and the Gram entries

$$\Sigma_{\ell,n}(x, x') = \frac{h^\ell(x)^\top h^\ell(x')}{n}, \qquad \Pi_{\ell,n}(x, x') = \frac{r^\ell(x)^\top r^\ell(x')}{n}.$$

For the input layer, set $\Sigma_{0,n}(x, x') = \Sigma_0(x, x') = d_{\text{in}}^{-1} x^\top x'$. The parameter derivatives are

$$\frac{\partial f_n}{\partial a_i} = \frac{h_i^L}{n}, \qquad \frac{\partial f_n}{\partial b_i^\ell} = \frac{r_i^\ell}{n},$$
$$\frac{\partial f_n}{\partial W_{ij}^\ell} = \frac{r_i^\ell h_j^{\ell-1}}{n\sqrt{n_{\ell-1}}}, \quad \frac{\partial f_n}{\partial c} = 1.$$

Multiplying by the prior blocks and summing gives the exact finite-width identity

$$K_n(x, x') = 1 + \Sigma_{L,n}(x, x') + \Pi_{1,n}(x, x')\big(1 + d_{\text{in}}\Sigma_0(x, x')\big) + \sum_{\ell=2}^{L} \Pi_{\ell,n}(x, x')\big(1 + \Sigma_{\ell-1,n}(x, x')\big). \tag{9}$$

Under the stated assumptions, Lemma 2 establishes joint almost-sure convergence of these Gram entries on every fixed finite input set. The kernel therefore converges by continuity of finite sums and products. Corollary 1 gives the extension to any fixed output dimension.

Stored-coordinate scales. Under an invertible linear change $\theta'_n = R_n \theta_n$, the prior and Jacobian transform as $S'_n = R_n S_n R_n^\top$ and $J'_n = J_n R_n^{-1}$, preserving $J_n S_n J_n^\top$. Absorbing the forward factors into $W^\ell_{\text{stored}} = W^\ell/\sqrt{n_{\ell-1}}$ and $a_{\text{stored}} = a/n$ gives input-weight scale n, internal-weight scales $n/n_{\ell-1}$ for $\ell \ge 2$, and readout scale $1/n$. For a readout $f = c_{\text{out}} w^\top h + c$, the weight scale is $1/(\text{fan_in}\, c_{\text{out}}^2)$. Table 18 lists the implementation scales.

Parameter block	Stored-coordinate $s_{g,n}$
Input weight	fan_out
Hidden weight	fan_out/fan_in
Hidden bias	fan_out
`MuReadout` weight	$1/(\text{fan_in}\, c_{\text{out}}^2)$
Output bias	1

Table 18: Prior-covariance scales for Linear/MLP parameters, with $c_{\text{out}} =$ `output_mult/width_mult`.

Implementation and checks. The adapter supports `nn.Linear` and `MuReadout` parameters. It supplies `laplace-torch` with precision $\lambda/s_{g,n}$ for every parameter in block g: a per-parameter vector for full or diagonal curvature, and one scalar per parameter tensor for KFAC. The structured-curvature results use undamped KFAC (`damping=False`), whose blocks are $Q_g + (\lambda/s_{g,n})I$. For `MuReadout`, `width_mult` is frozen before functional Jacobian evaluation, and extracted last-layer features include c_{out} to match the forward pass.

The prior-kernel coordinate identity above is checked numerically to machine precision. Full-curvature posterior-precision checks against $H_n + \lambda S_n^{-1}$ have relative Frobenius error below 2×10^{-8} for all-parameter regression, all-parameter classification, and last-layer classification. The diagonal and undamped-KFAC function-covariance identities in Proposition 1, under blockwise rescaling, are verified to relative errors below 2×10^{-15}.

K Proofs of the theoretical statements

This appendix states the formal versions of Theorems 1–4 (summarized informally in Section 5) and gives their proofs. For the first result, finite-set forward and backward Gram convergence is derived from TP-admissible finite-step μP training rather than assumed. The subsequent posterior and transfer results additionally assume stable data-space curvature, compact prior-precision search, convergence of the selected Laplace parameter subset's prior-weighted squared norm for fixed-checkpoint plug-in evidence-score transfer, and nonzero decision margins wherever exact argmax agreement is claimed.

Throughout, D denotes a finite training set, V a finite validation set, $\mathcal{P}$ a finite candidate pool, and $\mathcal{U}$ a fixed finite input collection containing all points needed by a decision score. For finite input sets A, B, write

$$K_n(A, B) = J_n(A) S_n J_n(B)^\top$$

for the width-normalized prior kernel used by $\sigma\text{Transfer}$, induced by the zero-mean weight prior $\theta_n \sim \mathcal{N}(0, \sigma_p^2 S_n)$. The Laplace posterior mean and covariance at prior precision $\lambda = \sigma_p^{-2}$ are

$$m_n^\lambda(X) = f_n(X; \hat{\theta}_n), \qquad C_n^\lambda(X, X') = J_n(X) \left(H_n + \lambda S_n^{-1}\right)^{-1} J_n(X')^\top.$$

The results after Theorem 1 are deterministic continuity statements applied pathwise. When their premises are supplied by that theorem, convergence and eventual agreement are understood almost surely. If analogous premises hold only in probability, the numerical conclusions hold in probability and the probability of the stated decision agreement tends to one.

K.1 Formal statement and proof of Theorem 1: prior-kernel stability

Assumption 1 (TP-admissible finite-step μP training)**.** *The depth, input and output dimensions, training and evaluation sets, and number of optimizer steps $T < \infty$ are fixed independently of width. The width sequence uses the Gaussian setup and general μP scaling of Yang & Littwin (2023, Setup 2.6.3 and Definition 2.9.12), up to the equivalent finite-width reparametrization of their Proposition 2.2.3. Wide parameter groups are independently initialized, zero-initialized biases and deterministic scalar parameters are handled as in their Remarks 2.9.10 and 2.9.27, and all widths are realized on a common probability space. The minibatch path and other external training choices are shared across widths. Apart from the ReLU derivative treated below, the loss-error and optimizer-history maps satisfy the regularity conditions of Yang & Littwin (2023, Assumption 2.9.17); in particular, the optimizer is entrywise and globally Lipschitz in each fixed finite gradient history. This*

includes SGD and positive-stabilizer Adam/AdamW with the prescribed μP gradient, or equivalently stabilizer, scaling (Yang & Littwin, 2023, Remark 2.2.6). When used, decoupled weight decay has width-independent multiplicative shrinkage as in their Section 2.10.1.

Assumption 2 (ReLU smoothing condition)**.** *Use the convention* $\mathrm{ReLU}'(0) = 0$*. At every $t \le T$, layer $\ell \le L$, and input x in the fixed training/evaluation set, the empirical preactivation law has no asymptotic atom at the ReLU kink:*

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \mathbf{1}\{|z^\ell_{i,n,t}(x)| \le \delta\} = 0 \qquad \textit{almost surely}.$$

For every fixed $p < \infty$, the empirical pth moments of all forward, backward, and optimizer-history coordinates through time T are almost surely bounded uniformly in n, using $n^{-1}\sum_i$ for vector histories and $n^{-2}\sum_{i,j}$ for matrix-indexed histories.

Lemma 1 (Finite-horizon ReLU stability)**.** *Under Assumptions 1 and 2, couple the exact-ReLU network with a network trained from the same initialization and along the same external training path, but with ReLU replaced by a smooth function φ_δ satisfying*

$$\sup_u |\varphi_\delta(u) - \mathrm{ReLU}(u)| \le C\delta, \qquad 0 \le \varphi'_\delta \le 1, \qquad \varphi'_\delta(u) = \mathbf{1}\{u > 0\} \quad \textit{when } |u| > \delta.$$

Along a fixed countable sequence $\delta \downarrow 0$, for every fixed $t \le T$, layer ℓ, and input x,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \left(\|h^{\ell,\delta}_{n,t}(x) - h^\ell_{n,t}(x)\|_{n,2} + \|r^{\ell,\delta}_{n,t}(x) - r^\ell_{n,t}(x)\|_{n,2} \right) = 0 \qquad \textit{almost surely},$$

where $\|v\|^2_{n,2} = n^{-1}\sum_i |v_i|^2$. The analogous discrepancies vanish in operator norm for each effective hidden matrix, in $\|\cdot\|_{n,2}$ for each wide vector parameter, and in absolute value for scalar parameters. Consequently, the trained outputs and the forward and backward Gram entries are asymptotically unchanged by smoothing.

Proof. The forward nonlinearity obeys

$$\|\varphi_\delta(u^\delta) - \mathrm{ReLU}(u)\|_{n,2} \le \|u^\delta - u\|_{n,2} + C\delta.$$

For the derivative gate, whenever $\epsilon > \delta$,

$$|\mathbf{1}\{u > 0\} - \varphi'_\delta(v)| \le \mathbf{1}\{|u| \le \epsilon\} + \mathbf{1}\{|u - v| \ge \epsilon - \delta\}.$$

After averaging, Assumption 2 controls the first term, whereas Markov's inequality bounds the second by $\|u - v\|^2_{n,2}/(\epsilon - \delta)^2$. Sending first $n \to \infty$, then $\delta \downarrow 0$, and finally $\epsilon \downarrow 0$ gives normalized L^2 convergence of the gates. The moment condition and Hölder's inequality give the same conclusion after multiplication by the corresponding reverse signals.

Updates preserve the coupling. For an effective hidden matrix, $\Delta\widetilde{W}_{ij} = \eta_\vartheta n^{-1} Q_{\vartheta,t}(G_{ij})$, where G_{ij} is its finite gradient history. Global Lipschitz continuity of $Q_{\vartheta,t}$ gives

$$\|\Delta\widetilde{W}^\delta - \Delta\widetilde{W}\|_{\mathrm{op}} \le \eta_\vartheta L_{\vartheta,t} \left[\frac{1}{n^2} \sum_{i,j} \|G^\delta_{ij} - G_{ij}\|_2^2 \right]^{1/2}.$$

Each history component is a finite sum of forward–reverse outer products, so the right-hand side vanishes by Cauchy–Schwarz and the preceding estimates. The same calculation, using $|Q(g)| \le |Q(0)| + L\|g\|_2$ and the moment bound, shows that every increment has $O(1)$ operator norm. Gaussian initialization supplies $O(1)$ effective-matrix norms (Yang & Littwin, 2023, Proposition 2.6.8); hence the trained effective matrices remain $O(1)$ through the fixed number of steps. Moreover,

$$\|W^\delta v^\delta - W v\|_{n,2} \le \|W^\delta\|_{\mathrm{op}} \|v^\delta - v\|_{n,2} + \|W^\delta - W\|_{\mathrm{op}} \|v\|_{n,2},$$

and likewise for transpose multiplication. Vector and scalar updates are analogous, and fixed multiplicative weight decay is linear.

These estimates close a finite induction through every forward, reverse, and update instruction. Finally, the normalized readout $f = n^{-1} a^\top h^L + c$ is stable by Cauchy–Schwarz, as are the empirical forward and backward inner products. This proves the lemma. □

Lemma 2 (Tensor-program Gram convergence). *Let $h^{\ell}_{n,t}(x)$ be the layer-ℓ activations and define the rescaled backward sensitivities*

$$r^{\ell}_{n,t}(x) = n_{\ell}\frac{\partial f_n(x;\theta_{n,t})}{\partial z^{\ell}_n(x;\theta_{n,t})}.$$

Under Assumptions 1 and 2, for every $t \le T$, $\ell \le L$, and pair x, x' in a fixed finite input set,

$$\Sigma^t_{\ell,n}(x,x') = \frac{1}{n_{\ell}}\sum_i h^{\ell}_{i,n,t}(x)h^{\ell}_{i,n,t}(x'), \qquad \Pi^t_{\ell,n}(x,x') = \frac{1}{n_{\ell}}\sum_i r^{\ell}_{i,n,t}(x)r^{\ell}_{i,n,t}(x')$$

converge jointly almost surely to finite limits. The corresponding outputs converge jointly as well.

Proof. Fix δ. The smoothed MLP is $\mathrm{NE}\otimes\mathrm{OR}^{\top}$-representable (Yang & Littwin, 2023, Section 2.6 and Example 2.9.6); its vector and scalar biases are admissible parameter types (Yang & Littwin, 2023, Definitions 2.9.1 and 2.9.7). The forward and scaled reverse computations form one total program (Yang & Littwin, 2023, Definitions 2.9.14 and 2.9.16), and the displayed $1/n$ readout makes the reverse variables precisely the sensitivities $r^{\ell}_{n,t}$. The smooth derivative and the loss and optimizer maps satisfy the required regularity conditions. Thus fixed-T training, including memoryful Adam/AdamW and decoupled weight decay, is covered by Yang & Littwin (2023, Theorem 2.9.30 and Section 2.10.1).

Append the coordinatewise products

$$h^{\ell}_{i,n,t}(x)h^{\ell}_{i,n,t}(x'), \qquad r^{\ell}_{i,n,t}(x)r^{\ell}_{i,n,t}(x'),$$

and their averaging instructions to this program. The product map is pseudo-Lipschitz, so the $\mathrm{NE}\otimes\mathrm{OR}^{\top}$ Master Theorem (Yang & Littwin, 2023, Theorem 2.6.10) gives joint almost-sure convergence of the two empirical averages and of the output for this fixed δ. Intersect the probability-one events over the chosen countable sequence $\delta \downarrow 0$.

Lemma 1 and Cauchy–Schwarz make the exact-ReLU Gram entries and outputs uniformly close, in the stated iterated limit, to their smoothed counterparts. The fixed-δ limits are therefore Cauchy as $\delta \downarrow 0$, and a three-ϵ argument gives the claimed limits for exact ReLU. There are only finitely many inputs, layers, and training steps, so the convergence is joint. □

Theorem (Formal statement of Theorem 1). *Consider the scalar-output, fixed-depth MLP and prior covariance S_n defined in Appendix J, with fixed input dimension and hidden widths $n_1 = \cdots = n_L = n$. Let $\hat{\theta}_n = \theta_{n,T}$ be obtained under Assumptions 1 and 2, and let J_n be the Jacobian at this checkpoint. On a fixed finite input set X, write $\Sigma_{\ell}(x,x')$ and $\Pi_{\ell}(x,x')$ for the almost-sure $t = T$ limits in Lemma 2, and set $\Sigma_0(x,x') = n_0^{-1}x^{\top}x'$. Then, jointly over $x, x' \in X$,*

$$K_n(x,x') = J_n(x)S_nJ_n(x')^{\top} \xrightarrow{\text{a.s.}} K_{\infty}(x,x'),$$

where

$$K_{\infty}(x,x') = 1 + \Sigma_L(x,x') + \Pi_1(x,x')\big(1 + d_{\mathrm{in}}\Sigma_0(x,x')\big) + \sum_{\ell=2}^{L}\Pi_{\ell}(x,x')\big(1 + \Sigma_{\ell-1}(x,x')\big).$$

The trained outputs $f_n(X;\theta_{n,T})$ converge jointly almost surely as well.

Proof. Equation (9) is an exact finite-width identity. By Lemma 2, all of its Gram factors converge jointly almost surely on the fixed finite set X. Since L and d_{in} are fixed, continuity of finite products and sums gives the displayed K_{∞} and almost-sure entrywise convergence of the finite kernel matrix. Output convergence is the final assertion of Lemma 2. □

Corollary 1 (Fixed-dimensional outputs). *Let the preceding network have a fixed output dimension q, a shared hidden trunk, and separate readout vectors and biases for its outputs. For output index a, define*

$$r^{\ell,a}_i(x) = n_{\ell}\frac{\partial f_{n,a}(x)}{\partial z^{\ell}_i(x)},$$

and, for output indices a, b, define

$$\Pi^{ab}_{\ell,n}(x,x') = \frac{1}{n_\ell}\sum_i r_i^{\ell,a}(x) r_i^{\ell,b}(x').$$

Under Assumptions 1 and 2, all these entries converge jointly almost surely. The prior-kernel blocks

$$K^{ab}_n(x,x') := J_{n,a}(x) S_n J_{n,b}(x')^\top \xrightarrow{\text{a.s.}} K^{ab}_\infty(x,x')$$

converge jointly, where

$$K^{ab}_\infty(x,x') = \delta_{ab}\big(1+\Sigma_L(x,x')\big) + \Pi^{ab}_1(x,x')\big(1+d_{\text{in}}\Sigma_0(x,x')\big) + \sum_{\ell=2}^{L}\Pi^{ab}_\ell(x,x')\big(1+\Sigma_{\ell-1}(x,x')\big).$$

Here δ_{ab} is the Kronecker delta. The q-dimensional trained outputs converge jointly almost surely as well.

Proof. Definition 2.9.16 of Yang & Littwin (2023) places the forward and backpropagation programs for all q output coordinates in one total program. Since q is fixed, the proof of Lemma 2 applies jointly to every output and output pair, giving both output convergence and the limits Π^{ab}_ℓ. Separate readout and output-bias blocks contribute the factor δ_{ab}; the shared hidden blocks give the displayed sum. □

K.2 Coordinate-invariant interpretation of Theorem 1

Let ϕ_n denote the width-normalized local perturbation coordinates used by $\sigma\mathrm{Transfer}$ and suppose the stored implementation parameters satisfy

$$\theta_n = T_n \phi_n,$$

where T_n is block diagonal. Then

$$J_{\phi,n}(X) = J_{\theta,n}(X) T_n.$$

An isotropic zero-mean prior in these width-normalized coordinates,

$$\phi_n \sim \mathcal{N}(0, \sigma_p^2 I),$$

induces in stored coordinates

$$\theta_n = T_n\phi_n \sim \mathcal{N}(0, \sigma_p^2 T_n T_n^\top).$$

Therefore

$$S_n = T_n T_n^\top.$$

The induced prior kernel is

$$J_{\theta,n}(X) S_n J_{\theta,n}(X')^\top = J_{\theta,n}(X) T_n T_n^\top J_{\theta,n}(X')^\top = J_{\phi,n}(X) J_{\phi,n}(X')^\top.$$

Thus the $\sigma\mathrm{Transfer}$ prior is equivalently an isotropic local prior in width-normalized perturbation coordinates. Prior-kernel stability is the statement that this normalized-coordinate Jacobian kernel has a stable large-width limit.

K.3 Formal statement and proof of Theorem 2: posterior-covariance stability

Theorem (Formal statement of Theorem 2)**.** *Fix finite sets D, X, X', and assume $S_n \succ 0$ for every n. Suppose that for every $A, B \subseteq D \cup X \cup X'$,*

$$K_n(A,B) = J_n(A) S_n J_n(B)^\top \longrightarrow K_\infty(A,B).$$

Assume the Laplace curvature has generalized Gauss–Newton form $H_n = J_n(D)^\top \Omega_n J_n(D)$, where $\Omega_n \succeq 0$ and $\Omega_n \to \Omega_\infty$ in operator norm. Then, for every compact interval $\Lambda = [\lambda_-, \lambda_+] \subset (0,\infty)$,

$$\sup_{\lambda\in\Lambda} \|C^\lambda_n(X,X') - C^\lambda_\infty(X,X')\|_{\text{op}} \to 0,$$

where

$$C^\lambda_\infty(X,X') = \lambda^{-1}K_\infty(X,X') - \lambda^{-2}K_\infty(X,D)\Omega^{1/2}_\infty\big(I + \lambda^{-1}\Omega^{1/2}_\infty K_\infty(D,D)\Omega^{1/2}_\infty\big)^{-1} \times \Omega^{1/2}_\infty K_\infty(D,X').$$

Proof. For any finite set A, define the whitened Jacobian

$$G_{A,n} = J_n(A)S_n^{1/2}.$$

Then

$$K_n(A, B) = G_{A,n}G_{B,n}^\top.$$

The GGN posterior precision can be written as

$$J_n(D)^\top \Omega_n J_n(D) + \lambda S_n^{-1} = S_n^{-1/2}\left(G_{D,n}^\top \Omega_n G_{D,n} + \lambda I\right) S_n^{-1/2}.$$

Therefore

$$\begin{aligned} C_n^\lambda(X, X') &= J_n(X)\left(J_n(D)^\top \Omega_n J_n(D) + \lambda S_n^{-1}\right)^{-1} J_n(X')^\top \\ &= G_{X,n}\left(G_{D,n}^\top \Omega_n G_{D,n} + \lambda I\right)^{-1} G_{X',n}^\top. \end{aligned}$$

Apply the Woodbury identity to

$$G_{D,n}^\top \Omega_n G_{D,n} + \lambda I.$$

With

$$M_n(\lambda) = I + \lambda^{-1}\Omega_n^{1/2} G_{D,n} G_{D,n}^\top \Omega_n^{1/2},$$

we obtain

$$\begin{aligned} C_n^\lambda(X, X') = \lambda^{-1} G_{X,n} G_{X',n}^\top \\ - \lambda^{-2} G_{X,n} G_{D,n}^\top \Omega_n^{1/2} M_n(\lambda)^{-1} \Omega_n^{1/2} G_{D,n} G_{X',n}^\top. \end{aligned}$$

Since

$$G_{A,n}G_{B,n}^\top = K_n(A, B),$$

this becomes the finite-dimensional data-space expression

$$\begin{aligned} C_n^\lambda(X, X') = \lambda^{-1} K_n(X, X') \\ - \lambda^{-2} K_n(X, D)\Omega_n^{1/2}\left(I + \lambda^{-1}\Omega_n^{1/2} K_n(D, D)\Omega_n^{1/2}\right)^{-1} \Omega_n^{1/2} K_n(D, X'). \end{aligned}$$

By assumption,

$$K_n(A, B) \to K_\infty(A, B)$$

for all finite $A, B \subseteq D \cup X \cup X'$, and

$$\Omega_n \to \Omega_\infty.$$

The square-root map is continuous on finite-dimensional positive semidefinite matrices, so

$$\Omega_n^{1/2} \to \Omega_\infty^{1/2}.$$

Moreover,

$$M_n(\lambda) = I + \lambda^{-1}\Omega_n^{1/2} K_n(D, D)\Omega_n^{1/2} \succeq I.$$

Hence $M_n(\lambda)$ is invertible and

$$\|M_n(\lambda)^{-1}\|_{\mathrm{op}} \leq 1.$$

The convergence of $M_n(\lambda)$ to

$$M_\infty(\lambda) = I + \lambda^{-1}\Omega_\infty^{1/2} K_\infty(D, D)\Omega_\infty^{1/2}$$

is uniform for $\lambda \in [\lambda_-, \lambda_+]$, because λ^{-1} is bounded and continuous on this compact interval. Since both $M_n(\lambda)$ and $M_\infty(\lambda)$ are bounded below by I,

$$\|M_n(\lambda)^{-1} - M_\infty(\lambda)^{-1}\|_{\mathrm{op}} \leq \|M_n(\lambda) - M_\infty(\lambda)\|_{\mathrm{op}},$$

so the inverses also converge uniformly. Substituting each uniformly convergent finite-dimensional factor into the Woodbury expression yields

$$\sup_{\lambda \in \Lambda} \|C_n^\lambda(X, X') - C_\infty^\lambda(X, X')\|_{\mathrm{op}} \to 0.$$

□

Corollary 2 (Last-layer Laplace stability)**.** *For the MLP of Appendix J, consider a last-layer Laplace approximation over only the readout weights a and output bias c. If the final-layer feature Gram matrix satisfies*

$$\frac{1}{n_L}\sum_{i=1}^{n_L} h_i^L(x)h_i^L(x') \to \Sigma_L(x,x'),$$

then the last-layer prior kernel

$$K_{LL,n}(x,x') = 1 + \frac{1}{n_L}\sum_{i=1}^{n_L} h_i^L(x)h_i^L(x')$$

converges to

$$K_{LL,\infty}(x,x') = 1 + \Sigma_L(x,x').$$

Consequently, the last-layer Laplace posterior covariance is stable under the assumptions of Theorem 2.

Proof. For the readout weights,

$$\frac{\partial f_n(x)}{\partial a_i} = \frac{h_i^L(x)}{n_L}, \qquad S_{a,n} = n_L I_a.$$

Therefore

$$n_L \sum_{i=1}^{n_L} \frac{\partial f_n(x)}{\partial a_i}\frac{\partial f_n(x')}{\partial a_i} = \frac{1}{n_L}\sum_{i=1}^{n_L} h_i^L(x)h_i^L(x').$$

The output-bias contribution is 1. Hence

$$K_{LL,n}(x,x') = 1 + \frac{1}{n_L}\sum_{i=1}^{n_L} h_i^L(x)h_i^L(x') \to 1 + \Sigma_L(x,x').$$

The posterior-covariance statement then follows by applying Theorem 2 with $K_n = K_{LL,n}$. □

Corollary 3 (Exact-Hessian residual condition)**.** *Under the kernel and curvature assumptions of Theorem 2, suppose the exact Hessian decomposes as*

$$H_n = J_n(D)^\top \Omega_n J_n(D) + R_n.$$

If

$$E_n = S_n^{1/2} R_n S_n^{1/2} \to 0$$

in operator norm, then the exact-Hessian Laplace posterior covariance has the same large-width limit as the GGN posterior covariance. The convergence is uniform for λ in every compact interval $\Lambda \subset (0,\infty)$.

Proof. In whitened coordinates, the exact-Hessian covariance is

$$\widetilde{C}_n^\lambda(X,X') = G_{X,n}\left(G_{D,n}^\top \Omega_n G_{D,n} + E_n + \lambda I\right)^{-1} G_{X',n}^\top.$$

Let

$$A_n(\lambda) = G_{D,n}^\top \Omega_n G_{D,n} + \lambda I.$$

Then

$$A_n(\lambda) \succeq \lambda_- I, \qquad \|A_n(\lambda)^{-1}\|_{\mathrm{op}} \le \lambda_-^{-1}$$

uniformly for $\lambda \in \Lambda$. For large n, $\|E_n\|_{\mathrm{op}} < \lambda_-/2$, so

$$A_n(\lambda) + E_n \succeq \frac{\lambda_-}{2} I$$

and

$$\|(A_n(\lambda) + E_n)^{-1}\|_{\mathrm{op}} \le \frac{2}{\lambda_-}.$$

The resolvent identity gives

$$\widetilde{C}_n^\lambda(X,X') - C_n^\lambda(X,X') = -G_{X,n}\left(A_n(\lambda) + E_n\right)^{-1} E_n A_n(\lambda)^{-1} G_{X',n}^\top.$$

Since

$$\|G_{X,n}\|_{\mathrm{op}}^2 = \|K_n(X,X)\|_{\mathrm{op}} = O(1)$$

and similarly for X', we obtain

$$\sup_{\lambda\in\Lambda}\|\widetilde{C}_n^{\lambda}(X,X') - C_n^{\lambda}(X,X')\|_{\mathrm{op}} \le O(1)\frac{2}{\lambda_-^2}\|E_n\|_{\mathrm{op}} \to 0.$$

Thus the exact-Hessian residual does not change the function-space covariance limit. □

K.4 Structured-curvature extensions: diagonal and KFAC

Theorem 2 uses the full GGN because its data-space form is determined entirely by K_n and Ω_n. Diagonal and Kronecker-factored approximations retain the width-normalized coordinate interpretation, but their width limits depend on more than K_n. We separate these two facts below.

Let $\theta_n = T_n\phi_n$ map width-normalized coordinates ϕ_n to stored coordinates θ_n, let $S_n = T_nT_n^\top$, and write $J_{\phi,n} = J_{\theta,n}T_n$. For an approximate positive-semidefinite curvature $\widetilde{H}_{\theta,n}$ in stored coordinates, define $\widetilde{H}_{\phi,n} = T_n^\top \widetilde{H}_{\theta,n}T_n$ and

$$\widetilde{C}_n^{\lambda}(A,B) = J_n(A)\left(\widetilde{H}_{\theta,n} + \lambda S_n^{-1}\right)^{-1} J_n(B)^\top.$$

Proposition 1 (Finite-width coordinate covariance)**.** *At every finite width, the following stored-coordinate computations equal the corresponding computations with an isotropic prior in the width-normalized coordinates.*

1. *For diagonal Laplace, the claim holds whenever T_n is diagonal.*
2. *For block-Kronecker Laplace, it holds whenever $T_n = \mathrm{blockdiag}_g(t_{n,g}I_g)$ with $t_{n,g} > 0$, the curvature blocks and prior groups are aligned, and the block approximation $\mathcal{Q}_g$ is positively homogeneous:*

$$\mathcal{Q}_g(t^2H) = t^2\mathcal{Q}_g(H).$$

The Kronecker statement concerns additive prior damping $\mathcal{Q}_g(H)+\lambda I_g$, not multiplicative factor damping.

The verified Linear/MLP width-normalized prior satisfies these structural conditions because its scale is constant within every weight or bias tensor.

Proof. The exact curvature transforms as $H_{\phi,n} = T_n^\top H_{\theta,n}T_n$. If T_n is diagonal, then

$$\mathrm{diag}(H_{\phi,n}) = T_n^\top \mathrm{diag}(H_{\theta,n})T_n.$$

Consequently,

$$\begin{aligned} &J_{\phi,n}(A)\left(\mathrm{diag}(H_{\phi,n}) + \lambda I\right)^{-1} J_{\phi,n}(B)^\top \\ &\quad = J_{\theta,n}(A)\left(\mathrm{diag}(H_{\theta,n}) + \lambda S_n^{-1}\right)^{-1} J_{\theta,n}(B)^\top. \end{aligned}$$

For a Kronecker block, $T_{n,g} = t_{n,g}I_g$ gives $H_{\phi,n,g} = t_{n,g}^2 H_{\theta,n,g}$. Homogeneity therefore gives $\mathcal{Q}_g(H_{\phi,n,g}) = t_{n,g}^2\mathcal{Q}_g(H_{\theta,n,g})$. In particular,

$$\begin{aligned} &J_{\phi,n,g}(A)\left(\mathcal{Q}_g(H_{\phi,n,g}) + \lambda I_g\right)^{-1} J_{\phi,n,g}(B)^\top \\ &\quad = J_{\theta,n,g}(A)\left(\mathcal{Q}_g(H_{\theta,n,g}) + \lambda t_{n,g}^{-2} I_g\right)^{-1} J_{\theta,n,g}(B)^\top. \end{aligned}$$

Summing the block contributions proves the second statement. □

Finite-width covariance does not by itself imply convergence across width. The following proposition states one sufficient condition that covers both diagonal and Kronecker curvature.

Proposition 2 (Conditional width stability for structured curvature)**.** *Work in the width-normalized coordinates and set $G_n(A) = J_{\theta,n}(A)T_n$. Let the structured curvature have spectral decomposition*

$$\widetilde{H}_{\phi,n} = \sum_{r=1}^{p_n} q_{n,r} u_{n,r} u_{n,r}^\top, \qquad q_{n,r} \geq 0.$$

For fixed finite sets A, B, define the matrix-valued spectral measure

$$\mathcal{M}_n^{A,B} = \sum_{r=1}^{p_n} \big(G_n(A)u_{n,r}\big)\big(G_n(B)u_{n,r}\big)^\top \delta_{q_{n,r}}.$$

For finite signed measures, write

$$d_{\mathrm{BL}}(\mu,\nu) = \sup_{\|h\|_\infty + \mathrm{Lip}(h) \leq 1} \left| \int h \, d(\mu - \nu) \right|.$$

Suppose that every scalar entry of $\mathcal{M}_n^{A,B}$ converges in d_{BL} to a finite signed measure $\mathcal{M}_\infty^{A,B}$ for all required A, B. Then, for every compact $\Lambda = [\lambda_-, \lambda_+] \subset (0,\infty)$,

$$\sup_{\lambda \in \Lambda} \left\| \widetilde{C}_n^\lambda(A,B) - \int_{[0,\infty)} \frac{1}{t+\lambda} \, d\mathcal{M}_\infty^{A,B}(t) \right\| \longrightarrow 0.$$

For diagonal Laplace, $q_{n,r}$ are the per-coordinate diagonal curvatures. For undamped KFAC, they are the eigenvalues of its block-Kronecker curvature.

Proof. The spectral theorem gives the exact representation

$$\begin{aligned} \widetilde{C}_n^\lambda(A,B) &= G_n(A)(\widetilde{H}_{\phi,n} + \lambda I)^{-1} G_n(B)^\top \\ &= \int_{[0,\infty)} \frac{1}{t+\lambda} \, d\mathcal{M}_n^{A,B}(t). \end{aligned}$$

For $h_\lambda(t) = (t+\lambda)^{-1}$ and $\lambda \in [\lambda_-, \lambda_+]$,

$$\|h_\lambda\|_\infty \leq \lambda_-^{-1}, \qquad \mathrm{Lip}(h_\lambda) \leq \lambda_-^{-2}.$$

The family $\{h_\lambda : \lambda \in \Lambda\}$ therefore has uniformly bounded norm in the bounded-Lipschitz class. Applying the assumed convergence entrywise and using equivalence of norms in the fixed-dimensional output space yields the stated uniform convergence. □

Corollary 4 (Structured-approximation validation-NLL and decision transfer)**.** *Under Proposition 2, if the corresponding predictive means converge uniformly and the assumptions of Theorems 3 and 4 hold, then their validation-NLL precision-transfer and decision-transfer conclusions hold with C_n^λ replaced by $\widetilde{C}_n^\lambda$.*

Proof. Proposition 2 supplies the uniform covariance convergence used in the proofs of Theorems 3 and 4; the remaining assumptions are unchanged. □

Status for the structured-curvature experiments. The finite-width coordinate-covariance identity is verified numerically, but the projected spectral-measure convergence condition of Proposition 2 is not directly tested. The diagonal/KFAC asymptotic extension is therefore conditional; the reported finite-width transfer is empirical evidence for its conclusion.

Why prior-kernel convergence alone is insufficient. Consider one training input d, one test input x, scalar output, and unit likelihood curvature. Two normalized Jacobian pairs are

	$G(d)$	$G(x)$
I	$(1,0)$	$(0,1)$
II	$2^{-1/2}(1,1)$	$2^{-1/2}(-1,1)$.

Both have exactly the same prior kernel: $K(d,d) = K(x,x) = 1$ and $K(d,x) = 0$. Yet diagonal curvature gives

$$\widetilde{C}_{\mathrm{I}}^\lambda(x,x) = \lambda^{-1}, \qquad \widetilde{C}_{\mathrm{II}}^\lambda(x,x) = (\lambda + \tfrac{1}{2})^{-1}.$$

An alternating width sequence therefore has a constant K_n but a nonconvergent diagonal posterior covariance. The example does not challenge Proposition 1: it uses a rotation, whereas the width-normalized prior is diagonal and block-scalar. It shows why the extra spectral condition in Proposition 2 is necessary. KFAC likewise depends on its retained block-factor spectrum, not only on the sum $G_n G_n^\top$.

Other approximations. Proposition 2 applies more generally to any positive-semidefinite curvature approximation whose projected spectral measures satisfy its assumption. The implementation checks cover full, diagonal, and KFAC curvature for Linear/MLP models.

K.5 Formal statement and proof of Theorem 3: inference-hyperparameter transfer

Theorem (Formal statement of Theorem 3)**.** *Let $\Lambda = [\lambda_-, \lambda_+] \subset (0, \infty)$ be the prior-precision search interval. Suppose that on a finite validation set V, the predictive mean and covariance converge uniformly in λ:*

$$\sup_{\lambda \in \Lambda} \|m_n^\lambda(V) - m_\infty^\lambda(V)\| \to 0, \qquad \sup_{\lambda \in \Lambda} \|C_n^\lambda(V,V) - C_\infty^\lambda(V,V)\| \to 0.$$

For scalar Gaussian regression with observation noise $\tau^2 > 0$, or for softmax classification with Gaussian logit posterior, the validation NLL curves converge uniformly: $\sup_{\lambda \in \Lambda} |\mathrm{NLL}_n(\lambda) - \mathrm{NLL}_\infty(\lambda)| \to 0$. If NLL_∞ has a unique minimizer $\lambda^\star$, then every sequence of minimizers satisfies $\lambda_n^\star \to \lambda^\star$. Moreover, for a proxy width n and target width N,

$$\mathrm{NLL}_N(\lambda_n^\star) - \min_{\lambda \in \Lambda} \mathrm{NLL}_N(\lambda) \le 2 \sup_{\lambda \in \Lambda} |\mathrm{NLL}_n(\lambda) - \mathrm{NLL}_N(\lambda)|.$$

Proof. Let

$$\mathcal{L}_n(\lambda) = \mathrm{NLL}_n(\lambda).$$

Because $\lambda \ge \lambda_- > 0$, the finite-width Laplace mean and covariance are continuous in λ; their uniform limits are therefore continuous as well. We first show uniform convergence of the validation NLL curves from uniform convergence of the predictive posterior mean and covariance.

Gaussian regression. For the usual sum of marginal predictive NLLs, put

$$s_{n,\lambda,i} = C_{n,\lambda}(x_i, x_i) + \tau^2, \qquad \tau^2 > 0.$$

Then

$$\mathcal{L}_n(\lambda) = \frac{1}{2} \sum_{i=1}^{|V|} \left[\frac{(y_i - m_{n,\lambda}(x_i))^2}{s_{n,\lambda,i}} + \log s_{n,\lambda,i} + \log(2\pi) \right].$$

Positive semidefiniteness gives $s_{n,\lambda,i} \ge \tau^2$. The assumptions imply uniform convergence of every $m_{n,\lambda}(x_i)$ and $s_{n,\lambda,i}$. Because the validation set and output dimension are fixed, the arguments remain in a compact subset of $\mathbb{R} \times [\tau^2, \infty)$, on which each displayed summand is uniformly continuous. Therefore

$$\sup_{\lambda \in \Lambda} |\mathcal{L}_n(\lambda) - \mathcal{L}_\infty(\lambda)| \to 0.$$

Softmax classification. For classification, let the Laplace posterior over logits at validation point x_i be

$$f_i \sim \mathcal{N}\left(m_{n,\lambda}(x_i), C_{n,\lambda}(x_i, x_i)\right).$$

The predictive probability of the observed class y_i is

$$p_{n,\lambda}(y_i \mid x_i) = \mathbb{E}_{f_i}\left[\mathrm{softmax}_{y_i}(f_i)\right].$$

The validation NLL is

$$\mathcal{L}_n(\lambda) = -\sum_{i=1}^{|V|} \log p_{n,\lambda}(y_i \mid x_i).$$

The map

$$f \mapsto \mathrm{softmax}_{y_i}(f)$$

is bounded and continuous. In finite dimension, expectations of bounded continuous functions under a Gaussian distribution depend continuously on the Gaussian mean and covariance. Because the posterior means and covariances converge uniformly over compact Λ, we get

$$\sup_{\lambda \in \Lambda} |p_{n,\lambda}(y_i \mid x_i) - p_{\infty,\lambda}(y_i \mid x_i)| \to 0$$

for each i. More explicitly, the limiting mean–covariance pairs form a compact image of Λ, and the finite-width pairs eventually remain in a compact neighborhood of that image; the Gaussian-expectation map is therefore uniformly continuous on the relevant set.

For every finite mean and covariance,

$$p_{\infty,\lambda}(y_i \mid x_i) > 0.$$

The map

$$\lambda \mapsto p_{\infty,\lambda}(y_i \mid x_i)$$

is continuous on compact Λ, so

$$\rho_i = \inf_{\lambda \in \Lambda} p_{\infty,\lambda}(y_i \mid x_i) > 0.$$

Uniform convergence then implies that for sufficiently large n,

$$p_{n,\lambda}(y_i \mid x_i) \geq \frac{\rho_i}{2}$$

for all $\lambda \in \Lambda$. Since $-\log t$ is uniformly continuous on $[\rho_i/2, 1]$,

$$\sup_{\lambda \in \Lambda} |-\log p_{n,\lambda}(y_i \mid x_i) + \log p_{\infty,\lambda}(y_i \mid x_i)| \to 0.$$

Summing over the finite validation set gives

$$\sup_{\lambda \in \Lambda} |\mathcal{L}_n(\lambda) - \mathcal{L}_\infty(\lambda)| \to 0.$$

The same argument covers a deterministic continuous probit approximation and a fixed common-random-number Monte Carlo rule, using the symmetric covariance square root for reparameterization. Independent fixed-sample Monte Carlo draws at each width retain simulation noise and are not covered without a growing sample count or an additional probabilistic error bound.

It remains to prove convergence of minimizers. Let

$$\lambda^\star = \arg\min_{\lambda \in \Lambda} \mathcal{L}_\infty(\lambda)$$

be the unique limiting minimizer, and let

$$\lambda_n^\star \in \arg\min_{\lambda \in \Lambda} \mathcal{L}_n(\lambda).$$

Fix $\varepsilon > 0$ and define

$$B_\varepsilon = \{\lambda \in \Lambda : |\lambda - \lambda^\star| \geq \varepsilon\}.$$

Since Λ is compact, B_ε is compact. Since $\mathcal{L}_\infty$ is continuous and has a unique minimizer at $\lambda^\star$,

$$\gamma_\varepsilon = \inf_{\lambda \in B_\varepsilon} [\mathcal{L}_\infty(\lambda) - \mathcal{L}_\infty(\lambda^\star)] > 0.$$

Uniform convergence gives, for sufficiently large n,

$$\sup_{\lambda \in \Lambda} |\mathcal{L}_n(\lambda) - \mathcal{L}_\infty(\lambda)| < \frac{\gamma_\varepsilon}{3}.$$

If $\lambda_n^\star \in B_\varepsilon$, then

$$\mathcal{L}_\infty(\lambda_n^\star) \geq \mathcal{L}_\infty(\lambda^\star) + \gamma_\varepsilon.$$

Therefore

$$\mathcal{L}_n(\lambda_n^\star) \geq \mathcal{L}_\infty(\lambda_n^\star) - \frac{\gamma_\varepsilon}{3} \geq \mathcal{L}_\infty(\lambda^\star) + \frac{2\gamma_\varepsilon}{3}.$$

But also

$$\mathcal{L}_n(\lambda^\star) \le \mathcal{L}_\infty(\lambda^\star) + \frac{\gamma_\varepsilon}{3}.$$

Thus

$$\mathcal{L}_n(\lambda_n^\star) > \mathcal{L}_n(\lambda^\star),$$

contradicting the fact that $\lambda_n^\star$ minimizes $\mathcal{L}_n$. Hence, for sufficiently large n,

$$|\lambda_n^\star - \lambda^\star| < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary,

$$\lambda_n^\star \to \lambda^\star.$$

Finally, for a proxy width n and target width N, define

$$\Delta_{n,N} = \sup_{\lambda\in\Lambda} |\mathcal{L}_n(\lambda) - \mathcal{L}_N(\lambda)|.$$

Let

$$\lambda_N^\star \in \arg\min_{\lambda\in\Lambda} \mathcal{L}_N(\lambda).$$

Because $\lambda_n^\star$ minimizes $\mathcal{L}_n$,

$$\mathcal{L}_n(\lambda_n^\star) \le \mathcal{L}_n(\lambda_N^\star).$$

Therefore

$$\begin{aligned}\mathcal{L}_N(\lambda_n^\star) - \mathcal{L}_N(\lambda_N^\star) &= [\mathcal{L}_N(\lambda_n^\star) - \mathcal{L}_n(\lambda_n^\star)] \\ &\quad + [\mathcal{L}_n(\lambda_n^\star) - \mathcal{L}_n(\lambda_N^\star)] \\ &\quad + [\mathcal{L}_n(\lambda_N^\star) - \mathcal{L}_N(\lambda_N^\star)] \\ &\le \Delta_{n,N} + 0 + \Delta_{n,N} = 2\Delta_{n,N}.\end{aligned}$$

This proves the transferred excess-NLL bound. □

Corollary 5 (Evidence-score precision transfer)**.** *Retain the fixed-training-set and full-GGN assumptions of Theorem 2. Use the zero-mean σTransfer prior $\theta_n \sim \mathcal{N}(0, \lambda^{-1}S_n)$; for each n, hold all likelihood hyperparameters fixed and do not refit $\hat\theta_n$ as λ varies. Fix a compact $\Lambda = [\lambda_-, \lambda_+] \subset (0,\infty)$. Let*

$$q_n = \hat\theta_n^\top S_n^{-1} \hat\theta_n.$$

Assume $q_n \to q_\infty < \infty$, and define

$$B_n = \Omega_n^{1/2} K_n(D,D) \Omega_n^{1/2}, \qquad B_\infty = \Omega_\infty^{1/2} K_\infty(D,D) \Omega_\infty^{1/2}.$$

Define the centered fixed-checkpoint plug-in Laplace evidence score used in our experiments by

$$\mathcal{E}_n(\lambda) = -\frac{\lambda}{2} q_n + \frac{1}{2}\log\det(\lambda S_n^{-1}) - \frac{1}{2}\log\det(H_n + \lambda S_n^{-1}), \qquad \lambda \in \Lambda.$$

Terms independent of λ, including the likelihood at $\hat\theta_n$, are omitted because they do not affect the selected precision within this fixed-checkpoint score. Let

$$\mathcal{E}_\infty(\lambda) = -\frac{\lambda}{2} q_\infty - \frac{1}{2}\log\det\left(I + \lambda^{-1} B_\infty\right).$$

Then

$$\sup_{\lambda\in\Lambda} |\mathcal{E}_n(\lambda) - \mathcal{E}_\infty(\lambda)| \to 0.$$

If $\mathcal{E}_\infty$ has a unique maximizer $\lambda_{\mathrm{ev}}^\star$, then every sequence $\lambda_{n,\mathrm{ev}}^\star \in \arg\max_{\lambda\in\Lambda} \mathcal{E}_n(\lambda)$ satisfies $\lambda_{n,\mathrm{ev}}^\star \to \lambda_{\mathrm{ev}}^\star$. Moreover, for proxy width n and target width N,

$$\max_{\lambda\in\Lambda} \mathcal{E}_N(\lambda) - \mathcal{E}_N(\lambda_{n,\mathrm{ev}}^\star) \le 2 \sup_{\lambda\in\Lambda} |\mathcal{E}_n(\lambda) - \mathcal{E}_N(\lambda)|.$$

For a fixed finite grid $G \subset \Lambda$, if $\lambda_{\mathrm{ev},G}^\star = \arg\max_{\lambda\in G} \mathcal{E}_\infty(\lambda)$ is unique, then

$$\arg\max_{\lambda\in G} \mathcal{E}_n(\lambda) = \lambda_{\mathrm{ev},G}^\star$$

for all sufficiently large n.

Proof. By definition, $\mathcal{E}_n$ is the λ-dependent part of the fixed-checkpoint plug-in score. Let $\widetilde{J}_n = J_n(D)S_n^{1/2}$. Factoring $S_n^{-1/2}$ from the two parameter-space determinants and then applying Sylvester's determinant identity gives

$$\begin{aligned}&\log\det(\lambda S_n^{-1}) - \log\det(H_n + \lambda S_n^{-1})\\ &\quad = -\log\det\Big(I + \lambda^{-1}\widetilde{J}_n^\top \Omega_n \widetilde{J}_n\Big) = -\log\det\big(I + \lambda^{-1}B_n\big)\,.\end{aligned}$$

Hence

$$\mathcal{E}_n(\lambda) = -\frac{\lambda}{2}q_n - \frac{1}{2}\log\det\big(I + \lambda^{-1}B_n\big)\,.$$

Each B_n is positive semidefinite, so the matrix inside the log-determinant is positive definite for every $\lambda > 0$, including when Ω_n is singular. Because the stacked training output has fixed finite dimension, convergence of $K_n(D, D)$ and Ω_n, together with continuity of the positive-semidefinite square-root map, implies $B_n \to B_\infty$ in operator norm. If r is the dimension of B_n, Weyl's inequality and the Lipschitz bound for $t \mapsto \log(1 + t/\lambda)$ give

$$\sup_{\lambda\in\Lambda}|\mathcal{E}_n(\lambda) - \mathcal{E}_\infty(\lambda)| \le \frac{\lambda_+}{2}|q_n - q_\infty| + \frac{r}{2\lambda_-}\|B_n - B_\infty\|_{\mathrm{op}} \to 0.$$

Applying the minimizer-consistency argument of Theorem 3 to $-\mathcal{E}_n$ proves convergence of the maximizers. Applying the same three-term comparison used for the excess-NLL bound proves the stated finite-width evidence-regret bound. The finite-grid conclusion follows from the same uniform convergence: once the uniform error is smaller than half the positive gap between the best and second-best limiting grid points, no other grid point can overtake the maximizer. □

Scope of the evidence result. The convergence of q_n is an additional assumption, not a consequence of prior-kernel or posterior convergence: likelihood-null parameter directions can change q_n and the plug-in evidence-score optimum without changing the predictive posterior. This condition can hold for normalized last-layer readouts, but generally fails for a zero-mean all-parameter prior containing $n \times n$ hidden matrices: in the explicit-coordinate convention, one such block with $O(1)$ entries contributes $n^{-1}\|\hat{W}_n\|_F^2 = \Theta(n)$ to q_n. The corollary covers only the full-GGN version of the fixed-checkpoint plug-in criterion on the chosen Laplace parameter subset; $\hat{\theta}_n$ and S_n are restricted to that subset. The experiments do not by themselves establish asymptotic convergence of q_n. This plug-in score is not the marginal likelihood obtained by refitting the posterior mode at every λ. Such refitted evidence, mode-relaxation corrections, and diagonal or KFAC curvature require further mode, gradient, or log-determinant stability assumptions and are not covered here. The approximate-optimization and finite-grid statements below also apply to this evidence criterion after replacing $\mathcal{L}_n$ by $-\mathcal{E}_n$.

Corollary 6 (Approximate hyperparameter optimization)**.** *If $\tilde{\lambda}_n$ is an approximate minimizer satisfying*

$$\mathcal{L}_n(\tilde{\lambda}_n) \le \inf_{\lambda\in\Lambda}\mathcal{L}_n(\lambda) + \eta_n, \qquad \eta_n \to 0,$$

then $\tilde{\lambda}_n \to \lambda^\star$ under the assumptions of Theorem 3.

Proof. The proof is the same as the argmin-consistency part of Theorem 3. For fixed $\varepsilon > 0$, use the same limiting margin $\gamma_\varepsilon > 0$. For sufficiently large n,

$$\sup_{\lambda\in\Lambda}|\mathcal{L}_n(\lambda) - \mathcal{L}_\infty(\lambda)| < \frac{\gamma_\varepsilon}{3}, \qquad \eta_n < \frac{\gamma_\varepsilon}{3}.$$

Any approximate minimizer outside the ε-ball around $\lambda^\star$ would then have objective value more than η_n above the best attainable value, contradicting the definition of $\tilde{\lambda}_n$. □

Corollary 7 (Finite-grid hyperparameter search)**.** *Let $G = \{\lambda_1, \dots, \lambda_R\} \subset \Lambda$ be a finite grid. If*

$$\max_{\lambda\in G}|\mathcal{L}_n(\lambda) - \mathcal{L}_\infty(\lambda)| \to 0$$

and the limiting grid optimum

$$\lambda_G^\star = \arg\min_{\lambda\in G}\mathcal{L}_\infty(\lambda)$$

is unique, then

$$\arg\min_{\lambda \in G} \mathcal{L}_n(\lambda) = \lambda_G^\star$$

for all sufficiently large n*.*

Proof. Define the grid margin

$$\gamma_G = \min_{\lambda \in G, \lambda \neq \lambda_G^\star} [\mathcal{L}_\infty(\lambda) - \mathcal{L}_\infty(\lambda_G^\star)] > 0.$$

Once

$$\max_{\lambda \in G} |\mathcal{L}_n(\lambda) - \mathcal{L}_\infty(\lambda)| < \frac{\gamma_G}{2},$$

no grid point can overtake $\lambda_G^\star$. Hence the grid minimizer agrees exactly for all sufficiently large n. □

K.6 Formal statement and proof of Theorem 4: decision and active-learning transfer

Theorem (Formal statement of Theorem 4)**.** *Let* $\mathcal{P} = \{u_1, \dots, u_M\}$ *be a finite candidate pool, let* $\mathcal{U} \supseteq \mathcal{P}$ *be a fixed finite input collection containing all candidate and auxiliary target/evaluation points used by the scores, and let* $\mathcal{A}$ *be a finite action set, such as* $\mathcal{A} = \mathcal{P}$ *for single-point acquisition or* $\mathcal{A} = \{S \subset \mathcal{P} : |S| = K\}$ *for batch acquisition. Fix* $\lambda > 0$*. Suppose* $m_n^\lambda(\mathcal{U}) \to m_\infty^\lambda(\mathcal{U})$ *and* $C_n^\lambda(\mathcal{U},\mathcal{U}) \to C_\infty^\lambda(\mathcal{U},\mathcal{U})$*. Let* $A_n^\lambda(a) = F_a(m_n^\lambda(\mathcal{U}), C_n^\lambda(\mathcal{U},\mathcal{U}), \lambda)$*, where each* F_a *is continuous at* $(m_\infty^\lambda(\mathcal{U}), C_\infty^\lambda(\mathcal{U},\mathcal{U}), \lambda)$*. Then* $\max_{a \in \mathcal{A}} |A_n^\lambda(a) - A_\infty^\lambda(a)| \to 0$*. If* $a_\infty^\star = \arg\max_{a \in \mathcal{A}} A_\infty^\lambda(a)$ *is unique with margin* $\gamma = A_\infty^\lambda(a_\infty^\star) - \max_{a \neq a_\infty^\star} A_\infty^\lambda(a) > 0$*, then for all sufficiently large* n*,* $\arg\max_{a \in \mathcal{A}} A_n^\lambda(a) = a_\infty^\star$*.*

Proof. For readability, write $A_{n,\lambda} = A_n^\lambda$ and use the same subscript convention for m and C throughout this proof. For each action $a \in \mathcal{A}$,

$$A_{n,\lambda}(a) = F_a\left(m_{n,\lambda}(\mathcal{U}), C_{n,\lambda}(\mathcal{U},\mathcal{U}), \lambda\right).$$

By assumption,

$$m_{n,\lambda}(\mathcal{U}) \to m_{\infty,\lambda}(\mathcal{U}), \qquad C_{n,\lambda}(\mathcal{U},\mathcal{U}) \to C_{\infty,\lambda}(\mathcal{U},\mathcal{U}).$$

Since F_a is continuous at the limiting posterior,

$$A_{n,\lambda}(a) \to A_{\infty,\lambda}(a)$$

for each fixed $a \in \mathcal{A}$. Because $\mathcal{A}$ is finite, pointwise convergence implies uniform convergence:

$$\delta_n := \max_{a \in \mathcal{A}} |A_{n,\lambda}(a) - A_{\infty,\lambda}(a)| \to 0.$$

Now let

$$a_n \in \arg\max_{a \in \mathcal{A}} A_{n,\lambda}(a),$$

and let

$$a_\infty^\star \in \arg\max_{a \in \mathcal{A}} A_{\infty,\lambda}(a).$$

Because a_n maximizes $A_{n,\lambda}$,

$$A_{n,\lambda}(a_n) \geq A_{n,\lambda}(a_\infty^\star).$$

Therefore

$$\begin{aligned} A_{\infty,\lambda}(a_\infty^\star) - A_{\infty,\lambda}(a_n) &= [A_{\infty,\lambda}(a_\infty^\star) - A_{n,\lambda}(a_\infty^\star)] \\ &\quad + [A_{n,\lambda}(a_\infty^\star) - A_{n,\lambda}(a_n)] \\ &\quad + [A_{n,\lambda}(a_n) - A_{\infty,\lambda}(a_n)] \\ &\leq \delta_n + 0 + \delta_n = 2\delta_n. \end{aligned}$$

Thus the proxy-selected action is asymptotically optimal for the limiting acquisition objective.

Finally assume the limiting maximizer is unique with margin

$$\gamma = A_{\infty,\lambda}(a_\infty^\star) - \max_{a \neq a_\infty^\star} A_{\infty,\lambda}(a) > 0.$$

For large enough n,

$$\delta_n < \frac{\gamma}{2}.$$

For the optimal action,

$$A_{n,\lambda}(a^\star_\infty) \geq A_{\infty,\lambda}(a^\star_\infty) - \delta_n.$$

For any $a \neq a^\star_\infty$,

$$A_{n,\lambda}(a) \leq A_{\infty,\lambda}(a) + \delta_n \leq A_{\infty,\lambda}(a^\star_\infty) - \gamma + \delta_n.$$

Therefore

$$A_{n,\lambda}(a^\star_\infty) - A_{n,\lambda}(a) \geq \gamma - 2\delta_n > 0.$$

So no suboptimal action can overtake $a^\star_\infty$, and

$$\arg\max_{a\in\mathcal{A}} A_{n,\lambda}(a) = a^\star_\infty$$

for all sufficiently large n. □

Corollary 8 (Decision transfer at proxy-selected precision)**.** *For each width j, define $A^\lambda_j(a) = F_a(m^\lambda_j(\mathcal{U}), C^\lambda_j(\mathcal{U},\mathcal{U}), \lambda)$. Let $n_k \to \infty$ be proxy widths and let $N_k \to \infty$ be target widths. Suppose $\lambda^\star_{n_k} \to \lambda^\star$ is selected on each proxy either by validation NLL under Theorem 3 or by the fixed-checkpoint plug-in evidence score under Corollary 5. Suppose that on the finite input collection $\mathcal{U}$,*

$$\sup_{\lambda\in\Lambda}\|m^\lambda_n(\mathcal{U}) - m^\lambda_\infty(\mathcal{U})\| \to 0, \qquad \sup_{\lambda\in\Lambda}\|C^\lambda_n(\mathcal{U},\mathcal{U}) - C^\lambda_\infty(\mathcal{U},\mathcal{U})\| \to 0,$$

that the limiting posterior is continuous in λ, and that each F_a is jointly continuous in (m, C, λ). Then, uniformly over the finite action set,

$$A^{\lambda^\star_{n_k}}_{n_k}(a) \to A^{\lambda^\star}_\infty(a), \qquad A^{\lambda^\star_{n_k}}_{N_k}(a) \to A^{\lambda^\star}_\infty(a).$$

If the limiting best action is unique with the margin of Theorem 4, the proxy decision and the target decision evaluated at the transferred proxy precision agree for all sufficiently large k.

Proof. Apply uniform posterior convergence along the proxy sequence and the target sequence. Together with $\lambda^\star_{n_k} \to \lambda^\star$ and joint continuity, this gives both displayed limits for each action. Finiteness of the action set makes both convergences uniform. The margin argument in the proof of Theorem 4 makes both maximizers equal to the unique limiting maximizer for all sufficiently large k. □

Corollary 9 (Proxy-to-target decision transfer)**.** *Let n be a proxy width and N a target width. Define*

$$\epsilon_{n,N} = \max_{a\in\mathcal{A}} |A_{n,\lambda}(a) - A_{N,\lambda}(a)|.$$

If

$$a_n \in \arg\max_{a\in\mathcal{A}} A_{n,\lambda}(a)$$

and

$$a^\star_N \in \arg\max_{a\in\mathcal{A}} A_{N,\lambda}(a),$$

then

$$A_{N,\lambda}(a^\star_N) - A_{N,\lambda}(a_n) \leq 2\epsilon_{n,N}.$$

If the target model has margin

$$\gamma_N = A_{N,\lambda}(a^\star_N) - \max_{a\neq a^\star_N} A_{N,\lambda}(a) > 0$$

and

$$2\epsilon_{n,N} < \gamma_N,$$

then

$$a_n = a^\star_N.$$

Proof. The proof is identical to the near-optimality and margin parts of Theorem 4, replacing the infinite-width score $A_{\infty,\lambda}$ by the target-width score $A_{N,\lambda}$. □

K.7 Examples of acquisition functions covered

Posterior-variance acquisition. For scalar regression and $x \in \mathcal{P}$,

$$A_{n,\lambda}(x) = C_{n,\lambda}(x, x).$$

Since $C_{n,\lambda}(\mathcal{U},\mathcal{U}) \to C_{\infty,\lambda}(\mathcal{U},\mathcal{U})$, the marginal variance converges uniformly over $x \in \mathcal{P}$. For vector-valued outputs, the same argument covers

$$A_{n,\lambda}(x) = \operatorname{tr} C_{n,\lambda}(x, x)$$

and

$$A_{n,\lambda}(x) = \log\det\left(C_{n,\lambda}(x, x) + \tau^2 I\right), \qquad \tau^2 > 0,$$

because trace and log-determinant on uniformly positive-definite matrices are continuous.

Gaussian information gain. For a candidate batch $S \subseteq \mathcal{P}$ and observation noise $\tau^2 > 0$, define

$$A_{n,\lambda}(S) = \frac{1}{2}\log\det\left(I + \tau^{-2} C_{n,\lambda}(S, S)\right).$$

The submatrix $C_{n,\lambda}(S, S)$ converges to $C_{\infty,\lambda}(S, S)$, and

$$I + \tau^{-2} C_{n,\lambda}(S, S) \succeq I.$$

Hence the log-determinant is continuous and

$$A_{n,\lambda}(S) \to A_{\infty,\lambda}(S).$$

Because there are finitely many K-batches in a finite pool, convergence is uniform over all batches.

Integrated variance reduction. For scalar regression, a candidate batch $S \subseteq \mathcal{P}$, and a fixed target/evaluation set $T \subseteq \mathcal{U}$, define

$$A_{n,\lambda}(S) = \sum_{x \in T} C_{n,\lambda}(x, S)\left(C_{n,\lambda}(S, S) + \tau^2 I\right)^{-1} C_{n,\lambda}(S, x), \qquad \tau^2 > 0.$$

Since

$$C_{n,\lambda}(S, S) + \tau^2 I \succeq \tau^2 I,$$

matrix inversion is stable. Every factor in every summand converges, and T is finite. Therefore

$$A_{n,\lambda}(S) \to A_{\infty,\lambda}(S).$$

Predictive information gain between candidate and target sets. Let $S \subseteq \mathcal{P}$ be a candidate batch and let $T \subseteq \mathcal{U}$ be a fixed target set. For Gaussian regression with observation noise $\tau^2 > 0$, define

$$Y_S = f_S + \varepsilon_S, \qquad Y_T = f_T + \varepsilon_T,$$

with independent

$$\varepsilon_S, \varepsilon_T \sim \mathcal{N}(0, \tau^2 I).$$

The covariance of (Y_S, Y_T) is

$$\Sigma_{n,\lambda}(S, T) = \begin{pmatrix} C_{n,\lambda}(S, S) + \tau^2 I & C_{n,\lambda}(S, T) \\ C_{n,\lambda}(T, S) & C_{n,\lambda}(T, T) + \tau^2 I \end{pmatrix}.$$

All blocks converge, and the noise terms make the covariance uniformly positive definite. Therefore any mutual-information expression written as a finite sum and difference of log-determinants of these covariance matrices converges.

BALD-style classification acquisition. For classification and $x \in \mathcal{P}$, suppose the Laplace posterior over logits at x is

$$f_n(x) \sim \mathcal{N}\left(m_{n,\lambda}(x), C_{n,\lambda}(x, x)\right).$$

The BALD score is

$$A_{n,\lambda}(x) = H\left(\mathbb{E}_{f_n(x)}[p(y \mid f_n(x))]\right) - \mathbb{E}_{f_n(x)}\left[H(p(y \mid f_n(x)))\right].$$

The softmax map is continuous, and categorical entropy is continuous and bounded on the probability simplex. Since Gaussian means and covariances converge in finite dimension, the corresponding Gaussian laws converge weakly. Expectations of bounded continuous functions therefore converge, so

$$A_{n,\lambda}(x) \to A_{\infty,\lambda}(x).$$

On a finite pool, this convergence is uniform over $x \in \mathcal{P}$. This statement concerns the exact Gaussian expectations. It also holds for a fixed common-random-number Monte Carlo rule by continuity of its finite reparameterized sum. Independent fixed-sample draws across widths require a separate Monte Carlo error bound.

Top-K ranking agreement. Suppose individual scores $a_{n,\lambda}(x)$ satisfy

$$\max_{x\in\mathcal{P}}|a_{n,\lambda}(x) - a_{\infty,\lambda}(x)| \to 0.$$

Let the limiting scores be ordered as

$$a_{\infty,\lambda}(x_{(1)}) \geq a_{\infty,\lambda}(x_{(2)}) \geq \cdots \geq a_{\infty,\lambda}(x_{(M)}).$$

If the top-K margin

$$\Delta_K = a_{\infty,\lambda}(x_{(K)}) - a_{\infty,\lambda}(x_{(K+1)}) > 0,$$

then for sufficiently large n the width-n top-K set equals the limiting top-K set. Indeed, once the uniform score error is below $\Delta_K/2$, every limiting top-K candidate still has larger width-n score than every non-top-K candidate.

Selective prediction and detection. Suppose a detection or rejection score is

$$s_{n,\lambda}(x) = g\left(m_{n,\lambda}(x), C_{n,\lambda}(x,x)\right),$$

where g is continuous. Then

$$s_{n,\lambda}(x) \to s_{\infty,\lambda}(x)$$

uniformly on finite $\mathcal{U}$. For a threshold rule

$$\text{flag } x \quad \Longleftrightarrow \quad s_{n,\lambda}(x) > t,$$

exact flag agreement requires the threshold margin

$$\eta = \min_{x\in\mathcal{U}}|s_{\infty,\lambda}(x) - t| > 0.$$

Once the uniform score error is below η, the width-n and limiting rules make identical flag/no-flag decisions on all of $\mathcal{U}$.

Finite-round sequential active learning. The one-step result extends to finitely many acquisition rounds by induction. Suppose the following convergence and margin conditions hold conditionally along the common limiting acquisition history: at each round $r = 1, \ldots, R$,

$$m^{(r)}_{n,\lambda}(\mathcal{U}) \to m^{(r)}_{\infty,\lambda}(\mathcal{U}), \qquad C^{(r)}_{n,\lambda}(\mathcal{U},\mathcal{U}) \to C^{(r)}_{\infty,\lambda}(\mathcal{U},\mathcal{U}),$$

and the limiting acquisition rule at that round has a nonzero margin. Then the round-1 selected action agrees for all sufficiently large n. Conditional on that agreement, the updated training set is the same for the width-n and limiting processes, so the round-2 hypotheses apply on their common history. Repeating for finitely many rounds gives agreement of the entire finite acquisition trajectory.