UNDERSTANDING QUANTIZATION-AWARE TRAINING: GRADIENTS AT QUANTIZED WEIGHTS BIAS TO THE LOW-LOSS BASIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) converts a trained full-precision model into low-bit weights without task-level retraining, while quantization-aware training (QAT) incorporates quantization into the training loop. Although PTQ is efficient and often accurate at moderate bitwidths, it can fail sharply at aggressive bitwidths; QAT is more expensive but can often recover the lost accuracy. We propose a unified geometric framework that explains both PTQ failure and QAT recovery. We model full-precision training as following a low-loss *river* inside a wider *valley*: a normal neighborhood of the river forms a nearly flat *basin*, while leaving this basin incurs a sharp loss increase. When the quantization grid is comparable to the basin width, local PTQ objectives, including rounding and Hessian-based second-order reconstruction, can select a high-loss deployed quantized point outside the basin even when nearby low-loss quantized points exist. In this regime, straight-through-estimator-based QAT has a useful bias: it evaluates gradients at the deployed quantized weights while updating latent full-precision weights, causing the gradient to sense the valley wall and acquire an inward component that steers subsequent quantized iterates back into the basin. We formalize this mechanism through a local landscape model, construct a geometric PTQ failure mode, and prove finite-time QAT recovery under the landscape and quantizer-compatibility assumptions. Experiments across vision and language models under multiple quantization schemes show PTQ degradation followed by QAT recovery, with landscape diagnostics consistent with this mechanism.

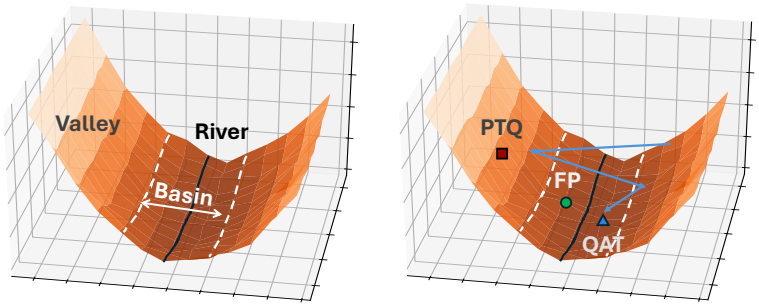
1 INTRODUCTION

Quantization reduces inference memory, latency, and energy consumption in large-scale models (Han et al., 2016; Sze et al., 2017; Jacob et al., 2018). Post-training quantization (PTQ) converts a pretrained full-precision model to low-bit weights without task-level fine-tuning, typically through local rounding or layerwise reconstruction objectives (Nahshan et al., 2021; Nagel et al., 2020; Frantar et al., 2023). PTQ is effective at moderate bitwidths, particularly around 8 bits (Krishnamoorthi, 2018; Nagel et al., 2020; Yuan et al., 2022). However, it often becomes fragile around 4 bits, especially when both weights and activations are quantized; even well-trained full-precision models can then suffer sharp performance degradation (Catalan-Tatjer et al., 2026; Liu et al., 2025b; Dremov et al., 2026).

A common remedy is quantization-aware training (QAT), which incorporates quantization into the training loop. In standard QAT with a straight-through estimator (STE), the forward pass uses quantized weights, while gradients evaluated at those weights update a latent full-precision copy. The STE bypasses the zero or undefined derivatives of discrete quantization (Bengio et al., 2013; Courbariaux et al., 2015; Jacob et al., 2018). [1] PTQ is fast and calibration-efficient but can lose accuracy at low precision; QAT is usually more robust but requires additional data, optimizer state, hyperparameter tuning, and compute. Practical pipelines therefore try PTQ first and use QAT fine-tuning if the quantized model misses the target accuracy.

[1] There are many variants of QAT beyond STE-based methods. In this paper, we focus on the standard STE-based formulation.

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.



(a) River–valley–basin landscape. (b) Quantization compatibility.

Figure 1: River–valley–basin loss geometry and quantization compatibility under QAT's gradient bias. (a) A 213M Llama loss landscape on SlimPajama shows a low-loss river inside a broader valley and basin. (b) PTQ can round a full-precision solution outside the basin; QAT can move the quantized model back.

Despite the success of PTQ–QAT pipelines, their mechanism remains less understood in the literature. When can PTQ fail even after full-precision training finds a good model? What does full-precision training contribute before QAT begins? Under the same training budget, why should QAT differ from continued full-precision training followed by PTQ? If QAT merely adds optimization steps, equal-budget full-precision fine-tuning followed by PTQ should reproduce its gains. If instead QAT corrects quantization-specific errors invisible to full-precision training, it can adapt latent weights to perform well after quantization where PTQ alone fails.

We study these questions through *quantization compatibility*: whether deployed low-bit weights lie in a low-loss region relative to the pretrained full-precision point. Our analysis uses a *river-valley-basin* view of the loss landscape (Figure 1(a)), supported by recent empirical and theoretical evidence (Ma et al., 2022; Wen et al., 2025; Chen et al., 2026; Bégout et al., 2026). The *river* is a low-loss manifold or trajectory along which full-precision training can move without large cost increases. Around the river lies a *basin* (dashed lines in Figure 1(a)) in which normal perturbations cause little loss change. Outside the basin, *valley walls* produce steep increases in normal directions.

PTQ can fail when the deployed quantized point lands on a valley wall outside the low-loss basin, sharply increasing loss (Figure 1(b)). When the grid resolution is comparable to basin width, PTQ may cross the boundary even though a nearby quantized point remains inside the basin with low loss. A nearly flat Hessian inside the basin can make local quadratic surrogates underestimate the boundary-crossing cost and miss basin-compatible quantized points. By contrast, QAT evaluates gradients at the quantized weights. Outside the basin, these gradients probe the steep valley wall, so the negative-gradient directions can have an inward normal component. Under suitable quantizer-compatibility conditions, applying these updates to the latent full-precision weights brings the quantized iterate into the basin in finite time. Ordinary full-precision fine-tuning may improve the unquantized model along the river while leaving the normal quantization error uncorrected. QAT can thus provide a quantization-specific correction by optimizing the deployed model, beyond the effect of additional compute. Our contributions are as follows:

- We formalize a river–valley–basin landscape and quantization-compatibility conditions. This identifies a PTQ failure mode: local rounding or second-order reconstruction can select a high-loss point outside the basin despite nearby low-loss quantized points.
- Under the stated landscape and quantizer-compatibility conditions, we prove finite-time entry of the quantized iterate into the basin, despite evaluating gradients at quantized weights and applying updates to latent full-precision weights.
- We evaluate the first-entry bound in a tractable toy model and study PTQ degradation, inward-gradient correction, and QAT recovery in the quantization of vision models and language models.

1.1 Related Work

PTQ and QAT. Early PTQ pipelines rely on rounding, calibration, and per-layer or per-channel scaling (Krishnamoorthi, 2018; Banner et al., 2019). More advanced methods minimize local

reconstruction objectives: loss-aware PTQ uses a local loss proxy (Nahshan et al., 2021), AdaRound learns rounding through a relaxed reconstruction objective (Nagel et al., 2020), and GPTQ applies a layerwise second-order reconstruction procedure for transformers (Frantar et al., 2023). Vision-transformer PTQ methods include twin uniform quantization in PTQ4ViT (Yuan et al., 2022) and Hessian-based reconstruction in APHQ-ViT (Wu et al., 2025). QAT instead quantizes the forward pass and typically uses an STE in the backward pass (Bengio et al., 2013; Courbariaux et al., 2015; Jacob et al., 2018). Empirically, QAT is often more robust at low precision, but it is more expensive than PTQ. Recent work studies hybrid PTQ–QAT pipelines and compute allocation between full-precision training and QAT fine-tuning (Catalan-Tatjer et al., 2026; Liu et al., 2025b; Dremov et al., 2026). Rather than propose a new quantization algorithm, we study when and why QAT can correct PTQ failure.

Theory of straight-through estimators. STE analyses typically impose structural assumptions on the quantizer, architecture, objective, or data distribution. For binary weights, Li et al. (2017) analyze STE-based SGD for finite-sum objectives and show convergence of averaged iterates to a neighborhood of the minimizer under strong convexity and smoothness conditions. Yin et al. (2019) study STE for two-layer networks with binary activations under Gaussian inputs and show that the expected coarse gradient can be a descent direction for the population objective. More recent work extends aspects of this analysis beyond population objectives (Jeong et al., 2025). Other work interprets quantized training through optimization and regularization. Bai et al. (2018) relate straight-through training to dual averaging for quantization-constrained optimization and propose ProxQuant, which uses proximal updates for regularized learning. Dockhorn et al. (2021) interpret BinaryConnect through dual averaging and generalized conditional gradients and introduce ProxConnect, a class of proximal-map quantizers. PARQ (Jin et al., 2025) uses convex piecewise-affine regularization to drive weights toward discrete values and aggregate proximal stochastic-gradient updates; it also casts STE as an asymptotic form of this framework. These results clarify STE's biased gradients and optimization interpretations but do not explain PTQ fragility after pretraining, QAT's pretrained initialization, or its advantage over equal-budget full-precision fine-tuning.

Neural network loss landscapes. Our landscape model builds on work on anisotropic neural network loss geometry. Ma et al. (2022) argue that multiscale, subquadratic growth can elude a single local quadratic approximation. Wen et al. (2025) propose a river-valley model to explain warmup-stable-decay learning rate schedules: high learning rates make progress along the river while oscillating across sharp valley walls, and learning rate decay reduces the off-river component. Chen et al. (2026) observe basin-like stability regions in large language models, where performance is nearly unchanged inside a basin but collapses outside it. Davis et al. (2025) show that a smooth function with subquadratic growth rate admits a smooth ravine manifold on which growth normal to the ravine is at least quadratic near the minimum. Bégout et al. (2026) establish that, near the solution, gradient descent aligns with the eigenvector corresponding to the smallest eigenvalue of the Hessian. Xu et al. (2025) empirically connect low-precision PTQ degradation to leaving the pretrained model's attractive basin and to misalignment between layerwise reconstruction and task loss; they also show improvements from quantization-aware fine-tuning. We analyze the latent-weight dynamics of STE-QAT and prove finite-time entry into a specified low-loss tube under landscape and quantizer-compatibility conditions.

2 River-valley-basin Landscape

Motivated by the empirical and theoretical evidence of the loss landscape of neural networks reviewed in Section 1.1, we now introduce a local model that isolates the geometry relevant to quantization. The model has three components: a low-loss *river* that represents the full-precision training trajectory, a surrounding *basin* in which moderate normal perturbations do not substantially change the loss, and sharper *valley walls* outside this basin that give the negative gradient an inward normal component. This abstraction is not intended as a global description of the full neural-network objective. Rather, it is a local description near a pretrained point, designed to separate tangential motion along the trained solution set from normal displacements induced by quantization. This separation is the key structure used later to explain both PTQ failure and QAT recovery.

We first introduce the local geometric notation. Let w_{fp} be the full-precision pretrained point, and let $\mathcal{M} \subset \mathbb{R}^d$ be an m-dimensional $C^{2,1}$ embedded manifold. For $\pi \in \mathcal{M}$, denote its tangent and normal spaces by $T_{\mathcal{M}}(\pi)$ and $N_{\mathcal{M}}(\pi)$. Let $\mathsf{H} \in \mathbb{R}^{d\times d}$ be symmetric positive definite, and write $\|x\|_{\mathsf{H}} = (x^\top \mathsf{H} x)^{1/2}$. Let $U = \{w \in \mathbb{R}^d \mid \mathrm{dist}(w, \mathcal{M}) < R_U\}$ be an open neighborhood of $\mathcal{M}$ containing w_{fp} on which the Euclidean nearest-point projection $P_{\mathcal{M}} : U \to \mathcal{M}$ is unique and L_P-Lipschitz, for some $L_P > 0$. For $R > 0$, define

$$\mathcal{T} = \{w \in U \mid \|w - P_{\mathcal{M}}(w)\|_{\mathsf{H}} \le R\}, \qquad \mathcal{T}(\pi) = \{w \in \mathcal{T} \mid P_{\mathcal{M}}(w) = \pi\}.$$

By definition, each displacement $(w - \pi)$ with $w \in \mathcal{T}(\pi)$ belongs to $N_{\mathcal{M}}(\pi)$.

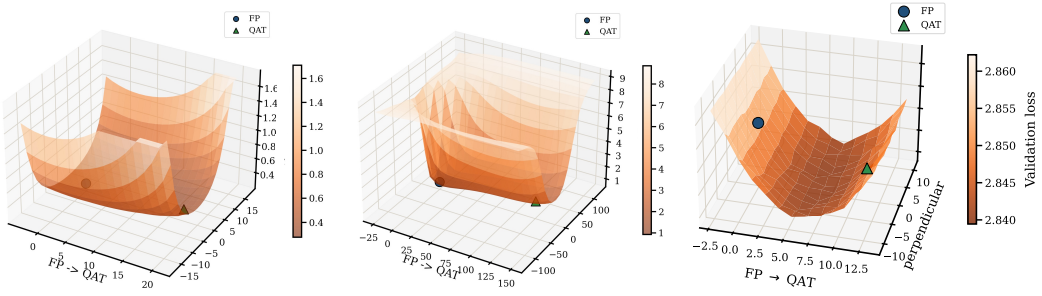
Assumption 1 (Local river-valley geometry)**.** *The function $f : U \to \mathbb{R}$ is L-smooth on U. The metric satisfies $\mathsf{H}N_{\mathcal{M}}(\pi) \subseteq N_{\mathcal{M}}(\pi)$ for every $\pi \in \mathcal{M}$. There exist constants $\epsilon_{\mathrm{flat}}, c_\perp > 0$ such that:*

- **(Anisotropic Basin)** *For every $\pi \in \mathcal{M}$ and $w \in \mathcal{T}(\pi)$, $|f(w) - f(\pi)| \le \epsilon_{\mathrm{flat}}$.*
- **(River center)** *For every $\pi \in \mathcal{M}$, $\nabla f(\pi) \in T_{\mathcal{M}}(\pi)$.*
- **(Sharp valley outside the basin)** *For every $w \in U \setminus \mathcal{T}$, defining the weighted normal direction $\nu_{\mathsf{H}}(w) = \mathsf{H}(w - P_{\mathcal{M}}(w))/\|\mathsf{H}(w - P_{\mathcal{M}}(w))\|$, we have $\langle \nabla f(w), \nu_{\mathsf{H}}(w)\rangle \ge c_\perp$.*

The *Anisotropic Basin* models the basin-like near-flatness observed in neural-network loss landscapes: within the anisotropic tube $\mathcal{T}(\pi)$ of radius R around a river point $\pi \in \mathcal{M}$, the loss changes by at most ϵ_{flat}. The metric H encodes anisotropic normal scaling, allowing some normal directions to be flatter than others; this is consistent with multiscale landscape structure (Ma et al., 2022) and with results that distinguish mild growth along a ravine from sharper transverse growth (Davis et al., 2025). The metric condition keeps weighted normal direction $\nu_{\mathsf{H}}(w)$ normal. It is automatic for $\mathsf{H} = cI$ with $c > 0$, but restricts a fixed anisotropic metric when the normal spaces rotate along a curved river. Its role is to give the inclusion $\nabla\big(\frac{1}{2}\|w - P_{\mathcal{M}}(w)\|_{\mathsf{H}}^2\big) = \mathsf{H}(w - P_{\mathcal{M}}(w)) \in N_{\mathcal{M}}(\pi)$.

The *River Center* condition requires the gradient on $\mathcal{M}$ to be tangential, formalizing the idea that full-precision training can continue moving along a low-loss river (Wen et al., 2025) and relating to talweg- and direction-based descriptions near minimizers (Bégout et al., 2026).

The *Sharp Valley* condition imposes sharp valley walls outside the basin: for every $w \in U \setminus \mathcal{T}$, the normal pairing $\langle \nabla f(w), \nu_{\mathsf{H}}(w)\rangle$ is uniformly bounded below by $c_\perp > 0$. This condition captures the sharp regime beyond basin stability observed in large language models (Chen et al., 2026) and provides the geometric input used later to prove finite-time recovery of the quantized iterate.



(a) ResNet on CIFAR-10 (b) ViT on ImageNet (c) Llama on SlimPajama

Figure 2: Local landscape diagnostics around pretrained checkpoints. Each panel plots the loss along one tangential and one normal direction relative to the QAT trajectory. The profiles are qualitatively consistent with Assumption 1: a near-flat basin in the normal direction and sharper growth outside it.

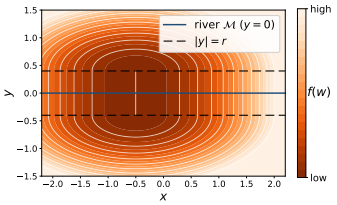
Figure 2 gives local landscape diagnostics around pretrained checkpoints for three architectures. The observed profiles are qualitatively consistent with the river-valley-basin geometry of Assumption 1. However, the two-dimensional slices in Figure 2 can only probe the *Anisotropic Basin* and *Sharp Valley* conditions along sampled directions; they do not provide a quantitative verification of other conditions such as *River Center*, nor do they determine the constants ϵ_{flat}, $c_\perp$, or R in the full parameter space.

To complement this empirical evidence, we present two analytically tractable examples. The first is a two-dimensional landscape that makes the river-valley-basin geometry visually transparent. The

second is an over-parameterized matrix factorization problem, which shows why an anisotropic normal metric naturally arises: some normal directions enter the loss quadratically, while others appear only through higher-order terms. For both examples, the detailed verification of Assumption 1 and the explicit constants are deferred to Appendices C.1 and D.1.

Example 1 (A two-dimensional river-valley-basin loss)**.** *Let* $w = (x, y) \in \mathbb{R}^2$*. Fix parameters* $u \in \mathbb{R}$*,* $\mu, r, \epsilon > 0$*, and define*

$$f(w) = \tfrac{1}{2}(x+u)^2 + \tfrac{\mu}{2}\big(\max\{|y| - r, 0\}\big)^2.$$



The river center is the horizontal line $\mathcal{M} = \{(x, 0) \mid x \in \mathbb{R}\}$*, and the basin is* $\mathcal{T} = \{(x, y) \in U \mid |y| \le R\}$ *with* $R = r + \sqrt{2\epsilon/\mu}$ *and* $U = \{(x, y) \in \mathbb{R}^2 \mid |y| < R_U\}$ *for any* $R_U > R$*. This example satisfies Assumption 1 on* U *with* $\mathsf{H} = \mathbb{I}$*,* $L_P = 1$*,* $L = \max\{1, \mu\}$*,* $\epsilon_{\text{flat}} = \epsilon$*, and* $c_\perp = \sqrt{2\mu\epsilon}$*.*

Notice that for fixed $\epsilon_{\text{flat}} = \epsilon$, increasing μ steepens the valley $c_\perp \propto \sqrt{\mu}$. That motivates calling μ the sharpness parameter of f. In section 3.1, we will use a rotated variant of this example to demonstrate failure of Hessian-based PTQ. The matrix factorization example below shows that the same geometry also appears in a standard over-parameterized nonconvex model, but with a genuinely anisotropic normal basin.

Example 2 (Over-parameterized matrix factorization)**.** *Assume* $M^\star = \mathrm{Diag}(D, \mathbf{0}) \in \mathbb{R}^{d\times d}$*, where* $D \in \mathbb{R}^{r\times r}$ *is positive definite and zero blocks have conformal dimensions. Consider*

$$f(X) = \|XX^\top - M^\star\|_F^2$$

with $X = \binom{P}{Z}$*, where* $P \in \mathbb{R}^{r\times k}$*,* $Z \in \mathbb{R}^{(d-r)\times k}$*, and* $k \ge r$*. Fix a full-row-rank matrix* P_0 *and a bounded open neighborhood* Ω *of* P_0 *such that all* P *in* $\overline{\Omega}$ *remain full-row-rank. Define*

$$\mathcal{M} = \left\{ \begin{pmatrix} P \\ \mathbf{0} \end{pmatrix} \,\middle|\, P \in \Omega \right\}, \qquad U = \left\{ \begin{pmatrix} P \\ Z \end{pmatrix} \,\middle|\, P \in \Omega,\ \|Z\|_F < R_U \right\}.$$

Expanding the objective gives $f(P, Z) = \|PP^\top - D\|_F^2 + 2\|PZ^\top\|_F^2 + \|ZZ^\top\|_F^2$*. Thus the normal geometry is anisotropic: the part of* Z *seen by the row space of* P *contributes quadratically through* $\|PZ^\top\|_F^2$*, whereas the row-null part first appears through the quartic term* $\|ZZ^\top\|_F^2$*. Freezing this row-space decomposition at* P_0*, one can choose a fixed positive definite metric* H *that weights the quadratic normal directions more heavily than the quartic ones. After possibly shrinking* Ω*, Assumption 1 holds on* U *with an anisotropic basin* $\mathcal{T}$*, tolerance* $\epsilon_{\text{flat}} = \epsilon$*, and some* $c_\perp > 0$*. A precise construction of* H*,* $\mathcal{T}$*, and* $c_\perp$ *is given in Appendix D.1.*

3 Quantization under the River-valley-basin Landscape

In this section, we analyze both PTQ and QAT under the river–valley–basin landscape introduced in Section 2. We first study how PTQ can fail when its deployed codeword crosses the basin boundary, then show how STE-based QAT can recover by updating latent full-precision weights using gradients evaluated at the quantized model.

3.1 Failure of Hessian-based PTQ

Many PTQ methods are motivated by a local second-order approximation around a fixed full-precision model. Given a pretrained point w_{fp} and a quantization codebook $\mathcal{Q}$, the idealized Hessian-based PTQ objective is

$$q_* \in \operatorname*{argmin}_{q\in\mathcal{Q}} \tfrac{1}{2}(q - w_{\text{fp}})^\top \nabla^2 f(w_{\text{fp}})(q - w_{\text{fp}}), \tag{1}$$

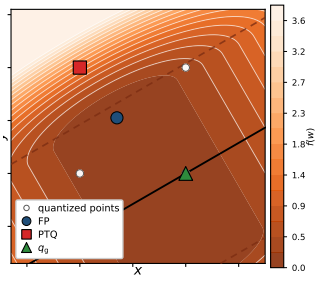
which is the second-order Taylor proxy for the loss function, with the linear term omitted because $\nabla f(w_{\text{fp}})$ is small near a well-trained model. In practice, directly solving (1) with the exact Hessian is typically computationally impractical: the Hessian is often replaced by a cheaper approximation, and the resulting discrete optimization over the quantization codebook is handled through relaxations or greedy/blockwise heuristics. For example, AdaRound (Nagel et al., 2020) derives a quadratic

rounding objective from a Taylor expansion and then optimizes a layerwise reconstruction relaxation, while GPTQ (Frantar et al., 2023) uses a layerwise output-reconstruction Hessian estimated from calibration data, together with damping and blockwise inverse-Hessian compensation. Thus, we treat (1) as a stylized model representing the core Hessian-based principle underlying these PTQ techniques. The failure mode we describe below, however, is not limited to this exact global objective; rather, it captures a broader limitation inherent to local PTQ surrogates.

When w_{fp} lies inside the flat basin, this quadratic proxy can be highly misleading. Within this band, the loss remains nearly constant in the normal direction, meaning the Hessian $\nabla^2 f(w_{\mathrm{fp}})$ has very small eigenvalues along that direction. As a result, the proxy treats normal-direction displacements as inexpensive, even though crossing the band boundary on the true landscape incurs a sharp increase in loss. At the same time, the Hessian may have substantial curvature along the tangent direction of the river, causing the proxy to penalize tangent movements toward downstream grid points that actually have low true loss. This also explains why simply spending more budget on ordinary full-precision training need not fix the PTQ failure. Once the full-precision iterate is inside the flat band, its gradient may mainly improve the along-river coordinate and need not provide an inward normal correction for the deployed quantized point. As a result, after additional full-precision steps, the final PTQ step can still select a quantized codeword outside the low-loss basin.

We rotate the function in Example 1 to show that the selected point of (1) can have arbitrarily larger loss than a nearby grid point inside the basin.

Revisiting Example 1: failure of Hessian-based PTQ. For any $\theta \in [\pi/6, \pi/4)$, let $R_\theta = \left(\begin{smallmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{smallmatrix}\right)$ denote the clockwise rotation matrix by θ and set $w_0 = (1, 0)$. Define $f_\theta(w) = f(R_\theta(w - w_0))$, where f is the loss in Example 1 with $\alpha = \cos\theta - \sin\theta$, $u \in (\frac{\alpha}{2}, \alpha)$, and $R < \cos\theta + \sin\theta$. The river center becomes $\mathcal{M} = \{w_0 + \lambda(\cos\theta, \sin\theta) \mid \lambda \in \mathbb{R}\}$ and the basin $\mathcal{T} = \{w \mid \mathrm{dist}(w, \mathcal{M}) \le R\}$. Consider uniform quantizer $\mathcal{Q} = \rho\{0, \pm 1, \cdots, \pm(2^{B-1} - 1)\}^2$ for $\rho = 1$ and $B \ge 2$. For any stationary point w_{fp} of f_θ at which the Hessian exists, the minimizer q_* of the Hessian proxy (1) lies outside the basin $\mathcal{T}$, even though $q_{\mathrm{g}} = (1, 0) \in \mathcal{T}$ is nearby. The loss gap $f_\theta(q_*) - f_\theta(q_{\mathrm{g}})$ grows at least linearly with the sharpness parameter μ.



Remark 1. *The failure holds throughout every stationary points satisfying $|z(w_{\mathrm{fp}})| < r$, not just a single w_{fp}. Appendix C.2 also proves the same failure for an open interval of grid spacings ρ around 1, so the failure mechanism is also robust to ρ.*

While the Hessian model (1) is intentionally stylized, it successfully isolates the local second-order logic common to practical PTQ algorithms. The previous example illustrates that this logic incorrectly ranks quantized candidates when the local curvature is nearly flat in the normal direction, completely ignoring the sharp rise in true loss that occurs just beyond the basin boundary.

3.2 QAT Recovers from PTQ Failure

We now give sufficient conditions for recovery when a fixed weight quantizer Q maps a pretrained checkpoint w_{fp} outside the basin. Starting from $w_0 = w_{\mathrm{fp}}$, consider the deterministic STE update:

$$w_{k+1} = w_k - \eta \nabla f(Q(w_k)).$$

The gradient is evaluated at the quantized weights $Q(w_k)$ and applied to the latent full-precision weights w_k. After T iterations, the deployed model is the quantized checkpoint $Q(w_T)$.

To proceed, we introduce the assumptions on the quantizer. Recall the notation for H and $P_{\mathcal{M}}$ in Assumption 1. We denote the largest and smallest eigenvalues of H by λ_1 and λ_d.

Assumption 2 (Quantizer compatibility)**.** *There exist constants $\rho, G > 0$ and $\kappa_{\mathcal{M}} \ge 0$ such that:*

(i) ***(Quantization error)*** *$\|Q(w) - w\| \le \rho$ for every $w \in U$.*

(ii) ***(Projection regularity and bounded gradients)*** *$\|\nabla f(w)\| \le G$ for every $w \in U$, and $P_{\mathcal{M}}$ is $C^{1,1}$ on U, i.e., there exists $\kappa_{\mathcal{M}} > 0$ such that*

$$\|\nabla P_{\mathcal{M}}(w) - \nabla P_{\mathcal{M}}(w')\|_{\mathrm{op}} \le \kappa_{\mathcal{M}} \|w - w'\|, \qquad \forall w, w' \in U.$$

(iii) ***(River width)*** $\rho(1+L_P)\sqrt{\lambda_1}\max\left\{\frac{G}{c_\perp}\sqrt{\frac{\lambda_1}{\lambda_d}},1\right\} < R < \sqrt{\lambda_d}(R_U-\rho)$, *where the parameters* L_P, $c_\perp$, R *and* R_U *are defined in Assumption 1.*

Here ρ bounds the Euclidean error of the entire weight vector. For coordinatewise nearest rounding with spacings h_j without clipping, a valid bound is $\rho = \frac{1}{2}\sqrt{\sum_j h_j^2}$. The result below does not require Q to be the Hessian-based selector of Section 3.1, but its quantizer-compatibility conditions (Assumption 2) must hold for the chosen map.

Theorem 1 (Recovery from PTQ failure under STE)**.** *Suppose Assumptions 1 and 2 hold. Let* $w_0 = w_{\mathrm{fp}} \in \mathcal{T}$ *with* $Q(w_0) \notin \mathcal{T}$*, and define* $T = \inf\{k \geq 0 \mid Q(w_k) \in \mathcal{T}\}$ *as the first time when the quantized iterate enters* $\mathcal{T}$*. If* $0 < \eta \leq \min\left\{\frac{c_\perp\sqrt{\lambda_d}R - \lambda_1(1+L_P)\rho G}{\lambda_1(1+L_P)G^2}, \frac{\rho}{G}\right\}$*, then we have*

$$T \leq 1 + \frac{\max\left\{\|w_0 - P_{\mathcal{M}}(w_0)\|_{\mathsf{H}}^2 - \left(R - \sqrt{\lambda_1}(1+L_P)\rho\right)^2,\ 0\right\}}{\eta\left(c_\perp\sqrt{\lambda_d}R - \lambda_1(1+L_P)\rho G\right)}.$$

The first-entry-time bound in Theorem 1 gives a cautious interpretation of scaling and bitwidth effects. If larger models have wider effective low-loss basins, i.e., larger R, as suggested by recent basin-visualization studies (Chen et al., 2026), then the upper bound for T predicts a shorter recovery phase, provided the other constants in the drift margin do not deteriorate. This offers a possible geometric explanation for empirical observations that larger models can be more tolerant to aggressive quantization (Dremov et al., 2026). Conversely, for a fixed quantizer family and dynamic range, lowering the bitwidth typically increases the quantization scale ρ, and the bound then predicts a longer first-entry time, and in the extreme case the drift margin may become nonpositive, in which case this theorem no longer guarantees recovery.

Tightness of the first-entry bound. Appendix C.3 evaluates this bound in a toy model. Across the reported configurations, the bound for T is between 2.47 and 3.81 times the observed entry time, and its η^{-1} dependence is attained. Holding the other constants and the initial normal distance fixed, the bound decreases with R and increases with ρ within the admissible regime.

To further interpret the quality of the recovered point, we separate into two regimes.

Corollary 1 (Near-stationary pre-trained checkpoint)**.** *Assume the conditions of Theorem 1. If* $\|\nabla f(\pi)\| \leq \epsilon$ *for every* $\pi \in \mathcal{M}$*, then for every stepsize* η *allowed by Theorem 1,*

$$f(Q(w_T)) \leq f(w_{\mathrm{fp}}) + 2\epsilon_{\mathrm{flat}} + \epsilon L_P(\eta T G + \rho).$$

Corollary 1 describes the common fine-tuning regime where the pretrained model has already been close to stationary on the river: after QAT brings the quantized iterate back into the basin, the recovered quantized loss is controlled by the basin flatness and the final quantization error. With the geometric and quantizer constants fixed, Theorem 1 keeps ηT bounded over admissible stepsizes, so the projected-loss contribution is $\mathcal{O}(\epsilon)$. The next corollary covers the case of an undertrained model, where the river direction still offers loss decrease.

Corollary 2 (Sub-optimal pre-trained checkpoint)**.** *Assume the conditions of Theorem 1. Let* $\mathcal{V} \subset \mathcal{M}$ *be a relatively open set containing* $P_{\mathcal{M}}(w_k)$ *and* $P_{\mathcal{M}}(Q(w_k))$ *for every* $0 \leq k < T$*. Suppose that there exists* $c_\parallel > 0$ *such that*

$$\langle \nabla f(P_{\mathcal{M}}(w)), \nabla f(w)\rangle > c_\parallel \|\nabla f(P_{\mathcal{M}}(w))\|, \quad \forall\, w \in U \text{ with } P_{\mathcal{M}}(w) \in \mathcal{V}.$$

For $\pi \in \mathcal{M}$*, denote the normalized gradient along the river by* $g(\pi) = \nabla f(\pi)/\|\nabla f(\pi)\|$*. If in addition,* $\eta \leq c_\parallel^2/\left[2LG^2(\kappa_{\mathcal{M}}(R/\sqrt{\lambda_d}+\rho)+1)^2\right]$*,* $\kappa_{\mathcal{M}} \leq \frac{c_\parallel}{4G(R/\sqrt{\lambda_d}+\rho)}$*, and there exists* κ *satisfying* $\kappa\rho L_P G \leq c_\parallel/4$ *such that* $\|g(\pi) - g(\pi')\| \leq \kappa\|\pi - \pi'\|$ *for every* $\pi, \pi' \in \mathcal{V}$*. Then,*

$$f(Q(w_T)) \leq f(w_{\mathrm{fp}}) + 2\epsilon_{\mathrm{flat}} + \rho L_P G - \tfrac{1}{4}\eta T c_\parallel^2.$$

The alignment condition in Corollary 2 implies $\|\nabla f(\pi)\| > c_\parallel$ on $\mathcal{M}$, so it describes a nonstationary sub-optimal regime rather than behavior through a near-stationary pretrained checkpoint. The negative term in Corollary 2 shows that, in this suboptimal regime, QAT can act as a descent method starting from w_{fp}, up to the flat-basin and quantization-error terms. This connects our geometric view to prior understanding of STE (Yin et al., 2019) as a descent direction for the population loss in two-layer networks with activation quantization; our result isolates a complementary mechanism in which the gradient is evaluated at the deployed quantized weights to correct quantization incompatibility.

4 EXPERIMENTS

We evaluate the PTQ-failure and QAT-recovery mechanism predicted by our landscape analysis on both vision and language tasks. Simulation results for the matrix factorization in Example 2 are deferred to Appendix D.2.

4.1 RESNET AND DEIT ON IMAGE CLASSIFICATION BENCHMARKS

We study ResNet-20/56 (He et al., 2016) on CIFAR-10 and DeiT-Tiny (Touvron et al., 2021) on ImageNet, quantizing the ResNet convolutional and linear weights and the DeiT transformer-block linear weights. We compare RTN, AdaRound (Nagel et al., 2020) or GPTQ, QAT, and continued FP + PTQ. RTN and QAT use the same PTQ-calibrated grid as the corresponding PTQ baseline. Detailed experimental settings and numerical results are in Appendices A and B.

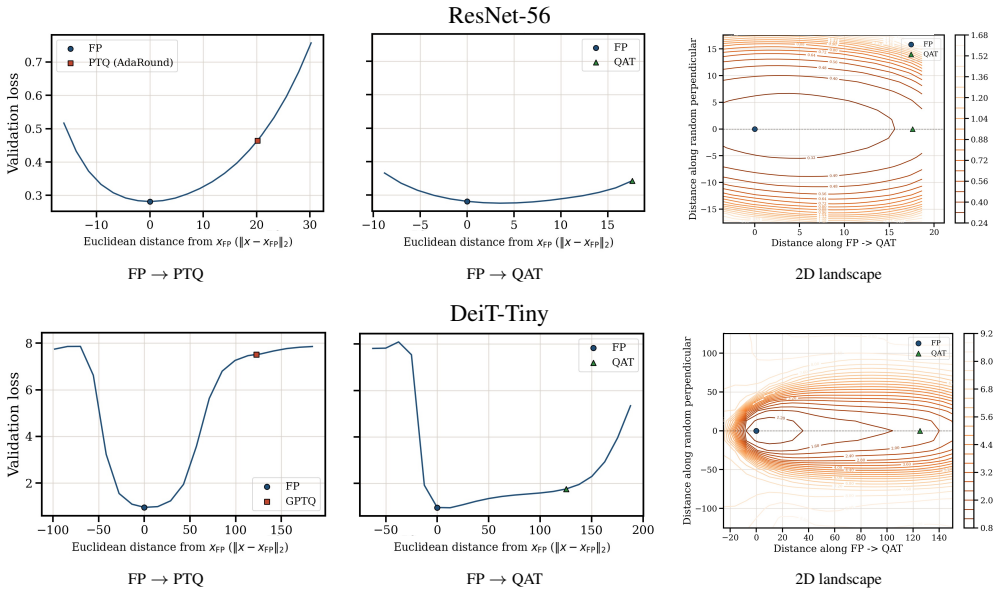


Figure 3: **ResNet/CIFAR-10 and DeiT/ImageNet landscape diagnostics under 2-bit quantization.** ResNet-56/CIFAR-10 (top) and DeiT-Tiny/ImageNet (bottom). Left: one-dimensional FP→PTQ loss profile (AdaRound or GPTQ, respectively). Middle: one-dimensional FP→QAT loss profile. Right: loss contour along FP→QAT and a randomly generated orthogonal direction.

At 2-bit precision, Figure 3 shows PTQ paths leaving the low-loss region while QAT follows lower-loss paths in both architectures. The flat-basin and sharp-valley pattern persists across five sampled normal directions (Appendix A). The 2-bit contrast is consistent with the quantization-specific correction predicted in Section 3.

Continued FP + PTQ. QAT's advantage over this control is largest at 2-bit precision and narrows at higher bitwidths; in the 4-bit ResNet-20 case, the control slightly outperforms QAT.

4.2 LLAMA ON SLIMPAJAMA

We test the same mechanism in language modeling with a Llama-style 213M model pretrained on SlimPajama for $60{,}000$ iterations under a cosine learning rate schedule. We apply weight-only E2M1 FP4 to the transformer-block linear layers, leaving the token embedding, language-model head, and RMSNorm parameters in bf16. GPTQ is used as the PTQ baseline. The full sequential GPTQ procedure gives a validation loss of 2.878. With its fitted FP4 quantizer held fixed, direct quantization of the same FP checkpoint gives QAT a step-zero loss of 2.893; after $3{,}000$ steps of QAT with a constant learning rate, the deployed loss reaches 2.843. Experimental details appear in Appendix A.

Figure 4 reports one-dimensional loss profiles from the converged full-precision checkpoint toward the GPTQ and QAT anchors, together with a two-dimensional contour plot around the FP→QAT direction. The QAT displacement is much shorter than the GPTQ displacement in parameter-space

Euclidean norm. The displacements are also nearly orthogonal (about 92°), so QAT follows a distinct adaptation direction rather than a small correction along GPTQ rounding. The FP→GPTQ profile rises monotonically, with no local minimum near deployment, consistent with rounding outside the basin. The FP→QAT profile drops below the full-precision baseline before rising past the QAT anchor, indicating a move into a lower-loss region before leaving it. In the two-dimensional slice, loss stays relatively flat along the QAT path and rises away from it.

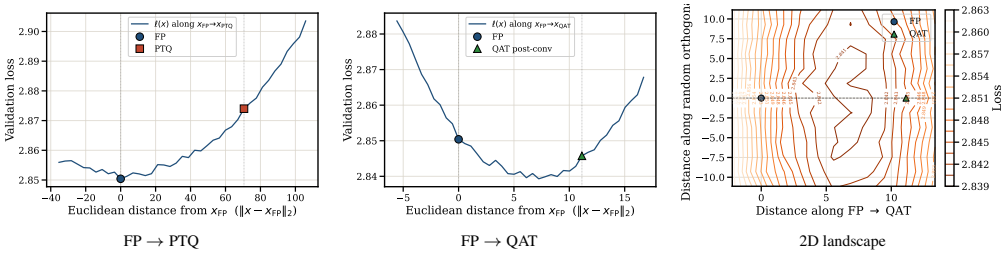


Figure 4: **Llama/SlimPajama landscape diagnostics at FP4.** Left: one-dimensional loss profile from the converged FP checkpoint toward the GPTQ anchor. Middle: one-dimensional loss profile from the converged FP checkpoint toward the post-convergence QAT anchor. Right: two-dimensional loss contours in the plane spanned by the FP→QAT direction and a randomly orthogonal direction.

Let w_k be the latent weights and $q_k = Q_0(w_k)$ the deployed weights. The measured pairing $s_k = \langle \nabla f(q_k), (q_k - w_k)/\|q_k - w_k\| \rangle$ is positive at all 11 logged checkpoints, while deployed validation loss falls from 2.893 to 2.845 by step 2,400 (Table 5 in Appendix B.3). If $q_k - w_k$ approximates the outward normal at steps outside the basin, this sign agrees with Assumption 1. Across three independently pretrained checkpoints and, separately, four calibration draws at one fixed checkpoint, GPTQ-FP4 increases validation loss relative to the matched FP model, while QAT reduces it below that baseline in every run (Appendix B.3).

Additional PTQ configurations. We also test AWQ (Lin et al., 2024) and SpinQuant (Liu et al., 2025a) alongside GPTQ at 4-bit weight-only precision. Across all four configurations, PTQ raises validation loss above the common pretrained FP checkpoint, whereas QAT with each fitted quantizer fixed lowers it below FP (Table 3 in Appendix B.3).

Continued FP + PTQ. Starting from the same pretrained checkpoint and matching QAT's additional training-token budget, continued FP training followed by GPTQ-FP4 yields validation loss 2.8593, versus 2.8430 for QAT (Appendix A). Continued FP training achieves 2.8283 before quantization, so QAT's advantage here concerns the deployed quantized model.

5 Discussion

Our theoretical and empirical analysis makes several testable predictions on the model quantization. First, PTQ failure should be most visible when the quantization grid is comparable to the basin width: coarser grids are more likely to cross the boundary, while sufficiently fine grids should remain in the flat band. Second, QAT can substantially reduce the loss of the quantized model while changing the latent full-precision loss only slightly, reflecting a correction of quantization compatibility rather than ordinary full-precision optimization. Third, equal-budget full-precision fine-tuning followed by PTQ can remain fragile when its updates move mainly along the river and do not reduce the normal quantization error.

Limitations. Our landscape model only gives sufficient local conditions for QAT recovery under deterministic STE updates with a fixed weight quantizer. Empirically, our experiments rely on slice-based diagnostics of the loss landscape, which illustrate the phenomenon but do not offer a complete verification of the global, high-dimensional geometry. Furthermore, our analysis of Hessian-based PTQ abstracts complex, practical layerwise techniques into a simplified local second-order proxy. Consequently, this model does not account for practical deployment factors such as activation quantization, dynamic scale learning, gradient clipping.

AI USE STATEMENT

We used ChatGPT (GPT-6 Sol) to help check and simplify proofs and improve the clarity and organization of the manuscript. In particular, it helped simplify the proofs of Theorem 1 and Corollary 1 and refine the example demonstrating failure of Hessian-based PTQ. The authors first constructed a toy model in which the loss gap between the Hessian-based PTQ solution and a nearby low-loss quantized point grows with the sharpness parameter μ, for a fixed full-precision reference point and quantization grid spacing. Through several rounds of discussion with ChatGPT, the authors extended the construction to show that the failure persists over a set of full-precision reference points and an interval of grid spacings. The authors reviewed and revised the AI-assisted material and take full responsibility for the final text, theoretical claims, proofs, and experimental results.

ETHICS STATEMENT

This work does not involve recruiting human participants or collecting new personal data. A better understanding of PTQ and QAT may help guide more resource-efficient model deployment.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 state the theoretical assumptions and results. The appendix provides proofs of the main results and verifies the landscape assumptions for the illustrative examples. Appendix A reports the settings for the vision and language experiments, including training hyperparameters, quantization schemes, calibration procedures, and hardware. Appendix B provides additional numerical results and comparisons across repeated runs, while Appendix D.2 describes the matrix-factorization simulations.

REFERENCES

Yu Bai, Yu-Xiang Wang, and Edo Liberty. Proxquant: Quantized neural networks via proximal operators. In *International Conference on Learning Representations*, 2018.

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pascal Bégout, Jérôme Bolte, Thomas Mariotti, and Francisco Silva. Gradient extremals, talwegs, valleys, and directional alignment for generic gradient descent. *arXiv preprint arXiv:2604.11213*, 2026.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Albert Catalan-Tatjer, Niccolò Ajroldi, and Jonas Geiping. Training dynamics impact post-training quantization robustness. In *International Conference on Learning Representations*, 2026.

Huanran Chen, Yinpeng Dong, Zeming Wei, Yao Huang, Yichi Zhang, Hang Su, and Jun Zhu. Unveiling the basin-like loss landscape in large language models. In *International Conference on Learning Representations*, 2026.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. BinaryConnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*, volume 28, pp. 3123–3131, 2015.

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Gradient descent with adaptive stepsize converges (nearly) linearly under fourth-order growth. *Mathematical Programming*, pp. 1–66, 2025.

Tim Dockhorn, Yaoliang Yu, Eyyüb Sari, Mahdi Zolnouri, and Vahid Partovi Nia. Demystifying and generalizing binaryconnect. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13202–13216, 2021.

Aleksandr Dremov, David Grangier, Angelos Katharopoulos, and Awni Hannun. Compute-Optimal Quantization-Aware Training. In *International Conference on Learning Representations*, 2026.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018.

Halyun Jeong, Jack Xin, and Penghang Yin. Beyond discreteness: Finite-sample analysis of straight-through estimator for quantization. *arXiv preprint arXiv:2505.18113*, 2025.

Lisa Jin, Jianhao Ma, Zechun Liu, Andrey Gromov, Aaron Defazio, and Lin Xiao. PARQ: Piecewise-affine regularized quantization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 28044–28062. PMLR, 2025.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.

Hao Li, Soham De, Zheng Xu, Christoph Studer, Hanan Samet, and Tom Goldstein. Training quantized nets: A deeper understanding. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6: 87–100, 2024.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In *International Conference on Learning Representations*, volume 2025, pp. 92009–92032, 2025a.

Zechun Liu, Changsheng Zhao, Hanxian Huang, Sijia Chen, Jing Zhang, Jiawei Zhao, Scott Roy, Lisa Jin, Yunyang Xiong, Yangyang Shi, et al. ParetoQ: Improving scaling laws in extremely low-bit LLM quantization. In *Advances in Neural Information Processing Systems*, 2025b.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *arXiv preprint arXiv:2204.11326*, 2022.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pp. 7197–7206, 2020.

Yury Nahshan, Brian Chmiel, Chaim Baskin, Evgenii Zheltonozhskii, Ron Banner, Alex M Bronstein, and Avi Mendelson. Loss aware post-training quantization. *Machine Learning*, 110(11):3245–3262, 2021.

Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Kaiyue Wen, Zhiyuan Li, Jason S Wang, David Leo Wright Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In *International Conference on Learning Representations*, 2025.

Zhuguanyu Wu, Jiayi Zhang, Jiaxin Chen, Jinyang Guo, Di Huang, and Yunhong Wang. APHQ-ViT: Post-training quantization with average perturbation Hessian based reconstruction for vision transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9686–9695, 2025.

Zifei Xu, Sayeh Sharify, Wanzin Yazar, Tristan Webb, and Xin Wang. Understanding the difficulty of low-precision post-training quantization for llms. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

P Yin, J Lyu, S Zhang, S Osher, YY Qi, and J Xin. Understanding straight-through estimator in training activation quantized neural nets. *International Conference on Learning Representations*, 2019.

Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. PTQ4ViT: Post-training quantization for vision transformers with twin uniform quantization. In *European Conference on Computer Vision*, pp. 191–207. Springer, 2022.

A Experimental Details

ResNet/CIFAR-10. Each full-precision ResNet is trained for 200 epochs with SGD, momentum 0.9, batch size 128, weight decay $2 \cdot 10^{-4}$, initial learning rate 0.1, and cosine annealing. We quantize all convolutional and linear weights with per-output-channel signed uniform scales. AdaRound uses 32 calibration batches of size 128, i.e. 4096 calibration images, and 10,000 rounding-optimization iterations. RTN uses the same AdaRound-fitted per-channel scales but no learned rounding. QAT starts from the epoch-180 checkpoint, keeps the AdaRound quantizer fixed, and trains for 20 epochs with SGD at learning rate 10^{-4}, dropping by 0.5 at epochs 10 and 16. The continued FP + PTQ control trains in full precision from epoch 180 to 200 and then applies AdaRound. All ResNet-20 and ResNet-56 runs use a single NVIDIA RTX A5000 GPU per run.

DeiT/ImageNet. For DeiT-Tiny, we quantize the transformer-block linear layers at 2-, 3-, and 4-bit precision and leave the classification head full precision. GPTQ uses 128 calibration images, group size 128, and no activation ordering. RTN uses the same fitted grid/group scales without GPTQ error correction. QAT starts from the pretrained full-precision checkpoint with the GPTQ quantizer frozen and runs for 10 epochs with AdamW at learning rate $5 \cdot 10^{-4}$. The continued FP + PTQ control resumes full-precision training from the same pretrained checkpoint with the same additional training budget as QAT, then refits GPTQ to the resulting weights. All DeiT/ImageNet runs use four NVIDIA A40 GPUs per run.

Vision landscape diagnostics. For the 2-bit diagnostics in Figure 3, five independently sampled directions orthogonal to FP→QAT showed similar flat-basin and sharp-valley profiles. The figure displays one of these directions.

Llama/SlimPajama-6B. We pretrain a 213M-parameter Llama-style decoder (24 layers, 12 heads, hidden size 768, RMSNorm) on SlimPajama-6B for 60,000 iterations with AdamW (weight decay 0.1, $\beta = (0.9, 0.95)$, gradient clip 1.0), a cosine schedule with base learning rate 10^{-3} and 300 warmup steps, sequence length 512, batch size 50×4 (50 per-device with gradient-accumulation 4, total batch 200) on a single H200, and bf16. GPTQ is applied weight-only to every `nn.Linear` inside the transformer blocks — the `q_proj`/`k_proj`/`v_proj`/`o_proj` attention projections and the `gate_proj`/`up_proj`/`down_proj` MLP projections — at FP4 (E2M1, with levels $\{0, \pm 0.5, \pm 1, \pm 1.5, \pm 2, \pm 3, \pm 4, \pm 6\}$) with per-output-channel signed scales, input-column groups of size 128, GPTQ block size 128, and percent-damp 0.01; the token embedding, the language-model head, and the RMSNorm parameters remain in bf16. Calibration draws 128 sequences of length 512 from the SlimPajama train split with a fixed seed. The GPTQ-fitted per-Linear FP4 quantizer (scales and group structure) is persisted and reused as the STE fake-quant in both QAT regimes, so PTQ and QAT share an identical FP4 grid. Continuation QAT resumes from the iter-48,000 (80%) FP checkpoint and runs the same cosine schedule (and the same data ordering, controlled by the seed) to iter 60,000, i.e. 12,000 STE iterations along the cosine LR tail. Post-convergence QAT initializes weights from the iter-60,000 FP checkpoint with a fresh optimizer, uses a constant learning rate of 10^{-5} with no warmup or decay, and trains for 3,000 STE iterations (5% of pretraining). The continued FP + PTQ control starts from the same iter-60,000 checkpoint, matches QAT's additional training-token budget, and then refits GPTQ-FP4. Both use the next-token-prediction loss without distillation; the control's final calibration tokens are additional to the matched training budget. Unless otherwise noted, Llama experiments use a single seed; the GPTQ-FP4/QAT repeats are detailed in Appendix B.3. The landscape figures show a representative run, and the river-cross diagnostic averages over three random perpendicular directions.

B Additional Experimental Results

B.1 ResNet/CIFAR-10

Table 1: ResNet/CIFAR-10 equal-budget comparison. Each method entry reports mean validation loss / top-1 accuracy over five seeds, with nonzero standard deviations in parentheses. The depth column reports the epoch-180 full-precision result. FP-200+AdaRound is the equal-budget full-precision control followed by AdaRound PTQ. RTN and QAT use the same AdaRound-calibrated weight grid as the AdaRound baseline. Bold marks the best quantized result in each row, separately for loss and accuracy.

Depth (FP-180)	Bits	RTN	AdaRound	QAT	FP-200 + AdaRound
20 (0.329 / 91.96)	2-bit	9.365 / 10.41	0.497 (0.017) / 86.62 (0.49)	**0.407** (0.022) / **87.92** (0.67)	0.576 (0.030) / 84.51 (0.85)
	3-bit	1.201 / 73.96	0.358 (0.004) / **91.19** (0.02)	**0.334** (0.003) / **91.19** (0.13)	0.362 (0.002) / 91.08 (0.09)
	4-bit	0.430 / 89.68	0.337 (0.001) / 91.79 (0.09)	0.335 (0.001) / 91.75 (0.07)	**0.330** (0.001) / **91.87** (0.08)
56 (0.281 / 93.50)	2-bit	5.212 / 13.88	0.448 (0.010) / 88.83 (0.30)	**0.337** (0.006) / **90.50** (0.17)	0.449 (0.015) / 89.07 (0.27)
	3-bit	0.618 / 86.21	0.314 (0.006) / 92.77 (0.15)	**0.288** (0.005) / 92.80 (0.10)	0.299 (0.005) / **93.04** (0.14)
	4-bit	0.334 / 92.60	0.289 (0.001) / 93.30 (0.09)	**0.278** (0.001) / 93.42 (0.04)	0.280 (0.001) / **93.46** (0.05)

Figure 5 provides the additional ResNet-20 seed-1 interpolation diagnostics for 3- and 4-bit precision, and Figure 6 provides the corresponding ResNet-56 profiles. The 2-bit diagnostics are shown in the main body in Figure 3.

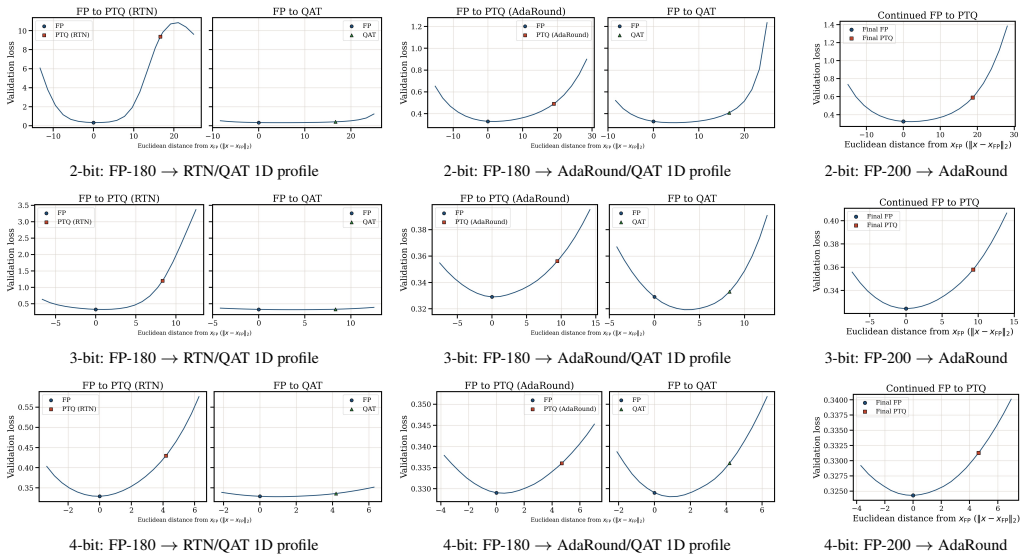


Figure 5: Additional ResNet-20 on CIFAR-10 interpolation diagnostics. 3- and 4-bit seed-1 interpolation profiles from the pretrained ResNet checkpoint toward PTQ, QAT, and final-PTQ endpoints for ResNet-20.

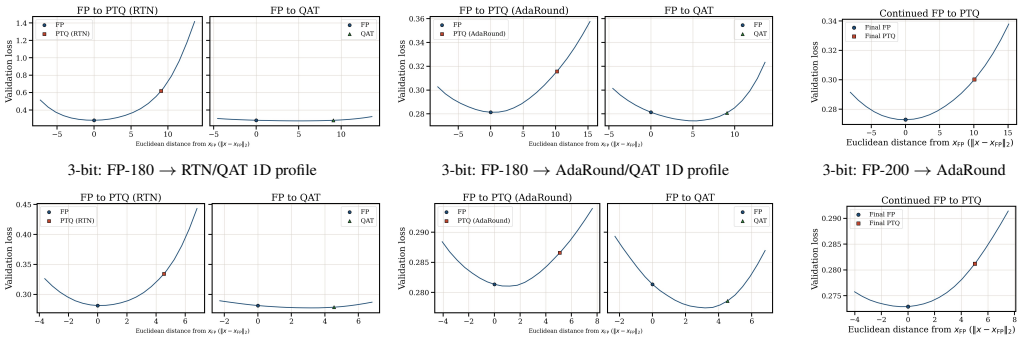


Figure 6: Additional ResNet-56 on CIFAR-10 interpolation diagnostics. 3- and 4-bit seed-1 interpolation profiles from the pretrained ResNet-56 checkpoint toward PTQ, QAT, and final-PTQ endpoints.

B.2 DeiT/ImageNet

Table 2: DeiT/ImageNet equal-budget comparison. Each method entry reports validation loss / top-1 accuracy. RTN, GPTQ, and QAT report means over three random seeds with standard deviations in parentheses, while FPFT+GPTQ is the available seed-0 equal-budget control. RTN and QAT use the same GPTQ-fitted weight grid as the GPTQ baseline. Bold marks the best quantized result in each row, separately for loss and accuracy.

pre-trained FP	Bits	RTN	GPTQ	QAT	fine-tuned FP + GPTQ
1.220 / 72.14	2-bit	7.216 (0.071) / 0.16 (0.05)	7.313 (0.084) / 0.19 (0.05)	**2.121** (0.011) / **52.78** (0.18)	7.328 / 0.22
	3-bit	3.166 (0.088) / 37.52 (1.28)	2.062 (0.029) / 56.78 (0.39)	**1.439** (0.003) / **66.60** (0.12)	2.021 / 56.73
	4-bit	1.402 (0.003) / 68.60 (0.08)	1.318 (0.002) / 70.17 (0.03)	**1.245** (0.003) / **71.01** (0.17)	1.261 / 70.86

Figure 7 provides the one-dimensional interpolation profiles for the 3- and 4-bit GPTQ and RTN anchors used in Table 2, together with the equal-budget FPFT+GPTQ control. The 2-bit diagnostics are shown in the main body in Figure 3.

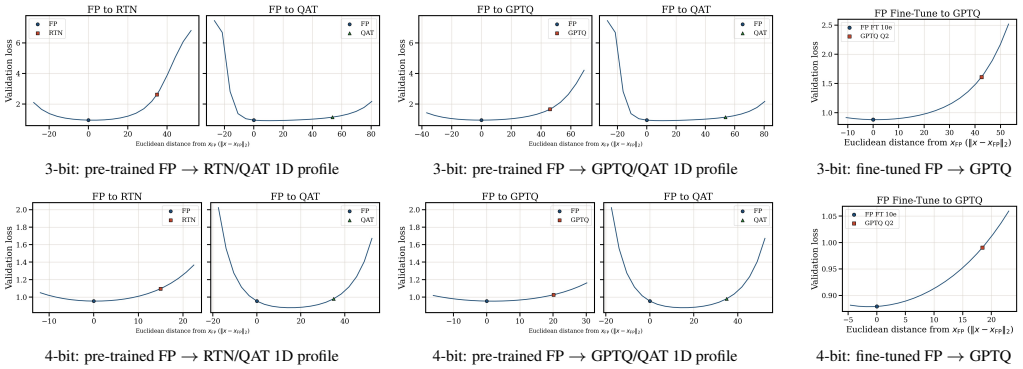


Figure 7: DeiT/ImageNet interpolation diagnostics. Rows correspond to 3- and 4-bit precision. The first column uses RTN, the second uses GPTQ, and the third column interpolates from the equal-budget full-precision fine-tuned checkpoint toward the refit GPTQ endpoint. These plots are diagnostics for the PTQ-failure/QAT-recovery pattern and are not used as direct proof of the full high-dimensional geometry.

B.3 LLAMA/SLIMPAJAMA

Table 3: **4-bit weight-only quantization on Llama-213M.** All configurations share the FP checkpoint (loss 2.8525), calibration draw, group size 128, and 3,000-step QAT budget. Each fitted quantizer remains fixed during QAT.

PTQ method	Weight format	PTQ loss ↓	QAT loss ↓
GPTQ	FP4 (E2M1)	2.8778	2.8430
GPTQ	INT4	2.8835	2.8465
AWQ	INT4	2.8910	2.8442
SpinQuant	INT4	2.9049	2.8482

Robustness across seeds and calibration draws. We repeat the GPTQ-FP4/QAT pipeline across three independently pretrained checkpoints, varying initialization and data order, and separately across four 128-sequence GPTQ calibration draws at one fixed checkpoint. Each fitted grid remains fixed during its paired QAT run. Table 4 reports the results: PTQ loss exceeds the matched FP loss and QAT loss falls below it in all three pretraining runs and all four calibration draws.

Table 4: **GPTQ-FP4/QAT robustness on Llama-213M.** Values are mean ± standard deviation; the FP checkpoint is fixed in the calibration-draw study. The angle $\theta = \angle(q_{\mathrm{PTQ}} - w_{\mathrm{fp}}, q_{\mathrm{QAT}} - w_{\mathrm{fp}})$ compares the two deployed-weight displacements from the starting FP weights.

Variation	FP loss	GPTQ-FP4 loss	QAT loss	θ
Pretraining seeds ($n = 3$)	2.8575 ± 0.0044	2.8810 ± 0.0029	2.8483 ± 0.0047	$92.0^\circ \pm 0.1^\circ$
Calibration draws ($n = 4$)	2.8525	2.8771 ± 0.0008	2.8430 ± 0.0003	$92.07^\circ \pm 0.01^\circ$

Recovery trajectory. Let w_k be the latent weights and $q_k = Q_0(w_k)$ their deployed weights at iterate k of QAT phase under the GPTQ-fitted quantizer. Table 5 tracks validation loss along the QAT trajectory.

Table 5: **GPTQ-FP4 QAT trajectory on Llama-213M.** Selected checkpoints are shown. The pairing $s_k = \langle \nabla f(q_k), (q_k - w_k)/\|q_k - w_k\| \rangle$ uses the raw task-loss gradient at the deployed weights. All loss columns report validation loss. $Q_{\mathrm{GPTQ}}(w_k)$ denotes reapplying full sequential GPTQ to the latent checkpoint. Values are independently rounded. The final step-3,000 QAT loss is reported in Table 3.

Step	$f(q_k)$	$f(q_k) - f(w_k)$	s_k	$f(Q_{\mathrm{GPTQ}}(w_k))$
0	2.8926	0.0401	1.4×10^{-3}	2.8778
50	2.8720	0.0177	8.3×10^{-4}	2.8744
300	2.8602	0.0083	6.1×10^{-4}	2.8699
1200	2.8501	0.0024	4.8×10^{-4}	2.8638
2400	2.8446	-0.0007	4.9×10^{-4}	2.8601

All 11 measured pairings are positive: the raw gradient at q_k has a component along $q_k - w_k$, so its negative opposes the quantization displacement. Across the displayed checkpoints, deployed loss falls from 2.8926 to 2.8446, and the deployed-minus-latent gap changes from 0.0401 to -0.0007. If $q_k - w_k$ approximates the outward normal at steps outside the basin, these signs agree with the sharp-valley condition.

At step 2,400, reapplying GPTQ to the latent weights gives loss 2.8601, compared with 2.8446 for the persisted-grid QAT model. Thus refitting GPTQ to the same latent weights does not reproduce the improvement obtained on the fixed grid.

C TOY EXAMPLES

Example 3 (Generalization of the two-dimensional river-valley-basin loss)**.** *Fix an angle* $\theta \in [\pi/6, \pi/4)$ *and write* $\tau = (\cos\theta, \sin\theta)$, $\nu = (-\sin\theta, \cos\theta)$, $\alpha = \cos\theta - \sin\theta$, *and* $\delta = \cos\theta + \sin\theta$.

Let $w_0 = (1, 0)$ and define $t(w) = \langle \tau, w - w_0 \rangle$ and $z(w) = \langle \nu, w - w_0 \rangle$. For parameters $\mu, r, \epsilon > 0$ and $u \in \mathbb{R}$, define $R = r + \sqrt{2\epsilon/\mu}$ and consider

$$f_\theta(w) = \tfrac{1}{2}\big(t(w) + u\big)^2 + \tfrac{\mu}{2}\big(\max\{|z(w)| - r, 0\}\big)^2. \tag{2}$$

The river center is $\mathcal{M} = \{w_0 + \lambda\tau \mid \lambda \in \mathbb{R}\}$. For any tubular neighborhood $U = \{w \in \mathbb{R}^2 \mid \mathrm{dist}(w, \mathcal{M}) < R_U\}$ with $R_U > R$, the basin is $\mathcal{T} = \{w \in U \mid |z(w)| \le R\}$.

C.1 Verification of Assumption 1

Claim: Example 3 satisfies Assumption 1 on $U = \{w \in \mathbb{R}^2 \mid \mathrm{dist}(w, \mathcal{M}) < R_U\}$ for $R_U > R$, with $\mathsf{H} = \mathbb{I}$, $L_P = 1$, $L = \max\{1, \mu\}$, $\epsilon_{\mathrm{flat}} = \epsilon$, and $c_\perp = \sqrt{2\mu\,\epsilon}$.

Proof. Since (τ, ν) is an orthonormal basis, the nearest-point projection onto $\mathcal{M}$ is $P_{\mathcal{M}}(w) = w_0 + t(w)\tau$, so $P_{\mathcal{M}}$ is affine and $L_P = 1$. The loss in (2) is L-smooth with $L = \max\{1, \mu\}$. For any $\pi = w_0 + \lambda\tau \in \mathcal{M}$, we have $t(\pi) = \lambda$ and $z(\pi) = 0$, hence $\nabla f_\theta(\pi) = (\lambda + u)\tau \in T_{\mathcal{M}}(\pi)$, which verifies the river-center condition. Next, if $w = \pi + \xi\nu$ with $|\xi| \le R$, then $t(w) = t(\pi)$ and $z(w) = \xi$. Therefore $|f_\theta(w) - f_\theta(\pi)| = \frac{\mu}{2}(\max\{|\xi| - r, 0\})^2 \le \frac{\mu}{2}(R - r)^2 = \epsilon$. For $w \in U \setminus \mathcal{T}$, i.e., $|z(w)| > R$, the outward normal direction is $\nu_{\mathsf{H}}(w) = \mathrm{sgn}(z(w))\nu$. Thus

$$\langle \nabla f_\theta(w), \nu_{\mathsf{H}}(w) \rangle = \mu(|z(w)| - r) \ge \mu(R - r) = \sqrt{2\mu\,\epsilon} = c_\perp.$$

This proves that the toy example satisfies Assumption 1. □

C.2 Failure of Hessian-based PTQ

Claim. Consider the function f_θ in Example 3, with $u \in (\alpha/2, \alpha)$ and $R < \delta$. For $\rho > 0$ and $B \ge 2$, let

$$\mathcal{Q}_{\rho,B} = \rho\{0, \pm 1, \ldots, \pm(2^{B-1} - 1)\}^2, \qquad q_{\mathrm{g}}(\rho) = (\rho, 0).$$

There is an open interval $\mathcal{I}_{\mathrm{fail}} \subset (0, \infty)$ containing 1 such that, for every $\rho \in \mathcal{I}_{\mathrm{fail}}$ and every $B \ge 2$:

(a) $q_{\mathrm{g}}(\rho) \in \mathcal{T}$. For any stationary checkpoint w_{fp} at which the Hessian exists, every minimizer $q_*(\rho)$ of (1) over $\mathcal{Q}_{\rho,B}$ lies outside $\mathcal{T}$.

(b) Every such minimizer satisfies

$$f_\theta(q_*(\rho)) - f_\theta(q_{\mathrm{g}}(\rho)) \ge \frac{\mu}{8}(\delta - r)^2 - \frac{1}{2}\big(u + \cos\theta(\rho - 1)\big)^2.$$

With parameters θ, u, r, ϵ fixed and $r < \delta$, the interval can be chosen independently of μ for all sufficiently large μ.

Proof. Write $c = \cos\theta$, $s = \sin\theta$, $\alpha = c - s$, and $\delta = c + s$. In the coordinates of Example 3,

$$t(w) = c(w_1 - 1) + sw_2, \qquad z(w) = -s(w_1 - 1) + cw_2,$$

so

$$f_\theta(w) = \frac{1}{2}(t(w) + u)^2 + \frac{\mu}{2}\big(\max\{|z(w)| - r, 0\}\big)^2.$$

The stationary points are exactly $t(w) = -u$, $|z(w)| \le r$. At $|z| = r$ the normal second derivative does not exist: its one-sided values are 0 and μ. At every stationary point with $|z| < r$, the Hessian is $\tau\tau^\top$, where $\tau = (c, s)$, and hence the proxy equals

$$S(q) = \frac{1}{2}(q - w_{\mathrm{fp}})^\top \tau\tau^\top (q - w_{\mathrm{fp}}) = \frac{1}{2}(t(q) + u)^2.$$

In particular, S does not depend on which admissible stationary checkpoint is chosen.

Set $\zeta = (R + \delta)/2$, so $r < R < \zeta < \delta$. For $(i, j) \in \mathbb{Z}^2$, write $q_{ij}(\rho) = \rho(i, j)$ and

$$t_{ij}(\rho) = c(\rho i - 1) + s\rho j, \qquad z_{ij}(\rho) = -s(\rho i - 1) + c\rho j.$$

Orthonormality of (c, s) and $(-s, c)$ gives

$$t_{ij}(\rho)^2 + z_{ij}(\rho)^2 = (\rho i - 1)^2 + (\rho j)^2.$$

The two points $q_{\mathrm{g}}(\rho) = (\rho, 0)$ and $q_{\mathrm{b}}(\rho) = (0, \rho)$ belong to $\mathcal{Q}_{\rho,B}$ for every $B \geq 2$. At $\rho = 1$, the strict inequalities $s|\rho - 1| < r$, $s + c\rho > R$, and $c - s\rho - u > 0$ hold because $r > 0$, $R < \delta$, and $u < \alpha$. By continuity, choose an open interval $\mathcal{I}_{\mathrm{fail}}$ around 1 on which all three remain true. Then

$$|z(q_{\mathrm{g}}(\rho))| = s|\rho - 1| < r, \qquad z(q_{\mathrm{b}}(\rho)) = s + c\rho > R,$$

and

$$|t(q_{\mathrm{b}}(\rho)) + u| = c - s\rho - u, \qquad S(q_{\mathrm{b}}(\rho)) = \frac{1}{2}(c - s\rho - u)^2.$$

We show that every grid point with $|z| < \zeta$ has proxy value strictly larger than $S(q_{\mathrm{b}}(\rho))$. First let $\mathcal{N} = \{(1,0), (0,0), (2,0), (1,1), (1,-1)\}$. Since $0 < u < \alpha < s < c$, at $\rho = 1$ the values of $|t_{ij}(1) + u|$ for these five pairs, in the listed order, are

$$u, \quad c - u, \quad c + u, \quad s + u, \quad s - u.$$

Here $s > \alpha$ follows from $\theta \geq \pi/6$, which gives $2s > c$. Each of the five values exceeds $\alpha - u$: for the first use $2u > \alpha$, and for the other four use $c > \alpha$ and $s > \alpha$. Since $\mathcal{N}$ is finite, continuity permits shrinking the interval so that

$$|t_{ij}(\rho) + u| > c - s\rho - u, \qquad \forall\, (i,j) \in \mathcal{N},\ \rho \in \mathcal{I}_{\mathrm{fail}}.$$

For $(i,j) \notin \mathcal{N}$, we have $\|(i-1, j)\| \geq \sqrt{2}$. Thus, simultaneously for all such pairs and every $\rho > 0$,

$$\|(\rho i - 1, \rho j)\| = \|\rho(i-1, j) + (\rho - 1, 0)\| \geq \rho\sqrt{2} - |\rho - 1|.$$

At $\rho = 1$, the identity $\delta^2 + \alpha^2 = 2$ and $\zeta < \delta$ give $\sqrt{2} > \sqrt{\zeta^2 + \alpha^2}$. Another shrinking of $\mathcal{I}_{\mathrm{fail}} \subset (0, \infty)$ ensures

$$\rho\sqrt{2} - |\rho - 1| > \sqrt{\zeta^2 + (c - s\rho)^2}, \qquad \forall\, \rho \in \mathcal{I}_{\mathrm{fail}}.$$

Consequently,

$$(\rho i - 1)^2 + (\rho j)^2 > \zeta^2 + (c - s\rho)^2, \qquad \forall\, (i,j) \notin \mathcal{N}.$$

If also $|z_{ij}(\rho)| < \zeta$, then $t_{ij}(\rho)^2 > (c - s\rho)^2$. Since $c - s\rho > u > 0$,

$$|t_{ij}(\rho) + u| \geq |t_{ij}(\rho)| - u > c - s\rho - u.$$

Combining the two cases yields

$$|z(q_{ij}(\rho))| < \zeta \quad \Longrightarrow \quad S(q_{ij}(\rho)) > S(q_{\mathrm{b}}(\rho)).$$

Every proxy minimizer $q_*(\rho)$ satisfies $S(q_*(\rho)) \leq S(q_{\mathrm{b}}(\rho))$ because $q_{\mathrm{b}}(\rho)$ is feasible. Therefore $|z(q_*(\rho))| \geq \zeta > R$, proving part (a). All interval choices above are independent of B.

Since $|z(q_{\mathrm{g}}(\rho))| < r$, its loss is $f_\theta(q_{\mathrm{g}}(\rho)) = \frac{1}{2}(u + c(\rho - 1))^2$. For every proxy minimizer, $f_\theta(q_*(\rho)) \geq \frac{\mu}{2}(\zeta - r)^2$. Finally, $\zeta - r = (R + \delta)/2 - r > (\delta - r)/2$ gives

$$f_\theta(q_*(\rho)) - f_\theta(q_{\mathrm{g}}(\rho)) \geq \frac{\mu}{8}(\delta - r)^2 - \frac{1}{2}(u + c(\rho - 1))^2,$$

which proves part (b).

For uniformity as μ increases, fix θ, u, r, ϵ with $r < \delta$, and choose μ_0 sufficiently large that $R_0 = r + \sqrt{2\epsilon/\mu_0} < \delta$. Carry out the preceding interval construction using R_0 and $\zeta_0 = (R_0 + \delta)/2$. For every $\mu \geq \mu_0$, the basin radius satisfies $R(\mu) = r + \sqrt{2\epsilon/\mu} \leq R_0$. The proxy S and the condition $|z(q_{\mathrm{g}}(\rho))| < r$ are unchanged, and every proxy minimizer still satisfies $|z(q_*(\rho))| \geq \zeta_0 > R_0 \geq R(\mu)$. Its loss therefore obeys the same lower bound in part (b). This proves the claimed uniform interval and linear lower bound. □

C.3 Tightness of the first-entry-time bound in the toy model

We evaluate the first-entry-time bound of Theorem 1 in an instance of Example 3. Fix the tangential coordinate at $t = -u$ and set $\mu = 1$. The resulting one-dimensional objective is

$$\phi(z) = \frac{1}{2}\big(\max\{|z| - r, 0\}\big)^2, \qquad z \in U_{\mathrm{sc}} = (-R_U, R_U).$$

For this scalar problem, take $\mathcal{M}_{\mathrm{sc}} = \{0\}$, $\mathcal{T}_{\mathrm{sc}} = [-R, R]$, and $0 < r < R < R_U$. The projection is constant, so valid constants are

$$\mathsf{H} = 1, \quad L_P = \kappa_{\mathcal{M}} = 0, \quad L = 1, \quad \epsilon_{\mathrm{flat}} = \frac{1}{2}(R-r)^2, \quad c_\perp = R - r, \quad G = R_U - r.$$

In particular, these are the constants of the scalar reduction. This comparison concerns recovery under a compatible fixed quantizer; the scalar quantizer below is distinct from the Hessian-proxy selector in Appendix C.2.

Quantizer and exact entry time. For a quantization-error radius $\rho > 0$, set

$$a = R + 0.1\rho, \qquad z_0 = R - 0.1\rho, \qquad Q_\rho(z) = a + 2\rho\ \mathrm{round}\left(\frac{z-a}{2\rho}\right),$$

with ties rounded to the smaller grid point. The grid spacing is 2ρ, and $|Q_\rho(z) - z| \le \rho$. Thus $z_0 \in \mathcal{T}_{\mathrm{sc}}$, whereas $Q_\rho(z_0) = a > R$. We run

$$z_{k+1} = z_k - \eta\phi'(Q_\rho(z_k)).$$

Before entry, $Q_\rho(z_k) = a$, and hence $z_k = z_0 - k\eta(R - r + 0.1\rho)$. The deployed iterate first enters $\mathcal{T}_{\mathrm{sc}}$ when the latent iterate reaches the rounding threshold $a - \rho$. For the admissible parameters below, $\eta G \le \rho$ prevents skipping the adjacent quantization cell, and $a - 2\rho \in [-R, R]$. Therefore the exact entry time is

$$T_{\mathrm{obs}} = \left\lceil \frac{0.8\rho}{\eta(R - r + 0.1\rho)} \right\rceil. \tag{3}$$

Upper bound in Theorem 1. Substituting the scalar constants into Theorem 1 gives

$$T_{\mathrm{bd}} = 1 + \frac{(R - 0.1\rho)^2 - (R - \rho)^2}{\eta\big[(R-r)R - \rho G\big]}. \tag{4}$$

The width and step-size conditions reduce to

$$\frac{\rho G}{R - r} < R < R_U - \rho, \qquad 0 < \eta \le \min\left\{\frac{(R-r)R - \rho G}{G^2},\ \frac{\rho}{G}\right\}.$$

All configurations in Table 6 satisfy these conditions.

Table 6: **Exact entry times and theorem bounds in the scalar toy model.** The baseline is $r = 0.2$, $R = 1$, $R_U = 1.5$, $\rho = 0.1$, and $\eta = 0.005$. Each row varies only the indicated parameter; the definitions of a and z_0 are updated accordingly. Ratios use unrounded bound values.

Parameter	Value	T_{obs}	T_{bd}	$T_{\mathrm{bd}}/T_{\mathrm{obs}}$
η	0.005	20	51.8	2.59
	0.010	10	26.4	2.64
	0.020	5	13.7	2.74
ρ	0.10	20	51.8	2.59
	0.20	40	119.7	2.99
	0.30	58	221.0	3.81
R	0.8	27	77.6	2.88
	1.0	20	51.8	2.59
	1.2	16	39.5	2.47

The bound is between 2.47 and 3.81 times the observed entry time in these configurations. For fixed r, R, R_U, ρ, both T_{obs} and T_{bd} grow as η^{-1} as $\eta \downarrow 0$, so the step-size dependence in Theorem 1 is attained in this example. This establishes sharp step-size dependence for this construction.

D Matrix Factorization

D.1 Local matrix-factorization geometry

We verify the geometric inequalities in Assumption 1 for Example 2 on its local product neighborhood. All constants depend on the chosen patch. The calculation establishes local normal geometry; the neighborhood differs from the full distance tube used in Theorem 1.

Proof. Write $X = \binom{P}{Z}$ with $P \in \mathbb{R}^{r\times k}$ and $Z \in \mathbb{R}^{(d-r)\times k}$. Then

$$f(P,Z) = \|PP^\top - D\|_F^2 + 2\|PZ^\top\|_F^2 + \|ZZ^\top\|_F^2.$$

Let

$$\mathcal{M} = \left\{ \begin{pmatrix} P \\ 0 \end{pmatrix} \,\middle|\, P \in \Omega \right\}, \qquad U = \left\{ \begin{pmatrix} P \\ Z \end{pmatrix} \,\middle|\, P \in \Omega,\ \|Z\|_F < R_U \right\}.$$

On this local product neighborhood, the nearest-point projection is $P_{\mathcal{M}}\binom{P}{Z} = \binom{P}{0}$, so $L_P = 1$ and $N_{\mathcal{M}}(P,0) = \binom{0}{\mathbb{R}^{(d-r)\times k}}$.

Since Ω is bounded and every $P \in \overline{\Omega}$ has full row rank, define $\mu = \inf_{P\in\overline{\Omega}} \sigma_{\min}(P) > 0$ and $M = \sup_{P\in\overline{\Omega}} \|P\|_{\mathrm{op}} < \infty$. For each full-row-rank matrix P, let $\Pi_P = P^\top(PP^\top)^{-1}P$ be the orthogonal projector onto the row space of P. Fix the base projector $\Pi_0 = P_0^\top(P_0P_0^\top)^{-1}P_0$. Choose positive weights $\lambda_{\mathrm{s}} \geq \lambda_{\mathrm{f}} > 0$ as follows:

$$\lambda_{\mathrm{f}} = \max\left\{ \frac{4R^2}{R_U^2}, \frac{2R^2}{\sqrt{\epsilon_{\mathrm{flat}}}} \right\}, \qquad \lambda_{\mathrm{s}} = \max\left\{ \lambda_{\mathrm{f}}, \frac{8M^2R^2}{\epsilon_{\mathrm{flat}}} \right\}.$$

Define $C_0 = \lambda_{\mathrm{s}}\Pi_0 + \lambda_{\mathrm{f}}(I - \Pi_0)$. After vectorizing matrix pairs, define the linear operator

$$\mathsf{H}(\Delta P, Z) = (\Delta P, ZC_0).$$

Since $C_0 \succeq \lambda_{\mathrm{f}} I$, the matrix H is symmetric positive definite. Moreover, it is block diagonal with respect to the splitting $(\Delta P, Z)$, so $\mathsf{H}N_{\mathcal{M}}(P,0) \subseteq N_{\mathcal{M}}(P,0)$.

Since the map $P \mapsto \Pi_P$ is continuous on the full-row-rank set, we may shrink Ω around P_0 so that $\sup_{P\in\Omega} \|\Pi_P - \Pi_0\|_{\mathrm{op}} \leq \delta_{\mathsf{H}}$, where $\delta_{\mathsf{H}} > 0$ will be chosen below. The bounds $\sigma_{\min}(P) \geq \mu_\Omega$ and $\|P\|_{\mathrm{op}} \leq M_\Omega$ remain valid on the smaller patch.

Smoothness. For the full variable X, $\nabla f(X) = 4(XX^\top - M^\star)X$. For any perturbation E,

$$\nabla^2 f(X)[E] = 4(EX^\top + XE^\top)X + 4(XX^\top - M^\star)E.$$

Thus, $\|\nabla^2 f(X)[E]\|_F \leq (12\|X\|_{\mathrm{op}}^2 + 4\|M^\star\|_{\mathrm{op}})\|E\|_F$. On U, we have $\|X\|_{\mathrm{op}}^2 \leq M^2 + R_U^2$ and $\|M^\star\|_{\mathrm{op}} = \|D\|_{\mathrm{op}}$. Therefore f is L-smooth on U with $L = 12(M^2 + R_U^2) + 4\|D\|_{\mathrm{op}}$.

River center. At a point on the river center, the river-center condition holds because

$$\nabla f(P,0) = \begin{pmatrix} 4(PP^\top - D)P \\ 0 \end{pmatrix} \in T_{\mathcal{M}}(P,0).$$

Metric comparison after freezing the row space. For each $P \in \Omega$, define the moving metric

$$C_P = \lambda_{\mathrm{s}}\Pi_P + \lambda_{\mathrm{f}}(I - \Pi_P).$$

Let $D_0(Z)^2 = \langle Z, ZC_0\rangle$ and $D_P(Z)^2 = \langle Z, ZC_P\rangle$. Because $C_0 - C_P = (\lambda_{\mathrm{s}} - \lambda_{\mathrm{f}})(\Pi_0 - \Pi_P)$, we have $\|C_0 - C_P\|_{\mathrm{op}} \leq \lambda_{\mathrm{s}}\delta_{\mathsf{H}}$. Also $C_P \succeq \lambda_{\mathrm{f}} I$, hence $\lambda_{\mathrm{f}}\|Z\|_F^2 \leq D_P(Z)^2$. Therefore

$$|D_0(Z)^2 - D_P(Z)^2| = |\langle Z, Z(C_0 - C_P)\rangle| \leq \lambda_{\mathrm{s}}\delta_{\mathsf{H}}\|Z\|_F^2 \leq \frac{\lambda_{\mathrm{s}}\delta_{\mathsf{H}}}{\lambda_{\mathrm{f}}} D_P(Z)^2.$$

Choosing $\delta_{\mathsf{H}} \leq \frac{\lambda_{\mathrm{f}}}{4\lambda_{\mathrm{s}}}$, we obtain the uniform comparison

$$\frac{3}{4}D_P(Z)^2 \leq D_0(Z)^2 \leq \frac{5}{4}D_P(Z)^2. \tag{5}$$

Anisotropic basin. Define the basin by

$$\mathcal{T} = \left\{ \begin{pmatrix} P \\ Z \end{pmatrix} \in U \,\middle|\, D_0(Z) = \|Z\|_{C_0} \leq R \right\}.$$

The choice $\lambda_{\mathrm{f}} \geq 4R^2/R_U^2$ guarantees that $D_0(Z) \leq R$ implies $\|Z\|_F \leq R/\sqrt{\lambda_{\mathrm{f}}} \leq R_U/2$, so $\mathcal{T} \subset U$. Suppose $D_0(Z) \leq R$. By the metric comparison (5), $D_P(Z)^2 \leq \frac{4}{3}R^2$. Decompose

$Z = Z_{\mathrm{s}} + Z_{\mathrm{f}}$ with $Z_{\mathrm{s}} = Z\Pi_P$ and $Z_{\mathrm{f}} = Z(I - \Pi_P)$. Since $P(I - \Pi_P) = 0$, we have $PZ_{\mathrm{f}}^\top = 0$, and hence

$$f(P, Z) - f(P, 0) = 2\|PZ_{\mathrm{s}}^\top\|_F^2 + \|ZZ^\top\|_F^2.$$

Using $\|P\|_{\mathrm{op}} \le M$, $D_P(Z)^2 \ge \lambda_{\mathrm{s}}\|Z_{\mathrm{s}}\|_F^2$, and $D_P(Z)^2 \ge \lambda_{\mathrm{f}}\|Z\|_F^2$, we obtain

$$\begin{aligned} f(P, Z) - f(P, 0) &\le 2M^2\frac{D_P(Z)^2}{\lambda_{\mathrm{s}}} + \frac{D_P(Z)^4}{\lambda_{\mathrm{f}}^2} \\ &\le \frac{8M^2R^2}{3\lambda_{\mathrm{s}}} + \frac{16R^4}{9\lambda_{\mathrm{f}}^2} \;\le\; \frac{\epsilon_{\mathrm{flat}}}{3} + \frac{4\epsilon_{\mathrm{flat}}}{9} = \frac{7\epsilon_{\mathrm{flat}}}{9} \end{aligned}$$

Thus $|f(P, Z) - f(P, 0)| \le \epsilon_{\mathrm{flat}}$ for all $(P, Z) \in \mathcal{T}$. This verifies the anisotropic-basin condition. It remains to verify the sharp-wall condition for the fixed metric C_0.

Step 1. We first prove the estimate for the moving metric C_P. Notice that

$$\begin{aligned} \langle \nabla_Z f(P, Z), ZC_P\rangle &= \langle 4ZP^\top P + 4ZZ^\top Z,\; ZC_P\rangle \\ &= 4\lambda_{\mathrm{s}}\langle Z_{\mathrm{s}}P^\top P,\; Z_{\mathrm{s}}\rangle + 4\langle ZZ^\top Z, ZC_P\rangle \\ &\ge 4\lambda_{\mathrm{s}}\|PZ_{\mathrm{s}}^\top\|_F^2 + 4\lambda_{\mathrm{f}}\|ZZ^\top\|_F^2. \end{aligned}$$

Since $\sigma_{\min}(P) \ge \mu$ on Ω, $\|PZ_{\mathrm{s}}^\top\|_F \ge \mu\|Z_{\mathrm{s}}\|_F$. Also, with $m_0 = \min\{d - r, k\}$, it follows from Cauchy-Schwarz inequality that $m_0\|ZZ^\top\|_F^2 \ge \|Z\|_F^4$. Thus

$$\langle \nabla_Z f(P, Z), ZC_P\rangle \ge 4\lambda_{\mathrm{s}}\mu^2\|Z_{\mathrm{s}}\|_F^2 + \frac{4\lambda_{\mathrm{f}}}{m_0}\|Z\|_F^4.$$

If $D_P(Z) = \sqrt{\lambda_{\mathrm{s}}\|Z_{\mathrm{s}}\|_F^2 + \lambda_{\mathrm{f}}\|Z_{\mathrm{f}}\|_F^2} > R$, then either $\sqrt{2\lambda_{\mathrm{s}}}\|Z_{\mathrm{s}}\|_F \ge D_P(Z)$ or $\sqrt{2\lambda_{\mathrm{f}}}\|Z_{\mathrm{f}}\|_F \ge D_P(Z)$. In the first case,

$$\frac{\langle \nabla_Z f(P, Z), ZC_P\rangle}{D_P(Z)} \ge 2\mu^2 D_P(Z) \ge 2\mu^2 R.$$

In the second case, $\|Z\|_F \ge \|Z_{\mathrm{f}}\|_F \ge D_P(Z)/\sqrt{2\lambda_{\mathrm{f}}}$, and therefore

$$\frac{\langle \nabla_Z f(P, Z), ZC_P\rangle}{D_P(Z)} \ge \frac{D_P(Z)^3}{m_0\lambda_{\mathrm{f}}} \ge \frac{R^3}{m_0\lambda_{\mathrm{f}}}.$$

Hence, whenever $D_P(Z) > R$,

$$\frac{\langle \nabla_Z f(P, Z), ZC_P\rangle}{D_P(Z)} \ge \min\left\{2\mu^2 R,\; \frac{R^3}{m_0\lambda_{\mathrm{f}}}\right\} \triangleq c_*.$$

Step 2. Next, we turn to the estimate for the fixed metric C_0. Now suppose $(P, Z) \in U \setminus \mathcal{T}$, i.e. $D_0(Z) > R$. By the metric comparison (5), $D_P(Z) \ge 2D_0(Z)/\sqrt{5} > 2R/\sqrt{5}$. Applying the moving-metric estimate with threshold $2R/\sqrt{5}$ gives

$$\frac{\langle \nabla_Z f(P, Z), ZC_P\rangle}{D_0(Z)} \ge \frac{16}{25}c_*.$$

We next control the error caused by replacing C_P with the frozen metric C_0. On U,

$$\|\nabla_Z f(P, Z)\|_F \le 4\|Z\|_F\|P\|_{\mathrm{op}}^2 + 4\|Z\|_F^3 \le 4R_U(M^2 + R_U^2) \triangleq G.$$

Since $D_0(Z) \ge \sqrt{\lambda_{\mathrm{f}}}\|Z\|_F$, we have

$$\left|\frac{\langle \nabla_Z f(P, Z), Z(C_0 - C_P)\rangle}{D_0(Z)}\right| \le \frac{\|\nabla_Z f(P, Z)\|_F\|Z\|_F\|C_0 - C_P\|_{\mathrm{op}}}{D_0(Z)} \le \frac{G\lambda_{\mathrm{s}}\delta_{\mathsf{H}}}{\sqrt{\lambda_{\mathrm{f}}}}.$$

By further shrinking Ω, we may ensure $\delta_{\mathsf{H}} \le \frac{c_*\sqrt{\lambda_{\mathrm{f}}}}{4G\lambda_{\mathrm{s}}}$. Therefore

$$\frac{\langle \nabla_Z f(P, Z), ZC_0\rangle}{D_0(Z)} \ge \left(\frac{16}{25} - \frac{1}{4}\right)c_* \ge \frac{1}{4}c_*.$$

Finally, Assumption 1 normalizes the sharp-wall direction as

$$\nu_{\mathsf{H}}(P,Z) = \frac{\mathsf{H}\big((P,Z) - P_{\mathcal{M}}(P,Z)\big)}{\big\|\mathsf{H}\big((P,Z) - P_{\mathcal{M}}(P,Z)\big)\big\|} = \frac{(0, ZC_0)}{\|ZC_0\|_F}.$$

Since $\|ZC_0\|_F \le \sqrt{\lambda_{\mathrm{s}}}\, D_0(Z)$, we obtain

$$\langle \nabla f(P,Z), \nu_{\mathsf{H}}(P,Z)\rangle = \frac{\langle \nabla_Z f(P,Z), ZC_0\rangle}{\|ZC_0\|_F} \ge \frac{c_*}{4\sqrt{\lambda_{\mathrm{s}}}}.$$

Thus the sharp-valley condition holds with $c_\perp = \frac{c_*}{4\sqrt{\lambda_{\mathrm{s}}}} > 0$.

Together, these estimates establish the basin, river-center, and sharp-valley conditions on the stated product neighborhood with the fixed metric H. □

D.2 Numerical Simulations of Matrix Factorization

We simulate the symmetric matrix factorization objective of Example 2,

$$f(X) = \|XX^\top - M^\star\|_F^2, \qquad M^\star = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \quad D \in \mathbb{R}^{r\times r} \text{ diagonal},\ D \succ 0,$$

with $X \in \mathbb{R}^{d\times k}$ and $k \ge r$, in the high-dimensional setting $d = 100$, $r = k = 5$ ($D = I_r$, $dk = 500$ parameters). The loss landscape is no longer drawable directly, so we apply the same family of diagnostics used on Llama in the main text: a 1D loss profile from FP toward each anchor, a 2D contour on the FP/PTQ/QAT affine plane, and a 3D river-cross surface in the plane spanned by FP→QAT and a random perpendicular direction.

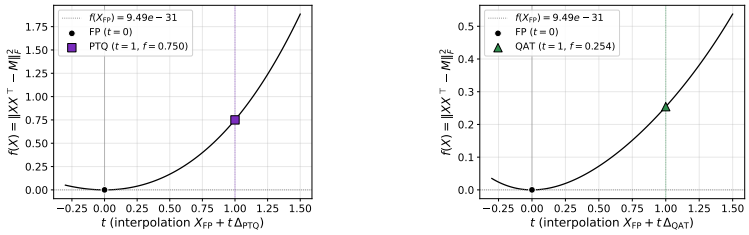
Pipeline. Stage 1 is full-precision gradient descent on $f(X)$ from a small random initialization, $X_0 \sim 0.30\,\mathcal{N}(0, I)$, with step size $\eta_{\mathrm{FP}} = 0.020$ for 4000 iterations; this drives $f(X_{\mathrm{fp}})$ to numerical zero ($\approx 10^{-30}$). Stage 2 applies PTQ by per-entry rounding, $X_{\mathrm{PTQ}} = Q(X_{\mathrm{fp}})$ with $Q(X) = \rho\,\mathrm{round}(X/\rho)$ and $\rho = 0.30$, chosen so that no FP optimum lies on the grid ($f(X) > 0$ for every $X \in \rho\,\mathbb{Z}^{d\times k}$). Stage 3 runs STE-QAT continuing from X_{fp},

$$X_{k+1} = X_k - \eta_k\, \nabla f\big(Q(X_k)\big),$$

for 1500 iterations with cosine-decayed step size $\eta_k = \frac{1}{2}\big(1 + \cos(\pi k/1500)\big)\cdot 0.010$, mirroring the cosine schedule used for the LLM runs. We report the deployed loss $f\big(Q(X_{\mathrm{QAT}})\big)$ on the final QAT iterate.

Results. The deployed PTQ loss is $f(X_{\mathrm{PTQ}}) = 0.7502$, a substantial gap from the FP optimum because the FP solution lies strictly off the grid in every coordinate. STE-QAT lowers the deployed loss to $f\big(Q(X_{\mathrm{QAT}})\big) = 0.2543$, a 66% reduction over the PTQ initialization at the *same* grid resolution and weight scope. The displacements $\Delta_{\mathrm{PTQ}} = X_{\mathrm{PTQ}} - X_{\mathrm{fp}}$ and $\Delta_{\mathrm{QAT}} = Q(X_{\mathrm{QAT}}) - X_{\mathrm{fp}}$ have Frobenius norms 0.474 and 0.594 and meet at an angle of 95.3° — they are nearly orthogonal, so QAT is not a small correction along the rounding direction but a separate, comparably-sized move.

Figure 8 shows one-dimensional loss profiles along these two directions. The FP→PTQ profile rises smoothly through the PTQ anchor at $t = 1$ to its $f = 0.7502$ value, while the FP→QAT profile reaches a lower $f = 0.2543$ at $t = 1$; both rise monotonically because X_{fp} is the global minimum, but the slope along Δ_{QAT} is markedly gentler, indicating a basin-aligned displacement compared to the rounding direction. Figure 9 renders the 2D loss landscape on the affine plane through the three anchors: FP, PTQ, and QAT fall at distinct corners with QAT placed inside the warm-color (low-loss) region while PTQ sits on a steeper part of the surface. Figure 10 tests the river-valley hypothesis directly: in the plane spanned by FP→QAT and a random perpendicular direction (rescaled to $\|\Delta_{\mathrm{FP}\to\mathrm{QAT}}\|$ and averaged over 5 random seeds, same construction as the ResNet/DeiT/Llama river-cross plots), the valley along $b = 0$ is narrow and deep; loss stays low along FP→QAT but rises sharply with $|b|$, consistent with Assumption 1 on this finite-dimensional matrix factorization instance.



(a) FP → PTQ 1D profile (b) FP → QAT 1D profile

Figure 8: **One-dimensional loss profiles for the high-dimensional Matrix Factorization simulation (Appendix D.2).** Loss along the linear interpolations $X_{\text{fp}} + t\,\Delta_{\text{PTQ}}$ (left) and $X_{\text{fp}} + t\,\Delta_{\text{QAT}}$ (right). The PTQ anchor at $t = 1$ sits at $f = 0.7502$; the QAT anchor at $t = 1$ sits at $f = 0.2543$. X_{fp} is the global minimum so both profiles start at zero, but the slope along Δ_{QAT} is gentler than along Δ_{PTQ} — the QAT direction follows a basin-aligned displacement rather than an off-river rounding move. Within each panel the dotted line marks $f(X_{\text{fp}})$ and the dashed vertical line marks the anchor location at $t = 1$.

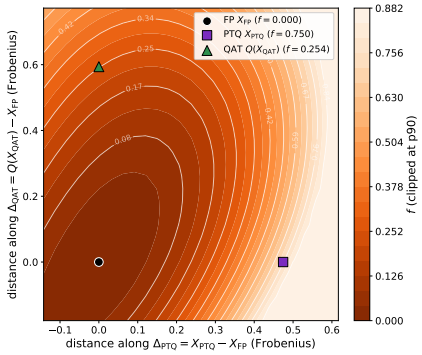


Figure 9: **FP/PTQ/QAT 2D loss landscape for the high-dimensional Matrix Factorization simulation.** Loss on the affine plane $X(a, b) = X_{\text{fp}} + a\,\Delta_{\text{PTQ}} + b\,\Delta_{\text{QAT}}$ through the three anchors (FP at $(0, 0)$, PTQ at $(1, 0)$, QAT at $(0, 1)$). Axes are in Frobenius distance units ($\|\Delta_{\text{PTQ}}\| = 0.474$, $\|\Delta_{\text{QAT}}\| = 0.594$, angle $95.3°$). PTQ sits on a steeper part of the surface than QAT; the QAT anchor lands inside the warm-color (low-loss) basin.

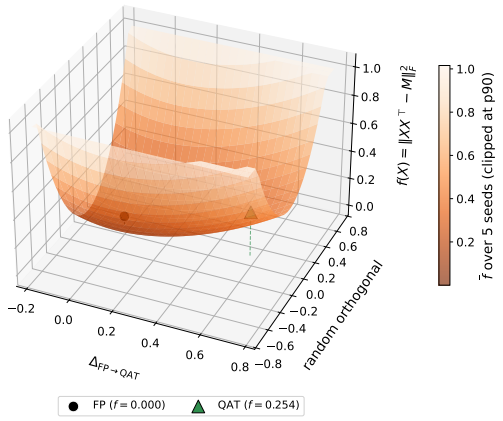


Figure 10: **River-cross 3D loss landscape for the high-dimensional Matrix Factorization simulation.** Loss f on the plane spanned by $\Delta_{\text{FP}\to\text{QAT}} = Q(X_{\text{QAT}}) - X_{\text{fp}}$ and a random direction projected orthogonal to it and rescaled to $\|\Delta_{\text{FP}\to\text{QAT}}\|$, averaged pointwise over 5 random seeds (same construction as the ResNet/DeiT/Llama river-cross plots in the main text). The valley along $b = 0$ is narrow and deep: loss stays low along $\Delta_{\text{FP}\to\text{QAT}}$ (the river axis) but rises sharply with $|b|$, directly consistent with Assumption 1 on this finite-dimensional matrix factorization instance.

E PROOFS IN SECTION 3

E.1 PROOF OF THEOREM 1

Proof of Theorem 1. Write $q_k = Q(w_k)$, $\pi_k = P_{\mathcal{M}}(w_k)$, and $d_k = \|w_k - \pi_k\|_{\mathsf{H}}$. Denote the line segment $\{(1-\delta)a + \delta b \mid \delta \in [0,1]\}$ connecting vectors a and b by $[a, b]$.

Consider $\Phi(w) = \frac{1}{2}\|w - \pi(w)\|_{\mathsf{H}}^2$ on U. Since $P_{\mathcal{M}}$ is C^1 on U and the range of $\nabla P_{\mathcal{M}}(w)$ lies in $T_{\mathcal{M}}(P_{\mathcal{M}}(w))$, whereas $\mathsf{H}(w - P_{\mathcal{M}}(w)) \in N_{\mathcal{M}}(P_{\mathcal{M}}(w))$ by metric compatibility. Hence,

$$\begin{aligned}\nabla\Phi(w) &= \mathsf{H}(w - P_{\mathcal{M}}(w)) - \nabla P_{\mathcal{M}}(w)^\top[\mathsf{H}(w - P_{\mathcal{M}}(w))] \\ &= \mathsf{H}(w - P_{\mathcal{M}}(w)).\end{aligned}$$

The L_P-Lipschitz continuity of $P_{\mathcal{M}}(\cdot)$ implies that $\nabla\Phi(\cdot)$ is $\lambda_1(1+L_P)$-Lipschitz on U.

Step 1. We first prove by induction that $w_k \in \mathcal{T}$ until first entry and that all segments used below lie in U. The base case is the assumed $w_0 \in \mathcal{T}$. Suppose $w_k \in \mathcal{T}$. For every $t \in [0,1]$,

$$\operatorname{dist}(w_k + t(q_k - w_k), \mathcal{M}) \le R/\sqrt{\lambda_d} + \rho < R_U.$$

Thus $[w_k, q_k] \subset U$, in particular $q_k \in U$ and $\|\nabla f(q_k)\| \le G$. Since $\eta G \le \rho$, we also have

$$\operatorname{dist}(w_k - t\eta\nabla f(q_k), \mathcal{M}) \le R/\sqrt{\lambda_d} + \eta G < R_U,$$

so $[w_k, w_{k+1}] \subset U$. If $k < T$, the sharp-valley condition at $q_k \in U \setminus \mathcal{T}$ gives

$$\begin{aligned}\langle\nabla\Phi(w_k), \nabla f(q_k)\rangle &\ge \langle\mathsf{H}(q_k - P_{\mathcal{M}}(q_k)), \nabla f(q_k)\rangle - \lambda_1(1+L_P)\rho G \\ &\ge c_\perp\|\mathsf{H}(q_k - P_{\mathcal{M}}(q_k))\| - \lambda_1(1+L_P)\rho G \\ &\ge c_\perp\sqrt{\lambda_d}\,\|q_k - P_{\mathcal{M}}(q_k)\|_{\mathsf{H}} - \lambda_1(1+L_P)\rho G \\ &\ge c_\perp\sqrt{\lambda_d}\,R - \lambda_1(1+L_P)\rho G.\end{aligned}$$

Since $\nabla\Phi(\cdot)$ is $\lambda_1(1+L_P)$-Lipschitz on U, we can apply the descent lemma to $\Phi(\cdot)$ along the contained step segment and use the stepsize upper bound $\eta \le \frac{c_\perp\sqrt{\lambda_d}R - \lambda_1(1+L_P)\rho G}{\lambda_1(1+L_P)G^2}$ to derive that

$$\begin{aligned}\Phi(w_{k+1}) &\le \Phi(w_k) + \langle w_{k+1} - w_k, \nabla\Phi(w_k)\rangle + \frac{\lambda_1(1+L_P)}{2}\|w_{k+1} - w_k\|^2 \\ &\le \Phi(w_k) - \eta\langle\nabla f(q_k), \nabla\Phi(w_k)\rangle + \frac{\lambda_1(1+L_P)}{2}\eta^2G^2 \\ &\le \Phi(w_k) - \eta\left[c_\perp\sqrt{\lambda_d}R - \lambda_1(1+L_P)\rho G\right] + \frac{\lambda_1(1+L_P)}{2}\eta^2G^2 \\ &\le \Phi(w_k) - \frac{\eta}{2}\left[c_\perp\sqrt{\lambda_d}R - \lambda_1(1+L_P)\rho G\right].\end{aligned} \tag{6}$$

Consequently, $d_{k+1}^2 \le d_k^2 - \eta\left[c_\perp\sqrt{\lambda_d}R - \lambda_1(1+L_P)\rho G\right] \le R^2$, so $w_{k+1} \in \mathcal{T}$. This completes the induction. It also establishes $[w_k, q_k] \subset U$ for $k \le T$ and $[w_k, w_{k+1}] \subset U$ for $k < T$.

Step 2. The bound for quantization error (Assumption 2(i)) and Lipschitz continuity of $P_{\mathcal{M}}$ give

$$\begin{aligned}\|q_k - P_{\mathcal{M}}(q_k)\|_{\mathsf{H}} &\le \|q_k - w_k\|_{\mathsf{H}} + \|w_k - \pi_k\|_{\mathsf{H}} + \|\pi_k - P_{\mathcal{M}}(q_k)\|_{\mathsf{H}} \\ &\le \|w_k - \pi_k\|_{\mathsf{H}} + \sqrt{\lambda_1}\big(\|q_k - w_k\| + \|P_{\mathcal{M}}(w_k) - P_{\mathcal{M}}(q_k)\|\big) \\ &\le d_k + \sqrt{\lambda_1}(1+L_P)\|q_k - w_k\| \\ &\le d_k + \sqrt{\lambda_1}(1+L_P)\rho.\end{aligned}$$

Thus $d_k \le R - \sqrt{\lambda_1}(1+L_P)\rho$ implies $q_k \in \mathcal{T}$. It follows from (6) that $T < +\infty$. Since $q_0 \notin \mathcal{T}$, $T \ge 1$, and at $k = T-1$, we have

$$\left(R - \sqrt{\lambda_1}(1+L_P)\rho\right)^2 < d_{T-1}^2 \le d_0^2 - (T-1)\eta\left(c_\perp\sqrt{\lambda_d}R - \lambda_1(1+L_P)\rho G\right).$$

Therefore,

$$T \le 1 + \frac{\max\left\{d_0^2 - \left(R - \sqrt{\lambda_1}\rho(1+L_P)\right)^2, 0\right\}}{\eta\left(c_\perp\sqrt{\lambda_d}R - \lambda_1(1+L_P)\rho G\right)}.$$

This completes the proof. □

E.2 Proofs of Corollary 1

Proof of Corollary 1. Write $q_k = Q(w_k)$ and $F = f \circ P_{\mathcal{M}}$. Theorem 1 and its proof give $T < +\infty$, $w_k \in \mathcal{T}$ for $0 \le k \le T$, $q_T \in \mathcal{T}$, and

$$[w_k, w_{k+1}] \subset U \quad \text{for } 0 \le k < T, \quad \text{and} \quad [w_T, q_T] \subset U.$$

Here $[a, b] = \{(1-\delta)a + \delta b \mid 0 \le \delta \le 1\}$ denotes the line segment connecting vectors a and b.

Since $P_{\mathcal{M}}$ is differentiable and L_P-Lipschitz on the open set U, $\|\nabla P_{\mathcal{M}}(w)\|_{\mathrm{op}} \le L_P$ for every $w \in U$. The chain rule and the near-stationary assumption on gradients along $\mathcal{M}$ therefore give

$$\begin{aligned}\|\nabla F(w)\| &= \left\|\nabla P_{\mathcal{M}}(w)^\top \nabla f(P_{\mathcal{M}}(w))\right\| \\ &\le \|\nabla P_{\mathcal{M}}(w)\|_{\mathrm{op}} \|\nabla f(P_{\mathcal{M}}(w))\| \le L_P \epsilon, \qquad \forall\, w \in U.\end{aligned}$$

This pointwise derivative bound will be integrated only along the segments already known to lie in U.

For $k < T$, let $\delta_k = w_{k+1} - w_k = -\eta \nabla f(q_k)$. By the fundamental theorem of calculus,

$$\begin{aligned}F(w_{k+1}) - F(w_k) &= \int_0^1 \langle \nabla F(w_k + t\delta_k), \delta_k \rangle \, dt \\ &\le \epsilon L_P \|\delta_k\| \le \epsilon L_P \eta G.\end{aligned}$$

Summing these inequalities gives

$$F(w_T) - F(w_0) \le \epsilon L_P \eta T G.$$

For the final rounding displacement $q_T - w_T$, the same argument on $[w_T, q_T]$ yields

$$F(q_T) - F(w_T) = \int_0^1 \langle \nabla F(w_T + t(q_T - w_T)), q_T - w_T \rangle \, dt \le \epsilon L_P \|q_T - w_T\| \le \epsilon L_P \rho.$$

Finally, basin flatness at w_0 and q_T (Assumption 1) implies

$$F(w_0) \le f(w_0) + \epsilon_{\mathrm{flat}}, \qquad f(q_T) \le F(q_T) + \epsilon_{\mathrm{flat}}.$$

Combining the four bounds and using $w_0 = w_{\mathrm{fp}}$ proves

$$f(q_T) \le f(w_{\mathrm{fp}}) + 2\epsilon_{\mathrm{flat}} + \epsilon L_P(\eta T G + \rho).$$

□

E.3 Proofs of Corollary 2

Before the proof, we first introduce a standard property of the nearest-point projection. For every $\pi \in \mathcal{M}$, it holds that

$$\nabla P_{\mathcal{M}}(\pi) = P_{T_{\mathcal{M}}(\pi)}. \tag{7}$$

Here $P_{T_{\mathcal{M}}(\pi)}$ denotes Euclidean orthogonal projection onto the tangent space $T_{\mathcal{M}}(\pi)$, whereas $P_{\mathcal{M}}$ projects onto the manifold.

Indeed, for any $z \in \mathbb{R}^d$, decompose $z = z_T + z_N$, where $z_T \in T_{\mathcal{M}}(\pi)$ and $z_N \in N_{\mathcal{M}}(\pi)$. For the tangent component, choose a smooth curve $\gamma \subset \mathcal{M}$ with $\gamma(0) = \pi$ and $\dot\gamma(0) = z_T$. Since $P_{\mathcal{M}}$ restricts to the identity map on $\mathcal{M}$, differentiating $P_{\mathcal{M}}(\gamma(t)) = \gamma(t)$ at $t = 0$ gives $\nabla P_{\mathcal{M}}(\pi) z_T = z_T$. For the normal component, since $P_{\mathcal{M}}(\pi + t z_N) = \pi$ for all sufficiently small t, differentiating at $t = 0$ gives $\nabla P_{\mathcal{M}}(\pi) z_N = 0$. Therefore $\nabla P_{\mathcal{M}}(\pi) z = z_T = P_{T_{\mathcal{M}}(\pi)} z$.

We also have Lemma 1 about projection linearization around the river.

Lemma 1. *Suppose $P_{\mathcal{M}} : U \to \mathcal{M}$ is $C^{1,1}$ and $\|\nabla P_{\mathcal{M}}(z) - \nabla P_{\mathcal{M}}(z')\|_{\mathrm{op}} \le \kappa_{\mathcal{M}} \|z - z'\|$ for any $z, z' \in U$. Let $w \in U$, $\pi = P_{\mathcal{M}}(w)$, and assume $\{w + t\delta \mid t \in [0,1]\} \subset U$. Then*

$$\left\|P_{\mathcal{M}}(w + \delta) - P_{\mathcal{M}}(w) - P_{T_{\mathcal{M}}(\pi)} \delta\right\| \le \kappa_{\mathcal{M}} \left(\|w - \pi\| \|\delta\| + \frac{1}{2} \|\delta\|^2 \right).$$

Proof of Lemma 1. Note that $P_{\mathcal{M}}(w+\delta) - P_{\mathcal{M}}(w) = \int_0^1 \nabla P_{\mathcal{M}}(w+t\delta)\delta\, dt$. Since $\pi = P_{\mathcal{M}}(w)$ and $\nabla P_{\mathcal{M}}(\pi) = P_{T_{\mathcal{M}}(\pi)}$, we have

$$P_{\mathcal{M}}(w+\delta) - P_{\mathcal{M}}(w) - P_{T_{\mathcal{M}}(\pi)}\delta = \int_0^1 \left[\nabla P_{\mathcal{M}}(w+t\delta) - \nabla P_{\mathcal{M}}(\pi)\right]\delta\, dt.$$

Using the $C^{1,1}$ regularity of $P_{\mathcal{M}}$ gives

$$\begin{aligned}\left\|P_{\mathcal{M}}(w+\delta) - P_{\mathcal{M}}(w) - P_{T_{\mathcal{M}}(\pi)}\delta\right\| &\le \int_0^1 \kappa_{\mathcal{M}}\|w+t\delta-\pi\|\|\delta\|\, dt \\ &\le \int_0^1 \kappa_{\mathcal{M}}\left(\|w-\pi\| + t\|\delta\|\right)\|\delta\|\, dt \\ &= \kappa_{\mathcal{M}}\left(\|w-\pi\|\|\delta\| + \frac{1}{2}\|\delta\|^2\right),\end{aligned}$$

completing the proof. □

Now we are ready to prove Corollary 2 with the help of the projection identity (7) and Lemma 1.

Proof of Corollary 2. Write $q_k = Q(w_k)$ and $\pi_k = P_{\mathcal{M}}(w_k)$. From Theorem 1 and its proof, we know that $\|w_k - \pi_k\| \le R/\sqrt{\lambda_d}$ for $k \le T$, $\|\nabla f(q_k)\| \le G$, $[w_k, w_{k+1}] \subset U$ for $k < T$, and $[w_T, q_T] \subset U$. Set

$$\Delta = \kappa_{\mathcal{M}}\left(R/\sqrt{\lambda_d} + \frac{1}{2}\eta G\right).$$

Step 1: Projected displacement and path length. Apply Lemma 1 with $w = w_k$ and $\delta = -\eta\nabla f(q_k)$. For $k < T$, it gives

$$\pi_{k+1} - \pi_k = -\eta P_{T_{\mathcal{M}}(\pi_k)}[\nabla f(q_k)] + r_k,$$

where

$$\begin{aligned}\|r_k\| &\le \kappa_{\mathcal{M}}\left(\eta\|w_k-\pi_k\|\cdot\|\nabla f(q_k)\| + \frac{1}{2}\eta^2\|\nabla f(q_k)\|^2\right) \\ &\le \eta G\kappa_{\mathcal{M}}\left(R/\sqrt{\lambda_d} + \frac{1}{2}\eta G\right) = \eta G\Delta.\end{aligned}$$

To compare projected losses, use the curve

$$x_k(t) = w_k - t\eta\nabla f(q_k), \qquad \gamma_k(t) = P_{\mathcal{M}}(x_k(t)), \qquad 0 \le t \le 1.$$

The first curve lies in U and the second in $\mathcal{M} \subset U$. Their projected endpoints are $\gamma_k(0) = \pi_k$ and $\gamma_k(1) = \pi_{k+1}$. By (7) and projection regularity,

$$\begin{aligned}\|\nabla P_{\mathcal{M}}(x_k(t))\|_{\mathrm{op}} &\le \|\nabla P_{\mathcal{M}}(\pi_k)\|_{\mathrm{op}} + \|\nabla P_{\mathcal{M}}(x_k(t)) - \nabla P_{\mathcal{M}}(\pi_k)\|_{\mathrm{op}} \\ &\le 1 + \kappa_{\mathcal{M}}(\|w_k-\pi_k\| + t\eta\|\nabla f(q_k)\|).\end{aligned}$$

Since $\dot\gamma_k(t) = -\eta\nabla P_{\mathcal{M}}(x_k(t))[\nabla f(q_k)]$, its length satisfies

$$\begin{aligned}\ell_k := \int_0^1 \|\dot\gamma_k(t)\|\, dt &\le \eta\|\nabla f(q_k)\|\left[1 + \kappa_{\mathcal{M}}\left(\|w_k-\pi_k\| + \frac{1}{2}\eta\|\nabla f(q_k)\|\right)\right] \\ &\le \eta G(1+\Delta).\end{aligned}$$

Step 2: Loss variation along the projected path. No containment of the straight chord $[\pi_k, \pi_{k+1}]$ is needed. Put $a_k(t) = \int_0^t \|\dot\gamma_k(s)\|\, ds$, so that $\|\gamma_k(t) - \pi_k\| \le a_k(t)$, $a_k(0) = 0$, and $a_k(1) = \ell_k$. The chain rule and the fundamental theorem of calculus imply

$$\begin{aligned}&f(\pi_{k+1}) - f(\pi_k) - \langle\nabla f(\pi_k), \pi_{k+1} - \pi_k\rangle \\ &= \int_0^1 \langle\nabla f(\gamma_k(t)) - \nabla f(\pi_k), \dot\gamma_k(t)\rangle\, dt\end{aligned}$$

$$\leq L \int_0^1 \|\gamma_k(t) - \pi_k\| \|\dot{\gamma}_k(t)\| \, dt$$
$$\leq L \int_0^1 a_k(t) a_k'(t) \, dt = \frac{L}{2} \ell_k^2.$$

The gradient comparison uses only points in U. River Center and symmetry of the tangent projector give

$$\big\langle \nabla f(\pi_k), P_{T_{\mathcal{M}}(\pi_k)}[\nabla f(q_k)] \big\rangle = \big\langle P_{T_{\mathcal{M}}(\pi_k)}[\nabla f(\pi_k)], \nabla f(q_k) \big\rangle = \langle \nabla f(\pi_k), \nabla f(q_k) \rangle.$$

Substituting Step 1 and bounding the remainder pairing by $\|\nabla f(\pi_k)\| \cdot \|r_k\|$ yields

$$f(\pi_{k+1}) \leq f(\pi_k) - \eta \langle \nabla f(\pi_k), \nabla f(q_k) \rangle + \eta G \Delta \|\nabla f(\pi_k)\| + \frac{L}{2} \eta^2 G^2 (1+\Delta)^2. \tag{8}$$

Step 3: Tangential alignment and descent. For $k < T$, both π_k and $P_{\mathcal{M}}(q_k)$ belong to $\mathcal{V}$ by assumption. Taking $w = \pi \in \mathcal{V}$ in the alignment assumption gives $\|\nabla f(\pi)\|^2 > c_{\|} \|\nabla f(\pi)\|$. The strict inequality excludes $\nabla f(\pi) = 0$, and division therefore shows $\|\nabla f(\pi)\| > c_{\|}$. Recall that $g(\pi) = \nabla f(\pi)/\|\nabla f(\pi)\|$ is the normalized gradient at point $\pi \in \mathcal{M}$ along the river. Taking $w = q_k$ in the same assumption and dividing by $\|\nabla f(P_{\mathcal{M}}(q_k))\| > 0$ gives

$$\langle g(P_{\mathcal{M}}(q_k)), \nabla f(q_k) \rangle > c_{\|}.$$

The assumed κ-Lipschitz continuity of $g(\cdot)$ and L_P-Lipschitz continuity of $P_{\mathcal{M}}(\cdot)$ imply

$$\|g(\pi_k) - g(P_{\mathcal{M}}(q_k))\| \leq \kappa \|P_{\mathcal{M}}(w_k) - P_{\mathcal{M}}(q_k)\| \leq \kappa L_P \rho.$$

Consequently,

$$\begin{aligned} \langle g(\pi_k), \nabla f(q_k) \rangle &\geq \langle g(P_{\mathcal{M}}(q_k)), \nabla f(q_k) \rangle - \|g(\pi_k) - g(P_{\mathcal{M}}(q_k))\| \cdot \|\nabla f(q_k)\| \\ &\geq c_{\|} - \kappa L_P \rho G. \end{aligned}$$

Since $\nabla f(\pi_k) = \|\nabla f(\pi_k)\| g(\pi_k)$, the inequality (8) becomes

$$f(\pi_{k+1}) \leq f(\pi_k) - \eta \left(c_{\|} - \kappa L_P \rho G - G\Delta \right) \cdot \|\nabla f(\pi_k)\| + \frac{L}{2} \eta^2 G^2 (1+\Delta)^2.$$

By $\eta G \leq \rho$ and the bounds on κ and $\kappa_{\mathcal{M}}$,

$$\kappa L_P \rho G \leq \frac{c_{\|}}{4}, \qquad G\Delta \leq G \kappa_{\mathcal{M}} (R/\sqrt{\lambda_d} + \rho) \leq \frac{c_{\|}}{4}.$$

Thus the coefficient of $\|\nabla f(\pi_k)\|$ is at least $c_{\|}/2$. Set $C = 1 + \kappa_{\mathcal{M}}(R/\sqrt{\lambda_d} + \rho)$, so $1 + \Delta \leq C$. The additional stepsize bound is $\eta \leq c_{\|}^2/(2LG^2C^2)$, which gives

$$\frac{L}{2} \eta^2 G^2 (1+\Delta)^2 \leq \frac{L}{2} \eta^2 G^2 C^2 \leq \frac{1}{4} \eta c_{\|}^2.$$

Using $\|\nabla f(\pi_k)\| > c_{\|}$ now yields

$$f(\pi_{k+1}) \leq f(\pi_k) - \frac{1}{2} \eta c_{\|}^2 + \frac{1}{4} \eta c_{\|}^2 = f(\pi_k) - \frac{1}{4} \eta c_{\|}^2.$$

Summing over $k = 0, \ldots, T-1$, we obtain

$$f(\pi_T) \leq f(\pi_0) - \frac{1}{4} \eta T c_{\|}^2. \tag{9}$$

Step 4: Return to the deployed loss. Let $F = f \circ P_{\mathcal{M}}$. The chain rule gives

$$\|\nabla F(w)\| \leq \|\nabla P_{\mathcal{M}}(w)\|_{\mathrm{op}} \cdot \|\nabla f(P_{\mathcal{M}}(w))\| \leq L_P G, \qquad \forall\, w \in U.$$

Integrating on $[w_T, q_T] \subset U$ therefore yields

$$\begin{aligned} f(P_{\mathcal{M}}(q_T)) - f(\pi_T) &= \int_0^1 \langle \nabla F(w_T + t(q_T - w_T)), q_T - w_T \rangle \, dt \\ &\leq L_P G \|q_T - w_T\| \leq L_P G \rho. \end{aligned} \tag{10}$$

Since $w_0, q_T \in \mathcal{T}$, basin flatness gives

$$f(\pi_0) \leq f(w_{\mathrm{fp}}) + \epsilon_{\mathrm{flat}}, \qquad f(q_T) \leq f(P_{\mathcal{M}}(q_T)) + \epsilon_{\mathrm{flat}}.$$

Combining these inequalities with (9) and (10) proves

$$f(q_T) \leq f(w_{\mathrm{fp}}) + 2\epsilon_{\mathrm{flat}} + \rho L_P G - \frac{1}{4} \eta T c_{\|}^2.$$

□