High-Dimensional Learning Dynamics of Quantized Models with Straight-Through Estimator

Yuma Ichikawa[1,2] **Shuhei Kashiwamura**[3,4] **Ayaka Sakata**[2,5]

[1]Fujitsu Limited [2]RIKEN Center for AIP [3]The University of Tokyo
[4]NTT Research, Inc. [5]Ochanomizu University

ICML 2026 · **Seoul** arXiv:2510.10693

Why study STE dynamics?

- Quantized training optimizes a **discrete, non-differentiable** objective.
- The standard fix: the **straight-through estimator (STE)** — pretend the quantizer is the identity in the backward pass.
- It works remarkably well, but: **how do bit width b and quantization range ω shape the learning dynamics?**
- Prior theory: weight-only *or* activation-only. **Joint weight–input quantization** — the regime of low-bit LLM inference — was unexplored.

A minimal but representative model

Teacher–student linear regression, jointly quantized:

$$y_\mu = \tfrac{1}{\sqrt{d}}\, \boldsymbol{x}_\mu^\top \boldsymbol{w}^* + \xi_\mu, \qquad \hat{y}(\boldsymbol{x}; \boldsymbol{w}) = \tfrac{1}{\sqrt{d}}\, \psi(\boldsymbol{w})^\top \psi(\boldsymbol{x}),$$

with a uniform b-bit quantizer ψ on $[-\omega, \omega]$ applied to **both** weights and inputs.

One-pass STE update:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \left[\frac{\hat{y}(\boldsymbol{x}^t; \boldsymbol{w}^t) - y^t}{\sqrt{d}}\, \psi(\boldsymbol{x}^t) + \frac{\lambda}{d}\, \psi(\boldsymbol{w}^t) \right]$$

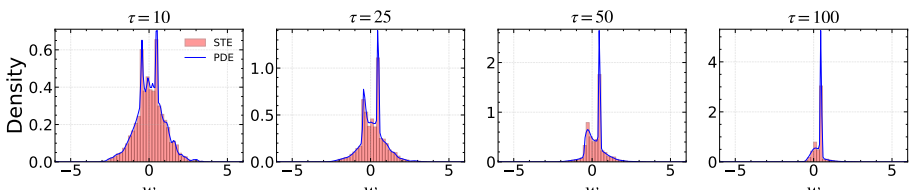
Structurally equivalent to the layer-wise objectives of LLM post-training quantization (GPTQ, AWQ).

Exact microscopic limit: an SDE for each weight

As $d \to \infty$ with $\tau = t/d$, weight coordinates follow

$$\mathrm{d}\hat{w}_\tau = \eta\big[\kappa_\psi \hat{w}^* - (\sigma_\psi^2 + \lambda)\psi_T(\hat{w}_\tau)\big]\mathrm{d}\tau + \eta\,\sigma_\psi\,\varepsilon_g^{1/2}(\tau)\,\mathrm{d}B_\tau$$

— noise level = the *instantaneous generalization error*.



STE simulation (d=3,000, red) vs. PDE prediction (blue) — no fitted parameters.

Two scalar ODEs predict the whole learning curve

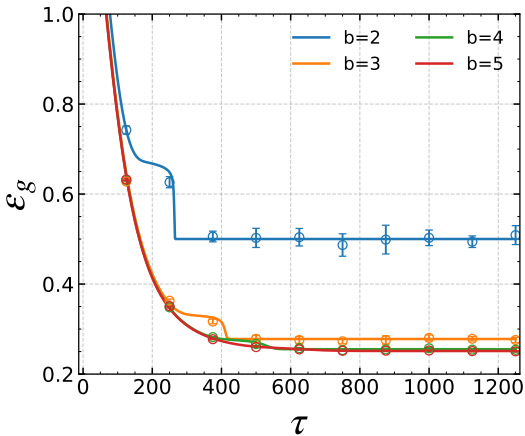
Isotropy closure: orthogonal weight components become Gaussian in high dimension $\Rightarrow$ closed equations for $\Psi = (m, q)$:

$$\frac{\mathrm{d}m}{\mathrm{d}\tau} = -\eta\big[(\sigma_\psi^2 + \lambda)m_\psi - \kappa_\psi \rho\big],$$
$$\frac{\mathrm{d}q}{\mathrm{d}\tau} = -2\eta\big[(\sigma_\psi^2 + \lambda)r_\psi - \kappa_\psi m\big] + \eta^2\sigma_\psi^2\,\varepsilon_g$$

Concentration: $\max_{0\le t\le dT} \mathbb{E}\|\Psi^t - \Psi(t/d)\| \le Cd^{-1/2}$.

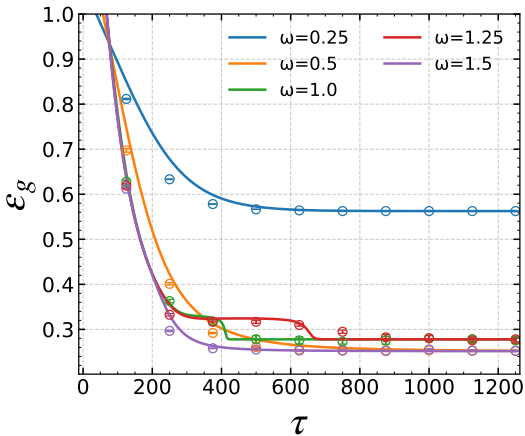
Key methodological step: extends high-dimensional SGD theory to **nonlinear transforms of both weights and inputs**.

Finding 1: a plateau, then a sharp drop



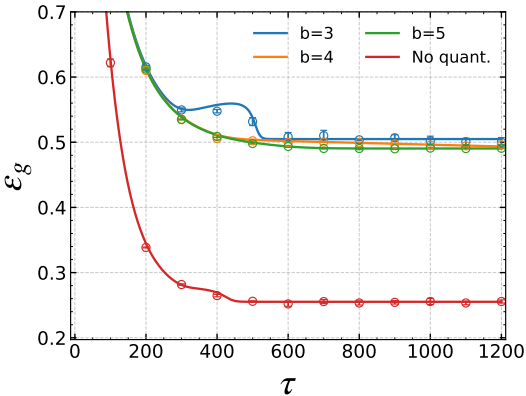
- Extended plateau $\rightarrow$ abrupt drop $\rightarrow$ stationary floor.
- Floor decreases with bit width b.
- **Drop time predicted exactly** by the ODEs (lines vs. markers).

Finding 2: the range ω controls the transition



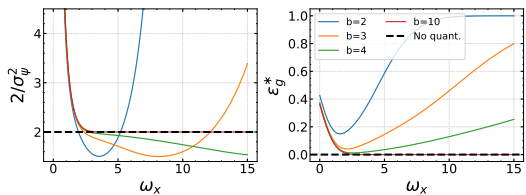
- Small ω: late transition, high error floor.
- Moderate-to-large ω: fast transition, low floor.
- ω is a **dynamical** hyperparameter — a case for *learned* scales in QAT.

Finding 3: inputs are the bottleneck



- Input quantization **raises the floor and slows convergence**.
- Low b_x: **non-monotonic**, oscillatory transients.
- Symmetric W4A4-style budgets are dynamically suboptimal.

Long-time behavior: stability and the quantization gap



Stable iff $0 < \eta < 2(\sigma_\psi^2 + \lambda)/\sigma_\psi^4$: for some ranges the stable region *exceeds* the unquantized baseline.

- Quantization can act as an **implicit regularizer**: larger admissible learning rates.
- Small-η gap to the unquantized model:

$$\varepsilon_g^* = \varepsilon_g^{(0)} + \sigma_\psi^2 \Delta^2 \, p(1-p) + o(\eta)$$

— explicit $\mathcal{O}(\Delta^2)$, non-monotonic in the grid offset p.

Takeaways

- **Exact dynamics:** high-dimensional STE training $=$ two deterministic ODEs.
- **Tune the range, not just the bits:** ω sets plateau escape and the error floor.
- **Inputs are the bottleneck:** low-bit inputs cause non-monotone, slower dynamics.
- **Quantization can stabilize training** — an implicit regularizer.
- **A toolbox** for nonlinear weight/input transforms; a step toward principled layer-wise PTQ for LLMs.

Thank you! Paper & code: arXiv:2510.10693