Revisiting Layer Normalization for Point Cloud Test Time Adaptation

Supplementary Material

6. Supplementary Material

6.1. Method: Proofs

Proof: [Proof of Lemma 1] Assume either $\sigma_d(x_t) > 0$ or $\varepsilon > 0$.

Shift. Since $\mathsf{C}(x_t + b\,\mathbb{1}) = \mathsf{C}x_t$ and $\sigma_d^2(x_t + b\,\mathbb{1}) = \sigma_d^2(x_t)$ (by Eq. (6)),

$$\widehat{\mathrm{LN}}(x_t + b\,\mathbb{1}) = \widehat{\mathrm{LN}}(x_t).$$

Scale. Let $r := d^{-1}\|\mathsf{C}x_t\|^2$ and $f(a; x_t) := \dfrac{a\sqrt{r+\varepsilon}}{\sqrt{a^2 r + \varepsilon}}$. From Eq. (7),

$$\widehat{\mathrm{LN}}(a\,x_t) = \sqrt{d}\,\frac{a\,\mathsf{C}x_t}{\sqrt{a^2\|\mathsf{C}x_t\|^2 + d\,\varepsilon}} = \frac{a}{|a|}\,\widehat{\mathrm{LN}}(x_t)\,f(|a|; x_t).$$

Hence for $a > 0$ we obtain Eq. (10), and for $a < 0$ we obtain Eq. (11). For the post-affine output $y(x_t) = \Gamma\,\widehat{\mathrm{LN}}(x_t) + \beta$,

$$y(a\,x_t) = \Gamma\,\widehat{\mathrm{LN}}(a\,x_t) + \beta = \frac{a}{|a|}\,f(|a|; x_t)\,\big(y(x_t) - \beta\big) + \beta.$$

□

Proof: [Proof of Proposition 1] Let $h = \widehat{\mathrm{LN}}(x_t)$; componentwise,

$$h_j = \frac{(x_t)_j - \mu_d(x_t)}{\sqrt{\sigma_d^2(x_t) + \varepsilon}}.$$

Since $\sum_{j=1}^{d}\big((x_t)_j - \mu_d(x_t)\big) = 0$, we have $\frac{1}{d}\sum_{j=1}^{d} h_j = 0$. For the second moment,

$$\frac{1}{d}\sum_{j=1}^{d} h_j^2 = \frac{1}{d}\cdot\frac{\sum_{j=1}^{d}\big((x_t)_j - \mu_d(x_t)\big)^2}{\sigma_d^2(x_t) + \varepsilon} = \frac{\sigma_d^2(x_t)}{\sigma_d^2(x_t) + \varepsilon},$$

which equals 1 when $\varepsilon = 0$, matching the norm identity in Eq. (9). □

Proof: [Proof of Theorem 1] *Step 1 (token-wise consistency of the random normalization).* For batch element b (with fixed but arbitrary token index t) define

$$\mu_{b,t}^{(d)} = \tfrac{1}{d}\sum_{i=1}^{d}(x_{b,t}^{(d)})_i, \qquad v_{b,t}^{(d)} = \tfrac{1}{d}\sum_{i=1}^{d}\big((x_{b,t}^{(d)})_i - \mu_{b,t}^{(d)}\big)^2,$$

and the j-th pre-affine normalized coordinate

$$\big(H_{b,t}^{(d)}\big)_j = \frac{(x_{b,t}^{(d)})_j - \mu_{b,t}^{(d)}}{\sqrt{v_{b,t}^{(d)}}}.$$

Finite fourth moments and the assumed Law of Large Numbers (LLN) across coordinates imply $\mu_{b,t}^{(d)} \xrightarrow{a.s.} \mu$ and $v_{b,t}^{(d)} \xrightarrow{a.s.} \sigma^2$ as $d \to \infty$. Let $g(y,a,b) = (y-a)/\sqrt{b}$, continuous on $\{b > 0\}$. By the continuous mapping theorem,

$$\big(H_{b,t}^{(d)}\big)_j = g\big((x_{b,t}^{(d)})_j, \mu_{b,t}^{(d)}, v_{b,t}^{(d)}\big) \;\Rightarrow\; \frac{(x_{b,t}^{(d)})_j - \mu}{\sigma}.$$

Uniform integrability (from finite fourth moments) yields moment convergence: $\mathbb{E}\big(H_{b,t}^{(d)}\big)_j \to 0$, $\mathrm{Var}\,\big(\big(H_{b,t}^{(d)}\big)_j\big) \to 1$.

Step 2 (batch LLN at fixed d). For each d, the variables $\{\big(H_{b,t}^{(d)}\big)_j\}_{b=1}^{B(d)}$ are i.i.d. with finite variance. Hence,

$$\begin{aligned} \overline{H}_{j,B(d)}^{(d)} &\xrightarrow{\text{a.s.}} \mathbb{E}(H_{1,t}^{(d)})_j, \\ \big(S_{j,B(d)}^2\big)^{(d)} &\xrightarrow{\text{a.s.}} \mathrm{Var}\big((H_{1,t}^{(d)})_j\big), \end{aligned} \qquad B(d) \to \infty.$$

Step 3 (diagonal limit). From Step 1, $\mathbb{E}\big(H_{1,t}^{(d)}\big)_j \to 0$ and $\mathrm{Var}\,\big(\big(H_{1,t}^{(d)}\big)_j\big) \to 1$ as $d \to \infty$. Combining with Step 2 and applying a diagonal argument (or Chebyshev in probability) yields

$$\overline{H}_{j,B(d)}^{(d)} \xrightarrow{\mathbb{P}} 0, \qquad \big(S_{j,B(d)}^2\big)^{(d)} \xrightarrow{\mathbb{P}} 1,$$

along any sequence $B(d) \to \infty$. □

Corollary 2 (Post-affine batch limits and $\varepsilon \geq 0$) *Let $H_{b,t} = \widehat{\mathrm{LN}}(x_{b,t})$ and $y_{b,t} = \Gamma H_{b,t} + \beta$ with $\Gamma = \mathrm{diag}(\gamma)$. Under the assumptions of Theorem 1, as $d \to \infty$ and $B = B(d) \to \infty$,*

$$\overline{y}_{j,B(d)} \xrightarrow{\mathbb{P}} \beta_j, \qquad S_{y,j,B(d)}^2 \xrightarrow{\mathbb{P}} \gamma_j^2\,\frac{\sigma^2}{\sigma^2 + \varepsilon},$$

where $\overline{y}_{j,B} = \frac{1}{B}\sum_{b=1}^{B}(y_{b,t})_j$ and $S_{y,j,B}^2 = \frac{1}{B}\sum_{b=1}^{B}\big((y_{b,t})_j - \overline{y}_{j,B}\big)^2$.

Proof: By the LLN across coordinates, $\mu_{b,t}^{(d)} \to \mu$ and $v_{b,t}^{(d)} \to \sigma^2$ a.s. as $d \to \infty$. Hence $(H_{b,t})_j = \big((x_{b,t})_j - \mu_{b,t}^{(d)}\big)/\sqrt{v_{b,t}^{(d)} + \varepsilon} \Rightarrow (X - \mu)/\sqrt{\sigma^2 + \varepsilon}$, so $\mathbb{E}(H_{b,t})_j \to 0$ and $\mathrm{Var}(H_{b,t})_j \to \sigma^2/(\sigma^2 + \varepsilon)$. Applying the batch LLN and mapping through $y_j = \gamma_j H_j + \beta_j$ gives the limits. □

6.2. Experiments

To test how our theoretical analysis of LN manifests in practice, we first evaluate the effect of the LN layer on domain information, followed by our backpropagation-free LN adaptation on three corrupted 3D point-cloud suites, ModelNet40-C [25], ShapeNet-C [5], and ScanObjectNN-C [26], following test-time protocols used in prior work [17, 29]. These benchmarks introduce diverse perturbations (e.g., noise, density shifts, geometric distortions) to stress robustness and cross-domain generalization. We report results under both *episodic* (reset after each test batch) and *online* (continuous stream) adaptation; because our method is a purely forward, stateless reparameterization of the LN affine (no gradients, no optimizers, no BN buffers), predictions are *identical* under the two protocols, there is no weight/buffer mutation and thus no temporal overfitting. For fair comparison, we start from the publicly released PointMAE backbone [21] as prepared in MATE [17] and reproduce baselines within the same evaluation harness [17, 29]. We follow TENT in resetting the BN statistics. Adaptation employs a batch size of 32, with a single forward pass per batch (no data augmentation or auxiliary adaptation passes), introduces no additional hyperparameters, and incurs only negligible compute/memory overhead; experiments are run on a single RTX 2060 GPU.

6.2.1. Datasets

We evaluate on three corrupted point-cloud benchmarks that share the same set of 15 test-time corruption types, covering geometric transformations, additive noise, and density variations, implemented by Sun et al. [25], which enables apples-to-apples robustness comparisons across datasets. **ModelNet40-C** [25] augments the original ModelNet40 test split with these 15 corruptions to probe resilience to realistic acquisition artifacts (e.g., sensor noise and LiDAR sparsity). **ShapeNet-C** is derived from ShapeNetCore-v2 [5] (51,127 shapes, 55 categories; 70%/10%/20% train/val/test) by applying the same 15 corruptions to the test set following the open-source pipeline of [25], as introduced in [17]. Finally, **ScanObjectNN-C** builds on the real-world ScanObjectNN dataset [26] (15 classes; 2,309 train / 581 test) by corrupting only the test set with the same 15 object-level perturbations, following the procedure of [17].

6.2.2. Batch-size Ablations.

We vary the test-time batch size B to assess practical sensitivity under a low batch schema. **LN-TTA** (ours) shows a smooth, monotonic gain with B and, crucially retains strong accuracy at small B, where TENT degrades sharply. This behavior stems from our design: since LN enforces token-wise zero mean and unit variance, the *post-affine* statistics are already well concentrated; our single-pass, backprop-free calibration merely recenters and rescales that operating point, requiring far less batch evidence than optimization-based TTA. At $B = 32$ (used in Tab. 3), **LN-TTA** attains **64.38%** mean accuracy on ShapeNet-C, outperforming TENT in the standard setting and maintaining a margin even relative to its online variant, all while remaining stateless (identical "Standard"/"Online" predictions), source-free, and hyperparameter-free. This justifies our choice of $B = 32$ as a balanced point between stability, throughput, and memory.

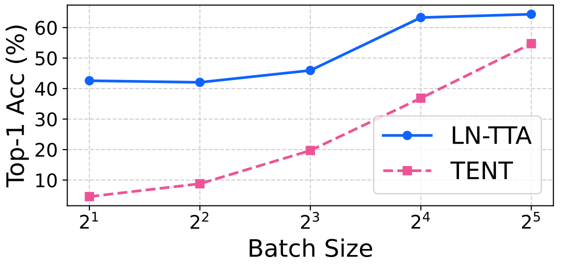

Figure 5. **Batch-size sensitivity on ShapeNet-C.** Top-1 accuracy (%) versus test-time batch size B for **LN-TTA (ours)** and TENT under the same backbone and corruption protocol. Both methods improve with larger B, but **LN-TTA** maintains a clear margin across the entire range and is notably more stable in the small-B regime.

Average per-feature marginals before and after LN. We synthesize batches whose coordinates are drawn from a mix of distributions {Normal, Laplace, Uniform, Student-t, Logistic, Exponential}, compute per-feature histograms across tokens, and average these histograms over features to obtain an "average marginal" (Sec. 3.3 viewpoint). We then compare pre-LN (dashed) and post-LN (solid) curves while (i) sweeping batch size B at fixed d and (ii) sweeping d at fixed B (Fig. 6). Post-LN curves align closely with the standard normal across settings, while pre-LN curves reflect the heterogeneous input laws. This outcome is consistent with our theory: LN enforces token-wise zero mean and unit variance (Proposition 1), and the per-feature, across-tokens marginals concentrate near $(0, 1)$ as (d, B) grow (Theorem 1). Consequently, re-standardizing "across the batch" is not the right lever for LN; the meaningful degrees of freedom lie in the angular structure within $\mathcal{H}$ and in the post-affine mean/scale, which motivates our post-affine test-time updates in Sec. 3.4.

6.2.3. LN Analyses

Per-channel marginals across batch sizes and feature dimensions. For each configuration with batch size $B \in \{32, 128, 512, 2048, 8192\}$ and feature dimension $d \in \{32, 128, 512, 2048, 8192\}$, we synthesize inputs whose coordinates are drawn from a heterogeneous mix of distributions {Normal, Laplace, Uniform, Student-t, Logistic, Ex-

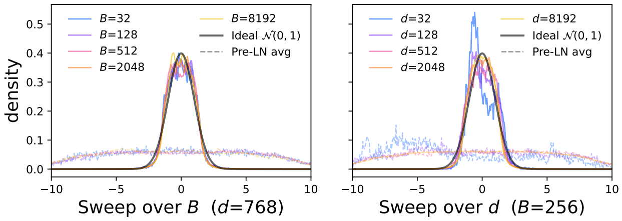

Figure 6. **Pre- vs. post-LN per-feature marginals.** We plot the *average* per-feature marginal density across tokens (dashed = pre-LN, solid = post-LN). **Left:** sweep batch size B at fixed d=768. **Right:** sweep feature dimension d at fixed B=256. The dashed black curve is the ideal $\mathcal{N}(0,1)$.

ponential}; see Figs. 7 to 9. For every feature, we compute its empirical marginal across tokens and overlay all per-feature curves together with their featurewise average, both *before* and *after* LN. Across all (B, d) pairs, the pre-LN panels exhibit wide, dataset-dependent variability (shifts, scales, and tail behaviours), whereas the post-LN panels collapse tightly toward the standard normal. Two trends are consistent throughout: (i) at fixed d, increasing B concentrates the per-feature, across-tokens marginals around mean 0 and variance 1; and (ii) at fixed B, increasing d further tightens the post-LN dispersion around $\mathcal{N}(0,1)$. These observations match our analysis: LN enforces token-wise zero mean and unit variance exactly (Proposition 1), and under mild moment assumptions the per-feature batch marginals concentrate near $(0,1)$ as $(d,B) \to \infty$ (Theorem 1). Consequently, batch/channel re-standardization is not the right lever for LN; the actionable degrees of freedom lie in the mean-free angular structure within $\mathcal{H}$ and in the post-affine mean/scale, which motivates our post-affine test-time updates in Sec. 3.4.

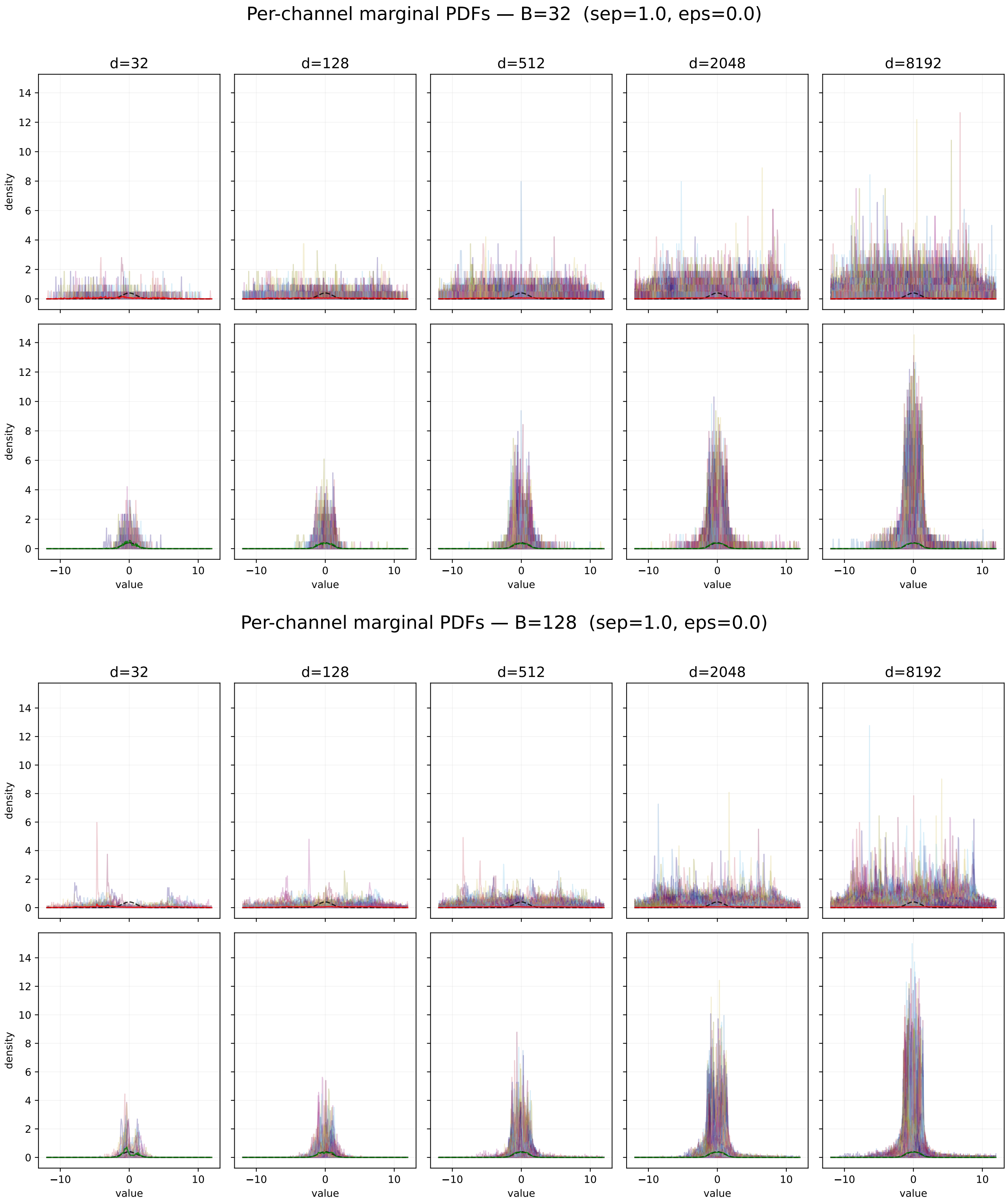

Figure 7. **Per-feature marginal PDFs before/after LN across batch sizes and feature dimensions.** Each panel (row) fixes a batch size $B \in \{32, 128\}$ and sweeps the feature dimension $d \in \{32, 128, 512, 2048, 8192\}$ across columns. Thin colored curves are per-feature empirical marginals across tokens; solid red is the featurewise average *pre*-LN, solid green is the average *post*-LN, and the dashed black curve is the ideal $\mathcal{N}(0,1)$ (here ε=0). Across all (B, d), pre-LN marginals are heterogeneous, while post-LN marginals collapse toward mean 0 and variance 1, with dispersion shrinking as B and d increase. This visualizes coordinate-wise concentration toward $(0, 1)$ as $(d, B) \to \infty$ (Theorem 1).

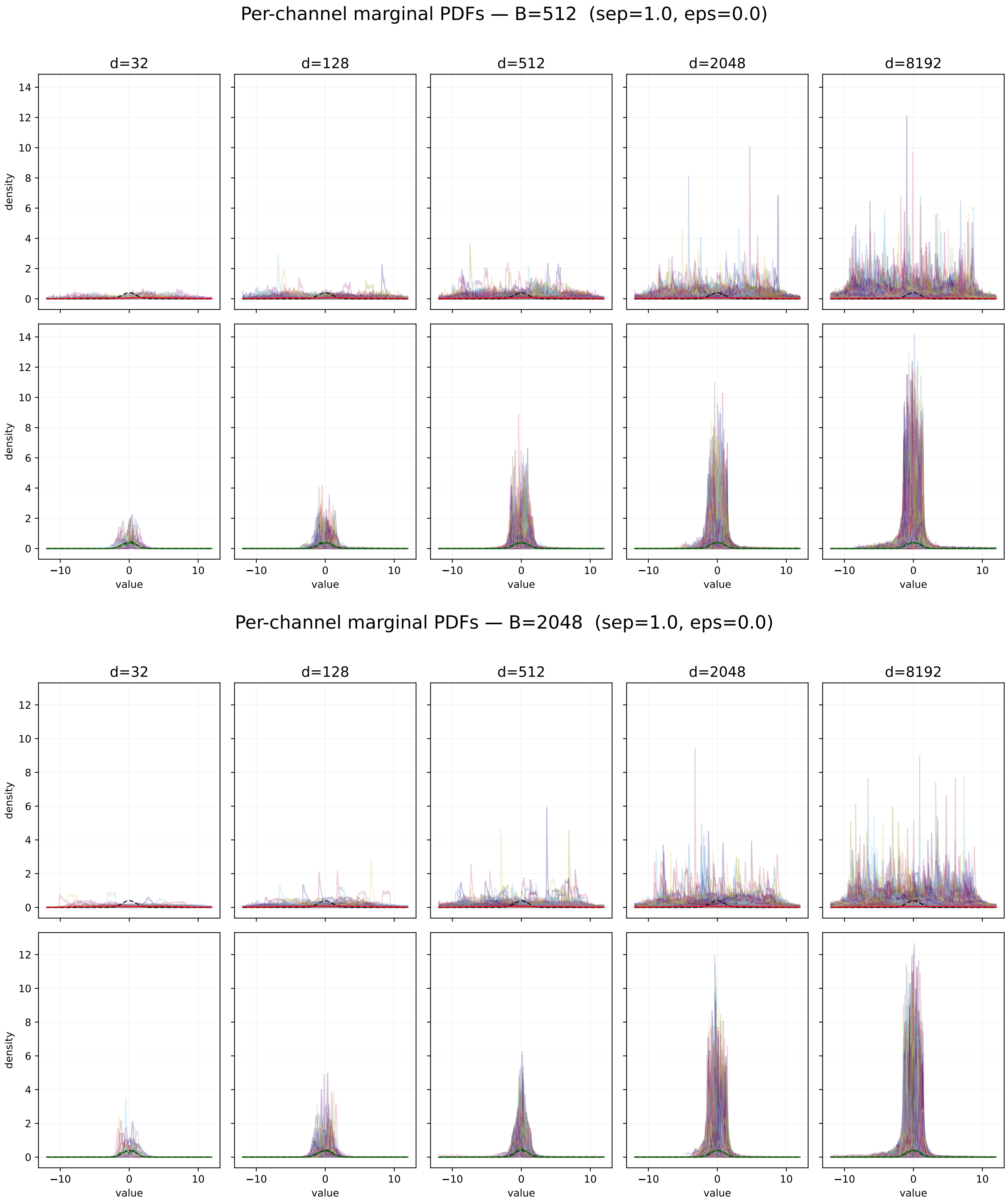

Figure 8. **Per-feature marginal PDFs before/after LN across batch sizes and feature dimensions.** Each panel (row) fixes a batch size $B \in \{512, 2048\}$ and sweeps the feature dimension $d \in \{32, 128, 512, 2048, 8192\}$ across columns. Thin colored curves are per-feature empirical marginals across tokens; solid red is the featurewise average *pre*-LN, solid green is the average *post*-LN, and the dashed black curve is the ideal $\mathcal{N}(0, 1)$ (here ε=0). Across all (B, d), pre-LN marginals are heterogeneous, while post-LN marginals collapse toward mean 0 and variance 1, with dispersion shrinking as B and d increase. This visualizes coordinate-wise concentration toward $(0, 1)$ as $(d, B) \to \infty$ (Theorem 1).

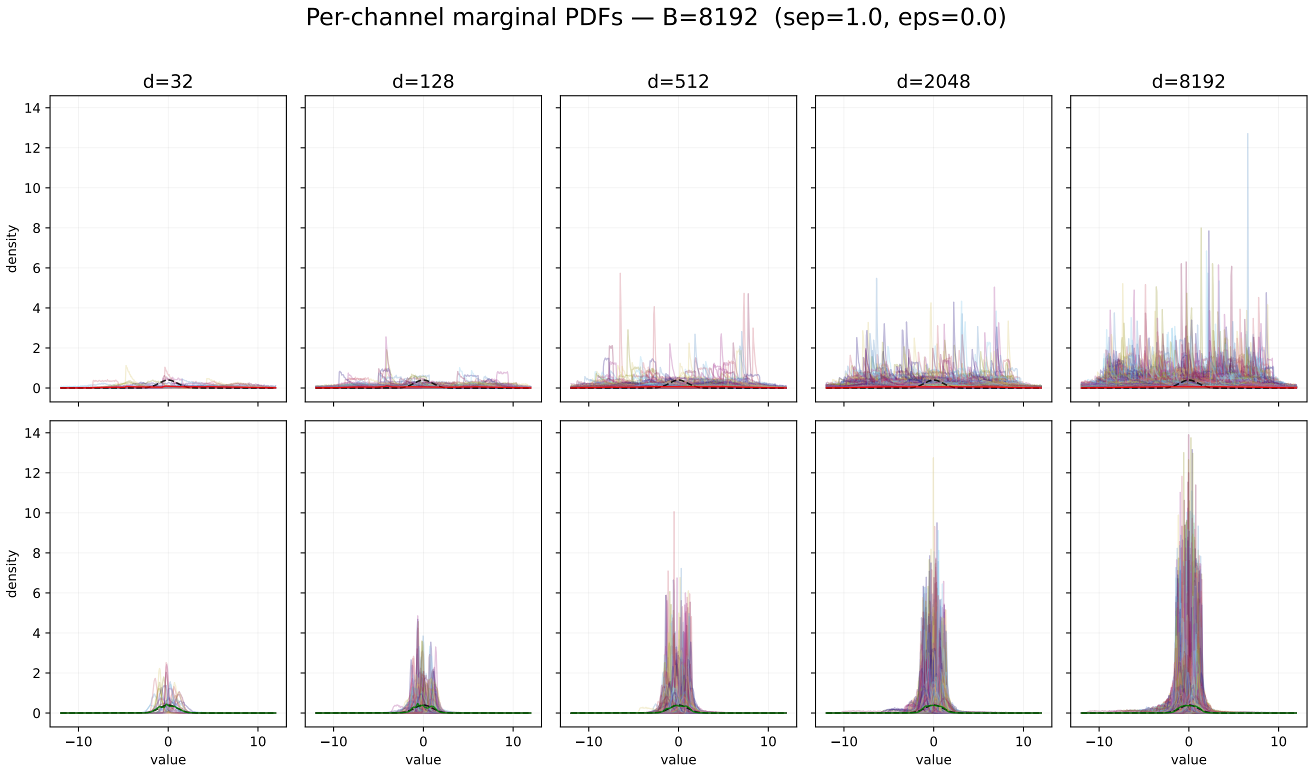

Figure 9. **Per-feature marginal PDFs before/after LN across batch sizes and feature dimensions.** Each panel (row) fixes a batch size $B = 8192$ and sweeps the feature dimension $d \in \{32, 128, 512, 2048, 8192\}$ across columns. Thin colored curves are per-feature empirical marginals across tokens; solid red is the featurewise average *pre*-LN, solid green is the average *post*-LN, and the dashed black curve is the ideal $\mathcal{N}(0,1)$ (here $\varepsilon{=}0$). Across all (B,d), pre-LN marginals are heterogeneous, while post-LN marginals collapse toward mean 0 and variance 1, with dispersion shrinking as B and d increase. This visualizes coordinate-wise concentration toward $(0,1)$ as $(d,B) \to \infty$ (Theorem 1).