ICML
International Conference
On Machine Learning

Private and Stable Test-Time Adaptation with Differential Privacy

Zefeng Li* Qiaoyue Tang* Mathias Lécuyer† Evan Shelhamer†
The University of British Columbia, Vector Institute

Why Adapt?

Deep Learning Works for Training and Testing on Similar Data

train

noise

blur

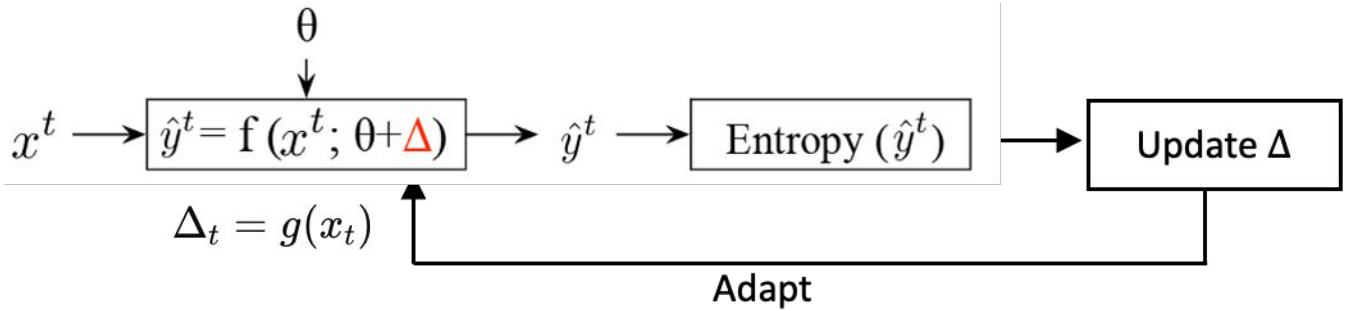
weather

compression

"goldfinch"

Test-time adaptation improves accuracy on corrupted data.

Test-Time Updates Can Leak Test Data



$x_1 \rightarrow$ update $\theta_1 \rightarrow x_2 \rightarrow$ update $\theta_2 \rightarrow \ldots \rightarrow \theta_T \quad \Rightarrow \quad \theta_T = h(x_1, x_2, \ldots, x_T)$

Test-time updates make the model a function of test data

If the adapted model is shared or queried, it may reveal information about individual test samples.

Making TTA Private: Clip Each Sample, Then Add Noise

$$B_t = \{x_i\}_{i=1}^{B} \longrightarrow g_i = \nabla_\theta \ell_{\text{TTA}}(x_i, \theta_t) \longrightarrow \bar{g}_i = \frac{g_i}{\max(1, \|g_i\|_2 / C)}$$

test batch

per-sample gradients

clip each sample

clipping bounds each sample's influence

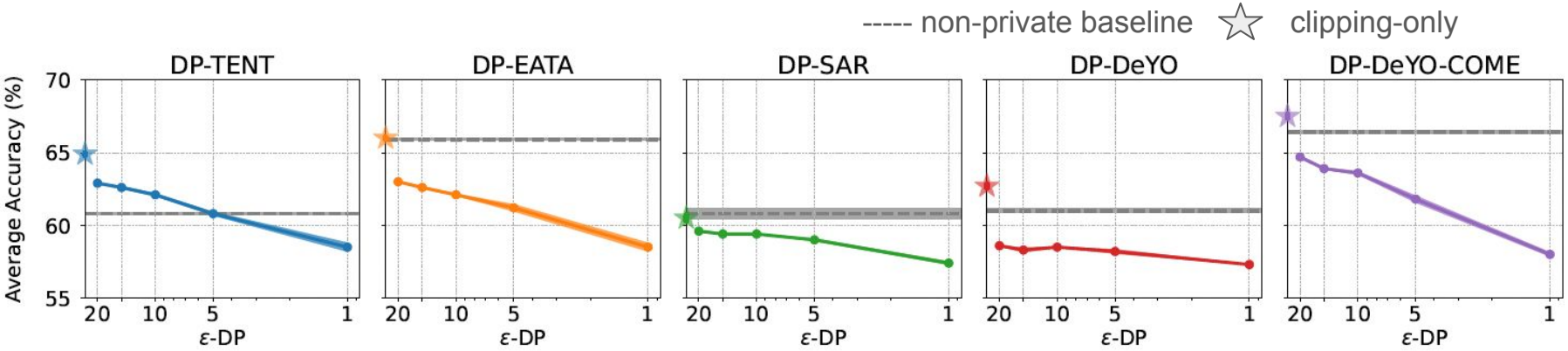
$$\Delta_t^{DP} = \frac{1}{B}\left(\sum_i \bar{g}_i + \mathcal{N}(0, C^2\sigma^2 I)\right) \longrightarrow \theta_{t+1} = \theta_t - \eta\Delta_t^{DP}$$

aggregate and add gaussian noise

private tta update

noise hides the aggregate update

DP-TTA Preserves Accuracy, and Clipping Alone Helps



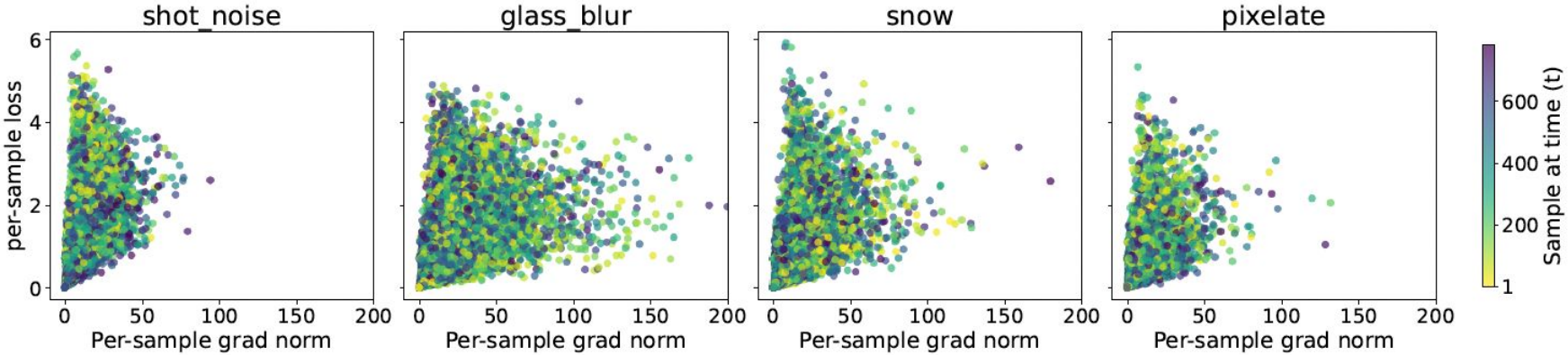
Smaller ϵ means stronger privacy but more accuracy drop.

Moderate privacy keeps useful adaptation performance.

★ Clipping without noise can improve stability and accuracy.

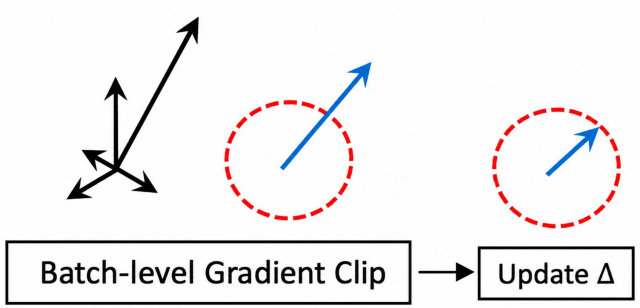
Tent: 60.8% → 64.9% DeYO-COME: 66.4% → 67.5%

Ablation: Why per-sample clipping?



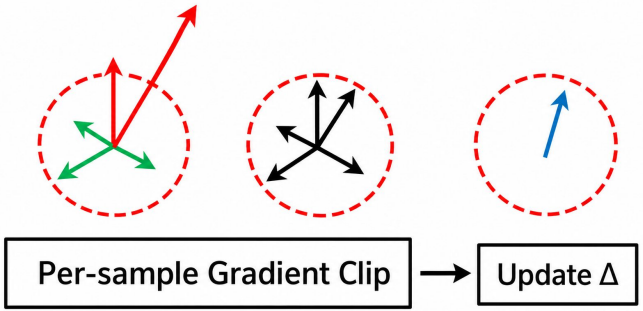
Entropy and gradient norm are weakly aligned, so entropy filtering can miss high-impact samples.

Ablation: Why per-sample clipping?



Clip after aggregate

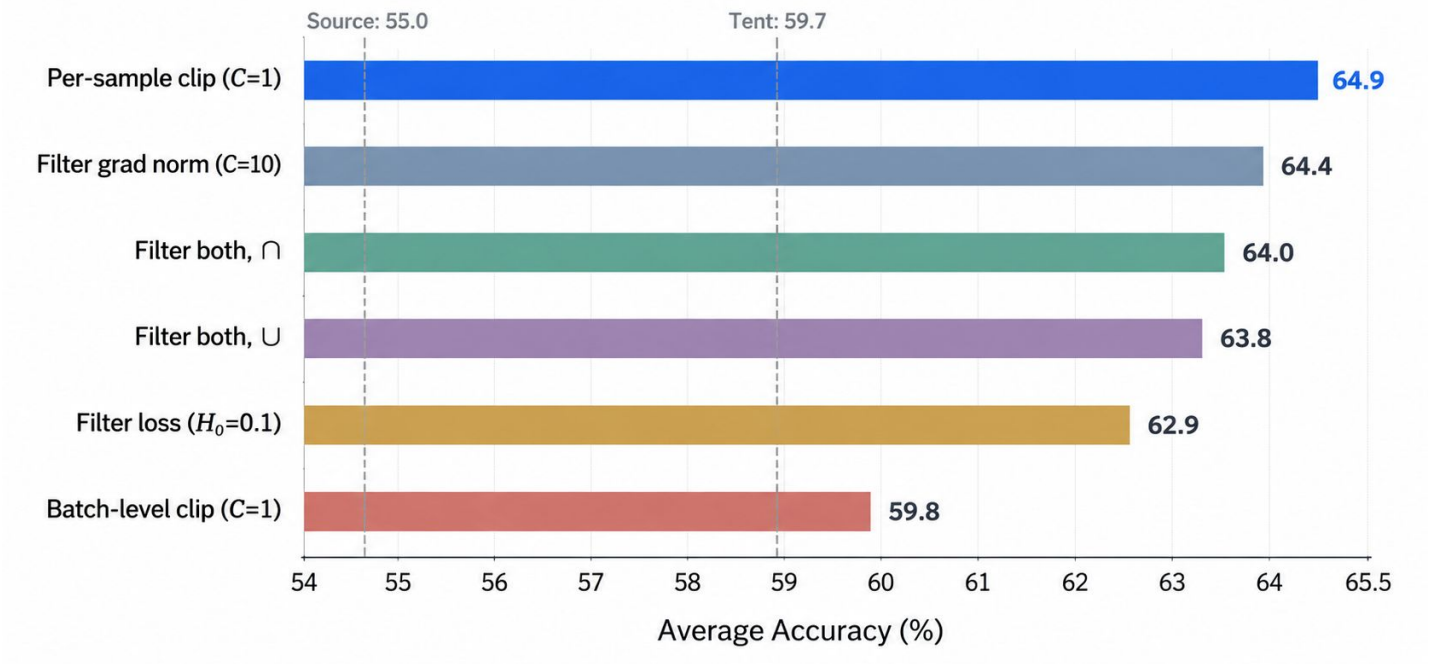
A single sample with an extremely large gradient can still **dominate** the direction of the update



Clip before aggregate

Independently limits the contribution of each **individual** sample, enforces sample-wise control in the update

Per-sample Clipping Performs Best



Compared with batch clipping and filtering, clipping each sample before aggregation gives the highest average accuracy.

Per-sample Clipping Performs Best

Method	Episodic	Continual
Source / No Update	55.0	55.0
Tent	63.0	60.8
+ Clip (Ours)	68.0	64.9
+ Regularizer	64.5	62.9
+ Filter-1	66.4	63.1
+ Filter-2	66.5	65.8
+ Filter-1 + Filter-2	65.5	63.7
+ SAM	62.5	60.7
+ SAM + Filter-1	63.2	60.9
+ SAM + Filter-3	63.0	60.0
+ SAM + Filter-1 + Filter-3	63.0	60.2
+ Adam	57.4	62.2
+ Adam + Clip	63.4	64.5

Clipping gives the best episodic accuracy and remains competitive in continual setting,

Clipping helps beyond optimizer rescaling.

Efficiency: Privacy and Accuracy Come with Modest Overhead

Method	Non-DP (ms)	DP (ms)	Clip-Only (ms)
Tent	174	189	189
EATA	178	193	195
SAR	340	362	342
DeYO	246	287	264
DeYO-COME	230	294	255

Clipping introduces only a small overhead (about **20–30 ms** per batch)

Even the slowest DP variant takes only **1.28 × time.**

Thank you!

TTA needs privacy: test-time updates depend on test data.

DP-TTA is practical: moderate privacy keeps useful accuracy with modest overhead.

Clipping is the surprise: per-sample clipping improves adaptation stability.




Visit our poster to discuss how to make test-time updates private and stable!