On the Learnability of Test-Time Adaptation

A Recovery Complexity Perspective

Zhi Zhou • Ming Yang • Shi-Yu Tian • Kun-Yang Yu • Lan-Zhe Guo • Yu-Feng Li
State Key Laboratory for Novel Software Technology, Nanjing University, China
{zhouz, liyf}@lamda.nju.edu.cn

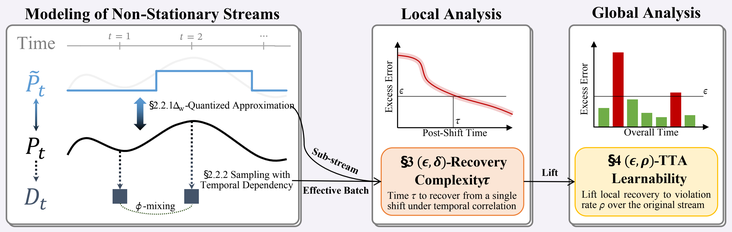
ICML International Conference On Machine Learning

NANJING UNIVERSITY

LAMDA Learning And Mining from DatA

TL;DR: The first learnability theory for TTA on non-stationary streams, defining (ϵ, δ)-**Recovery Complexity** (with matching bounds) and (ϵ, ρ)-**TTA Learnability**.

Framework at a Glance



Unified stream model → **local** recovery → **global** learnability.

Motivation & The Gap

Test-time adaptation (TTA) adapts a source model to distribution shifts using **only unlabeled** data via a *proxy loss* ψ. It is empirically successful yet can collapse under complex shifts.

Why existing theory falls short:

- **Online learning** controls *average* (regret) performance, not the **instantaneous, per-step reliability** TTA demands.
- Prior TTA bounds assume **special algorithms or label access**, offering no *global* guarantee for unlabeled supervision, non-stationarity, and correlation.

This work: a principled framework answering ***when is TTA learnable?***

Modeling the Non-Stationary Stream

A stream $\mathcal{S} = \{D_t\}_{t=1}^{T}$, $D_t \sim \mathcal{P}_t$, is captured along **two** axes.

(1) Global distribution shift: a W_1-quantized surrogate turns a continuous trajectory into a **piecewise-constant** stream (one knob Δ_W unifies **gradual & abrupt** shifts):

$$V_T = \sum_t W_1(\mathcal{P}_t, \mathcal{P}_{t+1}), \qquad \#\text{shifts } \tilde{K}_S(T) \le \left\lceil \tfrac{2V_T}{\Delta_W} \right\rceil$$

(2) Local temporal correlation: a ϕ-mixing process ($\phi(i) \le \varrho^i$) **shrinks the effective batch size** (C_ϕ=1 iff independent):

$$B_{\text{eff}} = \frac{B}{C_\phi} \le B, \qquad C_\phi = 1 + \frac{4\sqrt{\varrho}}{1-\sqrt{\varrho}}$$

A Well-Defined Target: Proxy-Optimal Competitor

Perfect adaptation is unattainable when $\psi \not\approx \ell$. We compete against the best **proxy-optimal** model in a local neighborhood $\mathcal{N}_r(\theta_1)$, matching the *constrained* adaptation of practical TTA methods:

$$\theta_t^\star \in \arg\min_{\theta \in \mathcal{N}_r(\theta_1)} \psi_t(\theta), \qquad R_t := \ell_t(\theta_t^\star), \quad \mathcal{E}_t := \ell_t(\theta_t) - R_t$$

The **excess risk** $\mathcal{E}_t$ is the central quantity throughout. It is driven down through the proxy only under (α, ζ)-**alignment**:

$$\langle \nabla\psi_t, \nabla\ell_t \rangle \;\ge\; \alpha \|\nabla\ell_t\|^2 - \zeta$$

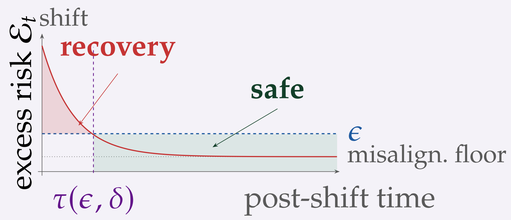
$\alpha > 0$ (alignment strength) gives a descent signal; $\zeta \ge 0$ (irreducible misalignment) sets an error floor that no amount of data can remove.

Local View: (ϵ, δ)-Recovery Complexity

The time after a distribution shift for excess risk to fall below ϵ (prob. $\ge 1-\delta$) and remain below that threshold thereafter:

$$\tau(\epsilon, \delta) := \inf\left\{ t : \sup_{u \ge t} \mathbb{P}(\mathcal{E}_u > \epsilon) \le \delta \right\}$$

Unlike regret, it controls the **length of unsatisfactory periods**.



Matching bounds (order-wise tight):

Lower (any algorithm, info. limit)
$$\tau \ge \Omega\left(\frac{C_\phi}{B\,\alpha(\sqrt{\zeta + 2\alpha\epsilon} + \sqrt{\zeta})^2} \right)$$

Upper (simple SGD baseline)
$$\tau \le O\left(\frac{C_\phi}{B\,\alpha^2\epsilon} \log \frac{\Delta_W + \epsilon}{\epsilon} \right)$$

Match up to log ⇒ **the baseline is near-minimax-optimal.**

What governs recovery:

- **Error floor:** $\zeta > 0$ leaves a non-vanishing term as $\epsilon \to 0$; misalignment caps accuracy.
- **Speed:** $\tau \propto 1/\alpha^2$ and $\tau \propto C_\phi/B$; Both alignment α, batch size B and temporal correlation matter.
- Lower & upper bounds share $C_\phi/(B\alpha^2\varepsilon)$; no algorithm can do better.

Global View: (ϵ, ρ)-TTA Learnability

(ϵ, ρ)-**learnable** if some algorithm keeps the violation fraction $\le \rho$:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{P}\big(\ell_t(\theta_t) - R_t > \epsilon\big) \le \rho$$

Recovery ⇒ Learnability transfer (Λ=L_x+$GL_{\nabla\psi}/\mu$):

$$\rho \;\le\; \delta + \frac{(\tilde{K}_S(T)+1)\,\tau(\epsilon', \delta)}{T}, \qquad \epsilon' = \epsilon - \Lambda\Delta_W$$

Strong transfer: faster recovery / fewer shifts ⇒ a smaller ρ.

Connection to dynamic regret: $\mathrm{Reg}(T) \le T(\epsilon + M\rho)$; persistent shifts ($\Delta_W$=$\Omega(1)$) force linear regret, while the benign regime recovers classical $o(T)$.

Empirical Validation

Synthetic Experiment (1-D recovery, vary α at B=16)

α	LB	τ	UB
0.05	59.0	322.0	311.9
0.10	15.6	77.0	78.0
0.20	4.1	19.0	19.5
0.50	0.7	4.0	3.1

- $\tau \cdot \alpha^2 \approx$ **const** ⇒ near-tight: τ tracks the predicted $O(1/\alpha^2)$ scaling across all α
- **LB** $\le \tau \le$ **UB up to a** log **factor:** τ stays sandwiched between the theoretical bounds

Real-world Experiment (ImageNet-C, temporal correlation $\beta\downarrow$)

β	$\tilde{\alpha}$	ΔAcc.	#Improve
0.001	−0.028	−12.7%	0/15
0.01	+0.090	+3.6%	12/15
0.1	+0.157	+13.7%	15/15
Uniform	+0.171	+14.7%	15/15

- **Alignment α is the single factor governing whether TTA succeeds.**

Takeaway: Recovery complexity → TTA learnability.
Pinning down *when* adaptation is possible and *what* limits it.

Scan for Paper · Telegram · WeChat · Slides · Homepage

Please feel free to contact **Zhi Zhou** zhouz@lamda.nju.edu.cn