Confidence-Guided Self-Training for Gradual Domain Adaptation

Akram Heidarizadeh[1] Akram Awad[1] HanQin Cai[2,3] George Atia[1,3]

[1]Dept. of Electrical and Computer Engineering, University of Central Florida, Orlando, FL, USA [2]School of Data, Mathematical, and Statistical Sciences, University of Central Florida, Orlando, FL, USA

[3]Dept. of Computer Science, University of Central Florida, Orlando, FL, USA

Problem & Motivation

- **Domain Adaptation (DA):** Transfers a model from a labeled *source* domain μ_0 to an unlabeled *target* domain μ_K under distribution shift.
- **Gradual DA (GDA):** Shift occurs through a sequence of intermediate domains $\mu_0, \mu_1, \ldots, \mu_K$, enabling smoother adaptation [1, 2].
- **Self-training** iteratively assigns pseudo-labels to unlabeled data. But pseudo-label **errors accumulate** across rounds, especially under large shift.

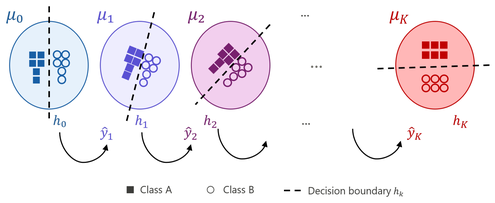


Figure 1. Gradual domain adaptation with self-training.

Open Question

- Theoretical effect of confidence/margin filtering on generalization in iterative GDA.
- Provable justification for percentile thresholding.

Setup & Notation

- $\mathcal{X} \subseteq \mathbb{R}^d$; label set $[C]$; classifier $h : \mathcal{X} \to \Delta_C$; $\hat{y}(x) = \arg\max_c h_c(x)$.
- At round k: receive unlabeled $S_k \sim \mu_k$, score each point, accept $\hat{S}_k$, train h_k via regularized ERM.

Filtering rules

$$\text{Confidence:} \quad \mathsf{A}_k(x) = \mathbf{1}\{c_{k-1}(x) \geq \tau_k\}, \quad c(x) = \max_c h_c(x)$$
$$\text{Margin:} \quad \mathsf{A}_k(x) = \mathbf{1}\{\text{margin}_{k-1}(x) \geq m_k\}$$

Coverage: $\rho_k := \mathbb{E}_{x\sim\mu_k}[\mathsf{A}_k(x)]$ (fraction of accepted samples).

Pseudo-labeling error on the accepted set:

$$\varepsilon_k := \max_{h\in\{h_k, h_{k-1}\}} \left| \overline{\mathcal{E}}_k(h) - \widehat{\mathcal{E}}_k(h) \right|$$

where $\overline{\mathcal{E}}_k$ is the masked true risk and $\widehat{\mathcal{E}}_k$ the masked pseudo-risk.

Main Result: Theorem 1 (Generalization Bound)

For any $\delta \in (0, 1)$, with probability $\geq 1 - \delta$:

$$\mathcal{E}_{\mu_K}(h_K) - \mathcal{E}_{\mu_0}(h_0) \leq \sum_{k=1}^{K} \left((1 - \rho_k) + 2\varepsilon_k + \frac{C_\delta}{\sqrt{n_k}} + LW_1(\mu_k, \mu_{k-1}) + \lambda B^2 \right),$$

Key features:

- Modular decomposition — isolates ε_k from shift and sampling terms.
- Filter-agnostic: works with confidence or margin.
- No stability assumptions; separates coverage penalty from substitution error — absent in prior work [1, 2, 3].

Contributions

1. **Generalization bound** explicitly tracking rejection rate $(1 - \rho_k)$ and accepted-set pseudo-label error ε_k at each round.
2. **Confidence- & margin-based control** of ε_k via calibration and Tsybakov-type margin decay.
3. **First theoretical justification** of percentile (quantile) thresholding: directly controls coverage while tightening ε_k, yielding $O(\log K)$ accumulation.
4. **CFSTDA** (Confidence-Filtered Self-Training for DA) and **MFSTDA** (Margin-Filtered Self-Training for DA): our two proposed methods.
5. **Empirical validation** on synthetic + 4 real GDA benchmarks, with observed and OT-generated intermediate domains.

Filtering-Based Control of ε_k

Lemma 1 (Confidence):

$$\varepsilon_k \leq \rho_k \, \phi_k(\tau_k) \leq \phi_k(\tau_k)$$

where $\phi_k(\tau) := \Pr_{(x,y)\sim\mu_k}\big[\hat{y}_{k-1}(x) \neq y \mid c_{k-1}(x) \geq \tau\big]$

Lemma 2 (Margin):

$$\varepsilon_k \leq \rho_k \, \zeta_k(m_k) \leq \zeta_k(m_k)$$

where $\zeta_k(m) := \Pr\big[\hat{y}_{k-1}(x) \neq y \mid \text{margin}_{k-1}(x) \geq m\big]$

Corollary (Log-accumulation): If $\phi_k(\tau_k) \leq C_a/k$ and $1 - \rho_k \leq c/k$, then

$$\sum_{k=1}^{K} \big[(1 - \rho_k) + 2\varepsilon_k\big] \;\leq\; (c + 2C_a)(1 + \log K) \;=\; O(\log K)$$

Tsybakov connection: Under power-law decay $\zeta_k(m) \leq Cm^{-\alpha}$, schedules $m_k \propto k^{1/\alpha}$ give $\varepsilon_k = O(k^{-\alpha})$ and $\sum_k \varepsilon_k = O(\log K)$ for $\alpha = 1$.

Percentile Thresholding Algorithm

Key idea: set θ_k as the empirical $(1 - q_k)$-quantile of scores on S_k, where $\text{score}_{k-1}(x) \in \{c_{k-1}(x),\ \text{margin}_{k-1}(x)\}$.

Define population coverage and conditional error at threshold θ:

$$\rho_k(\theta) := \Pr_{x\sim\mu_k}\big[\text{score}_{k-1}(x) \geq \theta\big],$$
$$\psi_k(\theta) := \Pr_{(x,y)\sim\mu_k}\big[\hat{y}_{k-1}(x) \neq y \;\big|\; \text{score}_{k-1}(x) \geq \theta\big].$$

Schedule: $q_k = 1 - c/k \;\Rightarrow\; 1 - \rho_k \leq c/k$ (shrinking rejected tail).

1. Compute scores $\{\text{score}_{k-1}(x)\}_{x\in S_k}$.
2. Set $\theta_k =$ empirical $(1 - q_k)$-quantile.
3. $\hat{S}_k = \{(x, \hat{y}_{k-1}(x)) : \text{score}_{k-1}(x) \geq \theta_k\}$.
4. Train h_k by regularized ERM on $\hat{S}_k$.

Proposition 1: Under Assumption 1 ($\psi_k(\theta) \leq C_t(1 - \rho_k(\theta))^\alpha$):

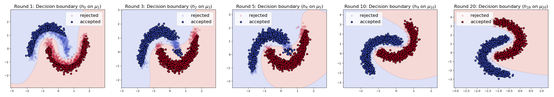
$$\varepsilon_k \;\leq\; C_t\Big(\frac{c}{k} + \sqrt{\frac{2\log(2K/\delta)}{M_k}}\Big)^\alpha \;=\; O(k^{-\alpha}).$$

Hence $\sum_k \varepsilon_k = O(\log K)$ for $\alpha = 1$.

Class balancing: minimum 10% per-class retention prevents minority collapse under imbalance.

Synthetic Dataset: Two Moons

90° rotation, 20 steps;
Schedule $q_k = 1 - 0.5/k$; 1000 samples/domain.



$1-\rho_k$=0.50 $1-\rho_k$=0.17 $1-\rho_k$=0.10 $1-\rho_k$=0.05 $1-\rho_k$=0.02

Figure 2. CFSTDA decision boundaries at rounds $k \in \{1, 3, 5, 10, 20\}$ of CFSTDA. Filled = accepted pseudo-labels; × = rejected. Rejection rate decreases as coverage grows.

Empirical Validation of Bounds

45 rounds, 2°/step rotation; top $q_k = 1 - c/k$ fraction retained by confidence (CFSTDA)

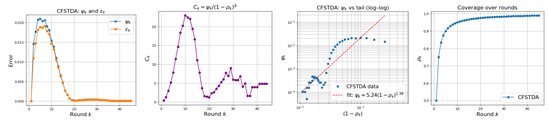


Figure 3. **Left:** conditional error ψ_k and pseudo-labeling error ε_k remain bounded; $\varepsilon_k \leq \psi_k$ (Lemmas 1–2). **Center-left:** ratio $\psi_k/(1 - \rho_k)^{\hat{\alpha}}$ proportional to the scaled errors $k^{\hat{\alpha}}\psi_k$ and $k^{\hat{\alpha}}\varepsilon_k$. (Assumption 1 holds). **Center-right:** log–log plot of ψ_k against $(1 - \rho_k)$ with fitted slope $\hat{\alpha} \approx 2.39$, **Right:** empirical coverage closely tracks q_k.

Ablation: Quantile Schedule Parameter c

Varying c in $q_k = 1 - c/k$: larger c is more selective in early rounds (lower coverage, lower ε_k), gradually relaxing as the model improves.

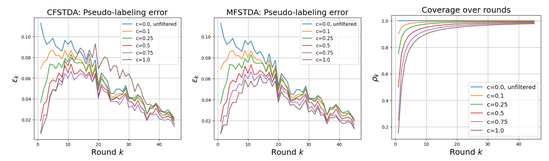


Figure 4. $c \in \{0, 0.1, 0.25, 0.5, 0.75, 1.0\}$. CFSTDA (left), MFSTDA (center), coverage ρ_k (right). Moderate-to-large c gives best coverage–noise trade-off.

Effect of Class Imbalance (30% vs. 70% prior)

Per-class minimum retention (10–20%) prevents minority collapse under class imbalance

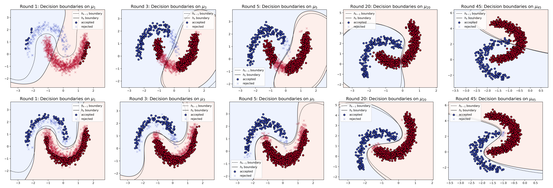


Figure 5. **Top:** without balancing — minority class collapses. **Bottom:** per-class retention preserves both classes across all rounds $k \in \{1, 3, 5, 20, 45\}$.

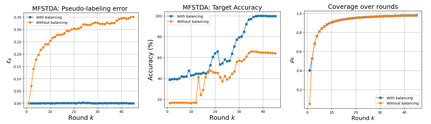


Figure 6. **Left:** pseudo-labeling error ε_k; **Center:** target-domain accuracy; **Right:** coverage ρ_k.

Real-World GDA Benchmarks

- **Rotated MNIST**: $0^\circ \to 45^\circ$, 8 intermediate domains.
- **Color-Shift MNIST**: pixel shift $[0, 1] \to [1, 2]$, 10 steps.
- **Portraits**: 18k yearbook photos 1905–2013 (temporal drift).
- **Cover Type**: tabular (54 features), sorted by distance to water.

Method	Rotated MNIST	Color-Shift	Portraits	Cover Type
Baseline	44.3±1.7	35.5±11.1	75.6±1.2	61.2±4.9
DANN [4]	48.0±6.4	37.9±22.6	76.8±3.2	66.0±2.6
DeepCORAL [5]	51.6±3.1	50.2±24.9	74.7±0.8	65.9±2.4
GST [1]	59.3±5.2	53.2±14.0	76.3±1.7	67.8±4.6
GOAT [2]	64.3±3.5	82.8±11.6	81.1±2.6	70.4±1.7
CFSTDA	**67.7**±5.4	88.8±5.5	79.8±3.1	71.7±2.9
MFSTDA	65.5±4.7	**91.1**±3.4	**84.2**±0.6	**72.5**±2.3

Table 1. Target accuracy (%). **Bold**=best; underline=second best.

Office-Home: Beyond Native GDA

Method	Office-Home Real → Product				Office-Home Art → Product			
Baseline	74.0±0.7				62.8±1.2			
DANN [4]	73.8±0.7				62.9±0.9			
DeepCoral[5]	74.0±0.4				63.2±1.1			
	0 gen	1 gen	2 gen	3 gen	0 gen	1 gen	2 gen	3 gen
GOAT [2]	74.2±1.0	74.7±1.0	73.1±0.6	72.2±0.4	63.5±0.9	62.8±0.9	61.0±1.3	57.9±1.2
CFSTDA	74.7±0.5	76.3±0.8	75.8±0.9	75.5±0.4	65.5±0.8	**67.6**±1.2	66.3±1.0	64.2±1.7
MFSTDA	75.0±0.8	**76.8**±1.4	76.1±1.7	75.8±0.7	65.9±1.0	67.2±1.2	67.0±1.2	64.1±1.1

Table 2. Evaluation on Office-Home under direct and generated gradual adaptation settings.

Conclusion

- **First modular bound** for iterative GDA self-training explicitly tracking ε_k and $(1 - \rho_k)$.
- **Logarithmic error accumulation** under percentile schedules: $\sum_k[(1 - \rho_k) + 2\varepsilon_k] = O(\log K)$.
- **CFSTDA & MFSTDA** consistently outperform GST and GOAT on all 4 benchmarks.
- Filtering benefits extend to standard UDA (Office-Home).

References

[1] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International conference on machine learning (ICML)*, pages 5468–5479. PMLR, 2020.

[2] Rui He, Chao Wang, Jiangchao Li, and Boqing Gong. Gradual domain adaptation: Theory and algorithms. *Journal of Machine Learning Research*, 25:1–40, 2024.

[3] Haoxiang Wang, Bo Li, and Han Zhao. Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. *Proceedings of Machine Learning Research*, 162:22784–22801, 2022.

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[5] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

This work was supported by DARPA under Agreement No. HR0011-24-9-0427 and NSF under Awards CCF-2106339 and DMS-2304489