When Does Curriculum Learning Help? A Theoretical Perspective

Raman Arora, Yunjuan Wang, Kaibo Zhang

Johns Hopkins University

NeurIPS 2025

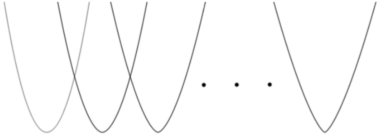
Motivation: Learning Like Humans

Curriculum Learning: Learn complex tasks by progressing through simpler ones

Key Question: When and why does curriculum learning provably help?

Our Approach:

- Develop theoretical framework for multi-task curriculum learning
- Identify conditions for a "good" curriculum
- Provide generalization guarantees (convex & non-convex)



Progressive task difficulty with similar loss landscapes

Our Framework: Biased RERM + (r, α) Condition

Biased Regularized ERM:

$$\hat{w}_t \in \arg\min_{w} \left\{ \hat{L}_{S_t}(w) + \frac{\mu_t}{2} \|w - \hat{w}_{t-1}\|_2^2 \right\}$$

The (r, α) Condition for "Good" Curricula

Tasks $t-1$ and t satisfy (r_t, α) if:

$$\inf_{w' : \|w' - w\|_2 \le r_t} \varepsilon_t(w') \le \alpha \varepsilon_{t-1}(w), \quad \alpha \in (0, 1)$$

Intuition: If predictor w has small excess risk on task $t-1$, there exists a nearby predictor (within radius r_t) with proportionally small excess risk on task t

$\Rightarrow$ Small r_t enables efficient knowledge transfer!

Main Theoretical Results

Theorem (Convex Lipschitz Tasks)

For convex, ρ_t-Lipschitz tasks satisfying (r_t, α) condition:

$$\mathbb{E}[\varepsilon_t(\hat{w}_t)] \leq \frac{2r_t\rho_t}{\sqrt{n_t}} + \alpha\mathbb{E}[\varepsilon_{t-1}(\hat{w}_{t-1})]$$

Theorem (Smooth Non-negative Convex Tasks)

For H_t-smooth tasks with optimal loss L_t^*:

$$\mathbb{E}[\varepsilon_t(\hat{w}_t)] \leq \sqrt{\frac{32L_t^*H_tr_t^2}{n_t}} + \frac{9H_tr_t^2}{(1-\alpha)n_t} + \frac{1+\alpha}{2}\mathbb{E}[\varepsilon_{t-1}(\hat{w}_{t-1})]$$

Fast rate when $L_t^* = 0$ (realizable): $O(1/n_t)$

Key insight: Sample complexity decreases with small r_t and good initialization from previous task!

Extensions and Additional Results

1. SGD-based Training

Same bounds achievable with efficient SGD (learning rate $\eta_t = \frac{r_t}{\rho_t \sqrt{n_t}}$)

2. Non-convex Settings

For Lipschitz non-convex losses via ERM with constraint $\|w - \hat{w}_{t-1}\|_2 \leq r_t$:

$$\varepsilon_t(\hat{w}_t) \leq 2\epsilon + \alpha\varepsilon_{t-1}(\hat{w}_{t-1}) \text{ with high probability}$$

3. Application to Adversarial Robustness

Results extend to adversarially robust learning by replacing standard loss with:

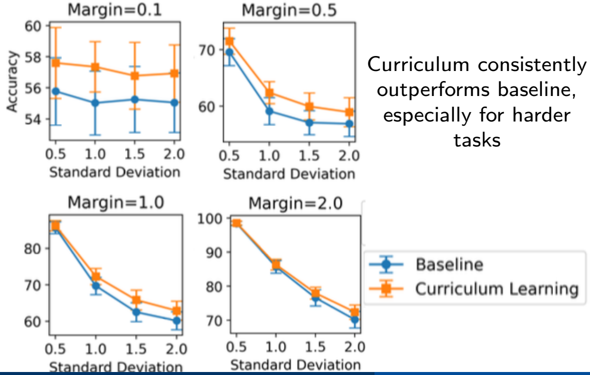
$$\ell_t^{\text{rob}}((x, y); w) := \sup_{\tilde{x} \in \mathcal{B}(x)} \ell_t((\tilde{x}, y); w)$$

Convexity and Lipschitzness preserved!

Experimental Validation

Synthetic Data (Gaussian Mixtures)

- Easy task: large margin ($\gamma = 3$), low variance ($\sigma = 0.5$)
- Hard task: varying $\gamma \in [0.1, 2.0]$, $\sigma \in [0.5, 2.0]$



Curriculum consistently outperforms baseline, especially for harder tasks

Adversarial Training on MNIST

- Progressive attack strength: $\alpha t/T$ for $t \in [T]$
- ℓ_2 regularization to previous model

α	$T = 1$ (Baseline) Nat	$T = 1$ (Baseline) PGD	$T = 3$ (Curriculum) Nat	$T = 3$ (Curriculum) PGD
0.1	99.18	96.07	**99.36**	**95.74**
0.3	98.27	92.77	**98.23**	**93.61**
0.4	11.35	11.35	**98.52**	**95.63**

Significant improvements for stronger attacks

Conclusion and Impact

Main Contributions

1. (r, α) **condition**: Simple, interpretable criterion for curriculum quality
2. **Theoretical guarantees**: Excess risk bounds for convex and non-convex settings
3. **Practical insights**: Guidelines for task ordering and regularization parameter selection
4. **Validated framework**: Synthetic and real data experiments confirm theory

Key Takeaway: Small r_t (nearby optimal solutions) enables reduced sample complexity through curriculum learning

Future Directions:

- Practical methods to verify (r, α) condition
- Data-driven curriculum design
- Extensions to deep learning architectures

Thank you!