On the Statistical Benefits of Curriculum Learning

Ziping Xu, Ambuj Tewari

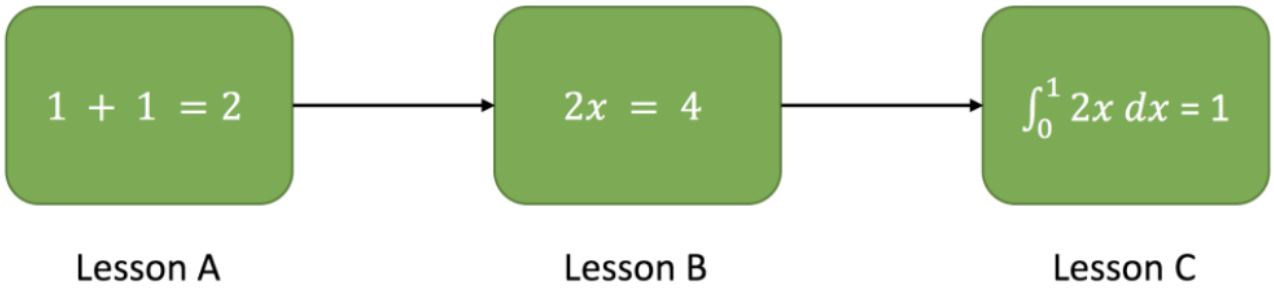
University of Michigan, Statistics

July 10, 2022

Background

Motivations:

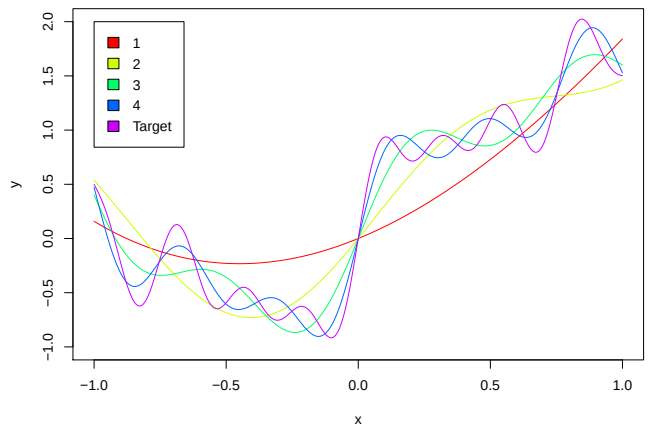
1. Multi-task learning can improve learning efficiency.
2. In some problems, we can decide the *order* of learning and the *number of observations* from each task.
3. Curriculum learning (CL) refers to any strategy that improves the performance with a better task scheduling [1].
4. The idea rooted deeply in the way human learns.



Two Benefits

There has been two understandings on the benefits of CL:

1. Optimization benefits: solutions for previous tasks serve as better initial points for later tasks that are highly nonconvex globally but convex locally.



2. Statistical benefits: any benefits except for the reduction in the optimization difficulties.
   Example: three identical linear regression tasks except for the different noise levels.

Problem Setup

We studied two setups for multitask linear regression tasks:

1. Unstructured linear regression: T linear regression tasks with a single target task. No structural information assumed.

$$\underset{\mathbb{R}^d}{Y_t} = \underset{\mathbb{R}^d}{X_t^T} \underset{\mathbb{R}^d}{\theta_t^*} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_t^2), X_t \sim \mathcal{N}(0, \Sigma^2).$$

2. Structured linear regression: T linear regression tasks sharing the same low dimensional representation function.

$$\underset{\mathbb{R}^d}{Y_t} = \underset{\mathbb{R}^d}{X_t^T} \underset{\mathbb{R}^{d\times k}}{B^*} \underset{\mathbb{R}^k}{\theta_t^*} + \epsilon_t, \text{where } B \text{ is the shared linear representation.}$$

Results Overview (Informal)

Unstructured linear regression:

1. The optimal selects the task that minimizes

$$\underbrace{\|\theta_t^* - \theta_{target}^*\|_2^2}_{\text{Transfer distance}} + \frac{d\sigma_t^2}{N}.$$

2. For task schedulers that adaptively learns to schedule one can achieve a MSE rate of

$$\underbrace{\frac{\sigma_T^2 \log(T)}{N}}_{\text{Extra error to identify good tasks}} + \min_t \left\{ \|\theta_t^* - \theta_{target}^*\|_2^2 + \frac{d\sigma_t^2}{N} \right\}$$

Structured linear regression:

1. Optimal curriculum schedules tasks such that the matrix does not degenerate

$$M = \sum_{i=1}^{N} \theta_{t_i}^* \theta_{t_i}^{*T}, \text{ where } t_i \text{ is the task scheduled at step } i.$$

2. Adaptive learnt curriculum can be as well as the optimal curriculum.


Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in **Proceedings of the 26th annual international conference on machine learning**, 2009, pp. 41–48.