Supplementary Materials: Self-paced Curriculum Learning

Lu Jiang[1], **Deyu Meng**[1,2], **Qian Zhao**[1,2], **Shiguang Shan**[1,3], **Alexander G. Hauptmann**[1]
[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, 15217
[2] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi, P. R. China, 710049
[3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P. R. China, 100190
lujiang@cs.cmu.edu, dymeng@mail.xjtu.edu.cn,
timmy.zhaoqian@gmail.com, sgshan@ict.ac.cn, alex@cs.cmu.edu

In this supplementary material, we present an illustrative toy example along with the proofs for Theorem 1 and Theorem 2 in the main body of our paper.

Example 1. *Given six samples a, b, c, d, e, f. In the current iteration, the losses for these samples are $\ell = [0.1, 0.2, 0.4, 0.6, 0.5, 0.3]$, respectively. A latent ground-truth curriculum is listed in the first row of the following table, followed by the curriculum of CL, SPL and SPCL. For simplicity, binary scheme is used in SPL and SPCL where $\lambda = 0.8333$. If two samples with the same weight, we rank them in ascending order of their losses, in order to break the tie. The Kendall's rank correlation is presented in the last column.*

Method	*Curriculum*	*Correlation*
Ground-Truth	a, b, c, d, e, f	-
CL	b, a, d, c, e, f	*0.73*
SPL	a, b, f, c, e, d	*0.46*
SPCL	a, b, c, d, e, f	*1.00*

The curriculum region used is a linear constraint $\mathbf{a}^T\mathbf{v} \leq 1$, where $\mathbf{a} = [0.1, 0.0, 0.4, 0.3, 0.5, 1.0]^T$. In the implementation, we add a small constant 10^{-7} in the constraints for optimization accuracy. The constraint follows Definition 2 in the paper. As shown, both CL and SPL yield the suboptimal curriculum, e.g. their correlations are only 0.73 and 0.46. However, SPCL exploits the complementary information in CL and SPL, and devises an optimal curriculum. Note that CL recommends to learn b before a, but SPCL disobeys this order in the actual curriculum. The final solution of SPCL is $\mathbf{v}^* = [1.00, 1.00, 1.00, 0.88, 0.47, 0.00]$.

When the predetermined curriculum is completely wrong, SPCL may still be robust to the inferior prior knowledge given reasonable curriculum regions are applied. In this case, the prior knowledge should not be encoded as strong constraints. For example, in the above example, we can use the following curriculum region to encode the completely incorrect predetermined curriculum: $\mathbf{a}^T\mathbf{v} \leq 6.0$, where $\mathbf{a} = [2.3, 2.2, 2.1, 2.0, 1.7, 1.5]^T$

Method	*Curriculum*	*Correlation*
CL	f, e, d, c, b, a	*-1.00*
SPL	a, b, f, c, e, d	*0.46*
SPCL	a, f, b, c, e, d	*0.33*

As we see, even though the predetermined curriculum is completely wrong (correlation -1.00), the proposed SPCL still obtains reasonable curriculum (correlation 0.33). This is because SPCL is able to leverage information in both prior knowledge and learning objective. The optimal solution of SPCL is $\mathbf{v}^* = [1.00, 0.91, 0.10, 0.00, 0.00, 1.00]$.

Theorem 1. *For training samples $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, given a curriculum $\gamma(\cdot)$ defined on it, denote $\mathbf{v} = [v_1, v_2, \cdots, v_n]^T$ as its weight variables in Eq. (3) of the maintext. The feasible region defined by*

$$\Psi = \{\mathbf{v} | \mathbf{a}^T\mathbf{v} \leq c\},$$

where $\mathbf{a} = [a_1, a_2, \cdots, a_n]^T$. Ψ is a curriculum region of $\gamma(\cdot)$ if 1) $\Psi \wedge \mathbf{v} \in [0, 1]^n$ is nonempty; 2) for any pair of samples $\mathbf{x}_i, \mathbf{x}_j$, if $\gamma(\mathbf{x}_i) < \gamma(\mathbf{x}_j)$, it holds that $\int_\Psi v_i \, d\mathbf{v} > \int_\Psi v_j \, d\mathbf{v}$, where $\int_\Psi v_i \, d\mathbf{v}$ calculates the expectation of v_i within Ψ; 3) if $\gamma(\mathbf{x}_i) = \gamma(\mathbf{x}_j)$, $\int_\Psi v_i \, d\mathbf{v} = \int_\Psi v_j \, d\mathbf{v}$.

Proof. (1) $\Psi \wedge \mathbf{v} \in [0, 1]^n$ is a nonempty convex set.

(2) For $\mathbf{x}_i, \mathbf{x}_j$ with $\gamma(\mathbf{x}_i) < \gamma(\mathbf{x}_j)$, denote $\Psi_{ij} = \{\mathbf{v}_{ij} | \mathbf{a}_{ij}^T\mathbf{v}_{ij} \leq c\}$, $\mathbf{a}_{ij}$/$\mathbf{v}_{ij}$ the sub-vector of $\mathbf{a}$/$\mathbf{v}$ by wiping off its ith and jth elements, respectively, we can then calculate the expected value of v_i on the region $\Psi = \{\mathbf{v} | \mathbf{a}^T\mathbf{v} \leq c\}$ as:

$$\begin{aligned} E(v_i) &= \int_\Psi v_i \, d\mathbf{v} \\ &= \int_{\Psi_{ij}} \int_0^{\frac{c - \mathbf{a}_{ij}^T\mathbf{v}_{ij}}{a_j}} \int_0^{\frac{c - \mathbf{a}_{ij}^T\mathbf{v}_{ij} - a_j v_j}{a_i}} v_i dv_i dv_j d\mathbf{v}_{ij} \\ &= \int_{\Psi_{ij}} \int_0^{\frac{c - \mathbf{a}_{ij}^T\mathbf{v}_{ij}}{a_j}} \frac{\left(c - \mathbf{a}_{ij}^T\mathbf{v}_{ij} - a_j v_j\right)^2}{2a_i^2} dv_j d\mathbf{v}_{ij} \\ &= \frac{\int_{\Psi_{ij}} \left(c - \mathbf{a}_{ij}^T\mathbf{v}_{ij}\right)^3 d\mathbf{v}_{ij}}{6a_i^2 a_j}. \end{aligned}$$

In the similar way, we can get that:

$$E(v_j) = \int_{\Psi} v_j \, d\mathbf{v} = \frac{\int_{\Psi_{ij}} \left(c - \mathbf{a}_{ij}^T \mathbf{v}_{ij}\right)^3 d\mathbf{v}_{ij}}{6a_j^2 a_i}.$$

We thus can get that

$$E(v_i) - E(v_j) = \frac{\int_{\Psi_{ij}} \left(c - \mathbf{a}_{ij}^T \mathbf{v}_{ij}\right)^3 d\mathbf{v}_{ij}}{6a_i^2 a_j^2}(a_j - a_i) > 0.$$

Similarly, we can prove that $\int_{\Psi} v_i \, d\Psi = \int_{\Psi} v_j \, d\Psi$ for $\gamma(\mathbf{x}_i) = \gamma(\mathbf{x}_j)$.

The proof is then completed. ■

Theorem 2. *The binary, linear, logarithmic and mixture scheme are self-paced functions.*

Proof. We first prove the above functions satisfying Condition 1 in Definition 3, i.e. they are convex with respect to $\mathbf{v} \in [0,1]^n$, where n is the number of samples. As binary, linear, logarithmic and mixture self-paced functions can be decoupled $f(\mathbf{v};\lambda) = \sum_{i=1}^n f(v_i;\lambda)$:

For binary scheme $f(v_i;\lambda) = -\lambda v_i$:

$$\frac{\partial^2 f}{\partial^2 v_i} = 0. \tag{1}$$

For linear scheme $f(v_i;\lambda) = \frac{1}{2}\lambda(v_i^2 - 2v_i)$:

$$\frac{\partial^2 f}{\partial^2 v_i} = \lambda > 0, \tag{2}$$

where $\lambda > 0$.

For logarithmic scheme $f(v_i;\lambda) = \zeta v_i - \frac{\zeta^{v_i}}{\log \zeta}$:

$$\frac{\partial^2 f}{\partial^2 v_i} = -\frac{1}{\log \zeta}\zeta^{v_i} > 0, \tag{3}$$

where $\zeta = 1 - \lambda$ and $\lambda \in (0,1)$.

For mixture scheme $f(v_i;\lambda) = -\zeta \log(v_i + \frac{1}{\lambda_1}\zeta)$:

$$\frac{\partial^2 f}{\partial^2 v_i} = \frac{\zeta\lambda_1^2}{(\zeta + \lambda_1 v_i)^2} > 0 \tag{4}$$

where $\lambda = [\lambda_1, \lambda_2]$, $\zeta = \frac{\lambda_1\lambda_2}{\lambda_1 - \lambda_2}$, and $\lambda_1 > \lambda_2 > 0$.

As the above second derivatives are non-negative, and the sum of convex functions is convex, we have $f(\mathbf{v};\lambda)$ for binary, linear, logarithmic and mixture scheme are convex.

We then prove the above functions satisfying Condition 2 that is when all variables are fixed except for v_i, ℓ_i, v_i^* decreases with ℓ_i

Denote $\mathbb{E}_{\mathbf{w}} = \sum_{i=1}^n v_i\ell_i + f(\mathbf{v};\lambda)$ as the objective with the fixed model parameters $\mathbf{w}$, where ℓ_i is the loss for the i^{th} sample. The optimal solution $\mathbf{v}^* = [v_1^*, \cdots, v_n^*]^T = \arg\min_{\mathbf{v}\in[0,1]^n} \mathbb{E}_{\mathbf{w}}$.

For binary scheme:

$$\begin{aligned} \mathbb{E}_{\mathbf{w}} &= \sum_{i=1}^n (\ell_i - \lambda)v_i; \\ \frac{\partial \mathbb{E}_{\mathbf{w}}}{\partial v_i} &= \ell_i - \lambda = 0; \\ \Rightarrow v_i^* &= \begin{cases} 1 & \ell_i < \lambda \\ 0 & \ell_i \geq \lambda. \end{cases} \end{aligned} \tag{5}$$

For linear scheme:

$$\begin{aligned} \mathbb{E}_{\mathbf{w}} &= \sum_{i=1}^n \ell_i v_i + \frac{1}{2}\lambda(v_i^2 - 2v_i); \\ \frac{\partial \mathbb{E}_{\mathbf{w}}}{\partial v_i} &= \ell + v_i\lambda - \lambda = 0; \\ \Rightarrow v_i^* &= \begin{cases} -\frac{1}{\lambda}\ell + 1 & \ell_i < \lambda \\ 0 & \ell_i \geq \lambda. \end{cases} \end{aligned} \tag{6}$$

For logarithmic scheme:

$$\begin{aligned} \mathbb{E}_{\mathbf{w}} &= \sum_{i=1}^n \ell_i v_i + \zeta v_i - \frac{\zeta^{v_i}}{\log \zeta}; \\ \frac{\partial \mathbb{E}_{\mathbf{w}}}{\partial v_i} &= \ell + \zeta - \zeta^{v_i} = 0; \\ \Rightarrow v_i^* &= \begin{cases} \frac{1}{\log \zeta}\log(\ell + \zeta) & \ell_i < \lambda \\ 0 & \ell_i \geq \lambda. \end{cases} \end{aligned} \tag{7}$$

where $\zeta = 1 - \lambda$ $(0 < \lambda < 1)$.

For mixture scheme:

$$\begin{aligned} \mathbb{E}_{\mathbf{w}} &= \sum_{i=1}^n \ell_i v_i - \zeta \log(v_i + \frac{1}{\lambda_1}\zeta); \\ \frac{\partial \mathbb{E}_{\mathbf{w}}}{\partial v_i} &= \ell - \frac{\zeta\lambda_1}{\zeta + \lambda_1 v_i} = 0; \\ \Rightarrow v_i^* &= \begin{cases} 1 & \ell_i \leq \lambda_2 \\ 0 & \ell_i \geq \lambda_1 \\ \frac{(\lambda_1 - \ell)\zeta}{\ell\lambda_1} & \lambda_2 < \ell_i < \lambda_1 \end{cases} \end{aligned} \tag{8}$$

where $\lambda = [\lambda_1, \lambda_2]$, and $\zeta = \frac{\lambda_1\lambda_2}{\lambda_1 - \lambda_2}$, $(\lambda_1 > \lambda_2 > 0)$.

By setting the partial gradient to zero we arrive the optimal solution of $\mathbf{v}$. It is obvious that v_i is decreasing with respect to ℓ_i in all functions. In all cases, we have that $\lim_{\ell_i \to 0} v_i^* = 1$, $\lim_{\ell_i \to \infty} v_i^* = 0$.

Finally, we prove that the above functions satisfying Condition 3 that is $\|\mathbf{v}\|_1$ increases with respect to λ, and it holds that $\forall i \in [1,n]$, $\lim_{\lambda \to 0} v_i^* = 0$, $\lim_{\lambda \to \infty} v_i^* = 1$.

It is easy to verify that each individual v_i^* increases with respect to λ in their closed-form solutions in Eq. (5), Eq. (6), Eq. (7) and Eq. (8) (in mixture scheme, let $\lambda = \lambda_1$ represent the model age). Therefore $\|\mathbf{v}\|_1 = \sum_{i=1}^n v_i$ also increases with respect to λ. In an extreme case, when λ approaches positive infinity, we have $\forall i \in [1,n] v_i = 1$, i.e. $\lim_{\lambda \to \infty} v_i^* = 1$ in Eq. (5), Eq. (6), Eq. (7) and Eq. (8). Similarly, when λ approaches 0, we have $\lim_{\lambda \to 0} v_i^* = 0$.

As binary, linear, logarithmic and mixture scheme satisfy the three conditions, they are all self-paced functions.

The proof is then completed. ■