Communications in Information and Systems
Volume 18, Number 1, 1–35, 2018

Understanding self-paced learning under Concave Conjugacy Theory

Shi-Qi Liu, Zi-Lu Ma, De-Yu Meng†,
Kai-Dong Wang, and Yong Zhang†

By simulating the easy-to-hard learning manners of humans/animals, the learning regimes called curriculum learning (CL) and self-paced learning (SPL) have been recently investigated and invoked broad interests. However, the intrinsic mechanism for analyzing why such learning regimes can work has not been comprehensively investigated. To this issue, this paper proposes a concave conjugacy theory for looking into the insight of CL/SPL. Specifically, by using this theory, we prove the equivalence of the SPL regime and a latent concave objective, which is closely related to the known non-convex regularized penalty widely used in statistics and machine learning. Beyond the previous theory for explaining CL/SPL insights, this new theoretical framework on one hand facilitates two direct approaches for designing new SPL models for certain tasks, and on the other hand can help conduct the latent objective of self-paced curriculum learning, which is the advanced version of both CL/SPL and possess advantages of both learning regimes to a certain extent. This further facilitates a theoretical understanding for SPCL, instead of only CL/SPL as conventional. Under this theory, we attempt to attain intrinsic latent objectives of two curriculum forms, the partial order and group curriculums, which easily follow the theoretical understanding of the corresponding SPCL regimes.

1. Introduction

Since being raised recently, self-paced learning (SPL)[1] and curriculum learning (CL)[2] have been grabbing attention in machine learning and artificial intelligence. Both learning paradigms are designed by simulating the learning principle of humans/animals, attempting to start learning from

Research supported by the NSFC projects under contracts 61661166011, 11690011,61603292, 61721002, National Key R&D Program of China (2018YFB 1004300) and Macau STDF under contract 003/2016/AFJ.

†De-Yu Meng and Yong Zhang are the corresponding authors of this paper.

easier examples and gradually including more complex ones into the training process. The CL regime [2–4] was formerly designed by setting a series of learning curriculums for ranking samples from easy to hard manually, and the SPL methodology [1] has then been latterly proposed to make this easy-to-hard learning manner automatically implementable by imposing a regularization term into a general learning object, which enables the learning machine to objectively evaluate the "easiness" of a sample and automatically learn the object in an adaptive way. This learning paradigm has been empirically verified to be helpful on alleviating the local-minimum issue for a non-convex optimization problem [5], while later on more comprehensively to be verified to be capable of making the learning method more robust to heavy noises/outliers [6]. Recently, such a new learning regime has been applied to many practical problems, such as multimedia event detection [7], neural network training [8], matrix factorization [5], multi-view clustering [9], multi-task learning [10], boosting classification [11], object tracking [12], person re-identification [13], face identification [14], object segmentation [15], and some related mechanisms have been applied to weakly supervised learning [16],[17], [18]. Furthermore, an intrinsic advanced version of CL/SPL, called self-paced curriculum learning (SPCL) [19], has been designed, which tends to inherit advantages of both SPL and CL and to have a broader application [20]. Besides, many variations of SPL realization schemes have also been constructed, like self-paced reranking [7], self-paced multiple instance learning [21, 22], self-paced learning with diversity [23], multi-objective self-paced learning [24], self-paced co-training [25] and etc.

For understanding the theoretical insights of the working mechanism underlying the CL/SPL strategy, some beneficial investigations have been made. Meng et al [6] proved that the alternative search algorithm generally used to solve the self-paced learning problem is equivalent to a majorization minimization algorithm implemented on a latent SPL object function, which is closely related to the non-convex penalty used in statistics and machine learning [6]. This follows a natural explanation for the intrinsic robustness of CL/SPL. Recently, they have further proved that SPL scheme converges to a critical point of the latent objective [26]. Afterwards, Fan et al. [27] explored an implicit regularization perspective of self-paced learning, which also conducts similar robust understandings for this learning regime. Recently, Li et al. [28] proposed a general way to find the desired self-paced functions, which is beneficial for constructing more variations of SPL forms in practice.

However, these investigations explore the SPL theory mainly through exploring the equivalence of the alternative search algorithm on the SPL objectives and other algorithms implemented on some latent objective functions, while not on the SPL objective function, as well as its self-paced regularizer, itself. This makes the theory not sufficiently insightful to the problem. For example, the intrinsic relationships between self-paced regularizers and the weighting scheme to measure the importance of training samples in a SPL model is generally implicit, and hard to be intuitively explained. Besides, after adding curriculum constraint in SPL regime to form a SPCL model, current theories cannot attain the latent function like under general SPL framework. The rationality of SPCL thus still rests on the intuitive level.

To alleviate these issues, this study mainly makes the following contributions: Firstly, we establish a systematic theoretical framework under concave conjugacy theory for understanding the CL/SPL/SPCL insights. We find that the concave conjugacy theory surprisingly tallies with the requirements of the SPL model. And under this framework, the relationship among self-paced regularizer, latent SPL object function and sample weights can be clarified in a theoretically sound manner. Besides, by using this theory, the redundancy of the original SPL axiom can be removed and simplified, and the influence of the age parameter can be interpreted. Secondly, we can render a general approach for designing the SPL regime by using this theory. Furthermore, one can easily embed the required prior knowledge directly to the sample weights under this framework to make it properly used in specific applications. Thirdly, the latent objective of SPCL can be obtained under this theory. We especially discuss the form of the latent objective functions of SPCL under the partial order and group curriculums. This theory is thus meaningful for providing generalizable explanation for more general CL/SPL variations.

The paper is organized as follows. Section 2 introduces the necessary concepts and theories on concave conjugacy. Section 3 proposes the concave conjugacy theory for understanding CL/SPL. Section 4 presents two general approaches for designing a specific SPL model. Section 5 provides the theoretical understanding for SPCL under this new theory, and discusses the latent objectives of two specific curriculums.

2. Related contents on concave conjugacy

In the following we use the bolded lower letter to denote a vector, and the non-bolded lower letter to denote a scaler. For $\mathbf{v}$ and u, denote $(\mathbf{v}, u)$ as a vector in $\mathcal{R}^{n+1}$ by arranging u after the last position of $\mathbf{v}$. The inequality

$\mathbf{v} \succeq \mathbf{u}$ means that satisfies $v_i \geq u_i$ for $i = 1, \ldots, n$; $\langle \mathbf{v}, \mathbf{l} \rangle = \mathbf{v}^T \mathbf{l}$ denotes the inner products of $\mathbf{v}$ and $\mathbf{l}$. For a concave function, we assume that it takes $-\infty$ out of its domain; for a convex function, we assume that it takes $+\infty$ out of its domain. Before giving more related concepts, we first presents the following definition.

Definition 1 (Increasing Function). A multivariate function $f(\mathbf{v})$ is increasing if $f(\mathbf{v}) \geq f(\mathbf{u})$ for all $\mathbf{v} \succeq \mathbf{u}$ lying in its domain denoted by $dom\ f$.

2.1. Conjugate

We first present some necessary concepts and their related properties on the conjugate theory.

Definition 2 (Hypograph). The **hypograph** associated with the function $g : \mathcal{R}^n \longrightarrow \mathcal{R}$ is the set of points lying on or below its graph:

$$hyp\ g = \{(\mathbf{v}, u) : \mathbf{v} \in \mathcal{R}^n, u \in \mathcal{R}, u \leq g(\mathbf{v})\} \subset \mathcal{R}^{n+1}.$$

Property 1 (Hypograph Correspondence[29]). The function $g(\cdot)$ and its hypograph satisfy the following correspondence:

$$g(\mathbf{v}) = \sup_{(\mathbf{v},u) \in hyp\ g} u.$$

Property 2 (Concave function). $g(\cdot)$ is a concave function if and only if $hyp\ g$ is a convex set.

Definition 3 (Closure of Function). The closure of the function $g(\cdot)$ is a function generated by the closure of its hypograph:

$$cl\ g = \sup_{(\mathbf{v},u) \in cl\ (hyp\ g)} u.$$

It yields

$$hyp\ (cl\ g) = cl\ (hyp\ g).$$

Definition 4 (Concave Conjugate). The concave conjugate of a function $g(\cdot)$ is defined as follows:

$$g^*(\mathbf{l}) = \inf_{\mathbf{v} \in \mathcal{R}^n} \{\langle \mathbf{v}, \mathbf{l} \rangle - g(v)\}.$$

Property 3 (Relation of Concave Conjugate and Convex Conjugate [29]). For a convex function $f(\mathbf{v}) = -g(\mathbf{v})$, it holds that:

$$g^*(\mathbf{l}) = -f^*(-\mathbf{l})$$

where $f^*(\mathbf{l})$ is the convex conjugate of $f(\mathbf{v})$ defined as:

$$f^*(\mathbf{l}) = \sup_{\mathbf{v}\in\mathcal{R}^n} \{\langle \mathbf{v}, \mathbf{l}\rangle - f(\mathbf{v})\}.$$

For notation convenience, in the follows we also use conjugate to represent concave conjugate.

Definition 5 (Proper Function). A concave function $g(\cdot)$ is proper if it takes value on $[-\infty, +\infty)$ and there is at least one $\mathbf{v}$ such that $g(\mathbf{v}) > -\infty$.

Following the proof given by W.Fenchel [30] regarding the property of the conjugate convex function, one can easily prove that if $g(\mathbf{v})$ is proper, then $g^*(v)$ is a closed concave function. The concave conjugacy inherits the following duality properties of convex conjugacy as well.

Property 4 (Duality[29]). If $g(\cdot)$ is a upper semi-continuous, concave and proper function,

$$g^{**}(\mathbf{v}) = g(\mathbf{v})$$

i.e.

$$g(\mathbf{v}) = \inf_{l\in\mathcal{R}^n} \{\langle \mathbf{v}, \mathbf{l}, \rangle - g^*(\mathbf{l})\}.$$

It can be observed that the concave conjugate presents a one-to-one correspondence for all closed proper concave functions defined on $\mathcal{R}^n$.

2.2. Additive properties

The additive properties of concave conjugacy are also required to prove the related theory for SPL. We thus introduce the following necessary definitions and properties.

Definition 6 (Sup-Convolution). The sup-convolution of functions $f(\cdot)$ and $g(\cdot)$ is defined as:

$$f \oplus g(\mathbf{v}) = \sup_{\mathbf{v}_1+\mathbf{v}_2=\mathbf{v}} \{f(\mathbf{v}_1) + g(\mathbf{v}_2)\}$$

The sup-convolution has the following properties:

Property 5 (Increasing and Concave Preserving). Let $h = f \oplus g$, and then

- if $f(\cdot)$ and $g(\cdot)$ are increasing function, so is h;
- if $f(\cdot)$ and $g(\cdot)$ are concave function, so is h.

The relationship between the sup-convolution and the concave conjugate can be well illustrated by the following result.

Property 6 (Additive Property). Let $g_1(\cdot), \ldots, g_m(\cdot)$ be proper concave functions defined on $\mathcal{R}^n$. Then we have:

$$\begin{aligned}(g_1 \oplus \cdots \oplus g_m)^* &= g_1^* + \cdots + g_n^*,\\ (cl\ g_1 + \cdots + cl\ g_m)^* &= cl(g_1^* \oplus \cdots \oplus g_m^*).\end{aligned}$$

If the relative interior of $(dom\ g_i), i = 1, \ldots, m$ have a point in common, the closure operation can be omitted from the above second formula, and

$$\begin{aligned}(g_1 + \cdots + g_m)^* &= g_1^* \oplus \cdots \oplus g_m^*,\\ (g_1 + \cdots + g_m)^*(\mathbf{l}) &= \sup_{\mathbf{l}^1 + \cdots + \mathbf{l}^m = \mathbf{l}} \{g_1^*(\mathbf{l}^1) + \cdots + g_m^*(\mathbf{l}^m)\},\end{aligned}$$

where for each $\mathbf{l}$ the supremum is attained.

The proof of this property can be referred to in [29].

2.3. Differential theory

The differential theory regarding the concave conjugate plays an important role in our SPL theory. Some necessary definitions and properties are thus introduced as follows.

Definition 7 (Subgradient). A vector $\mathbf{l}$ is a subgradient of a concave function $g(\cdot)$ at $\mathbf{v}$ if

$$g(\mathbf{z}) \leq g(\mathbf{l}) + \langle \mathbf{l}, \mathbf{z} - \mathbf{l} \rangle, \forall \mathbf{z} \in \mathcal{R}^n.$$

The set of all subgradients of $g(\cdot)$ at $\mathbf{v}$ is called the subdifferential of $g(\cdot)$ at $\mathbf{v}$ and is denoted by $\partial g(\mathbf{v})$.

Correspondingly, the subgradient $\mathbf{l}$ of a convex function $f(\mathbf{v}) = -g(\mathbf{v})$ at $\mathbf{v}$ if

$$f(\mathbf{z}) \geq f(\mathbf{v}) + \langle l, \mathbf{z} - \mathbf{v} \rangle, \forall \mathbf{z} \in \mathcal{R}^n.$$

The set of all subgradients of $f(\cdot)$ at $\mathbf{v}$ is called the subdifferential of $f(\cdot)$ at $\mathbf{v}$ and is denoted by $\partial f(\mathbf{v})$.

The above subdifferentials of $f(\cdot)$ and $g(\cdot)$ have the following relation

$$\partial g(v) = -\partial f(v).$$

Property 7 (Duality of Subdifferential [29]). For any closed proper concave function $g(\cdot)$ and any vector $\mathbf{v}$, the following conditions on a vector $\mathbf{l}$ are equivalent to each other:

- $\mathbf{l} \in \partial g(\mathbf{v})$;
- $\langle \mathbf{z}, \mathbf{l} \rangle - g(\mathbf{z})$ achieves its infmum in $\mathbf{z}$ at $\mathbf{z} = \mathbf{v}$;
- $g(\mathbf{v}) + g^*(\mathbf{l}) = \langle \mathbf{z}, \mathbf{l} \rangle$;
- $\mathbf{v} \in \partial g^*(\mathbf{l})$;
- $\langle \mathbf{v}, \mathbf{z} \rangle - g(\mathbf{z})$ achieves its infmum in $\mathbf{z}$ at $\mathbf{z} = \mathbf{l}$.

Property 8 (Structure of Subdifferential [29]). Let $g(\cdot)$ be a closed proper concave function such that *dom* g has a non-empty interior. Then

$$\partial g(\mathbf{x}) = cl(convS(\mathbf{x})) + K(\mathbf{x}) \ \forall \mathbf{x} \in \mathcal{R}^n,$$

where $K(\mathbf{x}) = \{\mathbf{x}^* | \langle \mathbf{y} - \mathbf{x}, \mathbf{x}^* \rangle \geq 0 \ \forall \mathbf{y} \in dom\ g\}$ is the normal cone to *dom* g at $\mathbf{x}$ and $S(\mathbf{x})$ is the set of all limits of sequences $(\nabla g(x_1), \nabla g(x_2), \dots)$ such that $g(\cdot)$ is differentiable at $\mathbf{x}_i$ and $\mathbf{x}_i$ converges to $\mathbf{x}$.

Theorem 1 (Duality of essential strictly convex and essentially smooth[29]). *A closed proper convex function is essential strictly convex if and only if its conjugate is essential smooth.*

Corollary 2. *If $f(\cdot)$ is a closed strictly convex function with bounded domain, then $f^*(\cdot)$ is a closed differentiable function on the whole space.*

Proof. Since $f(\cdot)$ is with bounded domain, we know $f(\cdot)$ is co-finite. And then we have that $f^*(\cdot)$ is defined on whole space [29].

Furthermore, since $f(\cdot)$ is strictly convex, we can deduce that it is essential strictly convex [29]. According to theorem 1, $f^*(\cdot)$ is essential

smooth on the whole space, meaning that $f^*(\cdot)$ is differentiable on the whole space [29]. □

2.4. Indicator function

The following theory illustrates that a restriction imposed on feasible region can be viewed as the addition of an indicator function of the restricted feasible region to the objective function.

Definition 8 (Indicator Function). The indicator function of a convex set $C \subset \mathcal{R}^n$ is defined by:

$$\delta(\mathbf{v}|C) = \begin{cases} 0 & \mathbf{v} \in C, \\ -\infty & \mathbf{v} \notin C. \end{cases}$$

The closure of $\delta(\mathbf{v}|C)$ satisfies $cl\ \delta(\mathbf{v}|C) = \delta(\mathbf{v}|cl\ C)$.

Definition 9. We call the conjugate of $\delta(\mathbf{v}|C)$ the support function of C :

$$\delta^*(\mathbf{l}|C) = \inf_{\mathbf{v}\in C} \langle \mathbf{v}, \mathbf{l} \rangle.$$

Based on the above definitions of indictor function and support function, the concave conjugate with constraint can be interpreted in a new way. Specifically, suppose $g(\cdot)$ is a upper semi-continuous, proper, concave function, Ψ is a closed convex set and the relative interior of *dom* g and Ψ have at lease a point in common. Then we have

$$\begin{aligned} \inf_{\mathbf{v}\in\Psi} \{\langle \mathbf{v}, \mathbf{l} \rangle - g(\mathbf{v})\} &= \inf_{\mathbf{v}\in\mathcal{R}^n} \{\langle \mathbf{v}, \mathbf{l} \rangle - g(\mathbf{v}) - \delta(\mathbf{v}|\Psi)\} \\ &= (g(\mathbf{v}) + \delta(\mathbf{v}|\Psi))^* = g^* \oplus \delta^*(\mathbf{l}|\Psi). \end{aligned}$$

This implies that a concave conjugate with domain constraint can be understood as the addition of two conjugate terms. This will help a lot to deduce the related theory on explaining SPCL. Details will be shown in Section 4.

Theorem 3 (Monotone Conjugate). *If $g(\mathbf{v})$ is a function defined on a closed set $\Psi \subset \mathcal{R}^n_+$, then*

$$g^*(l) = \inf_{\mathbf{v}\in\Psi} \{\langle \mathbf{v}, \mathbf{l} \rangle - g(\mathbf{v})\}$$

is increasing on $\mathcal{R}^n$.

The proof of this theorem can be seen in Appendix A.

3. Concave conjugate theory for SPL

3.1. SPL Regime

We first give a short review to the generally used SPL regime.

For a given data set $D = \{\mathbf{z}_i\}_{i=1}^n$, where $z_i = (\mathbf{x}_i, y_i)$ is a training sample with a datum and its corresponding label, SPL uses the following model for learning [5, 7]:

$$\inf_{f\in\mathcal{F},\mathbf{v}\in[0,1]^n} E(f,\mathbf{v};\lambda) = \inf_{f_w\in\mathcal{F},\mathbf{v}\in[0,1]^n} \sum_{i=1}^{n} v_i L(f,\mathbf{z}_i) + R_{SP}(\mathbf{v},\lambda) + R_{\mathcal{F}}(f), \tag{1}$$

where $\mathbf{v} = (v_1, v_2, \ldots, v_n) = (Weight(\mathbf{z}_1), \ldots, Weight(\mathbf{z}_n))$ represent the vector of weights imposed on all training samples, $R_{SP}(\mathbf{v},\lambda)$ is called self-paced regularizer which encodes the learning procedures following the principle from easy to hard, $R_{\mathcal{F}}(f)$ is the general regularizer for the model parameters to alleviate the overfitting problem, and λ is a parameter that controls the learning pace and guarantees the easy-to-complex learning procedure. By gradually increasing the age parameter, more samples can be automatically included with higher weights into training in a purely self-paced way. f is the decision function for the task, like a classifier or a regressor, $L(\cdot,\cdot)$ is the loss function (the function f is generally parameterized by parameters $\mathbf{w}$ and L is then the function with respect to $\mathbf{w}$ and $\mathbf{z}$). Let $\mathbf{l}$ denote the loss vector $(L(f,\mathbf{z}_1),\ldots,L(f,\mathbf{z}_n))^T$. This leads to a brief expression for the model:

$$\inf_{\mathbf{w}\in\mathcal{W},\mathbf{v}\in[0,1]^n} \langle \mathbf{v}, \mathbf{l}\rangle + R_{SP}(\mathbf{v},\lambda) + R_{\mathcal{W}}(\mathbf{w}).$$

A common way to solve the SPL model is to alternatively optimize the target function f and the weight vector $\mathbf{v}$ as follows:

- Optimize f:

$$f^k = \inf_{f\in\mathcal{F}} \langle \mathbf{v}^{k-1}, \mathbf{l}(f)\rangle + R_{\mathcal{F}}(f). \tag{2}$$

- Optimize $\mathbf{v}$:

$$\mathbf{v}^k = \inf_{\mathbf{v}\in[0,1]^n} \langle \mathbf{v}, \mathbf{l}(f^k)\rangle + R_{SP}(\mathbf{v},\lambda). \tag{3}$$

The SP-regularizer should satisfy necessary conditions to guarantee an expected easy-to-hard learning manner [5, 7]:

Definition 10 (SP-regularizer). $R_{SP}(\mathbf{v}, \lambda)$ is called a SP-regularizer, if

- $R_{SP}(\mathbf{v}, \lambda)$ is convex with respect to $\mathbf{v} \in [0, 1]^n$;
- $v_i(\lambda, l_i)$ decrease with respect to l, and it holds that $\forall i \in \{1, 2, \ldots, n\}$, $v_i(\lambda, l_i) \leq 1$ and $\lim_{l_i \to +\infty} v_i(\lambda, l_i) = 0$;
- $v_i(\lambda, l_i)$ increase with respect to λ, and it holds that $\forall i \in \{1, 2, \ldots, n\}$, $v_i(\lambda, l_i) \leq 1$ and $\lim_{\lambda \to 0} v_i(\lambda, l_i) = 0 (l_i > 0)$,

where $\mathbf{v}(\lambda, \mathbf{l}) = \arg \inf_{\mathbf{v} \in [0,1]^n} \{\langle \mathbf{v}, \mathbf{l} \rangle + R_{SP}(\mathbf{v}, \lambda)\}$.

By using such defined SP-regularizer, SPL can conduct the learning manner that imposes larger weights on easier samples while smaller on harder ones, and gradually increases the sample weights with the age parameter increasing.

3.2. Conjugate theory of SP-regularizer

We can prove the following conjugate result on a SP-regularizer $R_{SP}(\mathbf{v}, \lambda)$ as follows:

Theorem 4 (Conjugate Equivalence). *For arbitrary function $R_{SP}(\mathbf{v})$ satisfying $dom_{\mathbf{v}}\ R_{SP}(\mathbf{v}) \subset [0, 1]^n$, let $g(\mathbf{v}) = -R_{SP}(\mathbf{v})$, and then*

$$\begin{aligned} \inf_{\mathbf{v} \in [0,1]^n} \{\langle \mathbf{v}, \mathbf{l} \rangle + R_{SP}(\mathbf{v})\} &= \inf_{\mathbf{v} \in [0,1]^n} \{\langle \mathbf{v}, \mathbf{l} \rangle - g^{**}(\mathbf{v})\} \\ &= \inf_{\mathbf{v} \in [0,1]^n} \{\langle \mathbf{v}, \mathbf{l} \rangle + \overline{R_{SP}}(\mathbf{v})\}, \end{aligned}$$

where $\overline{R_{SP}}(\mathbf{v}) = -g^{**}(\mathbf{v})$.

The proof is provided in Appendix B.

From the above theorem, it can be found that there are redundancy in the definition of SP-regularizer, which can be simplified as follows:

Theorem 5 (SP-regularizer Simplification). *If $R_{SP}(\mathbf{v}, \lambda)$ satisfies*

- *$R_{SP}(\mathbf{v}, \lambda)$ is strictly convex in $\mathbf{v}$;*
- *$R_{SP}(\mathbf{v}, \lambda)$ is lower semi-continuous in $\mathbf{v}$;*

- *$dom_{\mathbf{v}}\ R_{SP}(\mathbf{v},\lambda) \subset [0,1]^n$ and $\boldsymbol{0}, \boldsymbol{1} \in cl(dom_{\mathbf{v}}\ R_{SP}(\mathbf{v},\lambda))$,*

then it holds that $\forall i \in \{1,2,\ldots,n\}$:

- *$v_i(\lambda, l_i)$ decrease with respect to l_i; $v_i(\lambda, l_i) \leq 1$; $\lim\limits_{l_i \to +\infty} v_i(\lambda, l_i) = 0$.*
- *If $R_{SP}(\mathbf{v},\lambda) = \lambda R_{SP}(\mathbf{v})$ where $R_{SP}(\mathbf{v})$ satisfy the above condition in $\mathbf{v}$, then $\forall i \in \{1,2,\ldots,n\}$, $v_i(\lambda, l_i)$ increases with respect to λ, $v_i(\lambda, l_i) \leq 1$, $\lim\limits_{\lambda \to 0} v_i(\lambda, l_i) = 0$ $(l_i > 0)$,*

where $\mathbf{v}(\lambda, \mathbf{l}) = \arg \inf\limits_{\mathbf{v} \in [0,1]^n} \{\langle \mathbf{v}, \mathbf{l} \rangle + R_{SP}(\mathbf{v}, \lambda)\}$.

The proof is presented in Appendix C.

This theorem shows that the conditions in $\mathbf{l}$ can be implied by the conditions being directly imposed on the SP-regularizer. According to simplification theorem, determining one easily handled representative in the equivalence class, the following assumption gives weaker conditions for a SP-regularizer.

Definition 11 (SP-regularizer simplification). $R_{SP}(\mathbf{v},\lambda)$ is called a self-paced regularizer with simplified conditions if:

1) $R_{SP}(\mathbf{v},\lambda)$ is convex in $\mathbf{v}$;
2) $R_{SP}(\mathbf{v},\lambda)$ is lower semi-continuous in $\mathbf{v}$;
3) $dom_{\mathbf{v}}\ R_{SP}(\mathbf{v},\lambda) \subset [0,1]^n$ and $\mathbf{0}, \mathbf{1} \in cl(dom_{\mathbf{v}}\ R_{SP}(\mathbf{v},\lambda))$.

3.3. Model Equivalence

Based on the concave conjugacy of SPL, its equivalent model can be derived as follows. For convenience, let $g_\lambda(\mathbf{v}) = -R_{SP}(\mathbf{v},\lambda)$, and then it holds that:

$$\begin{aligned} & \inf_{f \in \mathcal{F}, \mathbf{v} \in [0,1]^n} E(f, \mathbf{v}; \lambda) \\ \Longleftrightarrow \quad & \inf_{f \in \mathcal{F}} R_{\mathcal{F}}(f) + \inf_{\mathbf{v} \in [0,1]^n} \sum_{i=1}^{n} v_i L(f, z_i) + R_{SP}(\mathbf{v}, \lambda) \\ \Longleftrightarrow \quad & \inf_{f \in \mathcal{F}} g_\lambda^*(\mathbf{l}(f)) + R_{\mathcal{F}}(f) \Longleftrightarrow \inf_{f \in \mathcal{F}} F_\lambda(\mathbf{l}(f)) + R_{\mathcal{F}}(f) \end{aligned}$$

where $F_\lambda(\mathbf{l}) = g_\lambda^*(\mathbf{l})$. According to the property of the concave conjugate, $F_\lambda(\mathbf{l})$ is a proper closed concave function. Through this analysis, we can try to get more insights of SPL.

3.3.1. Latent SPL objective. Mostly, we can separate a SPL optimization model to multiple 1 dimension sub-problems:

$$\inf_{f\in\mathcal{F},\mathbf{v}\in[0,1]^n} E(f,\mathbf{v};\lambda) = \inf_{f\in\mathcal{F},\mathbf{v}\in[0,1]^n} \{\sum_{i=1}^{n}(v_i l_i + R_{SPi}(v_i,\lambda)) + R_{\mathcal{F}}(f)\}.$$

Then, the optimization on $\mathbf{v}$ can be reformulated as solving the following multiple subproblems on each of its component v_i:

$$\inf_{v\in[0,1]} E(w,v;\lambda) = \inf_{v\in[0,1]} \{vl + R_{SP}(v,\lambda)\}.$$

We denote

$$v(\lambda,l) = \arg\inf_{v\in[0,1]} \{\langle v,l\rangle + f(v,\lambda)\}$$

In [6], it is proved that the alternative search algorithm on the SPL objective is equivalent to the MM algorithm implemented on a latent objective

$$\int_0^l v(\lambda,j)\,dj$$

on l. We can get the similar result under concave conjugate theory as follows.

Theorem 6 (Model Equivalence). *If $R_{SP}(v,\lambda)$ satisfy the simplified conditions of SPL as defined in 11 and be strictly convex, then the latent SPL objective can be written as:*

$$F_\lambda(l) = \int_0^l v(\lambda,j)\,dj + C(\lambda),$$

where $C(\lambda)$ is a function in λ.

The proof is listed in Appendix D.

3.4. Relations

In the following theorem, we want to make the relations among the SP-regularizer $R_{SP}(v,\lambda)$, latent objective $F_\lambda(l)$, and the weight function $v(\lambda,l)$ clear.

Theorem 7. *If $R_{SP}(v,\lambda)$ satisfy the simplified conditions of SPL, then we have:*

$$\begin{aligned} l_\lambda(v) &= \partial_v(-R_{SP}(v,\lambda)),\\ v(\lambda,l) &= l_\lambda^{-1}(l),\\ v(\lambda,l) &= \partial F_\lambda(l),\\ F_\lambda(l) &= \langle v(\lambda,l),l\rangle + R_{SP}(v(\lambda,l),\lambda),\\ R_{SP}(v,\lambda) &= \langle v, l_\lambda(v)\rangle - R_{SP}(v,\lambda)(l_\lambda(v)). \end{aligned}$$

Furthermore, if $R_{SP}(v,\lambda)$ and $F_\lambda(l)$ is strictly convex in v and l, respectively and we can further obtain that

$$\begin{aligned} F_\lambda(l) &= \int_0^l v(\lambda,j)\,dj + C(\lambda),\\ R_{SP}(v,\lambda) &= -\int_0^v l_\lambda(j)\,dj + C(\lambda). \end{aligned}$$

The theorem is directly got from the Duality of Subdifferential Properties 7 and the latter two inequalities can be obtained based on Theorem 6.

According to Theorem 7, one can easily derive the weight function from the SP-regularizer through the differential and inverse step, which is empirically more convenient than through the arg-minimization analysis. We then discuss on how to specify the age parameter in the model.

3.5. On age parameter

An easy way to construct a SP-regularizer is first to generate a regularizer, denoted by $R_{SP}(\mathbf{v})$, satisfying the simplified conditions of SPL, and then use the SP-regularizer as $\lambda R_{SP}(\mathbf{v})$. The reason why it works can be interpreted as follows:

Let $g(\mathbf{v}) = -R_{SP}(\mathbf{v})$ and let the concave conjugate of $g^*(\mathbf{l}) = F(\mathbf{l})$. Then we have:

$$\begin{aligned} F_\lambda(\mathbf{l}) = (\lambda g(\mathbf{v}))^* &= \inf_{\mathbf{v}\in[0,1]^n}\{\langle \mathbf{v},\mathbf{l}\rangle - \lambda g(\mathbf{v})\}\\ &= \lambda \inf_{\mathbf{v}\in[0,1]^n}\{< \mathbf{v}, \lambda^{-1}\mathbf{l} > -g(\mathbf{v})\} = \lambda F(\lambda^{-1}\mathbf{l}). \end{aligned}$$

For simplicity, we assume $g(\mathbf{v})$ is strictly concave. As a result, $F(\mathbf{l})$ is differentiable and the original $\mathbf{v}(\mathbf{l}) = \nabla F(\mathbf{l})$, and then we have:

$$\mathbf{v}(\lambda,\mathbf{l}) = \nabla_{\mathbf{l}} F_\lambda(\mathbf{l}) = \lambda \nabla_{\mathbf{l}} F(\lambda^{-1}\mathbf{l}) = \mathbf{v}(\lambda^{-1}\mathbf{l}).$$

Thus, $v_i(\lambda, l_i)$ increase with respect to λ, and it holds that $\forall i \in \{1, 2, \ldots, n\}$, $\lim_{\lambda \to 0} v_i(\lambda, l_i) = \lim_{\lambda \to 0} v_i(\lambda^{-1} l_i) = 0$ and $\lim_{\lambda \to +\infty} v_i(\lambda, l_i) = \lim_{\lambda \to +\infty} v_i(\lambda^{-1} l_i) \leq 1$.

Besides, since $v(\lambda, l) = v(\lambda^{-1} l)$, the change of the λ stretches the shape of $v(\lambda, l)$. In particular, if the $v(l)$ is with threshold, then $v(\lambda, l)$ shifts the threshold through λ which reflexes the change of decision boundary regarding learning or not.

Then we give a discussion on how to specify a proper age parameter in the learning process.

Generally the SP-regularizer has the data screening properties, that is, there exists some λ^* such that $v(l \geq \lambda^*) = 0$. One can use two ways for specifying the age parameter. The first is suggested by [1]: first to choose a λ such that around half of example are used with positive weight, and then gradually increase the λ to include more samples into training. Another strategy is suggested in [7]: first calculate the loss of each example, and choose a age parameter such that a portion of samples with smaller loss is with positive weights and the other with zero weights; and then increase the portion number to implicitly increase the age parameter. Also some other variations [8] have also been discussed and can be considered in application.

4. Two methods for designing a SPL regime

By utilizing the aforementioned theoretical results, we can construct two methods for designing a general SPL regime in practice.

We call the first method as the vFlRλ method. The progress for one dimension sub-problem is provided as follows:

1) Design $v(l)$ satisfying $v(l)$ decrease with respect to l and

$$\lim_{l \to 0} v(l) = 1 \quad \lim_{l \to +\infty} v(l) = 0;$$

2) $F(l) = \int_0^l v(j)\, dj$;

3) $l(v) = v^{-1}(v)$;

4) $R_{SP}(v) = -\langle v, l(v) \rangle + F(l(v))$;

5) $R_{SP}(v, \lambda) = \lambda R_{SP}(v)$; $F_\lambda(l) = \lambda F(\lambda^{-1} l)$; $v(\lambda, l) = v(\lambda^{-1} l)$.

If $F(l)$ is given then $v(l) = \partial F(l)$ and the other steps are the same.

We can then provide an example for designing SPL by using this method.

1) $v(l) = (1 - l)_{[0,1]}$;

2) $F(l) = \int_0^l v(j)\,dj = \min(l - \frac{l^2}{2}, \frac{1}{2})$;

3) $l(v) = v^{-1}(v) = \begin{cases} 1 - v & v \in (0, 1], \\ [1, +\infty) & v = 0; \end{cases}$

4) $R_{SP}(v) = -\langle v, l(v)\rangle + F(l(v))$, whose component is computed by $\frac{(1-v)^2}{2}$;

5) $R_{SP}(v, \lambda) = \lambda R_{SP}(v) = \frac{\lambda(1-v)^2}{2}$; $F_\lambda(l) = \lambda F(\lambda^{-1} l) = \min(l - \frac{l^2}{2\lambda}, \frac{\lambda}{2})$; $v(\lambda, l) = v(\lambda^{-1} l) = (1 - \frac{l}{\lambda})_{(0,\lambda)}$.

In this example, linear SP-regularizer[7] is derivated from the weight function that linearly weights the sample whose loss is between 0 and λ.

The second method is called the flvFλ method. Its main process for one dimension sub-problem includes the following steps:

1) $R_{SP}(v)$ satisfy:
   - $dom\ R_{SP}(v) \subset [0, 1]^n$;
   - $0, 1 \in cl(dom\ R_{SP}(v))$;
   - $R_{SP}(v)$ is convex and continuous;

2) $l(v) = \partial(-R_{SP}(v))$;

3) $v(l) = l^{-1}(v)$;

4) $F(l) = \langle v(l), l\rangle + R_{SP}(v(l))$;

5) $R_{SP}(v, \lambda) = \lambda R_{SP}(v)$; $F_\lambda(l) = \lambda F(\lambda^{-1} l)$; $v(\lambda, l) = v(\lambda^{-1} l)$.

We also present an example for using this method to design SPL.

1) $R_{SP}(v) = -\log v\ v \in (0, 1]$;

2) $l(v) = \partial(-R_{SP}(v)) = \begin{cases} v^{-1} & v \in (0, 1), \\ (-\infty, 1] & v = 1; \end{cases}$

3) $v(l) = \min(1, l^{-1})$;

4) $F(l) = \langle v(l), l\rangle + R_{SP}(v(l)) = \begin{cases} 1 + \log l, & l \in (1, +\infty), \\ l, & l \in (-\infty, 1]; \end{cases}$

5)
   - $R_{SP}(v, \lambda) = \lambda R_{SP}(v) = -\lambda \log v\ v \in (0, 1]$;
   - $F_\lambda(l) = \lambda F(\lambda^{-1} l) = \begin{cases} \lambda + \log l - \log \lambda & l \in (\lambda, +\infty), \\ l & l \in (-\infty, \lambda]; \end{cases}$
   - $v(\lambda, l) = v(\lambda^{-1} l) = \min(1, \lambda l^{-1})$.

In this example, the weight function, which weights the sample by the minimal of 1 and λ times its loss reciprocal, is derived from the LOG-like SP-regularizer.

5. Concave conjugate theory for SPCL

In the conventional SPCL strategies, a curriculum region needs to be specified and added into a general SPL optimization as a constraint [19]. In this way, however, the latent objective of SPL as deduced in the previous sections is changed and cannot be obtained by the previous theory. We thus attempt to discuss this point, and provide explicit latent objective functions underlying SPCL for two specific curriculums. For notation convenience, in the following we omit λ in SPL functions.

5.1. Latent objective of SPCL

In the following theorem we propose the form of the latent objective underlying SPCL.

Theorem 8. *Suppose the self-paced regularizer is $R_{SP}(\mathbf{v})$ satisfying the simplified conditions of SPL. Let $F(\mathbf{l}) = \inf_{\mathbf{v}\in\mathcal{R}^n}\{\langle \mathbf{v}, \mathbf{l}\rangle + R_{SP}(\mathbf{v})\}$ denote the concave conjugate of $-R_{SP}(\mathbf{v})$ in $\mathbf{v}$. Ψ is closed convex set and $ri([0,1]^n \cap \Psi) \neq \emptyset$ and $\delta(\mathbf{v}|\Psi)$ is the indicator function. Then*

$$F^{new}(\mathbf{l}) \triangleq \inf_{\mathbf{v}\in\Psi}\{\langle \mathbf{v}, \mathbf{l}\rangle + R_{SP}(\mathbf{v})\} = F \oplus \delta^*(\cdot|\Psi)(\mathbf{l}),$$

and

$$\inf_{f\in\mathcal{F},\mathbf{v}\in[0,1]^n\cap\Psi} E(f,\mathbf{v}) = \inf_{f\in\mathcal{F}}\{R_{\mathcal{F}}(f) + F^{new}(\mathbf{l}(f))\}.$$

Proof.

$$\begin{aligned}\inf_{\mathbf{v}\in\Psi}\{\langle \mathbf{v}, \mathbf{l}\rangle + R_{SP}(\mathbf{v})\} &= \inf_{\mathbf{v}\in\mathcal{R}^n}\{\langle \mathbf{v}, \mathbf{l}\rangle + R_{SP}(\mathbf{v}) - \delta(\mathbf{v}|\Psi)\} \\ &= (-R_{SP}(\mathbf{v}) + \delta(\mathbf{v}|\Psi))^* = F \oplus \delta^*(\cdot|\Psi)(\mathbf{l}).\end{aligned}$$

□

From the theorem, we know that the latent objective of SPCL under certain curriculum region Ψ is the sup convolution of the original SPL latent objective without this constraint and the support function on it. There are several properties on this new objective $F^{new}(\mathbf{l})$.

Property 9. If the conditions of the theorem 8 hold, then $F^{new}(\mathbf{l})$ has the following properties.

- It is upper semi-continuous and concave since it is the concave conjugate.
- It is increasing according to the Theorem 3.
- $F^{new}(\mathbf{l}) \geq \max(F(\mathbf{l}), \delta^*(\mathbf{l}|\Psi))$ due to the property of sup convolution and the fact that $\delta^*(\mathbf{0}|\Psi)) \geq 0$.

Moreover, if $R_{SP}(\mathbf{v})$ is strictly convex, it yields that

- According to Corollary 2, $F^{new}(\mathbf{l})$ is differentiable.

5.2. Curriculum function

Through the above discussion, we may find that the curriculum region can be interpreted as a special family of curriculum function.

Suppose we provide the SPL model by adding a curriculum function $R_{CL}(\mathbf{v})$ which is a closed convex function and satisfies $ri(dom\ R_{CL}) \cap ri(dom\ R_{SP}) \neq \emptyset$. Then the new latent objective function can be obtained by the following:

$$F^{new}(\mathbf{l}) = \inf_{\mathbf{v}\in[0,1]^n} \{\langle \mathbf{v}, \mathbf{l}\rangle + R_{SP}(\mathbf{v}) + R_{CL}(\mathbf{v})\} = F \oplus (-R_{CL})^*(\mathbf{l}).$$

It can be seen that the curriculum properties depends on the conjugate of the curriculum function and the sup convolution step.

Suppose we have K curriculum functions which are proper closed convex functions, and let R_{CL0} denote R_{SP}. If they satisfy $\cap_{i=0}^{K} ri(dom\ R_{CLi}) \neq \emptyset$, then according to Property 6 the objective function of SPCL is

$$F^{new}(\mathbf{l}) = \oplus_{i=0}^{K} (-R_{CLi})^*(\mathbf{l}).$$

By introducing a new curriculum function R_{CL} into the model, new latent objective is obtained by sup convolution of original object function and conjugate of the curriculum function. The result can be viewed as the action of the new curriculum on the original latentive objective. We call this action **Curriculum Action** in the follows for convenience.

5.3. Basic curriculum region

Consider the following case that the feasible region of $\mathbf{v}$ is $\mathcal{R}^n$ and the SP-regularizer is 0, and then

$$\inf_{\mathbf{v}\in\mathcal{R}^n} \langle \mathbf{v}, \mathbf{l}\rangle = \delta(\mathbf{l}|\mathbf{0}),$$

which means that it takes finite value 0 when the component of $\mathbf{l}$ equals 0 and it takes $-\infty$ on $\mathcal{R}^n \setminus 0$.

For all proper concave function $f(\mathbf{l})$, it holds that

$$f(\mathbf{l}) \oplus \delta(\mathbf{l}|\mathbf{0}) = f(\mathbf{l}).$$

We can then give the following definition related to curriculums:

Definition 12 (Basic Curriculum Region). For the SPL model

$$\inf_{\mathbf{v}\in\mathcal{R}^n} \langle \mathbf{v}, \mathbf{l}\rangle + R_{SP}(\mathbf{v}),$$

we call the *dom* $R_{SP}(\cdot)$ the basic curriculum region.

The commonly discussed SP-regularizers are defined on $[0,1]^n$. Suppose the regularizer $g(v) = -R_{SP}(v)$ is a concave function being differentiable on $[0,1]^n$, and it can be extended to an open set which contains $[0,1]^n$. According to Property 8 the structure of subdifferential, we can obtain

$$\partial g(\mathbf{v}) = \begin{cases} \nabla g(\mathbf{1}) + \mathcal{R}^n_- & \mathbf{v} = \mathbf{1} \\ \nabla g(a^i) + < b^i_1 e^1, \ldots, b^i_n e^n > & \mathbf{v} = a^i \in V([0,1]^n) \\ \nabla g(\mathbf{0}) + \mathcal{R}^n_+ & \mathbf{v} = \mathbf{0} \\ \nabla g(\mathbf{v}) + K(\mathbf{v}) & \mathbf{v} \in \partial[0,1]^n / \{V([0,1])^n\} \\ \nabla g(\mathbf{v}) & \mathbf{v} \in (0,1)^n \end{cases}$$

where a^i is the vertex of the hypercube $[0,1]^n$, $b^i_j = \begin{cases} 1 & a^i_j = 0 \\ -1 & a^i_j = 1 \end{cases}$, $\langle b^i_1 e^1, \ldots, b^i_n e^n\rangle$ represents the cone generated by $b^i_1 e^1, \ldots, b^i_n e^n$ with positive coefficients and $V([0,1]^n)$ represents all the vertices of $[0,1]^n$.

By calculating the inverse of set-valued function $\partial g(\mathbf{v})$, the weight set-valued function $\mathbf{v}(\mathbf{l})$ can be obtained.

5.3.1. Linear Regularizer.

Definition 13 (Linear Regularizer). We call

$$R_{SP}(\mathbf{v}) = -\lambda^T \mathbf{v}$$

linear regularizer for the SPL model

Once we select the linear regularizer, we can obtain:

$$-R_{SP}(\mathbf{v}) = \lambda^T \mathbf{v}$$

$$\partial(-R_{SP})(\mathbf{v}) = \begin{cases} \lambda + \mathcal{R}_-^n & \mathbf{v} = \mathbf{1}, \\ \lambda + < b_1^i e^1, \ldots, b_n^i e^n > & \mathbf{v} = a^i \in V([0,1]^n)), \\ \lambda + \mathcal{R}_+^n & \mathbf{v} = \mathbf{0}, \\ \lambda + K(\mathbf{v}) & \mathbf{v} \in \partial[0,1]^n / \{V([0,1])^n\}, \\ \lambda & \mathbf{v} \in (0,1)^n. \end{cases}$$

According to the Property 7, we can obtain that

$$\partial F(\mathbf{l}) = \begin{cases} \mathbf{v} = \mathbf{1} & \mathbf{l} \in \lambda + (\mathcal{R}_-^n)^\circ \\ \mathbf{v} = a^i & \mathbf{l} \in \lambda + < b_1^i e^1, \ldots, b_n^i e^n >^\circ \quad \mathbf{v} = a^i \in V([0,1]^n)) \\ \mathbf{v} = \mathbf{0} & \mathbf{l} \in \lambda + (\mathcal{R}_+^n)^\circ \\ \cdots & \end{cases}$$

Hence, the domain of $\partial F(\mathbf{l}) = v(\mathbf{l})$ can be separated into 2^n part, each taking the same value corresponding to the vertex of the hypercube $[0,1]^n$.

5.4. Linear homogeneous curriculum

One of the most commonly used curriculum is the partial order curriculum. For instance, if one has the prior knowledge that example 1 is more important or reliable than example 2, it's reasonable to restrict their feasible region such that $v_1 \geq v_2$. In regard to $v_1 - v_2 \geq 0$, we call it linear homogeneous curriculum. Generally, those knowledge come as a series of linear inequalities and we call them partial order curriculum. For simplicity, in the following we consider the simple linear homogeneous curriculum and, for more curriculums, we can treat them one by one.

In order to avoid the disfunctional curriculum and to make analysis convenient, we render the following nonsingular assumption for the curriculum region.

Assumption 1 (Assumption for Curriculum Region). A curriculum region Ψ satisfies the following conditions:

- $int(\Psi) \cap int(dom\ R_{SP}) \neq \emptyset$,
- $\Psi \cap dom\ R_{SP} \neq dom\ R_{SP}$.

Definition 14 (Linear Homogeneous Curriculum). If $\Psi = \{\mathbf{v}|\mathbf{v}^T\mathbf{k} \geq 0\}$, we call Ψ a linear homogeneous curriculum and $\mathbf{k}$ the linear homogeneous curriculum direction.

We can then prove the following result:

Theorem 9. *Suppose $R_{SP}(v)$ satisfies Definition 11 and the curriculum as $\mathbf{v}^T\mathbf{k} \geq 0$ corresponding to $\Psi = \{\mathbf{v}|\mathbf{v}^T\mathbf{k} \geq 0\}$. If Ψ satisfies Assumption 1, then we have:*

$$F^{new}(\mathbf{l}) = F \oplus \delta(\cdot|\Psi^\circ)(\mathbf{l}) = \sup_{\mathbf{l}^1+\mathbf{l}^2=\mathbf{l}} \{F(\mathbf{l}^1)) + \delta(\mathbf{l}^2|\Psi^\circ)\} = \sup_{\mathbf{l}^1 \in \mathbf{l}-ray_{\mathbf{k}}} F(\mathbf{l}^1\rangle \geq 0\}$$

is another non-empty closed convex cone and $ray_{\mathbf{k}}$ denotes the ray starting from the origin in direction $\mathbf{k}$.

Proof in Appendix E.

Theorem 9 illustrates that the latent objective of SPCL is the supremum of the original objective function of SPL without curriculum constraint on the ray which starts from $\mathbf{l}$ to the direction $-\mathbf{k}$. We then give the theorem on the action of linear homogeneous curriculum.

Theorem 10 (Action of Linear Homogeneous Curriculum). *Suppose $R_{SP}(\mathbf{v})$ is essential strictly convex and satisfies Definition 11. Suppose we have the curriculum constraint $\mathbf{v}^T\mathbf{k} \geq 0$, corresponding to the curriculum region $\Psi = \{\mathbf{v}|\mathbf{v}^T\mathbf{k} \geq 0\}$. Then if Ψ satisfies assumption 1, it holds that $\nabla F^{new}(\mathbf{l})^T\mathbf{k} = v^{new}(\mathbf{l})^T\mathbf{k} \geq 0$ and*

$$F^{new}(\mathbf{l}) = \begin{cases} F(\mathbf{l}) & \mathbf{l} \in \partial(-R_{SP})(\mathbf{k}^\perp) - ray_{\mathbf{k}}, \\ \sup\limits_{\mathbf{l}' \in \mathbf{l}+line_{\mathbf{k}}} F(\mathbf{l}')(\geq F(\mathbf{l})) & \mathbf{l} \in \partial(-R_{SP})(\mathbf{k}^\perp) + ray_{\mathbf{k}}. \end{cases}$$

The proof is presented in Appendix F.

Theorem 10 illustrates the form of the latent objective of SPCL following the restriction imposed on the curriculum region. This naturally leads to the following concept of the critical region: We call $(-R_{SP})(k^\perp)$ the critical region for the new latent objective of SPCL.

According to Theorem 10, the most important thing for determining $F^{new}(l)$ is to determine the critical region $(-R_{SP})(k^\perp)$, since the critical region divides the $\mathcal{R}^n$ into two parts.

On one part, the linear homogeneous curriculum has no effects. On the other part, the more increase in the curriculum direction, the bigger penalization of the new latent function on the loss.

5.4.1. Partial Order Curriculum. We can then evaluate the insights of SPCL with partial order curriculum, where $v_1 \geq v_2$ encoding the prior knowledge that example 1 is more important and reliable than example 2. The SP-regularizer is chosen to be the exponential $R_{SP}(v) = v \log v - v + 1$. We can then deduce the following results by using the aforementioned theoretical results:

- The original SPL latent objective (without this curriculum) is $F(l) = 1 - e^{-l_1} + 1 - e^{-l_2}$;
- $l(v) = \partial(-R_{SP}(v)) = -(\log v_1, \log v_2)^T$;
- Curriculum direction is $k = (1, -1)^T$;
- The critical region is $line_{(1,1)^T}$;
- The new latent objective is

$$F^{new}(l) = \begin{cases} 1 - e^{-l_1} + 1 - e^{-l_2} & l_1 \leq l_2, \\ 2(1 - e^{-\frac{l_1+l_2}{2}}) & l_1 \geq l_2; \end{cases}$$

- The new weighting function

$$v^{new}(l) = \begin{cases} (e^{-l_1}, e^{-l_2})^T & l_1 \leq l_2, \\ (e^{-\frac{l_1+l_2}{2}}, e^{-\frac{l_1+l_2}{2}})^T & l_1 \geq l_2. \end{cases}$$

When hard SP-regularizer is chosen, the partial order curriculum can determine the learning order of each example [19].

5.4.2. Linear Curriculum. It's natural to extend the above discussion to the more general linear curriculum case, as defined in the following.

Definition 15 (Linear Curriculum). If $\Psi = \{\mathbf{v}|\mathbf{v}^T\mathbf{k} \geq b\}$, we call Ψ the linear curriculum and we call $\mathbf{k}$ the linear curriculum direction.

Then we can prove the following result:

Theorem 11. *Suppose $R_{SP}(\mathbf{v})$ is essential strictly convex and satisfies Definition 11. Suppose the curriculum is $\mathbf{v}^T\mathbf{k} \geq b$ corresponding to the curriculum region $\Psi = \{\mathbf{v}|\mathbf{v}^T\mathbf{k} \geq b\}$. If Ψ satisfies Assumption 1, then*

$$F^{new}(\mathbf{l}) = \begin{cases} F(\mathbf{l}) & \mathbf{l} \in \partial g(\mathbf{k}^{\perp} + \frac{b\mathbf{k}}{\|\mathbf{k}\|_2}) - ray_{\mathbf{k}}, \\ F(\mathbf{l} - \beta^0(\mathbf{l})\mathbf{k}) + \beta^0 b (\geq F(\mathbf{l})) & \mathbf{l} \in \partial g(\mathbf{k}^{\perp} + \frac{b\mathbf{k}}{\|\mathbf{k}\|_2}) + ray_{\mathbf{k}}, \end{cases}$$

where $\beta^0(\mathbf{l}) = \max \arg_{\beta}\{\nabla F(\mathbf{l} - \beta\mathbf{k})^T\mathbf{k} = 0\}$.

Proof in Appendix G.

Theorem 11 helps to obtain the latent objective under SPCL with linear curriculum, which shares the similar structure comparing the one in Theorem 10. The critical region here becomes $\partial g(k^{\perp} + \frac{bk}{\|k\|_2})$. It is easy to see that the linear curriculum punishes the loss on one side of the critical region and keeps the other side the same.

5.5. Group Curriculum

When we have the prior knowledge that some training samples are coming from the similar group with similar importance weights, we can group all samples into multiple categories and make each group share similar weights. In regard to weight sharing, we call it the group curriculum.

Suppose the original R_{SP} regularizer of each example is the same and SPL model can be separated into 1 dimesion sub-problem.

$$\begin{aligned} \inf_{\mathbf{v}\in[0,1]^n} E(f, \mathbf{v}; \lambda) = & \inf_{v_1\in[0,1]} \{v_1 l_1 + R_{SP}(v_1, \lambda)\} \\ & + \cdots + \inf_{v_n\in[0,1]} \{v_n l_n + R_{SP}(\mathbf{v}_n, \lambda)\} + R_{\mathcal{F}}(f). \end{aligned}$$

Suppose $\{i_1, \ldots, i_{s_i}\}$ are members of the group i and there are k groups. By adding the group curriculum, we obtain the new object function as:

$$\begin{aligned} & \inf_{v_1\in[0,1]} \left\{ s_1 R_{SP}(v_1, \lambda) + v_1 \sum_{j=1}^{s_1} l_{1j} \right\} \\ & \quad + \cdots + \inf_{v_k\in[0,1]} \left\{ s_k R_{SP}(v_k, \lambda) + v_k \sum_{j=1}^{s_k} l_{kj} \right\} + R_{\mathcal{F}}(f). \end{aligned}$$

Suppose the latent objective function in SPL without such curriculum constraint for one example is $\bar{F}_\lambda$, and then the new latent objective is

$$F^{new}(\mathbf{l}) = (s_1 \star \bar{F}_\lambda)\left(\sum_{j=1}^{s_1} l_{1j}\right) + \cdots + (s_k \star \bar{F}_\lambda)\left(\sum_{j=1}^{s_k} l_{kj}\right),$$

where $s_1 \star \bar{F}_\lambda(x) \triangleq s_1 \bar{F}_\lambda(s_1^{-1}x)$.

It can be seen that the group curriculum corresponds to a special curriculum region that weights in the same group are restricted to be the same. In regard to knowledge confidence of different group, the partial order curriculum can also be introduced together.

6. Conclusion

In this paper we have established a systematic theory for analyzing the SPL and SPCL via the concave conjugacy theory. We prove that the SPL model corresponds to optimizing a concave, increasing and continuous latent objective function. The relations among weight function, latent objective function and SP-regularizer can thus be obtained explicitly. Furthermore, two general methods for designing the SPL model has been rendered under this theoretical framework. Besides, the latent objective function of SPCL can be derived, instead of only those of SPL by some conventional methods. Such a study tends to be beneficial to facilitate deeper understandings and broader applications of SPL in the future research.

Appendix A

Theorem 3. If $g(\mathbf{v})$ is a function on a set $\Psi \subset R_+^n$, then

$$g^*(\mathbf{l}) = \inf_{\mathbf{v}\in\Psi}\{\langle \mathbf{v}, \mathbf{l}\rangle - g(\mathbf{v})\}$$

is increasing on $\mathcal{R}^n$.

Proof. Suppose $\mathbf{l}^1 \geq \mathbf{l}^2$. $\forall \varepsilon > 0, \exists \mathbf{v}^\varepsilon \in \Psi \subset \mathcal{R}_+^n$ such that

$$g^*(\mathbf{l}^1) + \varepsilon > \langle \mathbf{v}^\varepsilon, \mathbf{l}^1\rangle - g(\mathbf{v}^\varepsilon) \geq \langle \mathbf{v}^\varepsilon, \mathbf{l}^2 > -g(\mathbf{v}^\varepsilon) \geq g^*(\mathbf{l}^2)$$

Thus, it yields $g^*(\mathbf{l}^1) \geq g^*(\mathbf{l}^2)$. □

.

Appendix B

Theorem 4. For arbitrary function $R_{SP}(\mathbf{v})$ satisfying *dom* $R_{SP}(\mathbf{v}) \subset [0,1]^n$, let $g(\mathbf{v}) = -R_{SP}(\mathbf{v})$. Then

$$\begin{aligned}\inf_{\mathbf{v}\in[0,1]^n}\{\langle \mathbf{v},\mathbf{l}\rangle + R_{SP}(\mathbf{v})\} &= \inf_{v\in[0,1]^n}\{\langle \mathbf{v},\mathbf{l}\rangle - g^{**}(\mathbf{v})\} \\ &= \inf_{\mathbf{v}\in[0,1]^n}\{\langle \mathbf{v},\mathbf{l}\rangle + \overline{R_{SP}}(\mathbf{v})\} \\ \overline{R_{SP}}(\mathbf{v}) &= -g^{**}(\mathbf{v}).\end{aligned}$$

Proof. According to property 6, since *dom* $R_{SP}(\mathbf{v}) \subset [0,1]^n$, it yields

$$\begin{aligned}\inf_{\mathbf{v}\in[0,1]^n}\{\langle \mathbf{v},\mathbf{l}\rangle + R_{SP}(\mathbf{v})\} &= \inf_{\mathbf{v}\in\mathcal{R}^n}\{\langle \mathbf{v},\mathbf{l}\rangle + R_{SP}(\mathbf{v})\} = \inf_{\mathbf{v}\in\mathcal{R}^n}\{\langle \mathbf{v},\mathbf{l}\rangle - g(\mathbf{v})\} \\ &= g^*(\mathbf{v}) = (g^{**})^*(\mathbf{v}) = \inf_{\mathbf{v}\in\mathcal{R}^n}\{\langle \mathbf{v},\mathbf{l}\rangle - g^{**}(\mathbf{v})\}.\end{aligned}$$

According to the one-to-one correspondence of proper closed concave function, the hypograph of $g^{**}(\mathbf{v})$ is the closed convex hull of the *hyp* g.

It yields that *dom* $g^{**}(\mathbf{v}) \subset [0,1]^n$, and $g^{**}(\mathbf{v})$ is a proper concave and upper semi-continuous function. Thus, $\overline{R_{SP}}(\mathbf{v}) = -g^{**}(\mathbf{v})$ is a proper convex and lower semi-continuous function. □

Appendix C

Theorem 5. Suppose that v is a weight vector, l is the loss vector in $\mathcal{R}_+^n$, and λ is the age parameter.if $R_{SP}(v,\lambda)$ satisfy

1) $R_{SP}(v,\lambda)$ is strictly convex in v
2) $R_{SP}(v,\lambda)$ is lower semi-continuous in v
3) $dom_v\ R_{SP}(v,\lambda) \subset [0,1]^n$ and $\mathbf{0},\mathbf{1} \in cl(dom_v\ R_{SP}(v,\lambda))$

then **Condition in** l holds $\forall i \in \{1,2,\ldots,n\}$:

$v_i(\lambda,l)$ decrease with respect to l_i; $v_i(\lambda,l) \le 1$; $\lim_{l_i\to+\infty} v_i(\lambda,l) = 0$.

If $R_{SP}(v,\lambda) = \lambda R_{SP}(v)$ where $R_{SP}(v)$ satisfy the above condition in v, then **Condition in** λ holds $\forall i \in \{1,2,\ldots,n\}$:

$v_i(\lambda,l)$ increase with respect to λ; $v_i(\lambda,l) \le 1$; $\lim_{\lambda\to 0} v_i(\lambda,l) = 0$ $(l_i > 0)$ where $v(\lambda,l) = \arg\inf_{v\in[0,1]^n}\{\langle v,l\rangle + R_{SP}(v,\lambda)\}$.

Appendix D

Theorem 6. If $R_{SP}(v, \lambda)$ satisfy the simplified conditions on SPL and be strictly convex, then the latent SPL objective is of the form:

$$F_\lambda(l) = \int_0^{\mathbf{l}} \mathbf{v}(\lambda, j)\, dj + C(\lambda)$$

where $C(\lambda)$ is a function in λ.

Proof. According to property 7 and corollary 2, $v(\lambda, l) = \nabla F_\lambda(l)$. Therefore,

$$F_\lambda(l) = \int_0^{l} v(\lambda, j)\, dj + C(\lambda)$$

where $C(\lambda)$ is a function in λ. □

Appendix E

Theorem 9. Suppose $R_{SP}(v)$ satisfies Definition 11 and the curriculum as $\mathbf{v}^T\mathbf{k} \geq 0$ corresponding to $\Psi = \{\mathbf{v}|\mathbf{v}^T\mathbf{k} \geq 0\}$. If Ψ satisfies Assumption 1, then we have:

$$F^{new}(\mathbf{l}) = F \oplus \delta(\cdot|\Psi^\circ)(\mathbf{l}) = \sup_{\mathbf{l}^1+\mathbf{l}^2=\mathbf{l}} \{F(\mathbf{l}^1)) + \delta(\mathbf{l}^2|\Psi^\circ)\} = \sup_{\mathbf{l}^1 \in \mathbf{l}-ray_\mathbf{k}} F(\mathbf{l}^1\rangle \geq 0\}$$

is another non-empty closed convex cone and $ray_\mathbf{k}$ denotes the ray starting from the origin in direction $\mathbf{k}$.

Proof. If Ψ is a non-empty closed convex cone containing origin, then it holds that

$$\inf_{\mathbf{v}\in\Psi} \{\langle \mathbf{v}, \mathbf{l}\rangle - \delta(\mathbf{v}|\Psi)\} = \delta^*(\mathbf{l}|\Psi) = \delta(\mathbf{l}|\Psi^\circ),$$

where $\Psi^\circ = \{\mathbf{l}|\forall \mathbf{v} \in \Psi\ \langle \mathbf{v}, \mathbf{l}\rangle \geq 0\}$ is another non-empty closed convex cone. Based on this relationship, we have

$$F^{new}(\mathbf{l}) = F(\mathbf{l}) \oplus \delta(\mathbf{l}|\Psi^\circ) = \sup_{\mathbf{l}^1+\mathbf{l}^2=\mathbf{l}} \{F(\mathbf{l}^1)) + \delta(\mathbf{l}^2|\Psi^\circ)\} = \sup_{\mathbf{l}^1 \in \mathbf{l}-\Psi^\circ} F(\mathbf{l}^1).$$

Let $ray_\mathbf{k}$ denote the ray starting from the origin oriented to the direction k. Then

$$\Psi^\circ = \{\mathbf{l}|\forall \mathbf{v} \in \Psi\ \langle \mathbf{v}, \mathbf{l}\rangle \geq 0\} = \{\beta k|\beta \geq 0\} = ray_\mathbf{k}.$$

□

Appendix F

Theorem 10. Suppose $R_{SP}(\mathbf{v})$ is essential strictly convex, and satisfies Definition 11. Suppose we have the curriculum constraint $\mathbf{v}^T\mathbf{k} \geq 0$, corresponding to the curriculum region $\Psi = \{\mathbf{v}|\mathbf{v}^T\mathbf{k} \geq 0\}$. Then if Ψ satisfies Assumption 1, it holds that $\nabla F^{new}(\mathbf{l})^T\mathbf{k} = v^{new}(\mathbf{l})^T\mathbf{k} \geq 0$ and

$$F^{new}(\mathbf{l}) = \begin{cases} F(\mathbf{l}) & \mathbf{l} \in \partial(-R_{SP})(\mathbf{k}^\perp) - ray_{\mathbf{k}}, \\ \sup\limits_{\mathbf{l}' \in \mathbf{l}+line_{\mathbf{k}}} F(\mathbf{l}')(\geq F(\mathbf{l})) & l \in \partial(-R_{SP})(\mathbf{k}^\perp) + ray_{\mathbf{k}}. \end{cases}$$

Before proving the theorem, we first give two lemmas.

Lemma 1 (Limits of Direction Derivatives). *If $F(\mathbf{l})$ is a closed concave and differentiable function with the effective domain $\mathcal{R}^n$, and $\mathbf{v_k}(\mathbf{l})$ represents the directional derivative of $F(\mathbf{l})$ in direction $\mathbf{k}$, then*

$$F0^+(\mathbf{l}) = \lim_{\beta \to +\infty} \mathbf{v_k}(\mathbf{l} + \beta\mathbf{k})$$

where $F0^+(l)$ is the recession function of $F(\mathbf{l})$ determined by

$$hyp\ F0^+ = (\{y|y + hyp\ F \subset hyp\ F\}$$

Proof. Based on the results given by [25], it holds that

$$F0^+(\mathbf{k}) = \lim_{\beta \to +\infty} \frac{F(\mathbf{l} + \beta\mathbf{k}) - F(\mathbf{l})}{\beta}.$$

Since $F(\mathbf{l})$ is concave and differentiable, $F(\mathbf{l} + \beta\mathbf{k})$ is still concave and differentiable in β. Therefore, $v_{\mathbf{k}}(\mathbf{l} + \beta\mathbf{k})$ is decreasing in β, the Newton-Leibniz formula can be applied to getting:

$$F0^+(\mathbf{k}) = \lim_{\beta \to +\infty} \frac{F(\mathbf{l} + \beta\mathbf{k}) - F(\mathbf{l})}{\beta} = \lim_{\beta \to +\infty} \frac{\int_0^\beta v_{\mathbf{k}}(\mathbf{l} + t\mathbf{k})dt}{\beta},$$

$$v_{\mathbf{k}}(\mathbf{l}) \geq \lim_{\beta \to +\infty} \frac{\int_0^\beta v_{\mathbf{k}}(\mathbf{l} + t\mathbf{k})dt}{\beta} \geq \lim_{\beta \to +\infty} v_{\mathbf{k}}(\mathbf{l} + \beta\mathbf{k}).$$

Hence, we have:

$$\lim_{\beta \to +\infty} v_k(\mathbf{l} + \beta\mathbf{k}) = F0^+(\mathbf{k}).$$

□

Lemma 2 (Duality of Recession Function and Support Function). *The support function of $cl(dom\ R_{SP})$ is the recession function $F0^+(\mathbf{l})$ of $F(\mathbf{l})$, i.e.*

$$F0^+(\mathbf{k}) = \delta^*(\mathbf{k}|cl(dom\ R_{SP})),$$

where $F(\mathbf{l}) = \inf_{\mathbf{v}}\langle \mathbf{v}, \mathbf{l}\rangle + R_{SP}(\mathbf{v})$.

Proof. Let $*$ denote the concave conjugate, $'$ denote the convex conjugate, 0^+ denote the recession function generated by the recession cone of its hypograph and 0^- denote the recession function generated by the recession cone of its epigraph. Then it can be deduced that:

$$\begin{aligned} F0^+(\mathbf{k}) &= -((-F)0^-)(\mathbf{k}) \\ &= -(-\delta)'(\mathbf{k}|cl(-dom\ R_{SP})) = \delta^*(\mathbf{k}|cl(dom\ R_{SP})). \end{aligned}$$

The second equality holds due to Theorem 13.3 of the convex analysis [25] and the property 3, relation of concave conjugate and convex conjugate. □

Proof of the theorem. According to Theorem 9,

$$F^{new}(\mathbf{l}) = \sup_{\mathbf{l}^1 \in \mathbf{l} - ray_{\mathbf{k}}} F(\mathbf{l}^1) = \sup_{\beta \geq 0|\ \mathbf{l}^1 = \mathbf{l} - \beta\mathbf{k}} F(\mathbf{l}^1(\beta)).$$

Since $\mathbf{l}^1 = l - \beta\mathbf{k}$ is an affine mapping, $F(\mathbf{l}^1(\beta))$ is still a concave function in β. Take the derivative of $F(\mathbf{l}^1(\beta)$ in β, and we have

$$v_{-\mathbf{k}}(\mathbf{l} - \beta\mathbf{k}) = \nabla_\beta F(\mathbf{l}^1(\beta) = -v(\mathbf{l}^1(\beta))^T\mathbf{k}.$$

The *dom* R_{SP} reflects some properties of the latent objective as well as its direction derivative. According to Lemmas 1 and 2, we know that

$$\begin{gathered} \lim_{\beta \to +\infty} \mathbf{v}_{-\mathbf{k}}(\mathbf{l} - \beta k) = F0^+(-\mathbf{k}) \\ \lim_{\beta \to +\infty} v\mathbf{v}_{\mathbf{k}}(\mathbf{l} + \beta k) = -\lim_{\beta \to -\infty} \mathbf{v}_{-\mathbf{k}}(\mathbf{l} - \beta k) = F0^+(\mathbf{k}) \\ \lim_{\beta \to -\infty} \mathbf{v}_{-\mathbf{k}}(\mathbf{l} - \beta k) = -F0^+(\mathbf{k}). \end{gathered}$$

In order to consider the properties of the direction derivative more precisely, the situation can be divided into the following cases.

case 1. $0 \in (F0^+(-\mathbf{k}), -F0^+(\mathbf{k}))$

$F0^+(\mathbf{k}) = \delta^*(\mathbf{k}|cl(dom\ R_{SP})) < 0$ implies that there exists

$$\mathbf{v}_1 \in cl(dom\ R_{SP})$$

such that $\langle \mathbf{v}_1, \mathbf{k} \rangle\ < 0$.

$F0^+(-\mathbf{k}) = \delta^*(-\mathbf{k}|cl(dom\ R_{SP})) < 0$ implies that there exists

$$v_2 \in cl(dom\ R_{SP})$$

such that $\langle \mathbf{v}_2, \mathbf{k} \rangle\ > 0$.

Therefore,

$$0 \in (F0^+(-\mathbf{k}), -F0^+(\mathbf{k})) \Longleftrightarrow$$
$$\exists \mathbf{v}_1, \mathbf{v}_2 \in cl(dom\ R_{SP})\ s.t. \langle \mathbf{v}_1, \mathbf{k} \rangle\ < 0\ and \langle \mathbf{v}_2, \mathbf{k} \rangle\ > 0.$$

case 2. $0 \leq F0^+(\mathbf{k})$

$$\begin{aligned} &F0^+(\mathbf{k}) = \delta^*(\mathbf{k}|cl(dom\ R_{SP})) \geq 0 \\ \Longleftrightarrow\ &\langle \mathbf{v}, \mathbf{k} \rangle \geq 0\ \forall \mathbf{v} \in cl(dom\ R_{SP}) \\ \Longleftrightarrow\ &cl(dom\ R_{SP}) \subset \Psi \end{aligned}$$

case 3. $0 \leq F0^+(-\mathbf{k})$

$$\begin{aligned} &F0^+(-\mathbf{k}) = \delta^*(-\mathbf{k}|cl(dom\ R_{SP})) \geq 0 \\ \Longleftrightarrow\ &\langle \mathbf{v}, \mathbf{k} \rangle \leq 0\ \forall \mathbf{v} \in cl(dom\ R_{SP}) \\ \Longleftrightarrow\ &-(cl(dom\ R_{SP})) \subset \Psi \\ \Longrightarrow\ &int(\Psi) \cap int([0,1]^n) \cap int(dom\ R_{SP}) = \emptyset \end{aligned}$$

Hence, **case 2** and **case 3** have already been excluded by the assumptions regarding Ψ.

In **case 1**,

$$0 \in (\lim_{\beta \to +\infty} \mathbf{v}_{-\mathbf{k}}(\mathbf{l} - \beta\mathbf{k}), \lim_{\beta \to -\infty} \mathbf{v}_{-\mathbf{k}}(\mathbf{l} - \beta\mathbf{k}))$$

i.e.

$$0 \in (\lim_{\beta \to -\infty} \mathbf{v}_{\mathbf{k}}(\mathbf{l} - \beta\mathbf{k}), \lim_{\beta \to +\infty} \mathbf{v}_{\mathbf{k}}(\mathbf{l} - \beta\mathbf{k})).$$

Due to the Darboux theorem, there exists some $\beta^0(\mathbf{l}) \in (-\infty, +\infty)$ such that $\mathbf{v}_{\mathbf{k}}(\mathbf{l} - \beta^0(\mathbf{l})\mathbf{k}) = \nabla_\beta F(\mathbf{l}^1(\beta^0(\mathbf{l})) = 0$. Based on the monotonicity of $\nabla_\beta F(\mathbf{l}^1(\beta)$ and the continuity of $F(\mathbf{l}^1(\beta)$, the maximum must be attained

at a point or a closed finite interval. Thus, we set $\beta^0(\mathbf{l}) = \max \arg_\beta \{v_k(\mathbf{l} - \beta\mathbf{k}) = 0\}$

Let $l^*(\mathbf{l}) = l^1(\beta^0(\mathbf{l})) = \mathbf{l} - \beta^0(\mathbf{l})\mathbf{k}$ and it holds that $F(\mathbf{l})$ realizes the maximum on the line in direction $\mathbf{k}$ at point $l^*(\mathbf{l})$. Further, $\forall \beta \geq 0$ $\mathbf{l}^*(l) = \mathbf{l}^*(l + \beta\mathbf{k})$ takes the same value on the line in direction $\mathbf{k}$.

For the value of $F^{new}(\mathbf{l})$ on the line with direction $\mathbf{k}$, it holds that $F^{new}(\mathbf{l})$ equals to $F(\mathbf{l})$ on the side $l^*(\mathbf{l}) \notin \mathbf{l} - ray_{\mathbf{k}}$ while $F^{new}(\mathbf{l})$ equals to the constant $F(l^*(\mathbf{l}))$ on the other side $l^*(\mathbf{l}) \in \mathbf{l} - ray_{\mathbf{k}}$.

We can then obtain

$$F^{new}(\mathbf{l}) = \sup_{\beta \geq 0 |\ \mathbf{l}^1 = \mathbf{l} - \beta\mathbf{k}} F(l^1) = \begin{cases} F(\mathbf{l}) & l^*(\mathbf{l}) \notin \mathbf{l} - ray_{\mathbf{k}} \\ F(l^*(\mathbf{l})) & l^*(\mathbf{l}) \in \mathbf{l} - ray_{\mathbf{k}} \end{cases}$$

Since $F(\mathbf{l})$ is differentiable, $R_{SP}(\mathbf{v})$ is strictly concave on $[0,1]^n$ as well as $[0,1]^n \cap \Psi$. As a result of the previous analysis, $F^{new}(\mathbf{l})$ is differentiable.

For a fixed $\mathbf{l}$, consider

$$\nabla_\beta F^{new}(l^1(\beta)) = \begin{cases} -v^{new}(l^1(\beta))^T\mathbf{k} = -v(l^1(\beta))^T\mathbf{k} & l^*(l) \notin l - ray_{\mathbf{k}} \\ -v^{new}(l^1(\beta))^T\mathbf{k} = 0 & l^*(l) \in l - ray_{\mathbf{k}} \end{cases}$$

Since $F^{new}(l^1(\beta)) = \sup_{\alpha \geq 0 |\ l^1 = l - \beta\mathbf{k} - \alpha\mathbf{k}} F(l^1) = \sup_{\alpha \geq \beta |\ l^1 = l - \alpha\mathbf{k}} F(l^1)$ is a decreasing function in β, it holds that $\nabla_\beta F^{new}(l^1(\beta)) \leq 0$.

In case $l^*(\mathbf{l}) \notin \mathbf{l} - ray_{\mathbf{k}}$, by plugging $\beta = 0$ into it, it yields

$$v^{new}(\mathbf{l})^T\mathbf{k} \geq 0.$$

In case $l^*(l) \in l - ray_{\mathbf{k}}$, one can obtain that

$$v^{new}(l)^T\mathbf{l} = 0.$$

$$\begin{aligned} F^{new}(\mathbf{l}) = \sup_{\beta \geq 0 |\ l^1 = \mathbf{l} - \beta\mathbf{k}} F(l^1) &= \begin{cases} F(\mathbf{l}) & l^*(l) \notin \mathbf{l} - ray_{\mathbf{k}} \\ F(l^*(l))(\geq F(\mathbf{l})) & l^*(l) \in \mathbf{l} - ray_{\mathbf{k}} \end{cases} \\ &= \begin{cases} F(\mathbf{l}) & l - \beta^0(l)\mathbf{k} \notin l - ray_{\mathbf{k}} \\ F(\mathbf{l} - \beta^0(l)\mathbf{k})(\geq F(l)) & l - \beta^0(l)\mathbf{k} \in l - ray_{\mathbf{k}} \end{cases} \\ &= \begin{cases} F(\mathbf{l}) & \beta^0(l) < 0 \\ F(\mathbf{l} - \beta^0(l)\mathbf{k})(\geq F(\mathbf{l})) & \beta^0(l\mathbf{l}) \geq 0 \end{cases} \end{aligned}$$

The most important thing for determining $F^{new}(\mathbf{l})$ is to determine the critical region $\{\mathbf{l} | \beta^0(\mathbf{l}) = 0\}$, since the critical region divides the $\mathcal{R}^n$ into two parts.

- $F^{new}(\mathbf{l})$ becomes larger than $F(\mathbf{l})$ on the part in the direction $\mathbf{k}$ of the critical region.
- $F^{new}(\mathbf{l})$ keep the same value with $F(\mathbf{l})$ on the part in the direction $-\mathbf{k}$ of the critical region.

$$\beta^0(\mathbf{l}) = 0 \Longrightarrow 0 = \nabla_\beta F(\mathbf{l} - \beta\mathbf{k})|_{\beta=0} = -\nabla F(\mathbf{l})^T\mathbf{k}$$
$$\Longleftrightarrow \nabla F(\mathbf{l}) \in \mathbf{k}^\perp \Longleftrightarrow \mathbf{l} \in \partial(-R_{SP})(\mathbf{k}^\perp).$$

$\partial(-R_{SP})(\mathbf{k}^\perp)$[1] is the critical region of the problem with respect to the $\Psi = \{\mathbf{v}|\mathbf{v}^T\mathbf{k} \geq 0\}$.

Then we obtain

$$F^{new}(\mathbf{l}) = \begin{cases} F(\mathbf{l}) & \mathbf{l} \in \partial(-R_{SP})(\mathbf{k}^\perp) - ray_{\mathbf{k}} \\ F(\mathbf{l} - \beta^0(\mathbf{l})\mathbf{k})(\geq F(\mathbf{l})) & l \in \partial(-R_{SP})(\mathbf{k}^\perp) + ray_{\mathbf{k}} \end{cases}$$

The equivalence can then be illustrated as follows:

- $\mathbf{l} \in \partial(-R_{SP})(\mathbf{k}^\perp) - ray_k \Longleftrightarrow \exists\widetilde{\beta} \leq 0 \ s.t. v_k(l - \widetilde{\beta}k) = 0$

$$\Longrightarrow \exists\widetilde{\beta} \leq 0 \ s.t. \ \beta^0(\mathbf{l}) \geq \widetilde{\beta} \ and \ v_k(\mathbf{l} - \widetilde{\beta}k) = 0$$

  If $\beta^0(\mathbf{l}) < 0$ then we obtain the result that we need. If $\beta^0(\mathbf{l}) \geq 0$ then $0 \in [\widetilde{\beta}, \beta^0(\mathbf{l})]$. It still yields that $\sup_{\beta \geq 0|\ l^1 = l - \beta\mathbf{k}} F(l^1) = F(l - \beta\mathbf{k}) \ \beta \in [\widetilde{\beta}, \beta^0(\mathbf{l})]$. That means $F^{new}(\mathbf{l}) = F(\mathbf{l})$.

- $l \in \partial(-R_{SP})(\mathbf{k}^\perp) + ray_{\mathbf{k}} \Longleftrightarrow \exists\widetilde{\beta} \geq 0 \ s.t. \mathbf{v_k}(\mathbf{l} - \widetilde{\beta}\mathbf{k}) = 0$

$$\Longrightarrow \exists\widetilde{\beta} \geq 0 \ s.t. \ \beta^0(\mathbf{l}) \geq \widetilde{\beta} \geq 0$$

  Therefore, $F^{new} = F(\mathbf{l} - \beta^0(\mathbf{l})\mathbf{k})$.

Hence,

$$F^{new}(\mathbf{l}) = \begin{cases} F(\mathbf{l}) & \mathbf{l} \in \partial(-R_{SP})(\mathbf{k}^\perp) - ray_{\mathbf{k}}, \\ \sup_{\mathbf{l}' \in \mathbf{l} + line_{\mathbf{k}}} F(\mathbf{l}')(\geq F(\mathbf{l})) & l \in \partial(-R_{SP})(\mathbf{k}^\perp) + ray_{\mathbf{k}}. \end{cases}$$

□

[1]Notice

$$int(dom\ R_{SP} \cap [0,1]^n) = int(dom\ (-R_{SP})) \subset dom\ \partial(-R_{SP})$$
$$\subset dom\ (-R_{SP}) = dom\ R_{SP} \cap [0,1]^n.$$

Appendix G

Theorem 11. Suppose $R_{SP}(\mathbf{v})$ is essential strictly convex, and satisfies Definition 11. Suppose the curriculum is $\mathbf{v}^T\mathbf{k} \geq b$ corresponding to the curriculum region $\Psi = \{\mathbf{v}|\mathbf{v}^T\mathbf{k} \geq b\}$. If Ψ satisfies Assumption 1, then

$$F^{new}(\mathbf{l}) = \begin{cases} F(\mathbf{l}) & \mathbf{l} \in \partial g(\mathbf{k}^\perp + \frac{b\mathbf{k}}{\|\mathbf{k}\|_2}) - ray_{\mathbf{k}}, \\ F(\mathbf{l} - \beta^0(\mathbf{l})\mathbf{k}) + \beta^0 b (\geq F(\mathbf{l})) & \mathbf{l} \in \partial g(\mathbf{k}^\perp + \frac{b\mathbf{k}}{\|\mathbf{k}\|_2}) + ray_{\mathbf{k}}, \end{cases}$$

where $\beta^0(\mathbf{l}) = \max \arg_\beta \{\nabla F(\mathbf{l} - \beta\mathbf{k})^T\mathbf{k} = 0\}$.

Proof. Since $\Psi = \{\mathbf{v}^T\mathbf{k} \geq b\}$ satisfies assumption 1, we have that

$$\begin{aligned} \delta^*(\mathbf{l}|\Psi) &= \begin{cases} \frac{b\mathbf{l}^T\mathbf{k}}{\|\mathbf{k}\|_2} & \mathbf{l} \in ray_{\mathbf{k}}, \\ -\infty & \mathbf{l} \notin ray_{\mathbf{k}}, \end{cases} \\ F^{new}(\mathbf{l}) &= F(\mathbf{l}) \oplus \delta^*(\mathbf{l}|\Psi) = \sup_{\mathbf{l}^1+\mathbf{l}^2=\mathbf{l}} \{F(\mathbf{l}^1)) + \delta^*(\mathbf{l}^2|\Psi)\} \\ &= \sup_{\beta\geq 0|\mathbf{l}^1+\beta\mathbf{k}=l} F(\mathbf{l}^1) + \beta b = \sup_{\beta\geq 0} F(\mathbf{l} - \beta\mathbf{k}) + \beta b. \end{aligned}$$

It also holds that $\nabla_\beta(F(\mathbf{l} - \beta\mathbf{k}) + \beta b) = \mathbf{v}_{-\mathbf{k}}(\mathbf{l} - \beta\mathbf{k}) + b$ and $F0^+(-\mathbf{k}) = \lim_{\beta\to+\infty} \mathbf{v}_{-\mathbf{k}}(l - \beta\mathbf{k})$, $-F0^+(\mathbf{k}) = \lim_{\beta\to-\infty} \mathbf{v}_{-\mathbf{k}}(\mathbf{l} - \beta\mathbf{k})$.

$F0^+(\mathbf{k}) = \delta^*(\mathbf{k}|cl(dom\ R_{SP})) < b$ implies that there exists

$$\mathbf{v}_1 \in cl(dom\ R_{SP})$$

such that $\langle \mathbf{v}_1, \mathbf{k}\rangle < b$; and $F0^+(-\mathbf{k}) = \delta^*(-\mathbf{k}|cl(dom\ R_{SP})) < b$ implies that there exists $\mathbf{v}_2 \in cl(dom\ R_{SP})$ such that $\langle \mathbf{v}_2, \mathbf{k}\rangle > b$. Therefore,

$$\begin{gathered} -b \in (F0^+(-\mathbf{k}), -F0^+(\mathbf{k})) \Longleftrightarrow \\ \exists \mathbf{v}_1, \mathbf{v}_2 \in cl(dom\ R_{SP}) \quad \text{s.t. } \langle \mathbf{v}_1, \mathbf{k}\rangle < b \quad \text{and} \quad \langle \mathbf{v}_2, \mathbf{k}\rangle > b. \end{gathered}$$

Thus, the supremum can be attained on each line in direction $\mathbf{k}$. The critical region is obtained by:

$$\begin{gathered} 0 = \nabla_\beta F(\mathbf{l} - \beta\mathbf{k})|_{\beta=0} + b = -\nabla F(\mathbf{l})^T\mathbf{k} + b \\ \Longleftrightarrow \nabla F(l) \in \mathbf{k}^\perp + \frac{b\mathbf{k}}{\|k\|_2} \Longleftrightarrow \mathbf{l} \in \partial g\left(\mathbf{k}^\perp + \frac{b\mathbf{k}}{\|k\|_2}\right). \end{gathered}$$

By inheriting the analysis in the proof of Theorem 10, it can be obtained that:

$$F^{new}(\mathbf{l}) = \begin{cases} F(\mathbf{l}) & \mathbf{l} \in \partial g(\mathbf{k}^{\perp} + \frac{b\mathbf{k}}{\|\mathbf{k}\|_2}) - ray_{\mathbf{k}}, \\ F(\mathbf{l} - \beta^0(\mathbf{l})\mathbf{k}) + \beta^0 b (\geq F(\mathbf{l})) & \mathbf{l} \in \partial g(\mathbf{k}^{\perp} + \frac{b\mathbf{k}}{\|\mathbf{k}\|_2}) + ray_{\mathbf{k}}. \end{cases}$$

□

References

[1] M. P. Kumar, B. Packer, and D. Koller, *Self-paced learning for latent variable models*, in: Proceedings of the 23rd International Conference on Neural Information Processing Systems, NIPS'10, pages 1189–1197, Vancouver, British Columbia, Canada, Dec. 06-09, 2010.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, *Curriculum learning*, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML'09, pages 41–48, Montreal, Quebec, Canada, June 14-18, 2009.

[3] V. I. Spitkovsky, H. Alshawi, and D. Jurafsky, *From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing*, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT'10, pages 751–759, Los Angeles, California, June 02-04, 2010.

[4] A. Lapedriza, H. Pirsiavash, Z. Bylinskii, and A. Torralba, *Are all training examples equally valuable?* `arXiv:1311.6510 [cs.CV]`, (2013).

[5] Q. Zhao, D. Y. Meng, L. Jiang, Q. Xie, Z. B. Xu, and A. Hauptmann, *Self-paced learning for matrix factorization*, in: Proceedings of the 29th AAAI Conference on Artificial Intelligence, AAAI'15, pages 3196–3202, Austin, Texas, Jan. 25-30, 2015.

[6] D. Y. Meng, Q. Zhao, and L. Jiang, *A theoretical understanding of self-paced learning*, Information Sciences, **414** (2017), 319–328.

[7] L. Jiang, D. Y. Meng, T. Mitamura, and A. Hauptmann, *Easy samples first: Self-paced reranking for zero-example multimedia search*, in: Proceedings of the 22nd ACM international conference on Multimedia, MM'14, pages 547–556, Orlando, Florida, USA, Nov. 03-07, 2014.

[8] V. Avramova, *Curriculum learning with deep convolutional neural networks*, (2015).

[9] C. Xu, D. C. Tao, and C. Xu, *Multi-view self-paced learning for clustering*, In "Proceedings of the 24th International Joint Conference on Artificial Intelligence", IJCAI'15, pages 3974–3980, Buenos Aires, Argentina, July 25-31, 2015.

[10] C. S. Li, F. Wei, J. C. Yan, W. S. Dong, Q. S. Liu, and H. Y. Zha, *Self-paced multi-task learning*, `arXiv:1604.01474 [cs.LG]`, (2016).

[11] T. Pi, X. Li, Z. F. Zhang, D. Y. Meng, F. Wu, J. Xiao, and Y. T. Zhuang, *Self-paced boost learning for classification*, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, pages 1932–1938, New York, NY, USA, July 09-15, 2016.

[12] J. S. Supancic, and D. Ramanan, *Self-paced learning for long-term tracking*, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR'13, pages 2379–2386, Washington, DC, USA, June 23-28, 2013.

[13] S. P. Zhou, J. J. Wang, D. Y. Meng, X. M. Xin, Y. b. Li, Y. H. Gong, and N. N. Zheng, *Deep self-paced learning for person re-identification*, Pattern Recognition, **39** (2018), 739–751.

[14] L. Lin, K. Z. Wang, D. Y. Meng, W. M. Zuo, and L. Zhang, *Active self-paced learning for cost-effective and progressive face identification*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **40** (2018), no. 1, 7-19.

[15] D. W. Zhang, L. Yang, D. Y. Meng, D. Xu, and J. W. Han, *SPFTN: A self-paced fine-tuning network for segmenting objects in weakly labelled videos*, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR'17, pages 4429–4437, Honolulu, Hawaii, USA, July 21-26, 2017.

[16] X. D. Liang, S. Liu, Y. C. Wei, L. Q. Liu, L. Liang, and S. C. Yan, *Towards computational baby learning: A weakly-supervised approach for object detection*, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV'15, pages 999-1007, Santiago, Chile, Dec. 07-13, 2015.

[17] Y. C. Wei, X. D. Liang, Y. P. Chen, X. H. Shen, Y. P. Chen, J. S Feng, Y. Zhao, and S. C. Yan, *Stc: A simple to complex framework for*

weakly-supervised semantic segmentation, IEEE transactions on pattern analysis and machine intelligence, **39** (2017), no. 11, 2314–2320.

[18] X. D. Liang, Y. C. Wei, L. Liang, Y. P. Chen, X. H. Shen, J. C. Yang, and S. C. Yan, *Learning to segment human by watching Youtube*, IEEE transactions on pattern analysis and machine intelligence, **39** (2017), no. 7, 1462-1468.

[19] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann, *Self-paced curriculum learning*, in: Proceedings of the 29th AAAI Conference on Artificial Intelligence, AAAI'15, pages 2694–2700, Austin, Texas, Jan. 25-30, 2015.

[20] L. Jiang, S. Yu, D. Y. Meng, T. Mitamura, and A. Hauptmann, *Bridging the ultimate semantic gap: A semantic search engine for internet videos*, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ICMR'15, pages 27–34, Shanghai, China, June 23-26, 2015.

[21] D. W. Zhang, D. Y. Meng, C. Li, L. Jiang, Q. Zhao, and J. W. Han, *A Self-paced multiple-instance learning framework for co-saliency detection*, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV'15, pages 594–602, Santiago, Chile, Dec. 07-13, 2015.

[22] D. W. Zhang, D. Y. Meng, and J. W. Han, *Co-saliency detection via a self-paced multiple-instance learning framework*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **39** (2017), no. 5, 865-878.

[23] L. Jiang, D. Y. Meng, S. Yu, Z. Z. Lan, S. G. Shan, and A. Hauptman, *Self-paced learning with diversity*, in: Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS'14, pages 2078–2086, DecMontreal, Canada, Dec. 08-13, 2014.

[24] H. Li, M. G Gong, D. Y. Meng, and Q. G. Miao, *Multi-objective self-paced learning*, in: Proceedings of the 30th AAAI Conference on Artificial Intelligence, AAAI'16, pages 1802–1808, Phoenix, Arizona, Feb. 12-17, 2016.

[25] F. Ma, D. Y. Meng, Q. Xie, Z. N. Li, and X. Y. Dong, *Self-paced co-training*, in: Proceedings of the 34th International Conference on Machine Learning, ICML'17, pages 2275–2284, Sydney, NSW, Australia, Aug. 06-11, 2017.

[26] Z. Lu. Ma, S. Q. Liu, and D. Y. Meng, *On convergence property of implicit self-paced objective*, Information Sciences, 2018.

[27] Y. B. Fan, R. He, J. Liang, and B. G. Hu, *Self-paced learning: an implicit regularization perspective*, in: Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI'16, pages 1877–1883, San Francisco, California, USA, Feb. 04-09, 2017.

[28] C. S. Li, J. C. Yan, F. Wei, X. Y. Zhang, Q. S. Liu, and H. Y. Zha, *A self-paced regularization framework for multi-label learning*, IEEE Transactions on Neural Networks and Learning Systems, 2017. DOI:10.1109/TNNLS.2017.2697767.

[29] R. T. Rockafellar, *Convex analysis*, (1970).

[30] W. Fenchel, *On conjugate convex functions*, Canadian Journal of Mathematics, **1** (1949), no. 1, 73–77.

School of Mathematics and Statistics, Xi'an Jiaotong University
Xian Ning West Road, Xi'an, Shaan'xi, 710049, P.R. China
E-mail address: gujin4444@163.com
E-mail address: fancycold@163.com
E-mail address: dymeng@mail.xjtu.edu.cn
E-mail address: wangkd13@gmail.com
E-mail address: Yongzhang761@mail.xjtu.edu.cn