Information Sciences 414 (2017) 319–328

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A theoretical understanding of self-paced learning

Deyu Meng[a,*], Qian Zhao[a], Lu Jiang[b]

[a] School of Mathematics and Statistics and Ministry of Education Key Lab of Intelligent Networks and Network Security, Xian Jiaotong University, Xian 710049, PR China
[b] School of Computer Science, Carnegie Mellon University, USA

ARTICLE INFO

Article history:
Received 14 August 2016
Revised 23 May 2017
Accepted 25 May 2017
Available online 26 May 2017

MSC:
00-01
99-00

Keywords:
Self-paced learning
Curriculum learning
Multimedia event detection
Non-convex regularized penalty

ABSTRACT

Self-paced learning (SPL) is a recently proposed methodology designed by mimicking through the learning principle of humans/animals. A variety of SPL realization schemes have been designed for different computer vision and pattern recognition tasks, and empirically demonstrated to be effective in these applications. However, the literature is in lack of the theoretical understanding of SPL. Regarding this research gap, this study attempts to provide some new theoretical understanding of the SPL scheme. Specifically, we prove that the solution strategy on SPL accords with a majorization minimization algorithm implemented on an implicit objective function. Furthermore, we found that the loss function contained in this implicit objective has a similar configuration with the non-convex regularized penalty (NCRP) known in statistics and machine learning. Such connection inspires us to discover more intrinsic relationships between the SPL regimes and the NCRP forms, like smoothly clipped absolute deviation (SCAD), logarithmic penalty (LOG) and non-convex exponential penalty (EXP). The insight of the robustness under SPL can then be finely explained. We also analyze the capability of SPL regarding its easy loss-prior-embedding property, and provide an insightful interpretation of the effectiveness mechanism under current SPL variations. Moreover, we design a group-partial-order loss prior, which is especially useful for weakly labeled large-scale data processing tasks. By applying SPL with this loss prior to the FCVID dataset, which is currently one of the largest manually annotated video dataset, our method achieves state-of-the-art performance above existing methods, which further supports the proposed theoretical arguments.

1. Introduction

Since their inception, *curriculum learning* (CL) [2] and *self-paced learning* (SPL) [15] have been attracting increasing attention in the machine learning and pattern recognition fields. The idea under this learning paradigm is to simulate the learning principle of humans/animals, which generally starts by learning easier aspects of a learning task, and then gradually introduces more complex examples into training [13]. Instead of heuristically designing a curriculum by ranking samples based on manually presetting easiness measurements as in CL [18,24], SPL formulates this ad-hoc scheme as a concise model by introducing a regularizor into the learning objective. Such amelioration guides a sound SPL regime to automatically optimize an appropriate curriculum by the model itself, making it general enough to solving problems in various applications and avoid the subjective easiness measure setting problem [16,19,25,27]. Very recently, a variety of SPL realization schemes,

[*] Corresponding author.
E-mail addresses: dymeng@mail.xjtu.edu.cn (D. Meng), timmy.zhaoqian@gmail.com (Q. Zhao), lujiang@cs.cmu.edu (L. Jiang).

http://dx.doi.org/10.1016/j.ins.2017.05.043
0020-0255/© 2017 Published by Elsevier Inc.

like self-paced reranking (SPaR) [8] and self-paced multi-instance learning (SP-MIL) [36], have been proposed and shown to be effective for multiple computer vision and multimedia analysis tasks.

Albeit rational in intuition and effective in experience, there are only few investigations on the explanation of the underlying mechanism of SPL. Specifically, even though it is easy to prove that the SPL regime is convergent by adopting an alternative optimization strategy (AOS) on the SPL model, it is still unclear where this SPL iteration converges and why SPL is robust in solving the learning problems especially with highly noisy data. Such in-depth investigations, however, can be considerably necessary for future developments of CL, SPL and their related realizations, and will illuminate whether the SPL methodology is just an idealistic method occasionally performed on several datasets or a rigorous and solid scientific research field worthy of further exploration.

This study aims at understanding the theoretical insight under SPL. Our main results can be summarized as follows:

First, we prove that the AOS algorithm commonly utilized to solve the SPL problem is identical to a *majorization minimization* (MM) [28] algorithm implemented on an implicit SPL objective function. In the recent decade, MM has attracted much attention in machine learning and optimization, and many theories have been proposed. Such results facilitate an easy analysis on the properties underlying the SPL solving strategy, like convergence and stability, by utilizing the existing knowledge on MM.

Second, we prove that the loss function contained in this implicit SPL objective is closely related to the *non-convex regularized penalty* (NCRP). Specifically, we discover that multiple current SPL realizations exactly comply with some well known NCRP terms (e.g., the hard and linear SPL regimes are equivalent to the optimizations on implicit losses with the forms of capped-norm penalty (CNP) [7,35,38] and minimax concave plus penalty (MCP) [34], respectively). Such connection inspires us to discover more intrinsic relationship between SPL regimes and known NCRP forms, like smoothly clipped absolute deviation (SCAD) [5,23], logarithmic penalty (LOG) [32], and non-convex exponential penalty (EXP) [3].

Third, by connecting the SPL optimization with the NCRP loss minimization problems, we provide an easy explanation on why SPL is able to perform robust in the presence of outliers/heavy noises, and accordingly illustrate new insightful understandings of the intrinsic working mechanism under SPL. We also analyze the superiority of SPL regarding its easy loss-prior-embedding property beyond conventional learning strategies with a pre-fixed loss function. Such a property is expected to help a non-convex optimization problem better avert unreasonable local minima and makes SPL more compliant with the instructor-student-collaborative-learning mode in human education. Such understanding facilitates an easy interpretation for its intrinsic effective mechanism of SPL in previous applications [8–10,36,40].

We also propose a group-partial-order loss prior to facilitate better SPL performance in weakly labeled large-scale problems, and implement this SPL regime on the FCVID dataset, which is currently one of the biggest manually annotated video dataset. Our method achieves state-of-the-art performance as compared with previous methods. The results further support the theoretical arguments presented in this study.

The paper is organized as follows. Section 2 introduces the related work and Section 3 presents our main theoretical results, and clarifies the relationships between AOS and MM algorithms as well as SPL and NCRP problems. Section 4 introduces the group-partial-order loss prior and provides the related experimental results on the FCVID dataset. Finally, concluding remarks are made.

2. Related work

Curriculum Learning (CL). Inspired by the intrinsic learning principle of humans/animals, Bengio et al. [2] formalized the fundamental definition of CL. The core idea is to incrementally involve samples in learning, where easy samples are introduced first and more complex ones are then gradually included. These gradually included samples from easy to complex correspond to the curricula learned in different grown-up stages of humans/animals. This strategy, as supported by empirical evaluation, is helpful in alleviating the local optimum problem in non-convex optimization [1,22].

Self-paced Learning (SPL). Instead of using the heuristic strategies, Kumar et al. [15] formulated the key principle of CL as a concise SPL model. Formally, given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, in which $\mathbf{x}_i$ and y_i denote the ith observed sample and its label, respectively, $L(y_i, g(\mathbf{x}_i, \mathbf{w}))$ denotes the loss function that calculates the cost between the ground truth label y_i and the estimated one $g(\mathbf{x}_i, \mathbf{w})$, and $\mathbf{w}$ represents the model parameter in the decision function g. The SPL model includes a weighted loss term on all samples and a general self-paced regularizer imposed on sample weights, expressed as:

$$\min_{\mathbf{w}, \mathbf{v} \in [0,1]^n} \mathbf{E}(\mathbf{w}, \mathbf{v}, \lambda) = \sum_{i=1}^{n} (v_i L(y_i, f(\mathbf{x}_i, \mathbf{w})) + f(v_i, \lambda)), \tag{1}$$

where λ is the age parameter for controlling the learning pace, and $f(v, \lambda)$ represents the self-paced regularizer (SP-regularizer), whose intrinsic conditions have been theoretically abstracted by [8,40]. By jointly learning the model parameter $\mathbf{w}$ and the latent weight $\mathbf{v} = [v_1, \cdots, v_n]^T$ by AOS with gradually increasing age parameter, more samples can be automatically included in the training from easy to complex in a purely self-paced way.

Multiple variations of this SPL learning regime, like self-paced reranking [8], self-paced multiple instance learning [36,37], self-paced learning with diversity [9], and self-paced curriculum learning [10], have been proposed under the format (1). The effectiveness of this SPL paradigm, especially its robustness in highly corrupted data, has been empirically validated in various machine learning and computer vision tasks, such as object detector adaptation [27], specific-class segmentation learning [16], visual category discovery [19], concept learning [20], long-term tracking [25], and multimedia event detection [33].

There are few investigations, however, that attempt to theoretically explain the intrinsic mechanism under SPL. In this paper, we attempt to enhance the theoretical understanding on this learning paradigm.

Non-convex Regularized Penalty (NCRP). NCRP has been demonstrated to have attractive properties in sparse estimation (as a penalty term) [5,34,38] and robust learning (as a loss term) [26,30] both theoretically and practically, and attracted much attention in machine learning and statistics in recent years. Various NCRP realizations have also been proposed. Typical ones include the capped-norm based penalty (CNP) [7,35,38], the minimax concave plus penalty (MCP) [34], the smoothly clipped absolute deviation penalty (SCAD) [5], the logarithmic penalty (LOG) [32], and the non-convex exponential penalty (EXP) [3]. The mathematical forms of these NCRP terms in one dimension cases are listed as follows [12,39]:

$$\begin{aligned}
&\mathrm{CNP}:\ \mathrm{p}^{\mathrm{CNP}}_{\gamma,\lambda}(\mathrm{t}) = \gamma\min(|\mathrm{t}|,\lambda), \lambda > 0\\
&\mathrm{MCP}:\ p^{MCP}_{\gamma,\lambda}(t) = \begin{cases}\gamma(|t| - \frac{t^2}{2\gamma\lambda}), & \text{if } |t| < \gamma\lambda\\ \frac{\gamma^2\lambda}{2}, & \text{if } |t| \geq \gamma\lambda\end{cases}\\
&\mathrm{SCAD}:\ p^{SCAD}_{\gamma,\lambda}(t) = \begin{cases}\lambda|t|, & \text{if } |t| \leq \lambda\\ \frac{t^2-2\gamma\lambda|t|+\lambda^2}{2(1-\gamma)}, & \text{if } \lambda < |t| \leq \gamma\lambda\\ \frac{(\gamma+1)\lambda^2}{2}, & \text{if } |t| \geq \gamma\lambda\end{cases}\\
&\mathrm{LOG}:\ p^{LOG}_{\gamma,\alpha}(t) = \frac{1}{\gamma}\log(1+\alpha|t|)\\
&\mathrm{EXP}:\ p^{EXP}_{\gamma,\alpha}(t) = \frac{1}{\gamma}(1-\exp(-\alpha|t|)).
\end{aligned} \tag{2}$$

Albeit possessing elegant statistic properties and empirically verified to be effective in specific applications through finely designed solving strategies, involving such NCRP terms brings non-convexity to the model. This tends to result in the issue in which the algorithm easily becomes stuck at an undesired local minima of the problem [26,30].

In this work, we will construct the relationship between the NCRP terms and the SPL regimes and show that helpful loss prior knowledge can be easily embedded into the SPL framework, which is expected to facilitate a NCRP model possibly avoiding unreasonable local minima of the problem and attaining more rational ones that better comply with real states.

Majorization Minimization (MM) Algorithm. The MM algorithms have wide applications in machine learning and statistical inference [17]. These algorithms turn a complicated optimization problem into a tractable one by alternatively iterating the majorization and minimization steps. In particular, considering a minimization problem with the objective $F(\mathbf{w})$, given an estimate of $\mathbf{w}^k$ at the k^{th} iteration, a typical MM algorithm consists of the following two steps:

Majorization Step: Substitute $F(\mathbf{w})$ by a surrogate function $Q(\mathbf{w}|\mathbf{w}^k)$ such that:

$$F(\mathbf{w}) \leq Q(\mathbf{w}|\mathbf{w}^k)$$

with equality holding at $\mathbf{w} = \mathbf{w}^k$.

Minimization Step: Obtain the next parameter estimate $\mathbf{w}^{k+1}$ by solving the following minimization problem:

$$\mathbf{w}^{k+1} = \arg\min_{\mathbf{w}} Q(\mathbf{w}|\mathbf{w}^k).$$

It is easy to see that when the minimization of $Q(\mathbf{w}|\mathbf{w}^k)$ is tractable, the MM algorithm can then be very easily implemented, even when the original objective $F(\mathbf{w})$ might be difficult to optimize. Such a solving strategy has also been proven to own many good theoretical properties, like convergence and stability, under certain conditions.

3. SPL model and algorithm revisit

3.1. Axiomic definition of SP-regularizer

By mathematically abstracting the insightful properties underlying an SPL regime, [8,40] presented a formal definition for the SP-regularizer $f(v, \lambda)$ involved in the SPL model (1) as follows:

Definition 3.1 (SP-regularizer)**.** Suppose that v is a weight variable, ℓ is the loss, and λ is the age parameter. $f(v, \lambda)$ is called a self-paced regularizer, if

1. $f(v, \lambda)$ is convex with respect to $v \in [0, 1]$;
2. $v^*(\ell, \lambda)$ is monotonically decreasing with respect to ℓ, and it holds that $\lim_{\ell\to 0} v^*(\ell,\lambda) = 1$, $\lim_{\ell\to\infty} v^*(\ell,\lambda) = 0$;
3. $v^*(\ell, \lambda)$ is monotonically increasing with respect to λ, and it holds that $\lim_{\lambda\to\infty} v^*(\ell,\lambda) \leq 1$, $\lim_{\lambda\to 0} v^*(\ell,\lambda) = 0$;

where

$$v^*(\ell,\lambda) = \arg\min_{v\in[0,1]} v\ell + f(v,\lambda). \tag{3}$$

The three conditions in Definition 3.1 provide basic principles for constructing an SP-regularizer. Condition 2 indicates that the model inclines to select easy samples (with smaller losses) in favor of complex samples (with larger losses). Condition 3 states that when the model "age" (controlled by the age parameter λ) gets larger, it tends to incorporate more, probably complex, samples to train a "mature" model. The convexity in Condition 1 further ensures the soundness of this regularizer for optimization.

Under this definition, multiple SP-regularizers have been constructed. The following lists several typical ones, together with their closed-form solutions $v^*(\lambda, \ell)$ as defined in Definition 3.1:

$$\begin{aligned} & f^H(v,\lambda) = -\lambda v; \;\; v^*(\ell,\lambda) = \begin{cases} 1, & \text{if } \ell < \lambda \\ 0, & \text{if } \ell \geq \lambda \end{cases} \\ & f^L(v,\lambda) = \lambda(\tfrac{1}{2}v^2 - v); \;\; v^*(\ell,\lambda) = \begin{cases} -\ell/\lambda + 1, & \text{if } \ell < \lambda \\ 0, & \text{if } \ell \geq \lambda \end{cases} \\ & f^M(v,\lambda,\gamma) = \frac{\gamma^2}{v+\gamma/\lambda}; \;\; v^*(\ell,\lambda,\gamma) = \begin{cases} 1, & \text{if } \ell \leq \left(\frac{\lambda\gamma}{\lambda+\gamma}\right)^2 \\ 0, & \text{if } \ell \geq \lambda^2 \\ \gamma\left(\frac{1}{\sqrt{\ell}} - \frac{1}{\lambda}\right), & \text{otherwise.} \end{cases} \end{aligned} \tag{4}$$

The above Eq. (4) represents the hard, linear and mixture SP-regularizers proposed in [8,15], and [40], respectively. Using the AOS strategy to iteratively update **v** and **w** in the SPL regime (1) with the gradually increasing age parameter λ, a rational solution to the problem is expected to be progressively approached.

3.2. Revisit AOS algorithm for solving SPL

For convenience of notation, we briefly write $L(y_i, g(\mathbf{x}_i, \mathbf{w}))$ as $\ell_i(\mathbf{w})/\ell_i$ and $L(y, g(\mathbf{x}, \mathbf{w}))$ as $\ell(\mathbf{w})/\ell$ in the following equation. Given an SP-regularizer $f(v, \lambda)$, we can write the integrative function of $v^*(\ell, \lambda)$ calculated by Eq. (3) as:

$$F_\lambda(\ell) = \int_0^\ell v^*(l,\lambda)dl. \tag{5}$$

The following result can then be proved. The proof is listed in the appendix.

Theorem 1. *For $v^*(\ell, \lambda)$ conducted by an SP-regularizer and $F_\lambda(\ell)$ calculated by (5), given a fixed $\mathbf{w}^*$, it holds that:*

$$F_\lambda(\ell(\mathbf{w})) \leq Q_\lambda(\mathbf{w}|\mathbf{w}^*) = F_\lambda(\ell(\mathbf{w}^*)) + v^*(\ell(\mathbf{w}^*),\lambda)(\ell(\mathbf{w}) - \ell(\mathbf{w}^*)).$$

The theorem can be easily understood from the fact that: Since $v^*(\ell, \lambda)$ is monotonically decreasing in ℓ based on Condition 2 of SP-regularizer Definition 3.1, its integrative $F_\lambda(\ell)$ is concave with respect to ℓ, and thus it is easy to deduce that its Taylor series to the first order forms an upper bound of $F_\lambda(\ell)$.

Theorem 1 verifies that $Q_\lambda(\mathbf{w}|\mathbf{w}^*)$ represents a tractable surrogate for $F_\lambda(\ell(\mathbf{w}))$. Specifically, only considering the terms with respect to **w**, $Q_\lambda(\mathbf{w}|\mathbf{w}^*)$ simplifies $F_\lambda(\ell(\mathbf{w}))$, no matter how complicated its format is, as an easy weighted loss form $v^*(\ell(\mathbf{w}^*), \lambda)\ell(\mathbf{w})$. This constitutes the fundament of our new understanding of the AOS algorithm for solving SPL.

Based on Theorem 1, denote

$$Q_\lambda^{(i)}(\mathbf{w}|\mathbf{w}^*) = F_\lambda(\ell_i(\mathbf{w}^*)) + v^*(\ell_i(\mathbf{w}^*),\lambda)(\ell_i(\mathbf{w}) - \ell_i(\mathbf{w}^*),$$

and we can then easily attain the following:

$$\sum_{i=1}^n F_\lambda(\ell_i(\mathbf{w})) \leq \sum_{i=1}^n Q_\lambda^{(i)}(\mathbf{w}|\mathbf{w}^*). \tag{6}$$

Then, we can prove the equivalence between the AOS strategy for solving the SPL problem (1) and the MM algorithm for solving $\sum_{i=1}^n F_\lambda(\ell_i(\mathbf{w}))$ under surrogate function $\sum_{i=1}^n Q_\lambda^{(i)}(\mathbf{w}|\mathbf{w}^*)$ as follows:

If we denote $\mathbf{w}^k$ as the model parameters in the kth iteration of the AOS implementation on solving SPL, and then its two alternative search steps in the next iteration can be precisely explained as a standard MM scheme:

Majorization step: To obtain each $Q_\lambda^{(i)}(\mathbf{w}|\mathbf{w}^k)$, we only need to calculate $v^*(\ell_i(\mathbf{w}^k), \lambda)$ by solving the following problem under the corresponding SP-regularizer $f(v_i, \lambda)$:

$$v^*(\ell_i(\mathbf{w}^k),\lambda) = \min_{v_i \in [0,1]} v_i \ell_i(\mathbf{w}^k) + f(v_i,\lambda).$$

This exactly complies with the AOS step in updating **v** in (1) under fixed **w**.

Minimization step: We need to calculate the following:

$$\begin{aligned} \mathbf{w}^{k+1} &= \arg\min_{\mathbf{w}} \sum_{i=1}^n F_\lambda(\ell_i(\mathbf{w}^k)) + v^*(\ell_i(\mathbf{w}^k),\lambda)(\ell_i(\mathbf{w}) - \ell_i(\mathbf{w}^k)) \\ &= \arg\min_{\mathbf{w}} \sum_{i=1}^n v^*(\ell_i(\mathbf{w}^k),\lambda)\ell_i(\mathbf{w}), \end{aligned}$$

which is exactly equivalent to the AOS step in updating **w** in (1) under fixed **v**.

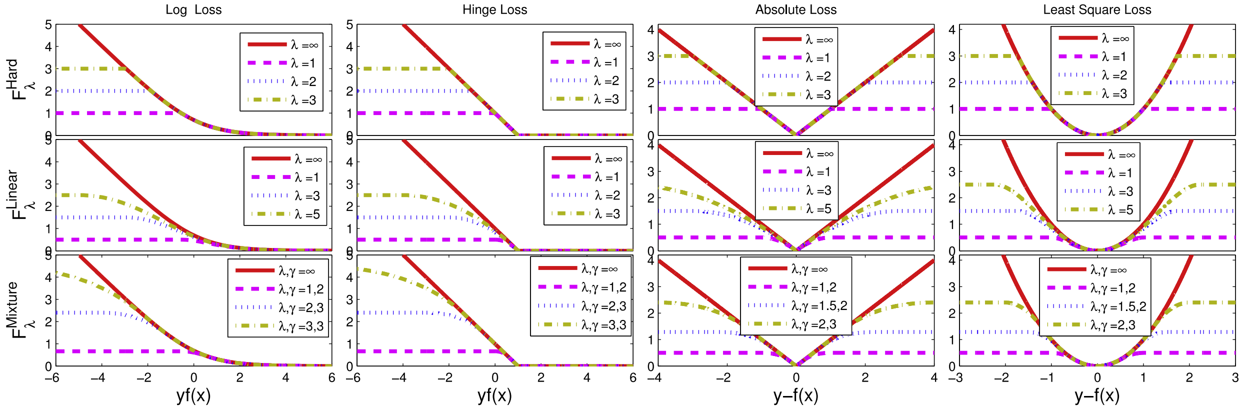

Fig. 1. Graphical illustration for implicit SPL losses $F_\lambda^{Hard}(\ell)$, $F_\lambda^{Linear}(\ell)$, and $F_{\lambda,\gamma}^{Mixture}(\ell)$ conducted by the hard, soft and mixture SP-regularizers on different loss functions, including the logistic loss, the hinge loss, the absolute loss, and the least square loss, under various pace parameters in 1-dimensional cases, respectively. Note that when $\lambda = \infty$ ($\lambda, \gamma = \infty$ in the mixture cases), the implicit SPL loss $F_\lambda(\ell)$ degenerates to the original loss ℓ.

It is then easy to see that the commonly utilized AOS strategy in previous SPL regimes is exactly the well known MM algorithm on a minimization problem of the *implicit SPL objective* $\sum_{i=1}^n F_\lambda(\ell_i(\mathbf{w}))$ with the *implicit SPL loss* $F_\lambda(\ell(\mathbf{w}))$. Various off-the-shelf theoretical results of MM can then be readily employed to explain the properties of such SPL solving strategies. For example, based on the MM theory, the lower-bounded implicit SPL objective is monotonically decreasing during MM/AOS iteration, and the convergence of the SPL algorithm can then be guaranteed.

The above theory provides a new viewpoint for understanding SPL insight. More in-depth knowledge on SPL is then expected to be extracted from it.

3.3. Revisit SPL model

Now we try to discover more interesting insights from the implicit SPL objective. To this aim, we first calculate the implicit SPL losses under hard, linear, and mixture SP-regularizers, as introduced in (4), by Eq. (5) as follows:

$$\begin{aligned}
F_\lambda^H(\ell) &= \begin{cases} \ell, & \ell < \lambda, \\ \lambda, & \ell \geq \lambda; \end{cases} \\
F_\lambda^L(\ell) &= \begin{cases} \ell - \ell^2/2\lambda, & \ell < \lambda, \\ \lambda/2, & \ell \geq \lambda; \end{cases} \\
F_{\lambda,\gamma}^M(\ell) &= \begin{cases} \ell, & \ell < \frac{1}{(1/\lambda+1/\gamma)^2}, \\ \gamma(2\sqrt{\ell} - \ell/\lambda) - \frac{\gamma}{(1/\lambda+1/\gamma)}, & \frac{1}{(1/\lambda+1/\gamma)^2} \leq \ell < \lambda^2, \\ \gamma(\lambda - \frac{1}{1/\lambda+1/\gamma}), & \ell \geq \lambda^2. \end{cases}
\end{aligned} \tag{7}$$

The configurations of these $F_\lambda(\ell)$s under different age parameters are depicted in Fig. 1 for easy observation.

Some common patterns under these implicit SPL losses can be easily observed from Fig. 1. For example, there is an evident suppressing effect of $F_\lambda(\ell)$ on large losses compared to the original loss function ℓ. When ℓ is larger than a certain threshold, $F_\lambda(\ell)$ will become a constant thereafter. This provides a rational explanation regarding why the SPL regime can perform robust in the presence of extreme outliers or heavy noise: The samples with loss values larger than the age threshold will have no influence on model training due to their zero gradients. Corresponding to the original SPL model, these large-loss samples will have zero importance weights v_i, and thus have no effect on the optimization of model parameters.

Now, we reexamine the intrinsic mechanism inside SPL implementation based on such understanding. In the beginning of SPL iteration, the age λ is small, and the implicit loss function $F_\lambda(\ell)$ has a significant suppressing effect on large losses and only allows a small number of high-confidence samples (with small loss values) into training; then, with gradually increasing λ, the suppressing effect of $F_\lambda(\ell)$ will gradually become weaker and relatively less informative samples tend to be included in the training. Through such robust guidance, increasingly faithful data knowledge tends to be incrementally learned by such learning schemes. Such a gradually changing tendency of implicit SPL loss $F_\lambda(\ell)$ can be easily understood by viewing Fig. 1.

3.4. Relationship with NCRP

An interesting observation is that the implicit SPL objective $F_\lambda(\ell)$ has a close relationship with NCRP that has been widely investigated in machine learning and statistics. For example, the hard and linear SPL objectives ($F_\lambda^H(\ell)$ and $F_\lambda^L(\ell)$) comply exactly with the forms of CNP and MCP, as defined in Eq. (2), which are imposed on ℓ by setting $\gamma = 1$, respectively, that is,

$$F_\lambda^H(\ell) = p_{1,\lambda}^{CNP}(\ell), \quad F_\lambda^L(\ell) = p_{1,\lambda}^{MCP}(\ell).$$

Furthermore, the form of $F_{\lambda,\gamma}^M(\ell)$ is almost similar to the SCAD term, both containing three phases of values, and the first and third of both are linear and constant, respectively. The only difference is in the second phase, where $F_{\lambda,\gamma}^M(\ell)$ is of linear+sqrt+constant form while SCAD is of a linear+square+constant form. It is easy to deduce that any $F_\lambda(\ell)$ led by an SP-regularizer is non-convex and has a very similar configuration with a general NCRP. Such a natural relationship provides a new viewpoint to see NCRP and facilitates more choices of NCRP formulations by virtue of $F_\lambda(\ell)$ obtained under various SP-regularizers and inspires us to borrow mature statistical tools and theoretical results regarding NCRP to further understand SPL insight in our future investigation.

This relationship is also helpful for finding self-paced formats of more typical NCRP terms. Here we also deduce the SP-regularizers of another two commonly utilized NCRP terms [39]: LOG and EXP (see Eq. (2)).

For LOG, we can construct the following SP-regularier:

$$f^{LOG}(v, \lambda, \alpha) = \frac{1}{\alpha} KL(1+\alpha\lambda, v) = \frac{1}{\alpha}\left((1+\alpha\lambda)\log\frac{1+\alpha\lambda}{v} - (1+\alpha\lambda) + v\right),$$

where $KL(x, y)$ denotes the Kullback-Leibler (KL) distance [39] between two variables. As calculated by Eq. (3), its optimal weight $v^*(\ell, \lambda, \alpha)$ is:

$$v^*(\ell, \lambda, \alpha) = \begin{cases} 1 & \ell \le \lambda, \\ \frac{1+\alpha\lambda}{1+\alpha\ell} & \ell > \lambda. \end{cases}$$

It is easy to prove that such a defined LOG SP-regularizer complies with the three conditions in Definition 3.1. By virtue of Eq. (5), we can obtain its implicit SPL loss with the form:

$$F^{LOG}(\ell, \lambda, \alpha) = \begin{cases} \ell & \ell \le \lambda, \\ p_{1+\alpha\lambda,\alpha}^{LOG}(\ell) + C_{\lambda,\alpha} & \ell > \lambda, \end{cases} \tag{8}$$

where $C_{\lambda,\alpha} = \lambda - \frac{1+\alpha\lambda}{\alpha}\log(1+\alpha\lambda)$ is a constant independent of ℓ and $p_{\gamma,\alpha}^{LOG}(\cdot)$ is defined as Eq. (2).

In addition, the EXP SP-regularizer can be constructed as:

$$f^{EXP}(v; \lambda, \alpha) = \frac{1}{\alpha} KL(v, \exp(\alpha\lambda)) = \frac{1}{\alpha}(v\log\frac{v}{\exp(\alpha\lambda)} - v + \exp(\alpha\lambda)).$$

Its optimal importance weight can be calculated by Eq. (3) as:

$$v^*(\ell, \lambda, \alpha) = \begin{cases} 1 & \ell \le \lambda, \\ \exp(-\alpha(\ell-\lambda)) & \ell > \lambda. \end{cases}$$

The corresponding implicit SPL loss is:

$$F^{EXP}(\ell, \lambda, \alpha) = \begin{cases} \ell & \ell \le \lambda, \\ p_{\alpha,\exp(-\alpha\lambda)}^{EXP}(\ell) + C_{\lambda,\alpha} & \ell > \lambda, \end{cases} \tag{9}$$

where $C_{\lambda,\alpha} = \lambda + \frac{1}{\alpha} - \frac{\exp(\alpha\lambda)}{\alpha}$ and $p_{\gamma,\alpha}^{EXP}(\cdot)$ is defined as Eq. (2).

It should be noted that LOG and EXP are different from CNP, MCP, and SCAD in their large-loss-suppressing effects. The latter suppress large losses as a constant (see Eq. (2)) while the former do this task by the gradually more slowly increasing property of logarithmic and negative exponential functions. It is easy to see that, in large loss cases, the LOG and EXP implicit SPL loss functions (8) and (9) degenerate to the conventional LOG and EXP terms, and thus possess similar robust mechanism for suppressing outliers/heavy noise. By properly adjusting age parameter λ, they are capable of adapting different noise extents in data.

3.5. Loss-prior-embedding property of SPL

An intrinsic property of SPL is to decompose the minimization of the robust but difficult-to-solve non-convex loss $F_\lambda(\ell(\mathbf{w}))$ into two easier optimization problems with respect to sample importance weights $\mathbf{v}$ (solved by the closed-form solution to an SP-regularizer) and model parameters $\mathbf{w}$ (solved by the weighted loss problem). Such decomposition not only simplifies the solving of the problem as an easy re-weighted strategy [4] but makes it feasible to embed helpful loss prior knowledge into an SPL scheme. Specifically, since the sample importance weight imposed on a sample in the SPL model reflects the extent of its loss value based on Condition 2 of Definition 3.1 for SP-regularizer (the larger the loss, the smaller the weight), the loss priors can be readily encoded as a regularization term or a constraint on $\mathbf{v}$ to deliver such knowledge in the SPL model.

In practical cases, some loss priors always exist which can be easily obtained from training data before the learning process. Here, we list some typical ones as follows:

1. Outlier prior: Some samples are significantly deviated from the main part of data sets, and thus they should show extremely large losses.
2. Spatial/temporal smoothness prior: Some spatially/temporally adjacent samples tend to show relatively similar large/small losses.
3. Sample importance order prior: A sample is pre-known to show smaller loss value (i.e., cleaner, easier, higher-confident) than others.
4. Diversity prior: Meaningful samples, which should be learned with small loss values (i.e., capable of being predicted accurately) for the learning task, should be scattered across the data range so that the learning can possibly include global-scale data knowledge.

All this loss prior knowledge can be embedded into an SPL scheme by properly encoding **v**. For example, Prior 1 can be realized by directly constraining the importance weights v_i of those outliers to be zeroes; Prior 2 can be formulated as a graph Laplacian term $\mathbf{v}^T\mathbf{L}\mathbf{v}$, where **L** is the Laplacian matrix on the data adjacent matrix [37]; Prior 3 can be encoded as the constraint $v_i > v_j$ if the ith sample is known to be more cleaner/easier than the jth sample [10]; and Prior 4 can be realized by a $-l_{2,1}$ norm [9] or $-l_{0.5,1}$ norm [36] (anti-group-sparsity) on **v**.

We review this loss-prior-embedding property from the perspective of NCRP. Currently various elegant solving strategies have been designed for solving a general or specific NCRP problem [31] so as to approach a local minimum or stationary point of the problem. In most of these strategies, however, whether the obtained solution complies with some evident loss prior knowledge has been neglected. For example, by using certain techniques, we might obtain a local minimum of the investigated non-convex problem. However, it might occur that the loss of Sample A predicted at this local minimum is larger than that of Sample B, while we have an intuitive or easily-obtained prior that A is more noisy than B. This implies that this solution, albeit being a local minimum, is an irrational one for the problem. If we transform this NCRP problem into an SPL regime based on the analysis provided in Section 3.4, and readily encode such loss priors (e.g., sample importance order loss prior) into the latent variables **v** as a regularizer or a constraint to the problem, the obtained solution is expected to more easily avoid such unreasonable local minima that violate the apriori loss priors. Such easy loss-prior-embedding capability thus tends to guide a sounder learning manner for NCRP as well as SPL, which might also provide a new viewpoint of alleviating the local-minimum-issue existing in NCRP problems.

3.6. SPCL revisit

Self-paced curriculum learning (SPCL) [10] was proposed to relate CL and SPL from the viewpoint of human learning. As opposed to “instructor-driven” and “student-driven” learning manners, as CL and SPL, respectively, through involving prior knowledge of sample importance into SPL iteration progress, SPCL is analogous to a more rational “instructor-student-collaborative” learning mode like practical human education.

The SPCL process actually illuminates the fundament of the SPL regime embedded with loss priors in the perspective of cognitive science. Specifically, the curriculum knowledge in SPCL complies with the loss priors in this study. That is, a teacher might know some curriculum information to guide the learning process of a student, e.g., some curricula are meaningless to learn (outlier prior), multiple curricula are closely related and should be learned jointly (smoothness prior), one curriculum is much more difficult than another and thus should be learned first (sample importance order prior), diverse curricula should be learned together to make the knowledge possibly comprehensive (diversity prior). Such relationship facilitates a natural interpretation for the empirical effectiveness of SPCL by the fact that the embedded loss priors (curricula) help alleviate the local-minimum-issue of the underlying non-convex optimization problems and guarantee a sound robust learning, and illustrate that an SPL scheme with properly specified loss priors corresponds more to a rational human education manner in real life compared to the pure CL or SPL strategies.

4. FCVID experiment

For a real large-scale pattern recognition task, with limited human labor and computation resources, in general we can only obtain weakly labeled samples, containing a large number of low-confidence annotations (with wrong or uncertain labels). This often leads to a noisy training dataset, which can hamper the robustness of the utilized learning algorithm. The SPL strategy is thus appropriate to be employed to alleviate this issue. Highly deviated samples (i.e., with relatively larger loss values) will be automatically screened out (i.e., with zero-valued v_i) from training and will not negatively influence the learning quality, while those samples with high-confidence annotations (i.e., with smaller loss values) tend to be selected and gradually rectify the learning performance. However, as we have analyzed, SPL intrinsically corresponds to solving a non-convex optimization problem, and useful loss priors are thus required to help avoid the problem stuck to an irrational local minimum. Through ameliorating the sample importance loss prior, we introduce an advanced group-partial-order loss prior, which is especially useful in such large-scale weakly labeled scenarios. Specifically, we can construct this loss prior in two steps: first group and rank data based on the difficulty of annotating them, and then impose a hierarchy loss structure by allowing samples located in groups in front of the ranking list (i.e., with higher-confidence labels) with larger weights than those in behind. The SPL regime is then expected to be soundly guided under such loss priors.

For verification, we use a real-world big dataset called the Fudan-columbia Video Dataset (FCVID) [11], which is by far one of the biggest annotated video sets [11] and thus is challenging for conventional concept detection techniques. Our

Table 1
Performance comparison of the proposed and baseline methods on FCVID.

Method	P@5	P@10	mAP
BatchTrain	0.782	0.763	0.469
Adaboost [6]	0.211	0.173	0.08
SPL [15]	0.793	0.754	0.414
GoogleHNM [29]	0.781	0.757	0.472
BabyLearning [21]	0.834	0.817	0.496
Ours	**0.889**	**0.874**	**0.5329**

goal is to learn detectors that can automatically recognize concepts occurring in the video content, such as people, objects, actions, etc. The FCVID contains 91,223 YouTube videos (4,232 h) from 239 categories. The class covers a wide range of concepts like activities, objects, scenes, sports, DIY, etc. Each video is manually labeled into one or more categories. As manually labeled videos are difficult to collect, we train concept detectors by only using the contexual information about the videos such as their titles, descriptions, and latent topics. For each concept, a video is automatically labeled as a positive sample if the concept name can be found in its video metedata. The generated weak labels are noisy and have both low accuracy and low recall: The labeled concepts may not be present in the video content, whereas concepts that are not in the web label may appear in the video. The ground truth labels are only used in testing to evaluate the performance. The performance is evaluated in terms of the precision of the top 5 and 10 ranked videos (P@5 and P@10) and the mean average precision (mAP) of 239 concepts.

We extract the Convolutional Neural Network (CNN), specifically AlexNet [14], features over each keyframe and create video-level feature by average pooling. The features are used across all methods. We adopt this feature since it has been widely substantiated that it can be reliably used in many computer vision tasks. We build our method on top of the CNN features and the l_2-regularized hinge loss is used as our loss function. The AOS algorithm is used to solve the optimization problem. To construct the group-partial-order loss prior knowledge into SPL on this task, we cluster the videos into a number of latent topics based on their metadata. Then, we rank these groups in the ascending order of the distances between cluster centers to the entire concept class center. Samples located in clusters in front of the ranking list should correspond to more high-confident ones and incline to have a larger weights v_i (i.e., with smaller loss values) than those ranking backwards. The hard SP-regularizer [15] was used in the SPL scheme. In implementation, we used 10% of top-ranked samples in the first iteration to get initialization and stopped increasing the model age λ after 100 iterations.

We compare our method against the following baseline methods, which cover both the classical and the state-of-the-art algorithms on the same problem. Moreover, the comparison to the baseline helps us understand the contribution of the loss prior knowledge on this problem. *BatchTrain* trains a single SVM model using all videos with noisy labels. *AdaBoost* is a classical ensemble approach that combines the sequentially trained base classifiers in a weighted fashion [6]. *Self-Paced Learning (SPL)* is the original SPL method without considering loss prior knowledge [15]. *BabyLearning* is a recently proposed method that simulates baby learning by starting with a few training samples and fine-tuning using more weakly labeled videos crawled from the search engine [21]. *GoogleHNM* is a hard negative mining method proposed by Google [29]. It utilizes hard negative mining to train the mixture of expert models according to the video's YouTube topics. The hyperparameters of all methods including the baseline methods are tuned on the same standard validation set.

Table 1 compares the precision and mAP of different methods. As we see, the SPL method with group-partial-order loss priors achieves the state-of-the-art result, which outperforms the recently proposed methods, BabyLearning [21] and GoogleHNM [29]. The average improvement over the baseline method on 239 classes are statistically significant at p-level of 0.05. As compared to classical methods, such as BatchTrain and Adaboost, the results empirically demonstrate the benefit of the robustness mechanism underlying SPL. Moreover, our improvement over the standard SPL method suggests that incorporating loss prior knowledge into learning yields a significant boost.

5. Conclusion

We have provided some new insightful understanding regarding the conventional SPL regime in this study. On one hand, we have shown that the AOS algorithm generally utilized for solving SPL exactly complies with the known MM algorithm on an implicit SPL objective, and on the other hand, we have verified that the loss function contained in this latent SPL objective precisely accords with the famous non-convex regularized penalty (NCRP). The effectiveness, especially its robustness to outliers/heavy noises, of SPL, as substantiated by previous experiences, can then be naturally explained under such understanding. We also analyzed the superiority of SPL on its easy loss-prior-embedding property, which provides a new methodology for alleviating the local-minimum-issue in general NCRP optimization problems. In our future investigation, we will attempt to employ the theories on MM and NCRP to more deeply explore the theoretical/statistical properties underlying the SPL regimes, and try to further extend the SPCL methodology and different loss priors to more application domains.

Acknowledgments

This research was supported by the NSFC projects with No. 61373114, 61661166011, 11690011, and 61603292, Macau Science and Technology Development Funds with No. 003/2016/AFJ and 973 Program of China with No. 3202013CB329404.

Appendix. Proof of Theorem 1

To prove the theorem, we need to show that

$$F_\lambda(\ell) \le F_\lambda(\ell_0) + v^*(\ell_0, \lambda)(\ell - \ell_0).$$

There are two cases required to examine.

1. $v^*(\ell, \lambda)$ is continuous with respect to ℓ.
   From Eq. (3), we know that
   $$v^*(\ell, \lambda) = F'_\lambda(\ell).$$
   By Definition 1, $v^*(\ell, \lambda) \ge 0$ when $\ell \ge 0$, and thus $F'_\lambda(\ell)$ is nondecreasing with respect to ℓ on $[0, \infty)$. Moreover, $v^*(\ell, \lambda)$ is monotonically decreasing with respect to ℓ. Therefore, we can conclude that $F_\lambda(\ell)$ is concave on $[0, \infty)$. Based on the property of a concave function, we obtain the following:
   $$F_\lambda(\ell) \le F_\lambda(\ell_0) + F'_\lambda(\ell_0)(\ell - \ell_0) = F_\lambda(\ell_0) + v^*(\ell_0, \lambda)(\ell - \ell_0).$$
2. $v^*(\ell, \lambda)$ is discontinuous with respect to ℓ.

Without loss of generality, suppose there is only one discontinuous $\tilde{\ell} \in [0, \infty)$. When $\ell, \ell_0 \in [0, \tilde{\ell})$ or $\ell, \ell_0 \in (\tilde{\ell}, \infty)$, following similar derivation, the inequality

$$F_\lambda(\ell) \le F_\lambda(\ell_0) + v^*(\ell_0, \lambda)(\ell - \ell_0)$$

holds.

Now suppose $\ell \in [0, \tilde{\ell})$ and $\ell_0 \in (\tilde{\ell}, \infty)$. Pick $\ell_1 \in [0, \tilde{\ell})$, and then:

$$F_\lambda(\ell) \le F_\lambda(\ell_1) + v^*(\ell_1, \lambda)(\ell - \ell_1),$$

and

$$F_\lambda(\tilde{\ell}) \le F_\lambda(\ell_0) + v^*(\ell_0, \lambda)(\tilde{\ell} - \ell_0).$$

Denote $v^*(\tilde{\ell}, \lambda)^- = \lim_{\ell \to \tilde{\ell}^-} v^*(\ell, \lambda)$, and let $\ell_1 \to \tilde{\ell}^-$. Since $F_\lambda(\ell)$ is continuous, we can deduce that

$$F_\lambda(\ell) \le F_\lambda(\tilde{\ell}) + v^*(\tilde{\ell}, \lambda)^-(\ell - \tilde{\ell}).$$

Therefore,

$$\begin{aligned} F_\lambda(\ell) - F_\lambda(\ell_0) &= F_\lambda(\ell) - F_\lambda(\tilde{\ell}) + F_\lambda(\tilde{\ell}) - F_\lambda(\ell_0) \\ &\le v^*(\tilde{\ell}, \lambda)^-(\ell - \tilde{\ell}) + v^*(\ell_0, \lambda)(\tilde{\ell} - \ell_0) \\ &\le v^*(\ell_0, \lambda)(\ell - \tilde{\ell}) + v^*(\ell_0, \lambda)(\tilde{\ell} - \ell_0) \\ &= v^*(\ell_0, \lambda)(\ell - \ell_0), \end{aligned}$$

where the second inequality holds because $\ell \le \tilde{\ell}$ and $v^*(\ell, \lambda) \ge 0$ is decreasing with respect to ℓ.

Similarly, if $\ell \in (\tilde{\ell}, \infty)$ and $\ell_0 \in [0, \tilde{\ell})$, the result also holds.

Now we consider the case $\ell_0 = \tilde{\ell}$. Suppose $\ell \in [0, \tilde{\ell})$ (derivation is similar for $\ell \in (\tilde{\ell}, \infty)$), and choose $\ell_1 \in [0, \tilde{\ell})$, and then:

$$F_\lambda(\ell) \le F_\lambda(\ell_1) + v^*(\ell_1, \lambda)(\ell - \ell_1).$$

Let $\ell_1 \to \tilde{\ell}^-$. Since $F_\lambda(\ell)$ is continuous,

$$F_\lambda(\ell) \le F_\lambda(\ell_0) + v^*(\lambda; \tilde{\ell})^-(\ell - \ell_0) \le F_\lambda(\ell_0) + v^*(\ell_0, \lambda)(\ell - \ell_0),$$

where the second inequality holds because $\ell \le \ell_0$ and $v^*(\ell, \lambda) \ge 0$ is decreasing with respect to ℓ.

From the above discussion, we can conclude that:

$$F_\lambda(\ell) \le F_\lambda(\ell_0) + v^*(\ell_0, \lambda)(\ell - \ell_0).$$

If we substitute ℓ and ℓ_0 with $\ell(\mathbf{w})$ and $\ell(\mathbf{w}^*)$, respectively, then Theorem 1 follows.

References

[1] S. Basu, J. Christensen, Teaching classification boundaries to humans, in: Proceeding of American Association for Artificial Intelligence, 2013, pp. 109–115.
[2] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: International Confenence on Machine Learning, 2009, pp. 41–48.
[3] P.S. Bradley, O.L. Mangasarian, Feature selection via concave minimization and support vector machines, in: International Confenence on Machine Learning, 1998, pp. 82–90.
[4] E.J. Candès, M.B. Wakin, S.P. Boyd, Enhancing sparsity by reweighted l1 minimization, J. Fourier Anal. Appl. 14 (5–6) (2008) 877–905.
[5] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Am. Stat. Assoc. 96 (456) (2001) 1348–1360.
[6] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (4) (2002) 367–378.
[7] P. Gong, C. Zhang, Z. Lu, J. Huang, J. Ye, A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems, in: International Confenence on Machine Learning, 2013, pp. 37–45.
[8] L. Jiang, D. Meng, T. Mitamura, A. Hauptmann, Easy samples first: self-paced reranking for zeroexample multimedia search, in: ACM Multimedia, 2014, pp. 547–556.
[9] L. Jiang, D.Y. Meng, S. Yu, Z.Z. Lan, S.G. Shan, A. Hauptman, Self-paced learning with diversity, in: Conference on Neural Information Processing Systems, 2014, pp. 2078–2086.
[10] L. Jiang, D.Y. Meng, Q. Zhao, S.G. Shan, A. Hauptman, Self-paced curriculum learning, in: Proceeding of American Association for Artificial Intelligence, 2015a, pp. 2694–2700.
[11] Y.G. Jiang, Z. Wu, J. Wang, X. Xue, S.F. Chang, Exploiting feature and class relationships in video categorization with regularized deep neural networks, 2015b, ArXiv preprint: 1502.07209.
[12] Y. Kang, Z. Zhang, W. Li, On the global convergence of majorization minimization algorithms for nonconvex optimization problems, 2015, ArXiv preprint: 1504.07791v2.
[13] F. Khan, X. Zhu, B. Mutlu, How do humans teach: on curriculum learning and teaching dimension, in: Conference on Neural Information Processing Systems, 2011, pp. 1449–1457.
[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.
[15] M. Kumar, B. Packer, D. Koller, Self-paced learning for latent variable models, in: Conference on Neural Information Processing Systems, 2010, pp. 1189–1197.
[16] M. Kumar, H. Turki, D. Preston, D. Koller, Learning specific-class segmentation from diverse data, in: International Conference on Computer Vision, 2011, pp. 1800–1807.
[17] K. Lange, D. Hunter, I. Yang, Optimization transfer using surrogate objective functions, J. Comput. Graph. Stat. 9 (1) (2000) 1–20.
[18] A. Lapedriza, H. Pirsiavash, Z. Bylinskii, A. Torralba, Are all training examples equally valuable?, 2013, In arXiv preprint: 1311.6510.
[19] Y. Lee, K. Grauman, Learning the easy things first: self-paced visual category discovery, in: Conference on Computer Vision and Pattern Recognition, 2011, pp. 1721–1728.
[20] J.W. Liang, L. Jiang, D.Y. Meng, A. Hauptman, Learning to detect concepts from webly-labeled video data, in: International Joint Conference on Artificial Intelligence, 2016, pp. 1746–1752.
[21] X. Liang, S. Liu, Y. Wei, L. Liu, S. Lin, L. Yan, Towards computational baby learning: a weakly-supervised approach for object detection, in: International Conference on Computer Vision, 2015, pp. 999–1007.
[22] E. Ni, C. Ling, Supervised learning with minimal effort, in: Advances in Knowledge Discovery and Data Mining, Springer, 2010, pp. 476–487.
[23] Z. Pan, C. Zhang, Relaxed sparse eigenvalue conditions for sparse estimation via non-convex regularized regression, Pattern Recognit. 48 (1) (2015) 231–243.
[24] V.I. Spitkovsky, H. Alshawi, D. Jurafsky, Baby steps: how "less is more" in unsupervised dependency parsing, in: NIPS Workshop: Grammar Induction, Representation of Language and Language Learning, 2009, pp. 1–10.
[25] J. Supančič III, D. Ramanan, Self-paced learning for long-term tracking, in: Conference on Computer Vision and Pattern Recognition, 2013, pp. 2379–2386.
[26] S. Suzumura, K. Ogawa, M. Sugiyama, I. Takeuchi, Outlier path: a homotopy algorithm for robust SVM, in: International Conference on Machine Learning, 2014, pp. 1098–1106.
[27] K. Tang, V. Ramanathan, F. Li, D. Koller, Shifting weights: adapting object detectors from image to video, in: Conference on Neural Information Processing Systems, 2012, pp. 638–646.
[28] F. Vaida, Parameter convergence for EM and MM algorithms, Stat. Sin. 15 (3) (2005) 831–840.
[29] B. Varadarajan, G. Toderici, S. Vijayanarasimhan, A. Natsev, Efficient large scale video classification, 2015, ArXiv preprint: 1505.06250.
[30] S. Wang, D. Liu, Z. Zhang, Nonconvex relaxation approaches to robust matrix recovery, in: International Joint Conference on Artificial Intelligence, 2013, pp. 1764–1770.
[31] Z. Wang, H. Liu, T. Zhang, Optimal computational and statistical rates of convergence for sparse nonconvex learning problems, Ann. Stat. 42 (2014) 2164–2201.
[32] J. Weston, A. Elisseeff, B. Schölkopf, M. Tipping, Use of the zero-norm with linear models and kernel methods, J. Mach. Learn. Res. 3 (2003) 1439–1461.
[33] S. Yu, L. Jiang, Z. Mao, e. al, CMU-informedia@TRECVID 2014 multimedia eventdetection (MED), TRECVID Video Retrieval Evaluation Workshop, 2014.
[34] C. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Stat. 38 (2) (2010) 894–942.
[35] C. Zhang, T. Zhang, A general theory of concave regularization for high-dimensional sparse estimation problems, Stat. Sci. 27 (4) (2012) 576–593.
[36] D. Zhang, D. Meng, J. Han, Co-saliency detection via a self-paced multiple-instance learning framework, in: International Conference on Computer Vision, 2015, pp. 594–602.
[37] D. Zhang, D. Meng, J. Han, Co-saliency detection via a self-paced multiple-instance learning framework, IEEE Trans. Pattern Anal. Mach. Intell. (2017), doi:10.1109/TPAMI.2016.2567393.
[38] T. Zhang, Analysis of multi-stage convex relaxation for sparse regularization, J. Mach. Learn. Res. 11 (2010) 1081–1107.
[39] Z. Zhang, B. Tu, Nonconvex penalization using laplace exponents and concave conjugates, in: Conference on Neural Information Processing Systems, 2012, pp. 611–619.
[40] Q. Zhao, D.Y. Meng, L. Jiang, Q. Xie, Z.B. Xu, A. Hauptman, Self-paced learning for matrix factorization, in: Proceeding of American Association for Artificial Intelligence, 2015, pp. 3196–3202.