Supplementary Materials for Self-Paced Learning: An Implicit Regularization Perspective

Yanbo Fan^{1,4} **, Ran He**^{1,2,3,4*} **, Jian Liang**^{1,2,4} **, Baogang Hu**^{1,4}

^{1}National Laboratory of Pattern Recognition, CASIA
^{2}Center for Research on Intelligent Perception and Computing, CASIA
^{3}Center for Excellence in Brain Science and Intelligence Technology, CAS
^{4}University of Chinese Academy of Sciences (UCAS)
{yanbo.fan, rhe, jian.liang, hubg}@nlpr.ia.ac.cn

In this supplementary material, we fist present an existing definition of self-paced regularizer along with their examples. Then we demonstrate the equivalency of Definition 1 and Definition 2 given in the main body of our paper.

Self-Paced Regularizer

Similar definitions of self-paced regularizer (or self-paced function) have been proposed in (Jiang et al. 2015; Zhao et al. 2015; Jiang et al. 2014a). The definition in (Zhao et al. 2015) is shown below.

Definition 3 ***(Self-Paced Regularizer)*** *(Zhao et al. 2015): Suppose that v is a weight variable, ℓ is the loss, and λ is the learning pace parameter. $g(\lambda, v)$ is called self-paced rgularizer, if*

1. $g(\lambda, v)$ is convex with respect to $v \in [0, 1]$;

2. $v^*(\lambda, \ell)$ is monotonically decreasing w.r.t. ℓ, and it holds that $\lim_{\ell \to 0} v^*(\lambda, \ell) = 1$, $\lim_{\ell \to \infty} v^*(\lambda, \ell) = 0$;

3. $v^*(\lambda, \ell)$ is monotonically increasing w.r.t. λ, and it holds that $\lim_{\lambda \to 0} v^*(\lambda, \ell) = 0$, $\lim_{\lambda \to \infty} v^*(\lambda, \ell) \leq 1$;

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

Table 1 tabulates some examples of self-paced regularizers $g(\lambda, v)$ and their corresponding $v^*(\lambda, \ell)$. We modify their original expressions for better comparison. It is still nontrivial to design self-paced regularizers or analyze their properties according to Definition 3. Besides, though shown to be effective in many applications experimentally, the underlying working mechanism of SPL is still unclear.

One attempt about the underlying working mechanism of SPL is (Meng and Zhao 2015). Starting from SPL regularizers and their minimizer functions, they show that the ASS method used for SPL accords with the *majorization minimization* (Vaida 2005) algorithm implemented on a latent SPL objective, and deduced the latent objective of hard, linear and mixture regulraizers. In contrast, we start from a latent loss function $\phi(\lambda, \ell)$ directly and propose self-paced implicit regularizer based on the convex conjugacy theory. We establish the relations between robust loss function $\phi(\lambda, \ell)$, self-paced implicit regularizer $\psi(\lambda, v)$ and minimizer function $\sigma(\lambda, \ell)$. According to Definition 1, $\psi(\lambda, v)$ and $\sigma(\lambda, \ell)$ are derived from latent loss function $\phi(\lambda, \ell)$, thus we can analyze their properties based on the development of $\phi(\lambda, \ell)$ (many loss functions have be widely studied in related areas). We further demonstrate that for SPL with the proposed implicit regularizer, its learning procedure actually associates with certain latent robust loss functions. Thus we can provide some inspirations for the working mechanism of SPL (e.g. its robustness to outliers and heavy noise). Moreover, by establishing the relations between $\phi(\lambda, \ell)$ and $\psi(\lambda, v)$, we can develop new SPL regularizers based on the development of robust loss functions. Specifically, we analyze the relations between self-paced implicit regularizer and HQ optimization. Many robust loss functions and their minimizer functions have been developed and widely used in HQ optimization, and they can be adjusted for self-paced implicit regularizers (some examples are given in Table 1 in main body).

Definition 1 and Definition 2

To show the equivalency of Definition 1 and Definition 2 in the main body, we first give the following proposition about Definition 2.

Proposition 2 *For any fixed λ, if $\phi(\lambda, t)$ in Definition 2 further satisfies the conditions referred in (Nikolova and Chan 2007), its minimizer function $\sigma(\lambda, t)$ is uniquely determined by $\phi(\lambda, t)$ and the analytic form of $\psi(\lambda, v)$ can be even unknown during the optimization.*

Proof. The proof sketch is similar to that in (Nikolova and Chan 2007). For ease of representation, we omit λ and use $\phi(t)$, $\psi(v)$ and $\sigma(t)$ for short. Some fundamental assumptions about $\phi(t)$ are: **H1:** $\phi : R_+ \to R$ is increasing with $\phi \not\equiv 0$ and $\phi(0) = 0$; **H2:** $t \to \phi(\sqrt{t})$ is concave; **H3:** $\phi(t)$ is C^1; **H4:** $\lim_{t \to \infty} \phi(t)/t^2 = 0$.

Put $\theta(t) = -\phi(\sqrt{t})$, then θ is convex by H2. Its convex conjugate is $\theta^*(v) = \sup_{t \geq 0} \{vt - \theta(t)\}$. By the Fenchel-Moreau theorem (Rockafellar 2015), the convex conjugate of θ^* is θ, that is $\theta(t) = (\theta^*)^*(t) = \sup_{v \leq 0} \{vt - \theta^*(v)\} = -\inf_{v \geq 0} \{vt + \theta^*(-v)\}$. Define $\psi(v) = \theta^*(-\frac{1}{2}v)$, we have

$$\psi(v) = \sup_{t \geq 0} \{-\frac{1}{2}vt - \theta(t)\} = \sup_{t \geq 0} \{-\frac{1}{2}vt^2 + \phi(t)\}. \quad (1)$$

$$\phi(t) = -\theta(t^2) = \inf_{v \geq 0} \{\frac{1}{2}vt^2 + \psi(v)\}. \quad (2)$$

Then the problem becomes how to achieve the supremum in (1) jointly with the infimum in (2). For any $\hat{v} > 0$, de-

Table 1: Recently Proposed Self-paced Regularizers $g(\lambda, v)$ and their Corresponding $v^*(\lambda, \ell)$

	$g(\lambda, v)$	$v^*(\lambda, \ell)$
(Kumar, Packer, and Koller 2010)	$-\lambda \sum_{i=1}^{n} v_i,\ \lambda > 0$	$\begin{cases} 1, & \ell_i < \lambda \\ 0, & otherwise \end{cases}$
(Jiang et al. 2014a; 2015)	$\frac{1}{2}\lambda \sum_{i=1}^{n}(v_i^2 - 2v_i),\ \lambda > 0$	$\begin{cases} 1 - \frac{1}{\lambda}\ell_i, & \ell_i < \lambda \\ 0, & otherwise \end{cases}$
(Jiang et al. 2014a; 2015)	$\sum_{i=1}^{n}(\zeta v_i - \frac{\zeta^{v_i}}{\log \zeta}),$ $\zeta = 1 - \lambda, 0 < \lambda < 1$	$\begin{cases} \frac{1}{\log \zeta}\log(\ell_i + \zeta), & \ell_i < \lambda \\ 0, & otherwise \end{cases}$
(Jiang et al. 2014a; 2015)	$-\zeta \sum_{i=1}^{n} \log(v_i + \frac{1}{\lambda_1}\zeta),$ $\zeta = \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2}, \lambda_1 > \lambda_2 > 0$	$\begin{cases} 1, & \ell_i \leq \lambda_2 \\ \frac{(\lambda_1 - \ell_i)\zeta}{\ell_i \lambda_1}, & \lambda_2 < \ell_i < \lambda_1 \\ 0, & \ell_i \geq \lambda_1 \end{cases}$
(Jiang et al. 2014b)	$-\lambda \sum_{i=1}^{n} v_i - \gamma \Vert\mathbf{v}\Vert_{2,1},\ \lambda > 0,\ \gamma > 0$	$\begin{cases} 1, & \ell_i \leq \lambda + \gamma \frac{1}{\sqrt{i} - \sqrt{i-1}} \\ 0, & otherwise \end{cases}$
(Xu, Tao, and Xu 2015)	$\sum_{i=1}^{n} \ln(1 + e^{-\lambda} - v_i)^{(1 + e^{-\lambda} - v_i)}$ $+ \ln(v_i)^{v_i} - \lambda v_i,\ \lambda > 0$	$\frac{1 + e^{-\lambda}}{1 + e^{\ell_i - \lambda}}$
(Zhao et al. 2015)	$\sum_{i=1}^{n} \frac{\lambda \gamma^2}{\lambda v_i + \gamma},\ \lambda > 0,\ \gamma > 0$	$\begin{cases} 1, & \ell_i \leq (\frac{\lambda\gamma}{\lambda + \gamma})^2 \\ 0, & \ell_i \geq \lambda^2 \\ \gamma(\frac{1}{\sqrt{\ell_i}} - \frac{1}{\lambda}), & otherwise \end{cases}$
(Zhang et al. 2015)	$-\lambda \sum_{k=1}^{K} \sum_{i=1}^{n_k} v_i^k - \gamma \sum_{k=1}^{K} \sqrt{\sum_{i=1}^{n_k} v_i^k},$ $\lambda > 0,\ \gamma > 0$	$\begin{cases} 1, & \ell_i^k < \lambda + \frac{\gamma}{2\sqrt{i}} \\ \frac{((\frac{\gamma}{2(\ell_i^k - \lambda)})^2 - (i-1)}{m}, & otherwise \end{cases}$

fine $f_{\hat{v}} : R_+ \to R$ by $f_{\hat{v}}(t) = \frac{1}{2}\hat{v}t + \theta(t)$, then we have $\psi(\hat{v}) = -\inf_{t \geq 0} f_{\hat{v}}(t)$ from (1). According to H1-H4, $f_{\hat{v}}$ is convex with $f_{\hat{v}}(0) = 0$ and $\lim_{t \to +\infty} f_{\hat{v}}(t) = +\infty$. Thus $f_{\hat{v}}$ can reach its unique minimum at a $\hat{t} \geq 0$, and $\psi(\hat{v}) = -\frac{1}{2}\hat{v}\hat{t}^2 + \phi(\hat{t})$ from (1). Hence equivalently the infimum in (2) is reached at $\hat{v}$ as $\phi(\hat{t}) = \frac{1}{2}\hat{v}\hat{t}^2 + \psi(\hat{v})$. Then we have $\hat{v} = \sigma(t) = -2\theta'(t^2) = \phi'(t)/t$. Thus the optimal v is uniquely determined by the minimizer function $\sigma(t)$ that is only related to $\phi(t)$. The analytic form of the dual potential function $\psi(v)$ could be unknown during the optimization. The proof is then completed.

Denote $\ell_i = L(y_i, f(\mathbf{x}_i, \mathbf{w}))$ and rewrite model (8) in the main body as

$$\min_{\mathbf{w}, \mathbf{v}} \mathbb{E}(\mathbf{w}, \mathbf{v}; \lambda) = \sum_{i=1}^{n} v_i (\sqrt{\ell_i})^2 + \psi(\lambda, v_i). \tag{3}$$

If we adopt $\psi(\lambda, v_i)$ with an implicit regularizer given in Definition 2 and use $v_i^* = \frac{1}{2}\sigma(\lambda, \sqrt{\ell_i})$, where $\sigma(\lambda, \sqrt{\ell_i})$ is the minimizer function in Definition 2, model (3) is optimizing a latent loss function $\sum_{i=1}^{n} \phi(\lambda, \sqrt{\ell_i})$ equivalently.

Now we demonstrate the equivalency of Definition 1 and Definition 2 in the main body. For easy of representation, we omit λ, and use $\{\phi_1(t), \psi_1(v), \sigma_1(t)\}$ and $\{\phi_2(t), \psi_2(v), \sigma_2(t)\}$ to refer to the functions in Definition 1 and Definition 2, respectively. Considering a simplified model

$$\min_{\mathbf{w}, v} vL(y, f(\mathbf{x}, \mathbf{w})) + \psi(v). \tag{4}$$

Denote $\ell = L(y, f(\mathbf{x}, \mathbf{w}))$. We show that for a same implicit regularizer $\psi(v) = \psi_1(v) = \psi_2(v)$, the optimal v^* and the latent loss function of model (4) derived from Definition 1 and Definition 2 are the same. Specifically, let $\psi_1(v) = \psi_2(v) = \sup_{t \geq 0} \{-vt + \phi_1(t)\}$ (where $\phi_1(t)$ satisfies conditions H1-H3 of Proposition 1 in the main body), it is easy to verify that its corresponding latent loss function is $\phi_1(\ell)$ and optimal $v^* = \sigma_1(\ell) = \phi_1'(\ell)$ according to Definition 1 and Proposition 1. Meanwhile, we have $\psi_2(v) = \sup_{t \geq 0} \{-vt + \phi_1(t)\} = \sup_{t \geq 0} \{-vt^2 + \phi_2(t)\}$, where $\phi_2(t) = \phi_1(t^2)$. Then model (4) can be considered to optimize a latent loss function $\phi_2(\sqrt{\ell}) = \phi_1(\ell)$ and the optimal $v^* = \frac{1}{2}\sigma_2(\sqrt{\ell}) = \phi_1'(\ell)$ according to Definition 2 and Proposition 2. Thus we show the equivalency of Definition 1 and Definition 2.

References

Jiang, L.; Meng, D.; Mitamura, T.; and Hauptmann, A. G. 2014a. Easy samples first: Self-paced reranking for zero-example multimedia search. In *ACM MM*, 547–556.

Jiang, L.; Meng, D.; Yu, S.-I.; Lan, Z.; Shan, S.; and Hauptmann, A. 2014b. Self-paced learning with diversity. In *NIPS*, 2078–2086.

Jiang, L.; Meng, D.; Zhao, Q.; Shan, S.; and Hauptmann, A. G. 2015. Self-paced curriculum learning. In *AAAI*, 2694–2700.

Kumar, M. P.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. In *NIPS*, 1189–1197.

Meng, D., and Zhao, Q. 2015. What objective does self-paced learning indeed optimize? *arXiv preprint arXiv:1511.06049*.

Nikolova, M., and Chan, R. H. 2007. The equivalence of half-quadratic minimization and the gradient linearization iteration. *TIP* 16(6):1623–1627.

Rockafellar, R. T. 2015. *Convex analysis*. Princeton university press.

Vaida, F. 2005. Parameter convergence for em and mm algorithms. *Statistica Sinica* 831–840.

Xu, C.; Tao, D.; and Xu, C. 2015. Multi-view self-paced learning for clustering. In *IJCAI*, 3974–3980.

Zhang, D.; Meng, D.; Li, C.; Jiang, L.; Zhao, Q.; and Han, J. 2015. A self-paced multiple-instance learning framework for co-saliency detection. In *ICCV*, 594–602.

Zhao, Q.; Meng, D.; Jiang, L.; Xie, Q.; Xu, Z.; and Hauptmann, A. G. 2015. Self-paced learning for matrix factorization. In *AAAI*, 3196–3202.