Self-Paced Robust Learning for Leveraging Clean Labels in Noisy Data: Supplementary Document

Xuchao Zhang[†], Xian Wu[‡], Fanglan Chen[†], Liang Zhao[§], Chang-Tien Lu[†]
[†]Discovery Analytics Center, Virginia Tech, Falls Church, VA, USA
[‡]University of Notre Dame, Notre Dame, IN, USA
[§]George Mason University, Fairfax, VA, USA
[†]{xuczhang, fanglanc, ctlu}@vt.edu, [‡]xwu9@nd.edu, [§]lzhao9@gmu.edu

A Related Work

The work related to this paper is summarized in three categories below.

A.1 Self-Paced Learning

In recent years, self-paced learning (Kumar *et al.* 2010) has received widespread attention for various applications in machine learning, such as image classification (Jiang *et al.* 2015), event detection (Jiang *et al.* 2014a) and object tracking (Supancic and Ramanan 2013; Zhang *et al.* 2016). Inspired by the learning processes used by humans and animals (Bengio *et al.* 2009), self-paced learning (*SPL*) (Kumar *et al.* 2010) considers training data in a meaningful order, from easy to hard, to facilitate the learning process. Unlike standard curriculum learning (Bengio *et al.* 2009), which learns the data in a predefined curriculum design based on prior knowledge, *SPL* learns the training data in an order that is dynamically determined by feedback from the learning process itself, which means it can be more extensively utilized in practice. Furthermore, a wide assortment of *SPL*-based methods (Pi *et al.* 2016; Ma *et al.* 2017a) have been developed, including self-paced curriculum learning (Jiang *et al.* 2015), self-paced learning with diversity (Jiang *et al.* 2014b), multi-view (Xu *et al.* 2015) and multi-task (Li *et al.* 2017a; Keerthiram Murugesan 2017) self-paced learning. In addition, several researchers have conducted theoretical analyses of self-paced learning. Meng et al. (Meng *et al.* 2015) provides a theoretical analysis of the robustness of *SPL*, revealing that the implicit objective function of *SPL* has a similar configuration to a non-convex regularized penalty. Such regularization restricts the contributions of noisy examples to the objective, and thus enhances the learning robustness. Ma et al. (Ma *et al.* 2017b) proved that the learning process of *SPL* always converges to critical points of its implicit objective under mild conditions. However, none of the existing self-paced learning approaches can be applied to our problem of leveraging clean labels in noisy data.

A.2 Robust Learning

A large body of literature on the robust learning problem has been established over the last few decades. Most of the studies aim to directly learn from noisy labels and focus on noise-robust algorithms. For instance, Chen et al. (Chen *et al.* 2013) proposed a robust algorithm based on trimmed inner product. McWilliams et al. (McWilliams *et al.* 2014) proposed a sub-sampling algorithm for large-scale corrupted linear regression. Bhatia et al. (Bhatia *et al.* 2015) and Zhang et al. (Zhang *et al.* 2017b) proposed hard-thresholding based methods with strong guarantees of coefficient recovery under a mild assumption on datasets. Another group of methods focuses on removing or correcting mislabeled data. For example, some work utilized heavy-tailed distributions (Zhu *et al.* 2013) such as Student t-distribution and Poisson distribution, to model the mislabeled data, while others detected these outliers based on Gaussian distribution (Solberg and Lahti 2005; Hodge and Austin 2004) under the assumption that outliers have a small probability of occurrence in the population. Some methods do not assume any prior knowledge on the data distribution based on kernel functions (Latecki *et al.* 2007; Roth 2006). These approaches utilize kernel functions to approximate the actual density distribution and declare the instances lying in the low probability area of the kernel density function as outliers. However, all these approaches typically jointly learn the clean and noisy data together, but cannot fully leverage the information contained in the clean set.

A.3 Weakly-Supervised Learning

Recently, some work in weakly-supervised learning (Medlock and Briscoe 2007) utilized additional clean labels in learning a noisy dataset. For instance, Azadi et al. (Zhang *et al.* 2017a) proposed an auxiliary image regularization to train a deep convolutional neural network in noisy labeled image data, in which a limited number of training examples are supplied with clean labels. In order to classify images from weakly labeled data, Li et al. (Li *et al.* 2017b) used not only a small clean dataset, but some other "side" information of label relations in a knowledge graph. Veit et al. (Veit *et al.* 2017) used millions of images with noisy annotations in conjunction with a small set of cleanly-annotated images to learn effective image representations, while Jiang et al. (Jiang *et al.* 2017) designed a curriculum paradigm to learn the instance weights in corrupted labels to prevent deep convolutional neural networks from overfitting. Compared to existing work utilizing the clean dataset, our methods are different in two ways. First, we consider the learning process from clean to noisy data in a self-paced manner, which hedges the risk of training corrupted data samples. Moreover, our model learns the instance weight determined by the feedback of the learner itself without using any additional prior knowledge, which means our methods can be applied to more general problems in practice.

B Proof of Theorem 1

Proof. Before we prove the convergence of Algorithm 1, we will first show that the value of objective function $\mathcal{J}$ is monotonically decreased. Objective function $\mathcal{J}$ has the following property:

$$
\begin{aligned}
&\mathcal{J}(\boldsymbol{w}^{t+1}, \boldsymbol{v}^{t+1}; \lambda^{t+1}) \\
&\quad\overset{(a)}{\leq} \sum_{i=1}^{k} \mathcal{L}\big(y_i, f(\boldsymbol{x}_i, \boldsymbol{w}^{t+1})\big) + \sum_{i=k+1}^{n} v_i^{t+1} \mathcal{L}\big(y_i, f(\boldsymbol{x}_i, \boldsymbol{w}^{t+1})\big) \\
&\qquad + \|\boldsymbol{w}^{t+1}\|_2^2 + \theta \|\boldsymbol{w}^{t+1} - \tilde{\boldsymbol{w}}\|_2^2 - \lambda^t \sum_{i=k+1}^{n} v_i^{t+1}
\end{aligned}
$$

The inequality follows from the fact that λ increases monotonically so that $\lambda^{t+1} \geq \lambda^t$ and $v_i^t \geq 0$. The optimization step in Line 7 in Algorithm 1 guarantees the following property:

$$
\begin{aligned}
&\sum_{i=k+1}^{n} v_i^{t+1} \mathcal{L}\big(y_i, f(\boldsymbol{x}_i, \boldsymbol{w}^{t+1})\big) - \lambda^t \sum_{i=k+1}^{n} v_i^{t+1} \\
&\qquad\qquad \leq \sum_{i=k+1}^{n} v_i^{t} \mathcal{L}\big(y_i, f(\boldsymbol{x}_i, \boldsymbol{w}^{t+1})\big) - \lambda^t \sum_{i=k+1}^{n} v_i^{t}
\end{aligned}
$$

Therefore, we have the following property:

$$
\begin{aligned}
&\mathcal{J}(\boldsymbol{w}^{t+1}, \boldsymbol{v}^{t+1}; \lambda^{t+1}) \\
&\quad\leq \sum_{i=1}^{k} \mathcal{L}\big(y_i, f(\boldsymbol{x}_i, \boldsymbol{w}^{t+1})\big) + \sum_{i=k+1}^{n} v_i^{t} \mathcal{L}\big(y_i, f(\boldsymbol{x}_i, \boldsymbol{w}^{t+1})\big) \\
&\qquad + \|\boldsymbol{w}^{t+1}\|_2^2 + \theta \|\boldsymbol{w}^{t+1} - \tilde{\boldsymbol{w}}\|_2^2 - \lambda^t \sum_{i=k+1}^{n} v_i^{t}.
\end{aligned}
$$

Similarly, the following inequality is satisfied since the optimizations step in Line 5 in Algorithm 1.

$$
\begin{aligned}
&\mathcal{J}(\boldsymbol{w}^{t+1}, \boldsymbol{v}^{t+1}; \lambda^{t+1}) \\
&\quad\leq \sum_{i=1}^{k} \mathcal{L}\big(y_i, f(\boldsymbol{x}_i, \boldsymbol{w}^{t})\big) + \sum_{i=k+1}^{n} v_i^{t} \mathcal{L}\big(y_i, f(\boldsymbol{x}_i, \boldsymbol{w}^{t})\big) \\
&\qquad + \|\boldsymbol{w}^{t}\|_2^2 + \theta \|\boldsymbol{w}^{t} - \tilde{\boldsymbol{w}}\|_2^2 - \lambda^t \sum_{i=k+1}^{n} v_i^{t} \\
&\quad= \mathcal{J}(\boldsymbol{w}^{t}, \boldsymbol{v}^{t}; \lambda^{t})
\end{aligned}
$$

Since the objective function is monotonically decreased and it has a lower bound according to Lemma 1, we have $\|\mathcal{J}^{t+1} - \mathcal{J}^t\|_2 < \varepsilon$ for $\forall \varepsilon > 0$. □

C Analysis of Parameter λ

Figure 1 shows the impact of parameter λ on both the robust regression and classification tasks. In Figure 1(a), the blue line depicts the relationship between parameter λ and the coefficient recovery error. As λ increases, the recovery error continues to decrease until it reaches a critical point, after which it increases. These results indicate that the training process will keep improving the model until parameter λ becomes so large that some corrupted samples are incorporated into the training data. The red line shows the value of the objective function $\mathcal{J}$ in terms of different values of parameter λ, leading us to conclude: 1) The value of objective function $\mathcal{J}$ monotonically decreases as λ increases. 2) The objective function $\mathcal{J}$ decreases much faster once λ reaches a critical point, following the same pattern as the recovery error shown in the blue line. In Figure 1(b), the blue line shows the values of the F1 score for the binary classification task. When λ increases, the F1 score increases quickly until it reaches a peak point. After that point, the score decreases because more corrupted data is incorporated into the training set. The red line shows the size of the training set. We can conclude: 1) When parameter λ increases, the size of the training set continuously increases until it reaches its maximum value. 2) When all the data is included into the training set, the F1 score also remains stable.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 774–782. JMLR Workshop and Conference Proceedings, May 2013.

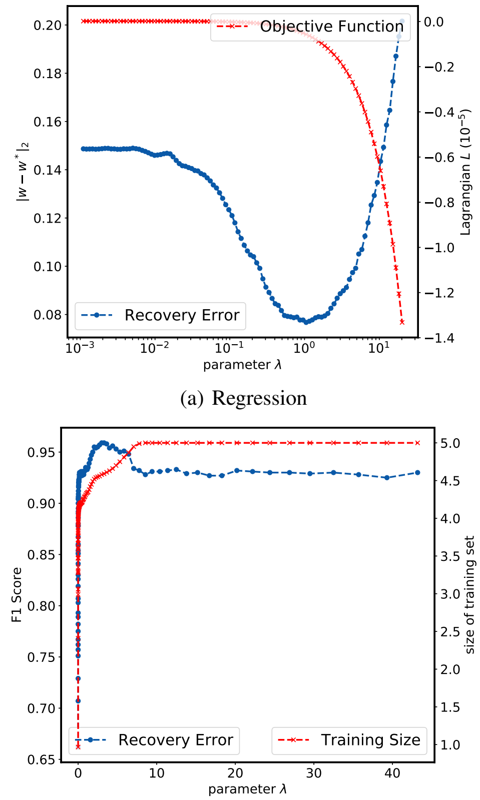
Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

Lu Jiang, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 547–556. ACM, 2014.

Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann. Self-paced learning with diversity. In *Advances in Neural Information Processing Systems*, pages 2078–2086, 2014.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. Self-paced curriculum learning, 2015.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Regularizing very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.



(a) Regression

(b) Classification

Figure 1: Impact of Parameter λ on Robust Regression and Classification Tasks

Jaime Carbonell Keerthiram Murugesan. Self-paced multitask learning with shared knowledge. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2522–2528, 2017.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

Longin Jan Latecki, Aleksandar Lazarevic, and Dragoljub Pokrajac. Outlier detection with kernel density functions. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 61–75. Springer, 2007.

Changsheng Li, Junchi Yan, Fan Wei, Weishan Dong, Qingshan Liu, and Hongyuan Zha. Self-paced multi-task learning. In *AAAI*, pages 2175–2181, 2017.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Jia Li. Learning from noisy labels with distillation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1928–1936, 2017.

Fan Ma, Deyu Meng, Qi Xie, Zina Li, and Xuanyi Dong. Self-paced co-training. In *International Conference on Machine Learning*, pages 2275–2284, 2017.

Zilu Ma, Shiqi Liu, and Deyu Meng. On convergence property of implicit self-paced objective. *arXiv preprint arXiv:1703.09923*, 2017.

Brian McWilliams, Gabriel Krummenacher, Mario Lucic, and Joachim M Buhmann. Fast and robust least squares estimation in corrupted linear models. In *Advances in Neural Information Processing Systems*, pages 415–423, 2014.

Ben Medlock and Ted Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 992–999, 2007.

Deyu Meng, Qian Zhao, and Lu Jiang. What objective does self-paced learning indeed optimize? *arXiv preprint arXiv:1511.06049*, 2015.

Te Pi, Xi Li, Zhongfei Zhang, Deyu Meng, Fei Wu, Jun Xiao, and Yueting Zhuang. Self-paced boost learning for classification. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1932–1938. AAAI Press, 2016.

Volker Roth. Kernel fisher discriminants for outlier detection. *Neural computation*, 18(4):942–960, 2006.

Helge Erik Solberg and Ari Lahti. Detection of outliers in reference distributions: performance of horn's algorithm. *Clinical chemistry*, 51(12):2326–2332, 2005.

James S Supancic and Deva Ramanan. Self-paced learning for long-term tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2379–2386, 2013.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 839–847, 2017.

Chang Xu, Dacheng Tao, and Chao Xu. Multi-view self-paced learning for clustering. In *IJCAI*, pages 3974–3980, 2015.

Dingwen Zhang, Deyu Meng, Long Zhao, and Junwei Han. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 3538–3544. AAAI Press, 2016.

X. Zhang, L. Zhao, A. P. Boedihardjo, and C. Lu. Online and distributed robust regressions under adversarial data corruption. In *2017 IEEE International Conference on Data Mining (ICDM)*, volume 00, pages 625–634, Nov. 2017.

Xuchao Zhang, Liang Zhao, Arnold P. Boedihardjo, and Chang-Tien Lu. Robust regression via heuristic hard thresholding. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, IJCAI'17. AAAI Press, 2017.

Hao Zhu, Henry Leung, and Zhongshi He. A variational bayesian approach to robust sensor fusion based on student-t distribution. *Information Sciences*, 221(Supplement C):201 – 214, 2013.