Advancing Non-Convex and Constrained Learning: Challenges and Opportunities

Tianbao Yang (The University of Iowa; tianbao-yang@uiowa.edu)
DOI: 10.1145/3362077.3362085

Introduction

As data gets more complex and applications of machine learning (ML) algorithms for decision-making broaden and diversify, traditional ML methods by minimizing an unconstrained or simply constrained convex objective are becoming increasingly unsatisfactory. To address this new challenge, recent ML research has sparked a *paradigm shift* in learning predictive models into non-convex learning and heavily constrained learning. Non-Convex Learning (NCL) refers to a family of learning methods that involve optimizing non-convex objectives. Heavily Constrained Learning (HCL) refers to a family of learning methods that involve constraints that are much more complicated than a simple norm constraint (e.g., data-dependent functional constraints, non-convex constraints), as in conventional learning. This paradigm shift has already created many promising outcomes: (i) non-convex deep learning has brought breakthroughs for learning representations from *large-scale structured data* (e.g., images, speech) (LeCun, Bengio, & Hinton, 2015; Krizhevsky, Sutskever, & Hinton, 2012; Amodei et al., 2016; Deng & Liu, 2018); (ii) non-convex regularizers (e.g., for enforcing sparsity or low-rank) could be more effective than their convex counterparts for learning *high-dimensional structured models* (C.-H. Zhang & Zhang, 2012; J. Fan & Li, 2001; C.-H. Zhang, 2010; T. Zhang, 2010); (iii) constrained learning is being used to learn predictive models that satisfy various constraints to *respect social norms* (e.g., fairness) (B. E. Woodworth, Gunasekar, Ohannessian, & Srebro, 2017; Hardt, Price, Srebro, et al., 2016; Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017; A. Agarwal, Beygelzimer, Dudík, Langford, & Wallach, 2018), to *improve the interpretability* (Gupta et al., 2016; Canini, Cotter, Gupta, Fard, & Pfeifer, 2016; You, Ding, Canini, Pfeifer, & Gupta, 2017), to *enhance the robustness* (Globerson & Roweis, 2006a; Sra, Nowozin, & Wright, 2011; T. Yang, Mahdavi, Jin, Zhang, & Zhou, 2012), etc. In spite of great promises brought by these new learning paradigms, they also bring emerging challenges to the design of computationally efficient algorithms for *big data* and the analysis of their statistical properties.

Non-Convex Learning

In this section, we describe some recent advances in non-convex learning with mentioning some of our recent related results. We will also describe their limitations and point out future directions. This article will focus on studies that are concerned with algorithm design and analysis for solving NCL and HCL problems instead of papers that are purely application-driven. It is notable that the references are not exhaustive due to a large volume of related works.

Non-Convex Minimization and Deep Learning. Deep learning can be formulated as the following non-convex minimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) := \mathrm{E}_{\mathbf{z}}[f(\mathbf{w};\mathbf{z})], \quad (1)$$

where $\mathbf{z}$ denotes a random data, and $\mathbf{w}$ denotes the parameters of the neural network to be learned, and $f(\mathbf{w};\mathbf{z})$ denotes the loss function. Due to the success of deep learning in many areas, this problem has attracted much attention from the community of mathematical programming and machine learning. Research has been conducted in the following directions.

- **Convergence to stationary points.** For general non-convex problems, it is NP-hard to find a global minimizer (Hillar & Lim, 2013). Hence, many studies have focused on finding stationary points of (1) (Nesterov & Polyak, 2006; N. Agarwal, Allen Zhu, Bullins, Hazan, & Ma, 2017; Carmon, Duchi, Hinder, & Sidford, 2016; P. Xu, Roosta-Khorasani, & Mahoney, 2017; Cartis, Gould, & Toint, 2011b, 2011a; Royer & Wright, 2017; M. Liu & Yang, 2017b, 2017a; Allen-Zhu, 2017; Kohler & Lucchi, 2017; Reddi et al., 2017). Typically, two types of stationary points are considered, namely first-order stationary point and second-order stationary point. A point $\mathbf{w}_*$ is called a first-order stationary point if it satisfies $\nabla F(\mathbf{w}_*) = 0$. A point $\mathbf{w}_*$ is called a second-order stationary if it satisfies $\nabla F(\mathbf{w}_*) = 0$ and $\nabla^2 F(\mathbf{w}_*) \succeq 0$. These studies concentrate on the complexity analysis of first or second-order methods. Many first-order methods (e.g., stochastic gradient descent (SGD)) have been proved to converge to first-order stationary points with a polynomial time complexity. In our study (Yan, Yang, Li, Lin, & Yang, 2018), we presented the first theoretical result showing that the commonly used stochastic heavy-ball (SHB) method and stochastic Nesterov's accelerated gradient (SNAG) method for deep learning converge to first-order stationary points, and also presented a unified framework that subsumes SGD, SHB and SNAG by varying a single parameter. Moreover, in (Y. Xu, Rong, & Yang, 2018) we presented a unified framework that can promote first-order algorithms to enjoy convergence to a second-order stationary point by using our proposed first-order negative curvature finding procedure named NEON.

- **Convergence to global minimizers.** Recently, several works have proved gradient descent or stochastic gradient descent can find global minimizers of minimizing an over-parameterized deep neural network under some mild conditions of input data (Allen-Zhu, Li, & Song, 2018; Arora, Cohen, & Hazan, 2018; Y. Li & Liang, 2018; Du, Zhai, Poczos, & Singh, 2018; Zou, Cao, Zhou, & Gu, 2018). Different from other studies that focus on general non-convex minimization problems, these recent works explored the properties for overparameterized deep neural networks and presented sharp analysis of (stochastic) gradient descent.

- **Smart Step Sizes or Learning Rates.** Step sizes or learning rates play an important role in an optimization algorithm for learning deep neural networks. Conventional polynomially decreasing step sizes are observed to be non-effective for deep learning. Smart step size schemes have been proposed including stagewise geometrically decreasing step size (Y. Xu, Lin, & Yang, 2017), and adaptive step sizes (Kingma & Ba, 2015; J. Chen & Gu, 2018; Zhou, Tang, Yang, Cao, & Gu, 2018; Zaheer, Reddi, Sachan, Kale, & Kumar, 2018; Luo, Xiong, Liu, & Sun, 2019; Z. Chen et al., 2019). A stagewise geometrically decreasing step size is usually adopted in SGD, SHB and SNAG for deep learning, which starts from a relatively large step size and decreases by a constant factor after a number of iterations. This step size scheme has achieved the state of the art result on the ImageNet classification task (He, Zhang, Ren, & Sun, 2016; Real, Aggarwal, Huang, & Le, 2019; Tan & Le, 2019). The idea of adaptive step size dates back to AdaGrad (Duchi, Hazan, & Singer, 2011), which was proposed for convex optimization. It has several variants with Adam (Kingma & Ba, 2015) being one of its most popular variants. The adaptive algorithms have been analyzed for non-convex optimization problems (X. Li & Orabona, 2018; J. Chen & Gu, 2018; Zhou et al., 2018; Zaheer et al., 2018; Luo et al., 2019).

Limitations and Future Directions. Although some nice results have been achieved in non-convex optimization and learning deep neural networks, there still remain many issues that require further investigation.

- **The gap between practice and theory.** There are several limitations of existing analysis: (i) most existing analysis of SGD uses a very small step size (Ghadimi & Lan, 2013; Yan et al., 2018; Davis & Drusvyatskiy, 2018), which is far from being practical; (ii) most theoretical analysis of non-convex optimization algorithms focus on optimization error; however, it is more important to consider the generalization performance of a stochastic optimization algorithm; (iii) global analysis of SGD imposes strong conditions on the level of over-parameterization (Allen-Zhu et al., 2018; Arora et al., 2018; Y. Li & Liang, 2018; Du et al., 2018; Zou et al., 2018), which is far from being practical. To address the first two limitations, we have conducted some preliminary study of SGD with a stagewise geometrically decreasing step size scheme by analyzing both the optimization error and the generalization error. Our analysis exhibits that the stagewise geometrically decreasing step scheme can leverage some nice properties of deep neural networks and enjoy faster convergence for both the training error and testing error than using a conventional polynomially decreasing step size. Some important theoretical questions that deserve more attention are (i) why do stochastic momentum methods exhibit better generalization performance than SGD (Yan et al., 2018); (ii) how does the adaptive learning rate affect the generalization performance; (iii) how can we derive much sharper analysis of practical SGD for finding a global minimizer of deep learning with good generalization performance.

- **Better stochastic algorithms for deep learning.** Beyond theoretical questions mentioned above, it is also important to design better stochastic algorithms for deep learning. While most recent studies focus on designing better adaptive learning rates, however, they have mostly ignored the role of stochastic gradients itself. The learning rate plays its role through multiplying with stochastic gradients. We believe that it is important to consider the properties of stochastic gradients, which essentially depend on the data.

Non-Convex Min-Max Optimization and Generative Adversarial Networks. Recently, non-convex non-concave min-max optimization has received increasing attention due to its application in generative adversarial networks (GAN) (Goodfellow et al., 2014; Radford, Metz, & Chintala, 2015; Arjovsky, Chintala, & Bottou, 2017). GAN has emerged to be an important paradigm of unsupervised learning. It learns a generator network and a discriminator network in a unified framework by solving a min-max problem of the following form:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{u}\in\mathcal{U}} \mathcal{L}(\mathbf{w}, \mathbf{u}),$$

where $\mathbf{w}$ denotes the parameter of the generator network and $\mathbf{u}$ denotes the parameter of the discriminator network. Although many variants of GAN have been investigated, the research on optimization algorithms for GAN remains rare. In practice, most studies use a primal-dual variant of Adam for optimization, which runs several steps of Adam for updating the discriminator network and then runs one step of Adam for updating the generator network. Theoretically, most existing results of min-max optimization algorithms for GAN are either asymptotic (Daskalakis, Ilyas, Syrgkanis, & Zeng, 2017; Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017; Nagarajan & Kolter, 2017; Cherukuri, Gharesifard, & Cortes, 2017) or their analysis require strong assumptions of the problem (Nagarajan & Kolter, 2017; Grnarova, Levy, Lucchi, Hofmann, & Krause, 2017) (e.g., the problem is concave in maximization). In our recent study (Lin, Liu, Rafique, & Yang, 2018), we proposed new stochastic algorithms based on the proximal point framework for solving the non-convex non-concave min-max problem of GAN, and established their complexities for finding approximate first-order stationary points without convex and concavity assumptions.

Future studies in this direction should answer the following questions (i) how can we analyze the generalization performance of stochastic min-max optimization algorithms for GAN? (ii) does GAN exhibit some nice properties as in deep learning that facilitates the design of better stochastic algorithms? (iii) why is the Adam algorithm more effective than SGD for GAN? (iv) how can we design faster stochastic algorithms for solving non-convex non-concave min-max problems with lower complexities?

Other Non-Convex Learning Problems. Beyond regular deep learning and GAN, non-convex learning also has some important applications in machine learning. Below, we will mention several of them.

- **Learning with Non-convex Regularizers.** Learning with a non-convex regularizer can be formulated as:

  $$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) := \mathrm{E}_{\mathbf{z}}[f(\mathbf{w}; \mathbf{z})] + R(\mathbf{w})$$

  where $R(\mathbf{w})$ denotes a regularizer, which includes the indicator function of a non-convex set. Commonly used non-convex regularizers that have been well studied include log-sum penalty (LSP) (Candès, Wakin, & Boyd, 2008), minimax concave penalty (MCP) (C.-H. Zhang, 2010), smoothly clipped absolute deviation (SCAD) (J. Fan & Li, 2001), capped ℓ_1 penalty (T. Zhang, 2010), transformed ℓ_1 norm (S. Zhang & Xin, 2014). However, there are many other interesting non-convex regularizations (Chartrand, 2012; Chartrand & Yin, 2016; Wen, Chu, Liu, & Qiu, 2018). For example, one can formulate learning a quantized neural network as a non-convex minimization with a non-convex constraint. Although non-smooth non-convex regularization has been considered in literature (Attouch, Bolte, & Svaiter, 2013; Bolte, Sabach, & Teboulle, 2014; Bot, Csetnek, & László, 2016; H. Li & Lin, 2015; Yu, Zheng, Marchetti-Bowick, & Xing, 2015; L. Yang, 2018; T. Liu, Pong, & Takeda, 2018; An & Nam, 2017; Zhong & Kwok, 2014), existing results are restricted to deterministic optimization and asymptotic or local convergence analysis. In our recent works (Y. Xu, Jin, & Yang, 2019; Y. Xu, Qi, Lin, Jin, & Yang, 2019), we have proposed new stochastic algorithms for tackling learning with a non-smooth non-convex regularizer, and established state-of-the-art non-asymptotic convergence rates.

- **DC programming.** Difference-of-Convex (DC) programming is to solve non-convex minimization problems of the following form:

$$\min_{\mathbf{w}} f(\mathbf{w}) - g(\mathbf{w})$$

where both f and g are convex functions. DC programming finds applications in many machine learning problems (Le Thi, Dinh, & Belghiti, 2014; Le Thi & Dinh, 2014; Nitanda & Suzuki, 2017; Thi, Le, Phan, & Tran, 2017; Khalaf, Astorino, d'Alessandro, & Gaudioso, 2017). For example, positive unlabeled learning problems can be formulated as a DC programming (Kiryo, Niu, du Plessis, & Sugiyama, 2017). In (Y. Xu, Qi, et al., 2019), we developed new stochastic DC algorithms for a broad family of DC problems, and established their complexities.

- **Distributionally Robust Optimization (DRO).** DRO is to solve the following min-max problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \max_{\mathbf{p}\in\mathcal{P}} \sum_{i=1}^{n} p_i f(\mathbf{w}, \mathbf{z}_i)$$

where $\mathcal{P} \subseteq \{\mathbf{p} \in \mathbb{R}^n, \sum_i^n p_i = 1, p_i \geq 0\}$ encodes some constraint that how far $\mathbf{p}$ deviates from the empirical distribution $\hat{p}_i = 1/n, i = 1, \dots, n$. DRO has found to be effective in handling imbalanced data (Namkoong & Duchi, 2016, 2017; Zhu, Li, Wang, Gong, & Yang, 2019; Y. Fan, Lyu, Ying, & Hu, 2017). It is also related to variance-based regularization and can yield smaller excess risk bounds (Namkoong & Duchi, 2017). When the loss function $f(\mathbf{w}, \mathbf{z})$ is non-convex in terms of $\mathbf{w}$, the above problem is non-convex and concave min-max problems. In (Rafique, Liu, Lin, & Yang, 2018), we have proposed efficient stochastic algorithms for solving the above min-max problems, and demonstrated that it gives better performance than SGD for learning a deep neural network in the presence of imbalanced data.

- **Learning with Truncated Losses.** Learning with truncated losses has long history in statistics (Wu & Liu, 2007; Belagiannis, Rupprecht, Carneiro, & Navab, 2015), which is more robust to outliers and can be formulated as

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \phi(f(\mathbf{w}, \mathbf{z}_i))$$

where $\phi(\cdot)$ is suitable concave truncation function (Y. Xu, Zhu, et al., 2019; Belagiannis et al., 2015). The above problem is a non-convex minimization problem. In (Y. Xu, Zhu, et al., 2019), we studied SGD for minimizing the above truncated losses and observed improved performance in the presence of various types of outliers and noise. However, it remains a question whether SGD converges to a global minimizer.

Heavily Constrained Learning

As ML is increasingly deployed in various domains, more and more problems are being formulated as constrained optimization problems where constraints are introduced to account for other factors/concerns beyond the prediction performance (B. Woodworth, Gunasekar, Ohannessian, & Srebro, 2017; Hardt et al., 2016; Globerson & Roweis, 2006b; Gupta et al., 2016; Canini et al., 2016; Globerson & Roweis, 2006b). Recently, there is much interest in measuring and ensuring fairness in ML, which is important in domains protected by anti-discrimination law (B. E. Woodworth et al., 2017; Hardt et al., 2016; Zafar et al., 2017; A. Agarwal et al., 2018). For example, a financial institution may want to use machine learning methods to predict whether a particular individual will pay back a loan or not for making a lending decision. In this case, it is morally and legally undesirable to discriminate based on the person's race and/or gender. A variety of notions of fairness has been considered in literature, including demographic parity, equality of opportunity, equalized odds, 80% rule, which can be modeled naturally as data dependent equality or inequality constraints (B. Woodworth et al., 2017; Hardt et al., 2016; Globerson & Roweis, 2006b).

Learning with data dependent constraints could also arise in *Interpretable learning*, which requires the prediction or the predictive model to be interpretable by a human. For example, if ML is used to predict whether a medication is effective for a client, then the client wants to know why it is effective in order to trust the medication. One way to achieve interpretable learning is to impose human-interpretable constraints into the learning process. For instance, for predicting an individual will pay back a loan or not, it is expected the probability of paying back is likely to increase as the person's income increases. It can be modeled as a constraint on the monotonicity of the predictive function respect to some features (Gupta et al., 2016).

Learning with complicated and complex constraints can find applications in other scenarios. In *Neyman-Pearson (NP) classification* paradigm (Rigollet & Tong, 2011), one needs to minimize false negative rate with an upper bound on false positive rate, where the upper bound on false positive rate is represented as a constraint. When the observed data are subject to some *uncertainty* (e.g, corruption, missing values, noise contamination), many studies have formulated the task as a constrained learning problem (Globerson & Roweis, 2006a; Sra et al., 2011). Recent works also found that imposing constrains on model parameters of neural networks can be more effective than using a regularization term in the objective for improving the prediction performance (Gouk, Frank, Pfahringer, & Cree, 2018; Ravi, Dinh, Lokhande, & Singh, 2018), and can improve the *robustness* of learned neural networks to adversarial examples (Cisse, Bojanowski, Grave, Dauphin, & Usunier, 2017). The robustness of a neural network is very important for applications in security critical domains (e.g., autonomous driving) (Carlini & Wagner, 2017; Tian, Pei, Jana, & Ray, 2018).

Constrained convex optimization has been studied extensively for a few decades and different methods, ranging from projected gradient methods, Frank-Wolfe methods (or conditional gradient methods), barrier methods, augmented Lagrangian methods, penalty methods, level-set methods to trust-region methods, have been developed and studied. However, the design of most existing constrained optimization algorithms suffers from severe scalability issues in the presence of big data and many complex constraints due to various reasons.

The general constrained learning problem can be formulated as:

$$\min_{\mathbf{x}\in\mathcal{X}} f_0(\mathbf{x}), \tag{2}$$

$$s.t.\ f_i(\mathbf{x}) \leq r_i, i = 1, \ldots, m \tag{3}$$

In (Mahdavi, Yang, Jin, & Zhu, 2012; T. Yang, Lin, & Zhang, 2017), we developed new theories of projection reduced (stochastic) first-order methods with only one or a logarithmic number of projections. In (Lin, Nadarajah, Soheili, & Yang, 2019), we developed new stochastic level-set methods for a family of finite-sum constrained convex optimization problems which can guarantee the exact feasibility of constraints. Recently, we proposed a class of subgradient methods for constrained optimization where the objective function and the constraint functions are non-convex (Ma, Lin, & Yang, 2019).

However, there still remain many challenging problems for heavily constrained learning.

- How to efficiently handle a large number of constraints?
- How do the constraints affect the generalization performance of a learned model?
- How to establish stronger convergence for a constrained optimization with non-convex objectives and non-convex constraints?

Acknowledgments

We acknowledge the support by National Science Foundation Career Award 1844403. Part of the our research cited in this paper was supported by National Science Foundation (IIS-1545995).

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In *Proceedings of the 35th international conference on machine learning (icml)* (pp. –).

Agarwal, N., Allen Zhu, Z., Bullins, B., Hazan, E., & Ma, T. (2017). Finding approximate local minima faster than gradient descent. In *Acm symposium on theory of computing (stoc)* (pp. 1195–1199).

Allen-Zhu, Z., Li, Y., & Song, Z. (2018). A convergence theory for deep learning via over-parameterization. *CoRR*, *abs/1811.03962*.

Allen-Zhu, Z. (2017). Natasha 2: Faster non-convex optimization than sgd. *CoRR*, */abs/1708.08694/v4*.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... Zhu, Z. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd international conference on international conference on machine learning (icml)* (pp. 173–182).

An, N. T., & Nam, N. M. (2017). Convergence analysis of a proximal point algorithm for minimizing differences of functions. *Optimization*, *66*(1), 129-147.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214–223).

Arora, S., Cohen, N., & Hazan, E. (2018). On the optimization of deep networks: Implicit acceleration by overparameterization. *arXiv preprint arXiv:1802.06509*.

Attouch, H., Bolte, J., & Svaiter, B. F. (2013). Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, *137*(1), 91–129.

Belagiannis, V., Rupprecht, C., Carneiro, G., & Navab, N. (2015). Robust optimization for deep regression. In *Proceedings of the ieee international conference on computer vision* (pp. 2830–2838).

Bolte, J., Sabach, S., & Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, *146*, 459–494.

Bot, R. I., Csetnek, E. R., & László, S. C. (2016, Feb 01). An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, *4*(1), 3–25.

Candès, E. J., Wakin, M. B., & Boyd, S. P. (2008, Dec 01). Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, *14*(5), 877–905.

Canini, K., Cotter, A., Gupta, M. R., Fard, M. M., & Pfeifer, J. (2016). Fast and flexible monotonic functions with ensembles of lattices. In *Proceedings of the 30th international conference on neural information processing systems (nips)* (pp. 2927–2935).

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)* (pp. 39–57).

Carmon, Y., Duchi, J. C., Hinder, O., & Sidford, A. (2016). Accelerated methods for non-convex optimization. *CoRR*, *abs/1611.00756*.

Cartis, C., Gould, N. I. M., & Toint, P. L. (2011a, Dec 01). Adaptive cubic regularisation methods for unconstrained optimization. part ii: worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, *130*(2), 295–319.

Cartis, C., Gould, N. I. M., & Toint, P. L. (2011b). Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, *127*(2), 245–295.

Chartrand, R. (2012). Nonconvex splitting for regularized low-rank+ sparse decomposition. *IEEE Transactions on Signal Processing*, *60*(11), 5810–5819.

Chartrand, R., & Yin, W. (2016). Nonconvex sparse regularization and splitting algorithms. In *Splitting methods in communication, imaging, science, and engineering* (pp. 237–249). Springer.

Chen, J., & Gu, Q. (2018). Closing the generalization gap of adaptive gradient methods in training deep neural networks.

arXiv preprint arXiv:1806.06763.

Chen, Z., Yuan, Z., Yi, J., Zhou, B., Chen, E., & Yang, T. (2019). Universal stage-wise learning for non-convex problems with convergence on averaged solutions. In *7th international conference on learning representations, ICLR 2019, new orleans, la, usa, may 6-9, 2019.*

Cherukuri, A., Gharesifard, B., & Cortes, J. (2017). Saddle-point dynamics: conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, *55*(1), 486–511.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 854–863).

Daskalakis, C., Ilyas, A., Syrgkanis, V., & Zeng, H. (2017). Training gans with optimism. *CoRR*, *abs/1711.00141*.

Davis, D., & Drusvyatskiy, D. (2018). Stochastic subgradient method converges at the rate o($k^{-1/4}$) on weakly convex functions. *arXiv preprint arXiv:1802.02988*.

Deng, L., & Liu, Y. (2018). *Deep learning in natural language processing*. Springer.

Du, S. S., Zhai, X., Poczos, B., & Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121–2159.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, Y., Lyu, S., Ying, Y., & Hu, B. (2017). Learning with average top-k loss. In *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, 4-9 december 2017, long beach, ca, USA* (pp. 497–505).

Ghadimi, S., & Lan, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, *23*(4), 2341–2368.

Globerson, A., & Roweis, S. (2006a). Nightmare at test time: Robust learning by feature deletion. In *Proceedings of the 23rd international conference on machine learning* (pp. 353–360).

Globerson, A., & Roweis, S. (2006b). Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on machine learning* (pp. 353–360).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Gouk, H., Frank, E., Pfahringer, B., & Cree, M. (2018). Regularisation of neural networks by enforcing lipschitz continuity. *arXiv preprint arXiv:1804.04368*.

Grnarova, P., Levy, K. Y., Lucchi, A., Hofmann, T., & Krause, A. (2017). An online learning approach to generative adversarial networks. *CoRR*, *abs/1706.03269*.

Gupta, M. R., Cotter, A., Pfeifer, J., Voevodski, K., Canini, K. R., Mangylov, A., ... Esbroeck, A. V. (2016). Monotonic calibrated interpolated look-up tables. *Journal of Machine Learning Research (JMLR)*, *17*, 109:1–109:47.

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315–3323).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems 30 (nips)* (pp. 6629–6640).

Hillar, C. J., & Lim, L.-H. (2013, November). Most tensor problems are np-hard. *Journal of ACM*, *60*(6), 45:1–45:39.

Khalaf, W., Astorino, A., d'Alessandro, P., & Gaudioso, M. (2017). A dc optimization-based clustering technique for edge detection. *Optimization Letters*, *11*(3), 627–640.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings.* Retrieved from http://arxiv.org/abs/1412.6980

Kiryo, R., Niu, G., du Plessis, M. C., & Sugiyama, M. (2017). Positive-unlabeled learning with non-negative risk estimator. In *Advances in neural information processing systems 30* (pp. 1675–1685).

Kohler, J. M., & Lucchi, A. (2017). Sub-sampled cubic regularization for non-convex optimization. In *Proceedings of the international conference on machine learning (icml)* (pp. 1895–1904).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (nips)* (pp. 1106–1114).

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Le Thi, H. A., & Dinh, T. P. (2014). Dc programming in communication systems: challenging problems and methods. *Vietnam Journal of Computer Science*, *1*(1), 15–28.

Le Thi, H. A., Dinh, T. P., & Belghiti, M. (2014). Dca based algorithms for multiple sequence alignment (msa). *Central European Journal of Operations Research*, *22*(3), 501–524.

Li, H., & Lin, Z. (2015). Accelerated proximal gradient methods for nonconvex programming. In *Proceedings of the 28th international conference on neural information processing systems - volume 1* (pp. 379–387).

Li, X., & Orabona, F. (2018). On the convergence of stochastic gradient descent with adaptive stepsizes. *arXiv preprint arXiv:1805.08114*.

Li, Y., & Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in neural information processing systems (neurips)* (pp. 8157–8166).

Lin, Q., Liu, M., Rafique, H., & Yang, T. (2018). Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*.

Lin, Q., Nadarajah, S., Soheili, N., & Yang, T. (2019). A data efficient and feasible level set method for stochastic convex optimization with expectation constraints. *CoRR*, *abs/1908.03077*.

Liu, M., & Yang, T. (2017a). On noisy negative curvature descent: Competing with gradient descent for faster non-convex optimization. *CoRR*, *abs/1709.08571*.

Liu, M., & Yang, T. (2017b). Stochastic non-convex optimization with strong high probability second-order convergence. *CoRR*, *abs/1710.09447*.

Liu, T., Pong, T. K., & Takeda, A. (2018, Sep 08). A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems. *Mathematical Programming*.

Luo, L., Xiong, Y., Liu, Y., & Sun, X. (2019). Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*.

Ma, R., Lin, Q., & Yang, T. (2019). Proximally constrained methods for weakly convex optimization with weakly convex constraints. *arXiv preprint arXiv:1908.01871*.

Mahdavi, M., Yang, T., Jin, R., & Zhu, S. (2012). Stochastic gradient descent with only one projection. In *Advances in neural information processing systems (nips)* (p. 503-511).

Nagarajan, V., & Kolter, J. Z. (2017). Gradient descent GAN optimization is locally stable. In *Advances in neural information processing systems 30 (nips)* (pp. 5591–5600).

Namkoong, H., & Duchi, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in neural information processing systems* (pp. 2208–2216).

Namkoong, H., & Duchi, J. C. (2017). Variance-based regularization with convex objectives. In *Advances in neural information processing systems* (pp. 2971–2980).

Nesterov, Y., & Polyak, B. T. (2006). Cubic regularization of newton method and its global performance. *Math. Program.*, *108*(1), 177–205.

Nitanda, A., & Suzuki, T. (2017). Stochastic difference of convex algorithm and its application to training deep boltzmann machines. In *Artificial intelligence and statistics* (pp. 470–478).

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rafique, H., Liu, M., Lin, Q., & Yang, T. (2018). Non-convex min-max optimization: Provable algorithms and applications in machine learning. *CoRR*, *abs/1810.02060*.

Ravi, S. N., Dinh, T., Lokhande, V. S. R., & Singh, V. (2018). Constrained deep learning using conditional gradient and applications in computer vision. *arXiv preprint arXiv:1803.06453*.

Real, E., Aggarwal, A., Huang, Y., & Le, Q. V. (2019). Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 4780–4789).

Reddi, S. J., Zaheer, M., Sra, S., Poczos, B., Bach, F., Salakhutdinov, R., & Smola, A. J. (2017). A generic approach for escaping saddle points. *arXiv preprint arXiv:1709.01434*.

Rigollet, P., & Tong, X. (2011, November). Neyman-pearson classification, convexity and stochastic constraints. *J. Mach. Learn. Res.*, *12*, 2831–2855.

Royer, C. W., & Wright, S. J. (2017). Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *CoRR*, *abs/1706.03131*.

Sra, S., Nowozin, S., & Wright, S. J. (2011). *Optimization for machine learning*. The MIT Press.

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

Thi, H. A. L., Le, H. M., Phan, D. N., & Tran, B. (2017). Stochastic dca for the large-sum of non-convex functions problem and its application to group variable selection in classification. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 3394–3403).

Tian, Y., Pei, K., Jana, S., & Ray, B. (2018). Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th international conference on software engineering* (pp. 303–314).

Wen, F., Chu, L., Liu, P., & Qiu, R. C. (2018). A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access*, *6*, 69883–69906.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., & Srebro, N. (2017). Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*.

Woodworth, B. E., Gunasekar, S., Ohannessian, M. I., & Srebro, N. (2017). Learning non-discriminatory predictors. In *Proceedings of the 30th conference on learning theory, COLT 2017, amsterdam, the netherlands, 7-10 july 2017* (pp. 1920–1953).

Wu, Y., & Liu, Y. (2007). Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, *102*(479), 974–983.

Xu, P., Roosta-Khorasani, F., & Mahoney, M. W. (2017). Newton-type methods for non-convex optimization under inexact hessian information. *CoRR*, *abs/1708.07164*.

Xu, Y., Jin, R., & Yang, T. (2019). Stochastic proximal gradient methods for nonsmooth non-convex regularized problems. *arXiv preprint arXiv:1902.07672*.

Xu, Y., Lin, Q., & Yang, T. (2017). Stochastic convex optimization: Faster local growth implies faster global convergence. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 3821–3830).

Xu, Y., Qi, Q., Lin, Q., Jin, R., & Yang, T. (2019). Stochastic optimization for DC functions and non-smooth non-convex regularizers with non-asymptotic convergence. In *Proceedings of the 36th international conference on machine learning, ICML 2019, 9-15 june 2019, long beach, california, USA* (pp. 6942–6951).

Xu, Y., Rong, J., & Yang, T. (2018). First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in neural information processing systems (neurips)* (pp. 5530–5540).

Xu, Y., Zhu, S., Yang, S., Zhang, C., Jin, R.,

& Yang, T. (2019). Learning with nonconvex truncated losses by SGD. In *Proceedings of the thirty-fifth conference on uncertainty in artificial intelligence, UAI 2019, tel aviv, israel, july 22-25, 2019* (p. 244).

Yan, Y., Yang, T., Li, Z., Lin, Q., & Yang, Y. (2018). A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI 2018, july 13-19, 2018, stockholm, sweden.* (pp. 2955–2961).

Yang, L. (2018). Proximal gradient method with extrapolation and line search for a class of nonconvex and nonsmooth problems. *CoRR*, *abs/1711.06831*.

Yang, T., Lin, Q., & Zhang, L. (2017). A richer theory of convex constrained optimization with reduced projections and improved rates. In *Proceedings of the 34th international conference on machine learning (icml)* (p. -).

Yang, T., Mahdavi, M., Jin, R., Zhang, L., & Zhou, Y. (2012). Multiple kernel learning from noisy labels by stochastic programming. In *Proceedings of the international conference on machine learning (icml)* (pp. 233–240).

You, S., Ding, D., Canini, K. R., Pfeifer, J., & Gupta, M. R. (2017). Deep lattice networks and partial monotonic functions. In *Advances in neural information processing systems 30 (nips)* (pp. 2985–2993).

Yu, Y., Zheng, X., Marchetti-Bowick, M., & Xing, E. P. (2015). Minimizing nonconvex non-separable functions. In *The* 17^{th} *international conference on artificial intelligence and statistics (AISTATS).*

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web* (pp. 1171–1180).

Zaheer, M., Reddi, S., Sachan, D., Kale, S., & Kumar, S. (2018). Adaptive methods for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31* (pp. 9793–9803). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/8186-adaptive-methods-for-nonconvex-optimization.pdf

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894 – 942.

Zhang, C.-H., & Zhang, T. (2012, 11). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, *27*(4), 576–593. doi: 10.1214/12-STS399

Zhang, S., & Xin, J. (2014). Minimization of transformed l_1 penalty: Theory, difference of convex function algorithm, and robust application in compressed sensing. *CoRR*, *abs/1411.5735*.

Zhang, T. (2010, March). Analysis of multistage convex relaxation for sparse regularization. *J. Mach. Learn. Res.*, *11*, 1081–1107.

Zhong, W., & Kwok, J. T. (2014). Gradient descent with proximal average for nonconvex and composite regularization. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence, july 27 -31, 2014, québec city, québec, canada.* (pp. 2206–2212).

Zhou, D., Tang, Y., Yang, Z., Cao, Y., & Gu, Q. (2018). On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*.

Zhu, D., Li, Z., Wang, X., Gong, B., & Yang, T. (2019). A robust zero-sum game framework for pool-based active learning. In *The 22nd international conference on artificial intelligence and statistics* (pp. 517–526).

Zou, D., Cao, Y., Zhou, D., & Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep relu networks. *CoRR*, *abs/1811.08888*.

Tianbao Yang is an assistant professor at the Computer Science Department at the University of Iowa since 2014. He was a researcher at NEC Laboratories America, and a researcher at GE Global Research before joining UIowa. He has won the best student paper award at COLT 2012 and the NSF Career Award. He is an associate editor for Neurocomputing Journal, and the Journal of Mathematical Foundations of Computing. He has served as senior program committee member for AAAI and IJCAI and reviewer for AISTATS, ICML and NeurIPS.