OPT2020: 12th Annual Workshop on Optimization for Machine Learning

PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimization

Zhize Li ZHIZE.LI@KAUST.EDU.SA
Hongyan Bao HONGYAN.BAO@KAUST.EDU.SA
Xiangliang Zhang XIANGLIANG.ZHANG@KAUST.EDU.SA
Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST)

Abstract

In this paper, we propose a novel stochastic gradient estimator—ProbAbilistic Gradient Estimator (PAGE)—for nonconvex optimization. PAGE is easy to implement as it is designed via a small adjustment to vanilla SGD: in each iteration, PAGE uses the vanilla minibatch SGD update with probability p or reuses the previous gradient with a small adjustment, at a much lower computational cost, with probability $1-p$. We give a simple formula for the optimal choice of p. We prove tight lower bounds for nonconvex problems, which are of independent interest. Moreover, we prove matching upper bounds both in the finite-sum and online regimes, which establish that PAGE is an optimal method. Besides, we show that for nonconvex functions satisfying the Polyak-Łojasiewicz (PL) condition, PAGE can automatically switch to a faster linear convergence rate. Finally, we conduct several deep learning experiments (e.g., LeNet, VGG, ResNet) on real datasets in PyTorch, and the results demonstrate that PAGE not only converges much faster than SGD in training but also achieves the higher test accuracy, validating our theoretical results and confirming the practical superiority of PAGE.

1. Introduction

Nonconvex optimization is ubiquitous across many domains of machine learning, including robust regression, low rank matrix recovery, sparse recovery and supervised learning [13]. Driven by the applied success of deep neural networks [21], and the critical place nonconvex optimization plays in training them, research in nonconvex optimization has been undergoing a renaissance [6, 8, 9, 25, 28, 47].

1.1. The problem

Motivated by this development, we consider the general optimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a differentiable and possibly nonconvex function. We are interested in functions having the *finite-sum* form

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{2}$$

where the functions f_i are also differentiable and possibly nonconvex. Form (2) captures the standard empirical risk minimization problems in machine learning [41]. Moreover, if the number of data samples n is very large or even infinite, e.g., in the online/streaming case, then $f(x)$ usually is modeled via the *online* form

$$f(x) := \mathbb{E}_{\zeta\sim\mathcal{D}}[F(x,\zeta)], \tag{3}$$

which we also consider in this work. For notational convenience, we adopt the notation of the finite-sum form (2) in the descriptions and algorithms in the rest of this paper. However, our results apply to the online form (3) as well by letting $f_i(x) := F(x,\zeta_i)$ and treating n as a very large number or even infinite.

1.2. Gradient complexity

To measure the efficiency of algorithms for solving the nonconvex optimization problem (1), it is standard to bound the number of stochastic gradient computations needed to find a solution of suitable characteristics. In this paper we use the standard term *gradient complexity* to describe such bounds. In particular, our goal will be to find a (possibly random) point $\widehat{x} \in \mathbb{R}^d$ such that $\mathbb{E}\|\nabla f(\widehat{x})\| \leq \epsilon$, where the expectation is with respect to the randomness inherent in the algorithm. We use the term *ϵ-approximate solution* to refer to such a point $\widehat{x}$.

Two of the most classical gradient complexity results for solving problem (1) are those for gradient descent (GD) and stochastic gradient descent (SGD). In particular, the gradient complexity of GD is $O(n/\epsilon^2)$ in this nonconvex regime, and assuming that the stochastic gradient satisfies a (uniform) bounded variance assumption (Assumption 1), the gradient complexity of SGD is $O(1/\epsilon^4)$. Note that although SGD has a worse dependence on ϵ, it typically only needs to compute a constant minibatch of stochastic gradients in each iteration instead of the full batch (i.e., n stochastic gradients) used in GD. Hence, SGD is better than GD if the number of data samples n is very large or the error tolerance ϵ is not very small.

There has been extensive research in designing gradient-type methods with an improved dependence on n and/or ϵ [8, 9, 32, 34]. In particular, the SVRG method of Johnson and Zhang [14], the SAGA method of Defazio et al. [5] and the SARAH method of Nguyen et al. [35] are representatives of what is by now a large class of *variance-reduced* methods, which have played a particularly important role in this effort. However, the analyses in these papers focused on the convex regime. Furthermore, several accelerated (momentum) methods have been designed as well [1, 18–20, 27, 30, 31, 33], with or without variance reduction. There are also some lower bounds given by [44, 45].

Coming back to problem (1) in the nonconvex regime studied in this paper, interesting recent development starts with the work of Reddi et al. [40], and Allen-Zhu and Hazan [2], who have concurrently shown that if f has the finite-sum form (2), a suitably designed minibatch version of SVRG enjoys the gradient complexity $O(n + n^{2/3}/\epsilon^2)$, which is an improvement on the $O(n/\epsilon^2)$ gradient complexity of GD. Subsequently, other variants of SVRG were shown to posses the same improved rate, including those developed by [7, 11, 23, 26, 39]. More recently, Fang et al. [6] proposed the SPIDER method, and Zhou et al. [47] proposed the SNVRG method, both of which improve the gradient complexity further to $O(n + \sqrt{n}/\epsilon^2)$. Further variants of the SARAH method (e.g., [12, 24, 25, 37, 43]) which also achieve the same $O(n+\sqrt{n}/\epsilon^2)$ gradient complexity have been developed. Particularly, Horváth et al. [12] proposed the Geom-SARAH to automatically adapt

different parameter settings via dynamic batch size and epoch length. Also there are some lower bounds given by [3, 6, 46]. See Table 1 for an overview of results.

2. Our Contributions

As we show in through this work, despite enormous effort by the community to design efficient methods for solving (1) in the nonconvex regime, there is still a considerable gap in our understanding. First, while optimal methods for (1) in the finite-sum regime exist (e.g., SPIDER [6], SpiderBoost [43], SARAH [37], SSRGD [25]), the known lower bound $\Omega(\sqrt{n}/\epsilon^2)$ [6] used to establish their optimality works only for $n \leq O(1/\epsilon^4)$, i.e., in the small data regime (see Table 1). Moreover, these methods are unnecessarily complicated, often with a double loop structure, and reliance on several hyperparameters. Besides, there is also no tight lower bound to show the optimality of optimal methods in the online regime.

Table 1: Gradient complexity for finding $\widehat{x}$ satisfying $\mathbb{E}\|\nabla f(\widehat{x})\| \leq \epsilon$ in nonconvex problems

Problem	Assumption	Algorithm or Lower Bound	Gradient complexity
Finite-sum (2)	Asp. 2	GD [34]	$O(\frac{n}{\epsilon^2})$
Finite-sum (2)	Asp. 2	SVRG [2, 40], SCSG [23], SVRG+ [26]	$O(n + \frac{n^{2/3}}{\epsilon^2})$
Finite-sum (2)	Asp. 2	SNVRG [47], Geom-SARAH [12]	$\widetilde{O}\left(n + \frac{\sqrt{n}}{\epsilon^2}\right)$
Finite-sum (2)	Asp. 2	SPIDER [6], SpiderBoost [43], SARAH [37], SSRGD [25]	$O(n + \frac{\sqrt{n}}{\epsilon^2})$
Finite-sum (2)	Asp. 2	PAGE (this paper)	$O(n + \frac{\sqrt{n}}{\epsilon^2})$
Finite-sum (2)	Asp. 2	Lower bound [6]	$\Omega(\frac{\sqrt{n}}{\epsilon^2})$ if $n \leq O(\frac{1}{\epsilon^4})$
Finite-sum (2)	Asp. 2	Lower bound (this paper)	$\Omega(n + \frac{\sqrt{n}}{\epsilon^2})$
Finite-sum (2)	Asp. 2 and 3	PAGE (this paper)	$O\left((n + \sqrt{n}\kappa)\log\frac{1}{\epsilon}\right)$ [1]
Online (3) [2]	Asp. 1 and 2	SGD [9, 15, 28]	$O(\frac{\sigma^2}{\epsilon^4})$
Online (3)	Asp. 1 and 2	SCSG [23], SVRG+ [26]	$O(b + \frac{b^{2/3}}{\epsilon^2})$
Online (3)	Asp. 1 and 2	SNVRG [47], Geom-SARAH [12]	$\widetilde{O}\left(b + \frac{\sqrt{b}}{\epsilon^2}\right)$
Online (3)	Asp. 1 and 2	SPIDER [6], SpiderBoost [43], SARAH [37], SSRGD [25]	$O(b + \frac{\sqrt{b}}{\epsilon^2})$
Online (3)	Asp. 1 and 2	PAGE (this paper)	$O(b + \frac{\sqrt{b}}{\epsilon^2})$ [3]
Online (3)	Asp. 1 and 2	Lower bound (this paper)	$\Omega(b + \frac{\sqrt{b}}{\epsilon^2})$
Online (3)	Asp. 1, 2 and 3	PAGE (this paper)	$O\left((b + \sqrt{b}\kappa)\log\frac{1}{\epsilon}\right)$

1. Note that PAGE can automatically switch to a faster *linear convergence* $O(\cdot \log \frac{1}{\epsilon})$ instead of sublinear $O(\cdot \frac{1}{\epsilon^2})$ by exploiting the local structure of the objective function via the PL condition (Assumption 3).
2. Note that we refer the online problem (3) as the finite-sum problem (2) with large or infinite n as discussed in the introduction Section 1.1. In this online case, the full gradient may not be available (e.g., if n is infinite), thus the bounded variance of stochastic gradient Assumption 1 is needed in this case.
3. In the online case, $b := \min\{\frac{\sigma^2}{\epsilon^2}, n\}$, and σ is defined in Assumption 1. If n is very large, i.e., $b := \min\{\frac{\sigma^2}{\epsilon^2}, n\} = \frac{\sigma^2}{\epsilon^2}$, then $O(b + \frac{\sqrt{b}}{\epsilon^2}) = O(\frac{\sigma^2}{\epsilon^2} + \frac{\sigma}{\epsilon^3})$ is better than the rate $O(\frac{\sigma^2}{\epsilon^4})$ of SGD by a factor of $\frac{1}{\epsilon^2}$ or $\frac{\sigma}{\epsilon}$.

In this paper, we resolve the above issues by designing a novel ProbAbilistic Gradient Estimator (PAGE) described in Algorithm 1 for achieving optimal convergence results in nonconvex optimization. Moreover, PAGE is very simple and easy to implement. In each iteration, PAGE uses minibatch SGD update with probability p_t, or reuses the previous gradient with a small adjustment (at a low computational cost) with probability $1-p_t$ (see Line 4 of Algorithm 1). We would like to highlight the following points:

• We prove that PAGE achieves the optimal rates for both nonconvex finite-sum problem (2) and online problem (3) (see Corollaries 2 and 4 [4]). We also provide tight lower bounds for these two problems to close the gap and show the optimality of PAGE (see Theorem 2 and Corollary 5). Our lower bounds are inspired and based by recent work [3, 6]. See Table 1 for a detailed comparison with previous work.

• Moreover, we show that PAGE can automatically switch to a faster *linear convergence* $O(\cdot \log \frac{1}{\epsilon})$ by exploiting the local structure of the objective function, via the PL condition (Assumption 3), although the objective function f is globally nonconvex. See the middle and the last row of Table 1. For example, PAGE automatically switches from the sublinear rate $O(n+\sqrt{n}/\epsilon^2)$ to the faster *linear rate* $O((n+\sqrt{n}\kappa)\log\frac{1}{\epsilon})$ for nonconvex finite-sum problem (2) (see the Remark after Corollary 6).

• PAGE is easy to implement via a small adjustment to vanilla minibatch SGD, and takes a lower computational cost than SGD (i.e., $p=1$ in PAGE) since $b' < b$. We conduct several deep learning experiments (e.g., LeNet, VGG, ResNet) on real datasets in PyTorch showing that PAGE indeed not only converges much faster than SGD in training but also achieves higher test accuracy. This validates our theoretical results and confirms the practical superiority of PAGE.

2.1. The PAGE gradient estimator

In this section, we describe PAGE, an SGD variant employing a new, simple and optimal gradient estimator (see Algorithm 1). In particular, PAGE was inspired by algorithmic design elements coming from methods such as SARAH [35], SPIDER [6], SSRGD [25] (usage of a recursive estimator), and L-SVRG [16] and SAGD [4] (probabilistic switching between two estimators to avoid a double loop structure). PAGE with constant probability p can be reduced to an equivalent form of the double loop algorithm with geometric distribution Geom-SARAH [12], but our single-loop PAGE is more flexible and also leads to simpler and better analysis.

Algorithm 1 ProbAbilistic Gradient Estimator (PAGE)

Input: initial point x^0, stepsize η, minibatch size b, $b' < b$, probability $\{p_t\} \in (0,1]$

1: $g^0 = \frac{1}{b}\sum_{i\in I}\nabla f_i(x^0)$ // I denotes random minibatch samples with $|I| = b$

2: **for** $t = 0, 1, 2, \ldots$ **do**

3: $x^{t+1} = x^t - \eta g^t$

4: $g^{t+1} = \begin{cases} \frac{1}{b}\sum_{i\in I}\nabla f_i(x^{t+1}) & \text{with probability } p_t \\ g^t + \frac{1}{b'}\sum_{i\in I'}(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) & \text{with probability } 1-p_t \end{cases}$

5: **end for**

Output: $\widehat{x}_T$ chosen uniformly from $\{x^t\}_{t\in[T]}$

In iteration t, the gradient estimator g^{t+1} of PAGE is defined in Line 4 of Algorithm 1, which indicates that PAGE uses the vanilla minibatch SGD update with probability p_t, and reuses the

4. All theorems and cocollaries and their proofs can be found in the full version of this paper [29].

previous gradient g^t with a small adjustment (which lowers the computational cost since $b' \ll b$) with probability $1 - p_t$. In particular, the $p_t \equiv 1$ case reduces to vanilla minibatch SGD, and to GD if we further set the minibatch size to $b = n$. We give a simple formula for the optimal choice of p_t, i.e., $p_t \equiv \frac{b'}{b+b'}$ is enough for PAGE to obtain the optimal convergence rates. More details can be found in the convergence theorems and corollaries in the full version of this paper [29].

3. Assumptions

Assumption 1 (Bounded variance) *The stochastic gradient has bounded variance if*

$$\exists \sigma > 0, \text{ such that } \mathbb{E}_i[\|\nabla f_i(x) - \nabla f(x)\|^2] \leq \sigma^2, \quad \forall x \in \mathbb{R}^d. \tag{4}$$

Assumption 2 (Average L-smoothness) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is average L-smooth if*

$$\exists L > 0, \text{ such that } \mathbb{E}_i[\|\nabla f_i(x) - \nabla f_i(y)\|^2] \leq L^2\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{5}$$

Moreover, we also prove faster linear convergence rates for nonconvex functions under the Polyak-Łojasiewicz (PL) condition [38].

Assumption 3 (PL condition) *A function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies PL condition if*

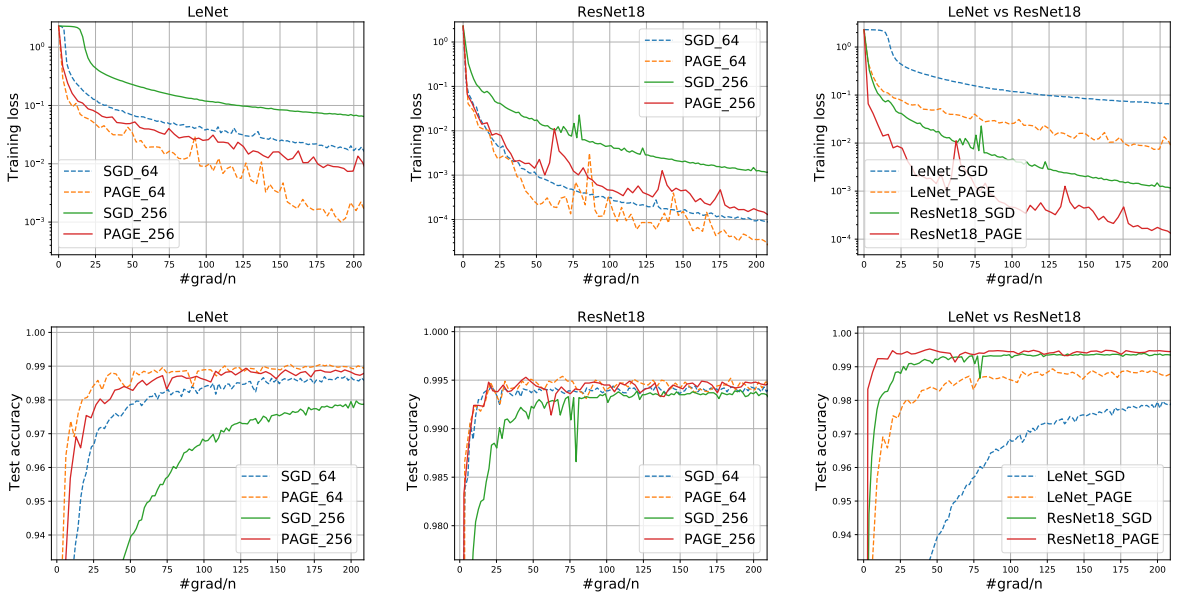
$$\exists \mu > 0, \text{ such that } \|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*), \ \forall x \in \mathbb{R}^d. \tag{6}$$

4. Experiments

We conduct several deep learning experiments for multi-class image classification. Concretely, we compare our PAGE algorithm with vanilla SGD by running standard LeNet [22], VGG [42] and ResNet [10] models on MNIST [22] and CIFAR-10 [17] datasets. We implement the algorithms in PyTorch [36] and run the experiments on several NVIDIA Tesla V100 GPUs.

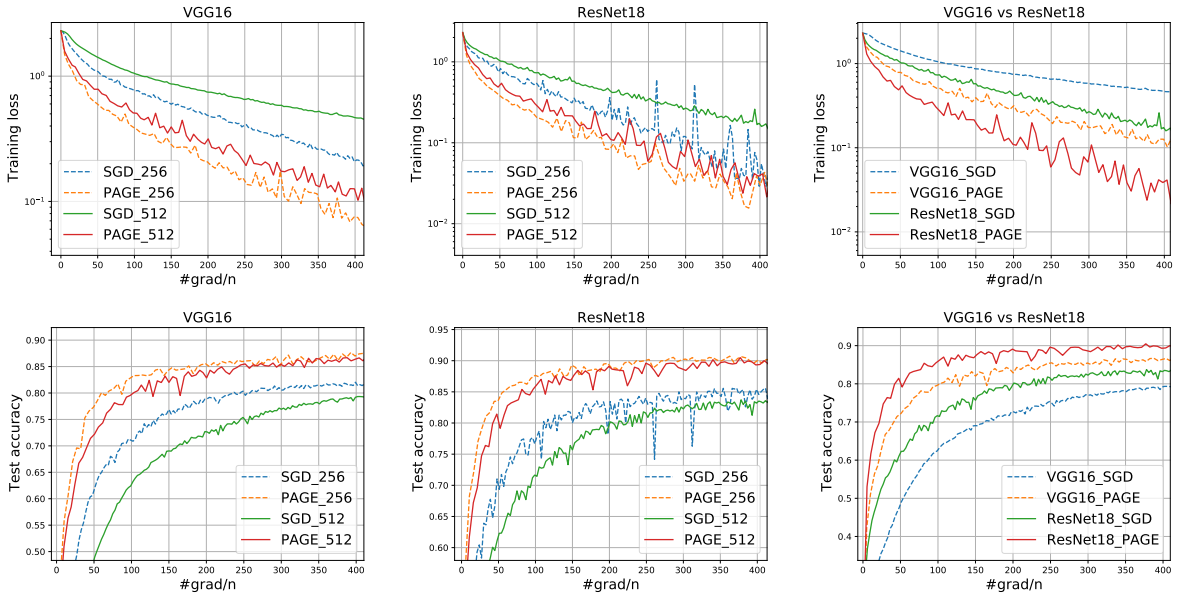
According to the update form in PAGE (see Line 4 of Algorithm 1), PAGE enjoys a lower computational cost than vanilla minibatch SGD (i.e., $p = 1$ in PAGE) since $b' < b$. Thus, in the experiments we want to show how the performance of PAGE compares with vanilla minibatch SGD under different minibatch sizes b. Note that we do not tune the parameters for PAGE, i.e., we set $b' = \sqrt{b}$ and $p = \frac{b'}{b+b'} = \frac{\sqrt{b}}{b+\sqrt{b}}$ according to our theoretical results (see e.g., Corollary 2 and 4).

Concretely, in Figure 1, we choose standard minibatch $b = 64$ and $b = 256$ for both PAGE and vanilla minibatch SGD for MNIST experiments. In Figure 2, we choose $b = 256$ and $b = 512$ for CIFAR-10 experiments. The first row of Figures 1 and 2 denotes the training loss with respect to the gradient computations, and the second row denotes the test accuracy with respect to the gradient computations. Both Figures 1 and 2 demonstrate that PAGE not only converges much faster than SGD in training but also achieves the higher test accuracy (which is typically very important in practice, e.g., lead to a better model). Moreover, the performance gap between PAGE and SGD is larger when the minibatch size b is larger (i.e, gap between solid lines in Figures 1a, 1b, 2a, 2b), which is consistent with the update form of PAGE, i.e, it reuses the previous gradient with a small adjustment (lower computational cost $b' = \sqrt{b}$ instead of b) with probability $1 - p_t$. The experimental results validate our theoretical results and confirm the practical superiority of PAGE.



(1a) Different minibatch size b (1b) Different minibatch size b (1c) Different neural networks

Figure 1: LeNet and ResNet18 on MNIST dataset



(2a) Different minibatch size b (2b) Different minibatch size b (2c) Different neural networks

Figure 2: VGG16 and ResNet18 on CIFAR-10 dataset

References

[1] Zeyuan Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages

1200–1205. ACM, 2017.

[2] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707, 2016.

[3] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

[4] Adel Bibi, Alibek Sailanbayev, Bernard Ghanem, Robert Mansel Gower, and Peter Richtárik. Improving SAGA via a probabilistic interpolation with gradient descent. *arXiv preprint arXiv:1806.05633*, 2018.

[5] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[6] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 687–697, 2018.

[7] Rong Ge, Zhize Li, Weiyao Wang, and Xiang Wang. Stabilized SVRG: Simple variance reduction for nonconvex optimization. In *Conference on Learning Theory*, pages 1394–1448, 2019.

[8] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[9] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] Samuel Horváth and Peter Richtárik. Nonconvex variance reduced optimization with arbitrary sampling. In *International Conference on Machine Learning*, pages 2781–2789, 2019.

[12] Samuel Horváth, Lihua Lei, Peter Richtárik, and Michael I Jordan. Adaptivity of stochastic gradient methods for nonconvex optimization. *arXiv preprint arXiv:2002.05359*, 2020.

[13] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3-4):142–336, 2017.

[14] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[15] Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

[16] Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.

[17] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[18] Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *arXiv preprint arXiv:1507.02000*, 2015.

[19] Guanghui Lan and Yi Zhou. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, 28(4):2753–2782, 2018.

[20] Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. In *Advances in Neural Information Processing Systems*, pages 10462–10472, 2019.

[21] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems*, pages 2345–2355, 2017.

[24] Bingcong Li, Meng Ma, and Georgios B Giannakis. On the convergence of sarah and beyond. In *International Conference on Artificial Intelligence and Statistics*, pages 223–233. PMLR, 2020.

[25] Zhize Li. SSRGD: Simple stochastic recursive gradient descent for escaping saddle points. In *Advances in Neural Information Processing Systems*, pages 1521–1531, 2019.

[26] Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 5569–5579, 2018.

[27] Zhize Li and Jian Li. A fast Anderson-Chebyshev acceleration for nonlinear optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1047–1057, 2020.

[28] Zhize Li and Peter Richtárik. A unified analysis of stochastic gradient methods for nonconvex federated optimization. *arXiv preprint arXiv:2006.07013*, 2020.

[29] Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. *arXiv preprint arXiv:2008.10898*, 2020.

[30] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning*, 2020.

[31] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in neural information processing systems*, pages 3384–3392, 2015.

[32] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4): 1574–1609, 2009.

[33] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

[34] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, 2004.

[35] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2613–2621. JMLR. org, 2017.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[37] Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *arXiv preprint arXiv:1902.05679*, 2019.

[38] Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[39] Xun Qian, Zheng Qu, and Peter Richtárik. L-SVRG and L-Katyusha with arbitrary sampling. *arXiv preprint arXiv:1906.01481*, 2019.

[40] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016.

[41] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: from theory to algorithms*. Cambridge University Press, 2014.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[43] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.

[44] Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in neural information processing systems*, pages 3639–3647, 2016.

[45] Guangzeng Xie, Luo Luo, and Zhihua Zhang. A general analysis framework of lower complexity bounds for finite-sum optimization. *arXiv preprint arXiv:1908.08394*, 2019.

[46] Dongruo Zhou and Quanquan Gu. Lower bounds for smooth nonconvex finite-sum optimization. *arXiv preprint arXiv:1901.11224*, 2019.

[47] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3925–3936, 2018.