Acceleration and Stability of the Stochastic Proximal Point Algorithm

Junhyung Lyle Kim JLYLEKIM@RICE.EDU
Department of Computer Science, Rice University
Panos Toulis PANOS.TOULIS@CHICAGOBOOTH.EDU
Booth School of Business, University of Chicago
Anastasios Kyrillidis ANASTASIOS@RICE.EDU
Department of Computer Science, Rice University

Abstract

Stochastic gradient descent (SGD) has emerged as the de-facto method for solving (unconstrained) stochastic optimization problems. However, it suffers from two fundamental limitations: (i) slow convergence due to inaccurate gradient approximation, and (ii) numerical instability, especially with respect to step size selection. To improve the slow convergence, accelerated variants such as stochastic gradient descent with momentum (SGDM) have been studied; however, the interference of gradient noise and momentum can aggravate the numerical instability. Proximal point methods, on the other hand, have gained much attention due to their numerical stability. Their stochastic accelerated variants though have received limited attention. To bridge this gap, we propose the stochastic proximal point algorithm with momentum (SPPAM), and study its convergence and stability. We show that SPPAM enjoys a better contraction factor compared to stochastic proximal point algorithm (SPPA), leading to faster convergence. In terms of stability, we show that SPPAM depends on problem constants more favorably than SGDM.

1. Introduction

In this paper, we are interested in the following unconstrained stochastic optimization problem:

$$\text{minimize}_{x\in\mathbb{R}^d} \; f(x) = \mathbb{E}_\xi[f(x;\xi)] \approx \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{1}$$

where the expectation is taken with respect to the random variable $\xi \in \mathcal{S}$ which represents the data.

Given the recent scale of datasets which reach millions and billions [8], stochastic gradient descent (SGD) has emerged as the main workhorse in machine learning community due to its computational efficiency [5, 6, 29]. Specifically, SGD iterates as follows:

$$x_{t+1} = x_t - \eta \nabla f_{i_t}(x_t), \tag{2}$$

where η is the step size, and i_t is drawn uniformly at random from $\{1, \dots, n\}$. While computationally efficient, it is well-known that stochastic methods suffer from two major limitations: (i) slow convergence and (ii) numerical instability. For instance, due to the noise present in the approximated gradient, SGD could take longer to converge, in terms of number of iterations [10, 19]. Moreover, SGD suffers from numerical instability both in theory [20] and practice [5], allowing only a small range of the step size η (which usually depend on unknown quantities) that leads to convergence [19].

With respect to the slow convergence, many variants of accelerated methods have been proposed, most notably Polyak's momentum [23] and Nesterov's acceleration [1, 21]. These methods allow faster (sometimes optimal) convergence rates, while having virtually the same computational cost as SGD. In particular, SGD with momentum (SGDM) iterates as follows:

$$x_{t+1} = x_t - \eta \nabla f_{i_t}(x_t) + \beta(x_t - x_{t-1}), \tag{3}$$

where $\beta \in (0, 1)$ is the momentum parameter. While there are many other acceleration schemes, much of the state-of-the-art performance have been achieved with SGDM [13–15].

On the other hand, to address the numerical stability, variants of SGD that utilize proximal updates have recently been proposed [2, 3, 25–28]. In particular, [28] introduced stochastic proximal point algorithms (SPPA) and analyzed its convergence and stability, which iterates as follows:

$$x_{t+1}^{+} = \arg\min_{x\in\mathbb{R}^n} \left\{ f(x) + \tfrac{1}{2\eta}\|x - x_t\|_2^2 \right\} = x_t - \eta\nabla f(x_{t+1}^{+}) \tag{4}$$

$$x_{t+1} = x_{t+1}^{+} - \eta\varepsilon_{t+1}. \tag{5}$$

Without the stochastic errors ε_{t+1}, Eq. (4) is known as the proximal point algorithm (PPA) [11, 24] or the implicit gradient descent (IGD), and is known to converge with minimal assumption [4, 22] in deterministic setting.

In this work, we bridge the two paths and study the convergence and stability of stochastic PPA with momentum (SPPAM):

$$\begin{aligned} x_{t+1}^{+} &= \arg\min_{x\in\mathbb{R}^p} \left\{ f(x) + \tfrac{1}{2\eta}\|x - x_t\|_2^2 - \tfrac{\beta}{\eta}\langle x_t - x_{t-1}, x\rangle \right\} \\ &= x_t - \eta\nabla f(x_{t+1}^{+}) + \beta(x_t - x_{t-1}) \end{aligned} \tag{6}$$

$$x_{t+1} = x_{t+1}^{+} - \eta\varepsilon_{t+1}. \tag{7}$$

In particular, we study if adding momentum results in faster convergence akin to SGDM, while preserving the numerical stability inherited by utilizing proximal updates.

Apart from the empirical success of SGDM, we motivate the inclusion of momentum in SPPA (among many alternatives of acceleration schemes) through the following geometric interpretation. First, for large η, the algorithm is minimizing the original function $f(x)$. On the other hand, for small η, the algorithm not only tries to stay local by minimizing the quadratic term, but also tries to minimize the inner product between x and the vector from x_t to x_{t-1}. By the definition of inner product, this means that the new parameter x_{t+1}, on top of minimizing $f(x)$ and staying to close to x_t, also tries to move along the direction from x_{t-1} to x_t. This intuition exactly aligns with that of Polyak's momentum [23].

2. Related Work

PPA was introduced to convex programming in [24], and was popularized in [11]. In particular, [11] proved that for convex function $f(\cdot)$, PPA satisfies

$$f(x_T) - f(x^\star) \le O\left(\tfrac{1}{\sum_{t=1}^{T}\eta_t}\right) \text{ for any } T \ge 1. \tag{8}$$

As can be seen, by setting the step size η_t to be large, PPA can converge "arbitrarily" fast. Due to this remarkable convergence property, PPA was soon considered in stochastic setting. In [25],

	Deterministic
PPA [11] / IGD	$x_{t+1} = \arg\min_x \left\{ f(x) + \frac{1}{2\eta}\|x - x_t\|_2^2 \right\}$ $\Leftrightarrow x_{t+1} = x_t - \eta \nabla f(x_{t+1})$
Catalyst [17, 18]	$x_{t+1} \approx \arg\min_x \left\{ f(x) + \frac{\kappa}{2}\|x - y_t\|_2^2 \right\}$ $y_t = x_t + \beta_t (x_t - x_{t-1})$ where $\alpha_t^2 = (1-\alpha_t)\alpha_{t-1}^2 + \frac{\mu}{\mu+\kappa}\alpha_t, \quad \beta_t = \frac{\alpha_{t-1(1-\alpha_{t-1})}}{\alpha_{t-1}^2+\alpha_t}$
	Stochastic
SPPAM (this work)	$x_{t+1} = x_t - \eta(\nabla f(x_{t+1}) + \varepsilon_{t+1}) + \beta(x_t - x_{t-1})$
SPI [25] / ISGD [26, 27]	$x_{t+1} = \arg\min_x \left\{ f_{i_t}(x) + \frac{1}{2\eta}\|x - x_t\|_2^2 \right\}$ $\Leftrightarrow x_{t+1} = x_t - \eta \nabla f_{i_t}(x_{t+1})$
APROX [2]	Set $f_{i_t}(x) := \max\left\{ f_{i_t}(x_t) + \langle \nabla f_{i_t}(x_t), x - x_t \rangle, \inf_z f_{i_t}(z) \right\}$ from SPI
Stochastic Catalyst [16]	$x_{t+1} \approx \arg\min_x \left\{ f(x) + \frac{\kappa}{2}\|x - y_t\|_2^2 \right\}$ $y_t = x_t + \beta_t (x_t - x_{t-1})$ where $f(x) := f(y_t) + \langle g_t, x - y_t \rangle + \frac{\kappa+\mu}{2}\|x - y_t\|_2^2$

Table 1: Comparison of different algorithms in Section 2.

Stochastic version of PPA dubbed as stochastic proximal iterations (SPI) was analyzed, where an approximation of $f(\cdot)$ using a single data point $f_i(\cdot)$ was considered. Later, the same algorithm was (statistically) analyzed under the name of implicit stochastic gradient descent (ISGD) in [26, 27]. It was also analyzed recently in [2, 3, 16] where $f_i(\cdot)$ was further approximated with simpler surrogate functions. While settings considered under which differ slightly, these works generally point to the same message: in the asymptotic regime, SGD and SPI/ISGD have the same convergence behavior, but in the non-asymptotic regime, SPI/ISGD outperforms SGD thanks to numerical stability provided by utilizing proximal updates.

In terms of acceleration, in deterministic setting, accelerated PPA was first proposed in [12], where Nesterov's acceleration [21] was applied to Eq. (4). However, Nesterov's acceleration requires setting an adequate schedule for the momentum parameter β on every iteration, and as can be seen in Eq. (8), in practice one can already achieve arbitrarily fast convergence (assuming PPM can be implemented exactly). Hence, following works studied the conditions under which the proximal step in Eq. (4) can be computed inexactly, while still exhibiting some acceleration [17, 18]. This was later extended to the stochastic setting in [16]. Acceleration of stochastic PPA was also considered in [7] where $f_i(\cdot)$ was further approximated with auxiliary functions, but similarly to the aforementioned works, a convoluted 3-step acceleration scheme was required. We summarize these algorithms Table 1. To the best of our knowledge, this is the first work that considers directly applying Polyak's momentum to stochastic PPA following the geometric intuition outlined at the end of Section 1, and studies its convergence and stability properties.

3. Acceleration and Stability of SPPAM

3.1. Acceleration

Here, we characterize whether and when SPPAM enjoys faster convergence than SPPA for strongly convex functions. We start with the iteration invariant bound:

Theorem 1 *For μ-strongly convex $f(\cdot)$, SPPAM in Eq. (7) satisfies the following iteration invariant bound:*

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1}-x^\star\|_2^2\right] \leq \frac{1-\beta}{1+2\eta\mu}\mathbb{E}\left[\|x_t-x^\star\|_2^2\right] \\ + \frac{\beta^2}{1+2\eta\mu}\left(\frac{2-\beta}{2-\beta(1+\beta)}\right)\mathbb{E}\left[\|x_{t-1}-x^\star\|_2^2\right] + \eta^2\mathbb{E}\left[\|\varepsilon_{t+1}\|_2^2\right].\end{aligned}$$

Moreover, its contraction factor is upper bounded by the following quantity :

$$\frac{1-\beta}{2(1+2\eta\mu)} + \frac{1}{2}\cdot\sqrt{\left(\frac{1-\beta}{1+2\eta\mu}\right)^2 + \frac{\beta^2}{1+2\eta\mu}\left(\frac{2-\beta}{2-\beta(1+\beta)}\right)}. \tag{9}$$

Remark 2 *Notice that for $\beta = 0$, the above contraction factor reduces to $\frac{1}{1+2\eta\mu}$, which exactly matches that of SPPA for strongly convex objective in [28].*

Based on the contraction factor in (9), it is not immediately obvious when SPPAM enjoys faster contraction than SPPA in Eq. (5). We characterize this condition in the following corollary:

Corollary 3 *For μ-strongly convex $f(\cdot)$, SPPAM in Eq. (7) converges faster than SPPA in Eq. (5) if the following condition holds:*

$$\frac{\beta(2-\beta)}{2-\beta(1+\beta)} < \frac{4}{1+2\eta\mu}.$$

In words, for a fixed η and the constant μ, there is a range of momentum parameter β that exhibits acceleration compared to SPPA. We showcase this behavior using linear regression and Poisson regression in Figure 1.

3.2. Stability

In this section, we study the stability of SPPAM in Eq. (7). Preliminary result is summarized in the following theorem:

Theorem 4 *Initial conditions of SPPAM in Eq. (7), $\|x_0-x^\star\|_2^2$ and $\|x_{-1}-x^\star\|_2^2$, exponentially discounts after T iterations with the factor*

$$\tau^{-1}\cdot\left(\frac{1-\beta}{1+2\eta\mu}+\tau\right)^T, \quad \textit{where} \quad \tau = \sqrt{\frac{1-\beta}{1+2\eta\mu} + \frac{\beta^2}{1+2\eta\mu}\left(\frac{2-\beta}{2-\beta(1+\beta)}\right)}.$$

We want the above contraction factor to be in $(0,1)$, which can be easily achieved by setting η sufficiently large. We plot the discount factor for $\eta = \mu = 1$ in the top-left plot of Figure 2. We conjecture that the discount factor can be bounded by exponentially decreasing function; we leave this for future work.

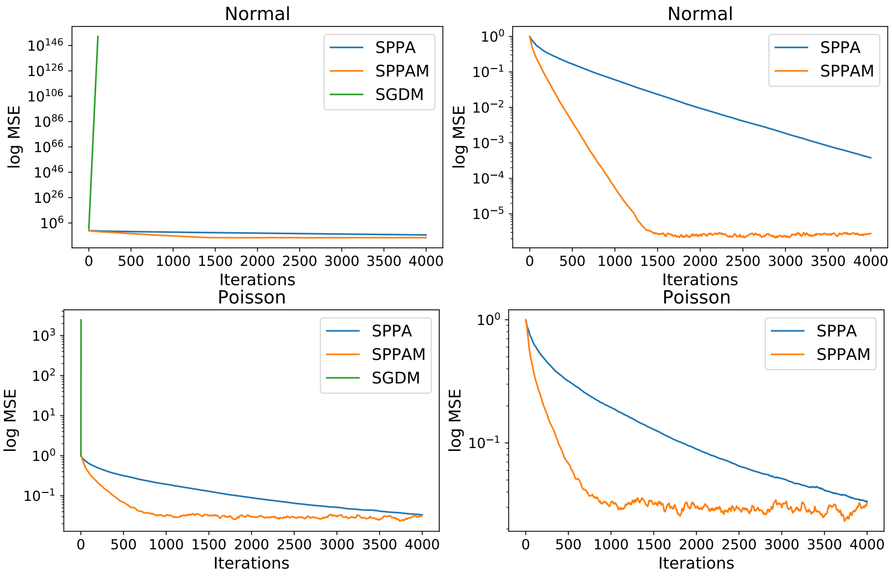

Figure 1: Illustration of acceleration and stability of SPPAM under linear/Poisson regressions. On the left panel, we plot SPPA in (5), SPPAM in (7), and SGDM in (3), all with the same constant step size (0.1 and 0.0001 for linear/Poisson regression respectively), batch size $= 10$, and $\beta = 0.8$ (when applicable). Note that SGDM diverges, exhibiting numerical instability. On the right panel, SPPA and SPPAM are plotted in the same setting, illustrating SPPAM's faster convergence. In both experiments, number of observations is 1000 while number of features is 100, with `1e-3` noise level.

3.3. Illustration of stability: quadratic model

In this section, for simplicity, we consider the quadratic optimization problem in deterministic setting, and derive the exact conditions that lead to convergence. Specifically, we consider the objective function

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x, \tag{10}$$

where the matrix $A \in \mathbb{R}^{n\times n}$ is positive semi-definite with eigenvalues $[\lambda_1, \ldots, \lambda_n]$. Below, we characterize the step size η and the momentum β that lead to convergence for different algorithms. Results for GD and GDM are from [9] but included for completeness.

Proposition 5 (GD [9]) *To minimize Eq. (10) with gradient descent, step size η needs to satisfy $0 < \eta < \frac{2}{\lambda_i}$, where λ_i is the i-th eigenvalue of A.*

Proposition 6 (PPA/IGD) *To minimize Eq. (10) with PPA, step size η needs to satisfy $\left|\frac{1}{1+\eta\lambda_i}\right| < 1$.*

Proposition 7 (GDM [9]) *To minimize Eq. (10) with gradient descent with momentum, step size η needs to satisfy $0 < \eta\lambda_i < 2 + 2\beta$ for $0 \le \beta \le 1$.*

Proposition 8 (PPAM) *Let $\delta_i = \left(\frac{\beta+1}{1+\eta\lambda_i}\right)^2 - \frac{4\beta}{1+\eta\lambda_i}$. To minimize Eq. (10) with PPA with momentum, step size η and momentum β need to satisfy:*

- $\eta > \frac{\beta-1}{\lambda_i} \quad if \;\; \delta_i \leq 0$
- $\frac{\beta+1}{1+\eta\lambda_i} + \sqrt{\delta_i} < 2 \quad if \;\; \delta_i > 0 \;\; and \;\; \frac{\beta+1}{1+\eta\lambda_i} \geq 0$
- $\frac{\beta+1}{1+\eta\lambda_i} - \sqrt{\delta_i} > -2 \quad if \;\; \delta_i > 0 \;\; and \;\; \frac{\beta+1}{1+\eta\lambda_i} < 0.$

Given above propositions, we can study the stability of different algorithms with respect to step size η and momentum β. Numerical simulation are illustrated in Figure 2, confirming our theory. In particular, for GD, only a small range of step size η leads to convergence (small white band); on the other hand, PPA/IGD converges in much wider choices of η. Similarly, GDM requires both η and β to be in a small region to converge, whereas PPAM converges in much wider choices of η and β; also note that the empirical convergent region (bottom-middle) almost exactly matches the region predicted by theory in Proposition 8.

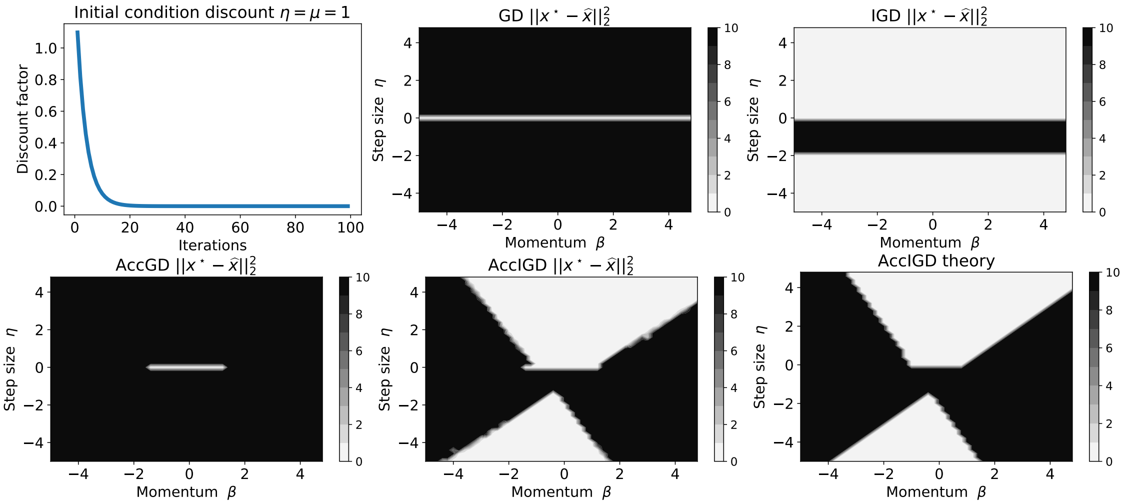

Figure 2: Top-Left: discount factor for $\eta = \mu = 1$ from Theorem 4; Rest: $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$ follow standard gaussian distribution. The condition number of A is 10. We sweep step size η and momentum μ from -5 to 5, and plot the final accuracy after 100 iterations. White region corresponds to convergence, and black region corresponds to divergence.

References

[1] Kwangjun Ahn. From Proximal Point Method to Nesterov's Acceleration. *arXiv:2005.08304 [cs, math]*, June 2020. URL http://arxiv.org/abs/2005.08304. arXiv: 2005.08304.

[2] Hilal Asi and John C. Duchi. Stochastic (Approximate) Proximal Point Methods: Convergence, Optimality, and Adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, January 2019. ISSN 1052-6234, 1095-7189. doi: 10.1137/18M1230323. URL http://arxiv.org/abs/1810.05633. arXiv: 1810.05633.

[3] Hilal Asi, Karan Chadha, and Gary Cheng. Minibatch Stochastic Approximate Proximal Point Methods. *34th Conference on Neural Information Processing Systems*, page 11, 2020.

[4] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[5] Léon Bottou. Stochastic Gradient Descent Tricks. In *Neural Networks: Tricks of the Trade*, volume 7700, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35288-1 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_25. URL http://link.springer.com/10.1007/978-3-642-35289-8_25. Series Title: Lecture Notes in Computer Science.

[6] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, January 2018. ISSN 0036-1445, 1095-7200. doi: 10.1137/16M1080173. URL https://epubs.siam.org/doi/10.1137/16M1080173.

[7] Karan Chadha, Gary Cheng, and John C. Duchi. Accelerated, Optimal, and Parallel: Some Results on Model-Based Stochastic Optimization. *arXiv:2101.02696 [cs, math, stat]*, January 2021. URL http://arxiv.org/abs/2101.02696. arXiv: 2101.02696.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[9] Gabriel Goh. Why momentum really works. *Distill*, 2017. doi: 10.23915/distill.00006. URL http://distill.pub/2017/momentum.

[10] Robert M Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. SGD: General Analysis and Improved Rates. *Proceedings of the 36 th International Conference on Machine Learning*, page 10, 2019.

[11] Osman Güler. On the Convergence of the Proximal Point Algorithm for Convex Minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, March 1991. ISSN 0363-0129. doi: 10.1137/0329022. URL https://epubs.siam.org/doi/10.1137/0329022. Publisher: Society for Industrial and Applied Mathematics.

[12] Osman Güler. New Proximal Point Algorithms for Convex Minimization. *SIAM Journal on Optimization*, 2(4):649–664, November 1992. ISSN 1052-6234. doi: 10.1137/0802032. URL https://epubs.siam.org/doi/abs/10.1137/0802032. Publisher: Society for Industrial and Applied Mathematics.

[13] 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.

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[16] Andrei Kulunchakov and Julien Mairal. A Generic Acceleration Framework for Stochastic Composite Optimization. *Advances in Neural Information Processing Systems*, 32, October 2019. arXiv: 1906.01164.

[17] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A Universal Catalyst for First-Order Optimization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3384–3392. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5928-a-universal-catalyst-for-first-order-optimization.pdf.

[18] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice. *Journal of Machine Learning Research*, 18:1–54, 2018.

[19] Eric Moulines and Francis R. Bach. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 451–459. Curran Associates, Inc., 2011.

[20] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, January 2009. ISSN 1052-6234, 1095-7189. doi: 10.1137/070704277. URL http://epubs.siam.org/doi/10.1137/070704277.

[21] Yurii Nesterov. *Lectures on Convex Optimization*. Springer Optimization and Its Applications. Springer International Publishing, 2 edition, 2018. ISBN 978-3-319-91577-7. doi: 10.1007/978-3-319-91578-4. URL https://www.springer.com/gp/book/9783319915777.

[22] Neal Parikh and Stephen Boyd. Proximal Algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, January 2014. ISSN 2167-3888, 2167-3918. doi: 10.1561/2400000003. URL https://www.nowpublishers.com/article/Details/OPT-003. Publisher: Now Publishers, Inc.

[23] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, January 1964. ISSN 0041-5553. doi: 10.1016/0041-5553(64)90137-5. URL http://www.sciencedirect.com/science/article/pii/0041555364901375.

[24] R. Tyrrell Rockafellar. Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, August 1976. ISSN 0363-0129. doi: 10.1137/0314056. URL https://epubs.siam.org/doi/abs/10.1137/0314056. Publisher: Society for Industrial and Applied Mathematics.

[25] Ernest K Ryu and Stephen Boyd. Stochastic Proximal Iteration: A Non-Asymptotic Improvement Upon Stochastic Gradient Descent. *Author website*, page 42, 2017.

[26] Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, August 2017. ISSN

0090-5364, 2168-8966. doi: 10.1214/16-AOS1506. Publisher: Institute of Mathematical Statistics.

[27] Panos Toulis, Jason Rennie, and Edoardo M Airoldi. Statistical analysis of stochastic gradient methods for generalized linear models. *International Conference on Machine Learning*, pages 667–675, 2014. URL http://proceedings.mlr.press/v32/toulis14.html.

[28] Panos Toulis, Thibaut Horel, and Edoardo M. Airoldi. The proximal Robbins–Monro method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1):188–212, 2021. ISSN 1467-9868. doi: 10.1111/rssb.12405. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12405.

[29] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.