Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains

Aymeric Dieuleveut aymeric.dieuleveut@epfl.ch
Ecole Polytechnique Fedérale de Lausanne
Alain Durmus durmus@cmla.ens-cachan.fr
Ecole Normale Supérieure de Cachan
Francis Bach francis.bach@ens.fr
Ecole Normale Supérieure de Paris, Inria

Abstract

We consider the minimization of an objective function given access to unbiased estimates of its gradient through stochastic gradient descent (SGD) with constant step-size. While the detailed analysis was only performed for quadratic functions, we provide an explicit asymptotic expansion of the moments of the averaged SGD iterates that outlines the dependence on initial conditions, the effect of noise and the step-size, as well as the lack of convergence in the general (non-quadratic) case. For this analysis, we bring tools from Markov chain theory into the analysis of stochastic gradient. We then show that Richardson-Romberg extrapolation may be used to get closer to the global optimum and we show empirical improvements of the new extrapolation scheme.

1 Introduction

We consider the minimization of an objective function given access to unbiased estimates of the function gradients. This key methodological problem has raised interest in different communities: in large-scale machine learning [4, 17, 18], optimization [10, 11], and stochastic approximation [6, 13, 16]. The most widely used algorithms are stochastic gradient descent (SGD), a.k.a. Robbins-Monro algorithm [15], and some of its modifications based on averaging of the iterates [13, 14, 19].

While the choice of the step-size may be done robustly in the deterministic case [see, *e.g.,* 3], this remains a traditional theoretical and practical issue in the stochastic case. Indeed, early work suggested to use step-size decaying with the number k of iterations as $O(1/k)$ [15], but it appeared to be non-robust to ill-conditioning and slower decays such as $O(1/\sqrt{k})$ together with averaging lead to both good practical and theoretical performance [1].

We consider in this paper constant step-size SGD, which is often used in practice. Although the algorithm is not converging in general to the global optimum of the objective function, constant step-sizes come with benefits: (a) there is single parameter value to set as opposed to the several choices of parameters to deal with decaying step-sizes, *e.g.,* as $1/(\square k + \triangle)^\circ$; the initial conditions are forgotten exponentially fast for well-conditioned (*e.g.,* strongly convex) problems [8, 9], and the performance, although not optimal, is sufficient in practice (in a machine learning set-up, being only 0.1% away from the optimal prediction often does not matter).

The main goals of this paper are (a) to gain a complete understanding of the properties of constant-step-size SGD in the strongly convex case, and (b) to propose provable improvements to get closer to the optimum when precision matters or in high-dimensional settings. We consider the iterates of the SGD recursion on $\mathbb{R}^d$ defined starting from $\theta_0 \in \mathbb{R}^d$, for $k \geq 0$, and a step-size $\gamma > 0$ by

$$\theta_{k+1}^{(\gamma)} = \theta_k^{(\gamma)} - \gamma\big[f'(\theta_k^{(\gamma)}) + \varepsilon_{k+1}(\theta_k^{(\gamma)})\big] \;, \tag{1}$$

where f is the objective function to minimize (in machine learning the generalization performance), $\varepsilon_{k+1}(\theta_k^{(\gamma)})$ the zero-mean statistically independent noise (in machine learning, obtained from a single i.i.d. observation of a data point). Following Bach and Moulines [2], we leverage the property that the sequence of iterates $(\theta_k^{(\gamma)})_{k\geq 0}$ is an *homogeneous Markov chain*.

This interpretation allows us to capture the general behavior of the algorithm. In the strongly convex case, this Markov chain converges exponentially fast to the unique stationary distribution π_γ highlighting the facts that

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

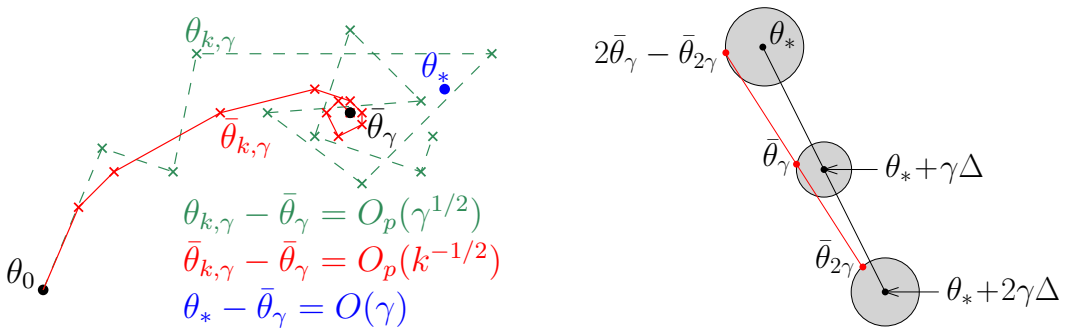

Figure 1: (Left) Convergence of iterates $\theta_k^{(\gamma)}$ and averaged iterates $\bar{\theta}_k^{(\gamma)}$ to the mean $\bar{\theta}_\gamma$ under the stationary distribution π_γ. (Right) Richardson-Romberg extrapolation, the disks are of radius $O(\gamma^2)$.

(a) initial conditions of the algorithms are forgotten quickly and (b) the algorithm does not converge to a point but oscillates around the mean of π_γ. See an illustration in Figure 1 (left). It is known that the oscillations of the non-averaged iterates have an average magnitude of $\gamma^{1/2}$ [12].

Consider the average process $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$ given for all $k \geq 0$ by $\bar{\theta}_k^{(\gamma)} = \frac{1}{k+1}\sum_{j=0}^k \theta_j^{(\gamma)}$. Then under appropriate conditions on the Markov chain $(\theta_k^{(\gamma)})_{k\geq 0}$, a central limit theorem on $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$ holds which implies that $\bar{\theta}_k^{(\gamma)}$ converges at rate $O(1/\sqrt{k})$ to $\bar{\theta}_\gamma := \int_{\mathbb{R}^d} \vartheta \, \mathrm{d}\pi_\gamma(\vartheta)$.

The deviation between $\bar{\theta}_k^{(\gamma)}$ and the global optimum θ_* is thus composed of a stochastic part $\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma$ and a deterministic part $\bar{\theta}_\gamma - \theta_*$.

For quadratic functions, it turns out that the deterministic part vanishes [2], that is, $\bar{\theta}_\gamma = \theta_*$ and thus averaged SGD with a constant step-size does converge. However, it is not true for general objective functions where we can only show that $\bar{\theta}_\gamma - \theta_* = O(\gamma)$, and this deviation is the reason why constant step-size SGD is not convergent.

The first main contribution of the paper is to provide an explicit asymptotic expansion that highlights all dependencies on initial conditions and noise variance, as achieved for least-squares [5], with an explicit decomposition into "bias" and "variance" terms: the bias term characterizes how fast initial conditions are forgotten and thus is increasing in a well-chosen norm of $\theta_0 - \theta_*$; while the variance term characterizes the effect of the noise in the gradient, independently of the starting point, and increases with the covariance of the noise.

Moreover, akin to weak error results for ergodic diffusions, we achieve a non-asymptotic weak error expansion in the step-size between π_γ and the Dirac at θ_*. Namely, we prove that for all functions $g : \mathbb{R}^d \to \mathbb{R}^q$, regular enough, $\int_{\mathbb{R}^d} g(\theta)\mathrm{d}\pi_\gamma(\theta) = g(\theta_*) + \gamma C + O(\gamma^2)$ for some $C \in \mathbb{R}^q$ independent of γ. Especially, for $g = \mathrm{Id}$, we get $\bar{\theta}_\gamma = \theta_* + \gamma\Delta + O(\gamma^2)$. Given this expansion, we can now use a very simple trick from numerical analysis, namely Richardson-Romberg extrapolation [20]: if we run two SGD recursions $(\theta_k^{(\gamma)})_{k\geq 0}$ and $(\theta_k^{(2\gamma)})_{k\geq 0}$ with the two different step-sizes γ and 2γ, then the averaged iterates $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$ and $(\bar{\theta}_k^{(2\gamma)})_{k\geq 0}$ will converge to $\bar{\theta}_\gamma$ and $\bar{\theta}_{2\gamma}$ respectively. Since $\bar{\theta}_\gamma = \theta_* + \gamma\Delta + O(\gamma^2)$ and $\bar{\theta}_{2\gamma} = \theta_* + 2\gamma\Delta + O(\gamma^2)$, for $\Delta \in \mathbb{R}^d$ independent of γ, the combined iterate $2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)}$ will converge to a point which is $\theta_* + O(\gamma^2)$ and we have thus gained one order in the convergence rate. See illustration in Figure 1 (right).

Contributions. Under simple assumptions, we prove the existence of a limit distribution π_γ, and convergence in distribution of $\theta_k^{(\gamma)}$ to π_γ at linear rate (§ 2.1). This allows to analyze and describe the position of $\bar{\theta}_\gamma$ with respect to θ_* (§ 2.2), and to provide an asymptotic expansion of the mean squared distance of the averaged SGD iterate to $\bar{\theta}_\gamma$, that outlines the dependence on initial conditions, the effect of noise and the step-size (§2.3). We illustrate empirical improvements of the extrapolation scheme on artificial datasets (§2.4).

2 Main results

Assumptions. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an objective function, satisfying the following assumptions: we assume the function f to be strongly convex with strong convexity constant μ, (*i.e.*, $f - \frac{\mu}{2}\|\cdot\|^2$ is convex); and that f is five times continuously differentiable with uniformly second to fifth bounded derivatives. Especially f is L-smooth: $\forall \theta \in \mathbb{R}^d$, the largest eigenvalue of $f''(\theta)$ is less than L.

Regarding the sequence of random functions $(\varepsilon_k)_{k\geq 1}$ (in Eq. (1)), we assume that there exists a filtration $(\mathcal{F}_k)_{k\geq 0}$ (*i.e.,* for all $k \in \mathbb{N}$, $\mathcal{F}_k \subset \mathcal{F}_{k+1}$) on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that for any $k \in \mathbb{N}$, for any $\theta \in \mathbb{R}^d$, $\varepsilon_{k+1}(\theta)$ is an $\mathcal{F}_{k+1}$-measurable random variable and $\mathbb{E}\left[\varepsilon_{k+1}(\theta)|\mathcal{F}_k\right] = 0$. In addition, $(\varepsilon_k)_{k\in\mathbb{N}^*}$ are independent and identically distributed (i.i.d.) random fields. Finally, $\varepsilon_k(\theta_*)$ admits bounded moments up to the order 4: $\mathbb{E}^{1/4}[\|\varepsilon_k(\theta_*)\|^4] < \infty$.

We observe a noisy gradient $f'_{k+1}(\theta_k^{(\gamma)}) = f'(\theta_k^{(\gamma)}) + \varepsilon_{k+1}(\theta_k^{(\gamma)})$ which is an unbiased estimator of f'. We assume that for any $k \in \mathbb{N}^*$, f'_k is almost surely L-co-coercive [22]: that is, for any $\eta, \theta \in \mathbb{R}^d$, $L\langle f'_k(\theta) - f'_k(\eta), \theta - \eta\rangle \geq \|f'_k(\theta) - f'_k(\eta)\|^2$.

Example: learning from i.i.d. observations. Our main motivation comes from machine learning; namely, we consider sets $\mathcal{X}, \mathcal{Y}$, and a convex loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$. The objective function is the generalization error $f_\ell(\theta) = \mathbb{E}_{X,Y}[\ell(X, Y, \theta)]$. Given $n \in \mathbb{N}$ i.i.d. observations $(x_k, y_k)_{k\in[\![1;n]\!]}$, for any $k \in [\![1; n]\!]$, we define $f_k(\cdot) = \ell(x_k, y_k, \cdot)$ the loss with respect to observation k. SGD then corresponds to following gradient of the loss on a single independent observation $(x_k, y_k)_{k\geq 1}$ at each step. Assumption on the noise is then satisfied with $\mathcal{F}_k := \sigma((x_j, y_j)_{1\leq j\leq k})$. In *least-squares regression*, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and the loss function is $\ell(X, Y, \theta) = (\langle X, \theta\rangle - Y)^2$, and the function f_ℓ is then quadratic. In *logistic regression*, $\ell(X, Y, \theta) = \log(1 + \exp(-Y\langle X, \theta\rangle))$.

2.1. Limit distribution. A first step is to prove the existence of a unique stationary distribution π_γ for the Markov chain $(\theta_k^{(\gamma)})_{k\geq 0}$ and convergence to π_γ. A fundamental tool in Markov chain theory is the *Markov kernel* [7], which is the equivalent for continuous spaces of the *transition matrix* in finite state spaces.

Markov kernel. We denote R_γ the *Markov kernel* associated with the SGD iterates $(\theta_k^{(\gamma)})_{k\geq 0}$. For all starting points $\theta \in \mathbb{R}^d$, for $k \in \mathbb{N}$, $\delta_\theta R_\gamma^k$ is the distribution of $\theta_k^{(\gamma)}$ starting at θ. Moreover, for a function $\varphi : \mathbb{R}^d \to \mathbb{R}^q$ we define $R_\gamma^k\varphi : \mathbb{R}^d \to \mathbb{R}^q$, such that $R_\gamma^k\varphi : \theta \mapsto \int_{\mathbb{R}^d} \varphi(\eta)\{\delta_\theta R_\gamma^k(\mathrm{d}\eta)\}$.

To show that $(\theta_k^{(\gamma)})_{k\geq 0}$ admits a unique stationary distribution π_γ and describe the convergence of $(\delta_\theta R_\gamma^k)_{k\geq 0}$ to π_γ, we use the Wasserstein distance W_2 [21].

Theorem 1. *For any step size $\gamma < L^{-1}$, the Markov chain $(\theta_k^{(\gamma)})_{k\geq 0}$, defined by the recursion (1), admits a unique stationary distribution π_γ, with finite second order moment. In addition for all $\theta \in \mathbb{R}^d$, $k \in \mathbb{N}$:*

$$W_2^2(\delta_\theta R_\gamma^k, \pi_\gamma) \leq (1 - 2\mu\gamma(1 - \gamma L))^k \int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 \,\mathrm{d}\pi_\gamma(\vartheta) \,.$$

To prove the existence of the limit, one shows that for any $\theta \in \mathbb{R}^d$, $(\delta_\theta R_\gamma^k)_{k\geq 0}$ is a Cauchy sequence in a particular Polish space. The existence of this limit allows to analyze the behavior of the chain under the limit distribution and the position of the limit point. Moreover, the speed of convergence allows so show the existence of solutions to the Poisson equation, that characterize the speed of convergence of the bias and variance term.

2.2. Behavior under the limit distribution, expansion of $\bar{\theta}_\gamma$ around θ_*, as $\gamma \to 0$. When f is a quadratic function, $\bar{\theta}_\gamma = \theta_*$. Indeed, since π_γ is invariant for $(\theta_k^{(\gamma)})_{k\geq 0}$, if $\theta_0^{(\gamma)}$ is distributed according to π_γ, then so is $\theta_1^{(\gamma)}$. Thus as $\theta_1^{(\gamma)} = \theta_0^{(\gamma)} - \gamma f'(\theta_0^{(\gamma)}) + \gamma\varepsilon_1(\theta_0^{(\gamma)})$ taking expectations on both sides leads to $\int_{\mathbb{R}^d} f'(\vartheta)\mathrm{d}\pi_\gamma(\vartheta) = 0$. For a quadratic function, the gradient is linear: $\int_{\mathbb{R}^d} f'(\vartheta)\mathrm{d}\pi_\gamma(\vartheta) = f'(\bar{\theta}_\gamma) = 0$, thus $\bar{\theta}_\gamma = \theta_*$. This highlights the crucial fact that for a quadratic function, the mean under the limit distribution is the optimal point, which explains why averaged least-mean squares algorithm does not saturate with constant step size [2]. If f is not quadratic but regular enough, we have the following first order development:

Theorem 2. *For $\gamma \to 0$, we have $\bar{\theta}_\gamma = \theta_* + \gamma\Delta + O(\gamma^2)$.*

Combining Theorems 1 and 2, we get that for γ small enough and all $k \geq 1$, $\mathbb{E}(\bar{\theta}_k^{(\gamma)} - \theta_*) = \frac{A(\theta_0,\gamma)}{k} + \gamma\Delta + O(\gamma^2) + O(e^{-k\mu\gamma})$. This expansion in the step size γ shows that a Richardson-Romberg extrapolation can be used to have better estimates of θ_*. Consider the average iterates $(\bar{\theta}_{2\gamma}^{(k)})_{k\geq 0}$ and $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$ associated with SGD with step size 2γ and γ respectively. Then

$$\mathbb{E}(2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)} - \theta_*) = \frac{2A(\theta_0, \gamma) - A(\theta_0, 2\gamma)}{k} + O(\gamma^2) + O(e^{-k\mu\gamma}) \,,$$

and therefore this very simple trick improves the convergence by a factor of γ. In practice, while the unaveraged gradient iterate $\theta_k^{(\gamma)}$ saturates rapidly, $\bar{\theta}_k^{(\gamma)}$ may already perform well enough to avoid saturation on real data-sets [2], and Richardson-Romberg extrapolated iterate $2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)}$ rarely reaches saturation.

2.3. Expansion for a given $\gamma > 0$ when k tends to $+\infty$. We also show that the convergence of $\bar{\theta}_k^{(\gamma)}$ to $\bar{\theta}_\gamma$, when $k \to \infty$, and the decay of $\mathbb{E}[\|\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\|^2]$ to 0 can be described very precisely. The expected squared distance decomposes as a sum of a bias term, that scales as k^{-2} and depends on the starting point θ_0, plus a variance term, that scales as k^{-1} and only depends on the asymptotic distribution π_γ, plus linearly decaying residual terms. The asymptotic bias and variance can be easily expressed as moments of solutions to several *Poisson equations*.

Poisson equation. For any (locally-) Lipschitz function $\varphi : \mathbb{R}^d \to \mathbb{R}^q$, there exists a function $\psi_\gamma : \mathbb{R}^d \to \mathbb{R}^q$ that satisfies $\pi_\gamma(\psi_\gamma) = 0$, and $(I - R_\gamma)\psi_\gamma = \varphi$. We call it *Poisson solution* associated with φ. Let ψ_γ, χ_γ^1 and χ_γ^2 be the Poisson solutions associated respectively to $\theta \mapsto \theta - \theta_*$, $\theta \mapsto \|\psi_\gamma(\theta)\|^2$ and $\theta \mapsto \|(\psi_\gamma - \varphi)(\theta)\|^2$.

Theorem 3 (Convergence of the Markov chain)**.** *Let $\gamma \in (0, 1/(2L))$. For any starting point $\theta_0 \in \mathbb{R}^d$, with $\rho := (1 - \gamma\mu)^{1/2}$:*

$$\mathbb{E}\left\|\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right\|^2 = \frac{\mathbb{E}_{\theta\sim\pi_\gamma}\left[\|\psi_\gamma(\theta)\|^2 - \|(\psi_\gamma - \varphi)(\theta)\|^2\right]}{k} + \frac{\|\psi_\gamma(\theta_0)\|^2 + \chi_\gamma^1(\theta_0) - \chi_\gamma^2(\theta_0)}{k^2} + O(\rho^k) \, .$$

When f_Σ is a quadratic function, it is possible, for any $\gamma > 0$, to compute ψ_γ and $\chi_\gamma^{1,2}$ explicitly; we exactly recover the result of Défossez and Bach [5].

2.4. Experiments. For the sake of illustration, we perform experiments on simulated data, for logistic regression, with $n = 10^7$ observations, for $d = 10$ (Fig. 2). We consider SGD with constant step-sizes $1/R^2$, $1/2R^2$ (and $1/4R^2$) with (plain lines) or without (dashed lines) averaging, with R^2 the smoothness constant. Without averaging , the chain saturates with an error proportional to γ . We consider Richardson Romberg iterates (red), which saturate at a much lower level, and performs much better than decaying step sizes (as $1/\sqrt{n}$) on the first iterations, as it forgets the initial conditions faster. We also propose an estimator that uses 3 different step sizes to perform a higher order interpolation. More precisely, we compute $\tilde{\theta}_k^3 := \frac{8}{3}\bar{\theta}_k^{(\gamma)} - 2\bar{\theta}_k^{(2\gamma)} + \frac{1}{3}\bar{\theta}_k^{(4\gamma)}$. With such an estimator, the *first 2* terms in the expansion vanish, scaling as γ and γ^2.

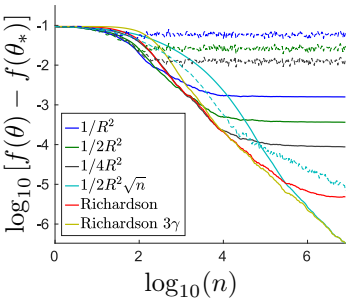

Figure 2

References

[1] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.*, 15(1):595–627, Jan. 2014.

[2] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[3] D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1995.

[4] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[5] A. Défossez and F. Bach. Averaged least-mean-squares: bias-variance trade-offs and optimal sampling distributions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2015.

[6] H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

[7] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[8] A. Nedić and D. Bertsekas. Convergence rate of incremental subgradient algorithms. In *Stochastic optimization: algorithms and applications*, pages 223–264. Springer, 2001.

[9] D. Needell, R. Ward, and N. Srebro. Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1017–1025. Curran Associates, Inc., 2014.

[10] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM J. on Optimization*, 19(4):1574–1609, Jan. 2009.

[11] Y. Nesterov and J. P. Vial. Confidence Level Solutions for Stochastic Programming. *Automatica*, 44(6):1559–1568, June 2008.

[12] G. C. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

[13] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

[14] A. Rakhlin, O. Shamir, and K. Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. *ArXiv e-prints*, Sept. 2011.

[15] H. Robbins and S. Monro. A stochastic approxiation method. *The Annals of mathematical Statistics*, 22(3):400–407, 1951.

[16] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[17] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2009.

[18] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 807–814, New York, NY, USA, 2007. ACM.

[19] O. Shamir and T. Zhang. Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes. *Proceedings of the 30^th International Conference on Machine Learning*, 2013.

[20] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*, volume 12. Springer Science & Business Media, 2013.

[21] C. Villani. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[22] D. L. Zhu and P. Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726, 1996.