A General Continuous-Time Formulation of Stochastic ADMM and Its Variants

Anonymous Author(s)
Affiliation
Address
email

Abstract

Stochastic versions of alternating direction method of multiplier (ADMM) and its variants play a key role in many modern large-scale machine learning problems. In this work, we introduce a unified algorithmic framework called generalized stochastic ADMM and investigate it via a continuous-time analylsis. The generalized framework widely include many stochastic ADMM variants including standard, linearized and gradient-based ADMM. Our continuous-time analysis provides with new insights on stochastic ADMM and variants. We show that its dynamics is approximated by a stochastic differential equations with small noise parameters. We rigorously proved that under some proper scaling, the trajectory of stochastic ADMM weakly converges to the trajectory of the stochastic differential equation. Our analysis also provides a simple proof of the empirical rule that the relaxation parameter must be between 0 and 2.

1 Introduction

For modern industrial scale machine learning problems, with the massive amount of data that are not only extremely large but often stored or even collected in a distributed manner, stochastic first-order methods almost become one of the default choices due to its excellent performance for online streaming large-scale dataset. Stochastic version of alternating direction method of multiplier (ADMM) algorithms is a popular approach to handle this distributed setting, especially for the regularized empirical risk minimization. Consider the following stochastic optimization problem:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad V(x) := f(x) + g(Ax) = \mathbb{E}_\xi\, f(x,\xi) + g(Ax), \tag{1}$$

where $f(x) = \mathbb{E}_\xi\, f(x,\xi)$ with $f(x,\xi)$ as the loss incurred on a sample ξ, $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$, $g : \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$, $A \in \mathbb{R}^{m\times d}$, and both f and g are convex and differentiable. The alternating direction method of multiplier (ADMM) [2] rewrites (1) as a constrained optimization problem

$$\begin{aligned} \underset{x\in\mathbb{R}^d, z\in\mathbb{R}^m}{\text{minimize}} \quad & \mathbb{E}_\xi\, f(x,\xi) + g(z) \\ \text{subject to} \quad & Ax - z = 0. \end{aligned} \tag{2}$$

Here and throughout the rest of the paper, we overload f to ease the notation, i.e. we adopt the two-argument $f(\cdot,\xi)$ for the stochastic instance and one-argument $f(\cdot)$ for its expectation. Note the classical setting of linear constraint $Ax + Bz = c$ can be reformulated as $z = Ax$ by a linear transformation when B is invertible. In the batch learning setting, $f(x)$ can be approximated by the empirical risk $f_{emp} = \frac{1}{N}\sum_{i=1}^N f(x,\xi_i)$. However, the computation cost for minimizing f_{emp} with a large amount of samples is significantly high and the efficiently is reasonably limited given the time and resource constraints. In the stochastic setting, at each step x_k is updated based on a small batch of samples (or even one) ξ_k instead of a large batch or full batch.

General Formulation for Stochastic ADMM Introducing stochasticity to ADMM [2] is in parallel to incorporating noise into gradient descent[55, 52, 42]. At iterate $k+1$, a sample ξ_{k+1} is randomly

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

drawn as an independent realization of ξ. Analogous to stochastic gradient descent, our **stochastic ADMM (sADMM)** introduced in [52, 42] performs the following updates:

$$x_{k+1} = \operatorname*{argmin}_x \left\{ f(x, \xi_{k+1}) + \frac{\rho}{2}\|Ax - z_k + u_k\|_2^2 \right\}, \tag{3a}$$

$$z_{k+1} = \operatorname*{argmin}_z \left\{ g(z) + \frac{\rho}{2}\|\alpha A x_{k+1} + (1-\alpha)z_k - z + u_k\|_2^2 \right\}, \tag{3b}$$

$$u_{k+1} = u_k + (\alpha A x_{k+1} + (1-\alpha)z_k - z_{k+1}). \tag{3c}$$

$\rho > 0$ is called the penalty parameter [2]. Here $\alpha \in (0, 2)$ is introduced as a *relaxation parameter* [2]. The algorithm is splitted into different regimes of α as follows:

- $\alpha = 1$, corresponding to the **standard stochastic ADMM**;
- $\alpha > 1$, corresponding to the **over-relaxed stochastic ADMM**;
- $\alpha < 1$, corresponding to the **under-relaxed stochastic ADMM**.

Historically, the above relaxation schemes were introduced and analyzed in [11], and experiments in [10, 12] suggest that over-relaxed regime of $\alpha > 1$ can improve the convergence rate. The acceleration phenomenon in over-relaxed ADMM has been discussed from the continuous perspective in deterministic setting [63, 16] , where relaxed ADMM algorithm accelerates by a factor of $\alpha \in (1, 2)$ in its convergence rate.

Since the emerging of ADMM, many variants have been introduced recently for solving a variety of optimization tasks. We focus on two stochastic ADMM variants to cater the need in application:

(i) In the linearized ADMM [20, 33, 43], (4a) replaces (3a) where

$$x_{k+1} := \operatorname*{argmin}_x \left\{ f(x, \xi_{k+1}) + \frac{\tau}{2}\left\| x - \left(x_k - \frac{\rho}{\tau} A^\top (A x_k - z_k + u_k) \right) \right\|_2^2 \right\}. \tag{4a}$$

In words, the augmented Lagrangian function is approximated by linearizing the quadratic term of x in (3a) plus the addition of a proximal term $\frac{\tau}{2}\|x - x_k\|_2^2$.

(ii) In the gradient-based ADMM [6, 54, 7, 32], (5a) replaces (3a) where

$$x_{k+1} := x_k - \frac{1}{\tau}\left(f'(x_k, \xi_{k+1}) + \rho A^\top (A x_k - z_k + u_k) \right). \tag{5a}$$

In words, in lieu to solving x-subproblem (3a) accurately, we apply one step of gradient descent with stepsize $1/\tau$.

In this paper we formulate a general scheme, called **Generalized stochastic ADMM (G-sADMM)**, to accommodate and unify all these variants of stochastic ADMM:

$$x_{k+1} = \operatorname*{argmin}_x \hat{\mathcal{L}}_{k+1}(x, z_k, u_k), \tag{6a}$$

$$z_{k+1} = \operatorname*{argmin}_z \left\{ g(z) + \frac{\rho}{2}\|\alpha A x_{k+1} + (1-\alpha)z_k - z + u_k\|_2^2 \right\}, \tag{6b}$$

$$u_{k+1} = u_k + (\alpha A x_{k+1} + (1-\alpha)z_k - z_{k+1}), \tag{6c}$$

where the objective $\hat{\mathcal{L}}_{k+1}$ in (6a) for x-subproblem is defined as

$$\begin{aligned} \hat{\mathcal{L}}_{k+1}(x, z_k, u_k) &= (1-\omega_1) f(x, \xi_{k+1}) + \omega_1 f'(x_k, \xi_{k+1})(x - x_k) \\ &\quad + (1-\omega)\cdot\frac{\rho}{2}\|Ax - z_k + u_k\|_2^2 + \omega\rho A^\top (A x_k - z_k + u_k)(x - x_k) + \frac{\tau}{2}\|x - x_k\|_2^2, \end{aligned} \tag{7}$$

with parameter $\rho, \tau \geq 0$, $\alpha \in (0, 2)$ and explicitness parameters $\omega_1, \omega \in [0, 1]$.

Main Results We study the continuous-time limit of the stochastic sequence $\{x_k\}$ defined in (6a) as $\rho \to \infty$. Following the recent line of work França et al. [14] [15] and Yuan et al. [63], we adopt a (sped-up) time rescaling of ρ^{-1}, namely the stepsize for (stochastic) ADMM, and hence we are in the small-stepsize regime as in (stochastic) gradient descent. Also, the proximal parameter τ in (7) grows linearly in ρ i.e. $\tau/\rho \to c$. In our setting of continuous limit theory, we focus on a fixed interval by T (referred to as "time"), corresponding to $\lfloor \rho T \rfloor$ steps (referred to as "step") in discrete time. Our main results can be concluded as follows:

(i) As $\rho \to \infty$ the sequence $\{x_{\lfloor \rho t \rfloor}\}$ for $1 \le t \le T$ admits a continuous limit $\{X_t : t \in [0, T]\}$, a stochastic process that solves the following stochastic differential equation (SDE):

$$\left[cI + \left(\frac{1}{\alpha} - \omega\right) A^\top A\right] dX_t = -\nabla V(X_t)dt + \sqrt{\frac{1}{\rho}}\sigma(X_t)dW_t, \tag{8}$$

where we recall from (1) that $V(x) = f(x) + g(Ax)$. The right hand of (8) consists of a deterministic part and a stochastic part, where $\sigma(x)$ is some diffusion matrix defined later in (9) and W_t is the standard Brownian motion in $\mathbb{R}^d$.[1] The continuous-limit approximation is rigorously characterized in weak convergence in the sense that applying a test function φ of a certain class, the expectation of $\varphi(x_k) - \varphi(X_{k/\rho})$ is uniformly convergent to zero in $k \in [0, \lfloor \rho T \rfloor]$ when ρ tends to infinity as stated in Theorem 1.[2] [3] The above SDE carries the small parameter $\sqrt{1/\rho}$ in its diffusion term and is also called *stochastic modified equation* (SME)[31], due to the historical reason in numerical analysis [38, 25].

(ii) We provide a continuous-time explanation of the effect of relaxation parameter $\alpha \in (0, 2)$ in stochatic ADMM, which is the first among stochastic ADMM work to our best knowledge. In light of our weak convergence of the discrete algorithm x_k, the residual $r_k = Ax_k - z_k$ roughly satisfies $r_{k+1} \approx (1 - \alpha)r_k$ and converges to 0 at a geometric rate $|1 - \alpha| < 1$. The rigorous statement is in Corollary 2.

Contributions Our contribution to this paper is the first continuous-time analysis of stochastic ADMM. First, we present a unified stochastic differential equation as the continuous-time model of variants of stochastic ADMM (standard ADMM, linearized ADMM, gradient-based ADMM) in the regime of large ρ under weak convergence. Second, we characterize the time evolution of $\mathrm{std}(x_k)$ and $\mathrm{std}(z_k)$ in their continuous-time counterparts and explicitly show that the standard deviation of stochastic ADMM variants has the scaling $\rho^{-1/2}$ in stochastic ADMM. Finally, our theoretical analysis of continuous model provides a simple justification of effect of relaxation parameter $\alpha \in (0, 2)$, and further sheds light on the principled choices of parameters c, α, ω.

2 Related Work

ADMM: ADMM is a widely used algorithm for solving problems with separable structures in machine learning, statistics, control etc. ADMM has a close connection with operator-splitting methods [8, 44, 18, 19]. But ADMM comes back to popularity due to several works like [5, 21, 56] and the highly influential survey paper [2]. Numerous modern machine learning applications are inspired by ADMM, for example, [40, 35, 62, 36, 51, 50, 45].

Variants of ADMM Linearized ADMM and gradient-based ADMM are widely used variants of ADMM. Linearized ADMM has been studied extensively, for example, in [3, 9, 22, 33, 34, 60–62, 65, 43]; gradient-based ADMM has been studied extensively too, for example, in [6, 54, 7, 32]. In this work, we present a general formulation for relaxed, linearized and gradient-based ADMM, and extended all of them to stochastic optimization.

Relaxation scheme for ADMM There are several important works studying the relaxation scheme of ADMM [10, 9, 11, 12] and propose to choose $\alpha \in (0, 2)$ with empirical suggestion for over relaxation $1 < \alpha < 2$. In this work, we give a simple and elegant proof of the theoretical reason that $\alpha \in (0, 2)$.

Continuous model for ADMM: The recent work in [14, 16] establishes the first deterministic continuous-time model for standard ADMM in the form of ordinary differential equation (ODE) for the smooth ADMM and [15, 63] extends its work to non-smooth ADMM, using the tool of differential inclusion, which is motivated by [53, 41].

Stochastic and online ADMM: The use of stochastic and online techniques for ADMM has recently drawn a lot of interest. [55] first proposed the online ADMM, which learns from only one sample (or a small mini-batch) at a time. [42, 52] proposed the variants of stochastic ADMM to attack the difficult nonlinear optimization problem inherent in $f(x, \xi)$ by linearization. Very recently, further accelerated algorithms for the stochastic ADMM have been developed in [67, 23].

[1] We refer the readers to Appendix A for backgrounds of SDE and the notion of weak convergence rate.

[2] We warn the reader that this is towards an orthogonal direction to the convergence of x_k to the optimizer x_* of the objective function as $k \to \infty$ for a *fixed* ρ,

[3] Appendix A is a summary of the background of the stochastic differential equation and the weak approximation, and the theory on the stochastic modified equation for interested readers.

Continuous models and analysis for stochastic optimization: The work in [49] is one seminal work of using continuous-time dynamical system to analyze discrete algorithms for optimization such as Nesterov's accelerated gradient method, with important extension to high resolution and symplectic structure in [47, 48, 24], to variational perspective in [59, 58]. The mathematical analysis of continuous formulation for stochastic optimization algorithms has become an important trend in recent years. We select an under-represented list out of numerous works: analyzing SGD from the perspective of Langevin MCMC [4, 66], analyzing stochastic mirror-descent [68], analyzing the SGD for the online algorithm of specific statistical models [28, 26, 30].

Stochastic modified equation: The mathematical connection between the stochastic gradient descent (SGD) and stochastic modified equation (SME) has been insightfully discovered in [29, 31]. This SME technique, originally arising from the numerical analysis of SDE [38, 25], has become the major mathematical tool for stochastic or online algorithms. The idea of using optimal control for stochastic continuous models to find adaptive parameters like step-size for stochastic optimization is used in [29] to controlled adaptive step size and momentum, and in [1] to improve the control on momentum, in [27] to choose batch size.

3 Weak Approximation to Stochastic ADMM

In this section, we show the weak approximation to the generalized family of stochastic ADMM variants (**G-sADMM**) (6). In the following sections, we define $\epsilon := \rho^{-1}$ to serve as an analog of stepsize for (stochastic) ADMM. The **G-sADMM** scheme (6) contains the relaxation parameter α, the parameter $c = \tau/\rho$ and the implicitness parameters ω, ω_1. This scheme (6) is very general and includes many existing variants as follows. It recovers all deterministic variants of ADMM for the setting $f(x,\xi) \equiv f(x)$. If $\omega_1 = \omega = \tau = 0$, then (6) is the standard stochastic ADMM (**sADMM**) or online ADMM [42, 67]. For the setting of parameters $\omega_1 = 0$, $\omega = 1$ and $c > 0$ it becomes stochastic version of the linearized ADMM for acceleration[9, 22, 65, 43]. If $\omega_1 = \omega = 1$ and $c > 0$, this scheme is the stochastic version of the gradient-based ADMM [6, 54, 7, 32]. In addition, the stochastic ADMM considered in [42] for non-smooth function corresponds to the setting here with $\alpha = 1, \omega_1 = 1, \omega = 0$ and $\tau = \tau_k \propto \sqrt{k}$.

Notations and Assumptions We use $\|\cdot\|$ to denote the Euclidean two norm if the subscript is not specified. And all vectors are referred to as column vectors. $f'(x,\xi)$, $g'(z)$ and $f''(x,\xi)$, $g''(z)$ refer to the first (gradient) and second (Hessian) derivatives w.r.t. x. The first assumption is **Assumption I**: $f(x)$, g and for each ξ, $f(x,\xi)$, are closed proper convex functions; A has full column rank.

Let $\mathcal{F}$ be the set of functions of at most polynomial growth. To apply the SME theory, we need the following **Assumptions II** on the smoothness:

(i) The second order derivative f'', g'' are uniformly bounded in x, and almost surely in ξ for $f''(x,\xi)$. $\mathbb{E}\,\|f'(x,\xi)\|_2^2$ is uniformly bounded in x.

(ii) $f(x)$, $f(x,\xi)$, $g(x)$ and the partial derivatives up to order 5 belong to $\mathcal{F}$.

(iii) $f'(x)$ and $f'(x,\xi)$ satisfy a uniform growth condition: $\|f'(x)\| + \|f'(x,\xi)\| \le C(1+\|x\|)$ for a constant C independent of ξ.

Main Theorem Given the noisy gradient $f'(x,\xi)$ and its expectation $f'(x) = \mathbb{E}_\xi\, f(x,\xi)$, we define the following matrix $\sigma(x) \in \mathbb{R}^{d\times d}$ by

$$\Sigma(x) = \sigma(x)\sigma(x)^\top = \mathbb{E}_\xi\left[\left(f'(x,\xi) - f'(x)\right)\left(f'(x,\xi) - f'(x)\right)^\top\right]. \tag{9}$$

Theorem 1 (SME for G-sADMM). *Let $\alpha \in (0,2)$, $\omega_1, \omega \in \{0,1\}$ and $c = \tau/\rho \ge 0$. Let $\epsilon = \rho^{-1} \in (0,1)$. $\{x_k\}$ denote the sequence of stochastic ADMM (6) with the initial choice $z_0 = Ax_0$. Define X_t as a stochastic process satisfying the SDE*

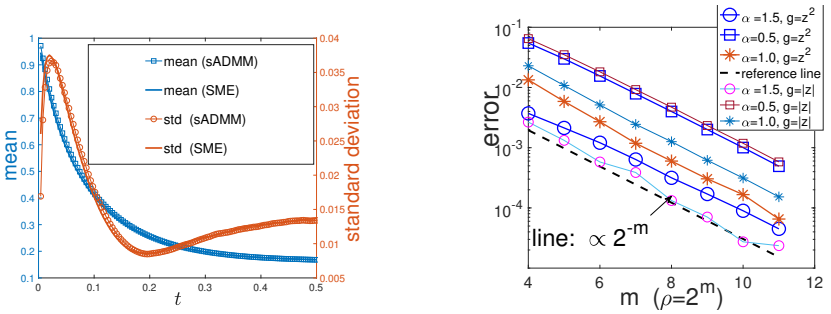
$$\widehat{M} dX_t = -\nabla V(X_t)dt + \sqrt{\epsilon}\sigma(X_t)dW_t \tag{10}$$

where the matrix

$$\widehat{M} := cI + \left(\frac{1}{\alpha} - \omega\right) A^\top A. \tag{11}$$

Then we have $x_k \to X_{k\epsilon}$ in weak convergence of order 1, with the following precise meaning: For any time interval $T > 0$ and for any test function φ such that φ and its partial derivatives up to order 4 belong to $\mathcal{F}$, there exists a constant C such that

$$|\mathbb{E}\,\varphi(X_{k\epsilon}) - \mathbb{E}\,\varphi(x_k)| \le C\epsilon, \quad \forall k \le \lfloor T/\epsilon \rfloor. \tag{12}$$



(a) The expectation (left axis) and standard deviation (right axis) of x_k and X_t. $\epsilon = 2^{-7}$. $\alpha = 1.5$.

(b) The weak convergence error err_m versus m for various α and ℓ_2, ℓ_1 regularization g.

Figure 1: The match between the stochastic ADMM and the SME and the verification of the first order weak approximation. The result is based on the average of 10^5 independent runs. The step size $\epsilon = 1/\rho = 2^{-m}T$ and $T = 0.5$ in (b). The details can be referred to in Appendix C.

The proof is in the next section. Here we highlight one useful corollary resulted from our proof of the theorem.

Corollary 2. *For generalized stochastic ADMM, when $\rho \to \infty$, the residual r_k satisfied the relation: $r_{k+1} = (1-\alpha)r_k$. From this relation, we can derive the condition on the relaxation parameter α for the convergence $r_k \to 0$ as $k \to \infty$: $|1-\alpha| < 1$, which matches the empirical range $\alpha \in (0, 2)$ proposed for standard ADMM with relaxation Boyd et al. [2].*

As an illustration of Theorem 1 on the first order weak convergence, Figure 1b verifies this rate for various α and even different g. To show that the SME does not only provide the expectation of the solution, but also provides the fluctuation of the numerical solution x_k for any given ϵ, Figure 1a plots the match of mean and standard deviation of X_t from the SME versus the x_k from the sADMM. One corollary of Theorem 1 is that the standard deviation of the stochastic ADMM x_k is $\mathcal{O}(\sqrt{\epsilon})$ and the rescaled two standard deviations $\epsilon^{-1/2}\,\mathrm{std}(x_k)$ and $\epsilon^{-1/2}\,\mathrm{std}(X_{k\epsilon})$ are close as the functions of the time $t_k = k\epsilon$. Other heuristics from Theorem 1 is about the continuous dynamics of the z_k variable and the residual r_k, which is discussed in Appendix B.

4 Proof of Main Results

Proof of Theorem 1. The ADMM scheme is the iteration of the triplet (x, z, λ) where $\lambda := \epsilon u$. In our proof, we shall show this triplet iteration can be effectively reduced to the iteration of x variable only in the form of $x_{k+1} = x_k + \epsilon\mathcal{A}(\epsilon, x_k, \xi_{k+1})$. The main approach for this reduction is the detailed analysis of the residual $r_k = x_k - Az_k$ below. This *one step difference* $x_{k+1} - x_k$ will play a central role in the theory of weak convergence, Milstein's theorem (Theorem 5 in Appendix A) [37].

For notational ease, we drop the random variable ξ_{k+1} in the scheme (6); the readers bear in mind that f and its derivatives do involve ξ and all conclusions hold almost surely for ξ.

The optimality conditions for the scheme (6) are

$$\begin{aligned}
&\omega_1\epsilon f'(x_k) + (1-\omega_1)\epsilon f'(x_{k+1}) + \epsilon A^\top\lambda_k \\
&\qquad + A^\top\left(\omega Ax_k + (1-\omega)Ax_{k+1} - z_k\right) + c(x_{k+1} - x_k) = 0 && \text{(13a)}\\
&\epsilon g'(z_{k+1}) = \epsilon\lambda_k + \alpha Ax_{k+1} + (1-\alpha)z_k - z_{k+1} && \text{(13b)}\\
&\epsilon\lambda_{k+1} = \epsilon\lambda_k + \alpha Ax_{k+1} + (1-\alpha)z_k - z_{k+1} && \text{(13c)}
\end{aligned}$$

Note that due to (13b) and (13c), the last condition (13c) can be replaced by $\lambda_{k+1} = g'(z_{k+1})$. So, without loss of generality, one can assume that $\lambda_k \equiv g'(z_k)$ for any integer k. The optimality conditions (13) now can be written only in the pair (x, z):

$$\begin{aligned}
&\omega_1\epsilon f'(x_k) + (1-\omega_1)\epsilon f'(x_{k+1}) + \epsilon A^\top g'(y_k) \\
&\qquad + A^\top\left(\omega Ax_k + (1-\omega)Ax_{k+1} - z_k\right) + c(x_{k+1} - x_k) = 0 && \text{(14a)}\\
&\epsilon g'(z_{k+1}) - \epsilon g'(z_k) = \alpha Ax_{k+1} + (1-\alpha)z_k - z_{k+1} && \text{(14b)}
\end{aligned}$$

We think of (x_{k+1}, z_{k+1}) in (14) as the function of x_k, z_k and ϵ and seek the asymptotic expansion of this function in the regime of small ϵ, by using Dominant Balance technique (see, e.g. [57]) with

the Taylor expansion for $f'(x_{k+1})$ and $g'(z_{k+1})$ around x_k and z_k, respectively. The expansion of the one step difference up to the order ϵ is

$$x_{k+1} - x_k = -M^{-1}A^\top r_k - \epsilon M^{-1}\left[f'(x_k) + A^\top g'(z_k)\right] - \epsilon N_k A^\top r_k + \epsilon^2 c_k, \tag{15a}$$

$$z_{k+1} - z_k = (1 - \epsilon g''(z_k))\alpha(A(x_{k+1} - x_k) + r_k) + \epsilon^2 c_k', \tag{15b}$$

where r_k is the *residual*

$$r_k := Ax_k - z_k \tag{16}$$

and the matrix M is

$$M := cI + (1-\omega)A^\top A. \tag{17}$$

I is the identity matrix. N_k is a matrix depending on the hessian $f''(x_k)$ and M. c_k and c_k' are uniformly bounded in k (and in random element ξ) due to the bound of functions f'' and g'' in **Assumption II**. Throughout the rest of the paper, we shall use the notation $\mathcal{O}(\epsilon^p)$ to denote these terms bounded by a (nonrandom) constant multiplier of ϵ^p, for $p = 1, 2, \ldots$.

Equation (15) clearly shows that the dynamics of the pair (x_{k+1}, z_{k+1}) depends on the residual $r_k = Ax_k - z_k$ in the previous step. In the limit of $\epsilon = 0$, the one step difference (i.e. *truncation error*) $x_{k+1} - x_k \to -M^{-1}A^\top r_k$ which does not vanish unless r_k is exact zero. We next turn to the analysis of the residual first. (15b) together with (15a) implies $z_{k+1} - z_k = -\alpha A M^{-1} A^\top r_k + \alpha r_k + \mathcal{O}(\epsilon)$ by absorbing ϵ order terms into $\mathcal{O}(\epsilon)$. By using (14b) for Ax_{k+1} , we have the recursive relation for residual r_{k+1}:

$$\begin{aligned} r_{k+1} = Ax_{k+1} - z_{k+1} &= \left(\frac{1}{\alpha} - 1\right)(z_{k+1} - z_k) + \frac{\epsilon}{\alpha}(g'(z_{k+1}) - g'(z_k)) \\ &= (1-\alpha)\left(I - AM^{-1}A^\top\right) r_k + \mathcal{O}(\epsilon). \end{aligned} \tag{18}$$

Since $r_0 = 0$ at the initial step by setting $z_0 = Ax_0$, then (18) shows that $r_1 = Ax_1 - y_1$ become $\mathcal{O}(\epsilon)$ after one iteration. Next, we have the following important property, particularly with assumption $\alpha = 1$, we have a stronger result than (18).

Proposition 3. *For any integer $k \geq 0$, if $r_k = \mathcal{O}(\epsilon)$, then*

$$r_{k+1} = (1 - \alpha + \epsilon\alpha g''(z_k))(r_k + A(x_{k+1} - x_k)) + \mathcal{O}(\epsilon^3). \tag{19}$$

If $\alpha = 1$, equation (19) reduces to the second order in ϵ:

$$r_{k+1} = \epsilon\alpha g''(z_k)(r_k + A(x_{k+1} - x_k)) + \mathcal{O}(\epsilon^3) = \mathcal{O}(\epsilon^2). \tag{20}$$

Proof of Proposition 3. Since $r_k = Ax_k - z_k = \mathcal{O}(\epsilon)$, then the one step difference $x_{k+1} - x_k$ and $z_{k+1} - z_k$ are both at order $\mathcal{O}(\epsilon)$ because of (15a) and (15b). We solve $\delta z := z_{k+1} - z_k$ from (14b) by linearizing the implicit term $g'(z_{k+1})$ and use the assumption that the third order derivative of g exits:

$$\epsilon g''(z_k)\delta z + \epsilon\mathcal{O}((\delta z)^2) + \delta z = \alpha(r_k + A\delta x).$$

where $\delta x := x_{k+1} - x_k = \mathcal{O}(\epsilon)$. Then since $\mathcal{O}((\delta z)^2) = \mathcal{O}(\epsilon^2)$, the expansion of $\delta z = z_{k+1} - z_k$ in ϵ is

$$z_{k+1} - z_k = \alpha(1 - \epsilon g''(z_k))\big(r_k + A(x_{k+1} - x_k)\big) + \mathcal{O}(\epsilon^3) \tag{21}$$

Then

$$\begin{aligned} r_{k+1} &= r_k + A(x_{k+1} - x_k) - (z_{k+1} - z_k) \\ &= \Big(1 - \alpha + \epsilon\alpha g''(z_k)\Big)(r_k + A(x_{k+1} - x_k)) + \mathcal{O}(\epsilon^3) \end{aligned}$$

□

We now focus on (15a) for $x_{k+1} - x_k$, which is rewritten

$$\begin{aligned} M(x_{k+1} - x_k) = &-A^\top r_k - \epsilon\left(f'(x_k) + A^\top g'(z_k)\right) \\ &+ \epsilon^2(1-\omega_1)f''(x_k)M^{-1}\left(f'(x_k) + A^\top g'(z_k) + \frac{1}{\epsilon}A^\top r_k\right) + \mathcal{O}(\epsilon^3). \end{aligned} \tag{22}$$

This expression does not contain the parameter α explicitly, but the residual $r_k = Ax_k - y_k$ significantly depends on α (see Proposition 3). If $\alpha = 1$, then r_k is on the order of ϵ^2, which leads to

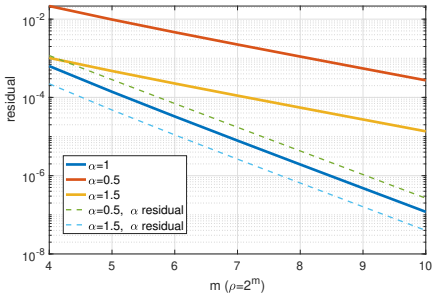


Figure 2: Illustration of the order of the residual r_k and the α-residual r_k^α defined in (23).

useful fact that there is no contribution from r_k to the weak approximation of x_k at the order 1. But for the relaxation case where $\alpha \neq 1$, r_k contains the first order term coming from $z_{k+1} - z_k$.

To obtain a second order for some "residual" for the relaxed scheme where $\alpha \neq 1$, we need a new definition, α-residual, to account for the gap induced by α. Motivated by (13b), we first define

$$r_{k+1}^\alpha := \alpha A x_{k+1} + (1-\alpha) z_k - z_{k+1}. \tag{23}$$

It is connected to the original residual r_{k+1} and r_k since it is easy to check that

$$r_{k+1}^\alpha = \alpha r_{k+1} + (\alpha - 1)(z_{k+1} - z_k) = \alpha r_k + \alpha A(x_{k+1} - x_k) - (z_{k+1} - z_k). \tag{24}$$

Obviously, when $\alpha = 1$, this α-residual r^α is the original residual $r = Ax - y$. In our proof, we need a modified α-residual, denoted by

$$\widehat{r}_{k+1}^\alpha := \alpha r_k + (\alpha - 1)(z_{k+1} - z_k) \tag{25}$$

We shall show that both r_{k+1}^α and $\widehat{r}_{k+1}^\alpha$ are as small as $\mathcal{O}(\epsilon^2)$:

$$r_{k+1}^\alpha = \mathcal{O}(\epsilon^2) \text{ and } \widehat{r}_{k+1}^\alpha = \mathcal{O}(\epsilon^2).$$

In fact, by the second equality of (24), (21) becomes $z_{k+1} - z_k = (1 - \epsilon g''(z_k))(r_{k+1}^\alpha + z_{k+1} - z_k) + \mathcal{O}(\epsilon^3)$, so,

$$r_{k+1}^\alpha = \epsilon(1 + \epsilon g''(z_k)) g''(z_k)(z_{k+1} - z_k) + \mathcal{O}(\epsilon^3)$$

which is $\mathcal{O}(\epsilon^2)$ since $z_{k+1} - z_k = \mathcal{O}(\epsilon)$.

The difference between $(z_{k+1} - z_k)$ and $(z_{k+2} - z_{k+1})$, is at the order ϵ^2 due to truncation error of the central difference scheme. Then we have the conclusion

$$\widehat{r}_{k+1}^\alpha = \alpha r_k + (\alpha - 1)(z_{k+1} - z_k) = \mathcal{O}(\epsilon^2) \tag{26}$$

by shifting the subscript k by one.

By (25) and the above proposition, we have $r_k = (\frac{1}{\alpha} - 1)(z_{k+1} - z_k) + \mathcal{O}(\epsilon^2)$. Furthermore, due to (21), $r_k = (\frac{1}{\alpha} - 1)(z_{k+1} - z_k) + \mathcal{O}(\epsilon^2) = (1-\alpha)(r_k + A(x_{k+1} - x_k)) + \mathcal{O}(\epsilon^2)$ which gives

$$r_k = \left(\frac{1}{\alpha} - 1\right)(z_{k+1} - z_k) + \mathcal{O}(\epsilon^2) = \left(\frac{1}{\alpha} - 1\right) A(x_{k+1} - x_k) + \mathcal{O}(\epsilon^2) \tag{27}$$

and it follows $z_{k+1} - z_k = A(x_{k+1} - x_k) + \mathcal{O}(\epsilon^2)$.

Remark 1. *As an illustration, the figure above for a toy example (see numerical example part in Appendix C) shows that $r_k \sim \mathcal{O}(\epsilon)$ for $\alpha \neq 1$ while $r_k \sim \mathcal{O}(\epsilon^2)$ for $\alpha = 1$ and the α-residual r_k^α is $\mathcal{O}(\epsilon^2)$ regardless of α (the solid lines are r_k and the dashed lines are r_k^α).*

With the above preparations for the residual analysis, we now apply Theorem 5 to show the main conclusion of our theorem. Combining equation (22) and 27, and noting the Taylor expansion of $g'(z_k)$: $g'(y_k) = g'(Ax_k - r_k) = g'(Ax_k) + \mathcal{O}(\epsilon)$ since $r_k = \mathcal{O}(\epsilon)$, and now putting back the random element ξ into f', we have

$$M(x_{k+1} - x_k) = -\epsilon \left(f'(x_k, \xi_{k+1}) + A^\top g'(Ax_k)\right) - \left(\frac{1}{\alpha} - 1\right) A^\top A(x_{k+1} - x_k) + \mathcal{O}(\epsilon^2) \tag{28}$$

For convenience, introduce the matrix

$$\widehat{M} := M + \frac{1-\alpha}{\alpha} A^\top A = c + \left(\frac{1}{\alpha} - \omega\right) A^\top A. \tag{29}$$

and let $\widehat{x}_k := \widehat{M} x_k$, and $\delta\widehat{x}_{k+1} = \widehat{M}(x_{k+1} - x_k)$ Then

$$\delta\widehat{x} = -\epsilon V'(x,\xi) + \epsilon^2 \left((1-\omega_1) f'' M^{-1} V'(x) - A^\top \theta\right) + \mathcal{O}(\epsilon^3).$$

The final step is to compute the momentums in the Milstein's theorem Theorem 5 as follows.

(i)

$$\mathbb{E}[\delta\widehat{x}] = -\epsilon\, \mathbb{E}\, V'(x,\xi) + \mathcal{O}(\epsilon^2) = -\epsilon V'(x) + \mathcal{O}(\epsilon^2) \tag{30}$$

(ii)

$$\begin{aligned}
\mathbb{E}[\delta\widehat{x}\,\delta\widehat{x}^\top] &= \epsilon^2\, \mathbb{E}\left(\left[f'(x,\xi) + A^\top g'(x)\right]\left[f'(x,\xi)^\top + g'(x)^\top A\right]\right) + \mathcal{O}(\epsilon^3) \\
&= \epsilon^2 \left(V'(x)V'(x)^\top\right) - \epsilon^2 \left(f'(x) + A^\top g'(x))(f'(x)^\top + g'(x)^\top A)\right) \\
&\quad + \epsilon^2\, \mathbb{E}\left(\left[f'(x,\xi) + A^\top g'(x)\right]\left[f'(x,\xi)^\top + g'(x)^\top A\right]\right) + \mathcal{O}(\epsilon^3) \\
&= \epsilon^2 \left(V'(x)V'(x)^\top\right) + \epsilon^2\, \mathbb{E}\left[\left(f'(x,\xi) - f'(x)\right)\left(f'(x,\xi) - f'(x)\right)^\top\right] + \mathcal{O}(\epsilon^3)
\end{aligned}$$

(iii) It is trivial that $\mathbb{E}[\Pi_{j=1}^s \delta x_{i_j}] = \mathcal{O}(\epsilon^3)$ for $s \geq 3$ and $i_j = 1, \ldots, d$.

So, Theorem 1 is proved.

□

Proof of Corollary 2. Based on Proposition 3, $r_{k+1} = (1-\alpha)(r_k + A(x_{k+1} - x_k)) + \epsilon\alpha g''(z_k)(r_k + A(x_{k+1} - x_k)) + \mathcal{O}(\epsilon^3)$. Since g'' are bounded and $x_{k+1} - x_k = \mathcal{O}(\epsilon)$, then $r_{k+1} = (1-\alpha)r_k + \mathcal{O}(\epsilon)$, with the leading term $(1-\alpha)r_k$. Therefore, when $\epsilon \to 0$, we need $|1-\alpha| < 1$ for $|r_k|$ to converge to zero when $k \to \infty$. □

5 Conclusion and Discussion

In this paper, we have developed the stochastic continuous dynamics in the form of stochastic modified equation (SME) to analyze the dynamics of a general family of stochastic ADMM, including the standard, linearized and gradient-based ADMM with relaxation α. Our continuous model (10) provides a unified framework to describe the dynamics of stochastic ADMM algorithms and particularly is able to quantify the fluctuation effect for a large penalty parameter ρ. Our analysis here generalize the existing works [14, 63] in the form of ordinary differential equation, but our results intrinsically encodes the impact of the noise in the stochastic ADMM. As the first order approximation to the stochastic ADMM trajectory, the solution to the continuous model can precisely describe the mean and the standard deviation (fluctuation) in stochastic trajectory.

One distinctive feature between the ADMM and its stochastic or online variants is highly similar to that between gradient descent and stochastic gradient descent [39, 29, 46]: there exists a transition time t_* after which the fluctuation (with a typical scale $\rho^{-1/2}$) starts to dominate the "drift" term, which means the traditional acceleration methods in deterministic case will fail to perform well, despite of their prominent performance in the early stage of training when the drift suppresses the noise. In our Appendix D, we briefly discuss a few new adaptive strategies such as for parameters ρ_t, τ_t, the batch size B_t and even relaxation parameters α_t, via the principle of control theory for the continuous model we obtained. For example, a large α (over-relaxation) is preferred before the transition time while a small α (under-relaxation) may be preferred later to help reduce the variance.

About the possible further development in theory, we note that our generalized family of ADMM methods is still restricted to the linearized ADMM and the gradient-based ADMM. Recently, there emerge many new efficient acceleration methods for stochastic ADMM such as [45, 67, 23, 42]. A potential future work is to extend our stochastic analysis to these new schemes for the understanding at the continuous level, even with non-smooth functions.

Broader Impact

Our work is a theoretic exploration of the existing and popular stochastic optimization with the ADMM. The continuous viewpoint to understand many machine learning algorithms is a powerful tool with a unifying conciseness and elegance, to complement the pure discrete-level analysis. It offers a unique bridge in theory between the applied mathematics and machine learning. Such a fusion between discrete world and continuum world can boost the beneficial communication between the algorithmic community and the applied math community. For ethical aspects and future societal consequences, the positive part is that large amount of machine learning algorithms are trained using ADMM and stochastic ADMM, our work would shed light for algorithm designing from theoretical point of view, that might lead to more efficient training algorithm that save the energy consumed in training from parallel search of training hyper-parameters. Due to the theoretical nature of our work, there are not much negative impact, the only related negative impact might come from algorithmic automation that might lead to unemployment of workers.

References

[1] Wangpeng An, Haoqian Wang, Qingyun Sun, Jun Xu, Qionghai Dai, and Lei Zhang. A pid controller approach for stochastic optimization of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8522–8531, 2018.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[3] G. Chen and M. Teboulle. A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64:81–101, 1994.

[4] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

[5] P. L. Combettes and Jean-Christophe Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):564–574, 2007.

[6] L. Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optimization Theory and Applications*, 158(2): 460–479, 2013.

[7] D. Davis and W. Yin. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4):829–858, 2017.

[8] J. Douglas and H. H. Rachford. On the numerical solution of the heat conduction problem in 2 and 3 space variables. *Transactions of the American Mathematical Society*, 82:421–439, 1956.

[9] J. Eckstein. Some saddle-function splitting methods for convex programming. *Optimization Methods and Software*, 4(1):75–83, 1994.

[10] Jonathan Eckstein. Parallel alternating direction multiplier decomposition of convex programs. *Journal of Optimization Theory and Applications*, 80(1):39–62, 1994.

[11] Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55 (1-3):293–318, 1992.

[12] Jonathan Eckstein and Michael C Ferris. Operator-splitting methods for monotone affine variational inequalities, with a parallel application to optimal control. *INFORMS Journal on Computing*, 10(2):218–235, 1998.

[13] W H Fleming and R W Rishel. *Deterministic and Stochastic Optimal Control*. Stochastic Modelling and Applied Probability. Springer New York, 1975. doi: 10.1007/978-1-4612-6380-7.

[14] Guilherme França, Daniel P Robinson, and René Vidal. ADMM and accelerated ADMM as continuous dynamical systems. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1559–1567, 2018.

[15] Guilherme Franca, Daniel P Robinson, and René Vidal. A dynamical systems perspective on nonsmooth constrained optimization. *arXiv preprint arXiv:1808.04048*, 2018.

[16] Guilherme França, Daniel P Robinson, and René Vidal. Relax, and accelerate: A continuous perspective on ADMM. *arXiv preprint arXiv:1808.04048*, 2018.

[17] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, New York, 3 edition, 2012.

[18] Daniel Gabay and Bertrand Mercier. *A dual algorithm for the solution of non linear variational problems via finite element approximation*. Institut de recherche d'informatique et d'automatique, 1975.

[19] Roland Glowinski and A Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2): 41–76, 1975.

[20] Donald Goldfarb, Shiqian Ma, and Katya Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming*, 141(1-2):349–382, 2013.

[21] T. Goldstein and S. Osher. The split Bregman method for L1-regularized problems. *SIAM J. Imaging Sci.*, 2:323–343, 2009.

[22] B. S. He, L. Liao, D. Han, and H. Yang. A new inexact alternating direction method for monotone variational inequalities. *Mathematical Programming*, 92:103–118, 2002.

[23] Feihu Huang, Songcan Chen, and Heng Huang. Faster stochastic alternating direction method of multipliers for nonconvex optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2839–2848, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/huang19a.html.

[24] Michael I Jordan. Dynamical, symplectic and stochastic perspectives on gradient-based optimization. *University of California, Berkeley*, 2018.

[25] P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer, New York, corrected edition, 2011. ISBN 9783662126165. URL https://books.google.com.hk/books?id=r9r6CAAAQBAJ.

[26] Chris Junchi Li, Zhaoran Wang, and Han Liu. Online ICA: Understanding global dynamics of nonconvex optimization via diffusion processes. In *Advances in Neural Information Processing System*, pages 4967–4975, 2016.

[27] Chris Junchi Li, Lei Li, Junyang Qian, and Jian-Guo Liu. Batch size matters: A diffusion approximation framework on nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

[28] Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Diffusion approximations for online principal component estimation and global convergence. In *Advances in Neural Information Processing System*, 2017.

[29] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *34th International Conference on Machine Learning, ICML 2017*, 34th International Conference on Machine Learning, ICML 2017, pages 3306–3340. International Machine Learning Society (IMLS), 1 2017.

[30] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. *arXiv preprint arXiv:1511.06251v3*, 2017.

[31] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

[32] T. Lin, S. Ma, and S. Zhang. An extragradient-based alternating direction method for convex minimization. *Foundations of Computational Mathematics*, 17(1):35–59, 2017.

[33] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in neural information processing systems*, pages 612–620, 2011.

[34] S. Ma. Alternating proximal gradient method for convex minimization. *Journal of Scientific Computing*, 68(2):546–572, 2016.

[35] Morteza Mardani, Qingyun Sun, David Donoho, Vardan Papyan, Hatef Monajemi, Shreyas Vasanawala, and John Pauly. Neural proximal gradient descent for compressive imaging. In *Advances in Neural Information Processing Systems*, pages 9596–9606, 2018.

[36] Morteza Mardani, Qingyun Sun, Vardan Papyan, Shreyas Vasanawala, John Pauly, and David Donoho. Degrees of freedom analysis of unrolled neural networks. *arXiv preprint arXiv:1906.03742*, 2019.

[37] G. N. Milstein. Weak approximation of solutions of systems of stochastic differential equations. *Theory of Probability & Its Applications*, 30(4):750–766, 1986. doi: 10.1137/1130095. URL https://doi.org/10.1137/1130095.

[38] G.N. Milstein. *Numerical Integration of Stochastic Differential Equations*, volume 313 of *Mathematics and Its Applications*. Springer, 1995. ISBN 9780792332138. URL https://books.google.com.hk/books?id=o2y80r_a4W0C.

[39] 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. URL http://papers.nips.cc/paper/4316-non-asymptotic-analysis-of-stochastic-approximation-algorithms-for-machine-learning.pdf.

[40] Robert Nishihara, Laurent Lessard, Benjamin Recht, Andrew Packard, and Michael I Jordan. A general analysis of the convergence of admm. *arXiv preprint arXiv:1502.02009*, 2015.

[41] Stanley Osher, Feng Ruan, Jiechao Xiong, Yuan Yao, and Wotao Yin. Sparse recovery via differential inclusions. *Applied and Computational Harmonic Analysis*, 41(2):436–469, 2016.

[42] Hua Ouyang, Niao He, Long Tran, and Alexander Gray. Stochastic alternating direction method of multipliers. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 80–88, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/ouyang13.html.

[43] Yuyuan Ouyang, Yunmei Chen, Guanghui Lan, and Eduardo Pasiliao Jr. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1): 644–681, 2015.

[44] D. H. Peaceman and H. H. Rachford. The numerical solution of parabolic elliptic differential equations. *SIAM Journal on Applied Mathematics*, 3:28–41, 1955.

[45] Clarice Poon and Jingwei Liang. Trajectory of alternating direction method of multipliers and adaptive acceleration. In *Advances in Neural Information Processing Systems 32*, pages 7357–7365. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8955-trajectory-of-alternating-direction-method-of-multipliers-and-adaptive-acceleration.pdf.

[46] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, pages 71–79, 2013.

[47] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

[48] Bin Shi, Simon S Du, Weijie Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, pages 5745–5753, 2019.

[49] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[50] Qingyun Sun and Stephen Boyd. Distributional robust kelly gambling. *arXiv preprint arXiv:1812.10371*, 2018.

[51] Qingyun Sun, Mengyuan Yan, David Donoho, and Stephen Boyd. Convolutional imputation of matrix networks. In *International Conference on Machine Learning*, pages 4818–4827, 2018.

[52] Taiji Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *International Conference on Machine Learning*, pages 392–400, 2013.

[53] Apidopoulos Vassilis, Aujol Jean-François, and Dossal Charles. The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case $b \leq 3$. *SIAM Journal on Optimization*, 28(1):551–574, 2018.

[54] B. C. Vu. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

[55] Huahua Wang and Arindam Banerjee. Online alternating direction method. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, Proceedings of the 29th International Conference on Machine Learning, ICML 2012, page 1699–1706, 10 2012. ISBN 9781450312851.

[56] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3):248–272, 2008.

[57] Roscoe B White. *Asymptotic Analysis of Differential Equations*. Imperial College Press, 2010.

[58] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, page 201614734, 2016.

[59] Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

[60] Y. Xu. Alternating proximal gradient method for sparse nonnegative Tucker decomposition. *Mathematical Programming Computation*, 7(1):39–70, 2015.

[61] J. Yang and X. Yuan. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation*, 82(281):301–329, 2013.

[62] Junfeng Yang and Yin Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM journal on scientific computing*, 33(1):250–278, 2011.

[63] Huizhuo Yuan, Yuren Zhou, Chris Junchi Li, and Qingyun Sun. Differential inclusions for modeling nonsmooth ADMM variants: A continuous limit theory. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7232–7241, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/yuan19c.html.

[64] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14:899–925, 2013.

[65] X. Zhang, M. Burger, X. Bresson, and S. Osher. Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM Journal on Imaging Science*, 3:253–276, 2010.

[66] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1702.05575*, 2017.

[67] Wenliang Zhong and James Kwok. Fast stochastic alternating direction method of multipliers. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 46–54, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/zhong14.html.

[68] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn. Stochastic mirror descent in variationally coherent optimization problems. In *Advances in Neural Information Processing Systems*, pages 7040–7049, 2017.