Proceedings of Machine Learning Research vol 144:1–63, 2021

Accelerated Learning with Robustness to Adversarial Regressors

Joseph E. Gaudio JEGAUDIO@MIT.EDU
Anuradha M. Annaswamy AANNA@MIT.EDU
José M. Moreu JMMOREU@MIT.EDU
Massachusetts Institute of Technology

Michael A. Bolender MICHAEL.BOLENDER@US.AF.MIL
Air Force Research Laboratory

Travis E. Gibson TEGIBSON@BWH.HARVARD.EDU
Brigham and Women's Hospital and Harvard Medical School

Abstract

High order momentum-based parameter update algorithms have seen widespread applications in training machine learning models. Recently, connections with variational approaches have led to the derivation of new learning algorithms with accelerated learning guarantees. Such methods however, have only considered the case of static regressors. There is a significant need for parameter update algorithms which can be proven stable in the presence of adversarial time-varying regressors, as is commonplace in control theory. In this paper, we propose a new discrete time algorithm which 1) provides stability and asymptotic convergence guarantees in the presence of adversarial regressors by leveraging insights from *adaptive control theory* and 2) provides non-asymptotic accelerated learning guarantees leveraging insights from convex optimization. In particular, our algorithm reaches an ϵ sub-optimal point in at most $\tilde{\mathcal{O}}(1/\sqrt{\epsilon})$ iterations when regressors are constant - matching lower bounds due to Nesterov of $\Omega(1/\sqrt{\epsilon})$, up to a $\log(1/\epsilon)$ factor and provides guaranteed bounds for stability when regressors are time-varying. We provide numerical experiments for a variant of Nesterov's provably hard convex optimization problem with time-varying regressors, as well as the problem of recovering an image with a time-varying blur and noise using streaming data.

1. Introduction

Iterative gradient-based optimization methods in machine learning commonly employ a combination of time-scheduled learning rates (Shalev-Shwartz, 2011; Hazan, 2016), adaptive learning rates (Duchi et al., 2011; Kingma and Ba, 2017; Wilson et al., 2017), and/or higher order "momentum" based dynamics (Polyak, 1964; Nesterov, 1983; Wibisono et al., 2016). Variants of the higher order update proposed by Nesterov (Nesterov, 1983) in particular, have received significant attention in the optimization (Nesterov, 2004; Beck and Teboulle, 2009a; Bubeck, 2015; Carmon et al., 2018; Nesterov, 2018) and neural network communities (Krizhevsky et al., 2012; Sutskever et al., 2013) due to their provable guarantees of accelerated learning for classes of convex functions. Empirical investigations for non-convex neural network training are also a topic of significant interest.

To gain insight into Nesterov's discrete time method (Nesterov, 1983), the authors in (Su et al., 2016) identified the second order ordinary differential equation (ODE) at the limit of zero step size. Still pushing further in the continuous time analysis of these higher order methods, several recent results have leveraged a variational approach showing that a larger class of higher order methods exist where one can obtain an arbitrarily fast convergence rate (Wibisono et al., 2016; Wilson et al., 2016).

Table 1: Comparison of gradient-based methods for a class of time-varying convex functions.

Algorithm	Equation	Constant Regressor # Iterations	Time-Varying Regressor
Gradient Descent Normalized	(3)	$\mathcal{O}(1/\epsilon)$	Stable
Gradient Descent Fixed	(20)	$\mathcal{O}(1/\epsilon)$	Unstable
Nesterov Acceleration Varying	(21)	$\mathcal{O}(1/\sqrt{\epsilon})$	Unstable
Nesterov Acceleration Fixed	(9)	$\mathcal{O}(1/\sqrt{\epsilon} \cdot \log(1/\epsilon))$	Unstable
This Paper	Alg 1	$\mathcal{O}(1/\sqrt{\epsilon} \cdot \log(1/\epsilon))$	Stable

An equivalent algorithm in discrete-time to these continuous-time results, which is implementable and has comparable convergence rates, is an active area of research (Betancourt et al., 2018; Wilson, 2018; Shi et al., 2019). It should be noted that, quite often, in many of these papers (Su et al., 2016; Wibisono et al., 2016; Wilson et al., 2016; Betancourt et al., 2018; Wilson, 2018; Shi et al., 2019), the analysis is performed for static features/regressors, with any dynamic components arising only due to a recursive update of the parameters.

There are many machine learning applications and paradigms where the features or inputs are time-varying. Examples include multi-armed bandits (Auer et al., 1995, 2002; Bubeck and Cesa-Bianchi, 2012), adaptive-filtering (Goodwin and Sin, 1984; Widrow and Stearns, 1985; Haykin, 2014), and temporal-prediction tasks (Dietterich, 2002; Kuznetsov and Mohri, 2015; Hall and Willett, 2015), to name a few. In addition, many models can be trained via adversarial learning (Shalev-Shwartz, 2011; Ben-David et al., 2009) which results in time-varying inputs during training (Auer et al., 1995; Cesa-Bianchi and Lugosi, 2006). Even if the application does not require time-varying inputs, the presence of large training data has necessitated online or stochastic training methods in several applications, bringing a dynamic component into the problem statement (Goodfellow et al., 2016; Shalev-Shwartz, 2011; Cesa-Bianchi et al., 2004; Bengio, 2012; Jain et al., 2018; Gitman et al., 2019). Online learning is another class of problems that requires an investigation of optimization (Zinkevich, 2003; Hazan et al., 2007, 2008; Hazan, 2016; Shalev-Shwartz, 2011; Raginsky et al., 2010). Online learning has had particular success in the development of state of the art gradient methods for training large neural networks (Duchi et al., 2011; Kingma and Ba, 2017).

Time variations in inputs become even more important in real-time applications, with potentially limited compute (Jordan and Mitchell, 2015), and in an area of machine learning which has now come to be referred to as continual/lifelong learning (Ben-David et al., 2009; Chen and Liu, 2018; Thrun and Mitchell, 1995; Thrun, 1998; Parisi et al., 2019). Continual/lifelong learning algorithms must be robust to adversarial features/inputs in addition to shifts in the distribution of the incoming data (Ben-David et al., 2009; Lopez-Paz and Ranzato, 2017), i.e. data is not necessarily independent and identically distributed from a fixed probability distribution (Pentina and Lampert, 2015). Such algorithms must also be able to incrementally learn for an indefinite amount of time without human intervention (Silver et al., 2013; Fei et al., 2016; Chen and Liu, 2018).[1] These notions are further important in robotics (Thrun and Mitchell, 1995; Thrun, 1998) and learning-based control theory (Sastry and Bodson, 1989; Narendra and Annaswamy, 2005; Goodwin and Sin, 1984; Ioannou and Sun, 1996) due to the requirement of continuously running in such applications.

1. In the online learning setting, when minimizing regret (Shalev-Shwartz, 2011), the learning rates decay over time.

This paper proposes a new discrete time algorithm for parameter updates that accommodates time-varying regressors and is of high order. This algorithm will be shown to achieve two objectives. The first objective is to demonstrate stability of this high-order parameter tuner in the presence of time-varying adversarial regressors. This is in contrast to many other iterative methods that cannot be proved to be stable in this setting. The second objective is to show an accelerated convergence rate when the regressors are constant. Our higher order tuner is based on a novel discretization of a continuous time higher order learning parameter update, and differs from the variational perspective-based high order tuner proposed in (Wibisono et al., 2016) which leverages time-scheduled hyperparameters. Unlike many of the papers listed above, we do not assume the requirement of an *a priori* bound in the time-varying regressor for stability of our algorithm, and directly deploy the gradient which utilizes the regressors. The non-asymptotic convergence rate, which corresponds to the second objective, will be shown to be a logarithm factor away from provable lower bounds due to Nesterov (Nesterov, 2018), with comparable constant factors.

Table 1 provides an overview of how these two objectives are realized with this algorithm in comparison to other iterative methods. The first objective ensures that our iterative algorithm remains stable and learns indefinitely as streaming data changes, even in an adversarial manner - crucial in learning for dynamical systems applications. The second objective demonstrates that the proposed learning algorithm retains fast convergence in the standard setting of constant regressors. That is, the significant benefit of provable stability of our algorithm in the presence of time-varying regressors does not come at the expense of large degradation in the rate of convergence - our proposed algorithm has a near-optimal convergence rate in the standard static regressor analysis setting as well.

The main contributions of this work are summarized as: (i) A new class of momentum/Nesterov-type iterative optimization algorithms, (ii) Accelerated learning guarantee a logarithm factor away from Nesterov (while remaining stable), (iii) Explicit stability conditions for the new algorithms with adversarial regressors, (iv) Connections to a Lagrangian variational perspective alongside an introduction of an adaptive systems-based normalization in machine learning, and (v) Numerical simulations demonstrating the efficacy of the proposed methods.

2. Problem setting

In this paper, we present the continuous time perspective with time t while discrete time steps are indexed by k. When in continuous time, the time dependence of variables may be omitted when it is clear from the context. The classes $\mathcal{L}_p$ and ℓ_p for $p \in [1, \infty]$ are described in (Gaudio et al., 2020, Appendix A), alongside definitions of (strong) convexity, smoothness, and Euler discretization techniques. Unless otherwise specified, $\|\cdot\|$ represents the 2-norm. We denote the discrete time difference of a function V as $\Delta V_k := V_{k+1} - V_k$. For notational clarity and to focus on multidimensional parameters/regressors we present the single output setting. The results of this paper trivially extend to multiple outputs.

We consider the setting of linear regression with time-varying regressors $\phi \in \mathbb{R}^N$ which are related in a linear combination with an unknown parameter $\theta^* \in \mathbb{R}^N$ to the output $y \in \mathbb{R}$ as $y_k = \theta^{*T}\phi_k$. Given that the parameter θ^* is unknown, an estimator $\hat{y}_k = \theta_k^T \phi_k$ is formulated, where $\hat{y} \in \mathbb{R}$ is the output estimate and $\theta \in \mathbb{R}^N$ is the parameter estimate. In this setting, the output error is defined as

$$e_{y,k} = \hat{y}_k - y_k = \tilde{\theta}_k^T \phi_k, \tag{1}$$

where $\tilde{\theta}_k = \theta_k - \theta^*$ is the parameter estimation error. The goal is to design an iterative algorithm to adjust the parameter estimate θ using streaming regressor-output data pairs $\mathcal{D}ata_k = (\phi_k, y_k)$ such that the prediction error e_y converges to zero with a provably fast non-asymptotic convergence rate when regressors ϕ_k are constant, and that stability and asymptotic convergence properties remain in the presence of time-varying regressors. An iterative gradient-based method is proposed to enable computational simplicity and accommodate data in real-time. To formulate the gradient-based methods of this paper, we consider the squared loss function using (1) of the form

$$L_k(\theta_k) = \frac{1}{2}e_{y,k}^2 = \frac{1}{2}\tilde{\theta}_k^T \phi_k \phi_k^T \tilde{\theta}_k, \tag{2}$$

where the subscript k in L_k denotes the regressor iteration number. At each iteration k, the gradient of the loss function is implementable as $\nabla L_k(\theta_k) = \phi_k e_{y,k}$. The Hessian of (2) can be expressed as $\nabla^2 L_k(\theta_k) = \phi_k \phi_k^T$, and thus $0 \leq \nabla^2 L_k(\theta_k) \leq \|\phi_k\|^2 I$. Therefore, the loss function can be seen to be (non-strongly) convex with a time-varying regressor-dependent smoothness parameter.

Remark 1 *The stability results of this paper will be shown to hold even for **adversarial** time-varying regressors ϕ_k. No bound on ϕ_k is required to be known and the prediction error $e_{y,k}$ is not assumed to be bounded a priori. This is in comparison to standard methods in online learning which assume knowledge of a bound on gradients and regressors for proving stability (Shalev-Shwartz, 2011; Hazan, 2016). Thus the algorithm proposed in this paper can be employed in the continual learning (Ben-David et al., 2009; Thrun and Mitchell, 1995; Thrun, 1998) and learning-based control theory (Sastry and Bodson, 1989; Narendra and Annaswamy, 2005; Goodwin and Sin, 1984; Ioannou and Sun, 1996) settings where such assumptions of a priori boundedness cannot be made.*

The starting point for our proposed algorithm comes from adaptive methods (see for example, (Goodwin and Sin, 1984, Ch. 3)) which leads to an iterative normalized gradient descent method

$$\theta_{k+1} = \theta_k - \gamma \nabla \bar{f}_k(\theta_k), \quad 0 < \gamma < 2, \tag{3}$$

where $\bar{f}_k(\cdot)$ corresponds to a normalized loss function defined as

$$\bar{f}_k(\theta_k) = \frac{L_k(\theta_k)}{\mathcal{N}_k}, \tag{4}$$

and $\mathcal{N}_k = 1 + \|\phi_k\|^2$ is a normalization signal employed to ensure boundedness of signals for any arbitrary regressor ϕ_k. Motivated by the normalized gradient method in (3), our goal is to derive a Nesterov-type higher order gradient method to ensure a provably faster convergence rate when regressors are constant while preserving stability of the estimation algorithm in the presence of adversarial time-varying ϕ_k.

3. Algorithms based on a higher order tuner

We begin the derivation of our discrete time Nesterov-type higher order tuner algorithm from the continuous time perspective, which provides insights into the underlying stability structure. This perspective further results in a representation of a second order differential equation as two first order differential equations which are used to certify stability in the presence of time-varying regressors by employing Lyapunov function techniques. Motivated by this continuous time representation, we provide a novel discretization to result in a discrete time higher order tuner which can be shown to be stable using the same Lyapunov function. The discrete time algorithm is then shown to be equivalent to the common Nesterov iterative method form when regressors are constant.

3.1. Continuous time higher order tuner

We begin the derivation of our algorithm using the variational perspective of Wibisono, Wilson, and Jordan (Wibisono et al., 2016). In particular, the Bregman Lagrangian in (Wibisono et al., 2016, Eq. 1) is re-stated with the Euclidean norm employed in the Bregman divergence as $\mathcal{L}(\theta(t), \dot{\theta}(t), t) = \mathrm{e}^{\bar{\alpha}_t + \bar{\gamma}_t} \left(\mathrm{e}^{-2\bar{\alpha}_t} \frac{1}{2} \|\dot{\theta}(t)\|^2 - \mathrm{e}^{\bar{\beta}_t} L_t(\theta(t)) \right)$. This Lagrangian weights potential energy (loss) $L_t(\theta(t))$, and kinetic energy $(1/2)\|\dot{\theta}(t)\|^2$, with an exponential term $\exp(\bar{\alpha}_t + \bar{\gamma}_t)$, which adjusts the damping. The hyperparameters $(\bar{\alpha}_t, \bar{\beta}_t, \bar{\gamma}_t)$ are commonly time-scheduled and result in different algorithms by appropriately weighting each component in the Lagrangian (see (Wibisono et al., 2016) for choices common in optimization for machine learning). It can be easily shown however, that time scheduling the hyperparameters can result in instability when regressors are time-varying. We thus propose the use of a regressor-based normalization $\mathcal{N}_t = 1 + \|\phi(t)\|^2$ with constant gains $\gamma, \beta > 0$ to parameterize the Lagrangian as

$$\mathcal{L}(\theta(t), \dot{\theta}(t), t) = \mathrm{e}^{\beta(t - t_0)} \left(\frac{1}{2} \|\dot{\theta}(t)\|^2 - \frac{\gamma\beta}{\mathcal{N}_t} L_t(\theta(t)) \right). \tag{5}$$

Using a Lagrangian, a functional may be defined as: $J(\theta) = \int_{\mathbb{T}} \mathcal{L}(\theta, \dot{\theta}, t) dt$, where $\mathbb{T}$ is an interval of time. To minimize this functional, a necessary condition from the calculus of variations (Goldstein et al., 2002) is that the Lagrangian solves the Euler-Lagrange equation: $\frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \dot{\theta}}(\theta, \dot{\theta}, t) \right) = \frac{\partial \mathcal{L}}{\partial \theta}(\theta, \dot{\theta}, t)$. Using (5), the second order differential equation resulting from the application of the Euler-Lagrange equation is: $\ddot{\theta}(t) + \beta \dot{\theta}(t) = -\frac{\gamma\beta}{\mathcal{N}_t} \nabla L_t(\theta(t))$. This differential equation can be seen to have the normalized gradient of the loss function as the forcing term parameterized with $\gamma\beta$, and constant damping parameterized with β. Crucial to the development of the results of this paper, this second order differential equation may be written as a *higher order tuner* given by

$$\begin{aligned} \dot{\vartheta}(t) &= -\frac{\gamma}{\mathcal{N}_t} \nabla L_t(\theta(t)), \\ \dot{\theta}(t) &= -\beta(\theta(t) - \vartheta(t)), \end{aligned} \tag{6}$$

which can be seen to take the form of a normalized gradient flow update followed by a linear time invariant (LTI) filter. This representation of a higher order tuner will be fundamental to prove stability with time-varying regressors using Lyapunov function techniques in Section 4.

3.2. Discretization of continuous time higher order tuner

We propose in this paper a specific discretization of the high-order tuner in (6), of the form

$$\begin{aligned} \text{Implicit Euler}: \vartheta_{k+1} &= \vartheta_k - \gamma \nabla \bar{f}_k(\theta_{k+1}), \\ \text{Explicit Euler}: \theta_{k+1} &= \bar{\theta}_k - \beta(\bar{\theta}_k - \vartheta_k), \\ \text{Extra Gradient}: \quad \bar{\theta}_k &= \theta_k - \gamma\beta \nabla \bar{f}_k(\theta_k), \end{aligned} \tag{7}$$

where $\bar{f}_k(\cdot)$ is given by (4), and the hyperparameters are γ and β. One can employ any number of techniques for the discretization of an ordinary differential equation, including Runge–Kutta, symplectic, and Euler methods (see for example methods in (Hairer et al., 2006; Betancourt et al., 2018)). The one employed in (7) can be viewed as a combination of implicit-Euler (for the variable

ϑ) and explicit-Euler method (for the variable θ). An important correction is introduced in the explicit-Euler component, which corresponds to the use of an extra gradient. It should also be noted that the extra gradient step only serves to adjust the direction of the update, but does not increase the order of the tuner beyond two.

3.3. Augmented objective function

As discussed in Section 2, the squared error loss function in (2) as well as its normalized version in (4) are non-strongly convex. In order to obtain accelerated learning properties similar to that of Nesterov's in (Nesterov, 2018), we now propose a new algorithm that builds on that in (7). For this purpose, we modify the normalized cost function in (4) to include L2 regularization as

$$f_k(\theta_k) = \bar{f}_k(\theta_k) + \frac{\mu}{2}\|\theta_k - \theta_0\|^2, \qquad (8)$$

where $\mu > 0$ is the regularization constant, θ_0 is the initial condition of the parameter estimate, and the subscript k in f_k denotes the regressor iteration number. Using (2), the Hessian of (8) can be expressed as $\nabla^2 f_k(\theta_k) = (\phi_k \phi_k^T)/\mathcal{N}_k + \mu I$, and thus it can be seen that $\mu I \leq \nabla^2 f_k(\theta_k) \leq (1+\mu) I$. Therefore, the objective function in (8) can be seen to be μ-strongly convex and $(1+\mu)$-smooth and has desirable properties of constant smoothness and strong convexity.

Algorithm 1 Higher Order Tuner Optimizer

1: **Input:** initial conditions θ_0, ϑ_0, gains γ, β, μ
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: **Receive** regressor ϕ_k, output y_k
4: Let $\mathcal{N}_k = 1 + \|\phi_k\|^2$
$\nabla f_k(\theta_k) = \frac{\nabla L_k(\theta_k)}{\mathcal{N}_k} + \mu(\theta_k - \theta_0)$,
$\bar{\theta}_k = \theta_k - \gamma\beta \nabla f_k(\theta_k)$
5: $\theta_{k+1} \leftarrow \bar{\theta}_k - \beta(\bar{\theta}_k - \vartheta_k)$
6: Let
$\nabla f_k(\theta_{k+1}) = \frac{\nabla L_k(\theta_{k+1})}{\mathcal{N}_k} + \mu(\theta_{k+1} - \theta_0)$
7: $\vartheta_{k+1} \leftarrow \vartheta_k - \gamma \nabla f_k(\theta_{k+1})$
8: **end for**

By replacing the objective function $\bar{f}_k(\theta_k)$ in (7) with $f_k(\theta_k)$ in (8), we now obtain Algorithm 1, the main higher order tuner optimizer introduced in this paper with hyperparameters γ, β, and μ. It should be noted that the gradient expressions for L_k in lines 4 and 6 of Algorithm 1 are given by $\nabla L_k(\theta_k) = \phi_k(\theta_k^T \phi_k - y_k)$ and $\nabla L_k(\theta_{k+1}) = \phi_k(\theta_{k+1}^T \phi_k - y_k)$ respectively. The stability properties of Algorithm 1, in the presence of a time-varying regressor ϕ_k, will be demonstrated in Section 4. In addition to the stability properties in the presence of ϕ_k, the additional advantage of Algorithm 1 is a fast minimization of (2) for constant regressors. This accelerated convergence property will be established in Section 5 by minimizing the augmented objective in (8). The relation between Algorithm 1 and Nesterov's method is summarized in the following proposition.

Proposition 1 *Algorithm 1 with a constant regressor $\phi_k \equiv \phi$ (and thus $f_k(\cdot) \equiv f(\cdot)$) may be reduced to the common form of Nesterov's equations (Nesterov, 2018, Eq. 2.2.22) with $\bar{\beta} = 1 - \beta$ and $\bar{\alpha} = \gamma\beta$ as*

$$\begin{aligned} \theta_{k+1} &= \nu_k - \bar{\alpha} \nabla f(\nu_k), \\ \nu_{k+1} &= \left(1 + \bar{\beta}\right) \theta_{k+1} - \bar{\beta} \theta_k. \end{aligned} \qquad (9)$$

Remark 2 *It is apparent from Proposition 1 that Algorithm 1 is Nesterov-type; it includes both averaging outside of the gradient evaluation, and an "extra-gradient" step to enable adjustments inside the gradient evaluation. The presence of both of these ingredients makes our Algorithm 1 similar to Nesterov-type rather than Heavy-ball type (Polyak, 1964) which only includes the first.*

Remark 3 *Similar to the derivation of Algorithm 1, which is Nesterov-type, we can derive another algorithm that is Heavy-Ball type (Polyak, 1964). We accomplish this in (Gaudio et al., 2020, Appendix A.3) and denote it as Algorithm 2. Algorithm 2 only includes the implicit-explicit mix and not the extra gradient step. As is shown in (Gaudio et al., 2020, Appendix B), Algorithm 2 is also stable in the presence of time-varying regressors. The disadvantage of Algorithm 2 over Algorithm 1 is simply due to the well known point of Heavy-ball type methods in comparison to Nesterov-type methods (c.f. (Lessard et al., 2016) for a clear example). This is the reason for our preference of Algorithm 1 over Algorithm 2.*

4. Stability and asymptotic convergence

In this section, we state the main results of stability and asymptotic convergence in the presence of time-varying regressors for the continuous time higher order tuner (6), discretized equations (7), and the main stability result of Algorithm 1. Proofs of all theorems and corollaries in this section are provided in (Gaudio et al., 2020, Appendix B) alongside more in-depth auxiliary results of stability. For completeness, complementary stability proofs of the normalized gradient method (3) in both continuous and discrete time are additionally provided in (Gaudio et al., 2020, Appendix B). We begin with the discussion of stability of the discretized equations in (7) (Algorithm 1 with $\mu = 0$), in the following theorem.

Theorem 4 *For the linear regression error model in (1) with loss in (2), with Algorithm 1 and its hyperparameters chosen as $\mu = 0$, $0 < \beta < 1$, $0 < \gamma \leq \frac{\beta(2-\beta)}{16+\beta^2}$, the following*

$$V_k = \frac{1}{\gamma}\|\vartheta_k - \theta^*\|^2 + \frac{1}{\gamma}\|\theta_k - \vartheta_k\|^2, \tag{10}$$

is a Lyapunov function with increment $\Delta V_k \leq -\frac{L_k(\theta_{k+1})}{\mathcal{N}_k} \leq 0$. It can also be shown that $V \in \ell_\infty$, and $\sqrt{\frac{L_k(\theta_{k+1})}{\mathcal{N}_k}} \in \ell_2 \cap \ell_\infty$. If in addition it is assumed that $\phi \in \ell_\infty$ then $\lim_{k\to\infty} L_k(\theta_{k+1}) = 0$.

We now proceed to the main stability theorem of Algorithm 1 with $\mu \neq 0$.

Theorem 5 *For the linear regression error model in (1) with loss in (2), with Algorithm 1 and its hyperparameters chosen as $0 < \mu < 1$, $0 < \beta < 1$, $0 < \gamma \leq \frac{\beta(2-\beta)}{16+\beta^2+\mu\left(\frac{57\beta+1}{16\beta}\right)}$, the function V in (10) can be shown to have increment $\Delta V_k \leq -\frac{L_k(\theta_{k+1})}{\mathcal{N}_k} - \mu c_1 V_k + \mu c_2$, for constants $0 < c_1 < 1$, $c_2 > 0$ (given in (Gaudio et al., 2020, Appendix B)). It can also be shown that $\Delta V_k < 0$ outside of the compact set $D = \left\{V \middle| V \leq \frac{c_2}{c_1}\right\}$. Furthermore, $V \in \ell_\infty$ and $V_k \leq \exp(-\mu c_1 k)\left(V_0 - \frac{c_2}{c_1}\right) + \frac{c_2}{c_1}$.*

The Lyapunov function in (10) which is employed in Theorems 4 and 5 was originally motivated by the continuous time higher order tuners in (Morse, 1992; Evesque et al., 2003). The continuous time equivalent of (10) is used in the following theorem to prove stability and asymptotic convergence properties for the continuous time higher order tuner in (6).

Theorem 6 *For continuous time equivalents to the linear regression model in (1) with loss in (2) (concretely, (17) and (18) in (Gaudio et al., 2020, Appendix A.2)), for the higher order tuner update in (6) with $\beta > 0$, $0 < \gamma \leq \beta/2$, the following*

$$V(t) = \frac{1}{\gamma}\|\vartheta(t) - \theta^*\|^2 + \frac{1}{\gamma}\|\theta(t) - \vartheta(t)\|^2, \tag{11}$$

is a Lyapunov function with time derivative $\dot{V}(t) \leq -\frac{L_t(\theta(t))}{\mathcal{N}_t} \leq 0$. *It can be shown that* $V \in \mathcal{L}_\infty$ *and* $\sqrt{\frac{L_t(\theta(t))}{\mathcal{N}_t}} \in \mathcal{L}_2 \cap \mathcal{L}_\infty$. *If in addition it assumed that* $\phi, \dot{\phi} \in \mathcal{L}_\infty$ *then* $\lim_{t\to\infty} L_t(\theta(t)) = 0$.

Remark 7 *The same function* V *is employed throughout, as motivated by the continuous time higher order tuner in Theorem 6. Note that the proofs of stability in the presence of adversarial time-varying regressors are enabled as the Lyapunov functions in (10) and (11) do not contain the regressor. In both the continuous and discrete time analyses, stability is proven by showing that the provided function* V *does not increase globally (Theorems 4 and 6) or at least does not increase outside a compact set containing the origin (Theorem 5).*

5. Non-asymptotic accelerated convergence rates with constant regressors

In this section, we state the main accelerated non-asymptotic convergence rate result for Algorithm 1 for the case of constant regressors, $\phi_k \equiv \phi$. All proofs in this section are provided in (Gaudio et al., 2020, Appendix C) alongside convergence rate proofs for first order gradient descent methods, and Nesterov's method with time-varying gains, as given in overview form in Table 1.

Given the constant smoothness and strong-convexity parameters of the augmented objective function in (8), the representation of Algorithm 1 as (9) may be used to provide a non-asymptotic rate for (8) in the following theorem due to Nesterov.

Theorem 8 (Modified from (Bubeck, 2015; Nesterov, 2018)) *For a* $\bar{L}$*-smooth and* μ*-strongly convex function* f*, the iterates* $\{\theta_k\}_{k=0}^{\infty}$ *generated by (9) with* $\theta_0 = \nu_0$, $\bar{\alpha} = 1/\bar{L}$, $\kappa = \bar{L}/\mu$, *and* $\bar{\beta} = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$ *satisfy* $f(\theta_k) - f(\theta^*) \leq \frac{\bar{L}+\mu}{2}\|\theta_0 - \theta^*\|^2 \exp\left(-\frac{k}{\sqrt{\kappa}}\right)$.

Leveraging the accelerated convergence rate for the augmented function f in (9) as provided by Theorem 8, we provide the following new lemmas to give accelerated non-asymptotic convergence rates for the normalized and unnormalized versions of the loss function in (2), as desired.

Lemma 9 *The iterates* $\{\theta_k\}_{k=0}^{\infty}$ *generated by (9) for the function in (8) with* $\theta_0 = \nu_0$, $\Psi \geq \max\{1, \|\theta_0 - \theta^*\|^2\}$, $\mu = \epsilon/\Psi$, $\bar{L} = 1 + \mu$, $\bar{\alpha} = 1/\bar{L}$, $\kappa = \bar{L}/\mu$, $\bar{\beta} = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$, *if*

$$k \geq \left\lceil \sqrt{1 + \frac{\Psi}{\epsilon}} \log\left(2 + \frac{\Psi}{\epsilon}\right) \right\rceil, \text{ then } \frac{L(\theta_k) - L(\theta^*)}{\mathcal{N}} \leq \epsilon. \tag{12}$$

Lemma 10 *The iterates* $\{\theta_k\}_{k=0}^{\infty}$ *generated by (9) for the function in (8) with* $\theta_0 = \nu_0$, $\Psi \geq \max\{1, \mathcal{N}\|\theta_0 - \theta^*\|^2\}$, $\mu = \epsilon/\Psi$, $\bar{L} = 1 + \mu$, $\bar{\alpha} = 1/\bar{L}$, $\kappa = \bar{L}/\mu$, $\bar{\beta} = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$, *if*

$$k \geq \left\lceil \sqrt{1 + \frac{\Psi}{\epsilon}} \log\left(2 + \frac{\Psi}{\epsilon}\right) \right\rceil, \text{ then } L(\theta_k) - L(\theta^*) \leq \epsilon. \tag{13}$$

Remark 11 *Lemmas 9 and 10 provide the provable number of iterations required to obtain an* ϵ *sub-optimal point of for the normalized and original loss function in (2) of* $\mathcal{O}(1/\sqrt{\epsilon} \cdot \log(1/\epsilon))$, *with all constants included. It can be noted that the constants are comparable to the constants for the gradient and Nesterov iterative methods shown in Table 2 and Figure 3 in (Gaudio et al., 2020, Appendix C).*

Remark 12 *It can be noted from both Lemmas 9 and 10 that the L2 regularization parameter μ in (8) is smaller than the ϵ sub-optimality gap. The L2 regularization parameter is present to ensure strong convexity of the augmented objective function in (8), such that Algorithm 1 can be reduced to Nesterov's iterative method with constant gains in (9), which results in the convergence rate for the augmented function in Theorem 8, which in turn lends to Lemmas 9 and 10.*

6. Numerical experiments

In this section, we analyze and compare the performance of our proposed algorithm in two different numerical experiment settings: a variant of Nesterov's provably hard smooth convex optimization problem (Nesterov, 2018, p. 69) and a variant of the image deblurring problem considered by Beck and Teboulle (Beck and Teboulle, 2009a). In each setting, we compare hyperparameters chosen in accordance with Theorem 5 and hyperparameters chosen optimally as per the standard Nesterov iterative method. All simulations are implemented in Python code available at link1 and link2. Videos demonstrating the real-time image deblurring results are furthermore available.

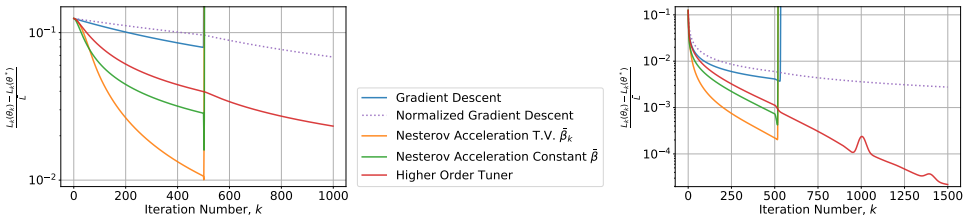
6.1. Nesterov's smooth convex function

In this section, we consider a modified version of Nesterov's provably hard smooth convex problem (Nesterov, 2018, p. 69) of the form $L_k(\theta) = \|\phi_k^T\theta\|^2 + B^T\theta$. In the experiments, the regressor ϕ_k changes, resulting in a change in the smoothness parameter $\bar{L}$. This problem was selected to demonstrate a lower bound of $\mathcal{O}(1/\sqrt{\epsilon})$ for iterative methods with gradient information (Nesterov, 2018, p. 69). (Gaudio et al., 2020, Appendix D.1) provides a detailed description.

In Figure 1, we compare two different experiments with hyperparameters selected in two ways, where the regressor ϕ_k changes at iteration $k = 500$. For the first experiment in Figure 1a, the hyperparameters for the high order tuner algorithm are chosen according to Theorem 5, and the hyperparameters of the other methods are chosen with the same step size and the momentum parameter as in Proposition 1. In Figure 1b, the hyperparameters are chosen optimally in accordance with the Nesterov iterative method with $\beta = 1 - \bar{\beta}$ and $\gamma = \bar{\alpha}/\beta$, as per Proposition 1. For $k < 500$, the convergence rate of the Higher Order Tuner algorithm can be seen to be comparable to the optimal Nesterov algorithm. After ϕ_k changes at iteration $k = 500$, all unnormalized algorithms become unstable. Even in the presence of a change in the regressor, the Higher Order Tuner algorithm can be seen to have a fast rate of convergence as compared to the normalized gradient descent method.

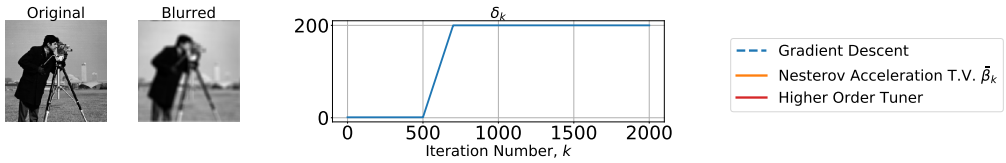
6.2. Image deblurring problem

In this section, we consider a variant of the image deblurring problem of Beck and Teboulle (Beck and Teboulle, 2009a), with a time-varying blur. All processing is done in the *frequency domain* in which the unknown true image is denoted as θ^*, the measured blurry version is represented as $y_k = \phi_k^T\theta^*$, where the blur is represented as the regressor ϕ_k. We consider the fully adversarial setting where ϕ_k may be affected due to issues in the communication system, changes in lighting, or any other adversarial effects. For the purposes of the presented results, we consider a scalar multiplicative perturbation δ_k, which results in a frequency domain blur representation as $\phi_k = \delta_k \odot \text{blur\_operator}(P_k)$, where P_k is a known point spread function. We employ the same squared loss function as in (2). A complete description of the problem formulation is provided in (Gaudio et al., 2020, Appendix D.2), alongside additional experiments with noisy measurements.

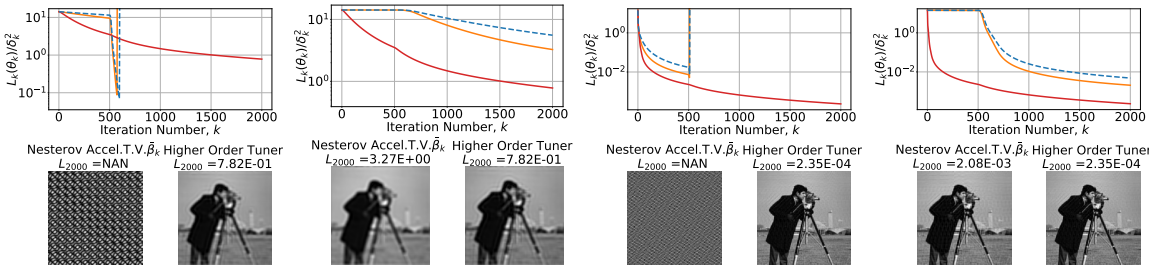


(a) (b)

Figure 1: A variant of Nesterov's smooth convex function (Nesterov, 2018, p. 69). (a) $\mu = 10^{-5}$, $\beta = 0.1$ and γ as in Theorem 5, $\bar{\beta} = 1 - \beta$ and $\bar{\alpha} = \gamma\beta$. At iteration $k = 500$, step change in $\bar{L}$ from 2 to 8000. (b) Hyperparameters chosen satisfying Lemma 10 at iteration $k = 0$ with $\epsilon = 0.001$, $\beta = 1 - \bar{\beta}$, and $\gamma = \bar{\alpha}/\beta$. At iteration $k = 500$, step change in $\bar{L}$, from 2 to 8.



(a)



(b) (c) (d) (e)

Figure 2: (a) Original and blurred images; Ramp increase of δ_k from 1 to 200 in 200 iterations, starting at $k = 500$. (b) and (c) Hyperparameters as in Theorem 5. (b) Loss values and reconstructed images when only ϕ_0 is known *a priori*. (c) Loss values and reconstructed images when all ϕ_k are known *a priori*. (d) and (e) Hyperparameters chosen optimally as per the Nesterov iterative method. (d) Loss values and reconstructed images when only ϕ_0 is known *a priori*. (e) Loss values and reconstructed images when all ϕ_k are known *a priori*. Please, see Figure 8 for the noisy case.

Figure 2 shows the loss values and reconstructed images at iteration 2000, for the ramp change in the regressor/blur in Figure 2a. We present numerical results for hyperparameters chosen in four different ways. In Figure 2b and Figure 2c, the higher order tuner hyperparameters are chosen according to Theorem 5, and $\bar{\alpha}$ is chosen as $\bar{\alpha} = \gamma\beta/\mathcal{N}_0$ in Figure 2b and $\bar{\alpha} = \gamma\beta/\max\mathcal{N}_k$ in Figure 2c. In Figure 2d and Figure 2e, the step size is chosen as $\bar{\alpha} = 1/\|\phi_0\|_2^2$ and $\bar{\alpha} = 1/\max\|\phi_k\|_2^2$ respectively, which results in hyperparameter choices for the higher order tuner as $\mu = 10^{-20}$, $\beta = 0.1$, and $\gamma = 1/\beta$, as per Proposition 1. The higher order tuner remains stable as opposed to the other methods which are unstable if the parameters are not chosen appropriately for all ϕ_k.

Acknowledgments

This work was supported by the Air Force Research Laboratory, Collaborative Research and Development for Innovative Aerospace Leadership (CRDInAL), Thrust 3 - Control Automation and Mechanization grant FA 8650-16-C-2642 and the Boeing Strategic University Initiative; cleared for release, case number 88ABW-2020-1889.

References

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*. IEEE Comput. Soc. Press, 1995. doi: 10.1109/SFCS.1995.492488.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2/3):235–256, 2002. doi: 10.1023/A:1013689704352.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, jan 2009a. doi: 10.1137/080716542.

Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, nov 2009b. doi: 10.1109/TIP.2009.2028250.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, oct 2009. doi: 10.1007/s10994-009-5152-4.

Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Lecture Notes in Computer Science*, pages 437–478. Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-35289-8_26.

Michael Betancourt, Michael I. Jordan, and Ashia C. Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. doi: 10.1561/2200000050.

Sébastien Bubeck and Nicolò Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*, volume 5. Now Publishers, 2012. doi: 10.1561/2200000024.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for NonConvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, jan 2018. doi: 10.1137/17M1114296.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, sep 2004. doi: 10.1109/TIT.2004.833339.

Zhiyuan Chen and Bing Liu. *Lifelong Machine Learning*. Morgan & Claypool Publishers, 2018.

Thomas G. Dietterich. Machine learning for sequential data: A review. In *Lecture Notes in Computer Science*, pages 15–30. Springer Berlin Heidelberg, 2002. doi: 10.1007/3-540-70659-3_2.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, July 2011.

S. Evesque, A. M. Annaswamy, S. Niculescu, and A. P. Dowling. Adaptive control of a class of time-delay systems. *Journal of Dynamic Systems, Measurement, and Control*, 125(2):186, 2003. doi: 10.1115/1.1567755.

Geli Fei, Shuai Wang, and Bing Liu. Learning cumulatively to become more knowledgeable. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. ACM Press, 2016. doi: 10.1145/2939672.2939835.

Joseph E. Gaudio, Anuradha M. Annaswamy, José M. Moreu, Michael A. Bolender, and Travis E. Gibson. Accelerated learning with robustness to adversarial regressors. *arXiv preprint arXiv:2005.01529*, 2020.

Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9633–9643. Curran Associates, Inc., 2019.

Herbert Goldstein, Charles Poole, and John Safko. *Classical Mechanics*. Addison Wesley, 2002.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Graham C Goodwin and Kwai Sang Sin. *Adaptive Filtering Prediction and Control*. Prentice Hall, 1984.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, 2006.

Eric C. Hall and Rebecca M. Willett. Online convex optimization in dynamic environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):647–662, jun 2015. doi: 10.1109/JSTSP.2015.2404790.

Per Christian Hansen, James G. Nagy, and Dianne P. O'leary. *Deblurring Images: Matrices, Spectra, and Filtering*. SIAM, 2006.

Simon Haykin. *Adaptive Filter Theory*. Pearson, 2014.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. doi: 10.1561/2400000013.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, aug 2007. doi: 10.1007/s10994-007-5016-8.

Elad Hazan, Alexander Rakhlin, and Peter L. Bartlett. Adaptive online gradient descent. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 65–72. Curran Associates, Inc., 2008.

Petros A. Ioannou and Jing Sun. *Robust Adaptive Control*. Prentice-Hall, 1996.

Prateek Jain, Praneeth Netrapalli, Sham M. Kakade, Rahul Kidambi, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 545–604. PMLR, 06–09 Jul 2018.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, jul 2015. ISSN 0036-8075. doi: 10.1126/science.aaa8415.

Diederik P. Kingma and Jimmy L. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 541–549. Curran Associates, Inc., 2015.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, jan 2016. doi: 10.1137/15m1009597.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6467–6476. Curran Associates, Inc., 2017.

David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, 1969.

A. S. Morse. High-order parameter tuners for the adaptive control of linear and nonlinear systems. In *Systems, Models and Feedback: Theory and Applications*, pages 339–364. Birkhäuser Boston, 1992. doi: 10.1007/978-1-4757-2204-8_23.

Kumpati S. Narendra and Anuradha M. Annaswamy. *Stable Adaptive Systems*. Dover, 2005.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27:372–376, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization*. Springer, 2004. doi: 10.1007/978-1-4419-8853-9.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2018. doi: 10.1007/978-3-319-91578-4.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, may 2019. doi: 10.1016/j.neunet.2019.01.012.

Anastasia Pentina and Christoph H. Lampert. Lifelong learning with non-i.i.d. tasks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1540–1548. Curran Associates, Inc., 2015.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, jan 1964. doi: 10.1016/0041-5553(64)90137-5.

V. M. Popov. *Hyperstability of Control Systems*. Springer-Verlag, 1973.

Maxim Raginsky, Alexander Rakhlin, and Serdar Yuksel. Online convex programming and regularization in adaptive control. In *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2010. doi: 10.1109/CDC.2010.5717262.

Benjamin Recht. Cs726 - lyapunov analysis and the heavy ball method. Online, October 2012.

Shankar Sastry and Marc Bodson. *Adaptive Control: Stability, Convergence and Robustness*. Prentice-Hall, 1989.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2011. doi: 10.1561/2200000018.

Bin Shi, Simon S. Du, Weijie J. Su, and Michael I. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5744–5752. Curran Associates, Inc., 2019.

Daniel Silver, Qiang Yang, and Lianghao Li. Lifelong machine learning systems: Beyond learning algorithms. In *AAAI Spring Symposium Series*, 2013.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. 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 1139–1147. PMLR, 2013.

Sebastian Thrun. Lifelong learning algorithms. In *Learning to Learn*, pages 181–209. Springer US, 1998. doi: 10.1007/978-1-4615-5529-2_8.

Sebastian Thrun and Tom M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 15(1-2):25–46, jul 1995. doi: 10.1016/0921-8890(95)00004-Y.

Ruxin Wang and Dacheng Tao. Recent progress in image deblurring. *arXiv preprint arXiv:1409.6838*, 2014.

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, nov 2016. doi: 10.1073/pnas.1614734113.

Bernard Widrow and Samuel D. Stearns. *Adaptive Signal Processing*. Prentice-Hall, 1985.

Ashia Wilson. *Lyapunov Arguments in Optimization*. PhD thesis, University of California, Berkeley, 2018.

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

Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4148–4158. Curran Associates, Inc., 2017.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.