NASA: Achieving Lower Regrets and Faster Rates via Adaptive Stepsizes

Hua Ouyang
Computational Science and Engineering
Georgia Institute of Technology
houyang@cc.gatech.edu

Alexander Gray
Computational Science and Engineering
Georgia Institute of Technology
agray@cc.gatech.edu

ABSTRACT

The classic Stochastic Approximation (SA) method achieves optimal rates under the black-box model. This optimality does not rule out better algorithms when more information about functions and data is available.

We present a family of *Noise Adaptive Stochastic Approximation* (*NASA*) algorithms for online convex optimization and stochastic convex optimization. NASA is an adaptive variant of Mirror Descent Stochastic Approximation. It is novel in its practical variation-dependent stepsizes and better theoretical guarantees. We show that comparing with state-of-the-art adaptive and non-adaptive SA methods, lower regrets and faster rates can be achieved under low-variation assumptions.

Categories and Subject Descriptors

F.1.2 [**Theory of Computation**]: Modes of Computation—*Online computation*; I.2.6 [**Artificial Intelligence**]: Learning—*Parameter learning*; G.1.6 [**Numerical Analysis**]: Optimization—*Gradient methods*

General Terms

Algorithms

Keywords

online learning, online convex optimization, adaptive learning, stochastic optimization

1. INTRODUCTION

Stochastic Approximation (*SA*) method originally proposed by [12] for solving root finding problems is a recursive first-order algorithm:

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \eta_t \mathbf{Y}_t, \qquad (1)$$

where $\eta_t \geq 0$ are called *stepsizes*. This simple paradigm has since then thrived in many fields such as adaptive signal processing, adaptive control, scientific computing and machine learning [6, 2]. In this paper we study SA in two specific and closely-related backgrounds: *online learning* and *stochastic optimization*.

SA is one of the main tools for stochastic optimization [14] which can expressed as:

$$\min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) := \mathbb{E}_{\boldsymbol{\xi}}[f(\mathbf{x}, \boldsymbol{\xi})], \qquad (2)$$

where $\boldsymbol{\xi}$ is a random vector with a fixed (and possibly unknown) probability distribution P. Taking $\mathbf{Y}_t$ in (1) as noisy first order information of Φ leads to the popular Stochastic Gradient Descent (SGD). Under the convexity assumption over f, SA's $O(1/\epsilon^2)$ iteration complexity (or equivalent, $O(1/\sqrt{T})$ rate of convergence for $\Phi(\mathbf{x})$) was shown to be optimal under the black-box (oracle) model [10]. With a strong convexity assumption over f, the $O(1/\sqrt{T})$ rate of convergence can be improved to $O(1/T)$ [9].

As an application of SA to online learning (specifically, online convex optimization), where no i.i.d. assumptions are made upon the sequential functions $\{f_t\}$, the Online Gradient Descent (OGD) algorithm proposed in [17] attains a $O(\sqrt{T})$ regret bound for minimizing general convex functions. When $\{f_t\}$ are H-strongly convex, logarithmic regrets can be attained using $1/Ht$ stepsizes [4].

SA is very sensitive to the selection of stepsizes which are crucial to its convergence. Classic analysis of SA relies on the assumptions that $\sum_{t\to\infty} \eta_t = \infty$ and $\sum_{t\to\infty} \eta_t^2 < \infty$ [6]. Despite the black-box-optimalities of SA, there is no unified theory that guides the selection of optimal stepsizes when more information is provided. Our motivation of this paper is that: a successful SA algorithm must explore the structures and properties of functions and data based on reasonable assumptions.

In a recent work [15], smoothness is considered as an important property for minimizing online nonnegative functions. The authors utilize properties of self-bounded functions termed in [13], and show that the regret of minimizing general convex functions is upper bounded by $O\big(\sqrt{\sum_{t=1}^T f_t(\mathbf{x}^*)}\big)$. This bound equals SA's $O(\sqrt{T})$ bound only in worst cases, and is much benign in low-noise (e.g. separable classification) applications. The underlying intuition is that, when functions are smooth, the gradient vanishes and the function value decreases as $\mathbf{x}_t$ is approaching the optimum $\mathbf{x}^*$. Hence instead of using a data-independent diminishing stepsize scheme (e.g. $1/\sqrt{t} \to 0$), it is preferable to adapt the current stepsize to the sum of (nonnegative) function values over previous iterations. In this paper we generalize this idea to nonsmooth and strongly convex functions and show that even faster rates and lower regret bounds can be attained using stepsizes .

We propose a Noise Adaptive Stochastic Approximation (NASA) scheme for online convex optimization (OCO) and stochastic convex optimization (SCO). It is an adaptive variant of Mirror Descent Stochastic Approximation (MDSA) [9]. NASA is novel in

KDD'12, August 12–16, 2012, Beijing, China.
Copyright 2012 ACM 978-1-4503-1462-6 /12/08 ...$15.00.

its variation-dependent stepsizes and better theoretical guarantees. We show that for OCO and SCO, very low regrets and fast rates can be achieved under low-variation assumptions.

In contrast to previous adaptive online learning works by [8] and [3], where the authors proposed to use constant stepsizes and data-dependent Mahalanobis distance $\psi(\mathbf{z}, \mathbf{x}_t) = (\mathbf{z}-\mathbf{x}_t)^T Q_t(\mathbf{z}-\mathbf{x}_t)$ as regulators (prox-functions), NASA's prox-function follows that of MDSA that is much more general than the Mahalanobis distance. Moreover, as we show in Section 4.1, NASA's regret bound is lower than AdaGrad [3] in most circumstances.

2. NOTATIONS AND ASSUMPTIONS

The subgradient of a function f is denoted as g. When f differentiable at $\mathbf{x}$, we also use ∇f to denote its gradient and $\nabla f(\mathbf{x}) = g(\mathbf{x})$. In OCO, we use subscripts t to denote the function and its (sub)gradients in round t. In SCO (2), we denote the (sub)gradient of Φ as $\Phi'(\mathbf{x}) := \mathbb{E}_{\boldsymbol{\xi}} g(\mathbf{x}, \boldsymbol{\xi})$. To unify some analysis which applies for both stochastic and online settings, we use $f_t(\mathbf{x})$ and $g_t(\mathbf{x})$ interchangeably with $f(\mathbf{x}, \boldsymbol{\xi}_t)$ and $g(\mathbf{x}, \boldsymbol{\xi}_t)$.

In the proposed noise-adaptive algorithms, *variations* play important roles in the analysis and results, and hence deserve their own notations. In OCO, we denote

$$\bar{g}_{1:t} := \frac{1}{t}\sum_{i=1}^{t} g_i(\mathbf{x}_i), \tag{3}$$

$$\bar{\delta}_t := g_t(\mathbf{x}_t) - \bar{g}_{1:t} \quad \text{and} \quad \gamma_t := \bar{g}_{1:t} - \bar{g}_{1:t-1}; \tag{4}$$

while in SCO with i.i.d. assumptions, we use

$$\Phi'(\mathbf{x}) := \mathbb{E}_{\boldsymbol{\xi}} g(\mathbf{x}, \boldsymbol{\xi}) \quad \text{and} \quad \delta_t := g_t(\mathbf{x}) - \Phi'(\mathbf{x}). \tag{5}$$

A norm $\|\cdot\|$ is endowed with an inner product $\langle \cdot, \cdot \rangle$. We denote the corresponding dual norm as $\|\cdot\|_*$. A function f is *H-strongly convex* in $\mathcal{X}$ w.r.t. $\|\cdot\|$ iff for any $\mathbf{a}, \mathbf{b} \in \mathcal{X}$, there is a constant $H > 0$ such that

$$f(\mathbf{a}) - f(\mathbf{b}) \geq \langle g(\mathbf{b}), \mathbf{a} - \mathbf{b} \rangle + \frac{H}{2}\|\mathbf{a} - \mathbf{b}\|^2. \tag{6}$$

A *distance generating function* ω is differentiable and α-strongly convex w.r.t. $\|\cdot\|$. The corresponding *Bregman divergence* ψ is defined as

$$\psi(\mathbf{a}, \mathbf{b}) := \omega(\mathbf{a}) - \omega(\mathbf{b}) - \langle \nabla\omega(\mathbf{b}), \mathbf{a} - \mathbf{b} \rangle. \tag{7}$$

Due to convexity of ω, Bregman divergence is always nonnegative.

In OCO, the *regret* is defined as $R(T) := \sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}^*)$, where $\mathbf{x}^* := \arg\min_{\mathbf{z}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{z})$. In SCO, $\mathbf{x}^* := \arg\min_{\mathbf{z}\in\mathcal{X}} \Phi(\mathbf{z})$.

The following notations appear frequently in our analysis and results:

$$\Delta_t := \max_{1\leq i\leq t} \|\bar{\delta}_i\|_*, \ \ \Gamma_t := \max_{1\leq i\leq t} \|\gamma_i\|_*. \tag{8}$$

The convexity assumption is made throughout the paper:

ASSUMPTION 2.1. *Function $f : \mathbb{R}^d \to \mathbb{R}$ is proper and convex, and $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex closed set. Moreover, f is M-Lipschitz continuous in $\mathcal{X}$:*

$$|f(\mathbf{a}) - f(\mathbf{b})| \leq M\|\mathbf{a} - \mathbf{b}\| \ \ \forall \mathbf{a}, \mathbf{b} \in \mathcal{X}, \ \text{where } M > 0.$$

We list three more assumptions that will be used either independently or jointly with others in the analysis.

ASSUMPTION 2.2. *Function $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable with L-Lipschitz continuous gradient ∇f w.r.t. $\|\cdot\|_*$:*

$$\|\nabla f(\mathbf{a}) - \nabla f(\mathbf{b})\|_* \leq L\|\mathbf{a} - \mathbf{b}\|, \ \ \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^d, \ \text{where } L > 0.$$

Functions satisfy this assumption are also called *L-Lipschitz smooth*.

ASSUMPTION 2.3. *Function $f : \mathbb{R}^d \to \mathbb{R}$ is H-strongly convex in $\mathcal{X}$ w.r.t. $\|\cdot\|_*$, where $H > 0$.*

ASSUMPTION 2.4. *Bregman divergence $\psi(\cdot, \cdot)$ grows Q-quadratically w.r.t. $\|\cdot\|_*$:*

$$\psi(\mathbf{a}, \mathbf{b}) \leq \frac{Q}{2}\|\mathbf{a} - \mathbf{b}\|_*^2, \ \ \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^d, \ \text{where } Q > 0.$$

3. NOISE ADAPTIVE STOCHASTIC APPROXIMATION (NASA)

Mirror Descent Stochastic Approximation (MDSA) is a non-Euclidean generalization of the classic SA. It is employed as our basis algorithm for both OCO and SCO due to its universality [16]. MDSA's updating rule can be expressed as minimizing a regularized first order approximation of f at the current solution:

$$\mathbf{x}_{t+1} \leftarrow \arg\min_{\mathbf{z}\in\mathcal{X}} \eta_t \langle g_t(\mathbf{x}_t), \mathbf{z} - \mathbf{x}_t \rangle + \psi(\mathbf{z}, \mathbf{x}_t), \tag{9}$$

where $\eta_t \geq 0$ is a stepsize which plays an important role in the convergence analysis. Classic SA and MDSA typically use a data-independent stepsize or a constant stepsize for all iterations. In all variations of NASA, we propose to use adaptive stepsizes $\eta_t(\delta_t)$, which is a function of variations δ_t of (sub)gradients over previous samples.

Section 4 is devoted to online learning, where we propose two variants of NASA for convex and strongly convex functions. The same adaptation idea is extended to stochastic optimization in Section 5.

4. NASA FOR ONLINE LEARNING

4.1 General Convex Functions

For general functions where we know nothing other than their convexities[1] (Assumption 2.1), Noise Adaptive Stochastic Approximation (NASA) uses MDSA's update scheme (9) together with the following stepsizes:

$$\eta_t = \frac{\sqrt{\alpha} D_{\psi,\mathcal{X}}}{\sqrt{\sum_{i=1}^{t} \left\|\bar{\delta}_i\right\|_*^2}}, \tag{10}$$

where we denote the upper bound of Bregman divergence:

$$D_{\psi,\mathcal{X}} := \max_t \sqrt{\psi(\mathbf{x}_t, \mathbf{x}^*)}. \tag{11}$$

Using updating scheme (9) and optimality conditions, we have the following key lemma that relates three vectors: $\mathbf{w}$, $\mathbf{w}^+$ and $\mathbf{v}$.

LEMMA 4.1. *Denote $\mathbf{w}^+ := \arg\min_{\mathbf{z}\in\mathcal{X}} \langle \mathbf{u}, \mathbf{z} - \mathbf{w} \rangle + \psi(\mathbf{z}, \mathbf{w})$, where ψ is defined in (7). For any $\mathbf{v}, \mathbf{w} \in \mathcal{X}$ and $\mathbf{u}, \bar{\mathbf{u}} \in \mathbb{R}^d$:*

$$\begin{aligned}\langle \mathbf{u}, \mathbf{w} - \mathbf{v} \rangle \leq{}& \psi(\mathbf{v}, \mathbf{w}) - \psi(\mathbf{v}, \mathbf{w}^+) \\ &+ \frac{\|\mathbf{u} - \bar{\mathbf{u}}\|_*^2}{2\alpha} + \langle \bar{\mathbf{u}}, \mathbf{w} - \mathbf{w}^+ \rangle.\end{aligned} \tag{12}$$

PROOF. Since $\langle \mathbf{u}, \mathbf{z} - \mathbf{w} \rangle + \psi(\mathbf{z})$ is convex and differentiable and $\mathcal{X}$ is convex, the necessary and sufficient optimality condition for $\mathbf{w}^+$ is

$$\langle \mathbf{u} + \nabla\psi(\mathbf{w}^+, \mathbf{w}), \mathbf{v} - \mathbf{w}^+ \rangle \geq 0, \ \ \forall \mathbf{v} \in \mathcal{X}.$$

[1] Even if we know that a function is strongly convex, the algorithms and analysis in this section still apply, and they are robust to the overestimation of strong-convexity constants.

Recalling the definition of ψ, we have $\nabla\psi(\mathbf{z},\mathbf{w}) = \nabla\omega(\mathbf{z}) - \nabla\omega(\mathbf{w})$ and the above can be written as

$$\langle \nabla\omega(\mathbf{w}^+) - \nabla\omega(\mathbf{w}) + \mathbf{u}, \mathbf{w}^+ - \mathbf{v}\rangle \le 0. \tag{13}$$

It follows that: $\psi(\mathbf{v},\mathbf{w}^+) - \psi(\mathbf{v},\mathbf{w})$

$$\begin{aligned}
&= \left[\omega(\mathbf{v}) - \omega(\mathbf{w}^+) - \langle \nabla\omega(\mathbf{w}^+), \mathbf{v} - \mathbf{w}^+\rangle\right] \\
&- \left[\omega(\mathbf{v}) - \omega(\mathbf{w}) - \langle \nabla\omega(\mathbf{w}), \mathbf{v} - \mathbf{w}\rangle\right] \\
&= \langle \nabla\omega(\mathbf{w}^+) - \nabla\omega(\mathbf{w}) + \mathbf{u}, \mathbf{w}^+ - \mathbf{v}\rangle + \langle \mathbf{u}, \mathbf{v} - \mathbf{w}^+\rangle \\
&- \left[\omega(\mathbf{w}^+) - \omega(\mathbf{w}) - \langle \nabla\omega(\mathbf{w}), \mathbf{w}^+ - \mathbf{w}\rangle\right] \qquad (14) \\
&\overset{(13)}{\le} \langle \mathbf{u}, \mathbf{v} - \mathbf{w}^+\rangle - \psi(\mathbf{w}^+, \mathbf{w}) \\
&= \langle \mathbf{u} - \bar{\mathbf{u}}, \mathbf{w} - \mathbf{w}^+\rangle - \psi(\mathbf{w}^+, \mathbf{w}) \\
&+ \langle \mathbf{u}, \mathbf{v} - \mathbf{w}\rangle + \langle \bar{\mathbf{u}}, \mathbf{w} - \mathbf{w}^+\rangle
\end{aligned}$$

By Hölder's Inequality,

$$\begin{aligned}
&\langle \mathbf{u} - \bar{\mathbf{u}}, \mathbf{w} - \mathbf{w}^+\rangle \le \|\mathbf{u} - \bar{\mathbf{u}}\|_* \|\mathbf{w} - \mathbf{w}^+\| \\
&\le \frac{\|\mathbf{u} - \bar{\mathbf{u}}\|_*^2}{2\alpha} + \frac{\alpha}{2}\|\mathbf{w} - \mathbf{w}^+\|^2 \le \frac{\|\mathbf{u} - \bar{\mathbf{u}}\|_*^2}{2\alpha} + \psi(\mathbf{w}^+, \mathbf{w}),
\end{aligned} \tag{15}$$

where the last inequality is due to the α-strong convexity of $\omega()$ and definition (7). Combining (14) and (15) and rearranging our claim follows. □

We now apply this lemma to online MDSA.

LEMMA 4.2. *Under Assumption 2.1, we have $\forall \mathbf{v} \in \mathcal{X}$:*

$$\begin{aligned}
\sum_{t=1}^{T} \left[f_t(\mathbf{x}_t) - f_t(\mathbf{v})\right] &\le \frac{D_{\psi,\mathcal{X}}^2}{\eta_T} \\
&+ \sum_{t=1}^{T} \left[\frac{\eta_t}{2\alpha}\|\bar{\delta}_t\|_*^2 + \langle \bar{g}_{1:t}, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle\right].
\end{aligned} \tag{16}$$

PROOF. By convexity and letting $\mathbf{u} = \eta_t g_t(\mathbf{x}_t)$, $\bar{\mathbf{u}} = \eta_t \bar{g}_{1:t}$, $\mathbf{w} = \mathbf{x}_t$ and $\mathbf{w}^+ = \mathbf{x}_{t+1}$ in Lemma 4.1 we have $\forall \mathbf{v} \in \mathcal{X}$: $f_t(\mathbf{x}_t) - f_t(\mathbf{v}) \le$

$$\begin{aligned}
&\langle g_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{v}\rangle \le \eta_t^{-1}\left[\psi(\mathbf{v},\mathbf{x}_t) - \psi(\mathbf{v},\mathbf{x}_{t+1})\right] \\
&+ \frac{\eta_t}{2\alpha}\|\bar{\delta}_t\|_*^2 + \langle \bar{g}_{1:t}, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle.
\end{aligned} \tag{17}$$

Summing up (17) over $t = 1, \ldots, T$ and using (11) the result follows. □

Note that this lemma differs from Theorem 1 of [17] in two places. One is that the squared magnitude of subgradient g is now replaced by $\|\bar{\delta}_t\|_*^2$; the other difference is the additional term $\langle \bar{g}_{1:t}, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle$. Before proceeding, we present a lemma that will help bounding this additional term.

LEMMA 4.3. *For any $t \ge 1$,*

$$\bar{g}_{1:t} - \bar{g}_{1:t-1} = \frac{\bar{\delta}_t}{t} + \frac{\gamma_t}{t}, \tag{18}$$

where $\bar{\delta}_t$ and γ_t are defined in (4) and we let $\bar{g}_{1:0} = 0$.

PROOF. For any $t \ge 1$ we have

$$\begin{aligned}
\bar{g}_{1:t+1} - \bar{g}_{1:t} &= \frac{1}{t+1}\sum_{i=1}^{t+1} g_i(\mathbf{x}_i) - \frac{1}{t}\sum_{i=1}^{t} g_i(\mathbf{x}_i) \\
&= \frac{1}{t(t+1)}\left[t g_{t+1}(\mathbf{x}_{t+1}) - \sum_{i=1}^{t} g_i(\mathbf{x}_i)\right] \\
&= \frac{1}{t+1}\left[g_{t+1}(\mathbf{x}_{t+1}) - \bar{g}_{1:t}\right] \\
&= \frac{1}{t+1}\left\{\left[g_{t+1}(\mathbf{x}_{t+1}) - \bar{g}_{1:t+1}\right] + \left[\bar{g}_{1:t+1} - \bar{g}_{1:t}\right]\right\}.
\end{aligned}$$

Combing with $\bar{g}_{1:1} = \frac{\bar{\delta}_1}{1} + \frac{\gamma_1}{1}$ our claim is proved. □

Rearranging (18) we have:

COROLLARY 4.4. *For any $t \ge 2$, $\gamma_t = \frac{\bar{\delta}_t}{t-1}$.*

This corollary is not used to obtain regret bounds, yet it helps to understand the physical meanings of our bounds. If we regard $\|\bar{\delta}_t\|_*$ as the *magnitude* of variations, then $\|\gamma_t\|_*$ can be regarded as the *rate* of variations.

We are now ready to present the very low regret bound.

THEOREM 4.5. *Under Assumption 2.1, NASA's regret has the following upper bound:*

$$\begin{aligned}
R(T) \le\ & \frac{2D_{\psi,\mathcal{X}}}{\sqrt{\alpha}}\sqrt{\sum_{t=1}^{T}\|\bar{\delta}_t\|_*^2 - C} \\
&+ B_{\mathcal{X}}\left[\Delta_T\left(1 + \log\sum_{t=s_1}^{T}\frac{\|\bar{\delta}_t\|_*}{\Delta_T}\right)\right. \\
&\left. + \Gamma_T\left(1 + \log\sum_{t=t_1}^{T}\frac{\|\gamma_t\|_*}{\Gamma_T}\right)\right]
\end{aligned} \tag{19}$$

where Δ_T and Γ_T are defined in (8), $D_{\psi,\mathcal{X}}$ is defined in (11), $B_{\mathcal{X}} := \max_t \|\mathbf{x}_t\|$, $C := \langle \bar{g}_{1:T}, \mathbf{x}_{T+1}\rangle$ and we denote $s_1 := \min\left\{\tau : \sum_{i=1}^{\tau}\|\bar{\delta}_i\|_*/\Delta_T \ge 1\right\}$ and
$t_1 := \min\left\{\tau : \sum_{i=1}^{\tau}\|\gamma_i\|_*/\Gamma_T \ge 1\right\}$.

PROOF. Replacing η_t in (16) with NASA's stepsize (10) and letting $\mathbf{v} = \mathbf{x}^*$ we have

$$\begin{aligned}
R(T) \le\ & \frac{D_{\psi,\mathcal{X}}}{\sqrt{\alpha}}\sqrt{\sum_{t=1}^{T}\|\bar{\delta}_t\|_*^2} + \frac{D_{\psi,\mathcal{X}}}{2\sqrt{\alpha}}\sum_{t=1}^{T}\frac{\|\bar{\delta}_t\|_*^2}{\sqrt{\sum_{i=1}^{t}\|\bar{\delta}_i\|_*^2}} \\
&+ \sum_{t=1}^{T}\langle \bar{g}_{1:t}, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle.
\end{aligned} \tag{20}$$

By Lemma A.1 in the appendix we can bound the second term on the RHS of above as

$$\frac{D_{\psi,\mathcal{X}}}{2\sqrt{\alpha}}\sum_{t=1}^{T}\frac{\|\bar{\delta}_t\|_*^2}{\sqrt{\sum_{i=1}^{t}\|\bar{\delta}_i\|_*^2}} \le \frac{D_{\psi,\mathcal{X}}}{\sqrt{\alpha}}\sqrt{\sum_{t=1}^{T}\|\bar{\delta}_t\|_*^2}. \tag{21}$$

By Lemma 4.3 we can bound the third term on the RHS of (20) as below.

$$\begin{aligned}
&\sum_{t=1}^{T}\langle \bar{g}_{1:t}, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle + C = \sum_{t=1}^{T}\langle \bar{g}_{1:t} - \bar{g}_{1:t-1}, \mathbf{x}_t\rangle \\
&\le \sum_{t=1}^{T}\|\bar{g}_{1:t} - \bar{g}_{1:t-1}\|_*\|\mathbf{x}_t\| \overset{(18)}{\le} B_{\mathcal{X}}\sum_{t=1}^{T}\left\|\frac{\bar{\delta}_t + \gamma_t}{t}\right\|_* \\
&\le B_{\mathcal{X}}\sum_{t=1}^{T}\left(\left\|\frac{\bar{\delta}_t}{t}\right\|_* + \left\|\frac{\gamma_t}{t}\right\|_*\right) \\
&\le B_{\mathcal{X}}\left[\Delta_T\sum_{t=1}^{T}\frac{\|\bar{\delta}_t\|_*/\Delta_T}{t} + \Gamma_T\sum_{t=1}^{T}\frac{\|\gamma_t\|_*/\Gamma_T}{t}\right] \\
&\le B_{\mathcal{X}}\Delta_T\left(1 + \sum_{t=s_1}^{T}\frac{\|\bar{\delta}_t\|_*/\Delta_T}{\sum_{i=1}^{t}\|\bar{\delta}_i\|_*/\Delta_T}\right) \\
&+ B_{\mathcal{X}}\Gamma_T\left(1 + \sum_{t=t_1}^{T}\frac{\|\gamma_t\|_*/\Gamma_T}{\sum_{i=1}^{t}\|\gamma_i\|_*/\Gamma_T}\right)
\end{aligned} \tag{22}$$

Applying Lemma A.3 in the appendix to the last inequality our claim follows. Note that in (19), if $T < s_1$ or $T < t_1$, then the corresponding logarithmic term vanishes. □

REMARK 4.6. We can compare NASA with Online Gradient Descent (OGD, [17]) and Online Adaptive Subgradient Method (AdaGrad, [3]). In OGD, $O(1/\sqrt{t})$ data-independent stepsizes are suggested. Denote $G_T := \max_{1\le t\le T}\|g_t(\mathbf{x}_t)\|_*$. Taking the optimal stepsize

$$\eta_t = \frac{\sqrt{\alpha}D_{\psi,\mathcal{X}}}{G_T\sqrt{t}} \tag{23}$$

we have

$$R_{\text{OGD}}(T) \le \frac{2D_{\psi,\mathcal{X}}G_T\sqrt{T}}{\sqrt{\alpha}}. \tag{24}$$

In AdaGrad, if we take an adaptive diagonal scaling matrix in its Mahalanobis distance prox-function [8], the equivalent stepsize can be written as:

$$\eta_t = \frac{\sqrt{\alpha}D_{\psi,\mathcal{X}}}{\sqrt{\sum_{i=1}^{t}\|g_i(\mathbf{x}_i)\|_*^2}} \tag{25}$$

and we have

$$R_{\text{AdaGrad}}(T) \le \frac{2D_{\psi,\mathcal{X}}}{\sqrt{\alpha}}\sqrt{\sum_{t=1}^{T}\|g_t(\mathbf{x}_t)\|_*^2}. \tag{26}$$

In an adversarial situation, e.g. the adversary has access to any mixed strategy such that $\|g_t(\mathbf{x}_t)\|_* = G\ \forall t = \{1 : t\}$, both OGD (24) and AdaGrad (26) retain $D_{\psi,\mathcal{X}}G\sqrt{T}$ bounds, while NASA's bound (19) becomes a constant. In benign cases, for example when $g_t(\mathbf{x}_t)$ are constants for all iterations except the k^{th} iteration:
$\|g_t(\mathbf{x}_t)\|_*^2 = T^{1/p-1}\ \forall t = \{1 : t\}\backslash k$ for some $p \ge 1$ and $\|g_k(\mathbf{x}_k)\|_*^2 = G^2$, then $R_{\text{OGD}}(T)$ is $O(T^{1/2})$, $R_{\text{AdaGrad}}(T)$ is $O(T^{1/(2p)})$, while NASA still retains a constant bound, providing that T is large enough to averaging out $\|\bar{\delta}_k\|_*$.

Besides, NASA's stepsize (10) is more practical than OGD's (23) where one needs to estimate G in addition to $D_{\psi,\mathcal{X}}$. A bad estimation of G, either too small or too large, will results in a even larger regret for OGD. In practice, even if one knows the form of function f_t and its subgradient, it is almost impossible to estimate G appropriately without seeing any data.

In contrast to AdaGrad, NASA is more flexible and it can be applied to any prox-function of a Bregman divergence style, while AdaGrad only applies to Mahalanobis distance prox-functions.

4.2 Strongly Convex Functions

In this section, we assume that the functions are strongly convex (Assumption 2.3) and the Bregman divergence $\psi()$ in MDSA grows Q-quadratically (Assumption 2.4).

For strongly convex functions, NASA uses MDSA's update scheme (9) together with the following stepsizes:

$$\eta_t = \begin{cases} \frac{Q}{H} & \text{if } t < r_1, \\ \frac{Q\Delta_t^2}{H\sum_{i=1}^{t}\|\bar{\delta}_i\|_*^2} & \text{otherwise,} \end{cases} \tag{27}$$

$$\text{where} \quad r_1 := \min\left\{\tau : \sum_{i=1}^{\tau}\|\bar{\delta}_i\|_*^2 \ge 1\right\}. \tag{28}$$

We can bound the cumulative function value reductions as follows.

LEMMA 4.7. *Under Assumption 2.1, 2.3 and 2.4, for any nonnegative stepsizes that satisfy $\frac{1}{\eta_1} \le \frac{H}{Q}$ and $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \le \frac{H}{Q}\ \forall t \ge 2$, we have $\forall \mathbf{v} \in \mathcal{X}$:*

$$\sum_{t=1}^{T}[f_t(\mathbf{x}_t) - f_t(\mathbf{v})] \le \sum_{t=1}^{T}\left[\frac{\eta_t}{2\alpha}\|\bar{\delta}_t\|_*^2 + \langle \bar{g}_{1:t}, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle\right]. \tag{29}$$

PROOF. By the definition of strong-convexity (6), Lemma 4.1 and Assumption 2.4 we have $\forall \mathbf{v} \in \mathcal{X}$:

$$\begin{aligned}
&f_t(\mathbf{x}_t) - f_t(\mathbf{v}) \le \langle g_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{v}\rangle - \frac{H}{2}\|\mathbf{x}_t, \mathbf{v}\|_*^2 \\
&\le \frac{1}{\eta_t}[\psi(\mathbf{x}_t, \mathbf{v}) - \psi(\mathbf{x}_{t+1}, \mathbf{v})] - \frac{H}{Q}\psi(\mathbf{x}_t, \mathbf{v}) \\
&+ \frac{\eta_t}{2\alpha}\|\bar{\delta}_t\|_*^2 + \langle \bar{g}_{1:t}, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle.
\end{aligned} \tag{30}$$

Summing up the first two terms of RHS over $t = 1, \ldots, T$ we have:

$$\begin{aligned}
&\sum_{t=1}^{T}\left[\frac{1}{\eta_t}[\psi(\mathbf{x}_t, \mathbf{v}) - \psi(\mathbf{x}_{t+1}, \mathbf{v})] - \frac{H}{Q}\psi(\mathbf{x}_t, \mathbf{v})\right] \\
&\le \sum_{t=2}^{T}\left[\psi(\mathbf{x}_t, \mathbf{v})\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \frac{H}{Q}\right)\right] \\
&- \frac{\psi(\mathbf{x}_{T+1}, \mathbf{v})}{\eta_T} + \frac{\psi(\mathbf{x}_1, \mathbf{v})}{\eta_1} - \frac{H}{Q}\psi(\mathbf{x}_1, \mathbf{v}) \le 0,
\end{aligned} \tag{31}$$

where the last inequality is due to the nonnegativity of Bregman divergences. Thus the claim is proved. □

We are now ready to present the low regret result for strongly convex OCO problems.

THEOREM 4.8. *Under Assumption 2.1, 2.3 and 2.4, NASA's regret has the following upper bound:*

$$\begin{aligned}
&R(T) \le \frac{Q}{2\alpha H}\left[1 + \Delta_T^2\log\sum_{t=r_1}^{T}\|\bar{\delta}_t\|_*^2\right] - C \\
&+ B_{\mathcal{X}}\left[\Delta_T\left(1 + \log\sum_{t=s_1}^{T}\frac{\|\bar{\delta}_t\|_*}{\Delta_T}\right)\right. \\
&\left.+ \Gamma_T\left(1 + \log\sum_{t=t_1}^{T}\frac{\|\gamma_t\|_*}{\Gamma_T}\right)\right]
\end{aligned} \tag{32}$$

where r_1 is defined in (28), and s_1, t_1, C are defined in Theorem 4.5.

PROOF. We first show that in the nontrivial case where $t \geq r_1$, NASA's stepsizes (27) satisfy the condition $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \leq \frac{H}{Q}$ in Lemma 4.7. For any fixed $t \geq 2$, if $\|\bar{\delta}_t\|_*^2 \geq \Delta_{t-1}^2$, then

$$\begin{aligned}\frac{Q}{H}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) &= \frac{\sum_{i=1}^{t} \|\bar{\delta}_i\|_*^2}{\|\bar{\delta}_t\|_*^2} - \frac{\sum_{i=1}^{t-1} \|\bar{\delta}_i\|_*^2}{\Delta_{t-1}^2} \\ &\leq \frac{\sum_{i=1}^{t} \|\bar{\delta}_i\|_*^2}{\|\bar{\delta}_t\|_*^2} - \frac{\sum_{i=1}^{t-1} \|g_i(\mathbf{x}_i)\|_*^2}{\|\bar{\delta}_t\|_*^2} = 1.\end{aligned}$$

If $\|\bar{\delta}_t\|_*^2 < \Delta_{t-1}^2$, then

$$\begin{aligned}\frac{Q}{H}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) &= \frac{\sum_{i=1}^{t} \|\bar{\delta}_i\|_*^2}{\Delta_{t-1}^2} - \frac{\sum_{i=1}^{t-1} \|\bar{\delta}_i\|_*^2}{\Delta_{t-1}^2} \\ &= \frac{\|\bar{\delta}_t\|_*^2}{\Delta_{t-1}^2} < 1.\end{aligned}$$

The regret bound follows immediately. Replacing $\mathbf{v}$ with $\mathbf{x}_*$ in (29) we have:

$$\begin{aligned}R(T) &\leq \frac{Q}{2\alpha H}\sum_{t=1}^{T} \frac{\Delta_t^2 \|\bar{\delta}_t\|_*^2}{\sum_{i=1}^{t} \|\bar{\delta}_i\|_*^2} + \sum_{t=1}^{T} \langle \bar{g}_{1:t}, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\ &\leq \frac{Q}{2\alpha H}\left(\sum_{t=1}^{r_1 - 1} \|\bar{\delta}_t\|_*^2 + \sum_{t=r_1}^{T} \frac{\Delta_t^2 \|\bar{\delta}_t\|_*^2}{\sum_{i=1}^{t} \|\bar{\delta}_i\|_*^2}\right) \qquad (33) \\ &+ \sum_{t=1}^{T} \langle \bar{g}_{1:t}, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle\end{aligned}$$

Due to the fact that $\Delta_t \leq \Delta_T$ and combine Lemma A.3 and (22) the claim is proved. As in Theorem 4.5, if $T < r_1$ or $T < s_1$ or $T < t_1$, the corresponding logarithmic term vanishes. □

REMARK 4.9. Although [8] and [3] are only proposed for general convex functions and there is no strongly convex counterparts, we can have an analogy to NASA (27) and let

$$\eta_t = \frac{QG_t^2}{H\sum_{i=1}^{t} \|g_i(\mathbf{x}_i)\|_*^2}, \qquad (34)$$

where $G_t := \max_{1 \leq i \leq t} \|g_i(\mathbf{x}_i)\|_*$. Let us call it SAdaGrad, and its regret bound is

$$R_{\text{SAdaGrad}}(T) \leq \frac{Q}{2\alpha H}\left[1 + G_T^2 \log \sum_{t=1}^{T} \|g_t(\mathbf{x}_t)\|_*^2\right]. \qquad (35)$$

We can compare NASA with SAdaGrad and strongly convex OGD [4] where logarithmic regret

$$R_{\text{OGD}}(T) \leq \frac{QG_T^2}{2\alpha H}(1 + \log T), \qquad (36)$$

is attained by taking the data-independent stepsize $\eta_t = \frac{Q}{Ht}$. In an adversarial case where $\|g_t(\mathbf{x}_t)\|_*^2 = G$, $\forall t = 1, \ldots, T$, NASA's bound (33) becomes a constant since $\|\bar{\delta}_t\|_* = 0$, while both R_{OGD} and R_{SAdaGrad} are of $O(\log T)$. In benign cases, for example when $g_t(\mathbf{x}_t)$ are constants for all iterations except the k^{th} iteration: $\|g_t(\mathbf{x}_t)\|_*^2 = T^{1/p-1}$ $\forall t = \{1 : t\} \backslash k$ for some $p \geq 1$ and $\|g_k(\mathbf{x}_k)\|_*^2 = G^2$, then $R_{\text{OGD}}(T)$ is $O(\log T)$, $R_{\text{SAdaGrad}}(T)$ is $O(\log T^{1/p})$, while NASA still remains a constant bound, providing that T is large enough to averaging out $\|\bar{\delta}_k\|_*$.

5. NASA FOR STOCHASTIC OPTIMIZATION

The idea of noise-adaptive SA can be naturally extended to stochastic convex optimization problems (SCO), where $\{f_t\}$ are i.i.d. We first quickly review two recent progresses in solving SCO.

RMDSA [9] is a stochastic counterpart of mirror descent for SCO. Taking time-varying stepsize policy $\eta_t = \frac{\sqrt{\alpha} D_{\psi,\mathcal{X}}}{G\sqrt{t}}$, RMDSA's rate of convergence for minimizing general convex functions is

$$\mathbb{E}[\Phi(\tilde{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)] \leq \frac{D_{\psi,\mathcal{X}} G}{\sqrt{\alpha}} \cdot \frac{3 + \ln T}{4(\sqrt{T+1} - 1)}, \qquad (37)$$

where $\tilde{\mathbf{x}}_T := \left(\sum_{t=1}^{T} \eta_t\right)^{-1} \sum_{t=1}^{T} \eta_t \mathbf{x}_t$ and $G := \sup_{\mathbf{x}} \mathbb{E}[\|g(\mathbf{x}, \boldsymbol{\xi})\|_*]$. This bound can be improved when smoothness assumptions are made. With Lipschitz gradients, Nesterov's accelerated gradient descent (AGD) is proved to be optimal for convex programming [11]. Stochastic versions of AGD [7], [5] are proposed and shown to outperform RMDSA. In following we show that when applying NASA to stochastic AGD, even faster rates can be attained for optimizing smooth or nonsmooth functions.

Algorithm 1 lists the accelerated MDSA (aka AC-SA) proposed in [7]. To unify the analysis for both smooth and nonsmooth functions we adopt the "composite" setting: $f_t(\mathbf{x}) = p_t(\mathbf{x}) + q_t(\mathbf{x})$, where p_t is a convex function satisfying Assumption 2.1, and q_t is a smooth function satisfying Assumption 2.2

Algorithm 1 Accelerated MDSA (AC-SA)

```
Initialize $\mathbf{x}_0^{\text{ag}} = \mathbf{x}_0 \in \mathcal{X}$
for $t = 1, \ldots, T$ do
    $\mathbf{x}_t^{\text{md}} \leftarrow \beta_t^{-1}\mathbf{x}_{t-1} + (1 - \beta_t^{-1})\mathbf{x}_{t-1}^{\text{ag}}$
    $\mathbf{x}_t \leftarrow \arg\min_{\mathbf{z}\in\mathcal{X}} \left\{\eta_t \left\langle g_t(\mathbf{x}_t^{\text{md}}), \mathbf{z}\right\rangle + \psi(\mathbf{z}, \mathbf{x}_{t-1})\right\}$
    $\mathbf{x}_t^{\text{ag}} \leftarrow \beta_t^{-1}\mathbf{x}_t + (1 - \beta_t^{-1})\mathbf{x}_{t-1}^{\text{ag}}$
end for
```

Two series of stepsizes β_t and η_t need to be specified to complete this algorithm. With the notation $\delta_t := g_t(\mathbf{x}_t^{\text{md}}) - \nabla\Phi(\mathbf{x}_t^{\text{md}})$, we propose the following stepsize policy for NASA:

$$\beta_t = \frac{t+1}{2}, \quad \eta_t = \frac{t+1}{2}\min\left\{\frac{\alpha}{2L}, \frac{\Lambda_t}{t+1}\right\}, \qquad (38)$$

where

$$\Lambda_t := \begin{cases} \sqrt{\alpha} D_{\psi,\mathcal{X}} & \text{if } t < s_1, \\ \frac{\sqrt{\alpha} D_{\psi,\mathcal{X}}}{\sqrt{\sum_{i=1}^{t} (2M + \|\delta_i\|_*)^2}} & \text{otherwise,} \end{cases}$$

$$s_1 := \min\left\{\tau : \sum_{i=1}^{\tau} (2M + \|\delta_i\|_*)^2 \geq 1\right\}. \qquad (39)$$

and

$$D_{\psi,\mathcal{X}} := \max_t \sqrt{\psi(\mathbf{x}_t, \mathbf{x}^*)}. \qquad (40)$$

We firstly present the key lemma of AC-SA, which is due to [7].

LEMMA 5.1. *[7] Assume that the stepsizes β_t and η_t satisfy $\beta_t \geq 1$ and $\alpha\beta_t > L\eta_t$. Under Assumption 2.1, 2.2 and 2.4, using AC-SA (Alg. 1), we have $\forall \mathbf{v} \in \mathcal{X}$:*

$$\begin{aligned}\beta_t\eta_t\left[\Phi(\mathbf{x}_t^{ag}) - \Phi(\mathbf{v})\right] &\leq (\beta_t - 1)\eta_t\left[\Phi(\mathbf{x}_{t-1}^{ag}) - \Phi(\mathbf{v})\right] \\ &+ \psi(\mathbf{x}_{t-1}, \mathbf{v}) - \psi(\mathbf{x}_t, \mathbf{v}) - \eta_t\left\langle\delta_t, \mathbf{x}_{t-1} - \mathbf{v}\right\rangle + U_t\end{aligned} \qquad (41)$$

$$\textit{where} \quad U_t := \frac{\beta_t\eta_t^2\left(2M + \|\delta_t\|_*\right)^2}{2(\alpha\beta_t - L\eta_t)}. \qquad (42)$$

NASA's convergence rate follows immediately.

THEOREM 5.2. *Under Assumption 2.1, 2.2, 2.4, taking NASA's stepsizes (38) we have:* $\mathbb{E}\left[\Phi(\mathbf{x}_T^{ag}) - \Phi(\mathbf{x}^*)\right] \leq$

$$\left[\frac{2LD^2_{\psi,\mathcal{X}}}{\alpha T(T+2)} + \frac{D_{\psi,\mathcal{X}}}{\sqrt{\alpha(T+2)}}\sqrt{\frac{\sum_{t=1}^{T+1}\sigma_t^2}{T+1}}\right]\left(5+\log\sum_{t=1}^{T}\sigma_t^2\right) \quad (43)$$

where $\sigma_t^2 := \mathbb{E}[(2M + \|\delta_t\|_*)^2]$ *and* s_1 *is defined in (39).*

PROOF. See Appendix B. □

6. EXPERIMENTAL RESULTS

6.1 Environment-Changing Ridge Regression

To demonstrate the adaptive capacity of NASA , we make a simulated dataset and perform experiments with it on the ridge regression problem: $f_t(\mathbf{x}) = (y_t - \mathbf{x}^T\mathbf{s}_t)^2 + \lambda\|\mathbf{x}\|_2^2$. Here $(\mathbf{s}_t, y_t)$ represents a data sample and its target response variable at round t. We let $\lambda = 0.0001$ in all the experiments. The data simulation is as follows. Instead of using a linear regressor with a fixed set of coefficients, we have two sets of "true" regression coefficients: $\mathbf{c}_1 = [0.31, -0.45, 0.92, 0.03, -0.75, -0.39]^T$ and $\mathbf{c}_2 = 2\mathbf{c}_1$. We generate $4,000$ data samples $\mathbf{s}_t \in \mathbb{R}^6$ where each dimension is drawn from the normal distribution $\mathcal{N}(2, 10)$. These samples are evenly divided into two subsets. For the first $2,000$ samples, $y_t = \mathbf{c}_1^T\mathbf{s}_t + n_t$, and for $t = 2,001 \sim 4,000$, $y_t = \mathbf{c}_2^T\mathbf{s}_t + n_t$, where $n_t \sim \mathcal{N}(0, 0.01)$ are noises.

We perform two sets of experiments. In the first set we assume a black-box model, hence the general convex NASA (10) is adopted. Strong-convexities are assumed in the second set of experiments, and (27) is adopted. It has been observed that the estimation of the strong-convexity constant H is crucial in the convergence of SA algorithms [1, 9]. Without tackling this drawback, in our experiments, H is searched within $\{1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\}$, and the one with the lowest $\sum_t f_t$ is reported.

The results are depicted in Fig.1. We can see that in both general and strongly convex settings, when $t > 2000$, NASA can very quickly adapt to the changes of the adversary's behavior. OGD can slowly recover from the changes and converges after a much longer time (Fig.1 Left). Comparing with NASA, OGD's total sum of function values for $T = 4000$ is 5-times higher. Strongly convex OGD even fails to converge due to the improperly estimated H.

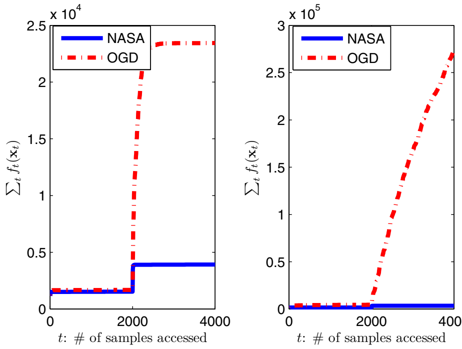

Figure 1: Online ridge regression. Left: convex; Right: strongly convex.

6.2 Real World Data

To evaluate the efficiency of NASA in real world applications, we take three datasets from `www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`. abalone is for regression, while svmguide1 and covtype are for binary classification. The ridge regression and SVM cost functions $f_t(\mathbf{x}) = [1 - y_t\mathbf{x}^T\mathbf{s}_t]_+ + \lambda\|\mathbf{x}\|_2^2$ are adopted. Both convex and strongly convex algorithms are tested, and as in Section 6.1, best H is searched within a fixed set.

The results are shown in Fig.2, 3 and 4. We can see that in all settings NASA outperforms OGD. Since the cost functions are strongly convex, the cumulated function values are slightly lower than when we use strongly convex OGD/NASA. This difference is not as prominent as that in Fig. 1. The reason might be that the loss function weakens the strong-convexity of the squared l_2 norm.

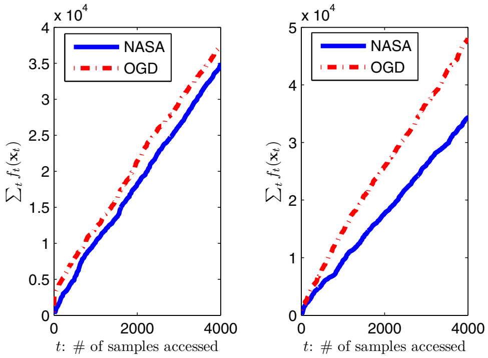

Figure 2: abalone. Left: convex; Right: strongly convex.

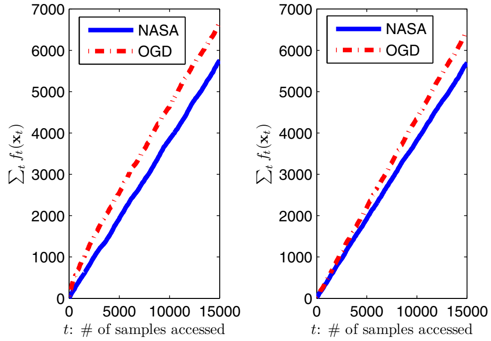

Figure 3: svmguide1. Left: convex; Right: strongly convex.

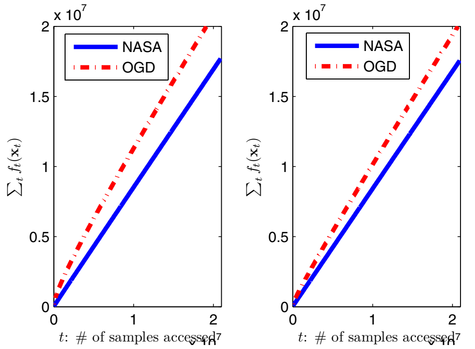

Figure 4: covtype. Left: convex; Right: strongly convex.

6.3 Comparisons to AdaGrad

6.3.1 Simulated Gradients

As discussed in the Remark 4.6, AdaGrad [3] is another way to adaptively take the advantage of previous gradient history, begging the question of whether it has similar behavior as NASA.

In the comparisons of upper bounds for OGD, AdaGrad and NASA presented in Remark 4.6 and 4.9, our statements are supported by extreme cases. In this section we use three sets of simulated experiments to further explore the advantages of NASA vs AdaGrad. Indeed, the comparison depends on the statistical properties of gradients. In these experiments, we assume that gradient vectors are from certain classes of statistical distributions. We observe that in most cases, NASA achieves a lower regret bound than AdaGrad.

In the first set of experiments, we focus on the general convex functions. We assume that each dimension of gradients $g_t(\mathbf{x}_t)$ are from a normal distribution $\mathcal{N}(\mu, \sigma)$. We let the dimension of our data to be 500, and plot the regret bounds for 1000 iterations.

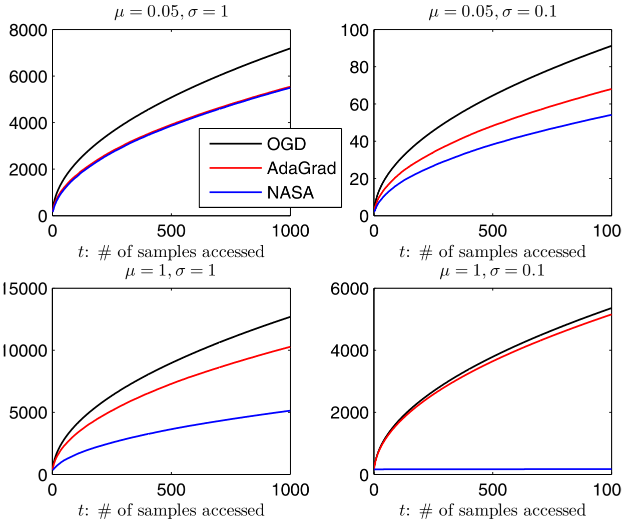

Figure 5: Comparing regret bounds.

Form Fig.5 we can observe that in all the 4 combinations of μ and σ, NASA outperforms both OGD and AdaGrad. When μ is relatively large (lower two figures), NASA achieves a significantly lower regret bound. This corresponds to difficult problems where the noise level is high. Moreover, when σ is relatively small (lower right), AdaGrad's performance is close to OGD, while NASA significantly outperforms its competitors. When μ is small (upper two figures), NASA's advantage is not as significant as that of large μ. This corresponds to the situations where the problem is relatively easy to solve. Moreover, for large σ (upper left), NASA's bound is very close to AdaGrad, but still lower.

In the second set of experiments we assume that each dimension of the gradient vectors has a uniform distribution, which corresponds to a stationary process. The simulation results are illustrated in Fig.6. We can observe that NASA's bound is significantly lower than the other two algorithms, while AdaGrad is marginally better than OGD.

In the third set of experiments we assume that each dimension of the gradient vectors is monotonically approaching 0 at the rate of $\frac{1}{t^p}$, where $p > 0$. In Fig.7 we plot regret bounds for $p = 0.1,\ 0.5, 1$ and 2. When $p = 0.5,\ 1$ and 2, we omit the result for OGD to magnify the differences between NASA and AdaGrad. We can observe that NASA's bound is lower than AdaGrad when $p <= 1$. For extremely easy problems where gradient vectors approaches 0 at a fast rate of $\frac{1}{t^2}$, AdaGrad achieves a slightly lower regret.

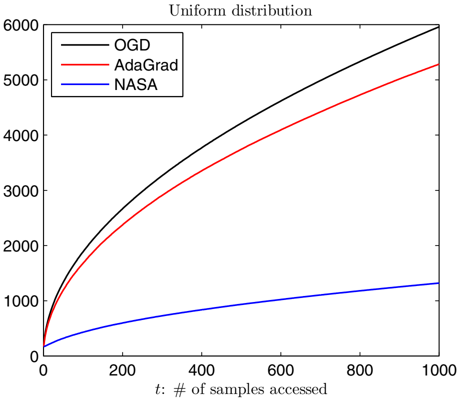

Figure 6: Comparing regret bounds.

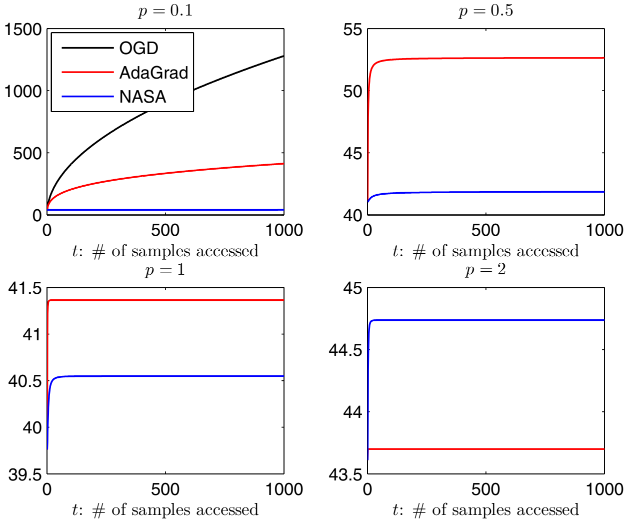

Figure 7: Comparing regret bounds.

6.3.2 Real World Data

To empirically compare the performance of NASA vs AdaGrad, we use a real-world text dataset RCV1 collected from Reuters news stories. We use unregularized hinge loss as our objective function, and do not assume strong convexity. The means and standard deviations of cumulated function values are shown in Table 1. Each value is calculated from 10 experiments under the same setting.

We can see that in all the settings, NASA's cumulated function values are lower than OGD and AdaGrad. When the number of iterations increases, the differences between NASA and AdaGrad

Table 1: Comparing $\sum_t f_t(\mathbf{x}_t)$

	OGD	AdaGrad	NASA
1 epoch	3461.9 ± 48.1	3260.5 ± 40.3	$\mathbf{3253.0 \pm 30.2}$
2 epochs	5268.7 ± 59.3	4688.5 ± 51.2	$\mathbf{4590.2 \pm 33.1}$
3 epochs	6814.1 ± 53.9	5809.5 ± 53.6	$\mathbf{5624.6 \pm 34.9}$
4 epochs	8050.0 ± 43.4	6722.1 ± 69.4	$\mathbf{6426.7 \pm 33.7}$

increases accordingly, which give the empirical evidence of the regret bound simulations presented in Section 6.3.1.

7. CONCLUSIONS

We propose a family of adaptive stochastic approximation algorithms, named NASA, for online convex optimization (both convex and strongly convex assumptions) and stochastic optimization. Their stepsizes are adaptive to the previous-seen data. These algorithms achieve lower regret bounds or faster rates of convergence, provided that the variations of (sub)gradients is low. NASA is under the mirror descent framework, hence is very flexible to various prox-functions. Experiments show that NASA is adaptive to the behaviors of the adversary. They outperform OGD and AdaGrad in both convex and strongly convex minimizations.

8. REFERENCES

[1] P. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Proceedings of NIPS*, 2007.

[2] L. Bottou and N. Murata. Stochastic approximations and efficent learning. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. Cambridge University Press, 2nd edition, 2002.

[3] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Proceedings of COLT*, 2010.

[4] E. Hazan, A. Agarwal, A. Kalai, and S. Kale. Logarithmic regret algorithms for online convex optimization. In *COLT*, 2006.

[5] C. Hu, J. Kwok, and W. Pan. Accelerated gradient methods for stochastic optimization and online learning. In *NIPS*, 2009.

[6] H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition, 2003.

[7] G. Lan. Efficient methods for stochastic composite optimization. Technical report, Georgia Institute of Technology, August 2008.

[8] H. B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of COLT*, 2010.

[9] A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[10] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983.

[11] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

[12] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[13] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

[14] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[15] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low-noise and fast rates. In *NIPS*, 2010.

[16] K. Sridharan, N. Srebro, and A. Tewari. On the universality of online mirror descent. In *Proceedings of NIPS 24*, 2011.

[17] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.

APPENDIX

A. TECHNICAL LEMMAS

LEMMA A.1. *For any $\delta_i \geq 0, i = 1, 2, \dots$ and integers $t \geq s \geq 1$ one has*

$$\sum_{i=s}^{t} \frac{\delta_i}{\sqrt{\sum_{j=1}^{i} \delta_j}} \leq 2 \left(\sqrt{\sum_{i=1}^{t} \delta_i} - \sqrt{\sum_{i=1}^{s-1} \delta_i} \right).$$

PROOF.

$$\sum_{i=s}^{t} \frac{\delta_i}{\sqrt{\sum_{j=1}^{i} \delta_j}} \leq \int_{\sum_{i=1}^{s-1} \delta_i}^{\sum_{i=1}^{t} \delta_i} \frac{1}{\sqrt{m}} \mathrm{d}m$$

$$= 2 \left(\sqrt{\sum_{i=1}^{t} \delta_i} - \sqrt{\sum_{i=1}^{s-1} \delta_i} \right).$$

□

COROLLARY A.2. *For any integers $t \geq i \geq s \geq 1$ one has*

$$2\left(\sqrt{t+1} - \sqrt{s}\right) \leq \sum_{i=s}^{t} \frac{1}{\sqrt{i}} \leq 2\left(\sqrt{t} - \sqrt{s-1}\right).$$

PROOF. For the LHS:

$$\sum_{i=s}^{t} \frac{1}{\sqrt{i}} \geq \int_{s-1}^{t} \frac{1}{\sqrt{m+1}} \mathrm{d}m = 2\left(\sqrt{t+1} - \sqrt{s}\right).$$

The RHS follows immediately from Lemma A.1. □

LEMMA A.3. *For any $\delta_i \geq 0, i = 1, 2, \dots$ and integers $t \geq s \geq 1$ such that $\sum_{i=1}^{s-1} \delta_i \geq 1$ one has*

$$\sum_{i=s}^{t} \frac{\delta_i}{\sum_{j=1}^{i} \delta_j} \leq \ln\left(\sum_{i=1}^{t} \delta_i\right) - \ln\left(\sum_{i=1}^{s-1} \delta_i\right).$$

PROOF. Since $\sum_{i=1}^{s-1} \delta_i \geq 1$,

$$\sum_{i=s}^{t} \frac{\delta_i}{\sum_{j=1}^{i} \delta_j}$$

$$\leq \int_{\sum_{i=1}^{s-1} \delta_i}^{\sum_{i=1}^{t} \delta_i} \frac{1}{m} \mathrm{d}m = \ln\left(\sum_{i=1}^{t} \delta_i\right) - \ln\left(\sum_{i=1}^{s-1} \delta_i\right).$$

□

COROLLARY A.4. *For any integers $t \geq i \geq s \geq 2$ one has*

$$\ln(t+1) - \ln(s) \leq \sum_{i=s}^{t} \frac{1}{i} \leq \ln t - \ln(s-1).$$

PROOF. For the LHS:

$$\sum_{i=s}^{t}\frac{1}{i}\geq\int_{s-1}^{t}\frac{1}{m+1}\mathrm{d}m=\ln(t+1)-\ln(s).$$

The RHS follows immediately from Lemma A.3. □

B. PROOF OF THEOREM 5.2

PROOF. We firstly show that (38) satisfies conditions in Lemma 5.1:

$$\alpha\beta_t=\frac{\alpha(t+1)}{2}\geq 2L\frac{t+1}{2}\frac{\alpha}{2L}\geq 2L\eta_t. \tag{44}$$

Next we show that

$$(\beta_{t+1}-1)\eta_{t+1}\leq\beta_t\eta_t. \tag{45}$$

It is easy to check that $\frac{\Lambda_{t+1}}{t+2}\leq\frac{\Lambda_t}{t+1}$ using the fact that $\|\delta_t\|_*\geq 0$. If $\frac{\Lambda_{t+1}}{t+2}\geq\frac{\alpha}{2L}$, then $\frac{\Lambda_t}{t+1}\geq\frac{\alpha}{2L}$, hence

$$(\beta_{t+1}-1)\eta_{t+1}=\frac{t^2+2t}{4}\frac{\alpha}{2L}\leq\frac{t^2+2t+1}{4}\frac{\alpha}{2L}=\beta_t\eta_t.$$

If $\frac{\Lambda_{t+1}}{t+2}<\frac{\alpha}{2L}$ and $\frac{\Lambda_t}{t+1}\geq\frac{\alpha}{2L}$, then

$$(\beta_{t+1}-1)\eta_{t+1}=\frac{t^2+2t}{4}\frac{\Lambda_{t+1}}{t+2}\leq\frac{t^2+2t+1}{4}\frac{\alpha}{2L}=\beta_t\eta_t.$$

If $\frac{\Lambda_{t+1}}{t+2}<\frac{\alpha}{2L}$ and $\frac{\Lambda_t}{t+1}<\frac{\alpha}{2L}$, then

$$(\beta_{t+1}-1)\eta_{t+1}=\frac{t^2+2t}{4}\frac{\Lambda_{t+1}}{t+2}\leq\frac{t^2+2t+1}{4}\frac{\Lambda_t}{t+1}=\beta_t\eta_t.$$

Hence (45) holds. Combine (41) and (45) we have the following:

$$\begin{aligned}&(\beta_{t+1}-1)\eta_{t+1}\left[\Phi(\mathbf{x}_t^{\mathrm{ag}})-\Phi(\mathbf{x}^*)\right]\\&\leq\beta_t\eta_t\left[\Phi(\mathbf{x}_t^{\mathrm{ag}})-\Phi(\mathbf{x}^*)\right]\\&\leq(\beta_t-1)\eta_t\left[\Phi(\mathbf{x}_{t-1}^{\mathrm{ag}})-\Phi(\mathbf{x}^*)\right]\\&+\psi(\mathbf{x}_{t-1},\mathbf{x}^*)-\psi(\mathbf{x}_t,\mathbf{x}^*)-\eta_t\left\langle\delta_t,\mathbf{x}_{t-1}-\mathbf{x}^*\right\rangle+U_t.\end{aligned}$$

By induction we have:

$$\begin{aligned}&\Phi(\mathbf{x}_T^{\mathrm{ag}})-\Phi(\mathbf{x}^*)\leq\frac{1}{(\beta_{T+1}-1)\eta_{T+1}}\\&\left(D_{\psi,\mathcal{X}}^2-\sum_{t=1}^{T}\eta_t\langle\delta_t,\mathbf{x}_{t-1}-\mathbf{x}^*\rangle+\sum_{t=1}^{T}U_t\right)\end{aligned}\tag{46}$$

Using Lemma A.3, $\sum_{t=1}^{T}U_t$ can be bounded as

$$\begin{aligned}&\sum_{t=1}^{T}U_t=\sum_{t=1}^{T}\frac{\beta_t\eta_t^2\left(2M+\|\delta_t\|_*\right)^2}{2(\alpha\beta_t-L\eta_t)}\\&\overset{(44)}{\leq}\sum_{t=1}^{T}\frac{\eta_t^2\left(2M+\|\delta_t\|_*\right)^2}{\alpha}\leq\frac{D_{\psi,\mathcal{X}}^2}{4}\left[\sum_{t=1}^{s_1-1}\left(2M+\|\delta_t\|_*\right)^2\right.\\&\left.+\sum_{t=s_1}^{T}\frac{\left(2M+\|\delta_t\|_*\right)^2}{\sum_{i=s_1}^{t}\left(2M+\|\delta_i\|_*\right)^2}\right]\\&\leq\frac{D_{\psi,\mathcal{X}}^2}{4}\left[1+\log\sum_{t=s_1}^{T}\left(2M+\|\delta_t\|_*\right)^2\right]\end{aligned}\tag{47}$$

Conditioned on $\{\boldsymbol{\xi}_1,\dots\boldsymbol{\xi}_{t-1}\}$, $\mathbf{x}_t$ is deterministic and the expectation of $\langle\delta_t,\mathbf{x}_t-\mathbf{x}^*\rangle$ is 0.

According to (38), if $\frac{\alpha}{2L}\leq\frac{\Lambda_t}{t+1}$, one will take $\eta_t=\frac{\alpha(t+1)}{4L}$. Substituting this stepsize into (46) we have

$$\begin{aligned}&\mathbb{E}\left[\Phi(\mathbf{x}_T^{\mathrm{ag}})-\Phi(\mathbf{x}^*)\right]\leq\frac{D_{\psi,\mathcal{X}}^2+\mathbb{E}[\sum_{t=1}^{T}U_t]}{\left(\frac{T+2}{2}-1\right)\frac{\alpha(T+2)}{4L}}\\&\overset{(47)}{\leq}\frac{2LD_{\psi,\mathcal{X}}^2}{\alpha T(T+2)}\left[5+\mathbb{E}\log\sum_{t=s_1}^{T}\left(2M+\|\delta_t\|_*\right)^2\right]\\&\leq\frac{2LD_{\psi,\mathcal{X}}^2}{\alpha T(T+2)}\left(5+\log\sum_{t=1}^{T}\sigma_t^2\right),\end{aligned}\tag{48}$$

where we use Jensen's Inequality in the last step. If $\frac{\alpha}{2L}>\frac{\Lambda_t}{t+1}$, one will take $\eta_t=\frac{\Lambda_t}{2}$ and we have

$$\begin{aligned}&\mathbb{E}\left[\Phi(\mathbf{x}_T^{\mathrm{ag}})-\Phi(\mathbf{x}^*)\right]\leq\mathbb{E}\left[\frac{1}{\left(\frac{T+2}{2}\frac{\Lambda_{T+1}}{2}\right)}\left(D_{\psi,\mathcal{X}}^2+\sum_{t=1}^{T}U_t\right)\right]\\&\leq\frac{4\mathbb{E}\left[\sqrt{\frac{\sum_{t=1}^{T+1}(2M+\|\delta_t\|_*^2)}{T+1}}\left(D_{\psi,\mathcal{X}}^2+\sum_{t=1}^{T}U_t\right)\right]}{D_{\psi,\mathcal{X}}\sqrt{\alpha(T+2)}}\\&\overset{(47)}{\leq}\frac{D_{\psi,\mathcal{X}}}{\sqrt{\alpha(T+2)}}\mathbb{E}\left[\sqrt{\frac{\sum_{t=1}^{T+1}(2M+\|\delta_t\|_*)^2}{T+1}}\right.\\&\left.\left(5+\log\sum_{t=s_1}^{T}\left(2M+\|\delta_t\|_*\right)^2\right)\right]\\&\leq\frac{D_{\psi,\mathcal{X}}}{\sqrt{\alpha(T+2)}}\sqrt{\frac{\sum_{t=1}^{T+1}\sigma_t^2}{T+1}}\left(5+\log\sum_{t=1}^{T}\sigma_t^2\right),\end{aligned}\tag{49}$$

where again the last step is due to Jensen's Inequality. Combining (48) and (49) our claim follows. If $T<s_1$, the logarithmic term vanishes. □