Accelerating Rescaled Gradient Descent: Fast Optimization of Smooth Functions

Ashia C. Wilson  
Microsoft Research  
ashia.wilson@microsoft.com

Lester Mackey  
Microsoft Research  
lmackey@microsoft.com

Andre Wibisono  
Georgia Tech  
wibisono@gatech.edu

Abstract

We present a family of algorithms, called descent algorithms, for optimizing convex and non-convex functions. We also introduce a new first-order algorithm, called rescaled gradient descent (RGD), and show that RGD achieves a faster convergence rate than gradient descent provided the function is strongly smooth—a natural generalization of the standard smoothness assumption on the objective function. When the objective function is convex, we present two frameworks for “accelerating” descent methods, one in the style of Nesterov and the other in the style of Monteiro and Svaiter, using a single Lyapunov function. Rescaled gradient descent can be accelerated under the same strong smoothness assumption using both frameworks. We provide several examples of strongly smooth loss functions in machine learning and numerical experiments that verify our theoretical findings. We also present several extensions of our novel Lyapunov framework, including deriving optimal universal higher-order tensor methods and extending our framework to the coordinate setting.

1 Introduction

We consider the optimization problem

$$
\min_{x \in \mathcal{X}} f(x)
$$

where $f : \mathcal{X} \to \mathbb{R}$ is a continuously differentiable function, and $\mathcal{X}$ is a finite-dimensional real vector space with inner product norm $\|v\|^2 := \langle v, Bv \rangle$ and dual inner product $\|s\|^2_* := \langle s, B^{-1}s \rangle$ on the dual space $s \in \mathcal{X}^*$, where $B : \mathcal{X} \to \mathcal{X}^*$ is a positive definite self-adjoint operator. We assume the minimum is attainable and define $x^* \in \arg \min_{x \in \mathcal{X}} f(x)$.

We study the performance of a family of discrete-time algorithms parameterized by $\delta > 0$ and an integer scalar $1 < p \leq \infty$, called $\delta$-descent algorithms of order $p$. These algorithms meet a progress condition that allows us to derive fast non-asymptotic convergence rate upper bounds, parameterized by $p$, for both nonconvex and convex instances of (1). For example, descent algorithms of order $1 < p < \infty$ satisfy the upper bound $f(x_k) - f(x^*) = O(1/(\delta k)^{p-1})$ for convex functions.

Using this framework we introduce a new method for smooth optimization called rescaled gradient descent (RGD),

$$
x_{k+1} = x_k - \eta \frac{1}{\|\nabla f(x_k)\|_B^{-\frac{1}{2}}} B^{-1} \nabla f(x_k), \quad \eta > 0, p > 1.
$$

We show that if (1) is sufficiently smooth, rescaled gradient descent is a $\delta$-descent algorithm of order $p$, and subsequently converges quickly to solutions of (1). RGD can be viewed as a natural generalization of gradient descent ($p = 2$) and normalized gradient descent ($p = \infty$), whose non-asymptotic behavior for quasi-convex functions has been well-studied ([11]).

When $f$ is convex, we present two frameworks for obtaining algorithms with faster convergence rate upper bounds. The first, pioneered in Nesterov [21, 22, 23, 24], shows how to wrap a $\delta$-descent method of order $1 < p < \infty$ in two sequences to obtain a method that satisfies $f(x_k) - f(x^*) = O(1/(\delta k)^p)$. The second, introduced by [17], shows how to wrap a $\delta$-descent method of order $1 < p < \infty$ in the same set of sequences and add a line search step to obtain a method that satisfies $f(x_k) - f(x^*) = O(1/(\delta k)^{2p-2})$. We provide a general description of both frameworks and show how they can be applied to RGD, as well as other descent methods of order $p$.

Our motivation also comes from several works (e.g., [26, 27]), many of which are recent (e.g., [29, 32, 13, 34, 4, 8, 31, 28, 30, 16]), that have utilized the connection between dynamical systems and optimization algorithms to develop new analyses and optimization methods. Zhang et al. [35], for example, introduce accelerated dynamics and apply Runge-Kutta integrators to them. They show that Runge-Kutta integrators converge quickly when the function is sufficiently smooth and when the order of the integrator is sufficiently large. We show accelerated RGD achieves a better convergence rate under a very similar smoothness assumption. Another example is Maddison et al. [16], who introduce conformal Hamiltonian dynamics, and show that if the objective function is sufficiently smooth, algorithms obtained by discretizing these dynamics converge at a linear rate. We show (accelerated) RGD also achieves a fast linear rate under similar smoothness conditions.

The remainder of this paper is organized as follows. Section 2 introduces $\delta$-descent algorithms and Section 2.1 describes several examples of descent algorithms that are popular in optimization. Section 2.2 introduces RGD and Section 3 presents two frameworks for accelerating $\delta$-descent methods and applies both to RGD. Section 5 describes several examples of strongly smooth objective functions as well as experiments to verify our findings. Finally, Section 6 discusses simple extensions of our framework, including deriving and analyzing optimal universal tensor methods for objective functions that have Hölder-continuous higher-order gradients and extending our entire framework and results to the coordinate setting.

## 2 Descent Algorithms

The focus of this section is a family of algorithms called $\delta$-descent algorithms of order $p$.

**Definition 1** An algorithm $x_{k+1} = A(x_k)$ is a $\delta$-descent algorithm of order $p$ for $1 < p \leq \infty$ if for some constant $0 < \delta < \infty$ it satisfies

\[
\frac{f(x_{k+1}) - f(x_k)}{\delta} \leq -\|\nabla f(x_k)\|_{\frac{p}{p-1}}^{\frac{p}{p-1}} \text{ for all } k \geq 0, \text{ or} \tag{2a}
\]

\[
\frac{f(x_{k+1}) - f(x_k)}{\delta} \leq -\|\nabla f(x_{k+1})\|_{\frac{p}{p-1}}^{\frac{p}{p-1}} \text{ for all } k \geq 0. \tag{2b}
\]

For $\delta$-descent algorithms of order $p$, it is possible to obtain non-asymptotic convergence guarantees for non-convex, convex and gradient dominated functions. Recall, a function is $\mu$-gradient dominated of order $p \in (1, \infty)$ if

\[
\frac{p-1}{p} \|\nabla f(x)\|_{\frac{p}{p-1}} \geq \mu^{\frac{1}{p-1}}(f(x) - f(x^*)), \quad \forall x \in \mathcal{X}. \tag{3}
\]

For the following three theorems, we use the shorthand $E_0 := f(x_0) - f(x^*)$ and assume $f$ is differentiable.

**Theorem 1** Any $\delta$-descent algorithm of order $p$ satisfies

\[
\min_{0 \leq s \leq k} \|\nabla f(x_s)\|_s \leq (E_0/(\delta k))^{\frac{p-1}{p}}. \tag{4}
\]

**Theorem 2** If $f$ is convex with $R = \sup_{x : f(x) \leq f(x_0)} \|x - x^*\| < \infty$, and $c_p := (\frac{1}{p-1})^{\frac{1}{p}}$, then any $\delta$-descent algorithm of order $p$ satisfies

\[
f(x_k) - f(x^*) \leq \left(2 \left(\frac{1}{E_0^{1/p}} + \frac{1}{R \gamma c_p^{1/p}} (\delta k)^{\frac{p-1}{p}}\right)^{-p}\right) = O(1/(1 + \frac{1}{R \gamma c_p^{1/p}} (\delta k)^{\frac{p-1}{p}})^p), \quad p < \infty \\
(2E_0 \exp(-\delta k/(R \gamma)), \quad p = \infty. \tag{5}
\]

where $\gamma = 1$ when (2a) is satisfied and $\gamma = (1 + \frac{1}{R \gamma c_p^{1/p}})^{\frac{p-1}{p}}$ when (2b) is satisfied.
Theorem 3 If \( f \) is \( \mu \)-gradient dominated of order \( p \), then any \( \delta \)-descent algorithm of order \( p \) satisfies
\[
f(x_k) - f(x^*) \leq E_0 \exp \left( -\frac{p}{p-1} \mu \frac{1}{\mu} \delta k \right).
\] (6)

The proof of Theorems 1 to 3 are all based on simple energy arguments and can be found in Appendix B. Bounds of the form (4) are common in the non-convex optimization literature and have previously been established for gradient descent (\( p = 2 \) see e.g. [25, Thm1]) and higher-order tensor methods (see e.g. [6]). Theorem 1 provides a more general description of algorithms that satisfy this kind of bound.

Typically, algorithms satisfy the progress condition (2) for specific smoothness classes of functions. For example, gradient descent with step-size \( 0 < \eta \leq 1/L \) is a \( \delta \)-descent method of order \( p = 2 \) with \( \delta = \eta/2 \) when \( \|\nabla^2 f\| \leq L \). We list several other examples.

2.1 Examples of descent algorithms

Theorems 1, 2 and 3 provide a seamless way to derive standard upper bounds for many algorithms in optimization.

Example 1 The universal higher-order tensor method,
\[
x_{k+1} = \arg\min_{x \in X} \left\{ f_{p-1}(x; x_k) + \frac{1}{p!} \|x - x_k\|^p \right\},
\] (7)
where \( f_{p-1}(y; x) = \sum_{i=0}^{p-1} \frac{1}{i!} \nabla^i f(x)(y - x)^i \) is the \((p - 1)\)-st order Taylor approximation of \( f \) centered at \( x \), has been studied by several works [3, 33, 20]. When \( f \) is convex and has Hölder-smooth \((p - 1)\)-st order gradients, namely \( \|\nabla^{p-1} f(x) - \nabla^{p-1} f(y)\| \leq L\|x - y\|^\nu \) for \( \nu = (0, 1) \), (7) with step size \( 0 < \eta \leq \frac{\sqrt{\nu(p-2)!}}{2L} \), is a \( \delta \)-descent method of order \( p - 1 + \nu \) with \( \delta = \eta^{\frac{1}{p-\nu}} \frac{1}{2L^{\nu-1}} \).

Example 2 The proximal method,
\[
x_{k+1} = \arg\min_{x \in X} \left\{ f(x) + \frac{1}{p!} \|x - x_k\|^p \right\},
\] (8)
was introduced by [18]. For any \( \eta > 0 \), the proximal method is a \( \delta \)-descent algorithm of order \( p \) with \( \delta = \eta^{\frac{1}{p-\nu}} / p \).

Example 3 Natural gradient descent,
\[
x_{k+1} = x_k - \eta \nabla^2 h(x_k)^{-1} \nabla f(x_k) = \arg\min_{x \in X} \left\{ \langle \nabla f(x_k), x \rangle + \frac{1}{2\eta} \|x - x_k\|^2 \right\},
\] (9)
where \( B = I \) and \( \|v\|_x = \langle v, \nabla^2 h(x)v \rangle \) was introduced by [2]. Suppose \( \|\nabla^2 f\| \leq L \) and \( mI \leq \nabla^2 h \leq MI \) for some \( m, L, M > 0 \). Then natural gradient descent with step size \( 0 < \eta \leq \frac{m^2}{ML} \) is a \( \delta \)-descent algorithm of order \( p = 2 \) with \( \delta = \frac{\eta}{2M} \).

Example 4 Mirror descent,
\[
x_{k+1} = \arg\min_{x \in X} \left\{ \langle \nabla f(x_k), x \rangle + \frac{1}{\eta} D_h(x, x_k) \right\},
\] (10)
where \( B = I \) was introduced by [19]. Here \( D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle \) is the Bregman divergence. Suppose \( \|\nabla^2 f\| \leq L \) and \( mI \leq \nabla^2 h \leq MI \) for some \( m, L, M > 0 \). Then mirror descent with step size \( 0 < \eta \leq \frac{m^2}{ML} \) is a \( \delta \)-descent algorithm of order \( p = 2 \) with \( \delta = \frac{\eta}{2M} \).

Example 5 The proximal Bregman method,
\[
x_{k+1} = \arg\min_{x \in X} \left\{ f(x) + \frac{1}{\eta} D_h(x, x_k) \right\},
\] (11)
where \( B = I \) was introduced by [7]). When \( mI \leq \nabla^2 h \leq MI \) the proximal Bregman method with step-size \( \eta > 0 \) is a \( \delta \)-descent algorithm of order \( p = 2 \) with \( \delta = \frac{mn}{2M\eta} \).

Details for these examples are contained in Appendix B.2.
2.2 Rescaled gradient descent

We end this section by discussing the function class for which rescaled gradient descent (RGD),

\[ x_{k+1} = x_k - \eta \frac{1}{p} B^{-1} \nabla f(x_k) = \arg \min_{x \in \mathcal{X}} \left\{ \langle \nabla f(x_k), x \rangle + \frac{1}{2} \frac{1}{\eta^p} ||x - x_k||^p \right\}, \tag{12} \]

is a \(\delta\)-descent method of order \(p\).

**Definition 2** A function \(f\) is strongly smooth of order \(p\) for some integer \(p > 1\), if there exist constants \(0 < L_1, \ldots, L_p < \infty\) such that for \(m = 1, \ldots, p-1\) and for all \(x \in \mathbb{R}^d\):

\[ |\nabla^m f(x)(B^{-1} \nabla f(x))^m| \leq L_m ||\nabla f(x)||^m_* + \frac{L-m}{p} \] \tag{13}

and moreover for \(m = p\), \(f\) satisfies the condition \(|\nabla^p f(x)(v)^p| \leq L_p ||v||^p\), \(\forall v \in \mathcal{X}\).

Here, \(\nabla^m f(x)(h)^m = \sum_{i_1, \ldots, i_m=1}^d \partial x_{i_1} \ldots x_{i_m} f(x) \prod_{j=1}^d h_{ij}\) where \(\partial x_i f\) is the partial derivative of \(f\) with respect to \(x_i\). We can always take \(L_1 = 1\). When \(p = 2\), (13) is the usual Lipschitz condition on the gradient of \(f\), but otherwise (13) is stronger. In particular, if \(f\) is strongly smooth of order \(p\), then the minimizer \(x^*\) has order at least \(p-1\), i.e., the higher gradients vanish: \(\nabla^m f(x^*) = 0\) for \(m = 1, \ldots, p-1\), whereas this is not implied under mere smoothness. An example of a strongly smooth function of order \(p\) is the \(p\)-th power of the norm \(f(x) = \|x\|_p^p\) with \(B = I\). We discuss other families of strongly smooth functions in Section 5. Finally, it is worth mentioning that for most of our results, the absolute value on the left hand side of (13) is unnecessary. We now present the main result regarding the performance of RGD on functions that satisfy (13):

**Theorem 4** Suppose \(f\) is strongly smooth of order \(p > 1\) with constants \(0 < L_1, \ldots, L_p < \infty\). Then rescaled gradient descent with step-size

\[ 0 < \eta \leq \min \left\{ 1, \frac{1}{2} \frac{1}{\sum_{m=2}^p \frac{1}{\eta^m}} \right\} \tag{14} \]

satisfies the descent condition (2a) with \(\delta = \eta^{1-p} / 2\).

The proof of Theorem 4 is in Appendix B.3. A corollary to Theorems 1-4 is the following theorem.

**Theorem 5** RGD with a step size that satisfies (14) achieves convergence rate guarantee, (4) when \(f\) is differentiable and strongly smooth of order \(p\), (5) when \(f\) is convex function and strongly smooth of order \(p\), and (6) when \(f\) is \(\mu\)-uniformly convex and strongly smooth of order \(p\), where \(\delta^{p-1} = \eta / 2^{p-1}\).

We now demonstrate how all the aforementioned examples of \(\delta\)-descent methods can be accelerated.

3 Accelerating Descent Algorithms

We present two frameworks for accelerating descent algorithms and apply them to RGD. The backbone of both frameworks is the Lyapunov function

\[ E_k = A_k(f(x_k) - f(x^*)) + D_k(x^*, z_k), \]

and two sequences (15) and (16) whose origins can be traced to the dynamical systems perspective introduced by Wibisono et al. [33], Wilson et al. [34]. We present a high-level description of both techniques in the main text and leave details of our analysis to Appendix C.

3.1 Nesterov acceleration of descent algorithms

In the context of convex optimization, the technique of “acceleration” has its origins in Nesterov [21] and refined in Nesterov [22]. In these works, Nesterov showed how to combine gradient descent with two sequences to obtain an algorithm with an optimal convergence rate. There have been many works since (as well as some frameworks, including [15, 1, 14, 34]) describing how to accelerate various other algorithms to obtain methods with superior convergence rates.
Wilson et al. [34], for example, show the following two schemes,

\[ x_k = \delta \tau_k z_k + (1 - \delta) \gamma_k \]  
\[ z_{k+1} = \arg \min_{z \in \mathcal{X}} \left\{ \alpha_k \langle \nabla f(x_k), z \rangle + \frac{\eta}{2} D_h(z, z_k) \right\} \]

where \( y_{k+1} \) satisfies the \( \delta \frac{p-1}{p} \)-descent condition \( f(y_{k+1}) - f(x_k) \leq -\delta \frac{p}{p-1} \| \nabla f(x_k) \|^{\frac{p}{p-1}} \); and

\[ x_k = \delta \tau_k z_k + (1 - \delta) \gamma_k \]
\[ z_{k+1} = \arg \min_{z \in \mathcal{X}} \left\{ \alpha_k \langle \nabla f(y_{k+1}), z \rangle + \frac{\eta}{2} D_h(z, z_k) \right\} , \]

where the update for \( y_{k+1} \) satisfies the \( (\delta \frac{p-1}{p}-\text{descent}) \) condition \( f(y_{k+1}) - f(x_k) \leq \langle \nabla f(y_{k+1}), y_{k+1} - x_k \rangle \leq -\delta \frac{p}{p-1} \| \nabla f(y_{k+1}) \|^{\frac{p}{p-1}} \), constitute an “accelerated method”. Their results can be summarized in the following theorem.

**Theorem 6** Assume for all \( x, y \in \mathcal{X} \), the function \( h \) satisfies the local uniform convexity condition \( D_h(x, y) \geq \frac{1}{p} \| x - y \|^{p} \). Then sequences (15) and (16) with parameter choices \( \alpha_k = (\delta/p)^{p-1} k^{(p-1)} \) (where \( k^{(p)} := k(k+1) \cdots (k+p-1) \) is the rising factorial) and \( \tau_k = \frac{\eta}{\delta(p+k)} = \Theta(\frac{1}{\delta k}) \) satisfy,

\[ f(y_k) - f(x^*) \leq \frac{p}{\delta k} D_h(x^*, z_0) = O \left( \frac{1}{\delta k^p} \right) . \]

Proof details are contained in Appendix C.1. Wilson et al. [34] call these new methods accelerated descent methods, given that Theorem 2 guarantees implementing just the \( y_{k+1} \) sequence (where we set \( x_k = y_k \)) satisfies \( f(y_k) - f(x^*) \leq O \left( \frac{1}{\delta k} \right) \), where \( \delta k^{-1} = \delta p \). The computational cost of adding sequences (15a) and (15b) (or (16a) and (16b)) to the descent method is at most an additional gradient evaluation.

**Remark 1 (Restarting)** If, in addition, \( f \) is \( \mu \)-gradient dominated of order \( p \), then algorithms (15) and (16) combined with a scheme for restarting the algorithm has a convergence rate upper bound \( f(y_k) - f(x^*) = O(\exp(-\mu \tau \delta k)) \). We consider this algorithm an accelerated method given the original descent method satisfies \( f(y_k) - f(x^*) = O(\exp(-\mu \tau \delta k)) \) under the same condition, where \( \delta k^{-1} = \delta p \). See Appendix C.2 for details.

To summarize, it is sufficient to establish conditions under which an algorithm is a \( \delta \)-descent algorithm of order \( p \) in order to (1) obtain a convergence rate and (2) accelerate the algorithm (in most cases).

**Accelerated rescaled gradient descent (Nesterov-style)** Using (15) we accelerate RGD.

**Algorithm 1** Nesterov-style accelerated rescaled gradient descent.

**Require**: \( f \) satisfies (13) and \( h \) satisfies \( D_h(x, y) \geq \frac{1}{p} \| x - y \|^{p} \)

1. Set \( x_0 = z_0, A_k = (\delta/p)^{p-1} k^{(p-1)}, \alpha_k = \frac{A_{k+1} - A_k}{\delta}, \tau_k = \frac{\alpha}{A_{k+1}}, \text{ and } \delta \tau^{p-1} = \eta^{\frac{p-1}{p}} / 2. \)
2. for \( k = 1, \ldots, K \) do
3. \( x_k = \delta \tau_k z_k + (1 - \delta) \gamma_k \)
4. \( z_{k+1} = \arg \min_{z \in \mathcal{X}} \left\{ \alpha_k \langle \nabla f(x_k), z \rangle + \frac{\eta}{2} D_h(z, z_k) \right\} \)
5. \( y_{k+1} = x_k - \frac{\eta^{\frac{p-1}{p}}}{B - 1} H^{-1} \nabla f(x_k) / \| \nabla f(x_k) \|^{\frac{p-1}{p}} \)
6. return \( y_k \).

We summarize the performance of Algorithm 1 in the following Corollary to Theorems 4 and 6:

**Theorem 7** Suppose \( f \) is convex and strongly smooth of order \( 1 < p < \infty \) with constants \( 0 < L_1, \ldots, L_p < \infty. \) Also suppose \( \eta \) satisfies (14). Then Algorithm 1 satisfies the convergence rate upper bound (17).

3.2 Monteiro-Svaiter acceleration of descent algorithms

Recently, Monteiro and Svaiter [17] have introduced an alternative framework for accelerating descent methods, which is similar to Nesterov’s scheme but includes a line search step. This framework
was further generalized by several more recent concurrent works [9, 12, 5] who demonstrate that higher-order tensor method (7) with the addition of a line search step obtains a convergence rate upper bound \( f(y_k) - f(x^*) = O(1/k^{3p-2}) \). When \( p = 2 \), this rate matches that of the Nesterov-style acceleration framework, but for \( p > 2 \) it is better. In this section, we present a novel, generalized version of the Monteiro-Svaiter acceleration framework. In particular, we use a simple Lyapunov analysis to generalize the framework and show that many other descent methods of order \( p \) can be accelerated in it, including the proximal method (8), RGD (12) and universal tensor methods.

**Proposition 8** Suppose \( h \) is 1-strongly convex and \( B = I \). Consider sequence (15) where in addition, we add a line search step which ensures the inequalities

\[
\begin{align*}
    a & \leq \frac{\lambda_{k+1}}{\delta} \| y_k - x_k \|^{p-2} \leq b, \quad 0 < a < b \quad \text{and} \\
    \| y_k - x_k + \lambda_{k+1} \nabla f(y_{k+1}) \| \leq \frac{1}{2} \| y_k - x_k \| 
\end{align*}
\]

hold for the pair \((\lambda_{k+1}, y_{k+1})\), where \( \lambda_{k+1} = \delta^2 \alpha_k^2 / A_{k+1} \). Then the composite sequence satisfies:

\[
    f(y_k) - f(x^*) \leq \frac{\delta^{p-2} D_h(x^*, x_0)^{\frac{p}{2}}}{(\delta k)^{\frac{p-2}{2}}} = O \left( \frac{1}{(\delta k)^{\frac{3p-2}{2}}} \right).
\]

The proof of Theorem 8 is in Appendix C.3. All the aforementioned concurrent works have demonstrated that the higher-order gradient method \( (\nu = 1) \) with the addition line search step satisfies (18). We show the same is true of the proximal method (8), rescaled gradient descent (12) and universal higher-order tensor methods. See Appendix C.5 for details. We conjecture that all methods that satisfy conditions (18a) and (18b) are descent methods of order \( p \) with an additional line search step.

**Remark 2 (Restarting)** If, in addition, \( f \) is \( \mu \)-gradient dominated of order \( p \), then (18) combined with a scheme for restarting the algorithm satisfies the convergence rate upper bound \( f(y_k) - f(x^*) = O(\exp(-\mu \frac{1}{k^{p-1}} \delta k)) \). See Appendix C.2 for details.

### 3.3 Accelerating rescaled gradient descent (Monteiro-Svaiter-style)

Monteiro-Svaiter accelerated rescaled gradient descent is the following algorithm.

**Algorithm 2** Monteiro-Svaiter-style accelerated rescaled gradient descent.

**Require:** \( f \) is strongly smooth of order 1 \(< p < \infty \) and \( h \) is 1-strongly convex, \( B = I \).

1. Set \( x_0 = z_0 = 0 \), \( A_0 = 0 \), \( \delta \cdot \frac{3p-2}{2} = \eta \), \( \frac{\mu}{\nu} \) \( \leq \min \{ \frac{2}{5p}, 1/(2 \sum_{m=2}^{\infty} \frac{L_m}{m!}) \} \)
2. for \( k = 1, \ldots, K \) do
3. Choose \( \lambda_{k+1} \) (e.g. by line search) such that \( \frac{3}{4} \leq \frac{\lambda_{k+1} \| y_{k+1} - x_k \|}{\eta} \leq \frac{5}{4} \), where

\[
    y_{k+1} = x_k - \eta \frac{1}{\| \nabla f(x_k) \|^{\frac{p-2}{2}}} \nabla f(x_k),
\]

and \( \alpha_k = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4 \lambda_k \lambda_{k+1}}}{2 \delta} \), \( A_{k+1} = \delta \alpha_k + A_k \), \( \tau_k = \frac{\alpha_k}{A_{k+1}} \) (so that \( \lambda_{k+1} = \delta \alpha_k / A_{k+1} \)) and

\[
    x_k = \delta \tau_k z_k + (1 - \delta \tau_k) y_k.
\]
4. Update \( z_{k+1} = \arg \min_{z \in X} \left\{ \alpha_k \langle \nabla f(y_{k+1}), z \rangle + \frac{1}{\delta} D_h(z, z_k) \right\} \)
5. return \( y_k \)

We summarize results on performance of Algorithm 2 in the following corollary to Theorem 8:

**Theorem 9** Assume \( f \) is convex and strongly smooth of order 1 \(< p < \infty \) with constants \( 0 < L_1, \ldots, L_p < \infty \). Then Algorithm 2 satisfies the convergence rate upper bound (19).

### 4 Related Work

Our acceleration framework is similar in spirit to many other acceleration frameworks that been proposed in the literature (e.g. Allen-Zhu and Orecchia [1], Lessard, Recht and Packard [14], Lin, Mairal...
and Harchaoui [15], Jelena and Orrechia [8]) but applies more generally to descent methods of order \( p > 2 \). In particular, it builds off the framework proposed Wilson, Recht and Jordan, but it (1) makes the connection to descent methods more explicit, and (2) incorporates a generalization and Lyapunov analysis of the Monteiro-Svaiter acceleration framework. These manifold generalizations crucially allow us to propose RGD and ARGD, which has superior theoretical and empirical performance to several existing methods on strongly smooth functions.

5 Examples and Numerical Experiments

We compare our result to several recent works that have shown that for some function classes, more intuitive first-order algorithms outperform gradient descent. In particular, both Zhang et al. [35] and Maddison et al. [16] obtain first-order algorithms by applying integration techniques to second-order ODEs. When the objective function is sufficiently smooth, both show their algorithm outperforms (accelerated) gradient descent. We show that Algorithms 1 and 2 achieves fast performance in theory and in practice on similar objectives.

Runge-Kutta Zhang et al. [35] show that if one applies an \( s \)-th order Runge-Kutta integrator to a family of second-order dynamics, then the resulting algorithm\(^1\) achieves a convergence rate \( f(x_k) - f(x^*) = O(1/k^{2s})^2 \) provided the function meet the following two conditions: (1) \( f \) satisfies the \( \text{gradient lower bound} \) of order \( p \geq 2 \), which means for all \( m = 1, \ldots, p - 1 \),

\[
f(x) - f(x^*) \geq \frac{1}{m} \|\nabla^m f(x)\|^{2s-p} \quad \forall x \in \mathbb{R}^n \tag{21}
\]

for some constants \( 0 < C_1, \ldots, C_{p-1} < \infty \); and (2) for \( s \geq p \) and \( M > 0 \), \( f \) is \( (s+2) \)-times differentiable and \( \|\nabla^{(i)} f(x)\| \leq M \) for \( i = p, p+1, \ldots, s+2 \). One can show that if \( f \) is strongly smooth of order \( p \), then \( f \) satisfies the gradient lower bound of order \( p \). The details of this result is in D. While are unable to prove that condition (21) is equivalent to strong smoothness, we have yet to find an example of a function that satisfies (21) and is not strongly smooth.

Hamiltonian Descent Maddison et al. [16] show explicit integration techniques applied to conformal Hamiltonian dynamics converge at a fast linear rate for a function class larger than gradient descent. The method entails finding a kinetic energy map that upper bounds the dual of the function. All examples for which we can compute such a map given by [16] are uniformly convex and gradient dominated functions; therefore, simply rescaling the gradient for these examples ensures a linear rate.

5.1 Examples

We provide several examples of strongly smooth functions in machine learning (see Appendix D.2 for details).

Example 6 The \( \ell_p \) loss function

\[
f(x) = \frac{1}{p} \|Ax - b\|_p^p, \tag{22}
\]

shown by Zhang et al. [35] to satisfy (21) of order \( p \), is strongly smooth of order \( p \).

Example 7 The logistic loss

\[
f(x) = \log(1 + e^{-yw^\top x}), \tag{23}
\]

shown by Zhang et al. [35] to satisfy (21) of order \( p = \infty \), is strongly smooth of order \( p = \infty \).

Example 8 The GLM loss,

\[
f(x) = \frac{1}{2} (y - \phi(x^\top w))^2 \quad \text{for} \quad \phi(r) = 1/(1 + e^{-r}), \quad y \in \{0, 1\}, \quad \text{and} \quad w \in \mathbb{R}^d, \tag{24}
\]

studied by Hazan et al. [11] is strongly smooth of order \( p = 3 \).

\(^1\)which requires at least \( s \) gradient evaluations per iteration
\(^2\)this matches the rate of Algorithm 1 in the limit \( s \to \infty \), where \( s \) is the order of the integrator.
Example 9  The $\ell_2$ loss to the $p$-th power

$$f(x) = \frac{1}{p} \|Ax - b\|_2^p,$$  \hspace{1cm} (25)

for which Hamiltonian descent [16] obtains a linear rate, is strongly smooth and gradient dominated of order $p$.

Example 10  The loss function,

$$f(x) = (x^{(1)} + x^{(2)})^4 + \frac{1}{16} (x^{(1)} - x^{(2)})^4,$$  \hspace{1cm} (26)

for which Hamiltonian descent [16] obtains a linear rate, is strongly smooth and gradient dominated of order $p = 4$.

5.2 Experiments

In this section, we perform a series of numerical experiments to compare the performance of (Nesterov-style accelerated) RGD with gradient descent (GD), Nesterov accelerated GD (NAG), and the state-of-the-art Runge-Kutta algorithms of Zhang et al. [35] (DD) on the logistic loss $f(x) = \sum_{i=1}^{10} \log(1 + e^{-w_i^T x y_i})$ and the $\ell_4$ loss. We use the same code, plots, and experimental methodology of Zhang et al. [35] (including data and step-size choice), adding to it (A)RGD. Specifically, for Fig. 1a-1d, the entries of $W \in \mathbb{R}^{10 \times 10}$ (and similarly $A$) are i.i.d. standard Gaussian, and the first five entries of $y$ (and $b$) are valued 0 and the rest are 1. For each experiment, a simple implementation of (A)RGD significantly outperforms the Runge-Kutta algorithm (DD), GD and NAG. The code for these experiments can be found here: https://github.com/aswilson07/ARGD.git.

Figure 1: Experimental results comparing (Nesterov-style A)RGD to gradient descent (GD), Nesterov accelerated GD (NAG) and Runge-Kutta (DD). The plots for Runge-Kutta use an $s = 2$ integrator which requires two gradient evaluations per iteration. We plot both iterations (Figs. 1a and 1b) and gradient evaluations (Figs. 1c and 1d). Fig. 1e shows the performance of A(RGD) on the Hamiltonian objective studied by [16]; for Fig. 1e, the largest step-size was chosen subject to the algorithm not diverging.
6 Additional Results and Discussion

This paper establishes broad conditions under which an algorithm will converge and its performance can be accelerated by adding momentum. We use these conditions to introduce (accelerated) rescaled gradient descent for strongly smooth functions, and showed it outperforms several recent first-order methods that have been introduced for optimizing smooth functions in machine learning.

There are (at least) two simple extensions of our framework. First, an analogous framework can be established for (accelerated) $\delta$-coordinate descent methods of order $p$. As an application, we introduce (accelerated) rescaled coordinate descent for functions that are strongly smooth along each coordinate direction of the gradient. We provide details in Appendix E.1. Second, with our generalization of the Monteiro-Svaiter framework, we derive optimal universal tensor methods for functions whose $(p - 1)$-st gradients are $\nu$-Hölder-smooth which achieve the upper bound $f(y_k) - f(x^*) = O(1/k^{\tilde{p}/2})$ where $\tilde{p} = p - 1 + \nu$. The matching lower bound for this class of functions was recently established by [10]. We present this result in Appendix E.3.

There are several possible directions for future work. We know that certain simple operations preserve convexity (e.g., addition), but what operations preserve strong smoothness? Understanding this could allow us to construct more complex examples of strongly smooth functions. Our results reveal an interesting hierarchy of smoothness assumptions which lead to methods that converge quickly – exploring this more is of significant interest. Finally, extending our analysis to the stochastic or manifold setting, studying the use of variance reduction techniques, and introducing other $\delta$-decent algorithms of order $p$ are all interesting directions for future work.

Acknowledgments

We would like to thank Jingzhao Zhang for providing us access to his code.

References


