SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

Understanding Convergence of Accelerated Gradient Descent Methods

av

Sebastian Fodor

2020 - No K22

Understanding Convergence of Accelerated Gradient Descent Methods

Sebastian Fodor

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Yishao Zhou

2020

Understanding Convergence of Accelerated Gradient Descent Methods

Sebastian Fodor

Abstract

In this paper we study a generalized version of Nesterovs Accelerated Gradient Method (NAG) that has three parameters instead of two, with one additional parameter for the amount of gradient correction. This gives the Generalized Nesterovs Accelerated Gradient Method (GNAG), of which both NAG and Polyak's heavy-ball method are special cases. We derive a differential equation that approximates this method and show that it converges with linear rate for functions of strong convexity and Lipschitz continuous gradients. We also consider GNAG as a dynamical system, and show that this dynamical system converges with linear rate to a steady state which is the optimal solution. We ask the question whether GNAG converges faster than NAG for certain choices of the gradient correction parameter, and by numerical examples arrive at the conclusion that a higher gradient correction parameter can lead to faster convergence.

1 Introduction

Convex optimization algorithms are an important part of numerical methods as they often offer fast and precise calculations. Since the rise of machine learning algorithms, there is a new interest in first order methods. Machine learning is often about solving large problems, where only the gradient of the objective function can be calculated within reasonable time, and the Hessian is unknown [7]. For this reason there has recently been a new-found interest in first order methods. The simplest first order method is the classical gradient descent method in which we always move in the direction opposite to the gradient of the objective function, that is we have the update scheme

$$x_{k+1} = x_k - \alpha \nabla f(x_k),$$

where α is called the step size. It can be shown that under the assumption that f is convex and α is sufficiently small, gradient descent converges to the global

minimum with rate $O(1/k)$. The first improvement to gradient descent is Polyak's heavy-ball method, with the intuition that instead of only moving in the steepest direction, we also have some momentum [5]. This gives us the update scheme

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha\nabla f(x_k),$$

where α is the step size and β is called the momentum coefficient. Unfortunately, this method does not come with any guaranteed global convergence, and even fails to converge on strongly convex objective functions with certain parameter choices, as shown in [2]. The next improvement comes from Nesterov and is often called Nesterovs Accelerated Gradient Method (NAG) [4]. It has the two step upgrade scheme

$$\begin{aligned} x_{k+1} &= y_k - \alpha\nabla f(y_k) \\ y_k &= x_k + \beta(x_k - x_{k-1}). \end{aligned}$$

It can be shown that NAG achieves quadratic convergence for convex functions, and converges with linear rate for strictly convex functions [2]. For further insight NAG can be condensed into one line as

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha\nabla f(x_k + \beta(x_k - x_{k-1})).$$

Written in this form, we can see that the only difference between the heavy-ball method and NAG is the so called gradient correction term $\beta(x_k - x_{k-1})$ inside the gradient term. One might ask whether it is crucial for this term to contain the same parameter β as the momentum term. This question leads to the Generalized Nesterovs Accelerated Gradient Method (GNAG), which has an additional parameter γ. Its update scheme is

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha\nabla f(x_k + \gamma(x_k - x_{k-1})). \tag{1}$$

Note that this method can be seen as a mix between the heavy-ball method and NAG, since $\gamma = 0$ and $\gamma = \beta$ corresponds to those two methods, respectively. For ease of notation we introduce the gradient correction coefficient $\Gamma = \gamma/\beta$. This generalized version of NAG has been proposed in both [7] and [6] but not yet examined properly in the literature.

In this paper we will study the convergence properties of GNAG, and also ask the question whether the choice $\gamma = \beta$ (as in NAG) really is the most efficient parameter choice, or whether we can get faster convergence through other choices of γ.

2 Preliminaries

Before we begin studying GNAG we discuss some useful theory that will be used in the later sections.

2.1 Studying Convergence

When studying convergence of gradient based methods some assumptions are often necessary. For example it is unreasonable to expect convergence to global optima when other local minima are present, as the method only has information about the local gradients. Therefore the supposition that the objective function f is convex is often made. Under this assumption it can be proved that simple gradient descent achieves $O(1/k)$ convergence and NAG achieves $O(1/k^2)$ convergence to the global optimum. Another assumption often made is that the objective function is strongly convex and has Lipschitz continuous gradients, as described in the next section. Under this assumption both gradient descent and NAG achieves linear convergence rate. It is proved in [2] that for m-strongly convex objective functions with L-Lipschitz continuous gradients NAG achieves fastest rate of convergence using the parameter choices $\alpha = 1/L$ and $\beta = (1-\sqrt{m\alpha})/(1+\sqrt{m\alpha})$. For this reason when studying convergence, these parameter choices are often made. It is important to note that in practice we might not know the parameters L and m, it is an interesting question how to set and update these parameters in that case, but this is a much more difficult topic and will not be dealt with in this paper.

2.2 Strong Convexity and Lipschitz Gradients

Throughout this paper we will mostly be interested in functions that are strongly convex and have Lipschitz continuous gradients. These two properties imply several useful inequalities about a function and its gradient. We first define strong convexity.

Definition 1. *A function $f(x)$ is strongly convex with parameter m if the function*

$$f(x) - \frac{m}{2}\|x\|^2$$

is convex.

Strong convexity implies multiple useful properties which we will use throughout.

Lemma 1. *The following statements are all equivalent*

(a) $f(x)$ *is strongly convex with parameter* m.

(b) $\nabla^2 f(x) - mI$ *is positive semidefinite for all* x.

(c) $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|x-y\|^2$ *for all* x *and* y.

(d) $(\nabla f(x) - \nabla f(y))^T (x-y) \geq m\|x-y\|^2$ *for all* x *and* y.

Proof Statements (b), (c) and (d) are the well known (see [1]) second order condition of convexity, the first order condition of convexity, and the monotone gradient condition of convexity on the function $f(x) - \frac{m}{2}\|x\|^2$. □

Having Lipschitz continuous gradients is defined similarly to strong convexity.

Definition 2. *A differentiable function* $f(x)$ *has Lipschitz continuous gradients with parameter* L *if the function*

$$\frac{L}{2}\|x\|^2 - f(x)$$

is convex.

Lipschitz gradients also imply multiple useful properties.

Lemma 2. *If* f *is convex the following statements are all equivalent*

(a) $f(x)$ *has Lipschitz continuous gradients with parameter* L,

(b) $LI - \nabla^2 f(x)$ *is positive semidefinite for all* x.

(c) $f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|x-y\|^2$ *for all* x *and* y.

(d) $(\nabla f(x) - \nabla f(y))^T (x-y) \leq L\|x-y\|^2$ *for all* x *and* y.

(e) $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|$ *for all* x *and* y.

(f) $(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$ *for all* x *and* y.

(g) $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$ *for all* x *and* y.

Proof Once again statements (b), (c) and (d) are the second order condition of convexity, the first order condition of convexity, and the monotone gradient condition of convexity on the function $\frac{L}{2}\|x\|^2 - f(x)$. To get from (a) from to (e) note that if f is convex and has Lipschitz continuous gradients with parameter L then the same can be said about the function $g(y) = f(y) - \nabla f(x)^T y$. So statement (c) on g is

$$f(z) - \nabla f(x)^T z \leq f(y) - \nabla f(x)^T y + (\nabla f(y) - \nabla f(x))^T (z-y) + \frac{L}{2}\|z-y\|^2.$$

Minimizing by z on both sides gives statement (e). To get from (e) to (f) formulate (e) for the pair (x, y) and for the pair (y, x), adding these two inequalities gives (f). We go from (f) to (g) to (d) using the Cauchy-Schwartz inequality. Thus we have $(a) \Leftrightarrow (b) \Leftrightarrow (c) \Leftrightarrow (d)$ and $(a) \Rightarrow (e) \Rightarrow (f) \Rightarrow (g) \Rightarrow (d)$ and so all the statements are equivalent. □

We denote by $S^2_{m,L}(\mathbb{R}^n)$ the set of twice differentiable function $\mathbb{R}^n \to \mathbb{R}$ that are both strongly convex with parameter m and have Lipschitz continuous gradients with parameter L.

2.3 Lyapunov Stability

We will study the convergence of GNAG using an Ordinary Differential Equation (ODE). One very useful tool when working with these is Lyapunov's second method of stability [3]. The idea of the method is that given an ODE $\dot{X} = f(X)$ we define a scalar function $V(X)$ that will act as an energy potential function. By showing that this function V decreases with time we can show that the differential equation converges, that is as $t \to \infty$, $X(t) \to x^\star$ where $x^\star$ is an equilibrium point. More formally we make the following definition.

Definition 3. *Let the ODE $\dot{X} = f(X)$ have equilibrium point $x^\star$. The function $V : \mathbb{R}^n \to \mathbb{R}$ is a Lyapunov function of the ODE if*

- $V(x) \geq 0$ *for all* x
- $V(x) = 0 \Leftrightarrow x = x^\star$
- *All sublevel sets $V_\alpha = \{x : V(x) \leq \alpha\}$ are bounded*
- $\frac{dV(X(t))}{dt} \leq 0$ *for all t and all trajectories of X.*
- $\frac{dV(X(t))}{dt} = 0 \Leftrightarrow X(t) = x^\star$

Lemma 3. *Let $\dot{X} = f(X)$ have equilibrium point $x^\star$ and let $V(x)$ be a smooth Lyapunov function of the ODE. Then for any starting point $X(0) = x_0$ we have $X(t) \to x^\star$ as $t \to \infty$.*

Proof Suppose $X(t)$ does not converge to $x^\star$. Since $V(X(t))$ is decreasing and is non-negative it converges to some $\epsilon < V(X(0))$. Let C be the closed and bounded, hence compact set

$$C = \{x : \epsilon \leq V(x) \leq V(X(0))\}.$$

Since C is compact and $V(x)$ is smooth we have

$$\max_{X(t)\in C} \frac{dV(X(t))}{dt} = a < 0.$$

Then for all T

$$V(X(T)) = V(X(0)) + \int_0^T \frac{dV(X(t))}{dt} dt \leq V(X(0)) + \int_0^T a dt = V(X(0)) + aT.$$

However as $a < 0$ this gives $V(X(T)) < 0$ for sufficiently large T, a contradiction. Hence our supposition is false and we have $X(t) \to x^\star$ as $t \to \infty$. □

In the case where we also want to prove that $X(t)$ converges with linear rate, we can construct a Lyapunov function which satisfies $dV(X(t))/dt < -cV(X(t))$. This method will be used later in the paper.

3 Convergence using ODEs

In order to study the convergence properties of GNAG we estimate the trajectory of the x_k with a continuous function and study the convergence of this estimate. Recall that the update is given by

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha\nabla f(x_k + \gamma(x_k - x_{k-1})). \tag{2}$$

Throughout this section we will be using the standard tuning of

$$\beta = \frac{1 - \sqrt{m\alpha}}{1 + \sqrt{m\alpha}}.$$

Let $Y(t)$ be a continuous smooth function such that it follows the trajectory of the points given by GNAG. Hence we set $t_k = k\sqrt{\alpha}$ and let $Y(t_k) = x_k$. Note now that $x_{k+1} = Y(t_{k+1}) = Y((k+1)\sqrt{\alpha}) = Y(t_k + \sqrt{\alpha})$. Taylor expansions give

$$x_{k+1} = Y(t_k) + \dot{Y}(t_k)\sqrt{\alpha} + \frac{1}{2}\ddot{Y}(t_k)\alpha + \frac{1}{6}\dddot{Y}(t_k)\alpha^{\frac{3}{2}} + O(\alpha^2), \tag{3}$$

$$x_{k-1} = Y(t_k) - \dot{Y}(t_k)\sqrt{\alpha} + \frac{1}{2}\ddot{Y}(t_k)\alpha - \frac{1}{6}\dddot{Y}(t_k)\alpha^{\frac{3}{2}} + O(\alpha^2), \tag{4}$$

$$\tag{5}$$

We will use these to find a differential equation describing $Y(t)$.

Divide both sides of (2) by $\alpha\beta$ and rearrange to get

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{\alpha} + \left(\frac{1}{\beta} - 1\right)\frac{x_{k+1} - x_k}{\alpha} + \frac{1}{\beta}\nabla f(x_k + \gamma(x_k - x_{k-1})) = 0.$$

Substitute (3) and (4) to arrive at the ODE

$$\ddot{Y} + O(\alpha) + \left(\frac{1}{\beta} - 1\right)\left(\frac{\dot{Y}}{\sqrt{\alpha}} + \frac{\ddot{Y}}{2} + O(\sqrt{\alpha})\right) + \frac{1}{\beta}\nabla f(Y + \gamma\sqrt{\alpha}\dot{Y} + O(\alpha)) = 0.$$

Note that due to our use of the standard tuning of β the term $(1/\beta - 1)$ is $O(\sqrt{\alpha})$. Together with a Taylor expansion of ∇f this gives

$$\ddot{Y} + \left(\frac{1}{\beta} - 1\right)\left(\frac{\dot{Y}}{\sqrt{\alpha}} + \frac{\ddot{Y}}{2}\right) + \frac{1}{\beta}(\nabla f(Y) + \nabla^2 f(Y)\dot{Y}\gamma\sqrt{\alpha}) + O(\alpha) = 0.$$

Rearrange and substitute the standard tuning for β to end up with the ODE

$$\ddot{Y} + 2\sqrt{m}\dot{Y} + \Gamma\sqrt{\alpha}\nabla^2 f(Y)\dot{Y} + (1 + \sqrt{m\alpha})\nabla f(Y) + O(\alpha) = 0.$$

While $Y(t)$ follows the trajectory of x_k it is difficult to prove convergence due to the $O(\alpha)$ term. By removing this term we instead get an approximation of the trajectory of x_k.

Definition 4. *The high resolution ODE of GNAG is given by*

$$\ddot{X} + 2\sqrt{m}\dot{X} + \Gamma\sqrt{\alpha}\nabla^2 f(X)\dot{X} + (1 + \sqrt{m\alpha})\nabla f(X) = 0. \tag{6}$$

With $X(0) = x_0$ and $\dot{X}(0) = \sqrt{\alpha}\nabla f(x_0)$. Recall $\Gamma = \gamma/\beta$.

3.1 Convergence of Smooth Approximation

We begin this section by proving a Lemma that will be useful later.

Lemma 4. *Let a smooth function $Z(t)$ satisfy $\dot{Z} \leq -cZ$. Then $Z(t) \leq Z(0)e^{-ct}$.*

Proof Since the exponential function is positive we have from the assumption that $0 \geq \left(\dot{Z} + cZ\right)e^{ct} = \frac{d}{dt}\left(Ze^{ct}\right)$. And hence $Z(0) = Z(0)e^{c0} \geq Z(t)e^{ct}$, and the Lemma follows. □

For proving the convergence of the high resolution ODE we will use a Lyapunov function. There is no standard method to find Lyapunov functions for most ODEs, and there are usually several different Lyapunov functions that might work. We will be using a modified version of the Lyapunov function used in [6].

Lemma 5. *The function*

$$\begin{aligned}\mathcal{E}(t) =&(1+\sqrt{m\alpha})(f(X)-f(x^*))+\frac{1}{4}\|\dot{X}\|^2\\&+\frac{1}{4}\|\dot{X}+2\sqrt{m}(X-x^*)+\Gamma\sqrt{\alpha}\nabla f(X)\|^2.\end{aligned}$$

is a Lyapunov function of the ODE (6)*.*

Proof Every term in the function is non-negative and only 0 when $X = x^*$. We will prove that the time derivative of $\mathcal{E}$ is negative in the proof of Theorem 6. □

The initial value of the Lyapunov function is

$$\begin{aligned}\mathcal{E}(0) =&(1+\sqrt{m\alpha})(f(x_0)-f(x^*))+\frac{1}{4}\|\sqrt{\alpha}\nabla f(x_0)\|^2\\&+\frac{1}{4}\|\sqrt{\alpha}\nabla f(x_0)+2\sqrt{m}(x_0-x^*)+\Gamma\sqrt{\alpha}\nabla f(x_0)\|^2.\end{aligned}$$

The Lipschitz gradient of f, together with a step size of $\alpha \leq 1/L$ gives

$$f(x_0)-f(x^*)\leq\frac{L}{2}\|x_0-x^*\|^2\leq\frac{1}{2\alpha}\|x_0-x^*\|^2, \tag{7}$$

$$\|\nabla f(x_0)\|^2\leq L^2\|x_0-x^*\|^2\leq\frac{1}{\alpha^2}\|x_0-x^*\|^2. \tag{8}$$

And hence the initial value of the Lyapunov function can be bounded as

$$\begin{aligned}\mathcal{E}(0) =&(1+\sqrt{m\alpha})(f(x_0)-f(x^*))+\frac{1}{4}\|\sqrt{\alpha}\nabla f(x_0)\|^2\\&+\frac{1}{4}\|\sqrt{\alpha}\nabla f(x_0)+2\sqrt{m}(x_0-x^*)+\Gamma\sqrt{\alpha}\nabla f(x_0)\|^2\\\leq&(1+\sqrt{m\alpha})(f(x_0)-f(x^*))+\frac{\alpha(2\Gamma^2+4\Gamma+3)}{4}\|\nabla f(x_0)\|^2+2m\|x_0-x^*\|^2\\\leq&\left(\frac{1+\sqrt{m\alpha}}{2}+\frac{2\Gamma^2+4\Gamma+3}{4}+2m\alpha\right)\frac{\|x_0-x^*\|^2}{\alpha}.\end{aligned}$$

And since $m\alpha \leq L\alpha \leq 1$ we get

$$\mathcal{E}(0)\leq\frac{2\Gamma^2+4\Gamma+15}{4}\frac{\|x_0-x^*\|^2}{\alpha}. \tag{9}$$

We are now ready to prove the convergence of (6).

Theorem 6. *For any $\alpha \geq 0$ and step size $0 < \alpha < 1/L$ the solution to the ODE* (6) *satisfies*

$$f(X(t)) - f(x^*) \leq \frac{C(\Gamma)\|x_0 - x^*\|^2}{\alpha} e^{-\frac{\sqrt{m}}{4(\Gamma+1)}t}.$$

Where $C(\Gamma) = (2\Gamma^2 + 4\Gamma + 15)/4$. Recall that $\Gamma = \gamma/\beta$.

Proof. Consider the Lyapunov function (7). The time derivative of this function is

$$\begin{aligned}\frac{d\mathcal{E}}{dt} =&(1+\sqrt{m\alpha})\nabla f(X)^T\dot{X} + \frac{1}{2}\dot{X}^T\ddot{X} \\ &+ \frac{1}{2}\left(\dot{X} + 2\sqrt{m}(X-x^*) + \Gamma\sqrt{\alpha}\nabla f(X)\right)^T\left(\ddot{X} + 2\sqrt{m}\dot{X} + \Gamma\sqrt{\alpha}\nabla^2 f(X)\dot{X}\right).\end{aligned}$$

Utilizing (6) we can rewrite the Lyapunov function as

$$\begin{aligned}\frac{d\mathcal{E}}{dt} =&(1+\sqrt{m\alpha})\nabla f(X)^T\dot{X} + \frac{1}{2}\dot{X}^T\left(-2\sqrt{m}\dot{X} - \Gamma\sqrt{\alpha}\nabla^2 f(X)\dot{X} - (1+\sqrt{m\alpha})\nabla f(X)\right) \\ &+ \frac{1}{2}\left(\dot{X} + 2\sqrt{m}(X-x^*) + \Gamma\sqrt{\alpha}\nabla f(X)\right)^T\left(-(1+\sqrt{m\alpha})\nabla f(X)\right) \\ =&-\sqrt{m}\left(\|\dot{X}\|^2 + (1+\sqrt{m\alpha})\nabla f(X)^T(X-x^*) + \Gamma\frac{\alpha}{2}\|\nabla f(X)\|^2\right) \\ &- \Gamma\frac{\sqrt{\alpha}}{2}\left(\|\nabla f(X)\|^2 + \dot{X}^T\nabla^2 f(X)\dot{X}\right) \\ \leq&-\sqrt{m}\left(\|\dot{X}\|^2 + (1+\sqrt{m\alpha})\nabla f(X)^T(X-x^*) + \Gamma\frac{\alpha}{2}\|\nabla f(X)\|^2\right).\end{aligned}$$

Due to the strong convexity of f we have the two inequalities

$$\begin{aligned}\nabla f(X)^T(X-x^*) &\geq f(X) - f(x^*) + \frac{m}{2}\|X-x^*\|^2, \\ \nabla f(X)^T(X-x^*) &\geq m\|X-x^*\|^2.\end{aligned}$$

Which together give the upper bound on the term

$$\begin{aligned}(1+\sqrt{m\alpha})\nabla f(X)^T(X-x^*) \geq&\frac{1+\sqrt{m\alpha}}{2}\nabla f(X)^T(X-x^*) + \frac{1}{2}\nabla f(X)^T(X-x^*) \\ \geq&\frac{1+\sqrt{m\alpha}}{2}(f(X) - f(x^*)) + \frac{3m}{4}\|X-x^*\|^2.\end{aligned}$$

So the derivative of the Lyapunov function can be further bounded as

$$\frac{d\mathcal{E}}{dt} \leq -\sqrt{m}\left(\frac{1+\sqrt{m\alpha}}{2}(f(X) - f(x^*)) + \|\dot{X}\|^2 + \frac{3m}{4}\|X-x^*\|^2 + \Gamma\frac{\alpha}{2}\|\nabla f(X)\|^2\right).$$

Rearranging and using the triangle inequality then yields

$$\begin{aligned}
-\frac{4(\Gamma+1)}{\sqrt{m}}\frac{d\mathcal{E}}{dt} \geq & 4(\Gamma+1)\left(\frac{1+\sqrt{m\alpha}}{2}(f(X)-f(x^*))+\|\dot{X}\|^2\right. \\
& \left.+\frac{3m}{4}\|X-x^*\|^2+\Gamma\frac{\alpha}{2}\|\nabla f(X)\|^2\right) \\
\geq & 2(1+\sqrt{m\alpha})(f(X)-f(x^*))+\|\dot{X}\|^2 \\
& +2m\|X-x^*\|^2+\Gamma^2\frac{\alpha}{2}\|\nabla f(X)\|^2 \\
\geq & (1+\sqrt{m\alpha})(f(X)-f(x^*))+\frac{1}{4}\|\dot{X}\|^2 \\
& +\frac{1}{4}\|\dot{X}+2\sqrt{m}(X-x^*)+\Gamma\sqrt{\alpha}\nabla f(X)\|^2=\mathcal{E}.
\end{aligned}$$

And so we have derived the inequality

$$-\frac{4(\Gamma+1)}{\sqrt{m}}\frac{d\mathcal{E}}{dt}\geq\mathcal{E}\Rightarrow\frac{d\mathcal{E}}{dt}\leq-\frac{\sqrt{m}}{4(\Gamma+1)}\mathcal{E}.$$

Which due to Lemma 4 proves the linear convergence of $\mathcal{E}$:

$$\mathcal{E}(t)\leq\mathcal{E}(0)e^{-\frac{\sqrt{m}}{4(\Gamma+1)}t}. \tag{10}$$

And due to the lower bound (9) on $\mathcal{E}(0)$ we further have

$$\mathcal{E}(t)\leq\frac{C(\Gamma)\|x_0-x^*\|^2}{\alpha}e^{-\frac{\sqrt{m}}{4(\Gamma+1)}t}.$$

Together with the fact that $f(X(t))-f(x^*)\leq\mathcal{E}(t)$ we get convergence of $X(t)$:

$$f(X(t))-f(x^*)\leq\frac{C(\Gamma)\|x_0-x^*\|^2}{\alpha}e^{-\frac{\sqrt{m}}{4(\Gamma+1)}t}.$$

□

We have thus shown that the smooth approximation of the trajectory of GNAG converges to the local minimum with linear rate, which is a new result. As this approximation is accurate for small step sizes α, we have reason to believe that for small α GNAG also converges with linear rate. Looking at Theorem 6 the exponent is $-\frac{\sqrt{m}}{4(\Gamma+1)}t$, this suggests that smaller choices of Γ yield faster convergence. However the rate in Theorem 6 is most likely not the best possible rate, throughout the proof we used multiple soft inequalities that can be made tighter. Furthermore the rate of convergence in Theorem 6 for $\Gamma=1$ is worse than the rate of convergence of NAG proven in [6]. However as the aim of this section is purely to prove linear convergence of the approximation we do not investigate this rate further. In the next section we will however prove some rates of convergence of GNAG for different choices of Γ.

4 Convergence using Dynamical Systems

One can formulate certain numerical methods as linear dynamical systems. We will state GNAG as such a system and also define an auxiliary system alongside it. By proving certain constraints on this auxiliary system we can then conclude convergence of GNAG. This procedure is similar to the one in [2].

4.1 Dynamical Systems

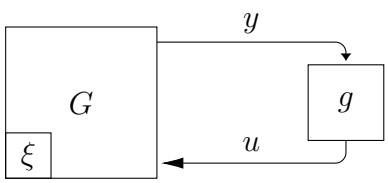
A linear dynamical system G with a nonlinear feedback g is a recursion of the form

$$\xi_{k+1} = A_G \xi_k + B_G u_k \tag{11a}$$
$$y_k = C_G \xi_k + D_G u_k \tag{11b}$$
$$u_k = g(y_k). \tag{11c}$$

Here ξ_k is called the state of the system at time k. The u_i constitute the inputs and the y_i constitute the outputs. The system can be expressed by the block matrix

$$\left[\begin{array}{c|c} A_G & B_G \\ \hline C_G & D_G \end{array}\right]$$

It can also be represented with a diagram.



4.2 GNAG as a Dynamical System

Recall that NAG can be written as the recursion

$$\begin{aligned} x_{k+1} &= y_k - \alpha \nabla f(y_k) \\ y_k &= (1+\beta)x_k - \beta x_{k-1}. \end{aligned}$$

To express it as a linear dynamical system with a nonlinear feedback we write it as

$$\begin{aligned} \xi_{k+1}^{(1)} &= (1+\beta)\xi_k^{(1)} - \beta \xi_k^{(2)} - \alpha u_k \\ \xi_{k+1}^{(2)} &= \xi_k^{(1)} \\ y_k &= (1+\beta)\xi_k^{(1)} - \beta \xi_k^{(2)} \\ u_k &= \nabla f(y_k). \end{aligned}$$

Here $\xi_k^{(1)}$ plays the role of x_k, and $\xi_k^{(2)}$ plays the role of x_{k-1}. Note that these two variables together make up the state of the system. We now see that NAG is represented by the matrix

$$\left[\begin{array}{c|c} A_G & B_G \\ \hline C_G & D_G \end{array}\right] = \left[\begin{array}{cc|c} (1+\beta)I & -\beta I & -\alpha I \\ I & 0 & 0 \\ \hline (1+\beta)I & -\beta I & 0 \end{array}\right]$$

We find a similar representation of GNAG with recursion

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f(x_k + \gamma(x_k - x_{k-1})).$$

Using the same idea of letting the state include both x_k and x_{k-1} we get the dynamical system

$$\xi_{k+1}^{(1)} = (1+\beta)\xi_k^{(1)} - \beta \xi_k^{(2)} - \alpha u_k \tag{12a}$$

$$\xi_{k+1}^{(2)} = \xi_k^{(1)} \tag{12b}$$

$$y_k = (1+\gamma)\xi_k^{(1)} - \gamma \xi_k^{(2)} \tag{12c}$$

$$u_k = \nabla f(y_k). \tag{12d}$$

With matrix representation

$$\left[\begin{array}{c|c} A_G & B_G \\ \hline C_G & D_G \end{array}\right] = \left[\begin{array}{cc|c} (1+\beta)I & -\beta I & -\alpha I \\ I & 0 & 0 \\ \hline (1+\gamma)I & -\gamma I & 0 \end{array}\right]$$

4.3 Auxiliary System

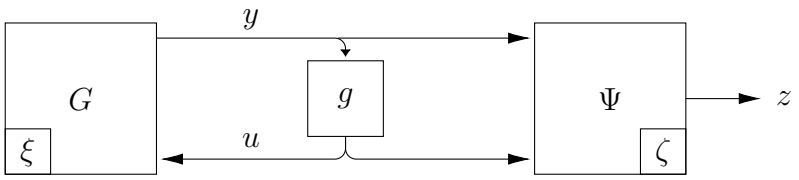
In order to prove that the system G converges we wish to use some properties of the nonlinearity g. By extending the system (11) with an auxiliary system Ψ we can formulate some constraints of g. Consider the following system:

$$\xi_{k+1} = A_G\xi_k + B_G u_k \tag{13a}$$
$$y_k = C_G\xi_k + D_G u_k \tag{13b}$$
$$u_k = g(y_k) \tag{13c}$$
$$\zeta_{k+1} = A_\Psi\zeta_k + B_\Psi y_k + C_\Psi u_k \tag{13d}$$
$$z_k = D_\Psi\zeta_k + E_\Psi y_k + F_\Psi u_k. \tag{13e}$$

Here ζ_k is the state of the auxiliary system Ψ at time k. The auxiliary system is expressed by the block matrix

$$\left[\begin{array}{c|c|c} A_\Psi & B_\Psi & C_\Psi \\ \hline D_\Psi & E_\Psi & F_\Psi \end{array}\right]$$

The linear dynamical system G together with the auxiliary system Ψ is the depicted by the diagram



Given a reference point $(u_\star, y_\star) = (g(y_\star), y_\star)$, and for a choice of A_Ψ such that 1 is not an eigenvalue of A_Ψ. there is a unique fixed point $(\zeta_\star, z_\star)$ of (13) that satisfy

$$\zeta_\star = A_\Psi\zeta_\star + B_\Psi y_\star + C_\Psi u_\star$$
$$z_\star = D_\Psi\zeta_\star + E_\Psi y_\star + F_\Psi u_\star.$$

By rearranging the above equations we find this fixed point to be

$$\zeta_\star = (I - A_\Psi)^{-1}(B_\Psi y_\star + C_\Psi u_\star)$$
$$z_\star = D_\Psi\zeta_\star + E_\Psi y_\star + F_\Psi u_\star.$$

4.4 Integral Quadratic Constraint (IQC)

We can now state some constraints on this auxiliary system, which mostly depend on the non-linearity g.

Definition 5 (ρ-Hard IQC)**.** *Suppose G is a linear dynamical system with the nonlinear feedback g and auxiliary system as given by* (13)*. Let $(u_\star, y_\star)$ be a given reference point and let $(\zeta_\star, z_\star)$ be a fixed point of the system. The nonlinearity satisfies the ρ-Hard IQC defined by $(\Psi, M, y_\star, u_\star)$ if for all sequences of y_i and for all $K \leq 0$,*

$$\sum_{k=0}^{K} \rho^{-2k} (z_k - z_\star)^T M (z_k - z_\star) \geq 0. \tag{14}$$

In our case we want to prove convergence of GNAG for strictly convex functions with Lipschitz continuous Gradients. Hence we are interested in IQCs for the nonlinearity ∇f. We will be using the weighted off-by-one IQC.

Lemma 7. *Suppose $f \in S^2_{m,L}(\mathbb{R}^n)$ with minimum at $y_\star$. Let*

$$\left[\begin{array}{c|c|c} A_\Psi & B_\Psi & C_\Psi \\ \hline D_\Psi & E_\Psi & F_\Psi \end{array}\right] = \left[\begin{array}{c|c|c} 0 & -LI & I \\ \hline \bar{\rho}^2 I & LI & -I \\ 0 & -mI & I \end{array}\right]$$

and

$$M = \left[\begin{array}{c|c} 0 & I \\ \hline I & 0 \end{array}\right]$$

Then for all $0 \leq \bar{\rho} \leq \rho \leq 1$ the nonlinearity ∇f satisfies the ρ-hard IQC defined by $(\Psi, M, y_\star, 0)$.

Proof. We prove $\bar{\rho}$-hardness, as this will imply ρ-hardness. The auxiliary system for this Ψ is given by

$$\begin{aligned} \zeta_{k+1} &= -Ly_k + u_k \\ z_k &= \begin{pmatrix} \bar{\rho}^2 I \\ 0 \end{pmatrix} \zeta_k + \begin{pmatrix} LI \\ -mI \end{pmatrix} y_k + \begin{pmatrix} -I \\ I \end{pmatrix} u_k. \end{aligned}$$

And so for $k \geq 1$.

$$z_k = \begin{pmatrix} L(y_k - \bar{\rho}^2 y_{k-1}) - (u_k - \bar{\rho}^2 u_{k-1}) \\ u_k - m y_k \end{pmatrix}$$

Also since $u_\star = 0$ we have

$$z_\star = \begin{pmatrix} L(y_\star - \bar{\rho}^2 y_\star) \\ -m y_\star \end{pmatrix}$$

And so for $k \geq 1$

$$z_k - z_\star = \begin{pmatrix} L((y_k - y_\star) - \bar{\rho}^2(y_{k-1} - y_\star)) - (u_k - \bar{\rho}^2 u_{k-1}) \\ u_k - m(y_k - y_\star) \end{pmatrix}$$

And for $k = 0$ we have

$$z_0 - z_\star = \begin{pmatrix} L(y_0 - y_\star) - u_0 \\ u_0 - m(y_0 - y_\star) \end{pmatrix}$$

And so the inequality (14) at hand is

$$\begin{aligned} &(u_0 - m(y_0 - y_\star))^T (L(y_0 - y_\star) - u_0) \\ &+ \sum_{k=1}^{K} \bar{\rho}^{-2k} (u_k - m(y_k - y_\star))^T (L((y_k - y_\star) - \bar{\rho}^2(y_{k-1} - y_\star)) - (u_k - \bar{\rho}^2 u_{k-1})) \geq 0. \end{aligned} \tag{15}$$

Define

$$\begin{aligned} s_k &= (u_k - m(y_k - y_\star))^T (L(y_k - y_\star) - u_k) \\ p_k &= (u_k - m(y_k - y_\star))^T (L(y_k - y_{k-1}) - (u_k - u_{k-1})). \end{aligned}$$

We can hence write the left hand side of (15) as

$$s_0 + \sum_{k=1}^{K} \bar{\rho}^{-2k} ((1 - \bar{\rho}^2) s_k + \bar{\rho}^2 p_k). \tag{16}$$

Define

$$h(x) = f(x) - f(y_\star) - \frac{m}{2} \|x - y_\star\|^2 \tag{17}$$

$$q_k = (L - m) h(y_k) - \frac{1}{2} \|\nabla h(y_k)\|^2. \tag{18}$$

It is clear that $h(x) \in S^2_{0,L-m}(\mathbb{R}^n)$ and that $h(x)$ has minimum at $y_\star$ where it equals 0. Due to this Lipschitz continuous gradients of $h(x)$ we see that $q_k \geq 0$. Also note that $\nabla h(y_k) = u_k - m(y_k - y_\star)$, and so

$$\begin{aligned} s_k &= \nabla h(y_k)^T ((L - m)(y_k - y_\star) - \nabla h(y_k)) \\ &= (L - m) \nabla h(y_k)^T (y_k - y_\star) - \|\nabla h(y_k)\|^2 \\ &\geq (L - m) h(y_k) - \frac{1}{2} \|\nabla h(y_k)\|^2 \\ &= q_k. \end{aligned}$$

Similarly

$$\begin{aligned} p_k &= (L-m)\nabla h(y_k)^T(y_k - y_{k-1}) - \nabla h(y_k)^T(\nabla h(y_k) - \nabla h(y_{k-1})) \\ &\geq (L-m)(h(y_k) - h(y_{k-1})) - \frac{1}{2}\|\nabla h(y_k)\|^2 + \frac{1}{2}\|\nabla h(y_{k-1})\|^2 \\ &= q_k - q_{k-1}. \end{aligned}$$

We can bound (16) as

$$\begin{aligned} & s_0 + \sum_{k=1}^{K} \bar{\rho}^{-2k}((1-\bar{\rho}^2)s_k + \bar{\rho}^2 p_k) \\ & \geq q_0 + \sum_{k=1}^{K} \bar{\rho}^{-2k}((1-\bar{\rho}^2)q_k + \bar{\rho}^2(q_k - q_{k-1})) \\ & = q_0 + \sum_{k=1}^{K} \bar{\rho}^{-2k} q_k - \bar{\rho}^{-2k+2} q_{k-1} = \bar{\rho}^{-2K} q_K \geq 0. \end{aligned}$$

And hence we have proven inequality (15). □

4.5 Proving Convergence Through IQCs

We will now see how we can use IQCs to prove the convergence of a dynamical system. Suppose we have a linear dynamical system G with $D_G = 0$. In this case (13) can be written as

$$\xi_{k+1} = A_G \xi_k + B_G u_k \tag{19a}$$
$$y_k = C_G \xi_k \tag{19b}$$
$$u_k = g(y_k) \tag{19c}$$
$$\zeta_{k+1} = A_\Psi \zeta_k + B_\Psi y_k + C_\Psi u_k \tag{19d}$$
$$z_k = D_\Psi \zeta_k + E_\Psi y_k + F_\Psi u_k. \tag{19e}$$

Substituting y with $C_G\xi$ in all equations gives

$$\begin{pmatrix} \xi_{k+1} \\ \zeta_{k+1} \\ z_k \end{pmatrix} = \begin{pmatrix} A_G & 0 & B_G \\ B_\Psi C_G & A_\Psi & C_\Psi \\ E_\Psi C_G & D_\Psi & F_\Psi \end{pmatrix} \begin{pmatrix} \xi_k \\ \zeta_k \\ u_k \end{pmatrix} \tag{20a}$$
$$u_k = g(C_G \xi_k). \tag{20b}$$

By partitioning up the matrix of (20a) into

$$A = \begin{pmatrix} A_G & 0 & B_G \\ B_\Psi C_G & A_\Psi & C_\Psi \end{pmatrix} \tag{21a}$$

$$B = \begin{pmatrix} E_\Psi C_G & D_\Psi & F_\Psi \end{pmatrix} \tag{21b}$$

We can formulate a theorem regarding the convergence of the linear dynamical system using a Linear Matrix Inequality (LMI).

Theorem 8. *Suppose that G and Ψ is given by* (19) *with a fixed point $(\xi_\star, \zeta_\star, y_\star, u_\star, z_\star)$ and that (A, B) is given by* (21)*. If g satisfies the ρ-hard IQC defined by $(\Psi, M, y_\star, u_\star)$ and there exists $P \succ 0$ such that*

$$A^T P A - \rho^2 \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} + B^T M B \preceq 0 \tag{22}$$

we have

$$\|\xi_K - \xi_\star\| \leq \rho^K \sqrt{\lambda_{max}/\lambda_{min}} \|\xi_0 - \xi_\star\|$$

for all K, where λ_{max} and λ_{min} are the largest and smallest eigenvalues of P.

Proof. Multiply (22) with $\begin{pmatrix} \xi_k - \xi_\star \\ \zeta_k - \zeta_\star \\ u_k - u_\star \end{pmatrix}$ from the right and by its transpose from the left. This gives

$$\begin{pmatrix} \xi_{k+1} - \xi_\star \\ \zeta_{k+1} - \zeta_\star \end{pmatrix}^T P \begin{pmatrix} \xi_{k+1} - \xi_\star \\ \zeta_{k+1} - \zeta_\star \end{pmatrix} - \rho^2 \begin{pmatrix} \xi_k - \xi_\star \\ \zeta_k - \zeta_\star \end{pmatrix}^T P \begin{pmatrix} \xi_k - \xi_\star \\ \zeta_k - \zeta_\star \end{pmatrix} + z_k^T M z_k \leq 0. \tag{23}$$

Multiply with ρ^{-2k} and sum over k, which gives

$$\rho^{-2K+2} \begin{pmatrix} \xi_K - \xi_\star \\ \zeta_K - \zeta_\star \end{pmatrix}^T P \begin{pmatrix} \xi_K - \xi_\star \\ \zeta_K - \zeta_\star \end{pmatrix} - \rho^2 \begin{pmatrix} \xi_0 - \xi_\star \\ \zeta_0 - \zeta_\star \end{pmatrix}^T P \begin{pmatrix} \xi_0 - \xi_\star \\ \zeta_0 - \zeta_\star \end{pmatrix} + \sum_{k=0}^{K-1} \rho^{-2k} (z_k - z_\star) M (z_k - z_\star) \leq 0.$$

Since the IQC is satisfied by assumption and since $\zeta_0 = \zeta_\star$ we have

$$\begin{pmatrix} \xi_K - \xi_\star \\ \zeta_K - \zeta_\star \end{pmatrix}^T P \begin{pmatrix} \xi_K - \xi_\star \\ \zeta_K - \zeta_\star \end{pmatrix} \leq \rho^{2K} \begin{pmatrix} \xi_0 - \xi_\star \\ 0 \end{pmatrix}^T P \begin{pmatrix} \xi_0 - \xi_\star \\ 0 \end{pmatrix}$$

$$\left\| \begin{pmatrix} \xi_K - \xi_\star \\ \zeta_K - \zeta_\star \end{pmatrix} \right\| \leq \rho^K \sqrt{\lambda_{max}/\lambda_{min}} \left\| \begin{pmatrix} \xi_0 - \xi_\star \\ 0 \end{pmatrix} \right\|$$

$$\|\xi_K - \xi_\star\| \leq \rho^K \sqrt{\lambda_{max}/\lambda_{min}} \|\xi_0 - \xi_\star\|.$$

□

4.6 GNAG with IQCs

Now let us consider the case where the dynamical system is (12) with matrix

$$\left[\begin{array}{c|c} A_G & B_G \\ \hline C_G & D_G \end{array}\right] = \left[\begin{array}{cc|c} (1+\beta)I & -\beta I & -\alpha I \\ I & 0 & 0 \\ \hline (1+\gamma)I & -\gamma I & 0 \end{array}\right]$$

and nonlinearity ∇f where $f \in S^2_{m,L}(\mathbb{R}^n)$. We then know from Lemma 7 that the nonlinearity satisfies the ρ-hard IQC defined by

$$\left[\begin{array}{c|c|c} A_\Psi & B_\Psi & C_\Psi \\ \hline D_\Psi & E_\Psi & F_\Psi \end{array}\right] = \left[\begin{array}{c|c|c} 0 & -LI & I \\ \hline \bar{\rho}^2 I & LI & -I \\ 0 & -mI & I \end{array}\right]$$

and

$$M = \left[\begin{array}{c|c} 0 & I \\ \hline I & 0 \end{array}\right]$$

The matrices A and B from Theorem 8 then are

$$A = \begin{pmatrix} A_G & 0 & B_G \\ B_\Psi C_G & A_\Psi & C_\Psi \end{pmatrix} = \begin{pmatrix} (1+\beta)I & -\beta I & 0 & -\alpha I \\ I & 0 & 0 & 0 \\ -L(1+\gamma)I & L\gamma I & 0 & I \end{pmatrix}$$
$$B = \begin{pmatrix} E_\Psi C_G & D_\Psi & F_\Psi \end{pmatrix} = \begin{pmatrix} L(1+\gamma)I & -L\gamma I & \bar{\rho}^2 I & -I \\ -m(1+\gamma)I & m\gamma I & 0 & I \end{pmatrix}$$

And so by Theorem 8 GNAG converges with some rate ρ if there is a matrix $P \succ 0$ and a constant $0 \leq \bar{\rho} \leq \rho \leq 1$ such that

$$A^T P A - \rho^2 \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} + B^T \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} B \preceq 0. \tag{24}$$

We note a few things about (24). First, a few calculations reveal that it is linear in both P and $\bar{\rho}$, meaning that it is an LMI feasibility problem that for given m, L, α, β and γ can be solved by well known methods. Second, note that if (24) is feasible for some $(m, L, \alpha, \beta, \gamma)$ it is also feasible for $(cm, cL, \alpha, \beta, \gamma)$, hence the feasibility depends on m and L only through the ratio L/m. Third, due to the block-wise diagonal structure of A and B, it can be shown that if the LMI holds for some $\bar{\rho}$ and P it also holds for some $\bar{\rho}$ and P with $P = \tilde{P} \otimes I$

where $\tilde{P} \in S^3_{++}$. Hence, no matter the dimension of the original system, we can determine convergence by studying the LMI (24) with

$$A = \begin{pmatrix} (1+\beta) & -\beta & 0 & -\alpha \\ 1 & 0 & 0 & 0 \\ -L(1+\gamma) & L\gamma & 0 & 1 \end{pmatrix}$$
$$B = \begin{pmatrix} L(1+\gamma) & -L\gamma & \bar{\rho}^2 & -1 \\ -m(1+\gamma) & m\gamma & 0 & 1 \end{pmatrix}$$

5 Numerical Results

5.1 Convergence rate by IQCs

Using the results from the previous section we can now study the convergence rate of GNAG. As noted in the previous section, it suffices to study the case when $f \in S^2_{1,L/m}(\mathbb{R}^n)$. We chose to use the standard values for the step size and the momentum coefficient, $\alpha = 1/(L/m)$ and $\beta = (\sqrt{L/m} - 1)/(\sqrt{L/m} + 1)$. For a given $L/m, \gamma$ and ρ, we can then determine the feasibility of (24) and hence determine whether GNAG with gradient correction coefficient γ achieves convergence rate of at least ρ for $f \in S^2_{1,L/m}(\mathbb{R}^n)$. We can then use bisection search to find the lowest ρ for which (24) is feasible, and hence determine the best guaranteed convergence rate of GNAG for given L/m and γ.

We first let L/m vary and study the convergence rate for some choices of the gradient correction coefficient $\Gamma = \gamma/\beta$. See Figure 1 for the results. As we can see the IQC approach proves a faster convergence rate for $\Gamma = 1.2$ than for $\Gamma = 1$, that is, than for the standard NAG. However, if we chose Γ to be a lot larger than 1 we cannot prove convergence for high values of L/m using IQCs. We also see that choices of Γ that are less than 1 do lead to convergence, but to a slower one than NAG. This is as expected, since these parameter choices give an algorithm that in terms of gradient correction lies between NAG and the heavy-ball method, which converges slower than NAG.

We also look at convergence rate depending on Γ for various values of L/m in order to see which choice of Γ gives the fastest convergence. See Figure 2 for the results. Here we can see $\Gamma = 1$ is most often not the parameter choice that guarantees the fastest convergence, and that the larger the condition ratio L/m of the function f is, the larger is the optimal choice of Γ. However we also see that as the choice of Γ becomes larger than its optimal value, the convergence rapidly becomes a lot slower.

5.2 GNAG on Some Classical Problems

In order to study the efficiency of GNAG and compare it to NAG and the heavy-ball method, we solve two classical convex optimization problems numerically using different choices of Γ. The two problems of choice are LASSO and logistic regression as these problems appear often in different machine learning environments and can only be solved numerically.

The results as presented in Figure 3 and Figure 4 show a similar picture. As expected NAG converges faster than the heavy-ball method, and the choice of $\Gamma = 0.5$ lies somewhere between the two. We also see that setting $\Gamma = 2$ and even $\Gamma = 4$ gives faster convergence in both cases. While larger choices of Γ eliminate most of the fluctuation of the standard NAG it does not lead to slower acceleration during the first few iterations. Setting Γ to be very large does however lead to divergence, which is consistent with the results of the previous subsection. In Figure 4 the parameter choice $\Gamma = 4$ gives faster convergence to the optimal point during the first couple of iterations but then fails to reach it. Instead, it oscillates due to repeated overshoot.

6 Discussion

In the paper we have studied the GNAG to solve strictly convex optimization problems $\min f(x)$. This new method is a generalized version of NAG, but instead has 3 parameters and update scheme (1). We derived an ODE that approximates the trajectory of GNAG well for small step sizes, and showed that this ODE converges with linear rate to the local minimum of f for all positive choices of the gradient correction coefficient Γ. We also formulated GNAG as a linear dynamical system with nonlinear feedback, and proved using IQCs that it achieves convergence with linear rate for some choices of Γ. We also saw through numerical examples that setting $\Gamma > 1$ can lead to faster convergence than setting $\Gamma = 1$, however very large choices of Γ lead to divergence.

A series of interesting questions are left to be researched. The IQC we used to prove convergence is only one version of the more general Zames-Falb IQC [2], and maybe using some other version of this more general constraint, convergence can be proven for more choices of Γ. Throughout the paper we mostly used the standard tuning of $\alpha = 1/L$ and $\beta = (\sqrt{L} - \sqrt{m})/(\sqrt{L} + \sqrt{m})$. Looking at the update scheme (1) of GNAG it seems that there is a close interplay between α and γ, it could be interesting to look at convergence when $\alpha \ll 1/L$ and γ is increased. We also saw that for some large choice of Γ we get accelerated convergence to the optimal point, but oscillation around the optimal point. As this does not occur for small choices of Γ one might suggest a method where Γ

is, according to some scheme, sequentially decreased with every iteration. This might also be useful in the case where the L and m of f are not known, as is often the case in practice. Finally, it can be interesting to look at a stochastic version of GNAG, as in many machine learning settings it is too expensive to calculate the gradient of the objective function at every iteration.

References

[1] S. BOYD AND L. VANDENBERGHE, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[2] L. LESSARD, B. RECHT, AND A. PACKARD, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM Journal on Optimization, 26 (2016), pp. 57–95.

[3] A. M. LYAPUNOV, *The General Problem of the Stability of Motion*, PhD thesis, University of Kharkov, 1892.

[4] Y. NESTEROV, *A method of solving a convex programming problem with convergence rate* $o(1/k^2)$, Soviet Mathematics Doklady, 27 (1983), pp. 372–376.

[5] B. T. POLYAK, *Introduction to optimization*, Optimization Software, Inc., 1987.

[6] B. SHI, S. S. DU, M. I. JORDAN, AND W. J. SU, *Understanding the acceleration phenomenon via high-resolution differential equations*. arXiv preprint arXiv:1810.08907, October 2018.

[7] G. STRANG, *Linear Algebra and Learning from Data*, Wellesley - Cambridge Press, Wellesley, Massachusetts, 2019.

A Figures

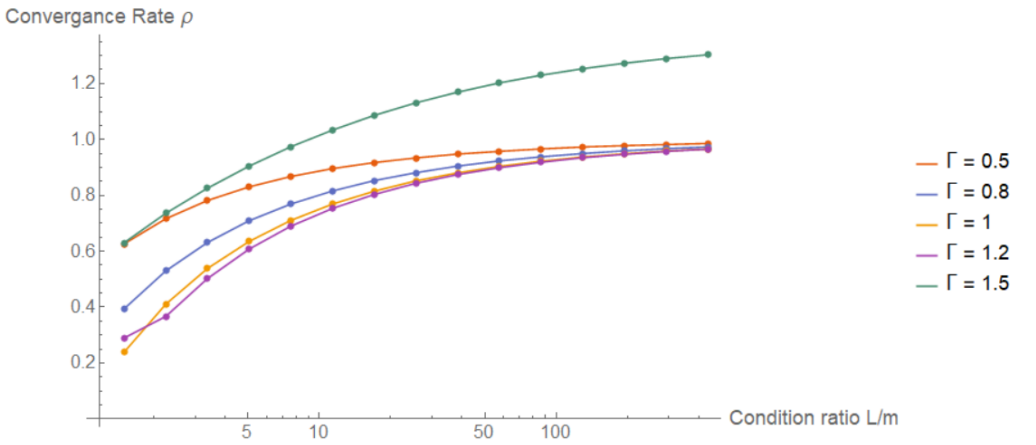


Figure 1: The convergence rate of GNAG for various values of the gradient correction coefficient Γ depending on the ratio of the objective functions Lipschitz Gradients and strict convexity.

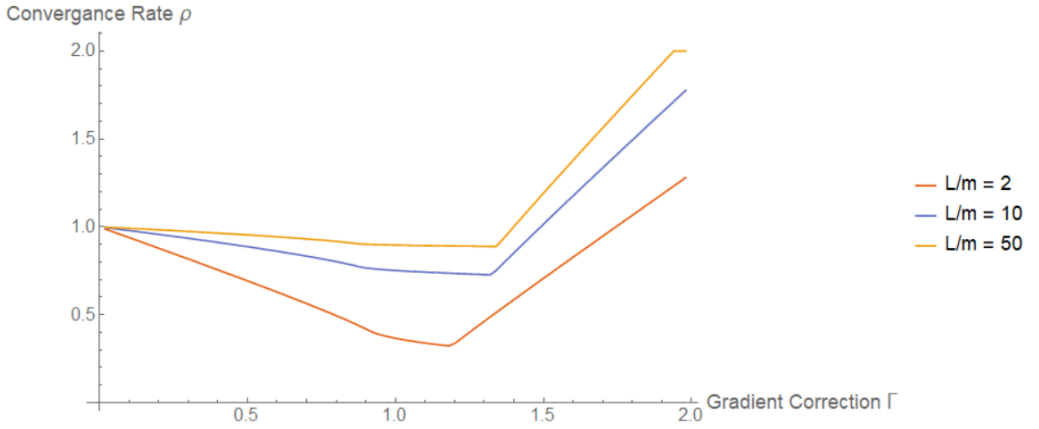


Figure 2: The convergence rate of GNAG for various condition ratios depending on the gradient correction coefficient Γ.

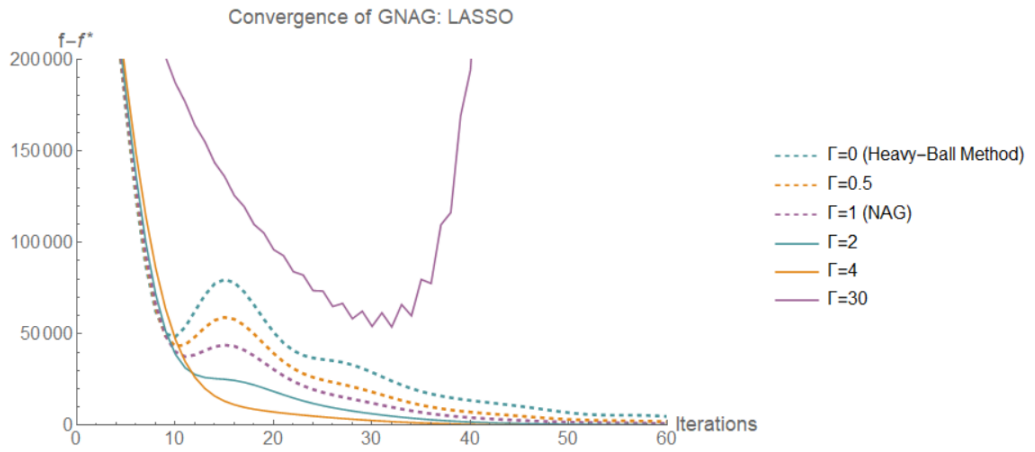


Figure 3: The convergence of GNAG for $f(x) = \|Ax - b\|^2 + \lambda\|x\|_1$ with A of size 100×200 with random entries and $\lambda = 4$.

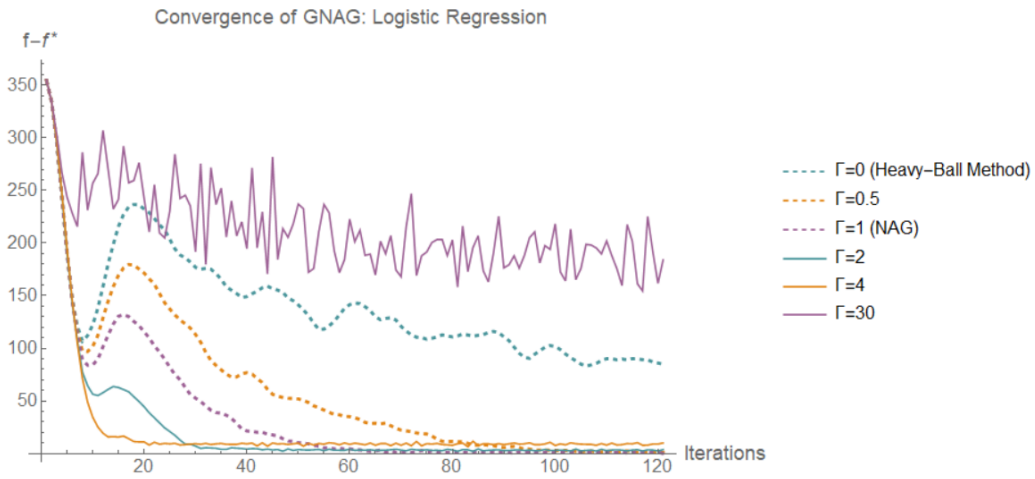


Figure 4: The convergence of GNAG for $f(x) = \sum_{i=1}^{n} -y_i a_i^T x + \log(1 + e^{a_i^T x}) + \lambda\|x\|_1$ with A of size 50×100 with random entries, y with random 0 or 1 entries, and $\lambda = 5$.