Research on 'A geometric alternative to Nesterov's accelerated gradient descent'

Abstract

The paper we focused on proposes a new algorithm for unconstrained optimization problems with a strongly convex and smooth objective function. Compared to gradient descent under Nesterov's acceleration, the new algorithm can reach the same optimal convergence rate. Inspired by the ellipsoid method, it is easier to interpret the new algorithm from geometric aspect. We also perform some numerical simulations to validate the advantage of new method.

1 Introduction

So far, Nesterov's acceleration has already been fundamental for optimization algorithms [1]. However, the intuition of Nesterov's accelerated gradient descent proves to be fairly difficult to explain. Therefore, plenty of recent researches aim to find a new way to interpret this algorithm [2, 3, 4, 5].

The paper we focused on proposes a new algorithm for unconstrained optimization problems with a α-strongly convex and β-smooth function [6]. Compared to gradient descent under Nesterov's acceleration, the new algorithm can reach the same optimal convergence rate. Moreover, it doesn't need the smoothness parameter and the number of iterations to guarantee that the value of objective function is strictly decreasing. The properties are useful for machine learning because the only required parameter α is always given. Furthermore, the combination of zeroth and first order information of the function makes the algorithm particularly well-suited in practice.

The organization of our work is as follows. In Section 2, we extract and summarize the proof analysis process outlined in the paper to demonstrate the theoretical feasibility and comprehensiveness of the proposed **GeoD**(Geometric Descent method) [6] method. In Section 3, we conduct some numerical simulations to validate the advantage of **GeoD**.

2 Method

2.1 Preliminaries

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a α-strongly convex and β-smooth function. Therefore, for $\forall x, y \in \mathbb{R}^n$,

$$f(x) + \nabla f(x)^T(y-x) + \frac{\alpha}{2}|y-x|^2 \le f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}|y-x|^2.$$

Let $\kappa = \frac{\beta}{\alpha}$ be the condition number, $||$ be the Euclidean norm in $\mathbb{R}^n$ and $\mathrm{B}(x, r^2) = \{y \in \mathbb{R}^n : |y-x|^2 \le r^2\}$.
And we denote

$$x^+ = x - \frac{1}{\beta}\nabla f(x), \quad \text{and} \quad x^{++} = x - \frac{1}{\alpha}\nabla f(x).$$

According to the α-strongly convex property, for $\forall y \in \mathbb{R}^n$,

$$f(x) + \nabla f(x)^T(y-x) + \frac{\alpha}{2}|y-x|^2 \le f(y).$$

Let $y = x^*$, we can get

$$\begin{aligned}
f(x) + \nabla f(x)^T(x^*-x) + \frac{\alpha}{2}|x^*-x|^2 &\le f(x^*) \\
\nabla f(x)^T(x^*-x) + \frac{\alpha}{2}|x^*-x|^2 &\le f(x^*) - f(x) \\
\frac{1}{\alpha}\nabla f(x)^T 2(x^*-x) + |x^*-x|^2 &\le \frac{2}{\alpha}(f(x^*) - f(x)) \\
|x^* - x + \frac{1}{\alpha}\nabla f(x)|^2 &\le \frac{|\nabla f(x)|^2}{\alpha^2} - \frac{2}{\alpha}(f(x) - f(x^*)) \\
|x^* - x^{++}|^2 &\le \frac{|\nabla f(x)|^2}{\alpha^2} - \frac{2}{\alpha}(f(x) - f(x^*))
\end{aligned}$$

Therefore,

$$x^* \in \mathrm{B}\left(x^{++}, \frac{|\nabla f(x)|^2}{\alpha^2} - \frac{2}{\alpha}(f(x) - f(x^*))\right).$$

Then substitute y with x^+ in β-smooth condition, we obtain

$$f(x^+) \le f(x) - \frac{1}{2\beta}|\nabla f(x)|^2.$$

With above inequality, substitute $f(x)$ to get a smaller ball,

$$x^* \in \mathrm{B}\left(x^{++}, \frac{|\nabla f(x)|^2}{\alpha^2}(1 - \frac{1}{\kappa}) - \frac{2}{\alpha}(f(x) - f(x^*))\right). \tag{2.1}$$

2.2 Suboptimal algorithm

Let $R_0 > 0$ and $x^* \in \mathcal{B}(x_0, R_0^2)$. Because $f(x^*) \leq f(x_0^+)$, we can omit the $-\frac{2}{\alpha}(f(x) - f(x^*))$ in (2.1).

$$x^* \in \mathcal{B}(x_0, R_0^2) \cap \mathcal{B}\left(x_0 - \frac{1}{\alpha}\nabla f(x_0), \frac{|\nabla f(x_0)|^2}{\alpha^2}(1 - \frac{1}{\kappa})\right), \tag{2.2}$$

It is obviously that for $\forall g \in \mathbb{R}^n$, $\epsilon \in (0, 1)$, there exists $x \in \mathbb{R}^n$ such that

$$\mathcal{B}(0, 1) \cap \mathcal{B}(g, |g|^2(1 - \epsilon)) \subset \mathcal{B}(x, 1 - \epsilon). \qquad \text{(Figure 1)} \tag{2.3}$$

Figure 1 shows the intersection shrinks at the same rate if only one of the ball shrinks.

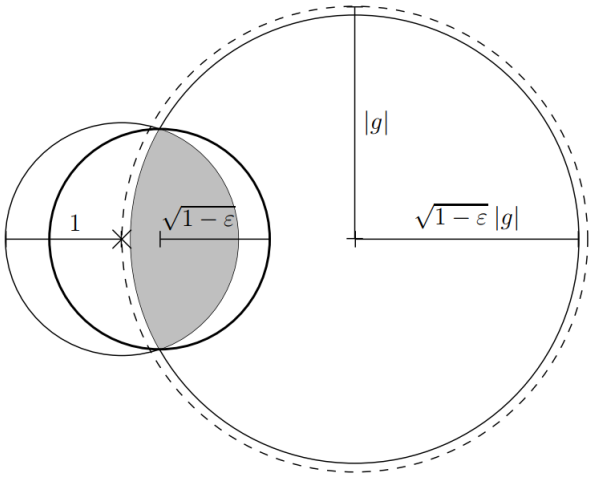

Figure 1: One ball shrinks.

Combing (2.2) and (2.3) we obtain that $\exists x_1 \in \mathbb{R}^n$ such that

$$x^* \in \mathcal{B}\left(x_1, R_0^2(1 - \frac{1}{\kappa})\right). \tag{2.4}$$

Consider the iteration sequence $\{x_k\}$,

$$\begin{aligned} x^* \in \mathcal{B}\left(x_1, R_0^2(1 - \frac{1}{\kappa})\right) &\iff |x^* - x_1|^2 \leq (1 - \frac{1}{\kappa})R_0^2 \\ &\iff |x^* - x_k|^2 \leq (1 - \frac{1}{\kappa})^k R_0^2, \end{aligned} \tag{2.5}$$

It means that the iterative method can achieve ϵ-close to minimizer x^* after $2\kappa \log(R_0/\epsilon)$ iterations.

2.3 Acceleration intuition

Choose $R_0 > 0$ such that $x^* \in \mathcal{B}(x_0, R_0^2 - \frac{2}{\alpha}(f(y) - f(x^*)))$ where $f(x_0) \leq f(y)$. According to α-strongly convex β-smoothness property of f that $f(x_0^+) = f(x_0 - \frac{1}{\beta}\nabla f(x_0)) \leq f(x_0) - \frac{1}{2\beta}|\nabla f(x_0)|^2 \leq$

$f(y) - \frac{1}{2\alpha\kappa}|\nabla f(x_0)|^2$, we have

$$x^* \in \mathcal{B}\left(x_0, R_0^2 - \frac{|\nabla f(x_0)|^2}{\alpha^2\kappa} - \frac{2}{\alpha}(f(x_0^+) - f(x^*))\right). \tag{2.6}$$

Meanwhile, from (2.1) we have

$$x^* \in \mathcal{B}\left(x_0^{++}, \frac{|\nabla f(x_0)|^2}{\alpha^2}(1 - \frac{1}{\kappa}) - \frac{2}{\alpha}(f(x_0^+) - f(x^*))\right). \tag{2.7}$$

It is intuitive to observe that both (2.6) and (2.7) are shrinkage of $\mathcal{B}(x_0, R_0^2)$. Intersecting them using **Lemma 1** below, we obtain that there exist $x_1 \in \mathbb{R}^n$ such that

$$x^* \in \mathcal{B}\left(x_1, R_0^2(1 - \frac{1}{\sqrt{\kappa}}) - \frac{2}{\alpha}(f(x_0^+) - f(x^*))\right), \qquad \text{(Figure 2)} \tag{2.8}$$

which gives an acceleration in shrinking of the radius since $\kappa > 1$ and $1 - \frac{1}{\sqrt{\kappa}} < 1 - \frac{1}{\kappa}$.

Figure 2 shows the intersection shrinks much faster if two balls shrinks at the same absolute amount

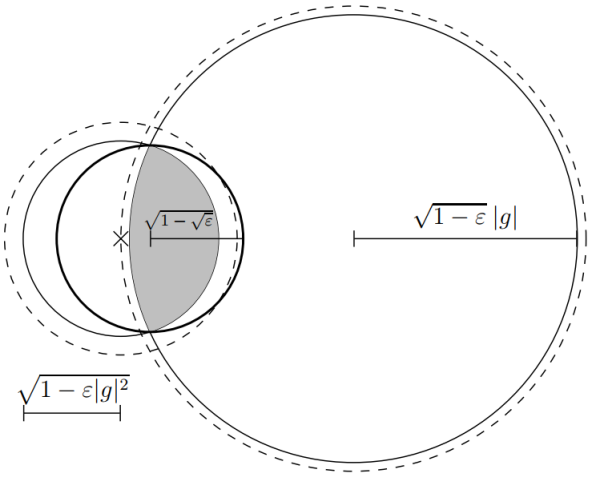

Figure 2: Two balls shrink.

2.4 Optimal algorithm

Algorithm 1: Minimum Enclosing Ball of the Intersection to Two Balls

Input: a ball centered at x_A with radius R_A and a ball centered at x_B with radius R_B.

if $|x_A - x_B|^2 \geq |R_A^2 - R_B^2|$ **then**

$c = \frac{1}{2}(x_A + x_B) - \frac{R_A^2 - R_B^2}{2|x_A - x_B|^2}(x_A - X_B)$. $R^2 = R_B^2 - \frac{\left(|x_A - x_B|^2 + R_B^2 - R_A^2\right)^2}{4|x_A - x_B|^2}$.

else if $|x_A - x_B|^2 < R_A^2 - R_B^2$ **then**

$c = x_B$. $R = R_B$.

else

$c = x_A$. $R = R_A$.[a]

end

Output: a ball centered at c with radius R.

[a] If we assume $|x_A - x_B| \geq R_B$ as in Lemma 1, this extra case does not exist.

Let $x_0 \in \mathbb{R}^n, c_0 = x_0^{++}$ and $R_0^2 = (1 - \frac{1}{\kappa})\frac{|\nabla f(x_0)|^2}{\alpha^2}$.

The iteration rule of x_k is

$$x_{k+1} = \text{line search}(c_k, x_k^+).$$

Theorem 1 For $\forall k \geq 0$, one has $x^* \in \mathrm{B}(c_k, R_k^2), R_{k+1}^2 \leq (1 - \frac{1}{\sqrt{\kappa}})R_k^2$, and

$$|x^* - c_k|^2 \leq (1 - \frac{1}{\sqrt{\kappa}})^k R_0^2.$$

Lemma 1 Let $a \in \mathbb{R}^n$ and $\epsilon \in (0, 1)$, $g \in \mathbb{R}_+$. Assume that $|a| \geq g$. Then $\exists c \in \mathbb{R}^n$ such that for any $\delta > 0$,

$$\mathrm{B}(0, 1 - \epsilon g^2 - \epsilon) \cap \mathrm{B}(a, g^2(1 - \epsilon) - \delta) \subset \mathrm{B}(c, 1 - \sqrt{\epsilon} - \delta).$$

Algorithm 1 calculate the minimum enclosing ball of the intersection to two balls. The correctness of **Algorithm 1** can be derived from **Theorem 1**.

Algorithm 2: Geometric Descent Method (GeoD)

Input: parameters α and initial points x_0.

$x_0^+ = \texttt{line\_search}(x_0, x_0 - \nabla f(x_0))$.

$c_0 = x_0 - \alpha^{-1}\nabla f(x_0)$.

$R_0^2 = \frac{|\nabla f(x_0)|^2}{\alpha^2} - \frac{2}{\alpha}\left(f(x_0) - f(x_0^+)\right)$.

for $i \leftarrow 1, 2, \cdots$ **do**

Combining Step:

$x_k = \texttt{line\_search}(x_{k-1}^+, c_{k-1})$.

Gradient Step:

$x_k^+ = \texttt{line\_search}(x_k, x_k - \nabla f(x_k))$.

Ellipsoid Step:

$x_A = x_k - \alpha^{-1}\nabla f(x_k)$. $R_A^2 = \frac{|\nabla f(x_k)|^2}{\alpha^2} - \frac{2}{\alpha}\left(f(x_k) - f(x_k^+)\right)$.

$x_B = c_{k-1}$. $R_B^2 = R_{k-1}^2 - \frac{2}{\alpha}\left(f(x_{k-1}^+) - f(x_k^+)\right)$.

Let $B(c_k, R_k^2)$ is the minimum enclosing ball of $B(x_A, R_A^2) \cap B(x_B, R_B^2)$.

end

Output: x_T.

Algorithm 2 (**GeoD**) is a more aggressive algorithm using line search instead of fixed step size. The correctness of **Algorithm 2** follows from a similar proof as **Theorem 1**.

3 Experiment

In this Section, we conduct a comparative experiment among **GeoD**, **SD**, and **AFG**. We evaluate the algorithms via a minimization problem

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{\beta}{2}\left((1-x_1)^2 + \sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + x_n^2\right) + \frac{1}{2}\sum_{i=1}^{n} x_i^2, \tag{3.1}$$

where β is the smoothness parameter. We choose $\epsilon = 10^{-8}$, $error = \|x_k - x^*\|_2^2$, and plot the variation of the *error* of **GeoD**, **SD**, and **AFG** with the number of iterations before reaching the ϵ-close to minimizer x^* ($error < \epsilon$). Then we figure out the convergence rate of **GeoD**, **SD**, and **AFG** from our graphics and validate the advantage of **GeoD** over others.

In background settings, our problem dimension is $d = 10^4$ with strongly convex parameter $\alpha = 0.5$ and smoothness parameter $\beta = 1$. Firstly, since (3.1) is strongly convex, we derive the optimal solution of (3.1) from solving linear equation $\nabla f(x^*) = 0$. Secondly, we implement **GeoD**, **SD**, and **AFG** separately under our background settings. For **GeoD**, we directly implement **Algorithm 2** [6] mentioned above, and use binary search to calculate next iteration point in function $line_search$. For **SD**, we update with $x_{k+1} = x_k - \eta\nabla f(x_k)$ where $\eta = 0.1$ is the learning rate. For **AFG**, we implement **Constant Step Scheme, II** [7] with $q = \alpha/\beta = 0.5$ and learning rate $\eta = 0.1$.

From **Figure 3**, we can see that **GeoD**, **SD**, and **AFG** all converge linearly since their *error* on

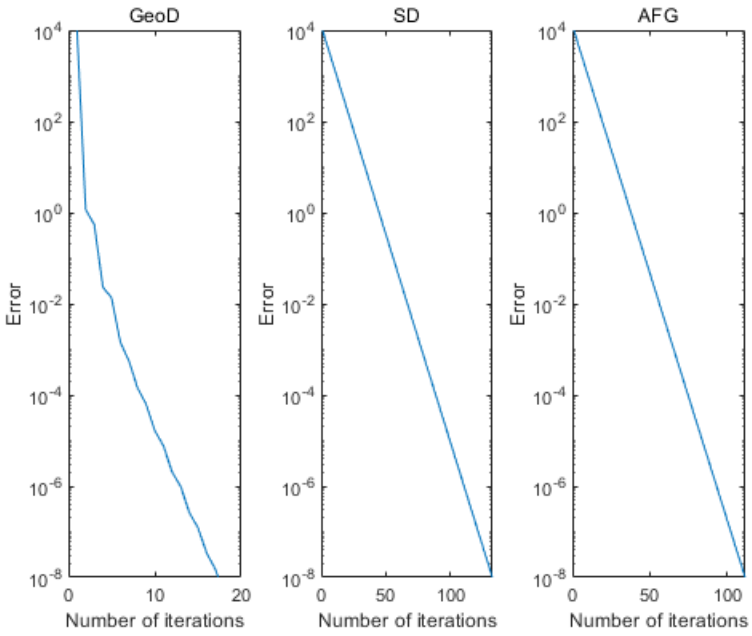

Figure 3: **GeoD**, **SD**, and **AFG** converge linearly

the logarithmic scale decrease approximately linearly with the increase of the number of iterations, which implies

$$\|x_{k+1} - x^*\|_2^2 = \rho \|x_k - x^*\|_2^2, \quad 0 < \rho < 1.$$

From **Figure 4**, **GeoD** achieves ϵ-close to minimizer x^* in 17 iterations, while **SD** and **AFG** achieve in 132 and 111 iterations. **GeoD** converges a constant multiple (i.e.$\times 6$) faster than **SD** and **AFG**.

4 Conclusion

References

[1] S. Bubeck, "Theory of convex optimization for machine learning," *arXiv preprint arXiv:1405.4980*, vol. 15, 2014.

[2] Z. Allen-Zhu and L. Orecchia, "Linear coupling: An ultimate unification of gradient and mirror descent," *arXiv preprint arXiv:1407.1537*, 2014.

[3] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 57–95, 2016.

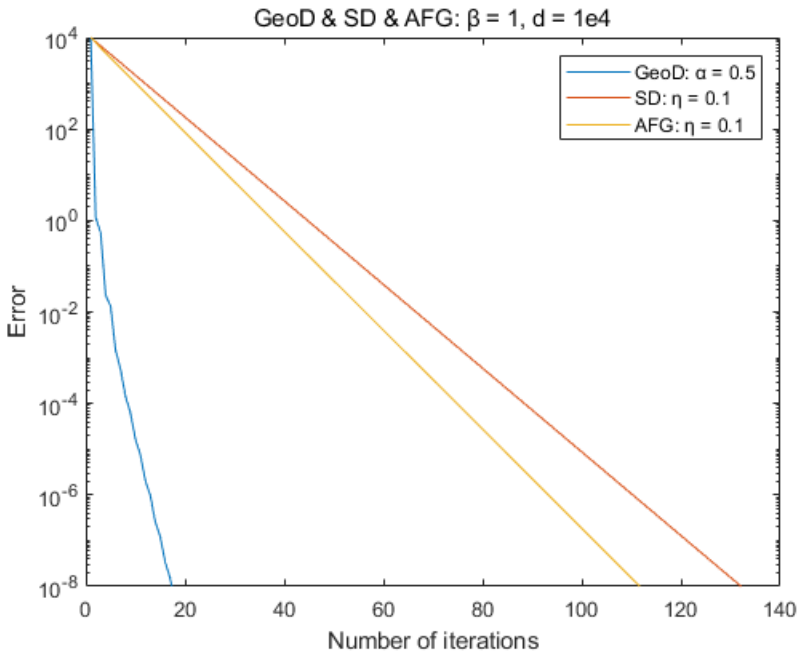

Figure 4: **GeoD** converge a constant multiple faster than **SD** and **AFG**

[4] W. Su, S. Boyd, and E. J. Candes, "A differential equation for modeling nesterov's accelerated gradient method: Theory and insights," *Journal of Machine Learning Research*, vol. 17, no. 153, pp. 1–43, 2016.

[5] N. Flammarion and F. Bach, "From averaging to acceleration, there is only a step-size," in *Conference on Learning Theory*. PMLR, 2015, pp. 658–695.

[6] S. Sachdeva and Y. Rubanova, "A geometric alternative to nesterov accelerated gradient descent," 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:53551438

[7] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004.