A Simple Proof for the Iteration Complexity of the Proximal Gradient Algorithm

N. Denizcan Vanli
MIT
denizcan@mit.edu

Mert Gürbüzbalaban
Rutgers University
mertg@mit.edu

Asu Ozdaglar
MIT
asuman@mit.edu

Abstract

We study the problem of minimizing the sum of a smooth strongly convex function and a non-smooth convex function. We consider solving this problem using the proximal gradient (PG) method, which at each iteration uses the proximal operator with respect to the non-smooth convex function at the intermediate iterate obtained using the gradient with respect to the smooth strongly convex function. We introduce a simple novel analysis and show that the PG algorithm attains a globally linear convergence rate provided that the step size is sufficiently small. Consequently, we obtain iteration complexity results for the PG method. We also extend our analysis to study an inexact proximal method, called the proximal incremental aggregated gradient method, and show that this method is globally convergent with a linear rate.

1 Introduction

We focus on *composite optimization problems*, where the objective function is given by the sum of a loss function f and a possibly non-differentiable regularization function r:

$$\min_{x\in\mathbb{R}^n} F(x) \triangleq f(x) + r(x). \tag{1}$$

We assume the loss function $f : \mathbb{R}^n \to (-\infty, \infty)$ is convex and continuously differentiable while the regularization function $r : \mathbb{R}^n \to (-\infty, \infty]$ is proper, closed, and convex but not necessarily differentiable. This formulation arises in many problems in constrained optimization, distributed optimization, machine learning, and signal processing. Examples include distributed optimization problems that arise in wireless sensor network as well as smart grid applications [8, 13], constrained and regularized least squares problems that arise in various machine learning [4, 6, 15] and signal processing [2, 3] applications .

The proximal gradient (PG) method is a popular method for solving the problem (1) [14, 17]. It uses the proximal operator with respect to the regularization function r at the intermediate iterate obtained using the gradient with respect to f, consisting of the iterations

$$x_{k+1} = \text{prox}_r^\eta(x_k - \eta\nabla f(x_k)), \quad k \geq 0, \tag{2}$$

where η is a constant step size and the proximal mapping is defined as follows

$$\text{prox}_r^\eta(y) = \arg\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2} ||x-y||^2 + \eta r(x) \right\}. \tag{3}$$

The convergence properties of the PG method is well studied in the literature under various conditions on the functions f and r, see e.g. [7, 10, 11, 23] where existing work on the proximal point algorithm such as [1, 18, 16] have been a building block for understanding the PG method. A particular

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

case of importance is when f is strongly convex and ∇f is Lipschitz continuous, which would for instance arise in regression problems. The classical analysis in the literature uses distance to the optimal solution (which is unique by strong convexity and will be denoted by x^*) as a Lyapunov function and shows that the distance of the iterates (generated by the PG method) to the optimal solution decreases with an exponential rate [20]. However, this result does not directly translate into a linear convergence result in the suboptimality of the objective values $F(x_k) - F(x^*)$ as the objective values $F(x_k)$ may not vary smoothly with respect to the iterates x_k due to the potentially non-differentiable regularization function r. Two recent papers [11] and [7] address this issue of deriving iteration complexity results for the suboptimality in objective values. In particular, in [11], the global linear convergence of the PG method is presented under the Polyak-Lojasiewicz (PL) condition, a weaker condition than the strong convexity of f. Similar to [11], in [7], it is shown that the iteration complexity of the PG method grows linearly with the condition number of the problem. The main idea is to show that under a quadratic growth condition (which is satisfied when f is strongly convex), the *error bound condition* holds, i.e., a multiple of the step length at each iteration bounds the distance to the solution set.[1]

Contributions. In this paper, we present a novel iteration complexity result for the PG algorithm in terms of suboptimality in function values using a simple and insightful analysis. Building on the properties of the proximal mapping and strong convexity, we first show that the error bound condition holds (Lemma 2.1). In this sense, our approach is different from the one in [11], while our results are asymptotically the same. Compared to [7], we make stronger assumptions (the strong convexity of f implies their quadratic growth condition), however in return we get better constants in the error bound condition (see Lemma 2.1 vs. [7, Cor 3.6]). Using this improved error bound, Lipschitzness of the gradients and a descent lemma, we show global linear convergence of the PG method and present the corresponding iteration complexity result (Theorem (2.2) and Corollary 2.3, respectively). Our simple analysis extends and allows us to study inexact proximal methods. We illustrate this by presenting a new iteration complexity result for the *proximal incremental aggregated gradient algorithm*, which is an inexact deterministic proximal gradient method that is typically much faster than the proximal gradient for problems when the loss function f has an additive structure of the form $f(x) = \frac{1}{m}\sum_i^m f_i(x)$ and m is large, a problem of particular interest in large-scale data processing and distributed optimization [4, 6, 15].

2 Convergence Analysis

We first start with some preliminaries on the PG algorithm. Defining $\phi(x) \triangleq \frac{1}{2}||x-y||^2 + \eta r(x)$ and letting $\partial\phi(x)$ denote the set of subgradients of the function ϕ at x, it follows from the optimality conditions of the problem in (3) that $0 \in \partial\phi(x_{k+1})$. This yields $x_{k+1} - (x_k - \eta\nabla f(x_k)) + \eta h_{k+1} = 0$, for some $h_{k+1} \in \partial r(x_{k+1})$. Hence, we can compactly represent our update rule as

$$x_{k+1} = x_k + \eta d_k, \tag{4}$$

where $d_k = -\nabla f(x_k) - h_{k+1}$ is the direction of the update at time k. Throughout this paper, we make the following assumptions on the objective function, which appears frequently for analyzing proximal methods in the literature [5, 6, 9, 19].

- **(A.1)** f is μ-strongly convex and ∇f is globally Lipschitz with constant L on $\mathbb{R}^n$,
- **(A.2)** r is proper, closed, convex and subdifferentiable everywhere in its effective domain.

A consequence of these assumptions is that the solution to (1) is unique, which we denote by x^*. The condition number of the problem is defined as the ratio

$$Q \triangleq L/\mu. \tag{5}$$

In the rest of the section, we show that the PG algorithm attains a global linear convergence rate with a constant step size provided that the step size is sufficiently small. Using this result, we obtain an iteration complexity result for the PG algorithm which depends linearly on the condition number Q. First, we introduce the following lemma, which can be interpreted as follows. For the special case $r(x) = 0$ for all $x \in \mathbb{R}^n$, i.e., when the PG algorithm reduces to the classical gradient descent algorithm, an identical result to this lemma simply follows from the strong convexity of the function

[1] See [12] for a rigourous definition of this condition.

f since $||x_k - x^*|| \leq \frac{1}{\mu} ||\nabla f(x_k) - \nabla f(x^*)||$ and $\nabla f(x^*) = 0$ due to the optimality condition of the problem. The following lemma indicates that even though we do not have such control over the subgradients of the regularization function (as the regularization function is neither strongly convex nor smooth), the properties of the proximal step yields a similar relation for the direction of update at the expense of constant $2/\mu$ (instead of $1/\mu$ compared to the $r(x) = 0$ case).

Lemma 2.1 *Under Assumptions (A.1)-(A.2), the distance of the iterates (generated by the PG algorithm) from the optimal solution is upper bounded as*

$$||x_k - x^*|| \leq \frac{2}{\mu} ||d_k||,$$

for any iteration $k \geq 0$ and step size $0 < \eta \leq \frac{1}{L}$.

Proof: The non-expansiveness property of the proximal map implies

$$||\text{prox}_r^\eta(x) - \text{prox}_r^\eta(y)||^2 \leq \langle \text{prox}_r^\eta(x) - \text{prox}_r^\eta(y), x - y \rangle,$$

for any $x, y \in \mathbb{R}^n$. Putting $x = x_k - \eta \nabla f(x_k)$ and $y = x^* - \eta \nabla f(x^*)$ in the above inequality, we obtain

$$\begin{aligned} ||x_k + \eta d_k - x^*||^2 &\leq \langle x_k + \eta d_k - x^*,\, x_k - \eta \nabla f(x_k) - x^* + \eta \nabla f(x^*) \rangle \\ &= \langle x_k + \eta d_k - x^*,\, x_k + \eta d_k - x^* \rangle + \langle x_k + \eta d_k - x^*,\, -\eta d_k + \eta \nabla f(x^*) - \eta \nabla f(x_k) \rangle, \end{aligned}$$

which implies

$$0 \leq \langle x_k + \eta d_k - x^*,\, -d_k + \nabla f(x^*) - \nabla f(x_k) \rangle.$$

This inequality can be rewritten as follows

$$\begin{aligned} \langle x_k - x^*,\, \nabla f(x_k) - \nabla f(x^*) \rangle &\leq \langle x_k - x^*,\, -d_k \rangle - \eta ||d_k||^2 + \eta \langle d_k,\, \nabla f(x^*) - \nabla f(x_k) \rangle \\ &\leq \langle x_k - x^*,\, -d_k \rangle + \eta \langle d_k,\, \nabla f(x^*) - \nabla f(x_k) \rangle \\ &\leq ||d_k|| \left(||x_k - x^*|| + \eta ||\nabla f(x^*) - \nabla f(x_k)|| \right) \\ &\leq ||d_k|| \left(||x_k - x^*|| + \eta L ||x_k - x^*|| \right) \\ &\leq 2 ||d_k|| \, ||x_k - x^*||, \end{aligned} \tag{6}$$

where the second inequality follows since $-||d_k||^2 \leq 0$, the third inequality follows by the Cauchy-Schwarz inequality, the fourth inequality follows from the L-smoothness of f, and the last inequality follows since $\eta \leq \frac{1}{L}$. Since μ-strong convexity of f implies

$$\mu ||x_k - x^*||^2 \leq \langle x_k - x^*,\, \nabla f(x_k) - \nabla f(x^*) \rangle, \tag{7}$$

then combining (6)-(7) and simplifying one of the distance terms, we obtain the desired inequality. □

Building on this lemma, in the following theorem, we show that the PG algorithm attains a global linear convergence rate for sufficiently small step sizes. A consequence of this theorem, Corollary 2.3 shows that the iteration complexity of the PG algorithm grows linearly with the condition number Q of the problem.

Theorem 2.2 *Under Assumptions (A.1)-(A.2), the PG algorithm with step size $0 < \eta \leq \frac{1}{L}$ is linearly convergent satisfying*

$$F(x_k) - F(x^*) \leq \left(1 + \eta \frac{\mu}{4}\right)^{-k} (F(x_0) - F(x^*)), \tag{8}$$

for any $k \geq 1$.

Proof: The difference in the function values of two consecutive iterations can be upper bounded as

$$F(x_{k+1}) - F(x_k) \leq \eta \langle \nabla f(x_k), d_k \rangle + \eta^2 \frac{L}{2} ||d_k||^2 + r(x_{k+1}) - r(x_k), \tag{9}$$

using the L-smoothness of f. We then observe that

$$\eta \langle \nabla f(x_k), d_k \rangle + r(x_{k+1}) - r(x_k) = -\eta ||d_k||^2 - \eta \langle h_{k+1}, d_k \rangle + r(x_{k+1}) - r(x_k) \leq -\eta ||d_k||^2, \tag{10}$$

where the equality follows by the definition $d_k = -\nabla f(x_k) - h_{k+1}$ and the inequality follows by the convexity of r. Using (10) in (9), we obtain

$$F(x_{k+1}) - F(x_k) \leq -\eta \left|\left|d_k\right|\right|^2 + \eta^2 \frac{L}{2} \left|\left|d_k\right|\right|^2 \leq -\frac{\eta}{2} \left|\left|d_k\right|\right|^2,$$

since $\eta \leq \frac{1}{L}$. Using Lemma 2.1 in the above inequality, we get

$$F(x_{k+1}) - F(x_k) \leq -\eta \frac{\mu}{4} \left|\left|d_k\right|\right| \left|\left|x_k - x^*\right|\right| \leq -\eta \frac{\mu}{4} \langle d_k, x^* - x_k \rangle, \tag{11}$$

where the last inequality follows by the Cauchy-Schwarz inequality. This inner product can be rewritten as

$$\begin{aligned} -\langle d_k, x^* - x_k \rangle &= \langle \nabla f(x_k) + h_{k+1}, x^* - x_k \rangle \\ &= \langle \nabla f(x_k), x^* - x_k \rangle + \langle h_{k+1}, x^* - x_{k+1} \rangle + \eta \langle h_{k+1}, d_k \rangle. \end{aligned} \tag{12}$$

Since f and r are convex, the right-hand side of (12) can be upper bounded by

$$-\langle d_k, x^* - x_k \rangle \leq f(x^*) - f(x_k) + r(x^*) - r(x_{k+1}) + \eta \langle h_{k+1}, d_k \rangle. \tag{13}$$

We then consider the inner product term $\eta \langle h_{k+1}, d_k \rangle$ in the above inequality and rewrite it as

$$\eta \langle h_{k+1}, d_k \rangle = -\eta \left|\left|d_k\right|\right|^2 - \eta \langle \nabla f(x_k), d_k \rangle = -\eta \left|\left|d_k\right|\right|^2 + \langle \nabla f(x_k), x_k - x_{k+1} \rangle.$$

Using the L-smoothness of f, we can upper bound the RHS of the above inequality as

$$\eta \langle h_{k+1}, d_k \rangle \leq f(x_k) - f(x_{k+1}) - \frac{\eta}{2} \left|\left|d_k\right|\right|^2, \tag{14}$$

since $\eta \leq \frac{1}{L}$. Using (14) in (13), we obtain

$$-\langle d_k, x^* - x_k \rangle \leq f(x^*) - f(x_{k+1}) + r(x^*) - r(x_{k+1}) - \frac{\eta}{2} \left|\left|d_k\right|\right|^2 \leq F(x^*) - F(x_{k+1}), \tag{15}$$

since $-\frac{\eta}{2} \left|\left|d_k\right|\right|^2 \leq 0$. Using (15) in (11) and after simple algebra, we get the inequality in (8). □

Corollary 2.3 *Under Assumptions (A.1)-(A.2), the PG algorithm with step size $\eta = \frac{1}{L}$ is guaranteed to return an ϵ-optimal solution after at most $5Q \log\left(\frac{F(x_0) - F(x^*)}{\epsilon}\right)$ iterations.*

3 Applications to Proximal Incremental Aggregated Gradient Method

The technique we present in Section 2 is applicable to analyze other algorithms that include a proximal step or an inexact proximal step. To illustrate this point further, we will next introduce the proximal incremental aggregated gradient method which is equivalent to an inexact PG method and derive novel iteration complexity results for its convergence.

When the loss function f has an additive form, i.e., $f(x) = \frac{1}{m} \sum_{i=1}^m f_i(x)$, and each f_i is L_i-smooth such that f is L-smooth with $L = \frac{1}{m} \sum_{i=1}^m L_i$. In such large-scale optimization problems (with large m), computing the full gradient $\nabla f(x_k)$ at each iteration and therefore the PG method becomes costly [5]. Instead, a prevalent approach is to calculate a gradient with respect to a single loss function f_i at each iteration and use the outdated gradients for the remaining loss functions. In particular, at each iteration k, we form an aggregated gradient $g_k = \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{\tau_{i,k}})$, where $\nabla f_i(x_{\tau_{i,k}})$ represents the gradient of the i-th component function sampled at time $\tau_{i,k}$. In this setup, the time delays in gradient computations are bounded, i.e., there exists an integer $K \geq 0$ such that $k - K \leq \tau_{i,k} \leq k$ for all $i \in \{1, \ldots, m\}$. Then, at each iteration k, the proximal incremental aggregated gradient algorithm takes a step along the approximate gradient descent direction $-g_k$ and apply the proximal mapping to this intermediate iterate, i.e., it takes the combined step $x_{k+1} = \text{prox}_r^\eta(x_k - \eta g_k)$. For this algorithm, using a similar methodology to Section 2, we obtain the following iteration complexity guarantee. The proof details can be found in [21] and [22].

Theorem 3.1 *The proximal incremental aggregated gradient algorithm with step size $\eta = \frac{16}{\mu}\left[\left(1 + \frac{1}{48Q}\right)^{\frac{1}{K+1}} - 1\right]$ is guaranteed to return an ϵ-optimal solution after at most $49Q(K + 1) \log\left(\frac{F(x_0) - F(x^*)}{\epsilon}\right)$ iterations.*

References

[1] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, 2009.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[3] A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery problems. In D. Palomar and Y. Eldar, editors, *Convex Optimization in Signal Processing and Communications*, pages 42–88, 2010.

[4] D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. *Optimization for Machine Learning*, 2010:1–38, 2011.

[5] D. P. Bertsekas. Incremental aggregated proximal and augmented lagrangian algorithms. *arXiv preprint arXiv:1509.09257*, 2015.

[6] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27*, pages 1646–1654, 2014.

[7] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *CoRR*, abs/1602.06661, 2016.

[8] F. Guo, C. Wen, J. Mao, and Y. D. Song. Distributed economic dispatch for smart grids with random wind power. *IEEE Transactions on Smart Grid*, 7(3):1572–1583, May 2016.

[9] M. Gurbuzbalaban, A. Ozdaglar, and P. Parrilo. On the convergence rate of incremental aggregated gradient algorithms. *arXiv preprint arXiv:1506.02081*, 2015.

[10] M. Kadkhodaie, M. Sanjabi, and Z.-Q. Luo. On the linear convergence of the approximate proximal splitting method for non-smooth convex optimization. *CoRR*, abs/1404.5350, 2014.

[11] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of proximal-gradient methods under the Polyak-Lojasiewicz condition. *CoRR*, abs/1608.04636, 2016.

[12] Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[13] A. Nedic, D. P. Bertsekas, and V. S. Borkar. Distributed asynchronous incremental subgradient methods. *Studies in Computational Mathematics*, 8:381–407, 2001.

[14] Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Applied Optimization. Springer, Boston, 2004.

[15] F. Niu, B. Recht, C. Re, and S. Wright. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24*, pages 693–701, 2011.

[16] N. Parikh and S. Boyd. Proximal algorithms. In *Foundations and Trends in Optimization*, volume 1, pages 123–231, 2013.

[17] B. T. Polyak. *Introduction to optimization*. Translations series in mathematics and engineering. Optimization Software, Publications Division, New York, 1987.

[18] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[19] N. L. Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems 25*, pages 2663–2671, 2012.

[20] M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *CoRR*, abs/1109.2415, 2011.

[21] N. D. Vanli, M. Gurbuzbalaban, and A. Ozdaglar. Global convergence rate of proximal incremental aggregated gradient methods. *CoRR*, abs/1608.01713, 2016.

[22] N. D. Vanli, M. Gurbuzbalaban, and A. Ozdaglar. A stronger convergence result on the proximal incremental aggregated gradient method. *CoRR*, abs/1611.08022, 2016.

[23] H. Zhang. The restricted strong convexity revisited: Analysis of equivalence to error bound and quadratic growth. *CoRR*, abs/1511.01635, 2015.