Finite convergence of the inexact proximal gradient method to sharp minima

Andrei Pătraşcu

[a] *Research Center for Logic, Optimization and Security (LOS), Department of Computer Science, Faculty of Mathematics and Computer Science, University of Bucharest, Academiei 14, Bucharest, Romania*

Abstract

Attractive properties of subgradient methods, such as robust stability and linear convergence, has been emphasized when they are used to solve nonsmooth optimization problems with sharp minima [12, 13]. In this letter we extend the robustness results to the composite convex models and show that the basic proximal gradient algorithm under the presence of a sufficiently low noise still converges in finite time, even if the noise is persistent.

Keywords: Inexact first order methods, Sharp Minima, Finite convergence

1. Introduction

In this letter we are interested in the following convex minimization problem:

$$F^* = \min_{x \in \mathbf{R}^n} \; F(x) := f(x) + h(x), \tag{1}$$

where f is the smooth convex component and h convex and lower-semicontinuous. In particular, Lipschitz continuous regularizers or indicator functions are covered by the above model. This nonsmooth model represents the topic of a large sector of recent optimization literature aiming to illustrate the behaviour of the proximal first order methods on minimizing composite structured functions. Linear convergence results are typically obtained under strong convexity or quadratic growth regularity [7, 4, 16], which may hold despite the nondifferentiability of the objective function F. However, as shown in several early papers of Polyak

Email address: andrei.patrascu@fmi.unibuc.ro (Andrei Pătraşcu)

[10, 11, 12, 13] and others [1, 5, 3], nondifferentiability may generate sharp minima. We say that F has weak sharp minima if there exists $\sigma_F > 0$ such that:

$$F(x) \geq F^* + \sigma_F \text{dist}_{X^*}(x) \quad \forall x \in \text{dom } F, \tag{2}$$

where X^* denotes the optimal set of (1) and σ_F the sharpness modulus. Well known works as [12, 5] emphasized the superstability of subgradient methods (SM) that minimize nonsmooth functions with sharp minima. On short, when inexact subgradients are evaluated and used at each iteration, if the possibly persistent noise is sufficiently low, then the variable stepsize SM converges linearly to the sharp optimum of (1).

Although (2) is more often met in nonsmooth models, sharp minima may also exist in the landscape of particular composite functions, see e.g. [3, 2]. When this happens, it is well-known that exact first-order methods typically converge in finite time [1, 2, 3, 8]. Up to our knowledge, complexity and stability bounds, under (2), on the inexact proximal gradient methods lacks from literature. Therefore, as a preliminary step in this direction, in this letter we derive finite termination of the basic inexact proximal gradient when the noise generated by the pertubed gradients of f is low and persistent.

Notations and preliminaries. Given the closed convex set X, let $\pi_X(\cdot)$ be the projection operator onto X. Thereby, the distance to X is $\text{dist}_X(x) = \|x - \pi_X(x)\|$. Let h be a closed, convex function, then its Moreau envelope and its proximal operator [14] are defined by:

$$h_\beta(x) = \min_z \; h(z) + \frac{1}{2\beta}\|z - x\|^2$$
$$\text{prox}_{h,\beta}(x) = \arg\min_z \; h(z) + \frac{1}{2\beta}\|z - x\|^2.$$

Thereby, given $z \in \mathbf{R}^n$, the following optimality condition holds:

$$\frac{1}{\beta}\left(z - \text{prox}_{h,\beta}(z)\right) \in \partial h(\text{prox}_{h,\beta}(z)). \tag{3}$$

The Moreau envelope h_β is smooth and ∇h_β Lipschitz continuous with constant $\frac{1}{\beta}$ [14]. Equivalently, the prox operator is nonexpansive:

$$\|\text{prox}_{h,\beta}(x) - \text{prox}_{h,\beta}(y)\| \leq \|x - y\| \qquad \forall x, y \in \text{dom} F. \tag{4}$$

A function f has Lipschitz gradient with constant L_f if:

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\| \quad \forall x, y \in \text{dom} F. \tag{5}$$

2. Inexact Proximal Gradient Method

Throughout this section, where we state our main results, we make the usual assumption that the smooth component f of (1) has Lipschitz gradient, i.e. it satisfies (5). The basic Proximal Gradient Method (PGM) is the following simple recurrence:

$$x^{k+1} := \text{prox}_{h,\alpha}(x^k - \alpha \nabla f(x^k)),$$

where $\alpha > 0$ is the constant stepsize. Let us denote the PGM update operator $\mathcal{T}(z;\alpha) := \text{prox}_{h,\alpha}(z - \alpha \nabla f(z))$ and gradient mapping $\mathcal{G}(z) = \frac{1}{\alpha}(\mathcal{T}(z;\alpha) - z)$. When exact information is used, the finite termination of PGM can be easily argued by a lower bounded distance between subsequent iterates. Indeed, assume that $\alpha \le \frac{2}{L_f}$, then by (3) we obviously have: $(\mathcal{T}(z;\alpha) - \alpha \nabla f(\mathcal{T}(z;\alpha))) - (z - \alpha \nabla f(z)) \in \partial F(\mathcal{T}(z;\alpha))$. Based on this fact, we derive on one hand the following bound:

$$\begin{aligned}
&\|\mathcal{G}(z)\|^2 = \|\frac{1}{\alpha}(\mathcal{T}(z;\alpha) - z)\|^2 \\
&= \frac{1}{\alpha^2}\| (\mathcal{T}(z;\alpha) - \alpha \nabla f(\mathcal{T}(z;\alpha))) - (z - \alpha \nabla f(z)) \\
&\qquad\qquad + \alpha (\nabla f(\mathcal{T}(z;\alpha)) - \nabla f(z)) \|^2 \\
&= \|F'(\mathcal{T}(z;\alpha))\|^2 + \|\nabla f(\mathcal{T}(z;\alpha)) - \nabla f(z)\|^2 \\
&+ \frac{2}{\alpha}\langle \nabla f(z) - \nabla f(\mathcal{T}(z;\alpha)), (z - \alpha \nabla f(z)) - (\mathcal{T}(z;\alpha) - \alpha \nabla f(\mathcal{T}(z;\alpha)))\rangle \\
&\ge \|F'(\mathcal{T}(z;\alpha))\|^2 + \left(\frac{2}{\alpha L_f} - 1\right) \|\nabla f(\mathcal{T}(z;\alpha)) - \nabla f(z)\|^2 \\
&\overset{\alpha \le \frac{2}{L_f}}{\ge} \|F'(\mathcal{T}(z;\alpha))\|^2 \ge \sigma_F^2.
\end{aligned} \tag{6}$$

On the other hand, there are standard easy arguments that may be used to show that $\|\mathcal{G}(x^k)\|^2$ vanishes as $k \to \infty$. Thereby, the lower bound from (6) yields the finite termination of exact PGM. This conclusion has been emphasized in several previous works [1, 3, 2].

Given any $x \in \text{dom}F$, then $g(\cdot)$ denotes the approximation of accuracy δ of the gradient $\nabla f(\cdot)$ with respect to

$$\|\nabla f(x) - g(x)\| \le \delta. \tag{7}$$

In several works as [9, 15, 16, 8] similar or more restrictive approximation measures are promoted to analyze inexact first order methods. The basic Inexact Proximal Gradient Method (IPGM) iteration is shortly described below.

Algorithm 1: Inexact Proximal Gradient Method

Data: x^0, α, δ

while *stopping criterion* **do**

  Compute $g(x^k)$ such that: $\|g(x^k) - \nabla f(x^k)\| \leq \delta$

  Set $x^{k+1} := \text{prox}_{h,\alpha}(x^k - \alpha g(x^k))$

  $k := k + 1$

end

The accuracy δ also illustrates a bound on the perturbation of exact PGM iteration:

$$\begin{aligned} \|x^{k+1} - \mathcal{T}(x^k;\alpha)\| &= \|\text{prox}_{h,\alpha}(x^k - \alpha g(x^k)) - \mathcal{T}(x^k;\alpha)\| \\ &\leq \alpha \|g(x^k) - \nabla f(x^k)\| \leq \alpha\delta, \end{aligned} \tag{8}$$

obtained by the nonexpansiveness of prox operator (4).

Lemma 1. *Let f have Lipschitz gradient with constant L_f and also let $y \in \mathbf{R}^n, \delta, \alpha > 0$. Then the following inequalities holds:*

$$\begin{aligned} (i)\ & \|\mathit{prox}_{h,\alpha}(y - \alpha g(y)) - \mathcal{T}(y;\alpha)\| \leq \alpha\delta \\ (ii)\ & \|F'(\mathit{prox}_{h,\alpha}(y - \alpha g(y))) - F'(\mathcal{T}(y;\alpha))\| \leq (1 + L_f\alpha)\delta \\ (iii)\ & \frac{\alpha}{4}\|\mathcal{G}(y)\|^2 \leq F(y) - F(\mathit{prox}_{h,\alpha}(y - \alpha g(y))) + (2 + L_f\alpha)\alpha\delta^2. \end{aligned}$$

Proof. Denote $\mathcal{T}_\delta = \text{prox}_{h,\alpha}(y - \alpha g(y))$. The first part (i) can be derived as:

$$\begin{aligned} \|\mathcal{T}_\delta - \mathcal{T}(y;\alpha)\| &\leq \|\text{prox}_{h,\alpha}(y - \alpha g(y)) - \text{prox}_{h,\alpha}(y - \alpha\nabla f(y))\| \\ &\overset{(4)}{\leq} \|y - \alpha g(y) - [y - \alpha\nabla f(y)]\| \leq \alpha\delta. \end{aligned}$$

The second inequality is obtained by using triangle inequality and property (i):

$$\begin{aligned} &\|F'(\mathcal{T}_\delta) - F'(\mathcal{T}(y;\alpha))\| \leq \|\nabla f(\mathcal{T}_\delta) - \nabla f(\mathcal{T}(y;\alpha))\| + \|h'(\mathcal{T}_\delta) - h'(\mathcal{T}(y;\alpha))\| \\ &\leq L_f\|\mathcal{T}_\delta - \mathcal{T}(y;\alpha)\| + \|\nabla h_\alpha(y - \alpha g(y)) - \nabla h_\alpha(y - \alpha\nabla f(y))\| \\ &\overset{(i)}{\leq} L_f\alpha\delta + \frac{1}{\alpha}\|y - \alpha g(y) - (y - \alpha\nabla f(y))\| \leq (1 + L_f\alpha)\delta. \end{aligned}$$

For the last relation, we use the convexity property of F, the Cauchy-Schwartz inequality and the previous relations as follows:

$$
\begin{aligned}
F(\mathcal{T}_\delta) &\leq F(\mathcal{T}(y;\alpha)) + \langle F'(\mathcal{T}_\delta), \mathcal{T}_\delta - \mathcal{T}(y;\alpha)\rangle \\
&\leq F(\mathcal{T}(y;\alpha)) + \|F'(\mathcal{T}_\delta)\|\|\mathcal{T}_\delta - \mathcal{T}(y;\alpha)\| \\
&\overset{(i)+TI}{\leq} F(\mathcal{T}(y;\alpha)) + \left(\|F'(\mathcal{T}_\delta) - F'(\mathcal{T}(y;\alpha))\| + \|F'(\mathcal{T}(y;\alpha))\|\right)\alpha\delta \\
&\overset{(ii)+(6)}{\leq} F(\mathcal{T}(y;\alpha)) + \left[(1+L_f\alpha)\delta + \|\mathcal{G}(y)\|\right]\alpha\delta. \qquad (9)
\end{aligned}
$$

The well-known descent lemma [7] states that:

$$
F(\mathcal{T}(y;\alpha)) \leq F(y) - \frac{\alpha}{2}\|\mathcal{G}(y)\|^2 \qquad (10)
$$

Finally, by combining (9) and (10) in the last inequality we obtain part (iii):

$$
\begin{aligned}
F(\mathcal{T}_\delta) &\leq F(y) - \frac{\alpha}{2}\|\mathcal{G}(y)\|^2 + \left[(1+L_f\alpha)\delta + \|\mathcal{G}(y)\|\right]\alpha\delta \\
&\leq F(y) - \frac{\alpha}{4}\|\mathcal{G}(y)\|^2 + (2+L_f\alpha)\alpha\delta^2,
\end{aligned}
$$

where, in the last inequality we used $ab \leq \frac{a^2}{4} + b^2$. $\square$

Lemma 2. *Let $\{x^k\}_{k\geq 0}$ be the sequence generated by IPGM, then the following recurrent inequality holds:*

$$
dist_{X^*}(x^{k+1}) \leq dist_{X^*}(x^k) - \alpha\frac{F(\mathcal{T}(x^k;\alpha)) - F^*}{dist_{X^*}(x^k)} + \alpha\delta.
$$

Proof. On one hand, using the composite structure of F we easily get for any z:

$$
\begin{aligned}
\|\mathcal{T}(x^k;\alpha) - z\|^2 &= \|x^k - z\|^2 + 2\langle \mathcal{T}(x^k;\alpha) - x^k, \mathcal{T}(x^k;\alpha) - z\rangle - \|\mathcal{T}(x^k;\alpha) - x^k\|^2 \\
&= \|x^k - z\|^2 + 2\alpha\langle h'(\mathcal{T}(x^k;\alpha)), z - \mathcal{T}(x^k;\alpha)\rangle + 2\alpha\langle \nabla f(x^k), z - x^k\rangle \\
&\quad - 2\alpha\left(\langle \nabla f(x^k), \mathcal{T}(x^k;\alpha) - x^k\rangle + \frac{1}{\alpha}\|\mathcal{T}(x^k;\alpha) - x^k\|^2\right) \\
&\leq \|x^k - z\|^2 - 2\alpha\left(F(\mathcal{T}(x^k;\alpha)) - F(z)\right). \qquad (11)
\end{aligned}
$$

On the other hand, by the triangle inequality we simply derive:

$$
\begin{aligned}
\|x^{k+1} - z\| &\leq \|\mathcal{T}(x^k;\alpha) - z\| + \|\mathcal{T}(x^k;\alpha) - x^{k+1}\| \\
&\overset{\text{Lemma } 1(i)}{\leq} \|\mathcal{T}(x^k;\alpha) - z\| + \alpha\delta. \qquad (12)
\end{aligned}
$$

Finally, by taking $z = \pi_{X^*}(x^k)$, then:

$$\begin{aligned}
\text{dist}_{X^*}(x^{k+1}) \le \|x^{k+1} - \pi_{X^*}(x^k)\| &\overset{(12)}{\le} \|\mathcal{T}(x^k;\alpha) - \pi_{X^*}(x^k)\| + \alpha\delta \\
&\overset{(11)}{\le} \sqrt{\text{dist}^2_{X^*}(x^k) - 2\alpha\left(F(\mathcal{T}(x^k;\alpha)) - F^*\right)} + \alpha\delta \\
&\le \text{dist}_{X^*}(x^k)\sqrt{1 - 2\alpha\frac{F(\mathcal{T}(x^k;\alpha)) - F^*}{\text{dist}^2_{X^*}(x^k)}} + \alpha\delta \\
&\le \text{dist}_{X^*}(x^k)\left(1 - \alpha\frac{F(\mathcal{T}(x^k;\alpha)) - F^*}{\text{dist}^2_{X^*}(x^k)}\right) + \alpha\delta,
\end{aligned}$$

where in the last inequality we used the fact $\sqrt{1-2a} \le 1 - a$, which led to the above result. □

For simplicity, let us further denote $\underline{x}^k = \arg\min_{1\le i\le k} F(\mathcal{T}(x^{i-1};\alpha))$. We now estimate the convergence rate IPGM measured in the residual $\|\mathcal{G}(\cdot)\|$.

Theorem 3. *Let $\{x^k\}_{k\ge 0}$ be the sequence generated by IPGM and $\alpha > 0$ be the constant stepsize, then the following inequality holds:*

$$\min_{0\le i\le k}\|\mathcal{G}(x^i)\|^2 \le 12\frac{dist^2_{X^*}(x^0)}{(k^2-2)\alpha^2} + \frac{29}{2}\delta^2 \qquad \forall k > 2.$$

Proof. By using the same notations as in Lemma 2, we recall:

$$\text{dist}_{X^*}(x^{k+1}) \le \text{dist}_{X^*}(x^k) - \alpha\frac{F(\mathcal{T}(x^k;\alpha)) - F^*}{\text{dist}_{X^*}(x^k)} + \alpha\delta. \tag{13}$$

Notice that (13) do not guarantee a recurrent descent of the residual $\text{dist}_{X^*}(x^k)$ as it is usually proved for exact PGM. Thus, let

$$K := \min\{k \ge 0 \;:\; \text{dist}_{X^*}(x^{k+1}) \ge \text{dist}_{X^*}(x^k)\}$$

be the smallest index of the iteration when the residual stops descending. Then (13) has two consequences. First, for all $k < K$, by multiplying both sides with $\text{dist}_{X^*}(x^k)$ and using the descent then:

$$\alpha\left[F(\mathcal{T}(x^k;\alpha)) - F^*\right] \le \text{dist}^2_{X^*}(x^k) - \text{dist}^2_{X^*}(x^{k+1}) + \alpha\delta\text{dist}_{X^*}(x^0),$$

where we used $\text{dist}_{X^*}(x^k) \le \text{dist}_{X^*}(x^0)$. By summing over the history we obtain:

$$
\begin{aligned}
F(\underline{x}^k) - F^* &= \min_{1 \le i \le k} F(\mathcal{T}(x^{i-1};\alpha)) - F^* \\
&\le \frac{1}{k}\sum_{i=1}^{k} F(\mathcal{T}(x^{i-1};\alpha)) - F^* \\
&\le \frac{\text{dist}^2_{X^*}(x^0)}{k\alpha} + \delta \text{dist}_{X^*}(x^0) \quad \forall k < K. \qquad (14)
\end{aligned}
$$

Second, by definition of K we have $\text{dist}_{X^*}(x^K) \le \text{dist}_{X^*}(x^{K-1}) \le \cdots \le \text{dist}_{X^*}(x^0)$ and further (13) implies:

$$
F(\underline{x}^k) - F^* \le F(\mathcal{T}(x^K;\alpha_K)) - F^* \le \delta \text{dist}_{X^*}(x^K) \le \delta \text{dist}_{X^*}(x^0), \qquad (15)
$$

for any $k \ge K+1$. By unifying both cases (14)-(15) we conclude that:

$$
\begin{aligned}
F(\underline{x}^k) - F^* &\le \frac{\text{dist}^2_{X^*}(x^0)}{k\alpha} + \delta \text{dist}_{X^*}(x^0) \\
&\le \frac{3\text{dist}^2_{X^*}(x^0)}{2k\alpha} + \frac{\delta^2 k\alpha}{2} \quad \forall k \ge 1. \qquad (16)
\end{aligned}
$$

We further denote $i_k = \arg\min_{1 \le i \le k} F(\mathcal{T}(x^{i-1};\alpha))$. By applying Lemma 1$(iii)$ with $y := \underline{x}^k$ then:

$$
\begin{aligned}
\frac{\alpha}{4}\|\mathcal{G}(\underline{x}^k)\|^2 &\le F(\underline{x}^k) - F(\text{prox}_{h,\alpha}(\underline{x}^k - \alpha g(\underline{x}^k))) + (2 + L_f\alpha)\alpha\delta^2 \\
&= F(\underline{x}^k) - F(x^{i_k+1}) + (2 + L_f\alpha)\alpha\delta^2 \\
&\overset{\alpha \le 1/L_f}{\le} F(\underline{x}^k) - F(\underline{x}^{k+1}) + 3\alpha\delta^2.
\end{aligned}
$$

Now when we sum over $i = \lceil \frac{k}{2} + 1 \rceil \cdots k$ and obtain:

$$
\begin{aligned}
\frac{1}{k+1-\lceil k/2+1 \rceil}\sum_{i=\lceil k/2+1 \rceil}^{k} \|\mathcal{G}(\underline{x}^i)\|^2 &\le 8\frac{F(\underline{x}^{\lceil \frac{k}{2}+1 \rceil}) - F(\underline{x}^{k+1})}{(k-2)\alpha} + 12\delta^2 \\
&\le 8\frac{F(\underline{x}^{\lceil \frac{k}{2}+1 \rceil}) - F^*}{(k-2)\alpha} + 12\delta^2 \\
&\overset{(16)}{\le} 12\frac{\text{dist}^2_{X^*}(x^0)}{(k^2-2)\alpha^2} + \frac{29}{2}\delta^2.
\end{aligned}
$$

In the first inequality we use a similar argument as in [6, Sec. I].

□

Remark 4. *Note that similar sublinear convergence rates of inexact proximal gradient, with respect to the same inexactness criterion as our, was also derived in [16]. The authors provide* $\mathcal{O}(\frac{1}{k})$ *estimates in averaged function value* $F\left(\frac{1}{k}\sum_{i=0}^{k-1} x^i\right) - F^*$*, while our finite convergence analysis strongly requires a convergence rate in* $\|\mathcal{G}(x^k)\|$*. Moreover, our proof seems to be shorter and use completely different arguments.*

As a conclusion, Theorem 3 states that after at most:

$$\mathcal{O}\left(\frac{\text{dist}_{X^*}(x^0)}{\alpha\delta}\right)$$

iterations, the norm of the gradients decreases below δ.

Corollary 5. *Let* $\{x^k\}_{k\geq 0}$ *be the sequence generated by IPGM and the noise* $g(x^k) - \nabla f(x^k)$ *be possibly persistent but bounded by* $\delta = \|g(x^k) - \nabla f(x^k)\| < \frac{\sigma_F}{5}$*. Then after at most* $k = \mathcal{O}\left(\frac{dist_{X^*}(x^0)}{\alpha\sigma_F}\right)$ *iterations, the sequence* $\{x^k\}_{k\geq 0}$ *reaches the optimal set, i.e.* $x^k \in X^*$*.*

Proof. Theorem 3 implies that after $\frac{3\text{dist}_{X^*}(x^0)}{4\alpha\delta}$ iterations, then $\min\limits_{0\leq i\leq k}\|\mathcal{G}(x^i)\| \leq 5\delta < \sigma_F$. This fact clearly proves that there exists $x^i \in X^*$ for $i \leq k$. $\square$

In conclusion, a sufficiently low perturbation guarantees finite termination of IPGM even the noise is persistent. In the future, convergence to sharp minima of a wider class of first-order algorithms will be investigated, such as stochastic proximal algorithms. Also, extensions of the results to manifold identification would be of considerable interest.

Acknowledgement. The research of A. Pătraşcu was supported by a grant of the Romanian Ministry of Education and Research, CNCS - UEFISCDI, project number PN-III-P1-1.1-PD-2019-1123, within PNCDI III.

References

[1] AS Antipin. On finite convergence of processes to a sharp minimum and to a smooth minimum with a sharp derivative. *Differential Equations*, 30(11):1703–1713, 1994.

[2] James V Burke and Michael C Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

[3] Michael C Ferris. Finite termination of the proximal point algorithm. *Mathematical Programming*, 50(1):359–366, 1991.

[4] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

[5] Angelia Nedić and Dimitri P Bertsekas. The effect of deterministic noise in subgradient methods. *Mathematical programming*, 125(1):75–99, 2010.

[6] Y. Nesterov. How to make the gradients small. *Optima*, 88, 2012.

[7] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[8] Andrei Patrascu and Paul Irofti. Computational complexity of inexact proximal point algorithm for convex optimization under holderian growth. *arXiv preprint arXiv:2108.04482*, 2021.

[9] Andrei Patrascu and Ion Necoara. On the convergence of inexact projection primal first-order methods for convex minimization. *IEEE Transactions on Automatic Control*, 63(10):3317–3329, 2018.

[10] B.T. Polyak. A general method of solving extremal problems. *Math. Doklady*, 8:593–597, 1967.

[11] B.T. Polyak. Minimization of unsmooth functionals. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 9:509–521, 1969.

[12] B.T. Polyak. Nonlinear programming methods in the presence of noise. *Mathematical Programming*, 14:87–97, 1978.

[13] B.T. Polyak. Introduction to optimization. *Optimization Software, Inc., Publications Division, New York*, 1987.

[14] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[15] Saverio Salzo and Silvia Villa. Inexact and accelerated proximal point algorithms. *Journal of Convex analysis*, 19(4):1167–1192, 2012.

[16] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 1458–1466, 2011.