Convergence rate estimates for the gradient differential inclusion

Osman Güler *

November 2003

Abstract

Let $f : H \to \mathbb{R} \cup \{\infty\}$ be a proper, lower semi–continuous, convex function in a Hilbert space H. The gradient differential inclusion is $x'(t) \in -\partial f(x(t))$, $x(0) = x$, where $x \in \overline{\mathrm{dom}(f)}$. If f is differentiable, the inclusion can be considered as the continuous version of the steepest descent method for minimizing f on H. Even if f is not differentiable, the inclusion has a unique solution $\{x(t) : t > 0\}$ which converges weakly to a minimizer of f if such a minimizer exists. In general, the inclusion can be interpreted as the the continuous version of the proximal point method for minimization f on H. There is a remarkable similarity between the behavior of the inclusion and its discrete counterparts as well in the methods used in both cases. As a simple consequence of our previous results on the proximal point method, we prove the convergence rate estimate $f(x(t)) - f(u) \le (1/2t)||u-x||^2 - (1/2t)||u-x(t)||^2 - (t/2)||\partial f^0(x(t))||^2$, where $\partial f^0(x(t))$ is the least norm element of $\partial f(x(t))$. If f has a minimizer x^*, this implies $f(x(t)) - f(x^*) = O(1/t)$, a result due to Brézis. If $x(t)$ converges strongly to x^*, we give a better estimate $f(x(t)) - f(x^*) \le 1/\big(\int_0^t ||x(s) - x^*||^{-2} ds\big) = o(1/t)$.

Key words. Global convergence rate, subdifferential, nonlinear semigroups, strong convergence.
Abbreviated title: Convergence rate estimates for the gradient inclusion.
AMS(MOS) subject classifications: primary 90C25, 90C60, 52A41; secondary 90C06, 52A40.

*Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland 21250, USA. E-mail: *guler@math.umbc.edu*. Research partially supported by the National Science Foundation under grant DMS–0075722.

1 Introduction

Let H be a real Hilbert space. Consider the optimization problem

$$\min_{x \in H} f(x), \tag{1.1}$$

where $f : H \to \mathbb{R} \cup \{\infty\}$ is a proper, lower semicontinuous (lsc) convex function, following the terminology in Aubin and Ekeland [1].

Almost any convex optimization problem in H may be posed as (1.1). If f is a finite valued, differentiable function, then the steepest descent method is a basic algorithm which solves (1.1), although it is slow. In the general case, (1.1) may be a constrained optimization problem with a possibly non–differentiable objective function. In this case, the proximal point method, or its dual the augmented Lagrangian method may be used. In both cases, there exist global convergence rate estimates for the objective gap $f(x_k) - \min f$.

It is well known that the continuous version of the steepest descent and the proximal point methods is the differential inclusion

$$\frac{dx(t)}{dt} \in -\partial f(x(t)), \quad x(0) = x, \tag{1.2}$$

where $\partial f(x)$ is the subdifferential of at x, and $x \in \overline{\mathrm{dom}(\partial f)}$. Although it looks complicated, it is remarkable that (1.2) is very well behaved, and has many interesting properties. Moreover, the two discrete algorithms and the continuous inclusion share many properties. One consequence of this is that it is often possible to gain insight into the discrete algorithm by studying the differential inclusion, or vice versa.

It is known [4, 1] that there exists a unique solution $x(t)$ to (1.2) which is Lipschitz continuous on every finite interval $[0, T]$ and satisfies (1.2) for almost all $t \geq 0$. Moreover, $x(t)$ is everywhere differentiable from the right, and

$$\frac{d^+x(t)}{dt} = -\partial f^0(x(t)), \quad \forall t \geq 0,$$

where $\partial f^0(x)$ is the least norm element of $\partial f(x)$. For this reason, $x(t)$ is sometimes called the 'slow' solution of the differential inclusion (1.2). The solutions $x(t)$ for different starting points $x \in \overline{\mathrm{dom}(f)}$ generate a one–parameter family of *contractive nonlinear semigroups* $\{S(t)\}_{t \geq 0}$, where $S(t) : \overline{\mathrm{dom}(f)} \to \overline{\mathrm{dom}(f)}$ is given by $S(t)x = x(t)$. Moreover, if $t > 0$, we have the so called *smoothing property* that $S(t)x \in \mathrm{dom}(\partial f)$.

There are essentially two distinct approaches to establishing the existence and properties of the semigroup $S(t)$. The first and the older approach introduced by Kōmura [13] is through the study of the differential equations

$$\frac{dx_\lambda(t)}{dt} = -\partial f_\lambda(x_\lambda(t)), \quad x(0) = x, \tag{1.3}$$

for $\lambda > 0$ and where $\partial f_\lambda = \left(I - (I + \lambda \partial f)^{-1}\right)/\lambda$ is a Yosida approximation of ∂f. The operator ∂f_λ is single-valued and maximal monotone, being the subdifferential of the function

$$f_\lambda(x) = \inf_{z \in H} \left\{ f(z) + \frac{1}{2\lambda}||z - x||^2 \right\},$$

and (1.3) is an ordinary differential equation. Thus, the existence and properties of $x_\lambda(t)$ can be studied using classical methods. The solution $x(t)$ to (1.2) is obtained by $x(t) = \lim_{\lambda \to 0} x_\lambda(t)$. An excellent treatment is given in Brézis [4]; see also Aubin and Ekeland [1], pp. 396–400.

The more recent approach, introduced by Crandall and Liggett [9], is to study the existence and properties of $x(t)$ through the backward Euler (implicit) approximation of the differential inclusion:

$$\frac{x_k - x_{k-1}}{t_k - t_{k-1}} \in -\partial f(x_k), \quad x_0 \in H, \ (k \geq 1), \tag{1.4}$$

where $t_0 = 0$. In this approach one establishes the existence and properties of $x(t)$ by showing that if $x_0 \in \overline{\mathrm{dom}(f)}$, then the path $\{x_k\}_{k \geq 0}$ in (1.4) converges to $x(t)$ in (1.2) as the mesh of the discretization $\{t_k\}_{k \geq 0}$ converges to 0. See [8, 9, 12, 7, 14] and [18], pp. 445–465 for more details.

In optimization theory, the discrete path is called the *proximal* path and (1.4) is called the *proximal point algorithm*. Many convex optimization problems can be solved using (1.4), see Rockafellar [15].

In [10], the author studies the properties of the proximal point algorithm (1.4) and establishes the estimate

$$f(x_n) - f(u) \leq \frac{||u - x_0||^2}{2t_n} - \frac{||u - x_n||^2}{2t_n} - \frac{t_n}{2}||y_n||^2, \tag{1.5}$$

where $u \in H$ is arbitrary. He deduces some interesting consequences from this estimate. In particular, if f has a minimizer, x^*, then

$$f(x_n) - f(x^*) \leq \frac{d(x_0, X^*)^2}{2t_n},$$

where $d(x_0, X^*)$ is the distance of x_0 to X^*. This implies $f(x_n) - f(x^*) = O(1/t_n)$.

In the case where f has a minimizer, it is well known that $\{x_n\}_{n \geq 0}$ converges weakly to a minimizer of f, (see Brézis and Lions [5], Theorem 9), as does $x(t)$ (see Bruck [6]). The author [10] shows that x_n need not converge strongly to a minimizer of f, and thereby establishing analog of Baillon's [2] counter–example to the strong convergence of the continuous path $x(t)$. However, if x_n *does* converge strongly to a minimizer x^*, he proves that

$$f(x_n) - f(x^*) \leq \frac{3}{2\sum_{k=1}^{n} \lambda_k ||x_{k-1} - x^*||^{-2}}, \tag{1.6}$$

and that (1.6) implies $f(x_n) - f(x^*) = o(1/t_n)$.

The purpose of the present paper is to extend the estimates (1.5) and (1.6) to the continuous path $x(t)$. These are proved here as simple consequences of our main results in [10]. We also derive some consequences from these estimates.

2 Convergence rate estimates

In this section we prove the continuous analogues of estimates (1.5) and (1.6), and derive some interesting consequences from them. We start by proving the analogue of the estimate (1.5).

Theorem 2.1. *Let $x_0 \in \overline{\mathrm{dom}(f)} = \overline{\mathrm{dom}(\partial f)}$. Then for any $u \in H$ and $t > 0$ the solution $S(t)x = x(t)$, $x(0) = x_0$ to the differential inclusion* (1.2) *satisfies the estimate*

$$f(x(t)) - f(u) \leq \frac{||u - x(0)||^2}{2t} - \frac{||u - x(t)||^2}{2t} - \frac{t}{2}||\partial f^0(x(t))||^2. \tag{2.1}$$

Proof. Consider a partition of the interval $[0, T]$ into subintervals with lengths $\{\lambda_k\}_1^n$, where $\lambda_k = t_k - t_{k-1}$, $t_0 = 0$, and run the proximal point algorithm with $x_0 = x$ and stepsizes $\{\lambda_k\}$. Taking a limit as the mesh of the intervals converge to zero, the sequence $\{x_k\}_1^n$ converges to $\{x(t) : 0 \leq t \leq T\}$, see [10] Corollary 4.1, for example. Also, y_k weakly converges to $\partial^0 f(x)$, see [17]. It follows from Theorem 1, p. 120 in [18] that $||\partial^0 f(x)|| \leq \liminf ||y_k||$. These imply the estimate (2.1). □

For reader's convenience and for completeness, in the Appendix we include an independent proof of this theorem which is essentially due to Brézis.

Corollary 2.2. *For any $u \in H$ and $t > 0$ the following estimate holds.*

$$f(x(t)) - f(u) \leq \frac{||u - x(0)||^2}{2t} - \frac{t}{2}||\partial f^0(x(t))||^2.$$

Thus, $f(x(t)) \to f^* := \inf_{z \in H} f(z)$. If $f^* > -\infty$ then $\lim_{t\to\infty} \sqrt{t}||\partial f^0(x(t))|| = 0$. If f has a minimizer, then

$$f(x(t)) - f^* \le \frac{d(x_0, X^*)^2}{2t} - \frac{t}{2}||\partial f^0(x(t))||^2,$$

where X^* is the set of minimizers of f.

Corollary 2.3. *For any $t > s > 0$, the following estimates hold.*

$$\begin{aligned} f(x(t)) &\le f(x(0)) - \frac{||x(t) - x(0)||^2}{2t} - \frac{t}{2}||\partial f^0(x(t))||^2 \\ f(x(t)) &\le f(x(s)) - \frac{||x(t) - x(s)||^2}{2(t-s)} - \frac{t-s}{2}||\partial f^0(x(t))||^2. \end{aligned}$$

Proof. The first estimate follows directly from Theorem 2.1, and the second one follows from the first one by the semigroup property of $x(t)$, that is, the path $\overline{x}(t)$ generated by the differential inclusion (1.2) with the initial value $\overline{x}(0) = x(s)$, coincides with the path $\{x(t)\}_s^\infty$. □

Corollary 2.4. ***(Brézis)*** *Let $u \in \mathrm{dom}(\partial f)$. The following estimates hold.*

$$\begin{aligned} ||\partial f^0(x(t))|| &\le \frac{||x(t) - x(0)||}{t}, \\ ||\partial f^0(x(t))|| &\le ||\partial f^0(u)|| + \frac{||u - x(0)||}{t} \end{aligned} \tag{2.2}$$

Proof. The first inequality follows by substituting $x(t)$ for u in (2.1). It remains to prove the second inequality (2.2). It follows from the convexity of f that $f(x(t)) \ge f(u) + \langle \partial f^0(u), x(t) - u \rangle$, which implies that

$$f(u) - f(x(t)) \le ||\partial f^0(u)|| \cdot ||x(t) - u||.$$

Using this inequality in (2.1), we obtain

$$\begin{aligned} t^2||\partial f^0(x(t))||^2 &\le ||u - x(0)||^2 - ||u - x(t)||^2 + 2t||\partial f^0(u)|| \cdot ||x(t) - u|| \\ &\le ||u - x(0)||^2 - ||u - x(t)||^2 + \left(t^2||\partial f^0(u)||^2 + ||x(t) - u||^2\right) \\ &= ||u - x(0)||^2 + t^2||\partial f^0(u)||^2 \\ &\le \left(||u - x(0)|| + t||\partial f^0(u)||\right)^2. \end{aligned}$$

The corollary is proved. □

We now extend estimate (1.6) to the continuous case.

Theorem 2.5. *Suppose that f has a minimizer. Let $x_0 \in \overline{\mathrm{dom}(f)}$ and consider the continuous path $x(t)$ generated by the solution to the differential inclusion* (1.2) *with the initial value* $x(0) = x_0$. *If* $x(t)$ *converges strongly to a minimizer* x^* *of* f, *then for any* $t > 0$ *we have the estimate*

$$f(x(t)) - f(x^*) \leq \left(\int_0^t ||x(s) - x^*||^{-2} ds \right)^{-1}. \tag{2.3}$$

Therefore,

$$f(x(t)) - f(x^*) = o\left(\frac{1}{t}\right),$$

that is, $\lim_{t\to\infty} t\,(f(x(t)) - f(x^*)) = 0$.

Proof. As in Theorem 2.1, we can obtain (2.3), but the right hand side multiplied by 3/2, from the estimate (1.6) by a limit argument. We prefer to give a direct, independent proof. Without losing generality, we may assume that $f(x^*) = 0$ and $f(x(t)) > 0$ for all $t > 0$; otherwise there exists some $T > 0$ such that $f(x(t)) = 0$ for all $t \geq 0$, and the path $\{x(t)\}$ reaches the optimal set in finite time. We may also assume that $x_0 \in \text{dom}(f)$, since the case $x_0 \in \overline{\text{dom}(f)}$ can be obtained by a limit argument. Recall that $f(x(t))$ is convex, nonincreasing, and $d^+ f(x(t))/dt = -||d^+ x(t)/dt||^2$ for $t \geq 0$. This implies

$$\frac{d^+}{dt}\left(f(x(t))^{-1}\right) = \frac{||d^+ x(t)/dt||^2}{f(x(t))^2}. \tag{2.4}$$

Since f is convex, we have

$$0 = f(x^*) \geq f(x(t)) - \langle d^+ x(t)/dt, x^* - x(t)\rangle,$$

so that

$$||\frac{d^+ x(t)}{dt}|| \geq \frac{f(x(t))}{||x(t) - x^*||}.$$

Using this estimate in (2.4) gives

$$\frac{d^+}{dt}\left(f(x(t))^{-1}\right) \geq ||x(t) - x^*||^{-2}.$$

Integrating this inequality over the interval $[0, t]$, we get

$$f(x(t))^{-1} \geq f(x(t))^{-1} - f(x(0))^{-1} \geq \int_0^t ||x(s) - x^*||^{-2} ds,$$

which implies (2.3). Since $||x(t) - x^*|| \downarrow 0$, we have $||x(t) - x^*||^{-1} \uparrow \infty$. It is easy to show that

$$t^{-1} \int_0^t ||x(s) - x^*||^{-2} ds \uparrow \infty.$$

Consequently, $f(x(t)) - f(x^*) = o(1/t)$. □

3 Discussion

In the case f has a minimizer, we have seen that the trajectory $\{x(t)\}_0^\infty$ of the differential inclusion (1.2) converges *weakly* to a minimizer of f, and we have given a global convergence rate estimate $f(x(t)) - \min f = O(1/t)$. In the case $\{x(t)\}_0^\infty$ converges *strongly* to a minimizer of f, then we are able to obtain a better estimate $f(x(t)) - \min f = 0(1/t)$. However, Baillon's counter–example [2] shows that strong convergence is not always attainable. Recently Solodov and Svaiter [16] and Bauschke and Combettes [3] have modified the proximal point algorithm, by introducing suitable projection steps, to force strong convergence. It is an interesting topic to investigate the implications of these developments for the continuous case, the differential inclusion (1.2).

In [11], the author develops a new proximal point algorithm which has a better convergence rate than the classical proximal point algorithm. This suggests that there may exist a differential inclusion with a faster convergence rate than the inclusion (1.2). This seems to us a promising future research subject.

4 Acknowledgments

This research was partially supported by the National Science Foundation under grant DMS–0075722.

5 Appendix

Here we give a direct proof of Theorem 2.1. A careful look at this proof and the proof of Lemma 2.2 in [10] will reveal many similarities.

It is well known that $x(t) \in \text{dom}(\partial f)$ for any $t > 0$ (see Brézis [4]) so that estimate (2.1) is meaningful. We have, since f is convex and $d^+x(t)/dt = -\partial f^0(x(t))$ from (2.1),

$$\begin{aligned} f(x(t)) - f(u) &\leq \langle \partial f^0(x(t)), x(t) - u\rangle \\ &= \langle \frac{d^+x(t)}{dt}, u - x(t)\rangle \\ &= -\frac{1}{2}\frac{d^+}{dt}||x(t) - u||^2. \end{aligned}$$

Integrating this inequality over the interval $[0, t]$ gives

$$\int_0^t f(x(s))ds - tf(u) \le \frac{1}{2}||x(0)-u||^2 - \frac{1}{2}||x(t)-u||^2. \qquad (5.1)$$

It is known (see Brézis [4], Theorem 3.2) that for $t \ge 0$, the function $t \mapsto f(x(t))$ is convex, non–increasing, and that $d^+f(x(t))/dt = -||d^+x(t)/dt||^2$. Since $||d^+x(s)/ds||$ is nonincreasing (see Brézis [4], Theorem 3.1), we have

$$\begin{aligned}\int_0^t s\frac{d^+}{ds}f(x(s))ds &= -\int_0^t s||d^+x(s)/ds||^2 ds \\ &\le -||d^+x(t)/dt||^2 \int_0^t sds \\ &= -\frac{t^2}{2}||\partial f^0(x(t))||^2.\end{aligned}$$

Integrating by parts the first term above, we obtain

$$\int_0^t s\frac{d^+}{ds}f(x(s))ds = tf(x(t)) - \int_0^t f(x(s))ds.$$

Thus,

$$tf(x(t)) - \int_0^t f(x(s))ds \le -\frac{t^2}{2}||\partial f^0(x(t))||^2.$$

Adding (5.1) and the above inequality yields

$$tf(x(t)) - tf(u) \le \frac{1}{2}||x(0)-u||^2 - \frac{1}{2}||x(t)-u||^2 - \frac{t^2}{2}||\partial f^0(x(t))||^2,$$

and the theorem follows.

References

[1] J.-P. Aubin and I. Ekeland. *Applied nonlinear analysis*. Pure and Applied Mathematics. John Wiley & Sons Inc., New York, 1984. A Wiley-Interscience Publication.

[2] J.-B. Baillon. Un exemple concernant le comportement asymptotique de la solution du problème $du/dt + \partial\varphi(u) \ni 0$. *J. Funct. Anal.*, 28(3):369–376, 1978.

[3] H. H. Bauschke and P. L. Combettes. A weak-to-strong convergence principle for Fejér-monotone methods in Hilbert spaces. *Math. Oper. Res.*, 26(2):248–264, 2001.

[4] H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland Publishing Co., Amsterdam, 1973. North-Holland Mathematics Studies, No. 5. Notas de Matemática (50).

[5] H. Brézis and P.-L. Lions. Produits infinis de résolvantes. *Israel J. Math.*, 29(4):329–345, 1978.

[6] R. E. Bruck, Jr. Asymptotic convergence of nonlinear contraction semigroups in Hilbert space. *J. Funct. Anal.*, 18:15–26, 1975.

[7] M. G. Crandall. Nonlinear semigroups and evolution governed by accretive operators. In *Nonlinear functional analysis and its applications, Part 1 (Berkeley, Calif., 1983)*, volume 45 of *Proc. Sympos. Pure Math.*, pages 305–337. Amer. Math. Soc., Providence, RI, 1986.

[8] M. G. Crandall and L. C. Evans. On the relation of the operator $\partial/\partial s + \partial/\partial \tau$ to evolution governed by accretive operators. *Israel J. Math.*, 21(4):261–278, 1975.

[9] M. G. Crandall and T. M. Liggett. Generation of semi–groups on nonlinear transformations on general Banach spaces. *Amer. J. Math.*, 93:265–298, 1971.

[10] O. Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM J. Control Optim.*, 29(2):403–419, 1991.

[11] O. Güler. New proximal point algorithms for convex minimization. *SIAM J. Optim.*, 2(4):649–664, 1992.

[12] Y. Kobayashi. Difference approximation of Cauchy problems for quasi-dissipative operators and generation of nonlinear semigroups. *J. Math. Soc. Japan*, 27(4):640–665, 1975.

[13] Y. Kōmura. Nonlinear semi–groups in Hilbert space. *J. Math. Soc. Japan*, 19:493–507, 1967.

[14] N. H. Pavel. *Nonlinear evolution operators and semigroups*, volume 1260 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1987. Applications to partial differential equations.

[15] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.*, 1(2):97–116, 1976.

[16] M. V. Solodov and B. F. Svaiter. Forcing strong convergence of proximal point iterations in a Hilbert space. *Math. Program.*, 87(1, Ser. A):189–202, 2000.

[17] L. Veron. Some remarks on the convergence of approximate solutions of nonlinear evolution equations in Hilbert spaces. *Math. Comp.*, 39:325–337, 1982.

[18] Kōsaku Yosida. *Functional analysis*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.