Journal of Convex Analysis
Volume 16 (2009), No. 1, 277–286

Asymptotic Convergence to the Optimal Value of Diagonal Proximal Iterations in Convex Minimization

Juan Peypouquet*

Departamento de Matemática, Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile
juan.peypouquet@usm.cl

Received: January 25, 2007
Revised manuscript received: April 9, 2008

Given an approximation $\{f_n\}$ of a given objective function f, we provide simple and fairly general conditions under which a diagonal proximal point algorithm approximates the value $\inf f$ at a reasonable rate. We also perform some numerical tests and present a short survey on finite convergence.

Introduction

Consider a function f to be minimized over a real Hilbert space H. Its effective domain D_f corresponds to the *feasible set*, and is given by constraints on the minimization problem. Let $\{f_n\}$ be a family of proper, lower-semicontinuous convex functions with a common domain D in H. The functions f_n are meant to approximate the *objective function* f in some sense. Since we do not require $D_f = D$, this accounts for several approximation, regularization, interior and exterior penalization schemes (see below). We assume that $\inf_H f = \inf_D f \geq -\infty$ and denote this quantity by f^*.

Take a starting point $x_0 \in H$ and two sequences of real parameters $\{\lambda_k\} \subset (0, \Lambda]$ and $\{\varepsilon_k\} \subset [0, \infty)$. A sequence $\{x_k\} \subset D$ is said to be an *inexact diagonal proximal sequence* generated by $(x_0, \{\lambda_k\}, \{f_k\}, \{\varepsilon_k\})$ if

$$y_k := \frac{x_{k-1} - x_k}{\lambda_k} \in \partial_{\varepsilon_k} f_k(x_k) \tag{1}$$

for all $k \geq 1$, where the approximate $\varepsilon-$subdifferential ∂_ε is defined by

$$\partial_\varepsilon g(u) = \{u^* \in H \mid g(v) \geq g(u) + \langle u^*, v - u\rangle - \varepsilon \ \forall v \in H\} \quad \varepsilon \geq 0.$$

In order to construct such a sequence it is necessary to solve approximately an optimization problem at each iteration, which is usually done by means of another algorithm (e.g. bundle methods) so the use of $\varepsilon-$subdifferentials is important for practical purposes.

If $f_n \equiv f$ and $\varepsilon_n \equiv 0$, this is the standard *proximal point algorithm*, introduced in [21] for solving variational inequalities. It has been used extensively in convex minimization (see [13] and [23]) in the search for the value f^* and, if there are any, the points at which it is

*Partly supported by MECESUP grant UCH0009.

ISSN 0944-6532 / $ 2.50

attained (also useful to find zeros of monotone operators as in [10] and [19]). For weak and strong convergence of the sequence $\{x_n\}$ see [24] and [7]. Several variations can be found in [14], [12] (relaxation) and [25] (hybrid projection-proximal point algorithm).

The term *diagonal* refers to the fact that the objective function is updated at each iteration, which is not a new approach. Many authors have combined approximation methods with the proximal point algorithm. The common feature is that a family $\{f_n\}$ is chosen so that the functions are better behaved in terms of smoothness, computability, coerciveness, etc. and converge to f in some sense. A pioneer work with interior penalties is [16] (see also [17]). For exterior penalties and variational convergence see [1]. For Tikhonov see [22]. For computation and bundle methods, see [4] and [5]. In the latter, a sequence of exterior penalties is considered and their algorithm is proved to converge in a finite number of steps in the linear case. In [2], [8] and [9] the authors study exponential penalty and log-barrier in linear programming using properties of the limit (unperturbed) dynamics.

Convergence to the optimal value is important on its own right but further information on this convergence is useful for several other reasons. First, when there are no minimizers or the convergence is only weak, there is no point in analyzing the whole sequence x_n. But even when the sequence x_n is known to converge, the rate at which $|f_n(x_n) - f^*|$ tends to 0 gives (along with the duality gap) good stopping rules, especially in large-scale problems.

Using simple estimations, we derive qualitative and quantitative convergence results for the values $f_n(x_n)$ provided $f_n \to f$ in a certain way. Unlike in the references cited above, our results do not depend explicitly on the type of approximation. Instead, we investigate on the minimal underlying structure that guarantees that one can approximate f^* by $f_n(x_n)$ at a reasonable rate. Some of our results generalize the corresponding ones in the classical paper [13], where $f_n \equiv f$. This paper is organized as follows: in Section 1, we study convergence to the optimal value f^*. The results hold even if the objective function has no minimizers and is not coercive (*cf* [1]). Section 2 deals with rates of convergence. When f has minimizers the relationship between f_n and the rate of convergence of the values becomes evident. A simple (academic) numerical example is presented. The case of summable stepsizes and algorithm termination are commented in Section 3.

1. Convergence of the values $f_n(x_n)$

Set $\sigma_n = \sum_{k=1}^n \lambda_k$ and $\sigma_0 = 0$. Unless otherwise stated we assume $\lim_{n\to\infty} \sigma_n = \infty$. The Kronecker Theorem (see, for example, [18], p. 129.) implies that if a sequence $\{p_n\}$ of positive numbers is in ℓ^1, then $\lim_{n\to\infty} \frac{1}{\sigma_n} \sum_{k=1}^n \sigma_k p_k = 0$. For a sequence $\{z_n\}$ denote by $\overline{z_n} = \frac{1}{\sigma_n} \sum_{k=1}^n \lambda_k z_k$ the sequence of its averages weighted by the stepsizes $\{\lambda_k\}$. Clearly $\lim_{n\to\infty} z_n = L$ implies $\lim_{n\to\infty} \overline{z_n} = L$. Let $\{p_n\}$ and $\{\varrho_n\}$ be two positive sequences. We write $p_n = \mathcal{O}(\varrho_n)$ if p_n/ϱ_n is bounded. Then $p_n = \mathcal{O}(\varrho_n)$ implies $\overline{p_n} = \mathcal{O}(\overline{\varrho_n})$. If $\lim_{n\to\infty} p_n/\varrho_n = 0$ we write $p_n = o(\varrho_n)$.

Lemma 1.1. *For any $u \in D$ we have*

$$\overline{f_n(x_n)} - \overline{f_n(u)} \leq \frac{\|u - x_0\|^2}{2\sigma_n} - \frac{1}{2\sigma_n} \sum_{k=1}^n \lambda_k^2 \|y_k\|^2 - \frac{\|u - x_n\|^2}{2\sigma_n} + \overline{\varepsilon_n}.$$

Proof. Let $u \in D$. The subdifferential inequality gives

$$f_k(u) - f_k(x_k) \geq \frac{1}{\lambda_k}\langle x_{k-1} - x_k, u - x_k\rangle - \varepsilon_k \tag{2}$$

and so

$$\begin{aligned} 2\lambda_k(f_k(u) - f_k(x_k)) &\geq 2\langle x_{k-1} - x_k, u - x_k\rangle - 2\lambda_k\varepsilon_k \\ &= \|x_{k-1} - x_k\|^2 + \|u - x_k\|^2 - \|u - x_{k-1}\|^2 - 2\lambda_k\varepsilon_k \\ &= \lambda_k^2\|y_k\|^2 + \|u - x_k\|^2 - \|u - x_{k-1}\|^2 - 2\lambda_k\varepsilon_k. \end{aligned} \tag{3}$$

Summing for $k = 1, \ldots n$ we get

$$2\sum_{k=1}^{n}\lambda_k f_k(u) - 2\sum_{k=1}^{n}\lambda_k f_k(x_k) \geq \sum_{k=1}^{n}\lambda_k^2\|y_k\|^2 + \|u - x_n\|^2 - \|u - x_0\|^2 - 2\sum_{k=1}^{n}\lambda_k\varepsilon_k. \tag{4}$$

Dividing by $2\sigma_n$ we obtain the desired inequality. □

Let $\mathcal{S} = \operatorname{Argmin} f$. For $r > 0$ define $\mathcal{S}_r = \{u \in D \cap D_f | f(u) \leq f^* + r\}$ whenever $f^* > -\infty$ and $\mathcal{S}_r = \{u \in D \cap D_f | f(u) \leq -r\}$ otherwise. Consider the following hypothesis:

Hypothesis $\mathbf{H_1}$: $\limsup_{n\to\infty} \overline{f_n(u)} \leq f(u)$ for all $u \in \mathcal{S}_r$ and some $r > 0$.

Finally set $f_n^* = \inf f_n$. An immediate consequence of the previous lemma is the following:

Corollary 1.2. *Let* $\lim_{n\to\infty}\overline{\varepsilon_n} = 0$ *and assume hypothesis* $\mathbf{H_1}$ *holds. We have*

i) $\limsup_{n\to\infty}\overline{f_n(x_n)} \leq f^*$;
ii) *If* $f^* \leq \liminf_{n\to\infty}\overline{f_n^*}$, *then* $\lim_{n\to\infty}\overline{f_n(x_n)} = f^*$;
iii) *If* $f^* \leq \liminf_{n\to\infty} f_n^*$, *then* $\liminf_{n\to\infty} f_n(x_n) = f^*$.

This result is similar to Theorem 2.1 in [6], though the hypotheses here are weaker and more concise. Also, we show three different "degrees" in the conclusion, which help understand more clearly the effect of averaging. The sequence $\overline{f_n(x_n)}$ approximates the optimal value f^* under fairly weak hypotheses on the sequence f_n and the errors ε_n.

Remark 1.3. Observe that in the special case where there is $x^* \in \mathcal{S} \cap D$ we have

$$\overline{f_n(x_n)} - f^* \leq \frac{\|x^* - x_0\|^2}{2\sigma_n} + \left[\overline{f_n(x^*)} - f^*\right] + \overline{\varepsilon_n}.$$

Suppose for simplicity that $f^* \leq \overline{f_n^*}$ so that the left-hand side is nonnegative. The rate of convergence of $\overline{f_n(x_n)}$ to f^* depends on three parameters, namely:

- σ_n, which seems to be intrinsic to the proximal iterations;
- ε_n, which has to do with computational precision;
- the behavior of f_n on the minimizing set $\mathcal{S}$.

If $\sum\lambda_k\varepsilon_k < \infty$ and $\overline{f_n(x^*)} - f^* = \mathcal{O}(\frac{1}{\sigma_n})$ for some $x^* \in \mathcal{S}$, then $\overline{f_n(x_n)} - f^* = \mathcal{O}(\frac{1}{\sigma_n})$. This speed of convergence is provided by Theorem 2.1 in [13] where $f_n \equiv f$ and $\varepsilon_n \equiv 0$. □

In order for the values $\{f_n(x_n)\}$ to converge we need further hypotheses on the way the sequence $\{f_n\}$ evolves:

Hypothesis $\mathbf{H_2}$: There exist a set $K \subseteq H$ containing the trajectory $\{x_n\}$ and a nonnegative sequence $\{a_n\}$ such that $f_n(x) \leq f_{n-1}(x) + a_{n-1}$ for all $n \geq 2$ and all $x \in K$.

The fact that the set K is chosen according to $\{x_n\}$ plays a key role in practice since it is possible to get *a priori* estimations for the sequence in many applications. For example, the verification of hypothesis $\mathbf{H_2}$ is considerably simplified under uniform coerciveness or inf-compactness assumptions on the family $\{f_n\}$. Hypothesis $\mathbf{H_2}$ holds trivially if the sequence $\{f_n\}$ is decreasing; which happens, for instance, for Tikhonov's approximation or penalization schemes as described in [9] using the log- or the inverse barrier. Hypothesis $\mathbf{H_2}$ is also true if the sequence converges uniformly.

Lemma 1.4. *Assume hypothesis* $\mathbf{H_2}$ *holds. For every* $u \in D$ *we have*

$$f_n(x_n) - \overline{f_n(u)} \leq \frac{\|u - x_0\|^2}{2\sigma_n} + \frac{1}{\sigma_n}\sum_{k=1}^{n} \sigma_k(a_k + \varepsilon_k) - \frac{1}{2\sigma_n}\sum_{k=1}^{n} \lambda_k(\sigma_{k-1} + \sigma_k)\|y_k\|^2 - \frac{\|u - x_n\|^2}{2\sigma_n}.$$

Proof. Multiply (2) by σ_{k-1} with $u = x_{k-1}$ and use hypothesis $\mathbf{H_2}$ to obtain

$$\sigma_{k-1}f_{k-1}(x_{k-1}) - \sigma_k f_k(x_k) + \lambda_k f_k(x_k) + \sigma_{k-1}a_{k-1} \geq \sigma_{k-1}\lambda_k\|y_k\|^2 - \sigma_{k-1}\varepsilon_k.$$

Summing for $k = 1, \ \ldots \ n$ we get

$$-\sigma_n f_n(x_n) + \sum_{k=1}^{n} \lambda_k f_k(x_k) + \sum_{k=2}^{n} \sigma_{k-1}a_{k-1} \geq \sum_{k=2}^{n} \sigma_{k-1}\lambda_k\|y_k\|^2 - \sum_{k=2}^{n} \sigma_{k-1}\varepsilon_k.$$

Adding twice this inequality to (4) we get

$$\begin{aligned} & 2\sum_{k=1}^{n} \lambda_k f_k(u) - 2\sigma_n f_n(x_n) + 2\sum_{k=2}^{n} \sigma_{k-1}a_{k-1} \\ \geq \ & 2\sum_{k=2}^{n} \sigma_{k-1}\lambda_k\|y_k\|^2 + \sum_{k=1}^{n} \lambda_k^2\|y_k\|^2 + \|u - x_n\|^2 - \|u - x_0\|^2 - 2\sum_{k=1}^{n} \sigma_k\varepsilon_k \\ = \ & \sum_{k=1}^{n} \lambda_k(\sigma_k + \sigma_{k-1})\|y_k\|^2 + \|u - x_n\|^2 - \|u - x_0\|^2 - 2\sum_{k=1}^{n} \sigma_k\varepsilon_k. \end{aligned}$$

Dividing by $2\sigma_n$ and rearranging the terms we obtain the desired inequality. □

In particular if $\mathbf{H_2}$ holds we have

$$f_n(x_n) - \overline{f_n(u)} \leq \frac{\|u - x_0\|^2}{2\sigma_n} + \frac{1}{\sigma_n}\sum_{k=1}^{n} \sigma_k(a_k + \varepsilon_k). \tag{5}$$

Rearranging the terms and using $\mathbf{H_2}$ once more we get

$$-\sum_{k=n}^{\infty} a_k \leq f_n(x_n) - f^* \leq \frac{1}{\sigma_n}\sum_{k=1}^{n} \sigma_k(a_k + \varepsilon_k) + \frac{\|u - x_0\|^2}{2\sigma_n} + \left[\overline{f_n(u)} - f^*\right]. \tag{6}$$

The first inequality in (6) gives

$$\liminf f_n(x_n) \geq f^*. \tag{7}$$

We shall prove that the sequence $f_n(x_n)$ approaches the optimal value f^* without any further assumptions. Convergence of the values $f_n(x_n)$ has been proved in a monotone setting (see Theorem 3.1 in [1] and Corollary 3.1 in [20]) or assuming that $\mathcal{S} \neq \emptyset$, that $f_n \to f$ in the sense of Mosco[1] and that $\{\lambda_n\}$ is bounded away from zero (see Theorems 3.2 and 3.3 in [1] and Theorem 2.2 in [6]).

Proposition 1.5. *Assume hypotheses* $\mathbf{H_1}$ *and* $\mathbf{H_2}$ *hold. If the sequences* $\{a_n\}$ *and* $\{\varepsilon_n\}$ *are in* ℓ^1 *then* $\lim_{n\to\infty} f_n(x_n) = f^*$. *If moreover* $f^* > -\infty$ *then*

$$i)\quad \lim_{n\to\infty} \overline{\sigma_n \|y_n\|^2} = 0; \qquad ii)\quad \lim_{n\to\infty} \frac{\|x_n\|^2}{\sigma_n} = 0; \quad \text{and} \quad iii)\quad \sum_{k=1}^{\infty} \lambda_k \|y_k\|^2 < \infty.$$

Proof. Recall from (7) that $\liminf_{n\to\infty} f_n(x_n) \geq f^*$. Take $r > 0$ and let $n \to \infty$ in inequality (5), so that hypothesis $\mathbf{H_1}$ and the Kronecker Theorem give $\limsup_{n\to\infty} f_n(x_n) \leq f(u)$ for all $u \in \mathcal{S}_r$. Hence $\limsup_{n\to\infty} f_n(x_n) \leq f^*$ and $\lim_{n\to\infty} f_n(x_n) = f^*$. Now assume $f^* > -\infty$.

i) For $\eta > 0$ take x_η such that $f(x_\eta) < f^* + \eta$. Using $\mathbf{H_1}$, the estimation given by Lemma 1.4 with $u = x_\eta$ and the fact that $\lim_{n\to\infty} f_n(x_n) = f^*$ we get

$$\limsup_{n\to\infty} \frac{1}{\sigma_n} \sum_{k=1}^{n} \lambda_k \sigma_k \|y_k\|^2 \leq f(x_\eta) - f^* < \eta.$$

ii) The proof is similar.

iii) Setting $u = x_{k-1}$ in (2) and using hypothesis $\mathbf{H_2}$ we find

$$\lambda_k \|y_k\|^2 \leq f_{k-1}(x_{k-1}) - f_k(x_k) + a_{k-1} + \varepsilon_k$$

for $k \geq 2$. Summing for $k = 2 \dots n$ we get

$$\sum_{k=2}^{n} \lambda_k \|y_k\|^2 \leq f_1(x_1) - f_n(x_n) + \sum_{k=2}^{n} a_{n-1} + \sum_{k=1}^{n} \varepsilon_k.$$

From the first part, the right-hand side converges as $n \to \infty$. □

Part *i*) implies $\lim_{n\to\infty} \overline{\|y_n\|^2} = 0$ and so $\lim_{n\to\infty} \overline{\|y_n\|} = 0$.[2] On the other hand, part *iii*) reveals that if $\lambda_n \geq \lambda > 0$ then $\lim_{n\to\infty} \|y_n\| = 0$.

Remark 1.6. Under epi-convergence, every cluster point of $\{x_n\}$ is a minimizer of f. More precisely, let τ be the strong or the weak topology in H. In addition to the hypotheses of Proposition 1.5, suppose that for all $x \in D_f$ and for all sequences $\{z_n\}$ in D converging to x for the topology τ we have $f(x) \leq \liminf_{n\to\infty} f_n(z_n)$. Then every $\tau-$cluster point of the proximal sequence $\{x_n\}$ is a minimizer of f. A similar result has been proved in [1] under different hypotheses including boundedness of the sequence $\{x_n\}$. □

[1]Epi-convergence both for the weak and the strong topologies (see below).

[2]Let $\psi : \mathbf{R}_+ \to \mathbf{R}_+$ be a continuous function with $\psi(0) = 0$ and let $\{\alpha_n\}$ be a sequence of positive numbers. Assume either that $\{\alpha_n\}$ is bounded or that ψ satisfies $\psi(x) \leq M(1 + x)$ for some $M > 0$. Then $\overline{\alpha_n} \to 0$ implies $\overline{\psi(\alpha_n)} \to 0$. This can be applied to $\psi(x) = \sqrt{x}$.

2. Some remarks on the rate of convergence

It is natural to expect the rate of convergence of $f_n(x_n)$ to depend on the way the sequence f_n converges to f. When the function f has minimizers we can give precise estimates on the rate of convergence of the values in terms of the rate of pointwise convergence of the sequence $\{f_n\}$ on the optimal set. The results that we present here do not depend explicitly on the type of approximation. Moreover, it is important to underscore the fact that the function f need not be coercive. We only require the set $\mathcal{S} \cap D$ to be nonempty. This is the case for exterior penalties and finite (but not infinite) barriers. Convergence may be slow when the jump is infinite.

Proposition 2.1. *Suppose* $\mathbf{H_2}$ *holds and there is* $x^* \in \mathcal{S} \cap D$ *with* $|\overline{f_n(x^*)} - f^*| = \mathcal{O}(\frac{1}{\sigma_n})$. *If* $\sum_{k=1}^{n} \sigma_k(a_k + \varepsilon_k) < \infty$ *then* $|f_n(x_n) - f^*| = \mathcal{O}(\frac{1}{\sigma_n})$. *Moreover,* $\sum_{k=1}^{\infty} \lambda_k \sigma_k \|y_k\|^2 < \infty$ *and the sequence* $\{x_n\}$ *is bounded.*

Proof. It suffices to apply inequality (6) noticing that $\sum_{k \geq n} a_k = \frac{1}{\sigma_n} \sum_{k \geq n} \sigma_n a_k \leq \frac{1}{\sigma_n} \sum_{k \geq n} \sigma_k a_k$. The rest follows easily from Lemma 1.4. □

Remark 2.2. In [13] the author studies the case $f_n \equiv 0$, $\varepsilon_n \equiv 0$ and proves that if the proximal sequence x_n happens to converge strongly to some $x^* \in \mathcal{S}$ then $f(x_n) - f^* = o(\frac{1}{\sigma_n})$. For a sequence f_n this need not be the case. However, it *is* the case if there is sufficiently fast convergence on $\mathcal{S}$. Indeed, from the subdifferential inequality one has

$$f_n(x^*) \geq f_n(x_n) + \frac{1}{\lambda_n} \langle x_{n-1} - x_n, x^* - x_n \rangle - \varepsilon_n.$$

Thus $f_n(x_n) - f^* \leq \|y_n\| \cdot \|x^* - x_n\| + [f_n(x^*) - f^*] + \varepsilon_n$. Assume the sequence f_n is bounded from below by f^* and that x_n happens to converge strongly to some $x^* \in \mathcal{S} \cap D$ such that $|f_n(x^*) - f^*| = o(\frac{1}{\sigma_n})$ (this depends only on the rate of approximation). If $\|y_n\| = \mathcal{O}(\frac{1}{\sigma_n})$ and $\varepsilon_n = o(\frac{1}{\sigma_n})$, then $|f_n(x_n) - f^*| = o(\frac{1}{\sigma_n})$. □

Observe that by virtue of Theorem 9 in [7], when $f_n \equiv 0$ and $\varepsilon_n \equiv 0$ we always have $\|y_n\| = \mathcal{O}(\frac{1}{\sigma_n})$, so that Theorem 3.1 in [13] is a consequence of Remark 2.2 above. It is important to point out that the proof in [13] uses a clever (but unnecessarily sophisticated) argument instead of the trivial observation $\|y_n\| = \mathcal{O}(\frac{1}{\sigma_n})$. In general, it is difficult to prove that $\|y_n\| = \mathcal{O}(\frac{1}{\sigma_n})$ because the speed depends strongly on the evolution of the sequence $\{f_n\}$ so one has to study each particular type of approximation.

Remark 2.3. Some comments are in order:

(1) If $\lambda_n \sigma_n \geq \lambda > 0$ then $\lim_{n \to \infty} \|y_n\| = 0$; and if $\lambda_n \geq \lambda > 0$, then $\lim_{n \to \infty} \sigma_n \|y_n\|^2 = 0$, thus $\|y_n\| = o(\frac{1}{\sqrt{\sigma_n}})$.

(2) From part *i*) we have $\overline{\sigma_n \|y_n\|^2} = \mathcal{O}(\frac{1}{\sigma_n})$, which implies that $\overline{\sigma_n^2 \|y_n\|^2}$ is bounded and so is $\overline{\sigma_n \|y_n\|}$. This does not fulfill the hypothesis in Remark 2.2, but is close.

(3) The rate of approximation is crucial for the boundedness of the sequence $\{x_n\}$. To see this, take $f \equiv 0$ on $D = [0, \infty)$ and $f_n(x) = \delta_n e^{-x}$, where the sequence $\{\delta_n\}$ decreases to 0. It is easy to verify that this approximation satisfies all our assumptions and if $\delta_n \to 0$ sufficiently slow, we will have $x_n \to \infty$. □

Example 2.4. Let $g_1, \dots, g_m \in \Gamma_0(H)$. A standard way to approximate $f = \max\{g_i\}_{i=1}^m$ is to set $F(u,\eta) = \eta \log \sum_{i=1}^m \exp\left(\frac{g_i(u)}{\eta}\right)$ for $\eta > 0$. If $f_n(u) = F(u,\eta_n)$ with $\eta_n \to 0$, then f_n decreases to f and Hypothesis $\mathbf{H_2}$ holds with $a_n \equiv 0$. If $x^* \in \mathcal{S}$ and $\sum \lambda_k \eta_k < \infty$, then $|\overline{f_n(x^*)} - f^*| = \mathcal{O}(\frac{1}{\sigma_n})$. Proposition 2.1 yields $|f_n(x_n) - f^*| = \mathcal{O}(\frac{1}{\sigma_n})$.

Let $g_1, g_2, g_3 \in \Gamma_0(\mathbf{R}^2)$ be given by $g_1(x,y) = x + 2y - 1$, $g_2(x,y) = 2x - y$ and $g_3(x,y) = 1 - x$. For $f = \max\{g_1, g_2, g_3\}$ we have $\mathcal{S} = \{(1/2, 1/2)\}$ and $f^* = 1/2$. Since the hypotheses of Corollary 1.6 also hold we must have $x_n \to (1/2, 1/2)$. Strong convergence then implies $\sigma_n f_n(x_n) \to 0$ as $n \to \infty$ by virtue of Theorem 2.2 provided $\eta_n \to 0$ fast enough.

Take $\lambda_n \equiv 1$, $\eta_n = 1/n^2$ and $x_0 = (0,0)$. Figure 2.1 shows the first 30 iterations of the algorithm, Figure 2.2 shows $f_n(x_n)$ vs n and Figure 2.3 shows $\sigma_n(f_n(x_n) - f^*)$ vs n. Figure 2.4, where we have $\sigma_n^2(f_n(x_n) - f^*)$ vs n, suggests that this quantity converges as $n \to \infty$. □

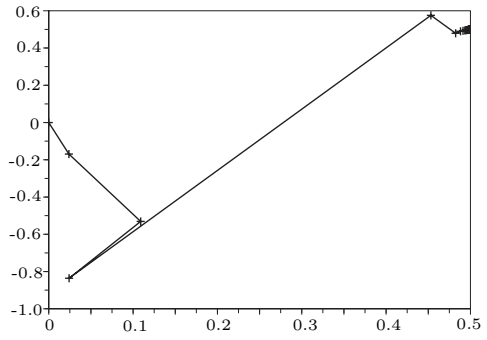

Figure 2.1

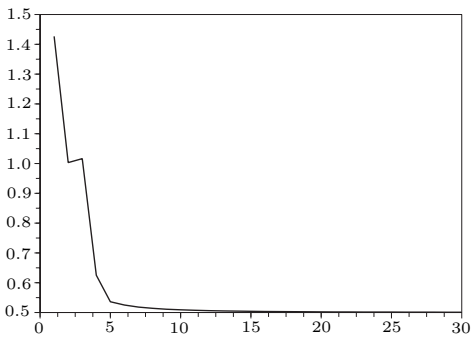

Figure 2.2

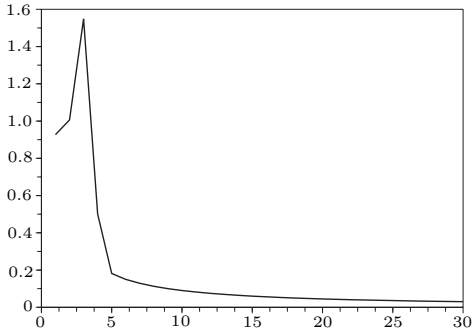

Figure 2.3

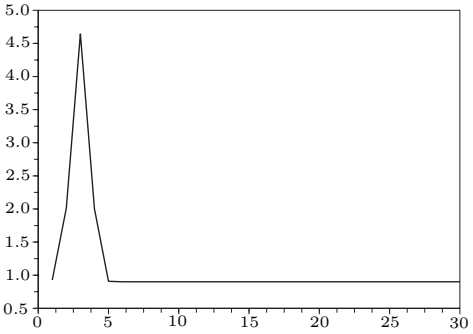

Figure 2.4

Remark 2.5. A different condition ensuring that the sequence of velocities converges to zero and does not depend on the choice of the stepsizes is the following: Assume there exists a sequence θ_n of nonnegative numbers such that for all $u_n^* \in \partial f_n(u_n)$ and $u_{n+1}^* \in \partial f_{n+1}(u_{n+1})$ we have

$$\langle\, u_{n+1}^* - u_n^*, u_{n+1} - u_n \,\rangle \geq -\theta_n \|u_{n+1} - u_n\|. \tag{8}$$

If hypotheses $\mathbf{H_1}$ and $\mathbf{H_2}$ hold, $f^* > -\infty$, the sequences $\{a_n\}$, $\{\varepsilon_n\}$ and $\{\theta_n\}$ are in ℓ^1 then $\lim_{n\to\infty} \|y_n\| = 0$. Indeed, inequality (8) implies $\|y_{n+1}\| \leq \|y_n\| + \theta_n$. Since $\{\theta_n\} \in \ell^1$ and $\|y_n\| \geq 0$ the limit $\lim_{n\to\infty} \|y_n\|$ exists. It must be 0 because $\lim_{n\to\infty} \overline{\|y_n\|} = 0$ by Proposition 1.5. An approximation fulfilling (8) is the following: Let g be a convex $\mathcal{C}^1$

function whose gradient is bounded by a constant B on D. Take a sequence $\{r_n\}$ of positive numbers and define $f_n = f + r_n g$. Then (8) is satisfied for $\theta_n = B|r_n - r_{n-1}|$. Also, if $\{g_n\}$ is a family of convex $\mathcal{C}^1$ functions such that $\{\nabla g_n\}$ converges uniformly, one can take $\theta_n = \|\nabla g_n - \nabla g_{n-1}\|_\infty$. □

3. Summable stepsizes and algorithm termination

For a moment assume $f_n \equiv f$ and $\varepsilon_n \equiv 0$. If the sequence $\{\lambda_n\}$ of stepsizes is in ℓ^1 ($\sigma_n \to \sigma < \infty$), the trajectory $\{x_n\}$ always converges strongly to a point x_∞ (observe that $\|x_n - x_m\| \leq \|y_1\| \cdot |\sigma_n - \sigma_m|$), no matter whether the function has minimizers. Even when the minimizing set is nonempty, if the distance between the initial point x_0 and the minimizing set $\mathcal{S}$ is greater than $\sigma\|y_1\|$, then x_∞ cannot be a point in $\mathcal{S}$. However, one would like to know whether or not it is possible to attain the minimizing set if the initial condition is close enough to $\mathcal{S}$ (alternatively, if σ is large enough). We give a partial answer in terms of the smoothness of the objective function around $\mathcal{S}$, which we assume to be nonempty. The following example includes the case where f is a differentiable function having a L-Lipschitz gradient in some η-neighborhood of $\mathcal{S}$ (consider the maximal monotone operator $A = \nabla f$):

Example 3.1. Let A be a maximal monotone operator on H. Set $\mathcal{S} = A^{-1}0$ and let $d(u, \mathcal{S})$ denote the distance from u to $\mathcal{S}$. Assume there exist $\eta > 0$ and $L > 0$ such that if $d(u, \mathcal{S}) < \eta$, then $\sup_{v \in Au} \|v\| \leq L\, d(u, \mathcal{S})$. It is clear that the minimizing set cannot be attained in a finite number of steps. If $d(x_N, \mathcal{S}) < \eta$ for some N, the proximal iteration yields $\|x_n - x_{n-1}\| \leq \lambda_n L d(x_n, \mathcal{S})$ for all $n > N$. Since $\|x_n - x_{n-1}\| \geq d(x_{n-1}, \mathcal{S}) - d(x_n, \mathcal{S})$ we have $(1 + \lambda_n L) d(x_n, \mathcal{S}) \geq d(x_{n-1}, \mathcal{S})$. Therefore

$$d(x_n, \mathcal{S}) \geq \left[\prod_{k=N+1}^{n} (1 + \lambda_k L)^{-1}\right] d(x_N, \mathcal{S}) \geq e^{-\sigma_n L} d(x_N, \mathcal{S}).$$

If $\sigma_n \to \sigma < \infty$, the sequence $\{x_n\}$ stays away from $\mathcal{S}$. □

On the other hand, if the function f is somehow *pointed* on $\partial\mathcal{S}$ the sequence $\{x_n\}$ will converge provided σ is large enough. To see this, let P denote the projection onto the nonempty closed convex set $\mathcal{S}$, take $r > 0$ and consider the following conditions:

(A) $\|v\| \geq r$ for all $v \in \partial f(x)$ such that $x \notin \mathcal{S}$;
(B) $f(x) \geq f^* + r\|x - Px\|$ for all $x \in D$; and

If (B) holds, then $f(x) \geq f^* + r\|x - Px\| \geq f(x) + \langle v, Px - x\rangle + r\|x - Px\|$ by the subdifferential inequality. Hence $r\|x - Px\| \leq \langle v, x - Px\rangle$ and we get (A).

Proposition 3.2. *Let $f_n \equiv f$, $\{\varepsilon_n\} \in \ell^1$ and $\sigma_n \to \sigma < \infty$.*

i) If (A) *holds for some $r > 0$ and $r^2\sigma > f(x_0) - f^* + \sum \varepsilon_k$, then there are $N \in \mathbf{N}$ and $x^* \in \mathcal{S}$ such that $x_n = x^*$ for all $n \geq N$.*

ii) If (B) *holds for some $r > 0$ and $r^2\sigma = f(x_0) - f^* + \sum \varepsilon_k$, then $\lim_{n\to\infty} d(x_n, \mathcal{S}) = 0$. If the sequence $\{\|y_n\|\}$ is bounded,[3] then x_n converges to a point in $\mathcal{S}$ as $n \to \infty$.*

Proof. If $x_k \notin \mathcal{S}$ for $k = 1, \ldots, n$, then $f(x_{k-1}) - f(x_k) + \varepsilon_k \geq \lambda_k \|y_k\|^2 \geq r^2 \lambda_k$ by (A). Summing up one gets $r^2\sigma_n \leq f(x_0) - f(x_n) + \sum \varepsilon_k$.

[3]For instance, if $\sum(\varepsilon_n + \varepsilon_{n-1})\lambda_n^{-1} < \infty$.

i) If $x_n \notin \mathcal{S}$ for all $n \in \mathbf{N}$ then $r^2\sigma \leq f(x_0) - f^* + \sum \varepsilon_k$, which is a contradiction.

ii) We have $\limsup_{n\to\infty} f(x_n) \leq f(x_0) - r^2\sigma + \sum \varepsilon_k = f^*$, so $\lim_{n\to\infty} f(x_n) = f^*$. By (B) we get $\lim_{n\to\infty} \|x_n - Px_n\| = 0$. Finally, if $\{\|y_n\|\}$ is bounded, then $\{x_n\}$ is a Cauchy sequence. Clearly its limit must be in $\mathcal{S}$. □

Condition (B) above was used in [11] to prove convergence in a finite number of steps. In the particular case where $\mathcal{S} = \{x^*\}$, it is equivalent to $\overline{B}(0,r) \subseteq \partial f(x^*)$, which is the assumption used in [24] with the same purpose in that specific case. In the cited works, H is assumed to be $\mathbf{R}^n$ and the sequence $\{\lambda_n\}$ to be bounded from below by a positive constant. The following example suggests that the hypotheses above are close to optimal.

Example 3.3. Let H be separable and define $H_m = \text{span}\{e_1, \ldots, e_m\}$, where $\{e_k\}$ is an orthonormal basis for H. The function

$$f(x) = \begin{cases} \sum \dfrac{1}{k} x^k & \text{if } x^k = \langle x, e_k \rangle \geq 0 \text{ for all } k \\ \infty & \text{otherwise} \end{cases}$$

satisfies condition (B) on each H_m with $r = 1/m$. Since $\cup H_m$ is dense in H, f is not too far from satisfying condition (B). The k-th component of the term x_n satisfies $(x_n)^k \geq (x_{n-1})^k - \lambda_n/k \geq (x_0)^k - \sigma_n/k$ and so $\liminf_{n\to\infty}(x_n)^k \geq (x_0)^k - \sigma/k$. If x_0 is selected so that $\sup_k k(x_0)^k = \infty$ we will have $\liminf_{n\to\infty}(x_n)^k > 0$ for some k no matter how large σ is. Hence x_n does not converge to 0, the unique minimizer of f. The initial condition x_0 can be chosen as small as desired because $\sup_k k(\varepsilon x_0)^k = \varepsilon \sup_k k(x_0)^k$. □

Remark 3.4. If $\{f_k\}$ is not constant, a similar argument leads us to

$$\sum_{k=1}^{n} \lambda_k \|y_k\|^2 \leq f_0(x_0) - f_n(x_n) + \sum_{k=1}^{n} (\varepsilon_k + a_{k-1}),$$

but no lower bound can be obtained here for $\|y_k\|$ because x_k can be a minimizer of f_k. □

References

[1] P. Alart, B. Lemaire: Penalization in non-classical convex programming via variational convergence, Math. Program., Ser. A 51 (1991) 307–331.

[2] F. Alvarez, R. Cominetti: Primal and dual convergence of a proximal point exponential penalty method for linear programming, Math. Program. 93 (2002) 87–96.

[3] H. Attouch, R. Cominetti: A dynamical approach to convex minimization coupling approximation with the steepest descent method, J. Differ. Equations 128 (1996) 519–540.

[4] A. Auslender: Numerical methods for nondifferentiable convex optimization, Math. Program. Study 30 (1987) 102–127.

[5] A. Auslender, J.-P. Crouzeix, P. Fedit: Penalty-proximal methods in convex programming, J. Optimization Theory Appl. 55 (1987) 1–21.

[6] M. A. Bahraoui, B. Lemaire: Diagonal bundle methods for convex minimization, unpublished.

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

[8] R. Cominetti: Coupling the proximal point algorithm with approximation methods, J. Optimization Theory Appl. 95 (1997) 581–600.

[9] R. Cominetti, M. Courdurier: Coupling general penalty schemes for convex programming with the steepest descent and the proximal point algorithm, SIAM J. Optim. 13 (2002) 745–765.

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

[11] M. Ferris: Finite termination of the proximal point algorithm, Math. Program., Ser. A 50 (1991) 359–366.

[12] E. G. Gol'shtein: A method for the modification of monotone mappings, Ehkon. Mat. Metody 11 (1975) 1144–1159.

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

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

[15] O. Güler: Convergence rate estimates for the gradient differential inclusion, Optim. Methods Softw. 20 (2005) 729–735.

[16] A. Kaplan: On a convex programming method with internal regularization, Sov. Math., Dokl. 19 (1978) 795–799.

[17] A. Kaplan, R. Tichatschke: Proximal methods in view of interior-point strategies, J. Optimization Theory Appl. 98 (1998) 399–429.

[18] K. Knopp: Theory and Application of Infinite Series, Blackie & Son, London (1951).

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

[20] B. Lemaire: About the convergence of the proximal method, Advances in Optimization (Lambrecht, 1991), W. Oettli et al. (ed.), Lect. Notes Econ. Math. Syst. 382, Springer, Berlin (1992) 39–51.

[21] B. Martinet: Régularisation d'inéquations variationnelles par approximations successives, Rev. Franç. Inform. Rech. Opér., Sér R-3 4 (1970) 154–158.

[22] A. Moudafi: Coupling proximal algorithm and Tikhonov method, Nonlinear Times Dig. 1 (1994) 203–210.

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

[24] R. T. Rockafellar: Monotone operators and the proximal point algorithm, SIAM J. Control Optimization 14 (1976) 877–898.

[25] M. Solodov, B. Svaiter: A hybrid projection-proximal point algorithm, J. Convex Analysis 6 (1999) 59–70.