Control and Cybernetics

vol. **31** (2002) No. 3

Optimizing properties of an inertial dynamical system with geometric damping. Link with proximal methods

by

H. Attouch, J. Bolte and P. Redont

ACSIOM-CNRS FRE 2311, Département de Mathématiques, case 51,
Université Montpellier II, Place Eugène Bataillon,
34095 Montpellier cedex 5, France

Abstract: The second-order dynamical system $\ddot{x} + \alpha\dot{x} + \beta\nabla^2\Phi(x)\dot{x} + \nabla\Phi(x) = 0$, $\alpha > 0$, $\beta > 0$, where the Hessian $\nabla^2\Phi(x)$ acts as a geometric damping, is introduced, mainly in view of the minimization of Φ. Minimizing Φ is a problem equivalent to the minimization of the functional $\Psi_{a,b}(x,y) = \frac{1}{b^2}\Phi(x) + \frac{1}{2}|ax + by|^2$, $a > 0$, $b > 0$. The latter naturally appears in the proximal regularization of Φ; it may also be viewed as an energy. The continuous steepest descent method applied to $\Psi_{a,b}$ yields a first-order system, which proves to be equivalent to the above-mentioned second-order system, when Φ is of class $\mathcal{C}^2$.

Keywords: dynamical systems in optimization, proximal regularization method, steepest descent method, entropic methods in optimization.

1. Introduction

Let H be a real Hilbert space and $\Phi : H \to \mathbb{R} \cup \{+\infty\}$ a proper, lower semicontinuous, convex function. Consider the convex minimization problem

$$(\mathcal{P}) \qquad \inf\{\Phi(x) : x \in H\}$$

and let $S := \operatorname{argmin} \Phi$ denote the solution set of $(\mathcal{P})$.

In relation with $(\mathcal{P})$, we wish to introduce a new dynamical system, called (DIN), which naturally arises and enjoys remarkable properties in convex optimization (its range of applications is much wider indeed). When Φ is a smooth $\mathcal{C}^2$ function, (DIN) assumes the following form

$$\text{(DIN)} \qquad \ddot{x}(t) + \alpha\dot{x}(t) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0$$

This dynamical system can be viewed from different perspectives.

The second derivative $\ddot{x}(t)$ (which induces inertial effects) may be considered as a singular perturbation, and in fact regularization, of the possibly degenerate classical continuous Newton dynamical system

$$\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0.$$

That is the origin of the terminology: (DIN) stands in short for Dynamical Inertial Newton-like system.

The system (DIN) also naturally derives from the Heavy Ball with Friction dynamical system (see Poliak, 1987, Antipin, 1994, Attouch-Goudou-Redont, 2000)

$$\text{(HBF)} \qquad \ddot{x}(t) + \alpha\dot{x}(t) + \nabla\Phi(x(t)) = 0.$$

The damping term $\alpha\dot{x}(t)$ confers optimizing properties on (HBF), but it acts isotropically and ignores the geometry of Φ. Adding a geometric damping term like $\beta\nabla^2\Phi(x(t))\dot{x}(t)$ puts down the possible oscillations of the trajectories and gives rise to (DIN).

Lastly, the system (DIN) is closely related to the minimization of the function

$$(x, y) \in H \times H \mapsto \psi(x, y) = \Phi(x) + \frac{1}{2\lambda}|x - y|^2$$

where λ is some fixed positive parameter. Indeed, the Continuous Steepest Descent method applied to ψ yields

$$\begin{cases} \dot{x}(t) + \nabla\Phi(x(t)) + \frac{1}{\lambda}(x(t) - y(t)) = 0 \\ \dot{y}(t) \qquad\qquad\qquad + \frac{1}{\lambda}(y(t) - x(t)) = 0. \end{cases}$$

Eliminating y, we obtain the following (DIN) system

$$\ddot{x}(t) + \frac{2}{\lambda}\dot{x}(t) + \nabla^2\Phi(x(t))\dot{x}(t) + \frac{1}{\lambda}\nabla\Phi(x(t)) = 0.$$

Introducing the function ψ is no contrived idea, since it naturally appears in two circumstances at least.

First, the proximal regularization method applied to $(\mathcal{P})$ (see Moreau, 1965, Martinet, 1972, Rockafellar, 1976) is nothing else than the iterated minimization of ψ alternatively with respect to the x and y variable. This point of view is set out in Section 2..

The function ψ appears, though, as well in the study of the following discrete analogue of the (HBF) system

$$\frac{1}{\qquad}(x_{\qquad} - x_{\qquad}) + \frac{\lambda}{\qquad}\nabla\Phi(x_{\qquad}) = 0$$

which is obtained by discretizing (HBF) with $\sqrt{\lambda}$ as a time-step. The investigation of the sequence (x_i) owes much to the discrete energy function $\psi(x_{i+1}, x_i) = \Phi(x_{i+1}) + \frac{1}{2}|\frac{x_{i+1}-x_i}{\sqrt{\lambda}}|^2$, the exact replica of the energy $\Phi(x(t)) + \frac{1}{2}|\dot{x}(t)|^2$ in the continuous case. By the way, in recent years there has been an increasing interest in studying the interaction between discrete and continuous dynamical systems in variational analysis and optimization (see Alvarez-Attouch, 2001, Attouch-Teboulle, to appear, Flam-Horvath, 1996, Antipin, 1994, Polyak, 1987, Lemaire, 1996, Cominetti, 1997).

The guideline of our introduction of (DIN) as a tool in optimization is the method of proximal regularization, which permits to solve general convex minimization problems with the help of well-posed convex minimization problems (without degeneracy of the conditioning).

2. From proximal regularization to (DIN)

In many situations of practical importance, the minimization problem $(\mathcal{P})$ is not well-posed, see for example Dontchev and Zolezzi (1993) for a thorough exposition of the notions of well-posedness and the presentation of various situations occurring in mathematical programming, calculus of variations, statistics, control theory, inverse problems, where well-posedness fails to be satisfied.

To regularize the problem $(\mathcal{P})$, a fruitful idea is to add a positive definite quadratic term, typically $\varepsilon|x|^2$, to $\Phi(x)$. This leads to various methods, like the Tikhonov approximation method, but in that case the conditioning becomes worse and worse as the approximation parameter ε goes to zero. By contrast, proximal regularization methods allow to preserve the conditioning away from zero.

The basic idea which lies behind the proximal methods is the following: take some $x^* \in S = \operatorname{argmin} \Phi$ and some $\lambda > 0$. Then, consider the minimization problem

$$(\mathcal{P}_*) \qquad \min\{\Phi(x) + \frac{1}{2\lambda}|x - x^*|^2 : x \in H\}.$$

Clearly, $(\mathcal{P}_*)$ is a well-posed convex minimization problem with x^* as unique solution and $\inf(\mathcal{P}) = \inf(\mathcal{P}_*)$. Unfortunately, this method is not constructive, since it makes use of some $x^* \in S$, which is unknown. Nevertheless, from a theoretical point of view, this method has proved to be quite fruitful. It was used by Barbu (1981) in the optimal control of variational inequalities, then Lions (1983) made a systematic use of it in the study of singular distributed control problems, in order to obtain optimality conditions.

The proximal method, which has been developed for numerical purposes consists in solving $(\mathcal{P}_*)$ not as a minimization problem (which is impossible, x^*

the minimization problem

$$(\mathcal{P}_y) \qquad \min\{\Phi(x) + \frac{1}{2\lambda}|x-y|^2 : x \in H\},$$

whose unique solution is denoted by $J_\lambda^\Phi(y)$. Clearly, x^* is a solution of $(\mathcal{P})$ if and only if $J_\lambda^\Phi(x^*) = x^*$. Taking advantage of $J_\lambda^\Phi(y)$ being a contraction (indeed, a firmly nonexpansive mapping), the proximal method consists in solving this fixed point problem by the successive approximation method. One obtains the following classical algorithm

$$(\mathcal{P}_k) \qquad \begin{array}{l} x_0 \text{ given} \\ x_k \to x_{k+1} = \operatorname{argmin}\{\Phi(x) + \frac{1}{2\lambda_k}|x-x_k|^2 : x \in H\}. \end{array}$$

This method, first introduced by Martinet (1972) in convex optimization, has been developed in a general framework by Rockafellar (1976) (see Lemaire, 1996, for a thorough exposition and further references). When writing the optimality condition for $(\mathcal{P}_k)$ one obtains

$$\lambda_k^{-1}(x_{k+1} - x_k) + \partial\Phi(x_{k+1}) \ni 0$$

which can be interpreted as the implicit discretization of the generalized continuous steepest descent method

$$\dot{x}(t) + \partial\Phi(x(t)) \ni 0.$$

Note that, in this continuous-discrete interaction, the property $\sum_{k=1}^{+\infty} \lambda_k = +\infty$ corresponds to $t \to +\infty$ (since $x(t_k) = x_k$, and $\lambda_k = t_{k+1} - t_k$). It is a remarkable property that both systems (discrete and continuous) enjoy a very similar asymptotical behaviour. In both cases, with Opial lemma one can prove that the trajectories converge weakly in H to an optimal solution. In the continuous case, this result has been obtained by Bruck (1975).

Let us notice, too, that the continuous dynamical system allows to treat parabolic PDEs like (nonlinear) heat equations, see Brézis (1973).

Let us now come to the original aspect of our approach. To that end, let us give a different formulation of the proximal regularization method. We are going to interpret it as a relaxation method applied to an energy-like function. Indeed, as we have already observed, the function of two variables

$$\begin{array}{l} \psi : H \times H \mapsto \mathbb{R} \cup \{+\infty\} \\ (x,y) \mapsto \psi(x,y) := \Phi(x) + (2\lambda)^{-1}|x-y|^2 \end{array}$$

plays the central role in the above results. In order to get some flexibility we introduce two other parameters:

DEFINITION 2.1 *Let $a, b \in \mathbb{R}$ be two real parameters, with $b \neq 0$. We define*

$$\psi_{a,b} : H \times H \mapsto \mathbb{R} \cup \{+\infty\}$$

by the following formula

$$\psi_{a,b}(x,y) = \Phi(x) + \frac{1}{2}|ax + by|^2.$$

It is called the energy function attached to the convex minimization problem $(\mathcal{P})$, *(with parameters* a *and* b*). The energy minimization problem* $(\mathcal{P}_{a,b})$ *is defined by*

$$(\mathcal{P}_{a,b}) \qquad \inf\{\Phi(x) + \frac{1}{2}|ax + by|^2 : (x,y) \in H\}.$$

Let us notice that we are not in the classical perturbation theory for convex problems since $\psi_{a,b}(x,0) = \Phi(x) + \frac{1}{2}|ax|^2$ is not equal to the original function Φ (unless $a = 0$). Let us make precise the connection between $(\mathcal{P}_{a,b})$ and $(\mathcal{P})$.

PROPOSITION 2.1 *For any values of* $a, b \in \mathbb{R}$, $b \neq 0$ *the following equalities hold:*

(i) $\inf\{\Phi(x) : x \in H\} = \inf\{\psi_{a,b}(x,y) : (x,y) \in H \times H\}$.

(ii) *If* x^* *is an optimal solution of* $(\mathcal{P})$, *then* $(x^*, -\frac{a}{b}x^*)$ *is an optimal solution of* $(\mathcal{P}_{a,b})$.

(iii) *Conversely, if* $(x^*, y^*) \in H \times H$ *is an optimal solution of* $(\mathcal{P}_{a,b})$, *then* $y^* = -\frac{a}{b}x^*$, *and* x^* *is an optimal solution of* $(\mathcal{P})$.

Proof. The statements are easy consequences of the following facts

$$\forall (x,y) \in H \times H, \Phi(x) \leq \psi_{a,b}(x,y),$$
$$\Phi(x) = \psi_{a,b}(x,y) \Leftrightarrow y = -\frac{a}{b}x. \quad \square$$

As a consequence, solving $(\mathcal{P})$ is equivalent to solving $(\mathcal{P}_{a,b})$. Note that $(\mathcal{P}_{a,b})$ is only partially well-conditioned. It is not globally well-conditioned because of the direction $y = -\frac{a}{b}x$, along which the quadratic form is degenerate.

Indeed, the strategy of the proximal method consists in minimizing $\psi_{a,b}$ by using a relaxation method making only use of directions, along which $(\mathcal{P}_{a,b})$ is well-conditioned, namely the x- and y-subspaces. Let us make this precise in the following statement

PROPOSITION 2.2 *The proximal method is the relaxation minimization method applied to* $\psi_{a,-a}$, *for* $a = \frac{1}{\sqrt{\lambda}}$. *More precisely*

$$(x_k, y_k = x_k) \rightharpoonup (x_{k+1}, y_{k+1}) : x_{k+1} = \operatorname{argmin}\{\psi_{a,-a}(x, y_k) : x \in H\}$$
$$y_{k+1} = \operatorname{argmin}\{\psi_{a,-a}(x_{k+1}, y) : y \in H\}.$$

Proof. By definition of the proximal method, by taking $a^2 = \frac{1}{\lambda}$

$$x_{k+1} = \operatorname{argmin}\{\Phi(x) + \frac{1}{2\lambda}|x - x_k|^2 : x \in H\}$$

since $y_k = x_k$. Next, when considering y_{k+1} as

$$y_{k+1} = \operatorname{argmin}\{\Phi(x_{k+1}) + \frac{a^2}{2}|x_{k+1} - y|^2 : y \in H\}$$

one clearly gets $y_{k+1} = x_{k+1}$. And so on. ■

From the numerical point of view it is tempting to minimize $\psi_{a,b}$ using better descent directions than those *a priori* given by the x and y directions. A natural candidate is the steepest descent method. Let us decribe it when it is applied to $\psi_{a,b}$. Indeed, it is convenient to consider the function

$$\Psi_{a,b}(x,y) = \frac{1}{b^2}\Phi(x) + \frac{1}{2}|ax + by|^2$$

in order to obtain a quite simple formulation (note that replacing Φ by $\frac{1}{b^2}\Phi$ does not change anything to the minimization problem $(\mathcal{P})$ and, like $\psi_{a,b}$, $\Psi_{a,b}$ may be called an energy associated to Φ).

THEOREM 2.1 *Let $\Phi : H \mapsto \mathbb{R} \cup \{+\infty\}$ be a convex, lower semicontinuous, proper function. Let a, b be real constants with $b \neq 0$.*

a) *The generalized continuous steepest descent method when applied to*

$$\Psi_{a,b}(x,y) = \frac{1}{b^2}\Phi(x) + \frac{1}{2}|ax + by|^2$$

provides the following system (energetical steepest descent)

$$\text{(ESD)} \qquad \begin{cases} \dot{x}(t) + \frac{1}{b^2}\partial\Phi(x(t)) + a[ax(t) + by(t)] \ni 0 & \textit{(ESD1)} \\ \dot{y}(t) + b[ax(t) + by(t)] = 0 & \textit{(ESD2)} \end{cases}$$

b) *When Φ is a smooth C^2 function, and $a \neq 0$, the above system (ESD) can equivalently be written (by eliminating the variable y)*

$$\ddot{x}(t) + (a^2 + b^2)\dot{x}(t) + \frac{1}{b^2}\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0. \tag{1}$$

c.1) *For any initial condition $x_0 \in \overline{\operatorname{dom}\Phi}$ and $y_0 \in H$, there exists a unique solution (x,y) of (ESD) in the following sense*

. *$x : [0,+\infty[\mapsto H$ is a continuous function, with $x(t) \in \operatorname{dom}\Phi$ $\forall t > 0$, Lipschitz continuous on $[\delta,+\infty[$ for every $\delta > 0$,*

. *$y : [0,+\infty[\mapsto H$ is a C^1 function, with a Lipschitz continuous derivative on $[\delta,+\infty[$ for every $\delta > 0$,*

. *(ESD1) is satisfied almost everywhere on $]0,+\infty[$,*

. *(ESD2) is satisfied for every $t \in]0,+\infty[$,*

. *$x(0) = x_0$ and $y(0) = y_0$.*

c.2) *As $t \to +\infty$, $\Phi(x(t))$ converges to $\inf\Phi$, whether the latter be finite or not.*

c.3) *If $S = \operatorname{argmin}\Phi \neq \emptyset$, then x and y weakly converge as $t \to +\infty$: $x(t) \overset{w-H}{\longrightarrow} x_\infty \in S$ and $y(t) \overset{w-H}{\longrightarrow} -\frac{a}{b}x_\infty$.*

Proof. a) The generalized continuous steepest descent (see Brézis, 1973) applied to $\Psi_{a,b}$ reads

$$(\dot{x}(t), \dot{y}(t)) + \partial\Psi_{a,b}(x(t), y(t)) \ni 0. \tag{2}$$

Making the inclusion above explicit yields (ESD).

b) Let (x, y) be a solution of (ESD). Since Φ is $\mathcal{C}^2$ we have

$$\dot{x} + \frac{1}{b^2}\nabla\Phi(x) + a[ax + by] = 0 \tag{3}$$

$$\dot{y} + b[ax + by] = 0. \tag{4}$$

Differentiate (3) to get

$$\ddot{x} + a^2\dot{x} + \frac{1}{b^2}\nabla^2\Phi(x)\dot{x} + ab\dot{y} = 0. \tag{5}$$

Perform a linear combination of (3), (4), (5) with b^2, $-ab$, 1 as coefficients to obtain (1).

Conversely, let x satisfy (1): define y by (3), which is legal since $ab \neq 0$. Differentiating (3) yields (5) as above. Perform a linear combination of (1), (3), (5) with -1, b^2, 1 as coefficients to obtain $ab\dot{y} + ab^2[ax + by] = 0$, which is equation (4).

c.1) The function $\Psi_{a,b}$ is proper, lower semicontinuous and convex; the point (x_0, y_0) belongs to $\overline{\operatorname{dom}\Phi} \times H = \overline{\operatorname{dom}\Psi_{a,b}}$. A theorem of Brézis (1973, Th. 3.2) then asserts the existence and uniqueness of a continuous function $(x, y) : [0, +\infty[\mapsto H \times H$, with $(x(t), y(t)) \in \operatorname{dom}\Psi_{a,b}$ for any t, $x(0) = x_0$, $y(0) = y_0$, which is Lipschitz continuous on $[\delta, +\infty[$ for every $\delta > 0$, and which satisfies (2) almost everywhere. This result readily entails the assertions.

c.2) After Lemaire (1996, Cor. 2.1) we have: $\Psi_{a,b}(x(t), y(t)) \to \inf\Psi_{a,b}$, as $t \to +\infty$. The inequalities $\inf\frac{1}{b^2}\Phi = \inf\Psi_{a,b} \leq \frac{1}{b^2}\Phi(x(t)) \leq \Psi_{a,b}(x(t), y(t))$ then entail the asserted convergence result.

c.3) If $\operatorname{argmin}\Phi \neq \emptyset$ then $\operatorname{argmin}\Psi_{a,b} \neq \emptyset$. It is now a theorem of Bruck (1975), which asserts the weak convergence of (x, y) towards a minimum point $(x_\infty, y_\infty) = (x_\infty, -\frac{a}{b}x_\infty)$ of $\Psi_{a,b}$ as $t \to +\infty$.

c.4) If Φ is even, then so is $\Psi_{a,b}$. Resorting once more to a theorem of Bruck (1975) yields the strong convergence. ∎

To keep with clarity, let us briefly sum up how (DIN) has been derived.

By analogy with the proximal regularization method, the minimization of the convex function Φ is replaced by the minimization of the convex function $\Psi_{a,b}(x, y) = \frac{1}{b^2}\Phi(x) + \frac{1}{2}|ax + by|^2$.

To that end, the continuous steepest descent method is applied to $\Psi_{a,b}$, which gives rise to system (ESD). Any solution (x, y) of the latter is such that $x(t)$ weakly converges to a minimum point of Φ as $t \to +\infty$.

If Φ is $\mathcal{C}^2$, then (ESD) is equivalent to a (DIN) system with $\alpha\beta > 1$ ($\alpha =$

3. Optimizing properties of (DIN) in general

In this part, the optimizing properties of (DIN) are examined with more generality than before, *i.e.* Φ need not be convex and $\alpha\beta > 1$ need not hold. Facts are stated without proofs, which may be found in F. Alvarez et al. (2002).

Let α, β, A, B, C be real constants, arbitrary for the moment. The system (DIN), which we recall

$$\text{(DIN)} \qquad \ddot{x}(t) + \alpha\dot{x}(t) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0$$

bears a strait relation with the following first order system

$$\text{(g-DIN)} \qquad \begin{cases} \dot{x} + C\nabla\Phi(x) + Ax + By = 0 \\ \dot{y} \qquad\qquad\quad + Ax + By = 0 \end{cases}$$

as the next proposition shows. In spite of their resemblance, (g-DIN) is not a gradient system (except if $A = B$) while (ESD) is. But the equivalence between (DIN) and (g-DIN) is more general than the equivalence between (DIN) and (ESD), which requires $\alpha\beta > 1$.

PROPOSITION 3.1 *Suppose $\Phi \in \mathcal{C}^2(H)$, and let the constants α, β, A, B, C satisfy*

$$\beta \neq 0,\ A = \alpha - \frac{1}{\beta},\ B = \frac{1}{\beta},\ C = \beta.$$

The systems (DIN) and (g-DIN) are equivalent in the sense that x is a solution of (DIN) if and only if there exists $y \in \mathcal{C}^2([0,+\infty[, H)$ such that (x, y) is a solution of (g-DIN).

Beyond being of first order in time, the system (g-DIN) is interesting because it does not involve the Hessian of Φ. As a first consequence, the numerical solution of (DIN) is highly simplified, since it may be performed on (g-DIN) and only requires approximating the gradient of Φ. As a second consequence, (g-DIN) allows to give a sense to (DIN) when Φ is of class $\mathcal{C}^1$ only, or when Φ is nonsmooth or involves constraints, provided that a notion of generalized gradient is available (e.g. the subdifferential set for a convex function Φ). But that remark would be of little utility if (g-DIN) did not have good existence and asymptotic convergence properties as $t \to +\infty$, under the sole assumption $\Phi \in \mathcal{C}^1(H)$. Actually (g-DIN) retains some of the optimizing properties of (DIN), at least if $\Phi \in \mathcal{C}^{1,1}(H)$.

THEOREM 3.1 *(optimizing properties of* (g-DIN)*)*

Assume that $\Phi : H \mapsto \mathbb{R}$ is bounded from below, differentiable with $\nabla\Phi$ Lipschitz continuous on the bounded subsets of H; assume further $C > 0$, $B > 0$,

(i) *For each* (x_0, y_0) *in* $H \times H$, *there exists a unique solution* (x, y) *of* (g-DIN) *defined on the whole interval* $[0, +\infty[$, *which belongs to* $C^1(0, \infty; H) \times C^2(0, \infty; H)$ *and satisfies the initial conditions* $x(0) = x_0$ *and* $y(0) = y_0$.

(ii) • $\dot{x}$ *and* $\nabla\Phi(x)$ *belong to* $L^2(0, +\infty; H)$,
• $\lim_{t\to+\infty} \Phi(x(t))$ *exists*,
• $\lim_{t\to+\infty}(\dot{x}(t) + C\nabla\Phi(x(t)) = 0$.

(iii) *Assuming, moreover, that* x *is in* $L^\infty(0, +\infty; H)$, *we have*
• $\dot{x}$, $\nabla\Phi(x)$ *are bounded on* $[0, +\infty[$,
• $\lim_{t\to+\infty} \nabla\Phi(x(t)) = \lim_{t\to+\infty} \dot{x}(t) = 0$.

In view of Proposition 3.1, when Φ belongs to $C^2(H)$, the conditions $C > 0$, $B > 0$, $B + A > 0$ for (g-DIN) are easily seen to be equivalent to $\alpha > 0$, $\beta > 0$ for (DIN). This readily implies the following corollary of Theorem 3.1.

COROLLARY 3.1 (optimizing properties of (DIN))
Assume that $\Phi : H \mapsto \mathbb{R}$ *is bounded from below, twice differentiable with* $\nabla^2\Phi$ *Lipschitz continuous on the bounded subsets of* H; *assume further* $\alpha > 0$, $\beta > 0$ *in* (DIN). *Then the following properties hold:*

(i) *For each* $(x_0, \dot{x}_0)$ *in* $H \times H$, *there exists a unique solution* x *of* (DIN) *defined on the whole interval* $[0, +\infty[$, *which belongs to* $C^2(0, \infty; H)$ *and satisfies the initial conditions* $x(0) = x_0$ *and* $\dot{x}(0) = \dot{x}_0$.

(ii) • $\dot{x}$ *and* $\nabla\Phi(x)$ *belong to* $L^2(0, +\infty; H)$,
• $\lim_{t\to+\infty} \Phi(x(t))$ *exists*,
• $\lim_{t\to+\infty}(\dot{x}(t) + \beta\nabla\Phi x(t)) = 0$.

(iii) *Assuming, moreover, that* x *is in* $L^\infty(0, +\infty; H)$, *we have*
• $\dot{x}$, $\nabla\Phi(x)$ *are bounded on* $[0, +\infty[$,
• $\lim_{t\to+\infty} \nabla\Phi(x(t)) = \lim_{t\to+\infty} \dot{x}(t) = 0$.

Let us finally state two convergence results (F. Alvarez et al., 2002).

THEOREM 3.2 *In addition to the hypotheses of Theorem* 3.1, *assume that* Φ *is convex, and that* $\operatorname{argmin}\Phi$, *the set of minimizers of* Φ *on* H, *is nonempty. Then for any solution* (x, y) *of* (g-DIN), $x(t)$ *weakly converges to a minimizer of* Φ *on* H *as* t *goes to infinity.*

THEOREM 3.3 *Assume that* $\Phi : \mathbb{R}^N \mapsto \mathbb{R}$ *is analytic, and let* x *be a bounded solution of* (DIN) *with* $\alpha > 0$, $\beta > 0$. *Then* $\dot{x}$ *belongs to* $L^1(0, +\infty; H)$ *and* $x(t)$ *converges towards a critical point of* Φ *as* $t \to \infty$.

4. An entropy-like version of the system (ESD)

From now on, H is assumed to be finite-dimensional, that is $H = \mathbb{R}^N$, $N \geq 1$.

A common feature in constrained optimization consists in replacing the quadratic kernel in the proximal point algorithm by a distance-like functional

set. If C is a non empty closed convex subset of $\mathbb{R}^N$, this leads to dynamics of the type

$$x^{k+1} \in \operatorname{argmin}\{\Phi(x) + \lambda_k d(x, x^k) : x \in C\}, \quad \lambda_k > 0,$$

where $d : C \times C \to \mathbb{R} \cup \{+\infty\}$ is strictly convex with respect to its first variable. Let us mention, for instance, the comprehensive survey of Kiwiel (1997) on generalized Bregman distances, the entropy-like algorithm using Czisàr φ divergences proposed in Iusem-Svaiter-Teboulle (1994), and also the recent logarithmic quadratic method of Auslender-Ben Tiba-Teboulle (1999).

Inspired by those fruitful ideas, and motivated by the properties of (DIN), we devote this section to the construction of an inertial method of the type (ESD), but with a φ divergence kernel - see formula (7) below- instead of the quadratic term $(x, y) \to \frac{1}{2}|ax + by|^2$.

The choice of this particular kernel is suggested by its remarkable jointly convex property, which naturally fits our energy-like descent method approach.

Let us now specify the setting. Consider the problem

$$(\mathcal{P}_+) \qquad \inf\{\Phi(x) : x \in \mathbb{R}^N_+\},$$

where the objective function $\Phi : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is assumed to be lower semicontinuous and convex with

$$\operatorname{dom}\Phi \cap \mathbb{R}^N_{++} \neq \emptyset, \quad \mathbb{R}^N_{++} = \{x \in \mathbb{R}^N | x_i > 0,\ \forall i \in \{1, \dots, N\}\}. \tag{6}$$

φ divergences are generated by the functions $\varphi : \mathbb{R}_{++} \to \mathbb{R}$ satisfying the following properties

$$(H)_\varphi \qquad \begin{cases} (\mathrm{i})_\varphi\ \varphi \text{ is continuous and nonnegative on } \mathbb{R}_{++}, \\ (\mathrm{ii})_\varphi\ \varphi \text{ is strictly convex}, \\ (\mathrm{iii})_\varphi\ \varphi(1) = 0. \end{cases}$$

Define the φ divergence $d_\varphi : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ as

$$d_\varphi(x, y) = \begin{cases} \sum_{i=1}^N y_i \varphi(y_i^{-1} x_i) \text{ if } (x, y) \in (\mathbb{R}^N_{++})^2, \\ +\infty \text{ elsewhere.} \end{cases} \tag{7}$$

EXAMPLE. As in Iusem et al. (1994), where many other examples are given, a particularly interesting example is provided by

$$\varphi_0(s) = s \log s - s + 1, \quad s \geq 0,$$

with the convention $0 \log 0 = 0$. The associated φ_0 divergence is the Kullback–Liebler entropy, that is

$$d_{\varphi_0}(x, y) = \sum^N x_i \log \frac{x_i}{} + y_i - x_i,\ \forall (x, y) \in \mathbb{R}^N_+ \times \mathbb{R}^N_{++}.$$

It is worthwile pointing out that d_{φ_0} can also be viewed as the D function of the Bregman function $\mathbb{R}^N_+ \ni x \to \sum_{i=1..N} x_i \log x_i$. This has relevant consequences in the asymptotic analysis of the proximal-like dynamics associated to d_{φ_0}, Iusem et al. (1994), Attouch and Teboulle, to appear.

The φ divergence d_φ need not be lower semicontinuous. In order to meet the classical assumptions in minimization problems, we introduce the lower semicontinuous regularization of d_φ, denoted by $\overline{d_\varphi}$. It is characterized by the following properties

(*a*) For all $(x, y) \in \mathbb{R}^N \times \mathbb{R}^N$, and for all sequences satisfying $(x^k, y^k) \to (x, y)$ as $k \to +\infty$,

$$\liminf_{k \to +\infty} d_\varphi(x^k, y^k) \geq \overline{d_\varphi}(x, y),$$

(*b*) For all $(x, y) \in \mathbb{R}^N \times \mathbb{R}^N$, there exists a sequence satisfying $(x^k, y^k) \to (x, y)$ as $k \to +\infty$, such that

$$\limsup_{k \to +\infty} d_\varphi(x^k, y^k) \leq \overline{d_\varphi}(x, y).$$

We have the following

LEMMA 4.1 *Let φ and d_φ satisfy $(H)_\varphi$ and (7). Then*

(i) $\overline{d_\varphi}$ *is a proper, lower semicontinuous, convex function,*
(ii) $\overline{d_\varphi} \geq 0$,
(iii) *For all* $(x, y) \in \mathbb{R}^N_{++} \times \mathbb{R}^N_{++}$, $\overline{d_\varphi}(x, y) = d_\varphi(x, y)$,
(iv) *For all* $(x, y) \in \mathbb{R}^N \times \mathbb{R}^N$, *the following separation property holds,* $\overline{d_\varphi}(x, y) = 0 \Leftrightarrow x = y,\ x \in \mathbb{R}^N_+$.

Proof. The convexity of d_φ, and therefore (i), comes from $(ii)_\varphi$ and the following fact:

$$g: R_{++} \to R \text{ is convex if and only if } (r, s) \in (\mathbb{R}_{++})^2 \to sg(s^{-1}r) \text{ is convex.}$$

One recognizes in $(r, s) \in (\mathbb{R}_{++})^2 \to sg(s^{-1}r)$ the Hörmander's perspective function of g; for a reference and further developments on the topic, see Maréchal (2001). The property (ii) follows from the fact that $d_\varphi \geq 0$, while (iii) is a consequence of $(i)_\varphi$.

To deal with (iv), let us examine the values of $\overline{d_\varphi}$. If $(x, y) \notin (\mathbb{R}^N_+)^2$ then, obviously, $\overline{d_\varphi}(x, y) = +\infty$ and by (iii) $\overline{d_\varphi}(x, y) = d_\varphi(x, y)$ as soon as $(x, y) \in (\mathbb{R}^N_{++})^2$.

To cope with the case of $(x, y) \in \partial(\mathbb{R}^N_+)^2$, where $\partial(\mathbb{R}^N_+)^2$ denotes the boundary of $(\mathbb{R}^N_+)^2$, let us first notice that the definition of d_φ, allows to restrict the requirement (a) to nonnegative sequences. Besides, in order to compute $\liminf_{k \to +\infty} d_\varphi(x^k, y^k)$, where (x^k, y^k) is a nonnegative sequence, observe that the structure of d_φ permits to argue on *each* coordinate, and thus it can be

For $(x,y) \in \partial\mathbb{R}^2_+ = \{(x,y) \in \mathbb{R}^2_+ | xy = 0\}$, let (x^k, y^k) $x^k, y^k > 0$ be a sequence converging to (x,y) as $k \to +\infty$. Three cases are distinguished,

• $x = 0, y \neq 0$. From $(i)_\varphi$, $(ii)_\varphi$ and $(iii)_\varphi$ it ensues that φ is non increasing on $(0,1)$ and achieves its minimum at $s = 1$. Therefore $d_\varphi(x^k, y^k) \to y \lim_{s\to 0^+} \varphi(s) > 0$, as $k \to +\infty$.

• $x \neq 0$ and $y = 0$. Fix $s_0 > 1$, and let us apply the convex inequality to φ, this gives for all $s \in \mathbb{R}_{++}$ and for all $g \in \partial\varphi(s_0)$

$$\varphi(s) \geq \varphi(s_0) + g.(s - s_0) \geq g.(s - s_0). \tag{8}$$

Observe that $(i)_\varphi$, $(ii)_\varphi$ and $(iii)_\varphi$ imply that all subgradients contained in $\partial\varphi(s_0)$ are positive. Hence (8) yields

$$d_\varphi(x^k, y^k) = y^k \varphi(\frac{x^k}{y^k}) \geq g x^k - g y^k s_0,$$

where $g \in \partial\varphi(s_0)$, $g > 0$, and thus $\liminf_{k\to+\infty} d_\varphi(x^k, y^k) \geq gx > 0$.

• $x = y = 0$. Just notice that $d_\varphi(\frac{1}{k}, \frac{1}{k}) = 0$ for all $k \geq 1$.

Applying the above results together with the properties $(i)_\varphi$, $(ii)_\varphi$, we easily deduce (iv). ■

Take d_φ as above and define for all $(x,y) \in \mathbb{R}^N \times \mathbb{R}^N$

$$\Psi_\varphi(x,y) = \Phi(x) + \overline{d_\varphi}(x,y). \tag{9}$$

By Lemma 4.1 and (6) this gives rise to a proper lower semicontinuous convex function. The optimality properties of Ψ_φ and Φ are linked in the following way:

LEMMA 4.2 *Let φ, d_φ and Ψ_φ satisfy $(\mathcal{H}_\varphi)$, (7) and (9). Then*

$$\inf_{\mathbb{R}^N\times\mathbb{R}^N} \Psi_\varphi = \inf_{\mathbb{R}^N_+} \Phi$$

$$\operatorname{argmin}\{\Psi_\varphi(x,y) : (x,y) \in \mathbb{R}^N \times \mathbb{R}^N\} = \{(x,x) : x \in \operatorname{argmin}_{\mathbb{R}^N_+} \Phi\}.$$

Proof. It relies on Lemma 4.1, and on the relations

$$\Psi_\varphi(x,x) = \Phi(x) \quad \forall x \in \mathbb{R}^N_+,$$
$$\Psi_\varphi(x,y) \geq \Phi(x) \quad \forall (x,y) \in \mathbb{R}^N_+ \times \mathbb{R}^N. \quad ■$$

For each $x, y \in \mathbb{R}^N$ let us set $X = (x,y) \in \mathbb{R}^N \times \mathbb{R}^N$. Following the lines of Section 2, let us define the dynamical system

$$(ESD)_\varphi \qquad \begin{cases} \dot{X}(t) + \partial\Psi_\varphi(X(t)) \ni 0 \text{ a.e. on } [0, +\infty[\\ X(0) = X_0 \end{cases}$$

where $X_0 = (x_0, y_0) \in \overline{\operatorname{dom}\Psi_\varphi}$ and $X(.)$ is the unique continuous solution

Ψ_φ a classical theorem concerning the subdifferential of a sum (Ekeland and Temam, 1973, Th. 5.6). Hence $(\mathrm{ESD})_\varphi$ can be rewritten

$$(ESD)_\varphi \qquad \begin{cases} \dot{x}(t) + \partial\Phi(x(t)) + \partial_x d_\varphi(x(t), y(t)) \ni 0 \text{ a.e. on } [0, +\infty[, \\ \dot{y}(t) + \partial_y d_\varphi(x(t), y(t)) = 0, \ \forall t \geq 0, \end{cases}$$

with $x_0 \in \overline{\mathrm{dom}\,\Phi} \cap \mathbb{R}^N_+$ and $y_0 \in \mathbb{R}^N_+$.

The dynamical system $(ESD)_\varphi$ presents the advantage of taking the constraints $\mathbb{R}^N_+$ into account without penalizing Φ, more precisely we have the following

THEOREM 4.1 *Let d_φ be a φ divergence, and Ψ_φ as in (9). Let $t \to X(t) = (x(t), y(t))$ be a solution of $(ESD)_\varphi$ then*

(i) $\Phi(x(t)) \to \inf\{\Phi(x) | x \in \mathbb{R}^N_+\}$ *as* $t \to +\infty$.

(ii) *If, moreover, $S_+ = \mathrm{argmin}\{\Phi(x) | x \in \mathbb{R}^N_+\}$ is non empty, then there exists $x^* \in S_+$ such that $(x(t), y(t)) \to (x^*, x^*)$ as $t \to +\infty$.*

Proof. It is a consequence of the previous lemma and of the results proved in Lemaire (1996) for (i), and in Brezis (1973), Bruck (1975) for (ii). ∎

REMARKS. 1. Parallelizing the derivation of (ESD) and $(\mathrm{ESD})_\varphi$ from $\Psi_{a,b}$ and Ψ_φ, respectively, via the continuous gradient method, we could also derive a nonautonomous version of (ESD) by considering the following family of functions: $\Psi_t(x, y) = \frac{1}{b^2(t)}\Phi(x) + \frac{1}{2}|a(t)x + b(t)y|^2$, where a and b are positive functions of t. This would lead to the following differential inclusion: $(\dot{x}(t), \dot{y}(t)) + \partial\Psi_t(x(t), y(t)) \ni 0$ (see Baillon, Cominetti, 2001, Furuya et al. 1986, for facts about this type of problems).

2. It would be interesting to know if $(\mathrm{ESD})_\varphi$ is a dynamical interior point method. Indeed its numerical treatment may be delicate if a trajectory happens to touch the boundary of $(\mathbb{R}^N_{++})^2$, since $\bar{d}_\varphi$ is liable to singularity there; choosing numerically good functions $\bar{d}_\varphi$ is not so easy. Yet, we presume that, under fairly general assumptions on Φ and φ, each trajectory starting from $(x_0, y_0) \in (\mathrm{dom}\,\Phi \cap \mathbb{R}^N_{++}) \times \mathbb{R}^N_{++}$ remains in the interior of the constraints. Certainly this question deserves further study.

References

ALVAREZ, F., and ATTOUCH, H. (2001) An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, **9**, Issue 1/2, 3–11.

ALVAREZ, F., ATTOUCH, H., BOLTE, J. and REDONT, P. (2002) A second-order gradient-like dissipative dynamical system with Hessian driven damping. *Journal de Mathématiques Pures et Appliquées*, **81**, 747–779.

ANTIPIN, A.S. (1994) Minimization of convex functions on convex sets by means

9, 1475–1486, Sep. 1994; (English translation: *Differential Equations*, **30**, n° 9, 1365-1375, 1994).

ATTOUCH, H., GOUDOU, X. and REDONT, P. (2000) The heavy ball with friction method, I. The continuous dynamical system: global exploration of the global minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, **2**, 1–34.

ATTOUCH, H. and TEBOULLE, M. A regularized Lotka-Volterra dynamical system as a continuous proximal-like method in optimization. *to appear.*

AUSLENDER, A., TEBOULLE, M. and BEN-TIBA, S. (1999) Interior proximal and multiplier methods based on second order homogeneous kernels. *Mathematics of Operation Research*, **24**, 645–668.

BAILLON, J.-B. and COMINETTI, R. (2001) A Convergence Result for Subgradient Evolution Equations and its Application to the Steepest Descent Exponential Penalty Trajectory in Linear Programming. *Journal of Functional Analysis*, **187**, 263–273.

BARBU, V. (1981) Necessary conditions for distributed control problems governed by parabolic variational inequalities. *SIAM J. on Control and Optimization*, **19**, 64–86.

BRÉZIS, H. (1973) Opérateurs maximaux monotones. *Mathematics Studies 5*, North-Holland-American Elsevier.

BRUCK, R.E. (1975) Asymptotic convergence of nonlinear contraction semigroups in Hilbert space. *Journal of Functional Analysis*, **18**, 15–26.

COMINETTI, R. (1997) Coupling the proximal point algorithm with approximation methods. *J. Optim. Theory Appl.*, **95**, 581–600.

DONTCHEV, A.L. and ZOLEZZI, T. (1993) Well-Posed optimization problems. *Lectures Notes in Mathematics*, **1543**, Springer.

EKELAND, I. and TEMAM, R. (1973) *Analyse convexe et problèmes variationnels.* Dunod, Paris.

FLAM, S.D. and HORVATH, CH. (1996) Network games; adaptations to Nash–Cournot equilibrium. *Annals of Operations Research*, **64**, 179–195.

FURUYA, H., MIYASHIBA, K. and KENMOCHI, N. (1986) Asymptotic Behavior of Solutions to a Class of Nonlinear Evolution Equations. *Journal of Differential Equations*, **62**, 73–94.

IUSEM, A.N., SVAITER, B.F. and TEBOULLE, M. (1994) Entropy-like proximal methods in convex programming. *Mathematics of Operation Research*, **19**, 4, 790–814.

KIWIEL, K.C. (1997) Proximal minimization methods with generalized Bregman functions. *SIAM J. of Control and Optimization*, **35**, 4, 1142–1168.

LEMAIRE, B. (1996) An asymptotical variational principle associated with the steepest descent method for a convex function. *Journal of Convex Analysis*, **3**, 1, 63–70.

LIONS, J.-L. (1983) Contrôle des systèmes distribués singuliers. *Méthodes Math-*

MARÉCHAL, P. (2001) On the convexity of the multiplicative potential and penalty functions and related topics. *Mathematical Programming*, Ser. A **89**, 505–516.

MARTINET, B. (1972) Détermination approchée d'un point fixe d'une application pseudo-contractante. Cas de l'application prox. *Comptes-Rendus de l'Académie des Sciences de Paris*, **274**, 163–165.

MOREAU, J.-J. (1965) Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, **93**, 273–299.

POLYAK, B.T. (1987) Introduction to Optimization. *Optimization Software*, New York.

ROCKAFELLAR, R.T. (1976) Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, **14**, 877–898.