ON GLOBAL AND LOCAL CONVERGENCE OF HALF-QUADRATIC ALGORITHMS

Marc Allain[†⋆], *Jérôme Idier*[†]

[†]CNRS-Supélec-Université Paris XI
Laboratoire des Signaux et Systèmes
3, rue Joliot-Curie
91192 Gif-sur-Yvette, France.
name@lss.supelec.fr

Yves Goussard[⋆]

[⋆]École Polytechnique
Department of Electrical Engineering
CP 6079, Succ. Centre ville
Montréal (Qc), H3C 3A7, Canada.
yves.goussard@polymtl.ca

ABSTRACT

This study gives original results on the global and local convergence properties of half-quadratic (HQ) algorithms resulting from the Geman and Yang (GY) and Geman and Reynolds (GR) primal-dual constructions. In particular, we show that the convergence domain of the GY algorithm can be extended with the benefit of an improved convergence rate.

1. INTRODUCTION

In the field of image processing, many methodological approaches lead to an optimization problem of the form:

$$\inf_{\boldsymbol{x}\in\mathbb{R}^N} J(\boldsymbol{x}) \tag{1}$$

where $J : \mathbb{R}^N \to \mathbb{R}$. The "penalized least-square" (PLS) approach in image processing is a typical example: J is then made up of a quadratic fidelity term with respect to some data set $\boldsymbol{y}$ and of an image model involving, for instance, pairwise pixel interactions, i.e.

$$J(\boldsymbol{x}) = ||\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}||^2 + \lambda\left[\sum_{i=1}^{I}\sum_{j=2}^{J}\phi\left(x_{ij} - x_{i,j-1}\right) + \sum_{i=2}^{I}\sum_{j=1}^{J}\phi\left(x_{ij} - x_{i-1,j}\right)\right]; \tag{2}$$

where ϕ is typically "edge-preserving" such as the Huber function. Matrix $\boldsymbol{H}$ is a linear observation operator such as a convolution matrix in image restoration or a projection operator in tomographic reconstruction. More specifically, the authors are interested in the minimization of penalized criteria for tridimensionnal reconstruction from helical tomographic data [1].

However, we underline that (1) is not restricted to optimization of PLS criteria optimization and can arise in many areas related to signal processing such as, for instance, robust estimation [2].

In a quite general setting, let us consider criteria (or objective functions) J of the form

$$J(\boldsymbol{x}) = J_0(\boldsymbol{x}) + \lambda\Phi(\boldsymbol{x}) \qquad \lambda \geq 0, \tag{3}$$

where J_0 is a quadratic convex form which reads,

$$J_0(\boldsymbol{x}) = \langle \boldsymbol{M}_0\boldsymbol{x}, \boldsymbol{x}\rangle - 2\langle \boldsymbol{m}_0, \boldsymbol{x}\rangle + \mu_0, \tag{4}$$

with $\boldsymbol{M}_0$ symmetric, and Φ a function defined by:

$$\Phi(\boldsymbol{x}) = \sum_{c=1}^{M}\phi\big(\langle \boldsymbol{d}_c, \boldsymbol{x}\rangle - w_c\big), \tag{5}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a scalar function; $w_c \in \mathbb{R}$ and $\boldsymbol{d}_c \in \mathbb{R}^N$ have a support regions restricted to clique c. In general, the solution to (1) must be sought iteratively. This can be done in an efficient way by resorting to Fenchel duality in the framework of convex analysis[1]: a *primal-dual* formulation of the *primal* problem (1) leads to "half-quadratic" (HQ) algorithms with remarkable structural properties. Two distinct constructions leading to HQ algorithms have been reported in the literature: the one due to Geman and Yang (GY) and the one due to Geman and Reynolds (GR). A synthetic presentation of these constructions as well as a global convergence study of the resulting algorithms within the convex analysis context can be found in [4].

The present study provides original convergence results for these algorithms by studying them as *constant stepsize descent algorithms*. In particular, the global convergence domain of the GY algorithm is extended and a faster form is deduced.

2. PRIMAL FORM OF SQ ALGORITHMS

For sake of notational simplicity, let us introduce:

$$\delta_c = \langle \boldsymbol{d}_c,\ \boldsymbol{x}\rangle,$$
$$\boldsymbol{D} = [\boldsymbol{d}_1|\cdots|\boldsymbol{d}_M],\ \boldsymbol{w} = [w_1, \cdots, w_M]^t,$$
$$\boldsymbol{\phi}'(\delta_\bullet - w_\bullet) = [\phi'(\delta_1) - w_1, \ldots, \phi'(\delta_M) - w_M]^t.$$

HQ algorithms perform "block coordinate descent" [5, p.267] on the primal-dual criterion by successive optimization on primal variables $\boldsymbol{x}$ and auxiliary (or dual) variables

[1] The reader is referred to [3] for a complete presentation of Fenchel duality.

l. For the GY HQ algorithm, the primal-dual update is defined by [4, §III]:

$$l^{(n)} = D^t x^{(n)} - w - \alpha\phi'(\delta_\bullet^{(n)} - w_\bullet) \quad (6a)$$

$$\bar{x}^{(n+1)} = (B_{\text{GY}}^{\alpha})^{-1}\left(2m_0 + \frac{\lambda}{\alpha} D l^{(n)}\right) \quad (6b)$$

where $B_{\text{GY}}^{\alpha} = 2M_0 + \frac{\lambda}{\alpha} DD^t$, and $\alpha > 0$ is a "scale factor". For the GR HQ algorithm, the update equations are given by [4, §IV]:

$$L^{(n)} = \text{diag}\left\{\phi'(\delta_c^{(n)} - w_c)/(\delta_c^{(n)} - w_c)\right\} \quad (7a)$$

$$\bar{x}^{(n+1)} = 2\left(B_{\text{GR}}^{(n)}\right)^{-1} m_0, \quad (7b)$$

with $B_{\text{GR}}^{(n)} = 2M_0 + \lambda D L^{(n)} D^t$. For these algorithms to be correctly defined, the inverse operator in (6) or (7) is assumed to exist and to be bounded. For both algorithms, the primal iterate $x^{(n+1)}$ reads:

$$x^{(n+1)} = (1-\theta)x^{(n)} + \theta\bar{x}^{(n+1)}; \quad (8)$$

where $\theta > 0$ is a relaxation factor.

Update (8), associated with (6) or (7), identifies with a constant stepsize primal algorithm: by substituting (6a) into (6b) and (6b) into (8) and after some manipulations, we get

$$x^{(n+1)} = x^{(n)} + \theta(B_\bullet^{(n)})^{-1}\nabla J(x^{(n)}) \quad (9)$$

with $B_\bullet^{(n)} = B_{\text{GY}}^{\alpha}$. Similarly, the GR algorithm can also read (9) with $B_\bullet^{(n)} = B_{\text{GR}}^{(n)}$. The update equation (9) characterizes a constant stepsize Newton-type method on J [6, p.501]. This equivalence result allows us to study the global and local convergence properties of HQ algorithms using classical tools of unconstrained optimization.

3. SUFFICIENT CONDITIONS FOR GLOBAL CONVERGENCE

The present study is restricted to convex, coercive and continuously differentiable (C^1) objective functions. As a consequence, the infimum of J is uniquely attained at, say, x^*. Given the primal form (3), these properties on J hold if $M_0 + DD^t$ invertible and if ϕ is chosen:

$$\begin{array}{l}(a)\ \text{strictly convex,}\\ (b)\ \text{coercive on } \mathbb{R},\\ (c)\ \text{at least } C^1.\end{array} \quad (10)$$

Now, let us describe the basic form of the algorithms of interest: given an initial vector $x^{(0)}$, one generates a sequence $\{x^{(n)}\}_{n=1,\ldots}$ defined by:

$$x^{(n+1)} = x^{(n)} + \theta^{(n)}\xi^{(n)}, \quad \theta^{(n)} > 0 \quad (11)$$

where $\xi^{(n)} \equiv \xi\left(x^{(n)}\right)$ and $\theta^{(n)} \equiv \theta\left(x^{(n)}\right)$ respectively denote the displacement direction and the stepsize for the current update. We also require the successive displacement directions $\xi^{(n)}$ to solve a linear system that reads

$$B^{(n)}\xi^{(n)} = -\nabla J(x^{(n)}), \quad (12)$$

where $B^{(n)} \equiv B\left(x^{(n)}\right) : \mathbb{R}^N \to \mathbb{R}^{N\times N}$ is a positive definite operator (PD). Finally, the following notations are adopted:

$$\mathcal{X} \equiv \{x^{(n)}\}, \quad \Xi \equiv \{\xi^{(n)}\},$$
$$\Theta \equiv \left\{\theta^{(n)}\right\}, \quad \mathcal{J} \equiv \{J\left(x^{(n)}\right)\}.$$

A triplet $(\mathcal{X}, \Xi, \Theta)$ that fulfills (11) and (12) will be called *iterative scheme* in the sequel. Let us recall now some useful global convergence results for unconstrained optimization.

3.1. Global convergence: general results

We now focus on "descent methods", that is, on iterative schemes $(\mathcal{X}, \Xi, \Theta)$ such that $\mathcal{J}$ is a nonincreasing sequence. However, we emphasize that this property does not suffice to ensure that $\mathcal{X}$ converges to x^*. Additional properties on Ξ and Θ have to be met. The fact that Ξ be *gradient related* to $\mathcal{X}$ is one of them [6, 14.3.1]: it guarantees that each direction allows a « significant » decrease of the objective function. The following proposition gives useful sufficient conditions (SC) to ensure this property.

Proposition 1 ([6, 14.4.1]) *Let J be C^1 on a compact set $D_0 \subset \mathbb{R}^N$, and $B : D_0 \to \mathbb{R}^{N\times N}$ be a PD operator for which $\exists\ \gamma_2 \geq \gamma_1 > 0$ such that $\forall u \in D_0, v \in \mathbb{R}^N$:*

$$\gamma_2||v||^2 \geq \langle B(u)v, v\rangle \geq \gamma_1||v||^2.$$

If Ξ is defined by (12), then Ξ is gradient related to $\mathcal{X}$.

The property of *Admissibility* for a given stepsize rule corresponding to Θ guarantees that successive stepsizes $\theta^{(n)}$ allow a "sufficient" decrease of the objective function; the Armijo rule and the Wolf conditions are the main stepsize rules. The following results are based on the Armijo rule, which proves simpler and more fruitful in the present context.

Definition 1 *Let J be a C^1 function and $(\mathcal{X}, \Xi, \Theta)$ an iterative scheme. The sequence Θ is admissible for $\mathcal{X}$ and Ξ in the sense of the Armijo rule if there exists $\omega \in (0;1)$ such that:*

$$\forall n\ J(x^{(n+1)}) - J(x^{(n)}) - \omega\theta^{(n)}\left\langle \nabla J(x^{(n)}), \xi^{(n)}\right\rangle \leq 0.$$

This stepsize rule, implemented using a "backtracking" technique [5, p.29], is sufficient to prove the convergence of iterative schemes, provided that sequence Ξ is gradient related to $\mathcal{X}$ [6, 14.3.2].

3.2. Global convergence and constant stepsize

Any iterative scheme $(\mathcal{X}, \Xi, \Theta)$ that fulfills (11) and (12) with $\theta^{(n)} = \theta \ \forall n$ is called a "constant stepsize iterative scheme". In this case, Θ is a *constant* sequence that is be denoted $\Theta = \theta$. Similarly, $(\mathcal{X}, \Xi, \theta)$ corresponds to a constant stepsize iterative scheme.

In general, a constant stepsize does not ensure global convergence of a gradient related sequence. However, ensuring that a given stepsize θ fulfills the Armijo condition for any initial vector $\boldsymbol{x}^{(0)} = \boldsymbol{u}$ provides the following SC:

Proposition 2 *Let J be C^1, strictly convex and coercive, and $(\mathcal{X}, \Xi, \theta)$ a constant stepsize iterative scheme. If Ξ is gradient related to $\mathcal{X}$ and if $\exists\, \omega \in (0;1) : \forall \boldsymbol{u} \in \mathbb{R}^N$,*

$$J(\boldsymbol{u} + \theta\xi(\boldsymbol{u})) - J(\boldsymbol{u}) - \omega\theta \langle \nabla J(\boldsymbol{u}), \xi(\boldsymbol{u}) \rangle \leq 0 \quad (13)$$

then $\mathcal{X} \rightarrow \boldsymbol{x}^*$.

In the sequel, a constant stepsize $\Theta = \theta$ will be referred to as admissible in the sens of Armijo rule if (13) is true for some $\omega \in (0;1)$.

4. GLOBAL CONVERGENCE STUDY FOR HQ ALGORITHMS

The existing sufficient conditions for global convergence of HQ algorithms have been deduced from a convexity study of the primal-dual criterion (see [4] and references therein). Here, an alternate convergence analysis is conducted for the GY and GR HQ algorithms, based on their equivalence with constant stepsize algorithms.

4.1. Global convergence of the GY algorithm

Application of Propositions 1 and 2 leads to the following result:

Theorem 1 *Let J be C^1, strictly convex and coercive. Assume also that J takes the form of (3) and that ϕ fulfills*[2]

$$\exists\, 0 < \widehat{\alpha} < \infty : \quad \widehat{g}(u) = u^2/2 - \widehat{\alpha}\phi(u) \quad \textit{convex.} \quad (14)$$

Consider an iterative scheme $(\mathcal{X}, \Xi, \theta)$ such that $\boldsymbol{B}^{(n)} \equiv \boldsymbol{B}_{GY}^{\alpha}$. If $(\theta, \alpha) \in (0;2) \times (0; 2\widehat{\alpha}/\theta)$ then $\mathcal{X} \rightarrow \boldsymbol{x}^*$.

Let us underline that Theorem 2 corresponds to substantially weakened conditions of global convergence compared to existing results. In particular, global convergence is guarantied not only within the interval $(0;\widehat{\alpha})$ introduced in [4, §III]; Figure 1(a) shows the extended part of the convergence domain brought by our study. Furthermore, the next section shows that this extension brings an increased local convergence rate which results in a faster form of the GY algorithm.

[2]Note that existence of a positive and finite $\widehat{\alpha}$ that renders $\widehat{g}$ convex is equivalent to the Lipschitz condition on ϕ' with constant $L = 1/\widehat{\alpha}$.

4.2. Global convergence of the GR algorithm

For the GR algorithm, the following result also relies on Propositions 1 and 2:

Theorem 2 *Let J be C^1, strictly convex and coercive. Assume also that J takes the form of (3), that ϕ fulfills (10a) and that*

$$\phi(\sqrt{u}) \textit{ concave on } \mathbb{R}^+, \quad (15)$$

$$\exists\, \infty > B > 0 : \forall u \in \mathbb{R}, \quad \phi'(u)/u \leq B, \quad (16)$$

$$\phi'(0) = 0. \quad (17)$$

Consider an iterative scheme $(\mathcal{X}, \Xi, \theta)$ such as $\boldsymbol{B}^{(n)} \equiv \boldsymbol{B}_{GR}^{(n)}$. If $\theta \in (0;2)$ then $\mathcal{X} \rightarrow \boldsymbol{x}^*$.

The conditions of Theorem 2 are slightly less restrictive than those previously available (c.f. [4, §IV]). In particular, global convergence is now guaranteed when ϕ is chosen as the Huber function.

5. ASYMPTOTIC BEHAVIOR OF HQ ALGORITHMS

Very few studies on local convergence properties of HQ algorithms are yet available. To our best knowledge, the convergence rate analysis for the GR algorithm presented in [7] remains the only significant contribution. From practical experience, we have found that the GR algorithm systematically achieves a better convergence rate than the GY algorithm. However, with respect to computational burden, GY is very attractive (the inverse of $\boldsymbol{B}_{GY}^{\alpha}$ need only be computed once) particularly when the size of the numerical problem becomes large. On the other hand, the free parameter α has a noticeable influence on the asymptotic behavior of the GY algorithm. This section shows that *faster forms of the GY iteration are obtained for values of α chosen in the "extended part" of the convergence domain*; a simulation is performed as an illustration.

In order to simplify the convergence rate analysis, let us consider that ϕ is C^2. The Hessian of J is then well defined and reads:

$$\boldsymbol{H}(\boldsymbol{x}) = 2\boldsymbol{M}_0 + \lambda \boldsymbol{D}\boldsymbol{\phi}''(\delta_\bullet - w_\bullet)\boldsymbol{D}^t$$

with $\boldsymbol{\phi}''(\delta_\bullet - w_\bullet) = \mathrm{diag}\{\phi''(\delta_c - w_c)\}$. From a qualitative viewpoint, the difference between the asymptotic behaviors of the GR and GY algorithms can be explained by their respective ability to provide a "good" approximation to the Hessian $\boldsymbol{H}$ at $\boldsymbol{x}^*$. Let us recall that, for an iterative scheme $(\mathcal{X}, \Xi, \theta)$ defined by (9), *the convergence order is linear and the corresponding rate is given by* [6, 10.1.4]:

$$\sigma = \rho\left(\boldsymbol{I} - \theta \boldsymbol{B}_\bullet(\boldsymbol{x}^*)^{-1}\boldsymbol{H}(\boldsymbol{x}^*)\right) \quad (18)$$

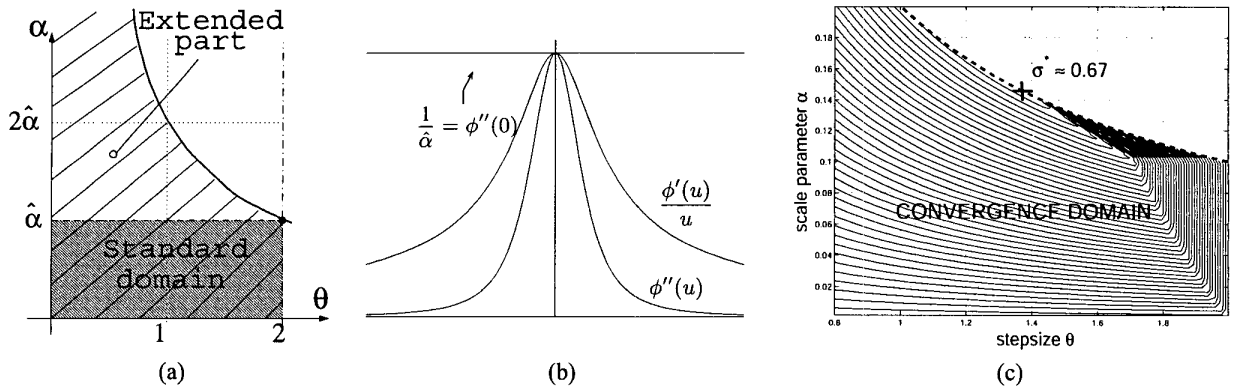

Fig. 1. (a) Extension of the convergence domains for GY algorithm; (b) drawing of $\phi''(u)$, $\phi'(u)/u$ and $1/\hat{\alpha}$ for $\phi(u) = \phi_H(u) = \sqrt{u^2+s^2}$; (c) estimation of the convergence rate $\sigma(\theta,\alpha)$ for GY algorithm when J takes the form (2) with $\phi = \phi_H$.

where $\rho(\boldsymbol{A})$ is the spectral radius of $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, and $\boldsymbol{B}_\bullet(\boldsymbol{x}^*)$ stands for either $\boldsymbol{B}_{\text{GY}}^{\alpha}$ or $\boldsymbol{B}_{\text{GR}}(\boldsymbol{x}^*)$. Remark that smaller values of σ correspond to faster asymptotic convergence. The fastest rate is obtained for $\theta^{-1}\boldsymbol{B}_\bullet(\boldsymbol{x}^*) = \boldsymbol{H}(\boldsymbol{x}^*)$, e.g. by Newton iterates. Algorithms with $\theta\boldsymbol{B}_\bullet(\boldsymbol{x}^*)^{-1}\boldsymbol{H}(\boldsymbol{x}^*)$ close to $\boldsymbol{I}$ are expected to have a low rate [3] σ. In this respect, since $\phi'(u)/u$ is a better approximation to $\phi''(u)$ than the constant value $1/\alpha$ (see Figure 1(b)), the GR algorithm is expected to converge faster than the GY one[4].

Obviously, σ depends either on θ alone (GR) or jointly on θ and α (GY), and there is a major interest in assessing the values θ^* and/or α^* that optimize σ. It is well known that θ^* is problem dependent. However, it is almost always chosen between 1.3 et 1.8 for image or signal processing problems. For the GY algorithm, empirical evidence indicates that the best rate is obtained at the boundary of the extended domain is depicted in Figure 1(a). We now illustrate this phenomenon with a synthetic example.

We used the GY algorithm to minimize a PLS criterion resulting from a tomographic reconstruction problem. The convergence rate was estimated according to relation (18). $\boldsymbol{x}^{\text{num}}$ was used instead of $\boldsymbol{x}^*$, $\boldsymbol{x}^{\text{num}}$ being the actual solution up to numerical precision (10^{-12}). Figure 1(c) shows a part of this simulation extracted from the convergence domain drawn in Figure 1(a). The maximum value of the convergence rate σ^* is reached at $\theta^* \approx 1.4$ and $\alpha^* = 2\widehat{\alpha}/\theta^*$. Finally, let us notice that the level contours follow the hyperbola in a large part of the convergence domain, i.e., σ remains constant for many couples (θ,α) such that $\alpha\theta = constant$.

[3] When $\sigma = 0$, the algorithm reaches a "superlinear" rate.

[4] Although in the quadratic case ($\phi(u) = u^2$), both algorithms are equivalent to Newton iterates.

In conclusion, extension of the usual convergence domain leads to a faster form of GY algorithm. We are currently carrying our study on from a theoretical ground, in order to assess that the fastest rate always satisfies $\alpha^* = 2\widehat{\alpha}/\theta^*$.

6. REFERENCES

[1] Marc Allain, Yves Goussard, and Jérôme Idier, "Approche régularisée en reconstruction tomographique 3D hélicoïdale," in *Actes 18e coll. GRETSI*, Toulouse, France, Sept. 2001.

[2] Rao Yarlagadda, J. Bee Bednar, and Terry L. Watt, "Fast algorithms for l_p deconvolution," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-33, no. 1, pp. 174–182, Feb. 1985.

[3] R. Tyrell Rockafellar, *Convex Analysis*, Princeton Univ. Press, 1970.

[4] Jérôme Idier, "Convex half-quadratic criteria and interacting auxiliary variables for image restoration," *IEEE Trans. Image Processing*, vol. 10, no. 7, pp. 1001–1009, July 2001.

[5] Dimitri P. Bertsekas, *Nonlinear programming*, Athena Scientific, Belmont, MA, 2 edition, 1999.

[6] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, NY, 1970.

[7] Tony F. Chan and Pep Mulet, "On the convergence of the lagged diffusivity fixed point method in total variation image restoration," *SIAM Journal of Numerical Analysis*, vol. 36, no. 2, pp. 354–367, 1999.