Pascal Joint Report I3S–LSE R-LP-2004-03

Asymptotic behaviour of a family of gradient algorithms in $\mathbb{R}^d$ and Hilbert spaces

Submitted to

Mathematical Programming

Luc Pronzato, Laboratoire I3S, Sophia Antipolis, France
Henry P. Wynn, London School of Economics, UK
Anatoly A. Zhigljavsky, Cardiff University, UK

Université Nice SOPHIA ANTIPOLIS

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Abstract

The asymptotic behaviour of a family of gradient algorithms (including the methods of steepest descent and minimum residues) for the optimisation of bounded quadratic operators in $I\!R^d$ and Hilbert spaces is analyzed. The results obtained generalize those of Akaike (1959) in several directions. First, all algorithms in the family are shown to have the same asymptotic behaviour (convergence to a two-point attractor), which implies in particular that they have similar asymptotic convergence rates. Second, the analysis also covers the Hilbert space case. A detailed analysis of the stability property of the attractor is provided.

1 Introduction

The paper generalizes the results presented in [16] to other optimisation algorithms of the gradient type. We introduce a class of algorithms, called P-gradient algorithms, that differ by the choice of the length of the step made in the gradient direction. The class includes in particular the usual steepest-descent algorithm and the method of minimal residues of Krasnosel'skii and Krein [10, 9]. We show that for a quadratic function, the worst asymptotic rate of convergence is the same for the whole class of algorithms considered. It is also true that, expressed in the right framework, all the algorithms in the class behave in a very similar fashion[1]. This analysis complements that presented in [1], [13, 14] and Chapter 7 of [15] which concerns steepest descent. Moreover, the analysis in [16] directly applies to all algorithms in the class considered, revealing the asymptotic behaviour for bounded quadratic operators not only in $I\!R^d$ but also in Hilbert spaces. The worst case behaviour exhibited is fundamental "bottom-line" in the study of optimisation whose understanding is critical for building more complex and faster algorithms.

The basic idea is renormalisation, as used throughout [15]. The main result in the finite dimension case is that for any algorithm in the class, in the renormalised space one observes convergence to a two-point attractor which lies in the space spanned by the eigenvectors corresponding to the smallest and largest eigenvalues of the matrix A of the quadratic operator. The proof for bounded quadratic operators in Hilbert space stems from the proof for $I\!R^d$ but is considerably more technical. In both cases, as in [1], the method consists of converting the problem to one containing a special type of operator on measures on the spectrum of the operator. The additional technicalities arise from the fact that in the Hilbert space case the measure, which is associated with the spectral measure of the operator, may be continuous. Another important result concerns bounds on convergence rates, named after Kantorovich, see [7]. For all algorithms in the family considered, the actual asymptotic rate of convergence, although satisfying Kantorovich bounds, depends on the starting point and is difficult to predict. This complex behaviour has consequences for the stability of the attractor, which are discussed following the main results.

The family of gradient algorithms we consider, called P-gradient algorithms, is introduced in Section 2. Renormalisation is presented there, which, together with the Lyapunov functions of Section 2.4, forms the core of the analysis to be conducted. The main results are presented in Section 3, first for the case $\mathcal{H} = I\!R^d$, then for the Hilbert space case. They rely on the convergence property of successive transformations of a probability measure, which is presented in Section 4. Again, the two cases $\mathcal{H} = I\!R^d$ and $\mathcal{H}$ a Hilbert space are distinguished, the exposition being much simpler in the former case. The stability of attractors is discussed in Section 5, only in the more general case of a Hilbert space, the case $\mathcal{H} = I\!R^d$ not allowing for a significant simplification of the presentation. Finally, Section 6 shows the asymptotic equivalence between several rates of convergence of gradient algorithms. All proofs and some important lemmas are collected in an appendix.

2 A family of gradient algorithms

2.1 P-gradient algorithms

Let A be a real bounded self-adjoint (symmetric) operator in a real Hilbert space $\mathcal{H}$ with inner product (x, y) and norm given by $(x, x) = \|x\|^2$. Assume that A is positive, bounded below, and denote its spectral boundaries by m and M:

$$m = \inf_{\|x\|=1} (Ax, x)\,, \;\; M = \sup_{\|x\|=1} (Ax, x)\,,$$

[1]Not all algorithms using the gradient direction belong to that class, which in particular does *not* include the spectral-gradient algorithm, see [3], proposed by Barzilai and Borwein in [2]. This method, which has been found in particular examples to allow significant improvement over standard steepest descent, see [17], thus requires a separate treatment.

with $0 < m < M < \infty$. The function to be minimized corresponds to the quadratic form

$$f(x) = \frac{1}{2}(Ax, x) - (x, y)\,. \tag{1}$$

It is minimum at $x^* = A^{-1}y$, its directional derivative at x in the direction u is

$$\frac{\partial f(x)}{\partial u} = \frac{(Ax - y, u)}{(u, u)^{1/2}}\,.$$

The direction of steepest descent at x is $-g$, with $g = g(x)$ the gradient at x, namely $g = Ax - y$. The minimum of f in this direction is obtained for the optimum step-length

$$\gamma = \frac{(g, g)}{(Ag, g)}\,,$$

which corresponds to the usual steepest-descent algorithm. One iteration of the steepest descent algorithm is thus

$$x_{k+1} = x_k - \frac{(g_k, g_k)}{(Ag_k, g_k)} g_k\,, \tag{2}$$

with $g_k = Ax_k - y$ and x_0 some initial element in $\mathcal{H}$. We define more generally the following class of algorithms.

Definition 1 *Let $P(\cdot)$ be a real function defined on $[m, M]$, infinitely differentiable, with Laurent series*

$$P(z) = \sum_{-\infty}^{\infty} c_k z^k\,,\ c_k \in I\!R \text{ for all } k\,,$$

such that $0 < \sum_{-\infty}^{\infty} c_k a^k < \infty$ for $a \in [m, M]$. The k-th iteration of a P-gradient algorithm is defined by

$$x_{k+1} = x_k - \gamma_k g_k \tag{3}$$

where the step-length γ_k minimizes $(P(A)g_{k+1}, g_{k+1})$ with respect to γ, with $g_{k+1} = g(x_{k+1}) = g(x_k - \gamma g_k)$.

Direct calculation gives

$$\gamma_k = \frac{(P(A)Ag_k, g_k)}{(P(A)A^2 g_k, g_k)}\,. \tag{4}$$

Note that $AP(A) = P(A)A$ and that the denominator and numerator of γ_k are linear in $P(A)$. Also, γ_k is scale-invariant in $P(A)$ and $\gamma_k \in [1/M, 1/m]$.

Taking $P(A) = A^{-1}$ gives the steepest-descent algorithm. Choosing $P(A) = I$, the identity operator, is equivalent to choosing the step-length that minimizes the norm of the gradient g_{k+1} at the next point. We then obtain the method of minimal residues introduced in [10] for the solution of linear equations. For any fixed $\alpha \in (0, 1)$, choosing γ_k that minimizes $\alpha f(x_k - \gamma g_k) + (1 - \alpha)(g(x_k - \gamma g_k), g(x_k - \gamma g_k))$ with respect to γ also gives an algorithm in the family. More generally, we show below how to construct P-gradient algorithms, with $P(\cdot)$ a polynomial in A, using evaluations of $f(\cdot)$ and $g(\cdot)$ only.

2.2 Practical construction when P is a polynomial

We consider the case where $P(A) = A^q$ for some integer $q \geq 1$. As mentioned, the cases $q = -1$ and $q = 0$ respectively correspond to the methods of steepest-descent and minimal residues. The extension to $P(\cdot)$ polynomial in A is straightforward (including also linear combinations with A^{-1}), using (4).

The minimisation of $(P(A)g_{k+1}, g_{k+1})$, or the calculation of γ_k in (4), requires the calculations of terms of the form $(A^n g, g)$, with $n = q$ or $n = q + 1, q + 2$. As shown below, they are easily obtained from evaluations of $g(\cdot)$ at different points. Notice that this construction implies that *one* iteration of the algorithm will require *several* evaluations of $g(\cdot)$. The construction proposed below is not necessarily the most economical one, and evaluations of $f(\cdot)$ and $g(\cdot)$ at different points could be combined to provide more efficient evaluations of terms $(A^n g, g)$. Our objective here is simply to show that the family of algorithms considered in the paper is not of purely theoretical interest, and that other algorithms than the steepest-descent and minimal residues could also be considered in practice.

Let $(A^n g, g)$ be the term to be evaluated, with $g = g(x)$ the gradient at the current point x. Define $x^{(0)} = x$ and

$$x^{(i+1)} = x^{(i)} - \beta g(x^{(i)}) \,,\ i \geq 0 \,,$$

with β a fixed positive number (for instance, β can be taken equal to the value of γ at previous iteration of the algorithm). We obtain

$$g^{(i)} = g(x^{(i)}) = (I - \beta A)^i g \,.$$

Define $P_i = (g, g^{(i)}) = (g, (I - \beta A)^i g)$. In matrix notation, $\mathbf{P}_n = \mathbf{Q}_n G_n$, where

$$\mathbf{P}_n = (P_0, P_1, \ldots, P_n)^\top \,,\ G_n = ((g,g), (Ag,g), \ldots, (A^n g, g))^\top$$

and the entries of the $(n+1) \times (n+1)$ matrix $\mathbf{Q}_n$ are the binomial coefficients,

$$\mathbf{Q}_n = \begin{pmatrix} 1 & & & & \\ 1 & -\beta & & & \\ 1 & -2\beta & \beta^2 & & \\ 1 & -3\beta & 3\beta^2 & -\beta^3 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} .$$

The value of $(A^n g, g)$ is then directly obtained from $G_n = \mathbf{Q}_n^{-1} \mathbf{P}_n$. The entries of $\mathbf{P}_n$, defined by $P_i = (g, g^{(i)})$, are also obtained more economically from

$$P_{2j} = (g^{(j)}, g^{(j)}) \,,\ P_{2j+1} = (g^{(j+1)}, g^{(j)}) \,.$$

Therefore, the evaluation of $\gamma_k = (P(A)Ag_k, g_k)/(P(A)A^2 g_k, g_k)$, with $P(\cdot)$ a polynomial of degree q, requires $\lceil q/2 \rceil + 2$ gradient evaluations (including the one at $x^{(0)} = x_k$).

2.3 Renormalisation

We can rewrite the iteration (3) as

$$(x_{k+1} - x^*) = (x_k - x^*) - \gamma_k g_k \,,$$

with $g_k = g(x_k) = A(x_k - x^*)$, so that

$$g_{k+1} = g_k - \gamma_k A g_k = g_k - \frac{(P(A)Ag_k, g_k)}{(P(A)A^2 g_k, g_k)} \, A g_k \,.$$

Define the renormalised variable

$$z_k = \frac{B g_k}{(P(A)Ag_k, g_k)^{1/2}} \,, \tag{5}$$

with B the positive square-root of $P(A)A$, so that $(z_k, z_k) = 1$. Also define

$$\mu_i^k = (A^i z_k, z_k) \,,\ i \in \mathbb{Z} \,, \tag{6}$$

so that $\mu_0^k = 1$ for any k and $\gamma_k = \mu_0^k / \mu_1^k = 1/\mu_1^k$. We obtain

$$\begin{aligned} z_{k+1} &= \frac{B g_{k+1}}{(P(A)Ag_{k+1}, g_{k+1})^{1/2}} = \frac{(I - \gamma_k A) B g_k}{((I - \gamma_k A) B g_k, (I - \gamma_k A) B g_k)} \\ &= \frac{(I - \gamma_k A) z_k}{((I - \gamma_k A) z_k, (I - \gamma_k A) z_k)} = \frac{(I - \gamma_k A) z_k}{(1 - 2\gamma_k \mu_1^k + \gamma_k^2 \mu_2)^{1/2}} \,, \end{aligned}$$

that is,

$$z_{k+1} = \frac{(I - A/\mu_1^k) z_k}{(\mu_2^k / (\mu_1^k)^2 - 1)^{1/2}} \,. \tag{7}$$

This gives the updating formula for the moments

$$\mu_i^{k+1} = (A^i z_{k+1}, z_{k+1}) = \frac{\mu_i^k - 2\mu_{i+1}^k / \mu_1^k + \mu_{i+2}^k / (\mu_1^k)^2}{\mu_2^k / (\mu_1^k)^2 - 1} \,. \tag{8}$$

In the special case where $\mathcal{H} = I\!R^d$, we can assume that A is already diagonalised, with eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$. We can then consider $[z_k]_i^2$, with $[z_k]_i$ the i-th component of z_k, as a weight on the eigenvalue λ_i, so that (7) defines a transformation on discrete probability measures ν_k supported on $\{\lambda_1, \ldots, \lambda_d\}$, with moments μ_i^k, $i \in Z\!\!Z$. The asymptotic behaviour of the sequence (z_k) generated by (7) was studied in [1], see also [5] and Chapter 7 of [15]. The main result is that, assuming $0 < \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_{d-1} < \lambda_d$, the sequence (z_k) converges to a two-dimensional plane, spanned by the eigenvectors e_1, e_d associated with λ_1 and λ_d. The attraction property is stated more precisely in Section 3, also in the Hilbert space case. It is already important to notice that although the results in the references above were obtained for the steepest-descent algorithm, the renormalisation (5), which depends on the chosen $P(\cdot)$, makes them applicable to any algorithm in the family considered. Also, using the renormalisation just defined we easily obtain (non asymptotic) results on the monotonicity of the algorithm along its trajectory.

2.4 Monotonicity of a rate of convergence

Consider the function $(P(A)g_{k+1}, g_{k+1})$ that γ_k minimizes, and compute the rate of convergence r_k of the algorithm at iteration k, defined by

$$r_k = \frac{(P(A)g_{k+1}, g_{k+1})}{(P(A)g_k, g_k)} \,. \tag{9}$$

Other rates of convergence will be considered in Section 6 where they will be shown to be asymptotically equivalent to r_k. Direct calculation gives $r_k = 1 - 1/L_k$, with

$$L_k = \mu_1^k \mu_{-1}^k$$

where the moments μ_i^k are defined by (6). Also, from (8), L_k satisfies

$$L_{k+1} - L_k = \frac{\mu_1^k}{D_k^2} \det \mathbf{M}_k$$

with

$$D_k = \mu_2^k - (\mu_1^k)^2$$

and

$$\mathbf{M}_k = \begin{pmatrix} \mu_{-1}^k & \mu_0^k & \mu_1^k \\ \mu_0^k & \mu_1^k & \mu_2^k \\ \mu_1^k & \mu_2^k & \mu_3^k \end{pmatrix} \,. \tag{10}$$

The moment matrix $\mathbf{M}_k$ is non-negative definite so that $\det \mathbf{M}_k \geq 0$, and thus $L_{k+1} \geq L_k$, that is, both L_k and the rate r_k are non decreasing along the trajectory followed by the algorithm. When $\mathcal{H} = I\!R^2$ ($d = 2$), $\det \mathbf{M}_k = 0$ and r_k is constant. When $d > 2$ or $\mathcal{H}$ is a Hilbert space, the rate is monotonically increasing for a typical x_0, indeed, for almost all $z_0 = z_0(x_0)$ with respect to the uniform measure on the unit sphere when $\mathcal{H} = I\!R^d$. Notice that if the rate is constant over two iterations ($\det \mathbf{M}_k = 0$), then the measure ν_k is supported on two points only, and the iteration (7) for the weights shows that this situation will continue: the rate will thus remain constant for all subsequent iterations.

Note that L_k and D_k are bounded (since ν_k has a bounded support), respectively by L^* and D^*, with $L^* = (M+m)^2/(4mM)$ and $D^* = (M-m)^2/4$, see Lemma 1 in Appendix A3. Therefore, since L_k is non decreasing it converges to some limit, and

$$\det \mathbf{M}_k = \frac{(L_{k+1} - L_k) D_k^2}{\mu_1^k} \leq \frac{(L_{k+1} - L_k)(D^*)^2}{m} \to 0 \,, \; k \to \infty \,. \tag{11}$$

In addition to L_k and r_k another quantity also turns out to be non decreasing along the trajectory. Consider

$$\frac{(P(A)Ag_{k+1}, g_{k+1})}{(P(A)A(x_{k+1} - x_k), (x_{k+1} - x_k))} = \frac{(P(A)Ag_{k+1}, g_{k+1})}{\gamma_k^2 (P(A)Ag_k, g_k)} = \mu_2^k - (\mu_1^k)^2 = D_k \,. \tag{12}$$

Direct calculation using (7) gives

$$D_{k+1} - D_k = \frac{1}{D_k^2} \det \mathbf{N}_k$$

with

$$\mathbf{N}_k = \begin{pmatrix} \mu_0^k & \mu_1^k & \mu_2^k \\ \mu_1^k & \mu_2^k & \mu_3^k \\ \mu_2^k & \mu_3^k & \mu_4^k \end{pmatrix}$$

and, again, $\mathbf{N}_k$ is non-negative definite and $\det \mathbf{N}_k \geq 0$ so that D_k is also non decreasing, converges to some limit and $\det \mathbf{N}_k$ converges to zero for the same reasons as above.

Substitution of $P(A)$ for a particular algorithm shows which quantities are monotonic. For the steepest-descent algorithm, $P(A) = A^{-1}$, $(A^{-1}g_k, g_k) = 2[f(x_k) - f(x^*)]$, and thus the ratios $r_k = [f(x_{k+1}) - f(x^*)]/[f(x_k) - f(x^*)]$ and $D_k = (g_{k+1}, g_{k+1})/((x_{k+1} - x_k), (x_{k+1} - x_k))$ are monotonically non decreasing. For the method of minimal residues, $P(A) = I$, and the ratios $r_k = (g_{k+1}, g_{k+1})/(g_k, g_k)$ and $D_k = (Ag_{k+1}, g_{k+1})/(A(x_{k+1} - x_k), (x_{k+1} - x_k))$ are monotonically non decreasing.

The monotonicity of L_k and D_k makes them suitable Lyapunov functions for studying the asymptotic behaviour of the algorithm. This is developed in the next and central section.

3 Asymptotic behaviour of gradient algorithms

Consider the case $\mathcal{H} = I\!R^d$, and assume that the minimal and maximal eigenvalues of A, $\lambda_1 = m$, $\lambda_d = M$, are simple. The attraction property can be stated as follows. Choose z_0, the renormalised variable defined by (5) at the initial point x_0, such that $(z_0, e_1) > 0$, $(z_0, e_d) > 0$, with e_1 and e_d the eigenvectors associated with λ_1 and λ_d respectively. Then

$$z_{2k} \to \sqrt{p}\, e_1 + \sqrt{1-p}\, e_d\,, \quad z_{2k+1} \to \sqrt{1-p}\, e_1 - \sqrt{p}\, e_d \quad \text{when} \quad k \to \infty\,,$$

where p is some number in $(0, 1)$, see Section 5 concerning the range of possible values for p. This property, stated in a more general framework in the theorem below, has important consequences for the asymptotic rate of convergence of the algorithm, see Section 6. The proof of the attraction property relies on the convergence of successive transformations of the probability measure ν_k defined by $[z_k]_i^2$. The approaches used in [1, 5] to study this convergence do not apply when $\mathcal{H}$ is infinite dimensional, and we shall present a more general proof in Section 4. It differs somewhat from the one in Chapter 7 of [15], in particular in the choice of the Lyapunov function, L_k instead of D_k.

The attraction theorem in $I\!R^d$ can be stated as follows. We can assume that A is diagonalised, and the probability measure ν_k is then discrete and puts weight $[z_k]_i^2$ at the eigenvalue λ_i. Notice that the updating rule (7) is identical for $[z_k]_i$ and $[z_k]_j$ associated with $\lambda_i = \lambda_j$, and the corresponding weights can thus be summed. We can therefore assume that all eigenvalues are different when studying the evolution of ν_k, see Theorem 3.

Theorem 1 *Let A be a positive definite symmetric matrix in $I\!R^d \times I\!R^d$, with minimum and maximum eigenvalues m and M such that $0 < m < M < \infty$ and apply a P-gradient algorithm, see Definition 1, for the minimisation of $f(x)$ given by (1), initialized at x_0, with $z_0 = z(x_0)$, see (5). Assume that*

$$E_1 z_0 \neq 0 \quad \textit{and} \quad E_d z_0 \neq 0\,, \tag{13}$$

where E_1 and E_d are the orthogonal projectors on the eigenspaces respectively associated with $\lambda_1 = m$ and $\lambda_d = M$. Then the asymptotic behaviour of the renormalised gradient $z_k = z(x_k)$ is such that

$$z_{2k} = \sqrt{p}\, u_{2k} + \sqrt{1-p}\, v_{2k}\,, \; z_{2k+1} = \sqrt{1-p}\, u_{2k+1} - \sqrt{p}\, v_{2k+1}\,,$$

with $\|u_n\| = \|v_n\| = 1$ $\forall n$, $\|Au_n - mu_n\| \to 0$, $\|Av_n - Mv_n\| \to 0$ as $n \to \infty$, and p, some number in $(0, 1)$, depending on z_0.

The proof is omitted since we prove later a more general property valid for $\mathcal{H}$ a Hilbert space. A more precise result is obtained when the eigenvalues λ_1 and λ_d are simple: the vector z_d converges to the two-dimensional plane defined by the eigenvectors e_1 and e_d associated with λ_1 and λ_d.

Corollary 1 *Let A be a positive definite symmetric matrix with ordered eigenvalues*

$$0 < m = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_{d-1} < \lambda_d = M$$

and let e_1, e_d be the eigenvectors associated with λ_1 and λ_d respectively. Apply a P-gradient algorithm, see Definition 1, for the minimisation of $f(x)$ given by (1), initialized at x_0 such that $z_0^\top e_1 \neq 0$ and $z_0^\top e_d \neq 0$, with $z_0 = z(x_0)$, see (5). Then the algorithm attracts to the plane Π spanned by e_1 and e_d in the following sense:

$$w^\top z_k \to 0\,,\ k \to \infty$$

for any nonzero vector $w \in \Pi^\perp$. Moreover, the sequence (z_k) converges to a two-point cycle.

This corollary is a straightforward consequence of Theorem 1: when λ_1 and λ_d are simple, with associated eigenvectors e_1 and e_d, u_n and v_n then respectively tend to e_1 and e_d. The result easily generalizes to the case when (13) is not satisfied. The algorithm then attracts to a two-dimensional plane defined by the eigenvectors e_i and e_j associated with the smallest and largest eigenvalues such that $z_0^\top e_i \neq 0$ and $z_0^\top e_j \neq 0$.

We state now the attraction theorem in the more general case where $\mathcal{H}$ is a Hilbert space. The proof is given in Appendix A1.

Theorem 2 *Let A be a bounded real symmetric operator in a Hilbert space $\mathcal{H}$, positive, with bounds m and M, such that $0 < m < M < \infty$ and apply a P-gradient algorithm, see Definition 1, for the minimisation of $f(x)$ given by (1), initialized at x_0, with $z_0 = z(x_0)$, see (5). Assume that z_0 is such that for any ϵ, $0 < \epsilon < (M-m)/2$,*

$$(E_{m+\epsilon} z_0, z_0) > 0 \quad \text{and} \quad (E_{M-\epsilon} z_0, z_0) < 1\,, \tag{14}$$

with (E_λ) the spectral family of projections associated with A. The asymptotic behaviour of the renormalised gradient $z_k = z(x_k)$ is such that

$$z_{2k} = \sqrt{p}\, u_{2k} + \sqrt{1-p}\, v_{2k}\,,\ z_{2k+1} = \sqrt{1-p}\, u_{2k+1} - \sqrt{p}\, v_{2k+1}\,, \tag{15}$$

with $\|u_n\| = \|v_n\| = 1\ \forall n$, $\|Au_n - mu_n\| \to 0$, $\|Av_n - Mv_n\| \to 0$ as $n \to \infty$, and p, some number in $(0,1)$, depending on z_0.

4 A property of successive transformations of a probability measure

The two properties established in this section form the cornerstones of the proofs of the Theorems of previous section. We consider first the case of a discrete measure with finite support, which in terms of convergence of a P-gradient algorithm corresponds to the case $\mathcal{H} = I\!R^d$. The proof is given in Appendix A2.

Theorem 3 *Let ν_0 be a discrete probability measure on $\{\lambda_1, \ldots, \lambda_d\}$ with*

$$0 < m = \lambda_1 < \lambda_2 < \cdots < \lambda_{d-1} < \lambda_d = M < \infty\,.$$

Let $[z_k]_i^2$ denote the weight placed at λ_i by ν_k, that is, $\nu_k(\lambda_i) = [z_k]_i^2$. Consider the transformation $T:\ \nu_k \to \nu_{k+1}$ defined by

$$[z_{k+1}]_i = \frac{(1-\lambda_i/\mu_1^k)[z_k]_i}{(\mu_2^k/(\mu_1^k)^2 - 1)^{1/2}} \tag{16}$$

with the moments μ_i^k defined by (6). Then, when $k \to \infty$,

$$[z_{2k}]_1^2 \to p\,,\ [z_{2k+1}]_1^2 \to 1-p \quad \text{and}\ [z_{2k}]_d^2 \to 1-p\,,\ [z_{2k+1}]_1^2 \to p \tag{17}$$

for some p depending on ν_0, $0 < p < 1$. Furthermore,

$$p = \frac{1}{2} \pm \frac{\rho+1}{\rho-1}\sqrt{\frac{1}{4} - \frac{\rho L}{(\rho+1)^2}}$$

with $\rho = M/m$ and $L = \lim_{k\to\infty} \mu_1^k \mu_{-1}^k$.

Note that the limiting value L depends on ν_0, so that the value of p that characterizes the attractor is difficult to predict. The range of possible values for p is discussed in Section 5.

We consider now the case of an arbitrary measure on an interval, which raises some additional difficulties compared to previous case. In terms of convergence of a P-gradient algorithm, it corresponds to the case where $\mathcal{H}$ is a Hilbert space: for E_λ the spectral family associated with the operator A, we define the measure ν_k by $\nu_k(d\lambda) = d(E_\lambda z_k, z_k)$, $m \leq \lambda \leq M$. The proof is given in Appendix A3.

Theorem 4 *Let ν_0 be a probability measure on the family $\mathcal{B}$ of Borel sets of $(0,\infty)$, with support $[m,M]$, so that*

$$\begin{aligned} m &= \text{ess inf}(\nu_0) = \sup(\alpha \ / \ \nu_0\{x\,,\, x<\alpha\}=0)\,, \\ M &= \text{ess sup}(\nu_0) = \inf(\alpha \ / \ \nu_0\{x\,,\, x>\alpha\}=0)\,. \end{aligned}$$

Assume that $0<m<M<\infty$. Consider the transformation $T:\ \nu_k \to \nu_{k+1}$ defined by

$$\nu_{k+1}(\mathcal{A}) = \int_{\mathcal{A}} \frac{(\lambda-\mu_1^k)^2}{D_k}\,\nu_k(d\lambda) \tag{18}$$

for any $\mathcal{A}\in\mathcal{B}$, where $\mu_1^k = \int \lambda\,\nu_k(d\lambda)$ and $D_k = \mu_2^k - (\mu_1^k)^2$, with $\mu_2^k = \int \lambda^2\,\nu_k(d\lambda)$. Then, as $k\to\infty$,

$$\nu_{2k}(\mathcal{A}) \to p\,,\ \nu_{2k+1}(\mathcal{A}) \to 1-p \tag{19}$$

for all $\mathcal{A}=[m,x)$, $m<x<M$, for some p depending on ν_0, $0<p<1$.

5 Stability of attractors

The range of possible values for p in the attraction Theorem 1 ($\mathcal{H}=I\!R^d$) is considered in Theorem 3 of [1] (see also Lemma 3.5 of [14]). This theorem states that when λ_i is not discarded at any iteration, that is, when $\mu_1^k \neq \lambda_i$ for any k, then $p\in[1/2-s(\lambda_i), 1/2+s(\lambda_i)]$, with $s(\lambda)$ given by (20) (note that this assumption cannot be checked). In this section we extend this result in two directions: (i) $\mathcal{H}$ is a Hilbert space, (ii) we study the stability of the attractor defined by p in Theorem 2. We shall use the following definition of stability, see [6] p. 444, [11], p. 7.

Definition 2 *A fixed point ν^* for a mapping $T(\cdot)$ on probability measures on a family $\mathcal{B}$ of Borel sets will be called stable if $\forall\epsilon>0$, $\exists\alpha>0$ such that for any ν_0 for which $d(\nu_0,\nu^*)<\alpha$, $d(T^n(\nu_0),\nu^*)<\epsilon$ for all $n>0$. A fixed point ν^* is unstable if it is not stable.*

We shall use the distance $d(\nu,\nu')$ given by the Lévy-Prokhorov metric, see [18] p. 349. In our case (measures supported on $[m,M]$), $d(\nu,\nu')$ becomes the Lévy distance between the distribution functions F,F' associated with ν,ν', which we denote

$$L(F,F') = \inf\{\epsilon : F'(x-\epsilon)-\epsilon \le F(x) \le F'(x+\epsilon)+\epsilon\,,\ \forall x\}\,.$$

In the case where one of the two measures is the discrete measure ν_p^* concentrated on m,M, with $\nu_p^*(m)=p$, $\nu_p^*(M)=1-p$, we get

$$\begin{aligned} d(\nu,\nu_p^*) &= L(F,F_p^*) \\ &= \inf\{\epsilon : F(x)\le p+\epsilon \text{ for } x<M-\epsilon \ \text{ and } \ p-\epsilon\le F(x) \text{ for } m+\epsilon\le x\}\,, \end{aligned}$$

with F_p^* the distribution function associated with ν_p^*. We then have proved the following, see Appendix A4.

Theorem 5 *Consider the situation of Theorem 4, with ν_0 any probability measure supported on some closed subset $\mathcal{S}_A$ of $[m,M]$ and* ess inf$(\nu_0)=m$, ess sup $(\nu_0)=M$.
(i) The measure ν_p^* is a fixed point for the mapping T^2.
(ii) Consider the set $\mathcal{I}_u$ defined by

$$\mathcal{I}_u = \left(0, \frac{1}{2}-s(\lambda^*)\right) \cup \left(\frac{1}{2}+s(\lambda^*), 1\right)\,,$$

where

$$s(\lambda) = \frac{\sqrt{(M-\lambda)^2+(\lambda-m)^2}}{2(M-m)}\,,\ \lambda^* = \min_{\lambda\in\mathcal{S}_A} s(\lambda)\,. \tag{20}$$

Any fixed point ν_p^* with p in $\mathcal{I}_u$ corresponds to an unstable fixed point for T^2.
(iii) Any point in the interval

$$\mathcal{I}_s = \left(\frac{1}{2}-s(\lambda^*), \frac{1}{2}+s(\lambda^*)\right) \tag{21}$$

corresponds to a stable ν_p^* for the mapping T^2.

Remark 1 *$d(\nu_k, \nu_p^*) \to 0$ is equivalent to weak convergence $\nu_k \xrightarrow{\text{w}} \nu_p^*$ in the usual sense. If z_k is associated with the spectral measure ν_k and z_p^* with ν_p^*, then, in the Hilbert space this is equivalent to $(z_k - z_p^*, y) \to 0$ for any $y \in \mathcal{H}$, whereas strong convergence would require $\|z_k - z_p^*\| \to 0$. For $I\!R^d$, the two types of convergence are equivalent, and thus Corollary 1 implies strong convergence. However, for $\mathcal{H}$ a Hilbert space the equivalence is false, and indeed strong convergence generally does* not *hold. The stability property (iii) is thus a weak statement for $\mathcal{H}$ a Hilbert space. The L_2 metric in $\mathcal{H}$ induces the Hellinger metric on the space of spectral measures, which defines the same topology as the distance in variation, see [18], p. 364. Strong convergence in $\mathcal{H}$ is thus related to convergence with distance in variation in the space of spectral measures, which is clearly difficult to obtain — except in the special situation where ν_0 has positive mass at $\{m\}$ and $\{M\}$ and presents a spectral gap: $\nu_0 = 0$ on $(m, m+\epsilon)$ and $(M-\epsilon, M)$.*

We have $\nu_{k+2}(d\lambda) = H(\nu_k, \lambda)\nu_k(d\lambda)$, with $H(\nu_k, \lambda)$ given by (29) in Appendix A4. One may then notice that when ν_0 is a discrete probability measure, the condition $H(\nu_p^*, \lambda) > 1$ used in the proof of the instability part of the theorem, see Appendix A4, corresponds to a condition on the eigenvalues of the Jacobian of the transformation T^2, see [15].

Note that the stability interval $\mathcal{I}_s$ always contains the interval

$$]\frac{1}{2} - \frac{1}{2\sqrt{2}},\ \frac{1}{2} + \frac{1}{2\sqrt{2}}[\ \approx\ ]0.14645,\ 0.85355[\,.$$

Numerical simulations for $\mathcal{H} = I\!R^3$, with A having eigenvalues $m < \lambda < M$, show that for any initial density of x_0 in $I\!R^d$ associated with a density of z_0 reasonably spread on the unit sphere, the density of the values of p corresponding to stable attractors ν_p^* can be approximated by

$$\varphi(p) = C \log \min\{1, H(\nu_p^*, \lambda)\} = \begin{cases} C \log H(\nu_p^*, \lambda) & \text{if } p \in \mathcal{I}_s \\ 0 & \text{otherwise}\,, \end{cases} \tag{22}$$

where C is a normalisation constant and $H(\nu_p^*, \lambda)$ is given by (30). Figure 1 shows the empirical density of attractors (full line) together with $\varphi(p)$ (dashed line) in the case $m = 1$, $\lambda = 4$, $M = 10$. The support of this density coincides with the stability interval $\mathcal{I}_s$ given by (21). When $d > 3$, the density of attractors depends on the initial density of x_0.

6 Rates of convergence

We first state a property showing that different definitions of rates of convergence are asymptotically equivalent, see Appendix A5.

Theorem 6 *Let W be a bounded positive self-adjoint operator in $\mathcal{H}$, with bounds c and C such that $0 < c < C < \infty$. Assume that W commutes with A (when $\mathcal{H} = I\!R^d$, W is a positive definite matrix in $I\!R^d \times I\!R^d$ with minimum and maximum eigenvalues respectively c and C). Define*

$$R_k(W) = \frac{(W g_{k+1}, g_{k+1})}{(W g_k, g_k)}$$

if $\|g_k\| \neq 0$ and $R_k(W) = 1$ otherwise. Apply a P-gradient algorithm (3), initialized at x_0, with γ_k given by (4), for the minimisation of $f(x)$ given by (1), with minimum value at x^*. Then the limit

$$R(W, x_0, x^*) = \lim_{n\to\infty} \left[\prod_{k=0}^{n-1} R_k(W)\right]^{1/n}$$

exists for all x_0, x^* in $\mathcal{H}$ and $R(W, x_0, x^*) = R(x_0, x^*)$ does not depend on W. In particular,

$$R(W, x_0, x^*) = \lim_{n\to\infty} \left(\prod_{k=0}^{n-1} r_k\right)^{1/n}$$

with r_k defined by (9).

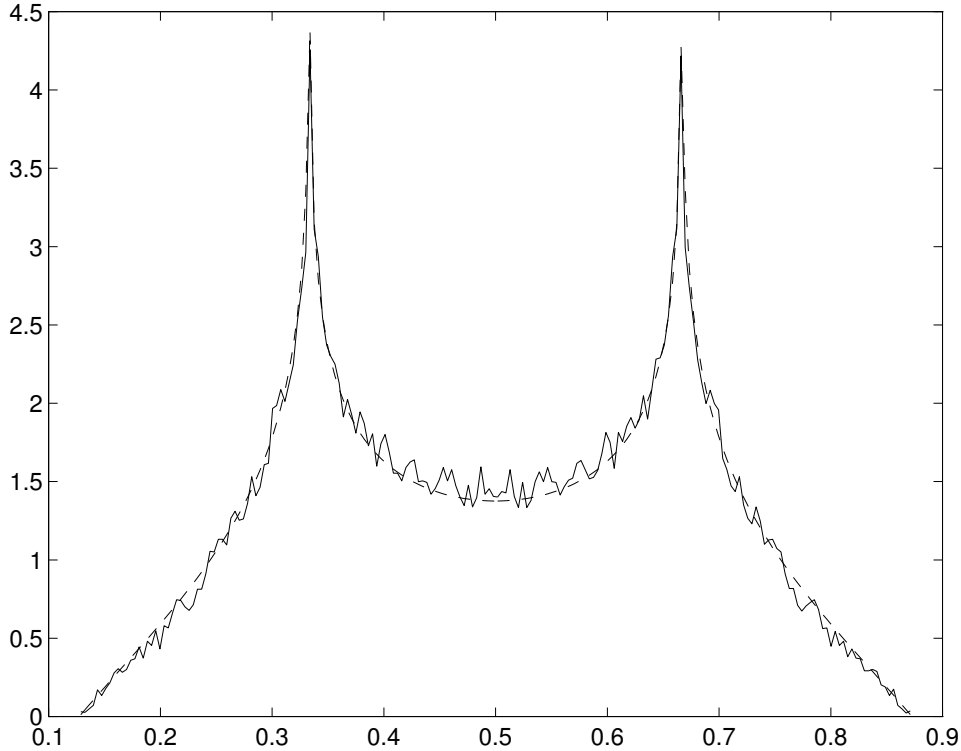

Figure 1: Empirical density of attractors (full line) and $\varphi(p)$, see (22), for $d = 3$ ($m = 1$, $\lambda = 4$, $M = 10$)

From the results of Section 3, we have

$$R(W, x_0, x^*) = r(p) = \frac{p(1-p)(\rho-1)^2}{[p+\rho(1-p)][(1-p)+\rho p]}$$

for any W, where p defines the attractor, see (15), and $\rho = M/m$ is the condition number of the operator. The function $r(p)$ is symmetric with respect to $1/2$ and monotonously increasing from 0 to $1/2$, see Figure 2. The worst asymptotic rate is thus obtained at $p = 1/2$:

$$R_{\max} = \left(\frac{\rho-1}{\rho+1}\right)^2 . \tag{23}$$

Note that $\forall k$, $r_k \leq R_{\max}$ since r_k is not decreasing, see Section 2.4. For a typical x_0 (such that the convergence is not finite, that is, such that $r(p) \neq 0$), the stability analysis of Section 5 shows that only values of p in $\mathcal{I}_s$ given by (21) may correspond to stable attractors. The range of possible values of $R(p)$ is thus $[R_{\min}, R_{\max}]$, where $R_{\max}$, given by (23), is obtained for $p = 1/2$ and

$$R_{\min} \leq R^*_{\min} = R(1/2 + 1/[2\sqrt{2}]) = \frac{(\rho-1)^2}{(\rho+1)^2 + 4\rho} .$$

Figure 3 presents the range $[R^*_{\min}, R_{\max}]$ as a function of $1/\rho$, the upper curve corresponding to $R_{\max}$ and the lower to $R^*_{\min}$. The maximum value of the range is $3 - 2\sqrt{2} \simeq 0.1716$, obtained at $\rho = 1 + 2\sqrt{2} + 2\sqrt{2+\sqrt{2}} \simeq 7.5239$. These results confirm the experimental observation that the rate of convergence of the gradient algorithm is generally close to its worst value $R_{\max}$, see [14]. The same property is true for any P-gradient algorithm.

Remark 2 *A similar analysis for D_k defined by (12), which is also not decreasing, shows that $D_k \to D(p) = p(1-p)(M-m)^2$ as $k \to \infty$, with $D_k \leq D^* = D(1/2) = (M-m)^2/4$ for all k. Also, for any typical x_0 such that $p \in \mathcal{I}_s$ given by (21), we have $D(p) \geq D(1/2 + 1/[2\sqrt{2}]) = (M-m)^2/8$.*

Another quantity of interest is given by

$$\Delta_N = \log(R_{\max}/R_{\min})/[\log(R_{\max})\log(R_{\min})] .$$

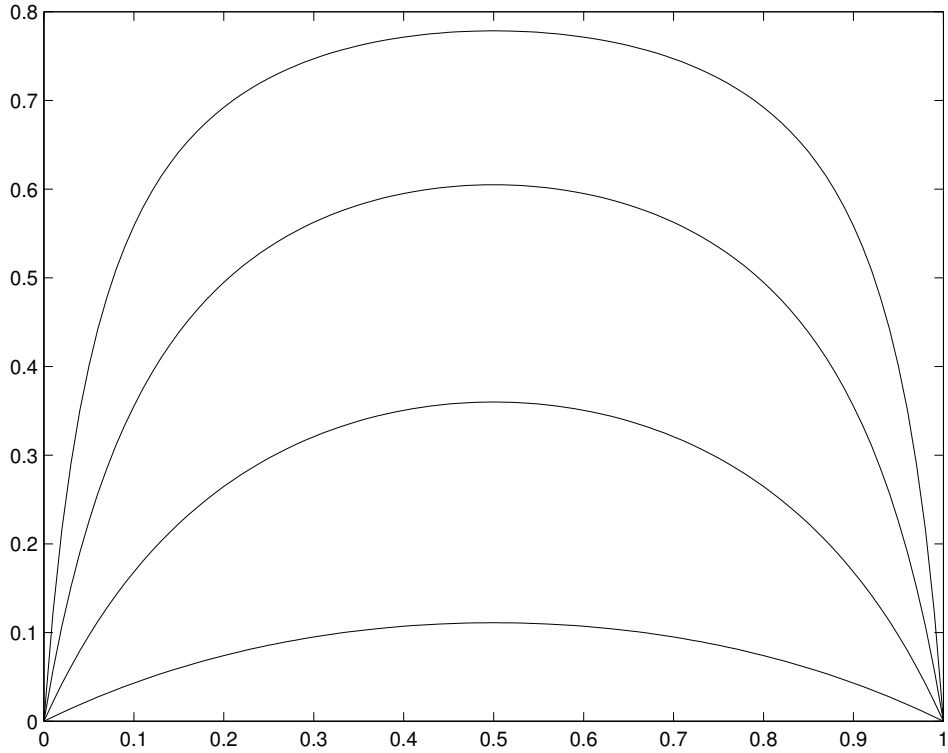

Figure 2: $r(p)$ as a function of p, for $\rho = 2$ (bottom curve), 4, 8 and 16 (top)

Indeed, for N large enough, $(Wg_N, g_N)/(Wg_0, g_0) \simeq r(p)^N$, the number N of iterations required for obtaining a ratio $(Wg_N, g_N)/(Wg_0, g_0) = \epsilon$ ($\epsilon \ll 1$) is approximately $\log(\epsilon)/\log[r(p)]$ and $\Delta_N|\log(\epsilon)|$ thus indicates the range of possible values for N due to the range of possible values for p. Direct calculation gives $\Delta_N \log(R_{\max}) < 1/2$ for any ρ and

$$\Delta_N = \rho/8 - 1/4 + O(1/\rho)\,,\ 1/\log(R_{\max}) = -\rho/4 + O(1/\rho)$$

for large ρ. Therefore, the number of iterations required by a P-gradient algorithm to achieve a given precision $\epsilon << 1$ varies at most by a factor 2 depending on the (typical) starting point x_0, factors of variation close to 2 being possible only when ρ is large.

The average value of $R(W, x_0, x^*)$ for $z_0 = z(x_0)$ uniformly distributed on the unit sphere is the same for any P-gradient algorithm, more generally, the distribution of $R(W, x_0, x^*)$ associated with a particular distribution of z_0 does not depend on the particular P-gradient algorithm considered. Moreover, numerical simulations show that the average value of $R(I, x_0, x^*)$ is the same for the steepest-descent ($P(A) = A^{-1}$) and minimum residues ($P(A) = I$) algorithms for x_0 uniformly distributed on the sphere $\|x_0 - x^*\| = 1$. The small deviations in average performance between different P-gradient algorithms can only be related to the fact that a fixed distribution for x_0 corresponds to different distributions for $z(x_0)$.

Remark 3 *It is known that the introduction of a relaxation coefficient γ, with $0 < \gamma < 1$, in the steepest-descent algorithm totally changes its behaviour, see, e.g., Chapter 7 of [15]; the algorithm (2) then becomes $x_{k+1} = x_k - \gamma[(g_k, g_k)/(Ag_k, g_k)]g_k$. For $\mathcal{H} = I\!R^d$ and a fixed A, depending on the value of γ, the renormalized process either converges to periodic orbits (the same for almost all starting points) or exhibits a chaotic behaviour, with the classical period-doubling phenomenon in the case $d = 2$. In higher dimensions, repeated numerical trials show that the process typically no longer converges to the 2-dimensional plane spanned by (e_1, e_d). A detailed analysis for $d = 2$ and experimental results for $d > 2$ also show that relaxation (with γ close to 1) considerably improves the rate of convergence. Similar results hold more generally for all P-gradient algorithms, with the iteration (3) transformed into $x_{k+1} = x_k - \gamma\gamma_k g_k$, with γ the (fixed) relaxation coefficient and γ_k given by (4).*

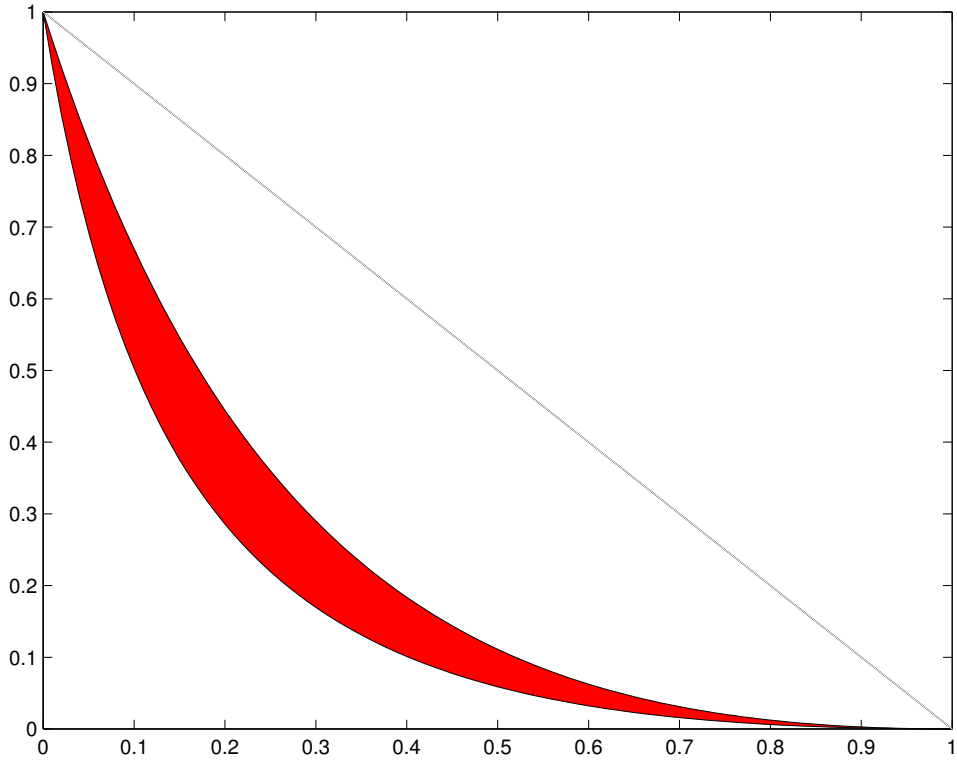

Figure 3: Range $[R^*_{\min}, R_{\max}]$ of possible values of the asymptotic rate $r(p)$ as a function of $1/\rho$

Appendix

A1. *Proof of Theorem 2.* The proof relies on Theorem 4 (Theorem 3 when $\mathcal{H} = I\!R^d$), which concerns successive transformations applied to a probability measure.

Since A is self-adjoint, its spectrum $\mathcal{S}_A$ is a closed subset of the interval $[m, M]$ of the real line and $m, M \in \mathcal{S}_A$. Let E_λ be the spectral family associated with A, and define the spectral measure ν_k by $\nu_k(d\lambda) = d(E_\lambda z_k, z_k)$, $m \le \lambda \le M$. Since $(z_k, z_k) = \int_m^M \nu_k(d\lambda) = 1$, ν_k is a probability measure on the Borel sets of $(0, \infty)$, with $\nu_k([m, M]) = 1\ \forall k$. This representation gives

$$\mu_1 = (Az_k, z_k) = \int \lambda\, \nu_k(d\lambda)\,,\ \ \mu_2 = (A^2 z_k, z_k) = \int \lambda^2\, \nu_k(d\lambda)$$

where integration is over $[m, M]$ unless otherwise specified. Therefore, for any Borel set $\mathcal{A}$ the transformation (7) gives in terms of ν_k:

$$\nu_{k+1}(\mathcal{A}) = \frac{\int_{\mathcal{A}} \left[\lambda - \int \lambda'\, \nu_k(d\lambda')\right]^2 \nu_k(d\lambda)}{\int \lambda'^2\, \nu_k(d\lambda') - \left[\int \lambda'\, \nu_k(d\lambda')\right]^2}\,.$$

The conditions (14) on z_0 are equivalent to $\operatorname{ess\,inf}(\nu_0) = m$ and $\operatorname{ess\,sup}(\nu_0) = M$, see Theorem 4, and the updating rule for ν_k can be written as (18). Theorem 4 then implies (19), which can be written as: $\forall \epsilon > 0$, $\epsilon \le \beta = (M - m)/2$,

$$(E_{m+\epsilon} z_{2k}, z_{2k}) \to p\,, \qquad (E_{M-\epsilon} z_{2k}, z_{2k}) \to p\,,$$
$$(E_{m+\epsilon} z_{2k+1}, z_{2k+1}) \to 1-p\,, \qquad (E_{M-\epsilon} z_{2k+1}, z_{2k+1}) \to 1-p\,,$$

as $k \to \infty$, where p depends on z_0, $0 < p < 1$. Define $p_{2k} = (E_{m+\beta} z_{2k}, z_{2k})$, $p_{2k+1} = 1 - (E_{m+\beta} z_{2k+1}, z_{2k+1})$, and the angles φ, φ_n by $\cos\varphi = \sqrt{p}$, $\sin\varphi = \sqrt{1-p}$, $\cos\varphi_n = \sqrt{p_n}$, $\sin\varphi_n = \sqrt{1-p_n}$, $\forall n$. Also define $s_{2k} = E_{m+\beta} z_{2k}/\cos\varphi_{2k}$, $s_{2k+1} = E_{m+\beta} z_{2k+1}/\sin\varphi_{2k+1}$, $t_{2k} = (z_{2k} - E_{m+\beta} z_{2k})/\sin\varphi_{2k}$, $t_{2k+1} = -(z_{2k+1} - E_{m+\beta} z_{2k+1})/\cos\varphi_{2k+1}$. This gives $p_n \to p$ as $n \to \infty$, $\|s_n\| = \|t_n\| = 1\ \forall n$, and $z_{2k} = \cos\varphi_{2k}\, s_{2k} + \sin\varphi_{2k}\, t_{2k}$, $z_{2k+1} = \sin\varphi_{2k+1}\, s_{2k+1} - \cos\varphi_{2k+1}\, t_{2k+1}$. Also,

$$\|As_n - ms_n\|^2 = \int (\lambda - m)^2\, d(E_\lambda s_n, s_n)\,,$$

which, for $n = 2k$ and any ϵ, $0 < \epsilon < \beta$, gives

$$\begin{aligned}
\|As_{2k} - ms_{2k}\|^2 &= \int_m^{m+\beta} \frac{(\lambda - m)^2}{p_{2k}}\, d(E_\lambda z_{2k}, z_{2k}) \\
&= \int_m^{m+\epsilon} \frac{(\lambda - m)^2}{p_{2k}}\, d(E_\lambda z_{2k}, z_{2k}) + \int_{m+\epsilon}^{m+\beta} \frac{(\lambda - m)^2}{p_{2k}}\, d(E_\lambda z_{2k}, z_{2k}) \\
&\le \frac{\epsilon^2}{p_{2k}} + \frac{\beta^2}{p_{2k}} \left[p_{2k} - \int_m^{m+\epsilon} d(E_\lambda z_{2k}, z_{2k}) \right] .
\end{aligned}$$

Since $p_{2k} \to p$ and $\int_m^{m+\epsilon} d(E_\lambda z_{2k}, z_{2k}) \to p$ as $k \to \infty$, $\|As_{2k} - ms_{2k}\| \to 0$ as $k \to \infty$. Similarly, $\|As_{2k+1} - ms_{2k+1}\| \to 0$ as $k \to \infty$ and $\|At_n - Mt_n\| \to 0$ as $n \to \infty$. Consider now

$$u_n = \cos\vartheta_n\, s_n + \sin\vartheta_n\, t_n\,, \;\; v_n = -\sin\vartheta_n\, s_n + \cos\vartheta_n\, t_n\,.$$

Straightforward calculations show that $\vartheta_n = \varphi_n - \varphi$ gives (15) with $\|u_n\| = \|v_n\| = 1\ \forall n$. Also

$$\|Au_n - mu_n\| \le |\cos\vartheta_n| \|As_n - ms_n\| + |\sin\vartheta_n|(M - m)\,,$$

and, since $\|As_n - ms_n\| \to 0$, $\vartheta_n \to 0$ as $n \to \infty$, $\|Au_n - mu_n\| \to 0$ as $n \to \infty$. Similarly, $\|Av_n - Mv_n\| \to 0$ as $n \to \infty$. ■

A2. *Proof of Theorem 3.* We first prove that the mass of ν_k tends to concentrate on two eigenvalues only. When ν_0 is non degenerate, $L_1 > 1$ from Jensen inequality, and thus, since (L_k) is non decreasing, see Section 2.4, $L_k \ge L_1 > 1$. Now, from Lagrange identity $(\sum a_i^2)(\sum b_i^2) = \sum_{i<j}(a_i b_j - a_j b_i)^2 + (\sum a_i b_i)^2$

$$\begin{aligned}
L_k &= \left(\sum_{i=1}^d \lambda_i [z_k]_i^2 \right) \left(\sum_{i=1}^d [z_k]_i^2 / \lambda_i \right) \\
&= \sum_{i<j} [z_k]_i^2 [z_k]_j^2 \left(\frac{\sqrt{\lambda_i}}{\sqrt{\lambda_j}} - \frac{\sqrt{\lambda_j}}{\sqrt{\lambda_i}} \right)^2 + \left(\sum_{i=1}^d [z_k]_i^2 \right)^2 \\
&= \sum_{i<j} [z_k]_i^2 [z_k]_j^2 \frac{(\lambda_i - \lambda_j)^2}{\lambda_i \lambda_j} + 1\,.
\end{aligned}$$

Let i_k and j_k denote the indices that achieve $\max_{i<j} [z_k]_i^2 [z_k]_j^2$. We have

$$L_k \le [z_k]_{i_k}^2 [z_k]_{j_k}^2 \sum_{i<j} \frac{(\lambda_i - \lambda_j)^2}{\lambda_i \lambda_j} + 1$$

and thus

$$[z_k]_{i_k}^2 [z_k]_{j_k}^2 \ge \delta = \frac{L_1 - 1}{\sum_{i<j} \frac{(\lambda_i - \lambda_j)^2}{\lambda_i \lambda_j}}\,.$$

Moreover, $[z_k]_{i_k}^2 + [z_k]_{j_k}^2 < 1$ gives

$$\delta < [z_k]_{i_k}^2 < 1 - \delta \;\text{ and }\; \delta < [z_k]_{j_k}^2 < 1 - \delta\,.$$

Consider the matrix $\mathbf{M}_k$ given by (10). Its determinant can be written as

$$\begin{aligned}
\det \mathbf{M}_k &= \sum_{i<j<l} [z_k]_i^2 [z_k]_j^2 [z_k]_l^2 \frac{(\lambda_i - \lambda_j)^2 (\lambda_i - \lambda_l)^2 (\lambda_j - \lambda_l)^2}{\lambda_i \lambda_j \lambda_l} \\
&\ge [z_k]_{i_k}^2 [z_k]_{j_k}^2 (\lambda_{i_k} - \lambda_{j_k})^2 \sum_{i \ne i_k, j_k} [z_k]_i^2 \frac{(\lambda_i - \lambda_{i_k})^2 (\lambda_i - \lambda_{j_k})^2}{\lambda_i \lambda_{i_k} \lambda_{j_k}} \\
&\ge \delta \frac{\delta_\lambda^6}{M^3} \sum_{i \ne i_k, j_k} [z_k]_i^2
\end{aligned}$$

where

$$\delta_\lambda = \min_{i,j} |\lambda_i - \lambda_j| \,.$$

Since $\det \mathbf{M}_k \to 0$ as $k \to \infty$, see (11), we get $\sum_{i \neq i_k, j_k} [z_k]_i^2 \to 0$ as $k \to \infty$. The mass thus tends to concentrate on $\lambda_{i_k}, \lambda_{j_k}$.

Next we prove that i_k and j_k eventually become fixed. From the result above, $\forall \epsilon > 0$, $\exists k_\epsilon$ such that $\sum_{i \neq i_k, j_k} [z_k]_i^2 < \epsilon$, $k > k_\epsilon$.

Consider the updating equation (16). We have for any i, $(\mu_1^k - \lambda_i)^2 \leq (M-m)^2$. Also, $D_k = \mu_2^k - (\mu_1^k)^2 \geq D_0$, see Section 2.4. This gives for $i \neq i_k, j_k$ and $k > k_\epsilon$

$$[z_{k+1}]_i^2 < \epsilon \frac{(M-m)^2}{D_0} \,.$$

Taking $\epsilon_1 = \delta D_0/(M-m)^2$ we obtain $[z_{k+1}]_i^2 < \delta$ for $i \neq i_k, j_k$ and $k > k_{\epsilon_1}$. Since $[z_{k+1}]_{i_{k+1}}^2 > \delta$ and $[z_{k+1}]_{j_{k+1}}^2 > \delta$, $i \notin \{i_k, j_k\}$ implies $i \notin \{i_{k+1}, j_{k+1}\}$, $k > k_{\epsilon_1}$ and thus $\{i_k, j_k\} = \{i^*, j^*\}$ for $k > k_{\epsilon_1}$.

We show now that $\{i^*, j^*\} = \{1, d\}$. Assume that $i^* < j^* < d$ (which implies $[z_k]_d^2 \to 0$, $k \to \infty$). We need to show that $(\lambda_d - \mu_1^k)^2 > (\lambda_{j^*} - \mu_1^k)^2$ for k large enough. We have

$$\mu_1^k = \lambda_{i^*}[z_k]_{i^*}^2 + \lambda_{j^*}[z_k]_{j^*}^2 + \sum_{i \neq i^*, j^*} \lambda_i [z_k]_i^2 \leq \lambda_{i^*}[z_k]_{i^*}^2 + \lambda_{j^*}[z_k]_{j^*}^2 + \lambda_d \sum_{i \neq i^*, j^*} [z_k]_i^2 \,.$$

Take $\epsilon_2 = \min\{\epsilon_1, \delta\delta_\lambda/\lambda_d\}$. For $k > k_{\epsilon_2}$ we have

$$\mu_1^k \leq \lambda_{i^*}[z_k]_{i^*}^2 + \lambda_{j^*}[z_k]_{j^*}^2 + \lambda_d \epsilon_2 \leq \lambda_{i^*}\delta + \lambda_{j^*}(1-\delta) + \lambda_d \epsilon_2 \leq \lambda_{j^*} - \delta\delta_\lambda + \lambda_d \epsilon_2 \leq \lambda_{j^*}$$

and thus $(\lambda_d - \mu_1^k)^2 > (\lambda_{j^*} - \mu_1^k)^2$. From (16), this gives for $k > k_{\epsilon_2}$

$$\frac{[z_{k+1}]_d^2}{[z_k]_d^2} = \frac{(\lambda_d - \mu_1^k)^2}{D_k} > \frac{(\lambda_{j^*} - \mu_1^k)^2}{D_k} = \frac{[z_{k+1}]_{j^*}^2}{[z_k]_{j^*}^2}$$

and thus

$$\frac{[z_{k+1}]_{j^*}^2}{[z_{k+1}]_d^2} < \frac{[z_k]_{j^*}^2}{[z_k]_d^2} \,.$$

We arrived at a contradiction since $[z_k]_d^2 \to 0$ and $[z_k]_{j^*}^2$ is bounded from below by δ. Therefore $j^* = d$. Similarly, $i^* = 1$.

Finally, let L denote $\lim_{k\to\infty} L_k$, see Section 2.4. There are only two discrete measures with nonzero weights on λ_1 and λ_d and such that $\mu_1\mu_{-1} = L$,

$$\nu^{(1)} = \left\{ \begin{array}{cc} \lambda_1 & \lambda_d \\ p & 1-p \end{array} \right\} \quad \text{and} \quad \nu^{(2)} = \left\{ \begin{array}{cc} \lambda_1 & \lambda_d \\ 1-p & p \end{array} \right\}$$

with

$$p = \frac{1}{2} - \frac{\rho+1}{\rho-1}\sqrt{\frac{1}{4} - \frac{\rho L}{(\rho+1)^2}}$$

and $\rho = M/m$. Direct calculation shows that $\nu_k = \nu^{(1)}$ gives $\nu_{k+1} = \nu^{(2)}$, hence the convergence of ν_k to the cyclic attractor $\nu^{(1)} \to \nu^{(2)} \to \nu^{(1)} \to \cdots$ ■

A3. The proof of Theorem 4 is more technical than that of Theorem 3 and relies on a series of lemmas stated below.

Lemma 1 *Let ν be any probability distribution on $[m, M]$, $0 < m \leq M < \infty$ with moments $\mu_i = \int \lambda^i \nu(d\lambda)$, $i \in \mathbb{Z}$ ($\mu_0 = 1$). Then,*

$$\mu_2 - \mu_1^2 \leq D^* = (M-m)^2/4 \tag{24}$$

$$\mu_1\mu_{-1} \leq L^* = (M+m)^2/(4mM) \,. \tag{25}$$

Proof. The proof relies on standard results in experimental design theory, see, e.g., [4, 19]. Consider the two linear regression models $\eta_1(\theta,\lambda) = \theta_0 + \theta_1\lambda$ and $\eta_2(\theta,\lambda) = \theta_0/\sqrt{\lambda} + \theta_1\sqrt{\lambda}$, with θ_0, θ_1 the model parameters and λ the design variable, $\lambda \in [m, M]$. D-optimum design (approximate theory) aims at determining a probability measure on $[m, M]$ that maximizes the determinant of the information matrix associated with a particular model, here respectively

$$\mathbf{I}_1(\nu) = \begin{pmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} \quad \text{and} \quad \mathbf{I}_2(\nu) = \begin{pmatrix} \mu_{-1} & \mu_0 \\ \mu_0 & \mu_1 \end{pmatrix} .$$

The function $\log\det \mathbf{I}(\nu)$ is concave on the set of probability measures on $[m, M]$, and its maximum is unique. The Kiefer-Wolfowitz General Equivalence Theorem [8] gives a characterization of the measure ν^* that maximizes $\det \mathbf{I}_1(\nu) = \mu_2 - \mu_1^2$ and $\det \mathbf{I}_2(\nu) = \mu_1\mu_{-1} - 1$. In this case it corresponds to the two point measure, supported at m and M, with both weights equal to $1/2$. Direct calculation gives (24,25). One may notice that (25) corresponds to the Kantorovich inequality, see [7] and [12], p. 151. (A fully development of this connection will be published elsewhere.) ■

Lemma 2 *Let ν be any probability distribution on $[m, M]$, $0 < m \le M < \infty$. Assume that there exists an interval $\mathcal{I} \subseteq [m, M]$, $|\mathcal{I}| \le \alpha$ and $\nu(\mathcal{I}) \ge 1 - \epsilon$, $\epsilon \in [0, 1]$. Then,* $\mathrm{Var}(\nu) \le \alpha^2/4 + 2\epsilon M^2$.

Proof. Define $\mu_1 = \int_{[m,M]} \lambda\, \nu(d\lambda)$, $\mu_{\mathcal{I}} = \int_{\mathcal{I}} \lambda\, \nu(d\lambda)$. Then $\mu_1 = \mu_{\mathcal{I}} + \int_{[m,M]\setminus\mathcal{I}} \lambda\, \nu(d\lambda)$. Therefore, $\mu_{\mathcal{I}} \le \mu_1 \le \mu_{\mathcal{I}} + \epsilon M$. We get

$$\begin{aligned} \mathrm{Var}(\nu) = \int (\lambda - \mu_1)^2\, \nu(d\lambda) \quad &\le \quad \int_{\mathcal{I}} (\lambda - \mu_1)^2\, \nu(d\lambda) + (M - m)^2\epsilon \\ &= \int_{\mathcal{I}} (\lambda - \mu_{\mathcal{I}})^2\, \nu(d\lambda) + (\mu_1 - \mu_{\mathcal{I}})^2 \nu(\mathcal{I}) + (M - m)^2\epsilon\,. \end{aligned}$$

Lemma 1 implies $\int_{\mathcal{I}} (\lambda - \mu_{\mathcal{I}})^2\, \nu(d\lambda) \le \alpha^2/4$ and $(\mu_1 - \mu_{\mathcal{I}})^2 \le \epsilon^2 M^2$ gives

$$\mathrm{Var}(\nu) \le \alpha^2/4 + \epsilon^2 M^2 + M^2\epsilon \le \alpha^2/4 + 2\epsilon M^2\,.$$

■

Lemma 3 *Let ν be any probability distribution on $[m, M]$, $0 < m \le M < \infty$. Assume that* $\mathrm{Var}(\nu) \le \epsilon$. *Then, there exist an interval $\mathcal{I}$ such that $|\mathcal{I}| \le \epsilon^{1/4}$ and $\nu(\mathcal{I}) \ge 1 - 4\sqrt{\epsilon}$*

Proof. Take $\mathcal{I} = [\mu_1 - \epsilon^{1/4}/2, \mu_1 + \epsilon^{1/4}/2]$, $\mu_1 = \int \lambda\, \nu(d\lambda)$, and apply the Chebyshev inequality. ■

Lemma 4 *Let ν be any distribution on $[m, M]$, $0 < m \le M < \infty$. Define $\mu_i = \int \lambda^i\, \nu(d\lambda)$ and*

$$\mathbf{M} = \begin{pmatrix} \mu_{-1} & \mu_0 & \mu_1 \\ \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \end{pmatrix} .$$

Assume that $L = \mu_{-1}\mu_1 > 1$ (which, by Jensen's inequality, holds when ν is not degenerate at a single point) and $\det \mathbf{M} < \epsilon$. Then, there exist two intervals $\mathcal{I}_1$ and $\mathcal{I}_2$ such that

$$\begin{aligned} &(i) \quad |\mathcal{I}_i| \le \frac{(M-m)\epsilon^{1/4}}{m^{9/4}(L-1)^{3/2}}\,,\ i = 1, 2\,,\ \nu(\mathcal{I}_1) + \nu(\mathcal{I}_2) \ge 1 - 4\sqrt{\epsilon}M^{3/2}\,, \\ &(ii) \quad \max_{x\in\mathcal{I}_i} |x - \mu_{-1}| > \frac{3(L-1)m^2}{4(M-m)}\,,\ i = 1, 2\,, \qquad (26) \\ &(iii) \quad \text{for } \epsilon < \epsilon_* = \frac{4(L-1)^8 M^8 m^{16}}{[32(L-1)^3 + M^4(M-m)^2]^2}\,, \\ &\qquad \nu(\mathcal{I}_i) \ge \frac{m^2(L-1)}{4M^2}\,,\ i = 1, 2\,, \qquad (27) \\ &\qquad \text{and} \quad \max_{x\in\mathcal{I}_1, y\in\mathcal{I}_2} |x - y| > m\sqrt{2(L-1)}\,. \end{aligned}$$

Proof.
(i) Consider the measure ν' defined by $\nu'(\mathcal{A}) = (1/\mu_{-1}) \int_{\mathcal{A}} (1/\lambda)\nu(d\lambda)$ for any Borel set $\mathcal{A} \subset [m, M]$, and denote its moments by $\mu'_i = (1/\mu_{-1}) \int \lambda^{i-1}\nu(d\lambda) = \mu_{i-1}/\mu_{-1}$. Note that for any Borel set $\mathcal{A}$

$$\frac{1}{M\mu_{-1}}\,\nu(\mathcal{A}) \leq \nu'(\mathcal{A}) \leq \frac{1}{m\mu_{-1}}\,\nu(\mathcal{A})\,.$$

We have

$$\mathbf{M}' = \begin{pmatrix} \mu'_0 & \mu'_1 & \mu'_2 \\ \mu'_1 & \mu'_2 & \mu'_3 \\ \mu'_2 & \mu'_3 & \mu'_4 \end{pmatrix} = \mathbf{M}/\mu_{-1}$$

and thus $\det \mathbf{M}' = \det \mathbf{M}/\mu_{-1}^3$. Also define $D' = \mu'_2 - (\mu'_1)^2$, $a = \sqrt{D'}$, $b = (\mu'_1\mu'_2 - \mu'_3)/\sqrt{D'}$, $c = a\mu'_2 + b\mu'_1 = [(\mu'_2)^2 - \mu'_1\mu'_3]/\sqrt{D'}$ (note that $a > 0$, $b < 0$ and $c < 0$) and $\eta = F(\lambda) = a\lambda^2 + b\lambda - c$, with λ having the distribution ν'. Direct calculation gives $E'\{\eta\} = \int \eta(\lambda)\nu'(d\lambda) = 0$ and $\mathrm{Var}'(\eta) = E'\{\eta^2\} - (E'\{\eta\})^2 = \det \mathbf{M}'$, so that $\det \mathbf{M} < \epsilon$ implies $\mathrm{Var}'(\eta) < \epsilon' = \epsilon/\mu_{-1}^3$. From Lemma 3, the interval $\mathcal{I} = [-(\epsilon')^{1/4}/2, (\epsilon')^{1/4}/2]$ is such that $\mathrm{Prob}\{\eta \in \mathcal{I}\} \geq 1 - 4\sqrt{\epsilon'}$. Also, from the mean-value theorem, there exist $\lambda_1 < \lambda_2$ such that $\lambda_i \in [m, M]$ and $a\lambda_i^2 + b\lambda_i - c = 0$, $i = 1, 2$. Direct calculation gives $F(\mu'_1) = F(\mu_{-1}) = a(\mu'_1)^2 + b\mu'_1 - c = -(D')^{3/2}$, and thus

$$m \leq \lambda_1 < 1/\mu_{-1} < \lambda_2 \leq M\,.$$

Take $\beta = (M - m)(\epsilon')^{1/4}/[2(D')^{3/2}]$, we get

$$\begin{aligned} F(\lambda_1 + \beta) &< -(\epsilon')^{1/4}/2\,, F(\lambda_1 - \beta) > (\epsilon')^{1/4}/2\,, \\ F(\lambda_2 + \beta) &> (\epsilon')^{1/4}/2\,, F(\lambda_2 - \beta) < -(\epsilon')^{1/4}/2\,, \end{aligned}$$

and $\nu(\mathcal{I}_1) + \nu(\mathcal{I}_2) \geq 1 - 4\sqrt{\epsilon}M^{3/2}$ when $\mathcal{I}_i = [\lambda_i - \beta, \lambda_i + \beta]$, $i = 1, 2$, with $|\mathcal{I}_i| = 2\beta = (M - m)\epsilon^{1/4}\mu_{-1}^{9/4}/(L - 1)^{3/2} \leq (M - m)\epsilon^{1/4}/[m^{9/4}/(L - 1)^{3/2}]$.
(ii) Define $y_1 = \mu'_1 - \lambda_1$, $y_2 = \lambda_2 - \mu'_1$, so that $\max_{x\in\mathcal{I}_1} |x - \mu_{-1}| > y_1$ and $\max_{x\in\mathcal{I}_2} |x - \mu_{-1}| > y_2$. We have $F(\lambda) = a(\lambda - \lambda_1)(\lambda - \lambda_2)$ and thus $y_1y_2 = -F(\mu'_1)/a = D'$. Also, $|y_2 - y_1| < y_1 + y_2 \leq M - m$, so that $D' > y_i(y_i + M - m)$, $i = 1, 2$, and thus

$$y_i > \frac{M - m}{2}\left[\sqrt{1 + \frac{4D'}{(M - m)^2}} - 1\right] > \frac{D'}{M - m}\left(1 - \frac{D'}{(M - m)^2}\right)\,,\ i = 1, 2\,.$$

Lemma 1 gives $D' < (M - m)^2/4$, so that $y_i > 3D'/[4(M - m)] > 3(L - 1)m^2/[4(M - m)]$, $i = 1, 2$.
(iii) Define $\gamma = \nu'(\mathcal{I}_2)$, part (i) implies $\nu'(\mathcal{I}_1) > 1 - 4\sqrt{\epsilon'} - \gamma$, and from Lemma 2

$$D' \leq \frac{(M - m)^2\sqrt{\epsilon'}}{4(D')^3} + 2(4\sqrt{\epsilon'} + \gamma)M^2\,,$$

which gives

$$\gamma \geq \frac{D'}{2M^2} - \sqrt{\epsilon'}\left[\frac{(M - m)^2}{8(D')^3M^2} + 4\right]\,,$$

and thus $\gamma \geq D'/(4M^2) > (L - 1)m^2/[4M^2]$ for $\epsilon < \epsilon_* < [4(D')^8]/[(M - m)^2 + 32(D')^3M^2]^2$, see (27).

Define now $\Delta = \max_{x\in\mathcal{I}_1, y\in\mathcal{I}_2} |x - y|$. Lemma 2 gives $D' \leq \Delta^2/4 + 8\sqrt{\epsilon'}M^2$, which implies $\Delta^2 \geq 4D' - 32M^2\sqrt{\epsilon'}$. Since $\epsilon < \epsilon_*$ implies $\sqrt{\epsilon'} < D'/(16M^2)$, we get $\Delta^2 > 2(L - 1)m^2$. ■

Proof of Theorem 4. The proof follows the same lines as that of Theorem 3 and is divided into four parts. In (i), we construct sequences of intervals $\mathcal{L}_k = [m_k, m_k + \delta]$ and $\mathcal{R}_k = [M_k - \delta, M_k]$ in which the measure ν_k will tend to concentrate. In (ii) we prove that $\mathcal{R}_k \cap \mathcal{R}_{k+1} \neq \emptyset$ and in (iii) that the sequence M_k is non decreasing. Finally, the limiting behaviour of ν_k is derived in (iv).
(i) We have seen in Section 2.4 that $\det \mathbf{M}_k \to 0$ as $k \to \infty$, with $\mathbf{M}_k$ given by (10). Therefore, given ϵ, $\exists K_\epsilon$ such that $\forall k > K_\epsilon$, $\det \mathbf{M}_k < \epsilon$. Define $L_k = \mu_1^k\mu_{-1}^k$ and note that $L_k > 1$ because no ν_k is degenerate at a single point. Using Lemma 4, for ϵ small enough, for any $k > K_\epsilon$ there exist two intervals $\mathcal{I}_1^k$, $\mathcal{I}_2^k$, with width at most

$$\delta = \delta(\epsilon) = \frac{(M - m)\epsilon^{1/4}}{m^{9/4}(L_0 - 1)^{3/2}}\,,$$

and such that $\nu_k(\mathcal{I}_1^k) + \nu_k(\mathcal{I}_2^k) \geq 1 - 4\sqrt{\epsilon}M^{3/2}$, $\nu_k(\mathcal{I}_1^k) \geq m^2(L_k - 1)/(4M^2)$, $\nu_k(\mathcal{I}_2^k) \geq m^2(L_k - 1)/(4M^2)$. Also, $\max_{x\in\mathcal{I}_1^k, y\in\mathcal{I}_2^k} |x - y| \geq m\sqrt{2(L_k - 1)}$. Without any loss of generality, assume that $\mathcal{I}_1^k$ is the interval on the left. Define $\mathcal{L}(x) = [x, x + \delta]$, $\mathcal{R}(x) = [x - \delta, x]$,

$$\begin{aligned}
\mathcal{X}_L^k &= \operatorname{Arg}\max_x \{\nu_k[\mathcal{L}(x)]\,,\, \mathcal{L}(x) \cap \mathcal{I}_1^k \neq \emptyset\}\,, \\
\mathcal{X}_R^k &= \operatorname{Arg}\max_x \{\nu_k[\mathcal{R}(x)]\,,\, \mathcal{R}(x) \cap \mathcal{I}_2^k \neq \emptyset\}\,,
\end{aligned}$$

and $m_k = \min \mathcal{X}_L^k$, $M_k = \max \mathcal{X}_R^k$, $\mathcal{L}_k = \mathcal{L}(m_k)$, $\mathcal{R}_k = \mathcal{R}(M_k)$; that is, M_k is the right endpoint of an interval $\mathcal{R}_k$, intersecting $\mathcal{I}_2^k$, with maximum measure, and similarly for m_k and $\mathcal{L}_k$. Note that $\nu_k(\mathcal{L}_k) + \nu_k(\mathcal{R}_k) \geq 1 - 4\sqrt{\epsilon}M^{3/2}$, $\nu_k(\mathcal{L}_k) \geq m^2(L_k - 1)/(4M^2)$ and $\nu_k(\mathcal{R}_k) \geq m^2(L_k - 1)/(4M^2)$. The situation is the same for the two sequences of intervals $(\mathcal{L}_k)$ and $(\mathcal{R}_k)$, and we concentrate on $(\mathcal{R}_k)$ in the rest of the proof.

(ii) We show now that $\mathcal{R}_k \cap \mathcal{R}_{k+1} \neq \emptyset$. Again for ϵ small enough $\mu_1^k \notin \mathcal{R}_k$ and $\lambda - \mu_1^k \geq M_k - \delta - \mu_1^k$ on $\mathcal{R}_k$ so that

$$\begin{aligned}
\nu_{k+1}(\mathcal{R}_k) &= \int_{\mathcal{R}_k} \frac{(\lambda - \mu_1^k)^2}{D_k}\, \nu_k(d\lambda) \geq \frac{\nu_k(\mathcal{R}_k)}{D_k}(M_k - \delta - \mu_1^k)^2 \\
&\geq \frac{m^2(L_k - 1)}{4M^2 D^*}(M_k - \delta - \mu_1^k)^2
\end{aligned}$$

with D^* the maximum possible value of D_k, $D^* = (M - m)^2/4$, see Lemma 1. By construction, $\max_{x\in\mathcal{I}_k^2} |x - \mu_1^k| \leq M_k + \delta - \mu_1^k$, and thus, from Lemma 4,

$$M_k - \mu_1^k + \delta > \frac{3m^2(L_k - 1)}{4(M - m)} \geq \frac{3m^2(L_0 - 1)}{4(M - m)} = C\,. \tag{28}$$

Choosing ϵ such that $\delta < C/4$ gives $M_k - \delta - \mu_1^k > C/2$ and thus

$$\nu_{k+1}(\mathcal{R}_k) > \frac{m^2(L_k - 1)}{4M^2 D^*}\,\frac{C^2}{4} \geq \nu_R^* = \frac{9m^6(L_0 - 1)^3}{16M^2(M - m)^4}\,.$$

Choosing now ϵ such that $4\sqrt{\epsilon}M^{3/2} < \nu_R^*$ we obtain $\mathcal{R}_k \cap \mathcal{R}_{k+1} \neq \emptyset$ for any $k > K_\epsilon$.

(iii) We prove now that the sequence (M_k) is not decreasing starting at some K_ϵ for ϵ small enough. Take $k > K_\epsilon$ and assume that $M_{k+1} = M_k - \beta$, $\beta > 0$. Then note that $\beta < \delta$ since $\mathcal{R}_k \cap \mathcal{R}_{k+1} \neq \emptyset$ by (ii) above. Consider the difference $\nu_{k+1}(\mathcal{R}_k) - \nu_{k+1}(\mathcal{R}_{k+1}) = \nu_{k+1}([M_k - \beta, M_k]) - \nu_{k+1}([M_k - \delta - \beta, M_k - \delta])$. Assume first that $\nu_{k+1}([M_k - \delta - \beta, M_k - \delta]) = 0$, then $\nu_{k+1}(\mathcal{R}_k) > \nu_{k+1}(\mathcal{R}_{k+1})$, which is impossible by construction. We can thus consider the following ratio

$$\begin{aligned}
\frac{\nu_{k+1}([M_k - \beta, M_k])}{\nu_{k+1}([M_k - \delta - \beta, M_k - \delta])} &= \frac{\int_{M_k-\beta}^{M_k} (\lambda - \mu_1^k)^2\, \nu_k(d\lambda)}{\int_{M_k-\delta-\beta}^{M_k-\delta} (\lambda - \mu_1^k)^2\, \nu_k(d\lambda)} \\
&\geq \frac{(M_k - \beta - \mu_1^k)^2}{(M_k - \delta - \mu_1^k)^2}\,\frac{\nu_k([M_k - \beta, M_k])}{\nu_k([M_k - \delta - \beta, M_k - \delta])}\,.
\end{aligned}$$

Since $M_k - \delta - \mu_1^k \geq C - 2\delta \geq 2\delta$ for $C > 4\delta$, see (28), and $\beta < \delta$, $(M_k - \beta - \mu_1^k)^2 > (M_k - \delta - \mu_1^k)^2$. Also, by construction,

$$\begin{aligned}
0 &\leq \nu_k([M_k - \beta, M_k]) - \nu_k([M_k - \delta - \beta, M_k - \delta]) \\
&= \nu_k([M_k - \beta, M_k]) - \nu_k([M_k - \delta - \beta, M_k - \delta])\,.
\end{aligned}$$

This gives

$$\frac{\nu_{k+1}([M_k - \beta, M_k])}{\nu_{k+1}([M_k - \delta - \beta, M_k - \delta])} > 1\,.$$

Therefore, $\beta > 0$ leads to $\nu_{k+1}(\mathcal{R}_k) > \nu_{k+1}(\mathcal{R}_{k+1})$, which is impossible. We thus obtain $M_{k+1} \geq M_k$ for $k > K_\epsilon$.

(iv) Since the sequence (M_k) is non decreasing and bounded from above (by M), it has a limit $M_* \geq M$. The same is true for m_k, and $m_k \to m_*$ as $k \to \infty$. We have thus proved that for any δ small enough and any k larger than some K_δ,

$$\nu_k([M_* - \delta, M_*]) + \nu_k([m_*, m_* + \delta]) \geq 1 - \frac{4M^{3/2}m^{9/2}(L_0 - 1)^3\delta^2}{(M - m)^2}\,.$$

Assume that $M_* < M$. This would imply $\nu_k([M-\delta, M]) \to 0$ as $k \to \infty$ for $\delta < M - M_*$. On the other hand,

$$\frac{\nu_{k+1}([M-\delta, M])}{\nu_{k+1}([M_*-\delta, M_*])} > \frac{\nu_k([M-\delta, M])}{\nu_k([M_*-\delta, M_*])},$$

which leads to a contradiction since $\nu_k([M-\delta, M])/\nu_k([M_*-\delta, M_*])$ is then increasing and $\nu_k([M_*-\delta, M_*])$ is bounded from below. Therefore, $M_* = M$, and similarly $m_* = m$, with, for δ small enough and any k larger than some K_δ, $\nu_k([m+\delta, M-\delta]) < 4M^{3/2}m^{9/2}(L_0-1)^3\delta^2/(M-m)^2$. Finally, from Helly's Theorem, see [18], p. 319, from the sequence (ν_k) we can extract a subsequence (ν_{k_i}) that is weakly convergent, and from the result above the associated limit has necessarily the form ν_p^*, where ν_p^* is the discrete measure concentrated on the two points m, M, with $\nu_p^*(m) = p$, $\nu_p^*(M) = 1-p$. Since L_{k_i} converges to some L, ν_p^* is such that the associated value of $\mu_1\mu_{-1}$ is equal to L, which only leaves two possibilities for p (and $1-p$):

$$p = \frac{1}{2} \pm \frac{\rho+1}{\rho-1}\sqrt{\frac{1}{4} - \frac{\rho L}{(\rho+1)^2}}$$

where $\rho = M/m$. Applying the transformation T, we get $\nu_{k_i+1} = T(\nu_{k_i}) \to T(\nu_p^*) = \nu_{1-p}^*$. ■

A4. *Proof of Theorem 5.*
(i) It is straightforward to check that $T^2(\nu_p^*) = \nu_p^*$, $\forall p \in (0,1)$.
(ii) We assume that $\mathcal{S}_A$ is not reduced to $\{m, M\}$ (otherwise $\mathcal{I}_u = \emptyset$). We have $\nu_{k+2}(d\lambda) = H(\nu_k, \lambda)\nu_k(d\lambda)$, with

$$H(\nu_k, \lambda) = \frac{(\lambda - \mu_1^k)^2(\lambda - \mu_1^{k+1})^2}{D_k D_{k+1}} \tag{29}$$

see (18), with μ_1^k, D_k defined as in Theorem 4. For $\nu_k = \nu_p^*$, it gives

$$H(\nu_p^*, \lambda) = \frac{[M(1-p) + mp - \lambda]^2[Mp + m(1-p) - \lambda]^2}{p^2(1-p)^2(M-m)^4}. \tag{30}$$

One can then check that for any $p \in \mathcal{I}_u$, $\max_{\lambda\in\mathcal{S}_A} H(\nu_p^*, \lambda) = H(\nu_p^*, \lambda^*) > 1$, with $\lambda^* = \min_{\lambda\in\mathcal{S}_A} s(\lambda)$. Therefore, for any $p \in \mathcal{I}_u$, one can choose ϵ small enough, such that $d(\nu_k, \nu_p^*) < \epsilon$ implies $\nu_{k+2}([a,b]) > K_p\nu_k([a,b])$, for some $K_p > 1$ and some a, b such that $m+\epsilon < a < b < M-\epsilon$ and $[a,b] \cap \mathcal{S}_A \neq \emptyset$. For any $\alpha > 0$, $\alpha < 1-p$, take an initial measure ν_0 putting weight p at m, $1-p-\alpha$ at M and α in the interval $[a,b]$. It satisfies $d(\nu_0, \nu_p^*) < \alpha$, and, for any m, either $d(\nu_{2m}, \nu_p^*) > \epsilon$ or $\nu_{2(m+1)}([a,b]) > K_p\nu_{2m}([a,b])$. The later case gives $\nu_{2m}([a,b]) > 2\epsilon$, and thus $d(\nu_{2m}, \nu_p^*) > \epsilon$, as soon as $m > \log(2\epsilon/\alpha)/\log(K_p)$, which shows that ν_p^* is unstable.
(iii) Part (a) concerns the case where a spectral gap is present, with point mass at m and M. The proof for the general situation is more technical and is sketched in part (b).

(a) Assume that the measure ν_0 presents a spectral gap: $\nu_0 = 0$ on $(m, m+s)$ and $(M-s, M)$ for some $s > 0$. Take $\gamma < s$ and assume that $d(\nu_0, \nu_p^*) < \alpha < \gamma$ with $p \in \mathcal{I}_s$. The arguments go as follows. First we bound $\nu_2\{(m+\gamma, M-\gamma]\}$ by $2K_0\alpha$ for some $K_0 < 1$, then we bound $\nu_2\{(M-\gamma, M]\}$ by $1-p+K_1\alpha$ for some $K_1 < \infty$. We show that $d(\nu_2, \nu_{p_2}^*) < K_0\alpha$ for some p_2 such that $|p_2 - p| < (K_0+K_1)\alpha$. Stability will then follow by an induction argument.

The maximum value of $H(\nu_0, \lambda)$ for λ varying in $[m+\gamma, M-\gamma]$ may be reached for some $\lambda^* \in (\mu_1^0, \mu_1^1)$ or at one of the two points $m+\gamma$, $M-\gamma$. Now, for α small enough $H(\nu_0, \lambda)$ will be close to $H(\nu_p^*, \lambda)$ given by (30), and $p \in \mathcal{I}_s$ implies

$$\max_{\lambda\in\mathcal{S}_A\cap(\mu_1^0,\mu_1^1)} H(\nu_0, \lambda) < 1. \tag{31}$$

Consider the situation at $M-\gamma$. We can write

$$H(\nu_0, M-\gamma) = H(\nu_p^*, M) - \gamma\,\frac{dH(\nu_p^*, \lambda)}{d\lambda}_{|\lambda=M} + F_H(\nu_p^*; \nu_0, M) + \mathcal{O}(\gamma^2), \tag{32}$$

with $F_H(\nu_p^*; \nu_0, M)$ the directional derivative of $H(\nu, M)$ at ν_p^* in the direction ν_0,

$$F_H(\nu_p^*; \nu_0, M) = \lim_{\beta\to 0^+} \frac{H[(1-\beta)\nu_p^* + \beta\nu_0, M] - H(\nu_p^*, M)}{\beta},$$

Define $F_H(\nu_p^*, x, \lambda) = F_H(\nu_p^*; \delta_x, \lambda)$ with δ_x the delta measure supported at x. We have

$$F_H(\nu_p^*; \nu_0, M) = \int_m^M F_H(\nu_p^*, x, M)\nu_k(dx)\,,$$

which we decompose in three parts:

$$\begin{aligned} F_H(\nu_p^*; \nu_0, M) &= \int_m^{m+\gamma} F_H(\nu_p^*, x, M)\nu_0(dx) + \int_{m+\gamma}^{M-\gamma} F_H(\nu_p^*, x, M)\nu_0(dx) \\ &\quad + \int_{M-\gamma}^{M} F_H(\nu_p^*, x, M)\nu_0(dx)\,. \end{aligned}$$

Direct calculation gives

$$F_H(\nu_p^*, x, M) = \frac{(x-m)^2(M-x)[x-m+(2p-1)(M-m)]}{p^2(1-p)^2(M-m)^4}$$

so that $F_H(\nu_p^*, m, M) = F_H(\nu_p^*, M, M) = 0$ and $F_H(\nu_p^*; \nu_0, M) < F^*\nu_0\{(m+\gamma, M-\gamma]\}$ with

$$F^* = \max_{p\in\mathcal{I}_s,\, x\in[m,M]} F_H(\nu_p^*, x, M) < \infty\,.$$

Also, $d(\nu_0, \nu_p^*) < \alpha$ implies $\nu_0\{(m+\gamma, M-\gamma]\} = \nu_0\{(m+\alpha, M-\alpha]\} < 2\alpha$, so that $F_H(\nu_p^*; \nu_0, M) < 2\alpha F^*$. Now,

$$H(\nu_p^*, M) = 1\,,\ \frac{dH(\nu_p^*, \lambda)}{d\lambda}_{|\lambda=M} = \frac{2}{p(1-p)(M-m)}\,,$$

which, together with (32) gives for γ small enough

$$H(\nu_0, M-\gamma) < 1 + 2\alpha F^* - \frac{\gamma}{p(1-p)(M-m)}$$

and thus

$$H(\nu_0, M-\gamma) < 1 - \frac{\gamma}{2p(1-p)(M-m)}$$

for $\alpha < \gamma/[4F^*p(1-p)(M-m)]$.

The situation is similar at $m+\gamma$. Together with (31) this implies for α small enough

$$\max_{\lambda\in\mathcal{S}_A\cap[m+\gamma, M-\gamma]} H(\nu_0, \lambda) < K_0 < 1$$

and therefore,

$$\nu_2\{(m+\gamma, M-\gamma]\} < 2K_0\alpha \tag{33}$$

with $K_0 < 1$ not depending on α.

Consider now the interval $(M-\gamma, M]$. We have

$$\nu_2\{(M-\gamma, M]\} = \nu_2(M) = H(\nu_0, M)\nu_0\{(M-\gamma, M]\} = H(\nu_0, M)\nu_0(M)\,,$$

with $d(\nu_0, \nu_p^*) < \alpha$ implying $\nu_0(M) < 1-p+\alpha$, and

$$H(\nu_0, M) = H(\nu_p^*, M) + F_H(\nu_p^*; \nu_0, M) + \mathcal{O}(\alpha^2) < 1 + 2\alpha F^* + \mathcal{O}(\alpha^2)\,.$$

This gives for α small enough

$$\nu_2\{(M-\gamma, M]\} < 1 - p + K_1\alpha$$

for some $K_1 < \infty$. Similarly, $\nu_2\{[m, m+\gamma]\} < p + K_1\alpha$.

Define $p_0 = p$, $p_2 = [\nu_2(m) - \nu_2(M) + 1]/2$, $\alpha_0 = \alpha$. We obtain

$$-K_0\alpha_0 < \nu_2(m) - p_2 < 0\,,\ -K_0\alpha_0 < \nu_2(M) - (1-p_2) < 0$$

which together with (33) implies

$$d(\nu_2, \nu_p^*) < \alpha_2 = K_0\alpha_0\,.$$

Moreover, $|p_2 - p_0| < (K_0 + K_1)\alpha_0$.

For α small enough, $p_2 \in \mathcal{I}_s$ and we can then repeat the same arguments. This gives for any m

$$d(\nu_{2m}, \nu^*_{p_{2m}}) < \alpha_{2m} = K_0^m \alpha$$

with

$$|\, p_{2m} - p \,| < (K_0 + K_1) \sum_{i=0}^{m-1} \alpha_{2i} = (K_0 + K_1) \frac{1 - K_0^m}{1 - K_0}\, \alpha < \frac{K_0 + K_1}{1 - K_0}\, \alpha$$

and $p_{2m} \in \mathcal{I}_s$, for α small enough. For any $p \in \mathcal{I}_s$ and any $\epsilon > 0$, taking ν_0 such that $d(\nu_0, \nu^*_p) < \alpha$ with α small enough thus implies $d(\nu_{2m}, \nu^*_p) < \epsilon$ for any m, and ν^*_p is thus stable.

(b) Consider now the general situation. The proof follows the same lines as in case (a), but more technicalities are required since we need to consider measures of intervals of the form $[m, m + \gamma]$ and $(M - \gamma, M]$, with γ decreasing in a suitable way as the number of iterations of the mapping T^2 increases.

Assume that

$$\begin{aligned}
&\nu_{2k}\{(m + \gamma_{2k}, M - \gamma_{2k}]\} < 2\alpha_{2k}\,, \\
&\nu_{2k}\{[m, m + \gamma_{2k}]\} < p_{2k} + \alpha_{2k}\,, \\
&\nu_{2k}\{(M - \gamma_{2k}, M]\} < 1 - p_{2k} + \alpha_{2k}\,.
\end{aligned}$$

for some $p_{2k} \in \mathcal{I}_s$ and some α_{2k}, γ_{2k}. Note that it implies $d(\nu_{2k}, \nu^*_{p_{2k}}) < \gamma_{2k}$ and that for $k = 0$, α_0, γ_0 can be chosen arbitrarily small, with $d(\nu_0, \nu^*_p) < \alpha_0$ for some $p \in \mathcal{I}_s$.

Consider one application of the mapping T^2 at a generic iteration k. We can write $H(\nu_{2k}, M) = H(\nu^*_{p_{2k}}, M) + F_H(\nu^*_{p_{2k}}; \nu_{2k}, M) + \mathcal{O}(\gamma_{2k}^2)$ with

$$\begin{aligned}
F_H(\nu^*_{p_{2k}}; \nu_{2k}, M) \;=\; & \int_m^{m+\gamma_{2k}} F_H(\nu^*_{p_{2k}}, x, M)\nu_{2k}(dx) \\
& + \int_{m+\gamma_{2k}}^{M-\gamma_{2k}} F_H(\nu^*_{p_{2k}}, x, M)\nu_{2k}(dx) \\
& + \int_{M-\gamma_{2k}}^{M} F_H(\nu^*_{p_{2k}}, x, M)\nu_{2k}(dx)\,.
\end{aligned}$$

The first integral term is of the order $\mathcal{O}(\gamma_{2k}^2)$ (since $F_H(\nu^*_{p_{2k}}, m, M) = 0$ and $dF_H(\nu^*_{p_{2k}}, z, M)/dz_{|z=m} = 0$), the second is bounded by $2\alpha_{2k}F^* + \mathcal{O}(\gamma_{2k}^2)$, as in case (a). For the third term, for which x is close to M, we can use the linear approximation

$$\begin{aligned}
F_H(\nu^*_{p_{2k}}, x, M) \;&=\; (x - M)\frac{dF_H(\nu^*_{p_{2k}}, z, M)}{dz}_{|z=M} + \mathcal{O}(\gamma_{2k}^2) \\
&=\; \frac{-2(x - M)}{p_{2k}(1 - p_{2k})^2(M - m)} + \mathcal{O}(\gamma_{2k}^2)
\end{aligned}$$

which gives

$$\int_{M-\gamma_{2k}}^{M} F_H(\nu^*_{p_{2k}}, x, M)\nu_{2k}(dx) = \frac{2}{p_{2k}(1 - p_{2k})^2(M - m)} I_{2k}(M) + \mathcal{O}(\gamma_{2k}^2)$$

where $I_{2k}(M) = \int_0^{\gamma_{2k}} z\nu'_{2k}(dz)$ with ν'_{2k} the measure obtained after applying the transformation $x \mapsto z = M - x$. We have thus obtained

$$H(\nu_{2k}, M) < 1 + 2\alpha_{2k}F^* + \frac{2I_{2k}(M)}{p_{2k}(1 - p_{2k})^2(M - m)} + \mathcal{O}(\gamma_{2k}^2)\,. \tag{34}$$

Consider now the behavior of $I_{2k}(M)$ as k increases. We assume that ν_{2k} remains in some neighborhood $\mathcal{V}(p)$ of ν^*_p, which we shall be able to guarantee afterwards. Define $A_{2k}(M) = I_{2k}(M)[\int_0^{\gamma_{2k}} \nu'_{2k}(dz)]^{-1}$. It satisfies $I_{2k}(M) < A_{2k}(M) < \gamma_{2k}$. Also, $\gamma_{2(k+1)} < \gamma_{2k}$ implies

$$A_{2(k+1)}(M) = \frac{\int_0^{\gamma_{2(k+1)}} zH(\nu_{2k}, M - z)\nu'_{2k}(dz)}{\int_0^{\gamma_{2(k+1)}} H(\nu_{2k}, M - z)\nu'_{2k}(dz)} < \frac{\int_0^{\gamma_{2k}} zH(\nu_{2k}, M - z)\nu'_{2k}(dz)}{\int_0^{\gamma_{2k}} H(\nu_{2k}, M - z)\nu'_{2k}(dz)}\,,$$

and, since $H(\nu_{2k}, M-z)$ decreases for z close to zero,

$$A_{2(k+1)}(M) < \frac{\int_0^{\gamma_{2k}} z \frac{H(\nu_{2k},M-z)}{H(\nu_{2k},M)} \nu'_{2k}(dz)}{\int_0^{\gamma_{2k}} \nu'_{2k}(dz)} .$$

We can bound the speed of decrease of $H(\nu_{2k}, M-z)$: $H(\nu, M-z)/H(\nu, M) < 1 - az$ for some $a > 0$, any z in $[0, \gamma_0]$ and any $\nu \in \mathcal{V}(p)$. This gives

$$A_{2(k+1)}(M) < \frac{\int_0^{\gamma_{2k}} z(1-az)\nu'_{2k}(dz)}{\int_0^{\gamma_{2k}} \nu'_{2k}(dz)} .$$

Repeating the same arguments we get for any $n > 0$,

$$A_{2(k+n)}(M) < \bar{A}_{2(k+n)}(M) = \frac{\int_0^{\gamma_{2k}} z(1-az)^n \nu'_{2k}(dz)}{\int_0^{\gamma_{2k}} \nu'_{2k}(dz)} ,$$

with $\bar{A}_{2(k+n)}(M)$ decreasing with n. Direct calculation gives $\sum_{n=0}^{\infty} \bar{A}_{2(k+n)}(M) = 1/a$, and therefore $I_{2k}(M) < \bar{A}_{2k}(M) = o(1/k)$.

Similarly to case (a), we can write

$$H(\nu_{2k}, M - \gamma_{2(k+1)}) = H(\nu_{2k}, M) - \frac{2\gamma_{2(k+1)}}{p_{2k}(1-p_{2k})(M-m)} + \mathcal{O}(\gamma_{2k}^2) ,$$

with $H(\nu_{2k}, M)$ bounded by (34). Assume that γ_{2k} is such that $\bar{A}_{2k}(M) = o(\gamma_{2k})$ and $\alpha_{2k} = o(\gamma_{2k})$. We obtain for p_{2k} close enough to p

$$H(\nu_{2k}, M - \gamma_{2(k+1)}) < \beta_{2(k+1)} = 1 - \frac{\gamma_{2(k+1)}}{p(1-p)(M-m)} . \tag{35}$$

We thus get the following bounds on the measure of subintervals of interest at the next iteration:

$$\nu_{2(k+1)}\{(m + \gamma_{2(k+1)}, M - \gamma_{2(k+1)}]\} < 2\max\{\beta_{2(k+1)}, K_0\}\alpha_{2k} \tag{36}$$

where $K_0 = \max_{\nu_{2k} \in \mathcal{V}(p)} \max_{\lambda \in \mathcal{S}_A \cap (\mu_1^{2k}, \mu_1^{2k+1})} H(\nu_{2k}, \lambda)$, and $K_0 < 1$ for p in $\mathcal{I}_s$ and $\mathcal{V}(p)$ small enough, see part (a);

$$\begin{aligned}
\nu_{2(k+1)}\{(M - \gamma_{2(k+1)}, M]\} \quad &< \quad \nu_{2(k+1)}\{(M - \gamma_{2k}, M]\} \\
&< H(\nu_{2k}, M)\nu_{2k}\{(M - \gamma_{2k}, M]\} \\
&< \left[1 + 2\alpha_{2k}F^* + \frac{2\bar{A}_{2k}(M)}{p_{2k}(1-p_{2k})^2(M-m)} + \mathcal{O}(\gamma_{2k}^2)\right] \nu_{2k}\{(M - \gamma_{2k}, M]\} \\
&\qquad < \nu_{2k}\{(M - \gamma_{2k}, M]\} + B\alpha_{2k} + C\bar{A}_{2k}(M) + D\gamma_{2k}^2
\end{aligned}$$

for some $B, C, D < \infty$. Similarly, we obtain

$$\nu_{2(k+1)}\{[m, m + \gamma_{2(k+1)}]\} < \nu_{2k}\{[m, m + \gamma_{2k}]\} + B\alpha_{2k} + C\bar{A}_{2k}(m) + D\gamma_{2k}^2$$

where $\bar{A}_{2k}(m)$ is defined similarly to $\bar{A}_{2k}(M)$. Define $p_{2(k+1)}$ as

$$p_{2(k+1)} = \frac{\nu_{2(k+1)}\{[m, m + \gamma_{2(k+1)}]\} - \nu_{2(k+1)}\{(M - \gamma_{2(k+1)}, M]\} + 1}{2} ,$$

it gives

$$\begin{aligned}
&0 < p_{2(k+1)} - \nu_{2(k+1)}\{[m, m + \gamma_{2(k+1)}]\} < \max\{\beta_{2(k+1)}, K_0\}\alpha_{2k} , \\
&0 < 1 - p_{2(k+1)} - \nu_{2(k+1)}\{(M - \gamma_{2(k+1)}, M]\} < \max\{\beta_{2(k+1)}, K_0\}\alpha_{2k} .
\end{aligned}$$

Together with (36) it implies $d(\nu_{2(k+1)}, \nu^*_{p_{2(k+1)}}) < \gamma_{2(k+1)} < \gamma_{2k}$, with

$$|p_{2(k+1)} - p_{2k}| < \Delta_{2k} = [B + 1 + \max\{\beta_{2(k+1)}, K_0\}]\alpha_{2k} + C\bar{A}'_{2k} + D\gamma_{2k}^2 ,$$

where $\bar{A}'_{2k} = \max\{\bar{A}_{2k}(m), \bar{A}_{2k}(M)\}$ and $\sum_k \bar{A}'_{2k} < \infty$.

Define $\alpha_{2(k+1)} = \max\{\beta_{2(k+1)}, K_0\}\alpha_{2k}$ and take $\gamma_{2k} = 1/k^q$ with $q < 1$, so that $\bar{A}'_{2k} = o(\gamma_{2k})$. From the definition of $\beta_{2(k+1)}$, see (35), $\sum_k \alpha_{2k} < \infty$ and $\alpha_{2k} = o(\gamma_{2k})$. Since $\sum_k \bar{A}'_{2k} < \infty$, taking $q > 1/2$ in the definition of γ_{2k} ensures $\sum_k \Delta_{2k} < \infty$. We can repeat the same argument, and $d(\nu_{2(k+n)}, \nu^*_{p_{2(k+n)}}) < \gamma_{2(k+n)}$ which tends to zero as n increases, with $|p_{2(k+n)} - p_{2k}|$ remaining finite. $\nu_{2(k+n)}$ thus remains in some neighborhood $\mathcal{V}(p)$ of ν^*_p for any n, and $\mathcal{V}(p)$ can be made arbitrarily small by choosing α_0 and γ_0 small enough. ■

A5. *Proof of Theorem 6.* Assume that x_0 is such that for some $k \geq 0$, $\|g_{k+1}\| = 0$ with $\|g_i\| > 0$ for all $i \leq k$ (that is, $x_{k+1} = x^*$ and $x_i \neq x^*$ for $i \leq k$). This implies $R_k(W) = 0$ for any W, and therefore $R(W, x_0, x^*) = R(x_0, x^*) = 0$.

Assume now that $\|g_k\| > 0$ for any k. Consider

$$V_n = \left[\prod_{k=0}^{n-1} R_k(W)\right]^{1/n} = \left[\prod_{k=0}^{n-1} \frac{(Wg_{k+1}, g_{k+1})}{(Wg_k, g_k)}\right]^{1/n} = \left[\frac{(Wg_n, g_n)}{(Wg_0, g_0)}\right]^{1/n} .$$

We have,

$$\forall z \in \mathcal{H}\,,\ c\|z\|^2 \leq (Wz, z) \leq C\|z\|^2\,,$$

and thus

$$(c/C)^{1/n}\left[\frac{(g_n, g_n)}{(g_0, g_0)}\right]^{1/n} \leq V_n \leq (C/c)^{1/n}\left[\frac{(g_n, g_n)}{(g_0, g_0)}\right]^{1/n} .$$

Since $(c/C)^{1/n} \to 1$ and $(C/c)^{1/n} \to 1$ as $n \to \infty$, $\liminf_{n\to\infty} V_n$ and $\limsup_{n\to\infty} V_n$ do not depend on W. Take $W = P(A)$; it gives $R_k(W) = r_k = 1 - 1/L_k$, see (9), which is not decreasing, and thus $\lim_{n\to\infty} V_n = 1 - 1/L$ for any W. ■

Acknowledgement: The work of Luc Pronzato and Henry P. Wynn has been supported by the Network of Excellence "PASCAL" (Pattern Analysis, Statistical Modelling and Computational Learning), nb. 2002-506778.

References

[1] H. Akaike. On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Ann. Inst. Statist. Math. Tokyo*, 11:1–16, 1959.

[2] J. Barzilai and J.M. Borwein. Two-point step size gradient methods. *IMA J. of Numerical Analysis*, 8:141–148, 1988.

[3] E.G. Birgin and J.M. Martínez. A spectral conjugate gradient method for unconstrained optimization. *Applied Mathematics and Optimization*, 43:117–128, 2001.

[4] V.V. Fedorov. *Theory of Optimal Experiments.* Academic Press, New York, 1972.

[5] G.E. Forsythe. On the asymptotic directions of the s-dimensional optimum gradient method. *Numerische Mathematik*, 11:57–76, 1968.

[6] J. Hale and H. Koçak. *Dynamics and Bifurcations.* Springer-Verlag, Heidelberg, 1991.

[7] L.V. Kantorovich and G.P. Akilov. *Functional Analysis.* Pergamon Press, London, 1982. Second edition.

[8] J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

[9] V.S. Kozjakin and M.A. Krasnosel'skii. Some remarks on the method of minimal residues. *Numer. Funct. Anal. and Optimiz.*, 4(3):211–239, 1982.

[10] M.A. Krasnosel'skii and S.G. Krein. An iteration process with minimal residues. *Mat. Sb. (in Russian)*, 31(4):315–334, 1952.

[11] J.P. LaSalle. *The Stability of Dynamical Systems.* SIAM, Philadelphia, 1976.

[12] D.G. Luenberger. *Introduction to Linear and Nonlinear Programming.* Addison-Wesley, Reading, Massachusetts, 1973.

[13] J. Nocedal, A. Sartenaer, and C. Zhu. On the accuracy of nonlinear optimization algorithms. Technical Report Nov. 1998, ECE Department, Northwestern Univ., Evanston, Il 60208, 1998.

[14] J. Nocedal, A. Sartenaer, and C. Zhu. On the behavior of the gradient norm in the steepest descent method. *Computational Optimization and Applications*, 22:5–35, 2002.

[15] L. Pronzato, H.P. Wynn, and A.A. Zhigljavsky. *Dynamical Search.* Chapman & Hall/CRC, Boca Raton, 2000.

[16] L. Pronzato, H.P. Wynn, and A.A. Zhigljavsky. Renormalised steepest descent in Hilbert space converges to a two-point attractor. *Acta Applicandae Mathematicae*, 67:1–18, 2001.

[17] M. Raydan. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM J. Optimization*, 7(1):26–33, 1997.

[18] A.N. Shiryaev. *Probability.* Springer, Berlin, 1996.

[19] S.D. Silvey. *Optimal Design.* Chapman & Hall, London, 1980.