Gradient algorithms for quadratic optimization with fast convergence rates

Luc Pronzato* and Anatoly Zhigljavsky †

Abstract We propose a family of gradient algorithms for minimizing a quadratic function $f(x) = (Ax, x)/2 - (x, y)$ in $\mathbb{R}^d$ or a Hilbert space, with simple rules for choosing the step-size at each iteration. We show that when the step-sizes are generated by a dynamical system with ergodic distribution having the arcsine density on a subinterval of the spectrum of A, the asymptotic rate of convergence of the algorithm can approach the (tight) bound on the rate of convergence of a conjugate gradient algorithm stopped before d iterations, with $d \leq \infty$ the space dimension.

Key words: Chebyshev polynomials, conjugate gradient, Krylov space, logistic map, quadratic operator, steepest descent.

1 Introduction

Consider the problem of minimizing a quadratic function $f(\cdot)$ defined either on $\mathbb{R}^d$ or a Hilbert space by

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

where $(\cdot, \cdot)$ denotes the inner product. We assume that A is either a symmetric positive-definite matrix or a self-adjoint operator, with

$$0 < m = \inf_{(x,x)=1} (Ax, x) < M = \sup_{(x,x)=1} (Ax, x) < \infty\,.$$

If A is a matrix, then m and M are the smallest and largest eigenvalues of A, respectively.

Consider a general gradient algorithm with iterations of the form

$$x_{k+1} = x_k - \gamma_k g_k\,,\ k = 0, 1, 2 \ldots \tag{2}$$

where $g_k = \nabla f(x_k)$ is the gradient of the objective function $f(\cdot)$ at point x_k. For the objective function (1), $\nabla f(x) = Ax - y$. The iteration (2) can be rewritten in terms of the gradients as

$$g_{k+1} = g_k - \gamma_k A g_k\,. \tag{3}$$

In a series of papers [10, 11, 12] and the monograph [9] many gradient algorithms have been shown to be equivalent to special algorithms for updating measures on the interval $[m, M]$. The central idea is that of renormalization applied to the gradient. For simplicity the presentation is made for the finite dimensional case where A is a matrix, which can be assumed, without loss of generality, to be diagonal

*Laboratoire I3S, CNRS - UNS, Les Algorithmes - Bât. Euclide B, 2000 route des Lucioles – B.P. 121, F-06903 Sophia Antipolis Cedex, France (pronzato@i3s.unice.fr)

†School of Mathematics, Cardiff University, Senghennydd Road, Cardiff, CF24 4YH, UK (ZhigljavskyAA@cf.ac.uk)

$A = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ with eigenvalues $m = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_d = M$. Extension to the Hilbert-space case will be considered in Section 5.

Write $z_k = g_k/\sqrt{(g_k, g_k)}$ for the normalized gradient at x_k and define

$$p_i^{(k)} = \{z_k\}_i^2 = \frac{\{g_k\}_i^2}{\sum_{j=1}^d \{g_k\}_j^2}\,, \quad i = 1, \dots, d\,,$$

as the i-th probability corresponding to vector z_k, where $\{v\}_i$ denotes the i-th component of vector v. Let ν_k denote the probability measure on the spectrum of A defined by the $p_i^{(k)}$'s, that is, $\nu_k(\lambda_i) = p_i^{(k)}$. The probability measure ν_{k+1} is defined by

$$p_i^{(k+1)} = \frac{\{g_{k+1}\}_i^2}{(g_{k+1}, g_{k+1})} \qquad \text{for } \ i = 1, \dots, d\,.$$

Note that (3) gives

$$(g_{k+1}, g_{k+1}) = (g_k, g_k) - 2\gamma_k (Ag_k, g_k) + \gamma_k^2 (A^2 g_k, g_k)\,, \tag{4}$$

so that

$$p_i^{(k+1)} = \frac{(1-\gamma_k\lambda_i)^2}{(g_k, g_k) - 2\gamma_k (Ag_k, g_k) + \gamma_k^2 (A^2 g_k, g_k)}\, \{g_k\}_i^2 = \frac{(1-\gamma_k\lambda_i)^2}{1 - 2\gamma_k \mu_1^{(k)} + \gamma_k^2 \mu_2^{(k)}}\, p_i^{(k)}\,, \tag{5}$$

where

$$\mu_\alpha^{(k)} = \mu_\alpha(\nu_k) = \frac{(A^\alpha g_k, g_k)}{(g_k, g_k)} \tag{6}$$

is the α-th moment of the measure ν_k. When two eigenvalues of A are equal, say $\lambda_j = \lambda_{j+1}$, the updating rules for $p_j^{(k)}$ and $p_{j+1}^{(k)}$ are identical so that the analysis of the behaviour of the algorithm remains the same when $p_j^{(k)}$ and $p_{j+1}^{(k)}$ are confounded. We may thus assume that all eigenvalues of A are distinct. Also, a zero weight remains equal to zero at all subsequent iterations, we thus assume that $\nu_0(\lambda_i) > 0$ for all i.

A common definition for the rate of convergence of the algorithm at iteration k is $r_k = (g_{k+1}, g_{k+1})/(g_k, g_k)$. The rate for n iterations is

$$\prod_{k=0}^{n-1} r_k = \frac{(g_n, g_n)}{(g_0, g_0)}\,;$$

therefore, the asymptotic rate of the algorithm can naturally be defined as

$$R = \lim_{n\to\infty} R_n\,, \quad \text{with } \ R_n = \left(\prod_{k=0}^{n-1} r_k\right)^{1/n}\,. \tag{7}$$

Of course, this rate may depend on the initial point x_0 or, equivalently, on g_0. Other rates which are asymptotically equivalent to $\{r_k\}$ can be considered as well, see [12] and Remark 6.

The most familiar gradient algorithm is the steepest-descent algorithm, for which the step-size γ_k at iteration k is chosen so as to minimize $f(x_k - \gamma g_k)$ with respect to γ, which gives $\gamma_k = (g_k, g_k)/(Ag_k, g_k) = 1/\mu_1^{(k)}$. Its asymptotic behaviour is well-known, see [1, 10]. In particular, its convergence is slow: the asymptotic rate R depends on the starting point but is never far from its worst value given by the Kantorovich bound

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

where $\rho = M/m$, the condition number of A. The asymptotic behaviour of the family of algorithms defined by $\gamma_k = \mu_\alpha^{(k)}/\mu_{\alpha+1}^{(k)}$ (which includes the method of minimum residues for $\alpha = 1$) is shown in [12] to be similar.

Obtaining a faster asymptotic rate of convergence for gradient algorithms requires to extend the possible choices for the step-size γ_k. Rewrite the updating rule (5) in terms of iteration applied to the probability measure ν_k,

$$\nu_{k+1}(\lambda) = \frac{(1-\gamma_k\lambda)^2}{1-2\gamma_k\mu_1^{(k)}+\gamma_k^2\mu_2^{(k)}}\,\nu_k(\lambda) = \frac{(\lambda-\beta_k)^2}{\beta_k^2-2\beta_k\mu_1^{(k)}+\mu_2^{(k)}}\,\nu_k(\lambda)\,, \tag{8}$$

where $\beta_k = 1/\gamma_k$ and $\nu_k(\lambda)$ is the weight assigned by the measure ν_k to the point λ. The roots β_k in (8) are the key control variables for a gradient algorithm. Different strategies for choosing β_k give different families of algorithms. Note that the only information about ν_k one has access to corresponds to its moments $\mu_\alpha^{(k)}$, $\alpha = 1, 2\ldots$ Many of the examples of algorithms presented in [6], with β_k a function of $\mu_1^{(k)}$ and $\mu_2^{(k)}$, exhibit a much faster asymptotic rate of convergence than $R_{\max}$ (it seems that allowing β_k to depend on more moments $\mu_\alpha^{(k)}$ does not yield further improvement in the rate of convergence). Fast convergence (small R) is observed for algorithms that exhibit a chaotic-type behaviour in $\mathbb{R}^d$, which makes their theoretical study difficult. The same is true for some algorithms for which β_k is allowed to depend on moments of several previous measures ν_{k-i}, $i = 1, \ldots, u$. For instance, in the Barzilai-Borwein algorithm [2], β_k is either $\mu_1^{(k-1)}$ or $\mu_2^{(k-1)}/\mu_1^{(k-1)}$.

Conjugate gradient, s-step optimal, MINRES and other algorithms based on Krylov spaces do not use gradient directions for their successive iterations, see, e.g., [8]. However, when analyzing their behaviour, one can construct an equivalent sequence of iterations following the gradient directions with control variables β_k depending on k and on moments of previous measures ν_{k-i}, $i = 0, 1, 2\ldots$ The conjugate gradient algorithm in $\mathbb{R}^d$ converges in d iterations. When d is large, preserving the conjugacy of successive directions is difficult and restarting the algorithm after each sequence of s iterations is recommended. This corresponds to the s-step optimal gradient algorithm, see [5, 13], which does not have finite convergence but whose guaranteed asymptotic rate of convergence is

$$R_s^* = \left(\frac{R_\infty^{s/2}+R_\infty^{-s/2}}{2}\right)^{-2/s} = T_s^{-2/s}\left(\frac{\rho+1}{\rho-1}\right) \tag{9}$$

where

$$R_\infty = \lim_{s\to\infty} R_s^* = \left(\frac{\sqrt{\rho}-1}{\sqrt{\rho}+1}\right)^2$$

and $T_s(\cdot)$ is the s-th Chebyshev polynomial:

$$T_s(t) = \cos[s\arccos(t)] = \frac{(t+\sqrt{t^2-1})^s+(t-\sqrt{t^2-1})^s}{2}\,.$$

In this paper we propose a family of gradient algorithms based on simple rules for choosing the sequence of control variables β_k. The main idea is to force $\nu_k(\lambda_j)$, $j = 2, \ldots, d-1$, to tend to zero as $k \to \infty$. The measure ν_k, which summarizes the state of the iterates at step k, is then almost fully characterized by $\nu_k(m)$, which facilitates the analysis of the asymptotic behaviour. Furthermore, we show that the sequence $\{\beta_k\}$ can be chosen independently of $\{\nu_k\}$ while ensuring that the asymptotic rate of convergence is arbitrarily close to R_∞. This independence of $\{\beta_k\}$ on $\{\nu_k\}$ makes the algorithms at the same time simple and robust with respect to the precision of calculations. Also, the step-sizes $\gamma_k = 1/\beta_k$, $k = 1, 2\ldots$ are simpler to calculate than those of the steepest-descent algorithm. Convergence rates close to R_∞ are obtained when the β_k's are constructed so that their asymptotic distribution is close to a distribution with the arcsine density.

The worst-case rate R_s^* can be reached for the s-step optimal gradient when $d > s$, in the sense that there exist eigenvalues λ_i and initial point x_0 for the algorithm such that the rate of convergence after s iterations is exactly R_s^* (and the behavior in terms of renormalized gradient z_k is then periodic with period s), see [5, 13]. The same is true for the conjugate gradient algorithm: for $s < d$ there exist eigenvalues λ_i and a starting point x_0 such that the convergence rate after s iterations is exactly R_s^*.

If d is large (relative to the total number of iterations), s is not very large and the eigenvalues of A are well-spread in the spectral interval $[m, M]$, then the actual rates (per one matrix-vector multiplication) of the MINRES and other optimal methods based on the use of s-dimensional Krylov spaces are very close to R_s^* and are often larger than R_∞. Bearing in mind that the asymptotic rates of the algorithms suggested below can be arbitrarily close to R_∞ and these algorithms are extremely simple and robust, these algorithms may be preferable to MINRES and other Krylov space based methods for large-scale quadratic optimization problems.

The paper is organized as follows. In Section 2 we show that for a suitable choice of the sequence $\{\beta_k\}$ the algorithm attracts to the plane spanned by the eigenvectors associated with $\lambda_1 = m$ and $\lambda_d = M$. In Section 3, we assume that the values of m and M are known and give the expression of the asymptotic rate of convergence of the algorithm in the case where the β_k's are generated by pairs symmetric with respect to $(m + M)/2$. Several examples are presented, some with a rate arbitrarily close to R_∞. The case where m and M are unknown is considered in Section 4 where a practical algorithm is suggested and some simulation results are presented. Finally, the infinite dimensional situation where $f(\cdot)$ is defined on a Hilbert space is considered Section 5.

2 Attraction of the sequence $\{\nu_k\}$ to the set of measures supported at m and M

Theorem 1 *Assume that $\beta_k > 0$, $\beta_k \notin \{m, M\}$ for all k and that the sequence $\{\beta_k\}$ has asymptotic distribution function $F(\beta)$ which is supported on an interval $[m', M']$ with $0 < m' \leq M' < \infty$. Suppose, moreover, that the limiting distribution satisfies*

$$\int \log(\beta-\lambda)^2 \, dF(\beta) < \max\left\{ \int \log(M-\beta)^2 \, dF(\beta), \int \log(\beta-m)^2 \, dF(\beta) \right\}, \quad \forall \lambda \in \{\lambda_2, \ldots, \lambda_{d-1}\}. \tag{10}$$

Then, the gradient algorithm associated with the sequence $\{\beta_k\}$ is such that $\lim_{k\to\infty} \nu_k(\lambda_i) = 0$ for all $i = 2, \ldots, d-1$. Furthermore, there exist constants $C > 0, k_0 > 0$ and $0 \leq \theta < 1$ such that

$$\sum_{i=2}^{d-1} \nu_k(\lambda_i) \leq C\theta^k \text{ for } k > k_0 \,. \tag{11}$$

Proof. The fact that the sequence $\{\beta_k\}$ has asymptotic distribution function $F(\beta)$ implies

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=0}^{k-1} h(\beta_j) = \int h(\beta) \, dF(\beta) \tag{12}$$

for any continuous function $h(\cdot)$ such that $\int |h(\beta)| \, dF(\beta) < \infty$, see [7]. Define

$$H_k(\lambda) = C_k \, (\lambda - \beta_0)^2 \, (\lambda - \beta_1)^2 \cdots (\lambda - \beta_{k-1})^2 \,, \tag{13}$$

with C_k a normalizing constant such that $\nu_k(\lambda) = H_k(\lambda)\nu_0(\lambda)$ in (8), and assume that

$$\int \log(M-\beta)^2 \, dF(\beta) \leq \int \log(\beta - m)^2 \, dF(\beta) \tag{14}$$

(if this inequality is not met, m should be replaced with M in all considerations below). Define the sum

$$S_k(\lambda, m) = \frac{1}{k} \log \frac{H_k(\lambda)}{H_k(m)} = -\frac{1}{k} \sum_{j=0}^{k-1} \log(\beta_j - m)^2 + \frac{1}{k} \sum_{j=0}^{k-1} \log(\lambda - \beta_j)^2 \tag{15}$$

and consider the first sum $I_k(m) = (1/k)\sum_{j=0}^{k-1} \log(\beta_j - m)^2$ in the right-hand side of (15) and the related integral $I(m) = \int \log(\beta - m)^2 \, dF(\beta)$. Since the c.d.f. $F(\cdot)$ is supported on a bounded interval $[m', M']$ we have $I(m) < \infty$. The assumptions (10) and (14) imply $I(m) > -\infty$ and the property (12) then gives the convergence $I_k(m) \to I(m)$ as $k \to \infty$.

Consider now the second sum $I_k(\lambda) = (1/k)\sum_{j=0}^{k-1} \log(\beta_j - \lambda)^2$ in the right-hand side of (15) and the related integral $I(\lambda) = \int \log(\beta - \lambda)^2 \, dF(\beta)$. Since the c.d.f. $F(\cdot)$ is supported on a bounded interval, the integral $I(\lambda)$ is properly defined but may equal $-\infty$ (for example, if the c.d.f. $F(\cdot)$ has a discontinuity at the point λ). If $I(\lambda) = -\infty$ then as $k \to \infty$ the sum $I_k(\lambda)$ tends to $-\infty$ too. If $I(\lambda) > -\infty$ then either $I_k(\lambda) = -\infty$ for all k large enough (when at least one β_j is equal to λ) or (12) implies that $I_k(\lambda)$ tends to $I(\lambda)$ as $k \to \infty$.

Therefore, from (10), $S_k(\lambda, m)$ tends to a negative value (possibly $-\infty$) as $k \to \infty$. This implies that there exists $k_0 \geq 0$ and $\delta > 0$ such that for all $k \geq k_0$ and $\lambda \in \{\lambda_2, \ldots, \lambda_{d-1}\}$

$$S_k(\lambda, m) = \frac{1}{k} \log \frac{H_k(\lambda)}{H_k(m)} \leq -\delta \, ; \tag{16}$$

that is, $H_k(\lambda)/H_k(m) \leq \theta^k$, where $\theta = \exp(-\delta) < 1$. This yields $\sum_{i=2}^{d-1} \nu_k(\lambda_i) \leq \theta^k \left(\sum_{i=2}^{d-1} \nu_0(\lambda_i)\right)/\nu_0(m)$ for $k > k_0$, hence (11). The result $\lim_{k\to\infty} \nu_k(\lambda_i) = 0$ for $i = 2, \ldots, d-1$ obviously follows from (11). ■

Remark 1 The sequence $\{\beta_k\}$ can be assumed random, for instance formed by independent and identically distributed random variables. In this case, all the statements are true with probability one. When the β_k's are simply independent, with $\{F_k\}$ the sequence of corresponding distribution functions and $(1/k)\sum_{j=0}^{k-1} F_j$ converging weakly to F as k tends to infinity, one may refer to [3, Th. 2.5.3, p. 36] for a property similar to (12).

Remark 2 Typically, the spectrum of A is unknown. In that case, the condition (10) can be replaced with the more restrictive one

$$\int \log(\beta - \lambda)^2 \, dF(\beta) < \max\left\{ \int \log(M - \beta)^2 \, dF(\beta), \int \log(\beta - m)^2 \, dF(\beta) \right\} \, , \quad \forall \lambda \in (m, M) \, . \tag{17}$$

Remark 3 If the distribution with c.d.f. $F(\cdot)$ is symmetric with respect to $(m+M)/2$, then we have $\int \log(M - \beta)^2 \, dF(\beta) = \int \log(\beta - m)^2 \, dF(\beta)$ and therefore the condition (17) simplifies to

$$\int \log(\beta - \lambda)^2 \, dF(\beta) < \int \log(\beta - m)^2 \, dF(\beta) \, , \quad \forall \lambda \in (m, M) \, . \tag{18}$$

Remark 4 Note that the support $[m', M']$ of the distribution with c.d.f. $F(\cdot)$ could be different from $[m, M]$ and does not have to be a subset of $[m, M]$.

Remark 5 The results of Theorem 1 also apply when β_k depends on the moments of previous measures ν_{k-i}, $i = 0, 1, 2 \dots$

Example 1 For the steepest-descent algorithm with $\beta_k = \mu_1^{(k)}$, the limiting measure for $\{\beta_k\}$ is the two-point measure assigning weights 1/2 at z and $m + M - z$ for some $z \in (m, M)$. The condition (17) then simply expresses the property that two successive iterations (8) of the algorithm asymptotically give a larger increase of the weights at the endpoints m and M than at any other point in the interval (m, M); that is,

$$(z - m)^2 \, (M - z)^2 > (z - \lambda)^2 \, (m + M - z - \lambda)^2 \, , \forall \lambda \in (m, M) \, . \tag{19}$$

Since for all z the only maximum of $(z - \lambda)^2 \, (m + M - z - \lambda)^2$ with respect to $\lambda \in (m, M)$ is at $\lambda^* = (m+M)/2$, the inequality (19) can be rewritten as $(z-m)^2 \, (M-z)^2 > (z-\lambda^*)^2 \, (m+M-z-\lambda^*)^2$, which gives

$$z \in \left(\frac{1}{2}(m + M) - \frac{1}{2\sqrt{2}}(M - m) \, , \frac{1}{2}(m + M) + \frac{1}{2\sqrt{2}}(M - m) \right) \, . \tag{20}$$

This corresponds to the definition of the stability interval for the attractor in [10, 12]. A similar result holds for all gradient-type algorithms from the family considered in [12].

Example 2 If we choose $\beta_k = \sqrt{\mu_2^{(k)}}$, then the limiting measure for $\{\beta_k\}$ is the delta-measure concentrated at the point $\lambda^* = (m + M)/2$; as a consequence, the asymptotic rate for the related gradient algorithm is $R_{\max}$. Proof of these facts can be found in [4] and [6], Sect. 2.7.

3 Asymptotic rate for symmetrically placed control variables

3.1 Main result

Theorem 2 *Assume that the conditions of Theorem 1 are satisfied and that, moreover, the control variables β_k are generated by symmetric pairs for large k; that is, $\beta_{2j+1} = M + m - \beta_{2j}$ for all $j \geq j_0$, with $\beta_{2j} \in [m + \varepsilon, M - \varepsilon]$ for some $\varepsilon \in (0, (M - m)/2)$. Then, the asymptotic rate R satisfies*

$$\log R = \int \log \left| \frac{(M - \beta) \, (\beta - m)}{\beta \, (m + M - \beta)} \right| \, dF(\beta) = \int \log \frac{(\beta - m)^2}{\beta^2} \, dF(\beta) \, . \tag{21}$$

Proof. First note that dividing (4) through by (g_k, g_k) gives the following expression for the rate r_k,

$$r_k = 1 - 2\gamma_k \frac{(Ag_k, g_k)}{(g_k, g_k)} + \gamma_k^2 \frac{(A^2 g_k, g_k)}{(g_k, g_k)} = 1 - 2\mu_1^{(k)}/\beta_k + \mu_2^{(k)}/\beta_k^2 \, . \tag{22}$$

Consider a measure ν with weights p and $1-p$ at m and M respectively, $0 < p < 1$. Apply two successive iterations (8) with control parameters β and $\beta' = m + M - \beta$ to this measure. The product of the two successive rates does not depend on p and is equal to $R_2^2(\beta) = (M - \beta)^2 (\beta - m)^2 / [\beta(m + M - \beta)]^2$.

According to Theorem 1, ν_k tends to be supported at m and M and the rate of convergence is exponential. We thus obtain for two successive iterations with control variables β_{2j} and $\beta_{2j+1} = m + M - \beta_{2j}$

$$R_2^2(\beta_{2j}) \left[1 - \frac{A \, \theta^{2j}}{R_2^2(\beta_{2j})} \right] < r_{2j} r_{2j+1} < R_2^2(\beta_{2j}) \left[1 + \frac{A \, \theta^{2j}}{R_2^2(\beta_{2j})} \right]$$

for some $A > 0$ and $j > k_0/2$, see Theorem 1. Since $\beta_{2j} \in [m+\varepsilon, M-\varepsilon]$, we have $R_2^2(\beta_{2j}) \geq R_2^2(m+\varepsilon) = \varepsilon(M - m - \varepsilon)/[(m + \varepsilon)(M - \varepsilon)] > 0$. Therefore,

$$\log R_2(\beta_{2j}) - B\theta^{2j} < \log \sqrt{r_{2j} r_{2j+1}} < \log R_2(\beta_{2j}) + B\theta^{2j} \, ,$$

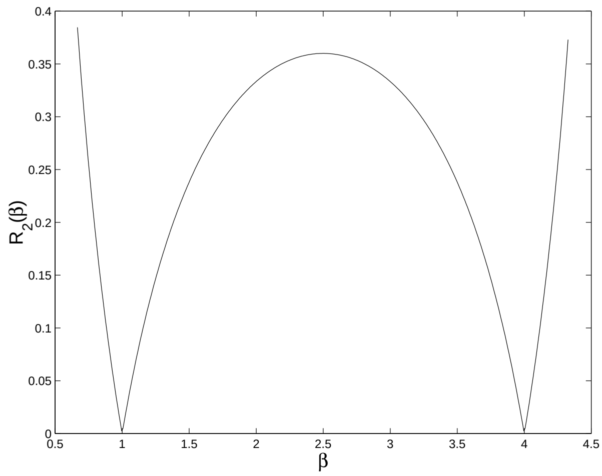

Figure 1: $R_2(\beta)$ for $m = 1$, $M = 4$

with $B = A/R_2^2(m+\varepsilon)$, for j large enough. Since $\sum_{j=0}^{\infty} \theta^{2j} = 1/(1-\theta^2) < \infty$, we obtain from (12),

$$\log R = \lim_{k\to\infty} \frac{1}{k} \sum_{j=0}^{k-1} \log \sqrt{r_{2j} r_{2j+1}} = \int \log R_2(\beta)\, dF(\beta)\,,$$

hence the first expression in (21). The second expression follows from the fact that the c.d.f. $F(\cdot)$ is symmetric with respect to $(m+M)/2$. ■

Example 3 *Uniform density.* Let the distribution with c.d.f. $F(\cdot)$ be uniform with density $p(\beta) = 1/(M'-m')$, $\beta \in [m', M']$, with $m' = m+\varepsilon$, $M' = M-\varepsilon$ and $0 < \varepsilon < (M-m)/2$. Then the asymptotic rate of convergence is

$$R_{\text{uniform},\,\varepsilon} = \exp\left\{ \frac{1}{M'-m'} \int_{m'}^{M'} \log \frac{(\beta-m')^2}{\beta^2}\, d\beta \right\} = (M'-m')^2 \exp\left\{ -2\, \frac{M' \log M' - m' \log m'}{M'-m'} \right\}. \tag{23}$$

Remark 6 One can easily check that the result stated in Theorem 2 holds for other definitions for the rate of convergence, see, e.g., [12, Th. 6]. For instance, the rate

$$r'_k = \frac{f(x_{k+1}) - f^*}{f(x_k) - f^*} = \frac{(A^{-1} g_{k+1}, g_{k+1})}{(A^{-1} g_k, g_k)}\,,$$

where $f^* = \min_x f(x)$, can be written as

$$r'_k = 1 - 2/(\mu_{-1}^{(k)}\, \beta_k) + \mu_1^{(k)}/(\mu_{-1}^{(k)}\, \beta_k^2) \tag{24}$$

and the corresponding asymptotic rate $R' = \lim_{n\to\infty} \left(\prod_{k=0}^{n-1} r'_k \right)^{1/n}$ is equal to R which can be computed by (21).

Remark 7 The shape of $R_2(\beta)$ as a function of β shows that fast convergence is obtained for β close to m or M, see Figure 1, hence the interest of taking ε small in Theorem 2.

Remark 8 When ν_k is a two-point measure supported at m and M, two iterations of (8) with $\beta_{k+1} = M + m - \beta_k$ give $\nu_{k+2} = \nu_k$. Under the conditions of Theorem 2 the measure ν_k thus converges to a measure $\bar{\nu}_k = p_k\delta_m + (1-p_k)\delta_M$ supported at m and M, with p_{2j} tending to a constant p_∞ as j tends to infinity. The limiting distribution of the sequence $\{p_{2j+1}\}$ depends on $F(\cdot)$ and p_∞, while the value of p_∞ depends on the initial measure ν_0 and the spectrum of A.

3.2 Finite collection of control variables

Assume that the points $\beta_0, \beta_1 \ldots$ are generated in repeated groups $B = \{\beta_0, \ldots, \beta_N\}$ of $N+1$ points in (m, M), $N \geq 0$. Additionally, the points in B are symmetric with respect to $(m+M)/2$. We may always assume that $\beta_0 \leq \ldots \leq \beta_N$. In this case, if N is even then $\beta_{N/2} = (m+M)/2$. The condition (18) now becomes

$$\sum_{j=0}^{N} \log(\beta_j - \lambda)^2 < \sum_{j=0}^{N} \log(\beta_j - m)^2\,, \quad \forall \lambda \in (m, M)\,. \tag{25}$$

If this condition is met then the asymptotic rate is

$$R = R_N = \left[\prod_{j=0}^{N} \frac{(\beta_j - m)^2}{\beta_j^2}\right]^{1/(N+1)}\,. \tag{26}$$

Example 4 *Uniform grid.* Assume that for some integer $N \geq 0$,

$$B = \{\beta_0, \ldots, \beta_N\} \text{ with } \beta_i = m + \frac{i+\frac{1}{2}}{N+1}(M-m), \;\; i = 0, 1, \ldots, N\,. \tag{27}$$

It is easy to see that the condition (25) is met. The rate R_N computed by (26) is given by

$$R_N = \left(\frac{\Gamma^2\left(N+3/2\right)\Gamma^2\left(\frac{m+M+2Nm}{2(M-m)}\right)}{\pi\,\Gamma^2\left(\frac{2\,NM+3\,M-m}{2(M-m)}\right)}\right)^{1/(N+1)}\,,$$

where $\Gamma(\cdot)$ is the gamma-function. The value of R_N for $m=1$, $M=4$ is plotted in Figure 2 as a function of N. Asymptotically, as $N \to \infty$, R_N approaches $R_{\text{uniform},0}$ defined in (23) (with $R_{\text{uniform},0} \simeq 0.2232$ for $m=1$, $M=4$). Instead of using the β_i's according to (27) for large N, one can generate the sequence $\{\beta_i\}$ using, for example, the Bernoulli shift:

$$H_B(t) = 2t \;[\text{mod } 1]\,, \; t \in (0,1)\,, \tag{28}$$

with β_0 randomly chosen in (m', M'), and for all $j = 0, 1, 2 \ldots$

$$\beta_{2j+1} = M' + m' - \beta_{2j}\,, \;\; \beta_{2j+2} = m' + (M'-m')\,H_B\left(\frac{\beta_{2j}-m'}{M'-m'}\right)\,,$$

with $m' = m + \varepsilon$, $M' = M - \varepsilon$ and $0 < \varepsilon < (M-m)/2$.

Example 5 *Nearly optimal $N+1$ points.* Consider first the case $N=1$. When the condition (25) is satisfied, the asymptotic rate is $R_2 = |M-\beta|\,|\beta-m|/[\beta\,|m+M-\beta|]$, see the proof of Theorem 2. For $\beta \in [m, M]$, R_2 improves when $|\beta - (m+M)/2|$ increases and reaches its minimum value, zero, at $\beta \in \{m, M\}$, see Remark 7. Condition (25) imposes that β belongs to the interval (20), by choosing β sufficiently close to $(m+M)/2 \pm (M-m)/(2\sqrt{2})$ one makes the rate arbitrarily close to R_2^*, with R_s^* defined by (9).

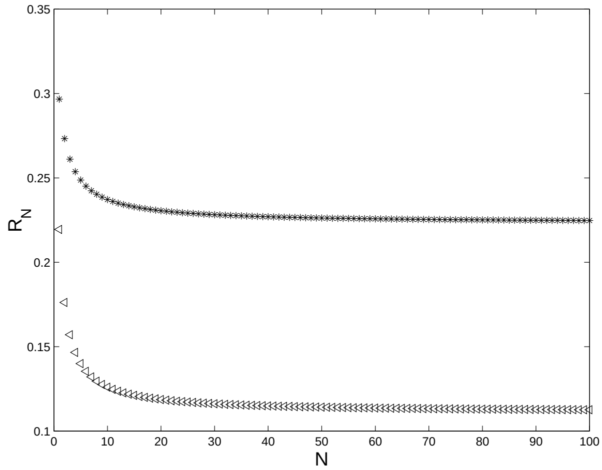

Figure 2: Asymptotic rate of convergence (26) for $m = 1$ and $M = 4$ when the β_j's are on the uniform grid (stars) and when they correspond to Chebyshev points (triangles, $\varepsilon = 10^{-6}$)

Take now $N = 2$, with $\beta_0 = \beta$, $\beta_1 = (m + M)/2$ and $\beta_2 = m + M - \beta$. Similarly to the previous case, condition (25) imposes that β belongs to the interval $((m + M)/2 - \sqrt{3}(M - m)/4, (m + M)/2 + \sqrt{3}(M - m)/4)$, with the rate R_3 getting close to R_3^* for β close to $(m + M)/2 \pm \sqrt{3}(M - m)/4$.

By induction, one can show that the rate R_N can be made arbitrarily close to the value R_s^* defined by (9), with $s = N + 1$, when the $N + 1$ points β_i are suitably chosen and are constructed from the roots of Chebyshev polynomials. This construction is considered in the next example. (Note that the fact that R_N can be made arbitrarily close to R_{N+1}^* is not a coincidence: the worst case analysis of the s-step optimal gradient algorithm, which yields the rate R_s^*, corresponds to the situation where the β_i's are rescaled roots of the s-th order Chebyshev polynomial, see [5].)

Example 6 *Chebyshev points.* Chebyshev points are defined by

$$t_k = \cos\left(\frac{\pi}{2}\frac{2k+1}{N+1}\right), \quad k = 0, \ldots, N.$$

and correspond to the roots of $T_{N+1}(x) = \cos((N + 1) \arccos x)$, the Chebyshev polynomials of the first kind. These points are symmetric on $(-1, 1)$. The asymptotic density of the points $\{t_k\}_0^N$, as $N \to \infty$, is $p(t) = 1/(\pi\sqrt{1 - t^2})$, $t \in (-1, 1)$.

Define

$$\beta_k = \frac{m+M}{2} + \frac{M-m-2\varepsilon}{2} t_k, \; k = 0, \ldots, N,$$

where $0 < \varepsilon < (M - m)/2$. These points belong to the interval $(m + \varepsilon, M - \varepsilon)$ and are symmetric with respect to $(m + M)/2$. As $\varepsilon > 0$, the condition (25) holds. The rate R_N computed by (26) is plotted in Figure 2 as a function of N for $m = 1$, $M = 4$ and $\varepsilon = 10^{-6}$. Asymptotically, as $N \to \infty$, R_N approaches $R_{\mathrm{arcsin},\varepsilon}$ defined below in (31).

3.3 Control variables with arcsine density on a subinterval of $[m, M]$

Let us assume that the distribution with c.d.f. $F(\cdot)$ has the density

$$p_\varepsilon(\beta) = \frac{1}{\pi\sqrt{(\beta - m')(M' - \beta)}}, \quad m' \le \beta \le M', \tag{29}$$

where $m' = m + \varepsilon$, $M' = M - \varepsilon$ and $0 < \varepsilon < (M - m)/2$. The density (29) is called the arcsine density on the interval $[m', M']$.

The sequence of points $\{\beta_i\}$ can be generated using, for example, the logistic map

$$H_L(x) = 4x(1-x)\,,\ x \in (0,1)\,, \tag{30}$$

with β_0 randomly chosen in (m', M'), and for all $j = 0, 1, 2 \ldots$

$$\beta_{2j+1} = M' + m' - \beta_{2j}\,,\ \beta_{2j+2} = m' + (M' - m')H_L\left(\frac{\beta_{2j} - m'}{M' - m'}\right)\,.$$

Note that the control variables β_j are placed symmetrically in the interval $[m, M]$. We show below that the condition (17) holds for each $\varepsilon > 0$. According to (21), the asymptotic rate of convergence is then

$$R_{\mathrm{arcsin},\,\varepsilon} = \exp\left\{\int_{m'}^{M'} \log\frac{(\beta - m)^2}{\beta^2}\, p_\varepsilon(\beta)\, d\beta\right\} \tag{31}$$

and we show below that

$$R_{\mathrm{arcsin},\,\varepsilon} = \left(\frac{M - m + 2\sqrt{\varepsilon(M - m - \varepsilon)}}{M + m + 2\sqrt{(M - \varepsilon)(m + \varepsilon)}}\right)^2\,. \tag{32}$$

For $\varepsilon = 0$ this gives $R_{\mathrm{arcsin},\,0} = R_\infty = (\sqrt{\rho} - 1)^2/(\sqrt{\rho} + 1)^2$ where $\rho = M/m$. However, we cannot choose $\varepsilon = 0$ as the condition (17) does not hold (we also show below that $I(\lambda) = \int \log(\beta - \lambda)^2\, dF(\beta) = 2\log(M - m) - 4\log 2$ for $\lambda \in [m, M]$). Since the condition does hold for any $\varepsilon > 0$, the rate of the algorithm can be made arbitrarily close to R_∞: for small $\varepsilon > 0$, we have

$$R_{\mathrm{arcsin},\,\varepsilon} = R_\infty\left(1 + 4\sqrt{\varepsilon(M - m)}\right) + O(\varepsilon),\ \ \varepsilon \to 0\,.$$

The rest of this section is devoted to the verification of (17) for $\varepsilon > 0$ and to the derivation of the formula (32) for the rate $R_{\mathrm{arcsin},\,\varepsilon}$. Define the integral

$$J(z, m', M') = \int_{m'}^{M'} \frac{\log(\beta - z)^2}{\pi\sqrt{(\beta - m')(M' - \beta)}}\, d\beta\,, \tag{33}$$

where $-\infty < z < \infty$. The changes of variables $t = -1 + 2(\beta - m')/(M' - m')$ and $x = -1 + 2(z - m')/(M' - m')$ in the integral (33) give

$$J(z, m', M') = 2\log\frac{M' - m'}{2} + \frac{1}{\pi} I_x, \ \text{ where } \ I_x = \int_{-1}^{1} \frac{\log(t - x)^2}{\sqrt{1 - t^2}}\, dt\,. \tag{34}$$

Assume first that $|x| \le 1$. By changing the variable $t = \cos\phi$ in the integral I_x, we obtain

$$I_x = \int_0^\pi \frac{\log(\cos\phi - x)^2}{\sin\phi} \sin\phi\, d\phi = \int_0^\pi \log(\cos\phi - x)^2\, d\phi\,.$$

As $\cos(\phi) = \cos(2\pi - \phi)\ \forall\phi$, we have $\int_0^\pi \log(\cos\phi - x)^2\, d\phi = \int_\pi^{2\pi} \log(\cos\phi - x)^2\, d\phi$, which implies $I_x = \frac{1}{2}\int_0^{2\pi} \log(\cos\phi - x)^2\, d\phi$. As we assume $-1 \le x \le 1$ we can set $\psi = \arccos x$ (so that $x = \cos\psi$). Using now the identity $\cos\phi - \cos\psi = 2\sin\frac{\psi - \phi}{2}\sin\frac{\phi + \psi}{2}$, we obtain

$$\begin{aligned}
I_x &= \frac{1}{2}\int_0^{2\pi} \log(\cos\phi - \cos\psi)^2\, d\phi = \frac{1}{2}\int_0^{2\pi} \log\left(2\sin\frac{\phi - \psi}{2}\sin\frac{\phi + \psi}{2}\right)^2 d\phi \\
&= \frac{1}{2}\left[\int_0^{2\pi} 2\,\log 2\, d\phi + \int_0^{2\pi} \log\left(\sin\frac{\phi - \psi}{2}\right)^2 d\phi + \int_0^{2\pi} \log\left(\sin\frac{\phi + \psi}{2}\right)^2 d\phi\right] \\
&= 2\pi\log 2 + \left[\int_0^\pi \log\left(\sin^2(\phi - \psi/2)\right) d\phi + \int_0^\pi \log\left(\sin^2(\phi + \psi/2)\right) d\phi\right]\,.
\end{aligned}$$

The function $t \to \sin^2 t$ is π-periodic and therefore for any ψ' we get

$$\int_0^\pi \log\left(\sin^2(\phi+\psi')\right)\, d\phi = \int_0^\pi \log\left(\sin^2(\phi)\right)\, d\phi = 2\int_0^\pi \log\left(\sin\phi\right)\, d\phi\,.$$

This implies

$$I_x = 2\pi\log 2 + 4\int_0^\pi \log(\sin\phi)\, d\phi = 2\pi\log 2 - 4\pi\log 2 = -2\pi\log 2\,, \quad \forall x\in[-1,1]\,. \tag{35}$$

Assume now that $|x|\geq 1$. From (35) we have $I_1 = -2\pi\log 2$ and differentiating I_x we get

$$I'_x = \left(\int_{-1}^1 \frac{\log(x-t)^2}{\sqrt{1-t^2}}dt\right)' = \frac{2\pi}{\sqrt{x^2-1}}\,.$$

Therefore, for $x>1$,

$$I_{-x} = I_x = I_1 + \int_1^x I'_z dz = -2\pi\log 2 + \int_1^x \frac{2\pi}{\sqrt{z^2-1}}dz = -2\pi\log 2 + 2\pi\log\left(\frac{x+\sqrt{x^2-1}}{2}\right)\,. \tag{36}$$

Combining (35) and (36) we obtain

$$I_x = \int_{-1}^1 \frac{\log(t-x)^2}{\sqrt{1-t^2}}\, dt = \begin{cases} -2\pi\log 2 & \text{if } |x|\leq 1 \\ 2\pi\log\left(|x|+\sqrt{x^2-1}\right) - 2\pi\log 2 & \text{if } |x|\geq 1\,, \end{cases}$$

together with (34), it gives

$$J(z,m',M') = \begin{cases} 2\log(M'-m') - 4\log 2 & \text{if } m'\leq z\leq M' \\ 2\log(M'-m') + 2\log\left(|t_z|+\sqrt{t_z^2-1}\right) - 4\log 2 & \text{if } z<m' \text{ or } z>M'\,, \end{cases} \tag{37}$$

where $t_z = -1 + 2(z-m')/(M'-m')$. Therefore, $J(\lambda,m',M') < J(m,m',M') = J(M,m',M')$ for all λ in (m,M) and (17) is satisfied. The expression (32) for the rate $R_{\mathrm{arcsin},\,\varepsilon}$ easily follows from (37) and the representation $R_{\mathrm{arcsin},\,\varepsilon} = \exp\left[J(m,m',M') - J(0,m',M')\right]$ with $m' = m+\varepsilon$ and $M' = M-\varepsilon$.

4 Estimation of m, M and a practical algorithm

4.1 Estimation of m, M and asymptotic behavior in the non symmetric case

The values of m and M can be easily estimated in the first iterations of the algorithm (3), for instance by computing the first moment $\mu_1^{(j)}$ for several values of $j=0,1,2\ldots$ and taking

$$m_k = \min\{\mu_1^{(j)}\,,\ j=0,\ldots,k\}\,, \quad M_k = \max\{\mu_1^{(j)}\,,\ j=0,\ldots,k\} \tag{38}$$

as estimates. We then necessarily have $m<m_k<M_k<M$ for $k\geq 1$.

Suppose that the estimation is stopped at some k_0, that is, $m_k = m_{k_0}$ and $M_k = M_{k_0}$ for all $k>k_0$. Then, under the conditions of Theorem 1 with $m' = m_{k_0}$ and $M' = M_{k_0}$ we have $\sum_{i=2}^{d-1}\nu_k(\lambda_i)\leq C\theta^k$ for k larger than some k_1 and constants $C>0$ and $0\leq\theta<1$. Suppose that the control variables are generated by pairs for $k>k_0$, as in Theorem 2, but with $\beta_{2k+1} = M_{k_0} + m_{k_0} - \beta_{2k}$, for all $k>k_0$.

If $M_{k_0} + m_{k_0} = M+m$, Theorem 2 applies and the asymptotic rate R satisfies (21). For instance, if the β_k's are generated as in Section 3.3, and have the arcsine density on $[m_{k_0}, M_{k_0}]$, the asymptotic rate is $R_{\mathrm{arcsin},\,\varepsilon}$ with $\varepsilon = m_{k_0} - m = M - M_{k_0}$. Consider now the standard situation where $M_{k_0} + m_{k_0} \neq M+m$

and suppose that $M - M_{k_0} > m_{k_0} - m$. The asymptotic distribution of the β_k's, symmetric in $[m_{k_0}, M_{k_0}]$, is then biased towards m and $\nu_k(m)$ tends to zero when $k \to \infty$. Following the same line as in the proof of Theorem 2, we obtain that the product of rates at two successive iterations for the delta measure at M, with control parameters respectively β and $\beta' = M_{k_0} + m_{k_0} - \beta$, is $R_2^2 = (M-\beta)^2(M-\beta')^2/(\beta\beta')^2$. The asymptotic rate then satisfies

$$\log R = \int \log \left| \frac{(M-\beta)\,(M+\beta-M_{k_0}-m_{k_0})}{\beta\,(M_{k_0}+m_{k_0}-\beta)} \right| \, dF(\beta)\,.$$

Similarly, supposing that $M - M_{k_0} < m_{k_0} - m$ gives an asymptotic distribution of the β_k's biased towards M, so that $\nu_k(m)$ tends to 1 as $k \to \infty$, and the asymptotic rate satisfies

$$\log R = \int \log \left| \frac{(\beta-m)\,(M_{k_0}+m_{k_0}-\beta-m)}{\beta\,(M_{k_0}+m_{k_0}-\beta)} \right| \, dF(\beta)\,.$$

Now, note that $\nu_k(m) \to 0$ implies that $\mu_1^{(k)} \to M$ and $\nu_k(m) \to 1$ implies that $\mu_1^{(k)} \to m$, $k \to \infty$, so that maintaining the adaptation of the estimation of m_k and M_k by (38) ensures that $m_k \to m$ and $M_k \to M$ as $k \to \infty$. This permits to recover the same asymptotic rates as Section 3.3, even in situations where m and M are unknown. Since the estimated values m_k and M_k quickly converge to m and M, see for instance Figure 3, we need to generate the control variable β_k in $[m_k + \varepsilon, M_k - \varepsilon]$ at iteration k. A practical algorithm is given below.

4.2 An algorithm based on the arcsine density

A possible algorithm is then as follows.

- Choose τ as a small positive number (e.g., $\tau = 10^{-6}$), set $z_0 = 0$;
- for $k = 0, 1$, set $\beta_k = \mu_1^{(k)}$ (steepest-descent) and set $m_1 = \min\{\mu_1^{(0)}, \mu_1^{(1)}\}$, $M_1 = \max\{\mu_1^{(0)}, \mu_1^{(1)}\}$;
- for $k > 1$, set $\varepsilon_k = \tau(M_{k-1} - m_{k-1})$ and generate the β_k's by pairs:
  - for $k = 2j$, set $z_j = \{\varphi + z_{j-1}\}$ and $\beta_{2j} = m_k + \varepsilon_k + (\cos(\pi z_j) + 1)(M_k - m_k - 2\varepsilon_k)/2$, where $\{t\}$ denotes the fractional part of t and $\varphi = (\sqrt{5} - 1)/2 \simeq 0.61803$;
  - for $k = 2j + 1$, set $\beta_{2j+1} = M_k + m_k - \beta_{2j}$;

  set $m_k = \min\{m_{k-1}, \mu_1^{(k)}\}$, $M_k = \max\{M_{k-1}, \mu_1^{(k)}\}$.

The sequence $z_1, z_2 \ldots$ is such that $z_j = \{j\varphi\}$ so that the sequence is asymptotically uniform on $[0, 1]$, see, e.g., [7]. This implies that the asymptotic distribution of the sequence β_k has the arcsine density on $[m + \varepsilon, M - \varepsilon]$ where $\varepsilon = \tau(M - m)$. From (32), the rate of the algorithm satisfies

$$\lim_{n\to\infty} R_n = R_{\mathrm{arcsin},\tau(M-m)} = R_\infty(1 + 4\sqrt{\tau}) + \mathcal{O}(\tau)\,,\ \tau \to 0\,.$$

The dynamical system $z_j = \{j\varphi\}$ generates a sequence in $[0, 1]$ with much better uniformity characteristics than sequences generated by the Bernoulli shift (28). Since the logistic map (30) corresponds to a transformation of the Bernoulli shift, the construction above, based on $z_j = \{j\varphi\}$, produces a sequence of control variables β_k with better distribution characteristics than sequences generated with (30).

Figures 3, 4 and 5 illustrate the typical behavior of the algorithm above in a large-dimensional badly conditioned problem. In the example presented, $d = 1000$, $m = 1$, $M = \rho = 1000$ and the eigenvalues λ_i

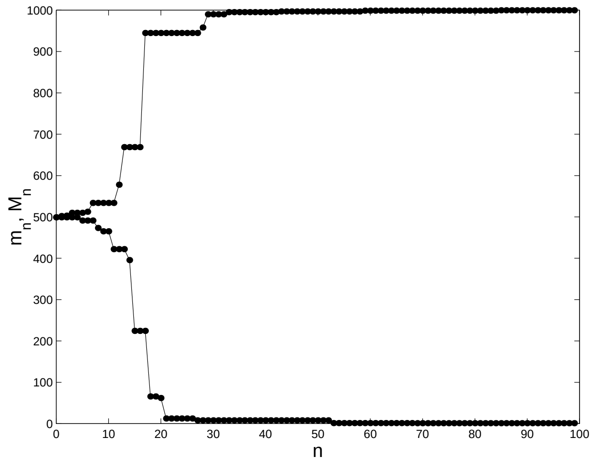

Figure 3: Convergence of the estimates m_n and M_n as functions of n ($m = 1, M = 1000$, $d = 1000$)

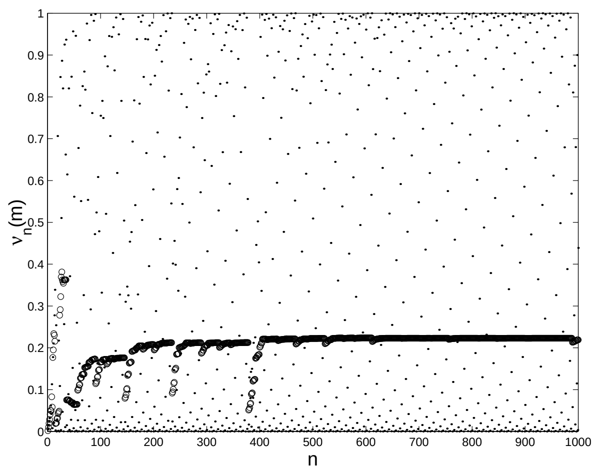

Figure 4: Values $\nu_n(m)$ as functions of n; circles for $n = 2j$, dots for $n = 2j + 1$ ($\rho = 1000$, $d = 1000$)

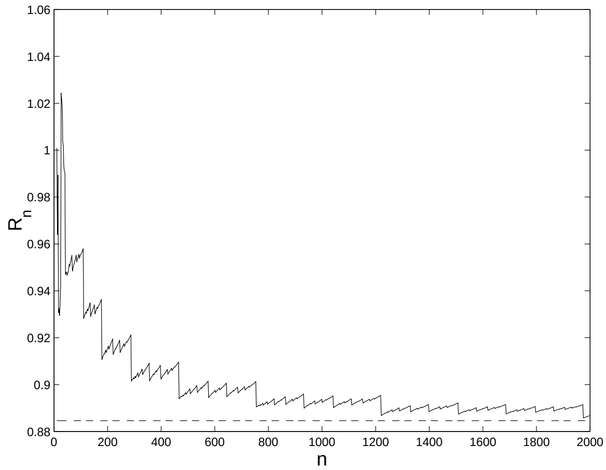

Figure 5: Rate of convergence R_n, see (7), as a function of n; the limiting value $R_{\text{arcsin},\tau(M-m)}$ is indicated by the dashed line ($m = 1, M = 1000$, $d = 1000$, $\tau = 10^{-6}$)

are random and uniformly distributed on the interval $[m, M]$ (one would obtain exactly the same plots if the eigenvalues were equally-spaced on $[m, M]$).

In terms of complexity of calculations, only the multiplications of d-dimensional vectors by the $d \times d$ matrix A are expensive. The steepest descent algorithm requires the calculation of $\beta_k = \mu_2^{(k)}/\mu_1^{k)} = (Ag_k, Ag_k)/(Ag_k, g_k)$ at iteration k. Having computed g_k and Ag_k, one may notice that next gradient g_{k+1} can be obtained as $g_{k+1} = g_k - (1/\beta_k)Ag_k$, so that only the computation of Ag_{k+1} is expensive at iteration $k + 1$. However, a long sequence of iterations of this type may produce an accumulation of rounding errors, and it is rather recommended to recalculate g_{k+1} from x_{k+1} by $g_{k+1} = Ax_{k+1} - y$, see (1). This then requires two multiplications by A at each steepest-descent iteration.

In the algorithm above, iteration k only requires the calculation of the gradient $g_k = Ax_k - y$, and thus only one multiplication by A. Notice that the estimation of m_k and M_k through the moments $\mu_1^{(j)} = (Ag_j, g_j)/(g_j, g_j)$, see (38), does not require the calculation of Ag_j at step k. Indeed, allowing a delay of one step in the estimation, we have $(g_j, g_{j+1}) = (g_j, g_j - (1/\beta_j)Ag_j)$ so that $\mu_1^{(j)}$ is obtained at next step from

$$\mu_1^{(j)} = \beta_j \left[1 - \frac{(g_j, g_{j+1})}{(g_j, g_j)}\right] .$$

Also, one may observe in Figure 3 that the convergence of m_n and M_n to m and M respectively is very fast, so that the estimation can be stopped after a few iterations. On the whole, it makes iterations with the algorithm above about twice simpler than steepest-descent iterations (even when m and M are estimated), with much faster convergence.

5 Hilbert space case

In the Hilbert-space case, A is a self-adjoint operator and its spectrum $\mathcal{S}_A$ is a closed subset of the interval $[m, M]$ of the real line, with $m, M \in \mathcal{S}_A$. Let E_λ be the spectral family associated with A and define the measure $\nu_k = d(E_\lambda z_k, z_k)$, $m \leq \lambda \leq M$, with $z_k = g_k/\sqrt{(g_k, g_k)}$ the normalized gradient at x_k. We have $(z_k, z_k) = 1 = \int_m^M \nu_k(d\lambda)$ and ν_k is a probability measure on the Borel sets of $(0, \infty)$, satisfying

$\nu_k([m, M]) = 1$ for all k and with moments still defined by (6). One iteration of a gradient algorithm with control variable β_k thus gives in terms of ν_k

$$\nu_{k+1}(\mathcal{A}) = \frac{\int_{\mathcal{A}} (\lambda - \beta_k)^2 \, \nu_k(d\lambda)}{\beta_k^2 - 2\beta_k \mu_1^{(k)} + \mu_2^{(k)}} \,,$$

for $\mathcal{A}$ any measurable subset of $[m, M]$, see (8). The properties obtained for the finite dimensional case remain valid and only a few adaptations are required.

Theorem 3 *Assume that the sequence $\{\beta_k\}$ has asymptotic distribution function $F(\beta)$ which is supported on an interval $[m', M'] = [m + \varepsilon, M - \varepsilon]$ with $0 < \varepsilon < (M - m)/2$. Suppose, moreover, that $I(\lambda) = \int \log(\beta - \lambda)^2 \, dF(\beta)$ is a continuous function of λ on (m', M') and that*

$$I(\lambda) < \max \left\{ \int \log(M - \beta)^2 \, dF(\beta), \int \log(\beta - m)^2 \, dF(\beta) \right\} \,, \quad \forall \lambda \in (m', M') \,, \tag{39}$$

and that $\nu_0\{[m, m + \gamma)\} > 0$ and $\nu_0\{(M - \gamma, M]\} > 0$ for all $\gamma > 0$. Then, the measure ν_k converges to a two-point measure supported at m and M, in the sense that there exists k_0 such that, for any function $g(\lambda)$ continuous on $[m, M]$ and any $\delta > 0$, there exists $\gamma > 0$ such that

$$\max \left\{ \left| \int_m^C g(\lambda)\nu_k(d\lambda) - g(m) \int_m^C \nu_k(d\lambda) \right| , \left| \int_C^M g(\lambda)\nu_k(d\lambda) - g(M) \int_C^M \nu_k(d\lambda) \right| \right\} < \delta + C_\gamma \alpha_\gamma^k \,, \ k > k_0 \,,$$

where $C = (m + M)/2$ and $C_\gamma > 0$, $\alpha_\gamma \in (0, 1)$ are constants depending on γ. If, moreover, the control variables β_k are generated by symmetric pairs for large k, that is, $\beta_{2j+1} = M + m - \beta_{2j}$ for all $j \geq j_0$, then the asymptotic rate R satisfies (21).

Proof. The proof of convergence of ν_k to a two-point measure follows the same arguments as for Theorem 1. Suppose that $F(\cdot)$ satisfies (14). We still have for the first term of the sum $S_k(\lambda, m)$ defined by (15)

$$I_k(m) = \frac{1}{k} \sum_{j=0}^{k-1} \log(\beta_j - m)^2 \to I(m) = \int \log(\beta - m)^2 \, dF(\beta) \,, \ k \to \infty \,.$$

Concerning the second term $I_k(\lambda) = (1/k) \sum_{j=0}^{k-1} \log(\beta_j - \lambda)^2$ we need now a bound uniform in λ, that is, we need to show that

$$\forall \epsilon > 0 \,, \ \exists K_0 \text{ such that:} \quad \sup_{\lambda \in (m', M')} I_k(\lambda) - I(\lambda) < \epsilon \,, \ \forall k > K_0 \,. \tag{40}$$

Take a ball $\mathcal{B}(\lambda_1, \delta) = \{\lambda : |\lambda - \lambda_1| \leq \delta\}$ and consider $\bar{a}_\delta(\beta) = \sup_{\lambda \in \mathcal{B}(\lambda_1, \delta)} \log(\beta - \lambda)^2$, which is an increasing function of δ, $\bar{a}_\delta(\beta) = 2 \log(|\beta - \lambda_1| + \delta)$. We have

$$\lim_{\delta \to 0} \int \bar{a}_\delta(\beta) \, dF(\beta) = \int [\lim_{\delta \to 0} \bar{a}_\delta(\beta)] \, dF(\beta) = I(\lambda_1)$$

and therefore, there exists $\delta_1 = \delta_1(\lambda_1)$ such that $\int \bar{a}_\delta(\beta) \, dF(\beta) < I(\lambda_1) + \epsilon/3$ for $\delta < \delta_1$. Now,

$$\sup_{\lambda \in \mathcal{B}(\lambda_1, \delta)} I_k(\lambda) \leq (1/k) \sum_{j=0}^{k-1} 2 \log(|\beta_j - \lambda| + \delta) < \int \bar{a}_\delta(\beta) \, dF(\beta) + \epsilon/3$$

for all k larger than some $K_1 = K_1(\lambda_1, \delta)$. Also, from the continuity of $I(\lambda)$, there exists $\delta_2 = \delta_2(\lambda_1)$ such that $\inf_{\lambda \in \mathcal{B}(\lambda_1, \delta)} I(\lambda) > I(\lambda_1) - \epsilon/3$ for $\delta < \delta_2$. Altogether it gives $\sup_{\lambda \in \mathcal{B}(\lambda_1, \delta)} I_k(\lambda) - I(\lambda) < \epsilon$ for

$\delta < \delta_0(\lambda_1) = \min(\delta_1, \delta_2)$ and $k > K_1$. It only remains to cover $[m', M']$ with a finite number of such balls $\mathcal{B}(\lambda_i, \delta)$, with $\delta < \min_i \delta_0(\lambda_i)$ to obtain the result (40). Since $\log(\beta - \lambda)^2$ is a decreasing (resp. increasing) function of λ in $[m, m']$ (resp. in $[M', M]$), together with the condition (39) it implies that for any set $\mathcal{S} \subset (m, M)$, $\limsup_{k\to\infty} \sup_{\lambda\in\mathcal{S}} S_k(\lambda, m) \le -\delta$ for some $\delta = \delta(\mathcal{S}) > 0$. Therefore, there exists k_0 such that, $\forall k > k_0$, $\sup_{\lambda\in(m',M')} H_k(\lambda)/H_k(m) \le \theta_\varepsilon^k$ where $\theta_\varepsilon = \exp(-\delta_\varepsilon) < 1$.

Consider now a function $g(\lambda)$ continuous on $[m, M]$ and define

$$\Delta_k = \left| \int_m^C g(\lambda)\nu_k(d\lambda) - g(m) \int_m^C \nu_k(d\lambda) \right| ,$$

where $C = (m + M)/2$. We show below that

$$\forall \delta > 0\,,\, \exists \gamma > 0 \text{ such that } \Delta_k < \delta + 2\, \frac{D_g}{\int_m^{m+\gamma} \nu_0(d\lambda)}\, \alpha_\gamma^k \;\text{ for all }\; k > k_0\,, \tag{41}$$

for some $\alpha_\gamma < 1$, where $D_g = \max_{\lambda\in[m,C]} |g(\lambda) - g(m)|$. We have $\Delta_k < \int_m^C |g(\lambda) - g(m)|\nu_k(d\lambda) = \Delta_{k,1} + \Delta_{k,2} + \Delta_{k,3}$, with

$$\Delta_{k,1} = \int_m^{m+2\gamma} |g(\lambda) - g(m)|\nu_k(d\lambda),\; \Delta_{k,2} = \int_{m+2\gamma}^{m'} |g(\lambda) - g(m)|\nu_k(d\lambda),\; \Delta_{k,3} = \int_{m'}^{C} |g(\lambda) - g(m)|\nu_k(d\lambda),$$

$\gamma < \varepsilon/2$. From the continuity of $g(\lambda)$, we can take γ small enough to have $\Delta_{k,1} < \delta \int_m^{m+2\gamma} \nu_k(d\lambda) \le \delta$. Next, $\Delta_{k,2} < D_g \int_{m+2\gamma}^{m'} \nu_k(d\lambda) = D_g \int_{m+2\gamma}^{m'} H_k(\lambda)\nu_0(d\lambda)$ with $H_k(\lambda)$ defined by (13). Since $\beta_k \in [m', M']$ for all k, $H_k(\lambda)$ is a decreasing function of λ for $\lambda \in [m, m']$, and for $m + 2\gamma < \lambda < m'$ it satisfies

$$H_k(\lambda) < H_k(m + 2\gamma) < H_k(m + \gamma) \left(\frac{M - m - \varepsilon - 2\gamma}{M - m - \varepsilon - \gamma} \right)^{2k} .$$

Since $\int_m^{m+\gamma} \nu_k(d\lambda) = \int_m^{m+\gamma} H_k(\lambda)\nu_0(d\lambda) \ge H_k(m + \gamma) \int_m^{m+\gamma} \nu_0(d\lambda)$, we obtain

$$\Delta_{k,2} < \frac{D_g}{\int_m^{m+\gamma} \nu_0(d\lambda)} \left[\frac{M - m - \varepsilon - 2\gamma}{M - m - \varepsilon - \gamma} \right]^{2k} .$$

We also obtain for the last term,

$$\Delta_{k,3} < D_g \int_{m'}^{C} H_k(\lambda)\nu_0(d\lambda) < D_g \theta_\varepsilon^k H_k(m) \int_{m'}^{C} \nu_0(d\lambda) < D_g \theta_\varepsilon^k H_k(m) \;\text{ for } k > k_0\,.$$

For $\lambda \in [m, m']$ we have $H_k(\lambda)/H_k(m) \ge (m' - \lambda)^{2k}/\varepsilon^{2k}$ so that

$$1 \ge \int_m^{m+\gamma} \nu_k(d\lambda) \ge H_k(m) \int_m^{m+\gamma} [(m' - \lambda)/\varepsilon]^{2k}\nu_0(d\lambda) > H_k(m)[(\varepsilon - \gamma)/\varepsilon]^{2k} \int_m^{m+\gamma} \nu_0(d\lambda)\,.$$

Therefore, for $k > k_0$,

$$\Delta_{k,3} < \frac{D_g}{\int_m^{m+\gamma} \nu_0(d\lambda)} \left[\frac{\theta_\varepsilon \varepsilon^2}{(\varepsilon - \gamma)^2} \right]^k .$$

We have $\theta_\varepsilon \varepsilon^2/(\varepsilon - \gamma)^2 < 1$ for $\gamma < \varepsilon(1 - \sqrt{\theta_\varepsilon})$ so that (41) is satisfied for $\alpha_\gamma = \max\{\theta_\varepsilon\varepsilon^2/(\varepsilon - \gamma)^2, (M - m - \varepsilon - 2\gamma)^2/(M - m - \varepsilon - \gamma)^2\}$ and $\alpha_\gamma < 1$ for γ small enough. One can show a similar property for $\Delta'_k = \left| \int_C^M g(\lambda)\nu_k(d\lambda) - g(M) \int_C^M \nu_k(d\lambda) \right|$.

Finally, we apply the property above to $g(\lambda) = \lambda$ and $g(\lambda) = \lambda^2$ and, following the same line as in the proof of Theorem 2, we then obtain for the product of rates at two successive iterations with control variables β_{2j} and $\beta_{2j+1} = m + M - \beta_{2j}$:

$$R_2^2(\beta_{2j}) \left[1 - \frac{A_\gamma\, \alpha_\gamma^{2j} + B\delta}{R_2^2(\beta_{2j})}\right] < r_{2j} r_{2j+1} < R_2^2(\beta_{2j}) \left[1 + \frac{A_\gamma\, \alpha_\gamma^{2j} + B\delta}{R_2^2(\beta_{2j})}\right] ,$$

for some $A_\gamma > 0$, $B > 0$ and $j > k_0/2$. Therefore,

$$\log R_2(\beta_{2j}) - A'_\gamma \alpha_\gamma^{2j} - B'\delta < \log \sqrt{r_{2j} r_{2j+1}} < \log R_2(\beta_{2j}) + A'_\gamma \alpha_\gamma^{2j} + B'\delta \,,$$

with $A'_\gamma = A_\gamma / R_2^2(m+\varepsilon)$ and $B' = B/R_2^2(m+\varepsilon)$, for j large enough. Since $\sum_{j=0}^\infty \alpha_\gamma^{2j} = 1/(1-\alpha_\gamma^2) < \infty$, we obtain from (12),

$$\left|\log R - \int \log R_2(\beta)\, dF(\beta)\right| = \left|\lim_{k\to\infty} \frac{1}{k} \sum_{j=0}^{k-1} \log \sqrt{r_{2j} r_{2j+1}} - \int \log R_2(\beta)\, dF(\beta)\right| < B'\delta \,.$$

Since δ is arbitrary, the asymptotic rate of convergence is thus the same as in the finite dimensional case.

■

Remark 9 Note that the condition $I(\lambda)$ being a continuous function of λ is satisfied for all examples considered in Section 3. It is also satisfied when the distribution function $F(\cdot)$ has density $\phi(\cdot)$ with derivative $\phi'(\cdot)$ uniformly bounded on (m', M'). Indeed, one can write $I(\lambda) = \int_{\lambda-M'}^{\lambda-m'} \phi(\lambda - t)\, \log t^2\, dt$ which has derivative $I'(\lambda) = \phi(m')\, \log(\lambda - m')^2 - \phi(M')\, \log(\lambda - M')^2 + \int_{m'}^{M'} \phi'(t)\, \log(\lambda - t)^2\, dt$; this derivative is bounded, which implies the continuity of $I(\lambda)$.

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 Journal of Numerical Analysis*, 8:141–148, 1988.

[3] H.J. Bierens. *Topics in Advanced Econometrics.* Cambridge University Press, Cambridge, 1994.

[4] Y. H. Dai and X. Q. Yang. A new gradient method with an optimal stepsize property. *Comput. Optim. Appl.*, 33(1):73–88, 2006.

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

[6] R. Haycroft, L. Pronzato, H.P. Wynn, and A.A. Zhigljavsky. Studying convergence of gradient algorithms via optimal experimental design theory. In L. Pronzato and A.A. Zhigljavsky, editors, *Optimal Design and Related Areas in Optimization and Statistics*, pages 13–37. Springer, 2009.

[7] L. Kuipers and H. Niederreiter. *Uniform Distribution of Sequences.* Wiley, New York, 1974.

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

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

[10] 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.

[11] L. Pronzato, H.P. Wynn, and A.A. Zhigljavsky. An introduction to dynamical search. In P.M. Pardalos and H.E. Romeijn, editors, *Handbook of Global Optimization*, volume 2, chapter 4, pages 115–150. Kluwer, Dordrecht, 2002.

[12] L. Pronzato, H.P. Wynn, and A.A. Zhigljavsky. Asymptotic behaviour of a family of gradient algorithms in $\mathbb{R}^d$ and Hilbert spaces. *Mathematical Programming*, A107:409–438, 2006.

[13] L. Pronzato, H.P. Wynn, and A.A. Zhigljavsky. A dynamical-system analysis of the optimum s-gradient algorithm. In L. Pronzato and A.A. Zhigljavsky, editors, *Optimal Design and Related Areas in Optimization and Statistics*, pages 39–80. Springer, 2009.