

Noisy Optimization: Fast Convergence Rates with Comparison-Based Algorithms

Marie-Liesse Cauwet, Olivier Teytaud

▶ **To cite this version:**

Marie-Liesse Cauwet, Olivier Teytaud. Noisy Optimization: Fast Convergence Rates with Comparison-Based Algorithms. Genetic and Evolutionary Computation Conference, Jul 2016, Denver, United States. pp.1101-1106. hal-01306636v2

HAL Id: hal-01306636
https://hal.science/hal-01306636v2

Submitted on 27 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Noisy Optimization: Fast Convergence Rates with Comparison-Based Algorithms

Marie-Liesse Cauwet Olivier Teytaud

TAO, Inria, Lri, Umr Cnrs 8623

Abstract

Derivative Free Optimization is known to be an efficient and robust method to tackle the black-box optimization problem. When it comes to noisy functions, classical comparison-based algorithms are slower than gradient-based algorithms. For quadratic functions, Evolutionary Algorithms without large mutations have a simple regret at best $O(1/\sqrt{N})$ when N is the number of function evaluations, whereas stochastic gradient descent can reach (tightly) a simple regret in $O(1/N)$. It has been conjectured that gradient approximation by finite differences (hence, not a comparison-based method) is necessary for reaching such a $O(1/N)$. We answer this conjecture in the negative, providing a comparison-based algorithm as good as gradient methods, i.e. reaching $O(1/N)$ - under the condition, however, that the noise is Gaussian. Experimental results confirm the $O(1/N)$ simple regret, i.e., squared rate compared to many published results at $O(1/\sqrt{N})$.

Keywords: Noisy continuous optimization; Comparison-based Algorithms

1 The black-box noisy optimization problem

In a real world optimization problem, the analytical form of the objective function is frequently unavailable. It is common in this field to obtain only the fitness values of the objective function: this is the black-box problem. In this setting, given a search point, an oracle returns the corresponding fitness value. Furthermore, due to stochastic effects or inaccurate measurements, the fitness values can be improper: this is called *noise*, and the optimization problem is then a noisy optimization problem. We here consider noisy optimization with constant additive Gaussian noise. Given an objective function F and a search point $x \in$

$\mathbb{R}^d$, the oracle outputs the fitness value $F_{\text{noisy}}(x)$:

$$F_{\text{noisy}}(x) = \mathcal{G}(F(x), b), \tag{1}$$

where $\mathcal{G}(a, b)$ is a Gaussian random variable with mean a and standard deviation $b > 0$.

Regarding some industrial applications, a call to the oracle might be expensive, requiring heavy computations. Thus, we aim to find an approximation of the optimum within a number of evaluations as small as possible. The algorithm spends N evaluations and then outputs an answer, which is an approximation - denoted $\hat{x}_N$ - of the minimum[1] x^* of F. With these notations, the simple regret after N evaluations is defined by:

$$SR_N = \mathbb{E}\left(F_{\text{noisy}}(\hat{x}_N) - F_{\text{noisy}}(x^*)\right) = \mathbb{E}F(\hat{x}_N) - F(x^*). \tag{2}$$

On the right-hand side of Eq. 2, the expectation operates on $\hat{x}_N$ which might be a random variable due to the stochasticity of the noisy evaluations or the possible internal randomization of the optimization algorithm.

Dupač [5] has shown that noisy quadratic strongly convex functions can be optimized with simple regret $O(1/N)$, when the budget (i.e. the number of evaluations) is N. Fabian [6] has broadened this result to a wider class of functions, but with only an approximation of this rate: for a function with arbitrarily many derivatives, a regret $O(1/N^\alpha)$ can be reached for $\alpha < 1$ arbitrarily close to 1. Furthermore, this bound $O(1/N)$ is optimal (see [3]). Shamir in [10] has improved the results, in terms of the non-asymptotic nature of some of these convergence, and in terms of explicit dependency in the dimension.

These rates are reached by algorithms introduced by Kiefer and Wolfowitz [8], which approximate the gradient using finite differences and thus using fitness values. However, as a refinement of the black-box problem, we might encounter some optimization problems where the fitness value itself is unknown. In this case, an oracle only provides a ranking of a given set of points, but not the fitness values of these points. For example in games, an operator can compare two agents, but not directly provide a level evaluation. In design, with the human in the loop, a user preference is a comparison between two search points. Searching a Pareto front might also involve a user providing his preferences. Comparison based algorithms such as Evolutions Strategies (ES), Differential Evolution (DE) or Particle Swarm Optimization (PSO) can handle this type of problem. The comparison oracle is also noisy in the sense that the points might be misranked.

Shamir in [10] has conjectured that the use of approximate gradients is necessary for fast rates (i.e. rates $O(1/N)$) in the noisy strongly convex quadratic case. In this case, the best known bounds for comparison-based algorithms are a simple regret $O(1/\sqrt{N})$ (see [1] for Evolution Strategies), which supports this conjecture. However, we show in the

[1] w.l.g. we assume that the optimum is a minimum.

present paper that, for noisy quadratic forms, a simple regret $O(1/N)$ can be reached by a comparison-based algorithm, combining the "mutate large inherit small" principle [2] and the use of large population sizes. The "mutate large inherit small" principle is used in the sense that we have long distances between current estimates of the optimum and search points, even when the estimate is close to the optimum.

Jamieson *et al.* in [7] have presented a bound for a comparison-based operator, using a number of comparisons quadratic $O\left(\frac{1}{\epsilon^2}\right)$ for ensuring precision ϵ in the simple regret - whereas we only need $O\left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)$ comparisons. More precisely, we fully rank $O\left(\frac{1}{\epsilon}\right)$ points; they can be sorted with $O\left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)$ comparisons.

Section 2 describes the key idea to get a fast comparison-based algorithm in a noisy setting. The theoretical aspects and a precise description of a fast optimization algorithm is given in Section 3 for the specific case of the sphere function. In this case, the technicality in the proof is lighter and allowed a good insight of what we will use when switching to a larger family of functions: the quadratic functions in Section 4. Last, we address the experimental aspects in Section 5.

2 Comparison Procedure

The main idea is to estimate the parameters of the objective function. The algorithm hence builds a model of the function and provides an approximation of the optimum. Specifically, comparing 2 search points N times provides an estimation at distance $O(1/\sqrt{N})$ of one parameter of the function. This estimation is made possible through the frequency at which the fitness values of one of the search points is better than the other. In particular, it is crucial to know the model of noise. Hence, the optimization algorithms of Sections 3 and 4 consist in a sequence of calls to Cop, given below.

Comparison Procedure (Cop).

procedure Cop(N, x, y, F_{noisy})
  $f \leftarrow 0$
  for $i = 1$ to N **do**
    $f_x^i \leftarrow F_{\text{noisy}}(x)$
    $f_y^i \leftarrow F_{\text{noisy}}(y)$
  end for
  $f \leftarrow \frac{1}{N^2} \sum_{1 \leq i,j \leq N} \mathbf{1}_{f_x^i < f_y^j}$
  return f
end procedure

Importantly, this operator can be computed faster than the apparent $O(N^2)$ complexity. Using sorting algorithm, the complexity is $O(N \log N)$.

3 Sphere function

3.1 In dimension 1

We first propose in Alg. 1 an algorithm (Cops1) achieving regret $O(1/N)$ on the noisy sphere problem in dimension 1.

Algorithm 1 Comparison Procedure for Sphere function in dimension 1 (Cops1).

Input:
an oracle $F_{\text{noisy}} : x \in \mathbb{R} \mapsto \mathcal{G}(|x - x^*|^2, 1)$
an even budget N

Output:
an approximation $\hat{x}$ of the optimum $x^* \in [-1, 1]$ of the objective function $F : \ x \mapsto |x - x^*|^2$

$K \leftarrow N/2$
$f \leftarrow \text{Cop}(K, 1, -1, F_{\text{noisy}})$
Define $\hat{x}$ such that $\mathbb{P}\left(\mathcal{G}(0,1) < \sqrt{8}\hat{x}\right) = f$
$\hat{x} \leftarrow \max(-1, \min(1, \hat{x}))$
return $\hat{x}$

Theorem 1 *Let $F_{noisy}(x) = |x - x^*|^2 + \mathcal{G}(0,1)$ be the noisy sphere function in dimension 1, where $x^* \in [-1, 1]$. Then the simple regret of* Cops1 *after N evaluations satisfies:*

$$SR_N = O(1/N). \tag{3}$$

Proof 1 *Consider* Cops1 *on such an objective function. By definition of F_{noisy} and F,*

$$\begin{aligned} p &= \mathbb{P}\left(F_{noisy}(1) < F_{noisy}(-1)\right) \\ &= \mathbb{P}\left(|1 - x^*|^2 + \mathcal{G}(0,1) < |-1 - x^*|^2 + \mathcal{G}(0,1)\right) \\ &= \mathbb{P}\left(\sqrt{2}\mathcal{G}(0,1) < (1 + x^*)^2 - (1 - x^*)^2\right) \\ &= \mathbb{P}\left(\mathcal{G}(0,1) < \sqrt{8}x^*\right). \end{aligned} \tag{4}$$

Step 1: Expectation and Variance of f.

With the notations of Cop*, let us define:*

$$\forall\ i, j \in \{1, \ldots, N\}^2,\ \mathbf{1}_{i,j} = \begin{cases} 1 \text{ if } f_1^i < f_{-1}^j \\ 0 \text{ otherwise} \end{cases}$$

$\mathbf{1}_{i,j}$ is Bernoulli distributed with probability of success p.

f is the output of the Cop *procedure. By definition,*

$$f = \frac{1}{K^2} \sum_{1 \leq i,j \leq K} \mathbf{1}_{i,j}.$$

The expectation and variance of f are then:

$$\begin{aligned} \mathbb{E}f &= p \\ Varf &= \frac{1}{K^4} \sum_{i=1}^{K} \sum_{j=1}^{K} Cov\left(\sum_{k=1}^{K} \mathbf{1}_{i,k}, \sum_{k'=1}^{K} \mathbf{1}_{j,k'} \right) \\ &= \frac{1}{K^4} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} \sum_{k'=1}^{K} Cov(\mathbf{1}_{i,k}, \mathbf{1}_{j,k'}) \end{aligned} \tag{5}$$

If $i \neq j$ and $k \neq k'$, $Cov(\mathbf{1}_{i,k}, \mathbf{1}_{j,k'}) = 0$ by independence. If $i = j$ (or $k = k'$), by Cauchy-Schwarz:

$$Cov(\mathbf{1}_{i,k}, \mathbf{1}_{i,k'}) \leq \sqrt{Var(\mathbf{1}_{i,k}) Var(\mathbf{1}_{i,k'})} \leq \frac{1}{4}$$

This together with Eq. 5 give:

$$\begin{aligned} Varf &= \frac{1}{K^4} \left(\sum_{i=1}^{K} \sum_{k=1}^{K} \sum_{k'=1}^{K} Cov(\mathbf{1}_{i,k}, \mathbf{1}_{i,k'}) + \right. \\ &\qquad \left. \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} Cov(\mathbf{1}_{i,k}, \mathbf{1}_{j,k}) \right), \\ &\leq \frac{1}{K^4} \times \frac{K^3}{2}, \\ &\leq \frac{1}{N}. \end{aligned}$$

Step 2: Lipschitz. *We denote by Φ the cumulative distribution function of the standard Gaussian: $\Phi(x) = \mathbb{P}\left(\mathcal{G}(0,1) < x\right)$ and m and M such that $\Phi^{-1}_{[m,M]} : [m, M] \to [-1, 1]$ is the inverse of Φ over these intervals. Let us define*

$$h(x) = \begin{cases} \Phi^{-1}_{[m,M]}(x) & \text{if } m \leq x \leq M \\ -1 & \text{if } x < m \\ 1 & \text{if } M < x \end{cases}$$

Let us evaluate the Lipschitz coefficient $L(h)$ of h. $\Phi^{-1}_{[m,M]}$ is differentiable over $[m, M]$ since Φ is differentiable over $[-1, 1]$ hence its Lipschitz $L(\Phi^{-1}_{[m,M]})$ is bounded. h is continuous, and h is constant over $(-\infty, m]$ and $[M, \infty)$; hence the Lipschitz of h is $L(\Phi^{-1}_{[m,M]})$ over $[m, M]$.

Step 3: Concluding. *We have, by definition of* Cops1 *for* $\hat{x}$ *and by Eq. 4 for* x^*,

$$\hat{x} = \frac{h(f)}{\sqrt{8}} \text{ and } x^* = \frac{h(p)}{\sqrt{8}}, \tag{6}$$

By definition of the simple regret in Eq. 2,

$$\begin{aligned} SR_N &= \mathbb{E}|\hat{x} - x^*|^2 \\ &\leq \mathbb{E}L(h)^2|f-p|^2/8 \text{ by } \textbf{Step 2} \\ &\leq \frac{L(h)^2}{8N} \text{ by } \textbf{Step 1}. \end{aligned}$$

Remark 1 *The result of Theorem 1 is based on the fact that the noise is a standard Gaussian. However, this result still holds as soon as the noise distribution has expectation 0, finite variance (possibly unknown, see Section 4) and a bounded Lipschitz. The distribution of the noise, on the other hand, must be known.*

3.2 Multidimensional sphere function

Alg. 2 (Cops) presents a straightforward extension to the noisy multidimensional sphere. $B_d(c, r)$ denotes the ball of center c and radius r in dimension d, and $\|.\|$ is the Euclidean norm.

Algorithm 2 Comparison procedure for the sphere function (Cops).

Input:
an oracle $F_{\text{noisy}} : x \in \mathbb{R}^d \mapsto \mathcal{G}(\|x - x^*\|^2, 1)$
a budget N (multiple of $2d$)

Output:
an approximation $\hat{x}$ of the optimum $x^* \in B_d(0, 1)$ of the objective function $F : x \mapsto \|x - x^*\|^2$

$K \leftarrow N/2d$
for $i = 1$ to d **do**
Apply Cops1 with a budget K on the unidimensional restriction of F_{noisy} to $\{0\}^{i-1} \times [-1, 1] \times \{0\}^{d-i}$
$\hat{x}_i$ be the obtained approximation of the optimum in $[-1, 1]$.
end for
return $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_d)$.

Theorem 2 *Let* $F_{noisy}(x) = \|x - x^*\|^2 + \mathcal{G}(0, 1)$ *be the noisy sphere function, with* $x^* \in B_d(0, 1) \subset \mathbb{R}^d$*. Then the simple regret of* Cops *after* N *evaluations is:*

$$SR_N = O(d/N).$$

Proof 2 *The conditions of Theorem 1 are verified for each application of* Cops1*. The simple regret for the multidimensional case is the sum of the simple regrets of each restrictions.*

4 General quadratic forms

Alg. 3 extends the principle of Section 3 to the optimization of a wider class of quadratic functions. $\|\cdot\|_2$ denotes the matrix norm induced by $\|\cdot\|$, i.e. $\|A\|_2 = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}$ and $\|\cdot\|_F$ is the Frobenius norm. (e_i) is the standard basis and A^t is the transpose of matrix A.

Algorithm 3 Comparison procedure for quadratic functions (CopQuad).

Input:
an oracle $F_{\text{noisy}} : x \in \mathbb{R}^d \mapsto \mathcal{G}(x^t Ax + Bx + C, D)$
a budget N (multiple of $d(d+3)-2$)

Output:
an approximation $\hat{x}$ of the optimum $x^* \in B_d(0,1)$ of the objective function $F : x \mapsto x^t Ax + Bx + C$

$K \leftarrow \frac{N}{d(d+3)-2}$
for $i = 1$ to d **do**
  $f_{-e_i,e_i} \leftarrow \text{Cop}(K, -e_i, e_i, F_{\text{noisy}})$
  Define $\hat{B}_i(D)$ such that $\mathbb{P}\left(\mathcal{G}(0,1) < \sqrt{2}\hat{B}_i(D)\right) = f_{-e_i,e_i}$
  $\hat{B}_i(D) \leftarrow \max(-5, \min(\hat{B}_i(D), 5))$ ▷ Estimate of B_i/D
  $f_{0,e_i} \leftarrow \text{Cop}(K, 0, e_i, F_{\text{noisy}})$
  Define $\theta_{ii}(D)$ such that $\mathbb{P}\left(\mathcal{G}(0,1) < \theta_{ii}(D)/\sqrt{2}\right) = f_{0,e_i}$
  $\theta_{ii}(D) \leftarrow \max(-5, \min(\theta_{ii}(D), 5))$
  $\hat{A}_{i,i}(D) \leftarrow \theta_{ii}(D) - \hat{B}_i(D)$ ▷ Estimate of $A_{i,i}/D$
end for
for $i = 1$ to d **do**
  for $j = i+1$ to d **do**
    $f_{0,e_i+e_j} \leftarrow \text{Cop}(K, 0, e_i + e_j, F_{\text{noisy}})$
    Define $\theta_{ij}(D)$ such that

$$\mathbb{P}(\mathcal{G}(0,1) < \theta_{ij}(D)/\sqrt{2}) = f_{0,e_i+e_j}$$

    $\theta_{ij}(D) \leftarrow \max(-5, \min(\theta_{ij}(D), 5))$
    $\hat{A}_{i,j}(D) \leftarrow \frac{1}{2}(\theta_{ij}(D) - \hat{B}_i(D)$
      $-\hat{A}_{i,i}(D) - \hat{B}_j(D) - \hat{A}_{j,j}(D))$
    $\hat{A}_{j,i}(D) \leftarrow \hat{A}_{i,j}(D)$ ▷ Estimate of $A_{i,j}/D$ and $A_{j,i}/D$
  end for
end for
$\hat{A}(D) \leftarrow (\hat{A}_{i,j}(D))$
$\hat{B}(D) \leftarrow (\hat{B}_i(D))$
if $\hat{A}(D)$ is not singular **then**
  $\hat{x} \leftarrow -\frac{1}{2}\hat{B}(D)^t \hat{A}(D)^{-1}$
else
  $\hat{x} \leftarrow 0$
end if
  return $\hat{x} \leftarrow$ projection of $\hat{x}$ on $B_d(0,1)$.

Theorem 3 *Let $\epsilon \in]0,1[$. Consider an objective function $F_{noisy}(x) = x^t Ax + Bx + C + D\mathcal{G}(0,1)$, with optimum x^* in $B_d(0, 1-\epsilon) \subset \mathbb{R}^d$, and $D > 0$. Assume that $\frac{1}{D}\|B\| \leq 1$ and*

$\frac{1}{D}|C| \leq 1$. *If A is symmetric positive definite such that its eigenvalues are lower bounded by some $c > 0$ and $\|\frac{1}{D}A\|_2 \leq 1$, then, when applying* CopQuad, $SR_N = O(\max((\lambda_{\max}/\lambda_{\min})^2, \lambda_{max}^2)D^2/N)$, *where $\lambda_{\max}$ is the maximum eigenvalue of $\frac{1}{D}A$, and $\lambda_{min} > \frac{1}{D}c$ is the minimum eigenvalue.*

Remark: Please note that $\lambda_{max} \leq 1$ by the assumptions in Theorem 3.

Proof 3 *Let x and y be two points to be compared in* CopQuad*: $(x, y) \in \mathcal{C} := \{(e_i, -e_i)_i, (0, e_i)_i, (0, e_i + e_j)_{i \neq j}\}$. We denote by $\Delta_{x,y}$ the value $\Delta_{x,y} := \mathbb{E}(F_{noisy}(y) - F_{noisy}(x)) = F(y) - F(x)$ and by $f_{x,y}$ the frequency $f_{x,y} := \frac{1}{K^2}\sum_{1 \leq i,j \leq K} \mathbf{1}_{f_x^i < f_y^j}$, where f_x^i and f_y^j are as in Section 2.*

Step 1: Mean Squared Error of frequencies.

Similarly to step 2 of Theorem 1, and using the notation $\Phi(x) = \mathbb{P}(\mathcal{G}(0,1) < x)$,

$$\mathbb{E}(f_{x,y}) = \Phi\left(\frac{\Delta_{x,y}}{\sqrt{2}D}\right)$$
$$\mathbb{E}\left(f_{x,y} - \Phi\left(\frac{\Delta_{x,y}}{\sqrt{2}D}\right)\right)^2 = Var(f_{x,y}) = O(1/N). \tag{7}$$

Step 2: Mean Squared Error of $\hat{A}(D)$ and $\hat{B}(D)$.

As in Step 3 of the proof of theorem 1, we denote by $\Phi^{-1}_{[\tilde{m},\tilde{M}]} : [\tilde{m}, \tilde{M}] \to [-5, 5]$ the inverse of Φ over these intervals:

$$\tilde{h}(x) = \begin{cases} \Phi^{-1}_{[\tilde{m},\tilde{M}]}(x) & \text{if } \tilde{m} \leq x \leq \tilde{M} \\ -5 & \text{if } x < \tilde{m} \\ 5 & \text{if } \tilde{M} < x \end{cases}$$

By assumption, $(x, y) \in \mathcal{C}$, $\frac{1}{D}\|A\|_2 \leq 1$ and $\frac{1}{D}\|B\| \leq 1$, $\Delta_{x,y}/\sqrt{2}D \in [-5, 5]$ and then, as in Step 3 and 4 of Theorem 1,

$$\begin{aligned} \mathbb{E}\left(\tilde{h}(f_{x,y}) - \frac{\Delta_{x,y}}{\sqrt{2}D}\right)^2 &\leq \mathbb{E}\left(\tilde{h}(f_{x,y}) - \tilde{h}\left(\Phi\left(\frac{\Delta_{x,y}}{\sqrt{2}D}\right)\right)\right)^2 \\ &\leq L(\tilde{h})^2\mathbb{E}\left(f_{x,y} - \Phi\left(\frac{\Delta_{x,y}}{\sqrt{2}D}\right)\right)^2 \\ &= O(1/N) \text{ by Eq. 7.} \end{aligned} \tag{8}$$

By applying Eq. 8, we then estimate the mean squared error of $\hat{A}(D)$ and $\hat{B}(D)$:

- *$\hat{B}_i(D) = \sqrt{2}\tilde{h}(f_{-e_i,e_i})/2$ and $B_i/D = \Delta_{-e_i,e_i}/2D$ $\forall i \in \{1, \ldots, d\}$, then $\mathbb{E}(\hat{B}_i(D) - B_i/D)^2 = O(1/N)$ by Eq. 8, hence $\mathbb{E}\|\hat{B}(D) - B/D\|^2 = O(1/N)$.*

- $\hat{A}_{i,i}(D) = \sqrt{2}\tilde{h}(f_{0,e_i}) - \hat{B}_i(D)$ *and* $A_{i,i}/D = \Delta_{0,e_i}/D - B_i/D$, *then* $\mathbb{E}(\hat{A}_{i,i}(D) - A_{i,i}/D)^2 = O(1/N)$ *using Eq. 8, and*

$$\mathbb{E}(\hat{B}_i(D) - B_i/D)^2 = O(1/N).$$

If $i \neq j$, *then*

$$\hat{A}_{i,j}(D) = \frac{1}{2}\left(\sqrt{2}\tilde{h}(f_{0,e_i+e_j}) \right.$$

$$\left. -\hat{B}_i(D) - \hat{A}_{i,i}(D) - \hat{B}_j(D) - \hat{A}_{j,j}(D)\right),$$

and

$$A_{i,j}/D =$$

$$1/2\left(\Delta_{0,e_i+e_j}/D - B_i/D - A_{i,i}/D - B_j/D - A_{j,j}/D\right)$$

hence, by proceeding as above,

$$\mathbb{E}(\hat{A}_{i,j}(D) - A_{i,j}/D)^2 = O(1/N)$$

and

$$\mathbb{E}\|\hat{A}(D) - A/D\|_F^2 = O(1/N).$$

Step 3: with probability at least $1-O(1/N)$**,** CopQuad **returns an estimate** $\hat{x}$ **solution of** $2\hat{x}\hat{A}(D) = -\hat{B}^t(D)$**.**

By definition of CopQuad , $2\hat{x}\hat{A}(D) \neq -\hat{B}^t(D)$ *only if* $\hat{x}$ *could not be properly defined because* $\hat{A}(D)$ *is singular or if we use the projection.*

The eigenvalues are continuous (see e.g. [11]); therefore in a neighborhood of A/D, $\hat{A}(D)$ *has eigenvalues lower bounded by some* $\delta > 0$. *Therefore,* $\hat{A}(D)$ *is singular only out of this neighborhood; this occurs, by Markov's inequality, with probability* $O(1/N)$. *Therefore, the first case occurs with probability at most* $O(1/N)$.

With probability at least $1 - O(1/N)$, *the solution* $\hat{x}$ *of* $2\hat{x}\hat{A}(D) = -\hat{B}^t(D)$ *is therefore the projection of* $-\frac{1}{2}\hat{B}(D)^t\hat{A}(D)^{-1}$. *For* $\hat{A}(D)$ *close enough to* A/D *and* $\hat{B}(D)$ *close enough to* B/D, *this is close to* x^*, *and therefore it is inside* $B_d(0, 1-\epsilon)$.

Step 4: concluding when $2\hat{x}\hat{A}(D) = -\hat{B}(D)^t$**.**

Define $B' = B/D - \hat{B}(D)$ *and* $A' = A/D - \hat{A}(D)$. *We have* $2x^*A = -B^t$ *and* $2\hat{x}\hat{A}(D) = -\hat{B}(D)^t$.

By substraction, we get

$$2(\hat{x}\hat{A}(D) - x^*A/D) = (B/D)^t - \hat{B}(D)^t$$

hence $2(\hat{x}A/D - \hat{x}A' - x^*A/D) = B'^t$, *using definitions of* A' *and* B'.

By step 2, all terms in A' *and* B' *have expected squared norm* $O(1/N)$; *and by step 3* $\hat{x}$ *is bounded, therefore*

$$2(\hat{x}A/D - x^*A/D) = B'^t + 2\hat{x}A'$$

has expected squared norm $O(1/N)$, *and*

$$(\hat{x} - x^*) = \frac{1}{2}uA^{-1}D$$

with $\mathbb{E}\|u\|^2 = O(1/N)$.

With $\lambda_{min} > 0$ *the smallest eigenvalue of* $\frac{1}{D}A$, *we get* $\mathbb{E}\|\hat{x} - x^*\|^2 = O(\lambda_{min}^{-2}/N)$.

Note that F *can be rewritten as*

$$F(x) = (x - x^*)^t A(x - x^*) + C',$$

where $x^* = -\frac{1}{2}B^t A^{-1}$ *and* $C' = C - x^{*t}Ax^*$.

$$\begin{aligned} \textit{Then } SR_N &= \|F(\hat{x}) - F(x^*)\|^2 = \|(\hat{x} - x^*)^t A(\hat{x} - x^*)\|^2 \\ &\leq \lambda_{max}^2 \|\hat{x} - x^*\|^2 \\ \textit{Hence } SR_N &= O\left(\left(\frac{\lambda_{max}}{\lambda_{min}}\right)^2 \frac{D^2}{N}\right), \textit{ which is the expected} \end{aligned}$$

result.

Step 5: General conclusion

Let us denote by $\mathcal{S}$ *the event "*CopQuad *returns an estimate* $\hat{x}$ *solution of* $2\hat{x}\hat{A}(D) = -\hat{B}(D)^t$*" and* $\bar{\mathcal{S}}$ *its complement. In the following, diam denotes the diameter. By definition,*

$$\begin{aligned} SR_N &= \mathbb{E}(F_{noisy}(\hat{x}) - F_{noisy}(x^*)) \\ &= \underbrace{\mathbb{E}(F_{noisy}(\hat{x}) - F_{noisy}(x^*)|\mathcal{S})}_{=O\left(\left(\frac{\lambda_{max}}{\lambda_{min}}\right)^2 \frac{D^2}{N}\right) \textbf{ by step } 4} \underbrace{\mathbb{P}(\mathcal{S})}_{\leq 1} \\ &+ \underbrace{\mathbb{E}(F_{noisy}(\hat{x}) - F_{noisy}(x^*)|\bar{\mathcal{S}})}_{\leq \lambda_{\max}^2 \times D^2 \times diam(B_d(0,1-\epsilon))} \underbrace{\mathbb{P}(\bar{\mathcal{S}})}_{=O(1/N) \textbf{by step } 3} \end{aligned}$$

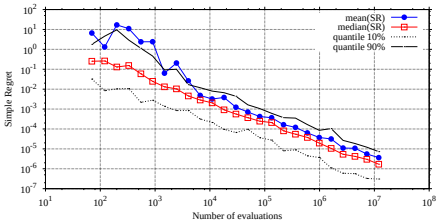
Hence the expected result.

5 Experiments

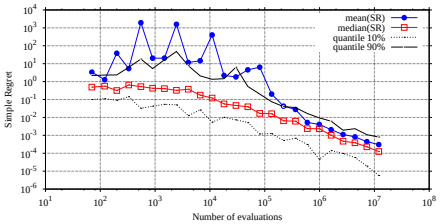
For each experiment, parameters A, B and C satisfying assumptions in Theorem 3 are randomly generated. CopQuad then returns an approximation of the optimum of the noisy quadratic function $F(x) = x^t Ax + Bx + C + D\mathcal{G}(0,1)$. Results are obtained over 50 runs.

CopQuad **to tackle strong noise.** Fig. 1 presents results of CopQuad in dimension 2 when the standard deviation D satisfies the assumptions in Theorem 3, i.e., $\|B\|/D \leq 1$, $|C|/D \leq 1$ and $\|A\|_2/D \leq 1$. The linear rate (in log-log scale) with slope -1 is clearly visible. We obtained similar graphs (not presented here) for dimensions 5.

CopQuad **with small noise.** Figure 2 then shows the case of a smaller noise D for dimension 2. Along with the theory ($\|A/D\|_2$ does not satisfy the assumptions), we lose the $O(1/N)$ rate. In the early stages, CopQuad still seems to converge, but it eventually stagnates around the optimum. It is counter-intuitive that an algorithm performs worse when noise decreases; nonetheless, in the case $\frac{1}{D}A \to 0$,



(a) $D = 1$



(b) $D = 10$

Figure 1: Dimension $d = 2$, over 50 runs. Mean, median and quantiles 10% and 90% are displayed.

the Cop operator always return 0 or 1, thus the estimated parameters are -5 or 5, and the algorithm does not converge. Incidentally, this is consistent with the bandit literature, where the hardest cases are when optimal arms have close values. Providing an algorithm able to cope with $D \leq \|A\|_2$ is possible - asymptotically, as for bandit algorithms mentioned above. Progressively widening the projection interval $[-b(N), b(N)]$ instead of keeping $[-5, 5]$ fixed makes this possible; if we have a slow enough function $b : N \mapsto b(N)$ for defining the interval $[-b(N), b(N)]$, then we get:

- e.g. $\log(\log(\log(N)))$ in Eq. 8,
- and asymptotically we still get a probability $1/N$ in Step 3 of Theorem 3.

So that, for $N > N_0$, we get Theorem 3 (up to the slight increase in the bound, depending on the choice of the b function) independently of $D \leq \|A\|_2$ - but N_0 depends on $\frac{1}{D}A$.

6 Conclusion

We have shown that comparison-based algorithms can reach a regret $O(1/N)$ on quadratic forms. This partially solves (negatively) a conjecture in [10], and improves results proposed in [4, 9]. Our main assumption is the Gaussian nature of the noise. We do not assume that the variance is known, but it is supposed to be constant.

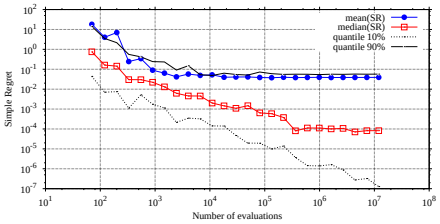


Figure 2: $d = 2$, $D = 0.65$.

Future work. We assume an exactly quadratic function; maybe rates in $O(1/N^{2/3})$ can be reached for non-quadratic functions under smoothness assumptions. Also we might extend the present results to non Gaussian noise.

References

[1] S. Astete-Morales, M.-L. Cauwet, and O. Teytaud. Evolution Strategies with Additive Noise: A Convergence Rate Lower Bound. In Foundations of Genetic Algorithms, Foundations of Genetic Algorithms, page 9, Aberythswyth, United Kingdom, 2015.

[2] H.-G. Beyer. Mutate Large, But Inherit Small! On the Analysis of Rescaled Mutations in $(\tilde{1}, \tilde{\lambda})$-ES with Noisy Fitness Data. In Parallel Problem Solving from Nature, 5, Heidelberg, 1998. Springer. in print.

[3] H. F. Chen, T. E. Duncan, and B. Pasik-Duncan. A stochastic approximation algorithm with random differences. In Proceedings of the 13th IFAC World Congress, volume H, pages 493–496, 1996.

[4] J. Decock and O. Teytaud. Noisy optimization complexity under locality assumption. In Proceedings of the twelfth workshop on Foundations of genetic algorithms XII, FOGA XII '13, pages 183–190, New York, NY, USA, 2013. ACM.

[5] V. Dupač. Notes on stochastic approximation methods. Czechoslovak Mathematical Journal, 08(1):139–149, 1958.

[6] V. Fabian. Stochastic Approximation of Minima with Improved Asymptotic Speed. Annals of Mathematical statistics, 38:191–200, 1967.

[7] K. G. Jamieson, R. Nowak, and B. Recht. Query complexity of derivative-free optimization. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 2672–2680. Curran Associates, Inc., 2012.

[8] J. Kiefer, J. Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics, 23(3):462–466, 1952.

[9] P. Rolet and O. Teytaud. Adaptive noisy optimization. In C. Di Chio, S. Cagnoni, C. Cotta, M. Ebner, A. Ekrt, A. Esparcia-Alcazar, C.-K. Goh, J. Merelo, F. Neri, M. PreuY, J. Togelius, and G. Yannakakis, editors, Applications of Evolutionary Computation, volume 6024 of Lecture Notes in Computer Science, pages 592–601. Springer Berlin Heidelberg, 2010.

[10] O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA, pages 3–24, 2013.

[11] M. Zedek. Continuity and location of zeroes of linear combinations of polynomials. Proc. Amer. Math. Soc., 16:78–84, 1965.