HAL
open science

Simple and Cumulative Regret for Continuous Noisy Optimization

Sandra Astete-Morales, Marie-Liesse Cauwet, Jialin Liu, Olivier Teytaud

To cite this version:

Sandra Astete-Morales, Marie-Liesse Cauwet, Jialin Liu, Olivier Teytaud. Simple and Cumulative Regret for Continuous Noisy Optimization. Theoretical Computer Science, 2015, 617, pp.12-27. 10.1016/j.tcs.2015.09.032 . hal-01194564

HAL Id: hal-01194564

https://inria.hal.science/hal-01194564v1

Submitted on 20 Sep 2015

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.

Simple and Cumulative Regret for Continuous Noisy Optimization

Sandra Astete-Morales, Marie-Liesse Cauwet, Jialin Liu, Olivier Teytaud
TAO, INRIA-CNRS-LRI, Univ. Paris-Sud, 91190 Gif-sur-Yvette, France
email:{firstname.lastname}@lri.fr

Preprint submitted to Elsevier *September 20, 2015*

Simple and Cumulative Regret for Continuous Noisy Optimization

Sandra Astete-Morales, Marie-Liesse Cauwet, Jialin Liu, Olivier Teytaud
TAO, INRIA-CNRS-LRI, Univ. Paris-Sud, 91190 Gif-sur-Yvette, France
email:{firstname.lastname}@lri.fr

Abstract

Various papers have analyzed the noisy optimization of convex functions. This analysis has been made according to several criteria used to evaluate the performance of algorithms: uniform rate, simple regret and cumulative regret.

We propose an iterative optimization framework, a particular instance of which, using Hessian approximations, provably (i) reaches the same rate as Kiefer-Wolfowitz algorithm when the noise has constant variance (ii) reaches the same rate as Evolution Strategies when the noise variance decreases quadratically as a function of the simple regret (iii) reaches the same rate as Bernstein-races optimization algorithms when the noise variance decreases linearly as a function of the simple regret.

Keywords: Noisy optimization, runtime analysis

1. Introduction

The term Noisy Optimization refers to the search for the optimum of a given stochastic objective function $f : (x, \omega) \mapsto f(x, \omega)$ where x is in the search domain $\mathcal{D} \in \mathbb{R}^d$ and ω is some random process. From now on, we assume the existence of some x^* such that $\mathbb{E}_\omega f(x^*, \omega)$ is minimum. Many results regarding the performance of algorithms at solving these problems and at the complexity of the problems themselves have been developed in the past, always trying to broaden the extent of them. In this paper we propose an optimization algorithm that allows us to generalize results in the literature as well as providing proofs for conjectured results.

We start by stating shortly the framework and definitions of the concepts we will review. Then we comment the state of the art related to stochastic optimization, which will bring us to the specific motivation of this paper. We finish this section with an outline of the reminder of the work.

1.1. Framework

When the gradient is available for the optimization process, there are algorithms developed in the literature that show a good performance: moderate numbers of function evaluations and good precision. Such is the case for Stochastic Gradient Descent, which is the stochastic version of the classic Gradient Descent Algorithm.

Nonetheless, having access to a gradient is a major assumption in real life scenario. Therefore, in this paper, we focus on a black-box case, i.e. we do not use any internal property of f, we only have access to function evaluations for points in the search space.

Indexation in the number of evaluations. A noisy black-box optimization algorithm at iteration $m \geq 1$: (i) chooses a new point x_m in the domain and computes its objective function value $y_m = f(x_m, \omega_m)$, where the ω_m are independent copies of ω; (ii) computes an approximation $\tilde{x}_m$ of the unknown optimum x^*.

Therefore, at the end of the application of the noisy optimization algorithm, we obtain several sequences: the search points $(x_m)_{m\geq 1}$, the function value on the search points $(y_m)_{m\geq 1}$ and the approximations $(\tilde{x}_m)_{m\geq 1}$. Let us note that each search point x_m is a computable function of the previous search point and their respective function values. But the computation of the search point involves random processes: the stochasticity of the function, and/or some specific random process of the algorithm, if the latter is randomized. The point $\tilde{x}_m$ is termed *recommendation*, and it represents the current approximation of x^*, chosen by the algorithm. Even though in many cases, the recommendation and the search points are exactly the same, we will make a difference here because in the noisy case it is known that algorithms which do not distinguish recommendations and search points can lead to poor results[1], depending on the noise level.

Indexation in the number of iterations. Depending on the study that one is carrying, there are arguments for indexing the sequences by *iter-*

[1]See Fabian (1967); Coulom (2012) for more on this.

ations or by *function evaluations*. It occurs often in the case of optimization of noisy functions, that it is more convenient to have multiple evaluations per iteration. In particular, a classical scheme is to generate a population of search points from a central point at each iteration (Fabian, 1967; Dupač, 1957; Coulom, 2012; Shamir, 2013). This mechanism is used in the present paper. Therefore, it is more convenient to introduce the iteration index rather than indexing the sequences by the number of evaluations, when we describe algorithms - but we use indexations by evaluations when we evaluate convergence rates, and in particular for the slopes of the convergence defined later. We then describe the "dual" notations, with iteration index, and we explain how to switch from an indexation to the other.

$x_{m,1}, \ldots, x_{m,r_m}$ denote the r_m *search points* at iteration m. When we need to access to the mth *evaluated search point*, we define x'_m the mth evaluated search point, i.e. $x'_m = x_{i,k}$ with $m = \sum_{j=1}^{i-1} r_j + k$ and $k \leq r_i$. On the other hand, x_m^{opt}, with only one subscript, is the *recommended point at iteration* m. $\tilde{x}_n$ will always denote the recommendation *after* n *evaluations*. Hence, when the approximations of the optimum are defined per iteration rather than per evaluation, the sequence of recommended points is redefined as follows: for all $n \geq 1$, $\tilde{x}_n = x_k^{\text{opt}}$, where k is maximal such that $\sum_{i=1}^{k-1} r_i \leq n$.

Now that we have defined the basic notations for the algorithms considered in this work, let us introduce the *optimization criteria* which will evaluate the *performance* of the algorithms. They allow us to compare the performance of the algorithm considering all search points or only the recommended points. The information we have on the *cost* of evaluating search points can be the dealbreaker when choosing what algorithm to use, as long as we have specialized optimization criteria to help us decide. We will consider three criteria: Uniform Rate (UR), Simple Regret (SR) and Cumulative Regret (CR), respectively defined in Eqs. 1, 2 and 3.

$$s(UR) = \limsup_i \frac{\log(UR_i)}{\log(i)} \tag{1}$$

$$s(SR) = \limsup_i \frac{\log(SR_i)}{\log(i)} \tag{2}$$

$$s(CR) = \limsup_i \frac{\log(CR_i)}{\log(i)} \tag{3}$$

where UR_i is the $1-\delta$ quantile of $\|x'_i - x^*\|$, SR_i is the $1-\delta$ quantile of $\mathbb{E}_\omega f(\tilde{x}_i, \omega) - \mathbb{E}_\omega f(x^*, \omega)$, CR_i is the $1-\delta$ quantile of $\sum_{j \leq i} \big(\mathbb{E}_\omega f(x'_j, \omega) - \mathbb{E}_\omega f(x^*, \omega)\big)$. $\|.\|$ stands for the Euclidean norm, x'_i de-

notes the ith evaluated search point and $\tilde{x}_i$ denotes the recommendation after i evaluations. We have expectation operators $\mathbb{E}_\omega$ above with respect to ω only, therefore $\mathbb{E}_\omega f(\tilde{x}_i, \omega)$ is not deterministic. Quantiles are with respect to all remaining stochastic parts such as noise in earlier fitness evaluations and possibly internal randomness of the optimization algorithm.

In Eqs. 1, 2 and 3, we consider the *slopes* in log-log graphs (x-axis: log of evaluation numbers; y-axis: log of UR or SR or CR). The use of the *slopes* turns out to be more convenient because it allow us to know with one single number how fast an algorithm is reaching the specific optimization criterion.

These quantities depend on the threshold δ, but in all cases below we get the same result independently of δ, therefore we will drop this dependency. Note that, for $s(UR)$ and $s(SR)$, 0 can be trivially reached by an algorithm with constant $(x'_m, \tilde{x}_m)$. Therefore, $s(UR)$ and $s(SR)$ are only interesting when they are less than 0. And $s(CR)$ is relevant when it is less than 1.

Finally, regarding to the objective functions, we investigate three types of noise models :

$$Var(f(x,\omega)) = O\left([\mathbb{E}_\omega f(x,\omega) - \mathbb{E}_\omega f(x^*,\omega)]^z\right) \qquad z \in \{0,1,2\} \tag{4}$$

We will refer to them respectively as the case where the variance of the noise is *constant*, *linear* and *quadratic* as a function of the simple regret.

1.2. State of the art

When solving problems in real life, having access to noisy evaluations of the function to be optimized, instead of the real evaluations, can be a very common issue. If, in this context, we have access to the gradient of the function, the Stochastic Gradient Method is particularly appreciated for the optimization, given its efficiency and its moderate computational cost (Bottou and Bousquet, 2011). However, the most general case consists in only having access to the function evaluations in certain points (i.e. black-box setting, as previously defined). This setting is specially relevant in cases such as reinforcement learning, where gradients are difficult and expensive to get (Sehnke et al., 2010). For example, Direct Policy Search, an important tool from reinforcement learning, usually boils down to choosing a good representation (Bengio, 1997) and applying black-box noisy optimization (Heidrich-Meisner and Igel, 2009).

Among the noisy optimization methods, we find in the work of Robbins and Monro (1951), the ground-break proposal to face the problem of having

noise in one or more stages of the optimization process. From this method derive other important methods as the type of stochastic gradient algorithms. On a similar track, Kiefer and Wolfowitz (1952) has also added tools based on finite difference inside of this kind of algorithms. In a more general way, Spall (2000, 2003, 2009) designed various algorithms which can be adapted to several settings, with or without noise, with or without gradient, with a moderate number of evaluations per iteration.

The tools based on finite differences are classical for approximating derivatives of functions in the noise-free case. Nonetheless, the use of finite differences is usually expensive. Therefore, for instance, quasi-Newton methods also use successive values of the gradients for estimating the Hessian (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970). And this technique has also been applied in cases in which the gradient itself is unknown, but approximated using successive objective function values (Powell, 2004, 2008). With regards to latter method, so-called NEWUOA algorithm, it presents impressive results in the black-box noise-free case but this results do not translate into the noisy case, as reported by Ros (2009).

In this work we refer to three optimization criteria to study the convergence of algorithms, so-called *uniform, simple and cumulative regret*, by taking into account the *slope* on the *log-log* graph of the criteria v/s the number of evaluations (see $s(UR)$, $s(SR)$ and $s(CR)$ defined in Eqs. 1, 2 and 3). The literature in terms of these criteria is essentially based on stochastic gradient techniques.

Sakrison (1964); Dupač (1957) have shown that $s(SR) = -\frac{2}{3}$ can be reached, when the objective function is twice differentiable in the neighborhood of the optimum when the noise has a bounded variance . This original statement has been broadened and specified. Spall (2000) obtained similar results with an algorithm using a small number of evaluations per iteration, and an explicit limit distribution. The small number of evaluations per iteration makes stochastic gradient way more practical than earlier algorithms such as Dupač (1957); Fabian (1967), in particular in high dimension where these earlier algorithms were based on huge numbers of evaluations per iterations[2]. In addition, their results can be adapted to noise-free settings and

[2]It must be pointed out that in the algorithms proposed in Dupač (1957); Fabian (1967), the number of evaluations per iteration is constant, so that the rate $O(1/n)$ is not modified by this number of evaluations. Still, the number of evaluations is exponential in the dimension, making the algorithm intractable in practice.

provide non trivial rates in such a setting.

Regarding the case of noise with constant variance (see case $z = 0$ in Eq. 4), Fabian (1967) made a pioneering work with a simple regret arbitrarily close to $O(1/n)$ after n evaluations, i.e. $s(SR) \simeq -1$, provably tight as shown by Chen (1988), when higher derivatives exist. Though they use a different terminology than recent papers in the machine learning literature, Fabian (1967) and Chen et al. (1996) have shown that stochastic gradient algorithms with finite differences can reach $s(UR) = -\frac{1}{4}$, $s(SR) = -1$ and $s(CR) = \frac{1}{2}$ on quadratic objective functions; the slopes $s(SR) = -1$ and $s(CR) = \frac{1}{2}$ are optimal in the general case as shown by, respectively, (Chen, 1988) (simple regret) and Shamir (2013) (cumulative regret). Shamir (2013) also extended the analysis in terms of dependency in the dimension and non-asymptotic results - switching to $-\frac{1}{2}$ for twice differentiable functions, in the non-asymptotic setting, as opposed to $-\frac{2}{3}$ for Dupač (1957) in the asymptotic setting.

Rolet and Teytaud (2009, 2010); Coulom et al. (2011) take into consideration functions with $\mathbb{E}_\omega f(x,\omega) - \mathbb{E}_\omega f(x^*,\omega) = C\|x - x^*\|^p$ $(p \geq 1)$ and different intensity on the perturbation; one with noise variance $\Theta(1)$ and the second with variance $Var(f(x,\omega)) = O(\mathbb{E}_\omega f(x,\omega) - \mathbb{E}_\omega f(x^*,\omega))$. In the case of "strong" noise (i.e. variance $\Theta(1)$), they prove that the optimal $s(UR)$ is in $[-\frac{1}{p}, -\frac{1}{2p}]$. In the case of $z = 1$ as well, $s(UR) = -\frac{1}{p}$ was obtained in Decock and Teytaud (2013), with tightness for algorithms matching some "locality assumption".

The study of noise variance decreasing quickly enough, for some simple functions, has been performed in Jebalia et al. (2010), where it is conjectured that one can obtain $s(UR) = -\infty$ and geometric convergence $(\|x_n\| = O(\exp(-\Omega(n))))$ - we prove this $s(UR) = -\infty$ for our algorithm.

Another branch of the state of the art involves Bernoulli random variables as objective functions: for a given x, $f(x,\omega)$ is a Bernoulli random variable with probability of failure $\mathbb{E}_\omega f(x,\omega)$. We wish to find x such that $\mathbb{E}_\omega f(x,\omega)$ is minimum. This framework is particularly relevant in games (Chaslot et al., 2008; Coulom, 2012) or viability applications (Aubin, 1991; Chapel and Deffuant, 2006) and it is a natural framework for $z = 1$ (when the optimum value $\mathbb{E}_\omega f(x^*,\omega)$ is 0) in the Bernoulli setting, the variance at x is linear as a function of the simple regret $\mathbb{E}_\omega f(x,\omega) - \mathbb{E}_\omega f(x^*,\omega)$ or for $z = 0$ (i.e. when the optimum value is > 0). In these theoretical works, the objective function at x is usually a Bernoulli random variable with parameter

depending on $a + b\|x - x^*\|^\zeta$ for some $\zeta \geq 1, a \geq 0, b > 0$. Some of the lower bounds below hold even when considering only Bernoulli random variables, while upper bounds usually hold more generally .

Notice that by definition the $s(UR)$ criterion is harder to be reached than $s(SR)$ because all search points must verify the bound, not only the recommended ones - for any problem, if for some algorithm, $s(UR) \leq c$, then for the same problem there is an algorithm such that $s(SR) \leq c$. There are also relations between $s(CR)$ and $s(UR)$, at least for algorithms with a somehow "smooth" behavior; we will give more details on this in our conclusions.

In the case of noise with constant variance, the best performing algorithms differ, depending on the optimization criteria (UR, SR, CR) that we choose. On the other hand, when variance decreases at least linearly, the algorithms used for reaching optimal $s(UR)$, $s(SR)$ and $s(CR)$ are the same. In all frameworks, we are not aware of differences between algorithms specialized on optimizing $s(UR)$ criterion and on $s(CR)$ criterion. An interesting remark on the differences between the criteria is the following: optimality for $s(SR)$ and $s(CR)$ can not be reached simultaneously. This so-called *tradeoff* is observed in discrete settings (Stoltz et al., 2011) and also in the algorithm presented in Fabian (1967), to which we will refer as Fabian's algorithm. Fabian's algorithm is a gradient descent algorithm, using finite differences for approximating gradients. Depending on the value of a parameter, namely γ_{Fabian} [3], we get a good $s(SR)$ or a good $s(CR)$, but never both simultaneously. In the case of quadratic functions with additive noise (constant variance $z = 0$):

- $\gamma_{Fabian} \to \frac{1}{4}$ leads to $s(SR) = -\frac{1}{2}$ and $s(CR) = \frac{1}{2}$;
- $\gamma_{Fabian} \to 0$ leads to $s(SR) = -1$ and $s(CR) = 1$.

The algorithms analyzed in this paper, as well as Fabian's algorithm or the algorithm described in Shamir (2013), present this tradeoff, and similar rates. A difference between the latter algorithm and ours is that ours have faster rates when the noise decreases around the optimum and proofs are included for other criteria (see Table 1). The cases with variance decreasing towards zero in the vicinity of the optimum are important, for example in the Direct Policy Search method when the fitness function is 0 for a success

[3] Fabian (1967) defines a sequence $c_n = cn^{-\gamma_{Fabian}}$, which specifies the sequence of finite difference widths used for the gradient approximation.

and 1 for a failure; if a failure-free policy is possible, then the variance is null at the optimum. Another example is parametric classification, when there exists a classifier with null error rate: in such a case, variance is null at the optimum, and this makes convergence much faster (Vapnik (1995)).

Importantly, some of the rates discussed above are for stronger convergence criteria than ours. For example, Fabian (1967) gets almost sure convergence. Shamir (2013) gets convergence in expectation. Spall (2000) gets asymptotic distributions. We get upper bounds on quantiles of various regrets, up to constant factors.

1.3. Motivation & key ideas

This section discusses the motivations for this paper. First, to obtain new bounds and recover existing bounds within a single algorithm (Section 1.3.1). Second, proving results in a general approximation setting, beyond the classical approximation by quadratic models; this is in line with the advent of many new surrogate models in the recent years (Section 1.3.2).

1.3.1. Generalizing existing bounds for noisy optimization

We here extend the state of the art in the case of $z = 2$ for all criteria, and $z = 1$ for more general families of functions (published results were only for sphere functions), and get all the results with a same algorithm. We also generalize existing results for UR or SR or CR to all three criteria. On the other hand, we do not get Fabian's $s(SR)$ arbitrarily close to -1 on smooth non-quadratic functions with enough derivatives, which require a different schema for finite differences and assumes the existence of a large number of additional derivatives.

1.3.2. Hessian-based noisy optimization algorithms and beyond

We propose and study a noisy optimization algorithm, which possibly uses a Newton-style approximation, i.e. a local quadratic model. Gradient-based methods (without Hessian) have a difficult parameter, which is the rate at which gradient steps are applied. Such a problem is solved when we have a Hessian; the gradient and Hessian provide a quadratic approximation of the objective function, and we can use, as next iterate, the minimum of this quadratic approximation. There are for sure new parameters, associated to the Hessian updates, such as the widths used in finite differences; however other algorithms, without Hessians, already have such parameters (e.g. Fabian (1967); Shamir (2013)). Such a model was already proposed in

Fabian (1971), a few years after his work establishing the best rates in noisy optimization (Fabian, 1967), but without any proof of improvement. Spall (2009) also proposed an algorithm based on approximations of the gradient and Hessian, when using the SPSA (simultaneous perturbation stochastic approximation) method (Spall, 2000). They provided some results on the approximated Hessian and on the convergence rate of the algorithm; Spall (2000) and Spall (2009) study the convergence rate in the search space, but their results can be converted in simple regret results and in this setting they get the same slope of simple regret $-\frac{2}{3}$ as Dupač (1957); the paper also provides additional information such as the limit distribution and the dependency in the eigenvalues. The algorithms in Spall (2000, 2009) work with a very limited number of evaluations per iteration, which is quite convenient in practice compared to numbers of evaluations per iteration exponential in the dimension in Fabian (1967).

In this paper, we propose a Hessian-based algorithm which provably covers all the rates above, including UR, SR and CR and $z = 0, 1, 2$ - except the -1 slope (by Fabian) for SR in the case $z = 0$ and infinitely many derivatives, which uses more assumptions than our results. Our results are summarized in Table 1. Importantly, our algorithm is not limited to optimization with black-box approximations of gradients and Hessians; we consider more generally algorithms with *low-squared error* (LSE) (Def. 2.1).

1.4. Summary of results

In this paper, we show general results on iterative noisy optimization, based on some properties of optimum estimates. Our proposed algorithm recovers most existing results, except the slope of simple regret -1 obtained by Fabian (1967) when arbitrarily many derivatives are supposed to exist. In particular, using a noisy evaluation of the gradient and Hessian, we get at best $s(SR) = -1$ or $s(CR) = \frac{1}{2}$ (not simultaneously; the former with parameters optimized for SR and the latter with parameters optimized for CR) for constant noise variance on quadratic positive definite functions, as well as $s(SR) = -\frac{2}{3}$ and $s(CR) = \frac{1}{2}$ (also not simultaneously; the former with parameters optimized for SR and the latter with parameters optimized for CR) for functions which have positive definite second order Taylor expansion, as in Fabian (1967); Shamir (2013) respectively.

We can also get $s(SR) = -\frac{2}{3}$ and $s(CR) = \frac{2}{3}$ (simultaneously) in the same setting. We get $s(SR) = -1$ and $s(CR) = 0$ (simultaneously) with linearly decreasing variance as in Rolet and Teytaud (2010), $s(SR) = -\infty$

and $s(CR) = 0$ (simultaneously) with quadratically decreasing variance as conjectured in Jebalia and Auger (2008) - for a different algorithm. In addition, our results are applicable with arbitrary surrogate models, provided that they verify the LSE assumption (Definition 2.1).

2. The Iterative Noisy Optimization Algorithm (Inoa)

Section 2.1 presents a general iterative algorithm, which uses a *sampling tool* and an *optimum estimator*. It relies on a *LSE assumption* (Def. 2.1) which is central in the assumptions for the main theorem. Section 2.2 provides examples of sampling tools, called Sampler functions, and examples of optimum estimators, given by Opt functions, which match the assumptions in Section 2.1.

2.1. General framework

The Iterative Noisy Optimization Algorithm (Inoa) is presented in Alg. 1. It uses a pair of functions (Sampler, Opt). Specific tasks and properties of these functions are described below and examples of such functions are given in Section 2.2.

Sampler is the element of the algorithm that provides new search points: given a point x in the search space, Sampler provides new search points that lie on the neighborhood of x. More precisely, $\forall i \in \{1, 2, \dots\}$, $\text{Sampler}(x, \sigma, i)$ outputs a point x_i such that it satisfies $\|x_i - x\| \leq 2\sigma$, with σ a given step-size. Notice that we do not make any assumptions on *how* the new search points are chosen, we only ask for them to be within a given maximal distance from the generator point x. Opt corresponds to the optimum estimator of the algorithm: given x, $x_1, \dots, x_r$ and $y_1, \dots, y_r$ with $y_i = f(x_i, \omega_i)$ (with ω_i independent copies of ω), Opt provides an estimate $x^{\text{opt}} := \text{Opt}(x, (x_i, y_i)_{i \in \{1,\dots,r\}})$ of x^*, the argmin of $\mathbb{E}f$. Additionally, for the sake of convergence, the pair (Sampler, Opt) verifies a property defined in Def. 2.1 and called the *Low squared error assumption (LSE)*.

The algorithm provides the sequence $(x_n^{\text{opt}})_{n \geq 1}$, indexed with the number of *iterations*, but the recommendations $(\tilde{x}_n)_{n \geq 1}$ in the definitions of Section 1.1 have to be indexed by the number of *evaluations*. Hence, for $m \geq 1$, the recommendation $\tilde{x}_m$ are defined by $\tilde{x}_m = x_{n(m)}$ with $n(m) = \max\{n; \sum_{i=1}^{n-1} r_i \leq m\}$, since there are r_i evaluations at iteration i.

Algorithm 1 Iterative Noisy Optimization Algorithm (INOA).

Input:
Step-size parameters $\alpha > 0$, $A > 0$
Number of revaluations parameters $\beta \geq 0$, $B > 0$
Initial points $x_1^{\text{opt}} = \tilde{x}_1$
A fitness function (also termed noisy objective function)
A sampler function $\textsc{Sampler}(\cdot)$
An optimizer function $\textsc{Opt}(\cdot)$
Output: approximations $(x_n^{\text{opt}})_{n\geq 1}$, recommendations $(\tilde{x}_m)_{m\geq 1}$, evaluation points $(x_{n,i})_{n\geq 1, i\in\{1,\ldots,r_n\}}$, fitness evaluations $(y_{n,i})_{n\geq 1, i\in\{1,\ldots,r_n\}}$
$n \leftarrow 1$
while The computation time is not elapsed **do**
    Compute step-size $\sigma_n = A/n^\alpha$
    Compute revaluations number $r_n = B\lceil n^\beta \rceil$
    for $i = 1$ to r_n **do**
        $x_{n,i} = \textsc{Sampler}(x_n^{\text{opt}}, \sigma_n, i)$
        $y_{n,i} =$ fitness evaluation at $x_{n,i}$
    end for
    Compute next approximation $x_{n+1}^{\text{opt}} = \textsc{Opt}(x_n^{\text{opt}}, (x_{n,i}, y_{n,i})_{i\in\{1,\ldots,r_n\}})$
    $n \leftarrow n+1$
end while

Definition 2.1 (Low squared error assumption (LSE)). *Given a domain $D \subseteq \mathbb{R}^d$, an objective function $f : D \to \mathbb{R}$ corrupted by noise. We assume that f is such that $\mathbb{E}_\omega f(x,\omega)$ has a unique optimum x^*. Let $C > 0$, $U > 0$, and $z \in \{0,1,2\}$. Then, we say that* (SAMPLER, OPT) *has a $(2z-2)$-low squared error for f, C, U, S if $\forall (r,\sigma) \in S$*

$$\|x - x^*\| \leq C\sigma \implies \textit{for any positive integer } r, \quad \mathbb{E}(\|x^{\text{opt}} - x^*\|^2) \leq U\frac{\sigma^{2z-2}}{r}, \tag{5}$$

where x^{opt} is provided by the OPT *function, which receives as input*

- *the given x,*
- *r search points $(x_i)_{i\in\{1,\ldots,r\}}$, outputs of* SAMPLER*,*
- *and their corresponding noisy fitness values.*

In the latter definition, z is related to the intensity of the noise. Recall that we consider three types of noise, namely *constant*, *linear* or *quadratic*

in function of the SR. More precisely, we consider that $Var(f(x,\omega)) = O\left([\mathbb{E}_\omega f(x,\omega) - \mathbb{E}_\omega f(x^*,\omega)]^z\right)$ with $z \in \{0,1,2\}$.

The rate $O(1/r)$ for a squared error is typical in statistics, when estimating some parameters from r samples. We will see in examples below that the scaling with σ is also relevant, as we recover, with the LSE as an intermediate property, many existing rates.

We can work with the additional assumption that $x^* = 0$ without loss of generality. Hence from now on, examples, proofs and theorems are displayed with $x^* = 0$.

2.2. Examples of algorithms verifying the LSE assumption

In this section we provide two examples of pairs (Sampler, Opt) which verify Def. 2.1. Not only Sampler and Opt are important, but also the type of functions we consider (conditions for expectation and variance on the properties that show the verification of LSE). The first example uses an estimation of the gradient of the function to produce an approximation to the optimum. The idea is simple: if we have x, a current approximation to the optimum, we sample around it and use these points to estimate the gradient and obtain the next approximation.

Let $(e_j)_{j=1}^d$ be the canonical orthonormal basis of $\mathbb{R}^d$. Sampler will output search points $x \pm \sigma e_j$ for some $j \in \{1,\dots d\}$. Therefore, the set of points that Sampler has access to is $E' := E'_+ \cup E'_-$ where $E'_+ = (x+\sigma e_j)_{j=1}^d$ and $E'_- = (x-\sigma e_j)_{j=1}^d$, and E' is ordered[4]. In this example, when Sampler is queried for the i-th time it will output the i-th point of E'. For the case $i > 2d = |E'|$, to simplify the notation we define a slightly different version of the usual modulo operation, denoted "$\overline{\text{mod}}$", such that for any i, d, $i \ \overline{\text{mod}}\ d = 1 + ((i-1) \bmod d)$. Therefore, when $i > 2d = |E'|$, Sampler will output the $(i \ \overline{\text{mod}}\ 2d)$-th point of E'. We assume that Sampler outputs at the end a sample of r points, all belonging to E'. Note that as soon as $r > 2d$ the search points are sampled several times. However, the values of the objective function of the same search point evaluated two or more times will differ due to the noise in the evaluation. On the other hand, Opt takes this regular repeated sample around x and its corresponding objective function values to compute an average value for each of the points in E'. Hence, the average is done over at least $\lfloor r/(2d) \rfloor$ function evaluations and it allows to

[4] $E' = \{x+\sigma e_1, \dots, x+\sigma e_d, x-\sigma e_1, \dots, x-\sigma e_d\}$

reduce the noise and obtain a more confident - still noisy - evaluation. With these averaged values, OPT computes the approximated optimum. Let us consider

$$Y_{j+} = \{\text{all evaluations of } x + \sigma e_j\} \text{ and } Y_{j-} = \{\text{all evaluations of } x - \sigma e_j\}$$

and use the notation $x^{(j)}$ to refer to the j-th coordinate of x. Also, when we use $\sum Y_{j+}$, with Y_{j+} a set, it will simply denote that we sum over all the elements of the multiset Y_{j+}.

Example 1 Gradient based method verifying the LSE assumption (Def. 2.1). Given $x \in \mathbb{R}^d$ and $\sigma > 0$, SAMPLER and OPT are defined as follows.

---

function SAMPLER(x, σ, i)

$$j \leftarrow i \ \overline{\text{mod}}\ 2d \tag{6}$$

$$x_i \leftarrow \text{the } j\text{-th point in } E' \tag{7}$$

  return x_i
end function

---

function OPT($x, (x_i, y_i)_{i \in \{1,\dots,r\}}$)
  for $j = 1$ to d **do**

$$\hat{y}_{j+} \leftarrow \frac{1}{|Y_{j+}|} \sum Y_{j+}, \qquad \hat{y}_{j-} \leftarrow \frac{1}{|Y_{j-}|} \sum Y_{j-} \tag{8}$$

$$\hat{g}^{(j)} \leftarrow \frac{\hat{y}_{j+} - \hat{y}_{j-}}{2\sigma} \tag{9}$$

  end for

$$x^{\text{opt}} \leftarrow x - \frac{1}{2}\hat{g}$$

  return x^{opt}
end function

Property 2.1 enunciates the fact that the pair (SAMPLER, OPT) defined in Example 1 satisfies the Low Squared Error assumption (Def. 2.1).

Property 2.1. (SAMPLER, OPT) in Example 1 satisfy $(2z-2)$-LSE for the sphere function.

Let f be the function to be optimized, and $z \in \{0, 1, 2\}$. We assume that:

Framework 1

$$\mathbb{E}_\omega f(x, \omega) = \|x\|^2 \tag{10}$$

$$Var(f(x, \omega)) = O(\|x\|^{2z}) \text{ for some } z \in \{0, 1, 2\} \tag{11}$$

Then there is $C > 0$, such that if x and σ verify $\|x\| \leq C\sigma$, then

$$\mathbb{E}(\|x^{\text{opt}}\|^2) = O(\sigma^{2z-2}/r). \tag{12}$$

where x^{opt} is the output of $\textsc{Opt}(x, (x_i, y_i)_{i\in\{1,\dots,r\}})$, $(x_i)_{i\in\{1,\dots,r\}}$ *is the output of* Sampler *and $(y_i)_{i\in\{1,\dots,r\}}$ their respective noisy fitness values.*

Proof: We know that $\mathbb{E}(\|x^{\text{opt}}\|^2) = \sum_{j=1}^{d} \left\{ Var(x^{\text{opt}(j)}) - (\mathbb{E}(x^{\text{opt}(j)}))^2 \right\}$. For all $j \in \{1, \dots d\}$, using the definition of $\hat{g}^{(j)}$ in Eq. 9 and using Eq. 10 we obtain

$$\mathbb{E}(\hat{g}^{(j)}) = 2x^{(j)} \Rightarrow \mathbb{E}(x^{\text{opt}(j)}) = 0$$

Now, using the variance of the noisy function in Eq. 11 and the fact that $\|x\| \leq C\sigma$,

$$Var(\hat{g}^{(j)}) = O\left(\frac{\sigma^{2z-2}}{r}\right) \Rightarrow \mathbb{E}(\|x^{\text{opt}}\|^2) = O\left(\frac{\sigma^{2z-2}}{r}\right)$$

□

The method using gradients described above is already well studied, as well as improved variants of it with variable step-sizes, (see Fabian (1967); Chen (1988); Shamir (2013)).

Therefore, we now switch to the second example, including the computation of the Hessian.

As in the Example 1, we consider a set of search points that are available for Sampler to output. Let us define $E'' = \{x \pm \sigma e_i \pm \sigma e_j; 1 \leq i < j \leq d\}$. And so the sample set will be E, which includes the set E'' defined above and the sample set E' defined for Example 1. Therefore, $|E| = 2d^2$ (E' has cardinal $2d$ and E'' has cardinal $2d(d-1)$). Also, we define naturally the sets of evaluations of the search points as follows:

$Y_{j+,k+} = \{$all evaluations of $x+\sigma e_j+\sigma e_k\}$, $Y_{j+,k-} = \{$all evaluations of $x+\sigma e_j-\sigma e_k\}$,

$Y_{j-,k+} = \{$all evaluations of $x-\sigma e_j+\sigma e_k\}$, $Y_{j-,k-} = \{$all evaluations of $x-\sigma e_j-\sigma e_k\}$.

Example 2 Noisy-Newton method verifying the $(2z-2)$-LSE assumption. Given $x \in \mathbb{R}^d$, $\sigma > 0$ and $c_0 > 0$, Sampler and Opt are defined as follows. $t(M)$ denotes the transpose of matrix M.

function Sampler(x, σ, i)

$$j \leftarrow i \ \overline{\text{mod}}\ 2d^2 \tag{13}$$

$$x_i \leftarrow \text{the } j\text{-th point in } E \tag{14}$$

return x_i
end function

function Opt($x, (x_i, y_i)_{i \in \{1,\dots,r\}}$)
for $j = 1$ to d **do**

$$\hat{y}_{j+} \leftarrow \frac{1}{|Y_{j+}|} \sum Y_{j+}, \quad \hat{y}_{j-} \leftarrow \frac{1}{|Y_{j-}|} \sum Y_{j-} \tag{15}$$

$$\hat{g}^{(j)} \leftarrow \frac{\hat{y}_{j+} - \hat{y}_{j-}}{2\sigma} \tag{16}$$

end for
for $1 \le j, k \le d$ **do**
$\hat{y}_{j+,k+} \leftarrow \frac{1}{|Y_{j+,k+}|} \sum Y_{j+,k+}, \qquad \hat{y}_{j+,k-} \leftarrow \frac{1}{|Y_{j+,k-}|} \sum Y_{j+,k-}$
$\hat{y}_{j-,k+} \leftarrow \frac{1}{|Y_{j-,k+}|} \sum Y_{j-,k+}, \qquad \hat{y}_{j-,k-} \leftarrow \frac{1}{|Y_{j-,k-}|} \sum Y_{j-,k-}$
$\hat{h}^{(j,k)} \leftarrow \frac{(\hat{y}_{j+,k+} - \hat{y}_{j-,k+}) - (\hat{y}_{j+,k-} - \hat{y}_{j-,k-})}{4\sigma^2}$
end for

$$\hat{h} \leftarrow \frac{\hat{h} + t(\hat{h})}{2}$$

if $\hat{h}$ is positive definite with least eigenvalue greater than c_0 **then**

$$x^{\text{opt}} \leftarrow x - (\hat{h})^{-1} \hat{g} \tag{17}$$

else

$$x^{\text{opt}} \leftarrow x \tag{18}$$

end if
return x^{opt}
end function

Note that in Example 2, the output of Sampler(x, σ, i) are equally distributed over E so that each of them is evaluated at least $\lfloor r/2d^2 \rfloor$ times. The pair (Sampler, Opt) defined in Example 2 verifies the LSE assumption (Property 2.2) when the noisy objective function is approximately quadratic (Eq 19) and the

noise follows the constraint given by Eq. 20.

Property 2.2. (Sampler, Opt) **in Example 2 satisfy LSE.** *Let f be the function to be optimized, $z \in \{0,1,2\}$. We assume that:*

$$\textbf{Framework 2} \left| \begin{array}{l} \mathbb{E}_\omega f(x,\omega) = \sum_{1\le j,k\le d} c_{j,k} x^{(j)} x^{(k)} \\ \qquad + \sum_{1\le j,k,l\le d} b_{j,k,l} x^{(j)} x^{(k)} x^{(l)} + o(\|x\|^3), \text{ with } c_{j,k} = c_{k,j} \qquad (19) \\ Var(f(x,\omega)) = O(\|x\|^{2z}) \text{ where } z \in \{0,1,2\}. \qquad (20) \end{array} \right.$$

Assume that there is some $c_0 > 0$ such that h is positive definite with least eigenvalue greater than $2c_0$, where h is the Hessian of $\mathbb{E}f$ at 0, i.e $h = (2c_{j,k})_{1\le j,k\le d}$.

Then there exists $\sigma_0 > 0$, $K > 0$, $C > 0$, such that for all σ that satisfies i) $\sigma < \sigma_0$ and ii) $\sigma^{6-2z} \le K/r$, and for all x such that

$$\|x\| \le C\sigma, \tag{21}$$

we have

$$\mathbb{E}\ \|x^{\text{opt}}\|^2 = O\left(\frac{\sigma^{2z-2}}{r}\right),$$

where x^{opt} is the output of $\text{Opt}(x,(x_i,y_i)_{i=1}^r)$, $(x_i)_{i=1}^r$ are the output of $\text{Sampler}(x,\sigma,i)$ and the $(y_i)_{i=1}^r$ are their respective noisy fitness values.

Proof: The event $\mathcal{E}_{\hat{h}}^{c_0}$ denotes the fact that the matrix $\hat{h}$ is positive definite with least eigenvalue greater than c_0, and $\overline{\mathcal{E}_{\hat{h}}^{c_0}}$ is the complementary event.

$$\mathbb{E}\ \|x^{\text{opt}}\|^2 = \underbrace{\mathbb{E}(\|x^{\text{opt}}\|^2|\overline{\mathcal{E}_{\hat{h}}^{c_0}})}_{A_1}\underbrace{\mathbb{P}(\overline{\mathcal{E}_{\hat{h}}^{c_0}})}_{A_2} + \underbrace{\mathbb{E}(\|x^{\text{opt}}\|^2|\mathcal{E}_{\hat{h}}^{c_0})}_{A_3}\underbrace{\mathbb{P}(\mathcal{E}_{\hat{h}}^{c_0})}_{\le 1}$$

where $A_1 \le (C\sigma)^2$ using Eqs. 18 and 21; $A_2 = O\left(\frac{\sigma^{2z-2}}{r\sigma^2}\right)$ by Lemma Appendix B.2; $A_3 = O\left(\frac{\sigma^{2z-2}}{r}\right)$ by Lemma Appendix B.3. Therefore $\mathbb{E}\ \|x^{\text{opt}}\|^2 = O\left(\frac{\sigma^{2z-2}}{r}\right)$, which is the expected result.□

Remark. Using the expressions of σ and r given by Inoa, if $(6-2z)\alpha \ge \beta$, and given $A > 0$, then there exists a constant $B_0 > 0$ such that if $B > B_0$ then the condition $\sigma_n^{6-2z} \le K/r_n$ is satisfied.

3. Convergence Rates of Inoa

Sections 3.1 and 3.2 provide, respectively, the main result and its applications, namely cumulative regret analysis and simple regret analysis for various models of noise. The special case of twice-differentiable functions is studied in Section 3.3.

3.1. Rates for various noise models

In this section, we present the main result, i.e. the convergence rates of INOA.

Theorem 3.1 (Rates for various noise models). *Consider some $A > 0$ and consider the iterative noisy optimization algorithm (*INOA*, Alg. 1, with parameters A, B, α, β). Assume that* (SAMPLER, OPT) *has a $(2z-2)$-low squared error assumption (LSE, Def. 2.1) for some f, C, U, S. Assume that $B > B_0$, where B_0 depends on α, β and A only. Let us assume that* INOA *provides (r_n, σ_n) always in S, and let us assume that*

$$1 < \beta + \alpha(2z-4), \tag{22}$$

Consider $\delta > 0$. Then there is $C > 0$, such that if $x_1^{\text{opt}} = \tilde{x}_1$ satisfies $\|x_1^{\text{opt}}\| \leq CA$, then with probability at least $1-\delta$,

$$\forall n, \quad \|x_n^{\text{opt}}\| \leq C\sigma_n \tag{23}$$

$$\forall n, \forall i \leq r_n, \quad \|x_{n,i}\| \leq (C+2)\sigma_n. \tag{24}$$

Remark. It is assumed that given x, SAMPLER provides a new search point x_i such that $\|x_i - x\| \leq 2\sigma$ (see Section 2.1). This together with Eq. 23 gives $\|x_{n,i}\| \leq \|x_{n,i} - x_n^{\text{opt}}\| + \|x_n^{\text{opt}}\| \leq (C+2)\sigma_n$. Hence Eq. 24 holds if Eq. 23 holds; we just have to show Eq. 23.

General organization of the proof of Eq. 23: Assume that Eq. 22 holds. Consider a fixed $C > 0$ and $1 > \delta > 0$. Consider hypothesis H_n: for any $1 \leq i \leq n$, $\|x_i^{\text{opt}}\| \leq C\sigma_i$ with probability at least $1 - \delta_n$, where

$$\delta_n = \sum_{i=1}^{n} c i^{-\beta-\alpha(2z-4)}. \tag{25}$$

c is chosen such that $\forall n \geq 1$, $\delta_n \leq \delta$. By Eq. 22, $\sum_{i=1}^{\infty} i^{-\beta-\alpha(2z-4)} = \Delta < \infty$, and $c = \delta/\Delta$ is suitable. We prove that for any positive integer n, H_n holds. The proof is by induction on H_n. H_1 is true since x_1^{opt} is chosen such that $\|x_1^{\text{opt}}\| \leq CA$, i.e $\|x_1^{\text{opt}}\| \leq C\sigma_1$.

Proof: Assume that H_n holds for a given integer n. We will show that H_{n+1} holds.

Step 1: concentration inequality for x_{n+1}.

By design of INOA, Alg. 1, Line 17, $x_{n+1}^{\text{opt}} = \text{OPT}(x_n^{\text{opt}}, (x_{n,i}, y_{n,i})_{i=1}^{r_n})$. When H_n is true, with probability at least $1 - \delta_n$, $\|x_n^{\text{opt}}\| \leq C\sigma_n$.

This together with the LSE imply that conditionally to an event with probability at least $1-\delta_n$,

$$\begin{aligned}
\mathbb{E}(\|x_{n+1}^{\text{opt}}\|^2) &\leq U\frac{\sigma_n^{2z-2}}{r_n} \\
&\leq U\left(\frac{A}{n^\alpha}\right)^{2z-2}\frac{1}{B\lceil n^\beta\rceil} \\
&\leq U\frac{A^{2z-2}}{B}\left(\frac{n}{n+1}\right)^{-\alpha(2z-2)-\beta}(n+1)^{-\alpha(2z-2)-\beta} \\
&\leq M(n+1)^{-\alpha(2z-2)-\beta} \qquad (26)\\
\text{where } M &= U\frac{A^{2z-2}}{B}\left(\sup_{n\geq 1}(\frac{n}{n+1})^{-\alpha(2z-2)-\beta}\right). \qquad (27)
\end{aligned}$$

Step 2: applying Markov's inequality. By Markov's inequality,

$$\mathbb{P}\left(\|x_{n+1}^{\text{opt}}\| > C\sigma_{n+1}\right) = \mathbb{P}\left(\|x_{n+1}^{\text{opt}}\|^2 > C^2\sigma_{n+1}^2\right) \leq \frac{\mathbb{E}\|x_{n+1}^{\text{opt}}\|^2}{C^2\sigma_{n+1}^2}.$$

We apply Eq. 26:

$$\begin{aligned}
\mathbb{P}\left(\|x_{n+1}^{\text{opt}}\| > C\sigma_{n+1}\right) &\leq \frac{M}{C^2A^2}(n+1)^{\alpha(2-(2z-2))-\beta} \\
&\leq c(n+1)^{-\beta-\alpha(2z-4)} = \epsilon_{n+1} \text{ if } B > B_0,
\end{aligned}$$

where $B_0 = \frac{UA^{2z-4}\left(\sup_{n\geq 1}(\frac{n}{n+1})^{-\alpha(2z-2)-\beta}\right)}{cC^2}$, using Eq. 27. Then, with probability $(1-\delta_n)(1-\epsilon_{n+1})$, $\|x_{n+1}^{\text{opt}}\| \leq C\sigma_{n+1}$. Hence with probability at least $1-\delta_n-\epsilon_{n+1} = 1-\delta_{n+1}$, $\|x_{n+1}^{\text{opt}}\| \leq C\sigma_{n+1}$. This is H_{n+1}. The induction is complete.□

3.2. Application: the general case

Theorem 3.1 ensures some explicit convergence rates for SR and CR depending on parameters α, β and z.

Corollary 3.2. *Consider the context and assumptions of Theorem 3.1, including some* (Sampler, Opt) *which has a* $(2z-2)$*-LSE (Def. 2.1) for some* f, C, U, S *such that for all* n, $(r_n, \sigma_n) \in S$, *and let us assume that* $\mathbb{E}_w f(x,\omega) - \mathbb{E}_\omega f(x^*,\omega) = O(\|x-x^*\|^2)$.

Then, the simple regret of Inoa *of has slope* $s(SR) \leq \frac{-\alpha(2z-2)-\beta}{\beta+1}$ *and the cumulative regret has slope* $s(CR) \leq \frac{\max(0,1+\beta-2\alpha)}{1+\beta}$.

Quadratic case: in the special case $z=0$ *and if* $\mathbb{E}f$ *is quadratic (i.e* $\mathbb{E}_\omega f(x,\omega) = \sum_{1\leq j,k\leq d} c_{j,k}x^{(j)}x^{(k)}$*, we get* $s(SR) \leq \frac{2\alpha-\beta}{\beta+1}$.

Proof: The number of evaluations until the end of iteration n, before recommending x_{n+1}^{opt}, is $m(n) = \sum_{i=1}^{n} r_i = O(n^{\beta+1})$.

- By assumption, $\mathbb{E}_\omega f(x,\omega) - \mathbb{E}_\omega f(x^*,\omega) = O(\|x-x^*\|^2)$. Markov's inequality applied to $\|x-x^*\|^2$ gives: $\mathbb{P}\left(\|x-x^*\|^2 > \frac{\mathbb{E}\|x-x^*\|^2}{\delta}\right) < \delta$. Hence, the simple regret SR_n after iteration n, when recommending $\tilde{x}_{m(n)} = x_{n+1}^{\text{opt}}$, is the $1-\delta$ quantile of $\|x_{n+1}^{\text{opt}} - x^*\|^2$, this is $O\left(\mathbb{E}\|x_{n+1}^{\text{opt}} - x^*\|^2\right)$. Using step 1 of Theorem 3.1, it follows that $SR_n = O\left(n^{-(2z-2)\alpha-\beta}\right)$.
- the cumulative regret CR_n until iteration n is the $1-\delta$ quantile of $\sum_{1\le i\le r_m, 1\le m\le n} \mathbb{E}_\omega f(x_{i,m},\omega) - \mathbb{E}_\omega f(x^*,\omega) = \sum_{1\le i\le r_m, 1\le m\le n} O\left(\|x_{i,m} - x^*\|^2\right)$. By Theorem 3.1, Eq. 24:

$$
\begin{aligned}
O\left(\sum_{i=1}^{n} r_i (C+2)^2/i^{2\alpha}\right) &= O\left(\sum_{i=1}^{n} i^{\beta}(C+2)^2/i^{2\alpha}\right) \\
&= \begin{cases} O\left(n^{1+\beta-2\alpha}\right) & \text{if } \beta - 2\alpha > -1, \\ O\left(\log(n)\right) & \text{if } \beta - 2\alpha = -1 \\ O(1) & \text{otherwise.} \end{cases}
\end{aligned}
$$

Dividing the log of simple regret at iteration n by the logarithm of the number of evaluations until iteration n leads to the expected result (slope) for simple regret. Dividing the log of cumulative regret until iteration n by the logarithm of the number of evaluations until iteration n leads to the expected result for cumulative regret.

3.3. Application: the smooth case

Table 1 presents optimal $s(SR)$ and $s(CR)$ in the more familiar case of smooth functions, with at least two derivatives. All results in this table can be obtained by Inoa with Opt and Sampler as in Example 2 and the provided parametrizations for α and β, except the result by Fabian (1967) assuming many derivatives.

In all cases except the quadratic case with $z = 0$, we assume $(6-2z)\alpha > \beta$, so that the LSE assumption holds for Inoa with Opt and Sampler as in Example 2 (see Property 2.2) and we assume $1 < \beta + \alpha(2z-4)$ so that Eq. 22 in Theorem 3.1 holds. Regarding the special case of $z = 0$ and quadratic function, the equation to satisfy is $1 < \beta - 4\alpha$. Please note that in this last case, the assumption $(6-2z)\alpha > \beta$ is not necessary. We then find out values of α and β such that good slopes can be obtained for CR and SR. Algorithms ensuring a slope $s(CR)$ in this table also

Table 1: $s(SR)$ and $s(CR)$ for INOA for various values of α and β, in the case of twice-differentiable functions. The references mean that our algorithm gets the same rate as in the cited paper. No reference means that the result is new.

z	optimized for CR		optimized for SR	
	$s(SR)$	$s(CR)$	$s(SR)$	$s(CR)$
0 (constant var)	$\alpha \simeq \infty, \beta \simeq 4\alpha + 1^+$		$\beta = 6\alpha, \alpha = \infty$	
	$-1/2$	$1/2$ Fabian (1967) Shamir (2013)	$-2/3$ Dupač (1957)	$2/3$
0 and ∞-differentiable.			-1 Fabian (1967)	
0 and "quadratic"			$\alpha = 0, \beta \simeq \infty$	
			-1 Dupač (1957)	
1 (linear var)	$\alpha \simeq \infty, \beta \simeq 2\alpha + 1^+$			
	-1 Rolet and Teytaud (2010)	0	-1	0
2 (quadratic var)	$\alpha \simeq \infty, \beta > 1$			
	$-\infty$	0	$-\infty$	0

ensure a slope $s(UR) = \frac{1}{2}(s(CR) - 1)$. It follows that the optimal parametrization for UR is the same as the optimal parametrization for CR.

We consider parameters optimizing the CR (left) or SR (right) - and both simultaneously when possible.These results are for B constant but large enough. Infinite values mean that the value can be made arbitrarily negatively large by choosing a suitable parametrization. X^+ denotes a value which should be made arbitrarily close to X by superior values, in order to approximate the claimed rate.

Results are not adaptive; we need a different parametrization when $z = 0$, $z = 1$, $z = 2$. Also, for $z = 0$, we need a different parametrization depending on whether we are interested in CR or SR.

4. Conclusion and further work

We have shown that estimating the Hessian and gradient can lead to fast convergence results. In fact, with one unique algorithm we obtain many of the rates presented by

- Spall (2009); Shamir (2013) in the case of a constant variance noise for simple regret and cumulative regret respectively.
- Rolet and Teytaud (2010); Coulom et al. (2011) ($z = 1$) and Jebalia and Auger (2008) ($z = 2$) for a larger space of functions than in these papers, where sphere functions are considered.

In summary, we observe on the Table 1 that results obtained here recover most previous results discussed in the introduction. And also the results presented here cover all the analyzed criteria: simple regret, cumulative regret, uniform rates.

Compared to Spall (2009), our algorithm uses more evaluations per iteration. This has advantages and drawbacks. The positive part is that it is therefore more parallel. For example, for $z = 0$, and an algorithm optimized for SR, we get $s(SR) = -2/3$; this rate is the same as the one in Spall (2009) in terms of number of evaluations, i.e. the number of evaluations is proportional to $(1/sr)^{2/3}$ for a simple regret sr, but our evaluations are grouped into a small number of iterations. On the other hand, it is far less convenient in a sequential setting as the optimization process starts only after an iteration is complete, which takes a significant time in our case. Our algorithm is proved for $z = 1$, $z = 2$; these cases are not discussed in Shamir (2013); Fabian (1967); Spall (2009).

Our algorithm is not limited to functions with quadratic approximations; quadratic approximations are a natural framework, but the success of various surrogate models in the recent years suggests that other approximation frameworks could be used. Our theorems are not specific for quadratic approximations and only require that the LSE approximation holds. The LSE assumption is natural in terms of scaling with respect to r - the $1/\sqrt{r}$ typical deviation is usual in e.g. maximum likelihood estimates, and therefore the method should be widely applicable for general surrogate models.

More generally, our results show a fast rate as soon as the estimator of the location of the optimum has squared error $O(\sigma^{2z-2}/r)$, when using r points sampled adequately within distance $O(\sigma)$ of the optimum.

Further work. In the theoretical side, further work includes writing detailed constants, in particular depending on the eigenvalues of the Hessian of the expected objective function at the optimum and the dimension of the search space. In the case of infinite slope (see Table 1, $z = 2$), we conjecture that the convergence is log-linear, i.e. the logarithm of the simple regret decreases as a function of the number of evaluations. In the other hand, future study consists of extensive experiments - but we refer to Cauwet et al. (2014) for significant artificial experiments and Liu and Teytaud (2014) for the application which motivated this work.

Part of the agenda is to extend the algorithm by providing other examples of estimators to be used for approximating the location of the optimum (other

than Examples 1 and 2, but verifying the LSE assumption); in particular, classical surrogate models, and applications to piecewise linear strongly convex functions as in Rolet and Teytaud (2010). A way to improve the algorithm is to use quasi-Newton estimates of the Hessian, from the successive gradients, rather than using directly finite differences. Last, making algorithms more adaptive by replacing the constants by adaptive parameters depending on noise estimates is under consideration.

Aubin, J.-P., 1991. Viability Theory. Birkhauser Boston Inc., Cambridge, MA, USA.

Bengio, Y., 1997. Using a financial training criterion rather than a prediction criterion. International Journal of Neural Systems 8 (4), 433–443, special issue on noisy time-series.
URL `http://www.iro.umontreal.ca/~lisa/pointeurs/profitcost.pdf`

Bottou, L., Bousquet, O., 2011. The tradeoffs of large scale learning. In: Sra, S., Nowozin, S., Wright, S. J. (Eds.), Optimization for Machine Learning. MIT Press, pp. 351–368.
URL `http://leon.bottou.org/papers/bottou-bousquet-2011`

Broyden, C. G., 1970. The convergence of a class of double-rank minimization algorithms 2. The New Algorithm. J. of the Inst. for Math. and Applications 6, 222–231.

Cauwet, M., Liu, J., Teytaud, O., 2014. Algorithm portfolios for noisy optimization: Compare solvers early. In: Learning and Intelligent Optimization - 8th International Conference, Lion 8, Gainesville, FL, USA, February 16-21, 2014. Revised Selected Papers. pp. 1–15.
URL `http://dx.doi.org/10.1007/978-3-319-09584-4_1`

Chapel, L., Deffuant, G., 2006. SVM viability controller active learning. In: Kernel machines and Reinforcement Learning Workshop - ICML 2006. United States.
URL `https://hal.archives-ouvertes.fr/hal-00616861`

Chaslot, G., Winands, M., I.Szita, van den Herik, H., 2008. Parameter tuning by cross entropy method. In: European Workshop on Reinforcement Learning.
URL `http://www.cs.unimaas.nl/g.chaslot/papers/ewrl.pdf`

Chen, H., Sep. 1988. Lower rate of convergence for locating the maximum of a function. Annals of statistics 16, 1330–1334.

Chen, H. F., Duncan, T. E., Pasik-Duncan, B., 1996. A stochastic approximation algorithm with random differences. In: Proceedings of the 13th IFAC World Congress. Vol. H. pp. 493–496.

Coulom, R., 2012. Clop: Confident local optimization for noisy black-box parameter tuning. In: Advances in Computer Games. Springer Berlin Heidelberg, pp. 146–157.

Coulom, R., Rolet, P., Sokolovska, N., Teytaud, O., 2011. Handling expensive optimization with large noise. In: Foundations of Genetic Algorithms, 11th International Workshop, FOGA 2011, Schwarzenberg, Austria, January 5-8, 2011, Proceedings. pp. 61–68.
URL `http://doi.acm.org/10.1145/1967654.1967660`

Decock, J., Teytaud, O., 2013. Noisy optimization complexity under locality assumption. In: Proceedings of the twelfth workshop on Foundations of genetic algorithms XII. FOGA XII '13. ACM, New York, NY, USA, pp. 183–190.
URL `http://doi.acm.org/10.1145/2460239.2460256`

Dupač, V., 1957. O Kiefer-Wolfowitzově aproximační Methodě. Časopis pro pěstování matematiky 082 (1), 47–75.
URL `http://eudml.org/doc/20601`

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

Fabian, V., 1971. Stochastic Approximation. SLP. Department of Statistics and Probability, Michigan State University.
URL `http://books.google.fr/books?id=a0aiMQEACAAJ`

Fletcher, R., 1970. A new approach to variable-metric algorithms. Computer Journal 13, 317–322.

Goldfarb, D., 1970. A family of variable-metric algorithms derived by variational means. Mathematics of Computation 24, 23–26.

Heidrich-Meisner, V., Igel, C., 2009. Hoeffding and bernstein races for selecting policies in evolutionary direct policy search. In: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, New York, NY, USA, pp. 401–408.

Jebalia, M., Auger, A., 2008. On multiplicative noise models for stochastic search. In: et a.l., G. R. (Ed.), Conference on Parallel Problem Solving from Nature

(PPSN X). Vol. 5199. Springer Verlag, Berlin, Heidelberg, pp. 52–61.
URL http://hal.inria.fr/docs/00/28/77/25/PDF/MohamedAnnePPSN08.ForHal.pdf

Jebalia, M., Auger, A., Hansen, N., 2010. Log-linear convergence and divergence of the scale-invariant (1+1)-ES in noisy environments. Algorithmica, 1–36Online first.
URL http://dx.doi.org/10.1007/s00453-010-9403-3

Kiefer, J., Wolfowitz, J., 1952. Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics 23 3, 462–466.
URL http://projecteuclid.org/euclid.aoms/1177729392.

Liu, J., Teytaud, O., 2014. Meta online learning: Experiments on a unit commitment problem. In: 22th European Symposium on Artificial Neural Networks, ESANN 2014, Bruges, Belgium, April 23-25, 2014.
URL http://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2014-139.pdf

Powell, M. J. D., 2004. The newuoa software for unconstrained optimization with derivatives. Tech. Rep. NA2004/08, Dept. of Applied Math. and Theoretical Physics.

Powell, M. J. D., February 2008. Developments of newuoa for minimization without derivatives. IMA J Numer Anal, drm047+.
URL http://dx.doi.org/10.1093/imanum/drm047

Robbins, H., Monro, S., 1951. A stochastic approximation method. The Annals of Mathematical Statistics 23 22, 400–407.
URL http://projecteuclid.org/euclid.aoms/1177729586

Rolet, P., Teytaud, O., 2009. Bandit-based estimation of distribution algorithms for noisy optimization: Rigorous runtime analysis. In: Proceedings of Lion4 (accepted); presented in TRSH 2009 in Birmingham. pp. 97–110.

Rolet, P., Teytaud, O., 2010. Adaptive noisy optimization. In: Di Chio, C., Cagnoni, S., Cotta, C., Ebner, M., Ekrt, A., Esparcia-Alcazar, A., Goh, C.-K., Merelo, J., Neri, F., PreuY, M., Togelius, J., Yannakakis, G. (Eds.), Applications of Evolutionary Computation. Vol. 6024 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 592–601.
URL http://dx.doi.org/10.1007/978-3-642-12239-2_61

Ros, R., 2009. Benchmarking the NEWUOA on the BBOB-2009 Noisy Testbed. In: GECCO. Montréal, Canada.
URL http://hal.inria.fr/inria-00377083

Sakrison, D., 1964. A continuous Kiefer-Wolfowitz procedure for random process. Ann. Math. Statist. 35, 590–599.

Sehnke, F., Osendorfer, C., Rückstiess, T., Graves, A., Peters, J., Schmidhuber, J., 2010. Parameter-exploring policy gradients. Neural Networks (23(4)), 551–559, intelligent Autonomous Systems.
URL http://tubiblio.ulb.tu-darmstadt.de/55396/

Shamir, O., 2013. 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. pp. 3–24.
URL http://jmlr.org/proceedings/papers/v30/Shamir13.html

Shanno, D. F., 1970. Conditioning of quasi-newton methods for function minimization. Mathematics of Computation 24, 647–656.

Spall, J., Oct 2000. Adaptive stochastic approximation by the simultaneous perturbation method. Automatic Control, IEEE Transactions on 45 (10), 1839–1853.

Spall, J., June 2009. Feedback and weighting mechanisms for improving jacobian estimates in the adaptive simultaneous perturbation algorithm. Automatic Control, IEEE Transactions on 54 (6), 1216–1229.

Spall, J. C., 2003. Introduction to stochastic search and optimization: Estimation, simulation, and control. John Wiley and Sons.

Stoltz, G., Bubeck, S., Munos, R., 2011. Pure exploration in finitely-armed and continuous-armed bandits. Theoretical Computer Science 412 (19), 1832–1852.

Vapnik, V. N., 1995. The Nature of Statistical Learning. Springer Verlag.

Zhan, X., 2005. Extremal eigenvalues of real symmetric matrices with entries in an interval. SIAM J. Matrix Analysis Applications 27 (3), 851–860.

Appendix A. Summary of Notations

Appendix A.1. General Notations

f	:=	objective function
x^*	:=	optimum of the objective function
d	:=	dimension of the search domain
$\mathcal{D} \subset \mathbb{R}^d$	:=	search domain of the objective function
x'_n	:=	n^{th} evaluated search point in a run
x_n	:=	search point used by the algorithm at iteration n
y_n	:=	fitness value of x_n, possibly noisy
$\tilde{x}_n$	:=	recommendation of the optimum after n evaluations
ω	:=	random variable
$\|\cdot\|$	:=	Euclidean norm
$\lvert\cdot\rvert$	:=	Absolute value when applied to a real number Cardinality when applied to a set
$\langle\cdot,\cdot\rangle$	:=	Inner product
$\mathbb{P}$, (resp. $\mathbb{P}(\cdot\|\cdot)$)	:=	Probability (resp. conditional probability)
$\mathbb{E}$, (resp. $\mathbb{E}(\cdot\|\cdot)$)	:=	Expectation (resp. conditional expectation)
$\mathbb{E}_\omega$	:=	Expectation on the random variable ω
Var	:=	Variance of a random variable
$Q_{1-\delta}$	:=	Quantile $1-\delta$ of a random variable
UR_i	:=	$Q_{1-\delta}\|x'_i - x^*\|$ Uniform Regret
SR_i	:=	$Q_{1-\delta}\left(\mathbb{E}_\omega f(\tilde{x}_i,\omega) - \mathbb{E}_\omega f(x^*,\omega)\right)$ Simple Regret
CR_i	:=	$Q_{1-\delta}\sum_{j\leq i}\left(\mathbb{E}_\omega f(x'_j,\omega) - \mathbb{E}_\omega f(x^*,\omega)\right)$ Cumulative Regret
$s(*R)$	:=	$\limsup_i \frac{\log(*R_i)}{\log(i)}$, $*R$ stand for SR, CR or UR, slope of the corresponding regret
z	:=	strength of the noise
$O(\cdot),\ o(\cdot),\ \Omega(\cdot)$	:=	Landau notations

Appendix A.2. In the algorithms and examples

$x_{i,n}$	:=	i^{th} evaluated search point at iteration n
$(e_j)_{j=1}^d$	:=	canonical orthonormal basis
$\lceil\cdot\rceil$ (resp. $\lfloor\cdot\rfloor$)	:=	ceiling (resp. floor) function
SAMPLER	:=	Routine for sampling the search space
OPT	:=	Routine for computing the next approxim. of x^*
σ (resp. σ_n)	:=	step-size (resp. step-size at iteration n)
r (resp. r_n)	:=	number of evaluations
		(resp. number of evaluations at iteration n)
$\alpha,\ A$	:=	parameters of the step-size
$\beta,\ B$	:=	parameters of the number of evaluations
$x^{(i)}$	:=	i^{th} coordinate of vector x
$\hat{g}$	:=	gradient approximated by finite differences
$\hat{h}$	:=	Hessian approximated by finite differences
x^{opt}	:=	output of OPT function
$\mathcal{E}_M^c$	:=	event : "the matrix M is positive definite
		with least eigenvalue greater than c"
$\lambda_1(M) \geq \cdots \geq \lambda_d(M)$	:=	d eigenvalues of matrix M
M^{-1}	:=	inverse of matrix M, when M is invertible
$t(M)$	:=	transpose of matrix M
$\overline{A}$	:=	complement of event A
$k \bmod l$	:=	remainder of the Euclidian division of k by l
$k \overline{\bmod} l$	:=	$1 + (k-1) \bmod l$

Appendix B. Proofs of Sections 2.2 and 3.2

The following Lemmas are used to prove property 2.2. In this Section, consider functions SAMPLER and OPT as is Example 2, and an objective function f described by Eqs. 19 and 20. h denotes the Hessian of $\mathbb{E}f$ at 0, i.e $h = (2c_{j,k})_{1\leq j,k\leq d}$. All results hold for C sufficiently small, and x verifying Eq. 21.

Lemma Appendix B.1 (Approximation lemma). *With the definitions of Example 2, for any $j \in \{1, \dots, d\}$, the approximate gradient verifies:*

$$\hat{g}^{(j)} = \begin{cases} 2 \sum_{1 \leq k \leq d} c_{j,k} x^{(k)} + \frac{\mathcal{N}_1}{\sqrt{r}\sigma} & \text{if } \mathbb{E}_\omega f(x,\omega) = \sum_{1 \leq j,k \leq d} c_{j,k} x^{(j)} x^{(k)} \\ 2 \sum_{1 \leq k \leq d} c_{j,k} x^{(k)} + O(\sigma^2) + \frac{\mathcal{N}_1}{\sqrt{r}\sigma} & \text{otherwise} \end{cases} \tag{B.1}$$

where $\mathcal{N}_1$ is an independent noise, with $\mathbb{E}(\mathcal{N}_1) = 0$ and $Var(\mathcal{N}_1) = O(\sigma^{2z})$, $z \in \{0, 1, 2\}$.

For any $(j,k) \in \{1, \dots, d\}^2$, the approximate Hessian verifies:

$$\hat{h}_{j,k} = \begin{cases} 2c_{j,k} + \frac{\mathcal{N}_2}{\sqrt{r}\sigma^2} & \text{if } \mathbb{E}_\omega f(x,\omega) = \sum_{1 \leq j,k \leq d} c_{j,k} x^{(j)} x^{(k)} \\ 2c_{j,k} + O(\sigma) + \frac{\mathcal{N}_2}{\sqrt{r}\sigma^2} & \text{otherwise} \end{cases} \tag{B.2}$$

where $\mathcal{N}_2$ is an independent noise, independent of $\mathcal{N}_1$, with $\mathbb{E}(\mathcal{N}_2) = 0$ and $Var(\mathcal{N}_2) = O(\sigma^{2z})$, $z \in \{0, 1, 2\}$.

Proof: We provide a proof for the natural gradient $\hat{g}^{(j)}$. The proof is similar for natural Hessian. There are $\lfloor r/(2d^2) \rfloor$ revaluations per point. $\langle ., . \rangle$ is the inner product. By definition of the noisy objective function f in Property 2.2,

$$f(x,\omega) = \mathbb{E}_\omega f(x,\omega) + \mathcal{N},$$

with $\mathbb{E}_\omega f(x,\omega)$ as in Eq. 19 and $\mathcal{N}$ a random variable s.t. $\mathbb{E}\mathcal{N} = 0$ and $Var(\mathcal{N}) = O(\|x\|^{2z})$, $z \in \{0, 1, 2\}$. In the same way,

$$\hat{g}^{(j)} = \mathbb{E}\hat{g}^{(j)} + \mathcal{N}'$$

where $\mathcal{N}'$ is a random variable s.t. $\mathbb{E}\mathcal{N}' = 0$ and variance to calculate. By definition of the natural gradient in Ex. 2,

$$\begin{aligned} \mathbb{E}\hat{g}^{(j)} &= \mathbb{E}\frac{\hat{y}_{j+} - \hat{y}_{j-}}{2\sigma} \\ &= \frac{1}{2\sigma} \left(\mathbb{E}_\omega f(x + \sigma e_j, \omega) - \mathbb{E}_\omega f(x - \sigma e_j, \omega) \right), \\ \mathbb{E}_\omega f(x + \sigma e_j, \omega) &= \sum_{1 \leq i,k \leq d} c_{i,k} x^{(i)} x^{(k)} + 2\sigma \sum_{1 \leq k \leq d} c_{j,k} x^{(k)} + c_{j,j}\sigma^2 \\ &+ \sum_{1 \leq i,k,l \leq d} b_{i,k,l} x^{(i)} x^{(k)} x^{(l)} + \sigma \sum_{1 \leq k,l \leq d} (b_{j,k,l} + b_{k,j,l} + b_{k,l,j}) x^{(k)} x^{(l)} \\ &+ \sigma^2 \sum_{1 \leq k \leq d} (b_{j,j,k} + b_{k,j,j} + b_{j,k,j}) x^{(k)} + \sigma^3 b_{j,j,j} + o(\|x + \sigma e_j\|^3) \end{aligned}$$

$$\mathbb{E}_\omega f(x+\sigma e_j,\omega) - \mathbb{E}_\omega f(x-\sigma e_j,\omega) = 4\sigma \sum_{1\le k\le d} c_{j,k}x^{(k)} + 2\sigma \underbrace{\sum_{1\le k,l\le d} (b_{j,k,l}+b_{k,j,l}+b_{k,l,j})x^{(k)}x^{(l)}}_{=O(\|x\|^2)} + 2\sigma^3 b_{j,j,j} + o(\|x+\sigma e_j\|^3) - o(\|x-\sigma e_j\|^3).$$

By Eq. 21, $O(\|x\|^2) = O(\sigma^2)$ and $o(\|x \pm \sigma e_j\|^3) = O(\sigma^3)$. Then,

$$\mathbb{E}\hat{g}^{(j)} = \begin{cases} 2\sum_{1\le k\le d} c_{j,k}x^{(k)} & \text{if } \mathbb{E}f(x,\omega) = \sum_{1\le j,k\le d} c_{j,k}x^{(j)}x^{(k)}, \\ 2\sum_{1\le k\le d} c_{j,k}x^{(k)} + O(\sigma^2) & \text{otherwise.} \end{cases}$$

Now compute $Var(\mathcal{N}') = Var(\hat{g}^{(j)})$.

$$\begin{aligned} Var(\hat{g}^{(j)}) &= Var\left(\frac{\hat{y}_{j+} - \hat{y}_{j-}}{2\sigma}\right) \\ &= \frac{Var\left(f(x+\sigma e_j,\omega) - f(x-\sigma e_j,\omega)\right)}{4\sigma^2\lfloor r/(2d^2)\rfloor} \\ &= \frac{Var\left(f(x+\sigma e_j,\omega)\right) + Var\left(f(x-\sigma e_j,\omega)\right)}{4\sigma^2\lfloor r/(2d^2)\rfloor} \end{aligned}$$

$Var\left(f(x+\sigma e_j,\omega)\right) \le R\|x+\sigma e_j\|^{2z},\ z\in\{0,1,2\}$ for a given $R>0$ by Eq. 20.

$$Var\left(f(x+\sigma e_j,\omega)\right) \le \begin{cases} R & \text{if } z=0, \\ R(\|x\|^2 + 2\langle x,\sigma e_j\rangle + \sigma^2) & \text{if } z=1, \\ R(\|x\|^4 + 4(\sigma^2+\|x\|^2)\langle x,\sigma e_j\rangle + \\ \quad 4\langle x,\sigma e_j\rangle^2 + 2\sigma^2\|x\|^2 + \sigma^4) & \text{if } z=2. \end{cases}$$

Hence using Eq. 21 and Cauchy-Schwarz inequality: $|\langle x,\sigma e_j\rangle| \le \|x\|\sigma$, $Var\left(f(x+\sigma e_j,\omega)\right) = O(\sigma^{2z})$ and $Var\left(f(x-\sigma e_j,\omega)\right) = O(\sigma^{2z})$. Then $Var(\mathcal{N}') = O(\frac{\sigma^{2z}}{r\sigma^2})$ and the expected result is obtained by putting $\mathcal{N}' = \frac{\mathcal{N}_1}{\sqrt{r}\sigma}$ with $Var(\mathcal{N}_1) = O(\sigma^{2z})$.$\square$

Lemma Appendix B.2. *There is $\sigma_0 > 0$ such that for all $\sigma < \sigma_0$, $\hat{h}$ is positive definite with its least eigenvalue greater than c_0 with probability at least $1 - O\left(\frac{\sigma^{2z-4}}{r}\right)$, where c_0 is as in the assumptions in Property 2.2.*

Proof: By Lemma Appendix B.1 (Eq. B.2), for any $(j,k) \in \{1,\dots,d\}^2$, $\mathbb{E}\ \hat{h}_{j,k} = h_{j,k} + O(\sigma)$, where $\hat{h}$ is the approximation of the Hessian h, defined in Example 2 and σ is the step-size. So there exists a function $\sigma \mapsto f_h(\sigma)$ and a constant $R > 0$ such that $\mathbb{E}\ \hat{h}_{j,k} = h_{j,k} - f_h(\sigma)$ with $f_h(\sigma) \leq R|\sigma|$. Let c_0 be as in the assumptions in Property 2.2 and d the dimension. Then,

$$\begin{aligned}
\mathbb{P}(|\hat{h}_{j,k} - h_{j,k}| \geq c_0/d) &= \mathbb{P}(|\hat{h}_{j,k} - h_{j,k} + f_h(\sigma) - f_h(\sigma)| \geq c_0/d) \\
&\leq \mathbb{P}(|\hat{h}_{j,k} - h_{j,k} + f_h(\sigma)| + f_h(\sigma) \geq c_0/d) \\
&= \mathbb{P}(|\hat{h}_{j,k} - \mathbb{E}\hat{h}_{j,k}| \geq c_0/d - f_h(\sigma)) \\
&\leq \frac{Var(\hat{h}_{j,k})}{(c_0/d - f_h(\sigma))^2} \text{ by applying Chebyshev's inequality}
\end{aligned}$$

Also by Lemma Appendix B.1 (eq. B.2), $Var(\hat{h}_{j,k}) = Var(\frac{\mathcal{N}_2}{\sqrt{r}\sigma^2}) = O(\frac{\sigma^{2z}}{r\sigma^4})$, where r is the number of revaluations. Furthermore, $(c_0/d - f_h(\sigma))^2 \geq (c_0/d - R|\sigma|)^2 > 0$ for all $\sigma < \frac{c_0}{dR} := \sigma_0$. So, $\forall (j,k) \in \{1,\dots,d\}^2$ we know that $\mathbb{P}(|\hat{h}_{j,k} - h_{j,k}| \geq c_0/d) = O(\frac{\sigma^{2z-4}}{r})$. Or equivalently

$$\mathbb{P}(|\hat{h}_{j,k} - h_{j,k}| \leq c_0/d) = 1 - O\left(\frac{\sigma^{2z-4}}{r}\right)$$

Since $\hat{h} - h$ is a symmetric matrix, then we deduce from Theorem 1 case (ii) in Zhan (2005) that $\lambda_d(\hat{h} - h) \geq -c_0$ with probability $1 - O\left(\frac{\sigma^{2z-4}}{r}\right)$, where we denote for any $d \times d$ matrix M its eigenvalues in decreasing order by $\lambda_1(M) \geq \cdots \geq \lambda_d(M)$. Using this and the fact that we assumed in Property 2.2 $\lambda_d(h) \geq 2c_0$, we have $\forall x \in \mathbb{R}^d$, $x \neq 0$,

$$\langle \hat{h}x, x\rangle = \langle (\hat{h} - h)x, x\rangle + \langle hx, x\rangle \geq -c_0\|x\|^2 + 2c_0\|x\|^2$$

$$\Leftrightarrow \frac{\langle \hat{h}x, x\rangle}{\langle x, x\rangle} \geq c_0$$

Since $\hat{h}$ is a Hermitian matrix, by the $\min - \max$ Theorem we know that $\lambda_d(\hat{h}) = \min\left\{\frac{\langle \hat{h}x,x\rangle}{\langle x,x\rangle}, x \neq 0\right\}$. Hence for all $\sigma > \sigma_0$, we have that $\lambda_d(\hat{h}) \geq c_0$ with probability $1 - O\left(\frac{\sigma^{2z-4}}{r}\right)$.□

Lemma Appendix B.3. (Good approximation of the optimum with the second order method) *Consider the context of Property 2.2. Then there exists a constant $K > 0$ such that for any pair of step size and number of revaluation (σ, r) that satisfies $\sigma^{6-2z} \leq K/r$, we have $\mathbb{E}(\|x^{\text{opt}}\|^2 | \mathcal{E}_{\hat{h}}^{c_0}) = O\left(\frac{\sigma^{2z-2}}{r}\right)$.*

Proof:

$$\begin{aligned}\mathbb{E}(\|x^{\text{opt}}\|^2|\mathcal{E}_{\hat{h}}^{c_0}) &= \mathbb{E}(\|\hat{h}^{-1}(\hat{h}x-\hat{g})\|^2|\mathcal{E}_{\hat{h}}^{c_0}) \text{ by definition of } x^{\text{opt}}, \\ &\leq (1/c_0)^2\mathbb{E}(\|\hat{h}x-\hat{g}\|^2|\mathcal{E}_{\hat{h}}^{c_0})\end{aligned}$$

using that $\forall x \in \mathbb{R}^d,\ x \neq 0,\ \|Mx\|^2 \leq (\lambda_1(M))^2\,\|x\|^2$ and $\lambda_1(M^{-1}) = \frac{1}{\lambda_d(M)}$, where M is a real symmetric matrix. Under $\mathcal{E}_{\hat{h}}^{c_0}$, using Eqs. B.1 and B.2,

$$\begin{aligned}\mathbb{E}(\|\hat{h}x-\hat{g}\|^2) &= \mathbb{E}\left\{\sum_{1\leq j\leq d}\left(\sum_{1\leq k\leq d}\hat{h}_{j,k}x^{(k)}-g^{(j)}\right)^2\right\}, \\ &= \mathbb{E}\left\{\sum_{1\leq j\leq d}\left(\frac{\mathcal{N}_2}{\sqrt{r}\sigma^2}\underbrace{\sum_{1\leq k\leq d}x^{(k)}}_{=O(\|x\|)}+O(\sigma)\underbrace{\sum_{1\leq k\leq d}x^{(k)}}_{=O(\|x\|)}+O(\sigma^2)+\frac{\mathcal{N}_1}{\sqrt{r}\sigma}\right)^2\right\} \\ &= d\mathbb{E}\left\{\left(O(\sigma^2)+\frac{\mathcal{N}_1}{\sqrt{r}\sigma}+\frac{\mathcal{N}_2}{\sqrt{r}\sigma}\right)^2\right\} \text{ using } \|x\|\leq C\sigma, \\ &= O(\sigma^4)+O\left(\frac{\sigma^{2z}}{r\sigma^2}\right) \text{ using } \mathbb{E}(\mathcal{N}_1)=\mathbb{E}(\mathcal{N}_2)=0, \\ &\quad Var(\mathcal{N}_1)=Var(\mathcal{N}_2)=O(\sigma^{2z}) \text{ and independence}, \\ &= O\left(\frac{\sigma^{(2z-2)}}{r}\right) \text{ if } \sigma^{4-(2z-2)}\leq K/r,\end{aligned}$$

which is the expected result.□

Remark. In Lemma Appendix B.3, if $\mathbb{E}f$ is simply quadratic, i.e $\forall(j,k,l) \in \{1,\ldots,d\}^3$, $b_{j,k,l}=0$, the assumption $\sigma^{6-2z}=O(1/r)$ is unnecessary.