A mathematically derived number of resamplings for noisy optimization

Jialin Liu, David L. Saint-Pierre, Olivier Teytaud

A mathematically derived number of resamplings for noisy optimization

Jialin Liu, David L. Saint-Pierre, Olivier Teytaud
TAO, Inria, Lri, UMR 8623 (CNRS - Univ. Paris-Sud), Univ. Paris-Sud, France

Categories and Subject Descriptors

G.1.6 [**Optimization**]: Unconstrained optimization

General Terms

Theory

Keywords

Noisy Optimization, Evolution Strategies

1. INTRODUCTION

This section presents noisy optimization, the motivations for studying resampling and the outline of the paper. Throughout the paper, log represents the natural logarithm and $\mathcal{N}$ denotes a standard Gaussian random variable.

Noisy optimization. Typical optimization problems consist in searching for the minimum x^* of some objective function $fitness : \mathbb{R}^d \to \mathbb{R}$, i.e. x^* is such that $\forall x,\ f(x^*) \leq f(x)$. $fitness$ is known as the fitness function and d represents the dimension of the problem. A black-box optimization consists in doing so by successive calls to $fitness$ only. $fitness$ is seen as an oracle; no internal property of $fitness$ is used, and the goal of the optimization algorithm consists in finding a good approximation of x^*, within a moderate number of calls to the objective function.

In the case of noisy optimization, the objective function is corrupted by a random process ω. By including noise ω in the objective function $fitness(x, \omega)$, we are interested in finding argmin $\mathbb{E}fitness(x, \omega)$. From here on out, to simplify the notation $fitness(x)$ denotes $fitness(x, \omega)$.

We use Simple Regret (SR) as a measure of performance. SR is the difference between the expectation of the fitness of a given x and the expectation of the fitness of x^*: $SR = \mathbb{E}fitness(x_m) - \mathbb{E}fitness(x^*)$, where x_m is the approximation of the optimum obtained by the optimization algorithm after m evaluations and x^* is the optimum. This index m takes into account multiple evaluations, as detailed later. Needless to say that the objective is to make the simple regret decrease to zero as fast as possible.

This decrease rate can be represented by $slope(SR)$ and is defined as follow:

$$slope(SR) = \limsup_m \frac{\log\left(\mathbb{E}fitness(x_m) - \mathbb{E}fitness(x^*)\right)}{\log m} \quad (1)$$

where x_m denotes the m^{th} search point at the m^{th} function evaluation. This number m takes into account multiple resamplings of a given search point.

[6] considers a noise model $+\sigma_\epsilon \mathcal{N}$, where σ_ϵ is the noise strength. [6] considers (i) constant noise variance, i.e. σ_ϵ does not depend on the current location and (ii) normalized noise variance, i.e. σ_ϵ decreases linearly with decreasing square distance to the optimum. [8] focuses on the latter, termed multiplicative noise. This assumption is convenient for mathematical analysis.

As a first step, we will focus on local noisy optimization. Local noisy optimization is the optimization of objective functions in which the main problem is noise, and not local minima. We will also restrict our attention to cases in which there is no conditioning issue. We will therefore study the simple sphere function, derive the structure of a resampling rule, and compare experimentally various parameters of this resampling rule. The goal of this paper is to provide a conclusive answer for non-adaptive resampling rules in this simple setting.

Resamplings In the noise-free case, Evolution Strategies can lead to *log-linear* convergence (i.e. the logarithm of the distance to the optimum typically scales linearly with the number of evaluations)[5],

$$\frac{\log \sigma_n}{n} \sim A < 0 \text{ and } \frac{\log ||x_n||}{n} \sim A' < 0. \quad (2)$$

where x_n is the recommendation and σ_n is the step-size at generation n. In the noisy case, noisy fitness values lead to a *log-log* convergence (i.e. the logarithm of the distance to the optimum typically scales linearly with the logarithm of number of evaluations)[9, 3].

In the noisy case, resamplings can be used in order to reduce the effect of noise[1, 2].Due to independence over multiple evaluations, the standard deviation of the noise is divided by $\sqrt{r}$ when working with r resamplings on a same search point. We refer to the number of function evaluations allowed for an optimization run as a “budget”. A large number of resamplings per individual may introduce a dissipation of budget (evaluations), hence the choice of the resampling number is important. The number of resamplings can be chosen by adaptive rules, such as estimating the noise

GECCO'14, July 12-16, 2014, Vancouver, BC, Canada.

level,possibly using Bernstein races[7], using the step-size[3, 4], or in a non-adaptive manner[3, 4]. [3] proved mathematically that (i) a non-adaptive rule with exponential number of resamplings and (ii) an adaptive number of resamplings depending on the step-size can lead to *log-log* convergence. [3] has also shown experimentally that (iii) a non-adaptive rule with polynomial number of resamplings can lead to the *log-log* convergence, i.e. $\frac{\log \|x_m\|^2}{\log m} \sim A'' < 0$, where x_m is the recommendation after m evaluations. **Comparing resampling rules.** We compare 8 resampling rules, that can be separated into 3 families. The first family is polynomial. It includes a linear, quadratic and cubic resampling rules. The second family is exponential. It contains 2 simple exponential resampling rules. The third family consists in 3 new resampling rules, which vary with both the current generation number n and the dimension d. All studied functions are given in Table 1.

Table 1: Resampling rules. d: dimension of search domain; n: current generation number.

Notation	r_n	Notation	r_n
linear	n	$exp/10$	$\lceil exp(\frac{n}{10}) \rceil$
quadratic	n^2	$math1$	$\lceil exp(\frac{4n}{5d})/d^2 \rceil$
cubic	n^3	$math2$	$\lceil exp(\frac{n}{10})/d^2 \rceil$
$2exp$	2^n	$math3$	$\lceil exp(\frac{n}{\sqrt{d}})/d^2 \rceil$

2. CONCLUSION

In all cases, the $math2$ formula, $r_n = \lceil exp(\frac{n}{10})/d^2 \rceil$ evaluations for each individual at the n^{th} generation, performs nearly optimally among our non-adaptive rules. The scaling with d is seemingly correct for saving up function evaluations: $2exp$ performs badly in large dimension and polynomial functions perform badly in small dimension. It seems that $math2$ has the best of both worlds, in the considered setting (noise standard deviation of the same order as fitness values). We do not claim that this conclusion holds in more general cases, compared to adaptive rules or different noise models; we just propose a conclusive answer for the simple case under work. We below present the main limitations and some further work.

Limitations. This work is restricted to non-adaptive rules. Such rules have robustness advantages: (i) we do not need bounds on fitness values (whereas Bernstein methods do), (ii) we have no problem with equal expected fitness values (whereas Bernstein rules can fall in infinite loops when expected values are equal), or rules based on empirical standard deviations have such problems[7]), (iii) no problem with step-size stagnation as in resampling rules based on the step-size[3]. But the results (both theoretical and experimental) are not relevant for easy cases, in which the noise standard deviation is very small.

Further work. Adaptive rules have their weaknesses, as they are sensitive to parameters and special cases. However, they can save up fitness evaluations. A natural further work is to use a combination of non-adaptive and adaptive rules:(i) Adaptive condition: If a rigorous statistical test concludes that there is no point in keeping resampling, we can stop. (ii) Non-Adaptive limit: never apply more than the non-adaptive rule, which is conservative. Such a combination is visible in [7]. The non-adaptive part might benefit from the scaling proposed in our rule $math2$.

Another possible further work is a different point of view, between adaptive methods (using Bernstein races or resampling numbers depending on step-sizes) and non-adaptive methods (as those studied in this paper). Results here suggest that our exponential formulas are asymptotically good. However, both the mathematical derivation and the experiments are based on the fact that the standard deviation of the noise is of the right order. The asymptotic behavior was sometimes reached very late. Maybe an exponential rule such as $math2$ or $math3$ but with adaptive constants make sense: keeping the scaling with n demonstrated in this paper, but adapting the parameters, in particular during early stages of the run. Instead of using, as proposed above, the minimum between the number of resamplings proposed by the adaptive rule and the number of resamplings proposed by the non-adaptive rule, we might introduce adaptivity in the parameters of the $math2$ formula.

3. REFERENCES

[1] D. V. Arnold and H.-G. Beyer. Local performance of the (1+1)-ES in a noisy environment. *IEEE Transactions on Evolutionary Computation*, 6(1):30–41, 2002.

[2] D. V. Arnold and H.-G. Beyer. A general noise model and its effects on evolution strategy performance. *IEEE Transactions on Evolutionary Computation*, 10(4):380–391, 2006.

[3] S. Astete-Morales, J. Liu, and O. Teytaud. log-log convergence for noisy optimization. In *Proceedings of EA 2013*, LLNCS, page accepted. Springer, 2013.

[4] S. Astete Morales, J. Liu, and O. Teytaud. Noisy optimization convergence rates. In *Proceeding of the fifteenth annual conference companion on Genetic and evolutionary computation conference companion*, pages 223–224, Amsterdam, Netherlands, 2013. ACM.

[5] A. Auger. Convergence results for the (1, λ)-sa-es using the theory of ϕ-irreducible markov chains. *Theoretical Computer Science*, 334(1):35–69, 2005.

[6] S. Finck, H.-G. Beyer, and A. Melkozerov. Noisy optimization: a theoretical strategy comparison of es, egs, spsa & if on the noisy sphere. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 813–820. ACM, 2011.

[7] V. Heidrich-Meisner and C. Igel. 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*, pages 401–408, New York, NY, USA, 2009. ACM.

[8] M. Jebalia, A. Auger, and N. Hansen. Log-linear convergence and divergence of the scale-invariant (1+ 1)-es in noisy environments. *Algorithmica*, 59(3):425–460, 2011.

[9] O. Teytaud, J. Decock, et al. Noisy optimization complexity. In *FOGA-Foundations of Genetic Algorithms XII-2013*, 2013.