Noisy Optimization Convergence Rates

Extended Abstract

Sandra Astete Morales, Jialin Liu, Olivier Teytaud
TAO, Inria, Lri, UMR 8623 (CNRS - Univ. Paris-Sud), Univ. Paris-Sud, France

ABSTRACT

We consider noisy optimization problems, without the assumption of variance vanishing in the neighborhood of the optimum. We show mathematically that evolutionary algorithms with simple rules and with exponential number of resamplings lead to a log-log convergence rate (log of the distance to the optimum linear in the log of the number of resamplings), as well as with number of resamplings polynomial in the inverse step-size.

Categories and Subject Descriptors

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

General Terms

Theory

Keywords

Noisy Optimization, Evolution Strategies

1. THEORETICAL ANALYSIS: EXPONENTIAL NON-ADAPTIVE RULES CAN LEAD TO LOG/LOG CONVERGENCE.

Scale-invaraint noisy optimization theorem: *Assume that without resampling ($r_n = 1$ in Alg. 1), for any $\delta > 0$, for some $\alpha > 0$, $\alpha' > 0$, with probability $1 - \delta/2$, with objective function $fitness(x) = ||x||$, $\exists C, C'$; $C' \exp(-\alpha' n) \leq ||x_n|| \leq C \exp(-\alpha n)$. Assume, additionally, that there is scale invariance: $\sigma_n = C''||x_n||$ for some $C'' > 0$. Then, for any $\delta > 0$, there is $K_0 > 0, \zeta_0 > 0$ such that for $K \geq K_0$, $\zeta > \zeta_0$, then the convergence above also holds with probability at least $1 - \delta$ for fitness function $f(z) = ||z||^p + \mathcal{N}$ and resampling rule as in Alg. 1.*

Remarks: Our theorem shows that if a scale invariant algorithm converges in the noise-free case, then it also converges in the noisy case with the exponential resampling rule, at least if parameters are large enough. We show a log-linear convergence rate as in the noise-free case, but at the cost of more evaluations per iteration. When normalized by the number of function evaluations, we get $\log(||x_n||)$ linear in the logarithm of the number of function evaluations, as detailed in Corollary 1. This is a log-log convergence when the results is properly normalized by the number of evaluations.

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

Algorithm 1 An evolution strategy, with exponential number of resamplings. If we consider $K = 1$ and $\zeta = 1$ we obtain the case without resampling. $\mathcal{N}$ is an arbitrary random variable with bounded density (each use is independent of others).

Parameters: $K > 0$,$\zeta \geq 0$, $\lambda \geq \mu > 0$, a dimension $d > 0$.
Input: an initial $x_1 \in \mathbb{R}^d$ and an initial $\sigma_0 > 0$.
$n \leftarrow 1$
while (true) **do**
  Generate λ individuals $i_1, \ldots, i_\lambda$ independently using $i_j = x_n + \sigma_{n,j}\mathcal{N}$.
  Evaluate each of them $r_n = \lceil K\zeta^n \rceil$ times and average their fitness values.
  Select the μ best individuals $j_1, \ldots, j_\mu$.
  Update: from x, σ_n, $i_1, \ldots, i_\lambda$ and $j_1, \ldots, j_\mu$, compute x_{n+1} and σ_{n+1}.
  $n \leftarrow n + 1$
end while

We have shown this property for an exponentially increasing number of resamplings, which is indeed similar to R-EDA[5], which converges with a small number of iterations but with exponentially many resamplings per iteration. Our experiments suggest that this also holds in the polynomial case.

In the corollary below, we get rid of the scale invariance and we extend to adaptive rules. In one corollary, we switch to both (i) adaptive resampling rule and (ii) no scale invariance; each change can indeed be proved independently of the other.

Algorithm 2 An evolution strategy, with number of resamplings polynomial in the step-size. The case without resampling means $Y = 1$ and $\eta = 0$. $\mathcal{N}$ is an arbitrary random variable with bounded density (each use is independent of others).

Parameters: $Y > 0$,$\eta \geq 0$, $\lambda \geq \mu > 0$, a dimension $d > 0$.
Input: an initial $x_1 \in \mathbb{R}^d$ and an initial $\sigma_0 > 0$.
$n \leftarrow 1$
while (true) **do**
  Generate λ individuals $i_1, \ldots, i_\lambda$ independently using $i_j = x_n + \sigma_{n,j}\mathcal{N}$.
  Evaluate each of them $r = \lceil Y {\sigma_n}^{-\eta} \rceil$ times and average their fitness values.
  Select the μ best individuals $j_1, \ldots, j_\mu$.
  Update: from x, σ_n, $i_1, \ldots, i_\lambda$ and $j_1, \ldots, j_\mu$, compute x_{n+1} and σ_{n+1}.
  $n \leftarrow n + 1$
end while

Corollary: adaptive resampling and no scale-invariance. *Assume that, in the noise free case, for any $\delta > 0$, there are constants $\rho > 0, V > 0, \rho' > 0, V' > 0$ such*

$p=1$				$p=2$				$p=4$			
d	ζ	K	Slope	d	ζ	K	Slope	d	ζ	K	Slope
2	1	1	-0.21	2	1	1	-0.11	2	1	1	-0.07
		2	-0.21			2	-0.08			2	-0.03
	2	1	-0.33		2	1	-0.14		2	1	-0.10
		2	-0.39			2	-0.19			2	-0.09
	3	1	-0.60		3	1	-0.24		3	1	-0.11
		2	-0.53			2	-0.26			2	-0.17
3	1	1	-0.38	3	1	1	-0.25	3	1	1	-0.09
		2	-0.45			2	-0.27			2	-0.14
	2	1	-0.50		2	1	-0.28		2	1	-0.14
		2	-0.55			2	-0.29			2	-0.14
	3	1	-0.49		3	1	-0.35		3	1	-0.13
		2	-0.56			2	-0.36			2	-0.15
4	1	1	-0.40	4	1	1	-0.20	4	1	1	-0.09
		2	-0.46			2	-0.25			2	-0.09
	2	1	-0.54		2	1	-0.30		2	1	-0.10
		2	-0.57			2	-0.26			2	-0.11
	3	1	-0.58		2	1	-0.32		3	1	-0.20
		2	-0.49			2	-0.35			2	-0.18
5	1	1	-0.41	5	1	1	-0.18	5	1	1	-0.05
		2	-0.44			2	-0.18			2	-0.06
	2	1	-0.56		2	1	-0.24		2	1	-0.07
		2	-0.58			2	-0.22			2	-0.09
	3	1	-0.64		3	1	-0.32		3	1	-0.10
		2	-0.58			2	-0.35			2	-0.12

Table 1: Estimated slope with $r_n = \lceil Kn^\zeta \rceil$ resamplings at iteration n, for dimension $2, 3, 4, 5$. Slopes are estimated on the second half of the curve.

that with probability at least $1-\delta$, $\forall n \geq 1, V' \exp(-\rho'_n) \leq \sigma_n \leq V \exp(-\rho n)$. Then the theorem still holds in the noisy case when the number of revaluations is $Y\left(\frac{1}{\sigma_n}\right)^\eta$ for Y and η sufficiently large. Individuals are still randomly drawn using $x_n + \sigma_n \mathcal{N}$ for some random variable $\mathcal{N}$ with bounded density.

Remark: The last remark is here for cases like self-adaptive algorithms, in which we do not use directly a Gaussian random variable, but a Gaussian random variable multiplied e.g. by $\exp(\frac{1}{\sqrt{d}})Gaussian$, with $Gaussian$ a standard Gaussian random variable. For example, SA-ES algorithms as in [1] are included in this proof because they converge log-linearly.This is a log-log convergence when the results is properly normalized by the number of evaluations.

Experiments are performed in a real setting, without scale invariance. These experiments consider a polynomial number of resamplings (see legend) rather than an exponential one or an adaptive rule depending on σ. Table 1 show the estimated slopes with $p = 1, 2, 4$ and $d = 2, 3, 4, 5$. In this table, $\mu = \min(d, \lceil \lambda/4 \rceil)$, $\lambda = \lceil d\sqrt{d} \rceil$. This is consistent with [2, 4]. We get larger slopes (faster convergence; maybe just non-asymptotically) than $-\frac{1}{2p}$, with $\zeta = 2$ or $\zeta = 3$. $\zeta = 0$ performs very poorly. R-EDA[5] reaches $-\frac{1}{2p}$; we seemingly get slightly better but this might be due to a non-asymptotic effect.

2. CONCLUSION

We have shown mathematically some log-log convergence (see Section 2.1) and studied experimentally the slope in this log-log convergence (see Section 2.2). Section 2.3 gives some research directions.

2.1 Log-log convergence

We have shown that the log-log convergence (i.e. linear convergence with x-axis the log of the number of evaluations and y-axis the log of the distance to the optimum) occurs in various cases:

- non-adaptive rules, with number of resamplings exponential in the iteration counter (mathematical proof); as shown by Corollary 2, this can be extended to non scale-invariant algorithms;
- adaptive rules, with number of resamplings polynomial in $1/\sigma_n$ with σ_n the step-size (mathematical proof; however, there is a strong sensitivity to constants Y and η with $Y\left(\frac{1}{\sigma_n}\right)^\eta$ resamplings per individual);
- non-adaptive rule, with polynomial number of resamplings; whereas this case is a quite convenient scheme experimentally, we have no proof in this case.

2.2 Slope in log-log convergence

Experimentally, the best slope in the log-log representation is often close to $-\frac{1}{2p}$ for fitness function $||x||^p + \mathcal{N}$. It is known that under modeling assumptions (i.e. the function is regular enough for being optimized by learning), it is possible to do better than that (the slope becomes $-1/2$ for infinitely differentiable cases, see [3] and references therein); under locality assumptions (if it is assumed that sampling far from the optimum can not bring additional information) there are still gaps between the upper and lower bounds, and the rates (the constants in the slope) of evolution strategies are not yet known, so we can only see experimentally a nearly good (not perfect) accordance with the $-\frac{1}{2p}$ model (for $fitness(x) = ||x||^p + \mathcal{N}$) and an accordance with the $-\frac{1}{p}$ bound (slopes are always at best $-\frac{1}{p}$ at least for large numbers of evaluations). The most stable results (linear slope on the log-log scale quickly visible) come from simple non-adaptive rules, e.g. number of revaluations per individual quadratic in the number of iterations.

2.3 Further work

The main further work is the mathematical analysis of the polynomial number of resamplings in the non-adaptive case. Also, a combination of adaptive and non-adaptive rules might be interesting; adaptive rules are intuitively satisfactory, but non-adaptive polynomial rules provide simple efficient solutions, with empirically easy (no tuning) results. If our life depended on a scheme, we would for the moment choose a simple polynomial rule with a number of revaluations quadratic in the number of evaluations, in spite of (maybe) moderate elegance due to lack of adaptivity.

3. REFERENCES

[1] A. Auger. Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible Markov chains. *Theoretical Computer Science*, 334(1-3):35–69, 2005.

[2] H.-G. Beyer. *The Theory of Evolution Strategies*. Natural Computing Series. Springer, Heideberg, 2001.

[3] R. Coulom. Clop: Confident local optimization for noisy black-box parameter tuning. In *Advances in Computer Games*. Springer Berlin Heidelberg, 2012.

[4] H. Fournier and O. Teytaud. Lower bounds for comparison based evolution strategies using vc-dimension and sign patterns. *Algorithmica*, January 2010.

[5] P. Rolet and O. Teytaud. Bandit-based estimation of distribution algorithms for noisy optimization: Rigorous runtime analysis. In *Proceedings of Lion4 (accepted); presented in TRSH 2009 in Birmingham*, 2009.