Theory of Evolution Strategies and Related Algorithms

Youhei Akimoto & Anne Auger
Shinshu University, Nagano, Japan
Inria Saclay Ile-de-France, France
y_akimoto@shinshu-u.ac.jp
anne.auger@lri.fr

http://www.sigevo.org/gecco-2015/

GECCO'15 Companion, July 11–15, 2015, Madrid, Spain.
ACM 978-1-4503-3488-4/15/07.
http://dx.doi.org/10.1145/2739482.2756585

Motivations

Evolution Strategies (ES): state-of-the-art methods for stochastic black-box optimization in continuous domain

in particular CMA-ES algorithm

Often argued in the EC field that theory lags behind "practice"

still true for ES … but less true than 15 years ago

Objectives of the tutorial

Give an overview of state-of-the-art theoretical results on ES

related to important practical properties of ES

Explain where and how theory is useful for algorithm design

Theory of Evolution Strategies

Basics: from discrete to continuous optimization

"interesting" theoretical questions and their relationship to practice

Linear convergence of adaptive algorithms

illustrate benefits and limitations of theory wrt experiments

Progress rate theory

provides "tight" lower bounds on convergence rates and give optimal parameter settings

Information geometry perspective

where theory sheds new light on "old" algorithms and gives new perspectives for algorithm design

Theory vs Experiments

Theory and experimental work complement each other very well

- theoretical results can hold for class of functions (infinite # of f)
- experiments done on single functions
- (often) on functions where theory cannot be tackled

  need theoretical results to generalize (like invariance)
- theory can reveal unexpected results that one would not have thought about (testing)
- theory finds inspiration in simulation / experiments
- simulations are useful to test quickly (promising) hypothesis
- for algorithm design: both theory and experiments are essential

Optimization in the Continuous World

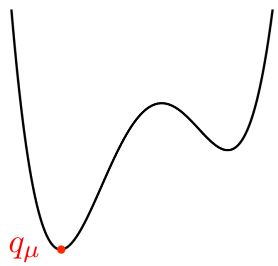
Minimize $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}^+$

i.e. find essential infimum $f(\mathbf{x}^*) = \operatorname{ess\,inf} f$

Essential infimum $q_\mu = \operatorname{ess\,inf} f$

$\Pr_{\mathbf{X}\sim\mu}(f(\mathbf{X}) < q_\mu) = 0$
$\Pr_{\mathbf{X}\sim\mu}(f(\mathbf{X}) < q_\mu + \epsilon) > 0$ for all ϵ

depends on μ



5

A Simple Continuous Algorithm (1+1)-ES

(1+1)-ES constant step-size

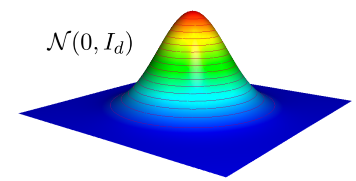
Given $f : \mathbb{R}^n \to \mathbb{R}^+$, $\sigma > 0$
Initialize $\mathbf{X}_0 \in \mathbb{R}^n$

While not happy

$\tilde{\mathbf{X}}_t = \mathbf{X}_t + \sigma\mathcal{N}(0, I_d)$
If $f(\tilde{\mathbf{X}}_t) \le f(\mathbf{X}_t)$
  $\mathbf{X}_{t+1} = \tilde{\mathbf{X}}_t$
$t = t + 1$

comparison-based algorithm

Sampling density



2-D multivariate normal distribution density

6

A Simple Continuous Algorithm (1+1)-ES

(1+1)-ES constant step-size

Given $f : \mathbb{R}^n \to \mathbb{R}^+$, $\sigma > 0$
Initialize $\mathbf{X}_0 \in \mathbb{R}^n$

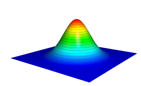
While not happy

$\tilde{\mathbf{X}}_t = \mathbf{X}_t + \sigma\mathcal{N}(0, I_d)$
If $f(\tilde{\mathbf{X}}_t) \le f(\mathbf{X}_t)$
  $\mathbf{X}_{t+1} = \tilde{\mathbf{X}}_t$
$t = t + 1$

This algorithm will never hit the optimum

$\forall \mathbf{x} \neq \mathbf{x}_0,\ \forall t > 0,\ \Pr(\mathbf{X}_t = \mathbf{x}) = 0$

because for a continuous random variable Y
$\Pr(\mathrm{Y} = \mathbf{x}) = 0$ for all $\mathbf{x}$
here $\mathrm{Y} = \mathcal{N}(0, I_d)$



7

A Simple Continuous Algorithm (1+1)-ES

(1+1)-ES constant step-size

Given $f : \mathbb{R}^n \to \mathbb{R}^+$, $\sigma > 0$
Initialize $\mathbf{X}_0 \in \mathbb{R}^n$

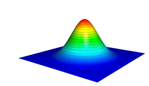
While not happy

$\tilde{\mathbf{X}}_t = \mathbf{X}_t + \sigma\mathcal{N}(0, I_d)$
If $f(\tilde{\mathbf{X}}_t) \le f(\mathbf{X}_t)$
  $\mathbf{X}_{t+1} = \tilde{\mathbf{X}}_t$
$t = t + 1$

This algorithm will never hit the optimum

$\forall \mathbf{x} \neq \mathbf{x}_0,\ \forall t > 0,\ \Pr(\mathbf{X}_t = \mathbf{x}) = 0$

instead
$\Pr(\mathrm{Y} \in B(\mathbf{x}, \epsilon)) > 0$ for all $\mathbf{x}$



8

A Simple Continuous Algorithm

(1+1)-ES

(1+1)-ES constant step-size

Given $f : \mathbb{R}^n \to \mathbb{R}^+$, $\sigma > 0$

Initialize $\mathbf{X}_0 \in \mathbb{R}^n$

While not happy

$$\tilde{\mathbf{X}}_t = \mathbf{X}_t + \sigma \mathcal{N}(0, I_d)$$

If $f(\tilde{\mathbf{X}}_t) \leq f(\mathbf{X}_t)$

$\mathbf{X}_{t+1} = \tilde{\mathbf{X}}_t$

$t = t + 1$

This algorithm will never hit the optimum

$$\forall \mathbf{x} \neq \mathbf{x}_0, \forall t > 0, \Pr(\mathbf{X}_t = \mathbf{x}) = 0$$

instead the algorithm can approximate the optimum with arbitrary precision

9

Discrete versus Continuous

Hitting Time

Discrete domain: hitting time of the optimum

$$T = \inf\{t \in \mathbb{N}, \mathbf{X}_t = \mathbf{x}^\star\}$$

Continuous domain: hitting time of epsilon-ball around optimum

fix an arbitrary ϵ, define

$$T_\epsilon = \inf\{t \in \mathbb{N}, \mathbf{X}_t \in B(\mathbf{x}^\star, \epsilon)\}$$
$$= \inf\{t \in \mathbb{N}, \|\mathbf{X}_t - \mathbf{x}^\star\| \leq \epsilon\}$$

(alternative) $T_\epsilon = \inf\{t \in \mathbb{N}, |f(\mathbf{X}_t) - f(\mathbf{x}^\star)| \leq \epsilon\}$

Note: depends also on dimension

$$T_\epsilon = \mathcal{T}(\epsilon, n)$$

10

Hitting Time versus Convergence

Finite hitting time for all epsilon

$$T_\epsilon = \inf\{t \in \mathbb{N}, \mathbf{X}_t \in B(\mathbf{x}^\star, \epsilon)\}$$
$$T_\epsilon < \infty \text{ for all } \epsilon > 0$$

$$\Longleftrightarrow$$

Convergence towards the optimum

$$\lim_{t \to \infty} \mathbf{X}_t = \mathbf{x}^\star$$

translate that an algorithm approximates the optimum with arbitrary precision

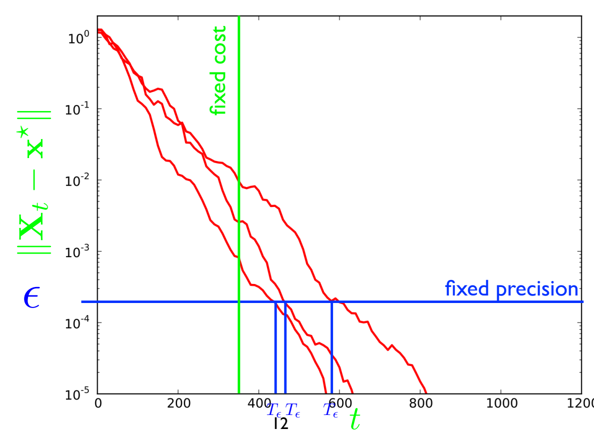
11

Hitting Time versus Convergence

two side of a coin, measuring

the hitting time T_ϵ given a fixed precision ϵ

the precision $\|\mathbf{X}_t - \mathbf{x}^\star\|$ (or ϵ) given the iteration number t



12

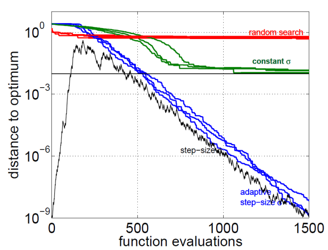
On Convergence alone ...

A theoretical convergence result is a "guarantee" that the algorithm will approach the solution in infinite time

$$\lim_{t\to\infty} \mathbf{X}_t = \mathbf{x}^\star$$

often the first/only question investigated about an optimization algorithm

But a convergence result alone is pretty meaningless in practice as it does not tell how fast the algorithm converges



need to quantify how fast the optimum is approached

13

Quantifying How Fast the Optimum is Approached

For a fixed dimension

convergence speed of $\mathbf{X}_t$ towards $\mathbf{x}^\star$ ⟷ dependency in ϵ of T_ϵ find $\epsilon \mapsto \tau(\epsilon, n)$

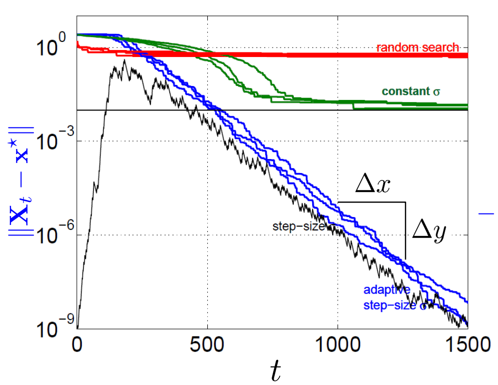
Scaling wrt the dimension

dependency of convergence rate wrt n ⟷ find $n \mapsto \tau(\epsilon, n)$

Compromises to obtain such results: asymptotic in n, in epsilon / t

14

Linear Convergence



$$-\frac{c}{n} = \frac{\Delta y}{\Delta x}$$

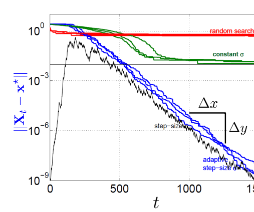
$$\frac{\|\mathbf{X}_{t+1} - \mathbf{x}^\star\|}{\|\mathbf{X}_t - \mathbf{x}^\star\|} \approx \exp\left(-\frac{c}{n}\right)$$

$$\log \frac{\|\mathbf{X}_{t+1} - \mathbf{x}^\star\|}{\|\mathbf{X}_t - \mathbf{x}^\star\|} \approx -\frac{c}{n}$$

$$\log \frac{\|\mathbf{X}_t - \mathbf{x}^\star\|}{\|\mathbf{X}_0 - \mathbf{x}^\star\|} \approx -\frac{c}{n} t$$

15

Linear Convergence



$$-\frac{c}{n} = \frac{\Delta y}{\Delta x}$$

$$\frac{\|\mathbf{X}_{t+1} - \mathbf{x}^\star\|}{\|\mathbf{X}_t - \mathbf{x}^\star\|} \approx \exp\left(-\frac{c}{n}\right)$$

$$\log \frac{\|\mathbf{X}_t - \mathbf{x}^\star\|}{\|\mathbf{X}_0 - \mathbf{x}^\star\|} \approx -\frac{c}{n} t$$

Different formal statements (not exactly equivalent)

almost surely

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{\|\mathbf{X}_t - \mathbf{x}^\star\|}{\|\mathbf{X}_0 - \mathbf{x}^\star\|} = -\frac{c}{n}$$

in expectation

$$\frac{\mathbb{E}\left[\|\mathbf{X}_{t+1} - \mathbf{x}^\star\|\right]}{\mathbb{E}\left[\|\mathbf{X}_t - \mathbf{x}^\star\|\right]} = \exp\left(-\frac{c}{n}\right)$$

$$\mathbb{E} \log \frac{\|\mathbf{X}_{t+1} - \mathbf{x}^\star\|}{\|\mathbf{X}_t - \mathbf{x}^\star\|} = -\frac{c}{n}$$

Connection with Hitting Time formulation

$$T_\epsilon \approx \frac{n}{c} \log \frac{\epsilon_0}{\epsilon}$$

16

Pure Random Search
Simple Convergence Rate Analysis

$f : \mathbf{x} \mapsto \|\mathbf{x} - \mathbf{x}^\star\|^2, \ \mathbf{x}^\star \in]0,1[^n$

Pure Random Search

sample $\mathbf{Y}_t \sim \mathcal{U}_{[0,1]^n}$ i.i.d.
$\mathbf{X}_t = \text{argmin}\{f(\mathbf{Y}_1), \dots, f(\mathbf{Y}_t)\}$

sample uniformly, keep best solution seen
blind algorithm

Convergence with probability one

$$\lim_{t \to \infty} \mathbf{X}_t = \mathbf{x}^\star \text{ almost surely}$$

proof ingredients: $\Pr(\|\mathbf{Y} - \mathbf{x}^\star\| \le \epsilon) \ge \delta (> 0)$

$\sum_t \Pr(\|\mathbf{X}_t - \mathbf{x}^\star\| > \epsilon) \le \sum_t (1-\delta)^t < \infty$ implies a.s. convergence (corollary of Borel Cantelli lemma)

17

Pure Random Search
Simple Convergence Rate Analysis

Formulation via hitting time

Theorem: For all ϵ such that $B(\mathbf{x}^\star, \epsilon) \subset]0,1[^n$

$$\mathbb{E}(T_\epsilon) = \frac{\Gamma(n/2+1)}{\pi^{n/2}} \frac{1}{\epsilon^n}$$

proof idea: T_ϵ follows a geometric distribution with parameter $p(\epsilon, n) = \Pr[\mathbf{Y} \in B(\mathbf{x}^\star, \epsilon)]$

$$\mathbb{E}[T_\epsilon] = \frac{1}{p(\epsilon, n)}$$

Formulation via convergence rate

$$\|\mathbf{X}_t - \mathbf{x}^\star\| \sim \frac{\Gamma(n/2+1)^{1/n}}{\sqrt{\pi}} \frac{1}{t^{1/n}}$$

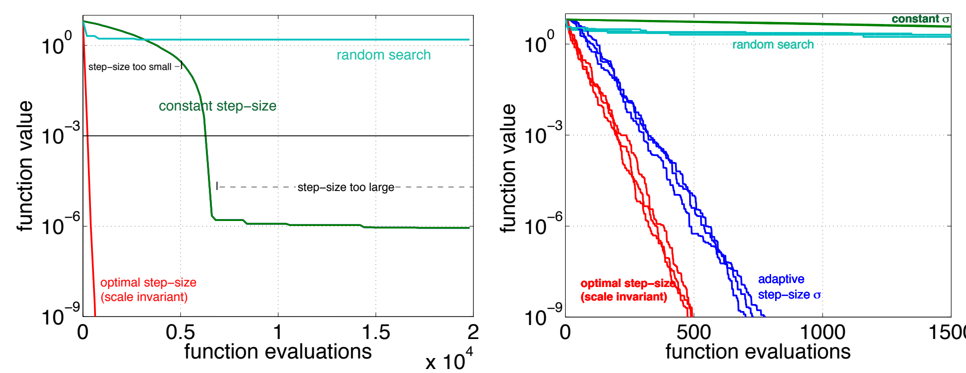
same convergence rate for (1+1)-ES with constant step-size

18

Convergence Rates - Hitting time
Wrap up

	Rate of convergence	Hitting time scaling
Pure Random Search **(1+1)-ES constant step-size**	$\frac{1}{t^{1/n}}$	$\frac{1}{\epsilon^n}$
Linear Convergence (fixed n) **+** **Linear dependence wrt n**	$\mathbb{E}[\|\mathbf{X}_t - \mathbf{x}^\star\|] = \exp\left(-\frac{c}{n}\right)^t \mathbb{E}[\|\mathbf{X}_0 - \mathbf{x}^\star\|]$ $\lim_{t\to\infty} \frac{1}{t} \log \frac{\|\mathbf{X}_t - \mathbf{x}^\star\|}{\|\mathbf{X}_0 - \mathbf{x}^\star\|} = -\frac{c}{n}$	$\frac{n}{c} \log \frac{\epsilon_0}{\epsilon}$

19



How to achieve linear convergence?

20

Adaptive Stochastic Search Algorithms

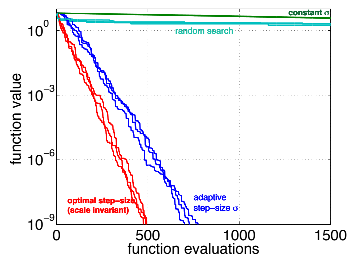
(1+1)-ES

Given $f : \mathbb{R}^n \to \mathbb{R}^+$, $\sigma > 0$
Initialize $\mathbf{X}_0 \in \mathbb{R}^n$

While not happy

$$\tilde{\mathbf{X}}_t = \mathbf{X}_t + \sigma \mathcal{N}(0, I_d)$$

If $f(\tilde{\mathbf{X}}_t) \leq f(\mathbf{X}_t)$
$\quad \mathbf{X}_{t+1} = \tilde{\mathbf{X}}_t$
$t = t + 1$



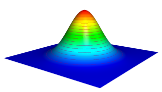
optimal step-size on $f(\mathbf{x}) = \|\mathbf{x}\|^2$

$$\sigma_t = \sigma^\star \|\mathbf{X}_t\|$$

the step-size σ needs to be adapted

adapt the scaling of the mutation

step-size proportional to the distance to the optimum



21

Adaptive Stochastic (Comparison-Based) Optimization Algorithms

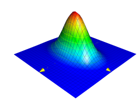
Step-size adaptive algorithms

Linear convergence on wide class of functions (ample empirical evidence)

Matyas, Random optimization, 1965
Schumer, Steiglitz, Adaptive step size random search, 1968
Devroye, The compound random search, 1972
Rechenberg, Evolution Strategies (ES), One-fifth success rule, 1973
Schwefel, Self-adaptive Evolution Strategies (SA-ES), 1981
Ostermeier, Hansen, Path-Length Control (CSA), 1994, 2001

Covariance matrix adaptive algorithms

Kjellström, Gaussian Adaptation, 1969
Hansen, Ostermeier, Covariance Matrix Adaptation ES, 2001 *State-of-the-art algorithm*
Glasmachers, Schaul, Yi, Wiestra, Schmidhuber, Exponential Natural ES, 2010



Learn second order information
solve efficiently ill-conditioned non-separable problems (ample empirical evidence) 22

(1+1)-ES with One-fifth Success Rule

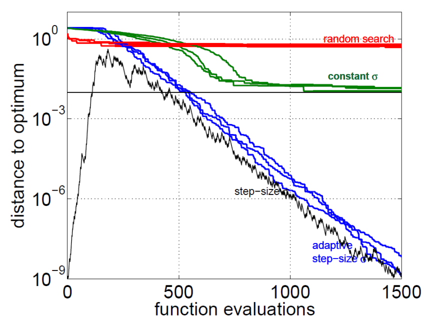
Step-size adaptive algorithm

Given $f : \mathbb{R}^n \to \mathbb{R}^+$
Initialize $\mathbf{X}_0 \in \mathbb{R}^n$, $\sigma_0 > 0$

While not happy

$$\tilde{\mathbf{X}}_t = \mathbf{X}_t + \sigma_t \mathcal{N}(0, I_d)$$

If $f(\tilde{\mathbf{X}}_t) \leq f(\mathbf{X}_t)$
$\quad \mathbf{X}_{t+1} = \tilde{\mathbf{X}}_t$
$\quad \sigma_{t+1} = \exp(1/3)\sigma_t$ — increase step-size if success
Else
$\quad \sigma_{t+1} = \exp(-1/3)^{1/4}\sigma_t$ — decrease step-size otherwise
$t = t + 1$



Rule of thumb: maintain a success probability of 1/5

23

Linear Convergence

General Lower Bounds

General Lower Bound (Jägersküpper, GECCO 2006)

Independently of how the mutation is adapted and on which function is optimized, the (1+λ) and (1,λ)-ES (λ > 1) need

$$\Omega(n \log(1/\epsilon)\lambda / \ln(\lambda))$$

function evaluations (w.o.p.) until the approximation error is at most an ε-fraction from the initial one.

Teytaud, Gelly PPSN 2006: general lower bounds for comparison-based algorithms

A, Hansen GECCO 2006, Jebalia, A, Liardet 2007: tight lower bounds, explicit asymptotic (in n) estimates

related to *progress rate theory* (Beyer, Arnold)
important for algorithm design

24

Linear Convergence - Upper bound
(1+1)-ES with one-fifth success rule

Upper Bound on the sphere (Jägersküpper, GECCO 2006)

Consider a (1+1)-ES with one-fifth success rule optimizing the SPHERE function $f(\mathbf{x}) = \|\mathbf{x}\|^2$, then the algorithm needs

$$\mathcal{O}(n \log(1/\epsilon)\lambda/\sqrt{\ln \lambda})$$

function evaluations until the approximation error is an ε-fraction from the initial one.

if λ is smaller than $O(n)$ then $\sqrt{\ln \lambda}$ faster

results on certain convex-quadratic functions where linear dependency in the condition numbers is proven (Jägersküpper, TCS 2006)

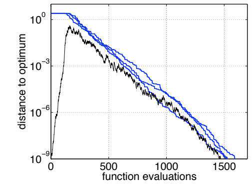
25

Linear Convergence on Scaling-Invariant Functions
Markov Chain Approach

Proof Idea

We want to prove that:

$$\frac{1}{t} \ln \frac{\|X_t\|}{\|X_0\|} \xrightarrow[t\to\infty]{} -\mathrm{CR} \quad ?$$



$$\frac{1}{t} \ln \frac{\|X_t\|}{\|X_0\|} = \frac{1}{t} \sum_{i=0}^{t-1} \underbrace{\ln \frac{\|X_{i+1}\|}{\|X_i\|}}_{= \mathcal{G}\left(Z_i := \frac{X_i}{\sigma_i}\right)}$$

homogeneous Markov chain on some functions

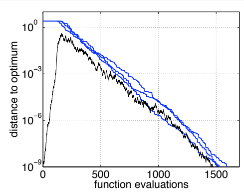
26

Linear Convergence on Scaling-Invariant Functions
Markov Chain Approach

Proof Idea

We want to prove that:

$$\frac{1}{t} \ln \frac{\|X_t\|}{\|X_0\|} \xrightarrow[t\to\infty]{} -\mathrm{CR} \quad ?$$



$$\frac{1}{t} \ln \frac{\|X_t\|}{\|X_0\|} = \frac{1}{t} \sum_{i=0}^{t-1} \ln \frac{\|X_{i+1}\|}{\|X_i\|}$$

$$= \frac{1}{t} \sum_{i=0}^{t-1} \mathcal{G}(Z_i)$$

if (Zt) "stable" enough (to satisfy LLN)

$$\xrightarrow[t\to\infty]{} \int \mathcal{G}(z)\pi(dz) =: -\mathrm{CR}$$

π invariant measure of (Zt)

27

On functions where (Z_t) is a "stable" Markov chain, we will have that for all X_0, σ_0

$$\frac{1}{t} \ln \frac{\|X_t\|}{\|X_0\|} \to \underbrace{-\mathrm{CR}}_{=\int \mathcal{G}(z)\pi(dz)} \leftarrow \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0}$$

Remaining questions

On which class of functions, for which algorithms do we have

1. $(\mathbf{Z}_t)$ is an homogeneous Markov chain?
2. $(\mathbf{Z}_t)$ is stable?

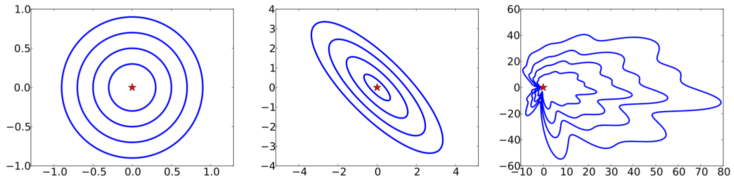
28

Answer to 1.

Class of functions: scaling-invariant functions

f is scaling-invariant if for all $\rho > 0$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$f(\mathbf{x}) \leq f(\mathbf{y}) \Leftrightarrow f(\mathbf{x}^\star + \rho(\mathbf{x} - \mathbf{x}^\star)) \leq f(\mathbf{x}^\star + \rho(\mathbf{y} - \mathbf{x}^\star)) \ .$$



Examples: if $f(\mathbf{x}) = g(\|\mathbf{x}\|)$ for any norm $\| \; \|$ and $g : \mathbb{R}^+ \to \mathbb{R}$ strictly increasing.
In particular all convex-quadratic functions are scaling invariants

Class of algorithms

Scale and translation invariant step-size adaptive randomized search
In particular step-size adaptive Evolution Strategies

Linear Convergence of Comparison-based Step-size Adaptive Randomized Search via Stability of Markov Chains, Auger, Hansen, 2014, http://arxiv.org/abs/1310.7697

29

Answer to 2.

The chain associated to the (1+1)-ES with one-fifth success rule is stable on positively homogeneous functions

$$f(\eta \mathbf{x}) = \eta^\alpha f(\mathbf{x})$$

Linear Convergence on Positively Homogeneous Functions of a Comparison Based Step-Size Adaptive Randomized Search: the (1+1) ES with Generalized One-fifth Success Rule, Auger, Hansen, 2014, http://arxiv.org/abs/1310.8397

The chain associated to the (1,λ)-ES with self-adaptation is stable on the SPHERE function (AA, TCS 2005)

presumably also on positively homogeneous functions

Presumably stability can be proven for many more algorithms

30

Benefits and Limitations of Theory
Linear CV of Adaptive Stochastic Search Algorithms

Convergence is proven on whole class of functions (pos. homogeneous functions) containing infinitely many functions

impossible to experiment on all those functions

proofs limited to a few algorithms (not CMA yet), not on all functions where we want to check the convergence

resort to experiments

Jägersküpper's proofs likely to be difficult to generalize to other algorithms (according to the author himself), not clear how much they generalize to other functions

MC approach does not allow to obtain explicit estimates for the convergence rate

31

Theory of Evolution Strategies

Basics: from discrete to continuous optimization
"interesting" theoretical questions and
their relationship to practice

Linear convergence of adaptive algorithms
illustrate benefits and limitations of theory wrt experiments

Progress rate theory
provides "tight" lower bounds on convergence rates and
give optimal parameter settings

Information geometry perspective
where theory sheds new light on "old" algorithms and
gives new perspectives for algorithm design

32

Definition: Progress Rate and Quality Gain

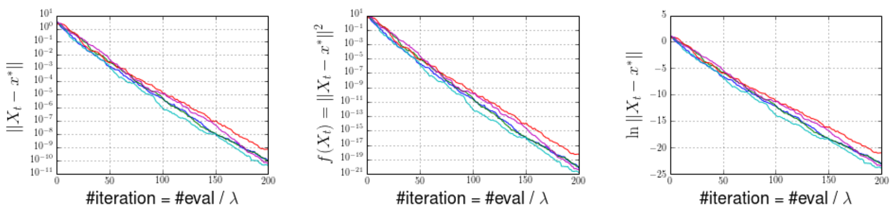
Progress Rate $\varphi^* := n\left(1 - \mathbb{E}\left[\frac{\|X_{t+1} - x^*\|}{\|X_t - x^*\|}\right]\right)$

one step expected progress in the search space

Quality Gain $\Delta^* := n\left(1 - \mathbb{E}\left[\frac{f(X_{t+1})}{f(X_t)}\right]\right)$

one step expected progress in the objective space

Log Progress $\varphi_{\ln} := \mathbb{E}[\ln\|X_t - x^*\| - \ln\|X_{t+1} - x^*\|]$



How do these quantities depend on the strategy and parameters?

Relation to Linear Convergence Rate

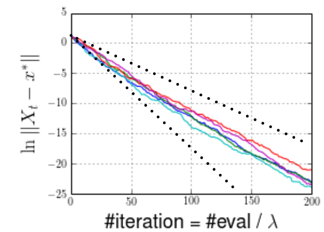
$\varphi_{\ln} := \mathbb{E}[\ln\|X_t - x^*\| - \ln\|X_{t+1} - x^*\|]$

If $\varphi_{\ln} \geqq c_u > 0$ for all $t \geqq t_0$,

$$\mathbb{E}[\ln\|X_t - x^*\|] = \mathbb{E}[\ln\|X_{t_0} - x^*\|] + \sum_{k=t_0}^{t-1}(\mathbb{E}[\ln\|X_{k+1} - x^*\|] - \mathbb{E}[\ln\|X_k - x^*\|])$$
$$\leq \mathbb{E}[\ln\|X_{t_0} - x^*\|] - \sum_{k=t_0}^{t-1} c_u$$
$$= \mathbb{E}[\ln\|X_{t_0} - x^*\|] - c_u(t - t_0)$$

If $\varphi_{\ln} \leqq c_l$ for all $t \geqq t_0$,

$$\mathbb{E}[\ln\|X_t - x^*\|] \geq \mathbb{E}[\ln\|X_{t_0} - x^*\|] - c_l(t - t_0)$$



The expected slope (in log-scale) is bounded

$$-c_l \leq \frac{1}{t - t_0} \frac{\mathbb{E}[\ln\|X_t - x^*\|]}{\mathbb{E}[\ln\|X_{t_0} - x^*\|]} \leq -c_u$$

Log-Progress of (1 + 1)-ES on Spherical Function

Def. (1+1)-ES

Initialize $X_0 \in \mathbb{R}^n, t = 0$
while not happy
compute σ_t
$Y_t = X_t + \sigma_t \mathcal{N}(0, I_n)$
$$X_{t+1} = \begin{cases} Y_t & \text{if } f(Y_t) \leq f(X_t) \\ X_t & \text{otherwise} \end{cases}$$
$t = t + 1$

Def. Spherical Function

$f(x) = g(\|x\|)$, where g increasing

Def. Scale-invariant step-size

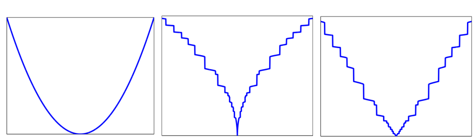
$\sigma_t = \sigma\|X_t\|$ for some $\sigma > 0$

not a practical step-size adaptation

Def. Conditional Log-Progress

$\varphi_{\ln}(X_t, \sigma_t)$
$:= \mathbb{E}[\ln\|X_t\| - \ln\|X_{t+1}\| \mid X_t, \sigma_t]$

independent on t since our algorithm is time-homogenous



they are equivalent for our algorithm

(1 + 1)-ES with adaptive step-size

Define $F_{1+1}(\sigma) = \mathbb{E}[\max(-\ln\|e_1 + \sigma\mathcal{N}\|, 0)]$ for $\sigma > 0$

- $e_1 = [1, 0, \ldots, 0]$
- $\mathcal{N} \sim \mathcal{N}(0, I_n)$ independently

Upper bound of the log-progress

For (1+1)-ES with adaptive σ_t,

$$\varphi_{\ln}(X_t, \sigma_t) \leq \sup_{\sigma \in [0,\infty)} F_{1+1}(\sigma)$$

Log-progress for scale-invariant σ_t

For (1+1)-ES with $\sigma_t = \sigma\|X_t\|$,

$$\varphi_{\ln}(X_t, \sigma_t) = F_{1+1}(\sigma)$$

⇨ The upper bound is reached by the scale-invariant step-size with $\sigma = \arg\max F_{1+1}(\sigma)$ [Jebalia et al. 2008]

$(1,\lambda)$-ES with adaptive step-size

Define $F_{(1,\lambda)}(\sigma) = -\mathbb{E}[\min_{1\le i\le\lambda} \ln\|e_1 + \sigma\mathcal{N}_i\|, 0)]$ for $\sigma > 0$

- $e_1 = [1, 0, \ldots, 0]$
- $\mathcal{N}_i \sim \mathcal{N}(0, I_n)$ independently

Upper bound of the log-progress

For $(1,\lambda)$-ES with adaptive σ_t,

$$\varphi_{\ln}(X_t, \sigma_t) \le \sup_{\sigma\in[0,\infty)} F_{(1,\lambda)}(\sigma)$$

Log-progress for scale-invariant σ_t

For $(1,\lambda)$-ES with $\sigma_t = \sigma\|X_t\|$,

$$\varphi_{\ln}(X_t, \sigma_t) = F_{(1,\lambda)}(\sigma)$$

Def. (1, λ)-ES on Spherical Function

Initialize $X_0 \in \mathbb{R}^n, t = 0$
while not happy
  compute σ_t
  for $i = 1, \ldots, \lambda$
    $Y_{t,i} = X_t + \sigma_t \mathcal{N}(0, I_n)$
  $X_{t+1} = \operatorname{argmin}_{x\in\{Y_{t,1},\ldots,Y_{t,\lambda}\}} f(x)$
  $t = t + 1$

The upper bound is reached by the scale-invariant step-size with $\sigma = \operatorname{argmax} F_{(1,\lambda)}(\sigma)$ [Auger et al. 2011]

37

Optimal σ for $(1,\lambda)$-ES in the limit $n \to \infty$

Let $\sigma^* = n\sigma$. For $n \to \infty$

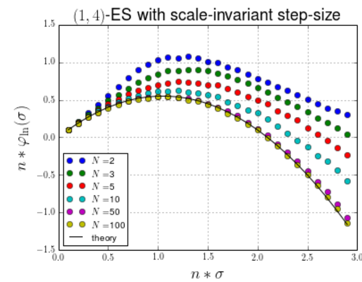
$$\lim_{n\to\infty} nF_{(1,\lambda)}(\sigma^*/n) = c_{1:\lambda}\sigma^* - \frac{(\sigma^*)^2}{2}$$

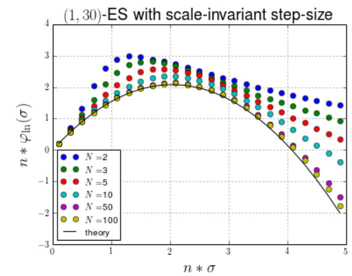
For a large n,

$$F_{(1,\lambda)}(\sigma) \approx c_{1:\lambda}\sigma - \frac{n\sigma^2}{2}$$

where $c_{1:\lambda} = \mathbb{E}[\max(\mathcal{N}_1, \ldots, \mathcal{N}_n)]$, $\mathcal{N}_i \sim \mathcal{N}(0,1)$

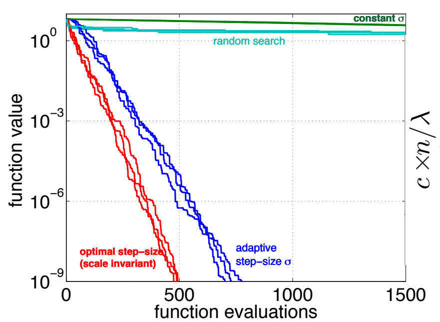
The RHS is maximized when $\sigma = \dfrac{c_{1:\lambda}}{n}$

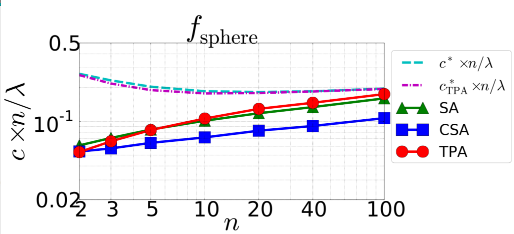




38

How helpful?





[Hansen et al. 2014]

- To evaluate how close your step-size adaptation is to the optimal one
- To design new step-size adaptation

39

$(\mu/\mu_w, \lambda)$-ES

Def. (μ/μw, λ)-ES

Given $w_i \in \mathbb{R}$
Initialize $X_0 \in \mathbb{R}^n, t = 0$
while not happy
  compute σ_t
  for $i = 1, \ldots, \lambda$
    $Y_{t,i} = X_t + \sigma_t \mathcal{N}(0, I_n)$
  sort $Y_{t,i}$ w.r.t. f and
   denote the ith best as $Y_{t,i:\lambda}$
  $X_{t+1} = \sum_{i=1}^{\lambda} w_i Y_{t,i:\lambda}$
  $t = t + 1$

- $(1, \lambda)$-ES is recovered when $w_1 = 1$ and $w_i = 0$ for $i > 1$
- How much can we gain by using all the information to update X_t?

40

Normalized Quality Gain for $(\mu/\mu_w,\lambda)$-ES

[Arnold 2005]

Normalized Quality Gain on Spherical Function

$$\Delta(X_t,\sigma_t) = \frac{n}{2}\mathbb{E}\left[\frac{f(X_t)-f(X_{t+1})}{f(X_t)} \mid X_t,\sigma_t\right] = \frac{n}{2}\left(1-\mathbb{E}\left[\frac{\|X_{t+1}\|^2}{\|X_t\|^2}\mid X_t,\sigma_t\right]\right)$$

Let $\sigma_t^* = \frac{n\sigma_t}{\|X_t\|}$. For $n\longrightarrow\infty$

$$\lim_{n\to\infty}\Delta(X_t,\sigma_t) = \sigma_t^*\sum_{i=1}^{\lambda} w_i c_{i:\lambda} - \frac{(\sigma_t^*)^2}{2}\sum_{i=1}^{\lambda} w_i^2$$

$c_{i:\lambda}$: the expectation of the ith largest among λ i.i.d. r.v. from $N(0,1)$

The RHS is maximized when

$$\sigma^* := \sigma_t^* = \frac{\sum_{i=1}^{\lambda} w_i c_{i:\lambda}}{\sum_{i=1}^{\lambda} w_i^2}$$, then $$\Delta^* := \lim_{n\to\infty}\Delta(X_t,\sigma_t) = \frac{\left(\sum_{i=1}^{\lambda} w_i c_{i:\lambda}\right)^2}{2\sum_{i=1}^{\lambda} w_i^2}$$

41

Optimal Recombination Weight for $(\mu/\mu_w,\lambda)$-ES

[Arnold 2005]

Let $\mu_w := \left(\sum_{i=1}^{\lambda} w_i^2\right)^{-1}$. Then $\Delta^* = \frac{\mu_w}{2}\left(\sum_{i=1}^{\lambda} w_i c_{i:\lambda}\right)^2$.

For an arbitrary $\mu_w > 0$, the optimal normalized quality gain is

$$\Delta^* = \frac{1}{2}\sum_{i=1}^{\lambda} c_{i:\lambda}^2$$

when $w_i = \frac{c_{i:\lambda}}{(\mu_w\sum_{i=1}^{\lambda} c_{i:\lambda}^2)^{1/2}}$ and $\sigma^* = (\mu_w\sum_{i=1}^{\lambda} c_{i:\lambda}^2)^{1/2}$

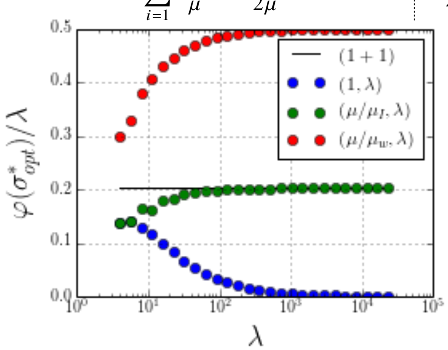
cf. for $(1,\lambda)$-ES ($w_1 = 1$, $w_i = 0$ for $i > 1$), $\Delta^* = \frac{c_{1:\lambda}^2}{2}$

⇒ we gain the factor $\frac{\sum_{i=1}^{\lambda} c_{i:\lambda}^2}{c_{1:\lambda}^2}$ by introducing weighted recombination

42

Comparison of Normalized Progress Rate

	φ^*	optimal σ^*	$\frac{\varphi^*(\sigma^*_{\mathrm{opt}})}{\lambda}$
(1+1)-ES	$\frac{\sigma^*}{\sqrt{2\pi}}\exp\left(-\frac{(\sigma^*)^2}{8}\right) - \frac{(\sigma^*)^2}{4}\left(1-\mathrm{erf}\left(\frac{\sigma^*}{\sqrt{8}}\right)\right)$	1.224	0.202
$(1,\lambda)$-ES	$\sigma^* c_{1:\lambda} - \frac{(\sigma^*)^2}{2}$	c	$\frac{c_{1:\lambda}^2}{2\lambda}$
$(\mu/\mu_w,\lambda)$-ES with optimal w	$\sigma^*\sum_{i=1}^{\lambda} w_i c_{i:\lambda} - \frac{(\sigma^*)^2}{2}\sum_{i=1}^{\lambda} w_i^2$	$\left(\frac{\sum_{i=1}^{\lambda} c_{i:\lambda}^2}{\sum_{i=1}^{\lambda} w_i^2}\right)^{1/2}$	$\frac{1}{2\lambda}\sum_{i=1}^{\lambda} c_{i:\lambda}^2$
$(\mu/\mu_I,\lambda)$-ES with μ = ⌊0.27λ⌋	$\sigma^*\sum_{i=1}^{\mu}\frac{c_{i:\lambda}}{\mu} - \frac{(\sigma^*)^2}{2\mu}$	$\sum_{i=1}^{\mu} c_{i:\lambda}$	$\frac{(\sum_{i=1}^{\mu} c_{1:\lambda})^2}{2\lambda\mu}$



Progress Rate Theory: Summary

More results on Noisy Sphere, Parabolic Ridge

H.-G. Beyer: The Theory of Evolution Strategies (Springer Verlag, 2001)
Hansen, N., D.V. Arnold, and A. Auger (2015). Evolution Strategies. To appear in Janusz Kacprzyk and Witold Pedrycz (Eds.): Handbook of Computational Intelligence, Springer

Used to design new algorithms

- Mirrored Sampling [Brockhoff et al. 2010]
- Median Success Rule (step-size adaptation) [Ait Elhara et al. 2013]

Limitations

- based on the approximation ($n\longrightarrow\infty$)
- sometimes based on other approximations (not easy to appraise the validity of the result)
- existence of the stationary distribution assumed
- scale-invariant step-size is not practical

Connexion to Markov chain approach for linear convergence:

In "progress rate" approach, it is assumed that $\frac{\|X_t\|}{\sigma_t}$ is constant by assuming $\sigma_t = \sigma\|X_t\|$ (remove stochasticity), while for a step-size adaptive algorithm it is the norm of a Markov chain.

44

Theory of Evolution Strategies

Basics: from discrete to continuous optimization
"interesting" theoretical questions and
their relationship to practice

Linear convergence of adaptive algorithms
illustrate benefits and limitations of theory wrt experiments

Progress rate theory
provides "tight" lower bounds on convergence rates and
give optimal parameter settings

Information geometry perspective
where theory sheds new light on "old" algorithms and
gives new perspectives for algorithm design

45

Black-Box Search Template

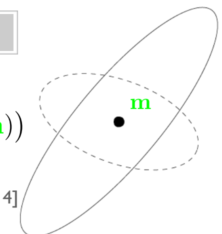
A black-box search template to minimize $f : \mathbb{R}^n \to \mathbb{R}$

Initialize distribution parameters θ, set population size λ
While not terminate

1. Sample distribution $p_\theta(x) : x_1, \ldots, x_\lambda \in \mathbb{R}^n$
2. Evaluate $x_1, \ldots, x_\lambda$ on f
3. Update parameters $\theta \leftarrow F(\theta, x_1, \ldots, x_\lambda, f(x_1), \ldots, f(x_\lambda))$

Example of p_θ on $\mathbb{R}^n$

multivariate normal distribution: $\mathbf{m} + \sigma \mathcal{N}(\mathbf{0}, \mathbf{C})$
density : $p_{\theta:=(\mathbf{m},\mathbf{C})}(x) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{C}|}} \exp\left(-\frac{1}{2}(x-\mathbf{m})^T \mathbf{C}^{-1}(x-\mathbf{m})\right)$



- Covariance Matrix Adaptation Evolution Strategies (CMA-ES) [N. Hansen et al, 2001-2014]
- Exponential Natural Evolution Strategies (xNES) [T. Glasmachers et al, 2010]

$\{x | (x-\mathbf{m})^T \mathbf{C}^{-1}(x-\mathbf{m}) = cst\}$

46

Change of Perspective: Optimization of θ

Natural Evolution Strategies (NES) [D. Wierstra et al, 2008, 2014]

Optimization of x → Optimization of θ

Search Space X → Statistical Manifold Θ
equipped with the Fisher metric I
Objective function f → Function J of θ

Objective of the update of θ

Expectation of f over P_θ: $$J(\theta) = \int_X f(x) p_\theta(x) dx$$

"adds one degree of smoothness"[T. Glasmachers. PGMO-COPI 2014]

- typically, $\inf_\theta J(\theta) = f^* = \operatorname{essinf}_x f(x)$
- by Markov inequality, $\Pr[|f(x) - f^*| < \epsilon] \geq 1 - \frac{J(\theta) - f^*}{\epsilon}$

minimization of J ⇒ minimization of f

47

Gradient Descent on $J(\theta)$

Natural Gradient [S. Amari, 1998]

Instead of taking the "vanilla" gradient $\nabla J(\theta) = \left[\frac{\partial J}{\partial \theta_1}, \ldots, \frac{\partial J}{\partial \theta_n} \right]^T$
that gives the steepest direction in the Euclidean sense

$$\frac{\nabla J(\theta)}{\|\nabla J(\theta)\|} = \lim_{\epsilon \to 0} \epsilon^{-1} \operatorname*{argmax}_{\|\Delta\| \leq \epsilon} J(\theta + \Delta)$$

taking the "natural" gradient $\tilde{\nabla} J(\theta) = \mathcal{I}(\theta)^{-1} \nabla J(\theta)$
that gives the steepest direction w.r.t. the KL-divergence

$$\frac{\tilde{\nabla} J(\theta)}{\|\tilde{\nabla} J(\theta)\|} = \lim_{\epsilon \to 0} \epsilon^{-1} \operatorname*{argmax}_{D_{\mathrm{KL}}(P_\theta \| P_{\theta+\Delta}) \leq \epsilon^2} J(\theta + \Delta)$$

considered also as the gradient on the differential manifold Θ
equipped with the Fisher metric in the given coordinate θ

48

Update of θ

Stochastic Natural Gradient Descent

$$
\begin{aligned}
\tilde\nabla J(\theta)\mid_{\theta=\theta^t} &= \tilde\nabla J(\theta)\mid_{\theta=\theta^t} \\
&= \tilde\nabla\left(\int f(x)p_\theta(x)dx\right)\mid_{\theta=\theta^t} \\
&= \int f(x)\tilde\nabla(p_\theta(x))\mid_{\theta=\theta^t} dx && \text{[exchange of int. and diff.]} \\
&= \int f(x)p_\theta(x)\tilde\nabla \ln(p_\theta(x))\mid_{\theta=\theta^t} dx && \nabla p_\theta(x) = p_\theta(x)\nabla\ln(p_{\theta^t}) \text{ [log-likelihood trick]} \\
&\approx \frac{1}{\lambda}\sum_{i=1}^{\lambda} f(x_i)\tilde\nabla\ln(p_\theta(x_i))\mid_{\theta=\theta^t} && x_1,\dots,x_\lambda \text{ are i.i.d. from } p_{\theta^t} \text{ [Monte-Carlo Approx.]}
\end{aligned}
$$

Parameter update

$$\theta^{t+1} = \theta^t + \eta\frac{1}{\lambda}\sum_{i=1}^{\lambda} f(x_i)\tilde\nabla\ln(p_\theta(x_i))\mid_{\theta=\theta^t}$$

η : learning rate (i.e., step-size)

$\tilde\nabla\ln(p_\theta(x))$ is analytically derivable for some probability models, e.g., normal distributions

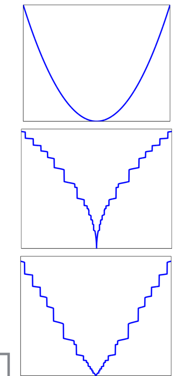
Information Geometric Optimization [Y. Ollivier et al. (2011)]

Not *invariant* to increasing transformations of f

not woking well without η adaptation because of this defect

$$\int f(x)p_\theta(x)dx \neq \int (g\circ f)(x)p_\theta(x)dx$$

$$\frac{1}{\lambda}\sum_{i=1}^{\lambda} f(x_i)\tilde\nabla\ln(p_\theta(x_i))\mid_{\theta=\theta^t} \neq \frac{1}{\lambda}\sum_{i=1}^{\lambda}(g\circ f)(x_i)\tilde\nabla\ln(p_\theta(x_i))\mid_{\theta=\theta^t}$$



Quantile-based Objective Transformation

$$f(x) \mapsto W^f_{\theta^t}(x) = w(P_{\theta^t}[X : f(X) \le f(x)])$$
$$\approx w(\#\{x_i : f(x_i) < f(x)\}/\lambda) \quad x_1,\dots,x_\lambda \sim P_{\theta^t}$$

- w: nonincreasing
- scaled in $[w(1), w(0)]$ at each iteration
- *invariant* to any increasing transformation, $(g\circ f)$

Parameter Update: $\theta^{t+1} = \theta^t + \eta\sum_{i=1}^{\lambda} w_{\mathrm{rk}(x_i)}\tilde\nabla\ln(p_\theta(x_i))\mid_{\theta=\theta^t}$

$$w_{\mathrm{rk}(x_i)} = \frac{1}{\lambda}w\left(\frac{\mathrm{rk}(x_i)-1/2}{\lambda}\right),\quad \text{where}\quad \mathrm{rk}(x_i) = \#\{x_j : f(x_j)\le f(x_i)\}$$

Instantiation

Multivariate Normal Distribution $N(m, C)$ [Glasmachers et al. 2010] [Akimoto et al. 2010]

$$m^{t+1} = m^t + \eta_m\sum_{i=1}^{\lambda} w_{\mathrm{rk}(x_i)}(x_i - m^t)$$

$$C^{t+1} = C^t + \eta_C\sum_{i=1}^{\lambda} w_{\mathrm{rk}(x_i)}[(x_i - m^t)(x_i - m^t)^{\mathrm{T}} - C^t]$$

= *Pure rank-μ update CMA-ES* [Hansen et al. 2003]

Multivariate Bernoulli Distribution with probability parameter θ [Ollivier et al. 2011]

$$\theta^{t+1} = \theta^t + \eta\sum_{i=1}^{\lambda} w_{\mathrm{rk}(x_i)}(x_i - \theta^t)$$

pmf: $p_\theta(x) = \prod_{i=1}^{d}\theta_i^{x_i}(1-\theta_i)^{1-x_i}$

= *Population Based Incremental Learning (PBIL)* [Baluja et al. 1995]

How is this perspective helpful? Theoretical Aspects

Twofold approximation of the solution to the ODE

$$\frac{d\theta(t)}{dt} = \tilde\nabla J_{\theta^t}(\theta)\mid_{\theta=\theta(t)}$$

Euler Discretization ($\eta \to 0$)

$$\theta^{t+\eta} = \theta^t + \eta\tilde\nabla J_{\theta^t}(\theta)\mid_{\theta=\theta^t}$$

Monte-Carlo Approx. ($\lambda \to \infty$)

$$\theta^{t+\eta} = \theta^t + \eta\sum_{i=1}^{\lambda} w_{\mathrm{rk}(x_i)}\tilde\nabla\ln(p_\theta(x_i))\mid_{\theta=\theta^t}$$

1. Convergence analysis of the ODE solution
   - variant with isotropic Gaussian [Akimoto et al. 2012][Glasmachers et al. 2012]
   - full Gaussian [Beyer 2014]
2. Convergence analysis of the infinite population model [Akimoto 2012]
   - Pure rank-mu update CMA with *fitness proportional weight*
   - $\lim_{t\to\infty}\mathrm{Cond}(C^t A) = 1$ *and its geometric convergence is proven on f(x)= $x^{\mathrm{T}}Ax$*

How is this perspective helpful?
Algorithm Design and U

Deriving algorithm variants from the same principle as CMA

- Linear time/space variants with restricted Gaussian for large scale problem
  - R1-NES [Sun et al. 2013]
  - VD-CMA [Akimoto et al. 2014]

Provide new interpretation of existing algorithms

- Active CMA [Jastrebski et al. 2006] is interpreted as the natural gradient estimation with *baseline* [Sun et al. 2009] (technique to reduce the estimation variance)
- Separable CMA [Ros et al. 2008] is derived from the IGO with Gaussian with diagonal covariance matrix [Akimoto et al. 2012]

Still, Information Geometric framework does not cover "many" relevant aspects for robust algorithm design:

- choice of some parameters (learning rate, …)
- cumulation, …

53

Not covered topics

Invariance

allow to generalize an empirical result on a function to a set of (infinitely many) functions

- invariance to order preserving transformation of f
- invariance to affine transformation of the search space X
  - translation
  - rotation
  - coordinate-wise scaling

Unbiasedness of the parameter update

Rapid divergence on a linear function

Maximal Likelihood Principle …

54