# A Blend of Invariance and Stochastic Stability for Proving Linear Convergence of Adaptive Evolution Strategies

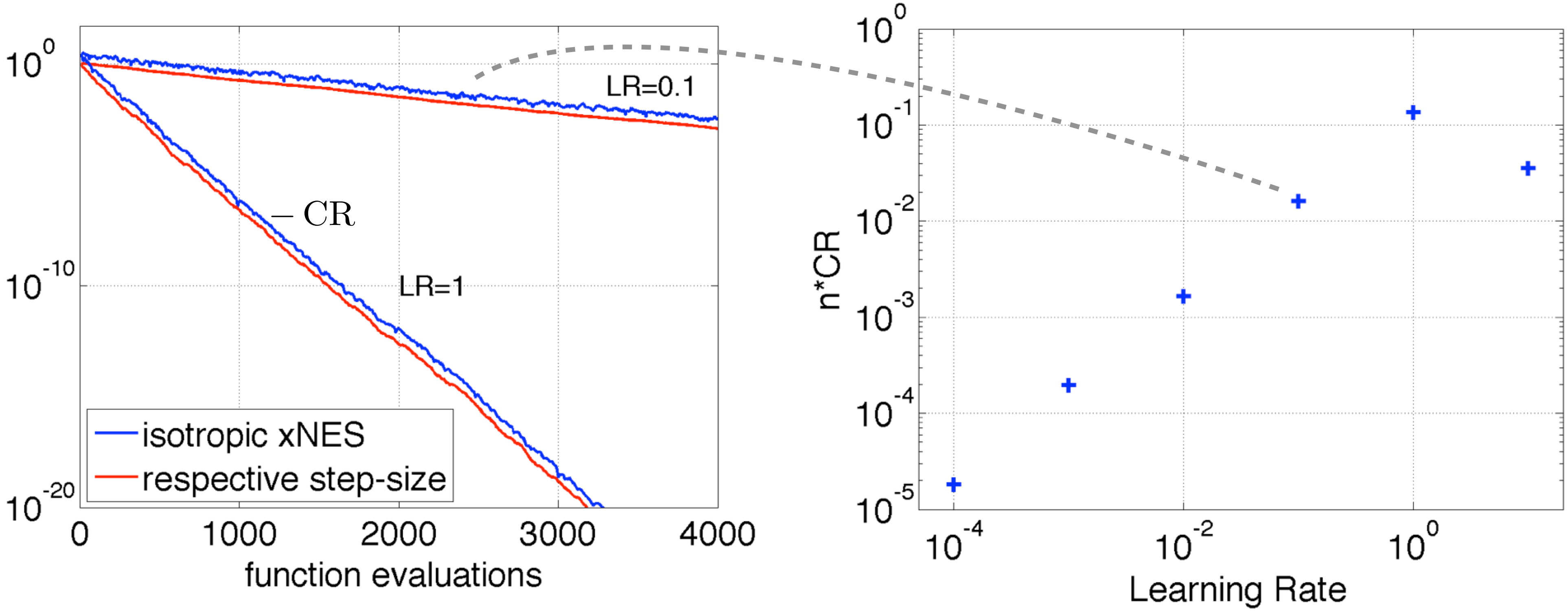
Anne Auger

joint work with N. Hansen

Optimization and Machine Learning Team (TAO)
INRIA Saclay-Ile-de-France

SIMINOLE Meeting, October 2012

# Linear Convergence for Different Learning Rates




**Linear or Exponential or Geometric convergence**

$\mathbf{x}_t$ deterministic

$$\ln \frac{\|\mathbf{x}_{t+1}\|}{\|\mathbf{x}_t\|} \to -\,\mathrm{CR} \qquad \Rightarrow \frac{1}{t} \ln \|\mathbf{x}_t\| \to -\,\mathrm{CR}$$

$$\frac{\|\mathbf{x}_{t+1}\|}{\|\mathbf{x}_t\|} \to \exp(-\,\mathrm{CR})$$

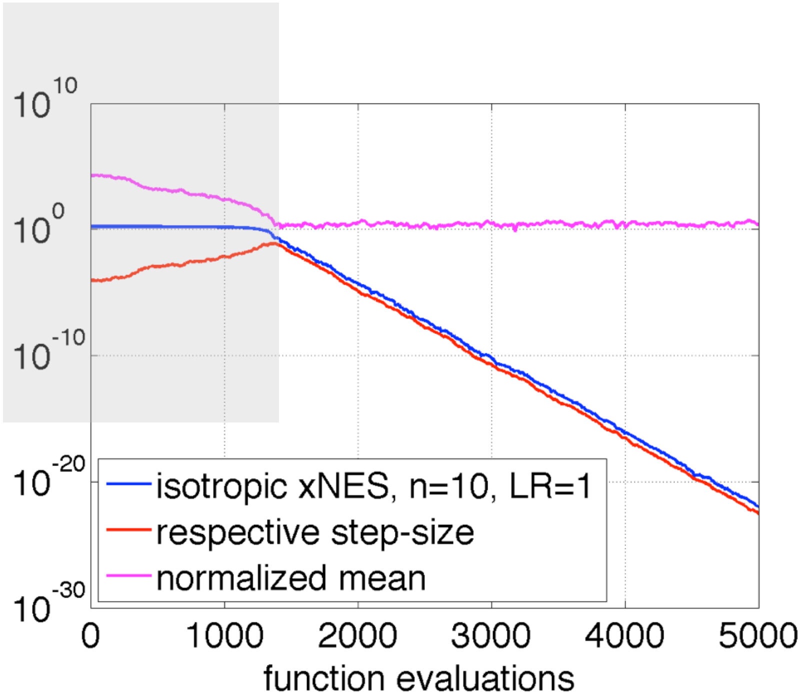
$f_{\mathrm{sphere}}(\mathbf{x}) = \|\mathbf{x}\|$, $n = 10$
isotropic xNES with CMA-ES default settings

xNES: Natural Evolution Strategies with exponential parametrization. Glasmachers et a.l. GECCO 2010

# In this Talk

Linear convergence of step-size adaptive ES without need of LR small *(needed for stochastic approximation approach)*

Reach stationary regime at geometric rate independent of starting point




Approach exploits invariances of algorithms and use stability analysis of underlying Markov chain

*hold on scaling-invariant functions*

# Motivating Example
# xNatural Evolution Strategies

xNES optimizing $f : \mathbb{R}^n \to \mathbb{R}$ $\quad (\mathbf{X}_t, \sigma_t) \in \mathbb{R}^n \times \mathbb{R}^+$

Sample $\lambda$ solutions: $\mathbf{X}_t + \sigma_t \mathbf{U}_t^i$, $\mathbf{U}_t^i \sim \mathcal{N}(0, I_d)$ i.i.d.

$\mathbf{U}_t = [\mathbf{U}_t^1, \ldots, \mathbf{U}_t^\lambda]$

Evaluate and rank solutions:

$$f(\mathbf{X}_t + \sigma_t \mathbf{U}_t^{1:\lambda}) \leq \ldots \leq f(\mathbf{X}_t + \sigma_t \mathbf{U}_t^{\lambda:\lambda})$$

$$\mathcal{S}el_{(\mathbf{X}_t, \sigma_t)} : \mathbf{U}_t \mapsto \mathbf{Y}_t := (\mathbf{U}_t^{1:\lambda}, \ldots, \mathbf{U}_t^{\mu:\lambda})$$

ranked-based selection

Update:

$$\mathbf{X}_{t+1} = \mathbf{X}_t + c_m \sigma_t \sum_{i=1}^{\mu} w_i \mathbf{Y}_t^i \qquad w_1 \geq \ldots \geq w_\mu$$

$$\sigma_{t+1} = \sigma_t \exp \underbrace{\left( \frac{c_\sigma}{2n} \left( \sum_{i=1}^{\mu} w_i (\|\mathbf{Y}_t^i\|^2 - n) \right) \right)}_{\eta^\star((\mathbf{X}_t, \sigma_t), \mathbf{Y}_t)}$$

# Motivating Example
# xNatural Evolution Strategies

$$(\mathbf{X}_t, \sigma_t) \in \mathbb{R}^n \times \mathbb{R}^+$$

$$\begin{aligned}(\mathbf{X}_{t+1}, \sigma_{t+1}) &= \mathcal{G}((\mathbf{X}_t, \sigma_t), \mathbf{Y}_t) \\ &= \mathcal{G}((\mathbf{X}_t, \sigma_t), \mathcal{S}el_{(\mathbf{X}_t, \sigma_t)}(\mathbf{U}_t))\end{aligned}$$

$\mathbf{U}_t$ i.i.d.

$$\mathcal{S}el_{(\mathbf{X}_t, \sigma_t)} : \mathbf{U}_t \mapsto \mathbf{Y}_t := (\mathbf{U}_t^{1:\lambda}, \ldots, \mathbf{U}_t^{\mu:\lambda})$$

$$\mathcal{G}((\mathbf{x}, \sigma), \mathbf{y}) = \begin{pmatrix} \mathbf{x} + \sigma c_m \sum_{i=1}^{\mu} w_i \mathbf{y}^i \\ \sigma \exp\left(\frac{c_\sigma}{2n} \sum_{i=1}^{\mu} w_i (\|\mathbf{y}^i\|^2 - n)\right) \end{pmatrix} = \begin{pmatrix} \mathcal{G}_1((\mathbf{x}, \sigma), \mathbf{y}) \\ \mathcal{G}_2(\sigma, \mathbf{y}) \end{pmatrix}$$

# Motivating Example
# xNatural Evolution Strategies

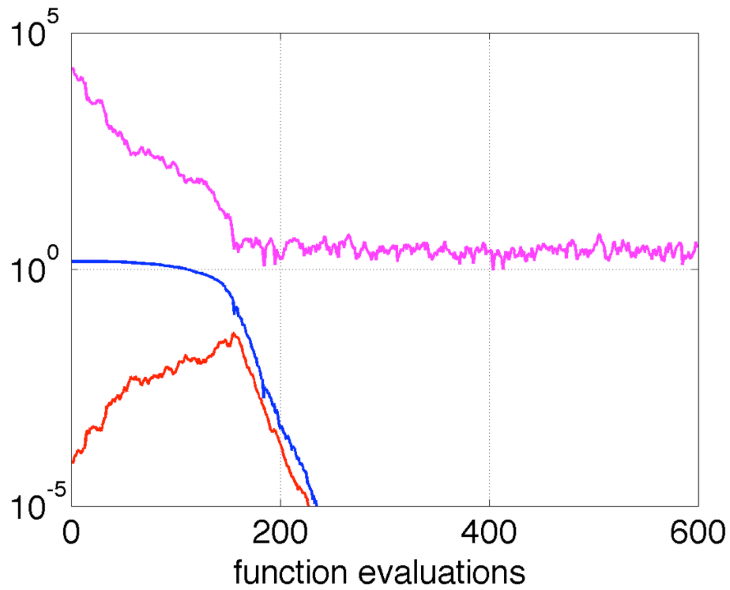
On scaling-invariant functions $f(\mathbf{x}) \leq f(\mathbf{y}) \Leftrightarrow f(\sigma\mathbf{x}) \leq f(\sigma\mathbf{y})$

w.l.g. def. w.r.t. 0

$\mathbf{Z}_t = \frac{\mathbf{X}_t}{\sigma_t}$ is an homogeneous Markov Chain

$$\mathbf{Y}_t = \mathcal{S}el_{(\mathbf{Z}_t,1)}(\mathbf{U}_t)$$

$$\mathbf{Z}_{t+1} = \frac{\mathbf{Z}_t + c_m \sum_{i=1}^{\mu} w_i \mathbf{Y}_t^i}{\underbrace{\exp\left(\frac{c_\sigma}{2n}\left(\sum_{i=1}^{\mu} w_i(\|\mathbf{Y}_t^i\|^2 - n)\right)\right)}_{\eta^\star(\mathbf{Z}_t,\mathbf{Y}_t)}}$$



# Linear convergence of xNatural Evolution Strategies

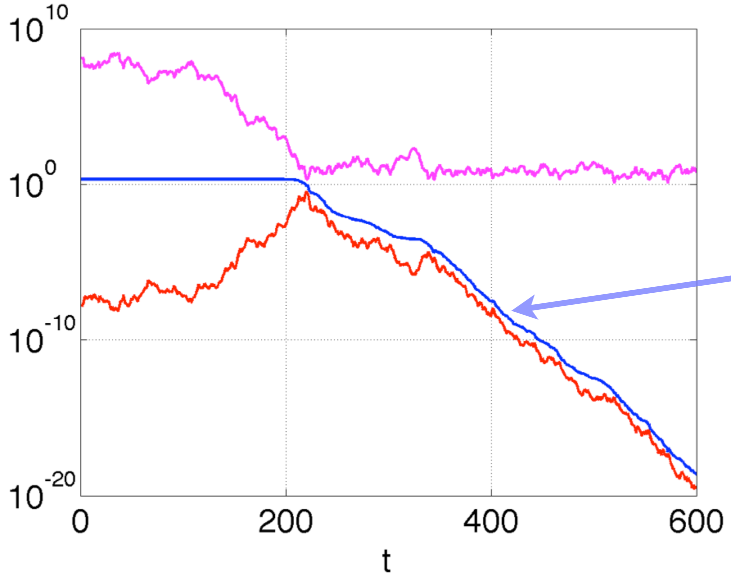
$$\frac{1}{t}\ln\frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} = \frac{1}{t}\sum_{k=0}^{t-1}\ln\frac{\|\mathbf{X}_{k+1}\|}{\|\mathbf{X}_k\|} = \frac{1}{t}\sum_{k=0}^{t-1}\ln\frac{\|\mathbf{X}_{k+1}\|}{\sigma_{k+1}}\frac{\sigma_k\eta^\star(\mathbf{Z}_k,\mathbf{Y}_k)}{\|\mathbf{X}_k\|}$$

$$= \frac{1}{t}\sum_{k=0}^{t-1}\ln\frac{\|\mathbf{Z}_{k+1}\|}{\|\mathbf{Z}_k\|}\eta^\star(\mathbf{Z}_k,\mathbf{Y}_k)$$

Law of Large Numbers for $\mathbf{Z}_t$

if $\mathbf{Z}_t$ stable enough

$$\int E[\ln\eta^\star(\mathbf{z},\mathbf{y})]\pi(d\mathbf{z})$$




$\pi$ stationary distrib. of $\mathbf{Z}_t$

# Linear convergence of xNatural Evolution Strategies

$$\frac{1}{t}\ln\frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} = \frac{1}{t}\sum_{k=0}^{t-1}\ln\frac{\|\mathbf{X}_{k+1}\|}{\|\mathbf{X}_k\|} = \frac{1}{t}\sum_{k=0}^{t-1}\ln\frac{\|\mathbf{X}_{k+1}\|}{\sigma_{k+1}}\frac{\sigma_k\eta^\star(\mathbf{Z}_k,\mathbf{Y}_k)}{\|\mathbf{X}_k\|}$$

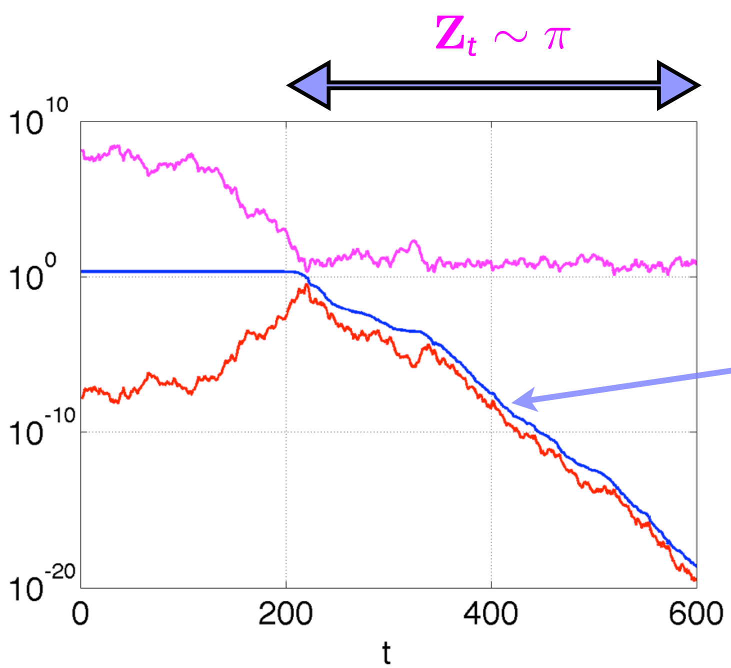
$$= \frac{1}{t}\sum_{k=0}^{t-1}\ln\frac{\|\mathbf{Z}_{k+1}\|}{\|\mathbf{Z}_k\|}\eta^\star(\mathbf{Z}_k,\mathbf{Y}_k)$$

LLN for $\mathbf{Z}_t$

if $\mathbf{Z}_t$ stable enough

$$\int E[\ln\eta^\star(\mathbf{z},\mathbf{y})]\pi(d\mathbf{z})$$

$\pi$ stationary distrib. of $\mathbf{Z}_t$




$\mathbf{Z}_t \sim \pi$ at a geometric rate ind. starting point

# Overview

Scale-invariance & step-size adaptive ESs

Construction of (homogeneous) normalized MC

Stability of normalized chain

Sufficient condition for geometric ergodicity

*step-size increase on linear functions*

# Step-size Adaptive ESs
# Definition

$\mathcal{S}el_{(\mathbf{x},\sigma)}$: ranked-based selection

$\mathcal{G} : (\mathbb{R}^n \times \mathbb{R}^+) \times \mathbb{R}^{n\times\mu} \mapsto (\mathbb{R}^n \times \mathbb{R}^+)$: update function

$$(\mathbf{X}_{t+1}, \sigma_{t+1}) = \mathcal{G}\left((\mathbf{X}_t, \sigma_t), \mathcal{S}el_{(\mathbf{X}_t,\sigma_t)}(\mathbf{U}_t)\right)$$

$\mathbf{U}_t$ i.i.d.

$$\mathbf{X}_{t+1} = \mathcal{G}_1\left((\mathbf{X}_t, \sigma_t), \mathcal{S}el_{(\mathbf{X}_t,\sigma_t)}(\mathbf{U}_t)\right)$$

$$\sigma_{t+1} = \mathcal{G}_2\left(\sigma_t, \mathcal{S}el_{(\mathbf{X}_t,\sigma_t)}(\mathbf{U}_t)\right)$$

Algorithms covered:

$(\mu/\mu_w, \lambda)$-ES with CMA-ES step-size adaptation (without cumulation)
xNES, self-adaptive ES

$(1+1)$ with 1/5 success rule

# Ranked-based selection
# Invariances

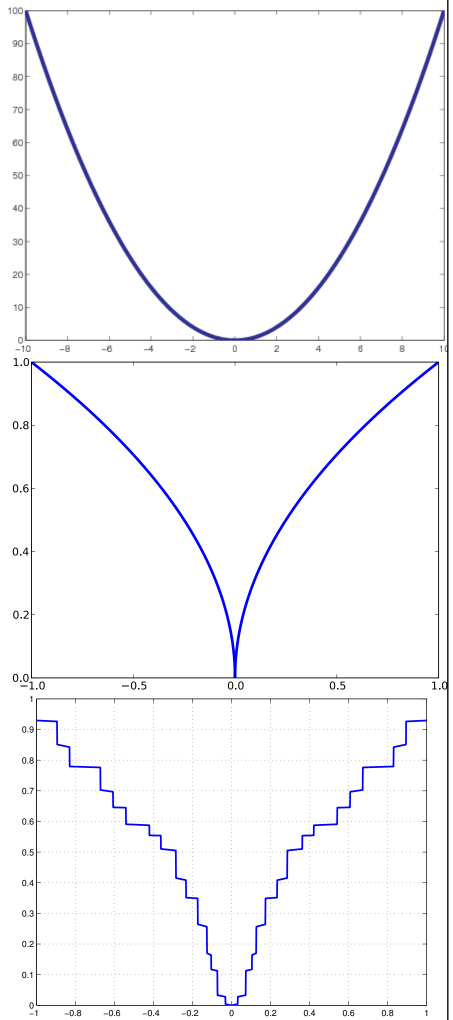
## Invariance to monotonic transformations

$$\mathcal{S}el^{f}_{(\mathbf{X}_t,\sigma_t)}(\mathbf{U}_t) = \mathcal{S}el^{g\circ f}_{(\mathbf{X}_t,\sigma_t)}(\mathbf{U}_t)$$

$g : \mathbb{R} \to \mathbb{R}$ increasing

$$f(\mathbf{X}_t + \sigma_t \mathbf{U}_t^{1:\lambda}) \leq \ldots \leq f(\mathbf{X}_t + \sigma_t \mathbf{U}_t^{\mu:\lambda})$$

$$g \circ f(\mathbf{X}_t + \sigma_t \mathbf{U}_t^{1:\lambda}) \leq \ldots \leq g \circ f(\mathbf{X}_t + \sigma_t \mathbf{U}_t^{\mu:\lambda})$$



## Scale-invariance

$$\mathcal{S}el^{f(\mathbf{x})}_{(\mathbf{X}_t,\sigma_t)}(\mathbf{U}_t) = \mathcal{S}el^{f(\alpha\mathbf{x})}_{(\frac{\mathbf{X}_t}{\alpha},\frac{\sigma_t}{\alpha})}(\mathbf{U}_t)$$

because $\quad f\left(\alpha\left(\frac{\mathbf{X}_t}{\alpha} + \frac{\sigma_t}{\alpha}\mathbf{U}_t^i\right)\right) = f(\mathbf{X}_t + \sigma_t \mathbf{U}_t^i)$

# Scale-invariance

*the algorithm has no intrinsic notion of scale*

**Morphism:**

$$\Phi : \alpha \in (\mathbb{R}^+, \cdot) \mapsto \Phi(\alpha) \qquad\qquad \Phi(\alpha)(\mathbf{x}, \sigma) \mapsto \left(\frac{\mathbf{x}}{\alpha}, \frac{\sigma}{\alpha}\right)$$

$$\Phi(\alpha_1 \cdot \alpha_2) = \Phi(\alpha_1) \circ \Phi(\alpha_2)$$

A step-size adaptive ES is scale-invariant if it satisfies the following commutative diagram

$$\begin{array}{ccc}
(\mathbf{X}_t, \sigma_t) & \xrightarrow{\mathcal{G}(., \mathcal{S}el^{f(\mathbf{x})}(.))} & (\mathbf{X}_{t+1}, \sigma_{t+1}) \\
\Phi(\alpha) \downarrow \uparrow \Phi(1/\alpha) = \Phi^{-1}(\alpha) & & \Phi(\alpha) \downarrow \uparrow \Phi^{-1}(\alpha) \\
(\underbrace{\mathbf{X}'_t}_{\frac{\mathbf{X}_t}{\alpha}}, \underbrace{\sigma'_t}_{\frac{\sigma_t}{\alpha}}) & \xrightarrow{\mathcal{G}(., \mathcal{S}el^{f(\alpha\mathbf{x})}(.))} & (\mathbf{X}'_{t+1}, \sigma'_{t+1})
\end{array}$$

# Scale-invariance (cont.)

A step-size adaptive ES is scale-invariant iff for all $\alpha > 0$, all $\mathbf{x}$, $\mathbf{y}$, $\sigma$

$$\mathcal{G}_1((\mathbf{x},\sigma),\mathbf{y}) = \alpha \mathcal{G}_1\left(\left(\frac{\mathbf{x}}{\alpha},\frac{\sigma}{\alpha}\right),\mathbf{y}\right) \quad \text{for all}$$

$$\mathcal{G}_2(\sigma,\mathbf{y}) = \alpha \mathcal{G}_2\left(\frac{\sigma}{\alpha},\mathbf{y}\right)$$

*homogeneity, scalability*

## Examples of Scale-invariant algorithms:

$(\mu/\mu_w,\lambda)$-ES with CMA-ES step-size adaptation (without cumulation)
xNES, self-adaptive ES

$(1+1)$ with 1/5 success rule

Update function for xNES:

$$\mathcal{G}((\mathbf{x},\sigma),\mathbf{y}) = \begin{pmatrix} \mathbf{x}+\sigma c_m \sum_{i=1}^{\mu} w_i \mathbf{y}^i \\ \sigma \exp\left(\frac{c_\sigma}{2n} \sum_{i=1}^{\mu} w_i (\|\mathbf{y}^i\|^2 - n)\right) \end{pmatrix} .$$

# Scaling-invariant Functions

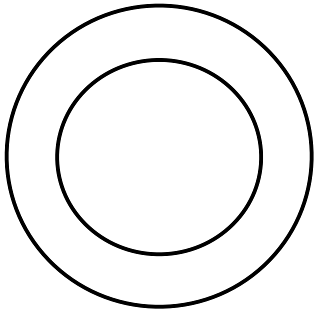
**Definition.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is scaling invariant (around 0) if for all $\sigma > 0$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$*

$$f(\sigma\mathbf{x}) \leq f(\sigma\mathbf{y}) \Leftrightarrow f(\mathbf{x}) \leq f(\mathbf{y})$$

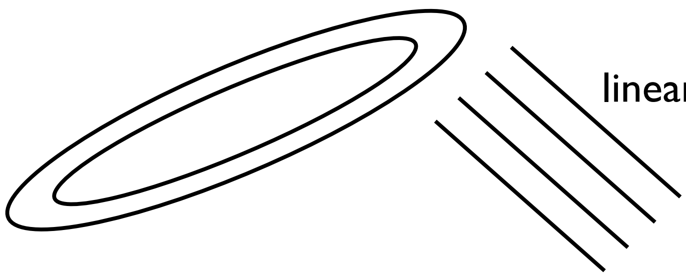
Implies $f(\sigma\mathbf{x}) = f(\sigma\mathbf{y}) \Leftrightarrow f(\mathbf{x}) = f(\mathbf{y})$

**Examples**: convex sublevel sets — non-convex sublevel sets

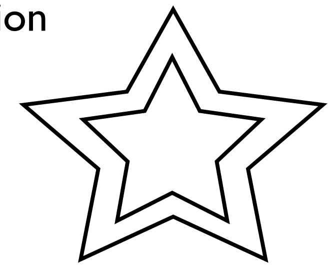
$f(\mathbf{x}) = g(\|\mathbf{x}\|)$ with $\|\ \|$ norm on $\mathbb{R}^n$, $g \in \mathcal{M}$
$\mathcal{M} = \{g : \mathbb{R} \to \mathbb{R}$ monotonically increasing$\}$

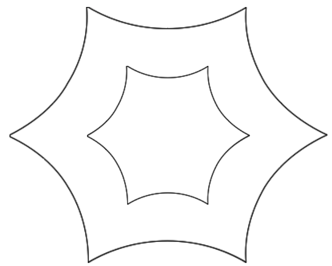


$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}}$




$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T H\mathbf{x}}$, $H$ SDP





# Scale-invariance on Scaling-invariant Functions ≡ homogeneous Markov chain

**Proposition 1** *Consider a scaling-invariant objective function $f$ optimized by a scale-invariant adaptive step-size ES, i.e.* $(\mathbf{X}_{t+1}, \sigma_{t+1}) = \mathcal{G}((\mathbf{X}_t, \sigma_t), \mathcal{S}el_{(\mathbf{X}_t,\sigma_t)}(\mathbf{U}_t))$*. Then* $\mathbf{Z}_t = \mathbf{X}_t/\sigma_t$ *is an homogeneous Markov Chain with update equation determined by*

$$\mathbf{Y}_t := Sel^f_{\mathbf{Z}_t}(\mathbf{U}_t) = \mathcal{S}el^f_{(\mathbf{Z}_t,1)}(\mathbf{U}_t) \text{ for all } \mathbf{U}_t \tag{1}$$

*and*

$$\mathbf{Z}_{t+1} = \frac{\mathcal{G}_1((\mathbf{Z}_t, 1), \mathbf{Y}_t)}{\mathcal{G}_2(1, \mathbf{Y}_t)} = \frac{\mathcal{G}_1((\mathbf{Z}_t, 1), \mathbf{Y}_t)}{\eta^\star(\mathbf{Y}_t)} \tag{2}$$

$$\mathbf{Z}_{t+1} = \frac{\mathbf{X}_{t+1}}{\sigma_{t+1}} = \frac{\mathcal{G}_1((\mathbf{X}_t, \sigma_t), \mathcal{S}el^f_{(\mathbf{X}_t,\sigma_t)}(\mathbf{U}_t))}{\mathcal{G}_2(\sigma_t, \mathcal{S}el^f_{(\mathbf{X}_t,\sigma_t)}(\mathbf{U}_t))} = \frac{\sigma_t \mathcal{G}_1((\mathbf{X}_t/\sigma_t, 1), \mathcal{S}el^{f(\sigma_t .)}_{(\mathbf{X}_t/\sigma_t,1)}(\mathbf{U}_t))}{\sigma_t \mathcal{G}_2(1, \mathcal{S}el^{f(\sigma_t .)}_{(\mathbf{X}_t/\sigma_t,1)}(\mathbf{U}_t))}$$

$$= \frac{\mathcal{G}_1((\mathbf{X}_t/\sigma_t, 1), \mathcal{S}el^{f(.)}_{(\mathbf{X}_t/\sigma_t,1)}(\mathbf{U}_t))}{\mathcal{G}_2(1, \mathcal{S}el^{f(.)}_{(\mathbf{X}_t/\sigma_t,1)}(\mathbf{U}_t))} .$$

**Reminder: chain for xNES**

$$\mathbf{Y}_t = \mathcal{S}el_{(\frac{\mathbf{X}_t}{\sigma_t},1)}(\mathbf{U}_t)$$

$$\mathbf{Z}_{t+1} = \frac{\mathbf{Z}_t + c_m \sum_{i=1}^{\mu} w_i \mathbf{Y}_t^i}{\exp\left(\frac{c_\sigma}{2n}\left(\sum_{i=1}^{\mu} w_i(\|\mathbf{Y}_t^i\|^2 - n)\right)\right)}$$

# Overview

Scale-invariance & step-size adaptive ESs

Construction of (homogeneous) normalized MC

Stability of normalized chain

Sufficient condition for geometric ergodicity

*step-size increase on linear functions*

# Stability of Normalized Markov Chain

Assumptions

$\bar{f} = g \circ f$ where $f$ is $C^1$ and homogeneous with degree $\gamma$ with $f(\mathbf{x}) > 0, \mathbf{x} \neq 0$

$$f(\sigma \mathbf{x}) = \sigma^{\gamma} f(\mathbf{x})$$

$\Rightarrow \mathbf{x}^{\star}$ unique in zero (W.L.G.)

$\mathbf{Z}_t$ is irreducible w.r.t. Lebesgue measure:

$$\forall A \text{ with } \mu_{\mathrm{Leb}}(A) > 0, \forall \mathbf{z}, \exists t_0 \text{ such that } P^{t_0}(\mathbf{z}, A) = \Pr(\mathbf{Z}_{t_0} \in A | \mathbf{Z}_0 = \mathbf{z}) > 0$$

$\mathbf{Z}_t$ is strongly aperiodic and compact are small sets

Small set: set $C$ such that $\exists \delta, t > 0$ a non-trivial measure $\nu_t()$: $P^t(\mathbf{z}, .) \geq \delta \nu_t(.) \ \ \forall \mathbf{z} \in C$

Strong aperiodicity: if $\exists$ a $\nu_1$ small set $C$ with $\nu_1(C) > 0$

Different proof technique depending on the algorithm:

★difficult to prove for "derandomized" algorithms (xNES, isotropic CMA without cumulation)

★easy for (1+1) with success rule or comma ES with self adaptation

# Stability of Normalized Markov Chain
# Geometric Ergodicity

We prove a sufficient condition for geometric drift, i.e. find $V \geq 1$ such that there exists $\alpha_0 < 1$

$$E[V(\mathbf{Z}_{t+1})|\mathbf{Z}_t = \mathbf{z}] \leq \alpha_0 V(\mathbf{z}), \quad \mathbf{z} \text{ outside a compact set}$$

$\Rightarrow$ existence of an invariant probability measure $\pi$:

$$\pi(A) = \int \pi(d\mathbf{z}) P(\mathbf{z}, A)$$

Equivalent to existence of $r > 1$ and $R < \infty$ such that for any starting point in the set $S_V = \{\mathbf{z} : V(\mathbf{z}) < \infty\}$

$$\sum_t r^t \| P^t(\mathbf{z}_0, .) - \pi \|_V \leq RV(\mathbf{z}_0) \tag{1}$$

where $\|\nu\|_V = \sup_{g:|g| \leq V} |\nu(g)|$

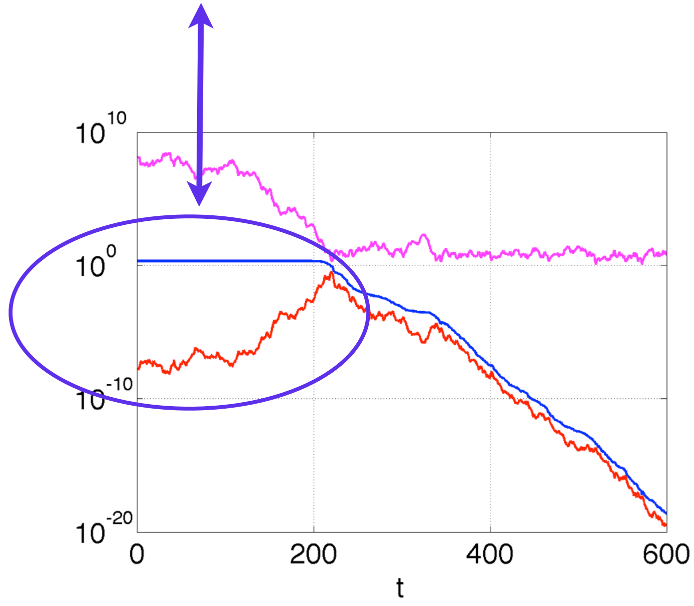
# Sufficient Condition for Geometric Drift
# Non elitist variants

If $\bar{f} = g \circ f$ where $f$ is $C^1$ and homogeneous with degree $\gamma$, $f(\mathbf{x}) > 0$ for $\mathbf{x} \neq 0$, the function $V(\mathbf{z}) = 1 + f^{\gamma'}(\mathbf{z})$ satisfy a geometric drift condition if

$$E\left[\frac{1}{{\eta^\star_{\text{linear}}}^{\gamma'}}\right] < 1$$

*increase of step-size on linear functions*

Remark: Drift different for (1+1), 0 is outside the domain
Need to control drift negative close to zero




Step-size increase on linear functions satisfied by
(1+1)-1/5 success rule, xNES, CSA and self-adaptation for lambda >2

Not satisfied by cross-entropy, EMNA

# Linear Convergence

Under the following assumptions, if sufficient condition for geometric drift satisfied:

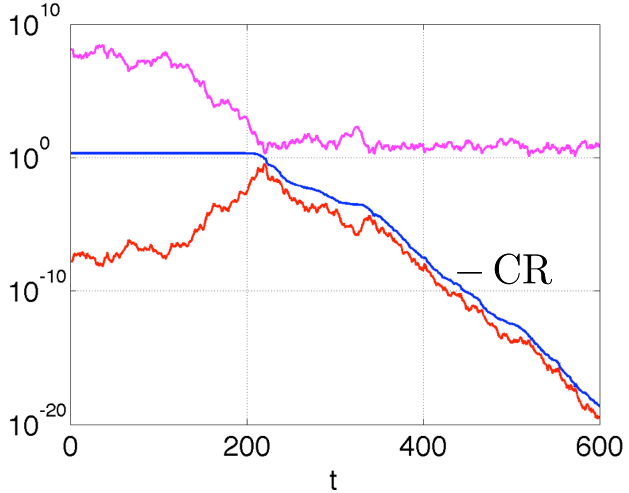
$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} = \int \ln \eta^\star(\mathbf{y}) q(\mathbf{z},\mathbf{y}) d\mathbf{y} \pi(d\mathbf{z}) =: -\mathrm{CR} \quad \text{a.s.} \qquad + \text{CLT}$$

density of $Sel_{\mathbf{z}}(\mathbf{y})$

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0} = -\mathrm{CR} \quad \text{a.s.}$$

From any starting point $(\mathbf{X}_0, \sigma_0)$

$$\lim_{t\to\infty} E_{(\mathbf{X}_0,\sigma_0)} \ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|} = -\mathrm{CR}$$




There exists $r > 1$, such that from any starting point $(\mathbf{X}_0, \sigma_0)$

$$\left| E_{\frac{\mathbf{x}_0}{\sigma_0}} \ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|} - (-CR) \right| \leq \frac{RV(\mathbf{X}_0/\sigma_0)}{r^t}$$

*consequence of geometric ergodicity*

# Open questions

- How much can we generalize those results
  - noisy objective function
  - cumulation for step-size
- Complex proof to prove irreducibility w.r.t. Lebesgue measure, aperiodicity with "derandomized" algorithms?
- proof of $\lim_{n \to \infty} n \, \mathrm{CR} = \mathrm{CR}_\infty$