GECCO '14, Jul 12-16 2014, Vancouver, BC, Canada
ACM 978-1-4503-2881-4/14/07.
http://dx.doi.org/10.1145/2598394.2605355

Information Geometry in Evolutionary Computation

Luigi Malagò[1] and Tobias Glasmachers[2]
[1] Università degli Studi di Milano, [2] Ruhr-Universität Bochum

GECCO Tutorial July 13, 2014

Optimization by Population-based EC

In EC, a common approach to find the optimum of a function is to evolve iteratively a population of candidate solutions by applying different operators which ensures a tradeoff between

- exploitation (e.g., selective pressure)
- exploration (e.g., variation, genetic diversity)

Many Evolutionary Algorithms (EAs) follow such paradigm, and can be defined as population-based, among others we have

- Genetic Algorithms (GAs)
- Ant Colony Optimization (ACO)
- Particle Swarm Optimization (PSO)
- Evolution Strategies (ES)
- ...

Population-based EC: Genetic Algorithms

Let us introduce some notation

- Ω the search space
- $f : \Omega \to \mathbb{R}$ the function to be optimized
- $\mathcal{P}_t = \{x \in \Omega\}$ a population of individuals at time t
- $\mathcal{P}_0$ the initial (e.g., random) population

The basic iteration of a naïve GA can be described as

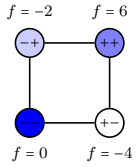
$$\mathcal{P}_t \xrightarrow{\text{selection}} \mathcal{P}_t^s \xrightarrow{\text{crossover}} \mathcal{P}_t^c \xrightarrow{\text{mutation}} \mathcal{P}_{t+1}$$

A Toy Example with 2 Binary Variables

Example: $\Omega = \{-1, 1\}^2$, $f(x) = x_1 + 2x_2 + 3x_1x_2$

$\mathcal{P}_0$		$f(x)$
1	-1	-4
-1	1	-2
1	-1	-4
-1	-1	0
1	1	6
-1	1	-2
-1	1	-2
1	1	6

Hypercube



GA: $\mathcal{P}_t \xrightarrow{\text{truncation selection}} \mathcal{P}_t^s \xrightarrow{\text{1-point crossover}} \mathcal{P}_t^c \xrightarrow{\text{mutation}} \mathcal{P}_{t+1}$

From Populations to Probability Distributions

A population $\mathcal{P}$ can be seen as a sample i. i. d. $\sim$ with respect to some probability distribution p in the probability simplex Δ

Let N denote the sample size. We have

$$\mathcal{P} \xrightarrow{\text{estimation}} \hat{p} \qquad \mathcal{P} \xleftarrow{\text{sampling}} p$$

For unbiased estimators and $N \to \infty$ (infinite population size analysis)

$$\mathcal{P} \xrightarrow{\text{estimation}} p$$

Such approach is at the basis of the theoretical analysis of Vose (1999) on SGA

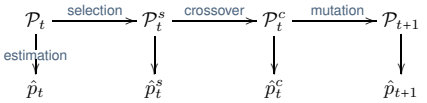
We can describe genetic operators as maps from the probability simplex to the the probability simplex itself, e.g.,

$$\text{selection} : \Delta \ni p \mapsto p^s \in \Delta$$

From Hypercubes to Probability Simplices

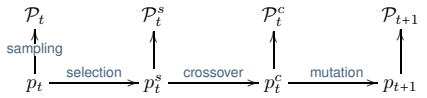
A run of a population-based EA identifies a sequence of points in Δ

Single run of the GA:



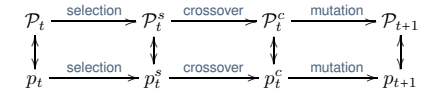
A run can be seen as a realization of the expected behavior of the algorithm

Expected behavior of the GA:



For unbiased estimators and $N \to \infty$, the map is one-to-one

Infinite population size analysis of the GA:

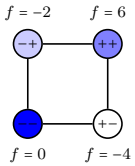


A Toy Example with 2 Binary Variables (cont.)

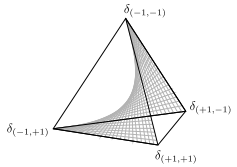
Example: $\Omega = \{-1, 1\}^2$, $f(x) = x_1 + 2x_2 + 3x_1x_2$

$\mathcal{P}_0$		$f(x)$
1	-1	-4
-1	1	-2
1	-1	-4
-1	-1	0
1	1	6
-1	1	-2
-1	1	-2
-1	1	6

Hypercube



Probability simplex



GA: $p_t \xrightarrow{\text{truncation selection}} p_t^s \xrightarrow{\text{1-point crossover}} p_t^c \xrightarrow{\text{mutation}} p_{t+1}$

Model-Based Optimization

In model-based optimization, the search for the optimum of f is performed explicitly in the space of probability distributions.

By updating the parameters of a probability distribution, iterative algorithms generate sequences of distributions.

Candidate solutions for the optimum of f can be obtained by sampling.

A model-based algorithm is expected to produce converging and minimizing sequence, however

- Which statistical model to choose?
- How to generate such sequence?

Examples of Model-based Algorithms

8

Evolutionary computation

- EDAs (Larrañaga and Lozano, 2002), DEUM framework (Shakya et al., 2005)

Gradient descent

- SGD (Robbins and Monro, 1951), CMA-ES (Hansen and Ostermeier, 2001), NES (Wierstra et al., 2008), SNGD (M. et al., FOGA 2011), IGO (Ollivier et al., 2011),

Boltzmann distribution and Gibbs sampler (Geman and Geman, 1984)

Simulated Annealing and Boltzmann Machines (Aarts and Korst, 1989)

The Cross-Entropy method (Rubinstein, 1997)

LP relaxation in pseudo-Boolean optimization (Boros and Hammer, 2001)

Methods of Moments (Meziat et al., 2001)

Model-based EC: Estimation of Distribution

9

In Estimation of Distribution Algorithms (EDAs) a statistical model is introduced to model interactions among variables of f

Genetic operators (crossover and mutation in GAs) are replaced by statistical operators such as estimation and sampling

Let us introduce some more notation

- $p(x,\theta)$ a probability distribution over Ω parametrized by θ
- $\mathcal{M} = \{p(x,\theta) : \theta \in \Theta\}$ a parametric statistical model

The basic iteration of an EDA can be described as

$$\mathcal{P}_t \xrightarrow{\text{selection}} \mathcal{P}_t^s \xrightarrow[\text{(model selection)}]{\text{estimation}} p_t \xrightarrow{\text{sampling}} \mathcal{P}_{t+1} \qquad p_t \in \mathcal{M}$$

From a model-based perspective, we have

$$p_t \xrightarrow{\text{sampling}} \mathcal{P}_{t+1} \xrightarrow{\text{selection}} \mathcal{P}_{t+1}^s \xrightarrow[\text{(model selection)}]{\text{estimation}} p_{t+1}$$

Estimation of Distribution Algorithms

10

Suppose we chose $\mathcal{M}$ to be the independence model for $x = (x_1, x_2)$

$$\mathcal{M} = \{p : p(x) = p_1(x_1)p_2(x_2)\}$$

and we parametrize it using marginal probabilities $p_i(x_i) = \mathbb{P}(X_i = x_i)$

Estimation of the parameters given a sample is obtained with a maximum likelihood estimator, i.e., we count occurrences

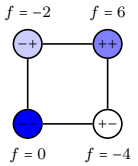
Back to the Toy Example with 2 Binary Variables

11

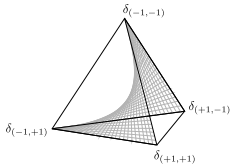
Example: $\Omega = \{-1,1\}^2$, $f(x) = x_1 + 2x_2 + 3x_1x_2$, independence model

$\mathcal{P}_0$		$f(x)$
1	-1	-4
-1	1	-2
1	-1	-4
-1	-1	0
1	1	6
-1	1	-2
-1	1	-2
-1	1	6

Hypercube



Probability simplex



EDA: $\mathcal{P}_t \xrightarrow{\text{truncation selection}} \mathcal{P}_t^s \xrightarrow{\text{estimation}} \mathcal{P}_t^c \xrightarrow{\text{sampling}} \mathcal{P}_{t+1}$

Expected Fitness Landscape

In model-based optimization, the search for the optimum in Ω is guided by a search in the space of the probability distributions.

A natural choice is to optimize the expected value of f over $\mathcal{M}$,

$$\mathbb{E}_p[f] : \mathcal{M} \to \mathbb{R}$$

which can be expressed as a function of ξ, given a parameterization for $p(x, \xi) \in \mathcal{M}$, i.e.,

$$\xi \mapsto \mathbb{E}_\xi[f]$$

Equivalent Parameterizations for the Independence Model $p(x) = p_1(x_1)p_2(x_2)$

Marginal probabilities $\mu = (\mu_1, \mu_2) \in [0,1]^2$

$p_i(x_i) = \mathbb{P}(X_i = x_i) \qquad p_i(1) = \mu_i \qquad p_i(-1) = 1 - \mu_i$

$p_i(x_i) = (2\mu_i x_i - x_i + 1)/2$

$\mathbb{E}_\mu[f] = \sum_{x \in \Omega} f(x) p_1(x_1) p_2(x_2) = -4\mu_1 - 2\mu_2 + 12\mu_1\mu_2$

Natural parameters $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ of the Exponential Family

$$p(x) = \exp\{\theta_1 x_1 + \theta_2 x_2 - \psi(\theta)\} \qquad \psi(\theta) = \ln \sum_{x \in \Omega} \exp\{\theta_1 x_1 + \theta_2 x_2\} = \ln Z(\theta)$$

$$p_i(x_i) = \frac{e^{\theta_i x_i}}{e^{\theta_i} + e^{-\theta_i}}$$

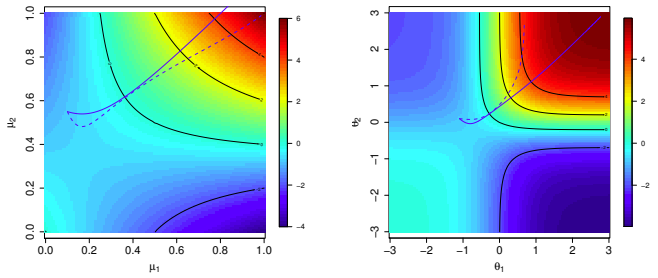
$$\mathbb{E}_\theta[f] = (-4e^{\theta_1 - \theta_2} - 2e^{-\theta_1 + \theta_2} + 6e^{\theta_1 + \theta_2})/Z(\theta)$$

The mapping between the two parameterizations is one-to-one for $p(x) > 0$

$$\theta_i = (\ln(\mu_i) - \ln(1 - \mu_i))/2 \qquad \mu_i = \frac{e^{\theta_i}}{e^{\theta_i} + e^{-\theta_i}}$$

Gradient Flows on the Independence Model

Marginal probabilities μ | Natural parameters θ



Gradient flows $\nabla \mathbb{E}_\xi[f]$ appear to depend on the parameters and do not always converge to the expected distribution

[Remark] We supposed the geometry of $\mathcal{M}$ to be Euclidean

Summary of the Intro

- There exists a common geometric framework to describe population- and model-based EAs
- Iterative algorithm generate sequences of distributions which can be compared to the gradient flow of $\mathbb{E}_p[f]$
- The choice of the statistical model and of the parameterization plays an important role
- The Euclidean geometry does not appear to be the proper geometry for $\mathcal{M}$

We need a more general mathematical framework, able to deal with non-Euclidean geometries, to define a unifying perspective on model-based optimization

Outline

Part I

- Stochastic relaxation of the fitness functions
- Introduction to the Information Geometry of statistical models
- Natural Gradient
- Fitness landscape and model selection

Part II

- Natural Evolution Strategies
- Stochastic Natural Gradient Descent
- Information Geometric Optimization
- Convergence theorems
- Practical performance

Part I

Stochastic Relaxation of f

Consider the following optimization problem

$$\text{(P)} \qquad \min_{x \in \Omega} f(x)$$

We define stochastic relaxation (SR) of f the function

$$F : p \mapsto \mathbb{E}_p[f]$$

Given a statistical model $\mathcal{M} = \{p(x)\}$, we look for the solution of (P) by generating minimizing sequences $\{p_t\}$ in $\mathcal{M}$ for $F(p)$

Let ξ be a parameterization for $\mathcal{M}$, i.e., $\mathcal{M} = \{p(x;\xi) : \xi \in \Xi\}$, the SR can be expressed as

$$\text{(SR)} \qquad \min_{\xi \in \Xi} F(\xi)$$

We move the search to the space of probability distribution

The parameters $\xi \in \Xi$ become the variables of the relaxed problem

Equivalence of (P) and (SR)

Let us introduce some notation

- $x^* \in \Omega^* = \arg\min_{x \in \Omega} f(x)$ the global optima of f
- $p_* \in \mathcal{M}^* = \arg\min_{p \in \overline{\mathcal{M}}} F(\xi)$ the global optima of F
- $\overline{\mathcal{M}}$ the topological closure of $\mathcal{M}$, i.e., $\mathcal{M}$ together all limit distributions of sequences $\{p_t\} \in \mathcal{M}$

Candidate solutions for (P) can be sampled by solutions of the (SR)

Distributions in $\mathcal{M}^*$ have reduced support and for discrete Ω corresponds to faces of Δ

(P) and (SR) and equivalent if and only if from a solution of (SR) we can sample points in Ω^* with $\mathbb{P}(X = x^*) = 1$

A sufficient condition is the inclusion of the Dirac distributions δ_{x^*} in $\overline{\mathcal{M}}$, i.e., there exists a sequence $\{p_t\} \in \mathcal{M}$ such that

$$\lim_{t \to \infty} F(p_t) = \min_{x \in \Omega} f(x)$$

The Exponential Family

20

In the following, we consider models in the Exponential Family $\mathcal{E}$

$$p(x,\theta) = \exp\left(\sum_{i=1}^{m} \theta_i T_i(x) - \psi(\theta)\right)$$

- sufficient statistics $T = (T_1(x), \ldots, T_m(x))$
- natural parameters $\theta = (\theta_1, \ldots, \theta_m) \in \Theta$
- log-partition function $\psi(\theta)$

Several statistical models belong to the exponential family, both in the continuous and discrete case

Information Geometry

21

The geometry of statistical models is not Euclidean, and we need tools from differential geometry to properly describe notions such as tangent vectors, shortest paths and distances between distributions

Information Geometry (IG), first propose by Amari, consists of the study of statistical models as manifolds of distributions endowed with the Fisher information matrix

Characterization of the Tangent Space of $\mathcal{E}$

22

A statistical model can be modeled as a manifold of distributions by introducing an affine chart in p such that any density q is locally represented w.r.t. to the reference measure p, i.e., $\frac{q}{p} - 1$

The tangent space at each point p is defined by

$$\mathsf{T}_p = \{v : \mathbb{E}_p[v] = 0\}$$

Consider a curve $p(\theta)$ such that $p(0) = p$, then $\frac{\dot{p}(\theta)}{p} \in \mathsf{T}_p$

In a moving coordinate system, tangent (velocity) vectors in $\mathsf{T}_{p(\theta)}$ to the curve are given by logarithmic derivatives

$$\frac{\dot{p}(\theta)}{p(\theta)} = \frac{d}{d\theta} \log p(\theta)$$

Characterization of the Tangent Space of $\mathcal{E}$

23

The one dimensional model

$$p(\theta) = \exp\{\theta T - \psi(\theta)\}$$

is a curve in the manifold, with tangent vector

$$T - \frac{d}{d\theta}\psi(\theta)$$

On the other side, given a vector field, at each p we have a vector $U(p)$ tangent to some curve, we obtain a differential equation

$$\frac{d}{d\theta} \log p(\theta) = U(p),$$

whose solution is a one dimensional model in $\mathcal{E}$

Geometry of the Exponential Family

In case of a finite sample space $\mathcal{X}$, we have

$$p(x;\theta) = \exp\left(\sum_{i=1}^{k} \theta_i T_i(x) - \psi(\theta)\right) \quad \theta \in \mathbb{R}^k$$

and

$$\mathsf{T}_\theta = \left\{ v : v = \sum_{i=1}^{k} a_i (T_i(x) - \mathbb{E}_\theta[T_i]), a_i \in \mathbb{R} \right\}$$

Since $\nabla \mathbb{E}_\theta[f] = \mathrm{Cov}_\theta(f, T)$, if $f \in \mathsf{T}_p$ the steepest direction is given by $f - \mathbb{E}_\theta[f]$ otherwise we take the projection of f onto T_p

$$\hat{f} = \sum_{i=1}^{k} \hat{a}_i (T_i(x) - \mathbb{E}_\theta[T_i]),$$

and obtain $\hat{f}$ by solving a system of linear equations

Geometry of Statistical Models

Since $f - \hat{f}$ is orthogonal to T_p

$$\mathbb{E}_\theta[(f - \hat{f}_\theta)(T - \mathbb{E}_\theta[T])] = \mathrm{Cov}_\theta(f - \hat{f}_\theta, T) = 0,$$

from which we obtain, for $i = 1, \ldots, k$,

$$\mathrm{Cov}_\theta(f, T_i) = \mathrm{Cov}_\theta(\hat{f}_\theta, T_i) = \sum_{j=1}^{k} \hat{a}_j \, \mathrm{Cov}_\theta(T_j, T_i)$$

As the Hessian matrix of $\psi(\theta)$ is invertible, we have

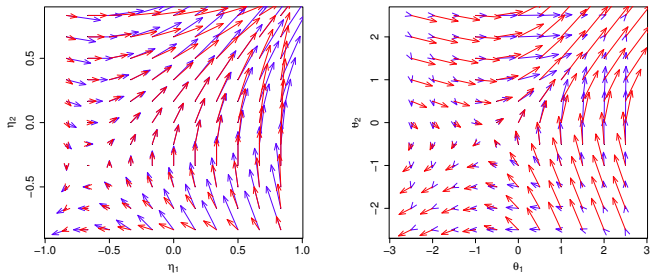
$$\hat{a} = [\mathrm{Cov}_\theta(T_i, T_j)]^{-1} \, \mathrm{Cov}_\theta(f, T) = I(\theta)^{-1} \nabla F(\theta)$$

In case $f \in \mathrm{Span}\{T_1, \ldots, T_k\}$, then $\hat{f}_\theta = f$

By taking projection to T_p, we obtained the natural gradient $\widetilde{\nabla} F$, i.e., the gradient evaluated w.r.t. the Fisher information metric

Vanilla Gradient vs Natural Gradient

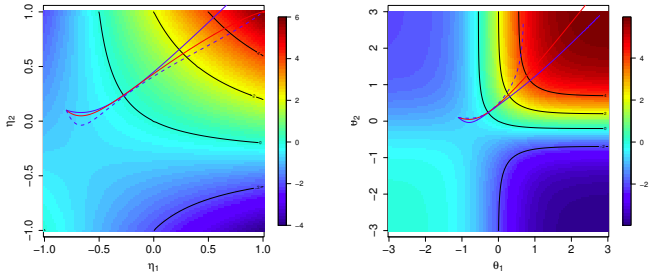
Expectation parameters η ($\lambda = 0.075$) Natural parameters θ ($\lambda = 0.15$)



Vanilla gradient ∇F vs Natural gradient $\widetilde{\nabla} F$

Vanilla Gradient vs Natural Gradient

Expectation parameters η Natural parameters θ



Vanilla gradient ∇F vs Natural gradient $\widetilde{\nabla} F$

There exist two basins of attraction, however natural gradient converges to δ_{x^*} for distributions close enough to the optimum

The Gibbs Distribution (Hwang, 1980; Geman and Geman, 1984)

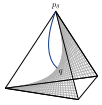
28

- The Gibbs or Boltzmann distribution is the following one dimensional exponential family

$$p(x;\beta) = \frac{qe^{-\beta f}}{\mathbb{E}_q[e^{-\beta f}]}, \quad \beta > 0$$

- The set of distributions is not weakly closed

$$\lim_{\beta\to 0} p(x;\beta) = q$$
$$\lim_{\beta\to\infty} p(x;\beta) = p_\delta$$

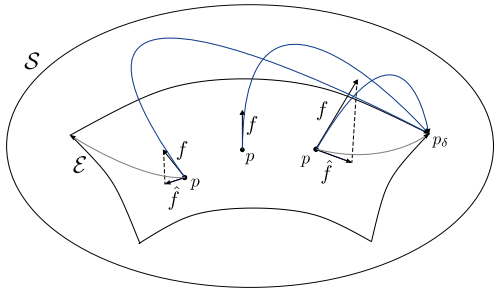


- The limit p_δ is the uniform distribution over the minima of f and since $\nabla \mathrm{E}_\beta[f] = -\mathrm{Var}_\beta(f) < 0$, $\mathrm{E}_\beta[f]$ decreases monotonically

Evaluating the partition function is computationally infeasible

The Big Picture

29



If $f \notin \mathrm{T}_p$, the projection $\hat{f}$ may vanish, and local minima may appear

Choice of $\mathcal{M}$

30

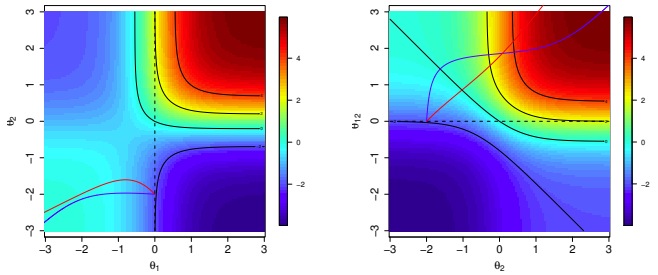
The choice of the statistical model $\mathcal{M}$ determines the landscape of F

Independence model, $\theta = (\theta_1, \theta_2, 0)$

$p(x) = \exp\{\theta_1 x_1 + \theta_2 x_2 - \psi(\theta)\}$

Exponential family, $\theta = (0, \theta_2, \theta_{12})$

$p(x) = \exp\{\theta_2 x_2 + \theta_{12} x_1 x_2 - \psi(\theta)\}$



Vanilla gradient ∇F vs Natural gradient $\widetilde{\nabla} F$

Generating minimizing sequences $\{p_t\}$

31

In model-based optimization, the relaxed problem (SR) can be approached with different techniques, among the other we have

- Estimation of distribution EDAs, see (Larrañaga and Lozano, 2002) for a review
- Covariance Matrix Adaptation CMA-ES (Hansen and Ostermeier, 2001)
- Fitness modelling DEUM framework (Shakya et al., 2005)
- Gradient descent NES (Wierstra et al., 2008), SNGD (M. et al., FOGA 2011), IGO (Arnold et al., 2011)

In the following we will show how a geometrical framework based on Information Geometry can be exploited to relate these different approaches

Part II

A General Framework for Algorithms

- In the first part we have seen natural gradients on distributions.
- Now we will derive concrete algorithms from this general framework.
- Design choices:
  - search space: discrete or continuous, structure?
  - statistical model?
  - stochastic relaxation?
  - efficient computation/estimation of $\tilde{\nabla} W_f(\xi)$?

The Log-Likelihood Trick

Assume the form $W_f(\xi) = \mathbb{E}_{x\sim P_\xi}\big[w\big(f(x)\big)\big]$.

$$\begin{aligned}
\nabla_\xi W_f(\xi) &= \nabla_\xi \mathbb{E}_{x\sim P_\xi}\big[w(f(x))\big] \\
&= \nabla_\xi \int_\Omega w(f(x)) p(x|\xi)\,\mathrm{d}x \\
&= \int_\Omega w(f(x)) \cdot \nabla_\xi p(x|\xi)\,\mathrm{d}x \\
&= \int_\Omega w(f(x)) \cdot \nabla_\xi p(x|\xi) \cdot \frac{p(x|\xi)}{p(x|\xi)}\,\mathrm{d}x \\
&= \int_\Omega w(f(x)) \cdot \frac{\nabla_\xi p(x|\xi)}{p(x|\xi)} \cdot p(x|\xi)\,\mathrm{d}x \\
&= \mathbb{E}_{x\sim P_\xi}\big[w(f(x)) \cdot \nabla_\xi \log\big(p(x|\xi)\big)\big]
\end{aligned}$$

The gradient of the expectation can be written as the expectation of a weighted gradient of the log likelihood.

The Log-Likelihood Trick

$$\nabla_\xi W_f(\xi) = \mathbb{E}_{x\sim P_\xi}\big[w(f(x)) \cdot \nabla_\xi \log\big(p(x|\xi)\big)\big]$$

- The expected value can be estimated efficiently.
- Its Monte Carlo estimate reads:

$$\nabla_\xi W_f(\xi) \approx \frac{1}{N} \sum_{x_1,\ldots,x_N\sim P_\xi} w(f(x_i)) \cdot \nabla_\xi \log\big(p(x_i|\xi)\big)$$

- Note: neither the gradient of W_f nor its approximation require the gradient of f.

Stochastic Gradient Descent (SGD)

- A simple way to optimize based on a gradient is **gradient descent (GD)**:

$$\xi \leftarrow \xi + \gamma \cdot \nabla_\xi W_f(\xi) \ .$$

- The parameter $\gamma > 0$ is called learning rate.
- Following an unbiased gradient estimate $G(\xi)$ (with $\mathbb{E}[G(\xi)] = \nabla_\xi W_f(\xi)$) is known as **stochastic gradient descent (SGD)**:

$$\xi \leftarrow \xi + \gamma \cdot G(\xi) \ .$$

- This is a well-established algorithm, e.g., in machine learning.
- Using $\tilde{\nabla} W_f(\xi)$ instead of $\nabla W_f(\xi)$ we have even stochastic **natural** gradient descent (SNGD).
- This approach gives a concrete scheme for iterative optimization of $W_f(\xi)$.

Continuous Optimization with NES

- In the context of evolution strategies such a scheme was first proposed by Wierstra et al. in 2008.
- Original **Natural Evolution Strategies (NES)** approach: apply SNGD to optimize expected fitness $W_f(\xi) = \mathbb{E}_{x \sim P_\xi}[f(x)]$ with multi-variate Gaussian search distributions $\mathcal{N}(m, C)$.
- SNGD with Gaussians is rather easy since the Fisher matrix is a closed form term:

$$\mathcal{I}_{i,j} = \frac{\partial m^T}{\partial \xi_i} C^{-1} \frac{\partial m}{\partial \xi_j} + \frac{1}{2} \operatorname{tr}\left(C^{-1} \frac{\partial C}{\partial \xi_i} C^{-1} \frac{\partial C}{\partial \xi_j} \right)$$

- Practical versions of NES apply many performance enhancing techniques like rank-based utilities and non-uniform learning rates that complement the SNGD approach.

Natural Gradients for Gaussian Distributions

- In $\mathbb{R}^d$ we start with the Gaussian density

$$p(x|\xi) = \frac{1}{\sqrt{(2\pi)^d \det(C)}} \exp\left(-\frac{1}{2}(x-m)^T C^{-1} (x-m) \right) \ .$$

Its natural logarithm is

$$\log\big(p(x|\xi)\big) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\operatorname{tr}(\log(C)) - \frac{1}{2}(x-m)^T C^{-1}(x-m) \ .$$

- It is often beneficial to represent the covariance matrix with a factor: $C = AA^T$. For $x \sim \mathcal{N}(m, C)$ we introduce the transformed sample $z = A^{-1}(x-m) \sim \mathcal{N}(0, \mathrm{I})$.
- In tailored coordinates

$$(m', A') = \left(m + A\delta, A\big(\mathrm{I} + \frac{1}{2}M\big) \right)$$

centered to the current distribution (m, A) the Fisher matrix w.r.t. the local parameters $\xi = (\delta, M)$ becomes the identity.

Natural Gradients for Gaussian Distributions

- The (natural) gradient of the log density at $(\delta, M) = 0$ is

$$\begin{aligned} \tilde{\nabla}_\delta \log(p(x|\xi)) &= A^{-1}(x-m) = z \\ \tilde{\nabla}_M \log(p(x|\xi)) &= \frac{1}{2}\big(A(x-m)(x-m)^T A - \mathrm{I}\big) = \frac{1}{2}(zz^T - \mathrm{I}) \end{aligned}$$

- The stochastic (natural) gradient of W_f becomes

$$\begin{aligned} G_\delta(\xi) &= \frac{1}{N}\sum_{i=1}^N f(x_i) \cdot z_i \\ G_M(\xi) &= \frac{1}{2N}\sum_{i=1}^N f(x_i) \cdot (z_i z_i^T - \mathrm{I}) \end{aligned}$$

- Tricks of the trade: replace “raw fitness” with “rank-based utility weights”

$$f(x_1) \le \cdots \le f(x_N) \quad \rightarrow \quad u_1 \ge \cdots \ge u_N$$

to achieve better invariance and faster convergence.

Natural Evolution Strategies (NES)

Canonical NES algorithm with Gaussians $\mathcal{N}(m, C = AA^T)$

`while` stopping criterion not met `do`

`// sample offspring`
`for` $i \in \{1, \ldots, N\}$ `do`
$z_i \leftarrow \mathcal{N}(0, I)$
$x_i \leftarrow m + A \cdot z_i$

`sort` $\{(z_i, x_i)\}$ w.r.t. $f(x_i)$

`// compute stochastic natural gradient`
$G_\delta \leftarrow \frac{1}{N} \sum_{i=1}^{N} u_i \cdot z_i$
$G_M \leftarrow \frac{1}{2N} \sum_{i=1}^{N} u_i \cdot (z_i z_i^T - I)$

`// apply update`
$m \leftarrow m + \gamma_m \cdot A \cdot G_\delta$
$A \leftarrow A \cdot (\mathrm{I} + \gamma_A \cdot G_M)$

`loop`

Natural Evolution Strategies (NES)

- NES is a CMA-ES-like algorithm from "first principles". It "explains" three aspects of ES from a single principle:
  - optimization – update of m
  - step size control – update of $\sigma = \sqrt[d]{\det(A)}$
  - shape control (CMA) – update of A (or of A/σ)
- However, it does not incorporate everything:
  - noise-countering techniques such as cumulation
- Moreover, "dirty tricks" are needed to make it fly:
  - rank-based utilities replace fitness values
  - different learning rates for mean and covariance
  - other techniques are subtraction of a fitness baseline and importance mixing to incorporate older samples

SNGD with Exponential Family

- Consider an exponential family

$$p(x|\theta) = \exp\left(\sum_{i=1}^{k} \theta_i T_i(x) - \psi(\theta)\right) \ .$$

  with sufficient statistics T_i.
- The derivative of the log density is simply

$$\frac{\partial \log(p(x|\theta))}{\partial \theta_i} = T_i(x) - \mathbb{E}[T_i(x)] \ .$$

- Hence also the gradient has a simple form:

$$\underset{x \sim P_\theta}{Cov}\Big(T(x), W_f(x)\Big) \ .$$

SNGD with Exponential Family

- The Fisher matrix has entries

$$\mathcal{I}_{ij}(\xi) = \underset{x \sim P_\theta}{Cov}\Big(T_i(x), T_j(x)\Big) \ .$$

- The natural gradient flow can be expressed solely in terms of covariances, between sufficient statistics and with the objective:

$$\nabla W_f(\theta) = \underset{x \sim P_\theta}{Cov}\Big(T(x), T(x)\Big)^{-1} \underset{x \sim P_\theta}{Cov}\Big(T(x), W_f\Big) \ .$$

SNGD with Exponential Family

44

- Now assume bitstrings $\Omega = \{0,1\}^n$.
- Then the probability simplex Δ and hence the statistical manifold Θ is finite dimensional.
- The sufficient statistics $T_i(x)$ are (square free) monomials.
- Each monomial characterizes a subset of bits the dependencies of which can be modeled.
- If the chosen model contains all interactions of variables in f then there is only one (global) optimum of W_f. The natural gradient will lead us there.

Information Geometric Optimization (IGO)

45

The **Information Geometric Optimization (IGO)** approach by Ollivier et al. introduces a unifying perspective:

- it emphasizes invariance properties as a means to reduce the number of arbitrary design choices,
- with a specific choice of W_f it explains the utility weights of NES from within the framework,
- it highlights the role of the gradient flow as the "pure form" of the EA, with the SNGD update being an approximation.

Dynamic Stochastic Relaxation

46

- Expected fitness $W_f(\xi) = \mathbb{E}_{x\sim P_\xi}[f(x)]$ is only one possible stochastic relaxation of f.
- We have already seen the generalization $W_f(\xi) = \mathbb{E}_{x\sim P_\xi}[w(f(x))]$ with a transformation w.
- In IGO the weight function depends on the f-quantile under the current distribution P_{ξ_0}: $w(f(x)) = \tilde{w}(q^{-1}_{\xi_0}(f(x)))$, where $q_{\xi_0} : [0,1] \to \mathbb{R}$ encodes the quantiles of the distribution of $f(x)$, $x \sim P_{\xi_0}$.
- E.g., $q_{\xi_0}(1/2)$ is the median of f-values, and $\tilde{w}(q) = 1$ for $q < 1/2$ and $\tilde{w}(q) = 0$ for $q \geq 1/2$ encodes truncation (selection): only the better half of the distribution enters the update equation.
- The dynamic choice of $w = w(\xi_0)$ rescales f to a locally relevant range. It emphasizes local improvements relative to the current f distribution.

Dynamic Stochastic Relaxation

47

- Benefit 1: W_f becomes invariant under rank-preserving transformations of fitness values.
- Benefit 2: the rank-based utility weights of NES are obtained automatically in a principled manner.
- Drawback: the objective function $W_f^{\xi_0}(\xi)$ becomes dependent on the current distribution $\xi = \xi_0$
- This means that the following situation may exist in principle:

$$W_f^{\xi_1}(\xi_2) > W_f^{\xi_1}(\xi_1)$$
$$W_f^{\xi_2}(\xi_3) > W_f^{\xi_2}(\xi_2)$$
$$W_f^{\xi_3}(\xi_1) > W_f^{\xi_3}(\xi_3)$$

  and the "optimization" turns around in circles...
- Provably, in important special cases this does not happen.

Vector Field, ODE, and Flow

48

- Algorithmic and hence discrete time optimization is only one possibility.
- The natural gradient $\tilde{\nabla} W_f(\xi)$ defines the vector field $V : \Xi \to T\Xi$ via $V(\xi) = \tilde{\nabla} W_f(\xi)$.
- This vector field is naturally associated with the differential equation $\dot{\gamma}(t) = V(\gamma(t))$ with solution curves $\gamma : \mathbb{R} \to \Xi$.
- The solution curves are collected in the flow $\varphi(\xi, t)$.
- Note 1: just like the natural gradient itself this flow is deterministic. This is achieved in the limit of infinite samples in the MC approximation, corresponding to infinite population size.
- Note 2: in each point the flow moves tangential to the vector field. This corresponds to re-evaluating the gradient after an infinitesimal step, or to an infinitesimal leaning rate in the gradient descent procedure.

SGD Algorithms

49

- An SGD algorithm is a two-fold approximation of the flow:
  - it discretizes time and performs Euler steps,
  - it relies on a stochastic gradient based on sampling.
- If the IGO objective is in use then such algorithms have been termed IGO algorithms.
- NES is a rather pure example of an IGO algorithm.
- Surprisingly many established algorithm fall into this category or have a close connection to SGD (or IGO) algorithms
- This derivation of randomized direct search algorithms opens a new perspective: the EA is an approximation of a much more simple and beautiful mathematical object, namely the deterministic, time continuous gradient flow, thought of as an idealized EA.

Connection to CMA-ES

50

- Modern CMA-ES is a versatile algorithm that comes in many variants, e.g., with mirrored sampling, a restart strategy, a noise handling mechanism, an elitist variant, constraint handling, ...
- Even the basic reference algorithm is rather complex.
- It updates the distribution mean with global weighted recombination:

$$m \leftarrow \sum_{i=1}^{\mu} w_i \cdot x_i \quad .$$

- For the purpose of simplified analysis we drop cumulative step size adaptation and consider only the so-called rank-μ update

$$C \leftarrow (1 - \gamma_C) \cdot C + \gamma_C \cdot \sum_{i=1}^{\mu} w_i \cdot (x_i - m)(x_i - m)^T \quad .$$

Connection to CMA-ES

51

- Both equations can be written as updates

$$m \leftarrow m + \gamma_m \cdot \sum_{i=1}^{N} w_i \cdot (x_i - m)$$

$$C \leftarrow C + \gamma_C \cdot \sum_{i=1}^{N} w_i \cdot \left((x_i - m)(x_i - m)^T - C \right)$$

  with fixed learning rate $\gamma_m = 1$ and $w_i = 0$ for $i > \mu$.
- The change of coordinates $C = AA^T$, $x = Az + m$ reveals:

$$m \leftarrow m + \gamma_m \cdot A \cdot \sum_{i=1}^{N} w_i \cdot z_i$$

$$C \leftarrow C + \gamma_C \cdot A \cdot \left(\sum_{i=1}^{N} w_i \cdot \left(z_i z_i^T - \mathrm{I} \right) \right) \cdot A^T$$

- This is essentially the IGO/NES SGD update (see Akimoto 2010).

Connection to Maximum Likelihood Estimation

- The CMA-ES update equations can be written as

$$m \leftarrow (1 - \gamma_m) \cdot m + \gamma_m \cdot \hat{m}_{\mathsf{ML}} \ ,$$
$$C \leftarrow (1 - \gamma_C) \cdot C + \gamma_C \cdot \hat{C}_{\mathsf{ML}} \ .$$

- Here $\hat{m}_{\mathsf{ML}} = \sum_{i=1}^{N} w_i \cdot x_i$ is the weighted Maximum Likelihood (ML) estimator of m.
- The term

$$\hat{C}_{\mathsf{ML}} = \sum_{i=1}^{\mu} w_i (x_i - m)(x_i - m)^T$$

  is the weighted ML estimator of C, *provided that* m *remains fixed*.
- For a learning rate of $\gamma_m = \gamma_C = 1$ we obtain $m \leftarrow \hat{m}_{\mathsf{ML}}$ and $C \leftarrow \hat{C}_{\mathsf{ML}}$.
- Due to the large learning rate this works in practice only for a large enough sample size μ.

Convergence Results

- Convergence of optimization algorithms to the optimum is more than “nice to have”.
- Example: consider a convex problem with minimum f^*. Convergence means that for any $\varepsilon > 0$ the algorithm with stopping criterion $f(x) < f^* + \varepsilon$ halts.
- Hence convergence guarantees let people sleep better.
- Convergence rates ($\mathcal{O}$ and also the actual constants) are equally important.
- This is because the fact that the optimization will stop at all is meaningless in practice; instead we need to know (roughly) how long it takes.

Convergence Results

- Convergence has been analyzed for many optimization algorithms. For most realistic evolutionary algorithm this is a hard task.
- One of the benefits of the gradient flow (as an idealization of actual algorithms) is that it is much easier to analyze.
- Question: Do all trajectories converge to the optimum?
- More formally, let δ_{x^*} denote the Dirac peak over an (isolated) optimum $x^* \in \Omega$. Does it hold $\lim_{t \to \infty} P_{\phi(\xi,t)} = \delta_{x^*}$ for all initial conditions ξ?
- Other notions of convergence are possible, e.g., convergence of expected fitness.
- Note: convergence of the flow does not directly imply convergence of stochastic approximate algorithms!

Representability: The Marginal Polytope

- Convergence of the gradient flow to the optimum has a hidden prerequisite:
- The statistical model must be sufficiently rich to focus the probability mass on an optimum, or more exactly, on a subset of the optimal set $\Omega^* \subset \Omega$.
- In general this property is a non-trivial prerequisite for optimization with a stochastic relaxation.
- As such it is not a prerequisite for convergence of the flow to an optimal distribution *within* (the closure of) the statistical model.
- However, it is a prerequisite for convergence to an optimal distribution, possibly outside the family, and hence to an optimum of the original problem $\min_{x \in \Omega} f(x)$.

Representability: The Marginal Polytope

- Assume an exponential family

$$p(x|\xi) = \exp\left(\sum_{i=1}^{k} \xi_i T_i(x) - \psi(\xi)\right)$$

  of distributions. Then this property has a geometric interpretation in distribution space.
- Distribution are restricted to the *marginal polytope*, which is the convex hull of the canonical statistics $T_i(x)$.
- For discrete search spaces Ω it is a sub-polytope of the probability simplex.
- The optimum of the stochastically relaxed problem describes an optimum of $f : \Omega \to \mathbb{R}$ iff a subset of $S \subset \Omega^*$ corresponds to an *exposed face* W_f of the marginal polytope, i.e., if $S = T^{-1}(W_f) \subset \Omega^*$.

Representability: The Marginal Polytope

- Consider the family of Gaussian distributions $\mathcal{N}(m, \sigma^2)$ with adaptive mean m and step size σ.
- The Dirac distribution δ_x of a point $x \in \mathbb{R}^d$ is obtained in the limit

$$\delta_x = \lim_{\sigma \to 0} \mathcal{N}(x, \sigma^2) \ .$$

- Hence the probability mass can be focused to an arbitrarily small neighborhood of an optimal point $x^* \in \Omega^* \subset \mathbb{R}^d$.
- As a consequence convergence is possible also for all supersets of distributions, i.e., multi-variate Gaussians.

Convergence Results

Theorem (Akimoto et al. 2012, Glasmachers 2012)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a strictly convex quadratic function with minimum x^*. Consider the class $\mathcal{N}(m, \sigma^2)$ of isotropic Gaussian search distributions. Then all trajectories of the IGO flow converge to the boundary point $m = x^*$ and $\sigma^2 = 0$ (corresponding to δ_{x^*}).

Corollary (Akimoto et al. 2012)

The same holds for monotonically transformed functions.

Corollary (Akimoto et al. 2012)

The same holds in the vicinity of any twice continuously differentiable local optimum.

Convergence Results with CMA

The following theorem was proven for IGO with weighted quantiles of a *fixed* (non-adaptive) reference distribution:

Theorem (Akimoto 2012)

Consider $f(x) = x^T Q x$ with strictly positive definite matrix Q and multivariate Gaussian search distributions $\mathcal{N}(m, C)$. Then it holds

$$m \to 0 \qquad C \to Q^{-1} \ .$$

Convergence Results with CMA

The following result is for the IGO flow with expected fitness:

Lemma (Beyer 2013)

For multivariate Gaussians $\mathcal{N}(m, C)$ on a convex quadratic objective $f(x) = x^T Q x$ the gradient flow is defined by the differential equation

$$\frac{dm(t)}{dt} = -2\,C(t)\,Q\,m(t)$$
$$\frac{dC(t)}{dt} = -2\,C(t)\,Q\,C(t)$$

A similar result was found by Akimoto for the IGO flow with fixed reference distribution.

Convergence Results with CMA

Theorem (Beyer 2013)

The non-linear ordinary differential equation (ODE) system

$$\frac{dC(t)}{dt} = -2\,C(t)\,Q\,C(t)$$

with initial condition $C(0) = C_0$ has the solution

$$C(t) = \left(C_0^{-1} + 2tQ\right)^{-1} .$$

$$\Rightarrow C(t) \to Q^{-1}$$
$$\Rightarrow \|C(t)Q^{-1}\| \in \mathcal{O}(1/t)$$

Beyer also obtains $\|m(t)\| \in \mathcal{O}(1/t)$.

Convergence Results with CMA

Theorem (Beyer 2013)

Under the assumption of (approximate) normality of fitness values the dynamics of IGO (with quantile-based objective) are

$$m(t) \approx \alpha \cdot \exp\left(-\sqrt{2/d} \cdot t\right) \cdot Q^{-1} C_0^{-1} m_0 \ ,$$
$$C(t) \approx \alpha \cdot \exp\left(-\sqrt{2/d} \cdot t\right) \cdot Q^{-1} \ .$$

- Normality of fitness values is exact for linear functions.
- It holds approximately for quadratic functions for large $d \gg 1$.
- The flow convergences at a linear rate, which is what we'd expect for an evolution strategy.
- The dependence on the dimension is surprisingly good; the square root is not achieved by SGD algorithms.

Convergence Results

Summary: convergence of the flow for continuous search spaces

- Proofs for isotropic and general multi-variate Gaussians.
- Results are restricted to (monotonically transformed) quadratic functions. This models convergence to twice differentiable local optima well.
- There are no results for more general problem classes such as all convex problems.
- Note once more: convergence results for the gradient flow do not imply convergence of the EA.
- The deviations of the algorithm from the flow due of stochasticity (finite populations) and finite step sizes (discrete time) are yet to be understood.

But honestly – does it really work?

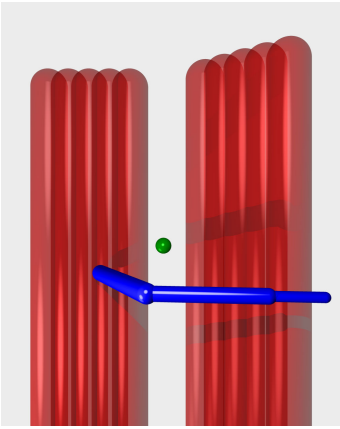
- This question is not well posed.
- Many EAs apply update equations that can be explained from information geometry.
- Statistical models and stochastic relaxations do not and will probably never cover all aspects of EAs.
- Realistic EAs can be built from two types of components:
  - update equations derived from information geometry,
  - classic tools for handling stochasticity.
- Each component must do its job.
- Information geometrical updates generally do a great job at improving the search distribution, provided that stochastic effects are sufficiently well controlled.

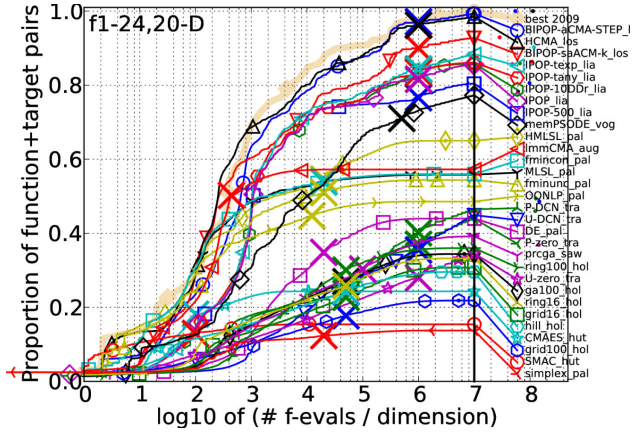
But honestly – does it really work?

As far as the above outlined role of information geometry in EAs is concerned the clear answer is:

Yes, it works great!

Case study: search for collision free robot arm trajectories





source: N. Hansen, A few overview results from the GECCO BBOB workshops

Summary

- Information geometry provides update equations for optimization from first principles.
- This often amounts to stochastic natural gradient descent applied to a stochastically relaxed problem.
- This is an approximation to an optimization flow.
- It can help the analysis of existing algorithms like CMA-ES.
- It is a generic design principle for optimization algorithms on any search space and any family of search distributions.
- Dedicated algorithms such as NES were built on this principle.
- Algorithms respecting the information geometry of their search distributions are among the top performers.

Summary

- However, randomized algorithms deviate widely from the idealized, deterministic gradient flow. These stochastic effects are outside the framework, they must be controlled by other means.
- Information geometric tools must be augmented with "orthogonal" tools for control of stochastic effects—together they provide a modern perspective on EA research.
- The same problem decomposition is a promising route for theoretical analysis: the gradient flow is becoming a well-investigated object, while more traditional tools (Markov chain analysis, etc.) may be necessary to connect it to real EAs.

References

1. Ollivier et al. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. *arXiv:1106.3708*, 2011.
2. Wierstra et al. Natural Evolution Strategies. *Congress on Evolutionary Computation (CEC)*, 2008.
3. Glasmachers et al. Exponential Natural Evolution Strategies. *Genetic and Evolutionary Computation Conference (GECCO)*, 2010.
4. Akimoto et al. Bidirectional relation between CMA evolution strategies and natural evolution strategies. *Parallel Problem Solving from Nature (PPSN) XI* 2010.
5. Akimoto et al. Convergence of the Continuous Time Trajectory of Isotropic Evolution Strategies on Monotonic C^2-composite Functions. *Parallel Problem Solving from Nature (PPSN) XII*, 2012.
6. Glasmachers. Convergence of the IGO-Flow of Isotropic Gaussian Distributions on Convex Quadratic Problems. *Parallel Problem Solving from Nature (PPSN) XII*, 2012.
7. Malagò et al. Natural Gradient, Fitness Modelling and Model Selection: A Unifying Perspective. *Congress on Evolutionary Computation (CEC)*, 2013.
8. Malagò et al. Towards the geometry of estimation of distribution algorithms based on the exponential family. *FOGA*, 2011.
9. Beyer. Convergence Analysis of Evolutionary Algorithms which are Based on the Paradigm of Information Geometry. *Personal communication.*

Thank you!