Introducing Information-Geometric Optimization, a distinct framework for randomized optimization

Nikolaus Hansen
INRIA, Research Centre Saclay
Machine Learning and Optimization Team, TAO
Univ. Paris-Sud, LRI

joint work with Yann Ollivier & Anne Auger

http://www.inria.fr

http://www.lri.fr/~hansen

based on

- Wierstra et al (2008). Natural Evolution Strategies, *IEEE Congress on Evolutionary Computation, CEC*.
- Yi et al (2009), Stochastic search using the natural gradient, *International Conference on Machine Learning, ICML*.
- Glasmachers et al (2010), Exponential natural evolution strategies. *ACM Genetic and Evolutionary Computation Conference, GECCO*.
- Akimoto et al (2010). Bidirectional relation between CMA Evolution Strategies and Natural Evolution Strategies. *Parallel Problem Solving from Nature, PPSN, proceedings.*
- Malagò et al (2011). Towards the geometry of Estimation of Distribution Algorithms based on the exponential family, *Foundations of genetic algorithms, FOGA*, workshop proceedings.
- Arnold et al (2011), Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles, *arXiv:1106.3708v1*

Black-Box Optimization (Search)

Minimize (or maximize) an objective (cost, loss, error, fitness) function

$$f : \mathcal{X} \to \mathbb{R}, \quad x \mapsto f(x)$$

in a black-box scenario (direct search)

where we have no specific assumptions on f

Stochastic optimization template

Given: a parametrized distribution $p(.|\theta)$ on $\mathcal{X}$ and $\lambda \in \mathbb{N}$
Initialize: parameter vector θ

e.g. mean and variance of p

While not *happy*

1. Sample $p(x|\theta) \to x_1, \ldots, x_\lambda \in \mathcal{X}$
2. Evaluate $x_1, \ldots, x_\lambda$ on $f \longrightarrow f(x_1), \ldots, f(x_\lambda)$
3. Update parameters

$$\theta \leftarrow \mathrm{Update}(\theta, x_1, \ldots, x_\lambda, f(x_1), \ldots, f(x_\lambda))$$

Return, e.g., the expectation or ML value of $p(.|\theta)$ as given in θ

Remark: this algorithm learns a probabilistic model of f

Stochastic optimization template

Given: a parametrized distribution $p(.|\theta)$ on $\mathcal{X}$ and $\lambda \in \mathbb{N}$
Initialize: parameter vector θ
While not *happy*

1. Sample $p(x|\theta) \to x_1, \ldots, x_\lambda \in \mathcal{X}$
2. Evaluate $x_1, \ldots, x_\lambda$ on $f \longrightarrow f(x_1), \ldots, f(x_\lambda)$
3. Update parameters

$$\theta \leftarrow \mathrm{Update}(\theta, x_1, \ldots, x_\lambda, f(x_1), \ldots, f(x_\lambda))$$

Return, e.g., the expectation or ML value of $p(.|\theta)$ as given in θ

How to choose the parametrized distribution $p(.|\theta)$?
How to update θ (point 3.)?

A new search problem

The stochastic search template implies a new search problem in θ-space (derived from the objective function $f : \mathcal{X} \to \mathbb{R}$):

$$J(\theta) = \mathbb{E}(W(f(x))), \quad x \sim p(.|\theta)$$

$$\overset{\mathcal{X} \text{ uncountable}}{=} \int_{\mathcal{X}} W(f(x)) p(x|\theta) \mathrm{d}x \qquad \text{to be maximized}$$

where $W \equiv W^f_{\theta_t} : \mathbb{R} \to \mathbb{R}$ depends on f and on the parameter θ_t at iteration t (but consider $W(f(x)) = -f(x)$ for the time being)

A new search problem

The stochastic search template implies a new search problem in θ-space (derived from the objective function $f : \mathcal{X} \to \mathbb{R}$):

$$J(\theta) = \mathbb{E}(W(f(x))), \quad x \sim p(.|\theta)$$

$$\overset{\mathcal{X} \text{ uncountable}}{=} \int_{\mathcal{X}} W(f(x)) p(x|\theta) \mathrm{d}x \qquad \text{to be maximized}$$

where $W \equiv W_{\theta_t}^f : \mathbb{R} \to \mathbb{R}$ depends on f and on the parameter θ_t at iteration t (but consider $W(f(x)) = -f(x)$ for the time being)

To improve J, we will consider the gradient

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}\left(W(f(x))\right)$$

$$= \mathbb{E}\left(W(f(x)) \nabla_\theta \log p(x|\theta)\right) \qquad \text{because } \nabla_\theta p = p \nabla_\theta \log p$$

$\nabla_\theta J$ is the direction of steepest ascend of J in θ

A new search problem

Let $x \sim p(.|\theta)$ the sample distribution. The new objective

$$J(\theta) = \mathbb{E}(W(f(x)), \quad x \sim p(.|\theta)$$

induces the time continuous gradient flow

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta_t = \nabla_\theta J(\theta)\Big|_{\theta=\theta_t} = \mathbb{E}\Big(\overbrace{W(f(x))}^{\text{in } \mathbb{R}} \underbrace{\nabla_\theta \log p(x|\theta)}_{\text{in } \mathbb{R}^{\dim(\theta)}} \Big)\Big|_{\theta=\theta_t}$$

discretized with λ samples and learning rate $\eta > 0$ the iteration

$$\underbrace{\theta_{t+1} - \theta_t}_{\downarrow\ \eta \frac{\mathrm{d}}{\mathrm{d}t}\theta_t \ (\lambda \to \infty)} = \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{W(f(x_k))}^{\text{preference weight}} \underbrace{\nabla_\theta \log p(x_k|\theta)\Big|_{\theta=\theta_t}}_{\text{direction for } x_k}, \quad x_k \sim p(.|\theta_t)$$

A new search problem

Let $x \sim p(.|\theta)$ the sample distribution. The new objective

$$J(\theta) = \mathbb{E}(W(f(x)), \quad x \sim p(.|\theta)$$

induces the time continuous gradient flow

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta_t = \nabla_\theta J(\theta)\Big|_{\theta=\theta_t} = \mathbb{E}\Big(\overbrace{W(f(x))}^{\text{in } \mathbb{R}} \underbrace{\nabla_\theta \log p(x|\theta)}_{\text{in } \mathbb{R}^{\dim(\theta)}} \Big)\Big|_{\theta=\theta_t}$$

discretized with λ samples and learning rate $\eta > 0$ the iteration

$$\underbrace{\theta_{t+1} - \theta_t}_{\downarrow \; \eta \frac{\mathrm{d}}{\mathrm{d}t}\theta_t \; (\lambda \to \infty)} = \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{W(f(x_k))}^{\text{preference weight}} \underbrace{\nabla_\theta \log p(x_k|\theta)\Big|_{\theta=\theta_t}}_{\text{direction for } x_k}, \quad x_k \sim p(.|\theta_t)$$

The update

$$\underbrace{\theta_{t+1} - \theta_t}_{\downarrow\ \eta \frac{\mathrm{d}}{\mathrm{d}t}\theta_t \ (\lambda \to \infty)} = \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{W(f(x_k))}^{\text{preference weight}} \underbrace{\nabla_\theta \log p(x_k|\theta)\Big|_{\theta=\theta_t}}_{\text{direction for } x_k}, \quad x_k \sim p(.|\theta_t)$$

We need to explain/compute

1. $W(f(x_k))$ — very simple to approximate in practice
2. $\nabla_\theta \log p(x|\theta)$ — heavily depends on $p(.|\theta)$

and start with 2.

The direction

$$\theta_{t+1} - \theta_t = \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{W(f(x_k))}^{\text{preference weight}} \underbrace{\nabla_\theta \log p(x_k|\theta)\Big|_{\theta=\theta_t}}_{\text{direction for } x_k}, \quad x_k \sim p(.|\theta_t)$$

∇_θ depends on a metric in θ-space

- why the Euclidean metric?
- which parametrization of p in θ?
- why not second order?

$\Longrightarrow$ invariance is a major design principle

A Reminder

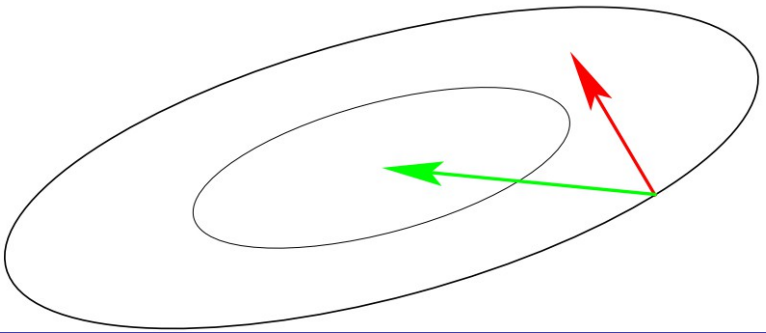
Let $x_k \in \mathbb{R}^n, \eta > 0$
In order to improve (reduce) $f(x_k)$, descend in gradient direction:

$$x_{k+1} = x_k - \eta \vec{\nabla} f(x_k) \qquad \text{with } \eta \text{ small}$$

or even better in Newton direction using the metric induced by the inverse Hessian, H^{-1}, of f:

$$x_{k+1} = x_k - \eta H^{-1} \vec{\nabla} f(x_k)$$

incorporating the curvature of f
Remark: H depends on f and might also depend on x



gradient direction $-f'(\boldsymbol{x})^{\mathrm{T}}$

Newton direction $-\boldsymbol{H}^{-1} f'(\boldsymbol{x})^{\mathrm{T}}$

A Metric for Probability Distributions

The *Fisher information metric* is the curvature of the entropy and implies an informational difference between probability distributions

The *natural gradient* $\widetilde{\nabla}_\theta = \mathcal{I}_\theta^{-1} \nabla_\theta$ uses the Fisher information metric (the respective inner product)

$$\mathcal{I}_{ij}(\theta) = -\mathbb{E} \frac{\partial^2 \log p(x|\theta)}{\partial \theta_i \partial \theta_j}$$

Among all gradients, the natural gradient is distinguished as being invariant under θ-re-parametrization and compliant with KL-divergence (relative entropy, informational difference)

Remark: all previous derivations hold for any gradient and are independent of the underlying problem f .

A Metric for Probability Distributions

The *Fisher information metric* is the curvature of the entropy and implies an informational difference between probability distributions

The *natural gradient* $\widetilde{\nabla}_\theta = \mathcal{I}_\theta^{-1} \nabla_\theta$ uses the Fisher information metric (the respective inner product)

$$\mathcal{I}_{ij}(\theta) = -\mathbb{E} \frac{\partial^2 \log p(x|\theta)}{\partial \theta_i \partial \theta_j}$$

Among all gradients, the natural gradient is distinguished as being invariant under θ-re-parametrization and compliant with KL-divergence (relative entropy, informational difference)

Remark: all previous derivations hold for any gradient and are independent of the underlying problem f .

The direction

$$\theta_{t+1} - \theta_t = \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{W(f(x_k))}^{\text{preference weight}} \underbrace{\widetilde{\nabla}_\theta \log p(x_k|\theta)\Big|_{\theta=\theta_t}}_{\text{direction for } x_k}, \quad x_k \sim p(.|\theta_t)$$

Examples:

- for the Bernoulli distribution in $x_k \in \{0,1\}^n$ with expectation $\theta \in [0,1]^n$, we have

$$\widetilde{\nabla}_\theta \log p(x_k|\theta) = x_k - \theta$$

- for the normal (Gaussian) distribution $x_k \sim \mathcal{N}(m, \mathbf{C})$ in $\mathbb{R}^n$, with $\theta = \begin{bmatrix} m \\ \mathbf{C} \end{bmatrix}$ we have

$$\widetilde{\nabla}_\theta \log p(x_k|\theta) = \begin{bmatrix} x_k - m \\ (x_k - m)(x_k - m)^{\mathrm{T}} - \mathbf{C} \end{bmatrix}$$

The update

$$\theta_{t+1} - \theta_t = \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{W(f(x_k))}^{\text{preference weight}} \underbrace{\widetilde{\nabla}_\theta \log p(x_k|\theta)\Big|_{\theta=\theta_t}}_{\text{direction for } x_k}, \quad x_k \sim p(.|\theta_t)$$

We need to explain/compute

1. $W(f(x_k))$

   very simple to approximate in practice

2. $\nabla_\theta \log p(x|\theta)$

   heavily depends on $p(.|\theta)$

Choice of W

$$\theta_{t+1} - \theta_t = \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{W(f(x_k))}^{\text{preference weight}} \underbrace{\widetilde{\nabla}_\theta \log p(x_k|\theta)\Big|_{\theta=\theta_t}}_{\text{direction for } x_k}, \quad x_k \sim p(.|\theta_t)$$

The intrinsic choice for W should be

- **f-compliant (monotone)**:
  $W(f(x_i)) \leq W(f(x_j)) \iff f(x_j) \leq f(x_i)$
- **robust** to f-outliers
- **invariant** under strictly monotonous increasing f-transformations (order-preserving transformations)
  for example $f \to f^3 + 10^9$

Defining W

We define

maximize $\mathbb{E}[W^f_{\theta_t}(f(x))\,|\,\theta]$ w.r.t. θ

$$W : y \mapsto W^f_{\theta_t}(y) = w(\underbrace{\Pr(f(X) \leq y \,|\, X \sim p(.|\theta_t))}_{\text{CDF of } f(X \sim p(.|\theta_t)) \text{ at point } y})$$

as

- the cumulative distribution function of $f(X)$ at y,
- the probability to get below value y (i.e. better) when sampling X according to $p(.|\theta_t)$,

transformed with a decreasing weight function $w : [0, 1] \to \mathbb{R}$

$W(f(x)) = w(\text{CDF}(f(x)))$, to be maximized

- is invariant under monotone f-transformations
- for $x \sim p(.|\theta_t)$ we have $W(f(x)) \sim w(\mathcal{U}[0, 1])$ independent of t, $p(.|\theta_t)$, and f
- results in "rank-based selection"

Defining W

We define

maximize $\mathbb{E}[W^f_{\theta_t}(f(x))\,|\,\theta]$ w.r.t. θ

$$W : y \mapsto W^f_{\theta_t}(y) = w(\underbrace{\Pr(f(X) \leq y \,|\, X \sim p(.|\theta_t))}_{\text{CDF of } f(X \sim p(.|\theta_t)) \text{ at point } y})$$

as

- the cumulative distribution function of $f(X)$ at y,
- the probability to get below value y (i.e. better) when sampling X according to $p(.|\theta_t)$,

transformed with a decreasing weight function $w : [0, 1] \to \mathbb{R}$

$W(f(x)) = w(\text{CDF}(f(x)))$, to be maximized

- is invariant under monotone f-transformations
- for $x \sim p(.|\theta_t)$ we have $W(f(x)) \sim w(\mathcal{U}[0, 1])$ independent of θ, t, p, θ_t, and f, J
- results in a comparison-based (rank-based) algorithm

Approximation of $W_{\theta_t}^f$

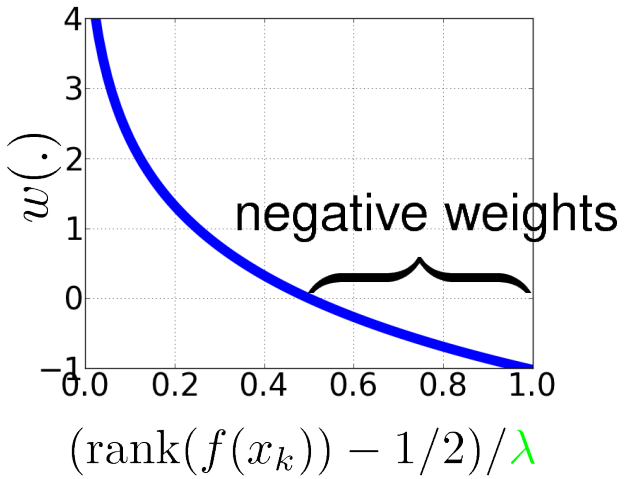
We can get a consistent approximation for

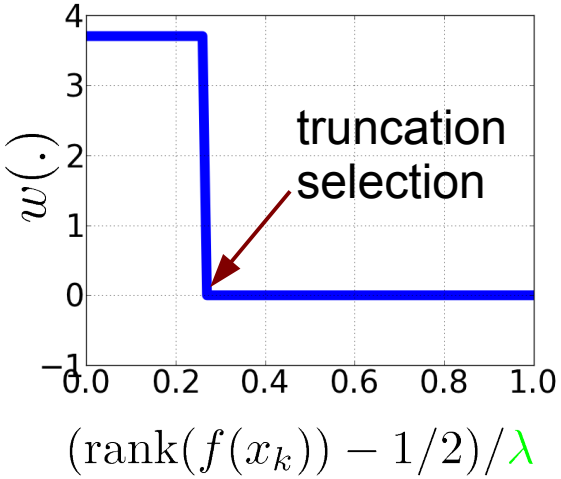
$$W_{\theta_t}^f(f(x_k)) = w(\Pr(f(X) \leq f(x_k), X \sim p(.|\theta_t)))$$

that is easy to compute only by sorting $f(x_1), \ldots, f(x_\lambda)$ as

$$W_{\theta_t}^f(f(x_k)) \approx w\left(\frac{\mathrm{rank}(f(x_k)) - 1/2}{\lambda}\right)$$

for $k = 1, \ldots, \lambda$, where w is monotonuously decreasing (and $w(0.5) = 0$), e.g.





Information-Geometric Optimization Algorithm

Given: search space $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$ to be minimized

Choose: $p(.|\theta)$ on $\mathcal{X}$, $\lambda \in \mathbb{N}$, $\eta > 0$, $w : [0, 1] \to \mathbb{R}$

Initialize: θ

While not *happy*

1. Sample $p(x|\theta) \to x_1, \ldots, x_\lambda \in \mathcal{X}$
2. Evaluate $x_1, \ldots, x_\lambda$ on $f \longrightarrow f(x_1), \ldots, f(x_\lambda)$
3. Update parameters

$$\theta \leftarrow \theta + \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{w\left(\frac{\mathrm{rank}(x_k) - 1/2}{\lambda}\right)}^{\text{preference weight}} \underbrace{\widetilde{\nabla}_\theta \log p(x_k|\theta)}_{\text{direction for } x_k}$$

Information-Geometric Optimization Algorithm

Given: search space $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$ to be minimized

Choose: $p(.|\theta)$ on $\mathcal{X}$, $\lambda \in \mathbb{N}$, $\eta > 0$, $w : [0, 1] \to \mathbb{R}$

Initialize: θ

While not *happy*

1. Sample $p(x|\theta) \to x_1, \ldots, x_\lambda \in \mathcal{X}$
2. Evaluate $x_1, \ldots, x_\lambda$ on $f \longrightarrow f(x_1), \ldots, f(x_\lambda)$
3. Update parameters

$$\theta \leftarrow \theta + \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{w\left(\frac{\mathrm{rank}(x_k) - 1/2}{\lambda}\right)}^{\text{preference weight}} \underbrace{\widetilde{\nabla}_\theta \log p(x_k|\theta)}_{\text{direction for } x_k}$$

not covered (but relevant in practice):

- different learning rates for different components of θ
- low pass filtering over several iteration steps
- the principle is insufficient for step-size control!?

Discrete case

let $x, x_k \sim p(.|\theta_t)$, then the update of θ_t reads

$$\theta_{t+1}$$

$$= \arg\max_{\theta} \left(\eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} W(f(x_k)) \log p(x_k|\theta) + (1-\eta) \underbrace{\mathbb{E}(\log p(x|\theta))}_{\text{cross entropy } \mathbb{E}(-\log p(x|\theta)) \text{ = entropy}(\theta_t) + \text{KL}(\theta_t \| \theta)} \right)$$

$$= \arg\max_{\theta} \left(\underbrace{\eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} W(f(x_k)) \log p(x_k|\theta)}_{\text{maximal if } p(.|\theta) \text{ resembles } W(f(.))} + (1-\eta) \underbrace{\int_{\mathcal{X}} p(x|\theta_t) \log p(x|\theta) \mathrm{d}x}_{\text{maximal for } \theta = \theta_t} \right)$$

$$= \theta_t + \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{W(f(x_k))}^{\text{preference weight}} \underbrace{\widetilde{\nabla}_{\theta} \log p(x_k|\theta)\Big|_{\theta=\theta_t}}_{\text{direction for } x_k} + \mathcal{O}(\eta^2) \quad \text{(for } \eta \text{ small enough)}$$

Key observation: trade off between minimal change of θ_t and bias towards $W(f(.))$
Cross entropy method (CEM) for $\eta = 1$

Summary

Given an objective function f and a family of probability distributions $p(.|\theta)$ on an arbitrary search domain

- we can derive a *randomized (stochastic)* search algorithm under a minimal amount of arbitrary decisions, based on invariance principles, in particular invariance under
  - (re-)parametrization
  - order-preserving f-transformations
- A key property: we get maximal improvement for minimal change of the distribution

Known algorithms

- Covariance Matrix Adaptation Evolution Strategy (CMA-ES)
- Population Based Incremental Learning (PBIL)

Current Work

- starting point for theoretical investigations
- guide to construct new algorithms, namely a “CMA” with linear number of parameters

IGO is a necessary but not sufficient requirement

Questions?

Furious activity is no substitute for understanding.

Albert Einstein