Mittagsseminar

Natural Gradient in Evolution Strategies

Sebastian U. Stich

ETH Zürich

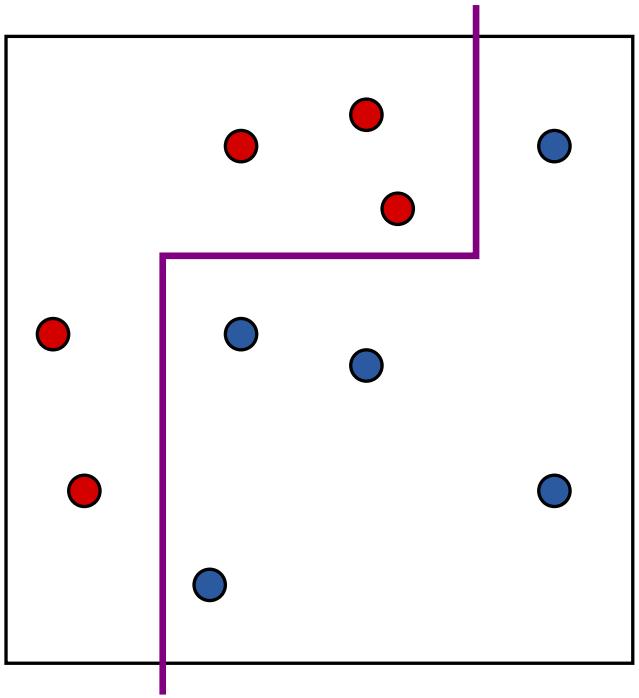
October 17, 2013

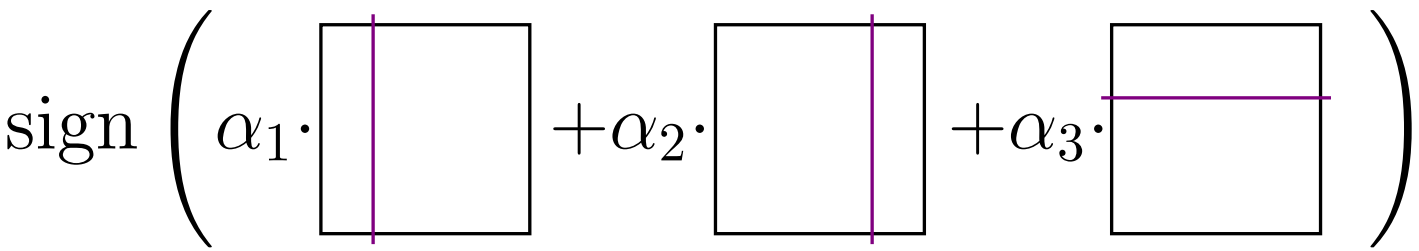
Abstract

In derivative-free optimization one aims at minimizing an unknown objective function. The only information accessible are algorithm-selected function measurements. Evolution Strategies (ES) are among the state of the art heuristics for this optimization problem. ES typically use parametrized probability distributions to generate correlated samples in promising regions.
Recently, it was shown that applying gradient descent in the parameter space of the search distribution leads to algorithms that are very similar to the most successful ES. The development of those so-called Natural Evolution Strategies (NES) provided new insights into the classical ES and started promising new theoretical investigations.
This still on-going research lead to the Information-Geometric-Optimization (IGO) framework, which tries to capture all ES in an unifying picture.
We present NES and give an introduction to the IGO framework.

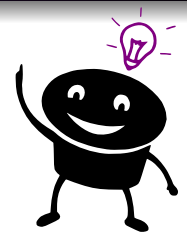
Table of contents

AdaBoost I [Freund, Schapire '97]





AdaBoost II



AdaBoost(B)

for $b = 1$ *to* B **do**

- Train c_b, using weights w
- $\epsilon_b \leftarrow$ weighted error
- $\alpha_b \leftarrow \log \frac{1-\epsilon_b}{\epsilon_b}$
- $w \leftarrow \text{UpdateWeights}(w, \alpha_b)$

return $\hat{c}_B = \text{sign}\left(\sum \alpha_b c_b\right)$

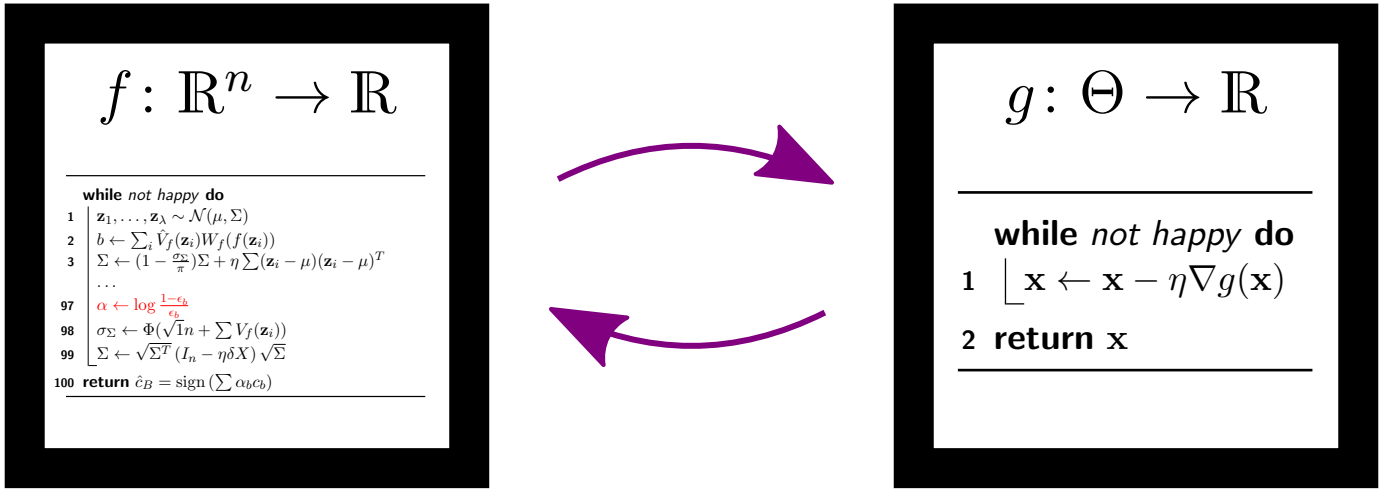
[MBBF'99]

"AdaBoost is performing a Gradient Descent on the cost function

$$C(\hat{c}_B) = \mathbb{E}\left[e^{-\mathbf{y}\hat{c}_B(\mathbf{x})}\right]$$

with step-size chosen by a line-search."

General Idea



Algorithm $\mathcal{A}$ Gradient Descent

- *Explain* algorithm with Gradient Descent (in a different space)
  - new insights, prove convergence, ...
- *Design* new algorithms based on simple principles

Gradient

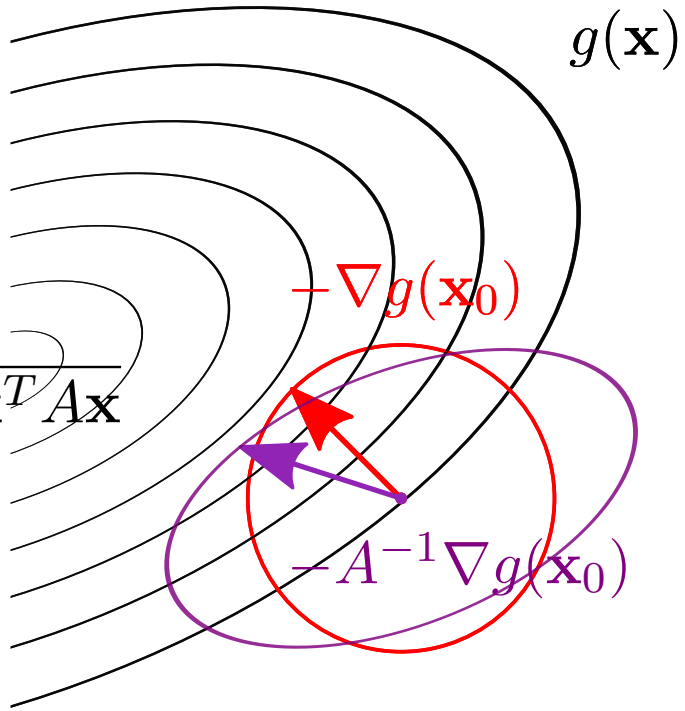
- $g\colon \mathbb{R}^n \to \mathbb{R}$, smooth, gradient $\nabla g(\mathbf{x}_0) \in \mathbb{R}^n$:

$$g(\mathbf{x}) = g(\mathbf{x}_0) + \langle \nabla g(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + o\left(\|\mathbf{x} - \mathbf{x}_0\|\right) \qquad \text{for } \mathbf{x} \to \mathbf{x}_0$$

- Direction of steepest ascent:

$$\frac{\nabla g(\mathbf{x}_0)}{\|\nabla g(\mathbf{x}_0)\|} = \lim_{\epsilon \to 0} \operatorname*{arg\,max}_{\|\mathbf{d}\| \leq 1 \|\mathbf{d}\| \leq 1} g(\mathbf{x}_0 + \epsilon \mathbf{d})$$

- Depends on the norm $\|\mathbf{x}\|$, $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^T A \mathbf{x}}$
- Hessian: $A = \nabla^2 g(\mathbf{x}_0)$
- Newton direction: $A^{-1} \nabla g(\mathbf{x}_0)$
- affine invariant



Parametrized Probability Distributions

- parametrized family of probability distributions P_θ, $\theta \in \Theta$
- Kullback-Leibler divergence defines a distance:

$$\mathrm{KL}(P_{\theta'} \| \, P_\theta) = \int_{\mathbf{x}} \log \frac{p_{\theta'}(\mathbf{x})}{p_\theta(\mathbf{x})} p_{\theta'}(\mathbf{x}) \mathrm{d}\mathbf{x}$$

- Fisher information matrix $F(\theta) = \left(\nabla^2_{\theta'} \mathrm{KL}(P_{\theta'} \| \, P_\theta) \right)_{\theta'=\theta}$
- $F(\theta)$ defines a metric
- Natural Gradient: $\tilde{\nabla}_\theta P_\theta = F(\theta)^{-1} \nabla_\theta P_\theta$

Natural Gradient

$$\tilde{\nabla}_\theta P_\theta$$

- KL invariant under parameter transformations $\Rightarrow$
- Natural gradient does not depend on the parametrization
- essentially the only way to obtain this property

Derivative-Free Optimization

Given: $f : \mathbb{R}^n \to \mathbb{R}$

$\mathbf{x} \to$  $\to f(\mathbf{x})$

Goal: $\mathbf{x}^* \in \{\mathbf{x} \mid f(\mathbf{x}) - \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \leq \epsilon\}$

- oracle access to $f(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^n$
- especially no access to gradients (may not even exist)
- no knowledge of the structure of f

Generic Algorithm

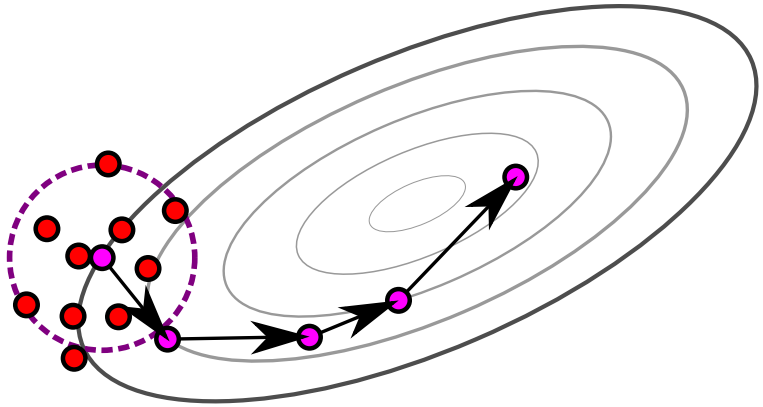
Stochastic Optimization Template

Given: parametrized distribution P_θ
initial $\theta \in \Theta$

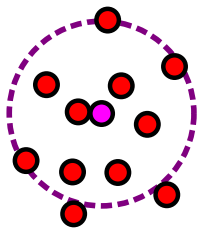
while *not happy* **do**

1. Sample $\mathbf{x}_1, \ldots, \mathbf{x}_\lambda \sim P_\theta$
2. Evaluate $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_\lambda)$
3. $\theta \leftarrow \text{Update}(\theta, \mathbf{x}_i, f(\mathbf{x}_i))$

return $\mathbb{E}_\theta[x]$



Internal Parameter



P_θ

$\theta \in \Theta$

Natural Gradient Descent

New objective $J(\theta) = \mathbb{E}[f(\mathbf{x})]\,, \quad \mathbf{x} \sim P_\theta$ to be minimized.
Consider the natural gradient

$$\tilde{\nabla}_\theta J(\theta) = \mathbb{E}\left[f(\mathbf{x})\tilde{\nabla}_\theta \log p_\theta(\mathbf{x})\right] \approx \frac{1}{\lambda}\sum_{i=1}^{\lambda} f(\mathbf{x}_i)\tilde{\nabla}_\theta \log p_\theta(\mathbf{x}_i)$$

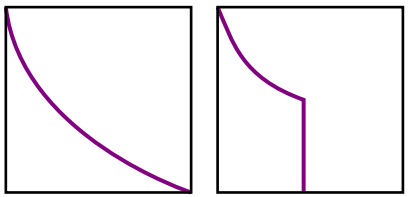
Gradient descent:

$$\theta_{k+1} = \theta_k - \eta\frac{1}{\lambda}\sum_{i=1}^{\lambda} f(\mathbf{x}_i)\tilde{\nabla}_\theta \log p_\theta(\mathbf{x}_i)$$

Information Geometric Optimization [Olliver et al.'13+]

$$J(\theta) = \mathbb{E}\left[W^f(f(\mathbf{x}))\right]$$

- **robust** to f-outliers
- **invariance** properties (order-preserving transformations of f)
- $W^f(\mathbf{x}) = w\left(\Pr[f(\mathbf{y}) \leq f(\mathbf{x}) \mid \mathbf{y} \sim P_\theta]\right)$
- w monotonously decreasing weight function
- $W^f(f(\mathbf{x}_i)) \approx w\left(\mathrm{rank}(f(\mathbf{x}_i))\right)$



NES/CMA-ES (variant)

Distribution: $P_\theta = \mathcal{N}(\mathbf{m}, C)$

Parameters: $\theta = (\mathbf{m}, C)$

Gradient: $$-\tilde{\nabla}_\theta \log p_\theta(\mathbf{x}) = \begin{bmatrix} \mathbf{x} - \mathbf{m} \\ (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T - C \end{bmatrix}$$

Weights: $\hat{w} = w(\mathrm{rank}(f(\mathbf{x}_i))$

Update:

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \eta \sum_{i=1}^{\lambda} \hat{w}_i (\mathbf{x}_i - \mathbf{m}_k)$$

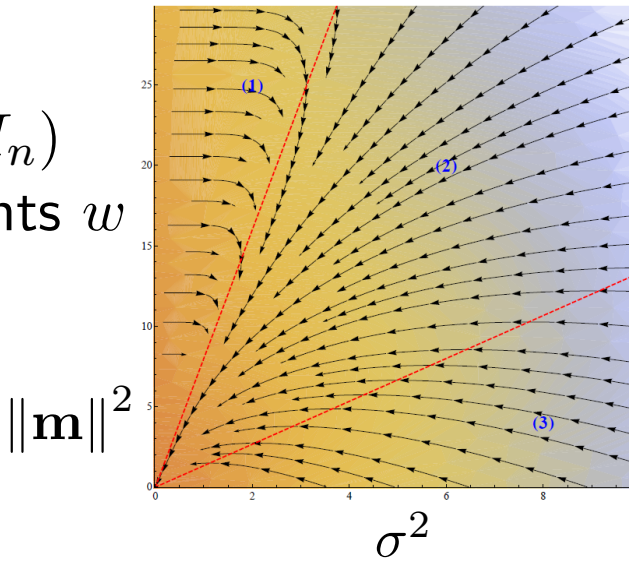
$$C_{k+1} = (1 - \eta) C_k + \eta \sum_{i=1}^{\lambda} \hat{w}_i \left((\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T \right)$$

Convergence Proofs

Note that the gradient field describes a flow

[Akimoto, Auger, Hansen, '12] IGO Flow, $P_\theta = \mathcal{N}(\mathbf{m}, C)$
on $f(\mathbf{x}) \approx \mathbf{x}^T A \mathbf{x}$ the fixed-point $\theta^*(\mathbf{x}^*, 0)$ is stable
$(\lambda \to \infty,\ \eta \to 0)$

[Schaul, '12] Radial NES, $P_\theta = \mathcal{N}(\mathbf{m}, \sigma^2 \cdot I_n)$
on $f(\mathbf{x}) = \mathbf{x}^T \mathbf{x}$ with clever chosen weights w
phase plane analysis $(\|\mathbf{m}\|^2,\ \sigma^2)$
(for dynamics: $\lambda \to \infty$)



Open Problems

- Extend the framework
  - different learning rates for the components of $\theta = (\mathbf{m}, C)$
  - more update rules/algorithms
- discrete dynamics
- "noisy" case ($\lambda \ll \infty$)
- multimodal objective functions

References


Mason, Baxter, Bartlett, Frean, Boosting Algorithms as Gradient Descent in Function Space, 1999.


Wierstra, Schaul, Peters, Schmidhuber, Natural Evolution Strategies, 2008.


Yi, Wierstra, Schaul, Schmidhuber, Stochastic Search using the Natural Gradient, 2009.


Glasmachers, Schaul, Yi, Wierstra, Schmidhuber, Exponential Natural Evolution Strategies, 2010.


Schaul, Natural Evolution Strategies Converge on Sphere Functions, 2012.


Ollivier, Arnold, Auger, Hansen, Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles, 2013+.