Information Geometric Optimization

How information theory sheds new light on black-box optimization

Anne Auger, Inria and CMAP
Main reference:

Y Ollivier, L. Arnold, A. Auger, N. Hansen, 
*Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles*, JMLR (accepted)
Black-Box Optimization

optimize $f : \Omega \mapsto \mathbb{R}$

discrete optimization $\Omega = \{0, 1\}^n$

continuous optimization $\Omega \subset \mathbb{R}^n$
Black-Box Optimization

optimize \( f : \Omega \mapsto \mathbb{R} \)

discrete optimization \( \Omega = \{0, 1\}^n \)

continuous optimization \( \Omega \subset \mathbb{R}^n \)

gradients not available or not useful
Adaptive Stochastic Black-Box Algorithm

\( \theta_t \): state of the algorithm

Sample candidate solutions

\[
X_{t+1}^i = Sol(\theta_t, U_{t+1}^i), \ i = 1, \ldots, \lambda
\]
\[
\{U_{t+1}, t \in \mathbb{N}\} \text{ i.i.d.}
\]

Evaluate solutions

\[
X_{t+1}^i \quad \text{f} \quad (X_{t+1}^i)
\]

Update state of the algorithm

\[
\theta_{t+1} = F \left( \theta_t, (X_{t+1}^1, f(X_{t+1}^1)), \ldots, (X_{t+1}^\lambda, f(X_{t+1}^\lambda)) \right)
\]
Sample candidate solutions

\[ X^i_{t+1} = Sol(\theta_t, U^i_{t+1}), i = 1, \ldots, \lambda \]

Evaluate and rank solutions

\[ f \left( X^{S(1)}_{t+1} \right) \leq \ldots \leq f \left( X^{S(\lambda)}_{t+1} \right) \]

\( S \) permutation with index of ordered solutions

Update state of the algorithm

\[ \theta_{t+1} = F \left( \theta_t, U^{S(1)}_{t+1}, \ldots, U^{S(\lambda)}_{t+1} \right) \]
Overview

1. Black-Box Optimization
   Typical difficulties

2. Information Geometric Optimization

3. Invariance

4. Recovering well-known algorithms
   CMA-ES
   PBIL, cGA
Information Geometric Optimization
Setting

- Family of probability distributions \((P_\theta)_{\theta \in \Theta}\) on \(\Omega\)

- \(\theta \in \Theta\) continuous multicomponent parameter
Information Geometric Optimization
Setting

- Family of probability distributions \((P_\theta)_{\theta \in \Theta}\) on \(\Omega\)
- \(\theta \in \Theta\) continuous multicomponent parameter
  \(\Theta\): statistical manifold

Example: \(\Omega = \mathbb{R}^n\)

\(P_\theta\) multivariate normal distribution
\(\theta = (m, C)\)
Changing Viewpoint I

- Transform original optimization problem on $\Omega$

$$\min_{x \in \Omega} f(x)$$

- Onto optimization problem on $\Theta$: Minimize

$$F(\theta) = \int f(x) P_\theta(dx)$$
Changing Viewpoint I

- Transform original optimization problem on $\Omega$
  \[
  \min_{x \in \Omega} f(x)
  \]

- Onto optimization problem on $\Theta$: Minimize
  \[
  F(\theta) = \int f(x) P_\theta(dx)
  \]

Minimizing $F \iff$ Find dirac-delta distribution concentrated on $\arg\min_x f(x)$

[Wiestra et al, 2014]
Changing Viewpoint I

- Transform original optimization problem on $\Omega$
  \[
  \min_{x \in \Omega} f(x)
  \]

- Onto optimization problem on $\Theta$ : Minimize
  \[
  F(\theta) = \int f(x)P_\theta(dx)
  \]

But not invariant to strictly increasing transformations of $f$
Changing Viewpoint II
Invariant under strictly increasing transformation of $f$

- Transform original optimization problem on $\Omega$
  \[
  \min_{x \in \Omega} f(x)
  \]

- Onto optimization problem on $\Theta$: Maximize
  \[
  J_{\theta t}(\theta) = \int W_{\theta t}^f(x) \ P_\theta(dx)
  \]
  \begin{align*}
    w(P_{\theta t}[y : f(y) \leq f(x)])
  \end{align*}
  with $w : [0, 1] \to \mathbb{R}$ decreasing weight function

**Rationale:** $f$ "small" $\iff$ $W_{\theta t}^f(x)$ "large"

[Ollivier et al.]
Maximizing $J_{\theta_t}(\theta)$

Information Geometric Optimization

- Perform natural gradient step on $\Theta$

\[
\theta^{t+\delta t} = \theta^t + \delta t \nabla_{\theta} \int W_{\theta_t}^f(x) P_{\theta}(dx)
\]
Maximizing $J_{\theta_t}(\theta)$
Information Geometric Optimization

- Perform natural gradient step on $\Theta$

$$\theta^{t+\delta t} = \theta^t + \delta t \tilde{\nabla}_\theta \int W_{\theta_t}^f(x) P_\theta(dx)$$
Natural Gradient
Fisher Information Metric

Natural gradient \( \tilde{\nabla}_\theta \) :

gradient wrt Fisher metric defined via Fisher matrix

\[
I_{ij}(\theta) = \int_x \frac{\partial \ln P_\theta(x)}{\partial \theta_i} \frac{\partial \ln P_\theta(x)}{\partial \theta_j} P_\theta(dx)
\]

\[
= - \int_x \frac{\partial^2 \ln P_\theta(x)}{\partial \theta_i \partial \theta_j} P_\theta(dx)
\]

\[
\tilde{\nabla} = I^{-1} \frac{\partial}{\partial \theta}
\]
Fisher Information Metric

Equivalently defined via second order expansion of KL

Kullback–Leibler divergence: measure of “distance” between distributions

\[ \text{KL}(P_{\theta'} \| P_{\theta}) = \int \ln \frac{P_{\theta'}(dx)}{P_{\theta}(dx)} P_{\theta}(dx) \]

Relation between KL divergence and Fisher matrix

\[ \text{KL}(P_{\theta + \delta \theta} \| P_{\theta}) = \frac{1}{2} \sum I_{ij}(\theta) \delta \theta_i \delta \theta_j + O(\delta \theta^3) \]
Natural Gradient
Fisher Information Metric

Natural gradient $\nabla_{\theta}$:

gradient wrt Fisher metric defined via Fisher matrix

$$I_{i,j}(\theta) = \int_x \frac{\partial \ln P_{\theta}(x)}{\partial \theta_i} \frac{\partial \ln P_{\theta}(x)}{\partial \theta_j} P_{\theta}(dx)$$

$$= - \int_x \frac{\partial^2 \ln P_{\theta}(x)}{\partial \theta_i \partial \theta_j} P_{\theta}(dx)$$

**intrinsic**: independent of chosen parametrization $\theta$ of $P_{\theta}$

*Fisher metric essentially the only way to obtain this property [Amari, Nagaoka, 2001]*
Maximizing $J_{\theta_t}(\theta)$
Information Geometric Optimization

- Perform natural gradient step on $\Theta$

$$
\theta^{t+\delta t} = \theta^t + \delta t \nabla_\theta \int W_{\theta_t}^f(x) P_\theta(dx)
$$
Maximizing $J_{\theta_t}(\theta)$
Information Geometric Optimization

Perform natural gradient step on $\Theta$

\[
\theta^{t+\delta t} = \theta^t + \delta t \tilde{\nabla}_\theta \int W_{\theta^t}(x) P_\theta(dx)
\]

\[
= \theta^t + \delta t \int W_{\theta^t}(x) \tilde{\nabla}_\theta \ln P_\theta(x)|_{\theta=\theta^t} P_\theta(dx)
\]
Maximizing $J_{\theta_t}(\theta)$
Information Geometric Optimization

- Perform natural gradient step on $\Theta$

$$\theta^{t+\delta t} = \theta^t + \delta t \tilde{\nabla}_\theta \int W_{\theta t}^f(x) P_\theta(dx)$$

$$= \theta^t + \delta t \int W_{\theta t}^f(x) \frac{\tilde{\nabla}_\theta P_\theta(x)}{P_{\theta t}(x)} P_{\theta t}(x)dx$$

$$= \theta^t + \delta t \int W_{\theta t}^f(x) \tilde{\nabla}_\theta \ln P_\theta(x)|_{\theta=\theta t} P_{\theta t}(dx)$$

$$= \theta^t + \delta t \int w(P_{\theta t}[y : f(y) \leq f(x)]) \tilde{\nabla}_\theta \ln P_\theta(x)|_{\theta=\theta t} P_{\theta t}(dx)$$

*does not depend on $\nabla f$*
Maximizing $J_{\theta_t}(\theta)$
Information Geometric Optimization

- Perform natural gradient step on $\Theta$

$$\theta^{t+\delta t} = \theta^t + \delta t \nabla_{\theta} \int W_{\theta_t}^f(x) P_{\theta}(dx)$$

$$= \theta^t + \delta t \int W_{\theta_t}^f(x) \frac{\nabla_{\theta} P_{\theta}(x)}{P_{\theta_t}(x)} P_{\theta_t}(x) dx$$

$$= \theta^t + \delta t \int W_{\theta_t}^f(x) \nabla_{\theta} \ln P_{\theta}(x)|_{\theta=\theta_t} P_{\theta_t}(dx)$$

$$= \theta^t + \delta t \int \omega(P_{\theta_t}[y : f(y) \leq f(x)]) \nabla_{\theta} \ln P_{\theta}(x)|_{\theta=\theta_t} P_{\theta_t}(dx)$$

1. IGO flow: $\delta t \rightarrow 0$ does not depend on $\nabla f$

2. IGO algorithms: discretization of integrals
set of continuous time trajectories in the $\Theta$- space defined by the ODE:

$$\frac{d\theta^t}{dt} = \int W^{f}_{\theta^t}(x) \tilde{\nabla}_{\theta} \ln P_{\theta}(x)|_{\theta=\theta^t} P_{\theta^t}(dx)$$
Information Geometric Optimization Algorithm
Information Geometric Optimization (IGO)

Monte Carlo Approximation of Integrals

Sample $X_i \sim P_{\theta t}$, $i = 1, \ldots N$

$$w(P_{\theta t}[y : f(y) \leq f(x)]) \approx w \left( \frac{\text{rk}(X_i)+1/2}{N} \right)$$

$$\text{rk}(X_i) = \# \{ j | f(X_j) < f(X_i) \}$$

IGO Algorithm

$$\theta^{t+\delta t} = \theta^t + \delta t \frac{1}{N} \sum_{i=1}^{N} w \left( \frac{\text{rk}(X_i) + 1/2}{N} \right) \nabla_{\theta} \ln P_{\theta}(X_i)|_{\theta=\theta^t}$$
IGO Algorithm

Monte Carlo Approximation of Integrals

Sample $X_i \sim P_{\theta^t}$, $i = 1, \ldots N$

$$w(P_{\theta^t}[y : f(y) \leq f(x)]) \approx w \left( \frac{\text{rk}(X_i) + 1/2}{N} \right)$$

IGO Algorithm

$$\theta^{t+\delta t} = \theta^t + \delta t \frac{1}{N} \sum_{i=1}^{N} w \left( \frac{\text{rk}(X_i) + 1/2}{N} \right) \nabla_{\theta} \ln P_{\theta}(X_i) |_{\theta=\theta^t}$$

$$= \theta^t + \delta t \sum_{i=1}^{N} \hat{w}_i \nabla_{\theta} \ln P_{\theta}(X_i) |_{\theta=\theta^t}$$

$$\hat{w}_i = \frac{1}{N} w \left( \frac{\text{rk}(X_i) + 1/2}{N} \right)$$

consistent estimator of integral

[Ollivier et al.]
Instantiation of IGO
Multivariate Normal Distributions

\( P_\theta \) multivariate normal distribution, \( \theta = (m, C) \)

IGO Algorithm

\[
m^{t+\delta t} = m^t + \delta t \sum_{i=1}^{N} \hat{w}_i (X_i - m^t)
\]

\[
C^{t+\delta t} = C^t + \delta t \sum_{i=1}^{N} \hat{w}_i ((X_i - m^t)(X_i - m^t)^T - C^t)
\]

Recovers the CMA-ES with rank-mu update algorithm

\( N = \lambda \)

\( \delta t \) learning rate for covariance matrix

additional learning rate for the mean

[Akimoto et al. 2010]
Instantiation of IGO
Bernoulli measures

\[ \Omega = \{0, 1\}^d \]

\[ P_\theta(x) = p_{\theta_1}(x_1) \ldots p_{\theta_d}(x_d) \] family of Bernoulli measures

Recovers

PBIL (Population based incremental learning)
[Baluja, Caruana 1995]

cGA (compact Genetic Algorithm) [Harick et al. 1999]
Conclusions

- Information Geometric Optimization framework: a unified picture of discrete and continuous optimization
- Theoretical foundations for existing algorithms
  - CMA-ES state-of-the-art in continuous bb optimization
- Some parts of CMA-ES algorithm not explained by IGO framework
  - Step-size adaptation, cumulation
- New algorithms: large-scale variant of CMA-ES based on IGO, …
References


[Hansen et al. 2003] N. Hansen, S.D. Müller, and P. Koumoutsakos, Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), ECJ 2003


