Information Geometric Optimization

How information theory sheds some 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, 2017

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

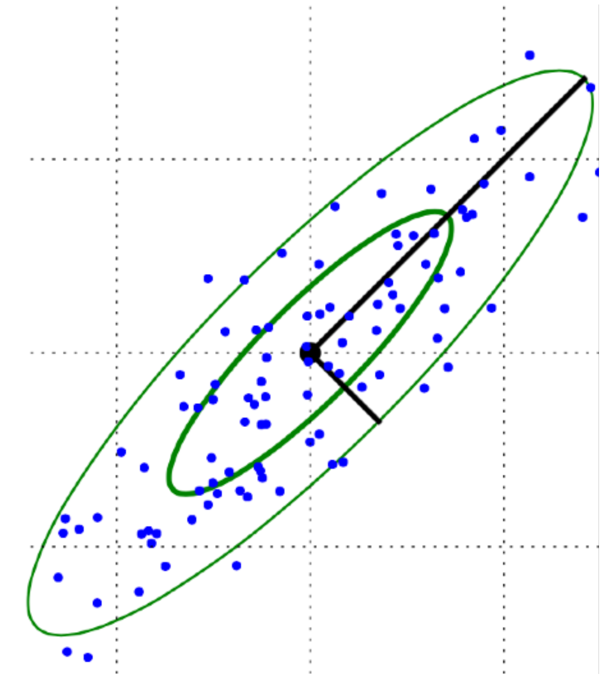
discrete optimization $\Omega = \{0, 1\}^n$

continuous optimization $\Omega \subset \mathbb{R}^n$

gradients not available or not useful

Adaptive Stochastic Black-Box Algorithm

$$\theta_t: \text{ state of the algorithm}$$



Sample candidate solutions

$$\mathbf{X}_{t+1}^i = \mathcal{S}ol(\theta_t, \mathbf{U}_{t+1}^i), i = 1, \ldots, \lambda$$

$$\{\mathbf{U}_{t+1}, t \in \mathbb{N}\} \text{ i.i.d.}$$

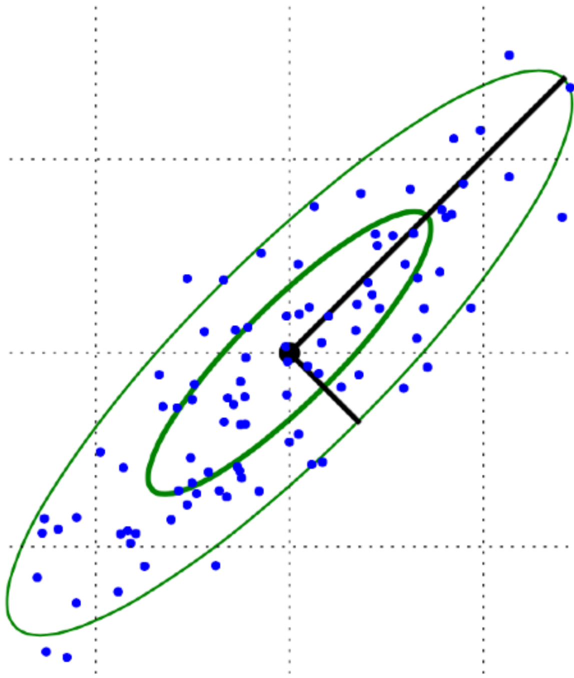
Evaluate solutions

Update state of the algorithm

$$\theta_{t+1} = \mathcal{F}\left(\theta_t, \left(\mathbf{X}_{t+1}^1, f(\mathbf{X}_{t+1}^1)\right), \ldots, \left(\mathbf{X}_{t+1}^\lambda, f(\mathbf{X}_{t+1}^\lambda)\right)\right)$$

Comparison-based Stochastic Algorithms

Invariance to strictly increasing transformations



Sample candidate solutions

$$\mathbf{X}_{t+1}^{i} = \mathcal{S}ol(\theta_t, \mathbf{U}_{t+1}^{i}), i = 1, \ldots, \lambda$$

Evaluate and rank solutions

$$f\left(\mathbf{X}_{t+1}^{\mathcal{S}(1)}\right) \leq \ldots \leq f\left(\mathbf{X}_{t+1}^{\mathcal{S}(\lambda)}\right)$$

$\mathcal{S}$ permutation with index of ordered solutions

Update state of the algorithm

$$\theta_{t+1} = \mathcal{F}\left(\theta_t, \mathbf{U}_{t+1}^{\mathcal{S}(1)}, \ldots, \mathbf{U}_{t+1}^{\mathcal{S}(\lambda)}\right)$$

Overview

❶ Black-Box Optimization

Typical difficulties

❷ Information Geometric Optimization

❸ Invariance

❹ Recovering well-known algorithms

CMA-ES

PBIL, cGA

Information Geometric Optimization
Setting

- Family of probability distributions $(P_\theta)_{\theta \in \Theta}$ on Ω
- $\theta \in \Theta$ continuous multicomponent parameter

Information Geometric Optimization
Setting

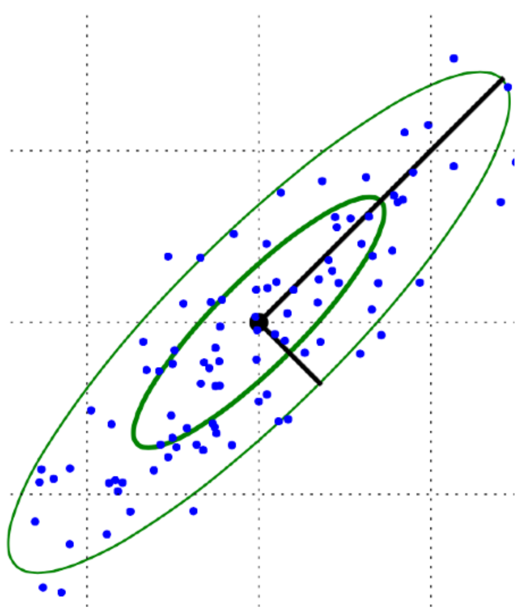
- Family of probability distributions $(P_\theta)_{\theta \in \Theta}$ on Ω
- $\theta \in \Theta$ continuous multicomponent parameter

Θ: statistical manifold

Example: $\Omega = \mathbb{R}^n$

P_θ multivariate normal distribution

$\theta = (m, C)$



Changing Viewpoint I

- Transform original optimization problem on Ω

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

- Onto optimization problem on Θ : Minimize

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

Changing Viewpoint I

- Transform original optimization problem on Ω

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

- Onto optimization problem on Θ : Minimize

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

Minimizing F $\Leftrightarrow$ Find dirac-delta distribution concentrated on $\operatorname{argmin}_x f(x)$

[Wiestra et al, 2014]

Changing Viewpoint I

- Transform original optimization problem on Ω

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

- Onto optimization problem on Θ : Minimize

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

N … tion

… $\text{in}_{\mathbf{x}} f(\mathbf{x})$

But **not** invariant to strictly increasing transformations of f

Changing Viewpoint II

Invariant under strictly increasing transformation of f

- Transform original optimization problem on Ω

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

- Onto optimization problem on Θ : Maximize

$$J_{\theta^t}(\theta) = \int \underbrace{W^f_{\theta^t}(x)}_{w(P_{\theta^t}[y\,:\,f(y) \leq f(x)])} P_\theta(dx)$$

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

Rationale: f "small" $\leftrightarrow$ $W^f_{\theta^t}(x)$ "large"

[Ollivier et al.]

Maximizing $J_{\theta_t}(\theta)$
Information Geometric Optimization

- Perform natural gradient step on Θ

$$\theta^{t+\delta t} = \theta^t + \delta t \, \widetilde{\nabla}_\theta \int W^f_{\theta^t}(x) \, P_\theta(dx)$$

Maximizing $J_{\theta_t}(\theta)$
Information Geometric Optimization

- Perform natural gradient step on Θ

$$\theta^{t+\delta t} = \theta^t + \delta t \, \widetilde{\nabla}_\theta \int W^f_{\theta^t}(x) \, P_\theta(dx)$$

Natural Gradient
Fisher Information Metric

Natural gradient $\widetilde{\nabla}_\theta$:

gradient wrt Fisher metric defined via Fisher matrix

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

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

$$\widetilde{\nabla} = I^{-1} \frac{\partial}{\partial \theta}$$

Fisher Information Metric

Equivalently defined via second order expansion of KL

Kullback–Leibler divergence:

measure of “distance” between distributions

$$\mathrm{KL}(P_{\theta'} \| P_\theta) = \int \ln \frac{P_{\theta'}(dx)}{P_\theta(dx)} P_\theta(dx)$$

Relation between KL divergence and Fisher matrix

$$\mathrm{KL}(P_{\theta+\delta\theta} \| P_\theta) = \frac{1}{2} \sum I_{ij}(\theta)\, \delta\theta_i \delta\theta_j + O(\delta\theta^3)$$

Natural Gradient
Fisher Information Metric

Natural gradient $\widetilde{\nabla}_\theta$:

gradient wrt Fisher metric defined via Fisher matrix

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

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

intrinsic: independent of chosen parametrization θ of P_θ

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 Θ

$$\theta^{t+\delta t} = \theta^t + \delta t \, \widetilde{\nabla}_\theta \int W^f_{\theta^t}(x) \, P_\theta(dx)$$

Maximizing $J_{\theta_t}(\theta)$
Information Geometric Optimization

- Perform natural gradient step on Θ

$$\begin{aligned}\theta^{t+\delta t} &= \theta^t + \delta t\, \widetilde{\nabla}_\theta \int W^f_{\theta^t}(x)\, P_\theta(dx) \\ &= \theta^t + \delta t \int W^f_{\theta^t}(x)\, \widetilde{\nabla}_\theta \ln P_\theta(x)|_{\theta=\theta^t} P_{\theta^t}(dx)\end{aligned}$$

Maximizing $J_{\theta_t}(\theta)$
Information Geometric Optimization

- Perform natural gradient step on Θ

$$
\begin{aligned}
\theta^{t+\delta t} &= \theta^t + \delta t\, \widetilde{\nabla}_\theta \int W^f_{\theta^t}(x)\, P_\theta(dx) \\
&= \theta^t + \delta t \int W^f_{\theta^t}(x)\, \frac{\widetilde{\nabla}_\theta P_\theta(x)}{P_{\theta^t(x)}} P_{\theta^t}(x) dx \\
&= \theta^t + \delta t \int W^f_{\theta^t}(x)\, \widetilde{\nabla}_\theta \ln P_\theta(x)|_{\theta=\theta^t} P_{\theta^t}(dx) \\
&= \theta^t + \delta t \int w(P_{\theta^t}[y : f(y) \leq f(x)])\, \widetilde{\nabla}_\theta \ln P_\theta(x)|_{\theta=\theta^t} P_{\theta^t}(dx)
\end{aligned}
$$

does not depend on ∇f

Maximizing $J_{\theta_t}(\theta)$
Information Geometric Optimization

- Perform natural gradient step on Θ

$$\theta^{t+\delta t} = \theta^t + \delta t\, \widetilde{\nabla}_\theta \int W^f_{\theta^t}(x)\, P_\theta(dx)$$

$$= \theta^t + \delta t \int W^f_{\theta^t}(x)\, \frac{\widetilde{\nabla}_\theta P_\theta(x)}{P_{\theta^t(x)}} P_{\theta^t}(x) dx$$

$$= \theta^t + \delta t \int W^f_{\theta^t}(x)\, \widetilde{\nabla}_\theta \ln P_\theta(x)|_{\theta=\theta^t} P_{\theta^t}(dx)$$

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

❶ IGO flow: $\delta t \to 0$ *does not depend on ∇f*

❷ IGO algorithms: discretization of integrals

IGO gradient flow
Information Geometric Optimization

set of continuous time trajectories in the Θ- space defined by the ODE:

$$\frac{d\theta^t}{dt} = \int W^f_{\theta^t}(x)\,\widetilde{\nabla}_\theta \ln P_\theta(x)\big|_{\theta=\theta^t} P_{\theta^t}(dx)$$

[Ollivier et al.]

Information Geometric Optimization Algorithm

Information Geometric Optimization (IGO)

Monte Carlo Approximation of Integrals

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

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

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

IGO Algorithm

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

IGO Algorithm

[Ollivier et al.]

Monte Carlo Approximation of Integrals

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

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

IGO Algorithm

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

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

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

consistent estimator of integral

Instantiation of IGO

Multivariate Normal Distributions

[Akimoto et al. 2010]

P_θ multivariate normal distribution, $\theta = (m, C)$

IGO Algorithm

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

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

Recovers the CMA-ES with rank-mu update algorithm

$N = \lambda$

δt learning rate for covariance matrix

additional learning rate for the mean

Instantiation of IGO
Bernoulli measures

$$\Omega = \{0,1\}^d$$
$$P_\theta(x) = p_{\theta_1}(x_1)\dots p_{\theta_d}(x_d) \text{ family of Bernoulli measures}$$

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

[Ollivier et al.] Y Ollivier, L. Arnold, A. Auger, N. Hansen, *Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles*, JMLR 2017

[Akimoto et al. 2010] Youhei Akimoto, Yuichi Nagata, Isao Ono, and Shigenobu Kobayashi *Bidirectional relation between CMA evolution strategies and natural evolution strategies,* PPSN 2010

[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

[Amari, Nagaoka 1993] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry,* 1993

[Baluja, Caruana 1995] Shumeet Baluja and Rich Caruana. *Removing the genetics from the standard genetic algorithm.* ICML, 1995.

[Harick et al. 1999] Georges R Harik, Fernando G Lobo, and David E Goldberg. *The compact genetic algorithm.* IEEE Trans EC, 1999.

[Wiestra et al, 2014] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. *Natural evolution strategies*. JMLR, 2014