A Gentle Introduction to Information Geometric Optimization (IGO)

Nikolaus Hansen
Inria
CMAP, CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, France

Presented in Dagstuhl 2022

Main Reference

Ollivier, Y., Arnold, L., Auger, A. and Hansen, N., 2017. Information-geometric optimization algorithms: A unifying picture via invariance principles. *Journal of Machine Learning Research*, *18*(18), pp.1-65.

Teaser: with invariance as a major design principle, there is a canonical way to turn any smooth parametric family of probability distributions (on an arbitrary search space) into a continuous-time black-box optimization method and into explicit “IGO algorithms” through time discretization.

Don’t

hesitate to interrupt

Context & Notations

We want to minimize

$$f : \mathcal{X} \to \mathbb{R}$$

- discrete optimization: $\mathcal{X} = \{0,1\}^n$
- continuous optimization: $\mathcal{X} = \mathbb{R}^n$

Specifically

- discrete: cGA [Harik et al 1999], PBIL [Baluja & Caruana 1995]
- continuous: Natural Evolution Strategies [Glasmachers et al 2010], major aspects of CMA-ES [Hansen et al 2003]

Context & Notations

Generic Randomized Search Template

Given $f : \mathcal{X} \to \mathbb{R}$, the *objective function*,

$P(x|\theta)$, a *parametrized distribution* on $x \in \mathcal{X}$,

θ, an initial (multi-)parameter (vector),

$\lambda \in \mathbb{N}$, a *sample size*.

While not happy

1. *Sample* distribution $P(.|\theta) \to x_1, \ldots, x_\lambda \in \mathcal{X}$
2. *Evaluate* samples on $f \to f(x_1), \ldots, f(x_\lambda)$
3. *Update parameters* $\theta \leftarrow \mathcal{U}(\theta, x_1, \ldots, x_\lambda, f(x_1), \ldots, f(x_\lambda))$

Generic Randomized Search Template

Given $f : \mathcal{X} \to \mathbb{R}$, the *objective function*,

$P(x|\theta)$, a *parametrized distribution* on $x \in \mathcal{X}$,

θ, an initial (multi-)parameter (vector),

$\lambda \in \mathbb{N}$, a *sample size*.

While not happy

1. *Sample* distribution $P(.|\theta) \to x_1, \dots, x_\lambda \in \mathcal{X}$
2. *Evaluate* samples on $f \to f(x_1), \dots, f(x_\lambda)$
3. *Update parameters*

$$\theta \leftarrow \theta + \delta t \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{w\left(\frac{\text{rank}(x_k) - 1/2}{\lambda}\right)}^{\text{preference weight for } x_k} \underbrace{\tilde{\nabla}_\theta \ln(P(x_k|\theta))}_{\theta\text{-direction of } x_k}$$

The IGO Algorithm

Context & Notations

Generic Randomized Search Template

Given $f : \mathcal{X} \to \mathbb{R}$, the *objective function*,

$P(x|\theta)$, a *parametrized distribution* on $x \in \mathcal{X}$,

θ, an initial (multi-)parameter (vector),

$\lambda \in \mathbb{N}$, a *sample size*.

While not happy

1. *Sample* distribution $P(.|\theta) \to x_1, \ldots, x_\lambda \in \mathcal{X}$
2. *Evaluate* samples on $f \to f(x_1), \ldots, f(x_\lambda)$
3. *Update parameters* $\theta \leftarrow \mathcal{U}(\theta, x_1, \ldots, x_\lambda, f(x_1), \ldots, f(x_\lambda))$

This suggests a…

Change of Viewpoint

- Instead of the original problem

$$\underset{x\in\mathcal{X}}{\arg\min}\, f(x)$$

- we consider the problem of finding the **arg min over** θ of

$$F : \theta \mapsto \mathsf{E}_x[f(x)\,|\,\theta]$$

Equivalence of Solutions

The optimal distribution

$$P(.\,|\,\underset{\theta}{\arg\min}\, F(\theta))$$

is the Dirac delta distribution in the solution $\arg\min_{x\in\mathcal{X}} f(x)$.

$$\arg\min_{x\in\mathcal{X}} f(x)$$

- we consider the problem of finding the **arg min over** θ of

$$F : \theta \mapsto \mathsf{E}_x[f(x) \,|\, \theta]$$

Assume that

- θ is from now on a continuous multi-parameter, for example $[\theta]_i = \Pr([x]_i = 1 \,|\, \theta) = \mathsf{E}([x]_i \,|\, \theta)$ when x is binary

- yet, even when θ is discrete (not considered here), we can construct an IGO update!

based on minimizing a weighted sum of a cross entropy and a "cross preference"

WTF—This Has Nothing to do With EC

Given θ is a continuous parameter(-vector), to optimize F iteratively, we could do a gradient descent on F (with step-size δt):

$$\theta^{t+\delta t} = \theta^t + \delta t\, \nabla_\theta F(\theta^t) \quad \text{or}$$

$$\begin{aligned}\theta^{t+\delta t} - \theta^t &= \delta t\, \nabla_\theta F(\theta^t) \\ &= \delta t\, \nabla_\theta \mathsf{E}_x[f(x)\,|\,\theta^t] \\ &= \dots \\ &= \delta t\, \mathsf{E}_x[f(x)\,\overbrace{\nabla_\theta \ln(P(x|\theta^t))}^{\text{an element of } \theta\text{-space}}\,|\,\theta^t]\end{aligned}$$

WTF—This Has Nothing to do With EC

$$\begin{aligned}
\theta^{t+\delta t} - \theta^t &= \delta t\, \nabla_\theta F(\theta^t) \\
&= \delta t\, \nabla_\theta \mathsf{E}_x[f(x)\,|\,\theta^t] \\
&= \delta t\, \nabla_\theta \int f(x) P(x|\theta^t) \mathrm{d}x \\
&= \delta t \int f(x) \nabla_\theta P(x|\theta^t) \mathrm{d}x \\
&= \delta t \int f(x) P(x|\theta^t) \nabla_\theta \ln(P(x|\theta^t)) \mathrm{d}x \\
&= \delta t\, \mathsf{E}_x[f(x)\, \underbrace{\nabla_\theta \ln(P(x|\theta^t))}_{\text{an element of } \theta\text{-space}}]
\end{aligned}$$

WTF—This Has Nothing to do With EC

Given θ is a continuous parameter(-vector), to optimize F iteratively, we could do a gradient descent on F (with step-size δt):

$$\theta^{t+\delta t} = \theta^t + \delta t \, \nabla_\theta F(\theta^t) \quad \text{or}$$

$$\begin{aligned} \theta^{t+\delta t} - \theta^t &= \delta t \, \nabla_\theta F(\theta^t) \\ &= \delta t \, \nabla_\theta \mathsf{E}_x[f(x) \,|\, \theta^t] \\ &= \dots \\ &= \delta t \, \mathsf{E}_x[f(x) \overbrace{\nabla_\theta \ln(P(x|\theta^t))}^{\text{an element of } \theta\text{-space}} \,|\, \theta^t] \end{aligned}$$

$$\begin{aligned}\theta^{t+\delta t} - \theta^t &= \delta t \, \nabla_\theta F(\theta^t) \\ &= \delta t \, \nabla_\theta \mathsf{E}_x[f(x) \,|\, \theta^t] \\ &= \ldots \\ &= \delta t \, \mathsf{E}_x[f(x) \overbrace{\nabla_\theta \ln(P(x|\theta^t))}^{\text{an element of } \theta\text{-space}} \,|\, \theta^t]\end{aligned}$$

- does not depend on ∇f,
- describes a gradient flow when $\delta t \to 0$ by the ODE
$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \int f(x) P(x \,|\, \theta) \, \nabla_\theta \ln(P(x \,|\, \theta)) \, \mathrm{d}x,$$
- describes a randomized search algorithm when the expected value is estimated as the average of a (small) sample,
- as we have chosen P to begin with, we might “know” $\nabla_\theta \ln(P(x \,|\, \theta))$, or even have a simple expression for it.

$$
\begin{aligned}
\theta^{t+\delta t} - \theta^{t} &= \delta t \, \nabla_\theta F(\theta^t) \\
&= \delta t \, \nabla_\theta \mathsf{E}_x[f(x) \,|\, \theta^t] \\
&= \dots \\
&= \delta t \, \mathsf{E}_x[f(x) \overbrace{\nabla_\theta \ln(P(x|\theta^t))}^{\text{an element of } \theta\text{-space}} \,|\, \theta^t]
\end{aligned}
$$

not so fast!

Thank you for you attention

$$\theta^{t+\delta t} - \theta^t = \delta t \, \nabla_\theta F(\theta^t)$$
$$= \delta t \, \nabla_\theta \mathsf{E}_x[f(x) \,|\, \theta^t]$$
$$= \ldots$$
$$= \delta t \, \mathsf{E}_x[f(x) \overbrace{\nabla_\theta \ln(P(x|\theta^t))}^{\text{an element of } \theta\text{-space}} \,|\theta^t]$$

not so fast!

(I) has an expectation (which seems “impractical” as algorithm)

(II) is ***not** a candidate* to describe update equations of algorithms which are invariant under order preserving f-transformations

(III) the (vanilla) gradient depends on how we parameterize P in θ (we may parameterize a probability $\in [0,1]$ by its logit value $\log(p/(1-p)) \in [-\infty, \infty]$)

$$= \delta t \, \mathsf{E}_x[f(x) \overbrace{\nabla_\theta \ln(P(x|\theta^t))} \, |\theta^t]$$

(I) has an expectation (which seems impractical as “algorithm”)

(II) is ***not** a candidate* to describe update equations of algorithms which are invariant under order preserving f-transformations

(III)the (vanilla) gradient depends on how we parameterize P in θ (we may parameterize a probability $\in [0,1]$ by its logit value $\log(p/(1-p)) \in [-\infty, \infty]$)

We address the three points in reverse order.

(III) to remove parametrization dependency we *replace the (vanilla) gradient* with the natural gradient

An Intrinsic Gradient

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

gradient w.r.t. the Fisher metric defined via Fisher matrix

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

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

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

intrinsic: independent of the parametrization of P in θ

the Fisher metric essentially the only way to obtain this property [Amari, Nagaoka, 2001]

An Intrinsic Gradient

Simply put, the natural gradient is (among all other gradients) the (unique) gradient that

- maximizes $||\nabla F||$, i.e. how much F improves when moving in gradient direction, under a given KL change of P, or
- minimizes, for a *given* improvement $||\nabla F||$, the KL change of P
- makes the "truly" smallest change with the largest effect.

"truly", because we use KL to measure change size

Algorithm Iteration Update After Fixing (III)

Given

$$F(\theta) := \int f(x)P(x|\theta)\mathrm{d}x = \mathsf{E}_x[f(x)\,|\,\theta]$$

to be minimized, we update θ like

$$\begin{aligned}\theta^{t+\delta t} - \theta^t &= \delta t\,\widetilde{\nabla}_\theta F(\theta^t)\\ &= \delta t \widetilde{\nabla}_\theta \mathsf{E}_x[f(x)\,|\,\theta^t]\\ &= \delta t\,\mathsf{E}_x[f(x)\widetilde{\nabla}_\theta \ln(P(x|\theta^t))\,|\,\theta^t]\end{aligned}$$

$$= \delta t \, \mathsf{E}_x[f(x) \widetilde{\nabla}_\theta \ln(P(x|\theta^t)) \,|\, \theta^t]$$

(I) has an expectation (which seems impractical as "algorithm")

(II) is ***not*** *a candidate* to describe update equations of algorithms which are invariant under order preserving f-transformations

✔ (III) the (vanilla) gradient depends on how we parameterize P in θ (we may parameterize a probability $\in [0,1]$ by its logit value $\log(p/(1-p)) \in [-\infty, \infty]$)

We address the three points in reverse order.

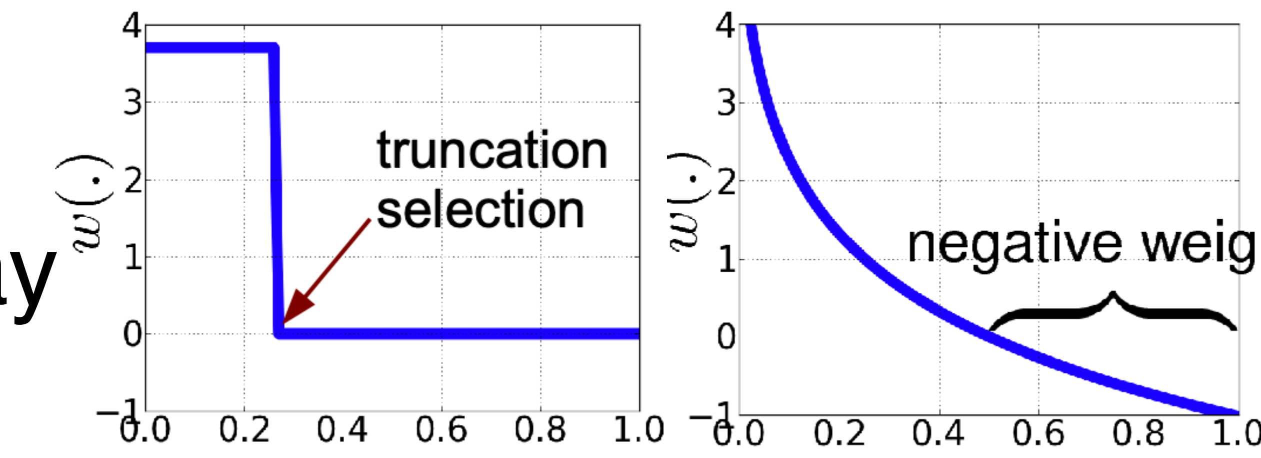
(II) Invariance under order-preserving f-transformations

$$\theta^{t+\delta t} - \theta^t = \delta t \widetilde{\nabla}_\theta \mathsf{E}_x[w^t(x) \,|\, \theta^t]$$

To obtain invariance to order-preserving f-transformations, we compose $f : \mathscr{X} \to \mathbb{R}$ like

$$w^t := w \circ \mathsf{CDF}^t \circ f : \mathcal{X} \to \mathbb{R}$$

where



☑ $w : [0,1] \to \mathbb{R}$ is a decreasing weight function (an algorithm parameter, say $w : z \mapsto 1/2 - z$ or $z \mapsto \mathrm{sign}(1/2 - z)$), and

☑ CDF^t is the cumulative distribution function of $f(x)$ when x is sampled according to $P(\,.\,|\,\theta^t)$:

$$\mathsf{CDF}^t : \mathbb{R} \to [0,1]$$
$$z \mapsto \Pr(f(y) \le z, y \sim P(.|\theta^t))$$

[Ollivier et al 2017]

☑ CDF^t is the cumulative distribution function of $f(x)$ when x is sampled according to $P(\,.\,|\,\theta^t)$:

$$\mathrm{CDF}^t : \mathbb{R} \to [0,1]$$
$$z \mapsto \Pr(f(y) \leq z, y \sim P(.|\theta^t))$$

- gives the *quantile* in [0,1] for the measured f value relative to the current distribution of $f(x)$ (where $x \sim P(\,.\,|\,\theta^t)$).

this is **the** *key feature of the construction of w^t replacing f in the definition of F*

- $\mathrm{CDF}(f) = 0$ is optimal, all other possible values are larger = worse.
- $\mathrm{CDF} \circ f$ is invariant under order-preserving transformations
- $w_t \equiv w \circ \mathrm{CDF} \circ f$ is to be maximized (w switches the sign by convention)

Algorithm Iteration Update After Fixing (II)-(III)

We replace $F : \theta \mapsto \mathsf{E}_x[f(x)\,|\,\theta]$ to be minimized with

$$W : \theta \mapsto \int w^t(x)P(x|\theta)\mathrm{d}x = \mathsf{E}_x[w^t(x)|\theta] \qquad w^t \equiv w \circ \mathsf{CDF}^t \circ f$$

to be maximized **for a fixed w^t**. Then we update θ^t like

$$\begin{aligned}\theta^{t+\delta t} - \theta^t &= \delta t\, \widetilde{\nabla}_\theta W(\theta^t) \\ &= \delta t \widetilde{\nabla}_\theta \mathsf{E}_x[w^t(x)\,|\,\theta^t] \\ &= \delta t\, \mathsf{E}_x[w^t(x) \widetilde{\nabla}_\theta \ln(P(x|\theta^t))\,|\,\theta^t]\end{aligned}$$

$$\theta^{t+\delta t} - \theta^t = \delta t\, \widetilde{\nabla}_\theta W(\theta^t)$$
$$= \delta t \widetilde{\nabla}_\theta \mathsf{E}_x[w^t(x)\,|\,\theta^t] \qquad x \sim P(.|\theta^t)$$
$$= \delta t\, \mathsf{E}_x[w^t(x) \widetilde{\nabla}_\theta \ln(P(x|\theta^t))\,|\,\theta^t]$$

Two final steps need to be taken to approximate E and w^t.

- Replace the expectation with an average
- We already knew the consistent approximation of the CDF^t in

$$w^t \equiv w \circ \mathsf{CDF}^t \circ f$$

because the constructing of w^t was based on its approximation.

Obtaining an IGO Algorithm

We replace the expected value with an average and replace

$$w^t \equiv w \circ \mathrm{CDF}^t \circ f$$

with the approximation

$$w\left(\frac{\mathrm{rank}(x_k) - 1/2}{\lambda}\right), \qquad k = 1, \ldots, \lambda$$

such that

$$\theta^{t+\delta t} - \theta^t = \delta t \int w^t(x) P(x|\theta^t)\, \widetilde{\nabla}_\theta \ln(P(x|\theta^t)) \mathrm{d}x$$

$$\approx \delta t \frac{1}{\lambda} \sum_{k=1}^{\lambda} w\left(\frac{\mathrm{rank}(x_k) - 1/2}{\lambda}\right) \underbrace{\widetilde{\nabla}_\theta \ln(P(x_k|\theta^t))}_{\text{an element of } \theta\text{-space}}, \quad x_k \sim P(.|\theta^t)$$

IGO Summary

- IGO flow, time continuous infinite population size model of the θ-change of a black-box EDA

$$\frac{\theta^{t+\delta t} - \theta^t}{\delta t} = \mathsf{E}_x[w^t(x)\widetilde{\nabla}_\theta \ln(P(x|\theta^t))] \qquad \text{where } w^t := w \circ \mathsf{CDF}^t \circ f$$

- IGO algorithm

$$x_k \sim P(.|\theta^t),\ k = 1, \ldots, \lambda$$

$$\theta^{t+\delta t} - \theta^t = \delta t \frac{1}{\lambda} \sum_{k=1}^{\lambda} \underbrace{w\left(\frac{\mathrm{rank}(x_k) - 1/2}{\lambda}\right)}_{\text{preference weight for } x_k} \underbrace{\widetilde{\nabla}_\theta \ln(P(x_k|\theta^t))}_{\theta\text{-direction to reinforce } x_k}$$

The IGO algorithm still depends on the parametrization of P (for $\delta t > 0$)!

the IGO-ML algorithm is the IGO algorithm that does not depend on the parametrization for $\delta t > 0$

The Common Theme: Invariance

- IGO flow is (for any $P(\,.\,|\,\theta)$ guarantied to be) invariant under
  - re-parameterization of the probability distribution
  - re-parametrization of the search space (provided the distribution family remains the same)

    for example, under exchanging zeros and ones for the Bernoulli distribution family or under affine transformations for the Gaussian distribution family
  - order-preserving f-transformations
- IGO algorithms are invariant under order-preserving f-transformations and as invariant as the flow at least for $\delta t \to 0$

Compact GA is an IGO Algorithm

$$x_k \sim P(.|\theta^t)$$

$$\theta^{t+\delta t} - \theta^t = \delta t \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{w\left(\frac{\mathrm{rank}(x_k) - 1/2}{\lambda}\right)}^{\text{preference weight for } x_k} \underbrace{\widetilde{\nabla}_\theta \ln(P(x_k|\theta^t))}_{\text{intrinsic } \theta\text{-direction to reinforce } x_k}$$

$$= \delta t \frac{1}{2}(x_{1:2} - \theta^t + -1(x_{2:2} - \theta^t))$$

$$= \delta t \frac{1}{2}(x_{1:2} - x_{2:2})$$

Define:

P : n-dimensional Bernoulli

$w : z \mapsto \mathrm{sign}(1/2 - z)$

$\lambda = 2$

δt

$\theta_i := \Pr(x_i = 1), i = 1 \dots n$

Compute:

$$I_{ii}^{-1} = \theta_i(1 - \theta_i)$$

$$\frac{\partial \ln P(x|\theta)}{\partial \theta_i} = \frac{x_i}{\theta_i} - \frac{1 - x_i}{1 - \theta_i}$$

$$[\widetilde{\nabla}_\theta \ln(P(x|\theta^t))]_i = I_{ii}^{-1} \frac{\partial \ln P(x|\theta^t)}{\partial \theta_i}$$

$$= \theta_i^t(1 - \theta_i^t)\left(\frac{x_i}{\theta_i^t} - \frac{1 - x_i}{1 - \theta_i^t}\right)$$

$$= x_i - \theta_i^t$$

Proving Convergence

- based on the time continuous model
  - Convergence of the flow [Akimoto et al 2012][Glasmachers 2012] [Ollivier et al 2017]
  - For geometric convergence of the IGO algorithm, we may show that stochastic deviations introduced by a finite population size and a finite step-size are small enough; we can even obtain convergence rates [Akimoto et al 2022]

The flow does not well reflect the behavior of Evolution Strategies.

Limitations

Currently poorly covered by the IGO framework:

- step-sizes and learning rates

  for example in CMA-ES, (i) different θ-parameters have different learning rates δt, and (ii) the step-size update is not an IGO algorithm

- cumulation / exponential smoothing / iterate averaging / momentum

  act as low pass filter or learning rate modulator

Generic Randomized Search Template

Given $f : \mathcal{X} \to \mathbb{R}$, the *objective function*,
$P(x|\theta)$, a *parametrized distribution* on $x \in \mathcal{X}$,
θ, an initial (multi-)parameter (vector),
$\lambda \in \mathbb{N}$, a *sample size*.

While not happy

1. *Sample* distribution $P(.|\theta) \to x_1, \dots, x_\lambda \in \mathcal{X}$
2. *Evaluate* samples on $f \to f(x_1), \dots, f(x_\lambda)$
3. *Update parameters* $\theta \leftarrow \mathcal{U}(\theta, x_1, \dots, x_\lambda, f(x_1), \dots, f(x_\lambda))$

Generic Randomized Search Template

Given $f : \mathcal{X} \to \mathbb{R}$, the *objective function*,

$P(x|\theta)$, a *parametrized distribution* on $x \in \mathcal{X}$,

θ, an initial (multi-)parameter (vector),

$\lambda \in \mathbb{N}$, a *sample size*.

While not happy

1. *Sample* distribution $P(.|\theta) \to x_1, \ldots, x_\lambda \in \mathcal{X}$
2. *Evaluate* samples on $f \to f(x_1), \ldots, f(x_\lambda)$
3. *Update parameters*

$$\theta \leftarrow \theta + \delta t \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{w\left(\frac{\text{rank}(x_k) - 1/2}{\lambda}\right)}^{\text{preference weight for } x_k} \underbrace{\tilde{\nabla}_\theta \ln(P(x_k|\theta))}_{\theta\text{-direction of } x_k}$$

The IGO Algorithm

A parametrization- and gradient-free IGO Algorithm

Definition 11 (IGO-ML algorithm) *The* IGO-ML algorithm *with step size* δt *updates the value of the parameter* θ^t *according to*

$$\theta^{t+\delta t} = \arg\max_{\theta} \left\{ \left(1 - \delta t \sum_i \widehat{w}_i\right) \int \ln P_\theta(x)\, P_{\theta^t}(\mathrm{d}x) \;+\; \delta t \sum_i \widehat{w}_i \ln P_\theta(x_i) \right\} \qquad (30)$$

where $x_1, \ldots, x_N$ *are sample points drawn according to the distribution* P_{θ^t}*, and* $\widehat{w}_i$ *is the weight* (14) *obtained from the ranked values of the objective function* f*.*

[Ollivier et al 2017]