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

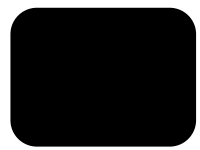
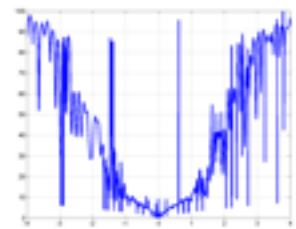
http://www.inria.fr http://www.lri.fr/~hansen

Problem Statement: Black-Box Optimization

Given an objective function

$$f : \mathcal{X} \cancel{\subset \mathbb{R}^n} \to \mathbb{R}$$

Minimize f in a *black-box scenario* (direct search, no gradients)

$x \longrightarrow$  $\longrightarrow f(x)$ 

Objective

- convergence to a global essential infimum of f as fast as possible

  linear convergence, $\mathcal{O}(n \log 1/\epsilon)$ black-box evaluations

- find $x \in \mathcal{X}$ with small $f(x)$ value using as few black-box calls as possible

The black box can

- be non-convex, multi-modal/rugged, discontinuous, noisy, dynamic
- take from milli-seconds to hours to evaluate

An Algorithm Template

Generic Randomized Search Template

Given: the *objective function*, $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$,
Choose: a *parametrized (search) distribution* on $\mathcal{X}$, $P(x|\theta)$,
an *initial* distribution, θ_0, and a *sample size*, $\lambda \in \mathbb{N}$.

for $t = 0, 1, 2, \ldots$

1. *Sample* distribution $P(x|\theta_t) \to x_1, \ldots, x_\lambda$
2. *Evaluate* samples on $f \to f(x_1), \ldots, f(x_\lambda)$
3. *Update parameters* $\theta_{t+1} = F(\theta_t, x_1, \ldots, x_\lambda, f(x_1), \ldots, f(x_\lambda))$

Open questions

- choice of $P(x|\theta)$
- choice of update function F

} algorithm design

- choice of λ and θ_0

} users choice

An Algorithm Template

Generic Randomized Search Template

Given: the *objective function*, $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$,
Choose: a *parametrized (search) distribution* on $\mathcal{X}$, $P(x|\theta)$,
an *initial* distribution, θ_0, and a *sample size*, $\lambda \in \mathbb{N}$.

for $t = 0, 1, 2, \ldots$

1. *Sample* distribution $P(x|\theta_t) \to x_1, \ldots, x_\lambda$
2. *Evaluate* samples on $f \to f(x_1), \ldots, f(x_\lambda)$
3. *Update parameters* $\theta_{t+1} = F(\theta_t, x_1, \ldots, x_\lambda, f(x_1), \ldots, f(x_\lambda))$

Open questions

- choice of $P(x|\theta)$ — **CMA-ES: family of multivariate normal distributions**
- choice of update function F
- choice of λ and θ_0 } users choice

An Algorithm Template

Generic Randomized Search Template

Given: the *objective function*, $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$,
Choose: a *parametrized (search) distribution* on $\mathcal{X}$, $P(x|\theta)$,
an *initial* distribution, θ_0, and a *sample size*, $\lambda \in \mathbb{N}$.

for $t = 0, 1, 2, \ldots$

1. *Sample* distribution $P(x|\theta_t) \to x_1, \ldots, x_\lambda$
2. *Evaluate* samples on $f \to f(x_1), \ldots, f(x_\lambda)$
3. *Update parameters* $\theta_{t+1} = F(\theta_t, x_1, \ldots, x_\lambda, f(x_1), \ldots, f(x_\lambda))$

Open questions

- choice of $P(x|\theta)$ — **BFGS: family of Dirac distributions**
- choice of update function F
- choice ~~of λ and~~ θ_0 } users choice

...a new search problem on θ ...

A new search problem

The algorithm template replaces the original search problem (defined in $\mathcal{X}$-space),

$$\arg\min_{x} \left(f(x)\right)$$

with a new search problem in Θ-space, the "stochastic relaxation"

$$\arg\max_{\theta} (J(\theta)) \quad \text{where} \quad J(\theta) = \mathrm{E}_{x\sim p(.|\theta)}\left[W_{\theta_t}(f(x))\right],$$

where W_{θ_t} is monotonous *decreasing*.

think of $W(f(x))$ as $-f(x)$ for the time being

Both problems have the same solution (same optimum):

$$P(x|\theta^*) = \delta(x - x^*) \quad \text{for all } W_{\theta_t}$$

i.e., $\Pr(x = x^* \,|\, \theta = \theta^*) = 1$.

A new search problem

The algorithm template replaces the original search problem (defined in $\mathcal{X}$-space),

$$\arg\min_x (f(x))$$

with a new search problem in Θ-space, the ”stochastic relaxation”

$$\arg\max_\theta (J(\theta)) \quad \text{where} \quad \boxed{J(\theta) = \mathrm{E}_{x \sim p(.|\theta)} \left[W_{\theta_t}(f(x))\right]},$$

where W_{θ_t} is monotonous *decreasing*.

think of $W(f(x))$ as $-f(x)$ for the time being

Both problems have the same solution (same optimum):

$$P(x|\theta^*) = \delta(x - x^*) \quad \text{for all } W_{\theta_t}$$

i.e., $\Pr(x = x^* \mid \theta = \theta^*) = 1$.

A new search problem

The algorithm template replaces the original search problem (defined in $\mathcal{X}$-space),

$$\arg\min_{x} \left(f(x)\right)$$

with a new search problem in Θ-space, the ”stochastic relaxation”

$$\arg\max_{\theta}(J(\theta)) \quad \text{where} \quad \boxed{J(\theta) = \mathrm{E}_{x \sim p(.|\theta)}\left[W_{\theta_t}(f(x))\right]},$$

where W_{θ_t} is monotonous *decreasing*.

think of $W(f(x))$ as $-f(x)$ for the time being

Both problems have the same solution (same optimum):

$$P(x|\theta^*) = \delta(x - x^*) \quad \text{for all } W_{\theta_t}$$

i.e., $\Pr(x = x^* \mid \theta = \theta^*) = 1$.

Time-continuous gradient flow

To improve J, we will consider the gradient

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

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

inducing the time-continuous gradient flow

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

IGO on one slide

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

$$\underbrace{\theta_{t+1} - \theta_t}_{\downarrow \; \frac{\partial}{\partial t}\theta_t \; (\lambda \to \infty \text{ and } \delta t \to 0)} = \eta \frac{1}{\lambda}\sum_{k=1}^{\lambda} \overbrace{W(f(x_k))}^{\text{preference weight}} \underbrace{\nabla_\theta \log p(x_k|\theta)\Big|_{\theta=\theta_t}}_{\text{direction for } x_k}, \quad x_k \sim p(.|\theta_t)$$

IGO on one slide

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

$$\underbrace{\theta_{t+1} - \theta_t}_{\downarrow \; \frac{\mathrm{d}}{\mathrm{d}t}\theta_t \; (\lambda \to \infty \text{ and } \delta t \to 0)} = \eta \frac{1}{\lambda} \sum_{k=1}^{\lambda} \overbrace{W(f(x_k))}^{\text{preference weight}} \underbrace{\nabla_\theta \log p(x_k|\theta)\Big|_{\theta=\theta_t}}_{\text{direction for } x_k}, \quad x_k \sim p(.|\theta_t)$$

The update

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

$\frac{\mathrm{d}}{\mathrm{d}t}\theta_t$ ($\lambda \to \infty$ and $\delta t \to 0$)

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

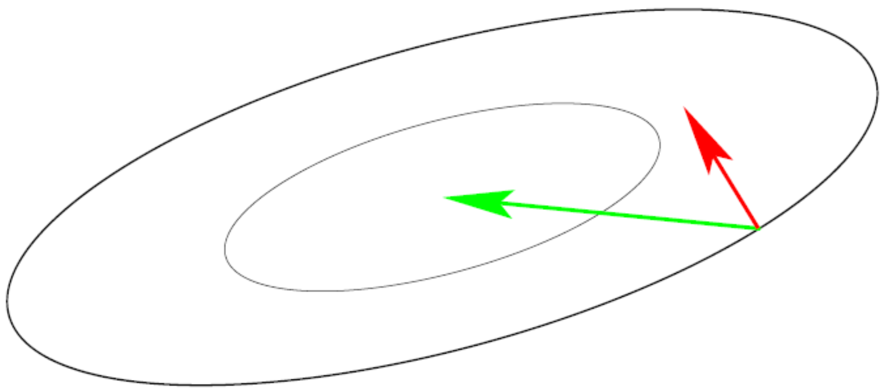
- why the Euclidean metric?
- which parametrization of p in θ?
- why not second order?

$\Longrightarrow$ invariance is a major design principle

Unique Steepest Ascent

a gradient ∇_θ is defined via a "small" change of θ, that is,
a small change of the probability distribution

what is "small" (what is the appropriate metric)?



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

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

- gradient and Newton direction use a different inner product or metric to define the "gradient":

$$\langle x, y\rangle = x^T I y \quad \text{versus} \quad \langle x, y\rangle_H = x^T H y$$

- only the Newton direction is invariant under *affine* coordinate transformations, hence *distinguished*

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)**:
  $W(f(x_i)) \leq W(f(x_j)) \iff f(x_j) \leq f(x_i)$ and
- **invariant**: there exists a monotonous decreasing weight function $w : \mathbb{R} \to \mathbb{R}$, s.t. for $x \sim p(.|\theta)$

$$W(f(x)) \sim w(\mathcal{U}[0,1]) \quad \forall f, \theta \;,$$

that is, the distribution of $W(f(x))$ values is independent of θ and f and only depends on a weight parameter w

resolving the question *how* they should depend on f and θ

Defining W

We define

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

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

as

- the cumulative distribution function of $f(X)$
- the probability to get below value y 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
- results in "rank-based selection" (invariant to t)
- 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

Defining W

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

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

as

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

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

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

- is invariant under monotone f-transformations
- results in "rank-based selection" (invariant to t)
- 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

Approximation of $W^f_{\theta_t}$

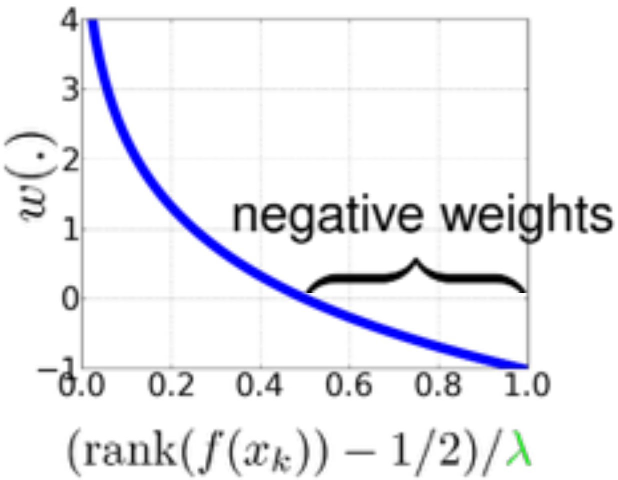
A consistent approximation for

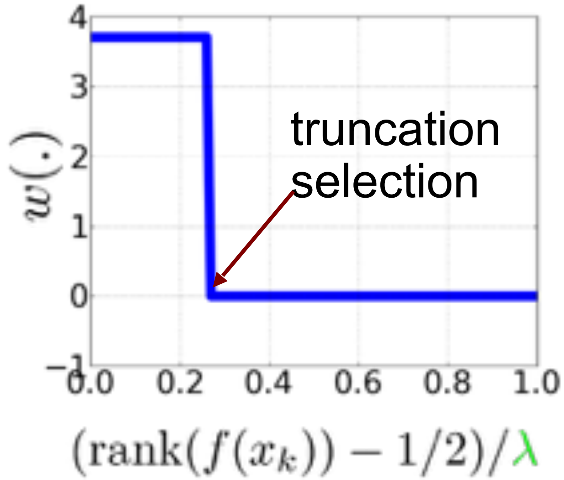
$$W^f_{\theta_t}(f(x_k)) = w(\Pr(f(X) \le f(x_k), X \sim p(.|\theta_t)))$$

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

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

for $k = 1, \dots, \lambda$, where w is monotonuously decreasing, 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

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

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 parametrization

let $x, x_k \sim p(.|\theta_t)$, then

$$\theta_{t+1}$$

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

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

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

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 and a (parametrized) probability distribution on an arbitrary search domain

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

Known algorithms that follow this derivation have been quite successful, in particular the Covariance Matrix Adaptation Evolution Strategy (CMA-ES)