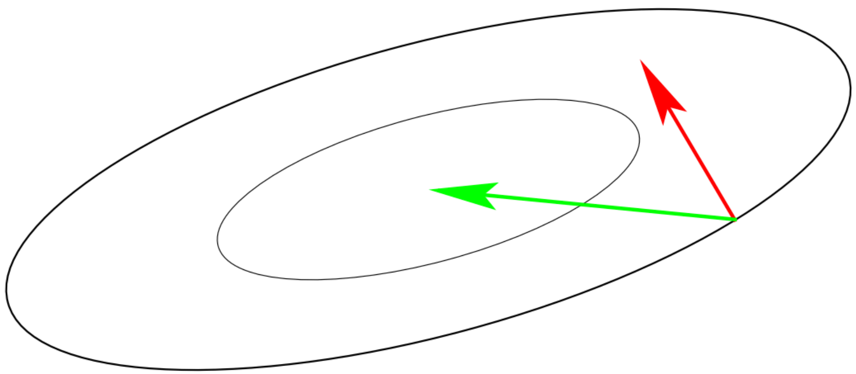
Ill-conditioned Problems

consider the curvature of the level sets of a function

ill-conditioned means "squeezed" lines of equal function value (high curvatures)



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

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

Condition number equals nine here. Condition numbers up to 10^{10} are not unusual in real world problems.

Part II: Algorithms

Landscape of Derivative Free Optimization Algorithms

Deterministic Algorithms

Quasi-Newton with estimation of gradient (BFGS) [Broyden et al. 1970]
Simplex downhill [Nelder and Mead 1965]
Pattern search, Direct Search [Hooke and Jeeves 1961]
Trust-region/Model Based methods (NEWUOA, BOBYQA) [Powell, 06,09]

Stochastic (randomized) search methods

Evolutionary Algorithms (continuous domain)

- Differential Evolution [Storn, Price 1997]
- Particle Swarm Optimization [Kennedy and Eberhart 1995]
- **Evolution Strategies, CMA-ES** [Rechenberg 1965, Hansen, Ostermeier 2001]
- Estimation of Distribution Algorithms (EDAs) [Larrañaga, Lozano, 2002]
- Cross Entropy Method (same as EDAs) [Rubinstein, Kroese, 2004]
- Genetic Algorithms [Holland 1975, Goldberg 1989]

Simulated Annealing [Kirkpatrick et al. 1983]

A Generic Template for Stochastic Search

Define $\{P_\theta : \theta \in \Theta\}$, a family of probability distributions on $\mathbb{R}^n$

Generic template to optimize $f : \mathbb{R}^n \to \mathbb{R}$

Initialize distribution parameter θ, set population size $\lambda \in \mathbb{N}$

While not terminate

1. Sample $x_1, \ldots, x_\lambda$ according to P_θ
2. Evaluate $x_1, \ldots, x_\lambda$ on f
3. Update parameters $\theta \leftarrow F(\theta, x_1, \ldots, x_\lambda, f(x_1), \ldots, f(x_\lambda))$

the update of θ should drive P_θ to concentrate on the optima of f

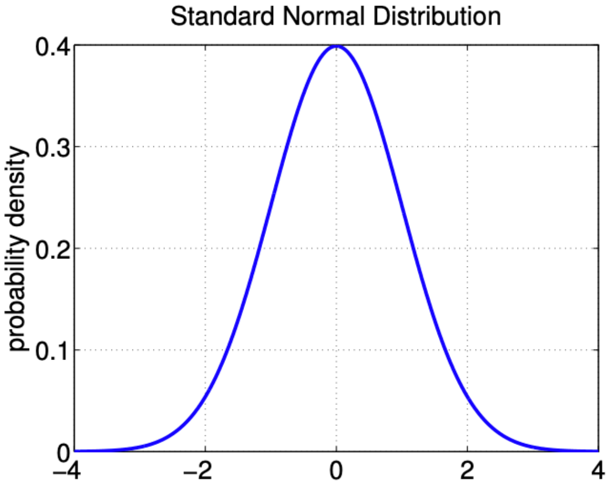
To obtain an optimization algorithm we need:

❶ to define $\{P_\theta, \theta \in \Theta\}$

❷ to define F the update function of θ

Which probability distribution to sample candidate solutions?

Normal distribution - 1D case



probability density of the 1-D standard normal distribution $\mathcal{N}(0,1)$

(expected (mean) value, variance) = (0,1)

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

General case

- Normal distribution $\mathcal{N}(\boldsymbol{m}, \sigma^2)$

  (expected value, variance) = $(\boldsymbol{m}, \sigma^2)$

  density: $p_{\boldsymbol{m},\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\boldsymbol{m})^2}{2\sigma^2}\right)$
- A normal distribution is entirely determined by its mean value and variance
- The family of normal distributions is closed under linear transformations: if X is normally distributed then a linear transformation $aX + b$ is also normally distributed
- **Exercice:** Show that $\boldsymbol{m} + \sigma\mathcal{N}(0,1) = \mathcal{N}(\boldsymbol{m}, \sigma^2)$

Generalization to n Variables: Independent Case

Assume X1 ~ $\mathcal{N}(\mu_1, \sigma_1^2)$ denote its density $p(x_1) = \frac{1}{Z_1} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 \right)$

Assume X2~ $\mathcal{N}(\mu_2, \sigma_2^2)$ denote its density $p(x_2) = \frac{1}{Z_2} \exp\left(-\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right)$

Assume X1 and X2 are **independent**, then (X1,X2) is a Gaussian vector with

$$p(x_1, x_2) =$$

Generalization to n Variables: Independent Case

Assume X1 ~ $\mathcal{N}(\mu_1, \sigma_1^2)$ denote its density $p(x_1) = \frac{1}{Z_1}\exp\Big(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2\Big)$

Assume X2~ $\mathcal{N}(\mu_2, \sigma_2^2)$ denote its density $p(x_2) = \frac{1}{Z_1}\exp\Big(-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2\Big)$

Assume X1 and X2 are **independent**, then (X1,X2) is a Gaussian vector with

$$p(x_1, x_2) = p(x_1)p(x_2) = \frac{1}{Z_1 Z_2}\exp\Big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big)$$

with $x = (x_1, x_2)^T$ $\mu = (\mu_1, \mu_2)^T$ $\Sigma = \begin{pmatrix}\sigma_1^2 & 0 \\ 0 & \sigma_2^2\end{pmatrix}$

Generalization to n Variables: Independent Case

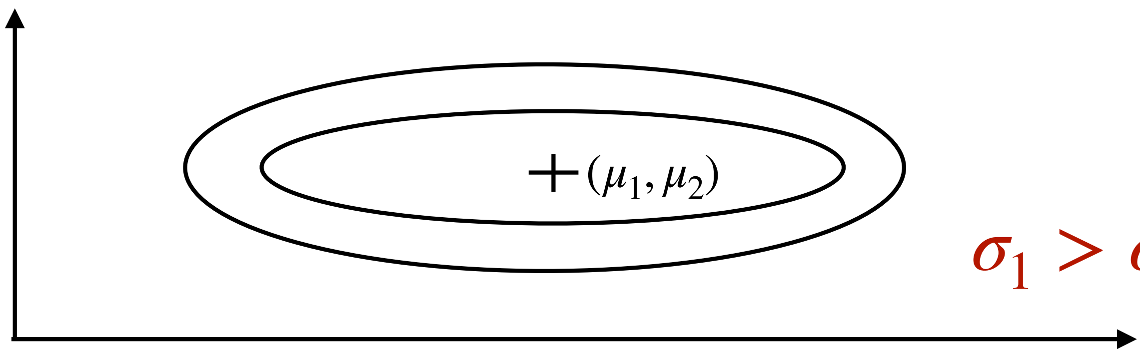
Assume X1 ~ $\mathcal{N}(\mu_1, \sigma_1^2)$ denote its density $p(x_1) = \frac{1}{Z_1}\exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2\right)$

Assume X2~ $\mathcal{N}(\mu_2, \sigma_2^2)$ denote its density $p(x_2) = \frac{1}{Z_1}\exp\left(-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2\right)$

Assume X1 and X2 are **independent**, then (X1,X2) is a Gaussian vector with

$$p(x_1, x_2) = p(x_1)p(x_2) = \frac{1}{Z_1 Z_2}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

with $x = (x_1, x_2)^T$ $\mu = (\mu_1, \mu_2)^T$ $\Sigma = \begin{pmatrix}\sigma_1^2 & 0 \\ 0 & \sigma_2^2\end{pmatrix}$



$\sigma_1 > \sigma_2$

Generalization to n Variables: General Case

Gaussian Vector - Multivariate Normal Distribution

A random vector $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is a Gaussian vector (or multivariate normal) if and only if for all real numbers $a_1, \ldots, a_n$, the random variable $a_1 X_1 + \ldots + a_n X_n$ has a normal distribution.

Gaussian Vector - Multivariate Normal Distribution

A random variable following a 1-D normal distribution is determined by its mean value $\boldsymbol{m}$ and variance σ^2.

In the n-dimensional case it is determined by its mean vector and covariance matrix

Covariance Matrix

If the entries in a vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ are random variables, each with finite variance, then the covariance matrix Σ is the matrix whose (i, j) entries are the covariance of (X_i, X_j)

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

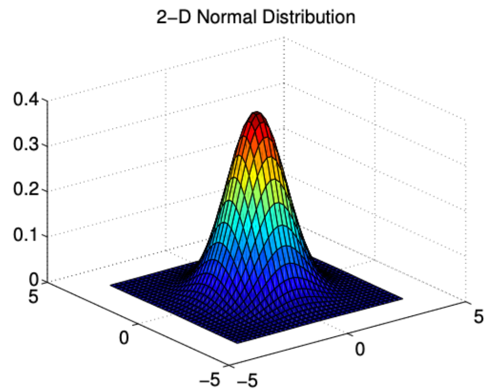
where $\mu_i = \mathrm{E}(X_i)$. Considering the expectation of a matrix as the expectation of each entry, we have

$$\Sigma = \mathrm{E}[(X - \mu)(X - \mu)^T]$$

Σ is symmetric, positive definite

Density of a n-dimensional Gaussian vector $\mathcal{N}(m, C)$:

$$p_{\mathcal{N}(m,C)}(x) = \frac{1}{(2\pi)^{n/2}|C|^{1/2}} \exp\left(-\frac{1}{2}(x-m)^\top C^{-1}(x-m)\right)$$



The mean vector m:

determines the displacement

is the value with the largest density

the distribution is symmetric around the mean

$$\mathcal{N}(m, C) = m + \mathcal{N}(0, C)$$

The covariance matrix:

determines the geometrical shape (see next slides)

Geometry of a Gaussian Vector

Consider a Gaussian vector $\mathcal{N}(m, C)$, remind that lines of equal densities are given by:

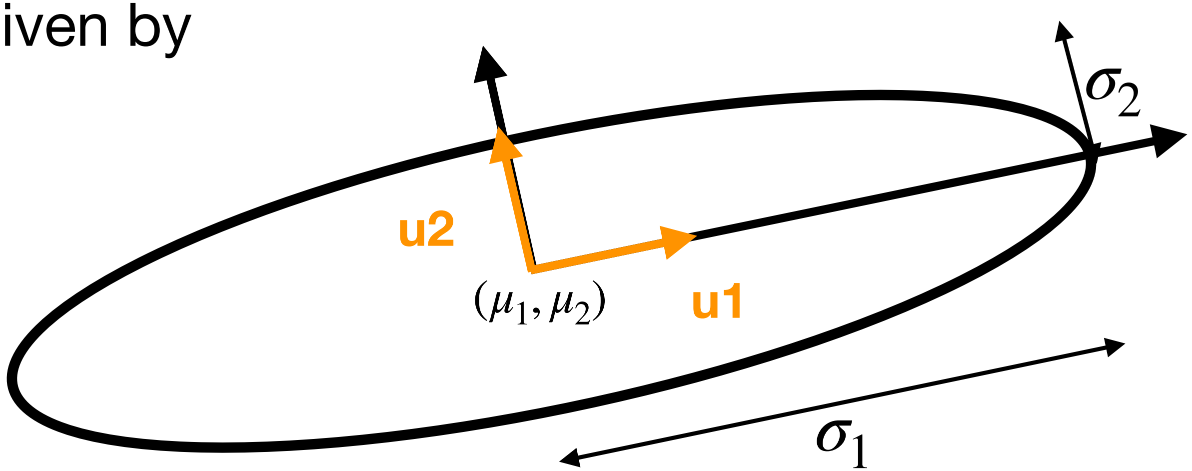
$$\{x \,|\, \Delta^2 = (x-m)^T C^{-1}(x-m) = \text{cst}\}$$

Decompose $C = U\Lambda U^\top$ with U orthogonal, i.e.

$$C = \begin{pmatrix} u_1 & u_2 \\ | & | \end{pmatrix} \begin{pmatrix} \sigma_1^2 & 0 \\ 0| & \sigma_2^2 \end{pmatrix} \begin{pmatrix} u_1 & - \\ u_2 & - \end{pmatrix}$$

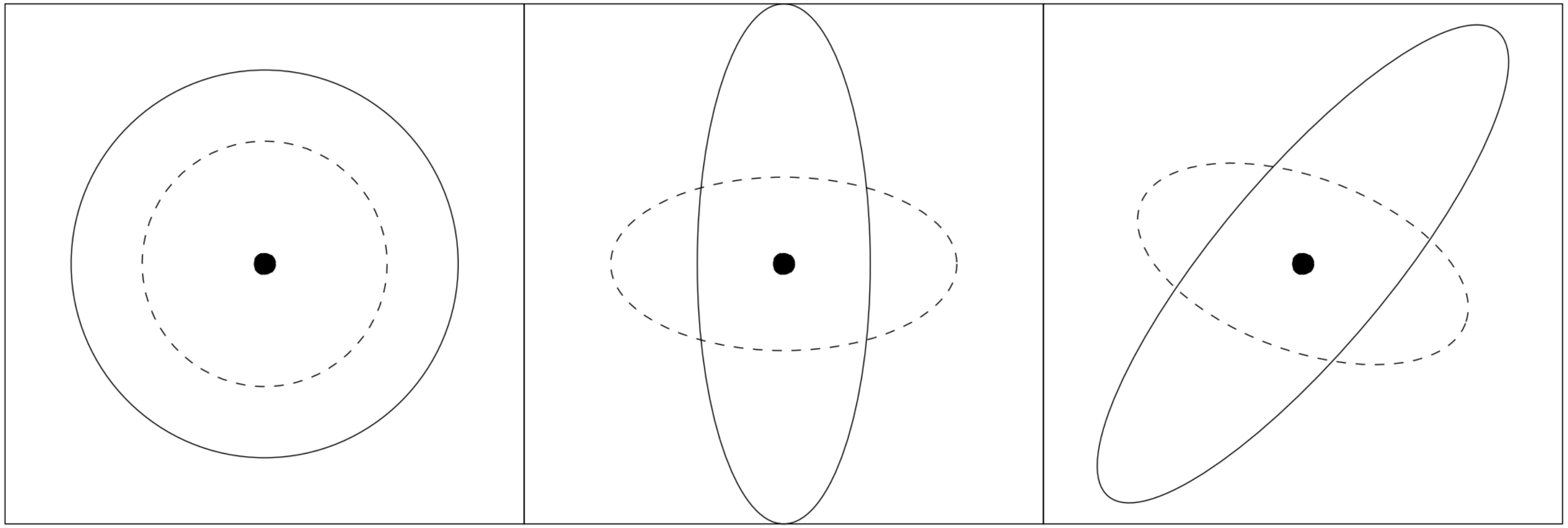
Let $Y = U^\top(x-m)$, then in the coordinate system, (u1,u2), the lines of equal densities are given by

$$\{x \,|\, \Delta^2 = \frac{Y_1^2}{\sigma_1^2} + \frac{Y_2^2}{\sigma_2^2} = \text{cst}\}$$



…any covariance matrix can be uniquely identified with the iso-density ellipsoid $\{\boldsymbol{x} \in \mathbb{R}^n \mid (\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}} \mathbf{C}^{-1}(\boldsymbol{x}-\boldsymbol{m}) = 1\}$

Lines of Equal Density



$\mathcal{N}(\boldsymbol{m}, \sigma^2\mathbf{I}) \sim \boldsymbol{m} + \sigma\mathcal{N}(\mathbf{0}, \mathbf{I})$
one degree of freedom σ
components are independent standard normally distributed

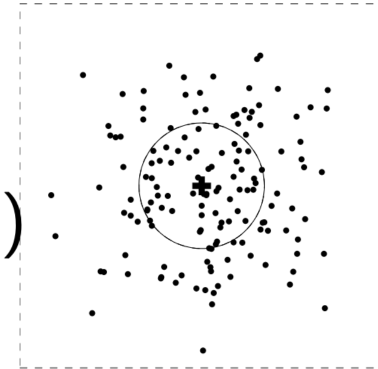
$\mathcal{N}(\boldsymbol{m}, \mathbf{D}^2) \sim \boldsymbol{m} + \mathbf{D}\,\mathcal{N}(\mathbf{0}, \mathbf{I})$
n degrees of freedom
components are independent, scaled

$\mathcal{N}(\boldsymbol{m}, \mathbf{C}) \sim \boldsymbol{m} + \mathbf{C}^{\frac{1}{2}}\mathcal{N}(\mathbf{0}, \mathbf{I})$
$(n^2+n)/2$ degrees of freedom
components are correlated

where $\mathbf{I}$ is the identity matrix (isotropic case) and $\mathbf{D}$ is a diagonal matrix (reasonable for separable problems) and $\mathbf{A} \times \mathcal{N}(\mathbf{0}, \mathbf{I}) \sim \mathcal{N}(\mathbf{0}, \mathbf{A}\mathbf{A}^{\mathrm{T}})$ holds for all $\mathbf{A}$.

Evolution Strategies

New search points are sampled normally distributed



$$\boldsymbol{x}_i = \boldsymbol{m} + \sigma\, \boldsymbol{y}_i \qquad \text{for } i = 1, \ldots, \lambda \text{ with } \boldsymbol{y}_i \text{ i.i.d.} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

as perturbations of $\boldsymbol{m}$, where $\boldsymbol{x}_i, \boldsymbol{m} \in \mathbb{R}^n$, $\sigma \in \mathbb{R}_+$, $\mathbf{C} \in \mathbb{R}^{n \times n}$

where

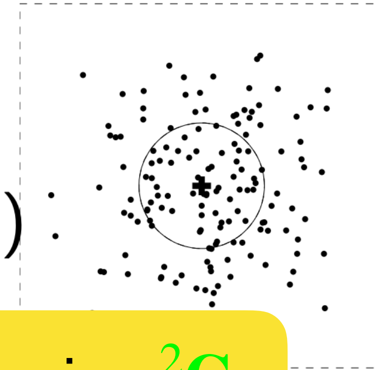
- the mean vector $\boldsymbol{m} \in \mathbb{R}^n$ represents the favorite solution
- the so-called step-size $\sigma \in \mathbb{R}_+$ controls the *step length*
- the covariance matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$ determines the shape of the distribution ellipsoid

here, all new points are sampled with the same parameters

Evolution Strategies

New search points are sampled normally distributed

$\boldsymbol{x}_i = \boldsymbol{m} + \sigma\, \boldsymbol{y}_i \qquad$ for $i = 1, \ldots, \lambda$ with $\boldsymbol{y}_i$ i.i.d. $\sim \mathcal{N}(\mathbf{0}, \mathbf{C})$



In fact, the covariance matrix of the sampling distribution is $\sigma^2\mathbf{C}$ but it is convenient to refer to $\mathbf{C}$ as the covariance matrix (it is a covariance matrix but not of the sampling distribution)

where

- the mean vector $\boldsymbol{m} \in \mathbb{R}^n$ represents the favorite solution
- the so-called step-size $\sigma \in \mathbb{R}_+$ controls the *step length*
- the covariance matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$ determines the shape of the distribution ellipsoid

here, all new points are sampled with the same parameters

How to update the different parameters $m, \sigma, \mathbf{C}$?

1. **Adapting the mean m**
2. Adapting the step-size σ
3. Adapting the covariance matrix C

Update the Mean: a Simple Algorithm the (1+1)-ES

Notation and Terminology:

one solution kept from one iteration to the next

(1+1)-ES

one new solution (offspring) sampled at each iteration

The + means that we keep the best between current solution and new solution, we talk about *elitist* selection

(1+1)-ES algorithm (update of the mean)

sample one candidate solution from the mean $\mathbf{m}$

$$\mathbf{x} = \mathbf{m} + \sigma \mathcal{N}(0, \mathbf{C})$$

if $\mathbf{x}$ is better than $\mathbf{m}$ (i.e. if $f(\mathbf{x}) \leq f(\mathbf{m})$), select $\mathbf{m}$

$$\mathbf{m} \leftarrow \mathbf{x}$$

The (1+1)-ES algorithm is a simple algorithm, yet:

- the elitist selection is not robust to outliers

we cannot loose solutions accepted by "chance", for instance that look good because the noise gave it a low function value

- there is no population (just a single solution is sampled) which makes it less robust

In practice, one should rather use a:

$$(\mu/\mu, \lambda)\text{-ES}$$

The μ best solutions are selected and recombined (to form the new mean)

λ solutions are sampled at each iteration

The $(\mu/\mu, \lambda)$-ES - Update of the Mean Vector

Given the i-th solution point $\boldsymbol{x}_i = \boldsymbol{m} + \sigma \underbrace{\boldsymbol{y}_i}_{\sim\mathcal{N}(\mathbf{0},\mathbf{C})}$

Let $\boldsymbol{x}_{i:\lambda}$ the i-th ranked solution point, such that $f(\boldsymbol{x}_{1:\lambda}) \leq \cdots \leq f(\boldsymbol{x}_{\lambda:\lambda})$.

Notation: we denote $\boldsymbol{y}_{i:\lambda}$ the vector such that $\boldsymbol{x}_{i:\lambda} = \boldsymbol{m} + \sigma \boldsymbol{y}_{i:\lambda}$

Exercice: realize that $\boldsymbol{y}_{i:\lambda}$ is generally not distributed as $\mathcal{N}(\mathbf{0}, \mathbf{C})$

The new mean reads

$$\boldsymbol{m} \leftarrow \sum_{i=1}^{\mu} w_i \, \boldsymbol{x}_{i:\lambda}$$

where

$$w_1 \geq \cdots \geq w_\mu > 0, \quad \textstyle\sum_{i=1}^{\mu} w_i = 1, \quad \frac{1}{\sum_{i=1}^{\mu} {w_i}^2} =: \mu_w \approx \frac{\lambda}{4}$$

The best μ points are selected from the new solutions (non-elitistic) and weighted intermediate recombination is applied.

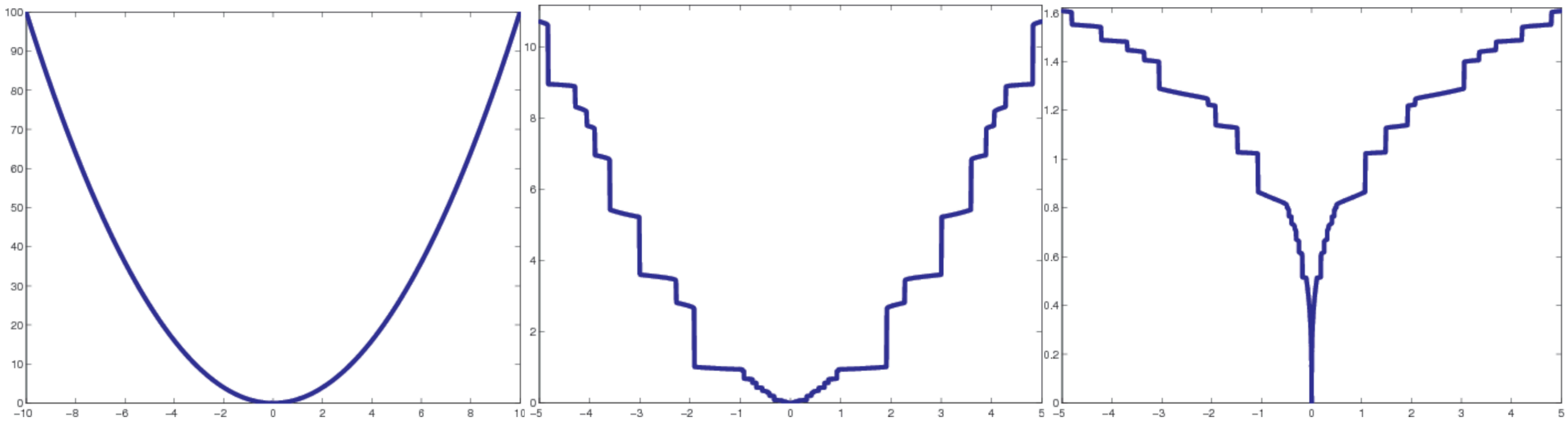
What changes in the previous slide if instead of optimizing f, we optimize $g \circ f$ where $g : \mathrm{Im}(f) \to \mathbb{R}$ is strictly increasing?

Invariance Under Monotonically Increasing Functions

Comparison-based/ranking-based algorithms:

Update of all parameters uses only the ranking:

$$f(x_{1:\lambda}) \leq f(x_{2:\lambda}) \leq \ldots \leq f(x_{\lambda:\lambda})$$



$$g(f(x_{1:\lambda})) \leq g(f(x_{2:\lambda})) \leq \ldots \leq g(f(x_{\lambda:\lambda}))$$

for all $g : \mathrm{Im}(f) \to \mathbb{R}$ strictly increasing

A Template for Comparison-based Stochastic Search

Define $\{P_\theta : \theta \in \Theta\}$, a family of probability distributions on $\mathbb{R}^n$

Generic template to optimize $f : \mathbb{R}^n \to \mathbb{R}$

Initialize distribution parameter θ, set population size $\lambda \in \mathbb{N}$

While not terminate

1. Sample $x_1, \ldots, x_\lambda$ according to P_θ
2. Evaluate $x_1, \ldots, x_\lambda$ on f
3. Rank the solutions and find π the permutation such

$$f(x_{\pi(1)}) \leq f(x_{\pi(2)}) \leq \ldots \leq f(x_{\pi(\lambda)})$$

4. Update parameters $\theta \leftarrow F(\theta, x_1, \ldots, x_\lambda, \pi)$