DERIVATIVE FREE OPTIMIZATION 2024/2025.

CLASS 2.

Pure Random search (PRS)

Assume $f : [-1,1]^n \to \mathbb{R}$, $x = (x_1, \dots, x_n) \to f(x)$

PRS :

Initialize $x_{best} = \text{Unif}([-1,1]^n)$ (uniform)

WHILE NOT HAPPY (while stop criterion not met)

Sample $x \sim \text{Unif}([-1,1]^n)$

If $f(x) \leq f(x_{best})$

$x_{best} \leftarrow x$

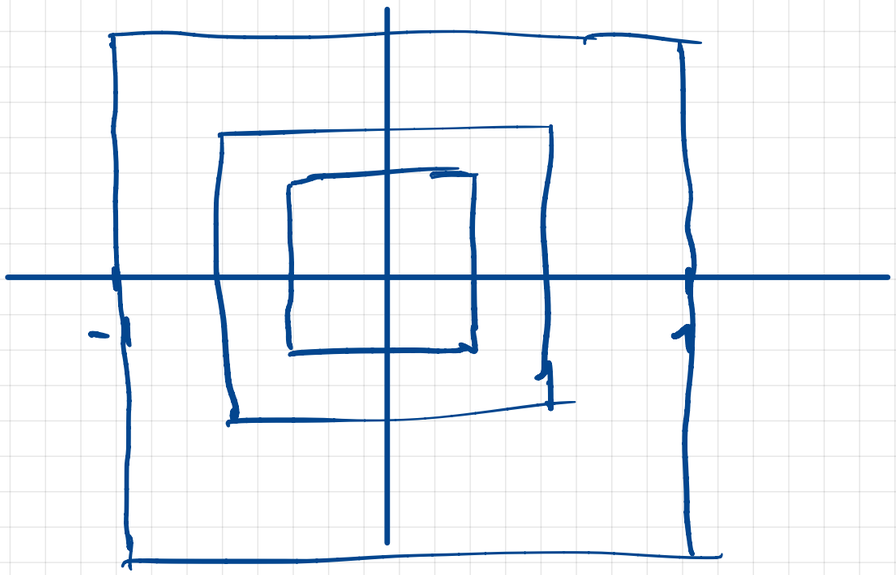
(See also Exercice 1)

Does this algorithm converge : Yes under mild assumptions on f
(need to have "volume" in a neighborhood of global optimum)

Simplified proof setting

$$f(x) = \|x\|_\infty = \max(|x_1|, \dots, |x_n|)$$

Level sets : $n = 2$



In exercice 1 : uniform samples $\{v_0, v_1, \dots, v_t, \dots\}$

$$v_i \sim \text{Unif}([-1, 1]^n)$$

x_t : best solution at iteration t

$$f(x_t) = \min\{f(v_1), \dots, f(v_t)\}$$

By induction.

Prove that $\forall \varepsilon > 0 \quad \lim_{t \to +\infty} \mathbb{P}_t(\|X_t\|_\infty \geq \varepsilon) = 0$

↳ give CV in probability.

$$\|X_t\|_\infty = \min\{\|V_1\|_\infty, \dots, \|V_t\|_\infty\}$$

$$\{\|X_t\|_\infty \geq \varepsilon\} = \{\forall \underset{\| \\ 1,\dots,t}{k} \ \|V_k\|_\infty \geq \varepsilon\} = \bigcap_{k=1}^{t} \{\|V_k\|_\infty \geq \varepsilon\}$$

$$\mathbb{P}_r(\{\|X_t\|_\infty \geq \varepsilon\}) = \mathbb{P}_r\left(\bigcap_{k=1}^{t} \{\|V_k\|_\infty \geq \varepsilon\}\right)$$

$$= \prod_{k=1}^{t} \mathbb{P}_r(\|V_k\|_\infty \geq \varepsilon) = \mathbb{P}_r(\|V_1\|_\infty \geq \varepsilon)^t$$

by ind of $\{V_i, i = 1, \dots k\}$

↑ because V_k are identically distributed.

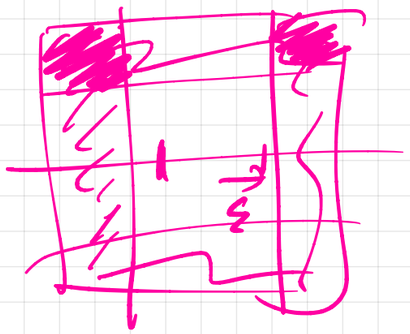
$= (1 - \varepsilon^n)^t$

$$\mathbb{P}r\left(\|U_1\|_\infty > \varepsilon\right) \overset{?}{=} 1 - \mathbb{P}r\left(\|U_1\|_\infty \leq \varepsilon\right)$$

$$= 1 - \frac{\mathrm{Vol}\left(\{x \mid \|x\|_\infty \leq \varepsilon\}\right)}{\mathrm{Vol}\left(\{x \mid \|x\|_\infty \leq 1\}\right)}$$

$$\not=$$

$$\prod_{\text{coord}} \left(\text{each coordinate} > \varepsilon\right) = 1 - \frac{(2\varepsilon)^n}{2^n} = 1 - \varepsilon^n$$



$$\mathbb{P}r\left(\|X_t\|_\infty > \varepsilon\right) = (1-\varepsilon^n)^t \xrightarrow[t \to +\infty]{} 0$$

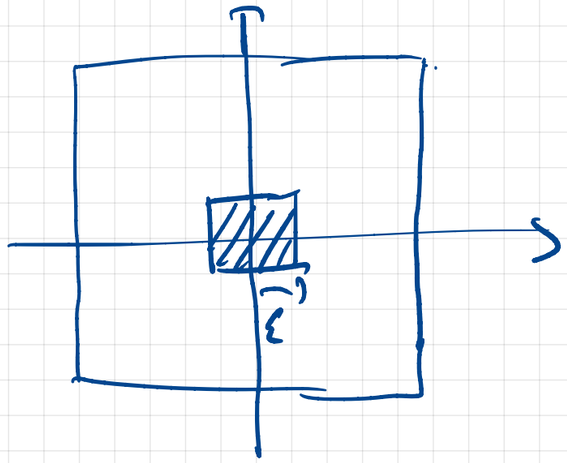
Hence RRS converges in probability to the optimum of f.

Let's look at how fast it converges.

$$T_\varepsilon = \inf\{ t \mid X_t \in B(x^*, \varepsilon) \}$$

(Ball for infinity norm)

$$B(x^*, \varepsilon) = \{x \mid \|x\|_\infty < \varepsilon\}$$



RRS is similar to game:

Sample v_t, win if $f(v_t) \leq \varepsilon$

$\Leftrightarrow v_t \in B(0, \varepsilon)$

Lose otherwise.

T_ε : time it takes to win this game.

Given a game with 2 outcomes win with proba p and lose with proba $(1-p)$, where the outcome is sampled randomly and independently [example : flip a coin), the time it takes to win a game is distributed according to a geometric distribution.

$$\mathbb{E}[T_\varepsilon] = \frac{1}{p}$$

Back to PRS. $p = \mathbb{P}r(\text{"win"}) = \mathbb{P}r(\|v_t\| \leq \varepsilon) = \varepsilon^n$

$$\mathbb{E}(T_\varepsilon) = \frac{1}{\varepsilon^n}$$

Is it fast ? No

To compare : linear CV $T_\varepsilon \sim n \log\left(\frac{1}{\varepsilon}\right)$ [gradient descent on L-smooth strongly convex]

The algorithm is "blind", does not take into account the information gathered on f, through the sampling of points to sample "better" solutions.

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 $x_1, \ldots, x_\lambda \in \mathbb{R}^n$ 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

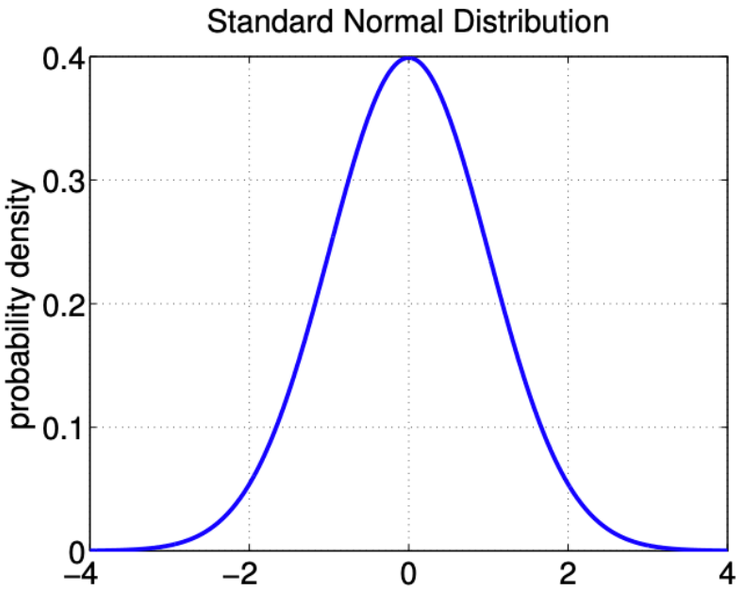
To obtain an optimization algorithm we need:

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 $\boldsymbol{m} + \sigma\mathcal{N}(0,1) = \mathcal{N}(\boldsymbol{m}, \sigma^2)$

$m + \sigma W(0,1)$ is normally distributed

We only need to identify its mean and variance:

$$\mathbb{E}\left(m + \sigma W(0,1)\right) = m + \sigma \underbrace{\mathbb{E}(W(0,1))}_{=0} = m$$

by linearity of $\mathbb{E}$

$$\begin{aligned}
&\mathrm{Var}(m + \sigma W(0,1)) \\
&= \mathbb{E}\left(\left(m + \sigma W(0,1) - m\right)^2\right) - \\
&= \mathbb{E}\left(\sigma^2 W(0,1)^2\right) \\
&= \sigma^2 \underbrace{\mathbb{E}(W(0,1)^2)}_{=1} = \sigma^2
\end{aligned}$$

$$\Rightarrow \quad m + \sigma W(0,1) \sim W(m, \sigma^2)$$

$$\mathbb{P}(m + \sigma\mathcal{N}(0,1) \le t) = \mathbb{P}_r\left(\mathcal{N}(0,1) \le \frac{t-m}{\sigma}\right)$$

$$= \int_{-\infty}^{\frac{t-m}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

$dy = \sigma dx$

$y = \sigma x + m$

$$= \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\left(\frac{y-m}{2\sigma}\right)^2\right) dy$$

$x = \frac{t-m}{\sigma}$ $\quad y = t$

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 $p(x_1) = \frac{1}{Z_1}\exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2\right)$

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

Generalization to n Variables: Independent Case

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

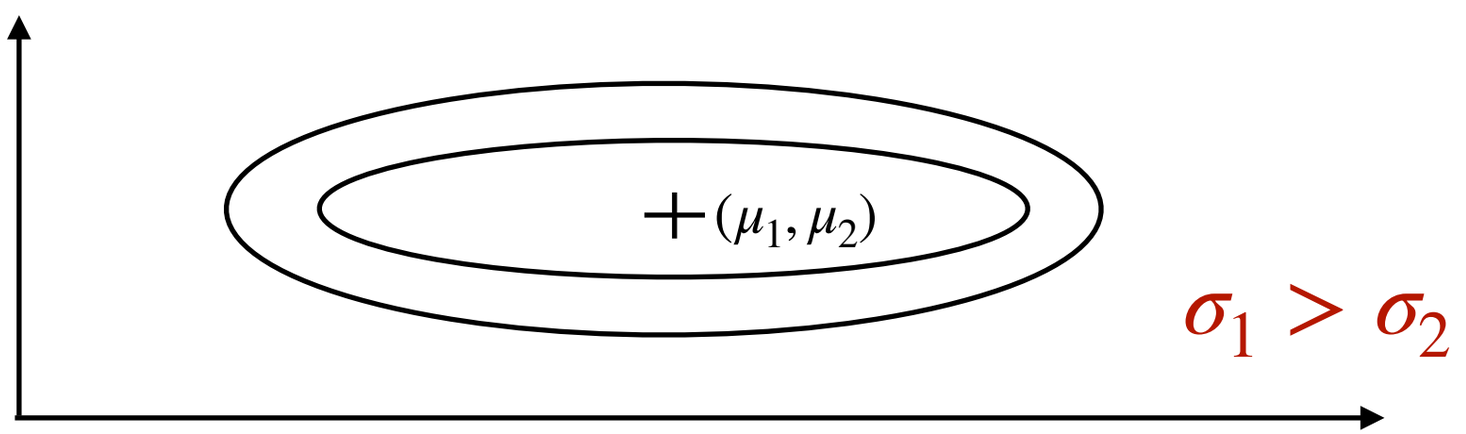
Assume X2~ $\mathcal{N}(\mu_2, \sigma_2^2)$ denote its density $\quad p(x_2) = \frac{1}{Z_1}\exp\left(-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2\right)$

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

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

$$\rightarrow -\frac{1}{2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]$$

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



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 $a_1X_1 + \ldots + a_nX_n$ 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)

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

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

$\Sigma_{ii} = \mathrm{Var}(X_i)$

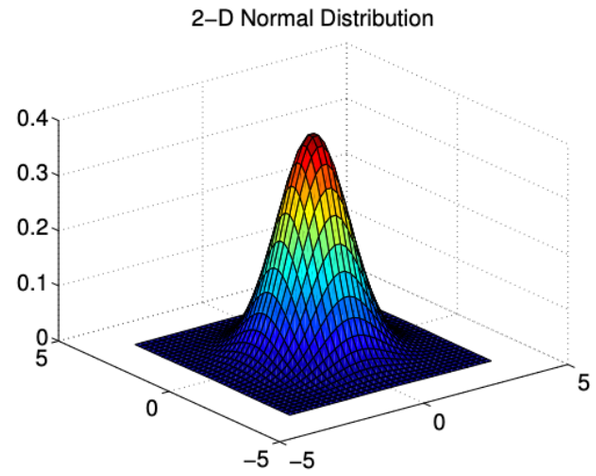
$$\Sigma = \mathrm{E}[(X - \mu)(X - \mu)^T]$$

Σ is symmetric, positive definite

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

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

$|C| = \det(C)$



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:

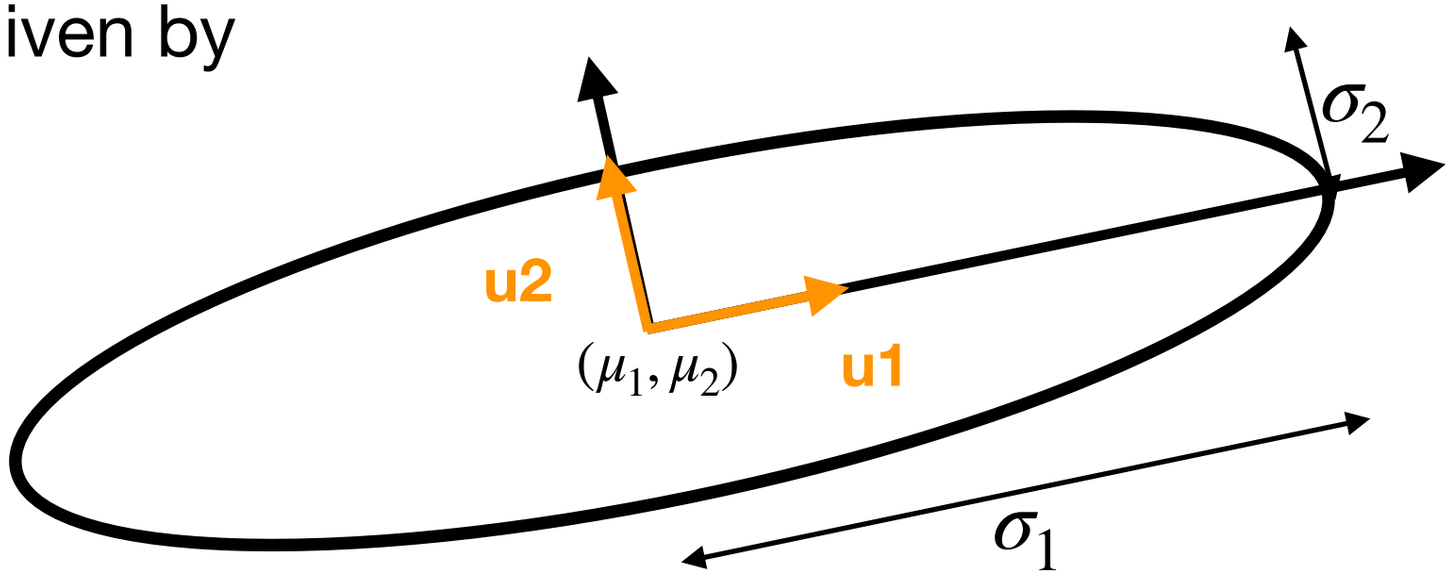
$$\{x \,|\, \Delta^2 = (x-m)^T C^{-1}(x-m) = \text{cst}\}$$

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

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

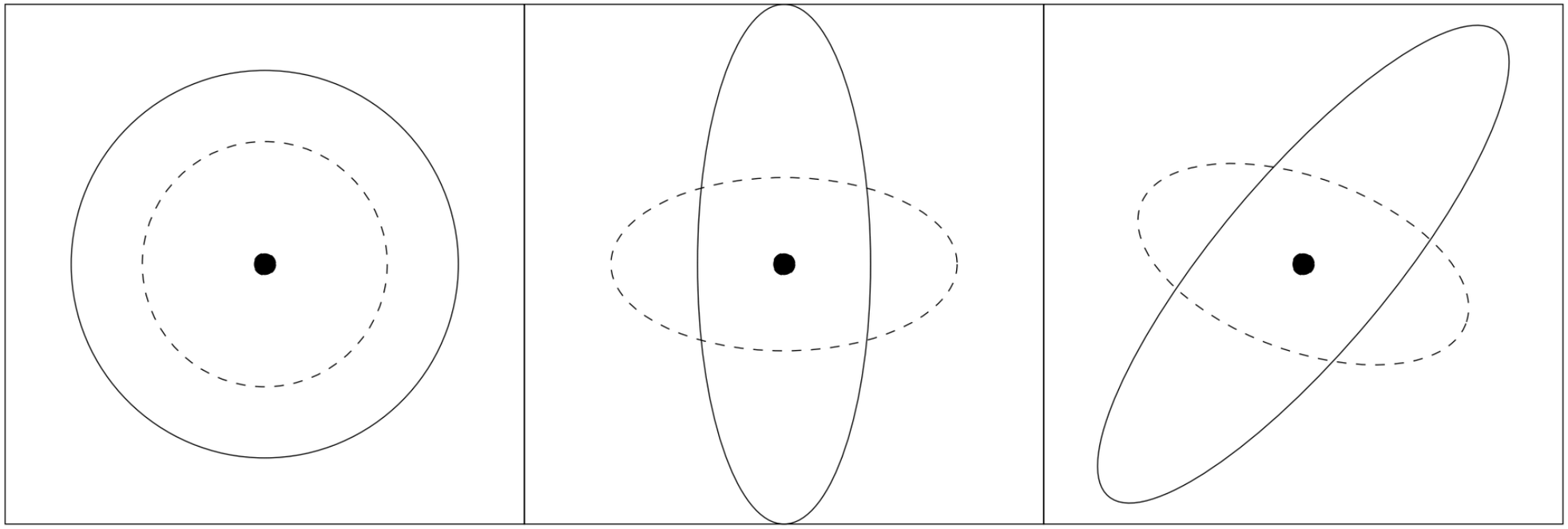
Let $Y = U^\top (x-m)$, then in the coordinate system, (u1,u2), the lines of equal densities are given by

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



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

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

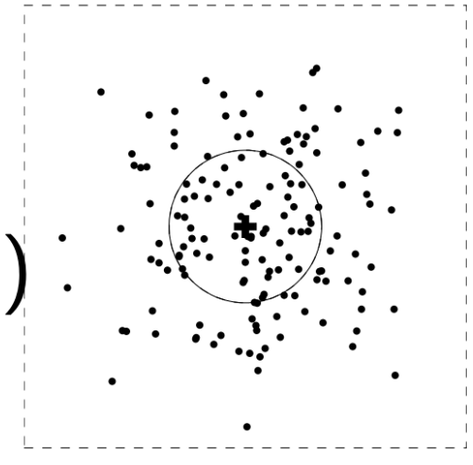
components are correlated

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

Evolution Strategies

New search points are sampled normally distributed

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



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

$x_i \sim \mathcal{N}(m, \sigma^2 C)$

where

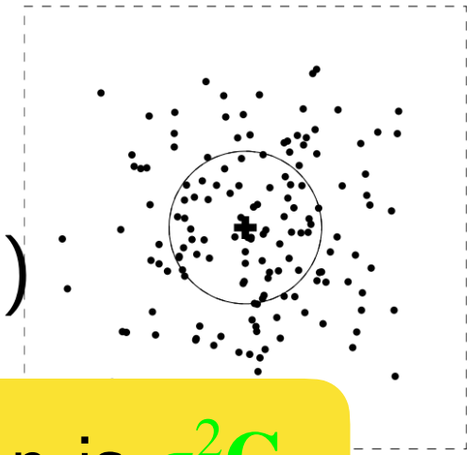
- the mean vector $\boldsymbol{m} \in \mathbb{R}^n$ represents the favorite solution
- the so-called step-size $\sigma \in \mathbb{R}_+$ controls the *step length*
- the covariance matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$ determines the shape of the distribution ellipsoid

here, all new points are sampled with the same parameters

Evolution Strategies

New search points are sampled normally distributed

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



a

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 $\mathbf{m}$

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