Mathématiques pour l'intelligence artificielle

M2 informatique parcours IAAA, Aix-Marseille Université

Thomas Schatz
Mardi 12 septembre 2023

Optimisation

1. Optimisation sans contraintes
2. Optimisation sous contraintes
3. Analyse convexe et dualité
4. Analyse des algorithmes d'optimisation

Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion

Probabilités

1. Notions de base
2. Calculs d’espérance et de variance
3. Loi gaussienne multivariée et autres distributions
4. Suites de variables aléatoires et notions de convergence

Univers et mesure de probabilité

Univers Ω

Evénement $A \subset \Omega$

Ensemble des événements (tribu) $\mathcal{F} \subset \mathcal{P}(\Omega)$ Stable par union dénombrable et par passage au complémentaire

Mesure de probabilité $P : \mathcal{F} \to \mathbf{R}$

$$\forall A \in \mathcal{F},\ P(A) \geq 0$$
$$P(\Omega) = 1$$

Pour tout ensemble dénombrable d'événements **disjoints** $A_1, A_2, \ldots,$

$$P\left(\bigcup_{i=1}^{+\infty} A_i\right) = \sum_{i=1}^{+\infty} P(A_i)$$

Variable aléatoire

Univers Ω

Ensemble des événements (tribu) $\mathcal{F} \subset \mathcal{P}(\Omega)$ Stable par union dénombrable et par passage au complémentaire

Mesure de probabilité $P : \mathcal{F} \to \mathbf{R}$

Variable aléatoire $X : \Omega \to E$

$\mathcal{T} \subset \mathcal{P}(E)$ Tribu sur l'espace d'arrivée

$\forall T \in \mathcal{T},\ X^{-1}(T) \in \mathcal{F}$

Loi de X (mesure de probabilité induite sur E) $\forall T \in \mathcal{T},\ P_X(T) = P(X^{-1}(T))$

Conditionnement, indépendance

Probabilité conditionnelle

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}$$

Indépendance mutuelle d'événements

$$\forall S \subset \{1, 2, \ldots, n\}, \quad P\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} P(A_i)$$

Loi de probabilité conditionnelle

$$X : \Omega \to E \quad Y : \Omega \to F \quad \mathcal{T} \subset \mathcal{P}(E) \quad \Sigma \subset P(F)$$

Origine des dépendances : Omega !

Tribus sur les espaces d'arrivée

$$T \in \mathcal{T} \quad S \in \Sigma$$

$$P_{X|Y}(T|S) = P(X^{-1}(T)|Y^{-1}(S)) = \frac{P(X^{-1}(T) \cap Y^{-1}(S))}{P(Y^{-1}(S))}$$

Variables aléatoires indépendantes

$$\forall T \in \mathcal{T}, \ \forall S \in \Sigma, \quad P_{X,Y}(T, S) = P_X(T) P_Y(S)$$

Variables aléatoires réelles

$$X : \Omega \to E \subset \mathbf{R}$$

$$p_X(x) = P_X(\{x\})$$ Fonction de masse

$$F_X(x) = P_X(\{y \mid y \leq x\})$$ Fonction de répartition

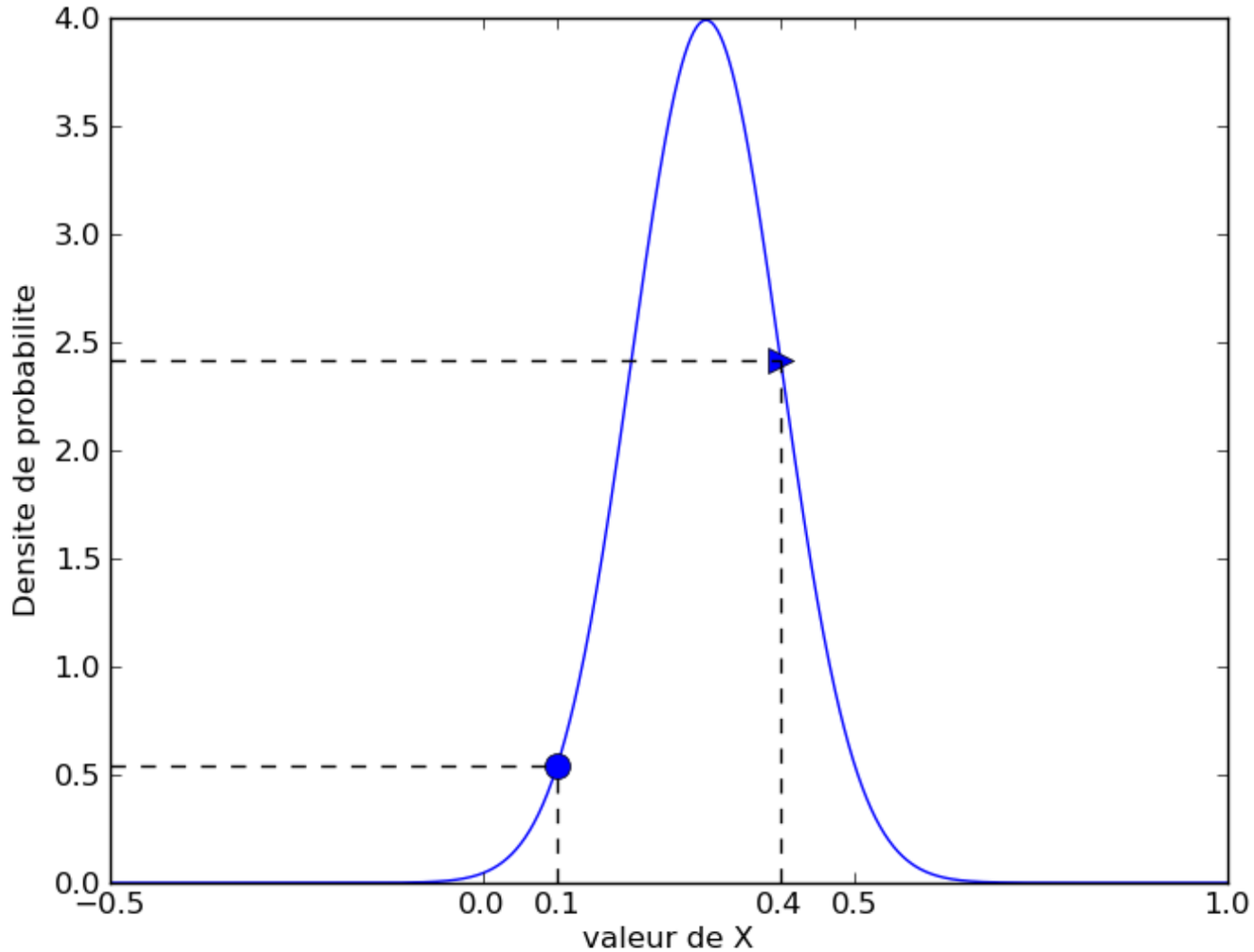
$$p_{dX} = \frac{dF_X}{dx}$$ Densité de probabilité

$$E[X] = \int x P_X(dx)$$ Espérance (intégrale de Lebesgue)

$$E[X] = \int x p_{dX}(x) dx$$ Cas continu

$$E[X] = \sum_x x p_X(x)$$ Cas discret

On a représente sur le graphe ci-dessous la densité de probabilité d'une variable aléatoire réelle X.

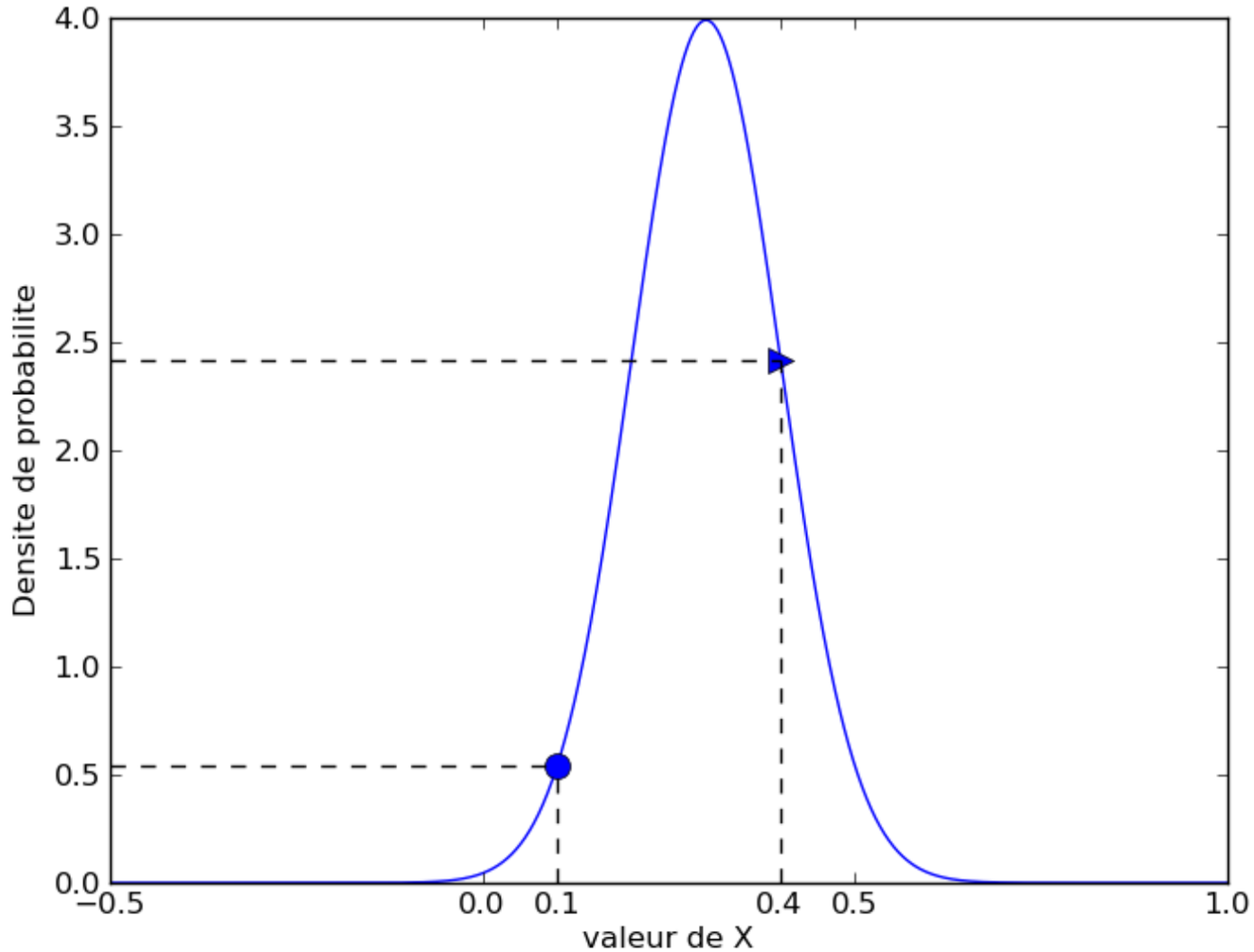


1) La densité de probabilité de X prend une valeur supérieure 1 en $X = 0.4$. Cela vous parait-il normal ? Justifiez votre réponse.

Soit x une réalisation de X.

2) Quelle est la probabilité d'avoir $x = 0.1$? Quelle est la probabilité d'avoir $x = 0.4$? Est il plus probable d'observer $x = 0.4$ ou $x = 0.1$? A quel point (approximativement) ?

On a représente sur le graphe ci-dessous la densité de probabilité d'une variable aléatoire réelle X.



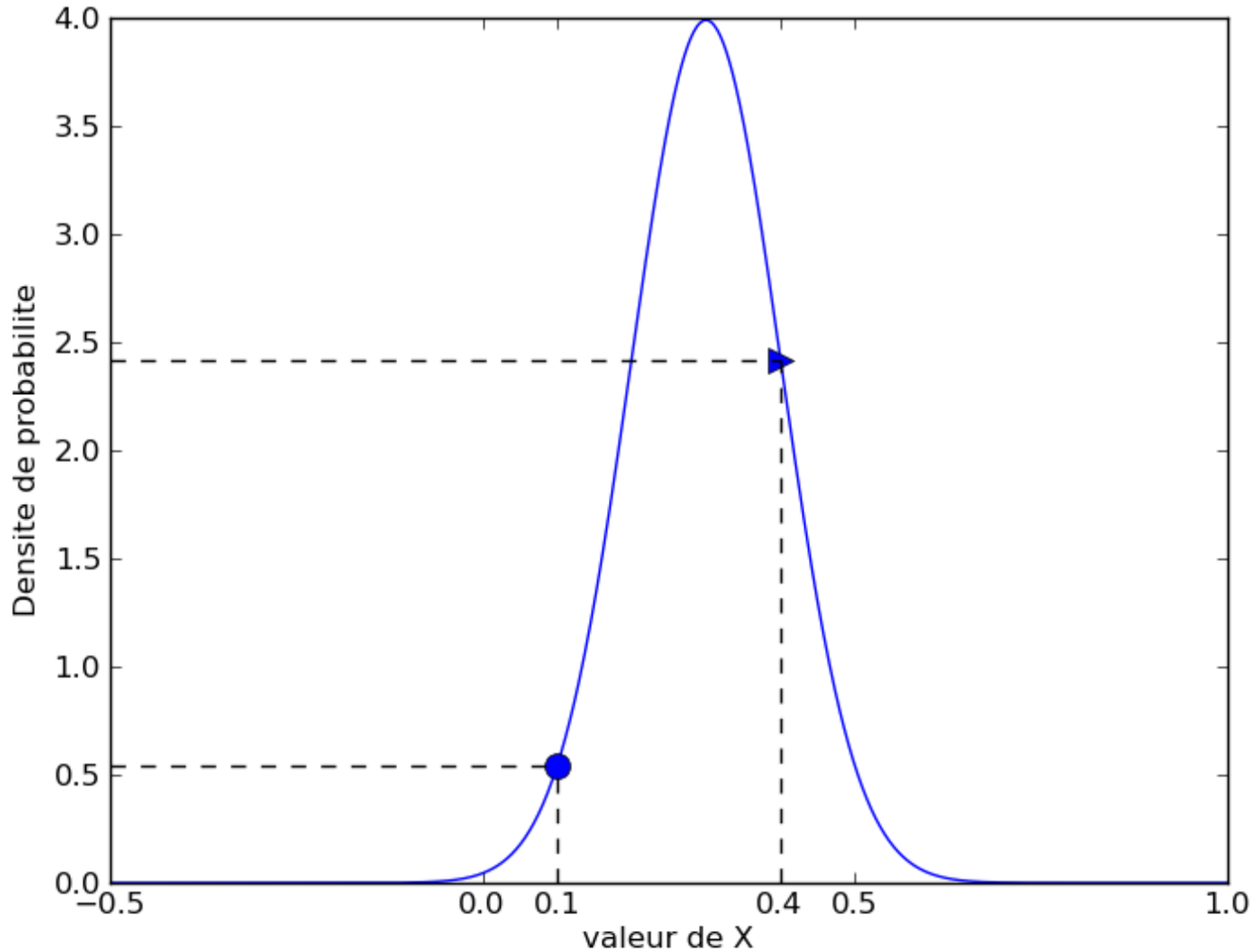
1) La densité de probabilité de X prend une valeur supérieure 1 en $X = 0.4$. Cela vous parait-il normal ? Justifiez votre réponse.

L'aire sous la courbe doit être égale à 1 mais la densité en un point particulier peut être supérieure et même arbitrairement grande.

Soit x une réalisation de X.

2) Quelle est la probabilité d'avoir $x = 0.1$? Quelle est la probabilité d'avoir $x = 0.4$? Est il plus probable d'observer $x = 0.4$ ou $x = 0.1$? A quel point (approximativement) ?

On a représente sur le graphe ci-dessous la densité de probabilité d'une variable aléatoire réelle X.



1) La densité de probabilité de X prend une valeur supérieure 1 en $X = 0.4$. Cela vous parait-il normal ? Justifiez votre réponse.

L'aire sous la courbe doit être égale à 1 mais la densité en un point particulier peut être supérieure et même arbitrairement grande.

Soit x une réalisation de X.

2) Quelle est la probabilité d'avoir $x = 0.1$? Quelle est la probabilité d'avoir $x = 0.4$? Est il plus probable d'observer $x = 0.4$ ou $x = 0.1$? A quel point (approximativement) ?

La probabilité que x soit 0.1 ou 0.4 est 0. Par contre, il est environ 5 fois plus probable d'observer $X = 0.4$ que $x = 0.1$.

Plusieurs variables aléatoires réelles

$$\mathbf{X} = (X_1, X_2, \ldots, X_n) : \Omega \to \mathbf{R}^n$$ Origine des dépendances : Omega !

$$p_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = P_{\mathbf{X}}(\{x_1, x_2, \ldots, x_n\})$$ Fonction de masse (jointe)

$$F_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = P_{\mathbf{X}}(\{y_1, y_2, \ldots, y_n \mid y_1 \leq x_1, y_2 \leq x_2, \ldots, y_n \leq x_n\})$$

Fonction de répartition (jointe)

$$p_{d\mathbf{X}} : x_1, x_2, \ldots, x_n \mapsto \frac{\partial F_{\mathbf{X}}}{\partial x_1 dx_2 \ldots dx_n}(x_1, x_2, \ldots, x_n)$$

Densité de probabilité (jointe)

Espérance $f : \mathbf{R}^n \to \mathbf{R}$

Cas continu $$E[f(\mathbf{X})] = \int_{x_1} \int_{x_2} \cdots \int_{x_n} f(x_1, x_2, \ldots, x_n) p_{d\mathbf{X}}(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n$$

Cas discret $$E[f(\mathbf{X})] = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} f(x_1, x_2, \ldots, x_n) p_{\mathbf{X}}(x_1, x_2, \ldots, x_n)$$

Modèles graphiques

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_{\pi_i})$$

Probabilités

1. Notions de base
2. Calculs d’espérance et de variance
3. Loi gaussienne multivariée et autres distributions
4. Suites de variables aléatoires et notions de convergence

Moments

Variance $\mathrm{Var}[X] = E[(X - E[X])^2]$

Moment $E[X^k]$

Moment centré $E[(X - E[X])^k]$

Covariance $\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$

Linéarité de l'espérance $$E\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i E(X_i)$$

Bilinéarité de la variance $$\mathrm{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} \sum_{i=1}^{n} a_i a_j \mathrm{Cov}(X_i, X_j)$$

Moments conditionnels

Espérance conditionnelle

$$f(y) = E[X|Y = y] = \int x P_{X|Y=y}(dx)$$

Variance conditionnelle

$$g(y) = \mathrm{Var}[X|Y = y] = E[(X - E[X])^2|Y = y]$$

Loi de l'espérance totale

$$E[X] = E[E[X|Y]]$$

Loi de la variance totale

$$\mathrm{Var}[X] = \mathrm{Var}[E[X|Y]] + E[\mathrm{Var}[X|Y]]$$

Probabilités

1. Notions de base
2. Calculs d'espérance et de variance
3. Loi gaussienne multivariée et autres distributions
4. Suites de variables aléatoires et notions de convergence

Example Distributions

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, ..., n$	np	$np(1-p)$
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, ...$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, ...$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a}$ for all $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

[2] Table reproduced from Maleki & Do's review handout by Koochak & Irvin

Random Vectors

Given n RV's $X_1, ..., X_n$, we can define a random vector X s.t.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Note: all the notions of joint PDF/CDF will apply to X.

Given $g : \mathbb{R}^n \to \mathbb{R}^m$, we have:

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix}, \mathbb{E}[g(X)] = \begin{bmatrix} \mathbb{E}[g_1(X)] \\ \mathbb{E}[g_2(X)] \\ \vdots \\ \mathbb{E}[g_m(X)] \end{bmatrix}.$$

Covariance Matrices

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose ij-th entry contains the covariance between X_i and X_j.

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & \dots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \dots & Cov[X_n, X_n] \end{bmatrix}$$

applying linearity of expectation and the fact that $Cov[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$, we obtain

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

Properties:

- Σ is symmetric and PSD
- If $X_i \perp X_j$ for all i, j, then $\Sigma = diag(Var[X_1], ..., Var[X_n])$

Multivariate Gaussian

The multivariate Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$, $X \in \mathbb{R}^n$:

$$p(x; \mu, \Sigma) = \frac{1}{det(\Sigma)^{\frac{1}{2}}(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

The univariate Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$, $X \in \mathbb{R}$ is just the special case of the multivariate Gaussian when $n = 1$.

$$p(x; \mu, \sigma^2) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Notice that if $\Sigma \in \mathbb{R}^{1\times 1}$, then $\Sigma = Var[X_1] = \sigma^2$, and so

- $\Sigma^{-1} = \frac{1}{\sigma^2}$
- $det(\Sigma)^{\frac{1}{2}} = \sigma$

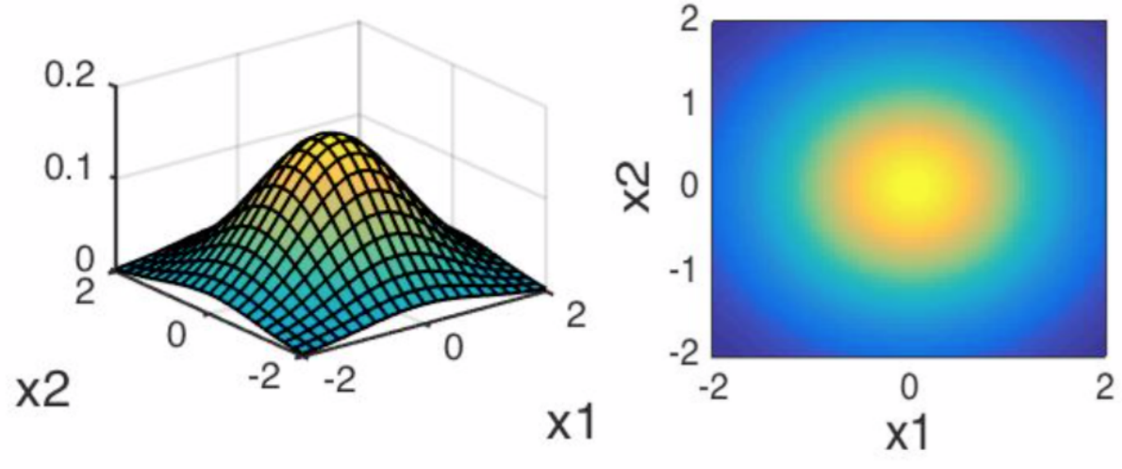
Some Nice Properties of MV Gaussians

- Marginals and conditionals of a joint Gaussian are Gaussian
- A d-dimensional Gaussian $X \in \mathcal{N}(\mu, \Sigma = diag(\sigma_1^2, ..., \sigma_n^2))$ is equivalent to a collection of d **independent** Gaussians $X_i \in \mathcal{N}(\mu_i, \sigma_i^2)$. This results in isocontours aligned with the coordinate axes.
- In general, the isocontours of a MV Gaussian are n-dimensional ellipsoids with principal axes in the directions of the eigenvectors of covariance matrix Σ (remember, Σ is PSD, so all n eigenvectors are non-negative). The axes' relative lengths depend on the eigenvalues of Σ.

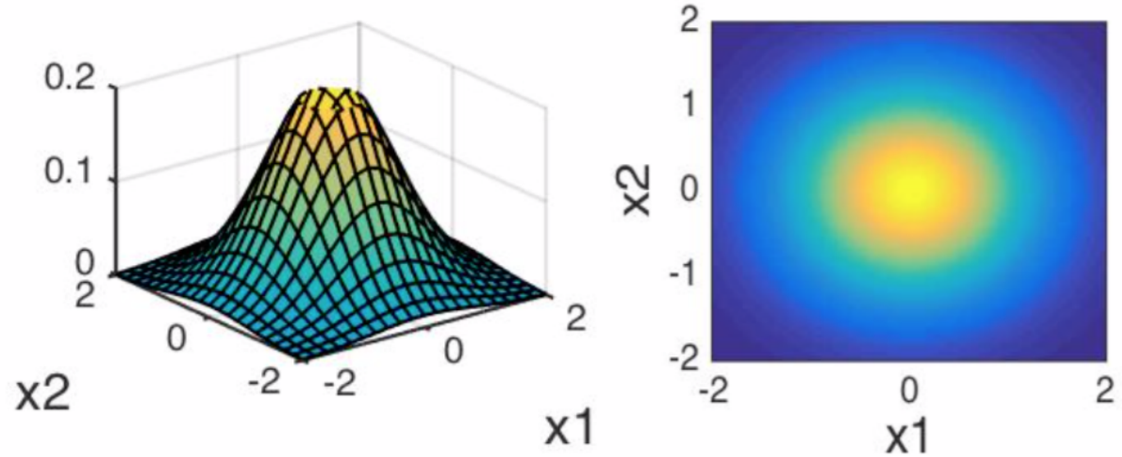
Visualizations of MV Gaussians

Effect of changing variance

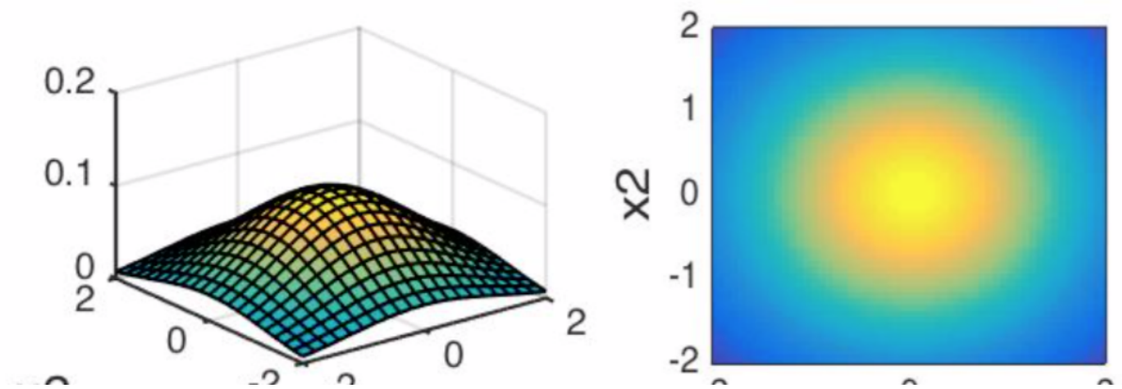
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 0.7 & 0 \\ 0 & 0.7 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 1.5 & 0 \\ 0 & 1.5 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$

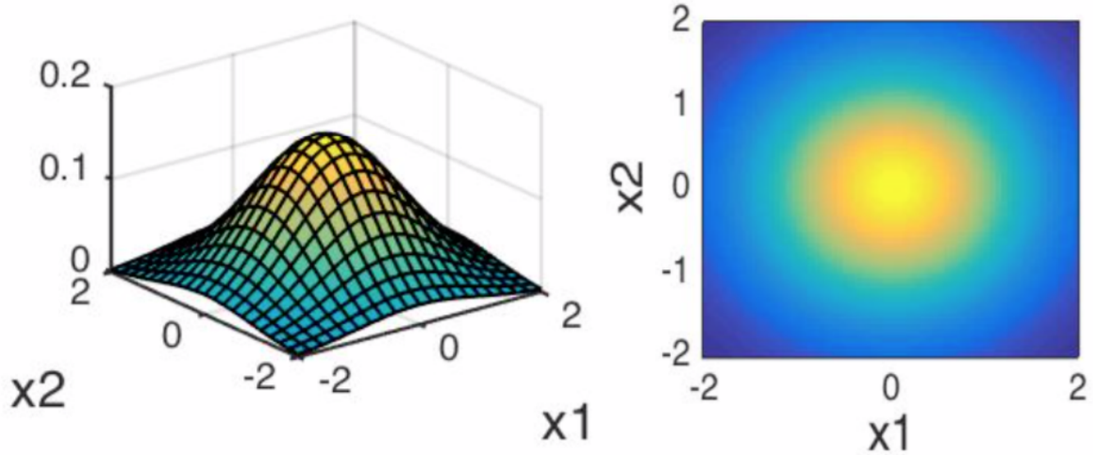


Visualizations of MV Gaussians

If $Var[X_1] \neq Var[X_2]$:

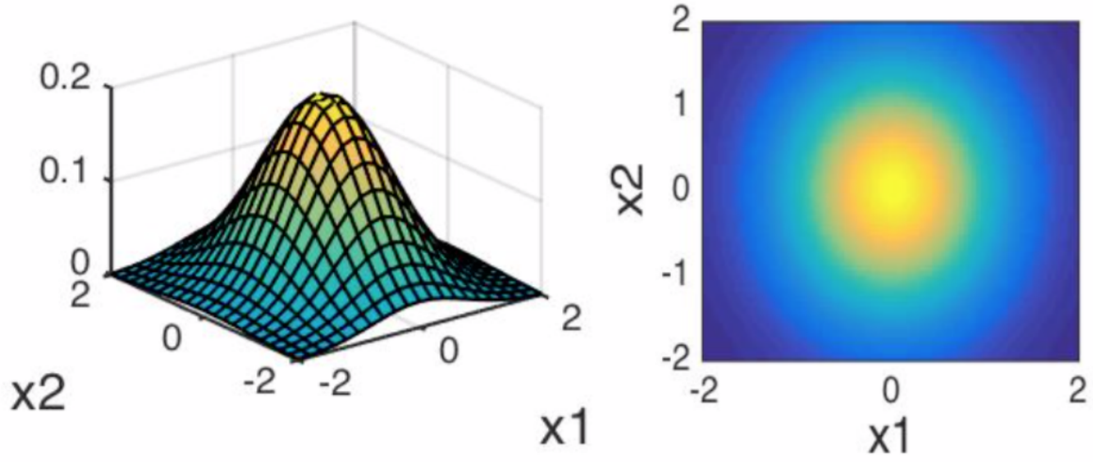
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



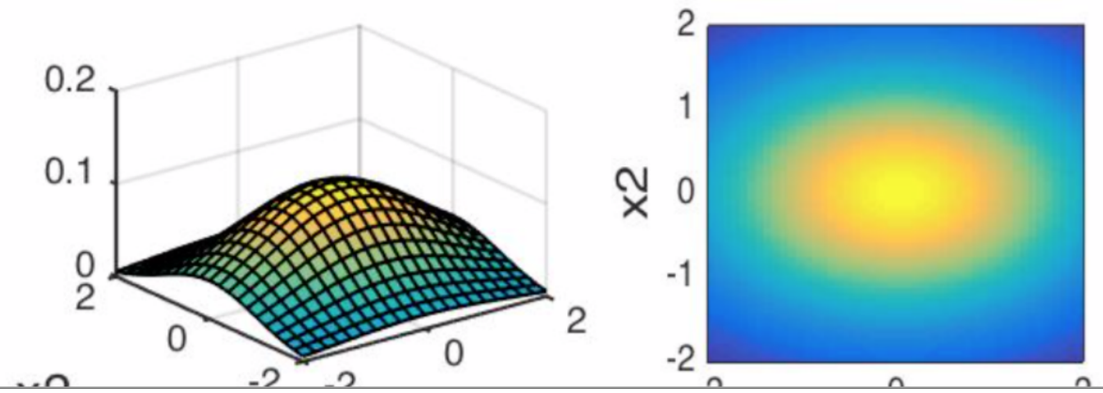
$$\Sigma = \begin{matrix} 0.6 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 2 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$

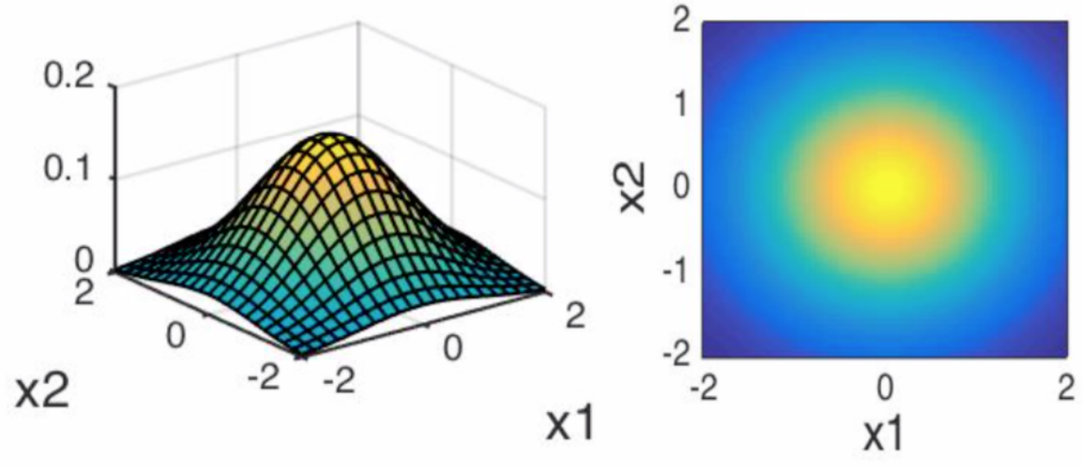


Visualizations of MV Gaussians

If X_1 and X_2 are positively correlated:

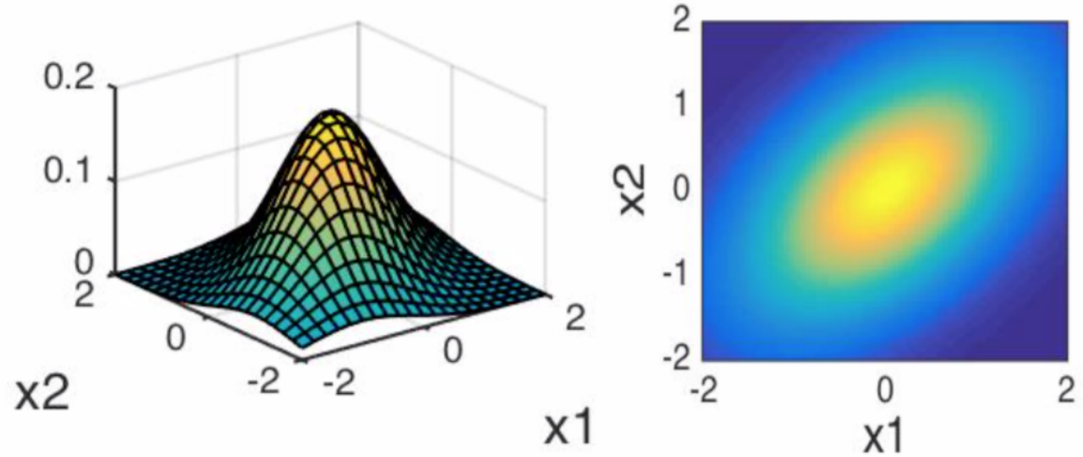
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



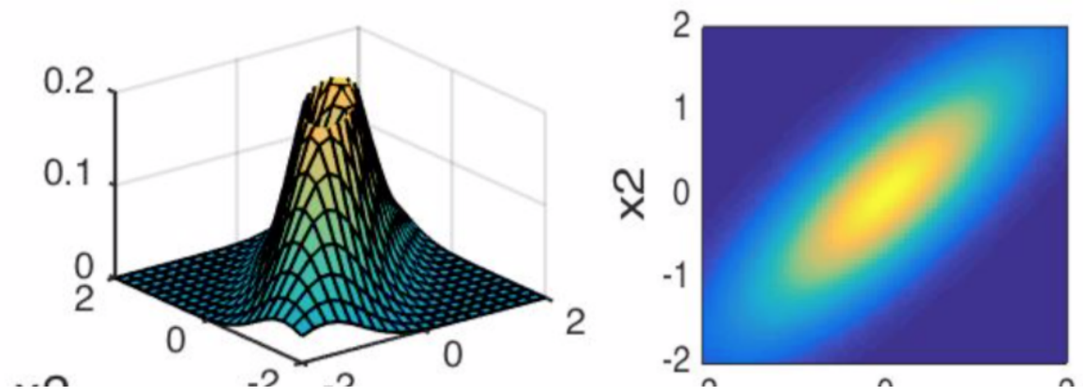
$$\Sigma = \begin{matrix} 1 & 0.5 \\ 0.5 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 1 & 0.8 \\ 0.8 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$

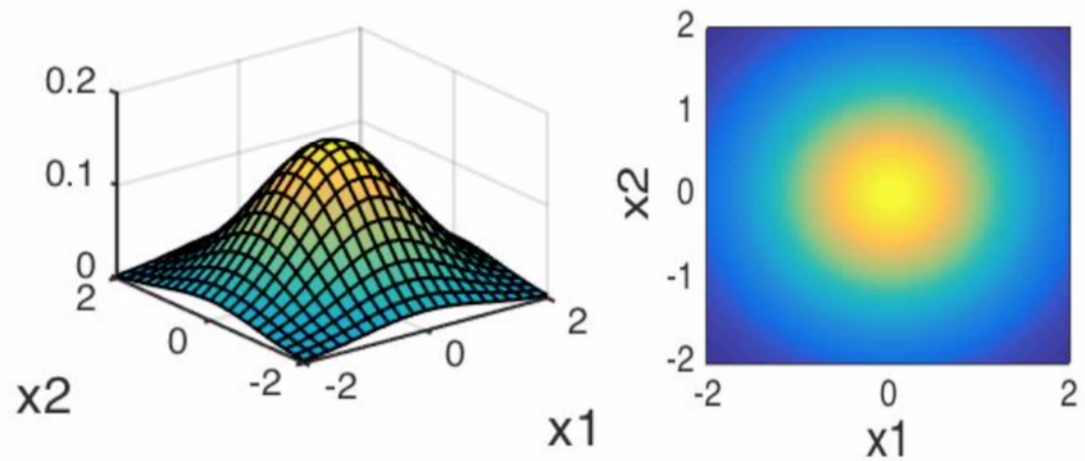


Visualizations of MV Gaussians

If X_1 and X_2 are negatively correlated:

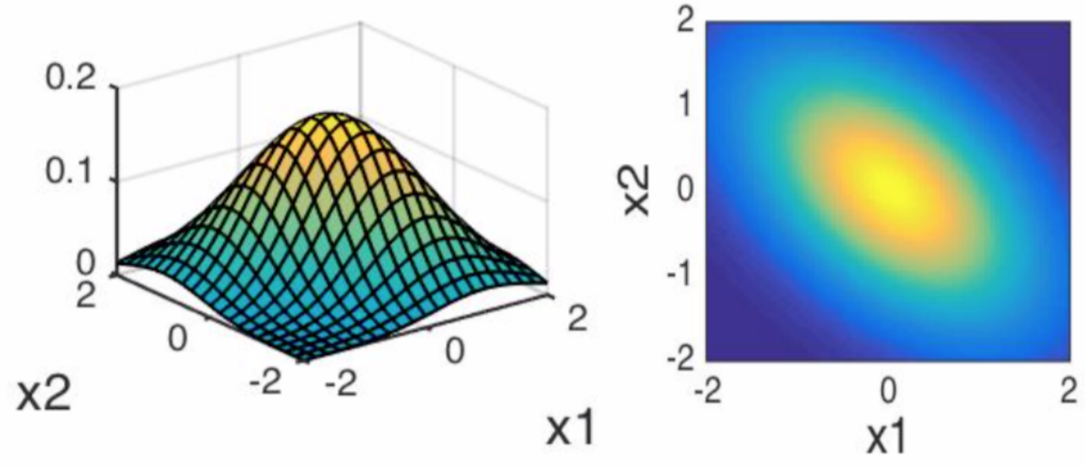
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



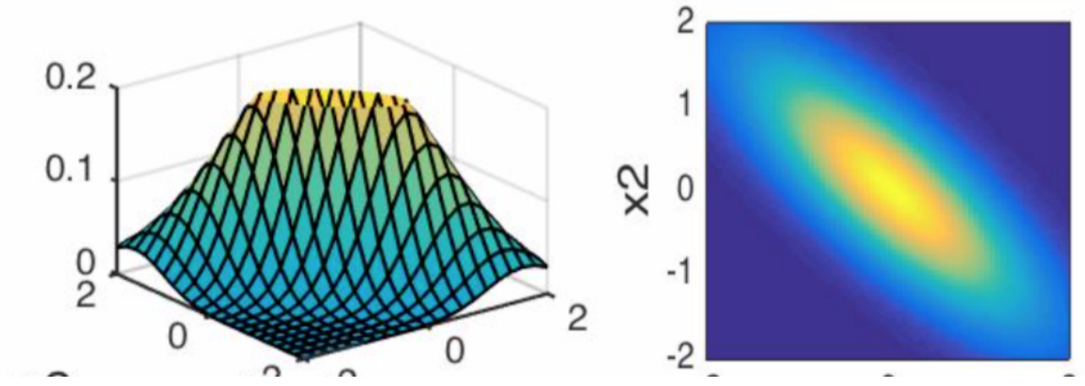
$$\Sigma = \begin{matrix} 1 & -0.5 \\ -0.5 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 1 & -0.8 \\ -0.8 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



Multivariate Gaussian

Définition générale

$$\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma}) \quad \Longleftrightarrow \quad \text{there exist } \mu \in \mathbb{R}^k, \boldsymbol{A} \in \mathbb{R}^{k \times \ell} \text{ such that } \mathbf{X} = \boldsymbol{A}\mathbf{Z} + \mu \text{ for } Z_n \sim \mathcal{N}(0,1), \text{i.i.d.}$$

Distributions conditionnelles

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}$$

$$p(\mathbf{x}_1 \mid \mathbf{x}_2 = \mathbf{a}) = \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\mathbf{\Sigma}}), with$$

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{+}(\mathbf{a} - \boldsymbol{\mu}_2)$$

$$\bar{\mathbf{\Sigma}} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{+}\mathbf{\Sigma}_{21}$$

Distributions marginales ?

Probabilités

1. Notions de base
2. Calculs d'espérance et de variance
3. Loi gaussienne multivariée et autres distributions
4. Suites de variables aléatoires et notions de convergence