

Dualité

Problème considéré

$$\min_{x \in X \subset \mathbf{R}^d} f(x) \text{ sous les contraintes } \begin{cases} e_i(x) = 0, & i = 1..r \\ c_i(x) \leq 0, & i = 1..s \end{cases}$$

$$f : X \rightarrow \mathbf{R}$$

$$r \in \mathbf{N}$$

$$s \in \mathbf{N}$$

$$e_i : X \rightarrow \mathbf{R}$$

$$c_i : X \rightarrow \mathbf{R}$$

$$\Omega = \{x \in X \mid e_i(x) = 0, \ c_j(x) \leq 0, \ i = 1..r, \ j = 1..s\}$$

$$f^* = \inf_{x \in \Omega} f(x)$$

$$\text{Hypothèses : } \begin{aligned} \Omega &\neq \emptyset \\ f^* &> -\infty \end{aligned}$$

Dualité

Lagrangien associé au problème

$$L(x, \mu, \lambda) = f(x) + \mu^T e(x) + \lambda^T c(x)$$

avec :

$$x \in X$$

$$c(x) = [c_1(x) \ c_2(x) \ \dots \ c_s(x)]^T$$

$$e(x) = [e_1(x) \ e_2(x) \ \dots \ e_r(x)]^T$$

$$\mu \in \mathbf{R}^r$$

$$\lambda \in \mathbf{R}^s$$

Dualité

Fonction duale $q(\mu, \lambda) = \inf_{x \in X} L(x, \mu, \lambda)$

Domaine $D = \{\mu, \lambda \mid q(\mu, \lambda) > -\infty\}$

Problème dual

$\max_{(\mu, \lambda) \in D} q(\mu, \lambda)$ sous les contraintes $\lambda_i \geq 0, i = 1..s$

Résultats

D est convexe et q est concave sur D

$q^* \leq f^*$ (dualité faible)

avec $q^* = \sup_{\mu \in \mathbf{R}^r, \lambda \in \mathbf{R}^s, \lambda_i \geq 0, i=1..s} q(\mu, \lambda)$

Dualité

Dualité et convexité

Dualité

Dualité forte, convexité et qualification des contraintes

Théorème 1:

Si X est polyédral, si f est convexe et si les contraintes e_i et c_j sont affines, alors il n'y a pas d'écart dual: $q^* = f^*$

(Un ensemble polyédral est un ensemble qui peut être obtenu comme l'intersection d'un ensemble fini de demi-espaces fermés)

Théorème 2 :

Slater's condition. Si X est convexe, si f est convexe, si les contraintes c_i sont convexes, si les contraintes e_j sont affines et s'il existe un point x_0 dans l'intérieur relatif de X tel que $c_i(x_0) < 0$ et $e_j(x_0) = 0$ pour $i = 1..s$ et $j = 1..r$, alors il n'y a pas d'écart dual: $q^* = f^*$

(L'intérieur d'un ensemble est cet ensemble privé de sa frontière. L'intérieur relatif d'un ensemble S est l'intérieur de cet ensemble au sein de son enveloppe affine, c'est à dire au sein du plus petit espace affine contenant S .)

Dualité

Retrouver la solution du primal à partir de la solution du dual en cas de dualité forte

En cas de dualité forte ($q^* = f^*$), les conditions :

- $x^* \in X$ et $e_i(x^*) = 0$ et $c_j(x^*) \leq 0$ (faisabilité primale)
- $\lambda^* \geq 0$ (faisabilité duale)
- $x^* = \arg \min_{x \in X} L(x, \mu^*, \lambda^*)$ (optimalité Lagrangienne)
- $\mu_j^* c_j(x^*) = 0, j = 1..s$ (complémentarité)

sont des conditions nécessaires et suffisantes pour que x^* soit une solution optimale du problème primal et (μ^*, λ^*) une solution optimale du problème dual.

Optimisation

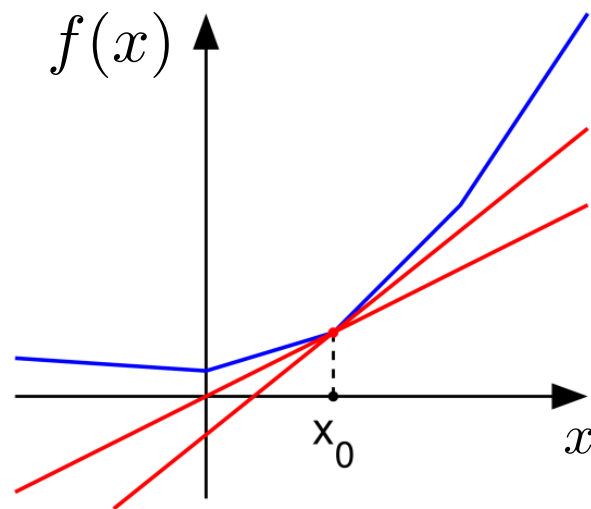
1. Optimisation lisse sans contraintes
2. Optimisation lisse avec contraintes
3. Analyse convexe et dualité Lagrangienne
4. Ouverture vers l'optimisation non-lisse
5. Algorithmes d'optimisation

Optimisation

1. Optimisation lisse sans contraintes
2. Optimisation lisse avec contraintes
3. Analyse convexe et dualité Lagrangienne
4. Ouverture vers l'optimisation non-lisse
5. Algorithmes d'optimisation

Un mot sur l'optimisation non lisse

Convexe



Sous-gradient (subgradient) $\partial f(x_0)$

Condition nécessaire de minimum local pour une fonction convexe (pas forcément lisse) :

Soit $f : \mathbf{R}^d \rightarrow \mathbf{R}$ une fonction convexe. Si x^* est un minimum local de f , alors : $0 \in \partial f(x^*)$

Version KKT disponible pour le cas convexe non-lisse sous contraintes

Non-convexe

Domaine de recherche ouvert

Cas particulier utile en apprentissage automatique :

Condition nécessaire de minimum local pour la somme d'une fonction lisse (pas forcément convexe) et d'une fonction convexe (pas forcément lisse) :

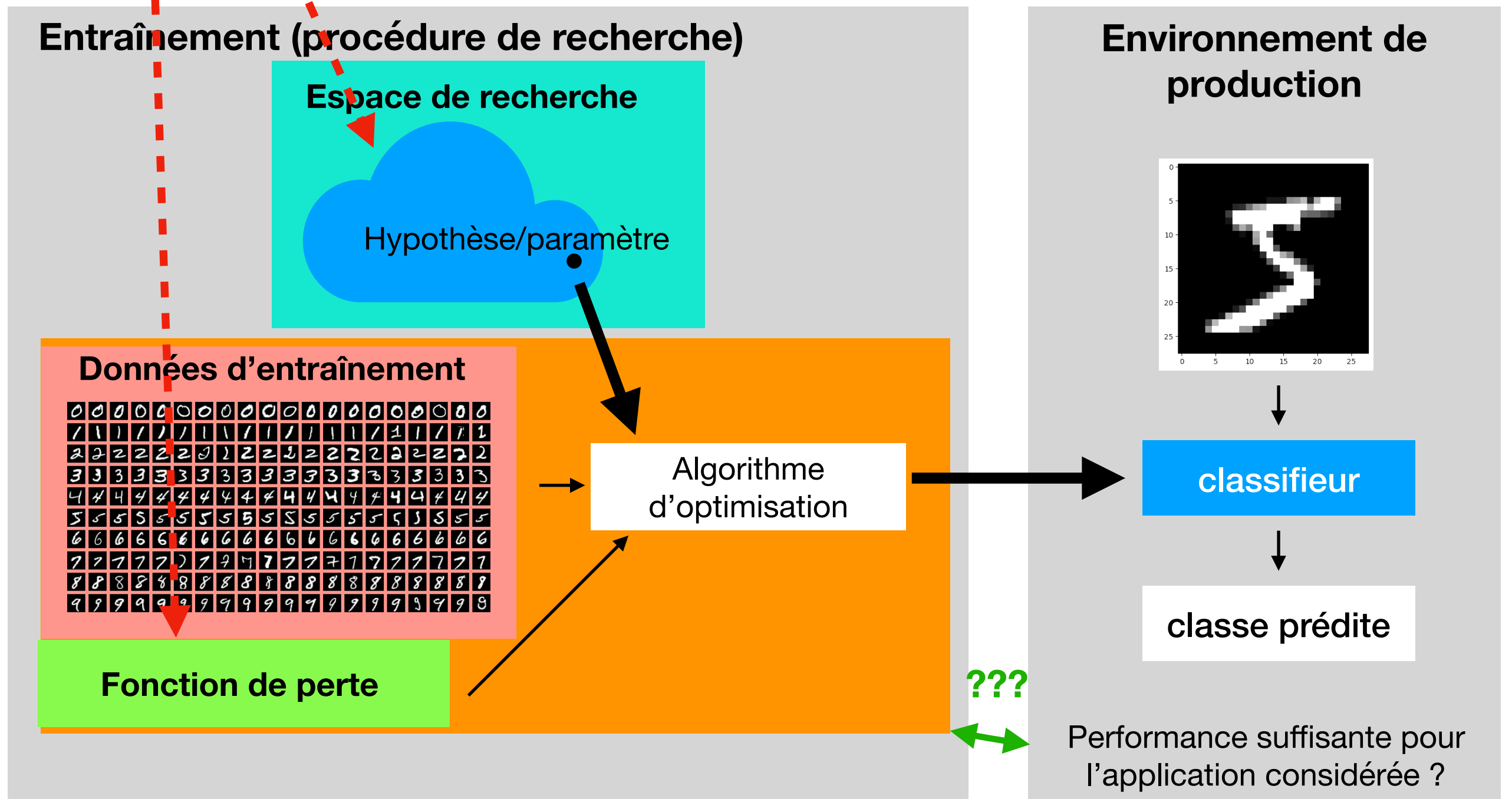
Soit $f : \mathbf{R}^d \rightarrow \mathbf{R}$ de classe $\mathcal{C}^1$ et $g : \mathbf{R}^d \rightarrow \mathbf{R}$ une fonction convexe. Si x^* est un minimum local de $f + g$, alors :

$$0 \in \nabla f(x^*) + \partial g(x^*) = \{\nabla f(x^*) + x \mid x \in \partial g(x^*)\}.$$

Probabilités

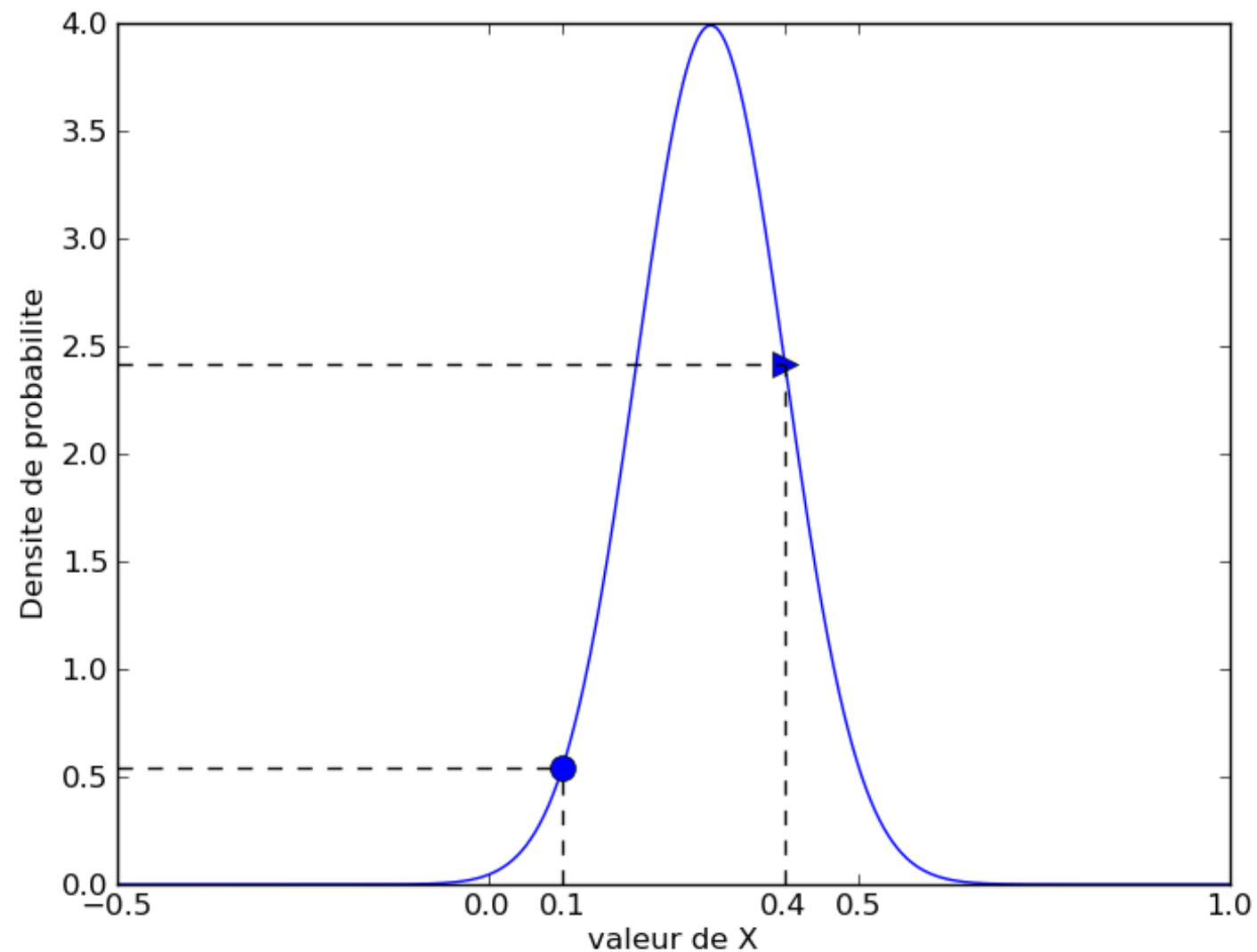
Introduction générale

Structure typique d'un problème d'apprentissage automatique



Continuous density interpretation

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

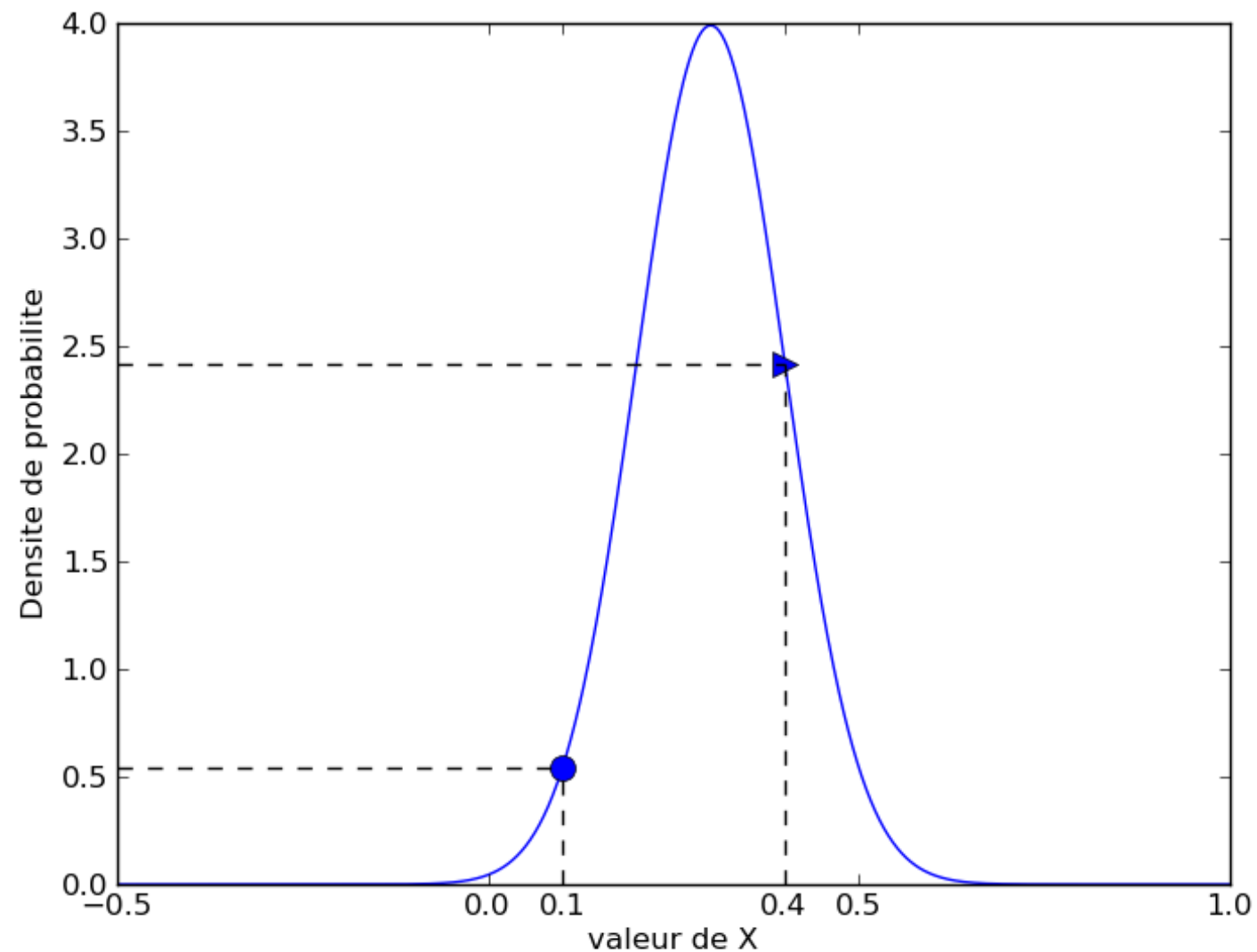


- 1) La densité de probabilité de X prend une valeur supérieure 1 en $X = 0.4$. Cela vous paraît-il normal? Justifiez votre réponse.

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



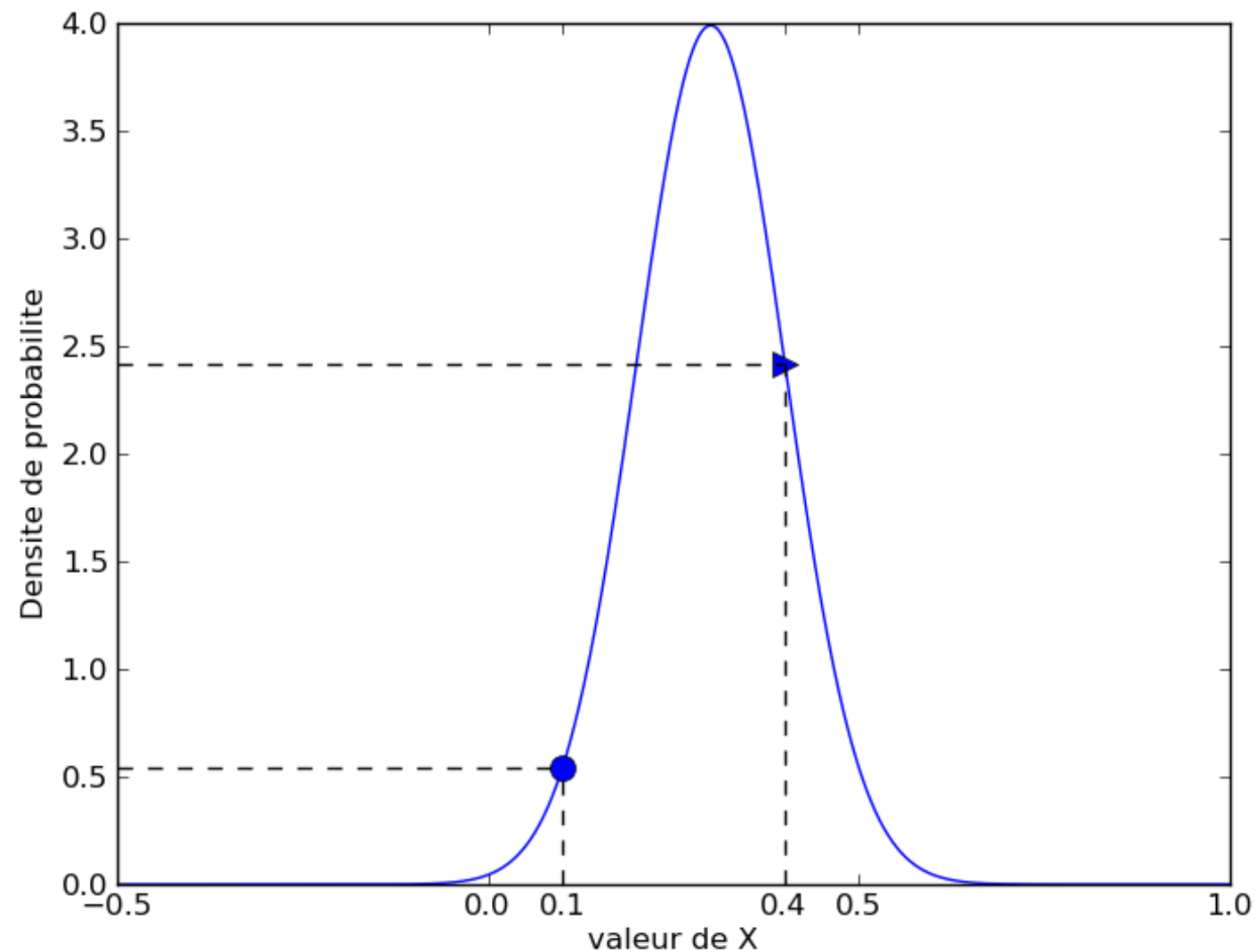
- 1) La densité de probabilité de X prend une valeur supérieure 1 en $X = 0.4$. Cela vous paraît-il normal? Justifiez votre réponse.

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 représenté 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 paraît-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 $x = 0.1$.

Calcul des probabilités

Modèles graphiques (dirigés)

Cadre très courant en apprentissage automatique : variables aléatoires $X_1, X_2, \dots, X_n$ avec densité de probabilité pour la loi jointe $p(X_1, X_2, \dots, X_n)$ se factorisant dans un modèle graphique dirigé $G = (1, 2, \dots, n, E)$, c'est à dire qu'on peut écrire :

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

où π_i indique les parents du nœud i dans le graphe G .

Rappels importants

Marginales

$$p(X_i) = \int p(X_1, X_2, \dots, X_n) dX_{j \neq i}$$

$$p(X_i, X_j) = \int p(X_1, X_2, \dots, X_n) dX_{k \notin \{i, j\}}$$

...

Conditionnelles

$$p(X_i | X_j) p(X_j) = p(X_i, X_j)$$

$$p(X_i | X_j, X_k) p(X_j, X_k) = p(X_i, X_j, X_k)$$

$$p(X_i, X_j | X_k) p(X_k) = p(X_i, X_j, X_k)$$

...

Calcul des probabilités

Problème considéré : calculer $p(X_{i_1}, X_{i_2}, \dots, X_{i_k} | X_{j_1}, X_{j_2}, \dots, X_{j_l})$

Méthode générale :

- exprimer la probabilité conditionnelle en termes de probabilités marginales
- utiliser les independences conditionnelles spécifiées par le modèle graphique pour calculer ces probabilités marginales

Loi Gaussienne multivariée

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 $k = 0, 1, \dots, n$	np	$np(1 - p)$
$Geometric(p)$	$p(1 - p)^{k-1}$ for $k = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$\frac{e^{-\lambda} \lambda^k}{k!}$ for $k = 0, 1, \dots$	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 $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

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

Random Vectors

Given n RV's $X_1, \dots, X_n$, 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 $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$, 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 .

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

applying linearity of expectation and the fact that $\text{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$, 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 $\Sigma = \text{diag}(\text{Var}[X_1], \dots, \text{Var}[X_n])$

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 $\Sigma = \text{Var}[X_1] = \sigma^2$, and so

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

Some Nice Properties of MV Gaussians

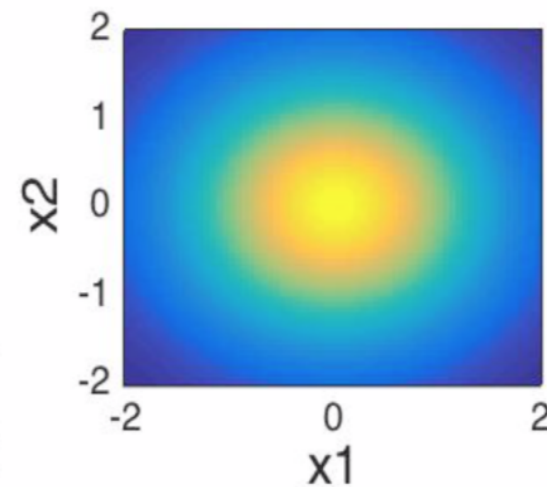
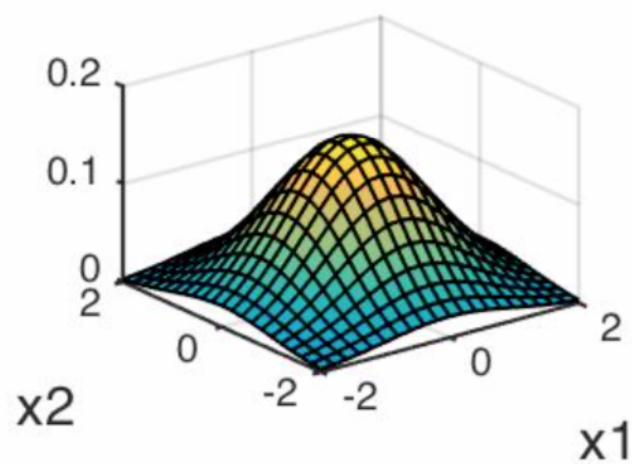
- ▶ Marginals and conditionals of a joint Gaussian are Gaussian
- ▶ A d -dimensional Gaussian $X \in \mathcal{N}(\mu, \Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2))$ is equivalent to a collection of d **independent** Gaussians $X_i \in \mathcal{N}(\mu_i, \sigma_i^2)$. This results in isocontours 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

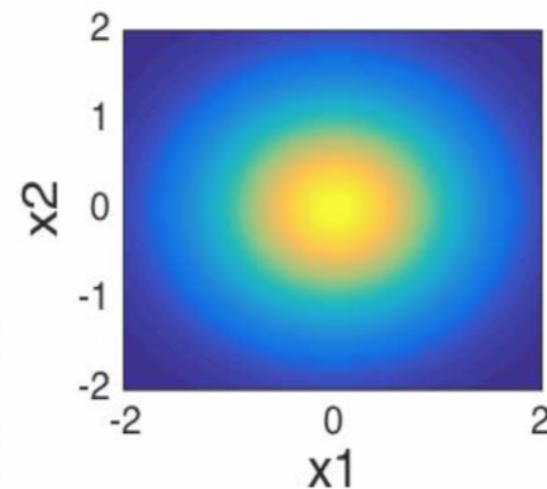
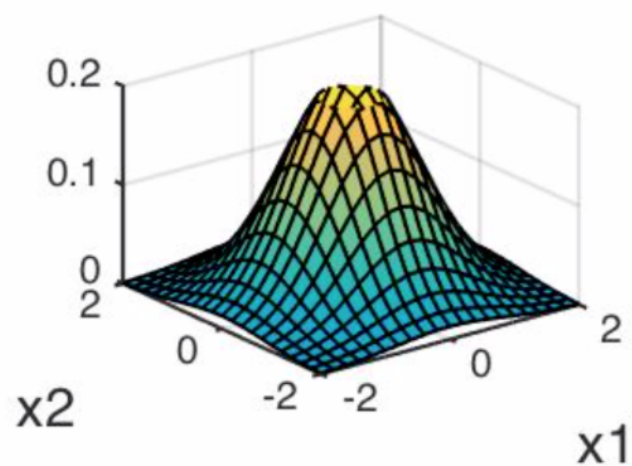
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

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



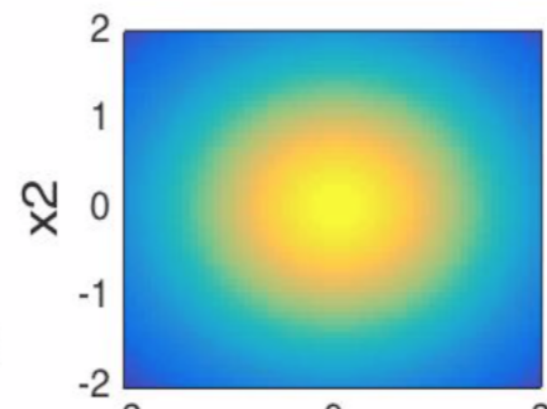
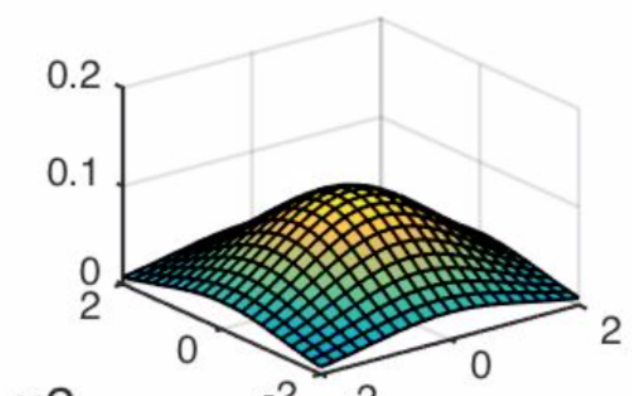
$$\Sigma = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.7 \end{bmatrix}$$

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



$$\Sigma = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}$$

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

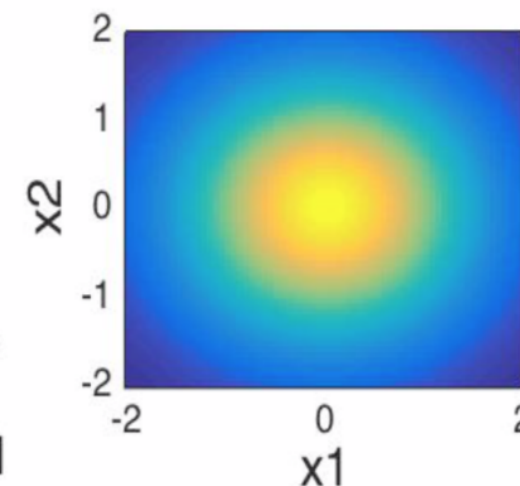
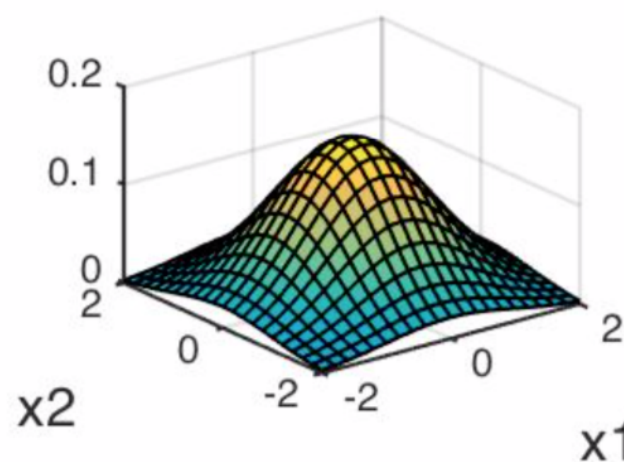


Visualizations of MV Gaussians

If $\text{Var}[X_1] \neq \text{Var}[X_2]$:

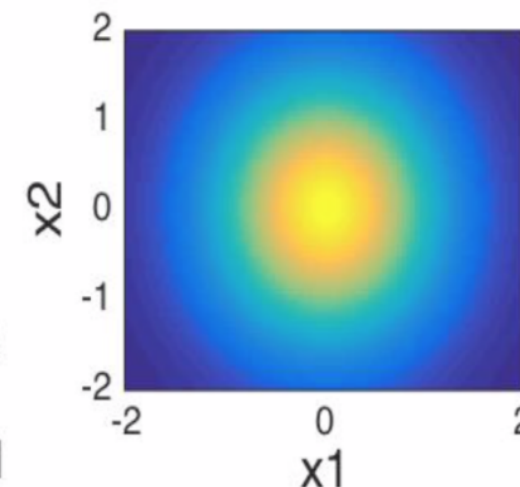
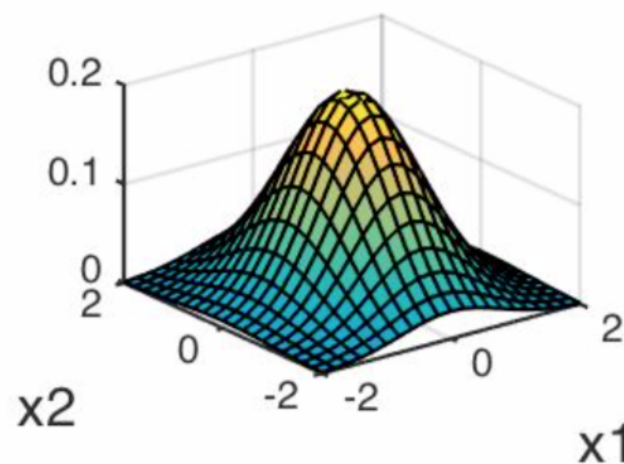
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

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



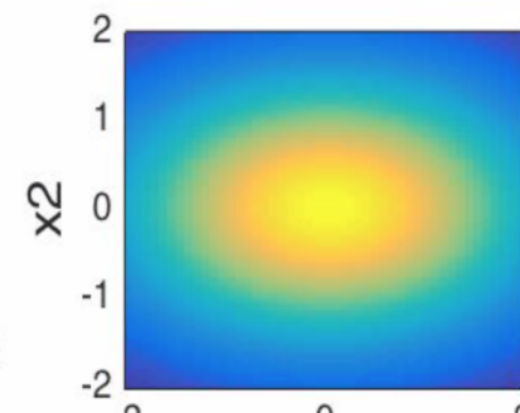
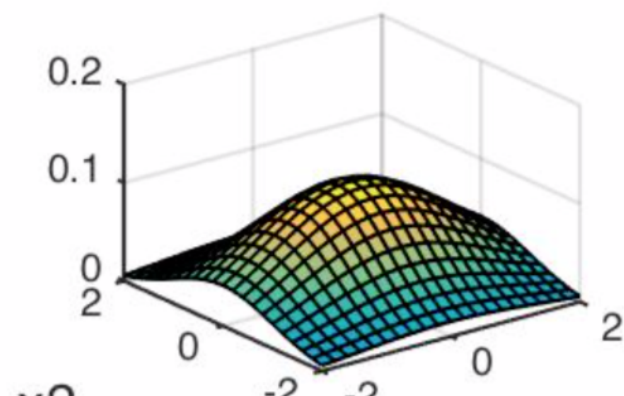
$$\Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$

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



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

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

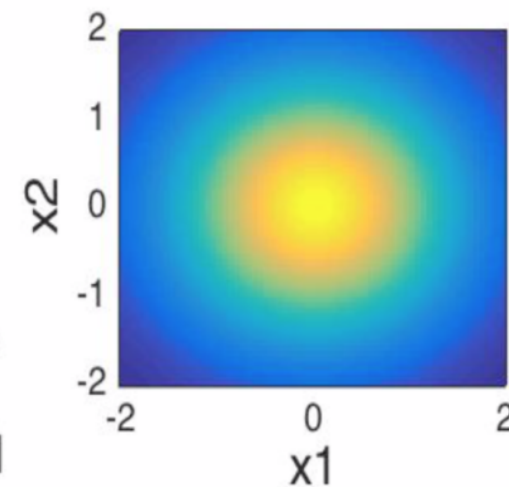
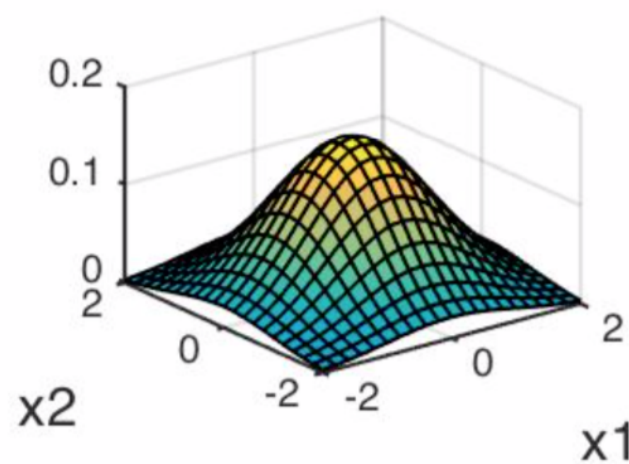


Visualizations of MV Gaussians

If X_1 and X_2 are positively correlated:

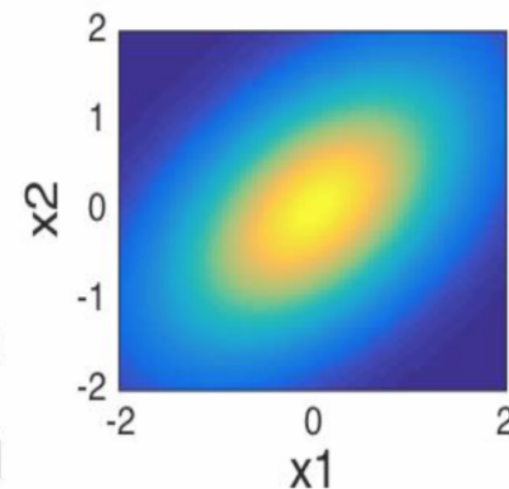
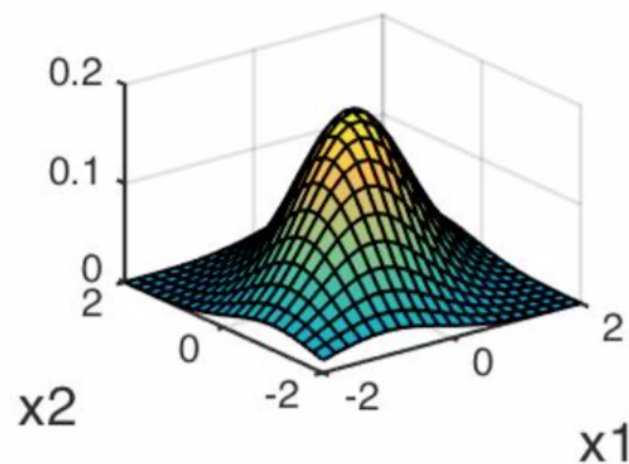
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

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



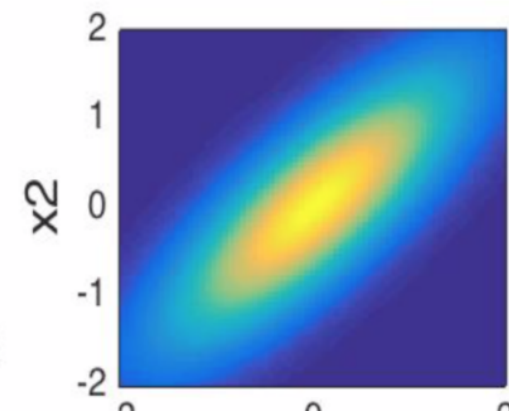
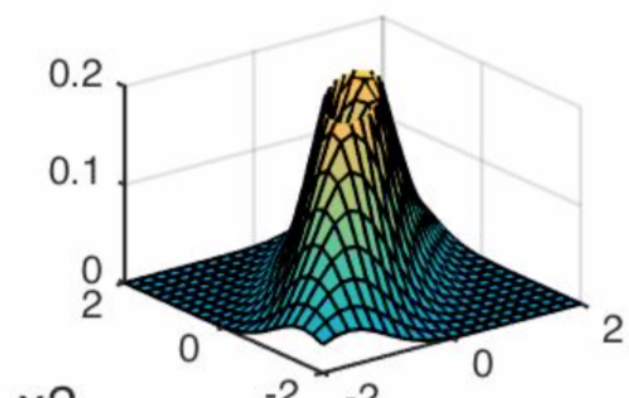
$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

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



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

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

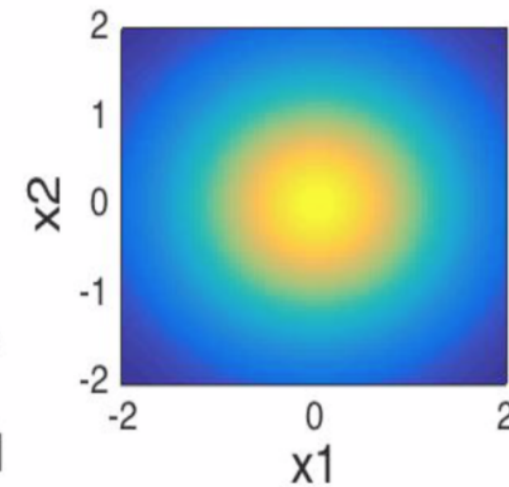
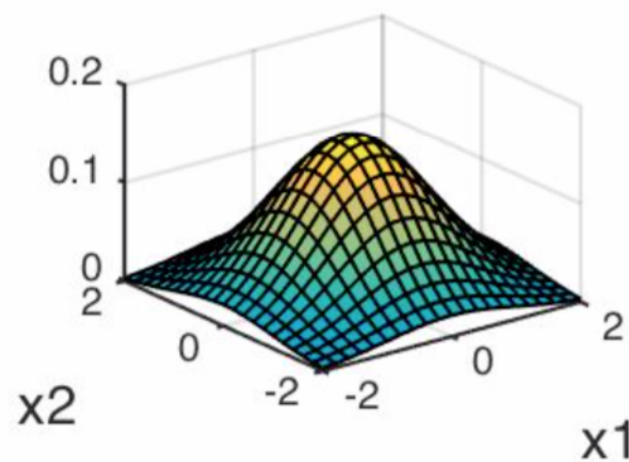


Visualizations of MV Gaussians

If X_1 and X_2 are negatively correlated:

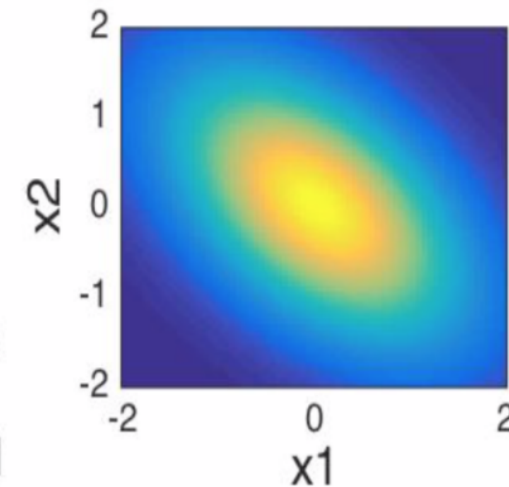
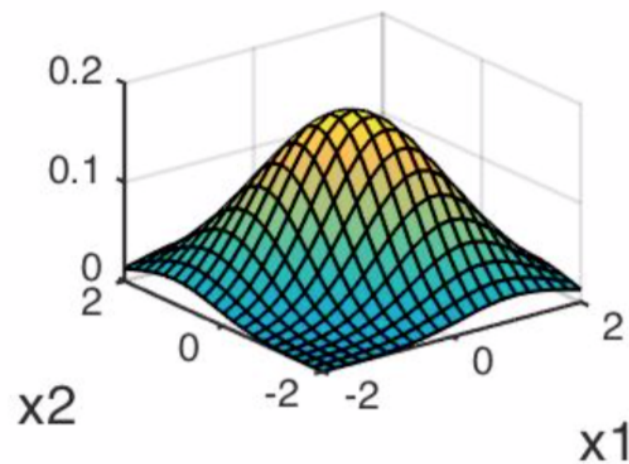
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

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



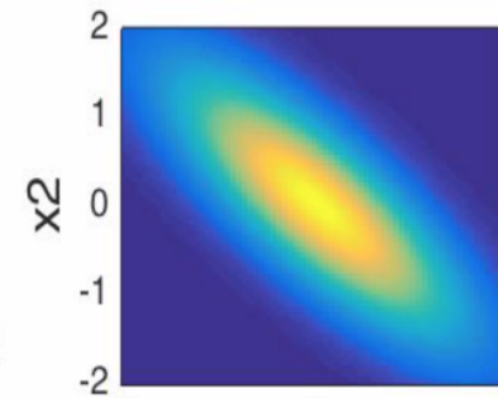
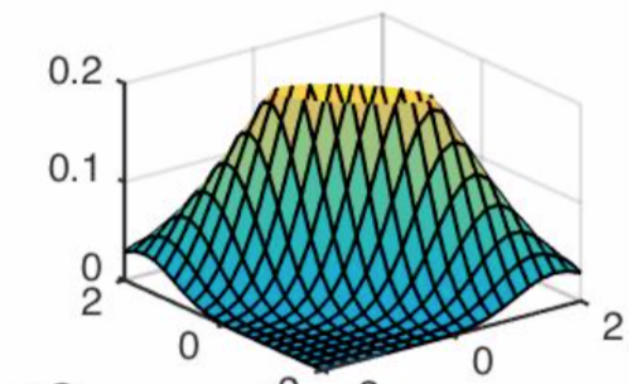
$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

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



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

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



Multivariate Gaussian

Définition générale

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

Distributions conditionnelles

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

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

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

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

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

Distributions marginales ?