Mathématiques pour l'intelligence artificielle

UE MIA, M2 IAAA, AMU, 2022-2023

Thomas Schatz

Vendredi 16 septembre 2022

Point sur les scribes :

- Possible de rendre un seul document par binôme
- Rendez le fichier .tex et le pdf
- Pour avoir le maximum de points : complet et self-contained (bonne base de révision pour vos camarades)

Plan du cours (prévisionnel)

9 séances de 3h

Partie 1 (DM1) : algèbre linéaire et probabilités

1. Notions de bases sur les preuves (+ Algèbre linéaire?)
2. Algèbre linéaire (+ Probabilités?)
3. Probabilités

Partie 2 (DM2): statistique et optimisation

4. Statistiques
5. Optimisation

Partie 3 (DM3):

6. Optimisation sous contraintes
7. Optimisation stochastique
8. Théorie de l'apprentissage

9. Putting it all together

Probabilités

1. Revue du calcul probabiliste, sur la base des transparents de Ding & Khani, Stanford CS229, April 2022 (revus , réordonnés et augmentés)

2. Quelques points plus avancés (intro théorie de la mesure et théorèmes de convergence)

Graphical model

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

Properties of Expectation

Law of Total Expectation
Given two RVs X, Y:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$$

N.B. $\mathbb{E}[X \mid Y] = \sum_{x \in Val(x)} x p_{X|Y}(x|y)$ is a function of Y.
See Appendix for details :)

Example of Law of Total Expectation

El Goog sources two batteries, A and B, for its phone. A phone with battery A runs on average 12 hours on a single charge, but only 8 hours on average with battery B. El Goog puts battery A in 80% of its phones and battery B in the rest. If you buy a phone from El Goog, how many hours do you expect it to run on a single charge?

Example of Law of Total Expectation

El Goog sources two batteries, A and B, for its phone. A phone with battery A runs on average 12 hours on a single charge, but only 8 hours on average with battery B. El Goog puts battery A in 80% of its phones and battery B in the rest. If you buy a phone from El Goog, how many hours do you expect it to run on a single charge?

Solution: Let L be the time your phone runs on a single charge. We know the following:

- $p_X(A) = 0.8$, $p_X(B) = 0.2$,
- $\mathbb{E}[L \mid A] = 12$, $\mathbb{E}[L \mid B] = 8$.

Then, by Law of Total Expectation,

$$\begin{aligned}\mathbb{E}[L] = \mathbb{E}[\mathbb{E}[L \mid X]] &= \sum_{X \in \{A,B\}} \mathbb{E}[L \mid X] p_X(X) \\ &= \mathbb{E}[L \mid A] p_X(A) + \mathbb{E}[L \mid B] p_X(B) \\ &= 12 \times 0.8 + 8 \times 0.2 = \boxed{11.2}\end{aligned}$$

Covariance

Intuitively: measures how much one RV's value tends to move with another RV's value. For RV's X, Y:

$$\begin{aligned} Cov[X, Y] &:= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

- If $Cov[X, Y] < 0$, then X and Y are negatively correlated
- If $Cov[X, Y] > 0$, then X and Y are positively correlated
- If $Cov[X, Y] = 0$, then X and Y are uncorrelated

Properties Involving Covariance

- If $X \perp Y$, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Thus,

$$Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$$

  This is unidirectional! $Cov[X, Y] = 0$ **does not imply** $X \perp Y$
- **Variance of two variables:**

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

  i.e. if $X \perp Y$, $Var[X + Y] = Var[X] + Var[Y]$.
- **Special Case:**

$$Cov[X, X] = \mathbb{E}[XX] - \mathbb{E}[X]\mathbb{E}[X] = Var[X]$$

Variance of a sum

$$\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}(X_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{Cov}(X_i, X_j)$$

Exercice : espérance et variance de la moyenne de n variables aléatoires i.id. ?

Variance of a sum

$$\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}(X_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{Cov}(X_i, X_j)$$

Exercice : espérance et variance de la moyenne de n variables aléatoires i.id. ?

Exercice : On a une procédure aléatoire pour entrainer un classificateur binaire, dont on obtient n échantillons (n classifieurs). Pour tester la qualité de la procédure d'entrainement, on a une procédure aléatoire de test qui produit une erreur de classification et qu'on applique m fois sur chacun des n classificateurs entrainé . Comment mesurer la variance de l'erreur de classification (par exemple pour savoir si elle est significativement en dessous du hasard) à partir des erreurs de classifications $(e_{i,j})_{1 \leq i \leq n, 1 \leq j \leq m}$?

Variance of a sum

$$\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}(X_i) + 2 \sum_{1 \le i < j \le n} \operatorname{Cov}(X_i, X_j)$$

Exercice : espérance et variance de la moyenne de n variables aléatoires i.id. ?

Law of total variance

$$\operatorname{Var}(Y) = \operatorname{E}[\operatorname{Var}(Y \mid X)] + \operatorname{Var}(\operatorname{E}[Y \mid X]).$$

Exercice : On a une procédure aléatoire pour entrainer un classificateur binaire, dont on obtient n échantillons (n classifieurs). Pour tester la qualité de la procédure d'entrainement, on a une procédure aléatoire de test qui produit une erreur de classification et qu'on applique m fois sur chacun des n classificateurs entrainé . Comment mesurer la variance de l'erreur de classification (par exemple pour savoir si elle est significativement en dessous du hasard) à partir des erreurs de classifications $(e_{i,j})_{1 \le i \le n, 1 \le j \le m}$?

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

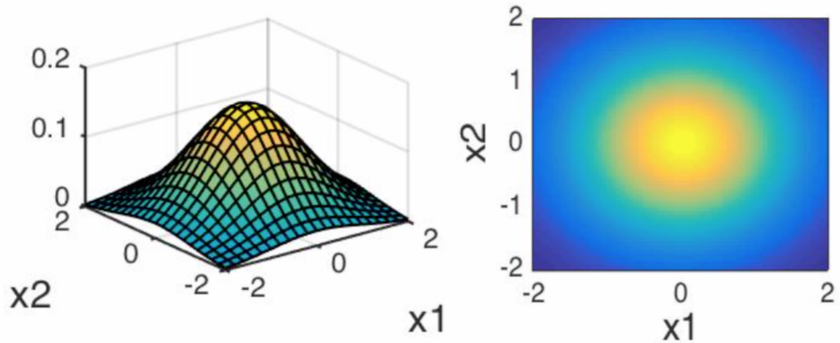
- Marginals and conditionals of a joint Gaussian are Gaussian
- A d-dimensional Gaussian $X \in \mathcal{N}(\mu, \Sigma = diag(\sigma_1^2, ..., \sigma_n^2))$ is equivalent to a collection of d **independent** Gaussians $X_i \in \mathcal{N}(\mu_i, \sigma_i^2)$. This results in isocontours aligned with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

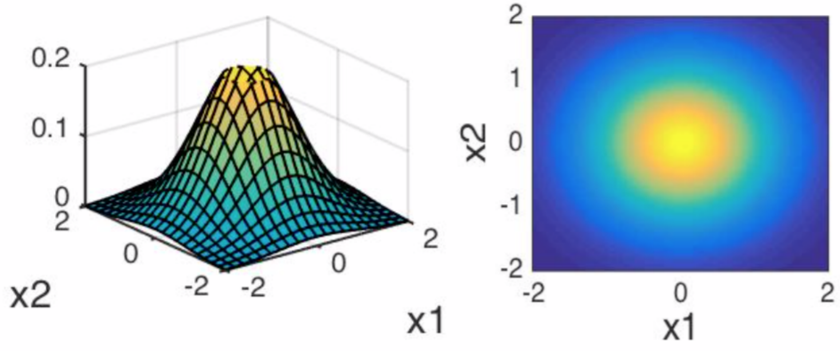
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



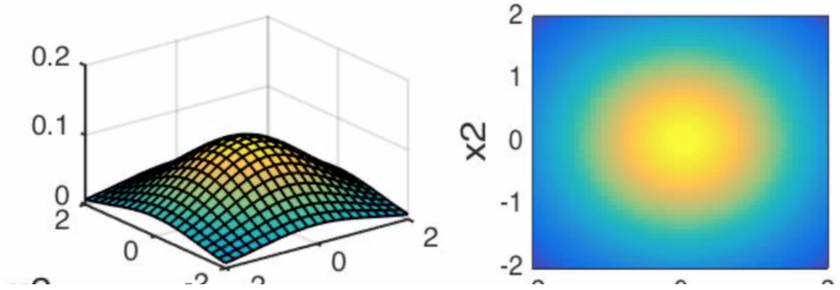
$$\Sigma = \begin{matrix} 0.7 & 0 \\ 0 & 0.7 \end{matrix}$$

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



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

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

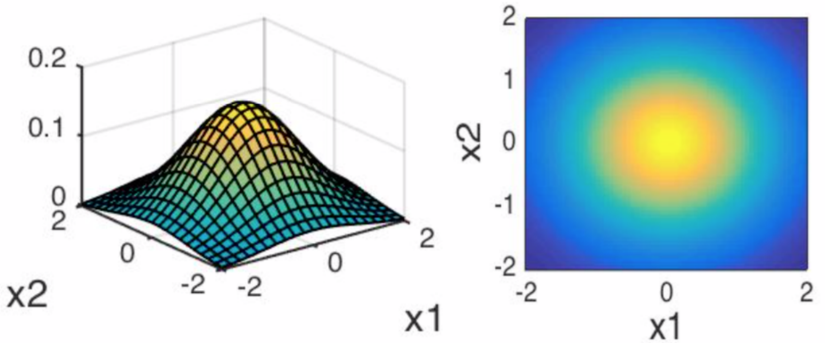


Visualizations of MV Gaussians

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

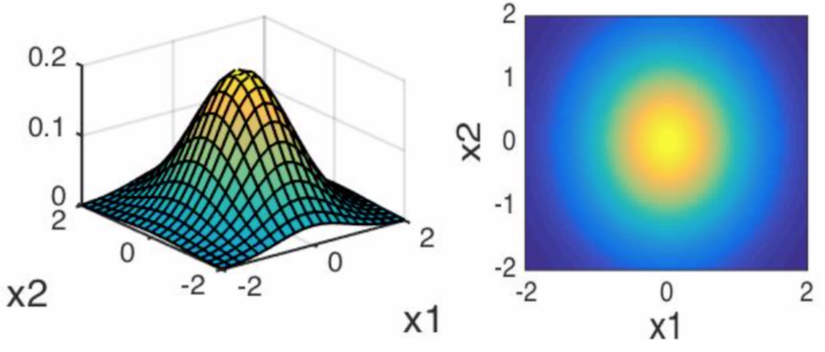
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



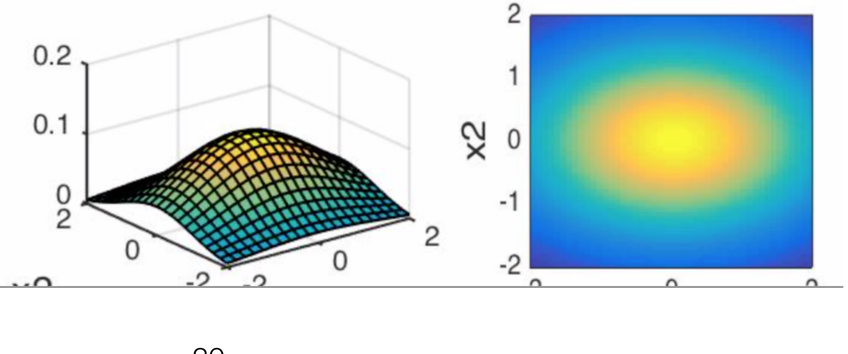
$$\Sigma = \begin{matrix} 0.6 & 0 \\ 0 & 1 \end{matrix}$$

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



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

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

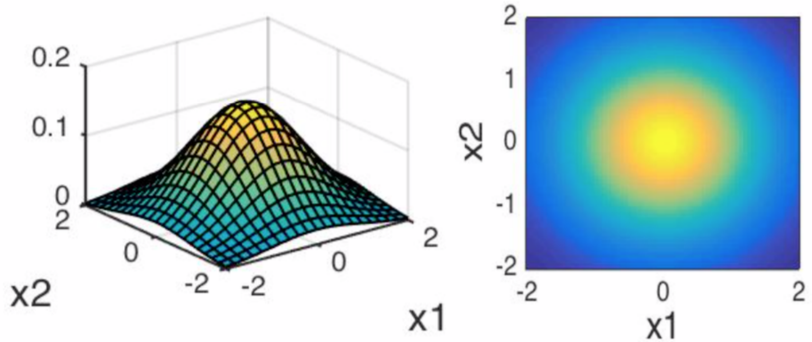


Visualizations of MV Gaussians

If X_1 and X_2 are positively correlated:

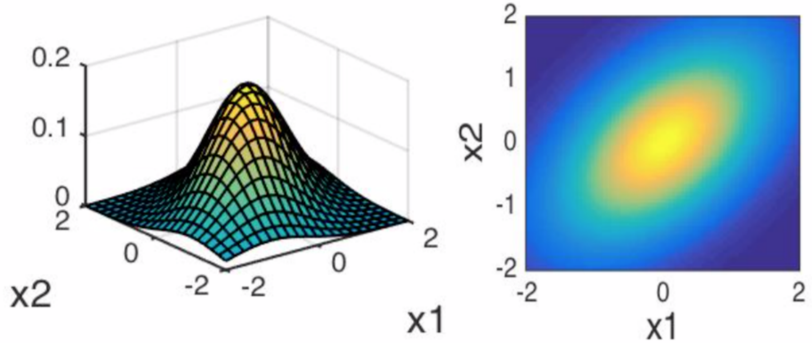
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



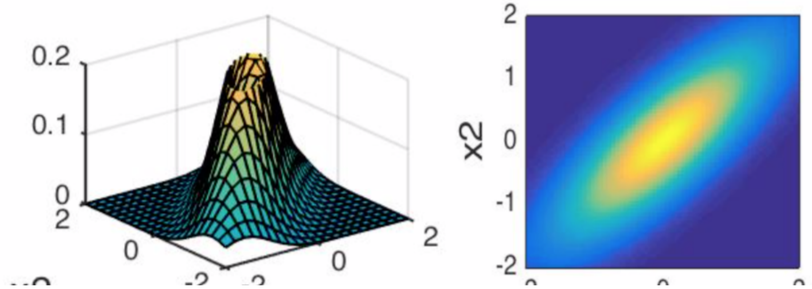
$$\Sigma = \begin{matrix} 1 & 0.5 \\ 0.5 & 1 \end{matrix}$$

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



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

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

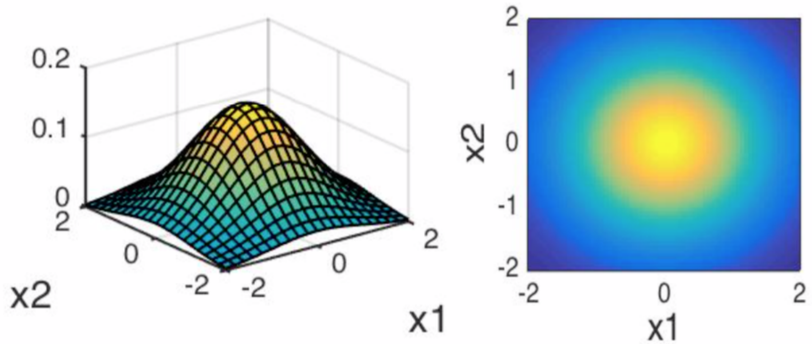


Visualizations of MV Gaussians

If X_1 and X_2 are negatively correlated:

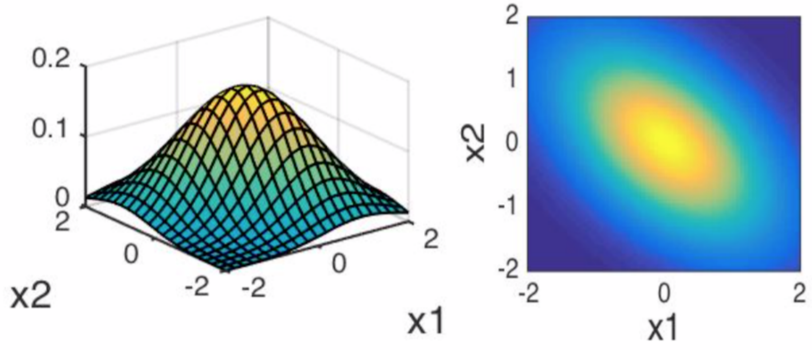
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



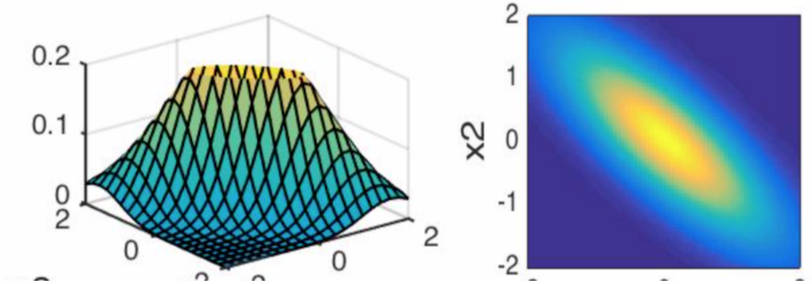
$$\Sigma = \begin{matrix} 1 & -0.5 \\ -0.5 & 1 \end{matrix}$$

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



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

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



Multivariate Gaussian

Définition générale

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

Distributions conditionnelles

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

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

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

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

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

Distributions marginales ?

Appendix: More on Total Expectation

Why is $\mathbb{E}[X|Y]$ a function of Y? Consider the following:

- $\mathbb{E}[X|Y = y]$ is a scalar that only depends on y.
- Thus, $\mathbb{E}[X|Y]$ is a random variable that only depends on Y. Specifically, $\mathbb{E}[X|Y]$ is a function of Y mapping $Val(Y)$ to the real numbers.

An example: Consider RV X such that

$$X = Y^2 + \epsilon$$

such that $\epsilon \sim \mathcal{N}(0, 1)$ is a standard Gaussian. Then,

- $\mathbb{E}[X|Y] = Y^2$
- $\mathbb{E}[X|Y = y] = y^2$

Appendix: More on Total Expectation

A derivation of Law of Total Expectation for discrete X, Y:[3]

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[\sum_x xP(X = x \mid Y)] \tag{1}$$

$$= \sum_y \sum_x xP(X = x \mid Y)P(Y = y) \tag{2}$$

$$= \sum_y \sum_x xP(X = x, Y = y) \tag{3}$$

$$= \sum_x x \sum_y P(X = x, Y = y) \tag{4}$$

$$= \sum_x xP(X = x) = \boxed{\mathbb{E}[X]} \tag{5}$$

where (1), (2), and (5) result from the definition of expectation, (3) results from the definition of cond. prob., and (5) results from marginalizing out Y.

[3] from slides by Koochak & Irvin

Appendix: A proof of Conditioned Bayes Rule

Repeatedly applying the definition of conditional probability, we have: [4]

$$\begin{aligned}
\frac{P(b|a,c)P(a|c)}{P(b|c)} &= \frac{P(b,a,c)}{P(a,c)} \cdot \frac{P(a|c)}{P(b|c)} \\
&= \frac{P(b,a,c)}{P(a,c)} \cdot \frac{P(a,c)}{P(b|c)P(c)} \\
&= \frac{P(b,a,c)}{P(b|c)P(c)} \\
&= \frac{P(b,a,c)}{P(b,c)} \\
&= P(a|b,c)
\end{aligned}$$

[4] from slides by Koochak & Irvin

Théorie de la mesure ?

- Quel intérêt : traitement rigoureux, unifié et général de la théorie des probabilité

  - Unifié : e.g. PMF et PDF deviennent deux cas particuliers d'un concept plus général, pas de différence à faire entre somme discrètes et intégrales continues

  - Général : variables mixtes (ni continues ni discrètes), variables plus compliquées (e.g. signal), théorèmes de convergence plus simples et plus forts qu'avec l'intégrale de Riemann (utile en statistique)

Théorie de la mesure ?

- Difficulté :
  - Plus abstrait
  - Le problème de la mesurabilité des fonctions

Elements of Probability

Sample Space Ω

$\{HH, HT, TH, TT\}$

Event $A \subseteq \Omega$

$\{HH, HT\}, \Omega$

Event Space $\mathcal{F}$

Probability Measure $P : \mathcal{F} \to \mathbb{R}$

$P(A) \geq 0 \quad \forall A \in \mathcal{F}$

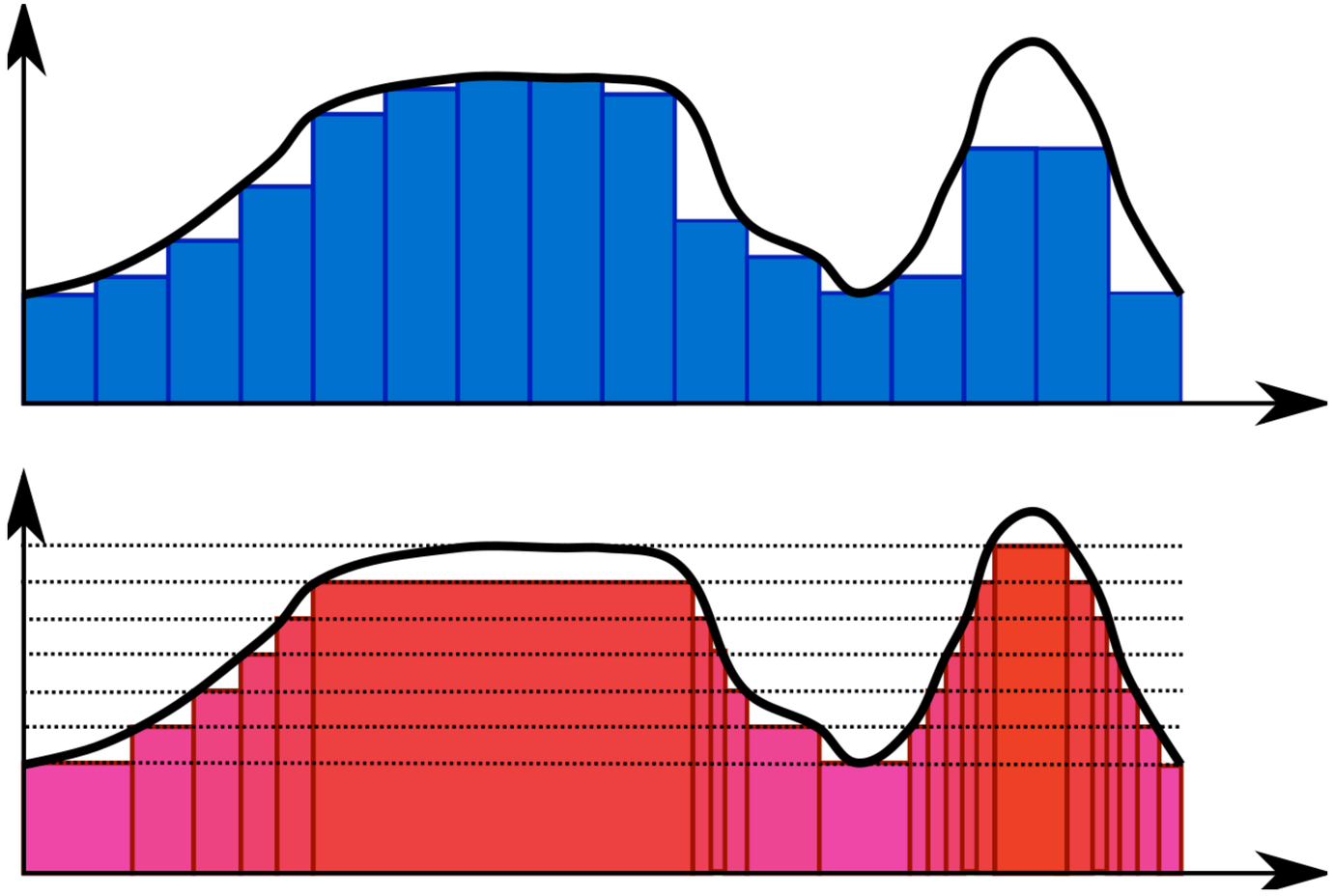
$P(\Omega) = 1$

countable

If $A_1, A_2, \ldots$ disjoint set of events ($A_i \cap A_j = \emptyset$ when $i \neq j$), then

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

Espérance d’une variable aléatoire: intégrale de Lebesgue



Convergence

- Convergence presque sûre ou presque partout
- Convergence en probabilité
- Convergence en loi

Loi forte des grands nombres

$X_1, X_2, \ldots$ variables aléatoires i.i.d.

(ii) (The SLLN). A necessary and sufficient condition for the existence of a constant c for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} c \tag{1.81}$$

is that $E|X_1| < \infty$, in which case $c = EX_1$

Théorème de la limite centrale

(Multivariate CLT). Let $X_1, ..., X_n$ be i.i.d. random k-vectors with a finite $\Sigma = \mathrm{Var}(X_1)$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - EX_1) \to_d N_k(0, \Sigma). \quad \blacksquare$$

Transformations continues et théorème de Slutsky

Theorem 1.10. Let $X, X_1, X_2, ...$ be random k-vectors defined on a probability space and g be a measurable function from $(\mathcal{R}^k, \mathcal{B}^k)$ to $(\mathcal{R}^l, \mathcal{B}^l)$. Suppose that g is continuous a.s. P_X. Then
(i) $X_n \to_{a.s.} X$ implies $g(X_n) \to_{a.s.} g(X)$;
(ii) $X_n \to_p X$ implies $g(X_n) \to_p g(X)$;
(iii) $X_n \to_d X$ implies $g(X_n) \to_d g(X)$. ∎

Theorem 1.11 (Slutsky's theorem). Let $X, X_1, X_2, ...,\ Y_1, Y_2, ...$ be random variables on a probability space. Suppose that $X_n \to_d X$ and $Y_n \to_p c$, where c is a fixed real number. Then
(i) $X_n + Y_n \to_d X + c$;
(ii) $Y_n X_n \to_d cX$;
(iii) $X_n / Y_n \to_d X/c$ if $c \neq 0$.

Inégalités de concentration

- Hoeffding’s bound