Chapter 1: Random variables and sampling

1.1. Basic probability theory

Ross Chaps 1-2

- See Ross for details
- Sample space : S or Ω
- Event : $E \subseteq \Omega$
- Probability :
  1- $0 \leq P(E) \leq 1$
  2- $P(S) = 1$
  3- $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$ $\quad E_i$ mutually exclusive

- Operations :

$$P(A \cup B) = P(A \text{ or } B)$$
$$P(A \cap B) = P(A \text{ and } B)$$
$$P(A^c) = 1 - P(A) = P(\text{not } A)$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Mutually exclusive
$P(A \cap B) = 0$
$A \cap B = \emptyset$

1.2. Conditional probabilities

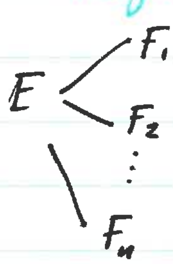
Ross Chap.3

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E, F)}{P(F)}, \qquad P(F) > 0$$

- Interpretations:
  1- Prob of E given F happens / is observed
  2- Prob of E in subsets of events in which F is satisfied

- Total probability: $P(E) = \sum_i P(E|F_i)\, P(F_i)$

marginal $\qquad$ conditional over alternatives



· Bayes's formula:

$$P(F|E) = \frac{P(E|F)\,P(F)}{P(E)}$$

F: event / hypothesis; E: evidence

· Interpretation: Hypothesis → evidence → update

$P(F)$ prior, $P(E)$, $P(F|E)$ posterior

· General formula:
$$P(F_j|E) = \frac{P(E|F_j)\,P(F_j)}{\sum_i P(E|F_i)P(F_i)}$$

· Bayesian inference / modelling:

System — Distribution, dyn. system, etc.

↓

Data — Observations, measurements (noisy)

↓

Model — Distribution, neural net, eqs of motion

Parameters: θ

$$P(\theta|D,M) = \frac{P(D|\theta,M)\,P(\theta,M)}{P(D|M)}$$

or just

$$\underbrace{P(\theta|D)}_{\text{Posterior}} = \frac{\overbrace{P(D|\theta)}^{\text{likelihood}}\,\overbrace{P(\theta)}^{\text{Prior}}}{P(D)}$$

$P(D)$ ← evidence / normalization

·Tasks:

1- Parameter estimation: Estimate θ from D : $p(\theta|D)$

2- Model comparison: $P(M_i|D)$

3- Prediction: Generate new (random) data from learnt model

- Independence: A, B independent if $A \perp\!\!\!\perp B$
  - $P(AB) = P(A \cap B) = P(A)P(B)$
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$

- Naive Bayes: $P(D_1, D_2, \ldots, D_n) = \prod_{i=1}^{n} P(D_i)$

Ross Chap. 4

1.3. Discrete random variables (RVs)

- RV X defined by
  - Set of possible values
  - Probability for each value

- Notation: $\underset{RV}{X} = \underset{value}{x}$ $\quad P(X=x)$ or $P\{X=x\}$ or $P(x)$ $\quad \sum_x P(x) = 1$

- Expectation: $E[X] = \sum_x x\, P(X=x) = \sum_x x\, P(x)$

$$E[g(X)] = \sum_x g(x)\, P(x)$$

- Variance: $\mathrm{var}(X) = E[(X-\mu)^2]$ $\quad \mu = E[X]$
  $= E[X^2] - E[X]^2 \geq 0$
  $\mathrm{var}(aX+b) = a^2\, \mathrm{var}(X)$

- Standard deviation: $\sigma(X) = \sqrt{\mathrm{var}(X)}$

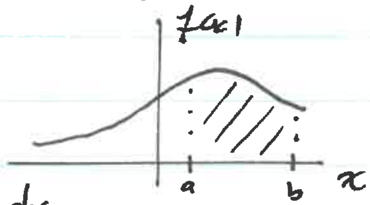
- Bernoulli RV: $X \in \{0,1\}$ $\quad P(1) = p \quad P(0) = 1-p$ $\quad X \sim \mathrm{Bern}(p)$
- Binomial RV:
  - Trial: Success / failure $\quad$ 1/0 $\quad$ true/false $\quad$ H/T
  - X = # successes in n independent trials
  - $X \in \{0, 1, 2, \ldots, n\}$
  - $P(x) = \binom{n}{x} p^x (1-p)^{n-x}$ $\quad X \sim \mathrm{Bin}(n,p)$

  - $E[X] = np \quad \mathrm{var}(X) = np(1-p)$

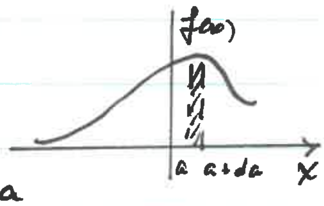
- Poisson RV: $X \sim \text{Poisson}(\lambda)$
  - $X \in \{0, 1, \dots\}$
  - $P(X=i) = e^{-\lambda} \dfrac{\lambda^i}{i!}$ $\lambda > 0$
  - $E[X] = \lambda$ $\text{var}(X) = \lambda$

1.4. Continuous random variables

- Probability density function (pdf): $p_X(x)$ or $f_X(x)$
  $p(x)$ $f(x)$
  - $X \in \mathbb{R}$
  - $f(x) \geq 0$
  - $\int_{\mathbb{R}} f(x)\, dx = 1$



  - $P(X \in [a,b]) = P(a \leq X \leq b) = \int_a^b f(x)\, dx$
  - $P(X \in A) = \int_A f(x)\, dx$



  - Interpretation: $P(X \in [a, a+da]) = f(a)\, da$

- Cumulative distribution function (cdf)
  $$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\, dy$$

- Expectation: $E[X] = \int_{\mathbb{R}} x\, f(x)\, dx$
  $E[g(X)] = \int_{\mathbb{R}} g(x) f(x)\, dx$

- Joint pdf: $p_{XY}(x,y)$ $p_X(x) = \int p_{XY}(x,y)\, dy$
  $p_Y(y) = \int p_{XY}(x,y)\, dx$
  $p(x,y) = p(x)\, p(y)$ if $X \perp\!\!\!\perp Y$

- Uniform RV: $X \sim \mathcal{U}[0,1]$
  - $X \in [0,1]$
  - $f(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$

- Exponential RV: $X \sim \text{Exp}(\lambda)$
  - $X \geq 0 \qquad X \in \mathbb{R}_+$
  - $f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

- Normal / Gaussian RV: $X \sim \mathcal{N}(\mu, \sigma^2)$
  - $X \in \mathbb{R}$
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-(x-\mu)^2/2\sigma^2}$
  - $E[X] = \mu \qquad \text{Var}(X) = \sigma^2$
  - Standardization: $Y = \frac{X-\mu}{\sigma^2} \sim \mathcal{N}(0,1)$

- Cauchy / Lorentz: $X \sim \text{Cauchy}(a, \sigma)$
  - $X \in \mathbb{R}$
  - $f(x) = \frac{1}{\pi} \, \frac{\sigma}{(x-a)^2 + \sigma^2}$
  - $E[X]$ undefined $\qquad \text{Var}(X) = \infty$

- Bivariate Gaussian: $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$
  - $\vec{X} = (X, Y) \in \mathbb{R}^2$
  - $p(x,y) = \frac{1}{2\pi \, |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right)$

$$\vec{\mu} = (\mu_x, \mu_y) \qquad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

Symmetric

1.5. Transformation of RVs

$$X \rightarrow \boxed{g} \rightarrow Y$$

- Proposition:
  - X continuous RV
  - $Y = g(X)$
  - g differentiable and monotonic ↗ g^{-1} exists

Then

$$p_Y(y) = \begin{cases} p_X(g^{-1}(y)) \left| \dfrac{d\, g^{-1}(y)}{dy} \right| & y = g^{-1}(x) \text{ exists} \\ 0 & \text{otherwise} \end{cases}$$

- Rem: $p_Y(y)\,dy = p_X(x)\,dx \Rightarrow p_Y(y) = p_X(x) \dfrac{dx}{dy}$

- General case:

$$p_Y(y) = \sum_{x \in g^{-1}(y)} p_X(g^{-1}(y)) \left| \frac{d\, g^{-1}(y)}{dy} \right|$$

pre-image

- Many dimensions:
  - $\vec{X}$
  - $\vec{Y} = g(\vec{X})$ differentiable / invertible

$$p_{\vec{Y}}(\vec{y}) = p_{\vec{X}}(\underbrace{\vec{x}(\vec{y})}_{g^{-1}(\vec{y})}) \underbrace{\left| \frac{\partial \vec{x}}{\partial \vec{y}} \right|}_{\text{det of Jacobian matrix}}$$

$$\frac{\partial \vec{x}}{\partial \vec{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots \\ \frac{\partial x_2}{\partial y_1} & \ddots & \\ \vdots & & \end{pmatrix}$$

Matrix of partial derivatives

- Example: $X \sim \mathcal{N}(\mu, \sigma^2)$ $\quad Y = aX + b$
  - $Y = g(X) = aX + b$
  - $X = g^{-1}(Y) = \dfrac{Y-b}{a} \Rightarrow \dfrac{dx}{dy} = \dfrac{d\, g^{-1}(y)}{dy} = \dfrac{1}{a}$

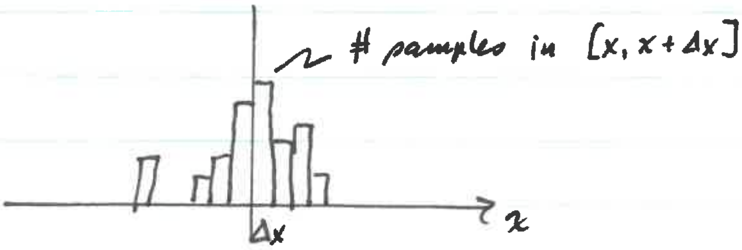
$$\Rightarrow p_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{\left(\frac{y-b}{a} - \mu \right)^2}{2\sigma^2} \right) \frac{1}{|a|}$$

$$= \frac{1}{\sqrt{2\pi\sigma'^2}} \exp\left(-(y-\mu')^2 / 2\sigma'^2 \right)$$

$\mu' = a\mu + b$
$\sigma' = |a|\sigma$

1.6. Histograms

- Data / sample : $\{X_i\}_{i=1}^{L}$ sample size $= L$



- $\hat{N}_{L,\Delta x}(x) = \#$ samples in $[x, x+\Delta x]$ $\sum_x \hat{N}_{L,\Delta x}(x) = L$

- $\hat{P}_{L,\Delta x}(x) = \dfrac{\hat{N}_{L,\Delta x}(x)}{L}$ $\sum_x \hat{P}_{L,\Delta x}(x) = 1$

- $\hat{f}_{L,\Delta x}(x) = \dfrac{\hat{N}_{L,\Delta x}(x)}{L\,\Delta x} = \dfrac{\hat{P}_{L,\Delta x}(x)}{\Delta x}$ $\sum_x \hat{f}_{L,\Delta x}(x)\,\Delta x = 1$
  $\approx \int f(x)\,dx = 1$

$\{X_i\} \rightarrow \hat{N}_{L,\Delta x}(x) \rightarrow \hat{P}_{L,\Delta x}(x) \rightarrow \hat{f}_{L,\Delta x}(x)$

Data	Histogram counts	Normalized histogram	Normalized density histogram
		Empirical distribution	Empirical density

See demo

- Own code :

```
def  myhist(data, a, b, dx):
     L = len(data)
     nbin = int((b-a)/dx)
     hN = np.zeros((1, nbin))
     for i in range(L):
          pos = ⌊(data[i]-a)/dx⌋
          hN[pos] += 1
     return hN/= L*dx
```

If time: demos on transformation of RVs

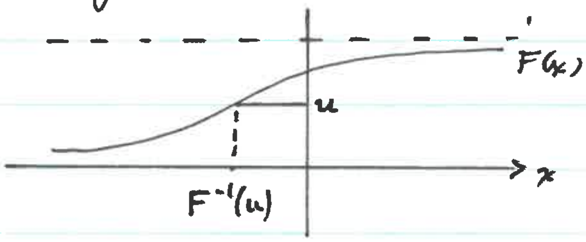
1.7. Pseudo random numbers

- $X \sim \mathcal{U}[0,1]$ uniform random float
- Python : np.random.random()
- Seed : $n_0 \to n_1 \to n_2 \to n_3 \to \cdots$

  np.random.seed(n)

1.8 Inversion method

- X continuous RV
- CDF : $F(x) = P(X \leq x)$
- pdf : $f(x) = F'(x)$
- Inverse of CDF :



$u \leq F(x) \iff F^{-1}(u) \leq x$

monotonic

- Proposition : If $U \sim \mathcal{U}[0,1]$, then $F^{-1}(U)$ has CDF F (and pdf f).

Proof :

$$P(F^{-1}(U) \leq x) = P(U \leq F(x))$$
$$= F(x) \quad \text{since } P(U \leq a) = a$$

$\blacksquare$

- Algorithm : ① Get CDF from pdf
  ② Invert CDF
  ③ $u \sim \mathcal{U}[0,1]$
  ④ Return $x = F^{-1}(u)$

- Example : $\mathrm{Exp}(\lambda)$ $\quad f(x) = \lambda e^{-\lambda x} \quad x \geq 0$

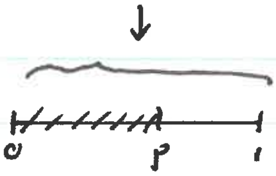
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\,dy = 1 - e^{-\lambda x}$$
$$\Rightarrow u = F(x) = 1 - e^{-\lambda x} \Rightarrow x = F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u)$$

$U \sim \mathcal{U}[0,1]$ ($U \sim 1-U$)

$$\Rightarrow F^{-1}(U) = -\frac{1}{\lambda}\ln U$$

- Example : Bern(p) (not continuous)
  - $X \sim \mathcal{U}[0,1]$
  - $Y = \mathbb{1}_{[0,p]}(X) = \begin{cases} 1 & \text{if } X < p \\ 0 & \text{otherwise} \end{cases}$
  - $P(Y=1) = P(X \in [0,p]) = p$

    $P(Y=0) = 1-p$



- Code :

```
def bern(p):
    r = np.random.random()
    if r < p:
        return 1
    else:
        return 0
```

- Example : Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$
  - $F(x) = P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right)$

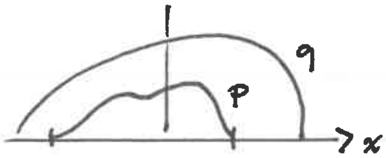
    $= P\left(Z \leq \frac{x-\mu}{\sigma}\right)$

    $= \Phi\left(\frac{x-\mu}{\sigma}\right)$

$\Rightarrow$ F^{-1} involves Φ^{-1}
  - Not known in closed form
  - Can be computed numerically but not efficient

More examples in CW1

1.9. Rejection method



- $X \sim p$ target
- $Y \sim q$ can be simulated easily

  $\exists\ m > 0\ \forall y \quad p(y) \leq m q(y)$

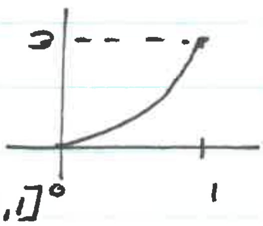
  $\dfrac{p(y)}{m q(y)}$ can be calculated easily

- Algorithm: ① Generate $Y \sim q$

  ② Accept variate with probability $\dfrac{p(y)}{m q(y)}$

  i.e. $U \sim \mathcal{U}[0,1]$ and accept if

  $U \leq \dfrac{p(y)}{mq(y)}$



- Example: $p(x) = 3x^2$ on $x \in [0,1]$

  $p(x) \leq 3 \Rightarrow$ Choose $q \simeq \mathcal{U}[0,1]$

  $\dfrac{p(y)}{mq(y)} = \dfrac{3y^2}{3} = y^2$

  Generate $y \sim \mathcal{U}[0,1]$

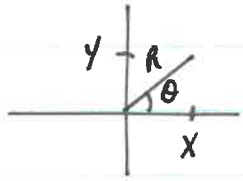
  $u \sim \mathcal{U}[0,1]$

  accept y if $u \leq y^2$

- Rem:
  - Many rejections where $p(x)$ is small
  - Certain rejection when $p(x) = 0$
  - $q = p \Rightarrow$ no rejection

1.10 Box-Muller method

· Exercise in CW1 : $X \sim N(0,1)$, $Y \sim N(0,1)$, $R = \sqrt{X^2 + Y^2}$, $\theta = \arctan \frac{Y}{X}$



$\Rightarrow$ $R \sim$ Rayleigh $p(r) = r e^{-r^2/2}$

$\theta \sim \mathcal{U}[0, 2\pi)$

· Box-Muller algorithm :

1- Generate $R \sim$ Rayleigh
2- Generate $\theta \sim \mathcal{U}[0, 2\pi)$
3- Return $(X, Y) = (R\cos\theta, R\sin\theta)$ 2 std RVs

· Step 1: $F(r) = P(R \leq r) = \int_0^r p(y)\,dy = 1 - e^{-r^2/2}$

$F^{-1}(u) = \sqrt{-2\ln(1-u)}$

$U_1 \sim \mathcal{U}[0,1]$

$\hookrightarrow R = \sqrt{-2\ln(1-U_1)}$ or $\sqrt{-2\ln U_1}$

· Step 2 : $U_2 \sim \mathcal{U}[0,1] \rightarrow \theta = 2\pi U_2 \sim \mathcal{U}[0, 2\pi)$

· Rem : · Must generate 2 uniform RVs
· Can't re-use U_1 for θ
· Generate 2 Gaussians. $\rightarrow$ return only one

1.11 Monte Carlo sampling

- RV: X
- Expectation: $\mu = E[X] = \sum_x x\, P(x)$ or $\int p(x)\, x\, dx$

$$\gamma = E[g(X)] = \sum_x g(x) P(x) \text{ or } \int g(x)\, p(x)\, dx$$

- MC estimation:
  - Generate sample $\{X_i\}_{i=1}^{L}$ $\quad X_i \sim p \quad$ iid
  - Estimator:

$$\hat{\gamma}_L = \frac{1}{L} \sum_{i=1}^{L} g(X_i)$$

  - Law of large numbers (LLN):

$$\lim_{L \to \infty} P(|\hat{\gamma}_L - \gamma| > \varepsilon) = 0$$

$$\hat{\gamma}_L \to \gamma \quad \text{in probability as } L \to \infty$$

  - For $L \gg 1$, $\hat{\gamma}_L \approx \gamma$
  - Take $\hat{\gamma}_L$ as estimate of γ

- Properties of estimators:

1- Unbiased: $E[\hat{\gamma}_L] = \gamma \quad \forall L$

2- Consistent: $\hat{\gamma}_L \to \gamma$ as $L \to \infty$ — LLN

3- Efficient: $\hat{\gamma}_L$ close to γ (for some L)
Small variance " " " — Central limit theorem CLT

- Rem: $\hat{\mu}_L = \frac{1}{L} \sum_{i=1}^{L} X_i \Rightarrow \hat{\mu}_L = \frac{X_L}{L} + \frac{L-1}{L} \hat{\mu}_{L-1}$

$$= a_L X_L + (1 - a_L) \hat{\mu}_{L-1}$$

Stochastic recursion — $a_L \to 0$
" approximation

1.12. Statistical errors

· Estimator: $\hat{\mu}_L = \frac{1}{L} \sum_{i=1}^{L} X_i$

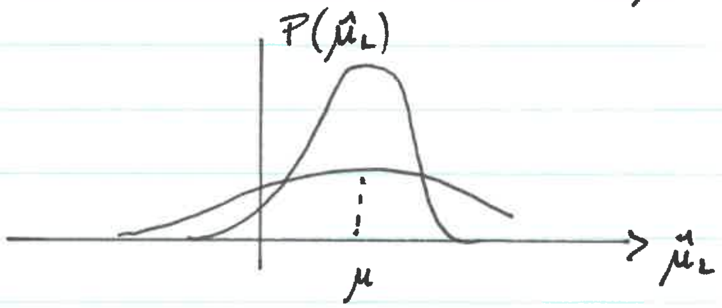
· Expectation: $E[\hat{\mu}_L] = \frac{1}{L} E[\sum_{i=1}^{L} X_i] = E[X] = \mu$ unbiased

· Variance: $\mathrm{var}(\hat{\mu}_L) = \mathrm{var}(\frac{1}{L} \sum_{i=1}^{L} X_i)$

$$= \frac{1}{L^2} \sum_{i=1}^{L} \mathrm{var}(X_i)$$

$$= \frac{\mathrm{var}(X)}{L} \sim \frac{1}{L}$$ consistent

· Standard deviation: $\mathrm{std}(\hat{\mu}_L) = \sigma(\hat{\mu}_L) = \sqrt{\mathrm{var}(\hat{\mu}_L)} \sim \frac{1}{\sqrt{L}}$



$P(|\hat{\mu}_L - \mu| > \varepsilon) \to 0$ LLN

$P(\hat{\mu}_L) \approx$ Gaussian for $L \gg 1$ near μ

$P(\hat{\mu}_L \in [\mu - \sigma_L, \mu + \sigma_L]) \approx 0.68$

· Error bars: $\underbrace{\hat{\mu}_L}_{\text{estimator}} \pm \underbrace{\sigma_L}_{\text{error bar}}$ confidence interval at 68%

$\hat{\mu}_L \pm 2\sigma_L$ " " " 95%

· Estimator of σ_L:

$$\sigma_L = \frac{\sqrt{\mathrm{var}(X)}}{\sqrt{L}} \Rightarrow \hat{\sigma}_L = \frac{\hat{\sigma}_X}{\sqrt{L}}$$

$$\sigma_X^2 = \underbrace{\frac{1}{L-1}}_{\approx \frac{1}{L}} \sum_{i=1}^{L} (X_i - \hat{\mu}_L)^2$$ ← in place estimator of μ

$$\Rightarrow \hat{\sigma}_L = \frac{1}{\sqrt{L}} \sqrt{\underbrace{\frac{1}{L} \sum_{i=1}^{L} X_i^2}_{\text{2nd moment estimator}} - \Big(\underbrace{\frac{1}{L} \sum_{i=1}^{L} X_i}_{\hat{\mu}_L}\Big)^2}$$

1.13. Direct vs MC integration

- Integral : $I = \int_{[0,1]^d} \varphi(x)\, dx$ in d-dim

- Direct numerical integration :
  - Uniform grid in d-dim cube
  - Spacing Δx
  - # pts $\sim \left(\frac{1}{\Delta x}\right)^d = n$
  - Convergence : $|I_n - I| \sim \frac{1}{n^s}$

- Example : 1-d
  - Riemann sum : $|I_n - I| \sim \frac{1}{n}$ $\quad \varphi \in C^1$
  - Trapezoidal method : $|I_n - I| \sim \frac{1}{n^2}$ $\quad \varphi \in C^2$
  - Simpson method : $|I_n - I| \sim \frac{1}{n^4}$ $\quad \varphi \in C^4$

    ⋮

- n calls of $\varphi \in C^s$, $\quad |I_n - I| \sim O(n^{-s/d})$

- MC integration : $I \approx \frac{1}{L} \sum_{i=1}^{L} \varphi(X_i)$ $\qquad X_i \sim \mathcal{U}[0,1]^d$

  $|I_n - I| \sim \frac{1}{\sqrt{n}}$ $\qquad$ independent of dim !