Robotics

Probabilities

Random variables, joint, conditional, marginal distribution, Bayes theorem, Probability distributions, Gauss, Dirac, Conjugate priors

Marc Toussaint
University of Stuttgart
Winter 2016/17

Lecturer: Peter Englert

Probability Theory

- Why do we need probabilities?

Probability Theory

- Why do we need probabilities?
  - Obvious: to express inherent (*objective*) stochasticity of the world

Probability Theory

- Why do we need probabilities?

  – Obvious: to express inherent (*objective*) stochasticity of the world

- But beyond this: (also in a “deterministic world”):
  – lack of knowledge!
  – hidden (latent) variables
  – expressing *uncertainty*
  – expressing *information* (and lack of information)
  – **Subjective Probability**

- Probability Theory: an information calculus

Outline

- Basic definitions
  - Random variables
  - joint, conditional, marginal distribution
  - Bayes' theorem
- Probability distributions:
  - Gauss
  - Dirac & Particles

Basic definitions

Probabilities & Sets

- **Sample Space/domain** Ω, e.g. $\Omega = \{1, 2, 3, 4, 5, 6\}$

- **Probability** $P: \; A \subset \Omega \mapsto [0, 1]$
  e.g., $P(\{1\}) = \frac{1}{6}, \quad P(\{4\}) = \frac{1}{6}, \quad P(\{2, 5\}) = \frac{1}{3},$

- Axioms: $\forall A, B \subseteq \Omega$
  – Nonnegativity $P(A) \geq 0$
  – Additivity $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$
  – Normalization $P(\Omega) = 1$

- Implications
  $0 \leq P(A) \leq 1$
  $P(\emptyset) = 0$
  $A \subseteq B \Rightarrow P(A) \leq P(B)$
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  $P(\Omega \setminus A) = 1 - P(A)$

Probabilities & Random Variables

- For a random variable X with discrete domain $\mathrm{dom}(X) = \Omega$ we write:
  $\forall_{x \in \Omega} : \; 0 \le P(X\!=\!x) \le 1$
  $\sum_{x \in \Omega} P(X\!=\!x) = 1$

  Example: A dice can take values $\Omega = \{1, .., 6\}$.
  X is the random variable of a dice throw.
  $P(X\!=\!1) \in [0, 1]$ is the probability that X takes value 1.

- A bit more formally: a random variable is a map from a measureable space to a domain (sample space) and thereby introduces a probability measure on the domain ("assigns a probability to each possible value")

Probabilty Distributions

- $P(X=1) \in \mathbb{R}$ denotes a specific probability
  $P(X)$ denotes the probability distribution (function over Ω)

Probabilty Distributions

- $P(X=1) \in \mathbb{R}$ denotes a specific probability
  $P(X)$ denotes the probability distribution (function over Ω)

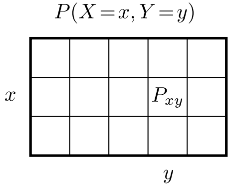
  Example: A dice can take values $\Omega = \{1, 2, 3, 4, 5, 6\}$.
  By $P(X)$ we discribe the full distribution over possible values $\{1, .., 6\}$. These are 6 numbers that sum to one, usually stored in a *table*, e.g.: $[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$

- In implementations we typically represent distributions over discrete random variables as tables (arrays) of numbers

- Notation for summing over a RV:
  In equations we often need to sum over RVs. We then write
  $$\textstyle\sum_X P(X) \;\cdots$$
  as shorthand for the explicit notation $\sum_{x \in \mathrm{dom}(X)} P(X=x) \;\cdots$

Joint distributions

Assume we have *two* random variables X and Y

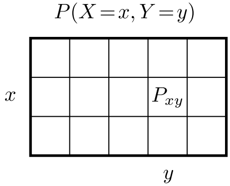


- Joint

  $P(X, Y)$

Joint distributions

Assume we have *two* random variables X and Y

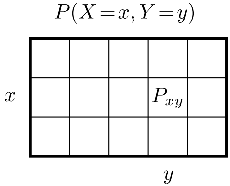


- Joint
  $P(X, Y)$

- Marginal (sum rule)
  $P(X) = \sum_Y P(X, Y)$

Joint distributions

Assume we have *two* random variables X and Y



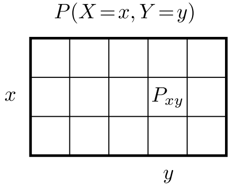
- Joint
  $P(X, Y)$

- Marginal (sum rule)
  $P(X) = \sum_Y P(X, Y)$

- Conditional:
  $P(X|Y) = \frac{P(X,Y)}{P(Y)}$
  The conditional is normalized: $\forall_Y : \sum_X P(X|Y) = 1$

Joint distributions

Assume we have *two* random variables X and Y



- Joint
  $P(X,Y)$

- Marginal (sum rule)
  $P(X) = \sum_Y P(X,Y)$

- Conditional:
  $P(X|Y) = \frac{P(X,Y)}{P(Y)}$
  The conditional is normalized: $\forall_Y : \ \sum_X P(X|Y) = 1$

- X is *independent* of Y iff: $P(X|Y) = P(X)$
  (table thinking: all columns of $P(X|Y)$ are equal)

Bayes' Theorem

- Implications of these definitions:

Product rule: $P(X, Y) = P(X)\,P(Y|X) = P(Y)\,P(X|Y)$

Bayes' Theorem

- Implications of these definitions:

Product rule: $P(X,Y) = P(X)\ P(Y|X) = P(Y)\ P(X|Y)$

Bayes' Theorem: $P(X|Y) = \frac{P(Y|X)\ P(X)}{P(Y)}$

Bayes' Theorem

- Implications of these definitions:

Product rule: $P(X,Y) = P(X)\ P(Y|X) = P(Y)\ P(X|Y)$

Bayes' Theorem: $P(X|Y) = \frac{P(Y|X)\ P(X)}{P(Y)}$

$$\text{posterior} \ = \ \frac{\text{likelihood} \ \cdot \ \text{prior}}{\text{normalization}}$$

Multiple RVs:

- Analogously for n random variables $X_{1:n}$ (stored as a rank n tensor)
  Joint: $P(X_{1:n})$
  Marginal: $P(X_1) = \sum_{X_{2:n}} P(X_{1:n})$,
  Conditional: $P(X_1 | X_{2:n}) = \frac{P(X_{1:n})}{P(X_{2:n})}$

- X is *conditionally independent* of Y given Z iff:
  $$P(X|Y,Z) = P(X|Z)$$

- Product rule and Bayes' Theorem:

$$P(X_{1:n}) = \prod_{i=1}^n P(X_i | X_{i+1:n})$$

$$P(X_1 | X_{2:n}) = \frac{P(X_2 | X_1, X_{3:n})\ P(X_1 | X_{3:n})}{P(X_2 | X_{3:n})}$$

$$P(X,Z,Y) = P(X|Y,Z)\ P(Y|Z)\ P(Z)$$

$$P(X|Y,Z) = \frac{P(Y|X,Z)\ P(X|Z)}{P(Y|Z)}$$

$$P(X,Y|Z) = \frac{P(X,Z|Y)\ P(Y)}{P(Z)}$$

Distributions over continuous domain

- Let x be a continuous RV. The **probability density function (pdf)** $p(x) \in [0, \infty)$ defines the probability

$$P(a \le x \le b) = \int_a^b p(x)\ dx \ \in [0, 1]$$

(In discrete domain: *probability distribution* and *probability mass function* $P(x) \in [0, 1]$ are used synonymously.)

Distributions over continuous domain

- Let x be a continuous RV. The **probability density function (pdf)** $p(x) \in [0, \infty)$ defines the probability

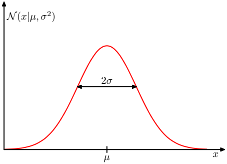
$$P(a \le x \le b) = \int_a^b p(x)\; dx \;\; \in [0, 1]$$

(In discrete domain: *probability distribution* and *probability mass function* $P(x) \in [0, 1]$ are used synonymously.)

- The **cumulative distribution function (cdf)** $F(y) = P(x \le y) = \int_{-\infty}^{y} p(x) dx \;\; \in [0, 1]$ is the cumulative integral with $\lim_{y \to \infty} F(y) = 1$

- Two basic examples:
Gaussian: $\mathcal{N}(x \,|\, \mu, \Sigma) = \frac{1}{|\, 2\pi\Sigma \,|^{\,1/2}}\; e^{-\frac{1}{2}(x-\mu)^\top\, \Sigma^{-1}\; (x-\mu)}$
Dirac or δ ("point particle") $\delta(x) = 0$ except at $x = 0$, $\int \delta(x)\; dx = 1$
$\delta(x) = \frac{\partial}{\partial x} H(x)$ where $H(x) = [x \ge 0] =$ Heavyside step function

Gaussian distribution



- 1-dim: $\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{|\, 2\pi\sigma^2 \,|^{1/2}}\ e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$
- n-dim Gaussian in *normal form*:

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{|\, 2\pi\Sigma \,|^{1/2}}\ \exp\{-\frac{1}{2}(x-\mu)^\top\, \Sigma^{\text{-}1}\ (x-\mu)\}$$

with **mean** μ and **covariance** matrix Σ. In *canonical form*:

$$\mathcal{N}[x \mid a, A] = \frac{\exp\{-\frac{1}{2}a^\top A^{\text{-}1} a\}}{|\, 2\pi A^{\text{-}1} \,|^{1/2}}\ \exp\{-\frac{1}{2}x^\top\, A\, x + x^\top a\} \tag{1}$$

with **precision** matrix $A = \Sigma^{\text{-}1}$ and coefficient $a = \Sigma^{\text{-}1}\mu$ (and mean $\mu = A^{\text{-}1}a$).

Gaussian identities

Symmetry: $\mathcal{N}(x \,|\, a, A) = \mathcal{N}(a \,|\, x, A) = \mathcal{N}(x - a \,|\, 0, A)$

Product:

$$\mathcal{N}(x \,|\, a, A)\; \mathcal{N}(x \,|\, b, B) = \mathcal{N}[x \,|\, A^{\text{-}1}a + B^{\text{-}1}b, A^{\text{-}1} + B^{\text{-}1}]\; \mathcal{N}(a \,|\, b, A + B)$$

$$\mathcal{N}[x \,|\, a, A]\; \mathcal{N}[x \,|\, b, B] = \mathcal{N}[x \,|\, a + b, A + B]\; \mathcal{N}(A^{\text{-}1}a \,|\, B^{\text{-}1}b, A^{\text{-}1} + B^{\text{-}1})$$

"Propagation":

$$\int_y \mathcal{N}(x \,|\, a + Fy, A)\; \mathcal{N}(y \,|\, b, B)\; dy = \mathcal{N}(x \,|\, a + Fb, A + FBF^\top)$$

Transformation:

$$\mathcal{N}(Fx + f \,|\, a, A) = \tfrac{1}{|\,F\,|}\; \mathcal{N}(x \,|\; F^{\text{-}1}(a - f),\; F^{\text{-}1}AF^{\text{-}\top})$$

Marginal & conditional:

$$\mathcal{N}\left(\begin{array}{c|c} x & a \\ y & b \end{array}, \begin{array}{cc} A & C \\ C^\top & B \end{array}\right) = \mathcal{N}(x \,|\, a, A) \cdot \mathcal{N}(y \,|\, b + C^\top A^{\text{-}1}(x \text{-} a),\; B - C^\top A^{\text{-}1}C)$$

More Gaussian identities: see

`http://ipvs.informatik.uni-stuttgart.de/mlr/marc/notes/gaussians.pdf`

Motivation for Gaussian distributions

- Gaussian Bandits
- Control theory, Stochastic Optimal Control
- State estimation, sensor processing, Gaussian filtering (Kalman filtering)
- Machine Learning
- rich vocabulary and easy to compute!
- etc.

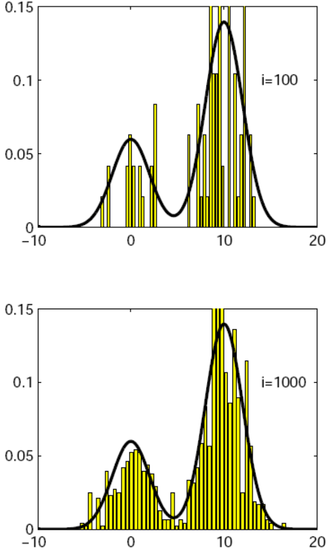
Dirac Delta / Point Particle

Dirac or δ ("point particle"): $\delta(x) = 0$ except at $x = 0$, $\int \delta(x)\, dx = 1$

$\delta(x) = \frac{\partial}{\partial x} H(x)$ where $H(x) = [x \geq 0]$ is the Heavyside step function

Particle Approximation of a Distribution

We approximate a distribution $p(x)$ over a continuous domain $\mathbb{R}^n$ with a histogram of particles:



Particle Approximation of a Distribution

We approximate a distribution $p(x)$ over a continuous domain $\mathbb{R}^n$

- A particle distribution $q(x)$ is a weighed set $\mathcal{S} = \{(x^i, w^i)\}_{i=1}^N$ of N particles
  - each particle has a "location" $x^i \in \mathbb{R}^n$ and a weight $w^i \in \mathbb{R}$
  - weights are normalized, $\sum_i w^i = 1$

$$q(x) := \sum_{i=1}^N w^i \; \delta(x - x^i)$$

  where $\delta(x - x^i)$ is the δ-distribution.

- Given weighted particles, we can estimate for any (smooth) f:

$$\langle f(x) \rangle_p = \int_x f(x) p(x) dx \;\approx\; \textstyle\sum_{i=1}^N w^i f(x^i)$$

  See *An Introduction to MCMC for Machine Learning*
  `www.cs.ubc.ca/~nando/papers/mlintro.pdf`

Motivation for particle distributions

- Numeric representation of “difficult” distributions
  - Very general and versatile
  - But often needs many samples
- Distributions over games (action sequences), sample based planning, MCTS
- State estimation, particle filters
- etc.

Conjugate priors

- Assume you have data $D = \{x_1, .., x_n\}$ with likelihood

$$P(D \mid \theta)$$

  that depends on an uncertain parameter θ
  Assume you have a prior $P(\theta)$

- The prior $P(\theta)$ is **conjugate** to the likelihood $P(D \mid \theta)$ iff the posterior

$$P(\theta \mid D) \propto P(D \mid \theta)\ P(\theta)$$

  is in the *same distribution class* as the prior $P(\theta)$

- Having a conjugate prior is very convenient, because then you know how to update the belief given data

Conjugate priors

likelihood	conjugate
Binomial $\mathrm{Bin}(D \mid \mu)$	Beta $\mathrm{Beta}(\mu \mid a, b)$
Multinomial $\mathrm{Mult}(D \mid \mu)$	Dirichlet $\mathrm{Dir}(\mu \mid \alpha)$
Gauss $\mathcal{N}(x \mid \mu, \Sigma)$	Gauss $\mathcal{N}(\mu \mid \mu_0, A)$
1D Gauss $\mathcal{N}(x \mid \mu, \lambda^{-1})$	Gamma $\mathrm{Gam}(\lambda \mid a, b)$
nD Gauss $\mathcal{N}(x \mid \mu, \Lambda^{-1})$	Wishart $\mathrm{Wish}(\Lambda \mid W, \nu)$
nD Gauss $\mathcal{N}(x \mid \mu, \Lambda^{-1})$	Gauss-Wishart $\mathcal{N}(\mu \mid \mu_0, (\beta\Lambda)^{-1})\ \mathrm{Wish}(\Lambda \mid W, \nu)$

Gaussian prior and posterior

- Assume we have data $D = \{x_1, .., x_n\}$, each $x_i \in \mathbb{R}^n$, with likelihood

$$P(D \,|\, \mu, \Sigma) = \textstyle\prod_i \mathcal{N}(x_i \,|\, \mu, \Sigma)$$

$$\operatorname*{argmax}_{\mu} \; P(D \,|\, \mu, \Sigma) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\operatorname*{argmax}_{\Sigma} \; P(D \,|\, \mu, \Sigma) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^\top$$

- Assume we are initially uncertain about μ (but know Σ). We can express this uncertainty using again a Gaussian $\mathcal{N}[\mu \,|\, a, A]$. Given data we have

$$P(\mu \,|\, D) \propto P(D \,|\, \mu, \Sigma) \; P(\mu) = \textstyle\prod_i \mathcal{N}(x_i \,|\, \mu, \Sigma) \; \mathcal{N}[\mu \,|\, a, A]$$
$$= \textstyle\prod_i \mathcal{N}[\mu \,|\, \Sigma^{-1} x_i, \Sigma^{-1}] \; \mathcal{N}[\mu \,|\, a, A] \propto \mathcal{N}[\mu \,|\, \Sigma^{-1} \sum_i x_i, \; n\Sigma^{-1} + A]$$

Note: in the limit $A \to 0$ (uninformative prior) this becomes

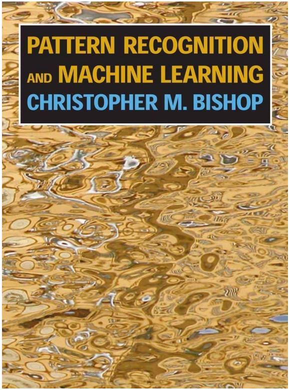
$$P(\mu \,|\, D) = \mathcal{N}(\mu \,|\, \frac{1}{n} \sum_i x_i, \frac{1}{n}\Sigma)$$

which is consistent with the Maximum Likelihood estimator

Some more continuous distributions*

Gaussian	$\mathcal{N}(x \mid a, A) = \frac{1}{\mid 2\pi A \mid^{1/2}}\ e^{-\frac{1}{2}(x-a)^\top\ A^{-1}\ (x-a)}$
Dirac or δ	$\delta(x) = \frac{\partial}{\partial x} H(x)$
Student's t (=Gaussian for $\nu \to \infty$, otherwise heavy tails)	$p(x;\nu) \propto [1 + \frac{x^2}{\nu}]^{-\frac{\nu+1}{2}}$
Exponential (distribution over single event time)	$p(x;\lambda) = [x \ge 0]\ \lambda e^{-\lambda x}$
Laplace ("double exponential")	$p(x;\mu,b) = \frac{1}{2b} e^{-\mid x-\mu \mid /b}$
Chi-squared	$p(x;k) \propto [x \ge 0]\ x^{k/2-1} e^{-x/2}$
Gamma	$p(x;k,\theta) \propto [x \ge 0]\ x^{k-1} e^{-x/\theta}$

Probability distributions



Bishop, C. M.: *Pattern Recognition and Machine Learning*.
Springer, 2006
`http://research.microsoft.com/en-us/um/people/cmbishop/prml/`