A Short Course on Graphical Models

1. Introduction to Probability Theory

Mark Paskin

mark@paskin.org

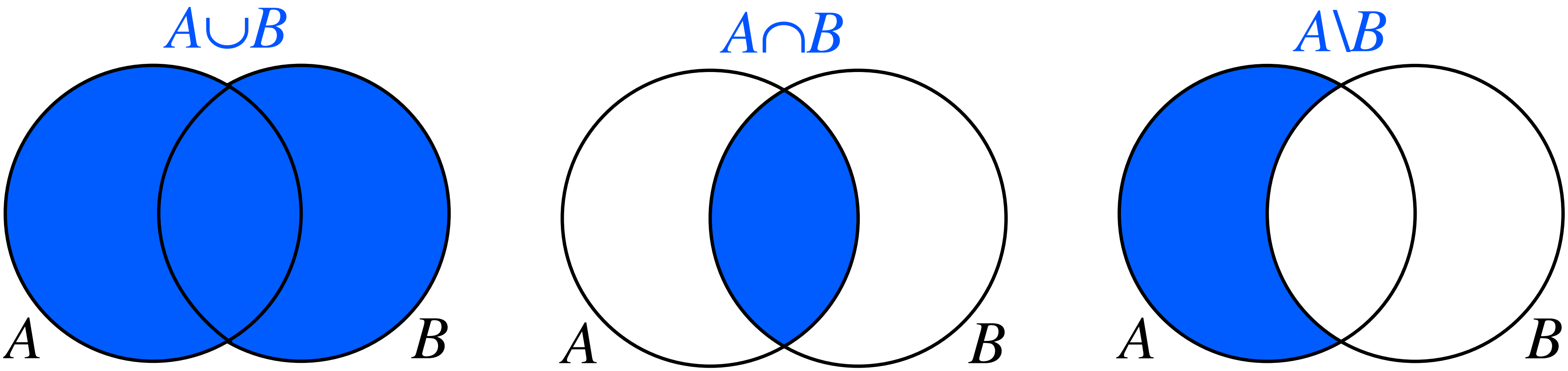
Reasoning under uncertainty

- In many settings, we must try to understand what is going on in a system when we have imperfect or incomplete information.
- Two reasons why we might reason under uncertainty:
  1. *laziness* (modeling every detail of a complex system is costly)
  2. *ignorance* (we may not completely understand the system)
- Example: deploy a network of smoke sensors to detect fires in a building. Our model will reflect both laziness and ignorance:
  – We are too *lazy* to model what, besides fire, can trigger the sensors;
  – We are too *ignorant* to model how fire creates smoke, what density of smoke is required to trigger the sensors, etc.

Using Probability Theory to reason under uncertainty

- Probabilities quantify uncertainty regarding the occurrence of events.
- Are there alternatives? *Yes, e.g., Dempster-Shafer Theory, disjunctive uncertainty, etc. (Fuzzy Logic is about imprecision, not uncertainty.)*
- Why is Probability Theory better? *de Finetti: Because if you do not reason according to Probability Theory, you can be made to act irrationally.*
- Probability Theory is key to the study of *action* and *communication*:
  - *Decision Theory* combines Probability Theory with Utility Theory.
  - *Information Theory* is "the logarithm of Probability Theory".
- Probability Theory gives rise to many interesting and important philosophical questions (which we will not cover).

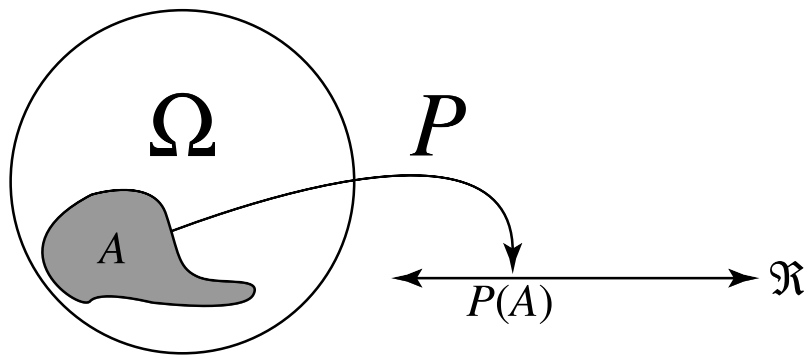
The only prerequisite: Set Theory



For simplicity, we will work (mostly) with finite sets. The extension to countably infinite sets is not difficult. The extension to uncountably infinite sets requires Measure Theory.

Probability spaces

- A *probability space* represents our uncertainty regarding an *experiment*.
- It has two parts:
  1. the *sample space* Ω, which is a set of *outcomes*; and
  2. the *probability measure* P, which is a real function of the subsets of Ω.



- A set of outcomes $A \subseteq \Omega$ is called an *event*. $P(A)$ represents how likely it is that the experiment's *actual* outcome will be a member of A.

An example probability space

- If our experiment is to deploy a smoke detector and see if it works, then there could be four outcomes:

  $$\Omega = \{(\mathit{fire}, \mathit{smoke}), (\mathit{no\ fire}, \mathit{smoke}), (\mathit{fire}, \mathit{no\ smoke}), (\mathit{no\ fire}, \mathit{no\ smoke})\}$$

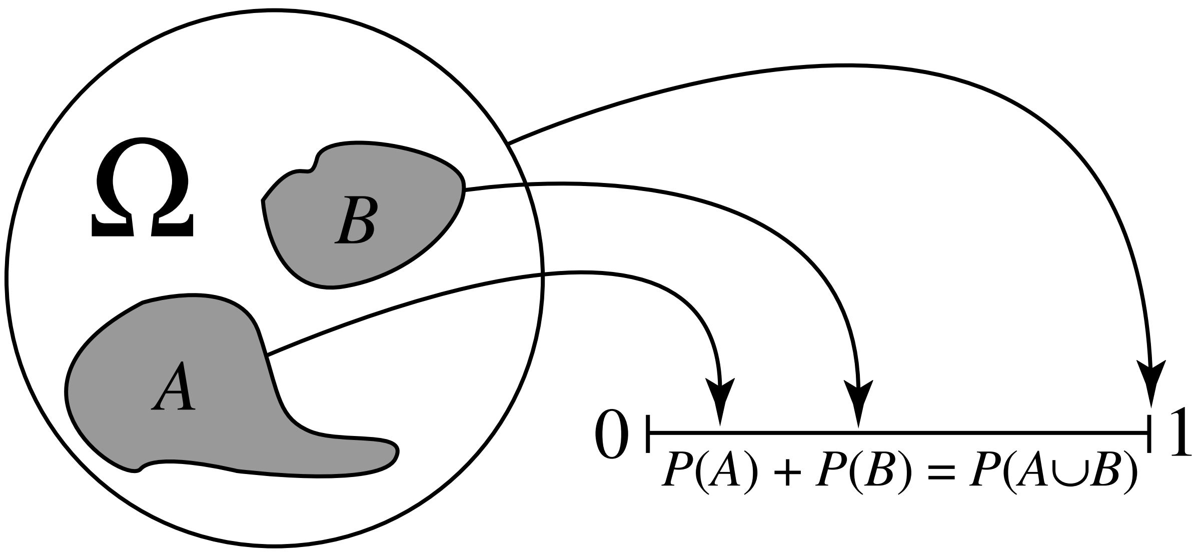
  Note that these outcomes are *mutually exclusive.*

- And we may choose:
  - $P(\{(\mathit{fire}, \mathit{smoke}), (\mathit{no\ fire}, \mathit{smoke})\}) = 0.005$
  - $P(\{(\mathit{fire}, \mathit{smoke}), (\mathit{fire}, \mathit{no\ smoke})\}) = 0.003$
  - $\ldots$

- Our choice of P has to obey three simple rules...

The three axioms of Probability Theory

1. $P(A) \geq 0$ for all events A
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ for disjoint events A and B



Some simple consequences of the axioms

- $P(A) = 1 - P(\Omega \backslash A)$
- $P(\emptyset) = 0$
- If $A \subseteq B$ then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) \leq P(A) + P(B)$
- $\ldots$

Example

- One easy way to define our probability measure P is to assign a probability to each outcome $\omega \in \Omega$:

	fire	*no fire*
smoke	0.002	0.003
no smoke	0.001	0.994

These probabilities must be non-negative and they must sum to one.

- Then the probabilities of all other events are determined by the axioms:

$$
\begin{aligned}
&P(\{(\textit{fire}, \textit{smoke}), (\textit{no fire}, \textit{smoke})\}) \\
&= P(\{(\textit{fire}, \textit{smoke})\}) + P(\{(\textit{no fire}, \textit{smoke})\}) \\
&= 0.002 + 0.003 \\
&= 0.005
\end{aligned}
$$

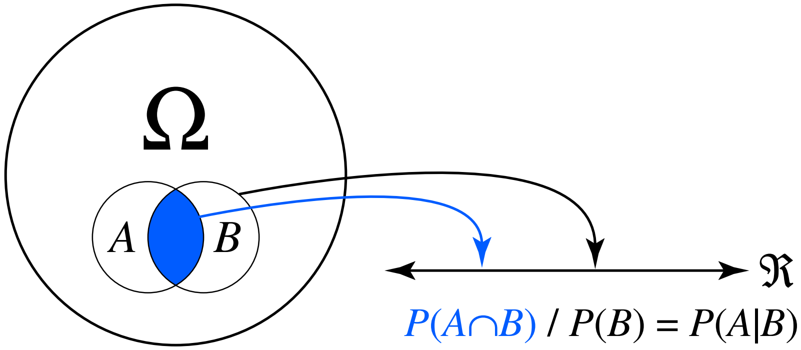
Conditional probability

- Conditional probability allows us to reason with *partial information.*
- When $P(B) > 0$, the *conditional probability of A given B* is defined as

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

This is the probability that A occurs, given we have *observed* B, i.e., that we know the experiment's actual outcome will be in B. It is the fraction of probability mass in B that also belongs to A.

- $P(A)$ is called the *a priori (or prior) probability* of A and $P(A \mid B)$ is called the *a posteriori probability* of A given B.



Example of conditional probability

If P is defined by

	fire	*no fire*
smoke	0.002	0.003
no smoke	0.001	0.994

then

$$\begin{aligned}
&P(\{(\mathit{fire}, \mathit{smoke})\} \mid \{(\mathit{fire}, \mathit{smoke}), (\mathit{no\ fire}, \mathit{smoke})\}) \\
&= \frac{P(\{(\mathit{fire}, \mathit{smoke})\} \cap \{(\mathit{fire}, \mathit{smoke}), (\mathit{no\ fire}, \mathit{smoke})\})}{P(\{(\mathit{fire}, \mathit{smoke}), (\mathit{no\ fire}, \mathit{smoke})\})} \\
&= \frac{P(\{(\mathit{fire}, \mathit{smoke})\})}{P(\{(\mathit{fire}, \mathit{smoke}), (\mathit{no\ fire}, \mathit{smoke})\})} \\
&= \frac{0.002}{0.005} = 0.4
\end{aligned}$$

The product rule

Start with the definition of conditional probability and multiply by $P(A)$:

$$P(A \cap B) = P(A)P(B \mid A)$$

The probability that A and B both happen is the probability that A happens times the probability that B happens, given A has occurred.

The chain rule

Apply the product rule repeatedly:

$$P\left(\cap_{i=1}^{k} A_i\right) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P\left(A_k \mid \cap_{i=1}^{k-1} A_i\right)$$

The chain rule will become important later when we discuss conditional independence in Bayesian networks.

Bayes' rule

Use the product rule both ways with $P(A \cap B)$ and divide by $P(B)$:

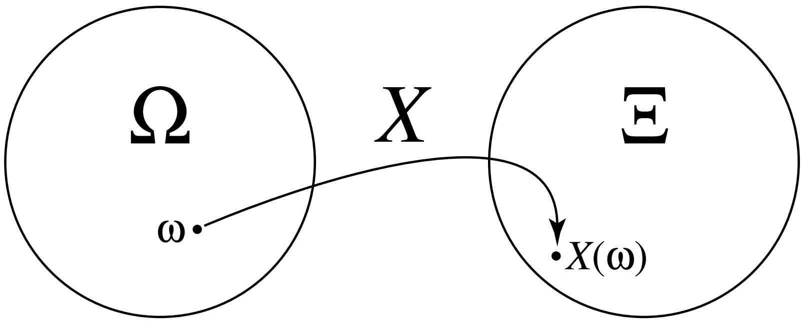
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Bayes' rule translates causal knowledge into diagnostic knowledge.

For example, if A is the event that a patient has a disease, and B is the event that she displays a symptom, then $P(B \mid A)$ describes a causal relationship, and $P(A \mid B)$ describes a diagnostic one (that is usually hard to assess). If $P(B \mid A)$, $P(A)$ and $P(B)$ can be assessed easily, then we get $P(A \mid B)$ for free.

Random variables

- It is often useful to "pick out" aspects of the experiment's outcomes.
- A *random variable* X is a function from the sample space Ω.



- Random variables can define events, e.g., $\{\omega \in \Omega : X(\omega) = \mathit{true}\}$.
- One will often see expressions like $P\{X = 1, Y = 2\}$ or $P(X = 1, Y = 2)$. These both mean $P(\{\omega \in \Omega : X(\omega) = 1, Y(\omega) = 2\})$.

Examples of random variables

Let's say our experiment is to draw a card from a deck:

$$\Omega = \{A\heartsuit, 2\heartsuit, \ldots, K\heartsuit, A\diamondsuit, 2\diamondsuit, \ldots, K\diamondsuit, A\clubsuit, 2\clubsuit, \ldots, K\clubsuit, A\spadesuit, 2\spadesuit, \ldots, K\spadesuit\}$$

random variable	example event
$H(\omega) = \begin{cases} \textit{true} & \text{if } \omega \text{ is a } \heartsuit \\ \textit{false} & \text{otherwise} \end{cases}$	$H = \textit{true}$
$N(\omega) = \begin{cases} n & \text{if } \omega \text{ is the number } n \\ 0 & \text{otherwise} \end{cases}$	$2 < N < 6$
$F(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is a face card} \\ 0 & \text{otherwise} \end{cases}$	$F = 1$

Densities

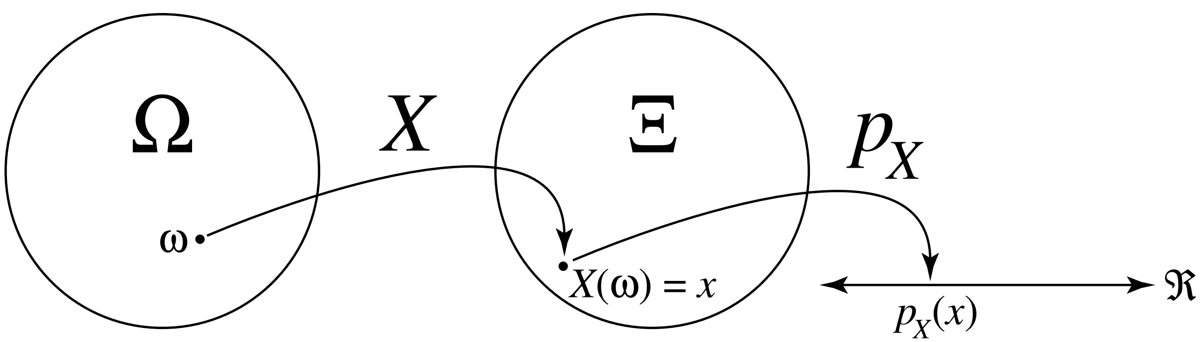
- Let $X : \Omega \to \Xi$ be a finite random variable. The function $p_X : \Xi \to \Re$ is the *density of* X if for all $x \in \Xi$:

$$p_X(x) = P(\{\omega : X(\omega) = x\})$$

- When Ξ is infinite, $p_X : \Xi \to \Re$ is the *density of* X if for all $\xi \subseteq \Xi$:

$$P(\{\omega : X(\omega) \in \xi\}) = \int_\xi p_X(x)\ \mathrm{d}x$$

- Note that $\int_\Xi p_X(x)\ \mathrm{d}x = 1$ for a valid density.



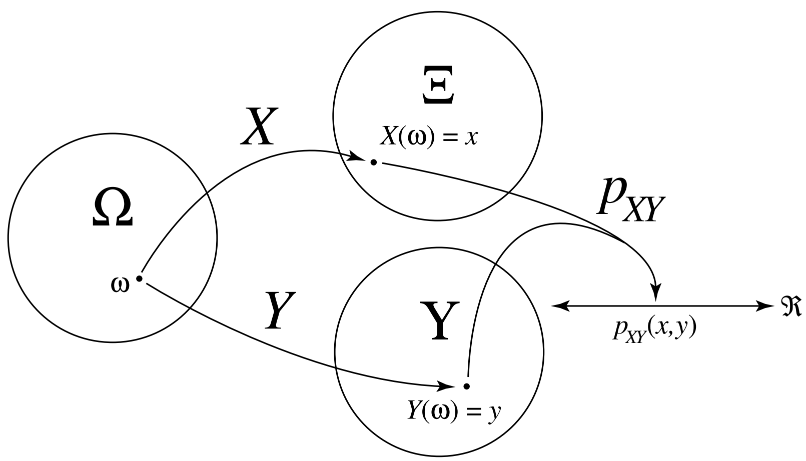
Joint densities

- If $X : \Omega \to \Xi$ and $Y : \Omega \to \Upsilon$ are two finite random variables, then $p_{XY} : \Xi \times \Upsilon \to \Re$ is their *joint density* if for all $x \in \Xi$ and $y \in \Upsilon$:

$$p_{XY}(x, y) = P(\{\omega : X(\omega) = x, Y(\omega) = y\})$$

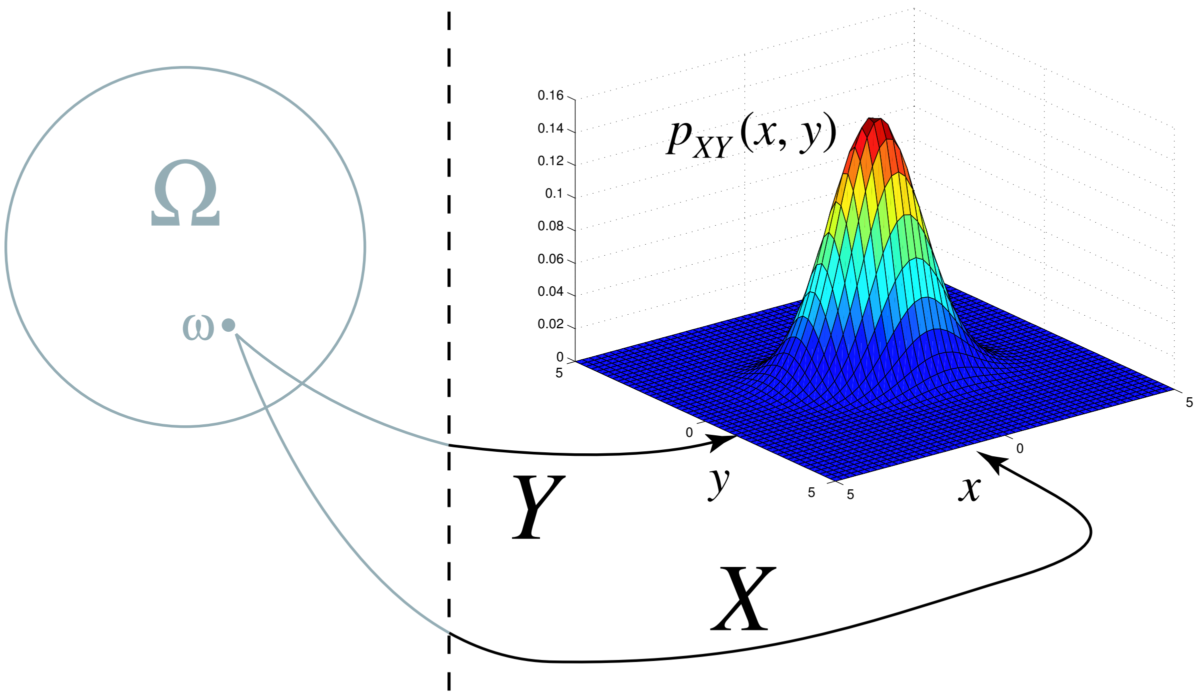
- When Ξ or Υ are infinite, $p_{XY} : \Xi \times \Upsilon \to \Re$ is the joint density of X and Y if for all $\xi \subseteq \Xi$ and $\upsilon \subseteq \Upsilon$:

$$\int_{\xi} \int_{\upsilon} p_{XY}(x, y) \, \mathrm{d}y \, \mathrm{d}x = P(\{\omega : X(\omega) \in \xi, Y(\omega) \in \upsilon\})$$



Random variables and densities are a layer of abstraction

We usually work with a set of random variables and a joint density; the probability space is implicit.



Marginal densities

- Given the joint density $p_{XY}(x, y)$ for $X : \Omega \to \Xi$ and $Y : \Omega \to \Upsilon$, we can compute the *marginal density* of X by

$$p_X(x) = \sum_{y \in \Upsilon} p_{XY}(x, y)$$

when Υ is finite, or by

$$p_X(x) = \int_{\Upsilon} p_{XY}(x, y) \, \mathrm{d}y$$

when Υ is infinite.

- This process of summing over the unwanted variables is called *marginalization*.

Conditional densities

- $p_{X|Y}(x, y) : \Xi \times \Upsilon \to \Re$ is the *conditional density of* X *given* $Y = y$ if

$$p_{X|Y}(x, y) = P(\{\omega : X(\omega) = x\} \,|\, \{\omega : Y(\omega) = y\})$$

  for all $x \in \Xi$ if Ξ is finite, or if

$$\int_{\xi} p_{X|Y}(x, y) \ \mathrm{d}x = P(\{\omega : X(\omega) \in \xi\} \,|\, \{\omega : Y(\omega) = y\})$$

  for all $\xi \subseteq \Xi$ if Ξ is infinite.

- Given the joint density $p_{XY}(x, y)$, we can compute $p_{X|Y}$ as follows:

$$p_{X|Y}(x, y) = \frac{p_{XY}(x, y)}{\sum_{x' \in \Xi} p_{XY}(x', y)} \qquad \text{or} \qquad p_{X|Y}(x, y) = \frac{p_{XY}(x, y)}{\int_{\Xi} p_{XY}(x', y) \ \mathrm{d}x'}$$

Rules in density form

- Product rule:

$$p_{XY}(x, y) = p_X(x) \times p_{Y|X}(y, x)$$

- Chain rule:

$$\begin{aligned} & p_{X_1 \cdots X_k}(x_1, \ldots, x_k) \\ & = p_{X_1}(x_1) \times p_{X_2|X_1}(x_2, x_1) \times \cdots \times p_{X_k|X_1 \cdots X_{k-1}}(x_k, x_1, \ldots, x_{k-1}) \end{aligned}$$

- Bayes' rule:

$$p_{Y|X}(y, x) = \frac{p_{X|Y}(x, y) \times p_Y(y)}{p_X(x)}$$

Inference

- The central problem of computational Probability Theory is the *inference problem*:

  Given a set of random variables $X_1, \ldots, X_k$ and their joint density, compute one or more conditional densities given observations.

- Many problems can be formulated in these terms. Examples:
  - In our example, the probability that there is a fire given smoke has been detected is $p_{F|S}(true, true)$.
  - We can compute the *expected position* of a target we are tracking given some measurements we have made of it, or the *variance* of the position, which are the parameters of a Gaussian posterior.

- Inference requires manipulating densities; how will we represent them?

Table densities

- The density of a set of finite-valued random variables can be represented as a table of real numbers.
- In our fire alarm example, the density of S is given by

$$p_S(s) = \begin{cases} 0.995 & s = \mathit{false} \\ 0.005 & s = \mathit{true} \end{cases}$$

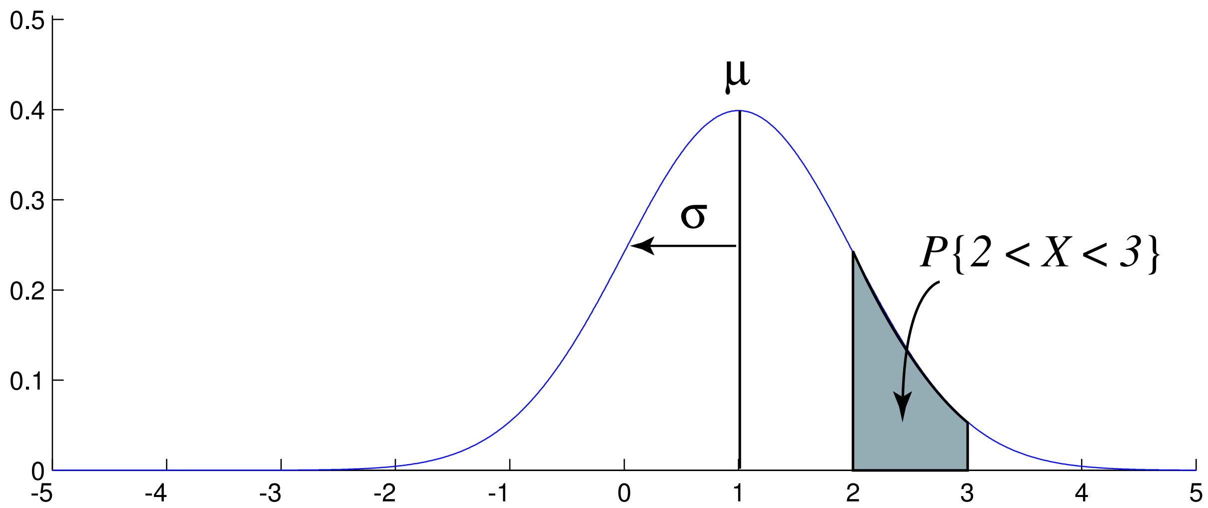
- If F is the Boolean random variable indicating a fire, then the joint density p_{SF} is represented by

$p_{SF}(s,f)$	$f = \mathit{true}$	$f = \mathit{false}$
$s = \mathit{true}$	0.002	0.003
$s = \mathit{false}$	0.001	0.994

- Note that the size of the table is *exponential* in the number of variables.

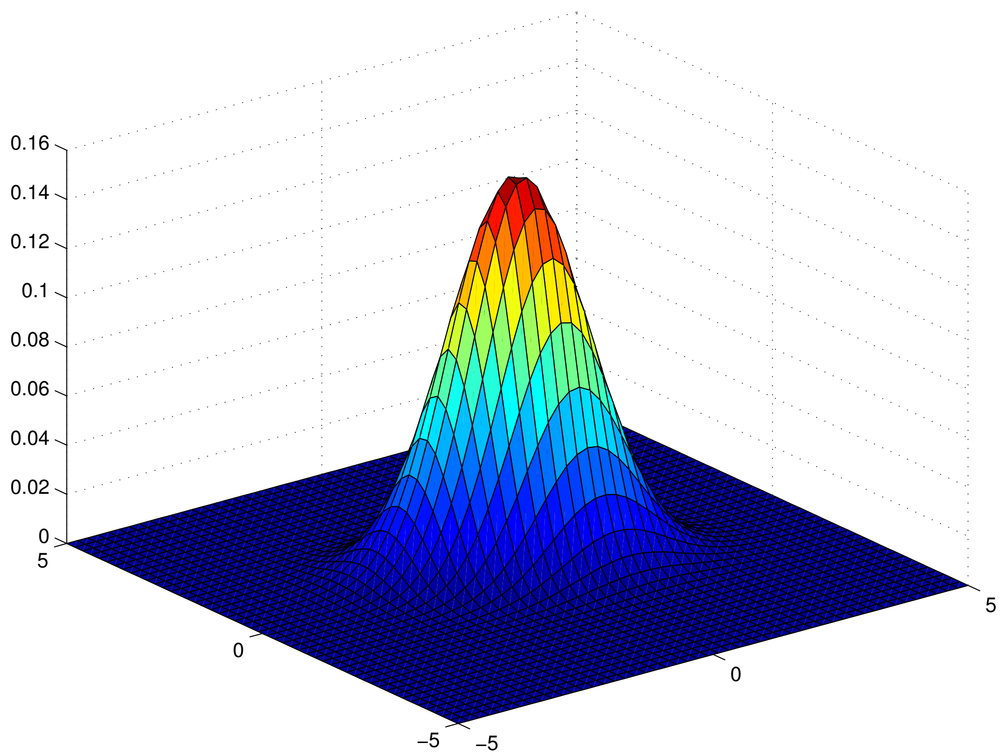
The Gaussian density

- One of the simplest densities for a real random variable.
- It can be represented by two real numbers: the mean μ and variance σ^2.



The multivariate Gaussian density

- A generalization of the Gaussian density to d real random variables.
- It can be represented by a $d \times 1$ *mean vector* μ and a symmetric $d \times d$ *covariance matrix* Σ.



Importance of the Gaussian

- The Gaussian density is the *only* density for real random variables that is "closed" under marginalization and multiplication.
- Also: a linear (or affine) function of a Gaussian random variable is Gaussian; and, a sum of Gaussian variables is Gaussian.
- For these reasons, the algorithms we will discuss will be tractable only for finite random variables or Gaussian random variables.
- When we encounter non-Gaussian variables or non-linear functions in practice, we will approximate them using our discrete and Gaussian tools. (This often works quite well.)

Looking ahead...

- Inference by enumeration: compute the conditional densities using the definitions. *In the tabular case, this requires summing over exponentially many table cells. In the Gaussian case, this requires inverting large matrices.*
- For large systems of finite random variables, representing the joint density is impossible, let alone inference by enumeration.
- Next time:
  - sparse representations of joint densities
  - Variable Elimination, our first efficient inference algorithm.

Summary

- A *probability space* describes our uncertainty regarding an *experiment*; it consists of a *sample space* of possible outcomes, and a *probability measure* that quantifies how likely each outcome is.
- An *event* is a set of *outcomes* of the experiment.
- A probability measure must obey three axioms: non-negativity, normalization, and additivity of disjoint events.
- *Conditional probability* allows us to reason with partial information.
- Three important rules follow easily from the definitions: the *product rule*, the *chain rule*, and *Bayes' rule*.

Summary (II)

- A *random variable* picks out some aspect of the experiment's outcome.
- A *density* describes how likely a random variable is to take on a value.
- We usually work with a set of random variables and their *joint density*; the probability space is implicit.
- The two types of densities suitable for computation are *table densities* (for finite-valued variables) and the *(multivariate) Gaussian* (for real-valued variables).
- Using a joint density, we can compute *marginal* and *conditional densities* over subsets of variables.
- *Inference* is the problem of computing one or more conditional densities given observations.