Lecture 2: Probability

AER1513: State Estimation

Timothy D. Barfoot

University of Toronto

Outline

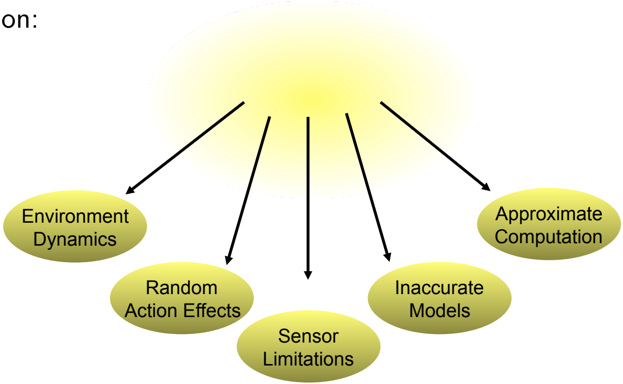
Lecture 2: Probability

- PDF Definitions
- Bayes Rule and Inference
- Gaussian PDFs
- Bayesian inference for Gaussians
- More Gaussian tools
- Quantifying Uncertainty
- Certainty Lower Bound

UNIVERSITY OF TORONTO

It's an uncertain world

– there are several sources of uncertainty we must overcome in estimation:



– we'll need some probabilistic machinery to acknowledge and manage uncertainty

UNIVERSITY OF TORONTO

Probability densities represent uncertainty in state

- we say that a random variable, x, is distributed according to a particular probability density function
- let $p(x)$ be a probability density function (PDF) for the random variable, x, over the interval $[a, b]$
- this is a non-negative function that satisfies the axiom of total probability:

$$\int_a^b p(x)\, dx = 1 \tag{1}$$

> We use PDFs to represent the likelihood of x being in all possible states in the interval, $[a, b]$.

UNIVERSITY OF TORONTO

Robot in a hallway

– this robot has a prior map of the hallway (i.e., knows where the doors are) and then detects a door

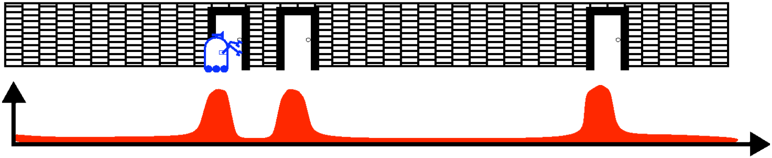
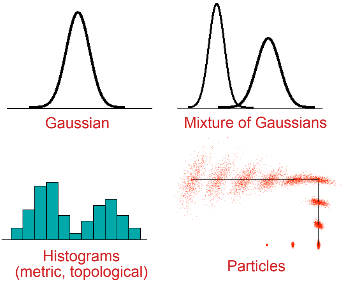


image: Thrun et al. (2006)

– it will have a PDF, $p(x)$, with three peaks representing that it is likely that it is near a door

Representations

- as we'll see, we can't represent PDFs perfectly in a computer so we need to choose an approximation



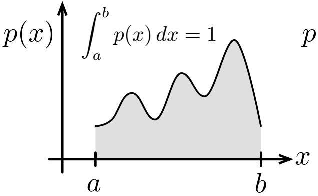
- each has its pros and cons, depending on the application

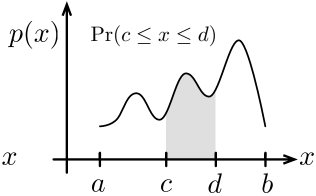
UNIVERSITY OF TORONTO

Probability is area under the curve

- note that probability density is not probability
- probability is given by the area under the density function
- for example, the probability that x lies between c and d, $\mathrm{Pr}(c \leq x \leq d)$, is given by

$$\mathrm{Pr}(c \leq x \leq d) = \int_c^d p(x)\, dx \tag{2}$$





Cumulative density and quantile functions

- the cumulative density function (CDF), $P(x)$, is given by

$$P(x) = \mathsf{Pr}(x' \leq x) = \int_{-\infty}^{x} p(x')\, dx' \tag{3}$$

  which is the area under the density function, $p(x)$, up to a particular x

- the quantile function, $Q(y)$, is the inverse (if it exists) of the CDF:

$$Q(y) = P^{-1}(y) \tag{4}$$

  where the range is between 0 and 1; we will use this later to sample from a distribution and also to perform statistical hypothesis tests

There are often strings attached

- we can introduce a conditioning variable to our PDFs
- let $p(x|y)$ be a PDF over $x \in \left[a, b\right]$ conditioned on $y \in \left[r, s\right]$ such that

$$(\forall y) \qquad \int_a^b p(x|y)\, dx = 1 \tag{5}$$

- this tells us about the likelihood of x given a particular value of y
- for example, x could be a robot position and y could be some sensor readings

UNIVERSITY OF TORONTO

The curse of dimensionality

- we may also denote joint probability densities for N-dimensional continuous variables in our framework as $p(\mathbf{x})$ where

$$\mathbf{x} = (x_1, \ldots, x_N) \tag{6}$$

  with $x_i \in \left[a_i,\, b_i\right]$

- note that we can also use the notation

$$p(x_1, x_2, \ldots, x_N) \tag{7}$$

  in place of $p(\mathbf{x})$

- sometimes we even mix and match the two and write

$$p(\mathbf{x}, \mathbf{y}) \tag{8}$$

  for the joint density of $\mathbf{x}$ and $\mathbf{y}$

Dimensions don't trump axioms

– in the N-dimensional case, the axiom of total probability requires

$$\int_{\mathbf{a}}^{\mathbf{b}} p(\mathbf{x})\, d\mathbf{x} = \int_{a_N}^{b_N} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} p\left(x_1, x_2, \ldots, x_N\right) dx_1\, dx_2 \cdots dx_N = 1 \quad (9)$$

where $\mathbf{a} = (a_1, \ldots, a_N)$ and $\mathbf{b} = (b_1, \ldots, b_N)$

UNIVERSITY OF TORONTO

Factoring a PDF

- we can always factor a joint probability density into a conditional and an unconditional factor:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \tag{10}$$

- in the specific case that $\mathbf{x}$ and $\mathbf{y}$ are statistically independent, we can factor the joint density as follows:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \tag{11}$$

  or put another way, $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y})$

Bayes' rule

- restating the factored expression:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \tag{12}$$

- Bayes' rule (Bayes, 1764) follows by rearranging:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \tag{13}$$

REV. T. BAYES

UNIVERSITY OF TORONTO

Bayesian inference

- we use Bayes' rule to infer one probability density function from another:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \tag{14}$$

- $p(\mathbf{x})$ is called the prior density (does not incorporate any data)
- $p(\mathbf{x}|\mathbf{y})$ is called the posterior density (incorporates data)
- $p(\mathbf{y}|\mathbf{x})$ is a generative model (e.g., sensor model)
- $p(\mathbf{y})$, the denominator, is discussed on the next slide

> Bayesian inference is the cornerstone of modern state estimation and machine learning.

UNIVERSITY OF TORONTO

On the margins

– we can compute the denominator in Bayes' rule using the idea of marginalization of a joint PDF,

$$p(\mathbf{x}, \mathbf{y}) \tag{15}$$

– if we integrate over all possible values of one of the joint variables, we can compute the density over only the other:

$$\int_{\mathbf{a}}^{\mathbf{b}} p(\mathbf{x}, \mathbf{y})\, d\mathbf{x} = \int_{\mathbf{a}}^{\mathbf{b}} p(\mathbf{x}|\mathbf{y})p(\mathbf{y})\, d\mathbf{x} = p(\mathbf{y}) \underbrace{\int_{\mathbf{a}}^{\mathbf{b}} p(\mathbf{x}|\mathbf{y})\, d\mathbf{x}}_{1} = p(\mathbf{y}) \tag{16}$$

– this is typically the most expensive step in Bayesian estimation

UNIVERSITY OF TORONTO

Just a moment

- when working with mass distributions (a.k.a., density functions) in dynamics, we often keep track of only a few properties called the moments of mass (e.g., mass, center of mass, inertia matrix)
- the same is true with probability density functions
- the zeroeth probability moment is always 1 owing to the axiom of total probability
- the first probability moment is known as the mean, $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = E\left[\mathbf{x}\right] = \int_{\mathbf{a}}^{\mathbf{b}} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x} \tag{17}$$

where $E[\cdot]$ denotes the expectation operator

UNIVERSITY OF TORONTO

To infinity and beyond!

- the second probability moment is known as the covariance matrix, $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right] = \int_{\mathbf{a}}^{\mathbf{b}} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\, p(\mathbf{x})\, d\mathbf{x} \tag{18}$$

- in $N = 1$ dimension, this is the familiar variance, σ^2, with σ the standard deviation
- the next two moments are called the skewness and kurtosis, but for the multivariate case, these get quite complicated and require tensor representations
- we will not need them here, but it should be mentioned that there are an infinite number of these probability moments

Sample mean and covariance

- we can draw samples (or realizations) from a PDF, which we denote as $\mathbf{x}_{\text{meas}} \leftarrow p(\mathbf{x})$; aside: how would we do this?
- to estimate the mean and covariance of random variable, $\mathbf{x}$, from N samples we use the sample mean and sample covariance:

$$\boldsymbol{\mu}_{\text{meas}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{i,\text{meas}} \tag{19a}$$

$$\boldsymbol{\Sigma}_{\text{meas}} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{x}_{i,\text{meas}} - \boldsymbol{\mu}_{\text{meas}}\right) \left(\mathbf{x}_{i,\text{meas}} - \boldsymbol{\mu}_{\text{meas}}\right)^T \tag{19b}$$

- the sample covariance uses $N-1$ rather than N in the denominator since it uses the sample mean, which is computed from the same measurements, resulting in a slight correlation

UNIVERSITY OF TORONTO

Statistically independent vs. uncorrelated

- two random variables, $\mathbf{x}$ and $\mathbf{y}$, are statistically independent if their joint density factors as follows:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})\, p(\mathbf{y}) \tag{20}$$

- two random variables are uncorrelated if

$$E\left[\mathbf{x}\mathbf{y}^T\right] = E\left[\mathbf{x}\right] E\left[\mathbf{y}\right]^T \tag{21}$$

- statistically independent always implies uncorrelated, but the reverse is not always true (but sometimes it is)

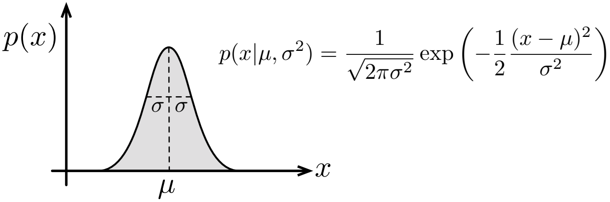
UNIVERSITY OF TORONTO

Gauss' namesake

- we'll be working with Gaussian probability density functions
- in one dimension, a Gaussian PDF is given by

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{22}$$

where μ is the mean and σ^2 is the variance (σ is called the standard deviation)



Multivariate Gaussian

- a multivariate Gaussian probability density function, $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, over the random variable, $\mathbf{x} \in \mathbb{R}^N$, may be expressed as

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \tag{23}$$

- $\boldsymbol{\mu} \in \mathbb{R}^N$ is the mean
- $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$ is the (symmetric positive-definite) covariance matrix

UNIVERSITY OF TORONTO

Gaussian moments

- for a Gaussian we must therefore have that

$$\boldsymbol{\mu} = E\left[\mathbf{x}\right] = \int_{-\infty}^{\infty} \mathbf{x} \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{\Sigma}}} \times \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) d\mathbf{x} \quad (24)$$

and

$$\begin{aligned} \boldsymbol{\Sigma} &= E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right] \\ &= \int_{-\infty}^{\infty} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{\Sigma}}} \times \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) d\mathbf{x} \end{aligned} \quad (25)$$

- check for yourself!

UNIVERSITY OF TORONTO

Gaussians are the new normal

- we may also write that $\mathbf{x}$ is normally (a.k.a., Gaussian) distributed using the following notation:

$$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \tag{26}$$

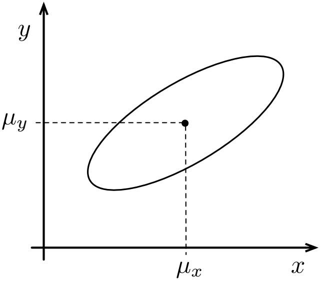
- we say a random variable is standard normally distributed if

$$\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{1}\right) \tag{27}$$

where $\mathbf{1}$ is an $N \times N$ identity matrix

UNIVERSITY OF TORONTO

Joint Gaussians



– we can also have a joint Gaussian over a pair of variables, $(\mathbf{x}, \mathbf{y})$, which we write as

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right) \tag{28}$$

with $\boldsymbol{\Sigma}_{yx} = \boldsymbol{\Sigma}_{xy}^T$

Statistically independent vs. uncorrelated

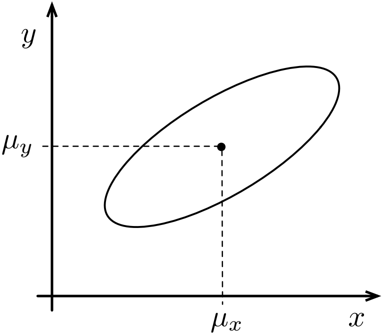
- for joint Gaussians, statistically independent implies uncorrelated and vice versa
- independence always implies uncorrelated (for any PDF)
- to go the other way, assume uncorrelated

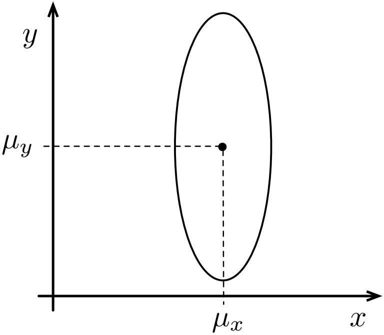
$$\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}^T = \mathbf{0} \tag{29}$$

which implies independence:

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right) \\ = \mathcal{N}\left(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}\right) = p(\mathbf{x}) \end{aligned} \tag{30}$$

Statistically independent vs. uncorrelated





correlated

$$E[(x-\mu_x)(y-\mu_y)] \neq 0$$

uncorrelated

$$E[(x-\mu_x)(y-\mu_y)] = 0$$

UNIVERSITY OF TORONTO

Gaussian inference

– recall that we can factor any joint PDF according to

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})\, p(\mathbf{y}) \tag{31}$$

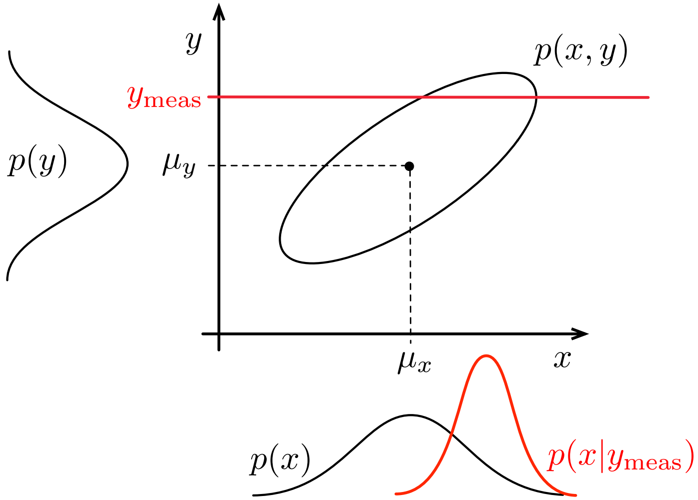
– in the case of a joint Gaussian, the factors are

$$\begin{aligned}
p(\mathbf{x}, \mathbf{y}) &= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix} \right) && \text{(32a)} \\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}\left(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} \right) && \text{(32b)} \\
p(\mathbf{y}) &= \mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{yy} \right) && \text{(32c)}
\end{aligned}$$

The expression for $p(\mathbf{x}|\mathbf{y})$ constitutes Bayesian inference for the case of Gaussian PDFs; we'll use this a lot!

UNIVERSITY OF TORONTO

Gaussian inference



How can you be Schur?

- to factor the joint Gaussian, we use the Schur complement:

$$\begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_{yy} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \times \begin{bmatrix} \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{yy} \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx} & \mathbf{1} \end{bmatrix} \tag{33}$$

- inverting this we have

$$\begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_{yy} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ -\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx} & \mathbf{1} \end{bmatrix} \times \begin{bmatrix} \left(\mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{yy}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1} & -\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \tag{34}$$

That's for Schur

– plugging this into the quadratic part of the Gaussian PDF we have

$$\log p(\mathbf{x}, \mathbf{y}) \propto \left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix} \right)^T \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}^{-1} \left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix} \right)$$
$$= \underbrace{\begin{matrix} \left(\mathbf{x} - \boldsymbol{\mu}_x - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)\right)^T \left(\boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right)^{-1} \\ \times \left(\mathbf{x} - \boldsymbol{\mu}_x - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)\right) \end{matrix}}_{\propto\ \log p(\mathbf{x}|\mathbf{y})}$$
$$+ \underbrace{\left(\mathbf{y} - \boldsymbol{\mu}_y\right)^T \boldsymbol{\Sigma}_{yy}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}_y\right)}_{\propto\ \log p(\mathbf{y})} \tag{35}$$

– recall that $\log(ab) = \log(a) + \log(b)$, for a and b scalar

Normalized product of Gaussians

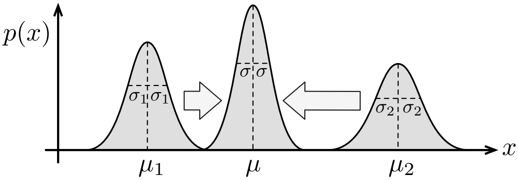
– the direct product of K Gaussian PDFs is also a Gaussian PDF:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{36}$$

where

$$\boldsymbol{\Sigma}^{-1} = \sum_{k=1}^{K} \boldsymbol{\Sigma}_k^{-1}, \qquad \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \sum_{k=1}^{K} \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k \tag{37}$$

– this gets used frequently in information fusion



$$\frac{1}{\sigma^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$$

$$\frac{\mu}{\sigma^2} = \frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}$$

UNIVERSITY OF TORONTO

Normalized product variation

- we also have that

$$\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$
$$\equiv \eta\prod_{k=1}^{K}\exp\left(-\frac{1}{2}(\mathbf{G}_k\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{G}_k\mathbf{x}-\boldsymbol{\mu}_k)\right), \quad (38)$$

where

$$\boldsymbol{\Sigma}^{-1}=\sum_{k=1}^{K}\mathbf{G}_k{}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{G}_k, \qquad \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}=\sum_{k=1}^{K}\mathbf{G}_k{}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k, \quad (39)$$

in the case that the matrices, $\mathbf{G}_k\in\mathbb{R}^{M_k\times N}$, are present, with $M_k\leq N$; η is a normalization constant

Gaussian transformation

- we now examine Gaussian transformation, namely computing

$$p(\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

  where we have that

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{g}(\mathbf{x}), \mathbf{R}\right) \\ p(\mathbf{x}) &= \mathcal{N}\left(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}\right) \end{aligned}$$

  and $\mathbf{g}(\cdot)$ is a nonlinear map: $\mathbf{g} : \mathbf{x} \mapsto \mathbf{y}$
- this is used, for example, in the denominator when carrying out full Bayesian inference using Gaussians
- the problem is that we can't compute this integral in general, so we need to approximate it

UNIVERSITY OF TORONTO

Transformation via linearization

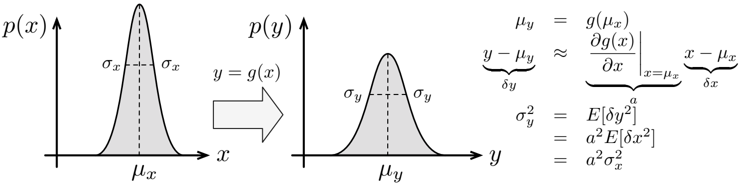
– we linearize the map such that

$$\mathbf{g}(\mathbf{x}) \approx \boldsymbol{\mu}_y + \mathbf{G}(\mathbf{x} - \boldsymbol{\mu}_x) \tag{40a}$$

$$\mathbf{G} = \left.\frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\boldsymbol{\mu}_x} \tag{40b}$$

$$\boldsymbol{\mu}_y = \mathbf{g}(\boldsymbol{\mu}_x) \tag{40c}$$

where $\mathbf{G}$ is the Jacobian of $\mathbf{g}$, with respect to $\mathbf{x}$



UNIVERSITY OF TORONTO

Transformation via linearization

- we have that

$$
\begin{aligned}
p(\mathbf{y}) &= \int_{-\infty}^{\infty} p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x} && (41)\\
&= \eta \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\mathbf{y}-(\boldsymbol{\mu}_y+\mathbf{G}(\mathbf{x}-\boldsymbol{\mu}_x))\right)^T\right. \\
&\qquad\qquad \times \quad \left.\mathbf{R}^{-1}\left(\mathbf{y}-(\boldsymbol{\mu}_y+\mathbf{G}(\mathbf{x}-\boldsymbol{\mu}_x))\right)\right) && (42)\\
&\qquad\qquad \times \quad \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x}-\boldsymbol{\mu}_x)\right)d\mathbf{x} && (43)\\
&= \rho\exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_y)^T\left(\mathbf{R}+\mathbf{G}\boldsymbol{\Sigma}_{xx}\mathbf{G}^T\right)^{-1}(\mathbf{y}-\boldsymbol{\mu}_y)\right) && (44)
\end{aligned}
$$

with ρ a normalization constant

- this is exactly Gaussian in $\mathbf{y}$:

$$
\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{yy}\right) = \mathcal{N}\left(\mathbf{g}(\boldsymbol{\mu}_x), \mathbf{R}+\mathbf{G}\boldsymbol{\Sigma}_{xx}\mathbf{G}^T\right) \tag{45}
$$

UNIVERSITY OF TORONTO

Sherman-Morrison-Woodbury identity

- the full derivation from the previous slide makes use of the Sherman-Morrison-Woodbury (SMW) identity, also sometimes called the matrix inversion lemma
- these identities are used frequently when manipulating expressions involving the covariance matrices associated with Gaussian PDFs:

$$
\begin{aligned}
(\mathbf{A}^{-1} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} &\equiv \mathbf{A} - \mathbf{A}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}\mathbf{B})^{-1}\mathbf{C}\mathbf{A} && \text{(46a)} \\
(\mathbf{D} + \mathbf{C}\mathbf{A}\mathbf{B})^{-1} &\equiv \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{C}(\mathbf{A}^{-1} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} && \text{(46b)} \\
\mathbf{A}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}\mathbf{B})^{-1} &\equiv (\mathbf{A}^{-1} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} && \text{(46c)} \\
(\mathbf{D} + \mathbf{C}\mathbf{A}\mathbf{B})^{-1}\mathbf{C}\mathbf{A} &\equiv \mathbf{D}^{-1}\mathbf{C}(\mathbf{A}^{-1} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} && \text{(46d)}
\end{aligned}
$$

- they were originally discovered when attempting to update the inverse of a big matrix with a few entries slightly changed

Quantifying uncertainty

- often in problems of estimation, we have estimated a PDF for some random variable and then want to quantify how certain we are in, for example, the mean of that PDF
- one method of doing this is to look at the negative entropy or Shannon information, H, which is given by

$$H(\mathbf{x}) = -E\left[\ln p(\mathbf{x})\right] = -\int_{\mathbf{a}}^{\mathbf{b}} p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x}$$

- we'll make this expression specific to Gaussian PDFs next

UNIVERSITY OF TORONTO

Gaussian information

- for a Gaussian PDF, we have for the Shannon information:

$$
\begin{aligned}
H(\mathbf{x}) &= -\int_{-\infty}^{\infty} p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x} && \text{(47a)} \\
&= -\int_{-\infty}^{\infty} p(\mathbf{x}) \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right. \\
&\qquad \left. - \ln \sqrt{(2\pi)^N \det \boldsymbol{\Sigma}} \right) d\mathbf{x} && \text{(47b)} \\
&= \frac{1}{2} \ln \left((2\pi)^N \det \boldsymbol{\Sigma} \right) + \int_{-\infty}^{\infty} \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\, p(\mathbf{x})\, d\mathbf{x} && \text{(47c)} \\
&= \frac{1}{2} \ln \left((2\pi)^N \det \boldsymbol{\Sigma} \right) + \frac{1}{2} E \left[(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] && \text{(47d)}
\end{aligned}
$$

- this term is exactly a squared Mahalonobis distance, which is like a squared Euclidean distance, but weighted in the middle by the inverse covariance matrix

UNIVERSITY OF TORONTO

Information manipulation

- a nice property of this quadratic function inside the expectation allows to rewrite it using the (linear) trace operator from linear algebra:

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right) \tag{48}$$

- since the expectation is also a linear operator, we may interchange the order of the expectation and trace arriving at:

$$\begin{aligned} E\left[(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] &= \mathrm{tr}\left(E\left[\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right]\right) \\ &= \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\underbrace{E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right]}_{\boldsymbol{\Sigma}}\right) = \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) = \mathrm{tr}\,\mathbf{1} = N \end{aligned}$$

which is just the dimension of the variable!

UNIVERSITY OF TORONTO

The bottom line

– finally, for our Shannon information expression we have

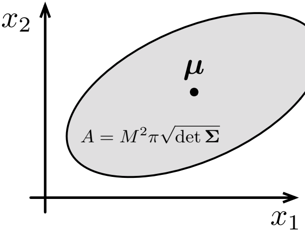
$$
\begin{aligned}
H(\mathbf{x}) &= \frac{1}{2}\ln\left((2\pi)^N \det \boldsymbol{\Sigma}\right) + \frac{1}{2}E\left[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] && \text{(49a)}\\
&= \frac{1}{2}\ln\left((2\pi)^N \det \boldsymbol{\Sigma}\right) + \frac{1}{2}N && \text{(49b)}\\
&= \frac{1}{2}\left(\ln\left((2\pi)^N \det \boldsymbol{\Sigma}\right) + N\ln e\right) && \text{(49c)}\\
&= \frac{1}{2}\ln\left((2\pi e)^N \det \boldsymbol{\Sigma}\right), && \text{(49d)}
\end{aligned}
$$

which is purely a function of $\boldsymbol{\Sigma}$, the covariance matrix of the Gaussian PDF

UNIVERSITY OF TORONTO

Uncertainty ellipsoid

- geometrically, we may interpret $\sqrt{\det \boldsymbol{\Sigma}}$ as the volume of the uncertainty ellipsoid formed by the Gaussian PDF



$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^2 e^{M^2} \det \boldsymbol{\Sigma}}}$$

$$H(\mathbf{x}) = \frac{1}{2} \ln\left((2\pi e)^2 \det \boldsymbol{\Sigma}\right)$$

- the geometric area inside the ellipse is

$$A = M^2\pi\sqrt{\det \boldsymbol{\Sigma}} \tag{50}$$

Equilikely contours

- note that along the boundary of the uncertainty ellipse, $p(\mathbf{x})$ is constant
- to see this, consider that the points along this ellipse must satisfy

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = M^2 \tag{51}$$

  where M is a factor applied to scale the nominal ($M = 1$) covariance
- in this case, we have that

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N e^{M^2} \det \boldsymbol{\Sigma}}} \tag{52}$$

  which is independent of $\mathbf{x}$ and thus constant

UNIVERSITY OF TORONTO

Drawing samples

- suppose we have a deterministic parameter, $\boldsymbol{\theta}$, that influences the outcome of a random variable, $\mathbf{x}$
- this can be captured by writing the PDF for $\mathbf{x}$ as depending on $\boldsymbol{\theta}$:

$$p(\mathbf{x}|\boldsymbol{\theta}) \tag{53}$$

- further suppose we now draw a sample, $\mathbf{x}_{\text{meas}}$, from $p(\mathbf{x}|\boldsymbol{\theta})$:

$$\mathbf{x}_{\text{meas}} \leftarrow p(\mathbf{x}|\boldsymbol{\theta}) \tag{54}$$

- the $\mathbf{x}_{\text{meas}}$ is sometimes called a realization of the random variable $\mathbf{x}$; we can think of it as a 'measurement'

UNIVERSITY OF TORONTO

As sure as sure can be

- the Cramér-Rao lower bound (CRLB) says that the covariance of any unbiased estimate, $\hat{\boldsymbol{\theta}}$ (based on the measurement, $\mathbf{x}_{\text{meas}}$), of the deterministic parameter, $\boldsymbol{\theta}$, is bounded by the Fisher information matrix, $\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}}$:

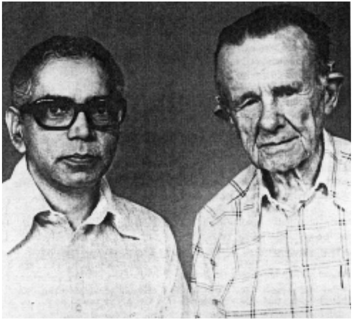
$$\text{cov}(\hat{\boldsymbol{\theta}}|\mathbf{x}_{\text{meas}}) = E\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right] \geq \boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}}^{-1}$$

  where 'unbiased' implies $E\left[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right] = \mathbf{0}$ and 'bounded' means $\text{cov}(\hat{\boldsymbol{\theta}}|\mathbf{x}_{\text{meas}}) - \boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}}^{-1}$ is positive-semi-definite
- the Fisher information matrix is given by

$$\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}} = E\left[\frac{\partial^2(-\ln p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^T\, \partial \boldsymbol{\theta}}\right]$$

UNIVERSITY OF TORONTO

What on Earth does that mean?



C. R. Rao, Harald Cramér, 1978

The Cramér-Rao lower bound says there is a fundamental limit on how certain we can be about an estimate of a parameter, given our measurements, regardless of the form of the estimator we use.

UNIVERSITY OF TORONTO

Our first estimation problem

- suppose that we have K samples (i.e., measurements), $\mathbf{x}_{\text{meas},k} \in \mathbb{R}^N$, drawn from a Gaussian PDF
- the K statistically independent random variables associated with these measurements are thus

$$(\forall k) \quad \mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{55}$$

- the term statistically independent implies that $E\left[(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_l - \bar{\mathbf{x}})^T\right] = \mathbf{0}$ for $k \neq l$
- now suppose our goal is to estimate the mean of this probability density function, $\boldsymbol{\mu}$, from the measurements, $\mathbf{x}_{\text{meas},1}, \ldots, \mathbf{x}_{\text{meas},K}$

UNIVERSITY OF TORONTO

Joint density

- for the joint density of all the random variables, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_K)$, we have

$$\ln p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}(\mathbf{x} - \mathbf{A}\boldsymbol{\mu})^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{A}\boldsymbol{\mu}) - \ln \sqrt{(2\pi)^{NK} \det \mathbf{B}} \tag{56}$$

where

$$\mathbf{A} = \underbrace{\begin{bmatrix} \mathbf{1} & \mathbf{1} & \cdots & \mathbf{1} \end{bmatrix}^T}_{K \text{ blocks}}, \quad \mathbf{B} = \text{diag}\underbrace{(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}, \ldots, \boldsymbol{\Sigma})}_{K \text{ blocks}} \tag{57}$$

- in this case, we have

$$\frac{\partial^2(-\ln p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}))}{\partial \boldsymbol{\mu}^T \partial \boldsymbol{\mu}} = \mathbf{A}^T \mathbf{B}^{-1} \mathbf{A} \tag{58}$$

UNIVERSITY OF TORONTO

Fisher information matrix

– the Fisher information matrix is

$$
\begin{aligned}
\boldsymbol{\mathcal{I}}_{\boldsymbol{\mu}} &= E\left[\frac{\partial^2(-\ln p(\mathbf{x}|\boldsymbol{\mu}))}{\partial\boldsymbol{\mu}^T\,\partial\boldsymbol{\mu}}\right] && \text{(59a)}\\
&= \mathbf{A}^T\mathbf{B}^{-1}\mathbf{A} && \text{(59b)}\\
&= K\boldsymbol{\Sigma}^{-1} && \text{(59c)}
\end{aligned}
$$

which we can see is just K times the inverse covariance of the Gaussian density

UNIVERSITY OF TORONTO

Cramér-Rao Lower Bound

- the CRLB then says

$$\text{cov}(\hat{\boldsymbol{\mu}}|\mathbf{x}_{\text{meas}}) \geq \frac{1}{K}\boldsymbol{\Sigma} \tag{60}$$

- in other words, the lower limit of the uncertainty in the estimate of the mean, $\hat{\boldsymbol{\mu}}$, becomes smaller and smaller the more measurements we have (as we would expect)
- note, in computing the CRLB we did not need to actually specify the form of the unbiased estimator at all; the CRLB is the lower bound for any unbiased estimator

UNIVERSITY OF TORONTO

Example unbiased estimator: mean...

- in this case, it is not hard to find an estimator that performs right at the CRLB:

$$\hat{\boldsymbol{\mu}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_{\text{meas},k} \tag{61}$$

- for the mean of this estimator we have

$$E\left[\hat{\boldsymbol{\mu}}\right] = E\left[\frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_k\right] = \frac{1}{K} \sum_{k=1}^{K} E[\mathbf{x}_k] = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\mu} = \boldsymbol{\mu} \tag{62}$$

which shows the estimator is indeed unbiased

UNIVERSITY OF TORONTO

...and the covariance

– for the covariance we have

$$\mathsf{cov}(\hat{\boldsymbol{\mu}}|\mathbf{x}_{\text{meas}}) \tag{63a}$$

$$= E\left[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\right] \tag{63b}$$

$$= E\left[\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k - \boldsymbol{\mu}\right)\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k - \boldsymbol{\mu}\right)^T\right] \tag{63c}$$

$$= \frac{1}{K^2}\sum_{k=1}^{K}\sum_{\ell=1}^{K}\underbrace{E\left[(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_\ell - \boldsymbol{\mu})^T\right]}_{\boldsymbol{\Sigma}\text{ when } k=\ell,\ \mathbf{0}\text{ otherwise}} \tag{63d}$$

$$= \frac{1}{K}\boldsymbol{\Sigma} \tag{63e}$$

which is right at the CRLB; we can do no better

UNIVERSITY OF TORONTO

References

Bayes, T., “Essay towards solving a problem in the doctrine of chances,” *Philosophical Transactions of the Royal Society of London*, 1764.

Thrun, S., Burgard, W., and Fox, D., *Probabilistic Robotics*, MIT Press, 2006.

UNIVERSITY OF TORONTO