Bayesian Inference and Data Assimilation

Prof. Dr.-Ing. Sebastian Reich

Universität Potsdam

12 April 2021

2 Introduction to Probability

There are at least three different methodologies of how to assign probabilities to events. These include:

(i) Conducting a large number of identical experiments and recording the *relative frequency* of an event A to occur. The probability of an event is then the large sample limit of this relative frequency.

(ii) Identifying equally likely alternatives and assigning equal probabilities to each of these alternatives.

(iii) Estimating probabilities based on perceived knowledge and previous experience of similar systems (this is necessarily subjective, and requires us to revise these estimates when new information is received).

Given a sample space Ω and an associated σ-algebra $\mathcal{F}$ (i.e. a set of subsets of Ω closed over complementation and countable unions) we can assign probabilities to events $E \in \mathcal{F}$ with the following properties:

Definition (Probability measure)

A probability measure is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ with the following properties:

(i) Total probability equals one: $\mathbb{P}(\Omega) = 1$, and

(ii) Probability is additive for disjoint events: If $A_1, A_2, \ldots, A_n, \ldots$ is a finite or countable collection of events $A_i \in \mathcal{F}$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i)$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space.

2.1 Random variables

Formally, when we consider random processes there will be an underlying abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ representing all sources of uncertainty. However, in practice we are usually only interested in the impact of these uncertainties on observables (such as heads or tails in the case of coin flipping). This means that most of the time we can focus on the concept of a random variable, which relates the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to induced uncertainties/probabilities in an observable (or measureable) quantity of interest.

Definition (Univariate random variable)

A function $X : \Omega \to \mathbb{R}$ is called a (univariate) random variable if the sets A_x, defined by

$$A_x = \{\omega \in \Omega : X(\omega) \leq x\},$$

are elements of the set of all events $\mathcal{F}$, *i.e.*, $A_x \in \mathcal{F}$ for all $x \in \mathbb{R}$. The (cumulative) probability distribution function of X is given by

$$F_X(x) = \mathbb{P}(A_x).$$

If X only takes finitely many values in $\mathbb{R}$, then we call X a discrete random variable, otherwise it is called a continuous random variable.

For a given random variable X on $\mathbb{R}$, we define an induced probability measure μ_X on $\mathbb{R}$ (not Ω) *via*

$$\mu_X(B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$$

for all intervals B in $\mathbb{R}$ (more precisely for all sets B in the Borel σ-algebra on $\mathbb{R}$).

Definition (Expectation)

A probability measure μ_X on $\mathcal{X}$ induces an associated (Lebesgue) integral over $\mathcal{X}$ and

$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)\mu_X(\mathrm{d}x)$$

is called the expectation of a function $f : \mathcal{X} \to \mathbb{R}$. Two important choices for f are $f(x) = x$, which leads to the mean $\bar{x} = \mathbb{E}[X]$ of X, and $f(x) = (x - \bar{x})^2$, which leads to the variance $\sigma^2 = \mathbb{E}[(X - \bar{x})^2]$ of X.

Remark (Lebesgue integral)

The integral is formally defined by first considering functions f which only take finitely many distinct values f_i in which case

$$\mathbb{E}[f(X)] = \sum_i f_i \, \mu_X(\{x \in \mathcal{X} : f(x) = f_i\}).$$

Such simple functions are then used to approximate general functions f and their expectation values.

Definition (Absolute continuity)

A probability measure μ_X on $\mathcal{X} = \mathbb{R}$ is called absolutely continuous (with respect to the standard Lebesgue integral $\mathrm{d}x$ on $\mathbb{R}$) if there exists a probability density function (PDF) $\pi_X : \mathcal{X} \to \mathbb{R}$ with $\pi_X(x) \geq 0$, and

$$\int_{\mathbb{R}} f(x)\mu_X(\mathrm{d}x) = \int_{\mathbb{R}} f(x)\pi_X(x)\mathrm{d}x.$$

The shorthand $\mu_X(\mathrm{d}x) = \pi_X(x)\mathrm{d}x$ is often adopted.

Remark

If a univariate random variable has PDF π_X, then the cummulative distribution function satisfies

$$F_X(x) = \int_{-\infty}^{x} \pi_X(x')\mathrm{d}x'$$

Remark (Multivariate random variables)

Univariate random variables naturally extend to the multivariate case, e.g., $\mathcal{X} = \mathbb{R}^{N_x}$, $N_x > 1$.

Consider, for example, a bivariate random variable $X : \Omega \to \mathbb{R}^2$ *with components* $X_1 : \Omega \to \mathbb{R}$ *and* $X_2 : \Omega \to \mathbb{R}$. *The sets*

$$A_x = \{\omega \in \Omega : X_1(\omega) \leq x_1,\ X_2(\omega) \leq x_2\},$$

$x = (x_1, x_2) \in \mathbb{R}^2$, *are assumed to be elements of the set of all events* $\mathcal{F}$. *The (cumulative) probability distribution function of* $X = (X_1, X_2)$ *is defined by*

$$F_X(x) = \mathbb{P}(A_x),$$

and the induced probability measure on $\mathbb{R}^2$ *is denoted by* μ_X. *If* μ_X *is absolutely continuous with respect to the Lebesgue measure on* $\mathbb{R}^2$, *then it has a PDF* $\pi_X(x)$, *which satisfies*

$$F_X(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \pi_X(x_1', x_2') \mathrm{d}x_1 \mathrm{d}x_2'.$$

Example (Gaussian random variables)

We use the notation $X \sim \mathrm{N}(\bar{x}, \sigma^2)$ to denote a univariate Gaussian random variable with mean $\bar{x}$ and variance σ^2. Its PDF is given by

$$\pi_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\bar{x})^2},$$

$x \in \mathbb{R}$. In the multivariate case, we use the notation $X \sim \mathrm{N}(\bar{x}, P)$ to denote a Gaussian random variable with PDF given by

$$\pi_X(x) = \frac{1}{(2\pi)^{N_x/2}|P|^{1/2}} \exp\left(-\frac{1}{2}(x-\bar{x})^T P^{-1}(x-\bar{x})\right),$$

for $x \in \mathbb{R}^{N_x}$. The covariance matrix is defined by

$$P = \mathbb{E}[(X-\bar{x})(X-\bar{x})^T] = \int_{\mathbb{R}^{N_x}} (x-\bar{x})(x-\bar{x})^T \pi_X(x)\mathrm{d}x,$$

and where we adopt the shorthand notation $|P| = |\det P|$ is the absolute value of the determinant of P. The PDF $\pi_X(x)$ of a Gaussian random variable $X \sim \mathrm{N}(\bar{x}, P)$ is denoted by $\mathrm{n}(x; \bar{x}, P)$.

Example (Laplace distribution and Gaussian mixtures)

The univariate Laplace distribution has PDF

$$\pi_X(x) = \frac{\lambda}{2} e^{-\lambda |x|},$$

$x \in \mathbb{R}$. This may be rewritten as

$$\pi_X(x) = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/(2\sigma)} \frac{\lambda^2}{2} e^{-\lambda^2 \sigma/2} \mathrm{d}\sigma.$$

Replacing the integral by a Riemann sum over a sequence of quadrature points $\{\sigma_j\}_{j=1}^J$, we obtain

$$\pi_X(x) \approx \sum_{j=1}^J \alpha_j \frac{1}{\sqrt{2\pi\sigma_j}} e^{-x^2/(2\sigma_j)}, \qquad \alpha_j \propto \frac{\lambda^2}{2} e^{-\lambda^2 \sigma_j/2} (\sigma_j - \sigma_{j-1}),$$

and the constant of proportionality is chosen such that the weights α_j sum to one. This finite sum approximation provides an example of a Gaussian mixture distribution, *i.e.* a weighted sum of Gaussians.

Definition (Point measure)

As a third example, we consider the point measure μ_{x_0} defined by

$$\int_{\mathcal{X}} f(x)\mu_{x_0}(\mathrm{d}x) = f(x_0).$$

The associated random variable X has the outcome $X(\omega) = x_0$ with probability 1. We call such a random variable *deterministic*, writing $X = x_0$ for short. Note that the point measure is not absolutely continuous with respect to the Lebesgue measure, *i.e.*, there is no corresponding PDF. Using the Dirac delta notation $\delta(\cdot)$, we shall nevertheless often formally write $\mu_{x_0}(\mathrm{d}x) = \delta(x - x_0)\mathrm{d}x$ or $\pi_X(x) = \delta(x - x_0)$. We find

$$\begin{aligned} F_{x_0}(x) &= \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) \\ &= \int_{-\infty}^{x} \mu_{x_0}(\mathrm{d}x) \\ &= \begin{cases} 1 & \text{if } x \geq x_0, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

for the cumulative distribution function of the point measure μ_{x_0}.

Remark (Best approximation of a random variable by a constant)

Let us assume that you wish to summarise the behaviour of a scalar random variable X with PDF π_X by a single number c^*. One way to achieve this is to minimise the expected distance between X and such constants c:

$$c^* := \arg\min L(c) \quad \textit{with} \quad L(c) := \frac{1}{2}\mathbb{E}[(X-c)^2].$$

Taking the derivative of $L(c)$ with respect to c yields

$$\frac{\mathrm{d}}{\mathrm{d}c}L(c) = -\mathbb{E}[X] + c$$

Hence c^* is obtained by setting this derivative to zero (critical point of $L(c)$), that is,

$$c^* = \mathbb{E}[X].$$

See Appendix 2.4 in the textbook and the Explain-Video for more details.

Lemma (Invertible transformations of random variables)

Let $X : \Omega \to \mathbb{R}^{N_x}$ *be a random variable with PDF* π_X*, and let* $Y : \Omega \to \mathbb{R}^{N_x}$ *be a random variable defined by* $Y = \Phi(X)$ *for smooth and invertible* $\Phi : \mathbb{R}^{N_x} \to \mathbb{R}^{N_x}$ *(i.e.,* Φ *is a diffeomorphism). Then,* Y *has PDF* π_Y *given by*

$$\pi_Y(y) = \pi_X(\Phi^{-1}(y))|J(y)|,$$

where Φ^{-1} *denotes the inverse of* Φ*,*

$$J(y) = D\Phi^{-1}(y) \in \mathbb{R}^{N_x \times N_x}$$

the Jacobian matrix of partial derivatives, and $|J|$ *the absolute value of the determinant of* J*, i.e.,* $|J| = |\det J|$*.*

Proof.

See Lemma 2.13 on pages 39 of the textbook. □

Definition (Marginals, independence, conditional probability distributions)

Let X_1 and X_2 denote two random variables on $\mathcal{X}$ with joint PDF $\pi_{X_1X_2}(x_1,x_2)$. The two PDFs

$$\pi_{X_1}(x_1) = \int_{\mathcal{X}} \pi_{X_1X_2}(x_1,x_2)\mathrm{d}x_2,$$

$$\pi_{X_2}(x_2) = \int_{\mathcal{X}} \pi_{X_1X_2}(x_1,x_2)\mathrm{d}x_1,$$

are called the marginal PDFs, *i.e.* $X_1 \sim \pi_{X_1}$ and $X_2 \sim \pi_{X_2}$. The two random variables are called independent if

$$\pi_{X_1X_2}(x_1,x_2) = \pi_{X_1}(x_1)\,\pi_{X_2}(x_2).$$

We also introduce the conditional PDFs

$$\pi_{X_1}(x_1|x_2) = \frac{\pi_{X_1X_2}(x_1,x_2)}{\pi_{X_2}(x_2)}$$

and

$$\pi_{X_2}(x_2|x_1) = \frac{\pi_{X_1X_2}(x_1,x_2)}{\pi_{X_1}(x_1)}.$$

Definition (Disintegration)

The two equivalent representations

$$\pi_{X_1X_2}(x_1, x_2) = \pi_{X_1}(x_1|x_2)\pi_{X_2}(x_2) = \pi_{X_2}(x_2|x_1)\pi_{X_1}(x_1),$$

are called disintegrations of the joint PDF $\pi_{X_1X_2}$. In the case of several random variables X_1, X_2, ..., X_n, this becomes, for example,

$$\pi_{X_1\cdots X_n}(x_1, \ldots, x_n) = \pi_{X_1}(x_1|x_2, \ldots, x_n)\pi_{X_2}(x_2|x_3, \ldots, x_n)\cdots\pi_{X_n}(x_n).$$

Lemma (Marginals as expectation of conditional PDFs)

Let X_1 and X_2 be two random variables with joint PDF $\pi_{X_1X_2}$ as above. Then

$$\pi_{X_1}(x_1) = \mathbb{E}\left[\pi_{X_1}(x_1|X_2)\right], \tag{1}$$

where the expectation is taken with respect to the random variable X_2.

Proof.

See Lemma 2.16 on page 41 of the textbook. □

Definition (Correlation)

The correlation between two univariate random variables X and Y is given by

$$\mathrm{corr}(X,Y) = \frac{\mathbb{E}[(X-\bar{x})(Y-\bar{y})]}{\sqrt{\mathbb{E}[(X-\bar{x})^2]\mathbb{E}[(Y-\bar{y})^2]}},$$

with $\bar{x} = \mathbb{E}[X]$ and $\bar{y} = \mathbb{E}[Y]$.

Remark

The normalisation factor is chosen so that $|\mathrm{corr}(X,Y)| \leq 1$ *and* $|\mathrm{corr}(X,Y)| \approx 1$ *indicates a high degree of correlation.*[a]

[a] However, it is important to note that high correlation between two variables does not indicate that there is a causal link. For example, the price of wheat in Angola may be highly correlated with the sales of laptops in Dakar, but this does not mean that more people in Dakar are buying laptops because the price of wheat in Angola has increased.

Example (Gaussian distribution)

A joint Gaussian distribution $\pi_{X_1X_2}(x_1,x_2)$, $x_1,x_2\in\mathbb{R}$, with mean $(\bar{x}_1,\bar{x}_2)$, covariance matrix

$$P=\left[\begin{array}{cc}\sigma_{11}^2 & \sigma_{12}^2\\ \sigma_{21}^2 & \sigma_{22}^2\end{array}\right],$$

and $\sigma_{12}=\sigma_{21}$ leads to a Gaussian conditional distribution

$$\pi_{X_1}(x_1|x_2)=\frac{1}{\sqrt{2\pi}\sigma_c}e^{-(x_1-\bar{x}_c)^2/(2\sigma_c^2)},$$

with conditional mean

$$\bar{x}_c=\bar{x}_1+\sigma_{12}^2\sigma_{22}^{-2}(x_2-\bar{x}_2)$$

and conditional variance

$$\sigma_c^2=\sigma_{11}^2-\sigma_{12}^2\sigma_{22}^{-2}\sigma_{21}^2.$$

See Example 2.18 on page 42 from the textbook.

Definition (independent and identically distributed)

Let X be a random variable with distribution π_X. A sequence of random variables $\{X_i\}_{i=1}^M$ with joint PDF $\pi_{X_1,\ldots,X_M}$ is called independent and identically distributed (i.i.d.) with distribution π_X if

1. the variables are mutually independent, *i.e.*,

$$\pi_{X_1,\ldots,X_M}(x_1,\ldots,x_M) = \pi_{X_1}(x_1)\pi_{X_2}(x_2)\cdots\pi_{X_M}(x_M),$$

   where π_{X_i} is the marginal distribution for X_i, $i = 1,\ldots,M$, and
2. the marginal distributions are all the same, *i.e.*, $\pi_{X_i}(x) = \pi_X(x)$.

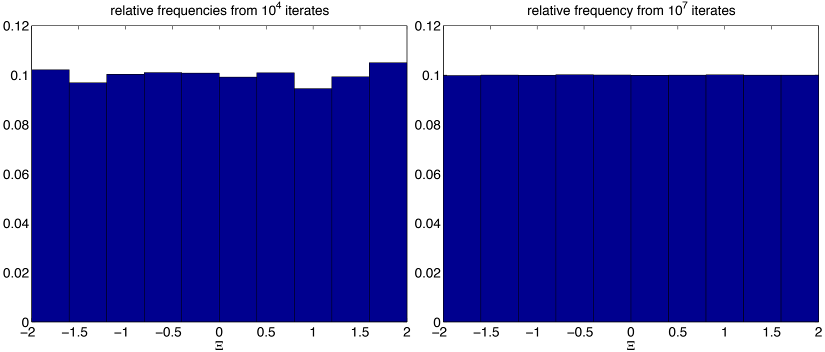
A particular realisation of an i.i.d. sequence of random variables of size M with distribution π_X, *i.e.*,

$$(x_1,\ldots,x_M) = (X_1(\omega),\ldots,X_M(\omega)),$$

is referred to as "M independent samples from the random variable X".

Example (deterministic processes as random variables)

We investigate the properties of the sequence $\{\Xi_i\}_{i=1}^{I}$ defined in Chapter 1 with $I = 10^4$ and $I = 10^7$, respectively. We display relative frequencies of the Ξ_i values in ten bins from the interval $[-a/2, a/2] = [-2, 2]$. Although the sequences $\{\Xi_i\}$ are generated in an entirely deterministic manner, these relative frequencies suggest that they behave like an identically distributed sequence of random variables with uniform distribution in $[-2, 2]$.



Example (continued)

An important aspect of the coin flipping experiment is that outcomes of successive flips are independent of each other. This clearly cannot be the case for the tent map process since successive Ξ_i values are, by definition, dependent. Since the dependence of a sequence of samples is difficult to assess numerically we instead compute the empirical (normalised) autocorrelations

$$C(\tau) = \frac{\sum_{i=1}^{I-\tau} \Xi_i \Xi_{i+\tau}}{\sum_{i=1}^{I-\tau} \Xi_i \Xi_i}, \qquad \tau \geq 0.$$

The variables Ξ_i and $\Xi_{i+\tau}$ are uncorrelated if $C(\tau) = 0$ for $\tau > 0$ and $I \to \infty$. For sample size $I = 10^7$, we find $C(0) = 1$ by definition and

$$C(1) = 0.1228 \times 10^{-3},\ C(2) = -0.1172 \times 10^{-3},\ C(3) = -0.1705 \times 10^{-3}.$$

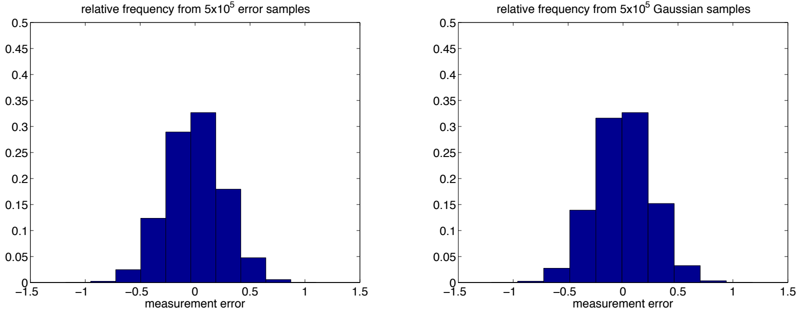
Hence we may conclude that, for all practical purposes, the samples Ξ_i are uncorrelated.

Example (continued)

We generate a total of 5×10^5 samples of the accumulated measurement error

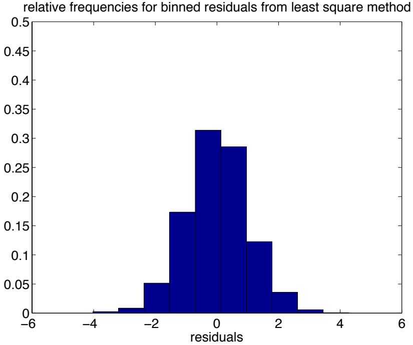
$$\delta = \sum_{i=1}^{20} \eta_i = \frac{1}{20} \sum_{i=1}^{20} \Xi_i .$$

and display their relative frequencies below. For comparison we also display the relative frequencies of an equal number of samples from a Gaussian distribution with mean zero and variance $\sigma^2 = 1/15$.



Example (deterministic processes as random variables)

We have considered the method of least squares in Chapter 1 in order to estimate the coefficients a_l in the autoregressive model of order five. We now use the data from Chapter 1 and analyse the resulting residuals r_j. The relative frequencies display a bell-shaped distribution.



Example (continued)

Our results suggests that the residuals follow a Gaussian distribution and that we may treat the residuals as realisations of a Gaussian random variable R. We verify this hypothesis by computing the empirical estimates for the mean $\bar{r} = \mathbb{E}[R]$, the variance $\sigma^2 = \mathbb{E}[(R - \bar{r})^2]$,

$$\text{skewness} = \frac{\mathbb{E}[(R - \bar{r})^3]}{\sigma^3}$$

and

$$\text{kurtosis} = \frac{\mathbb{E}[(R - \bar{r})^4]}{\sigma^4} - 3.$$

With $J = 2000$ data points we obtain the estimates $\bar{r}_J \approx 0.0060$, $\sigma_J^2 \approx 1.0184$, $\text{skewness}_J \approx 0.0031$, $\text{kurtosis}_J \approx 0.3177$. While the data is therefore not perfectly Gaussian, a Gaussian approximation with mean zero and variance one seems nevertheless justifiable.

Definition (Distance of probability measures)

Given two probability measures π_X and $\pi_{X'}$, a non-negative function $d(\pi_X, \pi_{X'}) \in \mathbb{R}_+$ is called a distance if

- $d(\pi_X, \pi_{X'}) = 0$ if and only if $\pi_X = \pi_{X'}$,
- $d(\pi_X, \pi_{X'}) = d(\pi_{X'}, \pi_X)$,
- $d(\pi_X, \pi_{X'}) \leq d(\pi_X, \pi_Z) + d(\pi_Z, \pi_{X'})$.

Example (TV-distance)

The total variation distance (TV-distance) is defined by

$$d_{\mathrm{TV}}(\pi_X, \pi_{X'}) = \int_{\mathbb{R}^N} |\pi_X(x) - \pi_{X'}(x)| \, \mathrm{d}x.$$

See Explain-Video for more details on the concept distances and divergences over probability spaces.

2.2 Coupling of measures and optimal transportation

Definition (Coupling of measures)

Let μ_{X_1} and μ_{X_2} denote two probability measures on a space $\mathcal{X}$. A coupling of μ_{X_1} and μ_{X_2} consists of a pair $Z = (X_1, X_2)$ of random variables such that $X_1 \sim \mu_{X_1}$, $X_2 \sim \mu_{X_2}$, and $Z \sim \mu_Z$. The joint measure μ_Z on the product space $\mathcal{Z} = \mathcal{X} \times \mathcal{X}$, is called the transference plan for this coupling. The set of all transference plans is denoted by $\Pi(\mu_{X_1}, \mu_{X_2})$.

Remark

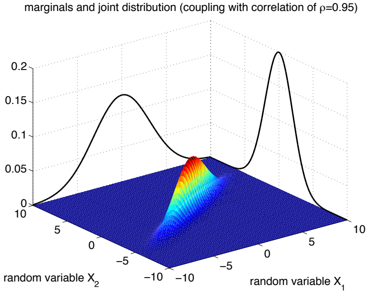
We will need to concept of coupling to understand modern algorithms for Bayesian inference which rely on a coupling between the prior and the posterior distributions. It is important to keep in mind that Bayes' theorem does **not** *specify such couplings (joint distributions between the prior and posterior distributions).*

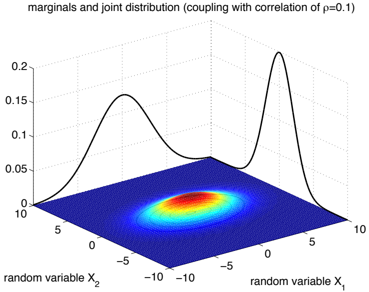
Example (Coupling univariate Gaussian measures)

We consider a pair of univariate Gaussian PDFs $\pi_{X_i}(x_i) = \mathrm{n}(x_i; \bar{x}_i, \sigma_{ii}^2)$, $i = 1, 2$. We seek a coupling in $z = (x_1, x_2)$ of the form $\pi_Z(z) = \mathrm{n}(z; \bar{z}, P)$. We find that $\bar{z} = (\bar{x}_1, \bar{x}_2)^T$ and the covariance matrix P must be of the form

$$P = \left(\begin{array}{cc} \sigma_{11}^2 & \rho\,\sigma_{11}\sigma_{22} \\ \rho\,\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{array} \right).$$

Since P has to be positive definite, the correlation between X_1 and X_2 has to satisfy $\rho^2 \leq 1$.


marginals and joint distribution (coupling with correlation of ρ=0.95)


marginals and joint distribution (coupling with correlation of ρ=0.1)

Example (continued)

We now investigate the conditional PDF $\pi_{X_2}(x_2|x_1)$ further. We have

$$\pi_{X_2}(x_2|x_1) = \mathrm{n}(x_2; \bar{x}_c, \sigma_c^2)$$

with

$$\bar{x}_c = \bar{x}_2 - \frac{\sigma_{21}^2}{\sigma_{11}^2}(\bar{x}_1 - x_1) = \bar{x}_2 - \rho\frac{\sigma_{22}}{\sigma_{11}}(\bar{x}_1 - x_1)$$

and

$$\sigma_c^2 = \sigma_{22}^2 - \sigma_{12}^4\sigma_{11}^{-2} = (1-\rho^2)\sigma_{22}^2.$$

We find that the limit $\rho \to 1$ leads to $\sigma_c \to 0$, which implies that the conditional probability density becomes a Dirac delta distribution centred about

$$\bar{x}_c = \bar{x}_2 - \frac{\sigma_{22}}{\sigma_{11}}(\bar{x}_1 - x_1).$$

Hence optimising the correlation between X_1 and X_2 has led us to a deterministic coupling

$$X_2 = \bar{x}_2 + \frac{\sigma_{22}}{\sigma_{11}}(X_1 - \bar{x}_1).$$

Definition (Deterministic coupling)

Assume that we have a random variable X_1 with law μ_{X_1} and a probability measure μ_{X_2}. A diffeomorphism $T : \mathcal{X} \to \mathcal{X}$ is called a transport map if the induced random variable $X_2 = T(X_1)$ satisfies

$$\mathbb{E}[f(X_2)] = \int_{\mathcal{X}} f(x_2)\mu_{X_2}(\mathrm{d}x_2) = \int_{\mathcal{X}} f(T(x_1))\mu_{X_1}(\mathrm{d}x_1) = \mathbb{E}[f(T(X_1))]$$

for all suitable functions $f : \mathcal{X} \to \mathbb{R}$. The associated coupling

$$\mu_Z(\mathrm{d}x_1, \mathrm{d}x_2) = \delta(x_2 - T(x_1))\mu_{X_1}(\mathrm{d}x_1)\mathrm{d}x_2,$$

where $\delta(\cdot)$ is the standard Dirac distribution, is called a deterministic coupling. Note that μ_Z is not absolutely continuous, even if both μ_{X_1} and μ_{X_2} are.

Example (Coupling univariate distributions)

Let $\pi_{X_1}(x)$ and $\pi_{X_2}(x)$ denote two PDFs on $\mathcal{X} = \mathbb{R}$. The associated cumulative distribution functions are defined by

$$F_{X_1}(x) = \int_{-\infty}^{x} \pi_{X_1}(x')\mathrm{d}x', \qquad F_{X_2}(x) = \int_{-\infty}^{x} \pi_{X_2}(x')\mathrm{d}x'.$$

The right inverse of F_{X_2} is given by

$$F_{X_2}^{-1}(p) = \inf\{x \in \mathbb{R} : F_{X_2}(x) \geq p\}$$

for $p \in [0,1]$. The inverse may be used to define a transport map that transforms X_1 into X_2 as follows,

$$X_2 = T(X_1) = F_{X_2}^{-1}(F_{X_1}(X_1)).$$

Example (continued)

For example, consider the case where X_1 is a random variable with uniform distribution $\mathrm{U}[0,1]$, and $X_2 \sim \mathrm{N}(0,1)$ is a standard Gaussian random variable. Then the transport map between X_1 and X_2 is simply the inverse of the cumulative distribution function

$$F_{X_2}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x')^2/2} \mathrm{d}x'.$$

This provides a standard tool for converting uniformly distributed random variables into Gaussian random variables.

The construction generalises to univariate measures μ_{X_2} which are not absolutely continuous. Consider, for example, the point measure μ_{x_0} with cumulative distribution function $F_{x_0}(x)$. Its right-continuous inverse is defined by

$$F_{x_0}^{-1}(p) = \inf\{x \in \mathbb{R} : F_{x_0}(x) \geq p\} = x_0$$

for all $p \in [0,1]$.

Example (Coupling multivariate Gaussian measures)

Consider two Gaussian distributions $\mathrm{N}(\bar{x}_1, P_1)$ and $\mathrm{N}(\bar{x}_2, P_2)$ in $\mathbb{R}^{N_x}$ with means $\bar{x}_1$ and $\bar{x}_2$ and covariance matrices P_1 and P_2, respectively. We first define the square root $P^{1/2}$ of a symmetric positive definite matrix P as the unique symmetric matrix which satisfies $P^{1/2}P^{1/2} = P$. Then the affine transformation

$$x_2 = T(x_1) = \bar{x}_2 + P_2^{1/2} P_1^{-1/2}(x_1 - \bar{x}_1)$$

provides a deterministic coupling. Indeed, we find that

$$(x_2 - \bar{x}_2)^T P_2^{-1}(x_2 - \bar{x}_2) = (x_1 - \bar{x}_1)^T P_1^{-1}(x_1 - \bar{x}_1)$$

under the suggested coupling. In contrast with the univariate case, deterministic couplings are not uniquely defined since

$$x_2 = T(x_1) = \bar{x}_2 + P_2^{1/2} Q P_1^{-1/2}(x_1 - \bar{x}_1),$$

where Q is any orthogonal matrix, also provides a deterministic coupling.

Remark (Maximising correlation)

One way to select a particular coupling is to choose the one that maximises the covariance. In addition, maximising the covariance for given marginals also has an important geometric interpretation. For simplicity, consider univariate random variables X_1 and X_2. Then we have

$$\begin{aligned}\mathbb{E}[(X_2 - X_1)^2] &= \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] - 2\mathbb{E}[X_1 X_2] \\ &= \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] - 2\mathbb{E}[(X_1 - \bar{x}_1)(X_2 - \bar{x}_2)] - 2\bar{x}_1\bar{x}_2 \\ &= \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] - 2\bar{x}_1\bar{x}_2 - 2cov(X_1, X_2).\end{aligned}$$

Hence, finding a joint measure μ_Z that minimises the expectation of $(X_1 - X_2)^2$ simultaneously maximises the covariance between univariate random variables X_1 and X_2. This geometric interpretation extends to multivariate random variables and leads to the celebrated Monge-Kantorovitch problem.

Definition (Monge-Kantorovitch problem)

A transference plan $\mu_Z^* \in \Pi(\mu_{X_1}, \mu_{X_2})$ is called the solution to the Monge-Kantorovitch problem with cost function $c(x_1, x_2) = \|x_1 - x_2\|^2$ if

$$\mu_Z^* = \arg \inf_{\mu_Z \in \Pi(\mu_{X_1}, \mu_{X_2})} \mathbb{E}[\|X_1 - X_2\|^2], \qquad \text{law}(Z = (X_1, X_2)) = \mu_Z,$$

where $\Pi(\mu_{X_1}, \mu_{X_2})$ denotes the set of all possible couplings between μ_{X_1} and μ_{X_2}. The associated functional

$$W(\mu_{X_1}, \mu_{X_2}) = \sqrt{\mathbb{E}[\|X_1 - X_2\|^2]},$$

is called the L^2-Wasserstein distance between μ_{X_1} and μ_{X_2}.

Example (linear transport problem)

In this example we consider couplings between two discrete random variables X_1, X_2 with domain given by the discrete set

$$\mathcal{X} = \{a_1, a_2, \ldots, a_M\}, \qquad a_i \in \mathbb{R}, \tag{2}$$

and probability distributions

$$\mathbb{P}(X_1 = a_i) = 1/M, \quad \mathbb{P}(X_2 = a_i) = w_i,$$

respectively, with $w_i \ge 0$, $i = 1, \ldots, M$, and $\sum_i w_i = 1$. Any coupling between these two probability distributions is characterised by a matrix $T \in \mathbb{R}^{M\times M}$ such that its entries $t_{ij} = (T)_{ij}$ satisfy $t_{ij} \ge 0$ and

$$\sum_{i=1}^{M} t_{ij} = 1/M, \qquad \sum_{j=1}^{M} t_{ij} = w_i. \tag{3}$$

These matrices characterise the set of all couplings Π in the definition of the Monge-Kantorovitch problem.

Example (continued)

Given a coupling T and the mean values

$$\bar{x}_1 = \frac{1}{M}\sum_{i=1}^{M} a_i, \qquad \bar{x}_2 = \sum_{i=1}^{M} w_i a_i,$$

the covariance between the associated discrete random variables X_1 and X_2 is defined by

$$\mathrm{cov}(X_1, X_2) = \sum_{i,j=1}^{M} (a_i - \bar{x}_2) t_{ij} (a_j - \bar{x}_1). \tag{4}$$

The particular coupling defined by $t_{ij} = w_i/M$ leads to zero correlation between X_1 and X_2.

Example (continued)

On the other hand, maximising the correlation leads to a linear transport problem in the M^2 unknowns $\{t_{ij}\}$. More precisely, the unknowns t_{ij} have to satisfy the inequality constraints $t_{ij} \geq 0$, the equality constraints (3), and should minimise

$$J(\{t_{ij}\}) = \sum_{i,j=1}^{M} t_{ij}(a_i - a_j)^2$$

which, following our previous discussion, is equivalent to maximising the correlation.

Example (coupling sorted samples)

We demonstrate how to obtain the optimal coupling for the (sorted) discrete target set $\mathcal{X}$ given by

$$a_i = \frac{1}{2M} + \frac{i-1}{M} \in [0,1]$$

with $M = 10$. The weights w_i are determined by $(0.2002, 0.2001, 0.1807, 0.149, 0.11, 0.073, 0.045, 0.025, 0.012, 0.005)$. The optimal coupling matrix is

$$T^* \approx \begin{pmatrix} 0.1 & 0.1 & 0.0002 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0998 & 0.1 & 0.0003 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0997 & 0.081 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.019 & 0.1 & 0.03 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.07 & 0.04 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.06 & 0.013 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.045 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.025 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.012 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.005 \end{pmatrix}.$$

Example (Optimal coupling univariate measures)

Consider two univariate PDFs π_{X_1} and π_{X_2} with cumulative distribution function F_{X_1} and F_{X_2}, respectively. Then, as already discussed, a coupling is achieved by

$$x_2 = T(x_1) = F_{X_2}^{-1}(F_{X_1}(x_1)),$$

which is also optimal under the $(\cdot)^2$ distance. Furthermore, the L^2-Wasserstein distance between π_{X_1} and π_{X_2} is given by

$$\begin{aligned} W(\pi_{X_1}, \pi_{X_2})^2 &= \int_{\mathbb{R}} \int_{\mathbb{R}} (x_1 - x_2)^2 \delta\left(x_2 - F_{X_2}^{-1}(F_{X_1}(x_1))\right) \pi_{X_1}(x_1) \mathrm{d}x_1 \mathrm{d}x_2, \\ &= \int_0^1 \int_{\mathbb{R}} (F_{X_1}^{-1}(p) - x_2)^2 \delta\left(x_2 - F_{X_2}^{-1}(p)\right) \mathrm{d}x_2 \mathrm{d}p, \\ &= \int_0^1 (F_{X_1}^{-1}(p) - F_{X_2}^{-1}(p))^2 \mathrm{d}p, \end{aligned}$$

with $p = F_{X_1}(x_1)$ and $\mathrm{d}p = \pi_{X_1}(x_1)\mathrm{d}x_1$.

Example (Optimal coupling of Gaussian measures)

Consider two Gaussian distributions $\mathrm{N}(\bar{x}_1, P_1)$ and $\mathrm{N}(\bar{x}_2, P_2)$ in $\mathbb{R}^{N_x}$, with means $\bar{x}_1$ and $\bar{x}_2$ and covariance matrices P_1 and P_2, respectively. We had previously discussed deterministic couplings. However, the induced affine transformation $x_2 = T(x_1)$ cannot be generated from a potential ψ since the matrix $P_2^{1/2} P_1^{-1/2}$ is not symmetric. Indeed, the optimal coupling in the sense of Monge-Kantorovitch with cost function $c(x_1, x_2) = \|x_1 - x_2\|^2$ is provided by

$$x_2 = T(x_1) := \bar{x}_2 + P_2^{1/2} \left[P_2^{1/2} P_1 P_2^{1/2} \right]^{-1/2} P_2^{1/2} (x_1 - \bar{x}_1) \,.$$

The associated Wasserstein distance between the two Gaussian distributions is

$$W(\mu_{X_1}, \mu_{X_2})^2 = \|\bar{x}_1 - \bar{x}_2\|^2 + \mathrm{trace}\left(P_1 + P_2 - 2 \left[P_2^{1/2} P_1 P_2^{1/2} \right]^{1/2} \right) .$$

The remaining material from these slides is advanced and isn't examinable for data science students.

Theorem (Optimal transference plan)

If the measures μ_{X_1}, μ_{X_2} on $\mathcal{X} = \mathbb{R}^{N_x}$ are absolutely continuous and have bounded second-order moments, then the optimal transference plan that solves the Monge-Kantorovich problem corresponds to a deterministic coupling with transfer map

$$X_2 = T(X_1) = \nabla_x \psi(X_1),$$

for some convex potential $\psi : \mathbb{R}^{N_x} \to \mathbb{R}$.

Proof.

See Theorem 2.30 from the textbook where the problem is reformulated in terms of a variational problem. A heuristic derivation of the theorem is discussed on the subsequent slides. □

In order to describe the geometric structure of optimal couplings (whether deterministic or not) we need to introduce the two concepts of cyclical monotonicity and subdifferentials of convex functions.

Definition (Support of a measure and cyclical monotonicity)

The support of a coupling μ_Z on $\mathcal{Z} = \mathcal{X} \times \mathcal{X}$ is the smallest closed set on which μ_Z is concentrated, *i.e.*

$$\mathrm{supp}\,(\mu_Z) := \bigcap \{S \subset \mathcal{Z} :\ S\,\text{closed and}\ \mu_Z(\mathcal{Z} \setminus S) = 0\}.$$

The support of μ_Z is called cyclically monotone if for every set of points $(x_1^i, x_2^i) \in \mathrm{supp}\,(\mu_Z) \subset \mathcal{X} \times \mathcal{X}$, $i = 1, \ldots, I$, and any permutation σ of $\{1, \ldots, I\}$, we have

$$\sum_{i=1}^{I} \|x_1^i - x_2^i\|^2 \le \sum_{i=1}^{I} \|x_1^i - x_2^{\sigma(i)}\|^2.$$

Remark (Cyclical monotonicity)

Note that the previous definition is equivalent to

$$\sum_{i=1}^{I} (x_1^i)^{\mathrm{T}} (x_2^{\sigma(i)} - x_2^i) \le 0.$$

We will find later that cyclical monotonicity implies that the support of a transference plan is contained in the subdifferential of an appropriate convex functional $\psi(x)$. If that convex function can be shown to be sufficiently regular then the transference plan is deterministic. This chain of arguments can be made rigorous, providing a proof of the previous theorem.

Example (Cyclical monotonicity and optimal couplings)

We return to the linear transport problem associated with the optimal coupling $\mu^*_{X_1X_2}$ of the two discrete random variables X_1 and X_2. The support of the optimal coupling T^* is defined by

$$(x_1 = a_j, x_2 = a_i) \in \operatorname{supp}\left(\mu^*_{X_1X_2}\right) \text{ if and only if } \quad t^*_{ij} > 0.$$

Let T now denote any coupling matrix, *i.e.*, a matrix with non-negative entries and

$$\sum_{i=1}^{M} t_{ij} = 1/M, \qquad \sum_{j=1}^{M} t_{ij} = w_i,$$

and take any two pairs $(a_{j_1}, a_{i_1}) \in \mathbb{R}^2$ and $(a_{j_2}, a_{i_2}) \in \mathbb{R}^2$ such that $t_{i_1j_1} > 0$ and $t_{i_2j_2} > 0$. If this pair does not satisfy cyclical monotonicity, *i.e.*,

$$(a_{i_1} - a_{j_1})^2 + (a_{i_2} - a_{j_2})^2 > (a_{i_1} - a_{j_2})^2 + (a_{i_2} - a_{j_1})^2,$$

then there is a T' with lower cost. See Example 2.34 from the book for details.

Theorem (Cyclical monotonicity)

If μ_Z^* is a solution to the Monge-Kantorovitch problem, then μ_Z^* has cyclically monotone support.

Proof.

See Theorem 2.35 on page 57 from the textbook for details. □

A fundamental theorem of convex analysis, Rockafellar's theorem, states that cyclically monontone sets $S \subset \mathbb{R}^{N_x} \times \mathbb{R}^{N_x}$ are contained in the subdifferential of a convex function $\psi : \mathbb{R}^{N_x} \to \mathbb{R}$.

Definition (Subdifferential)

The subdifferential $\partial\psi$ of a convex function ψ at a point $x \in \mathbb{R}^{N_x}$ is defined as the non-empty and convex set of all $m \in \mathbb{R}^{N_x}$ such that

$$\psi(x') \geq \psi(x) + m(x' - x)$$

for all $x' \in \mathbb{R}^{N_x}$. We write $m \in \partial\psi(x)$.

Example

Consider the convex function $\psi(x) = |x|$, which is differentiable away from $x = 0$ with $\psi'(x) = 1$ for $x > 0$ and $\psi'(x) = -1$ for $x < 0$. At $x = 0$ we need to find the set of all m such that

$$|x'| \geq mx'$$

and we find that $m \in [-1, 1]$. This is the interval spanned by the left and right derivative of ψ at $x = 0$. In summary, we obtain the set-valued subdifferential

$$\partial\psi(x) = \begin{cases} -1 & \text{for } x < 0, \\ [-1, 1] & \text{for } x = 0, \\ 1 & \text{for } x > 0. \end{cases}$$

Lemma

Given a cyclically monotone set $S \in \mathbb{R}^{N_x} \times \mathbb{R}^{N_x}$, a convex potential ψ such that $S \subset Graph(\partial\psi)$ can be defined as follows. Pick a particular pair $(x_1^0, x_2^0) \in S$. Then define

$$\psi(x) := \sup\{(x_2^I)^{\rm T}(x - x_1^I) + (x_2^{I-1})^{\rm T}(x_1^I - x_1^{I-1}) + \cdots + (x_2^0)^{\rm T}(x_1^1 - x_1^0)\}.$$

Here the supremum is taken over all possible families of pairs $\{(x_1^i, x_2^i) \in S\}_{i=1}^I$ with integer $I \geq 1$. Cyclical monotonicity of S implies that $\psi(x_1^0) \leq 0$ and the particular choice $I = 1$ with $x_1^1 = x_1^0$ and $x_2^1 = x_2^0$ leads to $\psi(x_1^0) = 0$.

Proof.

The proof of the Lemma can be found on page 82 of *Topics of Optimal Transportation* by Cedric Villani. See also the Explain Video for an example. □

Example (convex potential for cyclically monotone set)

If S is finite, then the previous proposition is constructive. Consider, for example, the coupling between X_1 and X_2 defined the example on page 36. The set S is defined by $(x_1 = a_j, x_2 = a_i) \in S$ if and only if $t^*_{ij} > 0$. Using $(x^0_1, x^0_2) = (a_1, a_1)$ in the definition of ψ leads to the convex potential:

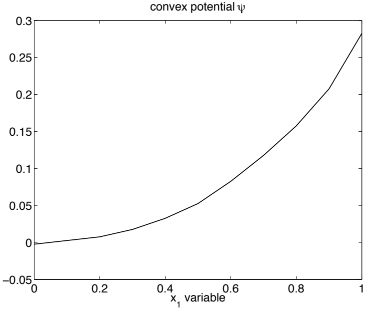


Figure: The potential is piecewise linear and its subdifferential $\partial\psi(x_1)$ is defined as the derivative of ψ whenever $x_1 \neq a_i$, $i = 1, \dots, M$, and as the interval spanned by the left and right derivatives of ψ whenever $x_1 = a_i$. These intervals include all possible cases $x_2 = a_i$ which are coupled to $x_1 = a_j$ by a non-zero entry in t^*_{ij}.

Remark (optimal coupling)

A compact statement of Rockafellar's theorem is that $S \subset \mathit{Graph}(\partial\psi)$ *for a suitable convex potential* ψ*. An optimal transport map* T *is obtained whenever* ψ *is sufficiently regular in which case the subdifferential* $\partial\psi(x)$ *reduced to the classic gradient* $\nabla_x\psi$ *and* $x_2 = T(x_1) = \nabla_x\psi(x_1)$*. This happens, for example, when the involved (marginal) measures are absolutely continuous with bounded second-order moments.*

While optimal couplings between continuous random variables are of broad theoretical and practical interest, their computational implementation is non-trivial. The case of discrete random variables with target space can be dealt with by linear transport algorithms. Furthermore, for scalar-valued discrete random variables the problem can easily be solved by sorting. See Section 5.8 for details.