

# Machine learning under physical constraints

## Introduction

Sixin Zhang  
([sixin.zhang@toulouse-inp.fr](mailto:sixin.zhang@toulouse-inp.fr))

# Outline

Course Content

Basic knowledge

# Outline

## Course Content

Basic knowledge

# Main problems

- ▶ Infer the state of dynamical system from observations: Machine learning and Data assimilation.
- ▶ Construct invariant representations: Machine learning and geometry of physics.

# Part I: Machine learning for Data assimilation (DA)

- ▶ Recurrent neural networks
- ▶ Data assimilation networks (DAN)
- ▶ Unsupervised learning in DA

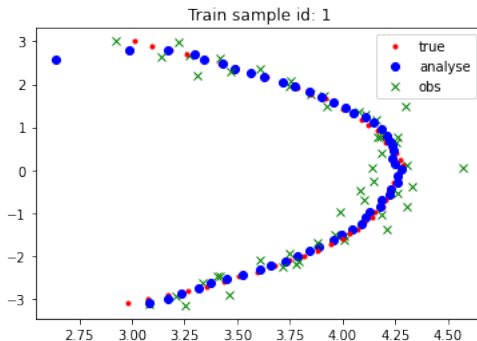
## Part II: Invariant representation for classification and regression

- ▶ Invariant properties in physical systems
- ▶ From Fourier to wavelet representation
- ▶ Wavelet scattering transform

# Course Evaluation

- ▶ Evaluation of TD: Basics
- ▶ Evaluation of TP: Implementation of DAN, wavelet scattering
- ▶ Kaggle project: Regression of molecular energy

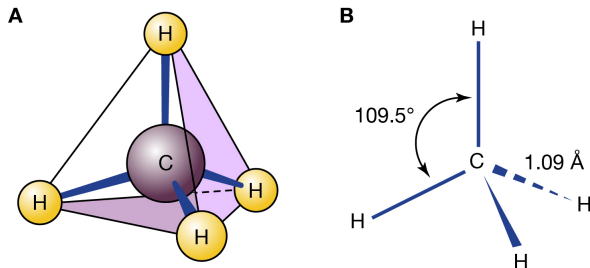
# Demo of TP: Data assimilation networks



**Figure:** The dynamics of  $x_t$  (true) and  $y_t$  (obs) in Linear 2d, together with the trajectories of the mean  $\mu_t^a$  of the analysis probability density  $q_t^a$ .



# Kaggle project: regression of molecular energy



© Encyclopædia Britannica, Inc.

**Figure:** Predict the molecular energy in 3d space based on its geometric structure. Image from <https://www.britannica.com/science/methane>.

# Outline

## Course Content

### Basic knowledge

# Recall

- ▶ Matrix calculus
- ▶ Probability and statistics
- ▶ Expectation-Maximization algorithm (EM)
- ▶ Reference: The Matrix Cookbook  
[<http://matrixcookbook.com>] by K. Petersen and M. Petersen, Pattern Recognition and Machine Learning by Christopher M. Bishop

# Matrix calculus: derivative

- ▶ Compute the derivative of a function of a vector or matrix
- ▶ Example 1: Assume  $A$  is a real symmetric matrix and  $x \in \mathbb{R}^d$ , then

$$\frac{\partial x^T A x}{\partial x} = 2Ax$$

- ▶ Example 2: Assume  $A$  and  $X$  are two matrices, then

$$\frac{\partial \text{tr}(X^T A)}{\partial X} = A$$

# Sherman-Morrison Formula

- ▶ For matrix inversion, it is useful to compute it progressively.
- ▶ Assume  $A \in \mathbb{R}^{n \times n}$  is an invertible matrix, and  $U \in \mathbb{R}^{n \times k}$  and  $V \in \mathbb{R}^{k \times n}$  such that  $I + VA^{-1}U$  is invertible, then

$$(A + UV)^{-1} = A^{-1} - A^{-1}U(I + VA^{-1}U)^{-1}VA^{-1}.$$

## Joint and Posterior distribution: Bayesian

Let  $v \in \mathbb{R}^n$ ,  $u \in \mathbb{R}^k$ ,  $L \in \mathbb{R}^{k \times n}$ ,  $m \in \mathbb{R}^m$ . Assume  $\Sigma \in \mathbb{R}^{k \times k}$  is positive definite, and  $v \sim \mathcal{N}(\mu, K)$ ,  $u|v \sim \mathcal{N}(Lv + m, \Sigma)$ , then



$$\begin{pmatrix} v \\ u \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \mu \\ L\mu + m \end{pmatrix}, \begin{pmatrix} K & (LK)^\top \\ LK & \Sigma + LKL^\top \end{pmatrix} \right)$$

# Maximum-likelihood estimation

- ▶ A principle to derive many common estimators in statistics, such as mean, variance, etc.
- ▶ Relation with KL divergence, one can rewrite

$$\max_{\theta} \mathbb{E}_{x \sim q}(\log p(x|\theta)) = \int \log p(x|\theta) q(x) dx$$

as

$$\min_{\theta} \int \log \frac{q(x)}{p(x|\theta)} q(x) dx$$

- ▶ Example: Gaussian model

# KL divergence

- ▶ Measure the difference between two densities  $p$  and  $q$ :

$$KL(q||p) = \int \log \frac{q(z)}{p(z)} q(z) dz$$

- ▶ It is **not a symmetric distance**, i.e.

$$KL(q||p) \neq KL(p||q)$$

- ▶ Jensen inequality shows that it is always positive

$$KL(q||p) \geq 0$$

with equality = holds i.f.f  $p(z) = q(z)$  a.e.



# EM algorithm: estimation in latent models

- ▶ A way to perform maximum-likelihood estimation (MLE) for latent variable models.
- ▶ Latent variable model:  $p(x, z|\theta)$ , only  $x$  is observed.
- ▶ Problem: estimate  $\theta$  from  $x$  by MLE

$$\max_{\theta} \log p(x|\theta) = \log \int p(x, z|\theta) dz$$

- ▶ Attention: we omit taking the expectation on  $x$  for simplicity.

# EM algorithm: principle

- ▶ Let

$$L(\theta, q) = \int \log \frac{p(x, z|\theta)}{q(z)} q(z) dz$$

- ▶ The maximization of  $\log p(x|\theta)$  can be solved by two steps in an alternative fashion,

$$\text{E-step : } \max_q L(\theta, q),$$

$$\text{M-step : } \max_{\theta} L(\theta, q).$$

- ▶ Justified by an important equality: for any density  $q(z)$ ,

$$\log p(x|\theta) = \int \log \frac{p(x, z|\theta)}{q(z)} q(z) dz - \int \log \frac{p(z|x, \theta)}{q(z)} q(z) dz$$

# EM algorithm: justification from Fisher equality

- ▶ Under suitable assumption, we have the following Fisher equality

$$\nabla_{\theta} \log p(x|\theta)|_{\theta=\theta_0} = \int \nabla_{\theta} \log p(x, z|\theta)|_{\theta=\theta_0} p(z|x, \theta_0) dz$$

- ▶ By following the gradient direction  $\nabla_{\theta} \log p(x|\theta)$ , one can increase the value of  $\log p(x|\theta)$ .

# EM algorithm: justification from log partition function

- ▶ Relation to physics: compute log partition function.
- ▶ Let  $f(x)$  be a function of  $x \in \mathbb{R}^d$ , how to compute

$$\log Z = \log \int e^{f(x)} dx \quad ?$$

- ▶ Let  $q$  be a probability density on  $\mathbb{R}^d$ . The key idea of EM is related to

$$\log Z = \max_q \int f(x)q(x)dx - \int \log q(x)q(x)dx$$

- ▶ The optimal  $q(x) = e^{f(x)}/Z$ .