Bayesian inference and related computational methods

Brief overview

L. Martino

May, 2022

- **Preambule - or a very long introduction from a signal processing point of view...**
- Bayesian Inference
- Computational methods for Bayesian Inference
- Monte Carlo sampling methods - brief overview

Basic, standard problem

- In many applications, we are interested in inferring a variable of interest,

$$\boldsymbol{\theta} = [\theta_1, ..., \theta_{d_\theta}] \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{d_\theta},$$

given a set of observations $\mathbf{y} \in \mathbb{R}^{d_Y}$.

- **We want to know θ given y:**

$$\boldsymbol{\theta} \Longrightarrow \mathbf{y}$$

Standard - typical approach

- Minimizing a cost function:

$$C(\boldsymbol{\theta}) = \texttt{Loss}(\boldsymbol{\theta}, \mathbf{y}),$$

obtaining

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} C(\boldsymbol{\theta}).$$

REGULARIZATION:

- fighting against overfitting,
- avoiding numerical problems and increasing the numerical stability.

Classical approach + regularization

Cost function:

$$C(\boldsymbol{\theta}) = \texttt{Loss}(\boldsymbol{\theta}, \mathbf{y}) + \texttt{Reg}(\boldsymbol{\theta})$$

Again, we minimize it (**optimization**).

- Why pass to a probabilistic domain/approach?
- **from optimization** $==>$ **to sampling**, why?
- a first answer in the next two slides.

Example: nonlinear model for regression

Regression problem: Assume that we have N data $\{x_i, y_i\}$.
We consider M bases

$$\phi_m(x) : \mathbb{R}^D \longrightarrow \mathbb{R}, \qquad m = 1, ..., M.$$

- We want that the solution has the following mathematical form:

$$\widehat{f}(x) = \sum_{m=1}^{M} \theta_m \phi_m(x).$$

- We want to find the θ_m's, we will consider

$$M \leq N.$$

Our nonlinear model

Assuming that we have N data $\{x_i, y_i\}$,

$$y_i = \sum_{m=1}^{M} \theta_m \phi_m(x_i) + error...$$

with

$$M \leq N.$$

Can we still associate a linear system?

Example: our observation model

the model is **non-linear**...but...

it is still linear...

with respect to the coefficients $\theta_1, \theta_2, ..., \theta_M$, — YES is still linear ! for this reason, it can be analytically solved

Then, we can construct some matrices and vectors...

Example: Rectangular linear system ($M \leq N$)

We can define:

$$\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_M]^\top, \quad M \times 1$$

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & ... & \phi_M(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & ... & \phi_M(x_2) \\ \vdots & \vdots & & \\ \phi_1(x_N) & \phi_2(x_N) & ... & \phi_M(x_N) \end{bmatrix} \quad N \times M,$$

$$\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top, \quad N \times 1.$$

Example: rectangular linear system

The system can be written as:

$$\mathbf{\Phi\theta = y}.$$

Check the dimensions

$$[N \times M] \times [M \times 1] = N \times 1.$$

Example: rectangular linear system

Since the system is rectangular, we cannot write

$$\boldsymbol{\theta} = \boldsymbol{\Phi}^{-1}\mathbf{y}, \quad \textit{NOOOO!!!}$$

the inverse matrix $\boldsymbol{\Phi}^{-1}$ does not exist ! since $\boldsymbol{\Phi}$ is rectangular !

Example

Cost function of the Regularized Least Squares:

$$C(\boldsymbol{\theta}) = ||\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}||^2 + \lambda||\boldsymbol{\theta}||^2.$$

we want to minimize $C(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

Example

Solution:

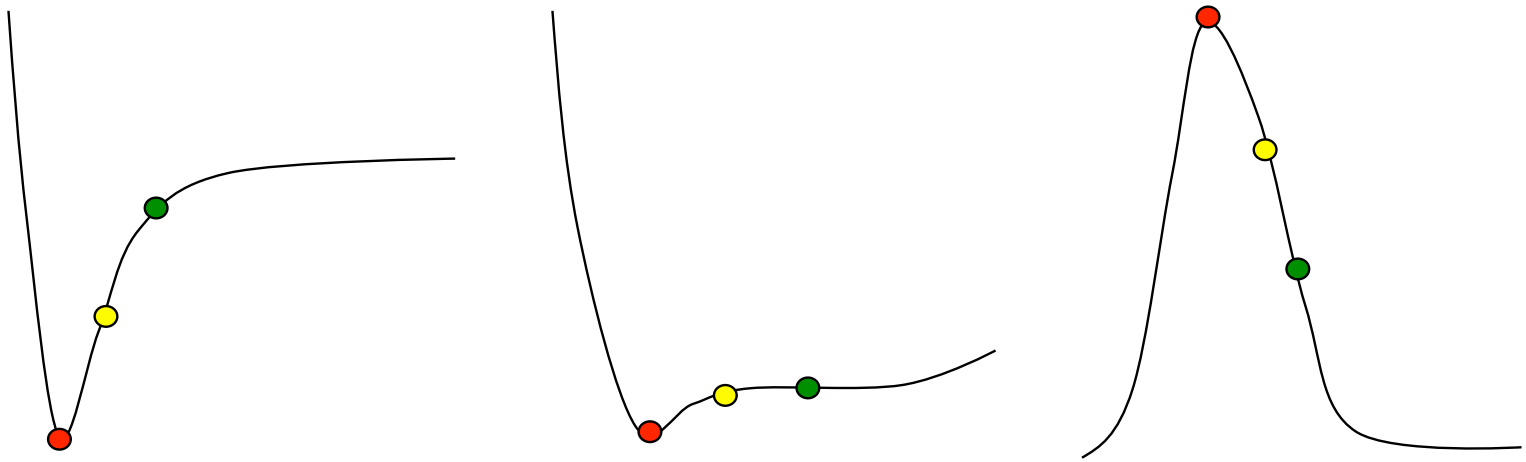
$$\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_1, ..., \widehat{\theta}_M]^\top = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \mathbf{I}_M)^{-1} \boldsymbol{\Phi}^\top \mathbf{y}. \quad (1)$$

Recall that solution has the following mathematical form:

$$\widehat{f}(x) = \sum_{m=1}^{M} \widehat{\theta}_m \phi_m(x).$$

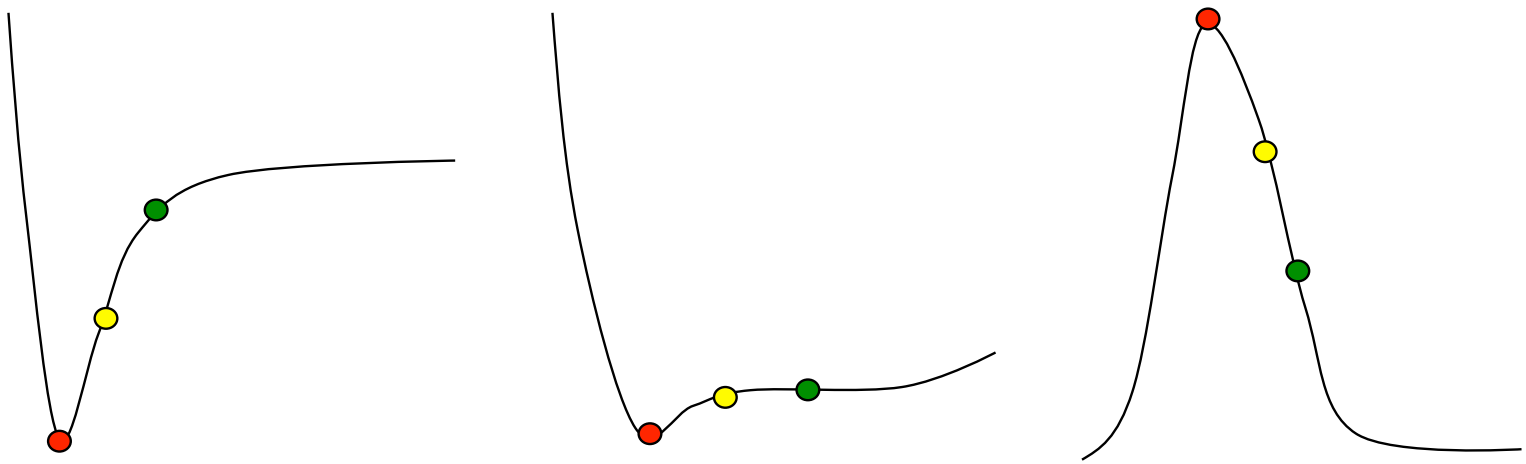
Probabilistic approach

(a) Differ among different cost functions, **(b)** different possible "points" for summarizing the cost function, **(c)** compute areas (e.g., for credible intervals) ...



Probabilistic approach

in the "probabilistic" approach: the MAP estimator, the MMSE estimator (*mean - expected value*), the median estimator are well-defined, and also the "areas" have a "meaning" and can be computed...



Probabilistic version...

Cost function:

$$C(\boldsymbol{\theta}) = \underbrace{\texttt{Loss}(\boldsymbol{\theta}, \mathbf{y})}_{\texttt{neg. log-likelihood}} + \underbrace{\texttt{Reg}(\boldsymbol{\theta})}_{\texttt{neg. log-prior}}$$

$$\texttt{likelihood} = p(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\left(-\texttt{Loss}(\boldsymbol{\theta}, \mathbf{y})\right),$$

$$\texttt{prior} = p(\boldsymbol{\theta}) \propto \exp\left(-\texttt{Reg}(\boldsymbol{\theta})\right).$$

Bayesian "slang"

Cost function:

$$C(\boldsymbol{\theta}) = \underbrace{\texttt{Loss}(\boldsymbol{\theta}, \mathbf{y})}_{\texttt{neg. log-likelihood}} + \underbrace{\texttt{Reg}(\boldsymbol{\theta})}_{\texttt{neg. log-prior}}$$

Bayesian Inference:

$$\texttt{posterior} \propto \exp\{-C(\boldsymbol{\theta})\} = \texttt{likelihood} \times \texttt{prior}$$

Bayesian Inference

Bayesian Inference:

$$\texttt{posterior} \propto \texttt{likelihood} \times \texttt{prior}.$$

where

$$\texttt{posterior} = p(\boldsymbol{\theta}|\mathbf{y}),$$

$$\texttt{likelihood} = p(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\left(-\texttt{Loss}(\boldsymbol{\theta}, \mathbf{y})\right),$$

$$\texttt{prior} = p(\boldsymbol{\theta}) \propto \exp\left(-\texttt{Reg}(\boldsymbol{\theta})\right).$$

Prior versus Regularization

$$\texttt{prior} = p(\boldsymbol{\theta}) \propto \exp\left(-\texttt{Reg}(\boldsymbol{\theta})\right).$$

Main difference: the prior density must/should be normalized ("normalizable")... (in some case, this condition can be also relaxed)

Normalization ==> since it represents probabilities ==> now we have more interpretability of different situations (think on different regularizations - e.g., previous figures - and different priors ...)

- Preambule - or a very long introduction from a signal processing point of view...
- **Bayesian Inference**
- Computational methods for Bayesian Inference
- Monte Carlo sampling methods - brief overview

Bayesian Inference:

1. **"Main Actors" in Bayesian inference**
2. Important considerations and consistency
3. Goals
4. Levels of inference
5. Type of Priors - choice of the priors
6. Reasons to be Bayesian

Problem Statement and main actors

- In many applications, we are interested in inferring a variable of interest,

$$\boldsymbol{\theta} = [\theta_1, ..., \theta_{d_\theta}] \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{d_\theta},$$

given a set of observations $\mathbf{y} \in \mathbb{R}^{d_Y}$.

Problem Statement and main actors

- The **posterior probability density function (pdf)** is

$$\bar{\pi}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})}{p(\mathbf{y})} \propto \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}), \qquad (2)$$

where

  - $\ell(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function (induced by the observation model);
  - $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ is the prior pdf,
  - $Z = p(\mathbf{y})$: marginal likelihood/Bayesian evidence.

(note that we have 2 conditionals $p(\boldsymbol{\theta}|\mathbf{y})$, $\ell(\mathbf{y}|\boldsymbol{\theta})$, and 2 marginals, $g(\boldsymbol{\theta})$ and $p(\mathbf{y})$ - with the prior and the likelihood, we create a joint pdf of $\boldsymbol{\theta}, \mathbf{y}$)

Marginal likelihood - Bayesian evidence

- Given $\mathbf{y}$, the marginal likelihood - Bayesian model evidence is an integral (a *normalizing constant*):

$$Z = p(\mathbf{y}) = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

  is fixed and, in general, is **unknown**.
- $Z =$ **Weighted average of the likelihood values !!**
- Note that

$$0 \leq \min p(\mathbf{y}|\boldsymbol{\theta}) \leq Z \leq \max p(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y}|\widehat{\boldsymbol{\theta}}_{\mathtt{ML}}).$$

- $Z = p(\mathrm{y})$ useful for model selection purposes.

Examples of "model selection"

Model selection: (some examples)

- Tuning of the parameters of the observation model (i.e., of the likelihood).
- Tuning of the parameters of prior.
- **Choose the best observation/measurement model among a set of possible models.**
- Select the order/complexity in a model (for instance, the order of a polynomial in a regression, or the order of an AR - FIR - filter etc.)
- Variable selection.
- etc.

Examples of "model selection"

Model selection: - The previous examples can divided in two main scenarios:

- Basic model selection
- Nested models

This classification is important for the possible choice of the priors.

- F. Llorente, L. Martino, E. Cuberlo, J. Lopez-Santiago, D. Delgado, "On the safe use of prior densities for Bayesian model selection", viXra:2110.0032 , 2021.

Unnormalized posterior

- Since Z is generally unknown: then, in many cases, we are only able to evaluate the unnormalized pdf

$$\pi(\boldsymbol{\theta}) = \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}) \propto \bar{\pi}(\boldsymbol{\theta}).$$

  Note that

$$\bar{\pi}(\boldsymbol{\theta}) = \frac{1}{Z}\pi(\boldsymbol{\theta}).$$

- Note that

$$Z = \int_{\Theta} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Break to explain: "Evaluation versus Sampling"

"Evaluation versus Sampling of a density" in this slides

- **Evaluation:** evaluate point-wise a function/density.
- **Sampling:** GENERATE RANDOM NUMBERS according to a density.

Typical example of application: Bayesian Inversion

- **Observation model - inducing likelihood:**

$$\mathbf{y} = \mathbf{G}(\boldsymbol{\theta}) + \mathbf{v},$$

where $\mathbf{G}(\boldsymbol{\theta})$ is a "physical" model (for instance) and $\mathbf{v}$ is an independent Gaussian noise (for instance).

- **Likelihood:**

$$p(\mathbf{y}|\boldsymbol{\theta}) = \ell(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\left(-\frac{||\mathbf{y} - \mathbf{G}(\boldsymbol{\theta})||^2}{2\sigma_v^2}\right).$$

- the "goal" is: virtually, $\boldsymbol{\theta} = \mathbf{G}^{-1}(\mathbf{y})$. For this reason, it is called "inversion"....

Bayesian Inference:

1. "Main Actors" in Bayesian inference
2. **Important considerations and consistency**
3. Goals
4. Levels of inference
5. Type of Priors - choice of the priors
6. Reasons to be Bayesian

Main differences with frequentist approach

- **The vector of data y is given and fixed.**
- $\boldsymbol{\theta}$ "should" be considered random, since we assume $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$.
- But "practical Bayesians" and/or "Bayesians with common sense", considers/knows that it exists a (fixed) $\boldsymbol{\theta}_{\texttt{true}}$, that we desire to infer.
- In fact, under mild conditions, **the Bayesian estimators are** *consistent* **as the number of data grows.**
- (note that bias, variance and MSE of an estimator are more frequentist ideas/quantities since consider expectation over **y**... we can extend these concepts here, considering different posteriors - one for each $\mathbf{y}'$ generated according to the model- and then make an average...)

Consistency

- Under mild conditions, **the Bayesian estimators are** *consistent* **as the number of data grows.**
- **Consistency:** As the number of data grows, the posterior becomes more tighter and tighter around $\theta_{\texttt{true}}$.

Bayesian Inference:

1. "Main Actors" in Bayesian inference
2. Important considerations and consistency
3. **Goals**
4. Levels of inference
5. Type of Priors - choice of the priors
6. Reasons to be Bayesian

Main goal

- Goal: **extract and summarize the statistical information contained in the posterior pdf $\bar{\pi}(\theta)$ and compute Z.**

Goal - quadrature problems

- More specifically, *in many cases*, our goal is to compute efficiently some integral involving π,

$$\mathbf{I} = \frac{1}{Z} \int_{\boldsymbol{\Theta}} \mathbf{f}(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{3}$$

where $\mathbf{f}(\boldsymbol{\theta}) : \boldsymbol{\Theta} \to \mathbb{R}^{\eta}$, and

$$Z = \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

In most of the cases, Z is also unknown and we have to estimate it (useful for model selection purposes).

Goal - quadrature problems - MOMENTS

- **Example:** If $\mathbf{f}(\theta) = \theta$, the integral $\mathbf{I} = \int_{\Theta} \theta \bar{\pi}(\theta) d\theta$ represents the MMSE estimator - the reason of this name required more hours of course...
- More generally, **all the moments of the posterior are:**

$$\mathbf{I}_k = \int_{\Theta} \theta^k \bar{\pi}(\theta) d\theta,$$

$k = 1, 2, 3...$

Model Selection - Marginal likelihood - again quadrature problem

- Marginal likelihood - Bayesian model evidence:

$$Z = p(\mathbf{y}) = \int_{\Theta} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

- $Z =$ **Weighted average of the likelihood values !!**
- Note that

$$0 \leq \min p(\mathbf{y}|\boldsymbol{\theta}) \leq Z \leq \max p(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y}|\widehat{\boldsymbol{\theta}}_{\mathtt{ML}}).$$

Goal

- Goal: extract and summarize the statistical information contained in the posterior pdf $\bar{\pi}(\boldsymbol{\theta})$ and compute Z.
- **The problem of extracting information from $\bar{\pi}(\boldsymbol{\theta})$ is *mainly* converted in a quadrature problem.**

Bayesian Inference:

1. "Main Actors" in Bayesian inference
2. Important considerations and consistency
3. Goals
4. **Levels of inference**
5. Type of Priors - choice of the priors
6. Reasons to be Bayesian

Levels of inference

In the standard/basic framework $(g(\boldsymbol{\theta}), \ell(\mathbf{y}|\boldsymbol{\theta}), Z)$:

- **Level 1:** Inference about $\boldsymbol{\theta}$.
- **Level 2:** Learning/obtaining Z.

different levels ==> different "rules", **in this sense different priors can be used or have different meanings...**

More levels: Hierarchical modeling

Hierarchical modeling: $h(\nu)$ prior over ν, $g(\theta|\nu)$ prior over θ

- **Level 0:** Inference about ν.
- **Level 1:** Inference about θ.
- **Level 2:** Learning/obtaining Z; in this case we have also several $Z|\nu$.

We can use Level 2 for learning ν or use a full-Bayesian solution. See Section of:

- F. Llorente, L. Martino, E. Cuberlo, J. Lopez-Santiago, D. Delgado, "On the safe use of prior densities for Bayesian model selection", viXra:2110.0032 , 2021.

Bayesian Inference:

1. "Main Actors" in Bayesian inference
2. Important considerations and consistency
3. Goals
4. Levels of inference
5. **Type of Priors - choice of the priors**
6. Reasons to be Bayesian

Type of priors !

- There are several type of priors. See as an example **Section 3.3**:

- F. Llorente, L. Martino, E. Cuberlo, J. Lopez-Santiago, D. Delgado, "On the safe use of prior densities for Bayesian model selection", viXra:2110.0032 , 2021.

Priors for Model Selection

- Marginal likelihood - Bayesian model evidence:

$$Z = p(\mathbf{y}) = \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

- Issues when the prior is improper...

- F. Llorente, L. Martino, E. Cuberlo, J. Lopez-Santiago, D. Delgado, "On the safe use of prior densities for Bayesian model selection", viXra:2110.0032 , 2021.

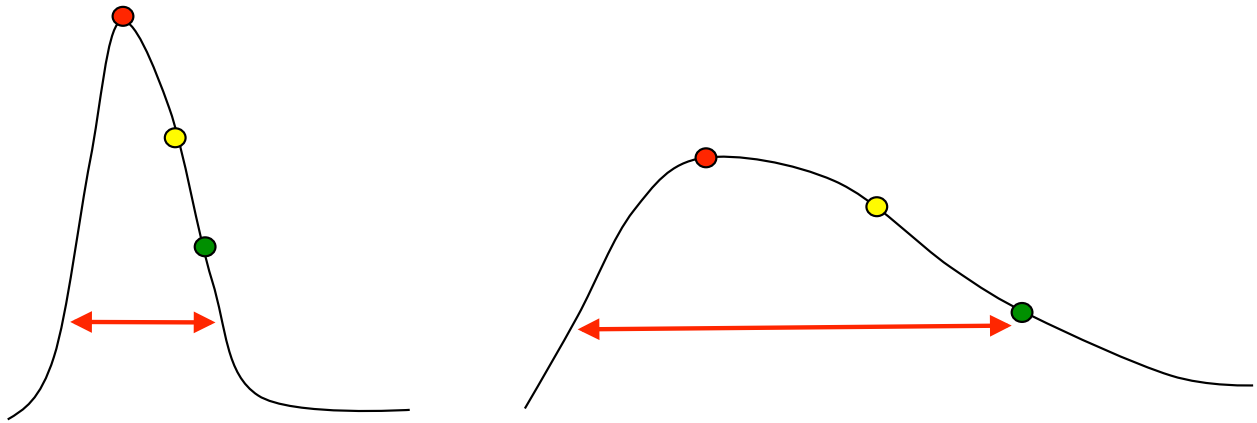
- F. Llorente, L. Martino, D. Delgado, J. Lopez-Santiago, "Marginal likelihood computation for model selection and hypothesis testing: an extensive review", (to appear) SIAM Review, 2022. arXiv:2005.08334

Bayesian Inference:

1. "Main Actors" in Bayesian inference
2. Important considerations and consistency
3. Goals
4. Levels of inference
5. Type of Priors - choice of the priors
6. **Reasons to be Bayesian**

Reasons to use a Bayesian approach

- include prior information in our model
- different possible point estimators (not only maximum...)



Reasons to use a Bayesian approach

- Provide complete posterior information:
  - **including information** by the prior density
  - **quantify uncertainties** (histograms)
  - credible intervals (areas)
  - quantiles (areas)
  - number of modes and modes
  - correlations among parameters (components of $\boldsymbol{\theta}$) (multi-dimensional histograms)
  - dependence/sensibility analysis of the model with respect to the components of $\boldsymbol{\theta}$ (related to the gradient/derivatives of the model/transformation and the variance of the histograms - marginal densities)

Reasons to use a Bayesian approach

- **easier application and interpretation of statistical quantities and procedures**: credible intervals (easier than confidence intervals), hypothesis testing, model selection etc... in my opinion, this is the main benefit with respect to the frequentist approach.
- regularization - numerical stability (due to the prior)
- **model selection** (marginal likelihood - Z)

Examples of $\boldsymbol{\theta}$

Inference about $\boldsymbol{\theta}$:

- $\boldsymbol{\theta}$ can be a vector of parameters of a model
- $\boldsymbol{\theta}$ can be a vector of hyper-parameters
- $\boldsymbol{\theta}$ can be a model index (model selection),
- $\boldsymbol{\theta}$ can include the number of parameters in a model (complexity of the model)
- More specifically, $\boldsymbol{\theta}$ can represent the position or the trayectory of a target, the volatility in a financial time series, velocity and direction of the wind etc.

- Preambule - or a very long introduction from a signal processing point of view...
- Bayesian Inference
- **Computational methods for Bayesian Inference**
- Monte Carlo sampling methods - brief overview

Computational methods for Bayesian Inference

- In many applications, we are not able to compute analytically $\mathbf{I}$ and Z.
- **Numerical approximations:**
  1. Deterministic quadrature rules - Gaussian-Hermite, Cubature rules...
  2. Monte Carlo techniques
  3. Variance Reduction - Quasi Monte Carlo (negative correlation)
  4. Variational inference techniques
  5. other modern quadrature rules

Computational methods for Bayesian Inference (2)

- The methods 1-2-3 of the previous list can considered **quadrature techniques**.

Computational methods for Bayesian Inference (2)

- The methods 1-2-3 of the previous list can be summarized with this formula:

$$\mathbf{I} = \frac{1}{Z}\int_{\Theta} \mathbf{f}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \approx \widehat{\mathbf{I}}_N = \sum_{n=1}^{N} \bar{w}_n \mathbf{f}(\boldsymbol{\theta}_n), \tag{4}$$

- **Error bounds:**

$$||\mathbf{I} - \widehat{\mathbf{I}}_N|| \leq V(\mathbf{f}) D_N(\boldsymbol{\theta}_{1:N}), \tag{5}$$

  - $V(\mathbf{f})$ depends on variation of $\mathbf{f}$ in Θ
  - $D_N(\boldsymbol{\theta}_{1:N})$ depends on choice of the nodes/samples $\boldsymbol{\theta}_{1:N}$
- **Clearly,** $D_N(\boldsymbol{\theta}_{1:N}) \to 0$ **when** $N \to \infty$

- Preambule - or a very long introduction from a signal processing point of view...
- Bayesian Inference
- Computational methods for Bayesian Inference
- **Monte Carlo sampling methods - brief overview**

Monte Carlo sampling methods

- They are **random number generators ==> from a generic density - given an available random source**
- that can be used for building/designing **stochastic quadrature rules in Bayesian inference.**

Important: Monte Carlo sampling methods are random number generators, they have life out of the "Bayesian world..." - but the main application is in Bayesian inference.

More than a quadrature rule and an optimizer...

with Monte Carlo:

- We can also optimize $\bar{\pi}(\boldsymbol{\theta})$ (or $\pi(\boldsymbol{\theta})$ is the same): global optimization, the "true" optimization !!
- But with optimization we just get one point - as I said before, we want to extract and summarize the statistical information contained in $\bar{\pi}(\boldsymbol{\theta})$.

SAMPLING >> OPTIMIZATION

- With a Monte Carlo sampling method, we have also an "optimizer".
- Optimization can be used for obtaining better "samplers".

Approximation of the "measure of the posterior"

- We obtain a particle approximation (with N samples)

$$\widehat{\pi}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \bar{w}_i \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}).$$

Standard Monte Carlo approximation

- If we can generate N i.i.d. random vectors $\boldsymbol{\theta}_n$ distributed according to $\bar{\pi}(\boldsymbol{\theta})$, $n = 1, \ldots, N$, then

$$\widehat{\mathbf{I}}_N = \frac{1}{N} \sum_{n=1}^{N} \mathbf{f}(\boldsymbol{\theta}_n) \approx \mathbf{I}, \qquad \boldsymbol{\theta}_n \sim \bar{\pi}(\boldsymbol{\theta}).$$

- However:
  - Often it is not possible to draw from $\bar{\pi}(\boldsymbol{\theta})$.
  - Even in this "ideal" case, it is not straightforward to approximate Z, i.e., to find $\widehat{Z} \approx Z$.

Sampling methods

If it is not possible to draw directly from the **target** pdf $\bar{\pi}(\boldsymbol{\theta})$:

- One can draw samples from a simpler **proposal pdf**, $q(\boldsymbol{\theta})$, and then *filter properly* these samples for obtaining an approximation of I and Z.

Sampling algorithm: all the steps corresponding to this "filtering operations".

$$\mathbf{z}_1, ..., \mathbf{z}_M \sim q(\boldsymbol{\theta}) \longrightarrow \boxed{\text{MC sampling}} \longrightarrow \widehat{\bar{\pi}}(\boldsymbol{\theta}) = \sum_{m=1}^{M} \bar{w}_m \delta(\boldsymbol{\theta} - \mathbf{z}_m)$$

$$M \geq N = \text{ effective number of samples}$$

$$\sum_{m=1}^{M} \bar{w}_m = 1 \Longrightarrow \begin{cases} \bar{w}_m \propto \{0,1\} & * \text{ accept/reject} \\ \bar{w}_m = \text{ by repetition} & * \text{ MCMC} \\ \bar{w}_m = \text{ generic} & * \text{ importance} \\ \bar{w}_m = \dfrac{1}{M} & * \text{ standard MC} \end{cases}$$

Sampling methods: classification

(STATIC SCENARIO)

4 main classes of algorithms:

1. Direct methods (based on random variable transformation).
   - **Independent samples.** (the best, almost)
   - computational effort: lowest.
   - applicability: low.
2. Rejection sampling.
   - **Independent samples.** (the best, almost)
   - computational effort: higher (depending on the acceptance rate).
   - applicability: wider of direct methods, but in general low.
3. Importance sampling (IS).
   - **Weighted samples.**
   - computational effort: low.
   - applicability: always. - **Easy approx of Z**
4. Markov chain Monte Carlo (MCMC).
   - **(positive) Correlated samples.**
   - computational effort: low.
   - applicability: always. - **Exploration of the space**

(the best scenario is: negative correlated samples =>, e.g., including deterministic "parts" within the method)

Sampling vs Variational/others approaches

Benefits:

- Applicability - flexible (easy to apply to different problems/framework) - the unique requirement is to be able to evaluate $\pi(\boldsymbol{\theta}) \propto \bar{\pi}(\boldsymbol{\theta})$; this condition can even relax.
- Complete approximation of the posterior that can be improved increasing the computational cost.

Drawbacks:

- "Slower" - more computational demanding (depending on the specific requirements of the considered application).

Speed up - Monte Carlo

To speed up:

- given the application, specific algorithm design:
  - better proposal choice
  - include more information of the posterior (e.g., gradient)
  - include determinism (when it is possible, in a proper way)

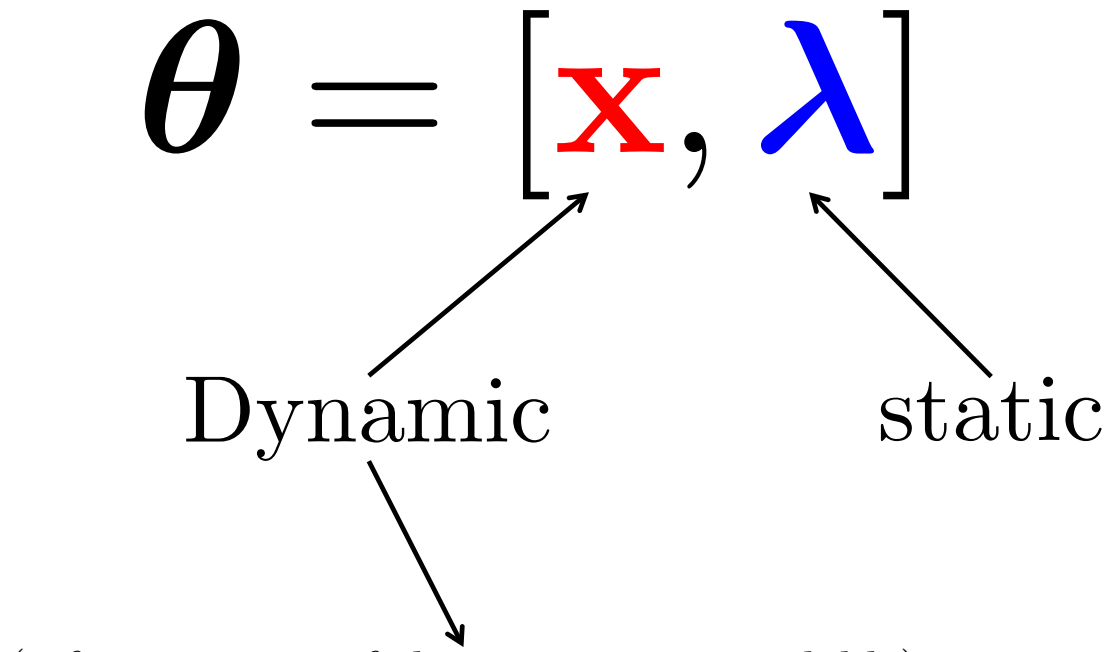
Black-box point of view:

- The performance strictly depends on the choice of $q(\boldsymbol{\theta})$.
- We desire $q(\boldsymbol{\theta}) \approx \bar{\pi}(\boldsymbol{\theta})$.

This is the reason for employing **adaptive techniques**.

- I worked (a lot) with adaptive MCMC and adaptive IS.

Static - dynamic parameters



$$\mathbf{x} = [x_1, ..., x_{d_x}] \in \mathbb{R}^{d_x},$$

$$\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_{d_\lambda}] \in \mathbb{R}^{d_\lambda},$$

$$d_\theta = d_x + d_\lambda.$$

Different frameworks - applications

Table: Scenarios where we able to evaluate $\pi(\boldsymbol{\theta}|\mathbf{y})$.

Computational Scenarios	**Monte Carlo approach**
$\pi(\boldsymbol{\lambda}\|\mathbf{y})$	static-batch scenario
$\pi(\lambda_1\|\mathbf{y}, \lambda_{2:d_\lambda} = \text{fixed})$, $\pi(\lambda_2\|\mathbf{y}, \lambda_1 = \text{fixed}, \lambda_{3:d_\lambda} = \text{fixed})$, $\pi(\lambda_3\|\mathbf{y}, \lambda_{1:2} = \text{fixed}, \lambda_{4:d_\lambda} = \text{fixed})$,...	static-batch scenario with **Gibbs approach**, component-wise approach
$\pi(\boldsymbol{\lambda}\|y_1)$,$\pi(\boldsymbol{\lambda}\|y_1, y_2)$, $\pi(\boldsymbol{\lambda}\|y_{1:\ell})$... $\pi(\boldsymbol{\lambda}\|y_{1:d_Y} = \mathbf{y})$.	**data tempering** or **online inference** (data streaming)
$\pi(\boldsymbol{\lambda}\|y_1, y_2, y_3)$,$\pi(\boldsymbol{\lambda}\|y_4)$, $\pi(\boldsymbol{\lambda}\|y_5, y_6)$...	**parallel - distributed - diffused estimation Big Data**
$\pi(x_1\|\mathbf{y})$,$\pi(x_1, x_2\|\mathbf{y})$, $\pi(x_1 x_2, x_3\|\mathbf{y})$...	the dimension of x's increases progressively (classifier and regressor chains)
$\pi(x_1\|y_1)$,$\pi(x_1, x_2\|y_1, y_2)$, $\pi(x_{1:3}\|y_{1:3})$... $\pi(x_{1:d_x} = \mathbf{x}\|y_{1:d_Y} = \mathbf{y})$.	**completely sequential scenario - HMM Kalman Filters; Particle Filters** state space models
$\pi(x_1, \boldsymbol{\lambda}\|y_1)$,$\pi(x_{1,2}, \boldsymbol{\lambda}\|y_{1,2})$, $\pi(x_{1:3}, \boldsymbol{\lambda}\|y_{1:3})$... $\pi(x_{1:d_x} = \mathbf{x}, \boldsymbol{\lambda}\|y_{1:d_Y} = \mathbf{y})$.	"Tracking and parameter estimation in state-space models"

Different frameworks - applications

Table: More "strange" scenarios; e.g., we cannot evaluate $\pi(\boldsymbol{\theta}|\mathbf{y})$.

Computational Scenarios	**Monte Carlo approach**
$Z_X(\boldsymbol{\theta}) = \int \ell(\mathbf{y}\|\boldsymbol{\theta})d\mathbf{y}$ unknown	methods for **"double intractable" posteriors** **pseudo-marginal MCMC**
costly likelihood, or impossible to evaluate the likelihood, or "too much" data	**Approximate Bayesian Computation (ABC)**, **pseudo-marginal MCMC**, **noisy MC** Monte Carlo for Big Data
unknown dimension d_θ of $\boldsymbol{\theta} = [\theta_1, ..., \theta_{d_\theta}]$	"tracking with unknown number of targets", "change point detection", inference also about d_θ, **Reversible Jump MCMC**, **Particle learning**
model selection	inference + choose the best model (related to the previous point)

- **Markov Chain Monte Carlo (MCMC)**

- L. Martino, V. Elvira. "Metropolis Sampling", Wiley StatsRef: Statistics Reference Online, 2017. arXiv:1704.04629

- L. Martino, "A Review of Multiple Try MCMC algorithms for Signal Processing", Digital Signal Processing, Volume 75, Pages: 134-152, 2018.

Markov Chain Monte Carlo (MCMC)

- Markov Chain Monte Carlo (MCMC) techniques yield an ergodic Markov chain

$$\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2 \to ....\boldsymbol{\theta}_t \to ....\boldsymbol{\theta}_{T-1} \to \boldsymbol{\theta}_T,$$

  with a stationary/invariant density, that is exactly the posterior $\bar{\pi}(\boldsymbol{\theta})$.

- There exists a $t_b < \infty$ (length of the **burn-in period**), such that

$$\boldsymbol{\theta}_t \sim \bar{\pi}(\boldsymbol{\theta}), \quad \text{for } t \geq t_b, \tag{6}$$

  i.e., the marginal pdf of $\boldsymbol{\theta}_t$ is the posterior.

Kernel of a MCMC algorithm

- An MCMC method is completely defined by the probability to obtain a new state θ_t given the previous one, θ_{t-1}.
- The corresponding conditional density $K(\theta_t|\theta_{t-1})$ is usually called **kernel**.
- $K(\theta_t|\theta_{t-1})$ summarizes all the steps of the MCMC algorithm.

Invariant/Stationary distribution

- Definition of invariant-stationary pdf $p_S(\mathbf{x}_t)$:

$$\int_{\Theta} K(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) p_S(\boldsymbol{\theta}_{t-1}) d\boldsymbol{\theta}_{t-1} = p_S(\boldsymbol{\theta}_t). \tag{7}$$

- **MCMC method:** design $K(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ in order to have

$$p_S(\boldsymbol{\theta}) = \bar{\pi}(\boldsymbol{\theta}). \tag{8}$$

Eigenfunctions

- This problem is related to the search of eigenvalues and eigenfunctions in the equation

$$\int_{\Theta} K(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})\phi(\boldsymbol{\theta}_{t-1})d\boldsymbol{\theta}_{t-1} = \mu\phi(\boldsymbol{\theta}_t) \tag{9}$$

where μ is an eigenvalue and $\phi(\cdot)$ is an eigenfunction (corresponding to μ).

Balance condition

- Use the definition of invariance is difficult, in general.
- The **balance condition**

$$K(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})\bar{\pi}(\boldsymbol{\theta}_{t-1}) = K(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t)\bar{\pi}(\boldsymbol{\theta}_t), \tag{10}$$

  is a sufficient condition to prove the invariance.
- If a density satisfies the balance condition, then is invariant w.r.t. the kernel $K(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$.
- In this case, the chain is **reversible**.

Metropolis-Hastings (MH) algorithm

- Recall $\pi(\boldsymbol{\theta}) \propto \bar{\pi}(\boldsymbol{\theta})$.
- Proposal pdf: $q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1})$.

MH algorithm:

- Choose $\boldsymbol{\theta}_0$.
- For $t = 1, ..., T$:
  1. Draw $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1})$.
  2. Set $\boldsymbol{\theta}_t = \boldsymbol{\theta}'$ with probability

$$\alpha = \min\left[1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}_{t-1})q(\boldsymbol{\theta}'|\boldsymbol{\theta}_{t-1})}\right]. \tag{11}$$

  Otherwise, set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$ (with probability $1 - \alpha$).

- Output: $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_T\}$.

Effective Sample Size (ESS) for MCMC

- The samples are **(positive) correlated**.
- Due to the ergodicity:

$$\widetilde{\mathbf{I}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}(\boldsymbol{\theta}_t) \approx \mathbf{I}. \tag{12}$$

(recall that we should consider only $t \geq t_b$)

- **Effective Sample Size (ESS):**

$$T_{eff} = T\frac{\text{var}\left[\widehat{\mathbf{I}}_T\right]}{\text{var}\left[\widetilde{\mathbf{I}}_T\right]} \approx \frac{T}{1+2\sum_{k=1}^{\infty}\rho_k}, \tag{13}$$

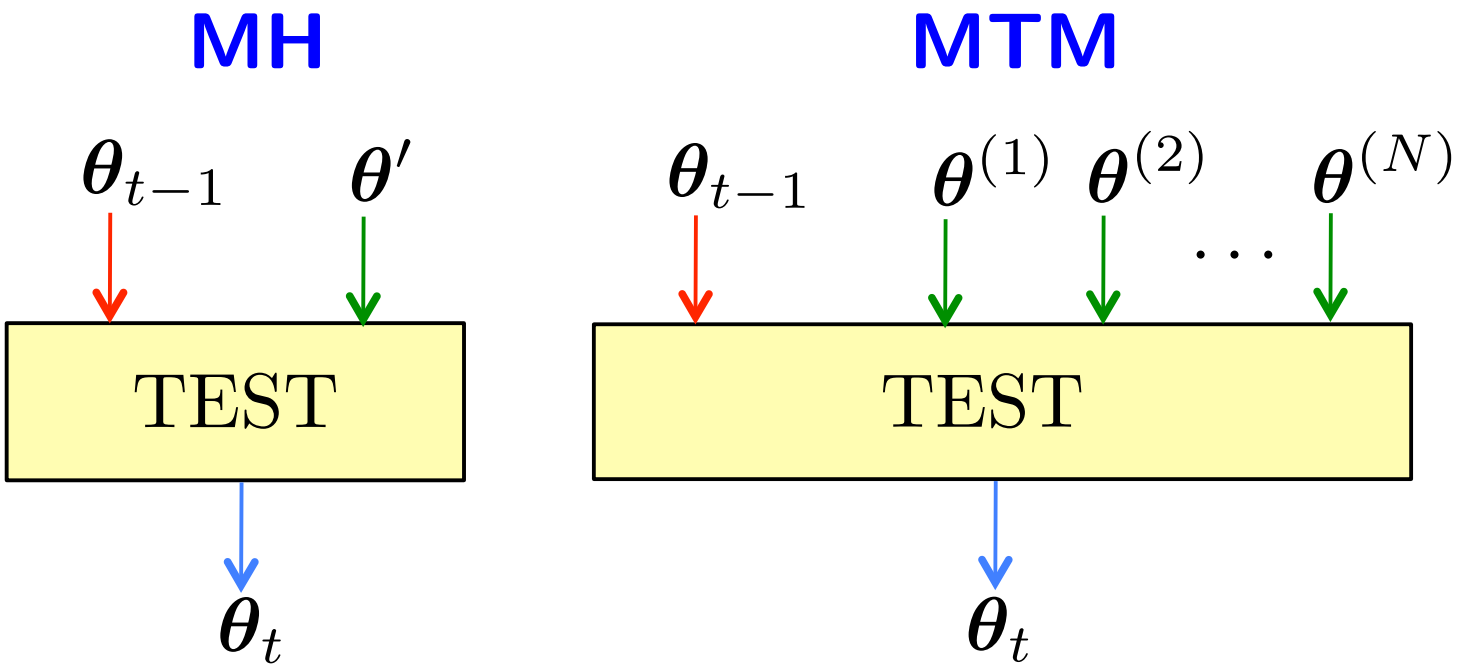
where $\rho_k = \frac{\text{cov}[f(\boldsymbol{\theta}_t), f(\boldsymbol{\theta}_{t+k})]}{\text{var}[f(\boldsymbol{\theta}_t)]}$.

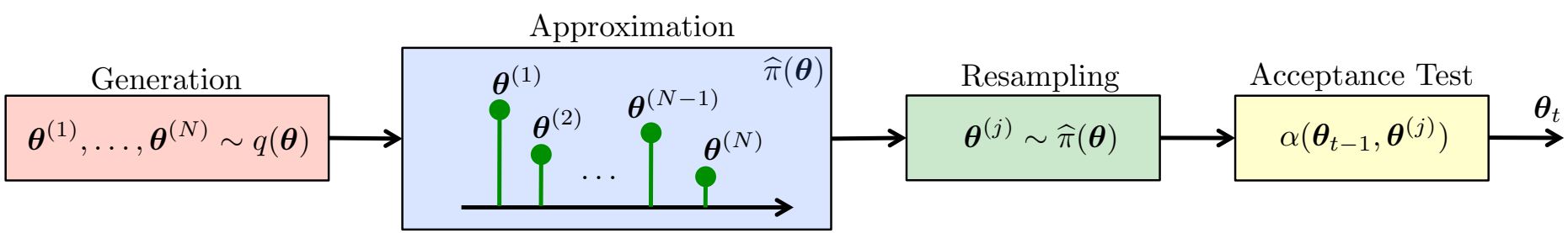
Improve performance

To reduce the correlation, speed up the convergence:

- MCMC with and adaptive proposal density
- Adding gradient information to the proposal pdf - Hamiltonian Monte Carlo
- Design more efficient algorithms: Multiple Try Metropolis (MTM)
- In high dimension, work component by component - Gibbs sampling
- Design non-reversible MCMC methods.

Multiple Try Metropolis (MTM)





- **Importance Sampling (IS)**

- V. Elvira, L. Martino, D. Luengo, M. F. Bugallo, "Generalized Multiple Importance Sampling", Statistical Science, Volume 34, Number 1, Pages 129-155, 2019.
- L. Martino, V. Elvira, F. Louzada, "Effective Sample Size for Importance Sampling Based on Discrepancy Measures", Signal Processing, Volume 131, Pages: 386-401, 2017

Importance Sampling (IS)

- Consider the following equality:

$$\begin{aligned} \mathbf{I} = E_{\bar{\pi}}[\mathbf{f}(\boldsymbol{\theta})] &= \int_{\boldsymbol{\Theta}} \mathbf{f}(\boldsymbol{\theta})\bar{\pi}(\boldsymbol{\theta})d\boldsymbol{\theta}, \\ &= \frac{1}{Z}\int_{\boldsymbol{\Theta}} \mathbf{f}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \\ &= \frac{1}{Z}\int_{\boldsymbol{\Theta}} \mathbf{f}(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}q(\boldsymbol{\theta})d\boldsymbol{\theta}, \\ &= \frac{1}{Z}\int_{\boldsymbol{\Theta}} \mathbf{f}(\boldsymbol{\theta})w(\boldsymbol{\theta})q(\boldsymbol{\theta})d\boldsymbol{\theta}, \\ &= \frac{1}{Z}E_q[\mathbf{f}(\boldsymbol{\theta})w(\boldsymbol{\theta})] \end{aligned}$$

  where $w(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}$.

- **Importance Sampling:** apply the standard MC procedure for approximating $E_q[\mathbf{f}(\boldsymbol{\theta})w(\boldsymbol{\theta})]$ (when Z is known).

IS ESTIMATORS (WITH UNKNOWN Z)

1. **Sampling:** N samples from the proposal $q(\boldsymbol{\theta})$

$$\boldsymbol{\theta}_n \sim q(\boldsymbol{\theta}), \qquad n = 1, ..., N.$$

2. **Weighting:** Each sample is "corrected" by the importance weight

$$w_n = \frac{\pi(\boldsymbol{\theta}_n)}{q(\boldsymbol{\theta}_n)}, \qquad n = 1, ..., N.$$

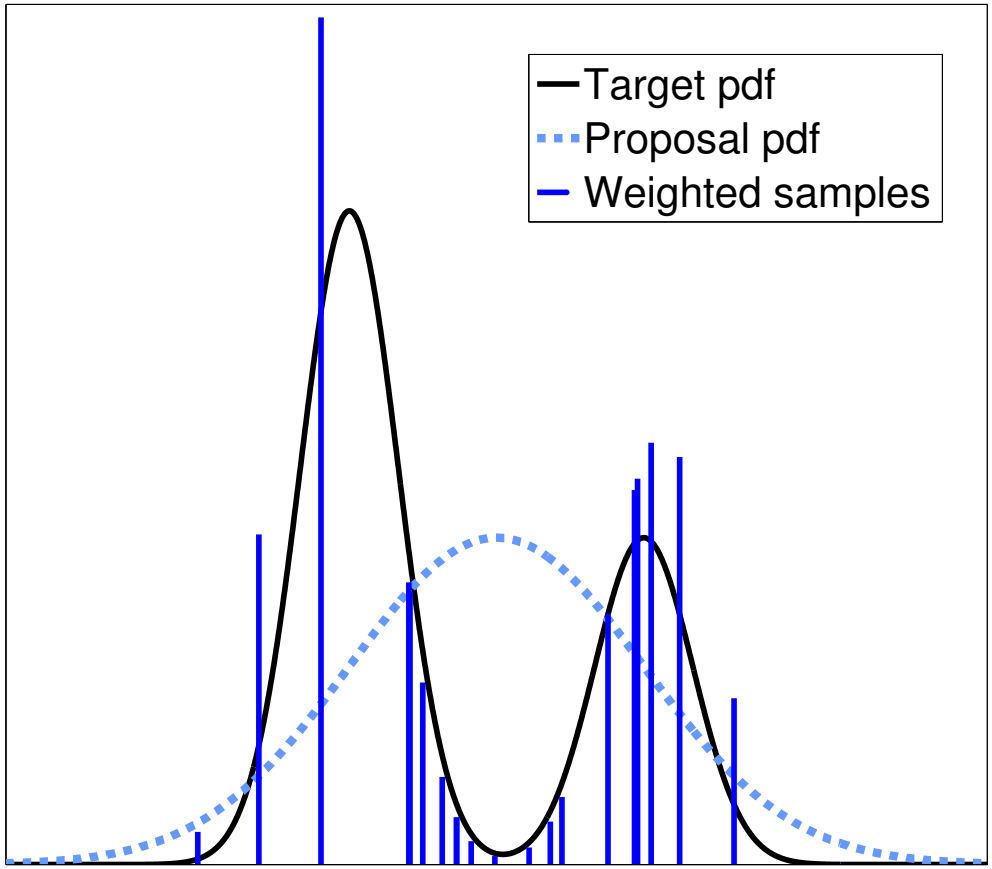
3. **Estimators**:

$$\widetilde{\mathbf{I}}_N = \sum_{n=1}^{N} \bar{w}_n \mathbf{f}(\boldsymbol{\theta}_n), \qquad \widehat{Z} = \frac{1}{N} \sum_{n=1}^{N} w_n,$$

where

$$\bar{w}_n = \frac{w_n}{\sum_{i=1}^{N} w_i} = \frac{w_n}{N\widehat{Z}}.$$

Example - IS



Proper weighting

- Is the previous weighting scheme unique? No.
- Consider an extended proposal pdf $q_e(\boldsymbol{\theta}, w) = q(w|\boldsymbol{\theta})q(\boldsymbol{\theta})$.
- **Properly weighted samples** with respect to $\bar{\pi}$:

$$E_{q_e}[W(\boldsymbol{\theta})\ \mathbf{f}(\boldsymbol{\theta})] = cE_{\bar{\pi}}[\ \mathbf{f}(\boldsymbol{\theta})], \tag{14}$$

  where $c > 0$ is a constant value.
- Different possible weighting schemes.
- Easy to see when different proposal densities are used jointly.

Multiple Importance Sampling (MIS)

Consider N proposal densities, $q_1(\boldsymbol{\theta}), ..., q_N(\boldsymbol{\theta})$.

- **Sampling:** $\boldsymbol{\theta}_n \sim q_n(\boldsymbol{\theta})$ with $n = 1, ..., N$.
- **Classical Weighting (CW):**

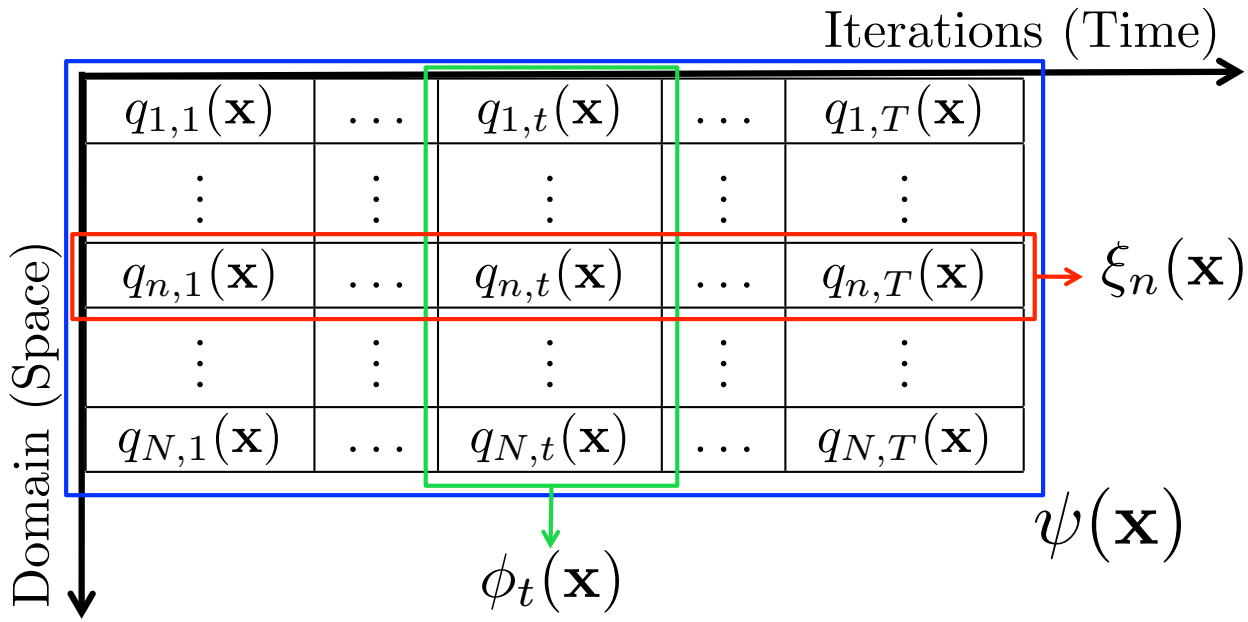
$$w_n = \frac{\pi(\boldsymbol{\theta}_n)}{q(\boldsymbol{\theta}_n)}, \qquad n = 1, ..., N.$$

- **Deterministic Mixture (DM) Weighting:**

$$w_n = \frac{\pi(\boldsymbol{\theta}_n)}{\frac{1}{N}\sum_{k=1}^{N} q_k(\boldsymbol{\theta}_n)}, \qquad n = 1, ..., N.$$

- The DM-IS estimators have lower variance than the CW-IS estimators (but more costly; a bit).
- There are even more possibilities: for instance, the partial DM weights.

Adaptation and MIS



- Trade-off: complexity - performance

- **Particle Filtering**

- L. Martino, J. Read, V. Elvira, F. Louzada, "Cooperative Parallel Particle Filters for on-Line Model Selection and Applications to Urban Mobility" Digital Signal Processing Volume 60, Pages: 172-185, 2017.

Particle filtering

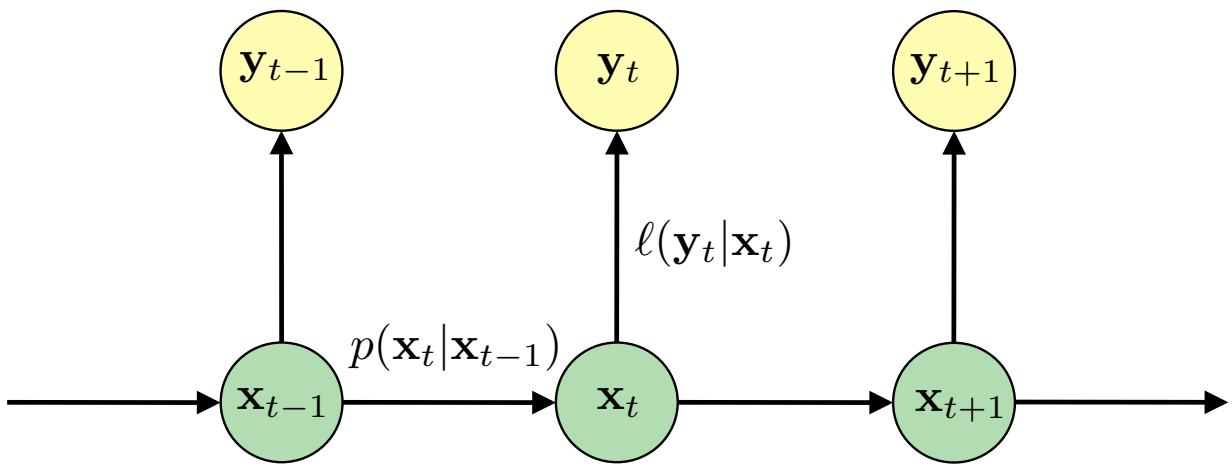
- **Particle Filtering = Sequential Importance Sampling + Resampling**
- In signal processing, mainly used in state-space models.

State-Space models

- $t \in \mathbb{N}$: discrete iteration index,
- $\mathbf{x}_t \in \mathbb{R}^{d_x}$: state variable that we want to infer,
- $\mathbf{y}_t \in \mathbb{R}^{d_y}$: observation at time t,

$$\mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{x}_{t-1}), \quad (\textbf{propagation}), \tag{15}$$

$$\mathbf{y}_t \sim \ell(\mathbf{y}_t | \mathbf{x}_t), \quad (\textbf{likelihood}). \tag{16}$$



Relationships with other methods

$\mathbf{x}_t \rightarrow$ continuous

$p, \ell \rightarrow$ both Gaussian

$\downarrow$

Kalman

$\mathbf{x}_t \rightarrow$ discrete

$\downarrow$

Hidden Markov Models (HMMs)

$\mathbf{x}_t \rightarrow$ continuous

$p, \ell \rightarrow$ generic

$\downarrow$

Particle Filtering

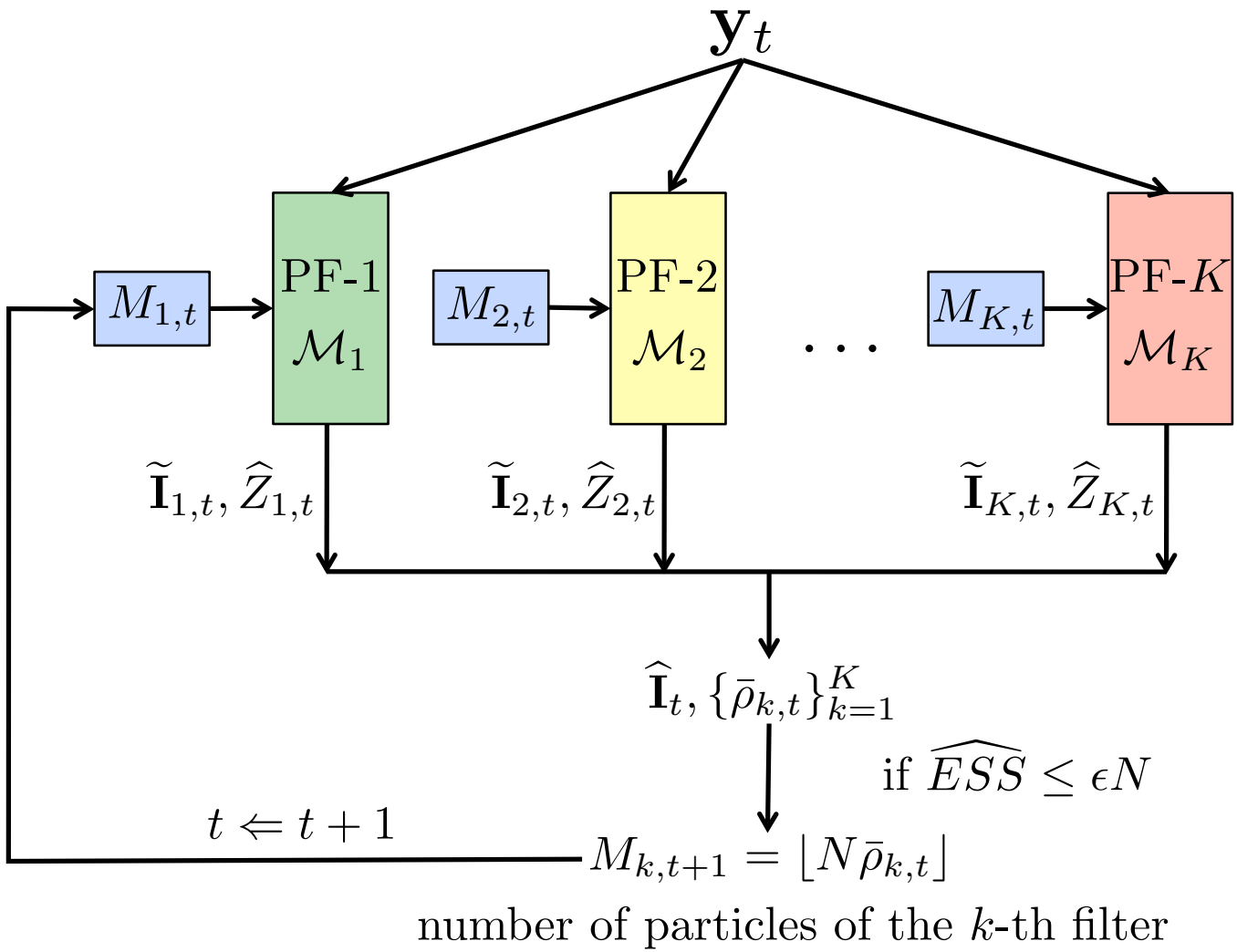
State-Space models with unknown parameter

$$\mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{x}_{t-1}, \boldsymbol{\lambda}_p), \quad (\textbf{propagation}), \tag{17}$$

$$\mathbf{y}_t \sim \ell(\mathbf{y}_t | \mathbf{x}_t, \boldsymbol{\lambda}_\ell), \quad (\textbf{likelihood}). \tag{18}$$

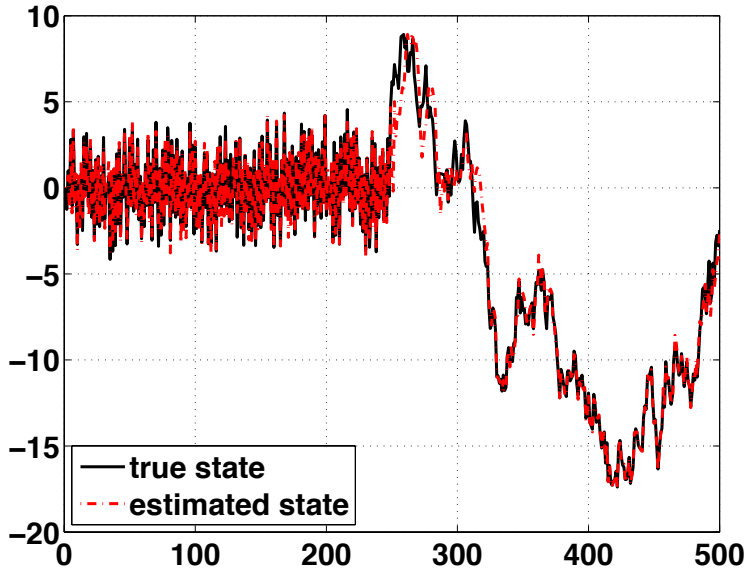
We can also select the best state-space model within a set of possible models.

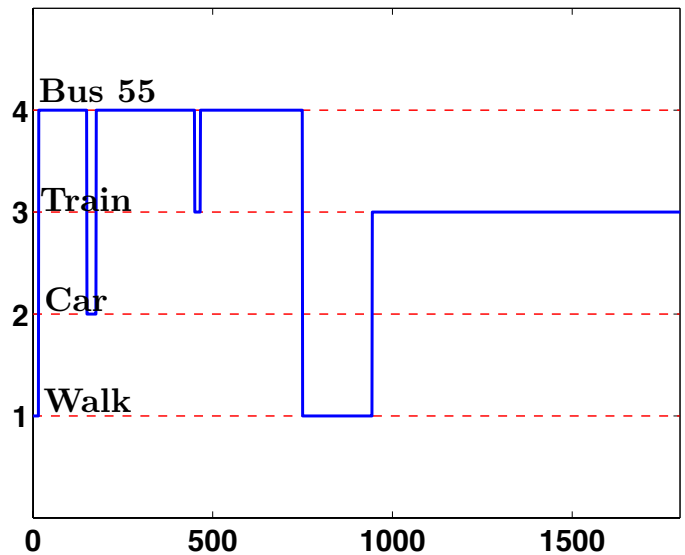
Cooperative Parallel Particle Filters



We also adapt the number of particles in each filter (but the sum of all particles is fixed)

Cooperative Parallel Particle Filters





- Thank you very much!
- Any questions?