Automatic tempered posterior distributions for Bayesian inversion problems

L. Martino[†], J. López-Santiago[*], J. Míguez[*]

[†] Universidad rey Juan Carlos (URJC), Madrid, Spain.

[*] Universidad Carlos III de Madrd (UC3M), Madrid, Spain.

September 5, 2020

Abstract

We propose a novel adaptive importance sampling scheme for Bayesian inversion problems where the inference of the variables of interest and the power of the data noise is split. More specifically, we consider a Bayesian analysis for the variables of interest (i.e., the parameters of the model to invert), whereas we employ a maximum likelihood approach for the estimation of the noise power. The whole technique is implemented by means of an iterative procedure, alternating sampling and optimization steps. Moreover, the noise power is also used as a tempered parameter for the posterior distribution of the the variables of interest. Therefore, a sequence of tempered posterior densities is generated, where the tempered parameter is automatically selected according to the actual estimation of the noise power. Numerical experiments show the benefits of the proposed approach.

1 Introduction

The estimation of unknown parameters from noisy observations is an essential problem in signal processing, statistics and machine learning [5, 2, 13, 4]. Within the Bayesian signal processing framework, these problems are addressed by constructing posterior probability distributions of the unknowns. Given the posterior, one often wants to make inference about the unknowns, e.g., if we are estimating parameters, finding the values that maximize their posterior, or the values that minimize some cost function given the uncertainty of the parameters. Unfortunately, obtaining closed-form solutions, usually expressed as integrals of the posterior, is infeasible in most practical applications. Hence, developing approximate computational techniques (such as importance sampling and MCMC algorithms) are often required [17, 9, 12].

The so-called *tempering of the posterior* is a well-known procedure for improving the performance of the Monte Carlo (MC) algorithms [8, 11, 6, 14]. The tempering is obtained by modulating an artificial scale parameter or by sequentially including new data. The reasons of the improvement in the performance are several: improving mixing, discovering modes, foster the exploration of the inference

space etc. In the first iterations of the MC scheme, a posterior density with a bigger scale is considered. The artificial scale parameter (often called "temperature") is reduced during the iterations, until considering the true posterior distribution. However, the user should decide a *temperature schedule*, i.e., a decreasing rule for the scale parameter, which is usually chosen in an heuristic way. In the literature, the tempering procedure has gained a particular attention for the estimation of the marginal likelihood (a.k.a., Bayesian model evidence) [6, 15, 10].

In this work, we design an adaptive importance sampling (AIS) scheme for Bayesian inversion problems, where an automatic tempering procedure is implemented. We consider that the vector of observations $\mathbf{y}$ is obtained by a nonlinear transformation $\mathbf{f}(\boldsymbol{\theta})$ of the variables of interest $\boldsymbol{\theta}$, perturbed by additive Gaussian noise with unknown power σ^2. The nonlinear mapping $\mathbf{f}(\boldsymbol{\theta})$ usually represents a complex physical model or a computer code etc. The resulting posterior densities are usually highly multimodal and complex distributions. Furthermore, the inference task in the a joint space $[\boldsymbol{\theta}, \sigma]$ is particularly challenging. Indeed, "wrong choices" of σ values can easily jeopardize the sampling of $\boldsymbol{\theta}$. We proposed a split strategy to tackle this problem. We consider an optimization approach over σ and a sampling scheme for $\boldsymbol{\theta}$. More specifically, we design an iterative procedure where this two tasks are alternated. Additionally, the actual maximum likelihood (ML) estimation of the noise power, $\widehat{\sigma}^2_{\texttt{ML}}$, is employed as a tempering parameter, starting from high values and then "cooling down" according to the ML estimations at each iterations. Therefore, the proposed scheme deals with a sequence of tempered posteriors according to the current estimation $\widehat{\sigma}^2_{\texttt{ML}}$. It is important to observe that, given a fixed vector $\boldsymbol{\theta}$, the ML estimation $\widehat{\sigma}^2_{\texttt{ML}}$ can be obtained analytically. The advantages of the proposed scheme are shown in two numerical experiments, one of them considering a complex astronomical model.

2 Problem Statement

Let denote the observed measurements as $\mathbf{y} = [y_1, ..., y_K]^\top \in \mathbb{R}^K$, and the variable of interest that we desire to infer, as $\boldsymbol{\theta} = [\theta_1, ..., \theta_M]^\top \in \boldsymbol{\Theta} \subseteq \mathbb{R}^M$. Furthermore, let consider the observation model

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}) + \mathbf{e}, \tag{1}$$

where we have a nonlinear mapping,

$$\mathbf{f}(\boldsymbol{\theta}) = [f_1(\boldsymbol{\theta}), ..., f_K(\boldsymbol{\theta})]^\top : \boldsymbol{\Theta} \subseteq \mathbb{R}^M \to \mathbb{R}^K, \tag{2}$$

and a Gaussian perturbation noise,

$$\mathbf{e} = [e_1, ..., e_K]^\top \sim \mathcal{N}(\mathbf{e}|\mathbf{0}, \sigma^2\mathbf{I}_K), \tag{3}$$

with $\sigma > 0$, and we have denoted the K-dimensional unit matrix as $\mathbf{I}_K$. The noise variance σ^2 is unknown, in general. The mapping $\mathbf{f}$ could be analytically unknown: the only assumption is that we are able to evaluate pointwise the nonlinear mapping $\mathbf{f}(\boldsymbol{\theta})$. The likelihood function is

$$\ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) = \frac{1}{(2\pi\sigma^2)^{K/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2\right), \tag{4}$$

$$= \frac{1}{(2\pi\sigma^2)^{K/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{K}(y_k - f_k(\boldsymbol{\theta}))^2\right). \tag{5}$$

Note that we have two types of variables of interest: the vector $\boldsymbol{\theta}$ contains the parameters of the nonlinear mapping $\mathbf{f}(\boldsymbol{\theta})$, whereas σ is a scale parameter of the likelihood function.

Goal. Given the vector of measurements $\mathbf{y}$, we desire to make infer regarding the hidden parameters $\boldsymbol{\theta}$ and the noise power σ^2, obtaining at least a point estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}^2$. Note also that the vector

$$\mathbf{r} = \mathbf{f}(\boldsymbol{\theta}) \in \mathbb{R}^K, \tag{6}$$

is a multivariate random variable obtained by the transformation of the random vector $\boldsymbol{\theta}$ trough the nonlinear mapping $\mathbf{f}$. Hence, an additional possible outcome is to obtain an smoothing version of the given observation vector $\mathbf{y} \in \mathbb{R}^K$, i.e., $\widehat{\mathbf{r}} \in \mathbb{R}^K$ (as well as uncertainty and correlation analysis between different y_k's). Finally, we are also interested in design efficient schemes in order to perform model selection, i.e., to compare, select or properly average different models.

Bayesian inference. We consider prior densities $g_\theta(\boldsymbol{\theta})$ and $g_\sigma(\sigma)$ over the unknowns. Hence, the complete posterior density is

$$\bar{\pi}(\boldsymbol{\theta}, \sigma|\mathbf{y}) = \frac{1}{p(\mathbf{y})}\pi(\boldsymbol{\theta}, \sigma|\mathbf{y}) = \frac{1}{p(\mathbf{y})}\ell(\mathbf{y}|\boldsymbol{\theta}, \sigma)g_\theta(\boldsymbol{\theta})g_\sigma(\sigma), \tag{7}$$

where $\pi(\boldsymbol{\theta}, \sigma|\mathbf{y}) = \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma)g_\theta(\boldsymbol{\theta})g_\sigma(\sigma)$ and note that $\bar{\pi}(\boldsymbol{\theta}, \sigma|\mathbf{y}) \propto \pi(\boldsymbol{\theta}, \sigma|\mathbf{y})$. The marginal likelihood $Z = p(\mathbf{y})$ is

$$Z = p(\mathbf{y}) = \int_{\mathbb{R}^+}\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}, \sigma|\mathbf{y})d\boldsymbol{\theta}d\sigma, \tag{8}$$

This quantity is often needed for model selection. Since $Z(\mathbf{y})$ is generally unknown, we can usually evaluate pointwise the unnormalized posterior $\pi(\boldsymbol{\theta}, \sigma|\mathbf{y})$. From now on, we remove the dependence on $\mathbf{y}$ to simplify the notation, using $\bar{\pi}(\boldsymbol{\theta}, \sigma)$, $\bar{\pi}(\boldsymbol{\theta}, \sigma)$, and Z. More generally, the computation of integrals of the form

$$I = \int_{\mathbb{R}^+}\int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta}, \sigma)\bar{\pi}(\boldsymbol{\theta}, \sigma)d\boldsymbol{\theta}d\sigma, \tag{9}$$

where $h(\cdot) : \boldsymbol{\Theta} \times \mathbb{R}^+ \to \mathbb{R}$ is an integrable function, is usually required. We consider a Monte Carlo quadrature approach for approximating the integral above and, more generally, provide a particle approximation of the posterior $\bar{\pi}(\boldsymbol{\theta}, \sigma|\mathbf{y})$.

Problem. Generally, generating efficiently samples from a complicated posterior in Eq. (7) and computing efficiently the integrals as in Eqs. (8)-(9) is an hard task. Moreover, this task becomes often more difficult when we try to perform a joint inference where are involved scale parameters, i.e., σ, and parameters of the nonlinearity, i.e., $\boldsymbol{\theta}$. Indeed, "wrong choices" of σ values can easily jeopardize the sampling of $\boldsymbol{\theta}$. Below, we describe the strategy that we propose to tackle this problem.

Conditional and marginal posteriors. In this scenario, the conditional posteriors are often con-

sidered, for instance,

$$\bar{\pi}_{\theta|\sigma}(\boldsymbol{\theta}|\sigma) = p(\boldsymbol{\theta}|\mathbf{y},\sigma) = \frac{p(\boldsymbol{\theta},\mathbf{y},\sigma)}{p(\mathbf{y},\sigma)} = \frac{\ell(\mathbf{y}|\boldsymbol{\theta},\sigma), g_\theta(\boldsymbol{\theta})g_\sigma(\sigma)}{p(\mathbf{y}|\sigma)g_\sigma(\sigma)}, \\ = \frac{\ell(\mathbf{y}|\sigma,\boldsymbol{\theta})g_\theta(\boldsymbol{\theta})}{p(\mathbf{y}|\sigma)}, \tag{10}$$

and the other one is $\bar{\pi}_{\sigma|\theta}(\sigma|\boldsymbol{\theta}) = \frac{\ell(\mathbf{y}|\sigma,\boldsymbol{\theta})g_\sigma(\sigma)}{p(\mathbf{y}|\boldsymbol{\theta})}$. The conditional marginal likelihood is obtained by integrating out one of the two variables, i.e.,

$$Z(\sigma) = p(\mathbf{y}|\sigma) = \int_{\mathcal{D}_\theta} \ell(\mathbf{y}|\boldsymbol{\theta},\sigma)g_\theta(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{11}$$

For computational reasons, or since we could be interested only in one of the two variables, we can consider the marginal posteriors defined as

$$\bar{\pi}_\sigma(\sigma) = p(\sigma|\mathbf{y}) = \frac{p(\mathbf{y}|\sigma)g_\sigma(\sigma)}{p(\mathbf{y})}, \tag{12}$$

$$\bar{\pi}_\theta(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})g_\theta(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{13}$$

where $p(\mathbf{y}|\boldsymbol{\theta}) = \int_{\mathbb{R}^+} \ell(\mathbf{y}|\boldsymbol{\theta},\sigma)g_\sigma(\sigma)d\sigma$. Generally, the integral (11) cannot be computed and an approximation $\widehat{Z}(\sigma) = \widehat{p}(\mathbf{y}|\sigma)$ is required. Finally note that the relationship among complete, conditional and marginal posteriors is give by

$$\bar{\pi}(\boldsymbol{\theta},\sigma) = p(\boldsymbol{\theta},\sigma|\mathbf{y}) = p(\boldsymbol{\theta}|\mathbf{y},\sigma)p(\sigma|\mathbf{y}), \quad \text{i.e.,} \quad \bar{\pi}(\boldsymbol{\theta},\sigma) = \bar{\pi}_{\theta|\sigma}(\boldsymbol{\theta}|\sigma)\bar{\pi}_\sigma(\sigma). \tag{14}$$

3 Suggested approach

The idea underlying the proposed scheme is to split the space $[\boldsymbol{\theta},\sigma]$, restricting the sampling problem only with respect to $\boldsymbol{\theta}$ and considering an optimization problem for with respect to σ. For the sake of simplicity, let us consider an improper uniform prior over $\boldsymbol{\theta}$. The proposed scheme described in the next section obtains the following three aims:

1. **MAP and ML estimation.** Given an approximation of the maximum likelihood (ML) estimator $\widehat{\sigma}^{(t)}_{\texttt{ML}}$ (at the t-th iteration of the proposed scheme), we also provide an approximation of the conditional maximum a-posteriori (MAP) estimator

$$\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}|\widehat{\sigma}^{(t)}_{\texttt{ML}} \approx \arg\max_{\boldsymbol{\theta}} \bar{\pi}_{\theta|\sigma}(\boldsymbol{\theta}|\widehat{\sigma}^{(t)}_{\texttt{ML}}). \tag{15}$$

   An important consideration is that, if $\widehat{\sigma}^{(t)}_{\texttt{ML}} \to \widehat{\sigma}_{\texttt{ML}}$ as t grows, then $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}|\widehat{\sigma}^{(t)}_{\texttt{ML}} \to \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$, where

$$\widehat{\boldsymbol{\theta}}_{\texttt{MAP}} \approx \arg\max_{\boldsymbol{\theta}} \bar{\pi}_{\theta|\sigma}(\boldsymbol{\theta}|\widehat{\sigma}_{\texttt{ML}}) = \arg\max_{\boldsymbol{\theta}} \bar{\pi}(\boldsymbol{\theta},\widehat{\sigma}_{\texttt{ML}}).$$

2. **Particle approximation.** We approximate the measure of the conditional posterior density after T iterations, i.e.,

$$\bar{\pi}_{\theta|\sigma}(\boldsymbol{\theta}|\widehat{\sigma}^{(T)}_{\texttt{ML}}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta}, \widehat{\sigma}^{(T)}_{\texttt{ML}}) g_{\theta}(\boldsymbol{\theta})}{p(\mathbf{y}|\widehat{\sigma}^{(T)}_{\texttt{ML}})}, \tag{16}$$

by a set of weighted samples $\{\boldsymbol{\theta}^{(n)}_t, \widetilde{w}^{(n)}_t\}$ for $t = 1, ..., T$ and $n = 1, ..., N$, i.e.,

$$\widehat{\pi}_{\theta|\sigma}(\boldsymbol{\theta}|\widehat{\sigma}^{(T)}_{\texttt{ML}}) = \sum_{t=1}^{T} \sum_{n=1}^{N} \widetilde{w}^{(n)}_t \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{(n)}_t).$$

3. **Marginal likelihood estimator.** We suggest to ways for estimating of the marginal likelihood $Z = p(\mathbf{y})$. One possible estimator is

$$\widehat{Z} = \widehat{Z}(\widehat{\sigma}_{\texttt{ML}}) \approx p(\mathbf{y}|\widehat{\sigma}_{\texttt{ML}}). \tag{17}$$

Other more sophisticated procedure is to approximate the integral

$$\widehat{Z} \approx Z = \int_{\mathbb{R}^+} p(\mathbf{y}|\sigma) g_{\sigma}(\sigma) d\sigma \approx \int_{\mathbb{R}^+} \widehat{Z}(\sigma) g_{\sigma}(\sigma) d\sigma, \tag{18}$$

where $\widehat{Z}(\sigma) = \widehat{p}(\mathbf{y}|\sigma)$ is an estimator for each possible value of σ (often called conditional marginal likelihood).

These three objectives are obtained by an iterative procedure. Thus, the resulting schemes are adaptive Monte Carlo algorithms which combines sampling schemes ad stochastic optimization. However, some part the conditional posterior of σ can be analytically maximized as shown below (jointly with some important considerations).

3.1 Analysis of the conditional posterior of σ^2

Note that the conditional marginal posterior with respect to σ is

$$\ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) \propto \frac{1}{\sigma^K} \exp\left(-\frac{\|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2}{2\sigma^2}\right) \tag{19}$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{K}{2}} \exp\left(-\frac{\|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2}{2\sigma^2}\right), \tag{20}$$

that, with respect to σ^2, has the form of an *Inverse Gamma* density. Then, we can focus on $\bar{\pi}_{\sigma^2|\theta}(\sigma^2|\boldsymbol{\theta})$ since this pdf can be studied analytically. For instance, it has a *unique* mode (maximum likelihood) at

$$\sigma^2_{\texttt{ML}}|\boldsymbol{\theta} = \frac{1}{K}\|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2, \tag{21}$$

where we have remarked that the expression above represents a MAP estimator *conditioned* to a specific value of $\boldsymbol{\theta}$. For $K > 4$, we can also obtain that

$$\text{var}(\sigma^2|\boldsymbol{\theta}) \propto \frac{1}{K^2}\|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^4.$$

The expression above shows that if Monte Carlo methods (MCMC or IS) is exploring a region where $\boldsymbol{\theta}$ does not provide a good fit with the data $\mathbf{y}$ (through the model $\mathbf{f}$) then the variance is proportional to the square of the MSE, i.e., it can be huge. This intuitively explains the issue of dealing with the joint sampling of $\boldsymbol{\theta}$ and σ.

Improper uniform prior over $\boldsymbol{\theta}$. Let us consider here an improper uniform prior over $\boldsymbol{\theta}$. In this scenario, we have $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}} = \widehat{\boldsymbol{\theta}}_{\texttt{ML}}$ which can be expressed as

$$\widehat{\boldsymbol{\theta}}_{\texttt{ML}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2.$$

As a consequence, $\widehat{\sigma}_{\texttt{ML}}$ and $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ are related as

$$\widehat{\sigma}_{\texttt{ML}} = \sqrt{\frac{1}{K}\|\mathbf{y} - \mathbf{f}(\widehat{\boldsymbol{\theta}}_{\texttt{MAP}})\|^2}, \tag{22}$$

where $\widehat{\sigma}_{\texttt{ML}}$ is the global ML estimator of σ. See Appendix A for further details.

4 Automatic Tempering Adaptive Importance Sampling (ATAIS)

In this section, we describe an adaptive importance sampler with an *automatic tempering* approach which follows the suggestions previously described. At each iteration t of the algorithm, we have an ML approximation of σ, i.e., $\widehat{\sigma}^{(t)}_{\texttt{ML}}$. Considering Eq. (10), we define the *tempered conditional posterior* at the t-th iteration,

$$\bar{\pi}_t(\boldsymbol{\theta}) \propto \pi_t(\boldsymbol{\theta}) = \pi_{\theta|\sigma}(\boldsymbol{\theta}|\widehat{\sigma}^{(t-1)}_{\texttt{ML}}) = \ell(\mathbf{y}|\boldsymbol{\theta}, \widehat{\sigma}^{(t-1)}_{\texttt{ML}}) g_\theta(\boldsymbol{\theta}), \tag{30}$$

At each iteration, we consider $\bar{\pi}_t(\boldsymbol{\theta}) \propto \pi_t(\boldsymbol{\theta})$ as a target distribution. The dependence on the iteration t is due to $\widehat{\sigma}^{(t)}_{\texttt{ML}}$ varies with t. The ATAIS algorithm is outlined in Table 1. Table 2 contains further details. It is important to remark that, if $\widehat{\sigma}^{(0)}_{\texttt{ML}}$ is bigger of the true ML value, we generate a non-increasing sequence of $\widehat{\sigma}^{(t)}_{\texttt{ML}}$, i.e., $\widehat{\sigma}^{(0)}_{\texttt{ML}} \geq \widehat{\sigma}^{(1)}_{\texttt{ML}} \geq ... \widehat{\sigma}^{(t)}_{\texttt{ML}} \geq \widehat{\sigma}^{(t+1)}_{\texttt{ML}}$ etc.

IS steps. A set of N samples $\{\boldsymbol{\theta}^{(n)}_t\}_{n=1}^N$ are drawn from a (normalized) proposal density $q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ with mean $\boldsymbol{\mu}_t$ and a covariance matrix $\boldsymbol{\Sigma}_t$. An importance weight

$$w^{(n)}_t = \frac{\pi_t(\boldsymbol{\theta}^{(n)}_t)}{q(\boldsymbol{\theta}^{(n)}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)},$$

is assigned to each sample.

Proposal adaptation. A particle approximation of the conditional MAP estimator of $\boldsymbol{\theta}$ is given by $\widehat{\boldsymbol{\theta}}_t = \arg\max_n \pi_t(\boldsymbol{\theta}^{(n)}_t)$. The value of current MAP approximation $\pi_t(\widehat{\boldsymbol{\theta}}_t)$ is then compared with the global MAP estimator obtained so far denoted as $\pi_{\texttt{MAP}}$. If $\pi_t(\widehat{\boldsymbol{\theta}}_t) \geq \pi_{\texttt{MAP}}$, all the global MAP estimators are updated and the proposal pdf is moved at $\widehat{\boldsymbol{\theta}}_t$, i.e., we set

$$\widehat{\boldsymbol{\theta}}_{\texttt{MAP}} = \widehat{\boldsymbol{\theta}}_t, \quad \pi^{(t)}_{\texttt{MAP}} = \pi_t(\widehat{\boldsymbol{\theta}}_t), \quad \boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_t, \tag{31}$$

Table 1: ATAIS: AIS with automatic tempering

1. **Initializations:** Choose N, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$, and obtain an initialization for $\widehat{\sigma}_{\texttt{ML}}^{(0)}$, $\pi_{\texttt{MAP}}$ (it can be done using a particle approximation at step $t = 0$).

2. **For** $t = 1, \ldots, T$**:**

   (a) **Sampling:**

   i. Draw $\boldsymbol{\theta}_t^{(1)}, ..., \boldsymbol{\theta}_t^{(N)} \sim q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$.

   ii. Assign to each samples the weights

$$w_t^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)} = \frac{\pi_{\theta|\sigma}(\boldsymbol{\theta}_t^{(n)}|\widehat{\sigma}_{\texttt{ML}}^{(t-1)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}, \qquad n = 1, ..., N. \tag{23}$$

   (b) **Current maximum estimations:**

   i. Obtain $\widehat{\boldsymbol{\theta}}_t = \arg\max_n \pi_t(\boldsymbol{\theta}_t^{(n)})$, and compute $\widehat{\mathbf{r}}_t = \mathbf{f}(\widehat{\boldsymbol{\theta}}_t)$ (for alternatives see Table 2).

   ii. Compute $\widehat{\sigma}_t = \sqrt{\frac{1}{K}\|\mathbf{y} - \widehat{\mathbf{r}}_t\|^2}$.

   (c) **Global maximum estimations:**

   i. If $\widehat{\sigma}_t \leq \widehat{\sigma}_{\texttt{ML}}^{(t-1)}$, then set $\widehat{\sigma}_{\texttt{ML}}^{(t)} = \widehat{\sigma}_t$. Otherwise, set $\widehat{\sigma}_{\texttt{ML}}^{(t)} = \widehat{\sigma}_{\texttt{ML}}^{(t-1)}$.

   ii. If $\pi_t(\widehat{\boldsymbol{\theta}}_t) \geq \pi_{\texttt{MAP}}$, then set $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}} = \widehat{\boldsymbol{\theta}}_t$ and $\pi_{\texttt{MAP}} = \pi_t(\widehat{\boldsymbol{\theta}}_t)$.

   (d) **Adaptation:** Set

$$\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}, \tag{24}$$

$$\boldsymbol{\Sigma}_t = \sum_{n=1}^{N} \bar{w}_t^{(n)} (\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t)^\top (\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t) + \delta \mathbf{I}_M, \tag{25}$$

   where $\bar{w}_t^{(n)} \frac{w_t^{(n)}}{\sum_{i=1}^N w_t^{(i)}}$ are the normalized weights, $\bar{\boldsymbol{\theta}}_t = \sum_{n=1}^N \bar{w}_t^{(n)} \boldsymbol{\theta}_t^{(n)}$ and $\delta > 0$.

3. **Output:** Return the final estimators $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$, $\widehat{\sigma}_{\texttt{ML}}$, and all the weighted samples $\{\boldsymbol{\theta}_t^{(n)}, \widetilde{w}_t^{(n)}\}$, for all t and n, with the corrected weights

$$\widetilde{w}_t^{(n)} = w_t^{(n)} \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}. \tag{26}$$

Otherwise, we keep the previous values $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$, $\pi_{\texttt{MAP}}$ and $\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1}$. The covariance matrix $\boldsymbol{\Sigma}_t$ is adapted by considering the empirical covariance of the weighted samples. Note that, we set $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ instead of using the empirical mean of the samples (as in other AIS schemes). This is due to we have no-

Table 2: Possible model approximations

1. **MAP:** Given $\widehat{\boldsymbol{\theta}}_t = \arg\max_n \pi_t(\boldsymbol{\theta}_t^{(n)})$, then set

$$\widehat{\mathbf{r}}_t = \mathbf{f}(\widehat{\boldsymbol{\theta}}_t), \tag{27}$$

2. **MMSE:** Given $\bar{w}_t^{(n)} \frac{w_t^{(n)}}{\sum_{i=1}^N w_t^{(i)}}$ and $\bar{\boldsymbol{\theta}}_t = \sum_{n=1}^N \bar{w}_t^{(n)} \boldsymbol{\theta}_t^{(n)}$, then set

$$\widehat{\mathbf{r}}_t = \mathbf{f}(\bar{\boldsymbol{\theta}}_t), \tag{28}$$

3. **Fully-Bayesian solution:** Given $\bar{w}_t^{(n)} \frac{w_t^{(n)}}{\sum_{i=1}^N w_t^{(i)}}$, then set

$$\widehat{\mathbf{r}}_t = \sum_{n=1}^N \bar{w}_t^{(n)} \mathbf{f}(\boldsymbol{\theta}_t^{(n)}). \tag{29}$$

tice that this choice provide better and more robust results, specially as the dimension of the problem grows.

Automatic tempering. As we show in the previous section, the current ML estimator of σ can be obtained analytically as

$$\widehat{\sigma}_t = \sqrt{\frac{1}{K}\|\mathbf{y} - \widehat{\mathbf{r}}_t\|^2}, \tag{32}$$

where $\widehat{\mathbf{r}}_t = \mathbf{f}(\widehat{\boldsymbol{\theta}}_t)$ (some alternatives are given in Table 2). If the current ML estimator $\widehat{\sigma}_t$ is smaller than a global one $\widehat{\sigma}_{\texttt{ML}}$, i.e., $\widehat{\sigma}_t < \widehat{\sigma}_{\texttt{ML}}^{(t-1)}$, then we update $\widehat{\sigma}_{\texttt{ML}}^{(t)} = \widehat{\sigma}_t$, Otherwise, we keep the value of $\widehat{\sigma}_{\texttt{ML}}^{(t)} = \widehat{\sigma}_{\texttt{ML}}^{(t-1)}$.

ATAIS outputs. After T iterations, a final correction of the weights is needed, i.e.,

$$\widetilde{w}_t^{(n)} = w_t^{(n)} \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}, \tag{33}$$

in order to obtain a particle approximation of the measure of the final conditional posterior $\bar{\pi}_{\theta|\sigma}(\boldsymbol{\theta}|\widehat{\sigma}_{\texttt{ML}})$. Thus, the algorithm returns the final estimators $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$, $\widehat{\sigma}_{\texttt{ML}}$, and all the weighted samples $\{\boldsymbol{\theta}_t^{(n)}, \widetilde{w}_t^{(n)}\}$, for all $n = 1, ..., N$ and $t = 1, ..., T$. Other outputs can be obtained with a post-processing of the weighted samples, as shown below.

Approximation of $Z(\sigma) = p(\mathbf{y}|\sigma)$. After the T iterations of ATAIS, we can also approximate the conditional marginal likelihood $Z(\sigma) = p(\mathbf{y}|\sigma)$ without additional evaluations of the target function. Indeed, saving the error values at each particles obtained for the computation of the likelihood function

during ATAIS,

$$e_t^{(n)} = \|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}_t^{(n)})\|^2,$$

then for a generic value of σ we can compute the IS weights,

$$\rho_t^{(n)}(\sigma) = \frac{\frac{1}{\sigma^K}\exp\left(-\frac{e_t^{(n)}}{2\sigma^2}\right) g_\theta(\boldsymbol{\theta})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}.$$

Thus, the IS estimator of the conditional marginal likelihood $Z(\sigma) = p(\mathbf{y}|\sigma)$ is given by the arithmetic mean of the weights $\rho_t^{(n)}(\sigma)$,

$$\widehat{Z}(\sigma) = \widehat{p}(\mathbf{y}|\sigma) = \frac{1}{NT}\sum_{t=1}^{T}\sum_{n=1}^{N}\rho_t^{(n)}(\sigma). \tag{34}$$

Furthermore, if we draw $\sigma^{(r)} \sim g_\sigma(\sigma)$, from $r = 1, ..., R$ then we can approximate the global marginal likelihood by applying the standard Monte Carlo to the integral in Eq. (18),

$$\widehat{Z} = \frac{1}{R}\sum_{r=1}^{R}\widehat{Z}(\sigma^{(r)}). \tag{35}$$

An approximation of the marginal posterior $\bar{\pi}_\sigma(\sigma) = \frac{p(\mathbf{y}|\sigma)g_\sigma(\sigma)}{p(\mathbf{y})}$ in Eq. (12) can be also obtained as

$$\widehat{\pi}_\sigma(\sigma) = \frac{\widehat{Z}(\sigma)g_\sigma(\sigma)}{\widehat{Z}}. \tag{36}$$

5 Simulations

We test the proposed scheme in two numerical examples. The first numerical experiment is a simple bidimensional example (which is easy to be reproduced). The second experiment considers a real-world application, i.e., a radial velocity models of exoplanet systems which is often employed in astronomy applications (with a dimension of the inference problem of 6 and 11).

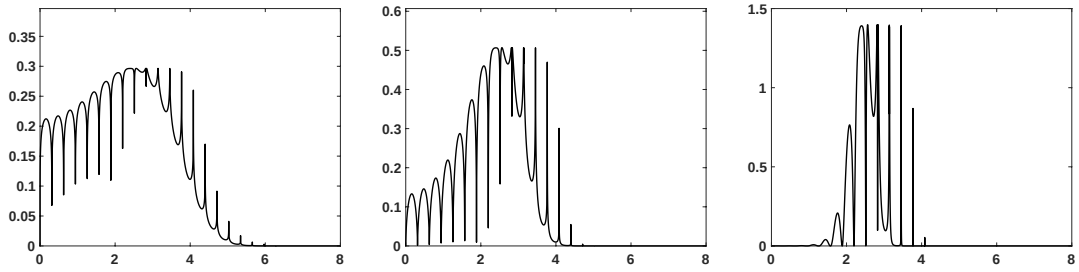
5.1 First numerical analysis

For the sake of simplicity, let us consider $\theta \in \mathbb{R}$ and an observation model given by the equation

$$y_k = \theta^2 - \log(|\sin(10\theta)|) + e_k,$$

where $e \sim \mathcal{N}(e_k|0, \sigma^2)$ and clearly $f(\theta) = \theta^2 + \log(|\sin(10\theta)|)$. We consider $\theta_{\texttt{true}} = 2.5$, and $\sigma_{\texttt{true}} = 4$. We generate $K = 8$ observations from the model above. We also consider a uniform prior for θ in $[0, 20]$. The conditional posterior $\bar{\pi}_{\theta|\sigma}(\theta|\sigma_{\texttt{true}})$ is shown in Figure 1(c). We can observe that $\bar{\pi}_{\theta|\sigma}$ is highly multimodal. Figure 1 also depict the conditional posteriors $\bar{\pi}_{\theta|\sigma}(\theta|\sigma)$ with $\sigma \in \{10, 20\}$. Considering also a uniform prior over σ in $[0, 20]$, we have also a bidimensional complete posterior

over $[\theta, \sigma]$, which is depicted in Fig. 2(a).
In this bidimensional example, it is possible to obtain the ground-truths using an expensive thin grid. The expected value and variance of the conditional posterior $\theta|\sigma_{\texttt{true}}$, approximated by a thin grid, $E[\theta|\sigma_{\texttt{true}}] = 2.48$ and $\text{var}[\theta|\sigma_{\texttt{true}}] = 0.11$. The MAP estimator of the conditional posterior $\theta|\sigma_{\texttt{true}}$ is $\theta_{\texttt{MAP}}|\sigma_{\texttt{true}} = 2.56$. The expected value and variance of marginal posteriors of θ and σ are $E[\theta] = 2.46$, $\text{var}[\theta] = 0.18$, $E[\sigma] = 4.32$ and $\text{var}[\sigma] = 2.43$ (these values coincide with the expected value and the diagonal of the covariance matrix of the complete posterior). The MAP estimator provided by the complete posterior is $[\theta_{\texttt{MAP-joint}}, \sigma_{\texttt{MAP-joint}}] = [2.56, 3.23]$. Since the prior over σ is uniform, the maximum likelihood of σ is $\sigma_{\texttt{ML}} = \sigma_{\texttt{MAP-joint}} = 3.23$. These values can be obtained sampling in the 2D space $[\theta, \sigma]$ and then considering the components of the drawn vectors. The MAP estimator of the marginal posteriors in Eq. (12) are $\theta_{\texttt{MAP-marg}} = 2.56$ and $\sigma_{\texttt{MAP-marg}} = 3.46$. The two marginal posteriors are shown in Figures 2(b)-(c).
We apply ATAIS with the goal of estimating the expected value and the variance of the posterior density with respect to θ. we consider a Gaussian proposal $q(\theta|\mu_t, \lambda_t)$ with $\mu_0 = 10$ and a starting variance of $\lambda_0 = 4$. Note that μ_0 is located in region that does not contains modes. We also start with $\widehat{\sigma}^{(0)}_{\texttt{ML}} = 20$ and $\pi_{\texttt{MAP}} = 0$ (initial conditions). The Mean Square Error (MSE) of ATAIS, averaged over 500 runs, in estimation of different moments and modes as function of N (and with $T = 10$), is given in Table 3. The ML estimation $\widehat{\sigma}^{(t)}_{\texttt{ML}}$ as function of the iteration t (with $N = 5$) for different runs, is given in Figure 3(a). Approximations $\widehat{\pi}_\sigma(\sigma)$ obtained as in Eq. (36) of the marginal posterior $\bar{\pi}_\sigma(\sigma)$, in one specific run, with different $N \in \{10, 100, 500\}$ and $T = 10$.

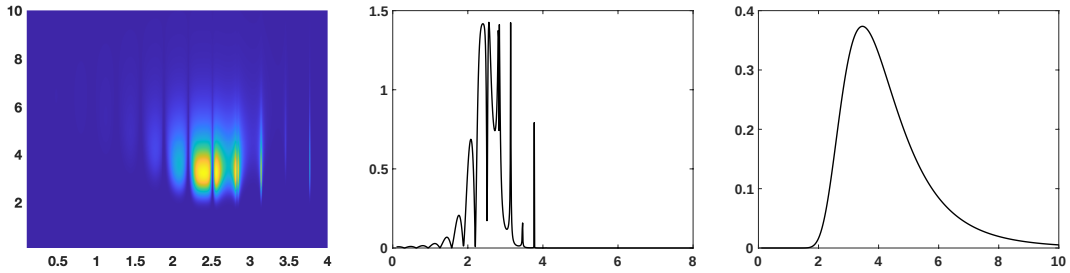


(a) $\sigma = 20$. (b) $\sigma = 10$. (c) $\sigma = \sigma_{\texttt{true}} = 4$.

Figure 1: Conditional posteriors corresponding to different value of σ: more specifically, **(a)** $\sigma = 20$, **(b)** $\sigma = 10$, **(c)** $\sigma = \sigma_{\texttt{true}} = 4$.

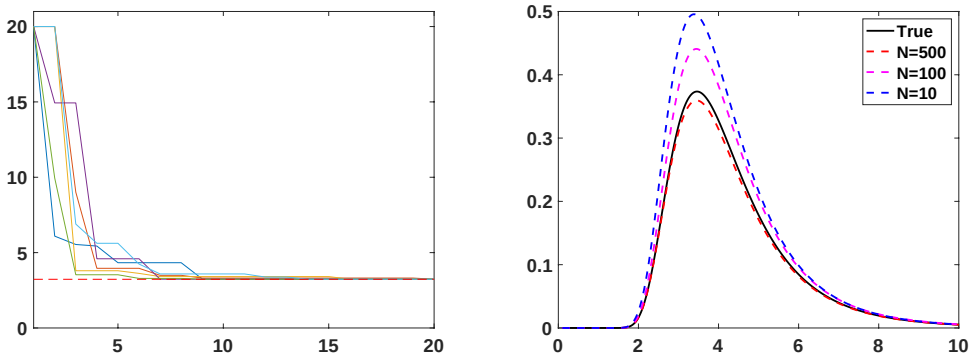
5.2 Radial velocity curves of exoplanets and binary systems

In this example, we consider an application in an astronomical model. In recent years, the problem of revealing objects orbiting other stars has acquired large attention. Different techniques have been proposed to discover exo-objects but, nowadays, the radial velocity technique is still the most used [7, 3, 1, 18]. The problem consists in fitting a model (the so-called radial velocity curve) to data acquired at different moments spanning during long time periods (up to years). The model is highly



(a) Complete posterior. (b) Marginal posterior $\bar{\pi}_\theta(\theta)$. (c) Marginal posterior $\bar{\pi}_\sigma(\sigma)$.

Figure 2: The bidimensional complete posterior $\bar{\pi}(\theta, \sigma)$ and the two marginal posteriors $\bar{\pi}_\theta(\theta)$, $\bar{\pi}_\sigma(\sigma)$ in Eq. (12), obtained by using a thin grid approximation.



(a) $\widehat{\sigma}^{(t)}_{\mathtt{ML}}$ (b) $\widehat{\pi}_\sigma(\sigma)$ for different N ($T = 10$).

Figure 3: **(a)** The ML estimation $\widehat{\sigma}^{(t)}_{\mathtt{ML}}$ (different runs) versus the iteration t, with $N = 5$. **(b)** The true marginal posterior $\bar{\pi}_\sigma(\sigma)$ and different approximations, in one specific run, $\widehat{\pi}_\sigma(\sigma)$ obtained as in Eq. (36) with different $N \in \{10, 100, 500\}$ and $T = 10$ (hence, the total number of samples are NT).

non-linear and it is costly in terms of computation time (specially, for certain sets of parameters). Obtaining a value to compare to a single observation involves numerically integrating a differential equation in time or an iterative procedure for solving to a non-linear equation. Typically, the iteration is performed until a threshold is reached or 10^6 iterations are performed. The problem of radial velocity curve fitting is applied in several related applications.

Observation model - likelihood. When analysing radial velocity data of an exoplanetary system, it is commonly accepted that the *wobbling* of the star around the centre of mass is caused by the sum of the gravitational force of each planet independently and that they do not interact with each other.

Table 3: MSE of ATAIS (averaged over 500 runs), in the estimation of different moments and modes as function of N and $T = 10$.

Moments-Modes	$N = 10$	$N = 100$	$N = 1000$	$N = 5000$
$E[\theta\|\sigma_{\texttt{true}}]$	0.0311	0.0098	0.0034	0.0024
$\text{var}[\theta\|\sigma_{\texttt{true}}]$	0.0474	0.0370	0.0298	0.0201
$\theta_{\texttt{MAP}}\|\sigma_{\texttt{true}}$	0.0410	0.0337	0.0285	0.0127
$E[\sigma]$	0.9233	0.0785	0.0097	0.0023
$\text{var}[\sigma]$	6.1869	0.2640	0.0035	0.0010
$\sigma_{\texttt{MAP-marg}}$	0.0056	0.0004	0.0001	$3 \cdot 10^{-5}$
$\sigma_{\texttt{ML}}$	$8 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$5 \cdot 10^{-7}$	$6 \cdot 10^{-9}$

Each planet follows a Keplerian orbit and the radial velocity of the host star is given by

$$y_{r,t} = V_0 + \sum_{i=1}^{S} K_i \left[\cos\left(u_{i,t} + \omega_i\right) + e_i \cos\left(\omega_i\right)\right] + \xi_t, \tag{37}$$

with $t = 1, \ldots, T$ and $r = 1, \ldots, R$. The number of objects in the system is S, that is consider known in this experiment (for the sake of simplicity). Both $y_{r,t}$, $u_{i,t}$ depend on time t, and then ξ_t is a Gaussian noise perturbation with variance σ^2. For the sake of simplicity, we consider this value known, $\sigma^2 = 1$. The likelihood function is defined by (37) and some indicator variables described below. The angle $u_{i,t}$ is the true anomaly of the planet i and it can be determined from

$$\frac{du_{i,t}}{dt} = \frac{2\pi}{P_i} \frac{\left(1 + e_i \cos u_{i,t}\right)^2}{\left(1 - e_i\right)^{\frac{3}{2}}} \tag{38}$$

As mentioned above, this equation has analytical solution. As a result, the true anomaly u_t can be determined from the mean anomaly M. However, the analytical solution contains a non linear term that needs to be determined by iterating. First, we define the mean anomaly $M_{i,t}$ as

$$M_{i,t} = \frac{2\pi}{P_i}\left(t - \tau_i\right), \tag{39}$$

where τ_i is the time of periastron passage of the planet i and P_i is the period of its orbit (see Table **??**). Then, through the Kepler's equation,

$$M_{i,t} = E_{i,t} - e_i \sin E_{i,t}, \tag{40}$$

where $E_{i,t}$ is the eccentric anomaly. Equation (40) has no analytic solution and it must be solved by an iterative procedure. A Newton-Raphson method is typically used to find the roots of this equation [16]. For certain sets of parameters this iterative procedure, can be particularly slow. We also have

$$\tan\frac{u_{i,t}}{2} = \sqrt{\frac{1 + e_i}{1 - e_i}} \tan\frac{E_{i,t}}{2}, \tag{41}$$

The variable of interest is then $\boldsymbol{\theta}$ is the vector

$$\boldsymbol{\theta} = [V_0, k_{a,1}, \omega_1, e_1, P_1, \tau_1, \ldots, k_{a,S}, \omega_S, e_S, P_S, \tau_S], \tag{42}$$

Then, for a single object (e.g., a planet or a natural satellite), the dimension of $\boldsymbol{\theta}$ is $M = 5 + 1 = 6$, with two objects the dimension of $\boldsymbol{\theta}$ is is $M = 11$ etc.

This example consists in a synthetic radial velocity curve of a planetary system with one planet or two planets (i.e., $S = 1$ or $S = 2$). More specifically, we generate simulated data with a model with two planets. The orbital parameters of the planets are listed in Table 4, where P is the period of the orbit, k_a is the amplitude of the curve, e is the eccentricity of the orbit, ω is the argument of perigee and τ is the last periastron passage. A mean velocity $v_0 = 5\ \mathrm{m\,s^{-1}}$ is assumed. A Gaussian noise perturbation is added with a standard deviation $\sigma = 3\ \mathrm{m\,s^{-1}}$. To simulate observations, a total of $K = 120$ points are selected from three, random time periods (and two planets in the system). Note that the amplitude of the radial velocity curve of the second planet is close to the noise level. We run ATAIS and a standard AIS scheme with the model with one planet and with the model with two planets. The purpose of this simulation is to check the ability of the method to detect the two planets (by approximating the model evidence).

Table 4: Main orbital parameters of the two exoplanets in the simulation.

Parameter	Planet 1	Planet 2
P	15 d	115 d
k_a	$25\ \mathrm{m\,s^{-1}}$	$5\ \mathrm{m\,s^{-1}}$
e	0.1	0.0
ω	35°	10°
τ	3 d	24 d

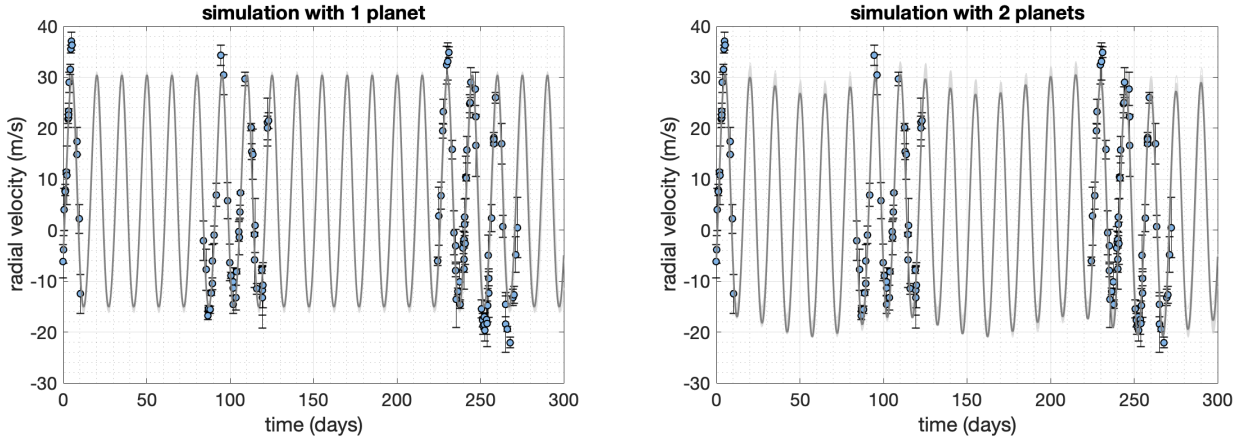


Figure 4: Comparison of the results of the ATAIS algorithm with the simulations (blue dots). Left panel shows, in grey, the radial velocity curve for $\widehat{\boldsymbol{\theta}}_{\mathtt{MAP}}$ using a model with one planet. Right panel is like left panel but considering a model with two planets.

We apply ATAIS and a standard AIS scheme [4] over the space $[\boldsymbol{\theta}, \sigma]$ for approximating the model evidence $Z = p(\mathbf{y})$ (marginal likelihood) of both models (one planet or two planets) with the given data (generated considering two planets). Uniform priors are considered for each parameter: $P \in [0, 365]$, $k_a \in [-20, 20]$, $e \in [0, 1]$, $\omega \in [0, 2\pi]$, and $\tau \in [0, 50]$ (moreover, $\sigma \in [0, 30]$ for the standard AIS scheme). The ATAIS algorithm and the standard AIS scheme has been run with $N = 10^5$ and $T = 50$ iterations for both, the model with one and two planets. In both case, we consider the same Gaussian proposal with a starting standard deviation of 5 for each component (note that the standard AIS scheme works in higher dimensional space due the inference over σ). To decide which model is more probable, the model evidence Z of each model is estimated. More specifically, we approximate the one-planet model $\widehat{Z}_1 = \widehat{p}_1(\mathbf{y})$ of the two-planets model $\widehat{Z}_2 = \widehat{p}_2(\mathbf{y})$ with the ATAIS algorithm and the standard AIS scheme. When $\widehat{Z}_1 > \widehat{Z}_2$ we select the first model otherwise if, $\widehat{Z}_1 < \widehat{Z}_2$, we select the second one. The true model is the two-planets model, since the simulated data are generated from the two-planets model. After 500 independent runs, the percentage of correct detection of the true model for ATAIS is $\approx 98\%$, whereas with the standard AIS scheme is only $\approx 56\%$. This due to the difficulty of making inference jointly over $[\theta, \sigma]$. In ATAIS, the ratio between the model evidences (averaged over the 500 runs) is $\widehat{Z}_1 = \widehat{p}_1(\mathbf{y})$ is $Z_2/Z_1 \approx 5 \cdot 10^3$. Therefore, for ATAIS, the model with two planets is clearly more probable than the model with one planet.

The fitted curves, corresponding to the vector of parameters $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ obtained with ATAIS. are shown in Fig. 4. From the figure, it is not clear which model fits better the simulated observations (blue points), although the model with two planets seems to fit better the observations in the time period from 200 to 300 days. The values of $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$, obtained in one specific run by ATAIS, is given in Table 5. We notice that ω and τ are highly correlated and more iterations may be needed to obtain the actual global maximum, but the remaining parameters obtained from $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ are similar to the simulated values. In addition, the amplitude of the curve of the second planet is close to the intensity of the noise, what makes difficult to derive the best fit for that planet. Summarizing, our results show the method is able to discriminate between a model with one planet (with 6 dimensions of the inference problem) and a model with two planets (with 11 dimensions of the inference problem), for this particular simulation. Finally, the evolution of the automatic tempering parameter $\widehat{\sigma}^{(t)}_{\texttt{ML}}$ is shown in Fig. 5. The dashed line is the evolution of $\widehat{\sigma}^{(t)}_{\texttt{ML}}$ for the single-planet model. The continuous line is the evolution of $\widehat{\sigma}^{(t)}_{\texttt{ML}}$ for the model with two planets. In this second model, the tempering parameters reaches a smaller value, as expected.

Table 5: The value of $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ for the 2-planets model

Parameter	Planet 1	Planet 2
P	14.99 d	110.39 d
K	23.78 m s^{-1}	3.50 m s^{-1}
e	0.05	0.00
ω	43.7°	0.39°
τ	6.8 d	7.96 d

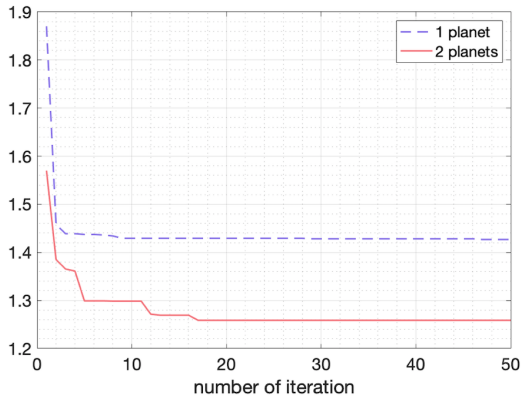


Figure 5: Evolution of the tempering parameter $\widehat{\sigma}_{\mathtt{ML}}^{(t)}$. The dashed line is the evolution for the model with one planet. The continuous line is the evolution of the two-planets model.

6 Conclusions

We have proposed a novel AIS scheme for Bayesian inversion problems where an automatic tempering procedure is implemented (called ATAIS). The inference of the variables of interest $\boldsymbol{\theta}$ and the noise power σ^2 is divided. A sampling strategy is considered for $\boldsymbol{\theta}$ and an optimization approach is employed for σ^2. Thus, ATAIS performs an iterative procedure, alternating sampling and optimization steps. Therefore, the proposed scheme deals with a sequence of tempered posteriors according to the current estimation of the noise power. We have also discussed to possibility of approximating the marginal posterior of σ without additional evaluation of the complex model (and of the posterior). Several simulations are provided and the application to a sophisticated astronomical model has been considered, where the number of planets in the system is detected by the analysis of the marginal likelihood. The results show the benefits of the proposed scheme. For instance, in the astronomical example, the percentage of correct detection of the true model obtained by ATAIS is $\approx 98\%$, whereas with the standard AIS scheme is only $\approx 56\%$. As future research, we plan to extend the ATAIS scheme in order to deal with an observation model with correlated noise perturbations (for instance, using a Gaussian Process).

References

[1] L. Affer, M. Damasso, G. Micela, E. Poretti, G. Scand ariato, J. Maldonado, A. F. Lanza, E. Covino, A. Garrido Rubio, J. I. González Hernández, R. Gratton, G. Leto, A. Maggio, M. Perger, A. Sozzetti, A. Suárez Mascareño, A. S. Bonomo, F. Borsa, R. Claudi, R. Cosentino, S. Desidera, P. Giacobbe, E. Molinari, M. Pedani, M. Pinamonti, R. Rebolo, I. Ribas, and B. Toledo-Padrón. HADES RV program with HARPS-N at the TNG. IX. A super-Earth around the M dwarf Gl 686. *arXiv:1901.05338*, 622:A193, February 2019.

[2] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

[3] S. C. C. Barros, D. J. A. Brown, G. Hébrard, Y. Gómez Maqueo Chew, D. R. Anderson, P. Boumis, L. Delrez, K. L. Hay, K. W. F. Lam, J. Llama, M. Lendl, J. McCormac, B. Skiff, B. Smalley, O. Turner, M. Vanhuysse, D. J. Armstrong, I. Boisse, F. Bouchy, A. Collier Cameron, F. Faedi, M. Gillon, C. Hellier, E. Jehin, A. Liakos, J. Meaburn, H. P. Osborn, F. Pepe, I. Plauchu-Frayn, D. Pollacco, D. Queloz, J. Rey, J. Spake, D. Ségransan, A. H. M. Triaud, S. Udry, S. R. Walker, C. A. Watson, R. G. West, and P. J. Wheatley. WASP-113b and WASP-114b, two inflated hot Jupiters with contrasting densities. *Astronomy and Aastrophysics*, 593:A113, 2016.

[4] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.

[5] W. J. Fitzgerald. Markov chain Monte Carlo methods with applications to signal processing. *Signal Processing*, 81(1):3–18, January 2001.

[6] N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(3):589–607, 2008.

[7] Philip C. Gregory. Bayesian re-analysis of the Gliese 581 exoplanet system. *Monthly Notices of the Royal Astronomical Society*, 415(3):2523–2545, August 2011.

[8] S. K. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[9] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2004.

[10] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *arXiv:2005.08334*, pages 1–59, 2020.

[11] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, 19(6):451–458, July 1992.

[12] L. Martino, D. Luengo, and J. Miguez. *Independent Random Sampling methods*. Springer, 2018.

[13] L. Martino and J. Míguez. Generalized rejection sampling schemes and applications in signal processing. *Signal Processing*, 90(11):2981–2995, November 2010.

[14] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[15] R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[16] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes in C++ : the art of scientific computing*. Springer, 2002.

[17] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[18] Trifon Trifonov, Stephan Stock, Thomas Henning, Sabine Reffert, Martin Kürster, Man Hoi Lee, Bertram Bitsch, R. Paul Butler, and Steven S. Vogt. Two Jovian Planets around the Giant Star HD 202696: A Growing Population of Packed Massive Planetary Pairs around Massive Stars? *The Astronomical Journal*, 157(3):93, March 2019.

A On the optimization of the likelihood function

Let us set $\delta = \sigma^2$ and consider to optimize of the likelihood function

$$\ell(\boldsymbol{\theta}, \delta) = \frac{1}{(2\pi\delta)^{K/2}} \exp\left(-\frac{V(\boldsymbol{\theta})}{\delta}\right).$$

Recall that, in our model, we have $V(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2$. We desire to obtain

$$[\boldsymbol{\theta}_{\mathtt{ML}}, \delta_{\mathtt{ML}}] = \arg\max \ell(\boldsymbol{\theta}, \delta).$$

We can write the gradient and equal to zero,

$$\begin{cases} \nabla_\theta \ell(\boldsymbol{\theta}, \delta) = -\dfrac{1}{\delta} \nabla_\theta V(\boldsymbol{\theta}) \left[\dfrac{1}{(2\pi\delta)^{K/2}} \exp\left(-\dfrac{V(\boldsymbol{\theta})}{\delta}\right)\right] = \mathbf{0} \Longrightarrow \nabla_\theta V(\boldsymbol{\theta}) = \mathbf{0}, \\ \dfrac{\partial \ell(\boldsymbol{\theta}, \delta)}{\partial \delta} = \dfrac{\mathrm{e}^{-\frac{V(\boldsymbol{\theta})}{\delta}} \ (2V(\boldsymbol{\theta}) - \delta K)}{2^{\frac{K}{2}+1} \delta^{\frac{K}{2}+2} \pi^{K/2}} = 0 \Longrightarrow \delta = \dfrac{2}{K} V(\boldsymbol{\theta}). \end{cases} \tag{43}$$

We have obtained that the ML solution is defined by the system of equations,

$$\begin{cases} \nabla_\theta V(\boldsymbol{\theta}_{\mathtt{ML}}) = \mathbf{0} \\ \delta_{\mathtt{ML}} = \dfrac{2}{K} V(\boldsymbol{\theta}_{\mathtt{ML}}). \end{cases} \tag{44}$$