UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

OPTIMAL SAMPLING METHOD FOR UNCERTAINTY REDUCTION OF ORBITAL PARAMETERS ON BINARY SYSTEMS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

FELIPE IGNACIO ALEJANDRO CÓRDOVA HUENUPIL

PROFESOR GUÍA:
JORGE SILVA SÁNCHEZ

PROFESOR CO-GUÍA:
RENÉ MÉNDEZ BUSSARD

MIEMBROS DE LA COMISIÓN:
MARCOS ORCHARD CONCHA
FRANCISCO FÖRSTER BURÓN

SANTIAGO DE CHILE
2021

RESÚMEN DE LA TÉSIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA, MENCIÓN ELÉCTRICA
RESÚMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: **FELIPE IGNACIO ALEJANDRO CÓRDOVA HUENUPIL**
FECHA: 2021
PROF. GUÍA: JORGE SILVA SÁNCHEZ

ESTRATEGIA DE MUESTREO ÓPTIMO PARA LA REDUCCIÓN DE INCERTEZA DE PARÁMETROS ORBITALES EN SISTEMAS ESTELARES BINARIOS.

La inferencia de parámetros orbitales para sistemas binarios es extremadamente importante para el estudio evolutivo del universo, puesto que dichos parámetros orbitales permiten inferir las masas individuales de los cuerpos celestes participantes. El presente trabajo busca reducir la incerteza en la estimación de parámetros orbitales mediante la selección de un instante óptimo de observación restringido a una agenda finita de eventos. Para lograr esto un enfoque estocástico es propuesto, el cual define un modelo probabilístico de observación para describir el acto fenomenológico de observar un sistema. Este enfoque probabilístico permite definir el problema de minimización mediante medidas provenientes de Teoría de la Información para caracterizar la reducción incerteza párametrica basado en el principio de *"Maximum Entropy Sampling"*. El marco metodológico propuesto en este trabajo es referido como *"Optimal Sampling Criterion"* y esta compuesto por una etapa de inferencia y una etapa de estimación de entropía diferencial. A través de experimentos utilizando observaciones simuladas y reales, se logra probar que la estrategia de muestreo propuesta es no solo capas de seleccionar los instantes de observación mas informativos por orden de prioridad, si no que también provee de sustento teórico a reglas heurísticas de muestreo utilizadas por la comunidad astronómica.

RESÚMEN DE LA TÉSIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA, MENCIÓN ELÉCTRICA
RESÚMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: **FELIPE IGNACIO ALEJANDRO CÓRDOVA HUENUPIL**
FECHA: 2021
PROF. GUÍA: JORGE SILVA SÁNCHEZ

OPTIMAL SAMPLING METHOD FOR UNCERTAINTY REDUCTION OF ORBITAL PARAMETERS ON BINARY SYSTEMS

The inference of orbital parameters from a binary system is extremely important to study the evolutionary process of the universe, mainly because allows to deduce the stellar masses of the involved celestial bodies. This work is focused on reducing the uncertainty of the parametric estimation by consciously selecting an optimal instant of observation from a plausible agenda of experiments. To achieve this, a stochastic approach is proposed in which a observation model is defined to describe the phenomenological act of observation. This probabilistic scheme allows to establish a minimization problem which uses Information Theoretic measures to characterize the reduction of uncertainty over the parameters space based on the principle of Maximum Entropy Sampling. The proposed framework is referred as Optimal Sampling Criteria model and is composed by an inference stage and a differential entropy estimation stage. Through experiments using simulated and real data, the proposed optimal sampling strategy is shown not only capable of accurately selecting the most informative measurements by priority, but also providing theoretical substance to heuristic intuition-based rules and identify probabilistic relationships between the orbital parameters and the observations.

En memoria de Negrito,
mi gatito.

Agradecimientos

A mi padres, por todo el apoyo y esfuerzo que me han brindado, por permitirme tener estudios superiores y aguantarme tantos años en la casa. Sinceramente muchas gracias y los amo mucho.

A mi gato Negrito, que vio el principio de este documento pero no el final, te extraño mucho. A mis gatitos Murci y Koala que me interrumpieron infinitas veces cruzándose en el monitor o para que les abriera la ventana, los quiero mucho a pesar de que se tomen mi té.

A mis profesores, por permitirme conocer el mundo de la investigación. Al profesor Jorge por guiarme en el campo de la teoría, al profesor Marcos por recordarme la bajada practica a la teoría y al profesor René por guiarme en el campo de la astronomía y por su risa contagiosa.

A mis compañeros de carrera, por acompañarme en este proceso, por los shawarmas y por la chelas tibias.

A mis compañeros del laboratorio IDS, por el agua embotellada, la cajita de los tés, las cápsulas nespresso, los cierres de transmisión y los vientos del Sebi.

A mis amigos por ayudarme a mantener la cordura a distancia durante esta pandemia. /spit al que le caiga.

A DualVision por invitarme a trabajar con ustedes en conjunto al desarrollo de este documento. Un saludo a los jefazos.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The study and observation of binary systems is central to understand the evolution of stellar systems. The stellar dynamics in those systems are fundamentally bounded to their physical properties, for example the individual masses of each star, and the orbital path made by the interaction of those celestial bodies. Consequently, the development of inference methods focused on estimating orbital parameters through observations is a key element for the area of astro-statistics. Many authors have addressed the estimation of orbital parameters from different angles using a wide variety of available observations. Most inference methods for orbital parameters utilize astrometric observations, informative about the position between the involved starts, and spectroscopic data, indicative of the radial velocity of the stars. Deterministic and stochastic approaches have been proposed to solve this problem, the last stochastic scenario is particularly relevant in this context because it allows to introduce the concept of parametric certainty when a Bayesian inference is performed. Naturally any method to infer the parameters of a binary system is fundamentally related to the quantity and the quality of the available data, without loss of generality it is possible to say that more observations often implies a reduction of uncertainty on the inferred parameters. Unluckily, astrometric and spectroscopic observation are extremely scarce in practice and obtaining new data is very difficult, due to restrictive schedules in observatories and monetary costs. Because of this, the astronomy community often selects instants of observation based on heuristics and suggestion by made repetition.

This works presents an statistical study on the dynamics of a binary system, the objective is to establish a method to suggest optimal instants of observation that minimize the expected posterior parametric certainty based on previous observations. The probabilistic analysis also includes a methodological approach to model the quality of the observation and its importance in the inference scheme. The proposed Optimal Sampling Criterion corresponds to a selection of optimal instant of observation which uses a theoretical framework to decide the best candidate from a finite amount of candidates in a observational agenda. The main benefit of this method is the theoretical background which supports the selection that do not include any biased information from the designer and can be easily automated to support the observatories to schedule observations.

1.1. Hypothesis

This work focuses on defining an Optimal Sampling Criterion. To accomplish this the present document aims to test the following hypothesis:

- The phenomenological relationship between observations (astrometric and spectroscopic) and orbital parameters allows to explicitly relate previous observations with future ones.
- Obtaining new observations of a binary system always implies an expected uncertainty reduction in the posterior distribution of the parameters.
- It is possible to minimize the expected uncertainty of the parametrical posterior distribution by selection optimal instant of observation.

1.2. Objectives

1.2.1. Main objective

The main objective of this work is to establish a theoretical and methodological framework that addresses the problem of optimal selection of observations, from a plausible agenda, which minimizes the expected uncertainty of the orbital parametrical inference of binary systems when pre-existing observations from the phenomena are already available.

1.2.2. Specific objectives

The specific objectives of this work can be listed as follows:

- Introduce Information Theory measures to define the concept of uncertainty and study the probabilistic relationship between parameter and observations using bayesian inference.
- Extend the Maximum Entropy Sampling [Sebastiani and Wynn, 2000] methodology for this problem, resulting in the definition of the Optimal Sampling Criterion.
- Establish a practical framework to estimate the Optimal Sampling Criterion by means of Markov Chain Monte Carlo particle simulation and differential entropy estimation, in a two stage workflow.
- Empirically prove the Optimal Sampling Criterion by means of an expectancy analysis using simulated and real observations.
- Discuss the results of the empirical analysis in a astronomical scope, looking for theoretical support to heuristic rules used by the astronomic community.

1.3. Structure

This work is documented and described by the following Chapters:

Chapter 2 of this work corresponds to the Theoretical Framework, where the dynamics of a binary are presented and the phenomenological equations are described. Also the observational model is completely defined in conjunction to the bayesian inference principals, which will serve as basis to the proposed sampling method.

Chapter 3 explores some related works regarding inference process and some previous attempts establish a optimal sampling selection founded in the literature. This will serve as motivation to address the problem of optimal sampling.

Chapter 4 derives and defines the Optimal Sampling Criterion and explores many applications, The previously mentioned stages are also presented with their respective embedded estimation techniques.

Chapter 5 explores in experimental settings the viability of the method in different scenarios. A simulated environment is used to widely study how the differential entropy estimation performs and the capability of the method to detect a priority list. A real data scenario showed a more practical approach since the parametrical ground-thruth is missing for this cases.

Chapter 6 summarizes the main result of this work and proposes future applications and researches regarding the defined Optimal Sampling Criterion.

Chapter 2

Theoretical Framework

2.1. Binary stars systems

A Binary stellar system is the main object of study of this work and, as the name suggest, it is composed by 2 stars that interact with each other due gravitational forces induced by their respective masses. This interaction provokes a elliptical dynamic between each star at the center mass of the whole system known as Keplerian orbit. In celestial mechanics there are 2 possibles interaction between 2 celestial bodies: a parabolic movement, which leads to a short interaction and an elliptical case where the system finds equilibrium and periodicity. This work is only interested on the latter case.

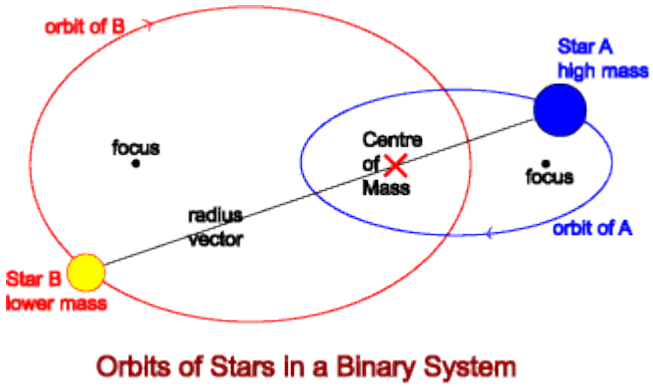


Figure 2.1: Orbits of Starts in a Binary System. Image Source: atnf.csiro.au

One important aspect of the binary system is the "dual representation" of the motion path, which is determined by the point of reference. For example Figure 2.1 shows the movement of each star when the centre of mass of the system is the reference and coincides with a focus point on both drawn ellipses for star A and B. In this context, the interaction can be expressed as a elliptical orbit of the lower mass star B having the high mass star A in a focal point. Both cases relate to each other by their respective semiaxes, which follows $a_1 + a_2 = a$ where a_1 and a_2 are the semi-major axis for the high mass star A and the lower mass star B orbits, respectively, and a is the semi-major axis of the relative orbit of star B with respect of star A.

The reference orbit is preferred not only for observational reasons (obtaining relative distance of stars towards an imaginary point in the sky is implausible), but also because it simplifies the movement description of both celestial bodies. Using geometric relationships given by conic sections mathematics and the fact that the orbit is periodic, it is possible to identify the exact position of the star B relative to A in any moment t using the set of parameters $[T, P, e, a]$ in which P is the period of the orbit, T describe moment in time where the primary star A and the secondary star B reach their minimum distance also known as periastron[1], e the eccentricity of the ellipse and a his semi-major axis. P and T can be thought as *temporal* parameters while e and a being *spacial* parameters. The method to describe the orbit on any instant t is through the Kepler Equation showed in Equation 2.1, where M is the Mean Anomaly that correspond to a projection angle over a circumference centred in the focus point of the ellipse. The equation implies that the periodical motion projects a constant angular velocity in the Mean Anomaly M. By the other hand, the Eccentric Anomaly E is the projection angle over a circumference centred between both focus points of the ellipse, This projection does not follow a simple rule but relates to the Mean Anomaly M through Equation 2.1

$$2\pi \frac{t-T}{P} = E(t) - e\sin E(t) = M(t) \tag{2.1}$$

The Eccentric Anomaly E is a useful reference to calculate the exact position of the star B relative to a massive star A because it allows to determine the True Anomaly ν which represents the angle projection of the real position of the star in his elliptical motion path as showed in Figure 2.2. This relationship is specified in Equation 2.2.

$$\tan\frac{\nu(t)}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{E(t)}{2} \tag{2.2}$$

Finally, the absolute distance between the secondary star B and the primary star A is calculated using conic section mathematics of a well defined ellipse by the parameters $[e, a]$. This is described in Equation 2.3.

$$r(t) = \frac{a(1+e^2)}{1+e\cos\nu(t)} \tag{2.3}$$

The resultant pair $(r, \nu)_t$ corresponds to the polar coordinates of the star B relative to star A, the convention says that the reference $\nu = 0$ is set in the periastron reached periodically in $t = T + kP \quad \forall k \in \mathbb{Z}$.

The ability to theoretically relate every point of an orbit with its parameters $[T, P, e, a]$ allows to estimate (via observation points) parameters of a binary system, more importantly this relationship grants the possibility to estimate important physical properties of the studied stars, one of the most remarkable examples is the estimation of the sum of the masses of the participants stars which connect with their orbit via the Kepler Third Law, shown in

[1] due the periodical nature can also changed by $T + kP \quad \forall k \in \mathbb{Z}$

Equation 2.4, where C is a constant proportional to a, P, m_1 and m_2.

$$\frac{a^3}{P^2} = C(m_1 + m_2) \tag{2.4}$$

In the next subsections, it will be discussed how measures made over the binary system relate to their orbital parameters through phenomenological equations.

2.1.1. Astrometric data

Astrometry is one of the most ancient disciplines of astronomy and is dedicated to measure as precise as possible the position of celestial bodies in the sky and correctly document their respective movements. In current days, this astrometrics observations are taken by extremely high resolution telescopes which allows to study the dynamics of binary systems undetectable on naked eye. The practical techniques needed to obtain this data points are beyond the methodological scope of this work, for further studies a presentation of this methods can be founded in [Reffert, 2009]. The astrometric observations are performed over the sky-plane of observation which differs with the true orbit plane. In the previous subsection was studied how the relative kinematics of both stars are described when the observer is located perpendicular to the plane of movement, this parametrical representation can be adjusted to a more generalized form for addressing the scenario when the orbital plane of the system is rotated in a 3D fashion with respect to the sky-plane of observation (defined perpendicular to the observer), allowing to completely parametrize all the observable binary systems founded in nature. In order to accomplish this generalized model 3 rotational angles must be introduced.

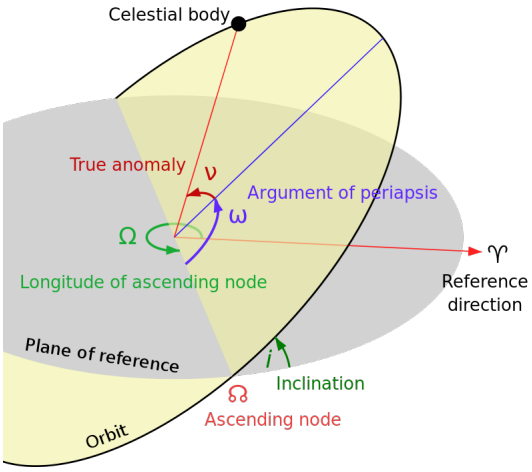


Figure 2.2: Orbital Parameters. Image Source: wikipedia.org

The Figure 2.2 depicts the projection of the orbit to a plane of reference where the observer is placed upside of the gray plane. This representation features three new angles to describe any possible orientation of the system relative to the sky-plane. In particular:

- i is the inclination angle and as its name suggests it determines the perpendicular

inclination of the orbit over the plane of reference.

- ω is the argument of periastron, its measures the angle discrepancy between the ascending node and the periastron.
- Ω is the Longitude of ascending node which relates the ascending node with the reference direction γ (Astronomical North) through an angle projected on the plane of reference itself.

The addition of new parameters to understand the astrometric observations induces an expansion of the parametrical space to represent the system leading to a new vector $[T, P, e, a, i, \omega, \Omega]$. Consequently by using the Thile-Innes representation it is possible to obtain the exact position of the observation at any given time t using an alternative parametrical vector $[T, P, e, A, B, F, G]$, being $[A, B, F, G]$ part of the Thiele-Innes representation. The derivation of this representation and the exact equations that determine the orbit are discussed in Appendix A.1 [2].

2.1.2. Spectroscopic data

Astrometric is not the only method used to measure the physical characteristics of a binary system. The eventual discrepancy of the orbital plane and the observational plane can suggest that the problem complexity will only rise with more data points and no new information of the system can be gathered, fortunately the spatial rotation itself allows to study movement of the system relative to the observer. At the start of Chapter 2 it was discussed the "dual representation" of the orbit and why a relative orbit is a better choice (due to mathematical complexity and simplicity of observation). Unfortunately the dual representation fuses the concepts of semi-major axes a_1 and a_2, described in Figure 2.1, in to a greater elliptical orbit of semi-major axis $a = a_1 + a_2$ resulting in to the loss of the ratio a_1/a_2, this missing information obfuscate the mass of each star making it unobtainable from only astrometric observations implying that in order to estimate these values a new type of observation must be introduced.

Spectroscopic observations are the measurement of the spectral shift of a celestial body induced by the radial velocity of the star with respect to an observer, this phenomenon is produced by the Doppler effect an directly affects the absorption and emission spectral lines location of the star. In Figure 2.3 both possible spectral shift are represented for the absorption spectrum of a star, the middle segment represents the natural spectral behavior for a star (derived from the known molecular composition of the star) an serve as basis to measure the shift. When the star moves away from the observer the perceived absorption lines move to a lower frequencies, this behavior is named red-shifting, on the other hand when the star is getting closer to the observer the absorption lines move to a higher frequencies and the phenomenon is named blue-shifting. In summary, measuring the magnitude and direction of the shifting effect induced by the Doppler effect allows to infer the radial velocity of the observed star.

[2] For further analysis on the described model the work of [van de Kamp, 1967] details all the matters regarding the definition of the Thile-Innes representation

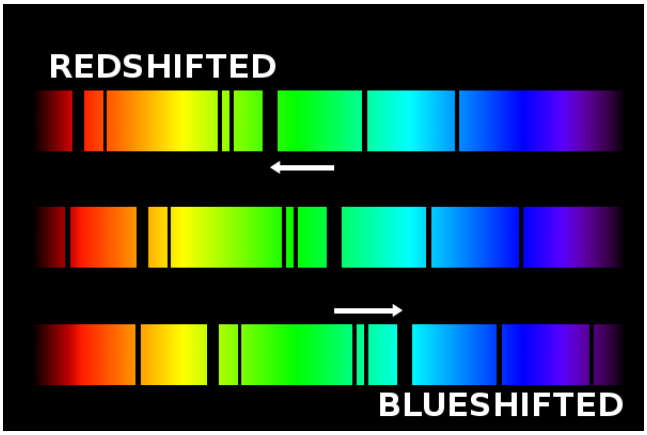


Figure 2.3: Doppler effect on absorption spectrum of a moving star. Image Source: wtamu.edu

Once the Radial Velocity of each star is estimated, it is possible to explicitly establish how the orbital parameters relate to the estimated values, this can be achieved by using vector derivatives of the model presented in Figure 2.2. By this approach, it is possible to find a closed form to express the Radial Velocity of each star as pictured in Equation 2.5. A detailed derivation of this results is described in [Claveria, 2017].

$$\begin{aligned} V_1(t) &= V_0 + \frac{2\pi a_1 \sin i}{P\sqrt{1-e^2}}[\cos(\omega + \nu(t)) + e\cos\omega] \\ V_2(t) &= V_0 - \frac{2\pi a_2 \sin i}{P\sqrt{1-e^2}}[\cos(\omega + \nu(t)) + e\cos\omega] \end{aligned} \tag{2.5}$$

To calculate exactly the velocity of each star is necessary to introduce three new parametrical variables: V_0 the constant velocity of the centre of mass of the system, π_p the parallax of the system, and q the quotient of mass between both participant bodies. The complete derivation of Equation 2.5 from the reference scheme presented in Figure 2.2 is described in the Appendix A.2. This new phenomenological equation results in a updated parametrical representation of the whole system dynamics composed as $[T, P, e, a, i, \omega, \Omega, V_0, \pi_p, q]$.

As presented in Section 2.1.1 an alternative vectorization for the system is possible by introducing the variables H and C, inducing a new parametrical vector $[T, P, e, i, \Omega, V_0, \pi_p, q, H, C]$, this representation will be important for future practical matters, one of his advantages is the linear dependence between the parameters, allowing to reduce the dimensional complexity of the representation into $[T, P, e, i, \Omega, \pi_p, q, H, C]$. This new vector is first introduced and derived in [Mendez et al., 2017] and plays an important role in the performance of the inference algorithm.

This work is focused on astrometric and spectroscopic observations, and how the parameters of the system interact with them. The binary system will be completely determined by its respective parametrical vector denoted by θ and the aim of this work is to study how these phenomenological equations determine the knowledge of the system.

2.2. Observation Model

In order to measure stochastic relationships between the data and the binary system is crucial to understand, or more importantly to model, the phenomenological act of observing the system in any of their physical states. The reason to set the problem in this way is to address the fact that noisy observations change the statistics of the parametrical space. The intuition says that a observation free from any source of noise is unpractical for real case scenarios, from this perspective a probabilistic approach allows relating the observations of the system to its parameters,

In this section it is proposed and formalized a probabilistic approach to model the act of observation by means of specific equations of the system exposed presented in Section 2.1 and the use of Central Limit Theorem.

2.2.1. Predictive Distribution

An observation model is proposed to obtain samples coming from a probability distribution that is function of the parameters of the system. The existing dependency means that every observation, as sample of a random variable, carries information about the parameter state of the system. Now in order to make explicit the relationship between the parameters and the observations we need to define the family of functions $\mathcal{M}$ that is indexed by an specific moment of time t. A function of this class at time t is a function and describe the function $M^t : \mathcal{O} \to \mathcal{Y}$, which maps a defined system a particular value for $\theta \in \mathcal{O}$ to a point in the observation space $\mathcal{Y}$. By this means $M^t(\theta)$ denotes the observation in the instant t for a system completely defined by $\theta \in \mathcal{O}$. As discussed before, we will consider two kinds of measurements, astrometric and spectroscopic which are defined in the observable plane (in *mas*) and RV spectre (in kms^{-1}) respectively. Each of them has its own associated physical equations, as mentioned in the previous section, and its consequently own respective family of functions M_* with a particular subindex $*$ to differentiate between them.

With the intent of maintaining abstraction in the representation of the observational model, we are going to refer to a generic family M which can be any kind of observation over a particular system; so any methodological analysis on this problem can be immediately translated to astrometric, spectroscopic or any other observation available from the studied system. In this context, it will be assumed that any indexation t over the family of functions $\mathcal{M}$ will have the same codomain $\mathcal{Y}$ for any case. A more generalized description for the family of functions $\mathcal{M}$ can be achieved by adding a super index on the codomain, this re-definition allows to introduce in the analysis multiple kinds of observation, which in a practical setting is often the case. In this initial analysis, and for the sake of simplicity and avoiding heavy mathematical notation, the codomain $\mathcal{Y}$ will be set as the same for any index t. In future practical analysis on astrometric and spectroscopic observations the results from this chapter will be extended to fit the practical setting.

Using a Bayesian approach requires that the objects of interest, in this case the parameters and observation, are defined as random variables. We denote the random variable for the parameters by $\Theta \in \mathcal{O}$ and its distribution p_Θ, the prior distribution p_Θ represents the previous

knowledge of the system over the parameter space before being measured or observed. In many cases this distribution is set uniform over $\mathcal{O}$ to avoid any bias on future estimations, in other words prior information on the parametrical space says that all possibles values for $\theta \in \Theta$ have the same probability of being the real parametrical state of the system. This uniform assumption is not mandatory, so under other assumptions the prior distribution can take any desirable shape. In similar fashion, we introduce a collection of random variables related to the physical observable states of the system as $Y := \{Y^t | t \in \mathcal{T}\}$, where each random variable $Y^t \in \mathcal{Y}$ and its distribution p_{Y^t} describes observations made over the system, p_{Y^t} relates to the knowledge of the system Θ through the phenomenological function M^t. Unlike the previous definition of Θ, every random variable $Y^t \in Y$ has its distribution p_{Y^t} free from any designer choice, this is because its shape is explicitly defined from p_Θ and the relationship between both random variables introduced as the conditional distribution $p_{Y^t|\Theta}$. In order to explicitly calculate p_{Y^t} it is necessary to define $p_{Y^t|\Theta}$. For this particular case of study the relationship between the parameters and observation is strictly deterministic, meaning that for each $\theta \in \Theta \quad \exists! y^t \in Y^t$ meaning that the definition of the conditional distribution $p_{Y^t|\Theta}$ must be established as shown in Equation 2.6, being this the observational case when no noise is perceived.

$$\forall t \in \mathcal{T} \qquad p_{Y^t|\Theta}(y^t|\theta) = \delta_{M^t(\theta)}(y^t) \tag{2.6}$$

where $M^t(\theta)$ is the real observable state given a system characterized by θ at the instant t and δ_c is the Dirac Delta function centered at c. The Equation 2.6 evidences a deterministic dependency of Y^t under Θ in a statistical approach. This behavior is observed due the uniqueness of state for a well defined system, in other words, a set system θ will only project one observable state at index t, this phenomenon is ruled by the deterministic function M^t and will take the specific value $M^t(\theta)$.

It is extremely important to understand that the reciprocal dependency only holds for those functions M^t where $M^{t^{-1}}$ exists, which is not often in astronomical settings. The functions M^t are mostly composed by non linear relationships between the parameters θ meaning that the link between Θ and Y^t is not unique. In general there exists a collection of $\{\theta_i\}_{i=1}^N$ denoted by $\mathcal{S}_{y_s,t}$, which project the same physical state $y_s \in \mathcal{Y}$ at the moment t holding the following expression.

$$M^t(\theta_i) = M^t(\theta_j) = y_s \quad \forall \theta_i, \theta_j \in \mathcal{S}_{y_s,t} \tag{2.7}$$

Pursuing the same line of thought it is possible to infer that the distributions p_θ and p_{Y^t}, related by this complex mapping M^t and the marginal distribution of Y^t (through Θ), can be explained via probabilistic marginalization, as shown in Eq 2.8.

$$\begin{aligned} p_{Y^t}(y^t) &= \int_{\mathcal{O}} p_{Y^t|\Theta}(y^t|\theta) p_\Theta(\theta) d\theta \\ &= \sum_{\theta \in \mathcal{S}_{y_s,t}} p_\Theta(\theta) \end{aligned} \tag{2.8}$$

Under this context p_{Y^t} is named as the Prior Predictive Distribution or Prior-PD and represents the prior knowledge of the system but expressed in the observational space $\mathcal{Y}$ instead of the parametric one $\mathcal{O}$. In a practical sense, this marginalization is extremely

complex or possibly untracteable because of the absence of a close expression for $\mathcal{S}_{y_s,t}$ in Equation 2.7. So a direct integration would be prohibitive, nonetheless the Equation 2.8 will be useful in further derivations.

2.2.2. Observational Predictive Distribution

Once all the natural relationships between the system and its projections over the observational space $\mathcal{O}$ has been made, we can to introduce the noise in the phenomenological act of observation. The intuition says that any measure made over a phenomena is always corrupted to a certain point by exogenous agents, which add non related information to the observation. This noise implies that the deterministic behavior described in Equation 2.6 is no longer faithful and a new random variable must be introduced in order to formalize this idea.

Let us denote the random variable $Y_\alpha^t \in \mathcal{Y}$ as the observation over the system at moment t. The uncertainty of this process is captured by a fixed parameter α (physical limitations and instrumental noise), which will parametrize the distribution $p_{Y_\alpha^t|Y^t}$. The most important difference between $p_{Y^t|\Theta}$ in Equation 2.6 and $p_{Y_\alpha^t|Y^t}$ is that the lost observational distribution can no longer be deterministic. Consequently we need to specify the shape on $p_{Y_\alpha^t|Y^t}$, this election is contained in α and will rule the theoretical behavior of the observer.

As observed in Equation 2.8 the characterization of $p_{Y_\alpha^t}$ is obtained via marginalization of the previously discussed distribution $p_{Y_\alpha^t|Y^t}$ as shown in Equation 2.9.

$$\begin{aligned} p_{Y_\alpha^t}(y_\alpha^t) &= \int_{\mathcal{Y}} p_{Y_\alpha^t|Y^t}(y_\alpha^t|y^t) p_{Y^t}(y^t) dy^t \\ &= \int_{\mathcal{O}} p_{Y_\alpha^t|\Theta}(y_\alpha^t|\theta) p_\Theta(\theta) d\theta \end{aligned} \tag{2.9}$$

The distribution of Y_α^t in Equation 2.9 shows two clear characterization of $p_{Y_\alpha^t|\Theta}$ through $p_{Y_\alpha^t|Y^t}$ and $p_{Y_\alpha^t|\Theta}$ respectively. In particular, we have a deterministic relationship between Θ and Y^t, as showed in Equation 2.6, meaning that the derivation of $p_{Y_\alpha^t|\Theta}$ can be performed by the definition of $p_{Y_\alpha^t|Y^t}$ and the mapping M^t through change of variable.

In formal terms, $p_{Y_\alpha^t}$ will be referred as the Prior Observational Predictive Distribution or Prior-OPD and it represents the actual belief of how the system will be observed at the moment t by a instrument paremetrized by α.

It is important to mention that the projection of the prior distribution p_Θ through observational variables, such as Y^t which are absolutely or partially dependant of our prior distribution, can be pictured as a generative model. The Figure 2.4 represents this projection of the prior distribution to observational variables by defining a graph diagram where the arrows that connect the random variables are the conditional distributions. These links hold a particular direction implying a non trivial relationship for the reciprocal statement (one-direction generative model). In fact the characterization of $p_{\Theta|Y_\alpha^t}$ it is a problem by itself that is addressed in the Bayesian Inference framework. One last remark about the observation model is that any modification regarding p will lead to different characterization of p_{Y^t} and $p_{Y_\alpha^t}$, meaning that any new information about the system presented in Figure 2.4

changes the probabilistic relationship between Θ and Y_α^t. This point will be crucial for future derivation for the Optimal Sampling Method.

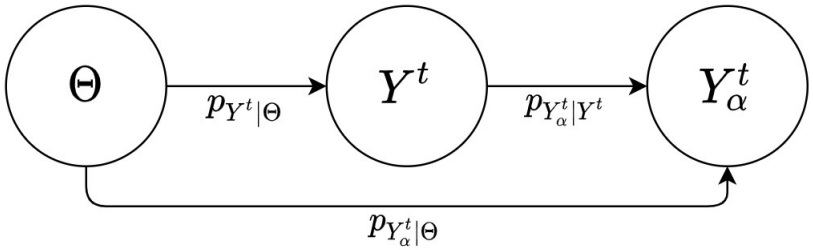


Figure 2.4: Observational Model.

Most of the time a particular election of $p_{Y_\alpha^t|Y^t}$ is considered. In instrumentation, the observational noise is usually modeled as the addition of N external sources of information which are statistically independent from the actual phenomena of interest . This model presents some mathematical challenges due to the addition of several external random variables, then in order to characterize this noise we need to determine the shape its distributions. An incorrect choice of this distribution could lead to a biased estimation of the instrumental noise. Importantly, the Central Limit Theorem states, in general, if we have N *i.i.d.* sources with a density distribution that holds a bounded variance, then as $N \to \infty$, the sum of those noised sources will follow a normal distribution. This theorem implies that assuming this general hypothesis and choosing a expected value 0 for each exogenous source will lead to a correct characterization of the instrument noise and no systematic effects. Therefore we can establish $p_{Y_\alpha^t|Y^t}(y_\alpha^t|y^t) \sim \mathcal{N}(y^t, \alpha)$ and alternatively $p_{Y_\alpha^t|\Theta}(y_\alpha^t|\theta) \sim \mathcal{N}(M(\theta)^t, \alpha)$ being α a diagonal co-variance matrix which implies independent sources of noise in each observational dimension.

2.3. Bayesian Framework

2.3.1. Bayesian Inference

Bayesian Inference (B.I.) is a statistical method used to model how the knowledge of a variable is affected after a observation of another variable has been performed. This theory is named after the Bayes Theorem which serves as the motor of this method. Bayesian Inference is relevant for this work because allows to relate the observational model on Section 2.2.2. with the acquisition of information through the explicit equations presented on Section 2.1.. The mathematical object that represents this relationship is the posterior distribution $p_{\Theta|D}$, the shape of this distribution can be inferred through the Bayes theorem which can give a closed characterization for $p_{\Theta|D}$.

Let $\Theta \in \mathcal{O}$ be the parametrical vector, the main focus of the inference, and let $D \in \mathcal{Y}^N$ be a vector of observations denoted by the collection $D = \{Y_{\alpha_i}^{t_i} | i \in [0, N]\}$. These two objects follow the observational model described in Figure 2.4 and will be statistically modeled in order to perform the inference. The main objective of the Bayesian inference is to characterize the posterior distribution $p_{\Theta|D}$ but, as mentioned in the previous section, the observational model does not define explicitly term, so a numerical method to approximate this value will be applied. The Equation 2.10 states the Bayes Theorem using the previously presented random variables Θ and D, where after a collection of observations has been made (given by a particular set of values d_i) the posterior distribution $p_{\Theta|D}(\theta|d) \ \ \forall \theta \in \mathcal{O}$ can be determined by means of Equation 2.10.

$$\begin{aligned} p_{\Theta|D}(\theta|d) &= \frac{p_{D|\Theta}(d|\theta)p_{\Theta}(\theta)}{p_D(d)} \\ &= \frac{\prod_{i=1}^{N} p_{Y_{\alpha_i}^{t_i}|\Theta}(d_i|\theta)p_{\Theta}(\theta)}{\prod_{i=1}^{N} p_{Y_{\alpha_i}^{t_i}}(d_i)} \\ &\propto \prod_{i=1}^{N} p_{Y_{\alpha_i}^{t_i}|\Theta}(d_i|\theta)p_{\Theta}(\theta) \end{aligned} \tag{2.10}$$

The conditional distribution $p_{\Theta|D}$ can be interpreted as a set of distributions where for each particular $\bar{d} \in \mathcal{D}$ a well defined distribution $p_{\Theta|D}(\theta|\bar{d}) \ \ \forall \theta \in \mathcal{O}$ is calculated. In other words, D can be thought as a indexing variable to obtain a distribution for Θ. Under a practical scope, a full characterization of $p_{\Theta|D} \quad \forall d \in \mathcal{D}$ is not much informative until a point for D takes a specific value $\bar{d}$. Therefore, the only relevant distribution, from a practical approach, is $p_{\Theta|D}(\theta|\bar{d})$, representing the knowledge acquired from $\bar{d}$ after the observation is made. The last important aspect of Bayesian inference is the last expression presented in Equation 2.10 where the denominator is replaced by a proportional constant. This is made for two reasons: first the observation sample $\bar{d}$ is constant for all θ, so there is no need to evaluate that expression to obtain a proportional term. Second the fact that we do not have a simple closed formula for $p_{Y_{\alpha_i}^{t_i}}$ being the only partial description the one presented in Equation 2.9. In discrete settings this last expression is crucial because the proportional term can be computed easily (by going through each value on $\mathcal{O}$ and then applying a simple normalization for each

case). In contrast, in the continuous setting, the parametric estimation of binary systems is particularly challenging due the intrinsic continuous nature of its probabilistic distributions, however for this kind of cases the existence of sampling method allows to perform a partial representation of the posterior distribution $p_{\Theta|D}$ through samples. Although this simulation does not provide a close mathematical expression for the Equation 2.10 the samples obtained enable the empirical estimation, such as expected value, variance or other metrics regarding the shape of $p_{|D}$.

The inference process previously discussed gave a methodological approach to obtain the posterior distribution $p_{\Theta|D}$ of a observable system. Following the idea of Section 2.2 a new predictive flow can be made using the same diagram featured in Figure 2.4 to generate the distributions $p_{Y^t|D}$ and $p_{Y^t_\alpha|D}$, which are denoted as the Posterior Predictive Distribution or Post-PD and the Posterior Observational Distribution or Post-OPD respectively.

2.3.2. Sequential Bayesian Inference

In order to address the main objective in this work, the selection of a new instant t of observation given previous data of the system. It is interesting to study another flavor of the bayesian framework. Once data have been acquired from the system, it is possible to process this information and transform it in the posterior distribution $p_{\Theta|D}$. Once a new observation over the system is performed it is important to rethink the inference as a sequential and recursive process.

$$\begin{aligned} p_{\Theta|D^{N-1},Y^{t_N}_{\alpha_N}}(\theta|d^{N-1}, y^{t_N}_{\alpha_N}) &= \frac{\prod_{i=1}^{N} p_{Y^{t_i}_{\alpha_i}|\Theta}(d_i|\theta)p_\Theta(\theta)}{\prod_{i=1}^{N} p_{Y^{t_i}_{\alpha_i}}(d_i)} \\ &= \frac{p_{Y^{t_N}_{\alpha_N}|\Theta}(y^{t_N}_{\alpha_N}|\theta)\prod_{i=1}^{N-1} p_{Y^{t_i}_{\alpha_i}|\Theta}(d_i|\theta)p_\Theta(\theta)}{p_{Y^{t_N}_{\alpha_N}}(y^{t_N}_{\alpha_N})\prod_{i=1}^{N-1} p_{Y^{t_i}_{\alpha_i}}(d_i)} \\ &= \frac{p_{Y^{t_N}_{\alpha_N}|\Theta}(y^{t_N}_{\alpha_N}|\theta)p_{\Theta|D^{N-1}}(\theta|d^{N-1})}{p_{Y^{t_N}_{\alpha_N}}(y^{t_N}_{\alpha_N})} \\ &\propto p_{Y^{t_N}_{\alpha_N}|\Theta}(y^{t_N}_{\alpha_N}|\theta)p_{\Theta|D^{N-1}}(\theta|d^{N-1}) \end{aligned} \tag{2.11}$$

The sequential inference is the process where in each step of inference, besides the first instance, a redefinition on the actual prior distribution is performed using the previously estimated posterior distribution. Therefore in each instant of inference the acquired knowledge of the system is updated to allow future estimations. The theoretical justification of the sequential inference is showed in Equation 2.11 where the final posterior distribution using all the data is proved proportional to the one-step bayesian inference formula but using an updated version of the prior distribution. This sequential approach is important because the problem of optimal sampling has inherently the same structure.

2.4. Information Theory

Information Theory is a statistics field widely use in communication applications among other scientific frameworks, it uses probabilistic and statistics theory to model information sources. This theory was originally proposed by [Shannon, 1948] and is considered one of the most revolutionaries scientific works of all time. The concepts presented and defined in this paper were crucial to understand of communications systems as we know them today. The main objects of study in Information Theory are information sources which are defined as random variables completely characterized by their respective distribution. This last mathematical concept rule the behavior of the random variable so by studying its shape will allow to understand the intrinsic properties of this information source.

The reason why Information Theory is relevant in this work is because allows to measure information transfer from the observations to the parameters in the context of Bayesian Inference. In order to determine an optimal sampling instant t a metric will be established to decide if a particular election t_i is a better choice than t_j, this will me performed by means of exploiting deep statistical relationship between the parametrical space and the observational space.

2.4.1. Entropy

The first object of information is the study of a random variable. The entropy is the most recognized concept in Information Theory and it is often interpreted as how much randomness has a variable. The entropy of a random variable $X \in \mathcal{X}$ is denoted as $H(X)$, the definition of entropy is presented in Equation 2.12, being $H(.)$ a mapping from the space of possibles distributions p_X to $\mathbb{R}$. This expression is upper bounded by the entropy of a random variable $X_u \in \mathcal{X}$ that has a distribution $p_{X_u} \sim \text{uniform}(\mathcal{X})$ and lower bounded by the entropy of a completely deterministic variable $X_d \in \mathcal{X}$ with distribution $p_{X_d} \sim 1_{x_c}(x) \quad x_c \in \mathcal{X}$. The Equation 2.12 is composed as the sum of a term dependent on the probability value in each element of the support of X.

$$\begin{aligned} H(X) &= -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) \\ 0 = H(X_d) &\leq H(X) \leq H(X_u) \end{aligned} \tag{2.12}$$

It is easy to notice that randomness of a variable and his entropy are tightly related in the extreme cases of absolute uncertainty (uniform distribution) or deterministic behavior (degenerate distribution). The non-extreme scenarios can be explained by other analysis, for example applications in the communication field. The problem of lossless source coding described in [Cover and Thomas, 2006] is defined as the search of an optimal rule to map $\mathcal{X}$ to a codebook $\mathcal{C}$, which contains codewords of different length, the optimal rule must guarantee the minimal expected length for the codewords. The result presented in [Cover and Thomas, 2006] says that the optimal expected length of a codeword for the description of an arbitrary discrete random variable X is proportional to the entropy $H(X)$, meaning that if X concentrates a big portion of his probability mass in certain symbols of

$\mathcal{X}$ the optimal codebook will assign short codewords to those symbols. The optimal code presented in [Cover and Thomas, 2006] can be interpreted as the relationship between the difficulty to understand the stochastic behavior of a random variable and the complexity of coding that information. This connection is one of the reason that entropy is recognized and used as a measure of uncertainty for a random variable in many applications.

The definition of entropy presented in [Shannon, 1948] only addresses the discrete scenario, this is because in communication applications the information sources are mostly modeled in discrete spaces. In contrast, the objective of this work is set in a continuous space following the intrinsic nature of the binary systems, then a extension of Equation 2.12 must be defined. The differential entropy corresponds to the natural extension of the entropy where the sum is replaced by an integral. This change can be interpreted as insignificant but unfortunately leads to the loss of some properties. The differential entropy is still upper bounded by differential entropy of a uniform distribution in $\mathcal{X}$ however the lower bound is lost, implying that the differential entropy can reach any negative value. This change hinders the interpretation of the differential entropy, unlike the discrete scenario where the entropy is always semi-positive and reaching the boundary case 0 when the random variable is degenerate. In contrast, the differential entropy lack of some of the practical interpretations, implying that any result derived from this concept will only serve theoretical proposes.

$$H(X) = -\int_{x\in\mathcal{X}} p_X(x)\log p_X(x)dx \tag{2.13}$$

The discrete entropy and the differential entropy relate a single random variable to a numerical value in $\mathbb{R}$ being only a self-description of a random object. An interesting analysis is to study two random variables interacting with each other and the effect of this relationship in a entropy-wise manner. For this we introduce the concept of conditional entropy as shown in Eq 2.14.

$$H(X|Z) = -\int_{\mathcal{Z}}\int_{\mathcal{X}} p_{X,Z}(x,z)\log p_{X|Z}(x|z)dxdz \tag{2.14}$$

$$\begin{aligned} H(X|Z) &= -\int_{\mathcal{Z}}\int_{\mathcal{X}} p_{X,Z}(x,z)\log p_{X|Z}(x|z)dxdz \\ &= \int_{\mathcal{Z}} p_Z(z)\underbrace{-\int_{\mathcal{X}} p_{X|Z}(x|z)\log p_{X|Z}(x|z)dx}_{H(X|z)}dz \\ &= \int_{\mathcal{Z}} p_Z(z)H(X|Z=z)dz \\ &= E_Z[H(X|Z=z)] \end{aligned} \tag{2.15}$$

The conditional entropy is most known by the first expression presented in Equation 2.14, but an alternative derivation is presented in Equation 2.15. This adds a intuitive interpretation to measure the statistical relationship between two random variables, it is crucial to understand that each particular value z of Z affects the distribution $p_{X|Z}(x|z)$, being this distribution the mathematical object that relates X to Z. Rhis implies that

the definition of $H(X|Z)$ forms of expectation over the random variable Z. Intuitively the conditional entropy is a measure of the uncertainty from the random variable X after Z has been observed.

2.4.2. Mutual Information

The characterization of the conditional differential entropy, presented in Equation 2.14, does not hold commutativity between X and Z. The intuition says that a measure of information should represent, symmetrically, the relationship between 2 information sources, so the next step must be the definition of expression which quantifies the difference in uncertainty of a variable X when an external variable Z is observed. The Mutual Information $I(X;Z)$ is a measure that represents the shared information between two random variables and how much one determines the other an vice versa. The explicit definition is presented in Equation 2.16.

$$\begin{aligned} I(X;Z) &= \int_{\mathcal{Z}} \int_{\mathcal{X}} p_{X,Z}(x,z) \log \frac{p_{X,Z}(x,z)}{p_X(x)p_Z(z)} dxdz \\ &= H(X) - H(X|Z) \\ &= H(Z) - H(Z|X) \geq 0 \end{aligned} \tag{2.16}$$

The Mutual Information is a fundamental measure in Information Theory, it has been used in many applications such as feature selection [Vergara and Estevez, 2014][Beraha et al., 2019], independence test [Berrett and Samworth, 2017], neural networks information bottleneck [Tishby et al., 2001] and image segmentation [Junmo Kim et al., 2005], among other cases. The Equation 2.16 shows two expressions for the Mutual Information which serve different proposes, the first one represents the measure by using the joint and marginals distributions of X and Z. This characterization is useful because it allows to connect this concept of information with the Kullback-Leibler divergence denoted as D_{KL} in Equation 2.17, this last term is a semi-measure of distance between two distributions which reaches its minimum value at 0 when both distributions are exactly equal. Thus the Mutual Information can be pictured as $D_{KL}(p_{X,Z}||p_X p_Z)$ this interpretation is heavily linked to the setting of statistical independence because measures the discrepancy of the $p_{X,Z}$ to its null hypothesis of independence, given by $p_X p_Z$.

$$D_{KL}(p_{X_1}||p_{X_2}) = \int \int p_{X_1}(x_1) \log \frac{p_{X_1}(x_1)}{p_{X_2}(x_2)} dx_1 dx_2 \tag{2.17}$$

On the other hand, the last expression in Equation 2.16 can also be interpreted as the difference between entropy of a variable and the same variable after an observation of the other has been made. One important aspect to discuss from $I(X;Z)$ is its extreme values, the Mutual Information is lower bounded by 0 and this condition is reached when both variables are independent (using the D_{kl} interpretation), mathematically this condition is expressed by $H(X) = H(X|Z)$. On the other extreme, the Mutual Information upper boundary is $\min[H(X), H(Z)]$, this happens when either $H(X|Z) = 0$ or $H(Z|X)$. The upper bound scenario happens when a deterministic relationship exists between X and Y and behaves exactly as Equation 2.6. This last case can be interpreted as one variable been absolutely

descriptive from the other variable.

The last important aspect to look at from the Mutual Information it is its semidefinite positivity, this implies that it is always expected a information gain between two information sources. However it is crucial to remember that the concept of conditional entropy, included in the definition of Mutual information in Equation 2.15, is defined as an expectation, therefore it is completely plausible that for a subset $\mathcal{Z}' \subset \mathcal{Z}$ that $H(X|Z = z) > H(X) \quad \forall z \in \mathcal{Z}'$. In a observational setting is completely possible that a new data point could increase the entropy of the inferred space even if this sampling moment was hand-picked to be optimal under a certain criteria. It is possible that samples obtained via well designed criterion, that uses measures based on expected value of a statistical object, could imply an increase in uncertainty in the inferred space despite being optimal before the optimal observation is made.

In summary, Information Theoretic measures are a set of mathematical concepts useful in the characterization of knowledge and certainty for many inference problems, being this the reason why we are interested in considering Mutual Information and conditional entropy for definition the optimal sampling strategy. We will use those concepts to formalize the optimization of our optimal sampling problem in orbital parameter estimation.

Chapter 3

Related Work

3.1. Inference

Obtaining the actual set of parameters from a binary system through observations has been the main focus of studying stellar astronomy for many years. Usually, in order to obtain a particular value for the parameters, an inference method using the available information and data is performed. Through the history of astronomy, many authors and researchers have been addressed this problem and proposed a large number of methods; those attempts vary both on mathematical tools and in the philosophical approach. This section will briefly discuss the history of parametrical inference for binary systems until reaching the current state of the art on this matter.

3.1.1. Deterministic Inference

Back in the early 20th century, the first widely accepted method to obtain the orbital parameters of a binary system was presented, often referred to as the Thiele-Innes-Van de Boss method inspired by the works in [Thiele, 1883][Thiele, 1926][den Bos W. H, 1926]. This scheme mainly uses three astrometric observations denoted by the polar coordinates of the less bright star relative to the brighter one (ρ, θ, t) and the estimation of the constant c, which corresponds to the areal velocity, related to the Kepler's Second Law obtained via external information. Using the proposed estimation was possible to obtain a parametrical representation for the system and consequently a solution candidate for the problem. Following the same approach as the Thiele-Innes-Van de Boss method, a new variant was proposed in [Cid Palacios, 1958], where instead of using the constant c a partial astrometric observation (considering only the angular and temporal components of the original tuple (θ, t)) was required to find the best suitable orbit. One of the last versions inspired in this principle is presented in [Docobo, 1985] where the extra piece of information besides the three tuples (ρ, θ, t) is an auxiliary angular variable denoted as V which induces a feasible set of orbits E according to the data. The three methods presented above share the same methodology in their approaches where the data $(\rho_i, \theta_i, t_i)_{i=1}^{3}$ and the extra measure (c, (θ_4, t_4) or V) will project a set of plausible orbits for the system. Under this paradigm, a set of parameters considered as a solution must project an orbit in the sky-plane that describes a perfect path

for all astrometric points measured. In section 2.2 it was extensively discussed how noise affected the observations of a system irremediably. Consequently, the statement that each orbit is solution must match perfectly the data can be picture a strong miss-modeling and naturally will lead to a not representative solution for the system.

The next generation of methods addressing the orbital parameter inference problem was mainly focused on modeling the reliability involved in each astrometric measure. This hypothesis allowed to infer orbital parameters for systems that had observations which differed in the grade of certainty between each other meaning that some data point were more accurate in comparison to others, the main theoretical focus for this kind of methods was the definition of a fitness function, being in each case parametrized for the observations (ρ_i, θ_i, t_i) itself and a constant w_i acting as a weighted certainty relative to the rest of data. The primary challenge for this setting was the inherent optimization problem associated with the search of an optimal solution relative to the fitness function, in [Tokovinin, 1992] for example, this problem is addressed by the Least Squares iterative method. At the same time, in [MacKnight and Horch, 2004] the Nelder-Mead method, also known as the downhill simplex method, is applied; these new approaches allowed to solve systems that, under the first generation of estimators, had contradictory data (observations that mismatch the predicted model). Naturally, this new focus on estimators is considered, by the community interested in binary systems, as an upgrade compared to earlier methods. Unfortunately, all the issues discussed in section 2.2 and 2.3 were not addressed in this scheme because the inference is still deterministic, meaning that a new approach for parametrical inference of binary systems must be established.

The work of [Pourbaix, 1994] was one of the first steps towards a stochastic characterization of the inference problem. The proposed method used the simulated annealing paradigm from [Metropolis et al., 1953] in order to optimize the fitness function. This approach did not use the full potential of simulated annealing [Metropolis et al., 1953], and the result lacked the properties of a Bayesian inference setting.

3.1.2. Probabilistic Inference

The Bayesian approach was not explored until [Ford, 2005] published a method based on Monte Carlo Markov Chains (MCMC) simulation to study the orbit of exoplanets, the particles generated by the MCMC method were taken as samples from the posterior distribution, and then an inference was possible. Due to the evident similarities between the study of orbits for exoplanets and binary systems, an extrapolation between both problems was natural, the works of [Lucy, 2014] and [Sahlmann et al., 2013] understood these similarities and adopted the same framework to address the problem of estimation of orbital parameters of binary systems. The main three advantage of this new simulation oriented method were: 1) robustness in solution for systems hard to estimate, 2) the probabilistic approach that gave a characterization of the posterior distribution instead of only a deterministic decision, and 3) the use of computational power to automatize the simulations quickly.

The individual masses of each star in a binary system are one of the most important physical aspects to study. In section 2.1 we discussed how astrometric data only allows inferring the sum of the masses, implying that in order to estimate these individual masses,

the spectroscopic measures must be introduced. One of the first works that involved a joint characterization of astrometric and spectroscopic data was [Pourbaix, 1998], which uses the same approach proposed in [Pourbaix, 1994] (simulated annealing). On the other hand, one of the most recent methods addressing the orbital parameter inference using both kinds of observation is presented in [Mendez et al., 2017], where through Gibbs-MCMC simulation, the posterior inference obtained is characterized by the generated samples, and consequently, a measure of certainty can be estimated using these particles.

The current state of the art for inference of orbital parameters is held by Bayesian approaches that use simulation techniques to obtain posterior distribution samples. This simulation methodology is implemented in several study cases, for example, using the samples to characterize the posterior distribution explicitly, allowing to study the relationship between the parametrical space and the observations [Lucy, 2014], analyzing the shape of the marginal distributions and evaluate the quality of the inference [Mendez et al., 2017]. The current methods and computational power available allow researchers to infer the parameters over hundreds of binary systems with a wide variety of observational qualities. This is the main reason why any observation available about the binary systems is crucial because it is the only method to improve the knowledge over any particular system that is observable. This search of more observations stimulates more recent works such as [Claveria, 2017], where imputation techniques are implemented in order to include partial or corrupted observations. From a different scope but under the same paradigm, the need for data encourages the development of new scheduling observation methods that exploit the posterior characterization of the inference and then better use the available observational resources.

3.2. New sample selection

3.2.1. Bayesian Experiment Design

In 1948 the mathematician and electrical engineer Claude Shannon published his work "A Mathematical Theory of Communication"[Shannon, 1948] and started a revolution in all fields related to communication. The metrics introduced in [Shannon, 1948] were not only relevant for this particular discipline of engineering, but the statistical approach presented in the work of Shannon was also interesting for mathematicians involved in Bayesian inference. For example the works [Blackwell, 1953] and [Blackwell and Girshick, 1979] were focused on modeling experiments and studying how two experiments can be considered as equivalent under the scope of information measures.

The increasing number of publications related to [Shannon, 1948] in conjunction with [Blackwell, 1953] and [Blackwell and Girshick, 1979] were the initial motivation to explore how information measures are an essential in Bayesian inference. As a result of this, the work [Chaloner and Verdinelli, 1956] presented the foundations of what would later be referred as Bayesian Experiment Design. The main focus of this paper was introducing the definition of information related to an experiment and how this value is closely dependent on the particular sample obtained in a realization of such experiment. The work starts by defining the intrinsic information $\mathcal{I}_0$ related to Θ, the variable to infer, and the information $\mathcal{I}_1$ over Θ after an observation of X was made, Both concepts are defined in Equation 3.1.

$$\begin{aligned} \mathcal{I}_0 &= \int_{\mathcal{O}} p_\Theta(\theta) \log p_\Theta(\theta) d\theta \\ \mathcal{I}_1(x) &= \int_{\mathcal{O}} p_{\Theta|X}(\theta|x) \log p_{\Theta|X}(\theta|x) d\theta \end{aligned} \tag{3.1}$$

Once the main measures regarding the Equation 3.1 are described, the next step is the mathematical definition of an experiment and how this object is related to an inference problem. In [Chaloner and Verdinelli, 1956] an experiment is denoted by the tuple $\epsilon = \{X, \mathcal{B}, \Theta, P\}$ being $X \in \mathcal{X}$ the observational random variable, its $\mathcal{B}$ σ-field subsets in $\mathcal{X}$, p_Θa prior distribution for Θ and P the set of probability measures $p_{X|\Theta}$ $\forall \theta \in \mathcal{O}$. This tuple describes the act of observation by using the likelihood of each possible observational outcome contained in the set P and the prior knowledge over the variable to infer p_Θ. The information gained in an experiment ϵ given an observational realization x is defined in Equation 3.2, which is the difference between the posterior information $\mathcal{I}_1(x)$ and the prior information $\mathcal{I}_0$ previously defined.

$$\mathcal{I}(\epsilon, p_\Theta(\theta), x) = \mathcal{I}_1(x) - \mathcal{I}_0 \tag{3.2}$$

It is worth pointing out that studying on experiments cannot be a function dependent of particular values x, because in a practical setting the outcome of an experiment cannot be known prior to its execution. So an expected value analysis must be performed; therefore, the definition of information obtained via a certain experiment ϵ is defined by the expected

value of the Equation 3.2 relative to X and is made explicit through Equation 3.3.

$$\mathcal{I}(\epsilon, p_\Theta(\theta)) = \mathbb{E}_X \left[\mathcal{I}_1(x) - \mathcal{I}_0\right] \tag{3.3}$$

The analysis made in [Chaloner and Verdinelli, 1956] regarding Equation 3.3 introduces cases of study for two or more candidate experiments, consequently also explores the properties existent in scenarios where these experiments are performed simultaneously, implying a study of joint information from such experiments. The similarities between mutual information of two random variables and information of an experiment are evident when comparing Equation 2.16 and Equation 3.3, this similarity can be noticed more evidently when observing the derivations presented in [Chaloner and Verdinelli, 1956] are known properties of the mutual information discussed in section 2.4. Lastly this work also features extensions for theorems presented in [Blackwell, 1953] and [Blackwell and Girshick, 1979]. On the other hand, the definitions presented in Equation 3.1 have no immediate interpretation; this happens because the values for each experiment analyzed are always negative and have no attachment to any intuition. Despite these drawbacks, the work presented in [Chaloner and Verdinelli, 1956] is one of the most solid theoretical backgrounds used in experiment design for actual practical approaches, and it offers a solid foundation for this field of study.

Most recent works inspired by Bayesian Experimental Design are focused on the theoretical results for a certain type of experimental configurations rather than addressing any specific practical approach. The main objective defined in most of these works is the maximization of a utility function taken as an expected value relative to Θ and X, the Equation 3.4 shows the most common generalization of the optimization problem and pictures the search of an optimal experimental setting considering prior knowledge over Θ and its prior projection over X. For this particular problem setting, the definition of the functional $U(.)$ will induce a particular optimization problem which consequently requires a different study in order to find a solution. Therefore for each definition of $U(.)$, a new optimization problem is defined. An excellent review about these different problem settings is presented in [Chaloner and Verdinelli, 1995] where the information gain is featured as one of the most addressed forms for $U(.)$. The results discussed in [Chaloner and Verdinelli, 1995] are purely theoretical and due to the complexity associated to the expression Equation 3.4, often implying non-traceable integers, most of the analysis presented in this matter will avoid any practical considerations.

$$U(\epsilon^*) = \max_{\epsilon} \int_{\mathcal{X}} \max_{d} \int_{\mathcal{O}} U(d, \theta, \epsilon, x) p_{\Theta|X,\epsilon}(\theta|x, \epsilon) p_{X|\epsilon}(x|\epsilon) d\theta dx \tag{3.4}$$

3.2.2. Maximum Entropy Sampling

The repercussions of [Shannon, 1948] and [Chaloner and Verdinelli, 1956] are still present in modern works. One of these extensions is the framework of Maximum Entropy Sampling, or MES, first introduced in [Shewry and Wynn, 1987] being an extremely short but condensed work where the MES criteria are defined. The main result presented in [Shewry and Wynn, 1987] is showed in Equation 3.5 where X is a collection of N random variables, X_s a subset of X and $X_{\bar{s}}$ its complement. This expression shows how the entropy of a variable X can be un-

derstood as the sum of the entropy of a subset of the original variable X_s and the conditional relationship between X_s and its complement $X_{\bar{s}}$.

$$H(X) = H(X_s) + \mathbb{E}_{X_s}[H(X_{\bar{s}}|X_s)] \tag{3.5}$$

This known result of Information Theory [Cover and Thomas, 2006] has important repercussions for experimental design. A revision of this Equation is made in [Sebastiani and Wynn, 2000] specifically to address the case of experiment selection and re-formulates the MES criteria to contextualize on the Bayesian Experiment Design framework. For this purpose, the Equation 3.5 is re-written and the random variables θ and X are newly introduced together with the experiment ϵ. The new expression is the following.

$$H(\Theta, Y|\epsilon) = H(Y) + \mathbb{E}_Y[H(\Theta|Y, \epsilon)] \tag{3.6}$$

The hypothesis made in [Sebastiani and Wynn, 2000] states that for practical reasons, for any experiment ϵ the term $H(\Theta, Y|\epsilon)$ can be considered as constant, implying that despite the selection of experiment ϵ (associated with their respective instrumental noise) the sum of $H(Y)$ and $\mathbb{E}_Y[H(\Theta|Y, \epsilon)]$ will be constant for any give experimental design ϵ. This assumption implies that the minimization of $\mathbb{E}_Y[H(\Theta|Y, \epsilon)]$ over ϵ inevitably results in the maximization of $H(Y)$ and vice-versa. In an optimal experiment selection, this relationship can be exploited to change the objective function of the optimization problem, [Sebastiani and Wynn, 2000] states that if the goal of the selection is to minimize the expected uncertainty of the parameters Θ when an observation is made over Y using the experiment ϵ, i.e. $\mathbb{E}_Y[H(\Theta|Y, \epsilon)]$ then the problem is equivalent to select the experiment that maximizes the entropy of the observation variable $H(Y)$. The Maximum Entropy Sampling (MES) is referred as such because it implies that the most informative experiment is the one that has the most entropic observation for Y.

The scope of the MES criteria in a practical setting is not discussed in [Shewry and Wynn, 1987] or [Sebastiani and Wynn, 2000] because these works are focused on theoretical settings where each variable and the experiment had a specific analytical expression. In order to apply this criterion in a practical scenario, an extensive verification of the hypothesis regarding $H(\Theta, Y|\epsilon)$ must be made. In the following sections, this assumption is studied and proven not true for the case of astrometric and spectroscopic observation in our problem, meaning that an extension for this principle must be done.

3.3. Optimal Scheduling State of the Art

The practical study of optimal experimental settings (to improve the performance of various inference methods) is not a new topic. Multiple theoretical approaches have been proposed through time, as shown in previous sections. This literature provides a strong mathematical framework for selecting appropriate experimental configurations. Different authors have proposed practical solutions for specifics problems in diverse scientific fields. The main shared feature among them is the use of a statistical model to address the phenomenological act of observation.

3.3.1. Posterior Variance Minimization

The problem of optimal scheduling in experimental settings can be described as the act of selecting a particular instant of observation t that gives the maximum benefit for a given metric. In [Vanlier et al., 2012] for example, the main objective is the variance minimization for a function z of the variable to infer Θ. In order to estimate this particular value of interest, a simulation-oriented method (in conjunction with an empirical estimation of expectations) is proposed. The first step of this method is an MCMC simulation to obtain samples of the posterior distribution $p_{\Theta|Y^D}$. This approach corresponds to a Bayesian inference analysis and coincides with the state of the art discussed in section 3.1.2. The result of this step is a collection of T samples $\{\theta_i\}_{i=1}^T$ coming from $p_{\Theta|Y^D}$, those samples will serve as a input for the estimation phase of the method.

$$p_{Y_n|Y^D}(y_n|y^D) = \int_{\mathcal{O}} p_{Y|\Theta}(y|\theta) p_{\Theta|Y^D}(\theta|y^D) d\theta \tag{3.7}$$

In the same fashion as Equation 2.9, the Post Observational Predictive distribution (OPD) samples are obtained by marginalization over the posterior distribution as shown in Equation 3.7. The authors of [Vanlier et al., 2012] postulate that through this distribution, the non-linear relationships between the observational and the parameters are preserved and constitute the main object to analyze for evaluating the performance of a certain experiment.

A study of the posterior distribution after an experiment is made presented in [Vanlier et al., 2012] which uses the expression in Equation 3.8, where the proportional value $p_{Y_n|\Theta}(y_n|\theta)p_{\Theta|Y^D}(\theta|y^D)$ is equivalent to the derivation presented in Equation 2.11 implying that this expression corresponds to a sequential Bayesian inference.

$$p_{\Theta|Y_n,Y^D}(\theta|y_n,y^D) = p_{Y_n|\Theta}(y_n|\theta) p_{\Theta|Y^D}(\theta|y^D) \frac{Z_1}{Z_2} \tag{3.8}$$

Through Equation 3.8, a normalization term for $p_{Y_n|\Theta}$ and a Monte Carlo approximation, the expected posterior value after a certain experiment for the function $z(\Theta)$ can be estimated as showed in Equation 3.9.

$$\mathbb{E}_\Theta[z|y_n,y^D] \approx \sum_{i=1}^{T} \frac{\tilde{p}_{Y_n|\Theta}(y_n|\theta_i)}{\sum_{j=1}^{T} \tilde{p}_{Y_n|\Theta}(y_n|\theta_j)} z(\theta_i) \tag{3.9}$$

The Equation 3.9 is function of the particular value that the experiment takes y_n, meaning that a new expectation with respect to Y_n must be performed. Under the same spirit, a Monte Carlo approximation is applied in Equation 3.10 being $G(\theta_i, \theta_r)$ part of the Gaussian model of noise selected and representative of $p_{Y_n|\Theta}$.

$$\begin{aligned} \mathbb{E}_{\Theta,Y_n}[z] &= \frac{1}{T}\sum_{r=1}^{T}\sum_{i=1}^{T}\frac{G(\theta_i,\theta_r)}{\sum_{k=1}^{T}G(\theta_k,\theta_r)}z(\theta_i) \\ G(\theta_a,\theta_b) &= \exp\left(-\frac{y(\theta_a)-y(\theta_b)}{2\sigma^2}\right) \end{aligned} \tag{3.10}$$

The function $z(\theta)$ in Equation 3.10 corresponds to the projection of θ to a space of interest Z. For simplicity, the choice of $z(\theta) = \theta$ can be made, meaning that the expectation presented in Equation 3.10 is not the main objective to minimize. In [Vanlier et al., 2012] a variance minimization is proposed for the function z, where Equation 3.11 specifies how the variance is calculated trough the expectations $\mathbb{E}[z^2]$ and $\mathbb{E}[z]^2$ also a Variance Reduction metric is also proposed in this work, where σ^2_{old} is the variance only using the Post OPD samples and σ^2_{new} is the variance for the new data obtained via experimentation.

$$\begin{aligned} \mathbb{V}\text{ar}[z] &= \mathbb{E}[z^2] - \mathbb{E}[z]^2 \\ \text{VarRedux} &= 1 - \left[\frac{\sigma^2_{\text{new}}}{\sigma^2_{\text{old}}}\right] \end{aligned} \tag{3.11}$$

In this fashion, the work [Vanlier et al., 2012] presents a structured method to select an experiment that reduces the variance of θ of interest. It is important to remark that this method is heavy, with a computational cost that is order $\mathcal{O}(T^2)$, being T is the number of samples generated by the MCMC method.

3.3.2. MES criteria via Monte Carlo Integration

The MES criteria is one of the most useful frameworks for optimal experimental setting, the closed and short statement made in 3.6 allows to implement this methodology in almost any scenario. An astronomy study made in [Loredo, 2004b] about exoplanets orbits used the Bayesian Experiment Design framework to select the best instant of observation which reduced the expected uncertainty on the parametrical space. The work started by reviewing the utility function presented in section 3.2.1 and quickly conclude about its lack of practicality (due to the intractability of the integrals presented in Equation 3.4), thus a MES criteria oriented method is proposed.

The main objective of [Loredo, 2004b] is the entropy maximization of the Post OPD subject to a proposed list of experiments. In order to estimate the differential entropy, a

Monte Carlo Integration is used to approximate $H(Y_n)$ using Equation 3.12.

$$\begin{aligned} H(Y_n) &= -\int_{\mathcal{Y}} p_{Y_n|Y^D}(y_n|y^D) \log p_{Y_n|Y^D}(y_n|y^D) dy_n \\ &= -\frac{1}{T}\sum_{i=1}^{T} \log p_{Y_n|Y^D}(y_n^{(i)}|y^D) \end{aligned} \tag{3.12}$$

The Equation 3.12 pictures how the distribution $p_{Y_n|Y^D}$ and samples coming from it allows to estimate the differential entropy. On this, the samples of the Post OPD $p_{Y_n|Y^D}$ are obtained via propagation of the samples coming from a MCMC simulation for the posterior distribution $p_{Y|Y^D}$ in the same way as Equation 3.7. The value that takes the distribution $p_{Y_n|Y^D}$ for a particular value y_n must be estimated. The Equation 3.7 can be reinterpreted again as an expectation and, consequently, a Monte Carlo Integration can be performed as shown in Equation 3.13.

$$\begin{aligned} p_{Y_n|Y^D}(y_n|y^D) &= \int p_{Y_n|\Theta}(y_n|\theta) p_{\Theta|Y_n}(\theta|y_n) d\theta \\ &\approx \frac{1}{T}\sum_{i=i}^{T} p_{Y_n|\Theta}(y_n|\theta_i) \end{aligned} \tag{3.13}$$

In summary, the method for optimal observation scheduling proposed in [Loredo, 2004a] (and later used in [Loredo et al., 2012] for exoplanets detection) can be performed by a nested Monte Carlo Integration over the samples coming from the Post OPD. This approach takes computational cost of order $\mathcal{O}(T^2)$ being T the number of samples generated by the MCMC simulation. The last aspect to remark from this work is the questionable hypothesis (from the MES criteria) in which is stated that the joint entropy $H(\Theta, Y_n|\epsilon)$ is constant for any experiment ϵ in the plausible agenda. In later chapters, this criterion will be discussed, and the cases where this hypothesis holds will be clarified.

Chapter 4

Optimal Sampling Criteria

4.1. MES Criteria Extension

In astronomical problems, obtaining observational data from a system is not an easy task. Multiple exogenous factors are present which, in the long run, determine the viability of an experimental setting, for example, the meteorological conditions that affect the instrument's accuracy, the observatory location and the date of the observation, even the schedule of the observatory itself are essential constraints in an experimental selection. For practical purposes, all of these exogenous difficulties attached to the practical restrictions of observation are reduced to a finite collection of possible measurements. On the other hand, a more general setting can be proposed for a continuous observation model; this new approach implies that the most informative instant should be perfectly selected for an exact moment of observation. Unluckily a continuous approach on the scheduling problem is avoided due to two main reasons: 1) Non-discrete approaches described in the literature are highly complex, often implying the derivation of untraceable integrals and non-closed solutions 2) In practice obtaining observational data is constrained by the exogenous factors such as the visibility of the system (by meteorological conditions or observability from the telescope location) and administrative viability of the observatory, these constraints will always lead to a finite number of viable days to observe. In conclusion, raising the mathematical complexity of the scheduling problem in the purse of perfectly selecting observational instant in a continuous setting does not serve any practical proposes when observing binary systems.

The optimal scheduling problem is defined as the search of an optimal experiment in a finite space $\mathcal{A}$. The optimal solution is basied on a functional $\mathcal{F}$ representative of the expected posterior distributions over the parameters after the experiments have been performed. The set of experiments $\mathcal{A}$ will be referred to as the Agenda and each element in it is defined as a set of experimental conditions (such as the instant of observation and the precision of the telescope) that completely characterizes the observational setting. This definition of the feasible space is shared in several related works as [Loredo, 2004b] and [Vanlier et al., 2012], being in each case the experiment ϵ identified in different forms. The main point of detachment between methods proposed in the literature is the definition of the functional $\mathcal{F}$, for example, [Vanlier et al., 2012] represents the school of thought focused in measure the different moments of the posterior distribution $p_{\Theta|D}$ such as the variance (2nd moment). In

this work, a new approach on $\mathcal{F}$ is studied.

Alternatively, objective functions coming from Information Theory for optimal scheduling have been proposed in multiple works regarding optimal scheduling and experimental selection[Sebastiani and Wynn, 2000][Chaloner and Verdinelli, 1995] [Loredo, 2004b]. The main result of those papers is the definition of the Maximum Entropy Sampling criteria (or MES criteria) which is portrayed in Equation 3.6. It is important to remark that this criterion gives a solution to the problem of conditional entropy minimization (under the hypothesis of the joint Entropy being constant), then concludes that the maximization of the marginal Entropy for the observational variable directly implies a minimization in the functional of interest. The definitions and nomenclature presented in Equation 3.6 are at least questionable; the manner that this work expresses the differences between experiments ϵ is ambiguous and lacks an explicit definition; it also does not explore the relationship between the observations and the parameters in an inference problem. Therefore, in order to clarify the notation of Equation 3.6, a new expression for the MES criteria is proposed using the observational model described in Section 2.2 and pictured in Figure 2.4.

$$H(\theta, Y_\alpha^t) = H(Y_\alpha^t) + \mathbb{E}_{Y_\alpha^t}[H(\theta|Y_\alpha^t = y_\alpha^t)] \tag{4.1}$$

This re-definition of 3.6 allows understanding more directly how the observational variable affects the MES criteria. In Equation 4.1 the experiment is no longer defined by ϵ, instead is completely captured by the observation Y_α^t which is indexed by the observation instant t and the observation instrument uncertainty α. This re-definition significantly clarifies the interaction between the parameters Θ and the observation Y_α^t and makes explicit the fact that changing the characterization of an experiment (the values of t and α) will naturally affect all the terms involved in Equation 4.1 and consequently changing the objective of minimization. It is important to notice that Θ is not explicitly denoted as the posterior given a collection of data D, the reason behind this is because all of the posterior knowledge acquired by observing D can be embedded in Θ already; this is achieved by replacing the prior distribution for the posterior distribution obtained via Bayesian Inference. In Section 2.3.2 it is discussed how, in a sequential inference approach, an update of the prior distribution can be made in order to simplify the calculation of future posterior distributions. Inspired by Section 2.3.2 and in the pursue of avoiding heavy notation, the random variable Θ will refer to the posterior distribution given data D, and consequently, Y^t and Y_α^t will refer to the Posterior Predictive Distribution (Post PD) and Posterior Observational Predictive Distribution (Post OPD) respectively, which are generated by Θ and using the marginalization technique shown in the observation model presented in Figure 2.4.

The redefinition of the MES criteria, in Equation 4.1, allows understanding the hypothesis and conclusions presented in [Sebastiani and Wynn, 2000] in the context of optimal scheduling for orbital parameters. The term $H(Y_\alpha^t)$ is the most straight forward concept present in the Equation 4.1, the full characterization of the post OPD Y_α^t via marginalization is specified in Equation 2.9 suggests that $H(Y_\alpha^t)$ completely depends on the posterior knowledge Θ and the observational indices t and α of the experiment to perform. The term $H(Y_\alpha^t)$ also corresponded to the main objective of MES presented in [Sebastiani and Wynn, 2000] and applied in [Loredo, 2004b], which can be conceptualized as the uncertainty in the post OPD

knowledge. This connection implies that obtaining a sample in the moment of major Entropy will lead to the largest reduction in expected uncertainty for a feasible Agenda $\mathcal{A}$. Intuitively observing the most uncertain instant of the system will coincide with the most informative observation. However, in order to give formal substance to this belief, the hypothesis of the joint Entropy $H(\Theta, Y_\alpha^t)$, presented in [Sebastiani and Wynn, 2000] and discussed in Section 3.2.2., must be mathematically explicit.

The main point of conflict in the approach presented in [Sebastiani and Wynn, 2000] is the verification of the joint entropy hypothesis. At first glance, a constant joint entropy regardless of the experiment chosen is a strong hypothesis. The joint entropy $H(\Theta, Y_\alpha^t)$ is a measure that captures deep probabilistic behavior between Θ and Y_α^t, suggesting that a constant behavior of $H(\Theta, Y_\alpha^t)$ for any selection of Y_α^t only occurs in a small set of inference settings. The joint entropy present in Equation 4.1 can be rephrased to identify the elements that conform $H(\Theta, Y_\alpha^t)$. This new expression is the following:

$$\begin{aligned}\mathbb{E}_{Y_\alpha^t}[H(\theta|Y_\alpha^t = y_\alpha^t)] &= H(\theta, Y_\alpha^t) - H(Y_\alpha^t) \\ &= \underbrace{H(\Theta) + \mathbb{E}_\Theta[H(Y_\alpha^t|\Theta = \theta)]}_{\text{definition of joint entropy}} - H(Y_\alpha^t)\end{aligned} \tag{4.2}$$

By using Equation 3.5 and inverting the order of the variables, a new expression for the MES criteria is obtained. Importantly, the new terms $H(\Theta)$ and $\mathbb{E}_\Theta[H(Y_\alpha^t|\Theta = \theta)]$ are explicitly related to our objective function $\mathbb{E}_{Y_\alpha^t}[H(\theta|Y_\alpha^t = y_\alpha^t)]$. The prior distribution p_Θ is not related with the instant of observation, meaning that, design-wise, the marginal entropy $H(\Theta)$ is constant, and consequently easily removed from the objective function. In the pursuit of improving the interpretability of the final objective function, a new treatment for Equation 4.2 is proposed in Equation 4.3 by relating the conditional entropy with the Mutual Information.

$$\begin{aligned}I(\theta; Y_\alpha^t) &= H(\Theta) - \mathbb{E}_{Y_\alpha^t}[H(\theta|Y_\alpha^t = y_\alpha^t)] \\ &= H(\Theta) - H(\Theta) - \mathbb{E}_\Theta[H(Y_\alpha^t|\Theta = \theta)] + H(Y_\alpha^t) \\ &= H(Y_\alpha^t) - \mathbb{E}_\Theta[H(Y_\alpha^t|\Theta = \theta)]\end{aligned} \tag{4.3}$$

Mutual Information is an indicator of the interaction between two random variables, as discussed in Section 2.4.2. In addition to its definition, the derived expression in Equation 4.3 establishes that this information metric is also indicative of entropy reduction in a Bayesian inference scenario, implying that the selection of a random variable from a feasible set, i.e. an experiment from the plausible agenda $\mathcal{A}$, will be the most informative when the expected posterior entropy for the parametrical variable finds his minimum value. It is important to remark that this new Equation follows the same principals as the original MES criteria presented in [Sebastiani and Wynn, 2000] [Loredo, 2004b], but now featuring the term $\mathbb{E}_\theta[H(Y_\alpha^t|\Theta = \theta)]$. This new extension implies that the original hypothesis on the joint entropy being constant is necessarily equivalent to a constant behavior of the conditional entropy $\mathbb{E}_\theta[H(Y_\alpha^t|\Theta = \theta)]$. This new connection rises the question about under which conditions, for the experimental setting and the problem framework itself, this constant $H(\Theta, Y_\alpha^t)$ assumption results valid. The simplest feasible scenario corresponds to the case where all

experiments in the plausible agenda $\mathcal{A}$ have the same observational model, defined in Figure 2.4, i.e all the distributions $p_{Y_\alpha^t|\Theta} \forall \theta \in \Theta$ for each entry in the Agenda $\mathcal{A}$ are the same. This proposition is true in several experimental settings, for example, let us take the case when a particular phenomenon can be observed by the same instrumental equipment but in different instances, being in each instance the observation function M is the same and consequently distribution $p_{Y_\alpha^t|\Theta}$ results equivalent in any experimental setting. Then the original MES criteria states that the moment of maximum marginal entropy on the post OPD variable Y_α^t is the instance that reaches the minimum expected new posterior entropy $H(\Theta|Y_\alpha^t)$. Due to the nature of the conditional entropy, the existence of exotics configurations for experiments in $\mathcal{A}$ that does not have the same conditional distribution $p_{Y_\alpha^t|\Theta}$ but still hold the statement that $\mathbb{E}_\theta[H(Y_\alpha^t|\Theta = \theta)]$ is constant are plausible, but the lack of a closed expression for the conditional entropy obscures any generalization of the MES criteria in a practical setting. In the search of a generalized discussion on the MES criteria, and addressing the practical restrictions of the inference problem on binary starts, a whole study of Equation 4.3 is suggested, meaning that the analysis of $\mathbb{E}_\theta[H(Y_\alpha^t|\Theta = \theta)]$ takes special importance. In the context of this work, a Gaussian observational noise is considered, exposed in Section 2.2, this assumption is important under the MES criteria scope because allows simplifying the Equation 4.3 by virtue of the Gaussian probability distribution and its properties.

$$
\begin{aligned}
I(\theta; Y_\alpha^t) &= H(Y_\alpha^t) - \mathbb{E}_\Theta[H(Y_\alpha^t|\Theta = \theta)] \\
&= H(Y_\alpha^t) - \int_\Theta p_\Theta(\theta) H(Y_\alpha^t|\Theta = \theta) d\theta \\
&= H(Y_\alpha^t) - \int_\Theta p_\Theta(\theta) \underbrace{\frac{1}{2} \ln[(2\,\pi\,e)^m \det \alpha]}_{H(Y_\alpha^t|\theta)} d\theta \\
&= H(Y_\alpha^t) - \frac{1}{2} \ln[(2\,\pi\,e)^m \det \alpha] \underbrace{\int_\Theta p_\Theta(\theta) d\theta}_{1} \\
&= H(Y_\alpha^t) - \frac{1}{2} \ln[(2\,\pi\,e)^m \det \alpha])
\end{aligned}
\tag{4.4}
$$

The main aspect about a multivariate Gaussian hypothesis for $p_{Y_\alpha^t}$ is that allows to derive a closed expression for the conditional Entropy which is only dependent on the covariance matrix α that holds all the instrumental uncertainty of the observation, the derivation of $H(Y_\alpha^t|\Theta = \theta)$ is present in Appendix B.1. The expression in Equation 4.4 allows to use the principals of the MES criteria but in a generalized framework, where the hypothesis of $H(\Theta, Y_\alpha^t)$ for all observations in $\mathcal{A}$ being constant is equivalent to consider an agenda $\mathcal{A}$ that considers only α_i and α_j that follows $\det\alpha_i = \det\alpha_j$

Using the Equation 4.4 as a main result of this section, under the mentioned assumptions, it is possible to articulate a new paradigm for optimal experimental selection in a wider range of experimental contexts than [Sebastiani and Wynn, 2000] and [Loredo, 2004b] initially proposed. Consequently, the next section aims to formalize and establish the definition of a Optimal Sampling criteria in the context of parametrical inference for binary systems.

4.2. Optimal Sampling Criterion

Let be $\Theta \in \mathcal{O}$ a random variable that represents the parametrical knowledge at certain point of a particular system, $\mathcal{M}_u$ a family of functions that can be indexed by $t \in \mathcal{T}$ resulting in a function $M_u^t : \mathcal{O} \to \mathcal{Y}_u$ mapping the parameters on $\mathcal{O}$ to the space of observations $\mathcal{Y}_u \subseteq \mathbb{R}^m$, and α a constant that represents all the noise regarding a particular observation of a system. For this particular case of study let us consider α the covariance matrix of the Gaussian conditional distribution $p_{Y_\alpha^t|\Theta} \sim \mathcal{N}(M_u^t(\theta), \alpha)$. Then, the triplet (u, t, α) represents an instance of the observational model graph presented in Figure 2.4 and also the setting for an observation of the system as a noisy experiment.

The problem of selecting an experiment (u, t, α) over the agenda $\mathcal{A} := \{(u_j, t_j, \alpha_j)|j \in [1, ..., J]\}$ that minimizes the posterior uncertainty Θ is the following:

$$\arg\max_j I(\theta; Y_{\alpha_j}^{t_j}) = \arg\max_j H(Y_{\alpha_j}^{t_j}) - \frac{1}{2}\ln[(2\,\pi\,e)^{m_j}\det\alpha_j]) \tag{4.5}$$

Solving Equation 4.5 implies an extensive search over the J available experiments. By looking at the Equation 4.5, this selection mostly depends on the non-trivial OPD marginal entropy $H(Y_{\alpha_j}^{t_j})$ induced by the relationship between Θ and Y^t through M_u^t and the instrumental noise α. The second terms works as a normalization constant for each entry on $\mathcal{A}$. As discussed in the previous section, this last term can be ignored if $\mathbb{E}_\theta[H(Y_\alpha^t|\Theta = \theta)]$ is constant over any observation in $\mathcal{A}$.

To contextualize the criterion on 4.5 for our binary system inference problem a detailed list of scenarios is presented.

- u represents the measurement that is performed. For this particular work Astrometric and Spectroscopic measurements are considered, both described in Section 2.1.
- t represents the moment in time when the observation is taken. This parameter captures the experimental plausibility considering all the practical restrictions associated to the act of observation. For example, the visibility of the system on the sky from a particular location on the observational schedule from the observatory.
- α models the instrumental and ambient noise associated to a certain case of observation. This constant is often constrained to the available instrumental equipment and in a practical setting has a physical dimension associated.

By utilizing this notation, the triplet (u, t, α) holds all the information regarding the experimental observation of the system. Given the generalized description of the optimization problem in Equation 4.5 a extension for experimental selection that considers more than one measurement can be specified. Let $A_j := \{(u_{k_j}, t_{k_j}, \alpha_{k_j})|k_j \in [1, ..., K_j]\}$ be a new candidate for the Agenda $\mathcal{A}$ being composed by K_j individual observations defined by the triplet $(u_{k_j}, t_{k_j}, \alpha_{k_j})$ (it is important to note that each observational setting has the freedom to take any desired practical configuration, regarding its uncertainty, and is not necessarily related to the other observations), then the generalized version of the Optimal Sampling

Criteria can be exposed as follows.

$$\arg\max_j I(\theta; \mathcal{A}_j) = \arg\max_j H(\mathcal{A}_j) - \underbrace{\sum_{k_j=1}^{K_j} \frac{1}{2} \ln[(2\,\pi\,e)^{m_{k_j}}\,\det\,\alpha_{k_j}]}_{\text{Conditional independence}} \tag{4.6}$$

To simplify Equation 4.6 the joint entropy $H(\{Y_{\alpha_{k_j}}^{t_{k_j}}\}_{k_j=1}^{K_j})$ of K_j observations for the jth entry on the Agenda $\mathcal{A}$ is represented by $H(A_j)$. In the same fashion the normalization term presented in Equation 4.5 is also updated for this new approach by considering that each observation is conditionally independent to each other (a decomposition of this conditional entropy is applied resulting in the sum all conditional entropy presented in Equation 4.6). The derivation of this last expression is presented in Appendix B.2. It is crucial to understand that despite the decomposition derived for the second term, the joint entropy featured in the first term, can not be decomposed. There is a strong dependency between the observational variables when not conditioned to a specific value θ for Θ. This implies that a full analysis on the newly joint space must be performed to obtain $H(\mathcal{A}_j)$ in Equation 4.6.

In summary, the MES criteria can be extended for experimental candidates that: 1) are not generated by the same noise model, 2) are not coming from the same phenomenological process and 3) considers more than one kind of observational sampling, This implies that the new criterion can select the optimal experiment from a wider collection of plausible phenomenons.

This new criterion is presented in Equation 4.5 for am agenda $\mathcal{A}$, and it is extended in Equation 4.6 for more general purposes. We call this Optimal Sampling Criteria which will be the main focus of study on this work.

4.2.1. On the Interpretability of Optimal Sampling Criteria

The Optimal Sampling Criteria not only provides a closed expression to select an experimental setting that minimizes the expected uncertainty of the parameters. The criterion also gives a relationship between the noise modeling and the post OPD variable using the entropy in each element. This definition can be interpreted as the information gain being proportional to the previous uncertainty attached to a certain experiment (the first term of Equation 4.5) but limited to the intrinsic noise associated to the phenomenological act of observation (Second term). This last interpretation results in a extension of the intuition behind the MES criteria discussed in [Sebastiani and Wynn, 2000], where the noise model takes a crucial part to precisely select an optimal experimental setting.

Under the scope of the Gaussian model utilized for observational noise modeling on this work, a further extension of the interpretation can be achieved. The main reason from this is that the Gaussian distribution has a well-known closed expression that clearly isolates the mean vector and covarianze matrix of Equation 4.4. To simplify this analysis let us take the Gaussian uni-variate scenario, meaning that the projected variable Y^t will be a scalar in conjunction with the observational variable Y_α^t. By using the marginalization technique discussed in Section 2.2 a expression for the distribution of the observational variable Y_α^t can

be accomplished by replacing the conditional distribution $p_{Y_\alpha^t|Y^t}$ (Gaussian model $\mathcal{N}(y_t, \alpha)$) a new expression for $p_{Y_\alpha^t}$ can be achieved.

$$\begin{aligned} p_{Y_\alpha^t}(y_\alpha^t) &= \int_{\mathcal{Y}} p_{Y_\alpha^t|Y^t}(y_\alpha^t|y^t) p_{Y^t}(y^t) dy^t \\ &= \int_{\mathcal{Y}} \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{(y_\alpha^t - y_t)^2}{2\alpha}} p_{Y^t}(y^t) dy^t \\ &= \int_{\mathcal{Y}} G(y_\alpha^t - y_t) p_{Y^t}(y^t) dy^t \\ &= (G * p_{Y^t})(y_\alpha^t) \\ \text{Being } G(x) &\sim \mathcal{N}(0, \alpha) \end{aligned} \tag{4.7}$$

In the context of Optimal Sampling Criterion, Equation 4.7 shows that the marginal distribution of Y_α^t can be seen as a convolutional filtering of the post PD Y^t with a Gaussian kernel of variance α. This new representation makes evident the change in the amplitude of the original distribution, being this variation induced by the value of α, after the filter is applied. In the particular case of a Gaussian window, all sudden changes in amplitude of the distribution p_{Y^t} (high frequency behavior) are smoothed after the filter is applied, implying that a more homogeneous distribution is induced. It is important to mention that the level of smoothness provoked by a Gaussian window is captured by α. Figure 4.1 shows the case for $\alpha = 1$ the resultant convolution over a uniform$[-2, 2]$ distribution features a more wide and evenly distributed result than using a tighter Gaussian window ($\alpha = 0.5$). For α values near to 0 the original distribution remain almost the same while for big α values the resulting distribution is significantly distorted to be more uniformly distributed.

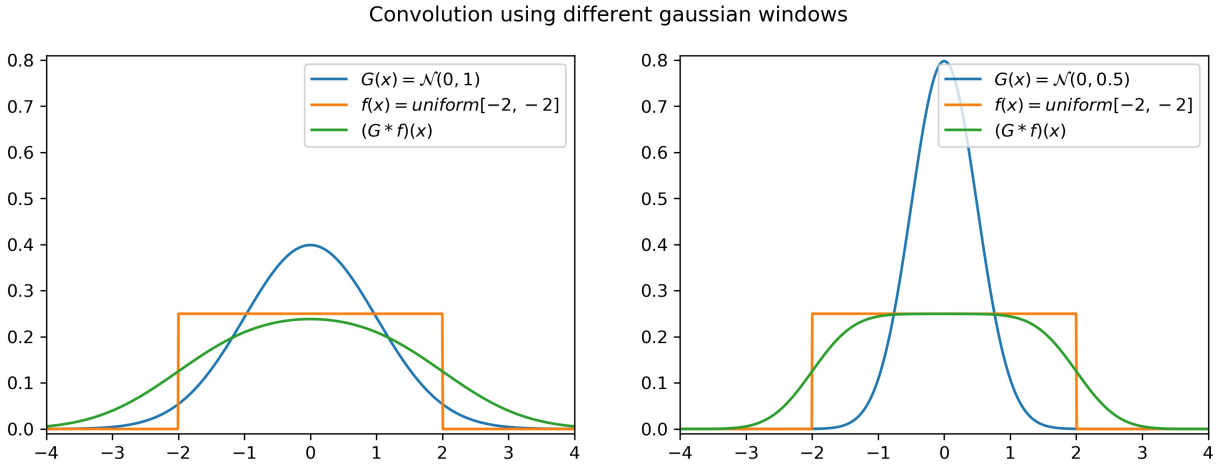


Figure 4.1: Convolution of a Gaussian window over $f(x) = \text{uniform}(-2, 2)$. a)$G(x) = \mathcal{N}(0, 1)$ and b) $G(x) = \mathcal{N}(0, 0.5)$

Regarding the Optimal Sampling Criteria and the convolution representation of $p_{Y_\alpha^t}$ 4.7 the following interesting cases can be mentioned:

- The deterministic posterior projection scenario: When $p_{Y^t}(y_t) = \delta_{y_c}(y^t)$, suggesting that

the post PD has no uncertainty associated when observing the system, then the post OPD is $p_{Y_\alpha^t} \sim \mathcal{N}(y_c, \alpha)$ implying that $H(Y_\alpha^t) = \mathbb{E}_\Theta[H(Y_\alpha^t|\Theta = \theta)]$ and consequently $I(\Theta; Y_\alpha^t) = 0$. This scenario aims to model what happens when a observation is performed in an instant that has not prior uncertainty (i.e. is predicted as deterministic). The intuition says that no new information should be acquired, which can be verified through the convolution analysis of the post OPD $p_{Y_\alpha^t}$.

- The infinitely precise measurement: When $\alpha \to 0$ implies $p_{Y_\alpha^t} \to p_{Y^t}$ resulting in $\mathbb{E}_\Theta[H(Y_\alpha^t|\Theta = \theta)] = 0$ and $I(\Theta; Y_\alpha^t) = H(Y_\alpha^t)$. This scenario shows what happens when a infinitely precise instrument is used to measure the system. The intuition says that all information of Y_α^t is obtained after the measurement, implying that all the uncertainty of Y_α^t at the instant t is removed from the posterior $p_{\Theta|Y_\alpha^t}$.

In summary, the Optimal Sampling Criteria can be interpreted by using the convolution of two probability distributions. The analysis of two boundary cases corroborate the intuition regarding the information gain from a certain experimental setting. By means of this study it is concluded that the selection of an optimal experiment relays on the trade off between the acquired knowledge of the system Θ, projected in a observational variable Y_α^t, and the noise uncertainty associated to the instrumental precision. It is important to remark that this relationship is non trivial because of the convoluted nature of Y_α^t and the complexity the noise (which is not Gaussian in general).

4.3. Estimation of Optimal Sample Criterion

The Optimal Sampling Criterion derived in Equation 4.5 is a closed expression that exactly calculates the expected information gain from a certain experiment. However in practical settings, the direct computation of $H(Y_\alpha^t)$ is very different, the intrinsic difficulties of the integrals regarding the entropy and the fact that the distribution $p_{Y_\alpha^t}$ is not known make a direct computation of Equation 4.5 unavailable. In order to obtain practical method to select an experiment from $\mathcal{A}$ for a system where the prior distribution of Θ and a collection of measurements D available, a simulation and empirical differential entropy estimation methods are proposed to solve Equation 4.8.

$$\arg\max_j \hat{I}(\theta; Y_{\alpha_j}^{t_j}) = \arg\max_j \hat{H}(Y_{\alpha_j}^{t_j}) - \frac{1}{2}\ln[(2\,\pi\,e)^{m_j}\det\alpha_j]) \tag{4.8}$$

4.3.1. Particle Simulation

The first stage of the OSM (Optimal Sampling Method) correspond to obtaining the posterior distribution $p_{\Theta|D}$ using the available observations D and a prior p_Θ. As discussed in Section 3.1.2 several works try to solve this by applying deterministic and probabilistic approaches. In order to obtain samples coming from the posterior distribution a particle simulation method proposed in [Mendez et al., 2017] and [Claveria et al., 2019] that uses Monte Carlo Markov Chains with Gibbs sampler is implemented. The implementation of the MCMC method is shown in the Algorithm 1, this variation of MCMC features the Gibbs sampling method along with the MCMC usual methodology. The method is performed as follows: First a random sample from the initialization distribution μ is obtained, then the particle simulation loop is started where the Gibbs sampling method proposes new particles by re-sampling one dimension of the previous samples. The proposed new particle θ' is generated via a proposal distribution q_Θ (in this case a Gaussian proposal is selected), finally in order to accept or reject this new sample a acceptance rate definition must be computed. In this last stage is where the MCMC method tries to imitate the posterior distribution by using the Bayes theorem (shown in Equation 2.10) in the form of a likelihood ration between the previous particle and the proposed one. The last step in the particle simulation loop is to sample a binomial distribution to decide if the new particle is accepted or is rejected, in the last case the new particle is set equal to the previous one. Finally, the algorithm ends when a number of desired samples N have been generated. The MCMC method generates that simulated particles follow the desired distribution, in this case the posterior distribution $p_{\Theta|D}$, when the number of samples is big enough. The obtained samples are denoted by

$\{\bar{\theta}^i\}_{i=1}^N$ following the desired posterior distribution $p_{\Theta|D}$.

Algorithm 1: Gibbs sampler MCMC

```
/* Initialization from μ                                                    */
θ^(1) = Sample θ ~ μ_θ
/* For each sample to generate                                              */
for i = 2, ..., N_steps do
    /* Retrieve value from the last iteration                               */
    θ^(i) = θ^(i−1)
    /* For each dimension of θ                                              */
    o for j = 1, ..., d do
        /* Assigns the candidate equal to the last state                    */
        θ' = θ^(i)
        /* Sample the proposal to update the j-dimension of the candidate   */
        θ'_j = Sample θ_j ~ q_θ = N(θ'_j, σ_j^2)
        /* Acceptance rate definition                                       */
        A = min{1, p_{D|Θ}(d|θ') / p_{D|Θ}(d|θ^(i))}
        /* Acceptance coin-flip                                             */
        u' = Sample u ~ uniform(0, 1)
        if u' < A then
            /* Accept the candidate as new sample                           */
            θ^(i) = θ'
        end
    end
end
```

By using the obtained samples of the MCMC method a full computation of the observational model graph shown in Figure 2.4 can be performed. In order to avoid unnecessary notation, the distribution $p_{\theta|D}$ will be denoted by p_θ using the sequential bayesian inference analysis presented in Section 2.3.2. The samples from the first node of the observational graph can be easily computed by using the deterministic mapping M_u^t existing between Θ and Y_α^t as shown in Equation 4.9.

$$\bar{Y}^i_{\sim p_{Y^t}(y^t)} = M_u^t(\bar{\theta}^i_{\sim p_\Theta(\theta)}) \qquad \forall i \in [0, N] \tag{4.9}$$

For obtaining the samples of the post OPD a sampling form the join distribution p_{θ,Y_α^t} can be made. This intermediate step is made because any joint distribution $p_{X,Z}$ satisfies that $p_{X,Z}(x,z) = p_X(x)p_{Z|X}(z)$, then by using the samples $\{\bar{\theta}^i_{\sim p_\Theta}\}_{i=1}^N$ in conjunction with samples of $p_{Y_\alpha^t|\Theta=\bar{\theta}^i}$ can be pack together to produce samples of the p_{θ,Y_α^t}. Equation 4.10 shows this samples.

$$(\bar{\theta}^i, \bar{Y}^i_{obs})_{\sim p_{\Theta,Y_\alpha^t}(\theta,y_\alpha^t)} = (\bar{\theta}^i_{\sim p_\Theta(\theta)}, \bar{Y}^i_{obs\sim p_{Y_\alpha^t|\Theta}(y_\alpha^t|\bar{\theta}^i)}) \qquad \forall i \in [0, N] \tag{4.10}$$

Samples of post OPD can be derived from p_{Y^t,Y_α^t}. The previously calculated samples $\{\bar{Y^t}^i_{\sim p_{Y^t}}\}_{i=1}^N$ can be utilized to produce $\bar{Y}^i_{obs\sim p_{Y_\alpha^t|Y^t}(y_\alpha^t|\bar{Y}^i)} = \bar{Y}^i_{\sim p_{Y^t}(y^t)} + \bar{\xi}^i_{\sim\mathcal{N}(0,\alpha)}$. Equation 4.11 shows the samples of $\bar{Y}^i, \bar{Y}^i_{obs}{}_{i=1}^N$.

$$(\bar{Y}^i, \bar{Y}^i_{obs})_{\sim p_{Y^t, Y^t_\alpha}(y^t, y^t_\alpha)} = (\bar{Y}^i_{\sim p_{Y^t}(y^t)}, \bar{Y}^i_{obs \sim p_{Y^t_\alpha|Y^t}(y^t_\alpha|\bar{Y}^i)}) \qquad \forall i \in [0, N] \tag{4.11}$$

In summary, the first stage of the Optimal Sampling produces an inference on the posterior parametrical distribution $p_{\Theta|Y^t_\alpha}$ via simulation. The presented MCMC method produces a set of samples i.i.d. from the post OPD $p_{Y^t_\alpha}$. This collection of samples is the final product of this stage and will be used in the next step of the Optimal Sampling method.

4.3.2. Differential Entropy Estimation

The second and last stage of the Optimal Sampling method corresponds to the differential entropy estimation phase for the two terms in Equation 4.5. The differential entropy estimating is a complex task that involves non-traceable integrals. There are few cases where a closed expression can be computed: one is the Gaussian case (shown in the Appendix B.1), by the other hand when the objective random variable has more complex probability distributions the differential entropy is often difficult to calculate an consequently an alternative approach is suggested, the sample based differential entropy estimators are a subject of study that has not been addressed in many contemporary researches this is because the lack of practical applications regarding the differential entropy, the main focus on the estimation research about Information Theory is in the mutual information measures, the majority of this works are data-driven estimators that suffers heavily on the dimensionality curse and consequently for high dimensional cases (generally considered from dimension 10) the performance is drastically affected, being this the main reason for avoid direct estimation of the mutual information instead of the proposed Equation 4.5.

Due to the general shortage of methods for differential entropy estimation a classical algorithm is selected for purposes of this work, the Kochazenko-Leonenko method proposed in [Kozachenko and Leonenko,] is one of the first documented approaches for estimation of the differential entropy of a random variable from which a collection of samples is available, also is one of the most cited works about this matter being studied under several scopes, this method have been used as a cornerstone for important works like the well-known Kraskov mutual information estimator [Kraskov et al., 2004]. The Kochazenko-Leonenko method, or shortly KL method, is based on a k-nearest-neighbor estimation of the objective probability distribution p_X, this assumption systematically induces a closed expression for the differential entropy $\hat{H}(X)$, the estimator can be expressed as the sum of four terms, as shown in Equation 4.12 being ψ the digamma function and c_m the volume of a m dimensional unit ball, the three first terms are constant and depends on the selected $k \in \mathbb{N}$ neighbor of interest and m the dimension of X, by the other hand the last term is dependent of the N particles available from p_X being $\epsilon(i)$ the double of the distance from the ith particle to his k-nearest-neighbor. The detailed derivation of the Equation 4.12 is documented in Appendix B.3.

$$\hat{H}(X) = -\psi(k) + \psi(N) + \log(c_m) + \frac{m}{N}\sum_{i=1}^{N}\epsilon(i) \tag{4.12}$$

Finally, by means of the KL method it is possible to estimate the marginal entropy for Y^t_α

using its collection of samples discussed in the Section 4.3.1.

The serialization of the inference stage and the entropy estimation stage allows to define a full method to empirically estimate the OS criteria for each experimental setting given a the plausible Agenda $\mathcal{A}$. This framework is represented by the diagram featured in Figure 4.2, where the computation of each node in the diagram can be summarised as follows:

- The prior distribution Θ, the collected data D and the likelihood distribution for the data $p_{D|\Theta}$ are set as input and parameters of the Inference Stage.
- A MCMC particle simulation is performed an the samples of the posterior distribution $p_{\Theta|D}$ are computed.
- By using the obtained samples from $p_{\Theta|D}$ and the plausible Agenda $\mathcal{A}$, a collection of observational particles for each jth experiment is computed.
- For each experiment a KL entropy estimation of parameter k using the observational samples and a closed evaluation of the Gaussian noise model are performed and then respectively the OS criteria is computed for each case.
- By using the obtained list of information gain, a priority list of experiments is defined, then the first element present in such list is selected as the optimal experiment that minimizes the expected uncertainty on the parametrical knowledge of the system.

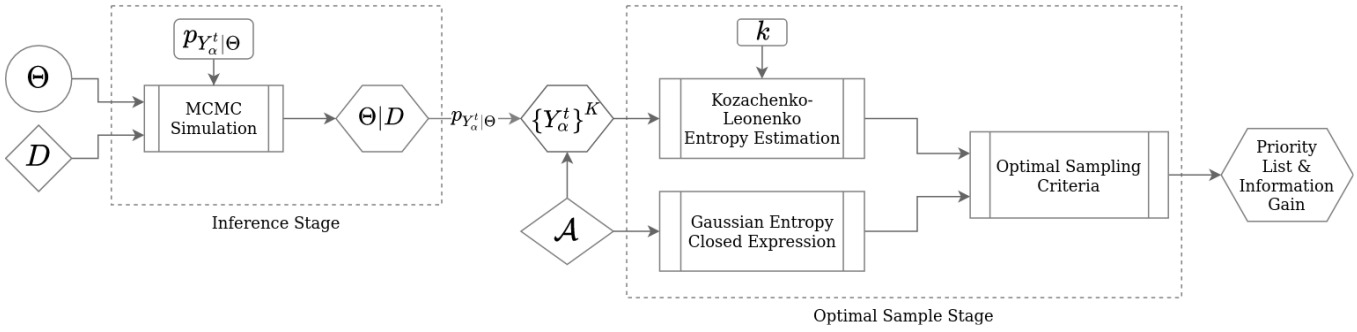


Figure 4.2: Full Optimal Sampling Criteria model. The circle represent the input prior distribution, the diamond a set of finite data and the hexagon a list of resultant data processed by the model

Chapter 5

Results and Analysis

The Optimal Sampling Criteria model gives a methodological framework which allows sorting by priority a set of possible observation on a system from a plausible agenda. Under the scenario where all the statistical information about the data and the system is perfectly defined, i.e. we know an expression for $p_{\Theta|D}$, the priority list given by our criteria is, in Equation 4.5, optimal for the set of experiments present in the agenda $\mathcal{A}$, under the objective of minimization of the expected conditional entropy of the distribution $p_{\Theta|D}$ with the new suggested data point $p_{Y_\alpha^t}$. On the other hand, when only estimations and raw data is available, i.e., only empirical estimations are available for $p_{\Theta|D}$ (as expected in a real practical scenario), the Optimal Sampling Criteria gives an approximation of the optimal list for a given agenda $\mathcal{A}$. The strong dependency of out criteria on the data motivates an extensive study for practical considerations. In order to obtain relevant information, all the tests must be performed for real-like observational scenarios using real and simulated data. The main objective of this section is to find empirical results to validate the Optimal Sampling Criteria for the task of predicting the best experimental setting.

For this empirical analysis, an Information Gain study is proposed. We study the joint entropy of parameters and observations and the resultant marginal entropy of the parameters obtained after a new observation has been made. It is expected that different observations will lead to different values for the posterior entropy, attributed to the quality of the observations. It is also expected that the OS Criterion will predict the experiments where the expected entropy is minimal for the set of observations in the plausible agenda $\mathcal{A}$. Two experimental scenarios are proposed:

- A simulated environment where the data is generated synthetically. The generator will be referred as the oracle and will produce data imitating the probabilistic phenomenon of observation. The parameters of the oracle will be set by the designer and it will serve as a reference to measure the performance of the posterior inference.
- A real experimental setting where the only information available is real data from a certain system. Naturally in this case the ground-truth parameters are missing meaning that only a entropy analysis on the parametrical space could be done.

5.1. Orbital data simulation

In a real observational scenario the true parameters of the system are unknown, this fact makes any performance analysis over an inference strategy non-feasible because the absolute error is immeasurable. On the other hand, in a controlled experimental environment, an artificial phenomenon can be simulated, implying that the ground-truth value for the system's parameters are available. This fact allows us to study the accuracy of an experimental scheduling criterion. The accessibility of the actual ground-truth state of the system and the possibility of synthetically generating data for any particular instant of observation makes the simulated experimental setting the most appropriate way to measure the expected behavior of an Optimal Sampling Criteria.

With the intention of testing the capabilities of the Optimal Sampling Criteria, the experimental setting will consider a collection of observations generated by a controlled noisy simulation from a given ground-truth set of parameters. By using the available observations and a given prior distribution p_{Θ}, a MCMC simulation is performed; the obtained particles from this process represent samples from $p_{\Theta|D}$. They will be utilized to empirically estimate the terms presented in Equation 4.8. In order to obtain the priority list from the agenda $\mathcal{A}$, an instance of the Observation Model present in Figure 4.2 is fed. This generates samples from the Post-OPD presented in Equation 2.9 for each entry in the agenda $\mathcal{A}$. Finally, using the expression for the differential entropy of a Gaussian distribution and the Kozachenko-Leonenko differential entropy estimator over the samples of each Post-OPD previously described, a list with the Optimal Sampling values for each case is created and sorted. This procedure will be applied for each experimental case, bringing a prediction of the optimal instant to observe the system.

In order to measure the actual change in the posterior inference of the system for each case, a set of n_{try} data points for each entry on the plausible Agenda are simulated. Consequently, the averaged response of the obtained inferences in each entry of $\mathcal{A}$ is reported and compared to the predicted behavior from the Optimal Sampling Criteria discussed before.

For this sub-section, a ground-truth set parameters coming from an artificial system are selected. The full set is defined on the first entry of Table 5.1 and Table 5.4. The chosen values correspond to a representative reference of the most common observable systems from the earth; also, the values are rounded to facilitate the analysis over the experimental cases. This section will consider two scenarios regarding the kind of data available and the plausible agenda:

- An astrometric case where the plausible agenda is defined as $[2020, 2020.6, 2021.5, 2022.4]$, or in terms of the parameters of the system $[T, T + 0.2P, T + 0.5P, T + 0.8P]$. All of the cases consider an instrumental noise with $\alpha = 0.003$. For each entry on the Agenda, $n_{\text{try}} = 5$ observations are simulated.
- A spectroscopic case where the plausible agenda is defined as $[2023, 2020.6, 2021.5, 2022.4]$, or in terms of the parameters of the system $[T + P, T + 0.2P, T + 0.5P, T + 0.8P]$. All of the cases consider an instrumental noise with $\alpha = 0.5$. For each entry on the Agenda $n_{\text{try}} = 5$ observations are simulated.

5.1.1. Astrometric data

The first experimental scenario corresponds to the inference of the simulated system using only astrometric observations. Due to the Equations discussed in Appendix A.1 only 7 parameters can be inferred from only astrometric observations. This set of parameters is presented in Table 5.1. To address this experimental case, five astrometric measures are generated randomly, but sparse enough to cover an important portion of the visible orbit. All of the available observations are measured in the interval of time $[T, T + P]$ and have an observational error characterized by $\alpha = 0.003$[arcsec] (typical value for nowadays observational instruments). The set of observations as well as the true orbit of the system are presented in Figure 5.1. It is important to mention that the base inference obtained is fairly close to oracle parameters. The only important difference comes from the angular values. This discrepancy is expected due to the dual representation of angles discussed in Appendix A.1 and the level of noise present our in the observations.

By using the base inference and its MCMC particles coming from $p_{\Theta|D}$, $n_{\text{try}} = 5$ sample observations are simulated for each entry in the agenda. Then an independent inference using the MCMC simulation is executed over a generated data point, in the same fashion as performed in the base inference. The resultant representative orbits and all the available observation points for the candidates are presented in Figure 5.2. Their respective predicted inference results via MAP rule over the posterior distribution, are reported in Table 5.1.

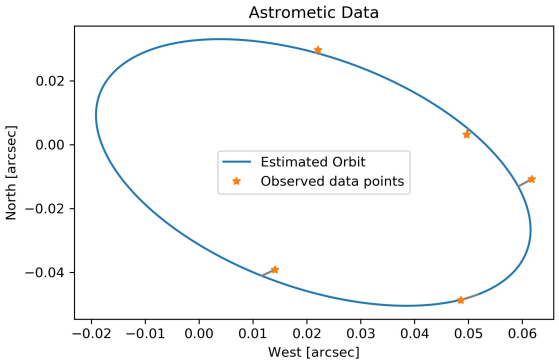


Figure 5.1: Astrometric observations from the synthetic system and the ground-truth orbit.

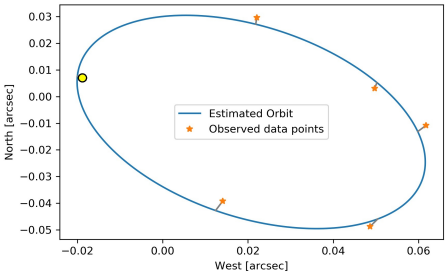
Table 5.1: Parameters from the oracle and inferred via MAP rule from a MCMC simulation. Astrometric only simulated scenario

Case	New data at	T	P	e	a	i	Ω	ω
Oracle	-	2020	3	0.5	0.05	45	35	40
Base Inf	-	2020.072	3.004	0.526	0.052	50.063	32.717	47.366
2020.0	T	2019.991	3.017	0.511	0.049	44.057	35.912	38.132
2020.6	T+0.2P	2020.067	3.018	0.517	0.052	50.808	35.595	45.637
2021.5	T+0.5P	2020.053	3.017	0.519	0.051	48.636	33.387	46.450
2022.4	T+0.8P	2019.998	3.059	0.455	0.053	50.494	42.069	36.669

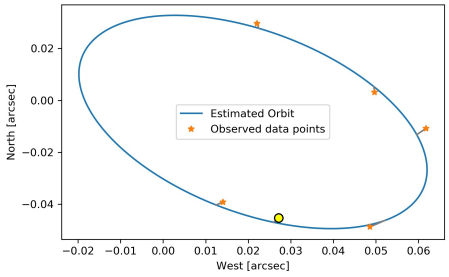
As it can be noticed in the Table 5.1, except for the agenda $\mathcal{A}$ is close to the base inference,

the noticeable difference is observed in the angular parameters. However, as previously discussed, this behavior is expected and it can be attributed to the angular ambiguity in the representation when only astrometric observation are available. A naive analysis over the inference problem by only studying the inferred values on Table 5.1 could conclude that no information was gain by adding a new observation point (simulated in this case) in different instants of time. A further analysis on the proposed Optimal Sampling Criteria will show that a substantial difference between the scenarios exist and it is measurable by the statistical tools described in Section 2.

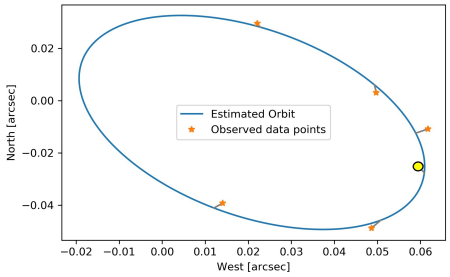
In Section 3.1.1 we discussed that a deterministic only analysis on the inferred parameters of the system does not bring information about the statistical properties on the inference and, consequently, does not study the entire phenomenon of inference. In a Bayesian framework, an inference is a decision made over a probabilistic object, such as $p_{\Theta|D}$. In the particular case of out study, a Maximum A Posteriori (MAP) rule is applied over the inference. This decision is not the only information about the parameter estimation. Indeed by re-introducing the concepts of Section 2.2.2 and the terms in 4.8, an entropy study is proposed here.



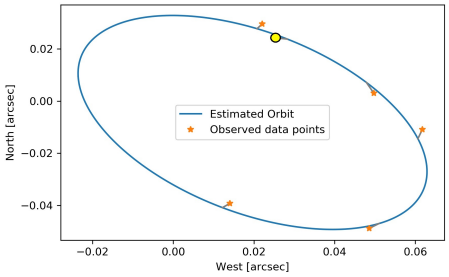
(a) New data point at instant T



(b) New data point at instant $T + 0.2P$



(c) New data point at instant $T + 0.5P$

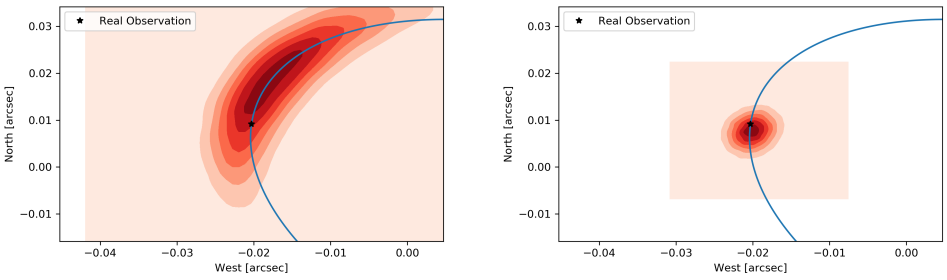


(d) New data point at instant $T + 0.8P$

Figure 5.2: Inferred orbits for each candidate, astrometric study. New observation is highlighted in each case.

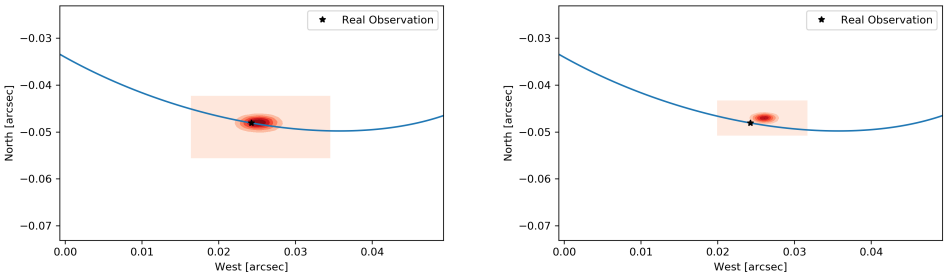
The Posterior Predictive Distribution (or Post PD) is the object obtained by projecting the posterior distribution $p_{\Theta|D}$ of the system in the observational Space, as described in Section 2.2.1. In this experimental analysis, two probabilistic objects will be introduced: The Prior Predictive Distribution (or Prior PD $p_{\Theta|D}$) and the Posterior Predictive Distribution (or Post PD $p_{Y^{t^i}|D,Y^{t_i}_{\alpha_i}=y^t_{i,j}}$) resultant from the jth simulation of the ith candidate in the agenda $\mathcal{A}$. Figures 5.3, 5.4, 5.5 and 5.6 present both Prior and Post PD and they show the effect of adding a new observation point in our inference. The drawn distributions in the Figures 5.3, 5.4, 5.5 and 5.6 are estimations obtained using a Gaussian kernel over all the particles

available of the Prior and Post PD. On the other hand, the pink area represents the smaller quadrilateral that contains all the particles of each distribution.



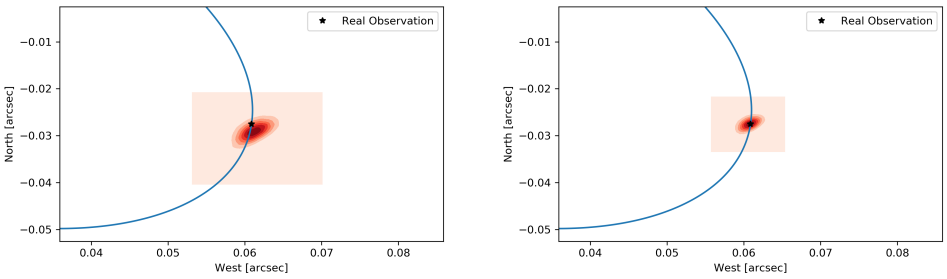
(a) Prior PD at T (b) Case 1 Post PD at T

Figure 5.3: PPD comparison for 1st case, astrometric study.



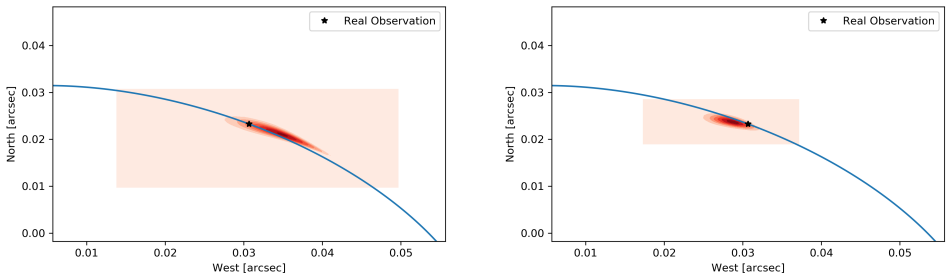
(a) Prior PD at $T + 0.2P$ (b) Case 2 Post PD at $T + 0.2P$

Figure 5.4: PPD comparison for 2nd case, astrometric study.



(a) Prior PD at $T + 0.5P$ (b) Case 3 Post PD at $T + 0.5P$

Figure 5.5: PPD comparison for 3rd case, astrometric study.



(a) Prior PD at $T + 0.8P$ (b) Case 4 Post PD at $T + 0.8P$

Figure 5.6: PPD comparison for 4th case, astrometric study.

As shown in Figures 5.3, 5.4, 5.5 and 5.6 for every case in the agenda a noticeable uncertainty reduction is observed. This reduction through each instance shows that new observations reduce the variance on the inferred observational space, implying a tighter distribution. The intuition, using Equation 4.8 for the Optimal Sampling Criteria, suggest that this variance reduction in the observational space will imply a significant reduction on the parametrical variable as well. By looking at the reductions in Figures 5.3, 5.4, 5.5 and 5.6, a preliminary optimal list can be sorted in the ordering $[1^{st}, 2^{nd}, 3^{rd}, 4^{th}]$ of the candidates in the agenda $\mathcal{A}$. It is clearly observable that the first candidate in Figure 5.3 presents the greatest uncertainty reduction, meanwhile an arguable ordering can be suggested for the rest of cases. The analysis over the Prior and Post PD on the agenda $\mathcal{A}$ shows empirically that regardless of the imperceptible change on the inferred parameters, a new observation from a system will imply a uncertainty reduction in the estimated posterior distribution.

A formal analysis of the entropy reduction can be achieved by estimating the joint and marginal differential entropy of the posterior distribution of each case and comparing it with the base inference reference. The observed Information Gain for the posterior distribution $p_{Y^t|D,Y^{t_i}_{\alpha_i}=y^t_{i,j}}$ is defined as the mean entropy measured through all the $n_{\text{try}} = 5$ simulations for each candidate in the agenda minus the entropy estimated for the base inference case. The obtained Information Gain for each case is reported in Table 5.2.

Table 5.2: Information Gain referred from the base inference for each new observation case. Astrometric only simulated scenario

Case	New data at	T	P	e	a	i	Ω	ω	joint
2020.0	T	2.46	1.67	2.01	1.29	0.91	0.95	1.12	9.45
2020.6	T+0.2P	3.79	-0.84	1.46	-0.82	0.50	0.44	0.94	5.64
2021.5	T+0.5P	0.05	0.56	0.36	0.11	0.04	0.18	0.08	2.9
2022.4	T+0.8P	0.01	0.79	0.84	0.18	0.1	0.29	0.06	0.78

An analysis on the joint Information gain of the Table 5.2 shows that the empirically estimated values attend the same ordering predicted by the visual inspection Figures 5.3, 5.4, 5.5 and 5.6. This implies that the relationship between the observational and parametrical distributions generated by the observation model in Figure 2.4, do in fact, hold a relationship that allows to relate the reduction of one with the other. Table 5.2 also describes the estimated values for the marginal Information Gain for each candidate. It is possible to identify that in most cases, a greater join IG implies a greater marginal IF for most of the parameters.

However, it is also possible to notice that for certain cases (e.g. 2020.6), a trade-off between the marginal IG occurs. We observe that parameters such as P and a increase their respective entropy in order to narrow the posterior distribution for T and consequently, inducing a greater value for the joint Information Gain. Unfortunately this particular behavior is not covered by the discussion in Section 4, suggesting further research on this matter for future works.

An important aspect to remark is the existence of negatives values for the marginal Information Gain. Two things can be said to this seemly anomalous behavior: First, the Optimal Sampling Criteria states that it is expected a joint positive Information Gain, but it does not guarantee the same for the marginal entropy, implying that a negative value it is plausible for a finite set of sample from $p_{Y_{\alpha_i}^{t_i}}$. Second, this entropy analysis is empirical, meaning that an estimation of the entropy via particles (i.i.d. samples) is made. The existence of an important estimation error is likely from a trial of $n_{\text{try}} = 5$ samples.

Once all the practical analyses from the synthetic simulation of the system have been made, it is pertinent to contrast the results with the prediction of the Optimal Sampling Criteria proposed in this work. In order to measure the performance of the OS Criteria, a prediction over the agenda $\mathcal{A}$ is made using only the estimated distribution $p_{\Theta|D}$ from the MCMC simulation on the base inference. The results are reported in Table 5.3.

Table 5.3: Joint Information Gain and Expected Information Gain for each new observation case. Astrometric only simulated scenario

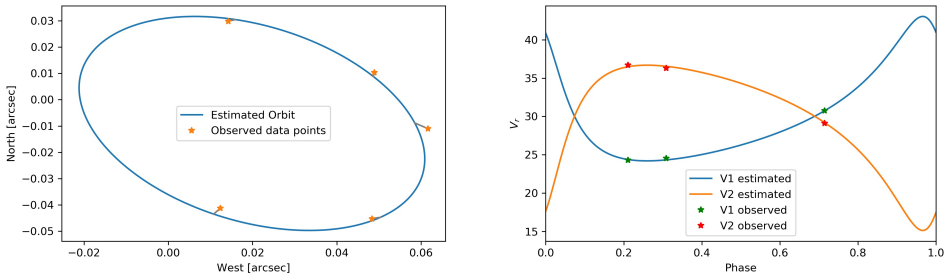
Case	New data at	IG	$\mathbb{E}[IG]$
2020.0	T	9.45	0.0102
2020.6	T+0.2P	5.64	0.0095
2021.5	T+0.5P	2.90	0.0078
2022.4	T+0.8P	0.78	0.0049

The Table 5.3 presents the estimated join values from Table 5.2 and contrast them with the expected Information Gain from the Optimal Sampling Criteria as presented in Equation 4.8 shows. It is noticeable that both columns in Table 5.3 present the same ordering, implying that the OS criteria is, in fact, capable of detecting the most relevant instant of observation on a finite agenda $\mathcal{A}$ by only analyzing the inferred distribution $p_{\Theta|D}$. It is also evident that the empirical estimation and the predicted Information Gain differ in magnitude, and this discrepancy does not correspond to a linear rescaling. This difference is mainly explained due to the difference in the dimension of the variable estimated. In the next section, it will be studied how and when these limitations occur.

5.1.2. Astrometric and RV data

With the purpose of studying our method in a different setting, an environment with Astrometric and Radial Velocity data, is considered. Following the same structure presented in Section 5.1.1, a set of observations from the synthetic system is simulated. On this occasion, an arrange of Astrometric and Spectroscopic observations is available. The main objective of this study is to analyze how two types of observations affect the inference and if it is possible

to apply the OS Criteria in this context. In Figure 5.7 the set of observations and the ground-truth orbits are presented. This arrangement is composed of five Astrometric observations with $\alpha = 0.005$[arsec] and three pairs of Radial Velocity observations with $\alpha = 1$[km/s]. The base inference for this experimental setting is obtained via MCMC simulation by using the phenomenological equations described in Section 2.1.1 and 2.1.2, in the same fashion as Section 5.1.1.

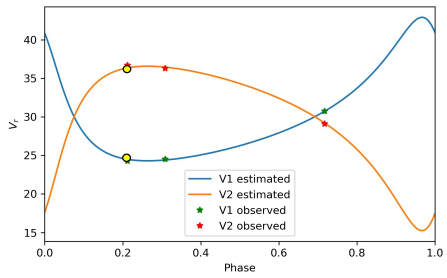


(a) Astrometric data and base inference

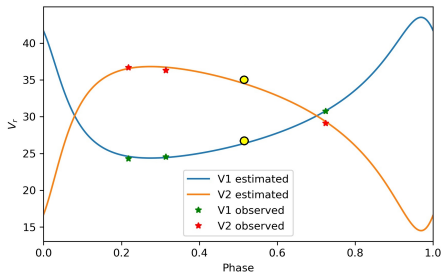
(b) RV data and base inference

Figure 5.7: Astrometric and RV from the synthetic system and the ground-truth orbit.

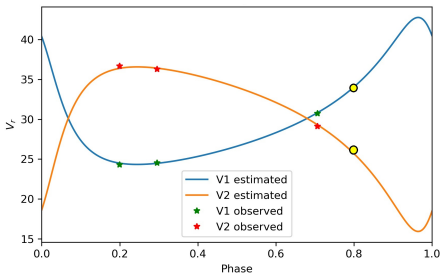
Following the same structure as the previous experimental scenario, four new candidates observations are proposed in a new agenda $\mathcal{A}$ to study the Information Gain over each case. In this setting, four pairs of observations for Radial Velocities are proposed in the time instant $[T + 0.2P, T + 0.5P, T + 0.8P, T + P]$ with $\alpha = 0.5$. By using the candidates in $\mathcal{A}$, $n_{\text{try}} = 5$ simulations for each case are generated and MCMC estimation of the posterior distribution $p_{\Theta|D,Y_{\alpha_i}^{t_i}=y^{t_{i,j}}}$ is performed. A representative iteration of each candidate and its resulting inferences are presented in Figure 5.8.



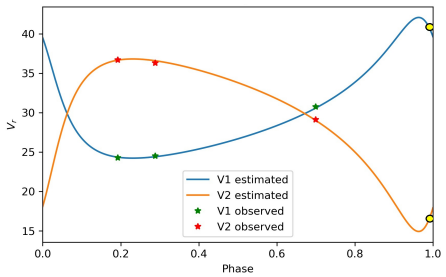
(a) New data point at instant $T + 0.2P$



(b) New data point at instant $T + 0.5P$



(c) New data point at instant $T + 0.8P$



(d) New data point at instant $T + P$

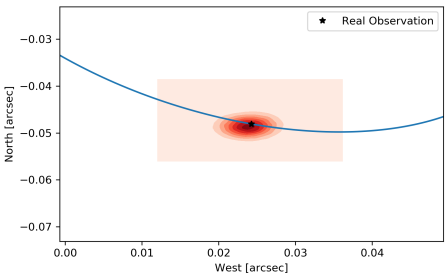
Figure 5.8: Inferred orbits for each candidate, combined study. New observation is highlighted in each case.

Under the same paradigm as the section before, the MAP inference of each candidate are averaged and presented in Table 5.4.

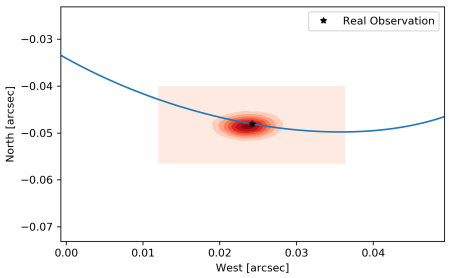
Table 5.4: Parameters from the oracle and inferred via MAP rule from a MCMC simulation. Double line simulated scenario

Case	New data at	T	P	e	a	i	Ω	ω	q	π	V_{cm}
Oracle	-	2020.0	3.0	0.5	0.05	45.0	35.0	40.0	0.8	0.02	30.0
Base Inf	-	2019.97	2.99	0.49	0.05	41.33	34.62	38.56	0.87	0.02	30.02
2020.6	T+0.2P	2019.98	2.99	0.49	0.05	41.37	35.13	37.82	0.87	0.019	30.02
2021.5	T+0.5P	2019.96	2.99	0.49	0.049	40.86	36.38	35.37	0.86	0.018	30.12
2022.4	T+0.8P	2020.017	2.99	0.49	0.05	45.54	35.35	41.10	0.89	0.02	30.11
2023.0	T+P	2020.04	2.99	0.50	0.05	47.58	34.81	43.43	0.82	0.02	29.88

The behavior presented in Table 5.4 follows the same paradigm as the previously discussed experimental in Sections 5.1.1. . All inferred parameters are fairly close to the true value, meaning that in a deterministic scope no much information has been gained by the new observations point. The only relevant difference across the inferred parameters is the $4th$ case of the agenda, where the mass ratio between the stars q reaches a closer value to true. From this result, it could be expected that the $4th$ candidate gives a greater Information Gain than the others. In order to visually inspect this intuition, a study on the PD projected on the orbital space is presented in Figures 5.9, 5.10, 5.11 and 5.12.

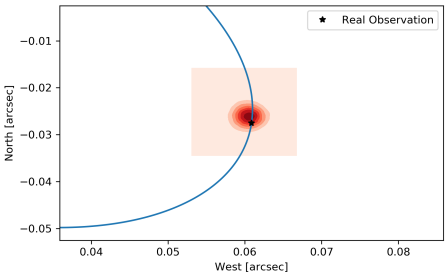


(a) Prior PPD at $T + 0.2P$

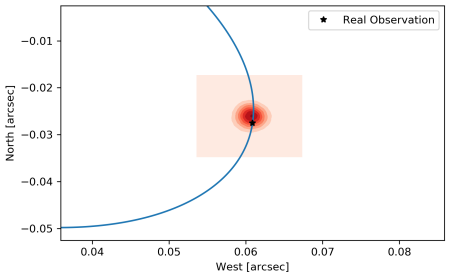


(b) Post PPD at $T + 0.2P$

Figure 5.9: PPD comparison for 1st case, combined study.

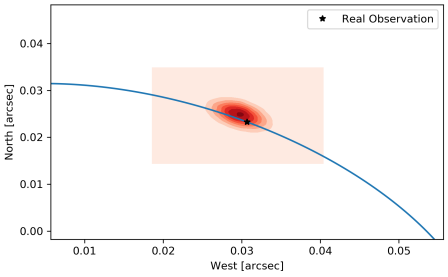


(a) Prior PPD at $T + 0.5P$

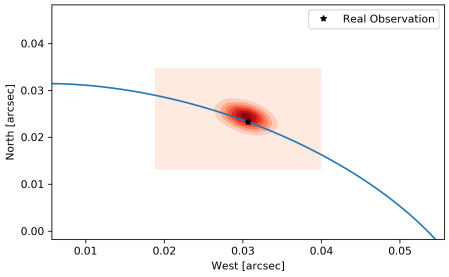


(b) Post PPD at $T + 0.5P$

Figure 5.10: PPD comparison for 2nd case, combined study.

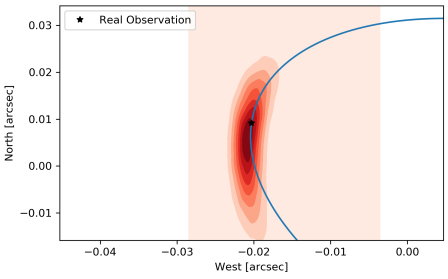


(a) Prior PPD at $T + 0.8P$

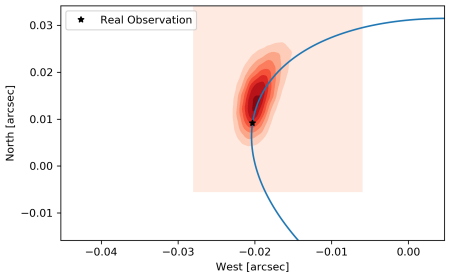


(b) Post PPD at $T + 0.8P$

Figure 5.11: PPD comparison for 3th case, combined study.



(a) Prior PPD at $T + P$



(b) Prior PPD at $T + P$

Figure 5.12: PPD comparison for 4th case, combined study.

The Figures 5.9, 5.10, 5.11 and 5.12 shows the difference in certainty for each case in the

agenda. The most evident difference in Figure 5.12 for the $4th$ candidate, where a noticeable reduction happens between Prior and Post PD in a astrometric projection. This behavior suggests that the $4th$ candidate will improve the quality of the inference when this observation is taken. For the $1st$, $2nd$ and $3rd$ candidates a arguable reduction can be detected through Figures 5.9, 5.10 and 5.11. From this result, a ordered list of Information Gain candidates could be guessed, as $[4th, 3rd, 1st, 2nd]$.

It is essential to notice that the newly added data correspond to a Spectroscopic observation of the primary and secondary star. So it could be expected that no changes will occur in the Post PD projected in the Astrometric Space. Nonetheless, a substantial reduction happens in most cases, meaning that underlying relationships between both kinds of observations exist.

In the same fashion as the previously discussed environment, an Information Gain Analysis is performed over each marginal parameter and the joint Information Gain of the whole set of simulations. The results are presented in Table 5.5.

Table 5.5: Information Gain referred from the base inference for each new observation case. Double line simulated scenario

Case	New data at	T	P	e	a	i	Ω	ω	q	π	V_{cm}	joint
2020.6	T+0.2P	0.39	0.55	0.11	0.02	-0.03	0.06	0.05	-0.01	0.12	-0.16	-0.04
2021.5	T+0.5P	0.71	1.13	1.01	0.11	0	0.26	0.29	0	0.41	-0.04	0.64
2022.4	T+0.8P	0.81	-0.14	0.38	0.14	0.31	0.06	0.06	0.37	0.2	1.03	2.71
2023.0	T+P	3.31	0.78	-3.12	0.09	0.44	0.14	0.07	1.27	0.23	2.66	4.31

The Table 5.5 describes the marginal and joint differential entropy estimation for each simulation round of the agenda. The presented values follow the same behavior as seen in Table 5.2, where some of them take negative values but others compensate this by acquiring greater marginal Information Gain. Nonetheless, for the case 2020.6, a negative value is reached for the joint Information Gain. This apparently contradicts the statement of semidefinite positivity of Equation 4.3. This negative behavior can occur because an empirical estimation of entropy is performed, meaning that an error can exist when the quantity of samples of the posterior distribution $p_{\Theta|D}$ and the amount of trials n_{try} is no infinite. In conclusion, negative value observed in Table 5.5 is considered zero for practical purposes, meaning that that particular instant observation does not carry novel information about the observed system.

In order to study the performance of the Optimal Sampling Criteria in a predictive selection, the man estimated values of Information Gain of Table 5.5 are presented in conjunction with the expected Information Gain estimated through the Optimal Sampling Criteria.

Table 5.6: Joint Information Gain and Expected Information Gain for each new observation case. Double line simulated scenario

Case	New data at	IG	$\mathbb{E}[IG]$
2020.6	T+0.2P	-0.04	0.129
2021.5	T+0.5P	0.64	0.170
2022.4	T+0.8P	2.71	0.253
2023.0	T+P	4.31	0.253

By observing the presented values of Table 5.6, it is possible to notice that both esti-

mated Information Gain and the expected information Gain can be sorted in the same order. This implies that, once again, the Optimal Sampling Criteria can detect the most crucial observation to reduce the posterior expected entropy in an inference. Similar to the behavior observed in Section 5.1.1, the periastron is the most informative instant of observation. Nonetheless, the dimensionality problems present in empirical estimation are still perceived in this analysis, meaning that it is not possible to predict the information gain when only few samples are available of the posterior distribution due the induced bias observable in Table 5.6, only a priority list can be predicted.

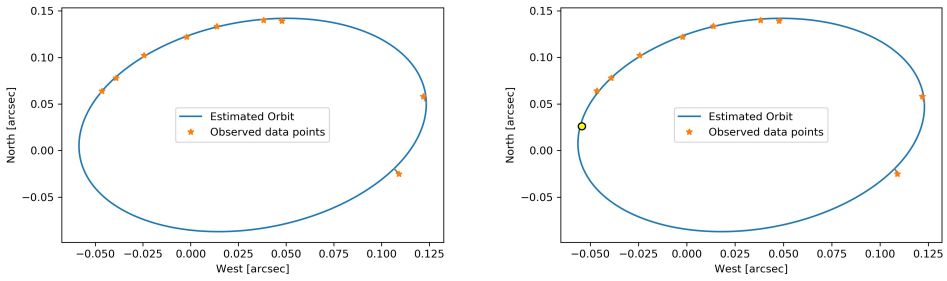
One important feature that can be noticed in this Radial Velocity simulated setting is that the instant of observation which reduces the posterior parametrical uncertainty the most occurs in the periastron of the orbit. This idea is commonly used (in works such as [Lucy, 2014]) to confidently determine the individual masses of the participant stars. This result shows that our Optimal Sampling Criterion gives a reasonable priority list of experiments and can support the idea of measuring the periastron of a system, regardless of the kind of observation.

5.2. Real Data

In order to measure the acquisition of information, once a new piece of data is observed for a particular system, a real data case is proposed given a collection of real observations. These observation points were collected for survey campaign purposes, and it will be utilized to measure how a statistical inference change when a new data point is added. The experimental setup is composed of 6 binary system (selected using the methodology presented in Appendix D) with only astrometric data available. The number of data points varies from case to case, but all of them includes an astrometric measure made at instant $t = 2020.824$ with $\alpha = 0.003$. In the same fashion as the section before, two inferences are performed for two different sets of data points: the first one is composed of all the available points but the one at $t = 2020.824$. The second set is the full collection of astrometric observations. The main objective of this study is the comparison of the acquisition of information, i.e. the reduction of joint parametrical entropy and the Expected Information Gain predicted from the Optimal Sampling Criterion.

Table 5.7 presents different inferences for the same system. The first entry corresponds to the reported inference from the catalog of binary orbits ORB6, where authors publish orbits found for certain systems; the best ones are reported. The entries with prefix TK are obtained from an inference method described and available in [Tokovini, 2020], which uses classical methods to find a plausible orbit: these results are deterministic. The third entries, denoted by the prefix U-1, utilizes the MCMC simulation as proposed in this work over the available data. The prefix U is set for the last one, which denotes an inference using the same MCMC simulation but with all the available data and a new data point. It is important to mention that, the first two entries correspond to a deterministic approach for obtaining the orbits, meaning that Information Theory measures cannot be applied. This happens because the knowledge on the parametrical space is not modeled as a random variable.

By analyzing the inferred values on Table 5.7 we note in general, with the exception of the period and eccentricity of the first case, that the predicted parameters are maintained through different methods and data supporting the idea that a sufficient quantity of data produces a stabilized inferences. Consequently a good behavior of the OS Criterion is observed, for example, in Figure 5.13 shows how stable are the inferences are for both cases.



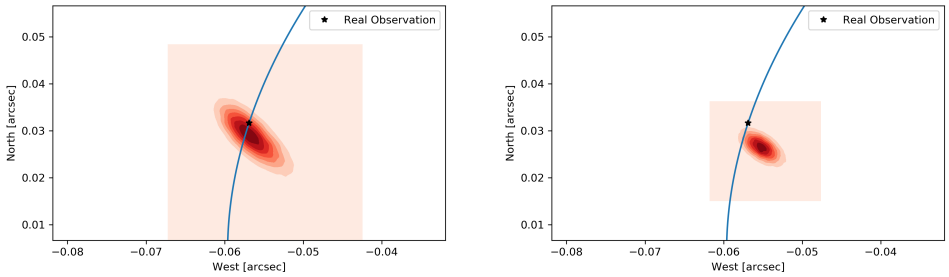
(a) U-1-HIP10885 case

(b) U-HIP10885 case

Figure 5.13: Comparison of predicted orbits, case HIP10885. New observation is highlighted.

As discussed previously, our Bayes approaches allow to model how the certainty of an

inference varies when new data is added. Figure 5.14 shows how the new data point at instant $t = 2020.824$ helps to reduce uncertainty regarding that instant t. Naturally, a reduction in the parametrical space is expected, but the value is not evident from just a visual inspection over the projected PPD.



(a) Prior PPD for case at instant $t = 2020.824$

(b) Post PPD for case at instant $t = 2020.824$

Figure 5.14: PPD comparisson for the case HIP10885.

Table 5.7: Orbital parameters inferred by various methods for 6 cases of real data.

System	T	P	e	a	i	Ω	ω
HDS866	2023.03	28.23	0.66	0.19	81.4	82.6	243.7
TK-HDS866	2026.81	75.09	0.35	0.16	76.82	86.19	217.37
U-1-HDS866	2027.36	76.4	0.37	0.16	77.6	86.06	219.55
U-HDS866	2027.09	64.59	0.43	0.16	78.4	86.56	234.1
HIP109634	2022.09	35.7	0.78	0.18	124.2	57.2	49.1
TK-HIP109634	2025.45	46.81	0.44	0.17	122.37	30.83	5.18
U-1-HIP109634	1991.25	29.95	0.99	0.76	96.43	66.61	83.79
U-HIP109634	1991.25	30.71	0.99	0.86	95.22	60.02	83.85
HIP9497	2006.39	14.4	0.06	0.1	152.5	120.5	227.2
TK-HIP9497	2006.69	14.16	0.08	0.1	151.58	131.81	241.86
U-1-HIP9497	2020.94	14.41	0.07	0.1	153.31	127.38	236.54
U-HIP9497	2020.85	14.18	0.08	0.1	151.7	131.7	241.52
HIP10885	1998.52	24.08	0.39	0.12	40.8	165.4	318.2
TK-HIP10885	1998.41	24.02	0.39	0.12	42.94	167	314.96
U-1-HIP10885	1998.37	24.05	0.4	0.12	41.96	167.93	313.89
U-HIP10885	1998.32	23.93	0.41	0.12	42.5	171.22	309.5
HIP14524	2021.66	38	0.55	0.3	140.6	46.1	258.2
TK-HIP14524	2021.01	34.51	0.65	0.32	131.88	22.7	79.11
U-1-HIP14524	2020.99	34.44	0.65	0.32	131.59	40.56	259.14
U-HIP14524	1991.25	29.75	0.71	0.32	127.88	26.25	257.35
HIP32767	2016.04	47.85	0.21	0.17	55.6	22.5	231.8
TK-HIP32767	2015.69	48.69	0.22	0.17	55.8	21.87	228.71
U-1-HIP32767	2015.94	47.97	0.21	0.17	55.63	22.45	230.79
U-HIP32767	2015.46	49.31	0.22	0.17	56.19	21.84	226.33

To measure the Information Gain empirically, an entropy estimation is realized for each stochastic inference available. Also, a prediction using the Optimal Sampling Criteria model

is made, the values obtained are reported in Table 5.8.

Table 5.8: Information Gain and Expected Information Gain

System	JE	JE+1	IG	$\mathbb{E}$[IG]
U-HDS866	11.449	9.017	2.432	0.348
U-HIP109634	-16.938	-44.543	27.605	0.049
U-HIP9497	-4.212	-19.435	15.223	0.058
U-HIP10885	-19.642	-40.585	20.943	0.008
U-HIP14524	-136.172	-179.414	43.242	0.01
U-HIP32767	-0.844	-2.177	1.333	0.095

As expected in each entry of the Table 5.8 an Information Gain is observed. The magnitude of this gain varies and naturally dependent of the observed system and the available data. On other hand the magnitude between the expected Information Gain from the OS Criterion and the observed is maintained shown in previous experiments. The main reason of this behavior is the difference of dimensions between the observational and parametrical space. Considering that the entropy estimation used in this work is based in measure of distance between particles, it is evident that this estimation suffers from the "curse of dimensionality"; i.e. the higher the dimension of a problem implies a higher complexity when comparing distance between points. The theory says that parametrical entropy and observational entropy must stay in the same magnitude, but when an empirical approximation is made this relationship is broke due dimensional issues. Therefore this behavior implies that no comparison can be made between different systems because practical reasons and then the possibility of a multiple system selection of Optimal Sampling is no available with the actual configuration of the model. Nonetheless due the oracle study made in Section 5.1. it is possible to affirm that in the context of comparing instants of observation for the same system, the OS Criteria model is capable of predict with high degree of certainty the priority order of a set of observations.

In summary, the experimental study on the Information Gain in the parametrical inference of binary systems, and the performance on detecting a priority list of the Optimal Sampling criteria, proves that the reduction of uncertainty in the parameters of a binary system is observable. The proposed method can detect the order of priority in measurements when a minimal quantity of data is accessible. Visual inspections on the Prior PPD can guide the selection of an optimal experiment. However, for accurate decisions, the OS Criterion has been shown to be an unbiased selector for realistic observation scenarios. Finally, the simulated environment has been an essential experimental setting to prove the capacities of the OS model due to the ability to generate noisy data and access to the ground truth parameters. This experience also validates the intuitive belief that the periastron is a crucial instant to observe. The OS criterion selects this sample instance even if not direct data of the periastron have been given or the base inference differs from the ground truth.

Chapter 6

Final Remarks and Future Work

This work presents an extensive description of how to characterize any observation on a system and how the prior and posterior knowledge of the parameters is function of available observations, implying that observation model can be constructed and the relationship between existing observations and new candidates is mathematically explicit. The observation model also shows theoretically that any new observation made over the system will always implies an expected reduction of uncertainty on the parameters, this statement is also proven practically by measuring the uncertainty difference in all cases of simulated and real data in Section 4.

The Optimal Sampling Criterion proposed in this work is a theoretical and practical framework for optimal scheduling observations for binary systems. This methodology is based on selecting an optimal experiment from a plausible agenda with a finite amount of candidates. Analysis using simulated and real data proves that the Optimal Sampling Criterion detects high informative data points from an agenda and helps the designer to find observations that reduce the conditional posterior entropy more consistently to enhance the inference. Our method is evaluated empirically by data simulation and expected entropy analyses, proving that in a practical approach the proposed Optimal Sampling Criterion. An analysis over the simulated data shows that the OS criterion supports in heuristic approaches, such as the importance of observing the periastron of a system, used astronomy to select instants of observation.

Both Optimal Sampling Criterion and observation model presented in this work are described in the most generalized form to serve as a basis for extension of other forms of observational phenomena.

The flexibility of the explored OS criterion and the observational model allows to implement this framework in a wide range of observational settings, for example, astronomical surveys such as LSST from the Rubin observatory or the SDSS from the Apache Point observatory. These observational campaigns represent a perfect scenario to implement the proposed framework due to the imperative necessity of automatizing the labor of selecting observations for each studied phenomenon. Naturally, the Optimal Sampling Criterion can also be used to study the complexity of specific systems or justify heuristic decisions when planing the observation of such systems.

As future works, the following interesting ideas are suggested:

- Improve the stability of the inference stage. Current research on new Monte Carlo-based method could be a source of huge improvement in the performance of put approach and also will provide a better samples from the target distribution, promoting better estimated values.
- Research on differential entropy estimators reduces the effect of dimensionality on the problem, resulting in better estimations for the implementation of our OS criterion.
- Exploring new frameworks to reduce the expected entropy of a random variable that is a function of the parameters of a studied system. A relevant example could be systematic minimization of the expected uncertainty of the individual masses of a system instead of their complete set of parameters.
- An systematic astronomic study on the different observable binary systems and definition of the notion of observational complexity. This could be achieved by studying the expected entropy reduction of the parameters from different observation strategies.

Bibliography

[Beraha et al., 2019] Beraha, M., Metelli, A. M., Papini, M., Tirinzoni, A., and Restelli, M. (2019). Feature selection via mutual information: New theoretical insights.

[Berrett and Samworth, 2017] Berrett, T. B. and Samworth, R. J. (2017). Nonparametric independence testing via mutual information.

[Blackwell, 1953] Blackwell, D. (1953). Equivalent comparisons of experiments. *Ann. Math. Statist.*, 24(2):265–272.

[Blackwell and Girshick, 1979] Blackwell, D. and Girshick, M. (1979). *Theory of Games and Statistical Decisions*. Dover Books on Mathematics. Dover Publications.

[Chaloner and Verdinelli, 1956] Chaloner, K. and Verdinelli, I. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005.

[Chaloner and Verdinelli, 1995] Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10:273–304.

[Cid Palacios, 1958] Cid Palacios, R. (1958). On the necessary and sufficient observations for determination of elliptic orbits in double stars. *AJ*, 63:395.

[Claveria, 2017] Claveria, R. (2017). Uncertainty characterization of orbital parameters in contexts of partial information. *Universidad de Chile*, pages 62–63.

[Claveria et al., 2019] Claveria, R. M., Mendez, R. A., Silva, J. F., and Orchard, M. E. (2019). Visual binary stars with partially missing data: Introducing multiple imputation in astrometric analysis. *Publications of the Astronomical Society of the Pacific*, 131(1002):084502.

[Cover and Thomas, 2006] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

[den Bos W. H, 1926] den Bos W. H, V. (1926). Orbital elements of binary systems. *Union Obs Circular*, 2:354.

[Docobo, 1985] Docobo, J. (1985). On the analytic calculation of visual double star orbits. *Celestial Mechanics*, 36(2):143–153.

[Ford, 2005] Ford, E. B. (2005). Quantifying the uncertainty in the orbits of extrasolar planets. *The Astronomical Journal*, 129(3):1706.

[Junmo Kim et al., 2005] Junmo Kim, Fisher, J. W., Yezzi, A., Cetin, M., and Willsky, A. S. (2005). A nonparametric statistical method for image segmentation using information

theory and curve evolution. *IEEE Transactions on Image Processing*, 14(10):1486–1502.

[Kozachenko and Leonenko,] Kozachenko, L. and Leonenko, N. Sample estimate of the entropy of a random vector". *Probl. Peredachi Inf*, 23(2):9–16.

[Kraskov et al., 2004] Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6).

[Loredo, 2004a] Loredo, T. J. (2004a). Accounting for source uncertainties in analyses of astronomical survey data. *AIP Conference Proceedings*.

[Loredo, 2004b] Loredo, T. J. (2004b). Bayesian adaptive exploration. *AIP Conference Proceedings*.

[Loredo et al., 2012] Loredo, T. J., Berger, J. O., Chernoff, D. F., Clyde, M. A., and Liu, B. (2012). Bayesian methods for analysis and adaptive scheduling of exoplanet observations. *Statistical Methodology*, 9(1-2):101–114.

[Lucy, 2014] Lucy, L. (2014). Mass estimates for visual binaries with incomplete orbits. *Astronomy & Astrophysics*, 563:A126.

[MacKnight and Horch, 2004] MacKnight, M. and Horch, E. (2004). Calculating visual binary star orbits with the downhill simplex algorithm (amoeba). 36:788.

[Mendez et al., 2017] Mendez, R. A., Claveria, R. M., Orchard, M. E., and Silva, J. F. (2017). Orbits for 18 visual binaries and two double-line spectroscopic binaries observed with hrcam on the ctio soar 4 m telescope, using a new bayesian orbit code based on markov chain monte carlo. *The Astronomical Journal*, 154(5):187.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

[Pourbaix, 1994] Pourbaix, D. (1994). A trial-and-error approach to the determination of the orbital parameters of visual binaries. *Astronomy and Astrophysics*, 290:682–691.

[Pourbaix, 1998] Pourbaix, D. (1998). Simultaneous least-squares adjustment of visual and spectroscopic observations of binary stars. *Astronomy and Astrophysics Supplement Series*, 131(2):377–382.

[Reffert, 2009] Reffert, S. (2009). Astrometric measurement techniques. *New Astronomy Reviews*, 53(11):329 – 335. Proceedings: VLTI summerschool.

[Sahlmann et al., 2013] Sahlmann, J., Lazorenko, P., Ségransan, D., Martín, E., Queloz, D., Mayor, M., and Udry, S. (2013). Astrometric orbit of a low-mass companion to an ultracool dwarf. *Astronomy & Astrophysics*, 556:A133.

[Sebastiani and Wynn, 2000] Sebastiani, P. and Wynn, H. (2000). Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society Series B*, 62:145–157.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

[Shewry and Wynn, 1987] Shewry, M. C. and Wynn, H. P. (1987). Maximum entropy sam-

pling. *Journal of Applied Statistics*, 14(2):165–170.

[Thiele, 1883] Thiele, T. N. (1883). Neue methode zur berechung von doppelsternbahnen. *Astronomische Nachrichten*, 104:245.

[Thiele, 1926] Thiele, T. N. (1926). On cowell's method of applying the newtonian law. *Astronomische Nachrichten*, 228:265.

[Tishby et al., 2001] Tishby, N., Pereira, F., and Bialek, W. (2001). The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, 49.

[Tokovini, 2020] Tokovini, A. (2020). Orbit: calculation of visual, spectroscopic, and combined orbits.

[Tokovinin, 1992] Tokovinin, A. (1992). The frequency of low-mass companions to k and m stars in the solar neighbourhood. *Astronomy and Astrophysics*, 256:121–132.

[van de Kamp, 1967] van de Kamp, P. (1967). Principles of astrometry. *American Journal of Physics*, 35(10):974–975.

[Vanlier et al., 2012] Vanlier, J., Tiemann, C., Hilbers, P., and van Riel, N. (2012). A bayesian approach to targeted experiment design. *Bioinformatics (Oxford, England)*, 28:1136–42.

[Vergara and Estevez, 2014] Vergara, J. and Estevez, P. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24.

Appendix A

Observational Data

A.1. Thiles Innes Representation

Thiele Innes representation is an alternative parametrical version to characterize the orbit of a binary system. The definition of the constants A, B, F and G in replace of a, ω, Ω and i is the main feature about the Thiles Innes representation. This new proposed constants are calculated as shown in Equation A.1.

$$\begin{aligned} A &= a(\cos\omega\cos\Omega - \sin\omega\sin\Omega\cos i) \\ B &= a(\cos\omega\sin\Omega + \sin\omega\cos\Omega\cos i) \\ F &= a(-\sin\omega\cos\Omega - \cos\omega\sin\Omega\cos i) \\ G &= a(-\sin\omega\sin\Omega + \cos\omega\cos\Omega\cos i) \end{aligned} \tag{A.1}$$

In this new representation, the parametrical vector is re-defined as $[T, P, e, A, B, F, G]$ instead of $[T, P, e, a, \omega, \Omega, i]$. In order to obtain the sky-plane relative position in a moment t using the Thiele Innes representation the auxiliary variables $\bar{x}(t)$ and $\bar{y}(t)$ must be calculated as shown in Equation A.2, where the Eccentric Anomaly $E(t)$ is defined in the Kepler's Equation (2.1) at instant t.

$$\begin{aligned} \bar{x}(t) &= \cos E(t) - e \\ \bar{y}(t) &= \sqrt{1-e^2}\sin E(t) \end{aligned} \tag{A.2}$$

Finally, the relative position of star B towards star A of the binary system in Cartesian coordinates (X, Y) is calculated as shown in Equation A.3.

$$\begin{aligned} X &= B\bar{x} + G\bar{y} \\ Y &= A\bar{x} + F\bar{y} \end{aligned} \tag{A.3}$$

The relationship between the standard representation $[T, P, e, a, \omega, \Omega, i]$ and the Thiele Innes representation $[T, P, e, A, B, F, G]$ is not bijective, i.e. there is no one-to-one relationship between both parametrical vectors, which means that multiples set of angles $[\omega, \Omega, i]$

are related to the same constants $[A, B, F, G]$. The Thiele Innes ambiguity is evidenced in Equation A.4.

$$\begin{aligned} \tan(\omega + \Omega) &= \frac{B - F}{A + G} \\ \tan(\omega - \Omega) &= \frac{-B - F}{A - G} \\ a^2(1 + \cos^2 i) &= A^2 + B^2 + F^2 + G^2 \\ a^2 \cos^2 i &= AG - BF \end{aligned} \tag{A.4}$$

Due to the not-invertibility of the tan() and $\cos^2()$, multiples solutions for ω, Ω and i results in the same Thiele Innes constants $[A, B, F, G]$ implying that only using astrometric observations does not provide enough information to determine the exact set of parameters $[T, P, e, a, \omega, \Omega, i]$ from the observed binary system. However, the Thile Innes representation is still one of the most utilized frameworks to represent the astrometric behavior of a binary system due to the linear relationship between the constants $[A, B, F, G]$ and the variables $\bar{x}(t)$ and $\bar{y}(t)$, allowing inference methods to exploit this properties and find solutions to the parametrical problem.

A.2. Radial Velocity Derivation

The radial vector z is definad as the perpendicular projection of the moving star B and the sky-plane. Figure 2.2 shows that the radial vector z and the inclination angle i, supported in the semi-major axis and displayed perpendicularly from it, forms a right triangle where the vector z is the cathetus opposite to the angle i and the hypotenuse s. On the other hand, the radial vector r and the vector s forms another right triangle, being in this time the hypotenuse r and s the cathetus opposite to the angle $\omega + \nu$. This geometric relationship leads to an expression of z dependent of r, presented in Equation A.5.

$$\sin(\omega + \nu) = \frac{s}{r} \quad \wedge \quad \sin i = \frac{z}{s} \implies z = r \sin(\omega + \nu) \sin i \tag{A.5}$$

After a simple vector derivation, the radial velocity $\dot{z}$ can be expressed in therms of $\dot{r}$ and $\dot{\nu}$ as shown in Equation A.6.

$$\dot{z} = \sin i(\dot{r} \sin(\omega + \nu) + r \cos(\omega + \nu)\dot{\nu}) \tag{A.6}$$

The term $\dot{r}$ is determined in A.7 by direct derivation of the closed expression for r described in Equation 2.3.

$$\dot{r} = r \frac{e \sin \nu}{1 + e \cos \nu} \dot{\nu} \tag{A.7}$$

The term $\dot{\nu}$ can be calculated by using the Kepler's Second Law and integrating over the

period P of the systems as shown in Equation A.8 . The area of an ellipse is calculated as $A = \pi a^2\sqrt{1-e^2}$.

$$\begin{aligned} \frac{dA}{dt} = \frac{r^2\dot{\nu}}{2} \quad &\xrightarrow{\text{Integration over P}} \quad \frac{A}{P} = \frac{r^2\dot{\nu}}{2} \\ \dot{\nu} &= \frac{2\pi a^2\sqrt{1-e^2}}{r^2 P} \end{aligned} \tag{A.8}$$

Finally, a closed expression for $\dot{z}$ is available by using Equations A.7 and A.8 in Equation A.6.

$$\dot{z} = \frac{2\pi a \sin i}{P\sqrt{1-e^2}}(\cos(\omega+\nu) + e\cos\omega) \tag{A.9}$$

This characterization of radial velocity can be directly translated to a referential frame where the central point is the centre of masses, by using the fact that the orbital paths presented in the Figure 2.1 are elliptical (with semi-major axis a_1 and a_2 for both starts). Finally it is possible to calculate the radial velocity of each star as presented in Equation 2.5.

Appendix B

Differential Entropy

B.1. Entropy of a Multivariate Gaussian variable

Let $X \in \mathbb{R}^m$ be a random vector with Gaussian distribution $\sim \mathcal{N}(\mu, \alpha)$ being $\mu \in \mathbb{R}^m$ his mean vector and $\alpha \in \mathbb{R}^{m \times m}$ his covariance matrix, symmetric by definition. The explicit formula for p_X is defined in Equation B.1.

$$p_X(x) = \frac{1}{(\sqrt{2\pi})^m \sqrt{\det \alpha}} e^{-\frac{1}{2}(x-\mu)^T \alpha^{-1}(x-\mu)} \tag{B.1}$$

Then the differential entropy for X can be expressed using the definition in Equation 2.13.

$$\begin{aligned} H(X) &= -\int_{\mathbb{R}}^{m} \frac{1}{(\sqrt{2\pi})^m \sqrt{\det \alpha}} e^{-\frac{1}{2}(x-\mu)^T \alpha^{-1}(x-\mu)} \ln \left[\frac{1}{(\sqrt{2\pi})^m \sqrt{\det \alpha}} e^{-\frac{1}{2}(x-\mu)^T \alpha^{-1}(x-\mu)} \right] dx \\ &= -\int_{\mathbb{R}}^{m} p_X(x) \log \left[\frac{1}{(\sqrt{2\pi})^m \sqrt{\det \alpha}} e^{-\frac{1}{2}(x-\mu)^T \alpha^{-1}(x-\mu)} \right] dx \\ &= -\int_{\mathbb{R}}^{m} p_X(x) * \ln \left[\frac{1}{(\sqrt{2\pi})^m \sqrt{\det \alpha}} \right] dx - \int_{\mathbb{R}}^{m} p_X(x) \ln \left[e^{-\frac{1}{2}(x-\mu)^T \alpha^{-1}(x-\mu)} \right] dx \end{aligned} \tag{B.2}$$

The First term can be simplified as shown in Equation B.3.

$$\begin{aligned} & -\int_{\mathbb{R}}^{m} p_X(x) * \ln \left[\frac{1}{(\sqrt{2\pi})^m \sqrt{\det \alpha}} \right] dx \\ &= \ln \left[(\sqrt{2\pi})^m \sqrt{\det \alpha} \right] \int_{\mathbb{R}}^{m} p_X(x) dx \\ &= \frac{1}{2} \ln \left[(2\pi)^m \det \alpha \right] \end{aligned} \tag{B.3}$$

For the second term of Equation B.2 an alternative representation is proposed in Equation

B.4 by using the properties of the matrix transpose.

$$
\begin{aligned}
&- \int_{\mathbb{R}}^{m} p_X(x) \ln \left[e^{-\frac{1}{2}(x-\mu)^T \alpha^{-1}(x-\mu)} \right] dx \\
&= \frac{1}{2} \int_{\mathbb{R}}^{m} p_X(x)(x-\mu)^T \alpha^{-1}(x-\mu) dx \\
&= \frac{1}{2} \int_{\mathbb{R}}^{m} p_X(x) \left[\left(\alpha^{-1}\right)^T (x-\mu) \right]^T (x-\mu) dx
\end{aligned}
\tag{B.4}
$$

In order to simply Equation B.4, the concept of trace of a matrix is introduced. Let $tr(A) = \sum_i^d A_{i,i}$ be the trace of the squared matrix $A \in \mathbb{R}^{d \times d}$, then following three properties hold $A \in \mathbb{R}^{d \times d}$:

- (1) $tr(A^T B) = tr(AB^T)$
- (2) $atr(A) = tr(aA) \forall a \in \mathbb{R}$
- (3) $\int tr(A(x))dx = tr(\int A(x)dx)$

By using the properties (1) and (2) the previous Equation B.4 can be rewritten as follows.

$$
\begin{aligned}
&\frac{1}{2} \int_{\mathbb{R}}^{m} p_X(x) \left[\left(\alpha^{-1}\right)^T (x-\mu) \right]^T (x-\mu) dx \\
&= \frac{1}{2} \int_{\mathbb{R}}^{m} p_X(x) tr \left(\left[\alpha^{-1}(x-\mu) \right]^T (x-\mu) \right) dx \\
&= \frac{1}{2} \int_{\mathbb{R}}^{m} p_X(x) tr \left(\alpha^{-1}(x-\mu)(x-\mu)^T \right) dx \\
&= \frac{1}{2} \int_{\mathbb{R}}^{m} tr \left(\alpha^{-1} p_X(x)(x-\mu)(x-\mu)^T \right) dx
\end{aligned}
\tag{B.5}
$$

Finally, by using the property (3) from the trace of a matrix a closed form for the integral can be computed.

$$
\begin{aligned}
&\frac{1}{2} \int_{\mathbb{R}}^{m} tr \left(\alpha^{-1} p_X(x)(x-\mu)(x-\mu)^T \right) dx \\
&= \frac{1}{2} tr \left(\alpha^{-1} \int_{\mathbb{R}}^{m} p_X(x)(x-\mu)(x-\mu)^T dx \right) \\
&= \frac{1}{2} tr \left(\alpha^{-1} \alpha \right) \\
&= \frac{1}{2} tr \left(I_m \right) \\
&= \frac{m}{2}
\end{aligned}
\tag{B.6}
$$

Then adding the Equations B.3 and B.6 a closed form for the entropy of the Gaussian

variable X is achieved.

$$\begin{aligned} H(X) &= \frac{1}{2}\ln\left[(2\pi)^m \det\alpha\right] + \frac{m}{2} \\ &= \frac{1}{2}\ln\left[(2\pi)^m \det\alpha\right] + \frac{m}{2}\ln(e) \\ &= \frac{1}{2}\ln\left[(2\pi)^m \det\alpha\right] + \frac{1}{2}\ln(e^m) \\ &= \frac{1}{2}\ln\left[(2\pi e)^m \det\alpha\right] \quad \square \end{aligned} \tag{B.7}$$

B.2. Joint Entropy of independent variables

Let $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$ be random variables with probability distribution p_X and p_Z, lets consider that X and Z are independent, i.e. $p_{X,Z}(x,z) = p_X(x)p_Z(z) \quad \forall x \in \mathcal{X} \wedge z \in \mathcal{Z}$, then the joint distribution can be calculated as follows.

$$\begin{aligned} H(X,Z) &= -\int_{\mathcal{X}}\int_{\mathcal{Z}} p_{X,Z}(x,z)\ln p_{X,Z}(x,z)dzdx \\ &= -\int_{\mathcal{X}}\int_{\mathcal{Z}} p_X(x)p_Z(z)\ln p_X(x)p_Z(z)dzdx \\ &= -\int_{\mathcal{X}}\int_{\mathcal{Z}} p_X(x)p_Z(z)\ln p_X(x)dzdx - \int_{\mathcal{X}}\int_{\mathcal{Z}} p_Z(z)p_X(x)p_Z(z)\ln p_Z(z)dzdx \\ &= -\int_{\mathcal{Z}} p_Z(z)dz\int_{\mathcal{X}} p_x(x)\ln p_X(x)dx - \int_{\mathcal{X}} p_X(x)dx\int_{\mathcal{Z}} p_Z(z)\ln p_Z(z)dz \\ &= -\int_{\mathcal{X}} p_X(x)dx - \int_{\mathcal{Z}} p_Z\ln p_Z(z)dz \\ &= H(X) + H(Z) \quad \square \end{aligned} \tag{B.8}$$

This expression can be trivially extended to the case of two random variables conditionally independent, this is because for any value c of the conditional variable C the probability distributions $p_{X|C}$ and $p_{Z|C}$ take a particular shape, then the Equation B.8 can be applied.

B.3. Kozachenko-Leonenko Differential Entropy Estimation

The Kozachenko-Leonenko method is a data-driven estimator to calculate the differential entropy from a collection of particles. The method is mainly based on interpreting the differential entropy as an expectation of X over the function $-\log p_X(x)$ as shown in Equation B.9.

$$\hat{H}(x) = \frac{-1}{N}\sum_{i=1}^{N}\log\hat{p_X}(x_i) \tag{B.9}$$

In order to directly use the Equation B.9, a estimation of p_X based on k nearest neighbour is proposed. Lets define $p_k(\epsilon)$ the probability distribution of the kth neighbor of x_i being at distance ϵ, this implies that $k-1$ elements must be in a distance minor to ϵ and consequently $N-K-1$ particles are further tan ϵ. Lets also define $p_i(\epsilon) = \int_{||x_i-\xi||<\epsilon/2} p_X(\xi)d\xi$ the probability mass contained in a sphere of radius $\epsilon/2$ centred on x_i, then the definition of p_k can be made explicit through Equation B.10 by using the trinomial formula.

$$p_k(\epsilon) = k\left(\frac{N-1}{k}\right)\frac{dp_i(\epsilon)}{d\epsilon}p_i^{k-1}(1-p_i)^{N-k-1} \tag{B.10}$$

Then the following expectation can be computed, being ψ the digamma function.

$$\begin{aligned}\mathbb{E}_{\epsilon\sim p_k}[\log p_i] &= \int_0^\infty p_k(\epsilon)\log p_i(\epsilon)d\epsilon \\ &= k\left(\frac{N-1}{k}\right)\int_0^1 p^{k-1}(1-p)^{N-k-1}\log p dp \\ &= \psi(k)-\psi(N)\end{aligned} \tag{B.11}$$

Aiming to simplify the objective distribution and make a analogy to an multivariate histogram approach, p_X is assumed constant in a ϵ-ball centred in each x_i, then this approximation leads to the relation between the defined probability p_i and p_X through Equation B.12 being c_m the unitary ball in a m dimensional space.

$$p_i(\epsilon) \approx c_m\epsilon^m p_X(x_i) \tag{B.12}$$

By isolating $p_X(x_i)$ and applying the log function a estimation of $\log p_X(x_i)$ can be made as shown in Equation B.13

$$\log p_X x_i \approx \psi(k)-\psi(N)-m\mathbb{E}[\log\epsilon]-\log c_d \tag{B.13}$$

Finally by taking a Monte Carlo estimation of the expectation of Equation **??** a closed form for the Kozachenko-Leonenko differential entropy estimator is obtained.

$$\hat{H}(X) = -\psi(k)+\psi(N)+\log(c_d)+\frac{m}{N}\sum_{i=1}^{N}\epsilon(i) \quad \square \tag{B.14}$$

$\epsilon(i)$ corresponds to twice the distance between the kth neighbor and the particle x_i.

Appendix C

Tables

Table C.1: Variance of the inferred orbital parameters by various methods for 6 cases of real data.

System	T	P	e	a	i	Ω	ω
HDS866	-	-	-	-	-	-	-
TK-HDS866	6.23	46.79		0.022	12.48	2.88	91.34
U-1-HDS866	11.45	27.45	0.242	0.106	4.18	2.77	44.31
U-HDS866	3.45	26.42	0.195	0.078	4.15	2.06	33.81
HIP109634	-	-	-	-	-	-	-
TK-HIP109634	-	10.09	0.19	0.015	3.51	11.69	33.26
U-1-HIP109634	0.018	8.16	0.241	0.137	2.97	12.19	1.62
U-HIP109634	0.001	0.26	0.007	0.112	1.31	2.34	1.52
HIP9497	0.58	0.89	0.051	0.005	4.3	13.2	16.9
TK-HIP9497	0.25	0.14	0.011	0.002	2.59	4.14	5.76
U-1-HIP9497	0.55	0.32	0.02	0.001	2.26	6.77	17.59
U-HIP9497	0.18	0.14	0.01	0.001	1.71	1.59	5
HIP10885	-	-	-	-	-	-	-
TK-HIP10885	0.53	0.53	0.019	0.003	2.32	2.4	3.41
U-1-HIP10885	0.53	0.53	0.034	0.029	2.03	30.96	37.35
U-HIP10885	0.49	0.49	0.026	0.003	1.77	29	33.1
HIP14524	-	-	-	-	-	-	-
TK-HIP14524	0.027	0.38	0.008	0.003	0.8	1.39	0.56
U-1-HIP14524	0.11	1.07	0.026	0.006	2.23	2.94	0.58
U-HIP14524	0.002	0.02	0.006	0.003	0.75	0.26	0.17
HIP32767	-	-	-	-	-	-	-
TK-HIP32767	1.121	3.33	0.029	0.009	1.13	1.71	11.45
U-1-HIP32767	1.071	2.92	0.026	0.008	1.74	1.72	11.42
U-HIP32767	0.958	2.8	0.028	0.008	1.65	1.44	10.41

Appendix D

ORB6 Study

The Sixth Catalog of Orbits of Visual Binary Stars(ORB6[1]) is a public compilation of binary stars orbits, where the inferred orbital parameters of multiple system are reported. The inferred parameters in the ORB6 catalog are classified by its precision, the grading goes from grade 5 (for indeterminate orbits) up to grade 1 (definitive orbits).

With the intention of make a real data analysis, as made in Section 5.2, a selection of candidates must be performed. To achieve this, an indicative value of the phase with respect to the periastron is proposed as a quick method to find systems that will pass near its periastron at certain moment of observation. The n_{check} indicator is defined as follows:

$$n_{\text{check}}(t_{\text{check}}) = \frac{t_{\text{check}} - T}{P} \tag{D.1}$$

being t_{check} the scheduled time to make the observation.

By using n_{check} for $t_{\text{check}} = 2021.824$ (i.e. October 27th of 2021) on each system in the ORB6 catalog, a selection of candidate systems is computed using the following additional filters:

- $|n_{\text{check}}| < 10\%$
- DEC < 20
- Grade 2, 3 or 4

A subset of the resultant query was selected to be observed. Such chosen systems are presented in tables 5.7, 5.8 and C.1.

The entire code to compute the queries on the ORB6 database can be accessed in this link[2] by using the Google Colaboratory platform.

[1] http://www.astro.gsu.edu/wds/orb6/orb6text.html
[2] https://colab.research.google.com/drive/1ExM3qO2LbxyCKMnR7Wtx$_C mb24pgaNLb?usp = sharing$