

Robust Orbit Determination and Classification: A Learning Theoretic Approach

Srinagesh Sharma* and James W. Cutler*

Orbit determination involves estimation of a non-linear mapping from feature vectors associated with the position of the spacecraft to its orbital parameters. The de facto standard in orbit determination in real-world scenarios for spacecraft has been linearized estimators such as the extended Kalman filter. Such an estimator, while very accurate and convergent over its linear region, is hard to generalize over arbitrary gravitational potentials and diverse sets of measurements. It is also challenging to perform exact mathematical characterizations of the Kalman filter performance over such general systems. Here we present a new approach to orbit determination as a learning problem involving distribution regression and, also, for the multiple-spacecraft scenario, a transfer learning system for classification of feature vectors associated with spacecraft, and provide some associated analysis of such systems.

I. Introduction

In recent years, there has been an increased interest in space systems and exploration. This has led to a rise in the number and scope of space missions, both near Earth and in deep space. Furthermore, since the development of the CubeSat standard [1], these small spacecraft have become increasingly crucial in science missions, technology demonstration, and in access to space. In fact, recently, such space missions have forayed into technology demonstrations for deep space technologies [2]. These spacecraft have a significantly smaller form factor and they are used for essentially high-risk, low-cost missions with short development periods [1]. A consequence is that the number of such spacecraft inserted into orbit per launch has increased. These spacecraft are highly power constrained systems, whose operations, including those of communications, are power constrained.

*University of Michigan, Ann Arbor

The research described in this publication was carried out under contract with the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. © 2015 All rights reserved. Government sponsorship acknowledged.

A second line of development, which follows directly from the requirement of increased capacity in communication systems, is a ground station network. Ground station networks allow increased downlink access times and, depending on orbit parameters, higher data rates at specific locations [3]. These have become increasingly popular due to the development of software-defined radios, which allow for autonomous ground station networks. A crucial requirement, both for spacecraft deployment and for communication, is identification and versatile, robust orbit determination.

A Keplerian orbit of a satellite around a spherical body can be described by 6 parameters, denoted generically as a 6-element vector γ and modeled here as a corresponding 6-element vector of random variables Γ . Examples of 6-parameter specifications of such orbits include the traditional Keplerian elements, equinoctial elements, simple r, v vectors in the required reference frame and, in the Hamiltonian setting, Poincaré elements. The position of the satellite at time t for a spherical body with uniform density is given by the general form of Kepler's equation [4]

$$\sqrt{\kappa}(t - t_P) = a \left[\chi - \sqrt{a} \sin \left(\frac{\chi}{\sqrt{a}} \right) \right] + \frac{\|r_P\| \|v_P\|}{\sqrt{\kappa}} a \left[1 - \cos \left(\frac{\chi}{\sqrt{a}} \right) \right] + r_P \sqrt{a} \sin \left(\frac{\chi}{\sqrt{a}} \right)$$

where $\kappa = \mathbb{GM}$, \mathbb{G} is the universal gravitational constant, \mathbb{M} is the mass of the Earth, $\dot{\chi} = \frac{\sqrt{\kappa}}{r}$, and r_P, v_P are the position and velocity vectors at the periapsis.

Furthermore, in case of the Earth, due to its oblateness and to orbital perturbations, the parameters associated with Keplerian orbits vary over time (although the variation is negligible over short periods such as a ground station pass). These perturbations are modeled by the Laplace equations for spherical harmonics [4]. In addition, there are also perturbations of the orbit due to air drag and solar radiation pressure. While there are infinite series summations that provide closed-form solutions for the position of the spacecraft as a function of time and orbit parameters for certain values of eccentricity in the case of purely Keplerian orbits (orbits that are described purely by Kepler's equations), the perturbations are usually computed using numerical integration techniques [4].

Orbit determination (OD) is a non-linear filtering problem that estimates the orbital parameters γ from observations of the highly non-linear trajectory resulting from Newton's laws and perturbations. Practical implementation of such estimation involves a two-step process of first obtaining an initial estimate through a few observations and then an asymptotic series of differential corrections to obtain orbital parameters [5, 6]. The standard technique for precision orbit determination is an extended Kalman filter (EKF) [4, 7]. The EKF is a suboptimal approximation of the Kalman filter for non-linear systems, which has been shown to converge asymptotically when the initial state of the system is in the linear region [8]. A large number of orbit determination implementations also use EKF techniques [5]. EKF techniques perform orbit determination of spacecraft over specific scenarios, and the mathematical computation of the gradient matrices governing OD systems is not easy to generalize (unless approximated with finite differencing, which is an approximation) over arbitrary gravitational potentials and

with a varying number of celestial bodies influencing the dynamics of the orbit. Secondly, the performance of EKF techniques is influenced by the extent of non-linearity associated with the system [9]. This makes modification of the estimation system for diverse sensor systems a difficult proposition, and it is not necessarily guaranteed to converge. As a consequence, it is also difficult to provide generalized bounds of convergence for the orbit estimation system as a whole. A second popular method of filtering involves particle filters and similar Bayesian techniques [10]. However, even in such cases, a Gaussian approximation regarding evolution of the filter is made.

Recent techniques in orbit determination involve estimation over observations on very short arcs [11–13]. These have become significantly more popular due to the requirements of estimating orbits of asteroids in celestial mechanics. For spacecraft in low Earth orbits, high-precision orbit determination can be performed using GPS techniques [14]; however, GPS techniques cannot be generalized for deep-space satellites.

This article serves as a proof of concept for a new and very general approach to the problem of orbit determination. We show that when the characteristics of the spacecraft and the general characteristics of the orbital parameters are known and observable, there exists a continuous mapping from a subset of the space of probability distributions of feature vectors associated with spacecraft position, such as RF transmissions observed at the ground station network, to the orbital parameters of the spacecraft. While it is computationally prohibitive to perform direct characterization of these probability distributions, it is possible to generate samples associated with these distributions, and perform two-stage sampled regression and classification to obtain the orbital parameters. The framework presented can also be applied for tracking deep-space missions with ground station networks as long as the orbit perturbations can be modeled accurately.

Specifically, the contributions of this article are threefold:

1. We present a novel modeling and method using techniques recently developed in machine learning to perform classification of RF transmissions and orbit determination of the spacecraft.
2. We present conditions under which such a system can be applied.
3. We present an experimental result of orbit determination of the GRIFEX¹ spacecraft using such a system.

For the purposes of intuition, we consider orbit determination of the GRIFEX spacecraft as

¹ The GEO-CAPE ROIC In-Flight Performance Experiment (GRIFEX) is a CubeSat developed by the Michigan Exploration Laboratory (MXL) supporting the proposed Geostationary Coastal and Air Pollution Events (GEO-CAPE) mission concept. GRIFEX performs engineering assessment of a JPL-developed all-digital in-pixel high-frame-rate Read-Out Integrated Circuit (ROIC).

a running example throughout this article. The GRIFEX CubeSat [15] was launched into orbit on January 31, 2015. It contains a 9.6 kbps radio which beacons approximately once every 10 seconds at 437.485 MHz with a random initial delay. The RF transmissions undergo Doppler shifts, whose parameters are dependent on the evolution of the spacecraft orbit in relation to a ground station. We have performed orbit determination of the GRIFEX CubeSat based on RF transmissions received over two passes at the FXB Ground Station, University of Michigan, Ann Arbor.

II. Background

A. Mathematical Preliminaries

In this section we introduce our mathematical notation, and some theoretical concepts in probability and set theory that will be useful for analysis of the system.

Notation Upper-case symbols are used to denote random variables or sets. Scalars or vectors are differentiated by context. Lower-case symbols are used to denote either instances of the random variable or known/observed constants. Script letters such as \mathcal{X}, \mathcal{Y} etc. are used to denote a measurable space with $\mathfrak{F}_{\mathcal{X}}, \mathfrak{F}_{\mathcal{Y}}$ etc., denoting the corresponding Borel σ -algebra and P_X, P_Y denoting the probability distributions, respectively. The symbol \hat{y} is used to denote an estimate of the corresponding true value y . Subscript T refers to the test system.

Set distance We shall define the distance between two measurable sets A and B to be $d_S(A, B) = \mathfrak{m}(A\Delta B)$, where \mathfrak{m} is the Lebesgue measure and $A\Delta B = A \setminus B \cup B \setminus A$.

Probability kernels and the Prokhorov metric The space of probability distributions on a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$ with Borel σ -algebra $\mathfrak{F}_{\mathcal{X}}$, is a metric space $(\mathcal{B}_{\mathcal{X}}, d_P)$ (weak topology), where d_P , the Prokhorov metric, is defined as

$$d_P(P_1, P_2) = \inf\{a : P_1(A) \leq P_2(A^a) + a \quad \forall A \in \mathfrak{F}_{\mathcal{X}} \text{ and vice versa}\} \quad (1)$$

where $A^a = \{s \in \mathcal{X} : d(s, A) < a\}$ and $d(s, A) = \inf\{d(s, s_A), s_A \in A\}$ For details see Chapter 2 of [16]. Also, for any two random variables X, Y defined on \mathcal{X} and \mathcal{Y} , the conditional probability $P(X|Y = y)$ is associated with a function $\mu : \mathcal{Y} \rightarrow \mathcal{B}_{\mathcal{X}}$ (see lemma 1.37 and Chapter 5 in [17]). We shall call this function a probability kernel function.

B. Recent Techniques from Machine Learning

Now we present a brief review of two machine learning techniques recently proposed in literature that we have applied to our problem: distribution regression [18] and transfer learning [19].

Distribution regression Distribution regression [18] is a technique to estimate the mappings from the space of distributions on a compact space \mathcal{X} , $\mathcal{B}_{\mathcal{X}}$ to \mathcal{S} , a compact subset of \mathbb{R} when the only access to the distribution is through samples drawn from it [18]. Say we are given N training samples $\{\{x_j^{(i)}\}_{j=1}^{n_i}, s_i\}_{i=1}^N$, drawn from a meta-distribution over $\mathcal{B}_{\mathcal{X}} \times \mathcal{S}$. The objective is to estimate a function $r : \Phi(\mathcal{B}_{\mathcal{X}}) \rightarrow \mathcal{S}$, where $\Phi(\mathcal{B}_{\mathcal{X}})$ is the image of $\mathcal{B}_{\mathcal{X}}$ under the mean embedding, such that

$$r^* = \arg \min_{r \in \mathcal{H}} \mathbb{E}[(r(\Phi(P_X)) - S)^2] + \xi_2 \|r\|_{\mathcal{H}}^2, \quad (2)$$

where ξ_2 is a slack variable, and the mean embedding is defined as $\Phi(P_X) = \int_{\mathcal{X}} k(\cdot, x) dP_X$ for a kernel k . The resulting optimization and predictor is

$$\hat{r}_{\xi_2} = \arg \min_{r \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N (r(\Phi(\hat{P}_X^{(i)})) - s_i)^2 + \xi_2 \|r\|_{\mathcal{H}}^2 \Rightarrow \hat{r}_{\xi_2}(\Phi(\hat{P}_X)) = s^\top (K + N\xi_2 I)^{-1} k_r, \quad (3)$$

where K is the kernel matrix and k_r is the kernel vector for the given distribution [18]. It has been shown that when the embedding is Hölder continuous with exponent h , this estimator is consistent and upper bounds for convergence can be obtained [18]. Lastly, when space \mathcal{X} is a Polish space, universal kernels that are dense in the space of continuous functions over compact metric spaces [20, 21] can be used. Distribution regression will be used to perform non-parametric estimation of orbits.

Transfer learning Consider a Polish space \mathcal{X} , a binary classification space $\mathcal{Y} = \{-1, +1\}$, and a loss function $L : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}_+$. Let the space of distributions over $\mathcal{X} \times \mathcal{Y}$ be $\mathcal{B}_{\mathcal{X} \times \mathcal{Y}}$. Assume that there exists a distribution λ over $\mathcal{B}_{\mathcal{X} \times \mathcal{Y}}$ such that $\lambda = \lambda_{Y|X} \lambda_X$ and $\lambda_{Y|X} = \delta_D$ almost everywhere, where δ_D is the Dirac-Delta function. For such functions, $\exists h : \mathcal{B}_{\mathcal{X}} \times \mathcal{X} \rightarrow \mathcal{Y}$ such that $y = h(P_X, x)$. Let \mathcal{H}_k be the reproducing kernel Hilbert space (RKHS) associated with kernel $k_1 : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. For $\Phi(\mathcal{B}_{\mathcal{X}})$, the set of mean embeddings associated with $\mathcal{B}_{\mathcal{X}}$, let \mathcal{H}_{k_P} be the RKHS associated with the kernel $k_P : \Phi(\mathcal{B}_{\mathcal{X}}) \times \Phi(\mathcal{B}_{\mathcal{X}}) \rightarrow \mathbb{R}$. We seek an estimate $h_{\mathcal{H}}$ of h such that the following criterion is satisfied:

$$h_{\mathcal{H}} = \arg \min_{h \in \mathcal{H}_{\bar{k}}} \mathbb{E}_{P_{XY} \sim \lambda, (X, Y) \sim P_{XY}} [L(h(\Phi(P_X), X), Y)], \quad (4)$$

where $\bar{k} : (\Phi(\mathcal{B}_{\mathcal{X}}) \times \mathcal{X}) \times (\Phi(\mathcal{B}_{\mathcal{X}}) \times \mathcal{X}) \rightarrow \mathbb{R}$ and $\bar{k}((P_X, X), (P_{X'}, X')) = k_P(\Phi(P_X), \Phi(P_{X'})) k_1(X, X')$. The resulting regularized problem has the formulation [19]

$$\hat{h}_{\xi_1} = \arg \min_{h \in \mathcal{H}_{\bar{k}}} \frac{1}{N} \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} L(h(\Phi(\hat{P}_X^{(i)}), X_{ij}), Y_{ij}) + \xi_1 \|h\|_{\mathcal{H}_{\bar{k}}}^2. \quad (5)$$

We shall use transfer learning to perform classification of feature vectors of multiple spacecraft.

III. Problem Setup

Consider a spacecraft with orbit parameters Γ drawn from the space \mathcal{J} according to a probability distribution P_{Γ} , which is known a priori and has compact support. There are

n_G synchronized ground stations that behave as sensors. The spacecraft produces vectors $F = [F_1 \ F_2 \ \dots \ F_{n_G} \ T_S]^T = [\tilde{F} \ T_S]^T$ over a compact set \mathcal{F} such that the connection between \tilde{F} and T_S is governed by a parameter z specific to the spacecraft and the system function associated with the dynamic system U described by

$$\begin{aligned}\dot{\tilde{\gamma}} &= g_0(\tilde{\gamma}) \\ \tilde{f} &= h_0(\tilde{\gamma}, z) \\ \tilde{\gamma}(0) &= \gamma,\end{aligned}\tag{6}$$

and $T_S \sim P_{T_S}(z)$, a continuous distribution over \mathcal{T} and a characteristic of the spacecraft which is known. The ground stations produce measurements of these vectors F to produce observations (feature vectors) $X = [\tilde{X}_1 \ \tilde{X}_2 \ \dots \ \tilde{X}_{n_G} \ T]^T$, where $\tilde{X}_k, k = 1, 2, \dots, n_G$ are the features extracted per transmission at ground station k , and T is a time stamp.

For example, in the GRIFEX orbit determination scenario, \tilde{F} represents the theoretical noiseless vector of frequencies at the ground station at time T_S , and \tilde{X} represents the Doppler shifts observed at the ground station. In this example, U is the dynamic system that controls the Doppler shift (range rate), and z represents the parameters of the communication system which are essential to draw samples of X , including bandwidth, thermal bias in crystal oscillator frequencies (which causes center frequency drift), and the probability of transmission over $[0, T_{max}]$.

With this scenario, the orbit determination problem can be stated as follows. Given P_T, U, z over the time interval \mathcal{T} , for feature vectors $\{x_1, x_2, \dots, x_{n_T}\}$, which are samples of X , we would like to estimate γ , the orbit parameters.

This problem can be extended to two spacecraft labeled $\{-1, +1\}$, with orbit distributions $P_{\Gamma_1, \Gamma_{-1}}$ and feature vector characteristics z_1, z_{-1} and corresponding system U , where given this setup and feature vectors $\{x_1, x_2, \dots, x_{n_T}\}$, we would like to estimate the corresponding labels $\{y_1, y_2, \dots, y_{n_T}\}$ and orbit parameters γ_{-1}, γ_{+1} . The estimation of the labels $\{y_i\}_{i=1}^{n_T}$ is the classification problem. Here we assume that there exists a mapping from the distribution of P_X to the label y . We consider two spacecraft, instead of a general n_S -spacecraft scenario, as it remains an open problem to develop the mathematical tools required for consistent classification of n_S spacecraft in the marginal transfer learning setting.

Remark Here we assume, for simplicity, that the feature vectors $\{x_1, x_2, \dots, x_{n_T}\}$ are independent and identically distributed (i.i.d.). In most practical systems involving RF transmission data, this is not necessarily the case as they are generated from a random process, where data can at most be exchangeable and not necessarily i.i.d. and is a possible source of sub-optimality.

IV. Non-Parametric Orbit Estimation and Classification

We will now present an equivalent mathematical model of the system described above which will provide insight into the characteristics of the system and the consequences of observability.

A. Probability Models

Consider the orbit determination scenario as detailed in section III. Given the communication system parameters $Z = z$ and the distribution of feature vectors with time $P(T_S|z)$, Γ induces a probability distribution on F , as detailed later in the section. Samples of F generate samples of X at the ground station network. It is to be noted that T_S and T_G are not necessarily the same, especially when accounting for propagation delays through the channel. This results in the graphical model shown in Figure 1(a).

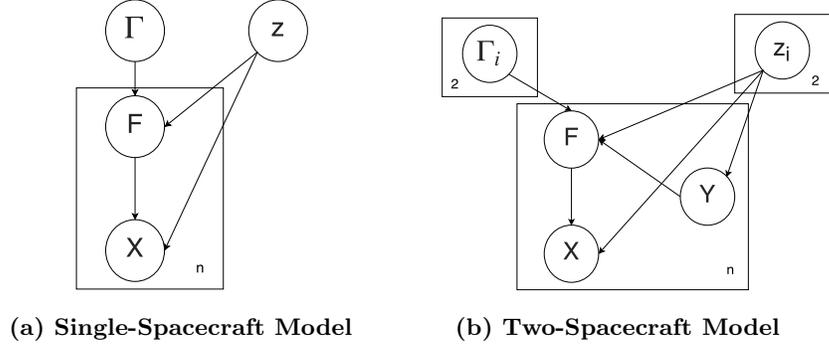


Figure 1. Graphical models.

The probability distribution can be split as

$$P(\Gamma, F, X|z) = P(\Gamma)P(F|\Gamma, z)P(X|F, z). \quad (7)$$

The conditional probability distribution $P(F|\Gamma, z)$ is shaped by a deterministic non-linear dynamic model operating on T_S . Consider the scenario where $\mathcal{T} = [0, T_{max}]$. For the moment, let us assume that $P(T_S|\gamma, z) = P(T_S|z)$. From the theories of astrodynamics [4], it is possible to describe the function that maps $[0, T_{max}]$ to $\tilde{\mathcal{F}}$, which is the space of frequencies observed at the ground stations, using the following dynamic model

$$\begin{aligned} \dot{\tilde{\gamma}} &= g_0(\tilde{\gamma}) \\ \tilde{f} &= h_0(\tilde{\gamma}, z) \\ \tilde{\gamma}(0) &= \gamma, \end{aligned} \quad (8)$$

where $\dot{\tilde{\gamma}}$ is the derivative of the orbital parameters with respect to time. We shall denote this

system as $U(\gamma)$. Such a dynamical system can be used to model perturbations associated with orbit parameters and can also be used to perform orbit propagation given a particular γ . For specific forms of these equations, see for example, Chapter 9 of [4].

The conditional distribution $P(F|\Gamma, z)$ of a system can be written as

$$P(\tilde{F} \in C, T_S \in B | \Gamma = \gamma, z) = P(T_S \in B \bigcap U(\gamma)^{-1}(C) | \Gamma = \gamma, z), \quad (9)$$

where C and B are sets in the corresponding σ -fields of \tilde{F} and T_S , and $U(\gamma)^{-1}$ is the pre-image of $U(\gamma)$.

B. Mathematical Analysis

Now, for the system defined from Equations (7) through (9), we make the following assumptions regarding existence and uniqueness :

- A I** U is observable and Lipschitz continuous over the support of P_Γ in \mathcal{T} .
- A II** The probability kernel function from $\mathcal{F} \rightarrow \mathcal{P}(\mathcal{X})$, for $\mathcal{P}(\mathcal{X}) \subseteq \mathcal{B}_\mathcal{X}$ is bijective and continuous.

We shall now discuss the origin and some physical consequences of these assumptions. Even though the assumption **A I** is a significantly strong assumption regarding observability, it is essential for the existence of a global estimator over the support of P_Γ . While we do not aim to prove this assumption holds for all low Earth orbits, we provide some evidence in such a direction. Since there exist linearized estimators, the corresponding non-linear system is also locally observable. In the setting of spacecraft, it also implies that the support of P_Γ cannot contain zero-eccentricity orbits, at least in this formulation using Keplerian elements, due to singularity. For such cases, higher-fidelity models or generalized orbital parameters (such as Poincaré elements), which guarantee observability and non-singularity, can be used. Note here however, that the learning algorithm will provide solutions which are optimal only specific to those elements and not necessarily in all of them due to the non-linear nature of the transformation between these elements. The assumption of Lipschitz continuity is a common assumption for existence and continuous dependence on initial parameters of the system in control theory literature, which hold for spacecraft systems since the equations governing perturbations of orbits due to the gravitational potential is a harmonic function [7]. This assumption mostly holds except in cases of the effect of eclipses on radiation pressure of large spacecraft (see section 14.3 of [7]), and even in such cases they can be modeled as a C^1 function with a very large Lipschitz constant [22]. Assumption **A II** is required for noise characteristics of the system. For narrowband communication systems, it has been shown that the correlation function $S(f)$ from which the feature vector X is obtained can be written as

$S(f) = Q(f_c + f_d) + \text{residual}$ [23,24], where f_c is the center frequency of the RF transmission and f_d is the Doppler, which satisfies the conditions of **A II**.

We now present a theorem that details the consequences of the assumptions on the framework:

Theorem IV.1 *If Equations (7) - (9) hold along with **A I** and **A II**, then there exists a continuous mapping $\lambda : R_{\mathcal{F}} \rightarrow \mathcal{J}$ for $R_{\mathcal{F}} \subseteq \mathcal{B}_{\mathcal{X}}$ on the topology $(\mathcal{B}_{\mathcal{X}}, d_P)$.*

Proof Sketch This follows from the observability and Lipschitz continuity properties of U . Due to Lipschitz continuity the corresponding probability distribution $P(F|\gamma, z)$ is continuous in $(\mathcal{B}_{\mathcal{F}}, d_P)$. For the function to exist, the probability kernel function $\mu(\gamma) = P(F|\gamma, z)$ has to be one-to-one. This holds as a result of observability of U and $P(T_S|z)$ being a continuous distribution over $[0, T_{max}]$. For detailed proof, see Subsection C of the Appendix.

This theorem can be extended to consider multiple passes, i.e., a set of time intervals in which the orbits may be observable. To do so, we will model the influence of the orbital parameters γ on P_{T_S} in a specific fashion. We will assume that the only effect γ has on T_S is

$$P(T_S|\gamma, z) := P(T_S|T_S \in \mathcal{T}(\gamma), z), \quad (10)$$

such that $P(T_S|\gamma, z)$ is continuous.

Generally the observed regions of the orbit are governed by the horizon of the ground stations. Let us denote the dynamic equation associated with the elevation as $V(\gamma)(t)$ over $[0, T_{max}]$. Then for a closed set O , $\mathcal{T}(\gamma)$ is defined as

$$\mathcal{T}(\gamma) = \{t \in [0, T_{max}] : V(\gamma)(t) \in O\}. \quad (11)$$

For γ to be estimable, we will need stronger assumptions on observability of U .

B I U is observable in $\mathcal{T}(\gamma)$, for every $\gamma \in \mathcal{J}$, and U is Lipschitz continuous over the support of P_{Γ} in \mathcal{T} .

Note that this is a significantly stronger assumption than observability over \mathcal{T} . It is, however, a weaker assumption compared to observability at every $t \in \mathcal{T}$. A simple example for this assumption in low Earth orbits occurs when estimating orbits with Doppler. In cases when the right ascension of the ascending nodes differ by ϵ with all other parameters being identical including the ground stations, there will exist regions where the doppler shifts are identical for significant chunks of the two passes. They will, however, be observable as the point of zero Doppler will differ in time.

Corollary IV.2 *If Equations (7) - (9), (10) - (11) hold for V continuous in γ and $\mathfrak{m}(\{t : \frac{\partial V}{\partial t} = 0\}) = 0$ along with assumptions **B I** and **A II**, then there exists a continuous mapping $\lambda : R_{\mathcal{F}} \rightarrow \mathcal{J}$ for $R_{\mathcal{F}} \subseteq \mathcal{B}_{\mathcal{X}}$ on the topology $(\mathcal{B}_{\mathcal{X}}, d_P)$.*

Proof Sketch The proof follows from the fact that if $\mathcal{T}(\gamma)$ is continuous in γ and $P(T_S|z)$ is absolutely continuous with respect to the Lebesgue measure then the mapping λ exists and is continuous. $\mathcal{T}(\gamma)$ is continuous in γ when V is continuous in γ and an inverse exists over all neighborhoods except on certain points of measure zero (which provides the condition $\mathfrak{m}(\{t : \frac{\partial V}{\partial t} = 0\}) = 0$). For detailed proof, see Subsection D of the Appendix.

The derivative conditions on V hold as a consequence of Newton's laws except in cases of geostationary orbits. However, for such orbits $\mathcal{T}(\gamma_{geo}) = [0, T_{max}]$, and therefore Theorem IV.1 holds.

This can now be extended to include the classification problem, with the label Y associated with each feature vector X . For two spacecraft with $Z_1 = z_1, Z_{-1} = z_{-1}$, the conditional distribution can be modeled as depicted in Figure 1(b):

$$P(\Gamma_1, \Gamma_{-1}, Y, F, X|z_1, z_{-1}) = P(\Gamma_1, \Gamma_{-1})P(Y|z_1, z_{-1})P(F|Y, \Gamma_1, \Gamma_{-1}, z_1, z_{-1})P(X|F, z_1, z_{-1}). \quad (12)$$

Consider the random variables $\Gamma_1, \Gamma_{-1}, X, Y$. There exists a probability kernel function $\nu : \mathcal{J} \times \mathcal{J} \rightarrow \mathcal{B}_{\mathcal{X} \times \mathcal{Y}}$. A probability distribution $P_{\Gamma_1, \Gamma_{-1}}$ induces the probability distribution $P_{\Gamma_1, \Gamma_{-1}} \circ \nu^{-1}$ on $\mathcal{B}_{\mathcal{X} \times \mathcal{Y}}$ such that $P_{\Gamma_1, \Gamma_{-1}} \circ \nu^{-1}(\xi) = P_{\Gamma_1, \Gamma_{-1}}\{(\gamma_1, \gamma_{-1}) \in \mathcal{J} \times \mathcal{J}; \nu(\gamma_1, \gamma_{-1}) \in \xi\}$ for $\xi \in \mathfrak{F}(\mathcal{B}_{\mathcal{X} \times \mathcal{Y}})$; this is a consequence of the disintegration theorem. Also, i.i.d. draws from $P_{\Gamma_1, \Gamma_{-1}}$ will result in i.i.d draws from $P_{\Gamma_1, \Gamma_{-1}} \circ \nu^{-1}$. Therefore, $P_{\Gamma_1, \Gamma_{-1}}$ induces a probability distribution on $\mathcal{B}_{\mathcal{X} \times \mathcal{Y}}$.

C. Observability and Probability Distribution Sampling

Theorem IV.1 and Corollary IV.2 imply that if the dynamic system controlling the evolution of the observed random variables is observable and Lipschitz continuous, then there exists a mapping from the probability distribution of the observations to the orbital parameters. In the GRIFEX setting, this implies that if the observability and continuity conditions in **B I** and **A II** are satisfied, then there exists a continuous mapping from the probability distributions of the RF transmissions observed to the orbital parameters. This continuous mapping also exists even when observations are spread across multiple ground stations in $[0, T_{max}]$.

While direct characterization of these probability distributions is prohibitive, it is possible to draw samples from these distributions and estimate the kernel embedding associated with the observed probability distributions and perform regression over the kernel embeddings as detailed in [18]. Secondly, since the mapping exists and is continuous over a complete topological space over a compact set, the convergence properties of continuous functions

for the $\mathcal{P}(b, c)$ class hold, and therefore it is possible to estimate the function mapping the probability distribution of the RF transmissions to orbital parameters γ arbitrarily close to its estimate in the corresponding RKHS. When Universal kernels are used, the RKHS will be dense in the space of continuous functions from \mathcal{B}_X to \mathcal{J} .

For the mathematical model described in IV-A and IV-B, we are given n_T observations $\{x_j^T\}_{j=1}^{n_T}$ and we would like to estimate the labels $\{y_j^T\}_{j=1}^{n_T}$ and the orbital parameters γ . Generally, for probabilistic graphical models of this nature, Bayesian inference is applied, either through direct computation of the posterior probabilities, the EM algorithm, or through techniques such as Markov chain Monte Carlo inference methods. However, in this case, there are two challenges. First, there exists no closed-form expression of the conditional distributions. Second, while it is possible to sample from the prior conditional distribution, direct parametric description of the posterior conditional is non-trivial and is also time variant. However, it is possible to perform forward sampling, and from Theorem IV.1 there exists a continuous function from the space of probability distributions of X onto Γ . Based on this, we propose the following. Perform sampling to generate $\{\{x_j^{(i)}\}_{j=1}^{n_i}, \gamma_i\}_{i=1}^N$ or $\{\{x_j^{(i)}, y_j^{(i)}\}_{j=1}^{n_i}, \gamma_{1,i}, \gamma_{-1,i}\}_{i=1}^N$. The resulting system satisfies the underlying assumptions of transfer learning [19] and distribution regression [18]. Perform transfer learning to obtain an estimate of $\{y_j^T\}_{j=1}^{n_T}$ and distribution regression to obtain estimates of γ .

The sampling can be performed from the probabilistic graphical models described in the mathematical formulation. It is crucial, however, that the probability distributions, especially with regards to bias, are samples as expected to be seen in the experimental data. This is not an impossibility as known sources can generally be characterized, especially artificial spacecraft. It is also crucial that the orbit dynamical models are accurate in the limit for consistency requirements to hold. For further details, see Section V.

D. Application of Distribution Regression

We use a vector extension to distribution regression. This technique is straightforward and follows the same technique as detailed in [18] and [25]. The resulting optimization problem is

$$r^* = \arg \min_{r \in \mathcal{H}} \mathbb{E}[\|r(\Phi(P_X)) - \Gamma\|_{\mathcal{J}}^2] + \xi_2 \|r\|_{\mathcal{H}}^2. \quad (13)$$

The kernel operator chosen on the embedding is in the form of a diagonal matrix, in which the six kernel bandwidths are tuned separately. For multiple spacecraft, two scenarios arise. When the probability distributions of Γ_1 and Γ_{-1} are independent of each other, two different functions $r_j^*, j = \{+1, -1\}$ can be estimated. However, in scenarios of constellation deployments, the kernel matrix can be used to take into account correlations associated with the orbit parameters. This scenario occurs frequently in CubeSat deployments, where multiple CubeSats are deployed with time offsets and small changes in deployment velocities. Note that by construction, the distribution induced by γ lies in the $\mathcal{P}(b, c)$ class (see Appendix)

and therefore convergence characteristics apply. The convergence properties of such systems to the fundamental limits of estimation are based on the number of training data sets and points available [18].

E. Application of Transfer Learning

The basic premise for the use of transfer learning is the following: When the constant characteristics of RF transmissions are identical, or when SNR values are low such that decodability is prohibitive, classification of a data point cannot be performed purely on the characteristics of that data point alone. The entire distribution induced by the orbit parameters over time period \mathcal{T} has to be taken into consideration while performing classification of the data point. We have training data such that

$$P^{(i)}(X, Y | \Gamma_1 = \gamma_{1,i}, \Gamma_{-1} = \gamma_{-1,i}) \sim \lambda \quad (14a)$$

$$(x_j^{(i)}, y_j^{(i)}) \sim P^{(i)}(X, Y | \Gamma_1 = \gamma_{1,i}, \Gamma_{-1} = \gamma_{-1,i}). \quad (14b)$$

Transfer learning can now be used to perform classification. It ought to be noted that this technique can also be used to classify and separate known local noise sources as well.

V. Test Results and Discussion

In this section we present our results from four test case scenarios. In one test we performed orbit determination on experimental data based on the recent launch of the GRIFEX satellite [15]. The other three test cases were created from synthetic orbital data. The first synthetic test case performs Doppler-based orbit determination using distribution regression over two ground stations. The second synthetic test case performs orbit determination on azimuth, elevation and range estimates over one ground station. The third synthetic test applies transfer learning over pooled classification for a scenario of RF transmissions from two satellites to one ground station. Before presenting our results from these tests, we briefly elaborate on the processes and practical aspects of the generation of training data required for the application of transfer learning and distribution regression.

Keplerian orbits were assumed for all four test scenarios, and the orbit parameters were taken to be the six traditional Keplerian elements, $\Gamma = (A, e, \omega, I, \Omega, M)$. In all of our test cases, the priors on these orbit parameters P_Γ were generally sufficiently broad to allow for significant initial uncertainties in the values of the orbit parameters. For example, in the GRIFEX orbit estimation scenario, the priors allowed the initial position vector to vary by around 1800 km. During post-launch orbit determination, the pre-launch orbit estimates can be used with distributions based off accuracy of orbit insertion, energy imparted during launch, and a uniform variation of perigee times based on launch window estimates. The orbit parameters can also be drawn from mixture distributions, allowing different distributions for different satellite systems.

The communication systems of artificial satellites are generally well characterized. The systems parameters Z can take into considerations of bias due to thermal effects of oscillators; if measurements exist for rate of de-tumbling, those can also be included for generation of training data. Furthermore, effects of solar radiation pressure, air drag, and, in the case of an attitude-determined spacecraft with pre-set orientation on orbit insertion, the corresponding RF attenuation due to antenna positioning, can also be taken into account. The systems parameters also take into consideration the statistics of the data generated, the communication systems parameters required for the generation of received signal samples, and the length of the interval of RF transmissions.

The data are generated as per the graphical model described in Figure 1(a) or 1(b). The training data are synthetically generated and are of the form $\{\{x_{ij}, y_{ij}\}_{j=1}^{n_i}, \gamma_{1,i}, \gamma_{-1,i}\}_{i=1}^N$. For a sample of the given prior P_Γ , the orbit parameters generate time intervals of observability on sight. Based on draws of time of transmission and resulting center frequency, the Doppler parameters are computed from the draw of γ_1, γ_{-1} using numerical techniques. Satellite propagators can be used for estimation of Doppler shifts at various frequencies.

A. Orbit Determination of the GRIFEX Spacecraft

Orbit estimation for the GRIFEX CubeSat was performed over two passes (200-minute interval) starting at 10 AM UTC, 2 Feb 2015, at the FXB ground station in Ann Arbor. The following priors were assumed:

$$\begin{aligned} A &\sim R_e + U(540, 570) \text{ km}, & e &\sim U(0.015, 0.02), \\ \omega &\sim U(330^\circ, 345^\circ), & I &\sim U(96^\circ, 101^\circ), \\ \Omega &\sim U(35^\circ, 45^\circ), & M &= U(255^\circ, 275^\circ) \end{aligned}$$

The orbit state being estimated was a TLE generated at 10 AM UTC, 2 Feb 2015. The GRIFEX spacecraft transmits at approximately a 10-second interval a beacon of GMSK modulation at 9.6 kbps with a random initial delay and therefore is stationary. A software-defined radio system was used to filter a 50-kHz band of raw in-phase and quadrature samples at 512 kHz at a baseband frequency of 437.479 MHz and then downsampled by a factor of 4. The techniques proposed in cognitive radio literature [26] were used to extract the Doppler shifts and the bandwidths. A further correction of 220.95 Hz for bias was performed by filtering and estimation of frequency at the point of maximum slope. The raw IQ data were corrupted by RF transmissions from the ground with significantly larger bandwidths and power. Simple thresholding was used to extract Doppler feature vectors for the GRIFEX spacecraft.

The training set consisted of 3000 orbits drawn from priors as described above. The training data was assumed to be obtained from points with a one second resolution, due to computational feasibility. In true test systems, the number of transmissions per orbit would be much

higher due to higher sampling rates. The number of orbits for training and number of transmissions per orbit were limited by the computational capability of a Xeon 4 core processing system. Five fold cross validation was performed. The resulting true and estimated Keplerian elements are shown in Table 1. We assume that the estimates of GRIFEX orbital parameters as given by Joint Space Operations Center (JSpOC) to be the ground truth.

Table 1. Keplerian elements of the GRIFEX spacecraft.

	A(km)	e	I(deg)	Ω (deg)	ω (deg)	M(deg)
True (JSpOC Estimates)	554.2014	0.0167974	99.1306	42.2363	334.6695	272.368
Estimated	551.0128	0.0161	99.3659	42.2911	339.9771	266.8452

B. Synthetic Test Case 1: Doppler-Based Orbit Determination

We performed orbit determination with synthetic data from two ground stations (Ann Arbor and Chicago) estimating Doppler of one spacecraft over two orbits. The prior P_{Γ} was chosen as follows:

$$\begin{aligned}
 A &\sim R_e + U(900, 1000) \text{ km}, & e &\sim U(0.07, 0.08), \\
 \omega &\sim U(16\pi/9, 33\pi/18), & I &\sim U(2\pi/9, 5\pi/18), \\
 \Omega &\sim 4\pi/3 + U(0, \pi/18), & M &= \pi/4,
 \end{aligned}$$

where U is the uniform distribution. With the stated prior, the maximum and average variation in the radius vector were 1086 km and 470 km respectively. The probability of transmission over the pass interval was chosen to be uniform. No sources of noise were added; 3000 orbits were used for training, with 100 orbits for testing orbit estimation. Probability of transmission was chosen to be 0.08. Coarse fivefold cross-validation was performed. The average error in the radius vector was 44 km. The root mean square radial, along-track and cross-track position errors were found to be 2.7541 km, 54.54 km, and 27.4290 km, respectively. Figure 2 shows the orbital parameters of the true values and estimates of the test orbits.

C. Synthetic Test Case 2: Orbit Determination Based on Range and Direction of Arrival (DOA)

Orbit determination was performed with direction of arrival and range estimates obtained from synthetic data from one ground station over two passes (4-hour interval). The prior on the orbit parameters was the same as that described in Synthetic Test Case 1, as were the testing and training setups. The average error in the radius vector was 26 km and the root mean square radial, along-track and cross-track position error estimates were found to be 2.17 km, 34.8 km, and 8.04 km.

True and estimated orbital parameters

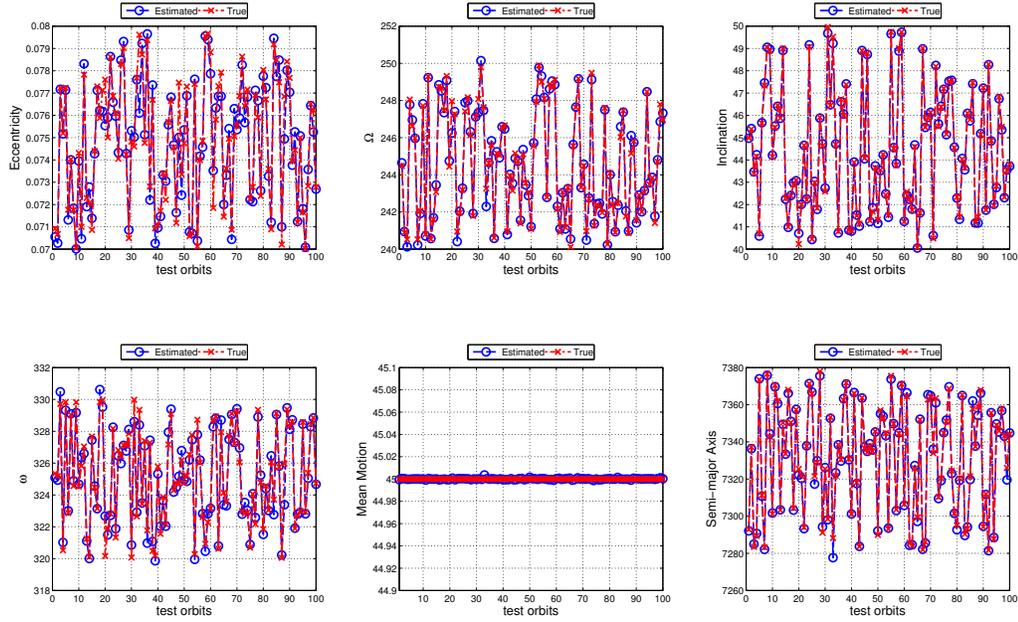


Figure 2. True and estimated orbital parameters for Doppler-based estimation.

D. Comparison of Results from the GRIFEX Spacecraft and Synthetic Test Cases 1, 2

Table 2 gives a comparison of along-track, cross-track and radial position errors for the three orbit determination test cases. Keplerian elements were used in all three estimates. For the GRIFEX scenario, a synthetic test based on 100 orbits drawn from the GRIFEX priors was also performed to obtain a better estimate of the errors seen for orbits with low eccentricity. It is to be noted that, while the total position error for the GRIFEX spacecraft was 29.74 km

Table 2. Comparison of GRIFEX test error estimates (position errors) and synthetic test RMS errors.

Error	GRIFEX Test: Real Data	GRIFEX Test: Synthetic Data	Synthetic Test: Case 1 (Doppler)	Synthetic Test: Case 2 (DOA)
Radial (km)	7.55	9.83	2.7541	2.17
Along-track (km)	17.23	61.03	54.54	34.8
Cross-track (km)	23.04	160.5	27.429	8.04
Total Error (km)	29.74	172	61.1109	35.7825

based on real observations, the RMS error for draws of orbital parameters with the stated priors and with training data of 3000 orbits with probability of transmission of 0.08 is 172 km, due to the use of Keplerian elements at low eccentricity and due to low observability of certain orbits in the Doppler domain.

Another point to note is that the along-track and cross-track errors are higher in the Doppler-based test cases. This is expected as this information is gained only from subtle variations in the Doppler shift during the pass, whereas the radial position information can be gained from the maximum Doppler shift variation.

E. Synthetic Test Case 3: Classification

This test case considers two satellites with identical modulation and data parameters of BPSK and 10 kbps data rates with center frequencies differing by 10 kHz. Propagation of the satellite orbits is performed using standard SGP4 propagators for a single ground station (Ann Arbor) over a single pass. Synthetic datasets were generated with 40 data sets for training and 10 data sets for testing. The priors on orbit parameters were as follows:

$$\begin{aligned}
 A &= 644.93 \text{ km}, \\
 e &\sim \frac{1}{E_1} \mathbf{1}_{\{1 \times 10^{-4} \leq e \leq 1 \times 10^{-3}\}} (\mathcal{N}(4 \times 10^{-4}, 1 \times 10^{-8}) + \mathcal{N}(-4 \times 10^{-4}, 1 \times 10^{-8})) \\
 \omega &\sim \mathcal{N}(16\pi/9, (\pi/18)^2), & I &\sim \mathcal{N}(2\pi/9, (\pi/1800)^2), \\
 \Omega &\sim \mathcal{N}(4\pi/3, (\pi/36)^2), & M &\sim \mathcal{N}(\pi/4, (\pi/1800)^2),
 \end{aligned}$$

where E_1 is the normalization factor. Note that for Gaussian priors the support was restricted to 5 times the standard deviation. Training of both the pooled classification and transfer learning systems was performed with fivefold cross-validation. All of the kernels chosen were Gaussian. A projection of the training data with the corresponding support vectors onto three dimensions is shown in Figure 3. It can be seen in Table 3 that the transfer learning system

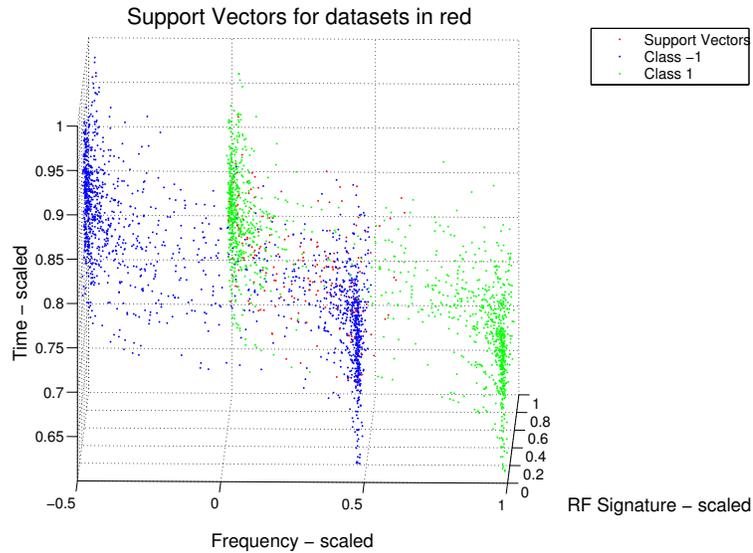


Figure 3. Training data for transfer learning system with support vectors marked.

produces a classification error of 2.8% whereas the pooled classification system produces an error of 6.83%.

Table 3. Classification error comparison.

Classification Method	% Error
Transfer Learning	2.8
Pooled Classification	6.83

VI. Conclusion

It is possible, when the orbital parameters of a spacecraft are observable, to perform orbit determination over orbital parameters defined over a compact space. The mapping will exist when observability conditions required for the particular orbit are satisfied, and using non-parametric estimation techniques, it is possible, theoretically, to perform orbit determination with errors arbitrarily small from the true estimates.

It can be seen from the test cases that it is possible to estimate orbital parameters even when they vary significantly. Linear region requirements are not necessary and initial estimates are not required as long as they lie in a compact space, the characteristics of the spacecraft are known and the parameters are observable. It is also possible to identify and classify transmissions of known spacecraft.

Acknowledgments

This work was done with the support of the Strategic University Research Partnership (SURP) Program, 2015, from the Jet Propulsion Laboratory, NASA. We also thank Jon Hamkins and Sam Dolinar for their support and encouragement in this effort.

References

- [1] H. Heidt, J. Puig-Suari, A. Moore, S. Nakasuka, and R. Twiggs, "Cubesat: A new generation of picosatellite for education and industry low-cost space experimentation," in *Proceedings of the 14th Annual/USU Conference on Small Satellites, Logan, USA, 2000*, 2000.
- [2] A. T. Klesh, J. D. Baker, J. Bellardo, J. Castillo-Rogez, J. Cutler, L. Halatek, E. G. Lightsey, N. Murphy, and C. Raymond, "Inspire: Interplanetary nanospacecraft pathfinder in relevant environment," in *Proceedings of the Annual Small Satellite Conference, Logan, USA*, 2013.
- [3] S. Spangelo, J. Cutler, A. Klesh, and D. Boone, "Models and tools to evaluate space communication network capacity," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 48, no. 3, pp. 2387–2404, 2012.

- [4] D. A. Vallado, *Fundamentals of astrodynamics and applications*, vol. 12. Springer, 2001.
- [5] J. R. Vetter, “Fifty years of orbit determination,” *Johns Hopkins APL technical digest*, vol. 27, no. 3, p. 239, 2007.
- [6] A. E. Roy, *Orbital motion*. CRC Press, 2004.
- [7] A. Milani and G. Gronchi, *Theory of orbit determination*. Cambridge University Press, 2010.
- [8] A. J. Krener, “The convergence of the extended kalman filter,” in *Directions in mathematical systems theory and optimization*, pp. 173–182, Springer, 2003.
- [9] B. F. La Scala, R. R. Bitmead, and M. R. James, “Conditions for stability of the extended kalman filter and their application to the frequency tracking problem,” *Mathematics of Control, Signals and Systems*, vol. 8, no. 1, pp. 1–26, 1995.
- [10] D.-J. Lee, *Nonlinear Bayesian filtering with applications to estimation and navigation*. PhD thesis, Texas A&M University, 2005.
- [11] A. Milani, G. F. Gronchi, M. d. Vitturi, and Z. Knežević, “Orbit determination with very short arcs. i admissible regions,” *Celestial Mechanics and Dynamical Astronomy*, vol. 90, no. 1-2, pp. 57–85, 2004.
- [12] L. Ansalone and F. Curti, “A genetic algorithm for initial orbit determination from a too short arc optical observation,” *Advances in Space Research*, vol. 52, no. 3, pp. 477–489, 2013.
- [13] D. A. Vallado and S. S. Carter, “Accurate orbit determination from short-arc dense observational data,” *The Journal of the astronautical sciences*, vol. 46, no. 2, pp. 195–213, 1998.
- [14] O. Montenbruck, T. Van Helleputte, R. Kroes, and E. Gill, “Reduced dynamic orbit determination using gps code and carrier measurements,” *Aerospace Science and Technology*, vol. 9, no. 3, pp. 261–271, 2005.
- [15] J. W. Cutler, C. Lacy, T. Rose, S.-h. Kang, D. Rider, and C. Norton, “An update on the grifex mission,” *Cubesat Developer’s Workshop*, 2015.
- [16] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, 2013.
- [17] O. Kallenberg, *Foundations of modern probability*. springer, 2002.
- [18] Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur, “Two-stage sampled learning theory on distributions,” in *AISTATS-Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, vol. 38, pp. 948–957, 2015.
- [19] G. Blanchard, G. Lee, and C. Scott, “Generalizing from several related classification tasks to a new unlabeled sample,” in *Advances in neural information processing systems*, pp. 2178–2186, 2011.
- [20] A. Christmann and I. Steinwart, “Universal kernels on non-standard input spaces,” in *Advances in neural information processing systems*, pp. 406–414, 2010.
- [21] I. Steinwart, “On the influence of the kernel on the consistency of support vector machines,” *The Journal of Machine Learning Research*, vol. 2, pp. 67–93, 2002.
- [22] J. T. Hwang, D. Y. Lee, J. W. Cutler, and J. R. Martins, “Large-scale multidisciplinary optimization of a small satellites design and operation,” *Journal of Spacecraft and Rockets*, vol. 51, no. 5, pp. 1648–1663, 2014.
- [23] W. A. Gardner, “Spectral correlation of modulated signals: Part i— analog modulation,” *Communications, IEEE Transactions on*, vol. 35, no. 6, pp. 584–594, 1987.

- [24] W. A. Gardner, W. Brown, and C.-K. Chen, "Spectral correlation of modulated signals: Part ii—digital modulation," *Communications, IEEE Transactions on*, vol. 35, no. 6, pp. 595–601, 1987.
- [25] A. Caponnetto and E. De Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 331–368, 2007.
- [26] M. Bkassiny, S. K. Jayaweera, Y. Li, and K. A. Avery, "Blind cyclostationary feature detection based spectrum sensing for autonomous self-learning cognitive radios," in *Communications (ICC), 2012 IEEE International Conference on*, pp. 1507–1511, IEEE, 2012.
- [27] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American mathematical society*, pp. 337–404, 1950.

Appendix

A. Kernels and Reproducing Kernel Hilbert Spaces

Let \mathcal{X} be a set. Consider the mapping $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. This mapping is said to be symmetric if for any two elements of \mathcal{X} , x_1, x_2 , we have $k(x_1, x_2) = k(x_2, x_1)$. It is said to be positive definite if, for any n elements, $x_1, x_2, \dots, x_n \in \mathcal{X}$, the matrix M with the i, j element $M(i, j) = k(x_i, x_j)$ (known as the Gram matrix) is positive definite. When the mapping k is symmetric and positive definite, it is called a kernel. For every such kernel, there exists a Hilbert space \mathcal{H} and a mapping $\phi : \mathcal{X} \rightarrow \mathcal{H}$ such that $\langle \phi(x_1), \phi(x_2) \rangle = k(x_1, x_2)$ [27] and the set $\mathcal{X}_f = \{\sum_i \alpha_i k(\cdot, x_i) | \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}\}$ is dense in \mathcal{H} . Note that the set \mathcal{X} can also be the set of probability distributions on a compact metric space [20]. For every function $f \in \mathcal{H}$, the following property holds: $\langle f, k(\cdot, x) \rangle = f(x)$ for $x \in \mathcal{X}$. Such a Hilbert space is known as a reproducing kernel Hilbert space (RKHS). RKHSs provide elegant ways of embedding probability distributions into Hilbert spaces.

B. Definition of $\mathcal{P}(b, c)$ Class

For the embedding $\Phi(P_X)$ and estimates γ with the joint probability distribution $\rho(\Phi(P_X), \gamma)$ if for estimate $f_{\mathcal{H}} \in \mathcal{H}$, given $1 < b \leq +\infty, 1 \leq c \leq 2$, the following conditions:

1. γ is square integrable with respect to $\rho(\Phi(P_X), \gamma)$ and $\exists \Sigma > 0, G > 0$ such that

$$\int_{\mathbb{R}} \left(e^{\frac{\|\gamma - f_{\mathcal{H}}(\Phi(P_X))\|_{\Gamma}}{G}} - \frac{\|\gamma - f_{\mathcal{H}}(\Phi(P_X))\|_{\Gamma}}{G} - 1 \right) dP(\gamma | \Phi(P_X)) \leq \frac{\Sigma^2}{2G^2}$$

2. For $T = \int_{\Phi(\mathcal{B}_{\mathcal{X}})} k_P(\cdot, \Phi(P_X)) \delta_{\Phi(P_X)}(\cdot) dP(\Phi(P_X))$, $\exists \tilde{f} \in \mathcal{H}$ such that $f_{\mathcal{H}} = T^{(c-1)/2} \tilde{f}$ with $\|\tilde{f}\|_{\mathcal{H}}^2 \leq R$, for a fixed constant R .
3. T is such that the residuals $\{t_i\}_{i=1}^I$ of T in \mathcal{H} satisfy

$$\alpha \leq i^b t_i \leq \beta$$

when $b < \infty$ and $I = +\infty$ and when $b = +\infty, I \leq \beta < \infty$.

are satisfied then we say that $\rho(\Phi(P_X), \gamma) \in \mathcal{P}(b, c)$.

C. Proof of Theorem IV.1

We need to prove continuity and uniqueness of the distribution for every $\gamma \in \mathcal{J}$. We prove continuity by using Lipschitz continuity of the function modulating a continuous probability distribution and uniqueness by the observability property.

If U is observable and Lipschitz continuous then there exists a continuous bijective mapping from \mathcal{J} onto $U(\mathcal{J})$. For such a bijective mapping U , let the probability kernel function associated with $P(F \in \cdot | \Gamma = \gamma, z) = \mu(\gamma)$. Then, by Lipschitz continuity, for a given δ we have ϵ such that: $\gamma_1, \gamma_2 \in \mathcal{J}$ with $\|\gamma_1 - \gamma_2\|_{\mathcal{J}} < \epsilon$ implies $\|U(\gamma_1)(t) - U(\gamma_2)(t)\| < \delta, \forall t \in [0, T_{max}]$.

Consider sets $C \in \mathfrak{F}_{\mathcal{F}}, B \in \mathfrak{F}_{\mathcal{T}_S}$. For any $t \in U(\gamma_1)^{-1}(C)$ we can find a point $f \in U(\gamma_2)(C^\delta)$ such that $\|U(\gamma_1)(t) - f\| < \delta$. This implies for every set $D = C \otimes B$, we have $\tilde{D} = C^\delta \otimes B \subseteq D^\delta$ such that $\mu(\gamma_1)(D) = \mu(\gamma_2)(\tilde{D})$ and $\mu(\gamma_1)(D) \leq \mu(\gamma_2)(D^\delta) + \delta$. Similarly $\mu(\gamma_2)(D) \leq \mu(\gamma_1)(D^\delta) + \delta$. This implies that

$$\begin{aligned} d_P(\mu(\gamma_1), \mu(\gamma_2)) &= \inf\{\alpha : \mu(\gamma_1)(D) \leq \mu(\gamma_2)(D^\alpha) + \alpha \text{ and } \mu(\gamma_2) \leq \mu(\gamma_1)(D^\alpha) + \alpha, \forall D \in \mathfrak{F}_{\mathcal{F}}\} \\ &< \delta \end{aligned} \tag{15}$$

Also, as U is observable over \mathcal{J} , if $\gamma_1 \neq \gamma_2$, then there exists $D \in \mathfrak{F}_{\mathcal{F}}$ such that $\mu(\gamma_1)(D) \neq \mu(\gamma_2)(D)$. Therefore, there exists a continuous function from $\tilde{R}_{\mathcal{F}}$ to \mathcal{J} for $\tilde{R}_{\mathcal{F}} \subseteq \mathcal{B}_{\mathcal{F}}$. Since the kernel function from \mathcal{F} to $\mathcal{B}_{\mathcal{X}}$ is bijective and continuous, the hypothesis holds. ■

D. Proof of Corollary IV.2

$P(T_S|z)$ is absolutely continuous with respect to the Lebesgue measure. If $\mathcal{T}(\gamma)$ is continuous in γ and can be expressed as a union of intervals, we have for a given δ , $\exists \epsilon_1, \epsilon_2$ such that $\|\gamma_1 - \gamma_2\| < \epsilon_1 \Rightarrow d_S(\mathcal{T}(\gamma_1), \mathcal{T}(\gamma_2)) < \epsilon_2 \Rightarrow d_P(\mu(\gamma_1), \mu(\gamma_2)) < \delta$.

Let $\mathcal{T}^\epsilon(\gamma) = \{t \in [0, T_{max}] : V(\gamma)(t) \in O^\epsilon\}$ and $\mathcal{T}^{-\epsilon}(\gamma) = \{t \in [0, T_{max}] : V(\gamma)(t) \in O^{-\epsilon}\}$ where $O^\epsilon = \{o \in \mathbb{R}^n | d(o, O) < \epsilon\}$ and $O^{-\epsilon} = ((O^c)^\epsilon)^c$. Since V is continuous in γ , we have $\mathcal{T}^{-2\epsilon}(\gamma_1) \subseteq \mathcal{T}^{-\epsilon}(\gamma_2) \subseteq \mathcal{T}(\gamma_1) \subseteq \mathcal{T}^\epsilon(\gamma_2) \subseteq \mathcal{T}^{2\epsilon}(\gamma_1)$. For a given γ , by definition, $\mathcal{T}(\gamma) \subseteq \mathcal{T}^\epsilon(\gamma)$. If the two sets $\mathcal{T}(\gamma)$ and $\mathcal{T}^\epsilon(\gamma)$ are equal for all γ , then the continuity condition is satisfied trivially and therefore we only need to consider the case when $\mathcal{T}(\gamma) \subset \mathcal{T}^\epsilon(\gamma)$. For a given ϵ consider $\mathcal{T}^{-\epsilon}(\gamma)\Delta\mathcal{T}^\epsilon(\gamma)$ i.e., the pre-image of $O^{-\epsilon}\Delta O^\epsilon = \bigcup_{p \in Bd(O)} B_\epsilon(p)$. We have, from the definition of the Lebesgue measure, $d_S(\mathcal{T}^{-\epsilon}(\gamma), \mathcal{T}^\epsilon(\gamma)) \leq \mathbf{m}(R_{O,\epsilon}) + \mathbf{m}(\bigcup C_i \cap V(\gamma)^{-1}(O^{-\epsilon}\Delta O^\epsilon))$, where $R_{O,\epsilon}$ is the set where the derivative of $V(\gamma)$ with respect to t is zero in $O^{-\epsilon}\Delta O^\epsilon$ and C_i is a countable covering of the set $[0, T_{max}] \setminus R_{O,\epsilon}$ over the neighborhoods of points where the implicit function theorem can be applied.

Therefore, we have that when $\mathbf{m}(R_{O,\epsilon}) = 0$, for any $\epsilon_2 > 0$, \exists an ϵ_3 such that $d_S(O^{-\alpha}, O^\alpha) < \epsilon_3$ implies $d_S(\mathcal{T}^{-\alpha}(\gamma), \mathcal{T}^\alpha(\gamma)) < \epsilon_2$, which implies continuity of $\mathcal{T}(\gamma)$ with respect to γ . ■