LEARNING TRANSFER OPERATORS BY KERNEL DENSITY ESTIMATION

Sudam Surasinghe [1,*], Jeremie Fish [1,†] and Erik M. Bollt [1,‡]
[1] *Clarkson Center for Complex Systems Science*
Department of Electrical and Computer Engineering
Clarkson University, 8 Clarkson Ave
Potsdam, New York 13699, USA
[*] *surasinc@clarkson.edu*
[†] *fishja@clarkson.edu*
[‡] *ebollt@clarkson.edu*

Inference of transfer operators from data is often formulated as a classical problem that hinges on the Ulam method. The usual description, which we will call the Ulam-Galerkin method, is in terms of projection onto basis functions that are characteristic functions supported over a fine grid of rectangles. In these terms, the usual Ulam-Galerkin approach can be understood as density estimation by the histogram method. Here we show that the problem can be recast in statistical density estimation formalism. This recasting of the classical problem, is a perspective that allows for an explicit and rigorous analysis of bias and variance, and therefore toward a discussion of the mean square error.

Keywords: Transfer Operators; Frobenius-Perron operator; probability density estimation; Ulam-Galerkin method;Kernel Density Estimation.

1. Introduction

Transfer operators play a vital role in the global analysis of dynamical systems. Abundant data from dynamical systems make those operators popular in data-driven analysis methods in complex systems. Hence, numerically estimating the transfer operators through data is key to success in global analysis. Frobenius-Perron operator is one such popular operator used for global analysis of dynamical systems, and the Ulam method [Ulam, 1960] is the most popular method to estimate it. However, as we point out here, there are tremendous opportunities to recast this problem as one of density estimation as would be stated in the statistics literature, specifically noting that there is a well matured analysis of variance and bias that allows for discussion of mean squared error. This langauge has been overlooked in the dynamical systems community. Therefore, here we introduce the probability density estimation viewpoint to estimating the Frobenius-Perron operator, and with it the rich analysis already developed in other mathematical communities and methods, notably kernel estimation which as it turns out is provably more efficient than the histogram methods used in the standard Ulam method.

A Frobenius-Perron operator evolves the density of ensembles of initial conditions of a dynamical system forward in time. This statement can also be re-interpreted in a Bayesian framework. In these terms, we have essentially a problem of density estimation, for the conditional probability density that is generally described as the Frobenius-Perron operator. The classical Ulam method is essentially a histogram method for estimation of this conditional density function by simple nonparametric means. Many, including one of the authors of this work, have described the approach as a projection onto basis functions as characteristic

functions, and in these terms, we described it as Ulam-Galerkin's method [Bollt & Santitissadeekorn, 2013], which covers many of the analyses of convergence since the original conjecture [Chiu *et al.*, 1992; Boyarsky & Góra, 1997; Dellnitz *et al.*, 2001; Guder *et al.*, 1997].

However, it is generally understood that histograms, while easy to describe, are a primitive variant amongst the approaches available to the problem of nonparametric density estimation. It has been said by Tukey [Tukey, 1961; Tukey & Tukey, 1981], that appearance of the traditional histogram is blocky, and difficult to balance smoothing, bandwidth, bias, and variance. Even in two dimensions, "blocky" variability of sampling, and details such as the simply choosing an appropriate orientation of the grid, become problematic. There are, however, more suitable methods that are reviewed here, especially the kernel density estimation which has many nice smoothing, analytic, and convergence properties. Additionally, density estimation is shifted from a question of density in space to expectation of points. Note also that the k-nearest neighbor (kNN) methods also have many of these advantages, but kernel methods allow for better tuning of smoothing parameters and good convergence statistics.

It is argued in [Dehnad, 1987], that the argument for kernel density estimation instead of a simple histogram method becomes stronger in more than one dimension, due to difficulties not only in histogram box size (bandwidth) but now also, in orientation and origin location that generally lead to a block appearance that becomes more difficult to interpret the joint and conditional probabilities. Tukey asserted [Tukey & Tukey, 1981], "...it is difficult to do well with bins in as few as two dimensions. Clearly, bins are for the birds!"

In this article, we show how to use the probability density estimation methods to approximate the Frobenius-Perron operator. Hence, KDE based method is a better approximation for the transfer operator. This analysis will demonstrate improvements in the accuracy of KDE based estimation compared to the current high popular Ulam-Galerkin's method. Understanding the Frobenius-Perron operator as the integration against a conditional distribution kernel is key in this demonstration. We will show that connection in section (3) then we will discuss the density estimation theory to demonstrate the theoretical advantages of KDE over the histogram-based method. Finally, we will numerically demonstrate the better accuracy of the KDE based method by using the chaotic logistic map example, which we show matches well to the rigorous analysis reviewed here for variance and bias.

2. Frobenius-Perron and the Classical Ulam-Galerkin Method for Estimation

First we briefly review a standard discussion of Frobenius-Perron operators for deterministic and then random maps, and flows are covered in as much as the maps discussed can be taken as derived from the flow by a Poincare' or stroboscopic mapping. Assuming a map,

$$\begin{aligned} f : X &\to X, \\ x &\mapsto f(x), \end{aligned} \tag{1}$$

the forward $orbit(x) = \{x, f(x), f^2(x), ...\}$ from an initial condition x is a subject of dynamical systems. However, if we consider an ensemble of many initial conditions, that are distributed by $\rho \in L^1(X)$, and we assume f is a nonsingular transformation and measurable relative to $(X, \mathcal{B}, \mu)$ on a Borel sigma-algebra of measurable sets $\mathcal{B} \subset X$, then follows the Frobenius-Perron operator that describes the orbit of ensembles, following our notation from [Bollt & Santitissadeekorn, 2013], and comparable to [Lasota & Yorke, 1982]. The linear map, $\mathcal{P}_f : L^1(X) \to L^1(X)$, follows the discrete continuity equation,

$$\int_{f(B)} \rho_{n+1} d\mu = \int_B \rho_n d\mu, \text{ for any } B \in \mathcal{B}. \tag{2}$$

For differentiable maps, this simplifies ,

$$P_f[\rho](x) = \sum_{y:x=f(y)} \frac{\rho(y)}{|Df(y)|}, \tag{3}$$

where if $f(y)$ is a single-variate function, then $|Df(y)| = |f'(y)|$ is the absolute value of the derivative, or it is the determinant of the Jacobain (matrix) derivative if multi-variate. The following equivalent form is

relevant for our purposes here,

$$\mathcal{P}_f[\rho](x) = \int_X \delta(x - f(y))\rho(y)dy, \tag{4}$$

in terms of the delta function. Also, we have specialized to Lebesgue measure on X from this point forward.

The Ulam-Galerkin method is a way to estimate the action of the Frobenius-Perron operator, given a (fine) finite topological cover of X by (usually rectangles, or boxes, or triangles, or other simple spatial elements) $\mathcal{B} = \{B_i\}_{i=1}^K$, $K > 0$. The estimator,

$$P_{i,j} = \frac{m(B_i \cap f^{-1}(B_j))}{m(B_i)}, \tag{5}$$

is stated in terms of Lebesgue measure, $m(B) = \int_B dx$. In fact, a simple estimate of this $K \times K$ matrix P follows if a large collection of input, output pairs are available, (x_n, x_{n+1}), as examples of $x \in B_i \cap f^{-1}(B_j)$ perhaps derived from a long orbit that samples the space. Note that f^{-1} denotes the pre-image of f which may well not be one-one.

$$P_{i,j} \sim \frac{\#x_n \in B_i, x_{n+1} \in B_j}{\#x_n \in B_i}. \tag{6}$$

Notice that the "$\cap$" notation for intersection of sets is coincident with the "," notation for "and" which denotes both events occur. This is useful for reinterpretation by a Bayesian discussion in the next section. Under the above construction, it can easily be seen that P is a stochastic matrix, which therefore has a leading eigenvalue of 1 and, if simple, a dominant eigenvector which describes the steady state of the corresponding Markov chain.

The original Ulam-conjecture [Ulam, 1960] described that, in a limit of a refining partition $\{B_i\}$, the dominant eigenvector of the discrete state Markov chain converges to invariant density of the original dynamical system. This conjecture was first proved by Li [Li, 1976] under hypothesis of bounded variation of one-dimensional maps, providing weak convergence.

An estimate such as Eqs. (5)-(6) has previously been called an Ulam-Galerkin estimate [Bollt & Santitissadeekorn, 2013; Ma & Bollt, 2013], a description which we made to intentionally separate the concept of the limit of long time iteration, as one does when considering ergodic averages, from short time considerations, such as that Eqs. (5)-(6) can simply be taken as an estimate of the action of the map on ensemble densities, between two time frames, or perhaps to be iterated a few times. The phrase Galerkin is stated in terms of projection of the action of the operator onto a basis of characteristic functions $\{\xi_{B_i}(x)\}$, supported over the grid elements, $\{B_i\}$,

$$\xi_{B_i}(x) = \begin{cases} 1, \text{if } x \in B_i \\ 0, \text{ otherwise} \end{cases}. \tag{7}$$

Then the Ulam-Galerkin estimate formally describes a projection, $R : L^2(X) \to \Delta_K$, for a finite linear subspace $\Delta_K \subset L^2(X)$, that is spanned by the collection of characteristic functions over the grid elements. Notice, for this description, this is in terms of $L^2(X)$, in order that an inner product structure makes sense, and then

$$P_{i,j} = \frac{(\xi(B_i), \xi(f^{-1}(B_j))}{\|\xi_{B_i}\|} = \frac{\int_X \xi(B_i)(x)\xi(f^{-1}(B_j)(x)dx}{\int_X \xi_{B_i}(x)^2 dx}. \tag{8}$$

If considering this finite rank transition for finite time discussion, then we worry only about the estimation of transitions by finite estimation by the basis functions as discussed in, [Bollt & Santitissadeekorn, 2013; Bollt *et al.*, 2002]. Infinite time questions are clearly more nuanced which is why the Ulam-conjecture remained a conjecture for almost twenty years. Our Bayesian discussion will likewise avoid the same.

In the more general case of a random dynamical system, Eq. (1) is recast,

$$x_{n+1} = f(x_n) + s_n, \tag{9}$$

which describes a deterministic part f together with a stochastic "kick" s which we assume is identically independently distributed by $x \sim \nu$. Consequently, the kernel integral form of the Frobenius-Perron transfer operator becomes,

$$\mathcal{P}_f[\rho](x) = \int_X \nu(x - f(y))\rho(y)dy, \tag{10}$$

which we see is closely related to the zero-noise case of Eq. (4) where the kernel in that case is a delta-function. For discussion in the next section, we will specialize further to the truncated normal distribution, $s \sim t\mathcal{N}(0, \sigma)$ to maintain perturbations within the bounded domain, unit square by avoiding unbounded tails.

$$\nu(x; \mu, \sigma, a, b) = \frac{1}{\sigma} \frac{\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}, \text{ where, } \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \Phi(z) = \frac{1 + erf(\frac{z}{\sqrt{2}})}{2}, \tag{11}$$

and we choose, $a = 0, b = 1, x - \mu = f(y)$.

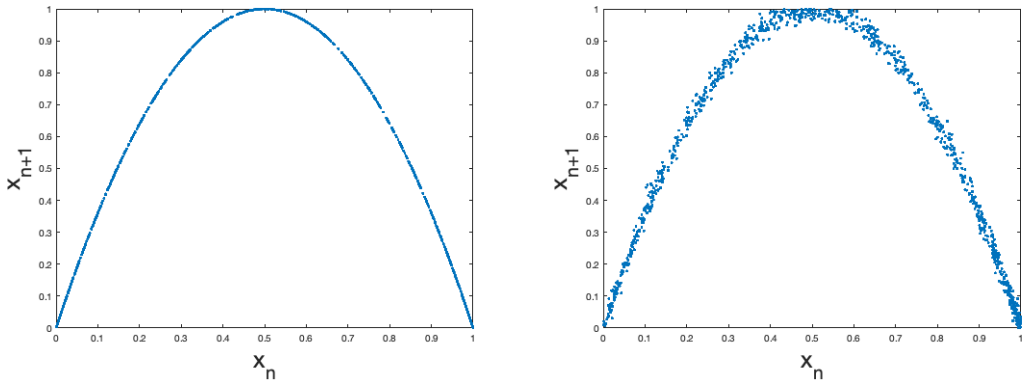

Fig. 1. Data consisting of $N = 1,000$ samples of (x_n, x_{n+1}) pairs sampled from: (Left) The logistic map, $x_{n+1} = f(x_n) = 4x_n(1 - x_n)$ along an orbit, following an initial transient so that the sample distribution closely approximates the invariant distribution, $p_X(x) = \frac{1}{\pi\sqrt{x(1-x)}}$. The joint distribution is a delta function, $p_{X'X}(x', x) = \delta(x' - f(x))$. (Right) A noisy logistic map $x_{n+1} = f(x_n) = 4x_n(1 - x_n) + s_n$, where s_n is chosen from an i.i.d. truncated normal distribution of standard deviation $\sigma = 0.02$. The "blur" of points roughly describes the joint distribution, $p_{X'X}(x', x) = \nu(x' - f(x))$, for ν is the truncated normal distribution, Eq. (11). See resulting kernels in Fig.2.

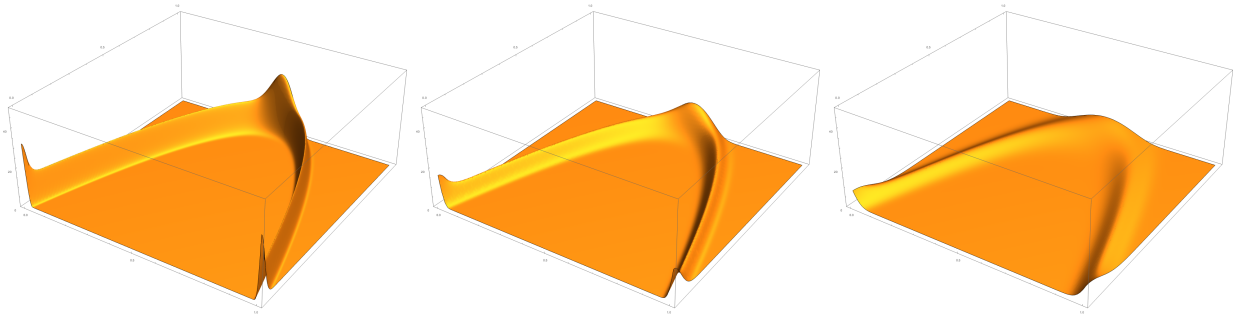

Fig. 2. Kernel of the Frobenius-Perron operator, in the case of a truncated normal distribution sampling, Eq. (11), with standard deviations $s = 0.025, s = 0.05$ and $s = 0.1$. Notice the "bumps" that appear as what would otherwise seem mostly like a normal distribution with tails on both sides of the peaks, becomes clamped, reflecting that the bounded domain variant must have this feature to remain a probability distribution, $\int p(x) = 1$ on the stated domain. See sample data in Fig. 1.

In [Bollt & Santitissadeekorn, 2013], we interpreted the random sampling associated with Eq. (6) as a Monte-Carlo integration estimate involved with projection onto basis functions, Eq. (8). Now in the

next sections, we will encode this same expression as a histogram based density estimator of a Bayesian interpretation of the transfer operator. This will open the door to considering a different kind of error analysis, as well as other estimators. A sample of data for the Logistic map, and a random map Logistic map perturbed by truncated normal distribution noise, is shown in Fig. 1 and 2. An Ulam-Galerkin estimate of the stochastic matrices Eq. (8) are shown in Fig. 3-4, with further interpretation as a histogram estimator as described in the next section of corresponding Bayesian estimators.

3. Bayesian Interpretation of the Transfer Operator

The Frobenius-Perron operator has a Bayesian interpretation as follows. Here we will write $x' = f(x)$ so that (x', x) are a output-input pair, of f, which we take as samples of random variables, X and X'. Considering, a statement of conditional and compound densities leads to an interpretation of the Frobenius-Perron operator as a Bayes update. Reviewing, the joint density, $p_{X'X}(x', x)$, of random variables X' and X marginalizes to,

$$p_{X'}(x') = \sum_{x:x'=f(x)} p_{X'X}(x', x) = \sum_{x:x'=f(x)} \frac{p_X(x)}{|Df(x)|}, \tag{12}$$

in terms of the summation over all pre-images of x'. Notice that the middle term is written as a marginalization across x of all those x that lead to x'. This Frobenius-Perron operator, as usual, maps densities of ensembles under the action of the map f. Comparing to the defining statement of a conditional density in terms of a joint density,

$$p_{X'X}(x', x) = p_{X'|X}(x'|x)p_X(x). \tag{13}$$

We reinterpret, in the noiseless case,

$$p_{X'|X}(x'|x) = \frac{1}{|Df(x)|}\delta(x' - f(x)). \tag{14}$$

In the language of Bayesian uncertainty propagation, $p_{X'|X}(x'|x)$ describes a likelihood function, interpreting future states x' as data and past states x as parameters, by the standard Bayes phrasing,

$$p(\Theta|\text{data}) \propto p(\text{data}|\Theta) \times p(\Theta), \tag{15}$$

for parameter Θ, or simply by standard names of the terms,

$$\text{posterior} \propto \text{likelihood} \times \text{prior}. \tag{16}$$

In these terms, comparing to Eq. (5), $P_{i,j}$ can be interpreted as a matrix of likelihood functions

$$P_{i,j} = P(x \in B_i | x' \in B_j) = \frac{P(x \in B_i, x' \in B_j)}{P(x \in B_i)} = \frac{m(B_i \cap f^{-1}(B_j))}{m(B_i)}. \tag{17}$$

Furthermore, the standard Ulam estimator, Eq. (6), can be taken as a histogram method to estimate, the joint and marginal probabilities, $p_{X'X}$ and p_X by occupancy counts in the related boxes, B_i and B_j with,

$$\begin{aligned} P(x \in B_i, x' \in B_j) &\sim \#x_n \in B_i, x_{n+1} \in B_j, \text{ and,} \\ P(x \in B_i) &\sim \#x_n \in B_i. \end{aligned} \tag{18}$$

The conditional follows by division, to the estimator of the matrix $P_{i,j}$ describing the likelihood function. In these terms, we are positioned to describe the statistical error of expressions such as Eq. (6) for the matrix $P_{i,j}$ estimator of the Frobenius-Perron operator, by the theory of density estimators, for $p_{X'X}(x', x)$ and $p_X(x)$ respectively. First, in the next section, we will discuss this histogram estimator, and then in following sections, we will consider other estimators, notably the kernel density estimator.

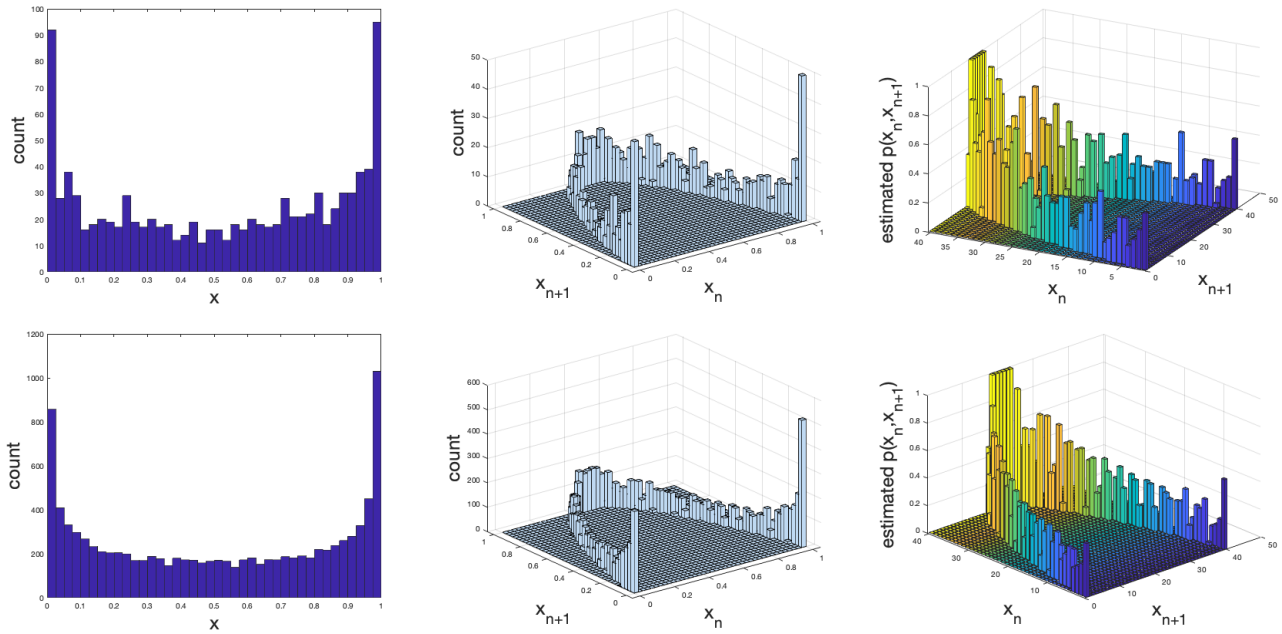

Fig. 3. Histogram estimates of (Top Row) The marginal, joint and conditional distributions $p_X(x), p_{X'X}(x', x), p_{X'|X}(x'|x)$of the $N = 1,000$ sample orbit illustrated in Fig. 1, using bandwidth of $K = 40$ and 40×40 cells. The rightmost estimate, $p_{X'|X}(x'|x)$ by Eqs. (5), (6), (8), (17), (18) is therefore an Ulam-Galerkin estimate of the Frobenius-Perron operator that can be understood as a transition matrix. (Bottom Row) Longer orbit of $N = 10,000$ iterates allows a better (smoother, with less variability) estimate of the true distributions. Compare to the true distribution shown in Fig. 2, sampled by truncated normal $t\mathcal{N}$.

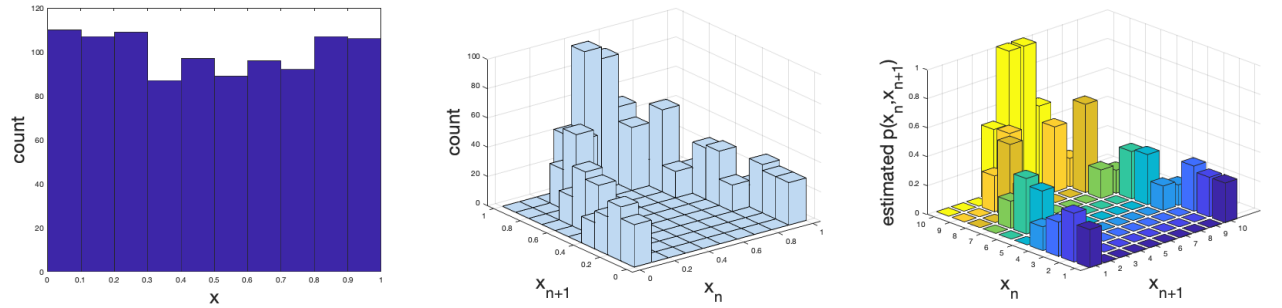

Fig. 4. Histogram distribution estimates, as Fig. 3, but sampling $N = 1,000$ points of x, not from an orbit, but i.i.d. from a uniform distribution, histogram estimates of the marginal, joint and conditional distributions $p_X(x), p_{X'X}(x', x), p_{X'|X}(x'|x)$, however more coarsely (wider bandwidth for less variability, more bias (smoothing) with $K = 10$ and 10×10 cells respectively. Compare to the true distribution shown in Fig. 2, sampled by truncated normal $t\mathcal{N}$.

4. Theory of Density Estimation

As we have argued in the previous section, the problem of estimation of an Ulam-Galerkin estimator of the Frobenius-Perron operator is equivalent to the Bayes computation of the conditional density $p_{X'|X}(x'|x)$, derived by histogram estimators of the joint and marginal densities, $p_{X'X}(x', x), p_X(x)$, respectively. Therefore in this section, we review what is classical theory from the statistics of density estimation, found in many excellent textbooks, such as [Silverman, 1999; Scott, 2015]. First we will review some details of histogram estimators, considering the central issues bias, variance and choice of bandwidth. Then in the subsequent subsection, we will re-cast the problem as one of kernel density estimation (KDE), for which those same three issues, bias, variance and bandwidth, suggest some advantages for KDE.

For each, we state a general random variable X that is distributed by $p_X(x)$, but for simplicity of presentation in this section, we assume a unit interval, $x \in [0, 1]$, and so is the support of $p_X(x)$. Likewise,

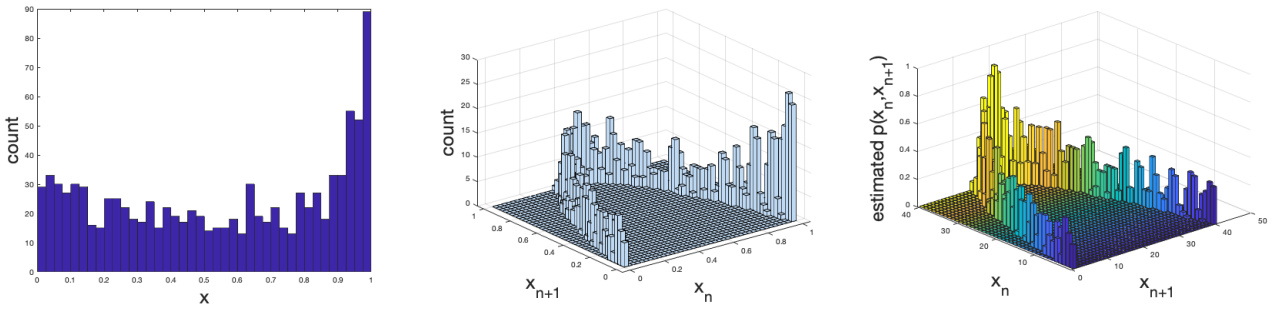

Fig. 5. Histogram distribution estimates, as Fig. 4, of $p_X(x)$, $p_{X'X}(x', x)$, $p_{X'|X}(x'|x)$, but of the random Logistic map orbit data by truncated normal distribution noise, $s = .02$ the orbit data shown in Fig. 1. These plots are similar in variability vs bias (smoothing) character as the no noise scenario of Fig.4(Top Row) even with the same smoothing, despite the differences of the true underlying distribution as the kernel cannot distinguish these properties. However, it is interesting that the marginal distribution estimating the invariant distribution is clearly different.Compare to the true distribution shown in Fig. 2, sampled by truncated normal $t\mathcal{N}$.

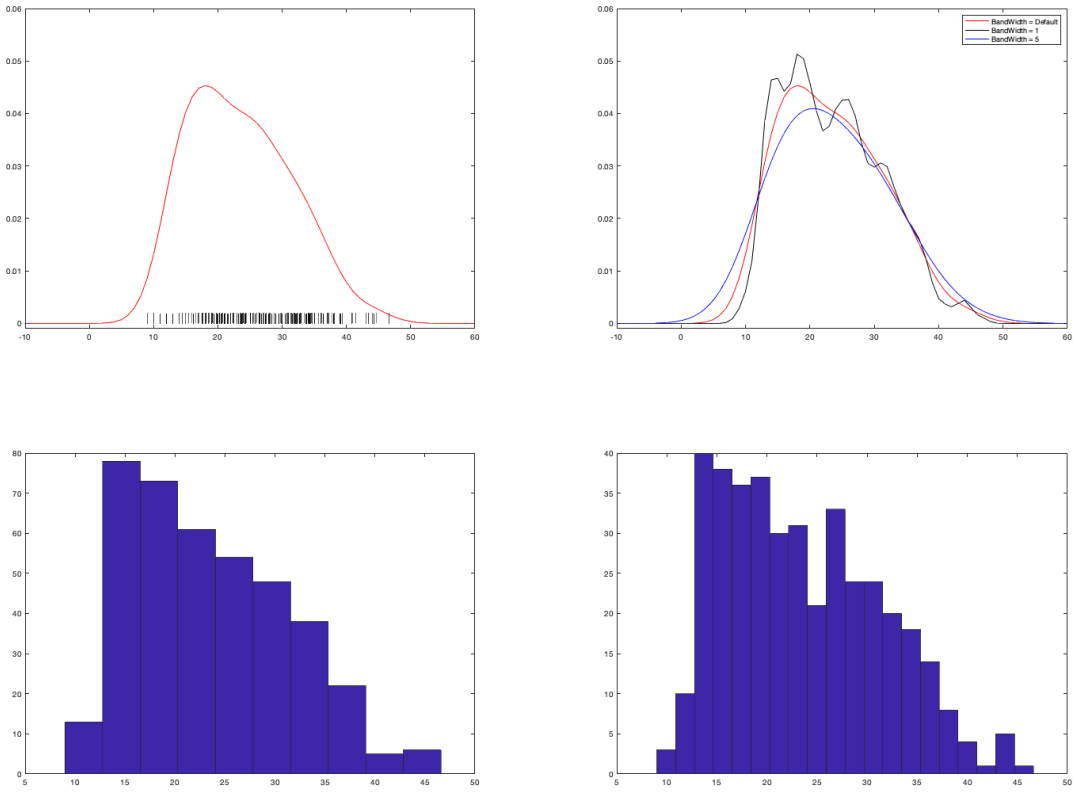

Fig. 6. Top raw shows the KDE estimation of of a given data set. Top right figure comperes the estimation with the kernel band width. Bottom raw demonstrate the density estimation for the same data set by histogram estimation method. Furthermore, bottom left and right figures compere the estimation with the bin size.

assume $(x', x) \in [0, 1] \times [0, 1]$, and $P_{X'X}(x', x)$ has support in the unit square. The standard theory of density estimation also assumes a smooth density function, $|p'_X(x)| \leq C_1$ and $|DP_{X'X}| \leq C_2$, for constants of uniform bound, $C1, C_2 \geq 0$. Note however, this is already a problem regarding perhaps the most popular

example for pedagogical study, the invariant density of the logistic map is $p_x(x) = \frac{1}{\pi\sqrt{x(1-x)}}$, is unbounded and has unbounded derivative, $p'_X(x) = \frac{2x-1}{2\pi\sqrt{x(1-x)^3}}$. Nonetheless, many have made practice of presenting the invariant distribution as estimated from sample orbits, Fig. 3(Left).

The key issue in any estimator is accuracy, versus the amount of data available. Generally, we want an analysis of mean square error, MSE of the estimator, which requires both bias and variance, since $MSE = bias^2 + Var^2$.

4.1. *Theory of Density Estimation for Histograms*

Here we review density estimation, closely following [Scott, 2015], first in one dimension, and then the multivariate scenario.

Considering a unit interval, $[0, 1]$, it may be divided into K cells (bins),

$$\mathcal{B} = \{B_i\}_{i=1}^K, B_i = [\frac{i-1}{K}, \frac{i}{K}), i = 1, 2, .., K, \tag{19}$$

which is a uniform topological partition, meaning interiors are mutually disjoint and the union closure covers. Similarly, a multivariate histogram is a topological partition into bins, usually rectangles (but other shapes, especially tessellations are not uncommon). Otherwise, continuing with the discussion of the single variate estimator, given a sample $\{x_n\}_{n=1}^N$, suppose that $x \in B_i$. Then,

$$\overline{p_{N,K}}(x) = \frac{\#x_n \in B_i}{N} \times \frac{1}{m(B_i)} = \frac{K}{N}\sum_{j=1}^{N} \xi_{B_i}(x_j). \tag{20}$$

The key part of density estimators is the analysis of the bias of the estimator, [Scott, 2015], continuing with $\overline{p_{N,K}}(x)$ for a point $x \in B_i$ that need not be assume to be one of the data points x_j. Consider the probability that the sample $x_j \in B_i$, $P(x_j \in B_j)$,

$$\mathbb{E}(\overline{p_{N,K}}(x) = KP(x_j \in B_j)) = K\int_{\frac{i-1}{K}}^{\frac{i}{K}} p_X(s)ds = p_X(\tilde{x}), \text{ for } \tilde{x} \in B_j, \tag{21}$$

by mean value theorem and fundamental theorem of calculus. Therefore, the bias of the estimator is,

$$bias_{hist}[\overline{p_{N,K}}(x))] = \mathbb{E}(\overline{p_{N,K}}(x) - p_X(x)) = p_X(\tilde{x}) - p_X(x) \leq |p'_X(\hat{x})||\tilde{x} - x| \leq \frac{C_1}{K}. \tag{22}$$

The first inequality follows again by the mean value theorem, this time for a value $\hat{x} \in (x, \tilde{x})$ (or perhaps the opposite order), and by product of absolute values, and the fact that $\tilde{x}, \hat{x} \in B_i$. Bias is a question of balancing the derivative $p'_X \leq C_1$ versus a good choice of the number of bins, K.

The variance can be computed with Eq. (20),

$$\begin{aligned} Var_{hist}(\overline{p_{N,K}(x)}) &= K^2 Var(\frac{K}{N}\sum_{j=1}^{N} \xi_{B_i}(x_j)) \\ &= \frac{K^2 P(x_j \in B_i)(1 - P(x_j \in B_i))}{N} = \frac{K^2(\frac{p_X(\tilde{x})}{K})(1 - \frac{p_X(\tilde{x})}{K})}{N} \\ &= \frac{Kp_X(\hat{x}) + p_X^2(\hat{X})}{N} \end{aligned} \tag{23}$$

Variance is a question of balancing the number of bins K versus the data count N, but relative to the unknown density p_X.

Therefore the mean square error for the density estimation $p_{N,K}(x)$ at an arbitrary point $x \in [0, 1]$ follows,

$$MSE_{hist}(\overline{p_{N,K}}(x)) = bias^2_{hist}(\overline{p_{N,K}}(x)) + Var_{hist}(\overline{p_{N,K}}(x)) \leq \frac{C_1^2}{K^2} + \frac{Kp_X(\hat{x}) + p_X^2(\hat{X})}{N}. \tag{24}$$

To interpret, when a fixed data set (size) N is given, from an unknown distribution p_X, we can only choose K and this choice is called bandwidth selection. From Eq. (24), large K (more bins) yields decreased bias (the first term), but the variance (the second term) will tend to be large. Thus demonstrates the balancing struggle between bias and variance in choosing the number of bins, the bandwidth. Figs. 3-5 demonstrate this bandwidth selection balancing act. Again, we reiterate that formally the analysis requires $C_1 \geq 0$ be bounded whereas the derivative of the invariant density of the logistic map is not of bounded derivative type in [0,1], still many estimates exist in the literature ([Hall & Wolff, 1995; Bollt, 2000; Nie & Coca, 2018] etc.), including ourselves, which we now call a "typical sin." An argument that it may not be fatal for practical problems is the fact that in real world dynamical systems that there is always noise, which has the effect of smoothing (e.g., noise sampled from a smooth distribution serves as a mollifier that can bring even a singular distribution "blurred" into a C^∞ distribution) or rather producing invariant densities that are smooth after all. See Fig. 3 for histogram density estimations for the Fig. 1(Right) randomly perturbed logistic map data.

Note that analysis of MSE in the theory of multivariate histogram estimators is similar in methodology, to which we refer to [Scott, 2015]. The important point to this stage of the paper is that the famous Ulam-Galerkin estimation of the transfer operators by formula Eq. (6) amounts to problems of density estimations of the marginal and joint distributions $p_X(x)$ and $p_{X'X}(x', x)$ leading to the estimation of the conditional distribution $p_{X'|X}(x'|x)$. This said, we can now contrast that this discussion is not a description of the Ulam problem (vs the Ulam-Galerkin estimation), since the Ulam problem describes that these estimates are stochastic matrices each with dominant eigenvector describing the invariant state of the corresponding Markov chain, and that converges weakly to the invariant distribution of the original dynamical system; and conditions for when this is in fact true were given as a theorem under hypothesis of bounded total variation first in [Li, 1976].

Now we pursue to other, perhaps more favorable density estimators of the transfer operator, notably kernel density estimation.

4.2. *Theory of Kernel Density Estimation*

Another major category of data-driven nonparameteric density estimators is the Kernel Density Estimator, or KDE. It is a data-driven estimator based on mixing simpler densities. These are defined in terms of a kernel function, $\mathcal{K}$, which is itself a real density function. Stating for single-variate data, $\mathcal{K} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, and such that: 1) $\mathcal{K}(x)$ is symmetric, 2) $\int_{\mathbb{R}} \mathcal{K}(x)dx = 1$, 3) $lim_{x\to\pm\infty}\mathcal{K}(x) = 0$. These are sufficient to guarantee that the KDE estimator built out of convex sums of sampling $\mathcal{K}$ at data points, itself is a density,

$$\overline{p_{N,\delta}}(x) = \frac{1}{\delta N}\sum_{i=1}^{N} \mathcal{K}(\frac{x_i - x}{\delta}), \tag{25}$$

where $\delta > 0$ is the bandwidth that controls the range or extent of influence of a given data point x_i and is a primary parameter choice, just as was the bin size for the histogram method. There are several favorite kernels, but we mention especially the Gaussian kernel, $\mathcal{K}(x) \propto exp(-x^2/2)$, and the Epanechnikov kernel, $\mathcal{K}(x) =\propto 1 - x^2$.

Similarly, for multivariate data, $\overline{p_{N,\Sigma}}(x) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{K}_\Sigma(x_i - x)$, using the common compact "scaled" kernel notation, and $\mathcal{K}_\Sigma(z) = |\Sigma^{-1/2}|K(\Sigma^{-1/2}z)$ which for the most commonly used Gaussian kernel, $K(z) = (2\pi)^{-d/2}exp(-z^Tz/2)$. The matrix Σ serves the role of a variance-covariance in the case of a Gaussian with mean x_i.

A crucial difference is whereas histograms are centered on spatial positions, the location of the bins, and data would occupy those positions, a kernel density estimator is centered only where there is data. This can be a real savings when considering sparsely sampled data from a distribution with a relatively small support, especially in higher dimensions where the curse of dimensionality prohibits covering the space with boxes, many of which may be empty of data if the support of the density zero.

To analyze MSE, we must again state the bias and variance of the estimator.

$$bias_{kde}(\overline{p_{N,\delta}}(x)) = \mathbb{E}(\frac{1}{\delta N}\sum_{i=1}^{N}\mathcal{K}(\frac{x_i - x}{\delta}) - p_X(x))$$
$$= \frac{1}{N}\int \mathcal{K}(\frac{y-x}{\delta})p(y)dy - p_X(x) = \int \mathcal{K}(z)p(x+\delta z)dz - p_X(x). \quad (26)$$

By substituting a Taylor series, $p(x+\delta z) = p_X(x) + \delta z p'_X(z) + \frac{1}{2}\delta^2 z^2 p''_X(x) + o(\delta^2)$, it follows [Scott, 2015] that,

$$bias_{kde}(\overline{p_{N,\delta}}(x)) = \frac{c}{2}\delta^2 p''_X(z) + o(\delta^2), \quad (27)$$

where $c = \int z^2\mathcal{K}(z)dz$ is the second moment of the kernel.

Analysis of variance follows similarly.

$$Var_{kde}(\overline{p_{N,\delta}}(x)) = Var(\frac{1}{\delta N}\sum_{i=1}^{N}\mathcal{K}(\frac{x_i - x}{\delta})) \leq \frac{1}{\delta^2 N}\mathbb{E}(\mathcal{K}^2(\frac{x_i - x}{\delta})) \quad (28)$$

$$= \frac{1}{\delta N}\int \mathcal{K}^2(y)(p_X(x) + \delta y p'_X(x) + o(\delta))dy = \frac{1}{\delta N}(p_X(x))\int \mathcal{K}^2(y)dy + o(\delta)) = \frac{1}{\delta N}p_X(x)d + o(\frac{1}{\delta N}), \quad (29)$$

with $d = \int \mathcal{K}^2(y)dy$.

So follows the MSE, combining Eqs. (27) and (28).

$$MSE_{kde}(x) = \frac{c^2}{4}\delta^4|p''_X(x)|^2 + \frac{d}{\delta N}p_X(x) + o(\delta^4) + o((\delta N)^{-1}). \quad (30)$$

Or,

$$MSE = O(\delta^4) + O((\delta N)^{-1}) \quad (31)$$

moderates the MSE relative to bandwidth δ choice. We see the in the role of bandwidth choice, bias dominate for larger $\delta > 0$ proportionally to $p''_X(x)$ (or curvature), or variance dominating for smaller δ.

4.3. *Optimal MSE*

The choice of bandwidth tailored to a given data set size is the key question in using a given nonparameteric estimator. While both the histogram discussion and KDE discussion each have unknown to us constants, depending on either $p_X(x)$ or derivatives of the same, which not knowing p, these are inaccessible, all we can assert is bandwidth and data set size. For histograms,

$$MSE_{hits}(\overline{p_{N,K}}(x)) \sim \mathcal{O}(\frac{1}{K^2}) + \mathcal{O}(\frac{K}{N}), \quad (32)$$

but for kernel density estimation,

$$MSE_{hits}(\overline{p_{N,K}}(x)) \sim \mathcal{O}(\delta^4) + \mathcal{O}(\frac{1}{\delta N}), \quad (33)$$

each balances large bias when the bandwidth is too large, versus large variance large variance when the bandwidth times data set size is too small, but at different rates. The asymptotic mean square error can be shown to be optimal when,

$$\delta_{opt;KDE} = \frac{C}{N^{1/5}}, \quad (34)$$

where C is a constant related to the unknown density function, $C = \frac{4p(x)d}{c^2|p''(x)|^2}$. Similarly, for histograms, an optimal bandwidth selection is described by,

$$K_{opt;hist} = (\frac{NC_1^2}{p(\tilde{x})})^{\frac{1}{3}}, \quad (35)$$

(and note that bandwidth for a histogram is considered to be as $1/K$). So we see that asymptotically, cubic versus quintic scaling and the KDE may be better when best used (optimal bandwidth), but in practice that also depends on the constants, and one depends largely on the p_X and the other also on P''_X. The most relevant quantity when choosing a method is that for a KDE, MSE when using the optimal bandwidth is, [Scott, 2015]

$$MSE_{\delta_{opt;KDE}}(\overline{p_{\delta,N}}(x)) = \mathcal{O}(\frac{1}{N^{\frac{4}{5}}}). \tag{36}$$

However, all we can do is selection in practice, since we will not know p_X, is to inspect the scaling as we will do in the results Sec. 5. Beyond 1-dimensional density estimation, multivariate KDE has a slower bandwidth rate,

$$\delta_{opt;KDE} = \frac{C}{N^{\frac{1}{4+D}}}, \tag{37}$$

in $D \geq 1$ dimensions. For example, the density estimation problem associated with $P_X(x)$ is $D = 1$ for a transfer operator of the logistic map, but the joint density $P_{X'X}(x', x)$ is $D = 2$ for the same.

5. Results and Discussion

In this section, we focus on estimating the Frobenius-Perron operator by using the previously discussed density estimation methods. In other words, we calculate the P matrix in eq. (17) by density estimation methods. We use the logistic map example to demonstrate the results. In this demonstration, uniformly distributed $N = 10^6$ initial conditions were used and evolved using a logistic map $x' = 4x(1 - x)$ for a relatively long time which was used to approximate the invariant density $\rho(x) = \frac{1}{\pi\sqrt{x(1-x)}}$. Now our goal is to estimate the Frobenius-Perron operator by evaluating the probability density function $p(x'|x) = \frac{p(x,x')}{p(x)}$. For this calculation, we estimated the $p(x)$ and the joint probability density $p(x, x')$ by using the data through density estimation methods. In this section, we analyze the estimation of the $p(x)$ by the density estimation methods in detail and compare it to the theoretical explanation in the section 4. Then we demonstrate the estimation P matrix of the Frobenius-Perron operator by discretized probability density function $p(x'|x)$.

5.1. *Histogram Estimation vs Kernel Density Estimation for logistic map example*

Histogram Estimation is based on the number of samples (N) and the number of bins (K) (See details in section 4). Here, we demonstrate the effect of the bin size. The estimation of the invariant density $\rho(x)$ by the histogram method is denoted by $\overline{\rho_{N,K}}$. Figure(7) shows the changes in the estimation with parameter K. Since the true density is known, the mean squared error(MSE) and the upper bound (UB) of the MSE can be calculated for this example by Eq. (24). Note that the notations

$$MSE = (\rho(x) - \overline{\rho_{N,K}}(x))^2$$

$$UB = C_1^2 \frac{1}{K^2} + \frac{\rho(\hat{x})}{N} K + \frac{\rho^2(\hat{x})}{N}$$

are similar to the Eq. (24) and the constants are evaluated by the true density function $\rho(x)$. Furthermore, by analyzing the UB and MSE, we can get an idea of the optimal bin size $K_{opt;hist}$ (see figure (8) and section (4.3)).

As we discussed in the section (4.2), KDE is based on the number of data points (N) and the kernel bandwidth (δ). In this section, we numerically demonstrate the effect of the bandwidth. Notice that, a change in the bandwidth will result in a change in the MSE (see figure (9)). The upper bound (UB) of MSE for the kernel density estimation is given in Eq. (30) and the following results are calculated by the Eq. (30). Error analysis and the optimal MSE is demonstrated in the figure (10). Furthermore, note that the optimal MSE can be achieved with a bandwidth of approximately $\delta_{opt;KDE} = 0.0011$.

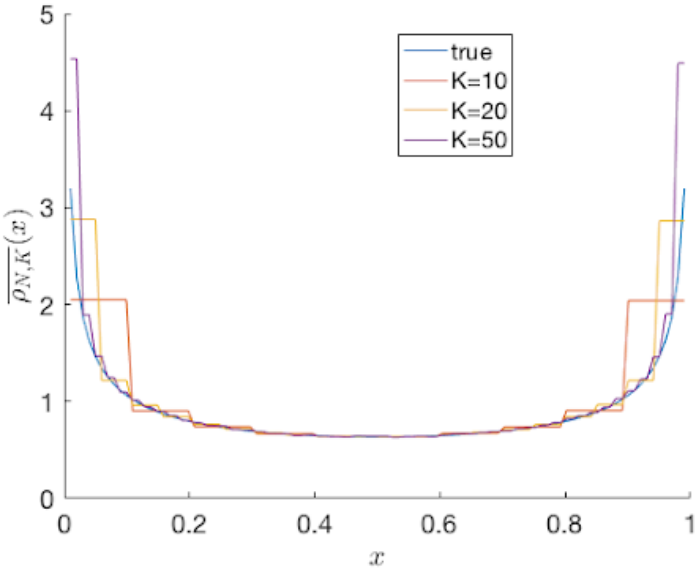

Fig. 7. The histogram based density estimation $\overline{\rho_{N,K}}$ for the invariant density ρ of the logistic Map. This figure demonstrate the effects of the number of bins K for the estimation $\overline{\rho_{N,K}}$ with sample size $N = 10^6$.

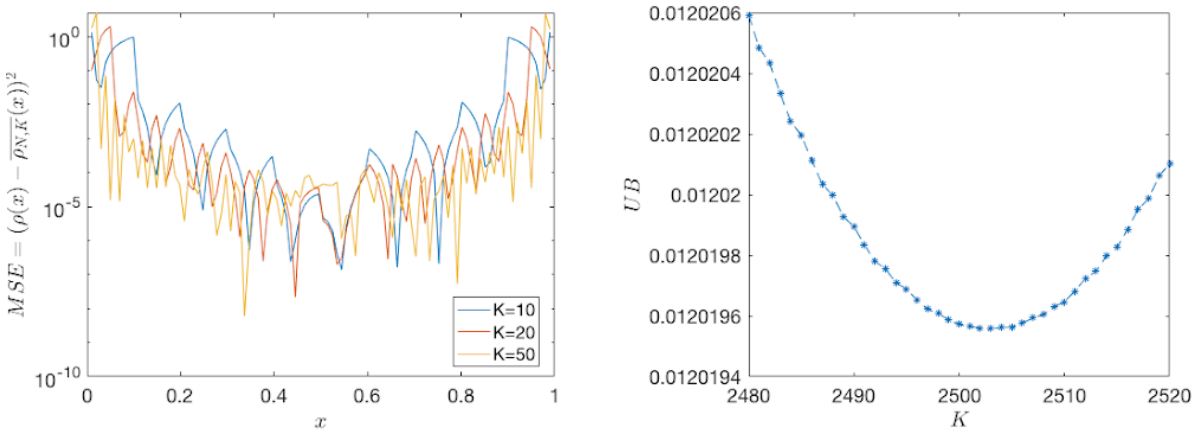

Fig. 8. Left figure shows the MSE of the estimated density function on the equaly spaced 100 grid points on the interval $[0.01, 0.99]$. Furthermore it shows the behavior of the MSE with bin size K. Analysis of UB can be found in the right figure. It shows the changes of the UB with the bin size for logistic map example and optimal MSE can be achieved around the bin size $K_{opt} = 2503$.

Due to the unboundedness of the density function, both estimation methods have higher estimation errors closer to the endpoints of the interval $[0, 1]$. In general, KDE has issues when estimating a probability density with finite support. However, the overall estimating error is much lower for KDE when compared to the histogram method. Figure (11) demonstrates that MSE for KDE is comparatively lower than the MSE for histrogram method with their optimal parameter values.

5.2. *Estimating the Frobenius-Perron operator*

Now, we will numerically investigate the theoretical discussion presented in section (3). We have shown the popular classical Ulam-Galerkin method as a histogram estimation of the $p(x'|x)$. Hence, we argued that estimating Frobenius-Perron operator through conditional probability density $p(x'|x)$ can be extended by using any density estimation method. In this article, we presented the KDE as an alternative density

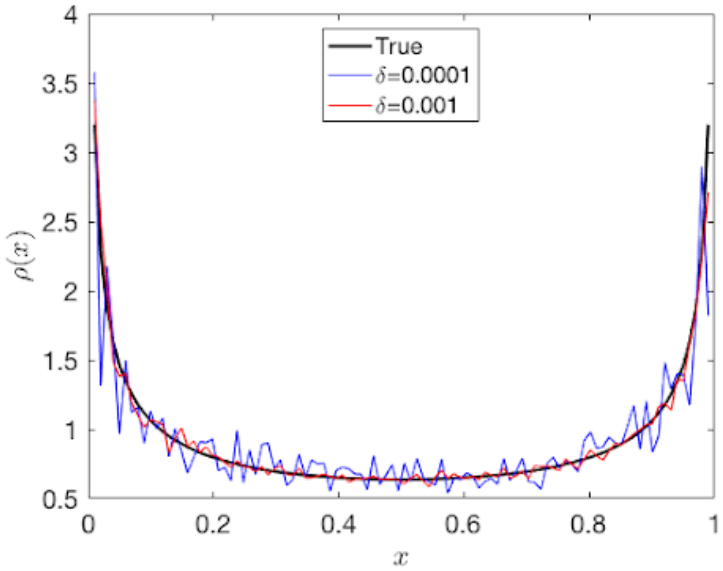

Fig. 9. The KDE for the invariant density ρ of the logistic map. This figure shows the effect of kernel bandwidth δ for the KDE with sample size $N = 10^6$.

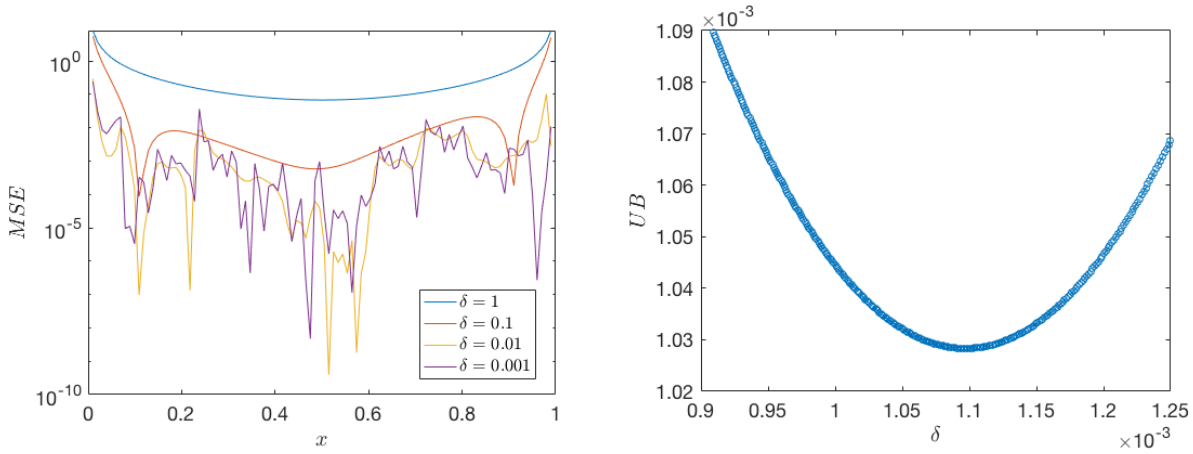

Fig. 10. Left shows the MSE of the estimated density function on the equaly spaced 100 grid points on the interval $[0.01, 0.99]$. Furthermore it shows the behavior of the MSE with bandwidth δ. Analysis of the UB can be found in the right figure. It shows the changes of the UB with the δ for the logistic map example and optimal bandwidth can be identified as being around $\delta_{opt} = 0.0011$.

estimation to the histogram method.

The finite-dimensional estimation of Frobenius-Perron operator can be represented using a matrix (P). The comparison of the estimated P matrix by each method can be found in the figure (12). Furthermore, left eigenvector corresponds to the eigenvalue 1 of the P matrix can be used to estimate the invariant density of the map. Figure (13) shows the left eigenvector correspond to the eigenvalue 1 of matrix P calculated by histogram and KDE method.

6. Conclusion

The main contribution of this paper is the probability density viewpoint to the estimating Frobenius–Perron operator that enables us to incorporate the existing rich analysis of statistical density estimation formalism to find an efficient estimator. Furthermore, the theory suggests that the kernel density estimation is more efficient than the histogram methods used in the standard Ulam method. Additionally, this paper

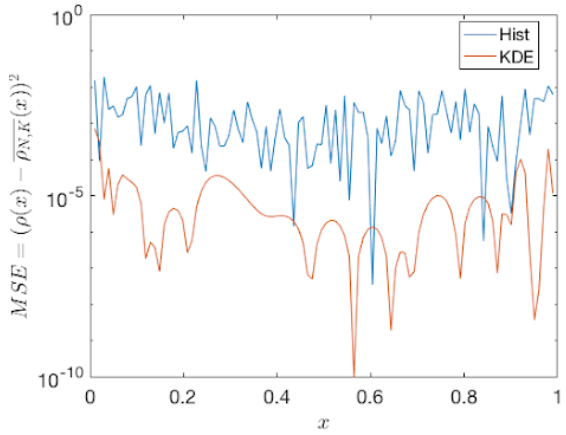

Fig. 11. Comparison of the MSE of KDE and histogram methods with their optimal parameter values.

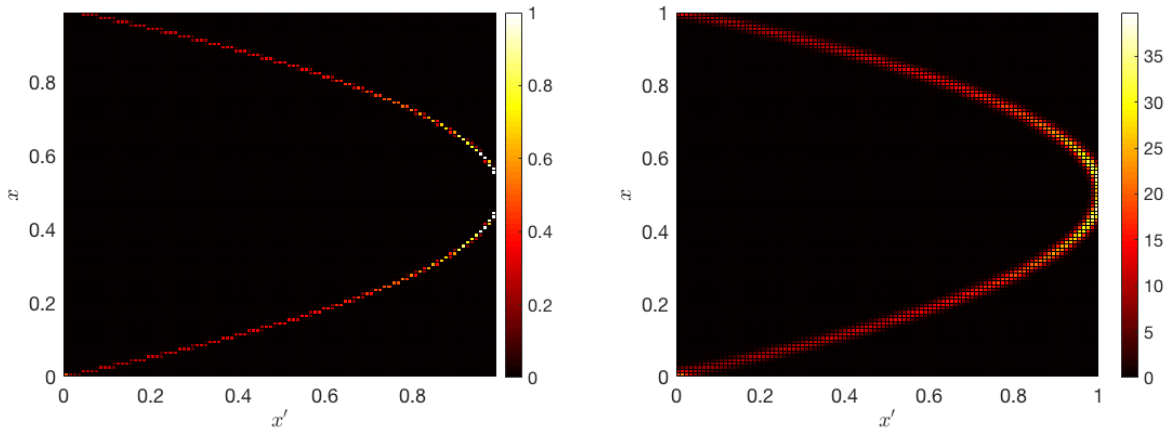

Fig. 12. The matrix P which estimates the Frobenius-Perron operator in finite domain. The P matrix (left) is calculated by the histogram method and (right) the P matrix as calculated by the KDE.

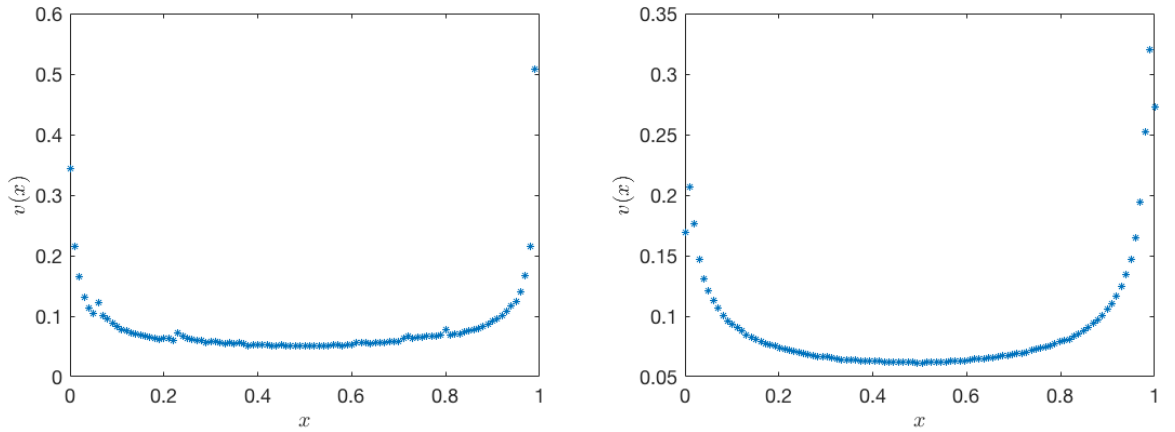

Fig. 13. The left eigenvector correspond to the eigenvalue 1 of matrix P which can used estimates the invariant density of the logistic map. The eigenvector of P matrix which is calculated by the histogram method (left). The eigenvector of P matrix which is calculated by the KDE (right).

discusses a kernel density estimation method to estimate the transition probability that estimates the Frobenius–Perron operator, from empirical time series ensemble data of a dynamical system. To date, the literature mostly used the Ulam method for estimating the transfer operator but this study offers a more accurate estimation based on KDE. Our Bayesian interpretation of the Frobenius–Perron operator is important to identify the operator in terms of conditional probability density because it allows us to bring density estimation theory into play. It is shown at the beginning of this article how the Ulam-Galerkin method can be interpreted as a histogram density estimation method. Theory and numerical results have been presented which suggest that KDE is a better approximation for estimating probability densities. Hence, is is also shown that KDE may be used for finite approximation for the Frobenius–Perron operator. Finally we showed that the KDE-based approximation is a better estimator for the operator than the histogram-based current Ulam-Galerkin method.

As a result of conducting this research, we propose the possibility of introducing any density estimation methods to this field. It would be fruitful to pursue further research about the KDE-based estimation of the Frobenius–Perron operator for high-dimensional maps to analyze this method.

Acknowledgments

Erik Bollt gratefully acknowledges funding from the Army Research Office (ARO), the Defense Advanced Research Projects Agency (DARPA), the National Science Foundation (NSF) and National Institutes of Health (NIH) CRNS program, and the Office of Naval Research (ONR) during the period of this work.

References

Bollt, E. M. [2000] "Controlling chaos and the inverse frobenius–perron problem: global stabilization of arbitrary invariant measures," *International Journal of Bifurcation and Chaos* **10**, 1033–1050.

Bollt, E. M., Billings, L. & Schwartz, I. B. [2002] "A manifold independent approach to understanding transport in stochastic dynamical systems," *Physica D: Nonlinear Phenomena* **173**, 153–177.

Bollt, E. M. & Santitissadeekorn, N. [2013] *Applied and computational measurable dynamics* (SIAM).

Boyarsky, A. & Góra, P. [1997] "Laws of chaos : Invariant measures and dynamical systems in one dimension," .

Chiu, C., Du, Q. & Li, T. [1992] "Error estimates of the markov finite approximation of the frobenius-perron operator," *Nonlinear Analysis: Theory, Methods & Applications* **19**, 291–308, doi:https://doi.org/10.1016/0362-546X(92)90175-E, URL https://www.sciencedirect.com/science/article/pii/0362546X9290175E.

Dehnad, K. [1987] "Density estimation for statistics and data analysis," *Technometrics* **29**.

Dellnitz, M., Froyland, G. & Junge, O. [2001] "The algorithms behind gaio — set oriented numerical methods for dynamical systems," *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, ed. Fiedler, B. (Springer Berlin Heidelberg, Berlin, Heidelberg), ISBN 978-3-642-56589-2, pp. 145–174.

Guder, R., Dellnitz, M. & Kreuzer, E. [1997] "An adaptive method for the approximation of the generalized cell mapping," *Chaos, Solitons & Fractals* **8**, 525–534, doi:https://doi.org/10.1016/S0960-0779(96)00118-X, URL https://www.sciencedirect.com/science/article/pii/S096007799600118X, nonlinearities in Mechanical Engineering.

Hall, P. & Wolff, R. C. L. [1995] "Properties of invariant distributions and lyapunov exponents for chaotic logistic maps," *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 439–452, URL http://www.jstor.org/stable/2345972.

Lasota, A. & Yorke, J. A. [1982] "Exact dynamical systems and the frobenius-perron operator," *Transactions of the american mathematical society* **273**, 375–384.

Li, T.-Y. [1976] "Finite approximation for the frobenius-perron operator. a solution to ulam's conjecture," *Journal of Approximation Theory* **17**, 177–186, doi:https://doi.org/10.1016/0021-9045(76)90037-X, URL https://www.sciencedirect.com/science/article/pii/002190457690037X.

Ma, T. & Bollt, E. M. [2013] "Relatively coherent sets as a hierarchical partition method," *International Journal of Bifurcation and Chaos* **23**, 1330026.

Nie, X. & Coca, D. [2018] "A matrix-based approach to solving the inverse frobenius–perron problem using sequences of density functions of stochastically perturbed dynamical systems," *Communications in Nonlinear Science and Numerical Simulation* **54**, 248–266, doi:10.1016/j.cnsns.2017.05.011.

Scott, D. W. [2015] *Multivariate density estimation: Theory, practice, and visualization* (John Wiley & Sons, Inc.).

Silverman, B. W. [1999] *Density Estimation for statistics and data analysis* (Chapman and Hall).

Tukey, J. W. [1961] "Curves as parameters, and touch estimation," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (University of California Press), pp. 681–695.

Tukey, P. A. & Tukey, J. W. [1981] "Graphical display of data sets in 3 or more dimensions," *Interpreting multivariate data* **29**.

Ulam, S. M. [1960] *A collection of mathematical problems*, 8 (Interscience Publishers).