AFRL-AFOSR-JP-TR-2017-0042

Variety Preserved Instance Weighting and Prototype Selection for Probabilistic Multiple Scope Simulations

Takashi Washio
Osaka University

05/30/2017
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 30-05-2017	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* 22 Apr 2015 to 21 Apr 2017

4. TITLE AND SUBTITLE
Variety Preserved Instance Weighting and Prototype Selection for Probabilistic Multiple Scope Simulations

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA2386-15-1-4008

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Takashi Washio

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Osaka University
1-1 Yamadaoka
Suita, Osaka, 565-0871 JP

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-JP-TR-2017-0042

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Studies of probabilistic modeling and simulation focused on some special situations and behaviors, including rare events and scenarios, occurred in large scale and complex systems have been performed extensively. However, few studies have explored the modeling and simulation techniques to seamlessly cover multiple scopes including both major (frequent) and minor (rare) situations and behaviors embedded in a given large data set. A main obstacle to develop such techniques comes from the difficulty to capture the situations and the behaviors having highly contrasted probabilities in a unique model of the data distribution. Two technical issues must be addressed for overcoming this obstacle; (a) weighting instances in a given large data set and (b) selecting prototypes from a given large data set. Particularly, the latter is for the modeling from massive data to which the thorough access is not feasible. A method to address these two issues preserves the variety of instance distributions of the data set and provide the basis of the seamless simulations over the multiple scopes. In the first year, we performed a mathematical analysis to derive the required conditions on our targeting method and designed a principle of the method which largely alleviate the obstacle by efficiently sampling the required prototypes with their appropriate weights. In the second year, we implemented the designed principle of the targeting method to an algorithm in computers and evaluated its generic performance preserving the variety of instance distributions of the data set in the selected prototypes. The mathematical analysis, the designed principle, the implemented algorithm and its performance evaluation presented in this report is the first work in worldwide for the seamless and comprehensive probabilistic simulations of the large scale and complex systems.

15. SUBJECT TERMS
Data Mining

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			KNOPP, JEREMY
Unclassified	Unclassified	Unclassified	SAR	28	**19b. TELEPHONE NUMBER** *(Include area code)* 315-227-7006

Final Report for AOARD Grant FA2386-15-1-4008

"Variety Preserved Instance Weighting and Prototype Selection for Probabilistic Multiple Scope Simulations"

April/21st/2017

Name of Principal Investigators (PI and Co-PIs): Takashi Washio

- e-mail address : washio@ar.sanken.osaka-u.ac.jp
- Institution : The Institute of Scientific and Industrial Research, Osaka University
- Mailing Address : 8-1, Mihogaoka, Ibarakishi, Osaka, 567-0047, Japan
- Phone : +81-6-6879-8540
- Fax : +81-6-6879-8544

Period of Performance: April/22nd/2015 – April/21st/2017

Abstract: Studies of probabilistic modeling and simulation focused on some special situations and behaviors, including rare events and scenarios, occurred in large scale and complex systems have been performed extensively. However, few studies have explored the modeling and simulation techniques to seamlessly cover multiple scopes including both major (frequent) and minor (rare) situations and behaviors embedded in a given large data set. A main obstacle to develop such techniques comes from the difficulty to capture the situations and the behaviors having highly contrasted probabilities in a unique model of the data distribution. Two technical issues must be addressed for overcoming this obstacle; (a) weighting instances in a given large data set and (b) selecting prototypes from a given large data set. Particularly, the latter is for the modeling from massive data to which the thorough access is not feasible. A method to address these two issues preserves the variety of instance distributions of the data set and provide the basis of the seamless simulations over the multiple scopes. In the first year, we performed a mathematical analysis to derive the required conditions on our targeting method and designed a principle of the method which largely alleviate the obstacle by efficiently sampling the required prototypes with their appropriate weights. In the second year, we implemented the designed principle of the targeting method to an algorithm in computers and evaluated its generic performance preserving the variety of instance distributions of the data set in the selected prototypes. The mathematical analysis, the designed principle, the implemented algorithm and its performance evaluation presented in this report is the first work in worldwide for the seamless and comprehensive probabilistic simulations of the large scale and complex systems.

Introduction: Needs of probabilistic simulations of large scale and complex systems are rapidly increasing. Various situations and behaviors occur under very special conditions with some low probability in such systems because of their huge and complex structures. For example, a particular protein folding structure is known to occur under very limited conditions. A giant typhoon is also induced by special conditions consisting of various meteorological factors.

Accurate probabilistic simulations of all situations and behaviors of the large scale and complex systems are not usually tractable, because the systems and their behaviors are too complex to be well modeled and computed by using our background knowledge. State of the art to overcome this difficulty is to combine a divide and conquer approach and a data-driven approach. The divide and conquer approach limits the scope of the modeling and the simulations to our interested situations and behaviors. For example, we focus the simulations on a protein folding structure by limiting biological and physical conditions to make it occur. For the giant typhoon, we also limit the simulation to some specific rare conditions. The data-driven approach introduces more empirical modeling based on the observed data. For example, probabilistic free energy models of the protein molecule include many empirical parameters. Some parts of a probabilistic typhoon model are also derived by empirically observed data. This framework reduces the difficulty of the system modeling and achieves the sufficient accuracy in the simulations with their tractability.

However, the divide and conquer approach provides the models and the simulations fragmented into the individual scope, and their interpretability are limited. This drawback of the current framework also reduces its applicability to practical problems including analysis, control and management of the systems across their multiple scopes. These difficulties will become more significant in near future, since the scale and the complexity of the systems and the problems are significantly increasing in our modern society. On the other hand, amount of data observed from the systems is rapidly increasing, and this provides “big data”. Since the big data is acquired under various situations and behaviors of the systems, it tends to cover their many scopes. Accordingly, we may obtain better models for the simulations in a data-driven manner. However, such models are not easily derived, since thorough access to all instances in the big data is not tractable.

Aim and Goals

In this study, we aim to overcome the drawbacks of the current probabilistic simulation framework and also to address the issue of the modeling which uses the big data. We target the following research goals under these aims:

(1) We investigate the mathematical principle for the data-driven probabilistic modeling to capture variety of the instance distribution in a given data set for covering multiple scopes of our objective system in a seamless manner.

(2) Based on the investigated mathematical principle, we characterize the required prototype distributions in a subsample data set selected from the given data set to provide such a seamless probabilistic model.

(3) We develop an instance weighing and sub-sampling algorithm for the prototype selection preserving the variety of the instance distribution in a large data set. This enables highly tractable probabilistic modeling of the objective system over its multiple scopes by efficiently extracting instances representing each scope of the system from the big data.

(4) Performance of the developed method for the instance weighing and the prototype selection is evaluated and confirmed through some simulation examples.

In the first year, we theoretically investigated the goals (1) and (2). For the goal (1), we introduced our problem setting at first, and second, we defined some measures to characterize the optimal subsample distributions, and third, we seek some mathematical analysis method for the characterization. For the goal (2), by applying the mathematical analysis method, we characterize the required distributions of instances in a subsample data set of a given original large data set. The instances having the distributions in the subsample data set minimize these measures and preserve the varieties of the distribution of the original data set in a compact fashion by the nature of these measures.

In the final year, we worked on the goals (3) and (4). For the goal (3), an algorithm which efficiently subsamples the instances as prototypes by following the required distribution from the large original data set while minimizing the measures were designed based on a principle of the instance weighting. Then, for the goal (4), the basic performance of the designed algorithm was evaluated through some generic simulations, and its promising performance was confirmed.

Problem Setting: The population distribution of a very large data set is denoted as $f(x)$, where x is a continuous variable on $\mathcal{R}$, and the support of f(x) is limited to [x_{min}, x_{max}]. Let $\mathrm{X} = \{x_i, i = 1, \cdots, N\}$ with very large N be samples independently generated from $f(x)$. An extra auxiliary distribution is defined as $g(x)$, and another subsample set $\mathrm{Y} = \{\tilde{x}_1, \cdots, \tilde{x}_n\}$ （$n << N$） is independently sampled from X according to their sampling weights $\pi(\tilde{x}_i) = 1/\omega(\tilde{x}_i)$ making the population distribution of Y be $g(x)$. This $\omega(\tilde{x}_i)$ is given by

$$\omega(\tilde{x}_i) = \frac{f(\tilde{x}_i)}{g(\tilde{x}_i)}. \tag{1}$$

In Eq. (1), we assume $g(x) \neq 0$ whenever $f(x) \neq 0$.

Let an estimator of $f(x)$: $\hat{f}(x \mid g(x), \mathrm{X}, \mathrm{Y})$ is the following weighted Kernel density estimator using Y [1].

$$\hat{f}(x|g(x), X, Y) = \frac{1}{nh}\sum_{i=1}^{n}\omega(\tilde{x}_i)\mathrm{K}(\frac{x-\tilde{x}_i}{h}) = \frac{1}{nh}\sum_{i=1}^{n}\frac{f(\tilde{x}_i)}{g(\tilde{x}_i)}\mathrm{K}(\frac{x-\tilde{x}_i}{h}). \tag{2}$$

Our problem is to derive an optimum auxiliary distribution $g_{opt}(x)$, which minimizes a given error measure $M(g, X, Y)$ between $f(x)$ and $\hat{f}(x \mid g(x), \mathrm{X}, \mathrm{Y})$ under the data set X independently sampled from $f(x)$. Note that $g_{opt}(x)$ derived here defines the optimum $\omega(\tilde{x}_i)$ by Eq.(1). Our further problem is to design efficient algorithm for subsampling Y from X according to the weight $\pi(\tilde{x}_i) = 1/\omega(\tilde{x}_i)$.

Measures: Our candidate measures $M(g, X, Y)$ chosen for preserving the varieties of the distribution of X in Y are Mean Integrated Square Percentage Error (MISPE) and Alpha Divergence.

MISPE

The simplest candidate measure is the following Mean Integrated Square Percentage Error (MISPE) [2].

$$\begin{aligned} MISPE[f(x)|\hat{f}(x \mid g(x), \mathrm{X}, \mathrm{Y})] &= E\left[\int_{\infty}^{\infty}\left(\frac{f(x) - \hat{f}(x \mid g(x), \mathrm{X}, \mathrm{Y})}{f(x)}\right)^2 dx\right] \\ &= \int_{\infty}^{\infty} E\left[\left(\frac{f(x) - \hat{f}(x \mid g(x), \mathrm{X}, \mathrm{Y})}{f(x)}\right)^2\right] dx \end{aligned} \tag{3}$$

Because $f(x)$ in the denominator weights the error between $f(x)$ and $\hat{f}(x \mid g(x), \mathrm{X}, \mathrm{Y})$, $\hat{f}(x \mid g(x), \mathrm{X}, \mathrm{Y})$ minimizing this measure reflects $f(x)$ more when it is smaller. Hence, the resultant Y and its $\omega(\tilde{x}_i)$ is supposed to capture the varieties of $f(x)$.

Alpha Divergence

A divergence has been considered as a dissimilarity measure. Some of its properties [3] allow us to minimize alpha-divergence to find the best approximating distribution. Firstly, alpha-divergence is zero if $f = \hat{f}$ and positive otherwise, so it satisfies the basic property of an error measure. The property follows from the fact that alpha-divergences are convex with respect to f and $\hat{f}$ [4]. The alpha-divergence being used as an error measure has the variant structure with different selection of parameter α. For example, the case with $\alpha = 0.5$ is known as Hellinger distance. The case with $\alpha = -1$ is considered as a measure similar to the mean integrated squared percentage error (MISPE). The case with $\alpha \to 0$ or 1 is defined as KL-divergence.

Let an alpha-divergence $D^{(\alpha)}(f(x)|\hat{f}(x|g(x),\mathrm{X},\mathrm{Y}))$ be an error measure between $f(x)$ and $\hat{f}(x|g(x),\mathrm{X},\mathrm{Y})$. In our analysis, we design $g(x)$ to minimize this divergence measure. The original definition of the alpha-divergence is

$$D^{(\alpha)}(f(x)|\hat{f}(x|g(x),X,Y))=\frac{\int_{-\infty}^{\infty}(f^{\alpha}(x)\hat{f}^{1-\alpha}(x|g(x),X,Y)-\alpha f(x)+(\alpha-1)\hat{f}(x|g(x),X,Y))dx}{\alpha(\alpha-1)},\alpha\neq 0,1. \quad (4)$$

The case with $\alpha\to 0$ or 1 is defined as KL-divergence, which is given by

$$\lim_{\alpha\to 0}D^{(\alpha)}(f(x)|\hat{f}(x|g(x),X,Y))=KL(\hat{f}(x|g(x),X,Y)|f(x)) \quad (5)$$

$$\lim_{\alpha\to 1}D^{(\alpha)}(f(x)|\hat{f}(x|g(x),X,Y))=KL(f(x)|\hat{f}(x|g(x),X,Y)) \quad (6)$$

Eq. (4) can be reformulated to a simpler expression as

$$D^{(\alpha)}(f(x)|\hat{f}(x|g(x),X,Y))=\frac{1+\int_{-\infty}^{\infty}f^{\alpha}(x)\hat{f}^{1-\alpha}(x|g(x),X,Y)dx}{\alpha(\alpha-1)},\alpha\neq 0,1,\ f(x)\neq 0. \quad (7)$$

In Eq. (7), $f(x)$ causes singularity of alpha divergence when $f(x)=0$. Therefore, without loss of generality, we exclude the area of $f(x)=0$ from the integral and assume $f(x)\neq 0$ in the following analysis. $\hat{f}(x)$ has the similar effect on the alpha divergence, and we assume that $\hat{f}(x)\neq 0$ whenever $f(x)\neq 0$.

Calculus of Constrained Variations and Characterization of Optimal Auxiliary Distributions: For deriving $\omega(\tilde{x}_i)$ of the instances in the subsample data set using Eq.(1), we need to know the optimum auxiliary distribution $g_{opt}(x)$ minimizing $M(g,X,Y)$. Since $g_{opt}(x)$ is a probability density function which integral over the entire R is unity, this optimization problem of $g(x)$ with its integral constraints is generally falls into the following "Calculus of Constrained Variations" [5] and is represented as

$$\begin{aligned}&\text{Minimize}: J=\int_a^b L(x,y(x),y'(x))dx,\\ &\text{Subject to}: I=\int_a^b G(x,y(x),y'(x))dx,\end{aligned} \quad (8)$$

where y(x) is to be optimized, and *I* is a known constant. To solve this problem, $L(x,y(x),y'(x))$ is extended to $\tilde{L}(x,y(x),y'(x),\lambda)$ which includes the Lagrange multiplier.

$$\tilde{L}(x,y(x),y'(x))=L(x,y(x),y'(x))+\lambda\{I-G(x,y(x),y'(x))\}, \quad (9)$$

Where λ is a constant. Then, the optimization problem of $L(x,y(x),y'(x))$ is transformed to the following standard "Calculus of Variations" of the extended $\tilde{L}(x,y(x),y'(x),\lambda)$ without any constraint [6] and [7].

$$\text{Minimize}: J=\int_a^b \tilde{L}(x,y(x),y'(x))dx. \quad (10)$$

The optimium y(x) is known to be the solution of the following partial differential equation.

$$\frac{\partial\tilde{L}}{\partial y(x)}-\frac{\mathrm{d}}{\mathrm{dx}}\left(\frac{\partial\tilde{L}}{\partial y'(x)}\right)=0. \quad (11)$$

In case of MISPE (See Appendix A),

$$L(x,g(x)) = MISPE[f(x) | \hat{f}(x \mid g(x),\mathrm{X},\mathrm{Y})] = \int_{\infty}^{\infty} E\left[\left(\frac{f(x) - \hat{f}(x \mid g(x),\mathrm{X},\mathrm{Y})}{f(x)}\right)^2\right] dx$$

with the integral constraint $\int_{-\infty}^{\infty} g(x)dx = 1$ is to be minimized. Thus, by formulating

$$\tilde{L}(x,g(x)) = L(x,g(x)) + \lambda(1 - \int_{-\infty}^{\infty} g(x)dx) \,, \tag{12}$$

we solve Eq.(11). This results the following optimum g(x).

$$g_{opt}(x) = \sqrt{\frac{\sigma_t}{\lambda nh}} = \begin{cases} \dfrac{1}{x_{max} - x_{min}} & x \in [x_{min}, x_{max}], \\ 0 & otherwise. \end{cases} \tag{13}$$

In case of the alpha-divergence (See Appendix B),

$$L(x,g(x)) = D^{(\alpha)}(f(x) \mid \hat{f}(x \mid g(x),X,Y)) = \begin{cases} \{1 - \int_{-\infty}^{\infty} f^{(\alpha)}(x)\widehat{f}^{1-\alpha}(x \mid g(x),X,Y)dx\}/\{\alpha(\alpha-1)\} & \text{for } \alpha \neq 0,1, \\ KL(f(x) \mid \hat{f}(x \mid g(x),X,Y)) & \text{for } \alpha \to 1, \\ KL(\hat{f}(x \mid g(x),X,Y) \mid f(x)) & \text{for } \alpha \to 0, \end{cases} \tag{14}$$

with the integral constraint $\int_{-\infty}^{\infty} g(x)dx = 1$ is to be minimized. By applying Eq.(9) and Eq.(11), we obtained the following solutions.

(A) Case: $\alpha \neq 0,1,2$

$$g_{opt}(x) = \frac{[\exp(u(x))\delta(x)]^{\frac{1}{\alpha-1}}}{\int_{-\infty}^{\infty} [\exp(u(x))\delta(x)]^{\frac{1}{\alpha-1}} dx}, \tag{15}$$

where $u(x) = \alpha \ln\left(f(x)\right) + \frac{\alpha}{\alpha-1}\ln\left(x - E_g[\tilde{X}]\right) - \frac{\alpha}{h}\frac{K'(0)}{K(0)}\int dx$ and $\delta(x) = \int \frac{\exp\left(-u(x)\right)}{\left(x - E_g[\tilde{X}]\right)} dx$.

(B) Case: $\alpha = 2$

$$g_{opt}(x) = \frac{[\exp(u(x))\delta(x)]}{\int_{-\infty}^{\infty} [\exp(u(x))\delta(x)] dx}, \tag{16}$$

where $u(x) = 2\ln\left(f(x)\right) + 2\ln\left(x - E_g[\tilde{X}]\right) - \frac{2}{h}\frac{K'(0)}{K(0)}\int dx$ and $\delta(x) = \int \frac{\exp\left(-u(x)\right)}{\left(x - E_g[\tilde{X}]\right)} dx$.

(C) Case: $\alpha \to 1$ (The solution of $\alpha \to 0$ is similarly obtained.)

$$g_{opt}(x) = \frac{[\exp(u(x))\delta(x)]^{-1}}{\int_{-\infty}^{\infty} [\exp(u(x))\delta(x)]^{-1} dx}, \tag{17}$$

where $u(x) = \ln\left(x - E_g[\tilde{X}]\right) + C$ and $\delta(x) = \int \frac{\exp\left(-u(x)\right)}{\left(x - E_g[\tilde{X}]\right)} \frac{1}{f(x)} dx$.

Algorithm to Select Prototypes: The optimal auxiliary distributions $g_{opt}(x)$ given by the alpha divergence depends on f(X), and thus it is not easily computed because f(X) is unknown. In contrast, $g_{opt}(x)$ based on the MISPE measure is uniquly given as the uniform distribution shown in Eq.(13). Accordingly, we focus on this simpler case in our further study. The sampling algorithm of Y from X must be designed to sample each instance from X which follows f(x) over $\mathcal{R}$ with a uniform probability density $g_{opt}(x)$ over $\mathcal{R}$ by applying the sampling weight $\pi(\tilde{x}_i) = 1/\omega(\tilde{x}_i)$ designated by Eq.(1), so that the samples in Y drawn from X are uniformly distributed everywhere in the support of f(x).

An issues to design this algorithm is the lackness of the infornation on true f(x) which is needed to directly compute correct $\pi(\tilde{x}_i) = 1/\omega(\tilde{x}_i)$. We introduce an iterative approximation alogoritthm named "Wang-Landau algorithm" [8] to derive the subsample set Y following the uniform distribution $g_{opt}(\tilde{x}_i)$. The key idea of this algorithm is to use a histogram of the prototype set Y over $[x_{min}, x_{max}]$ and gradually modify $\pi_i(\tilde{x}_i) = 1/\omega(\tilde{x}_i)$ at each bin of the histogram in iterative computations. Let $X_j \subset [x_{min}, x_{max}]$ $(j = 1, ..., B)$ be a bin of the histogram of the subsample data set Y. The histogram consists of B bins partitioning $[x_{min}, x_{max}]$. We define a weight of each bin as $\hat{\pi}(X_j)$, and let the weight of each $\tilde{x}_i$ in Y be $\hat{\pi}(\tilde{x}_i) = \hat{\pi}(X_j)$ subject to $\tilde{x}_i \in X_j$.

Initially, we give a tentative arbitrary weight value to $\hat{\pi}(X_j)$ of each bin. This is equivalent to assume some arbitrary density $\hat{f}(\tilde{x}_i)$ for $\tilde{x}_i \in X_j$ as

$$\hat{f}(\tilde{x}_i) = \frac{g_{opt}(\tilde{x}_i)}{\hat{\pi}(\tilde{x}_i)} = \frac{g_{opt}(\tilde{x}_i)}{\hat{\pi}(X_j)} = \frac{c}{\hat{\pi}(X_j)} \text{ where } c = \frac{1}{x_{max} - x_{min}}. \tag{18}$$

After each trial to draw $\tilde{x}$ from X, we select $\tilde{x}$ into Y as $\tilde{x}_i$ in probability which is proportional to the weight $\hat{\pi}(\tilde{x}) = \hat{\pi}(X_j)$ subject to $\tilde{x} \in X_j$ as

$$p(\tilde{x}) = \frac{\hat{\pi}(X_j)h(X_j)}{\sum_{j=1}^{B} h(X_j)\hat{\pi}(X_j)}, \tag{19}$$

where $h(X_j)$ is a frequency of the instances included in a bin X_j.

Subsequently, $\hat{\pi}(X_j)$ is lessen by multiplying a constant factor 0 <*F* < 1 to it. This procedure reduces the weight of the instance $\tilde{x}$ belonging to a bin having a large frequency in the histogram of Y, *i.e.*, the instance $\tilde{x}$ frequently selected into Y from X. The instances more frequently sampled in Y become less sampled by their lowered weights, and the instances less frequently sampled in Y become more sampled by their relatively lifted weights. Accordingly, this procedure has an effcet to flatten the histogram's shape.

These instance selection and weight update are repeated and then posed when the histogram becomes almost flat. At this point, Y and the frequency $h(X_j)$ of its all histogram bins are reset to empty while keeping their latest weights $\hat{\pi}(X_j)$. In addition, the factor *F* is changed by $F \leftarrow \sqrt{F}$ to make it closer to 1. This updated factor *F* enables finer tuning of the weights to more precisely flatten the histogram. Then, the instance selection into Y, the weight update and the factor update are further repeated until we obtain Y having a sufficiently flattened histogram, *i.e.*, a uniform distribution.

This Wang-Landau algorithm indirectly reflects the population distribution of X, f(X), to the weights of the histogram bins, *i.e.*, the weights of $\hat{\pi}(\tilde{x}_i) = \hat{\pi}(X_j)$ subject to $\tilde{x}_i \in X_j$, by iteratively modify them while computing the frequencies of the bins, and achieves the sufficiently uniform $g_{opt}(\tilde{x}_i)$ of Y. A strong advantage of this algorithm is that we can efficiently derive the prototype set Y having a sufficiently uniform distribution without assessing entire distribution of X, *i.e.*, accessing the entire data set X which can be very huge.

The following is a pseudo-code of this algorithm. Here, we use logarithmic weights $LP(X_j) = \log \hat{\pi}(X_j)$ in place of $\hat{\pi}(X_j)$, and further we define logarithmic factor $LF = -\log F$ (i.e., with a minus sign) in place of *F*. These are because the amplitudes of the weights can vary over many orders of magnitude.

Wang-Landau algorithm:

```
1. Initialize $LP(X_j)$ and LF; set other parameters.
   – Set $LP(X_j)$ = 0 for j = 1, . . . , B.
   – Set LF > 0 (e.g., LF = −log(1/e) = 1).
   – Set the maximum number of iterations $K_{max}$ (e.g., $K_{max}$ = 18).
   – Set the counter of iterations K to 0.
2. Initialize a prototype set Y and its histogram H.
   – If K > $K_{max}$, end.
   – Make the prototype set Y empty.
   – Set $h(X_j) = 0$ for j = 1, . . . , B.
3. Selection of an instance from X into Y.
   – Randomly sample $\tilde{x}$ from X and select it into Y in probability proportional to
      $\pi(X_j) = \exp(LP(X_j))$ subject to $\tilde{x} \in X_j$.
4. Modify $LP(X_j)$ and update the histogram H.
   – $LP(X_j) \leftarrow LP(X_j) - LF$ subject to $\tilde{x} \in X_j$.
   – $h(X_j) \leftarrow h(X_j) + 1$ subject to $\tilde{x} \in X_j$.
5. Check whether H is "sufficiently flat."
   – If so, $LF \leftarrow LF/2$, K = K + 1 and go to Step 2.
   – Otherwise go to Step 3.
```

The application of this Wang-Landau algorithm is not limited to derive the prototype set Y having the uniform distribution. By changing the condition at the step 5; Check whether H is "sufficiently flat" to any target distribution, this algorithm can derive Y having the distribution. Accordingly, this is also applicable to Y given under the alpha-divergence measure, if Eq.(15), (16) and (17) are represented by some analytical forms or numerically solved.

Experiment and Results: We applied the Wang-Landau algorithm with the MISPE measure to a virtual big data set X having the following Weibull distribution.

$$f(x;\lambda,k) = \begin{cases} \dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

where λ=1.0 and k=1.5. This distribution is known as a typical example distribution having a significant long tail, i.e., a large portion of the probability belongs to a wide range of the probability variable with very low probability. Therefore, this distribution produces a data set X which large part consists of rare instances. The big data set X was not generated at the step 3 of the pseudo-code in reality. Rather, the direct draw of each instance $\tilde{x}$ from $f(x;\lambda,k)$ was performed. This is equivalent to have an infinite data set X. We limited the range of $\tilde{x}$ to a finite interval [0,5] for avoiding computational divergence. The initial values of the logarithmic weights $LP(X_j)=\log\hat{\pi}(X_j)$ and the logarithmic factor $LF=-\log F$ were set at 0 and 1, respectively. $K_{max}=18$ was used for the number of the factor updates and the histogram reconstructions.

We assessed the performance of our proposed method by checking if such rare instances are efficiently sampled over the long tail of the distribution. Figure 1 shows the histogram of the samples in the prototype set Y together with the plot of the Weibull distribution having λ=1.0 and k=1.5. This clearly shows that Y almost uniformly includes variety of samples over the entire range of $\tilde{x}$, and the rare prototypes in the range [4,5] having almost negligible probability *f(x)* form a significant portion of Y. The similar prototype set can be obtained by accessing the entire data set X and selectively acquiring the rare prototypes. However, the efficiency of this approach depends on the sizes of X and Y. For example, if the size of given X is 160,000 and the required size of Y is 100, its efficiency to obtatin an prototype in Y is $100/160{,}000=6.25\times10^{-4}$. This efficiency is comparable with that of our proposed prototype sampling, 6.48×10^{-4}, which is the ratio of the sample population acquired in Y over the total number of the draws from X, *i.e.*, $f(x;\lambda,k)$. On the other hand, if the size of given X is 16,000,000 and the required size of Y is 100, its efficiency $100/16{,}000{,}000=6.25\times10^{-6}$ is far worse than the efficiency of our method. In other words, our approach is more efficient than the thorough access to X to acquire 100 prototypes in Y, if the size of X is more than 160,000. This comparison demonstrates the advantage of our proposed approach for the prototype sampling from a big data set.

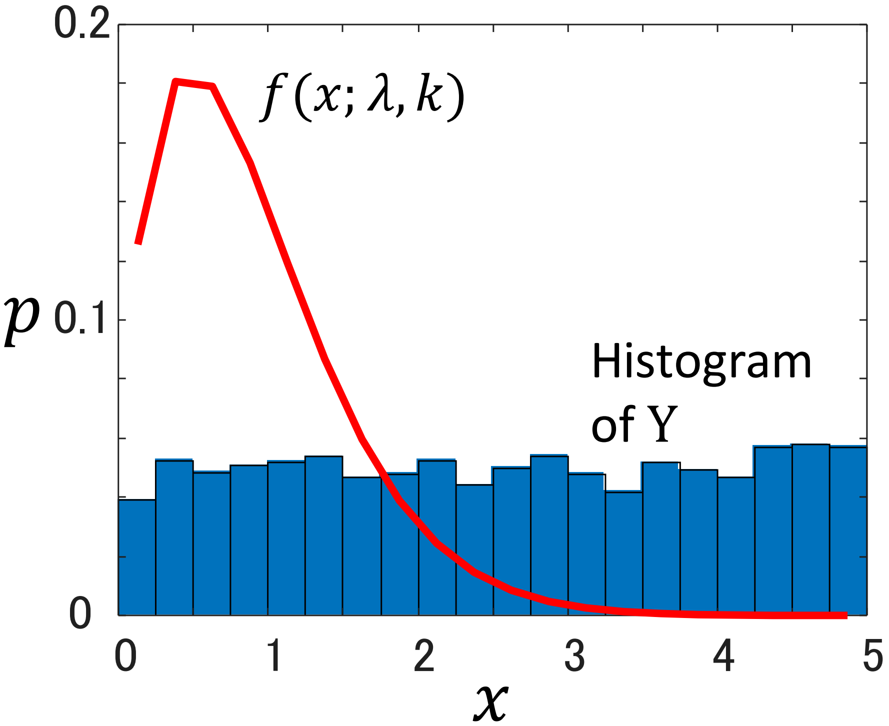

Figure 1. Weibull distribution f(x;λ,k) having λ=1.0 and k=1.5 and histogram of Y.

Next, we compared MISPE and MISE between our approach and the standard random sampling. In our approach, the probability density f(x;λ,k) of the Weibull distribution is estimated by using Eq.(1) as

$$\hat{f}(\tilde{x};\lambda,k) = g_h(X_j)\hat{\omega}(X_j) = \frac{g_h(X_j)}{\hat{\pi}(X_j)} \text{ subject to } \tilde{x} \in X_j,$$

where $g_h(X_j)$ and $\hat{\pi}(X_j)$ are the normalized frequency and the weight of the histogram bin X_j, respectively. In the standard random sampling, $\hat{f}(\tilde{x};\lambda,k)$ is simply computed by the histogram of the randomly sampled prototype set Y. Once $\hat{f}(\tilde{x};\lambda,k)$ is estimated in the both methods, their MIPSE are computed by Eq.(3), and their MISE are computed by the standard mean integrated square error between $\hat{f}(\tilde{x};\lambda,k)$ and f(x;λ,k). Figure 2 show the comparison of their MIPSE. By repeating the factor updates and the histogram reconstruction in our approach, its MIPSE got far smaller than that of the standard random sampling. Figure 3 depicts the comparison of their MISE. Though the MISE of our approach reduces along the iterations, it does not reach the level of the MISE of the standard method. These results are consistent with our theoretical analysis presented earlier.

Discussion and Conclusion: In this study, we have theoretically investigated principles of data-driven probabilistic modeling to capture variety of the instance distribution in a given data set for covering multiple scopes of our objective system in a seamless manner. We analyzed two error measures, MISPE and alpha-divergence, to derive the required distributions of instances in a subsample data set of a given original large data set, and we presented a mathematical principle to derive the optimal distributions of the instances in the subsample data set required to minimize these measures. The instances in these subsample data sets are expected to minimize these measures and preserve the varieties of the distribution of the original data set in compact fashions by the nature of these measures.

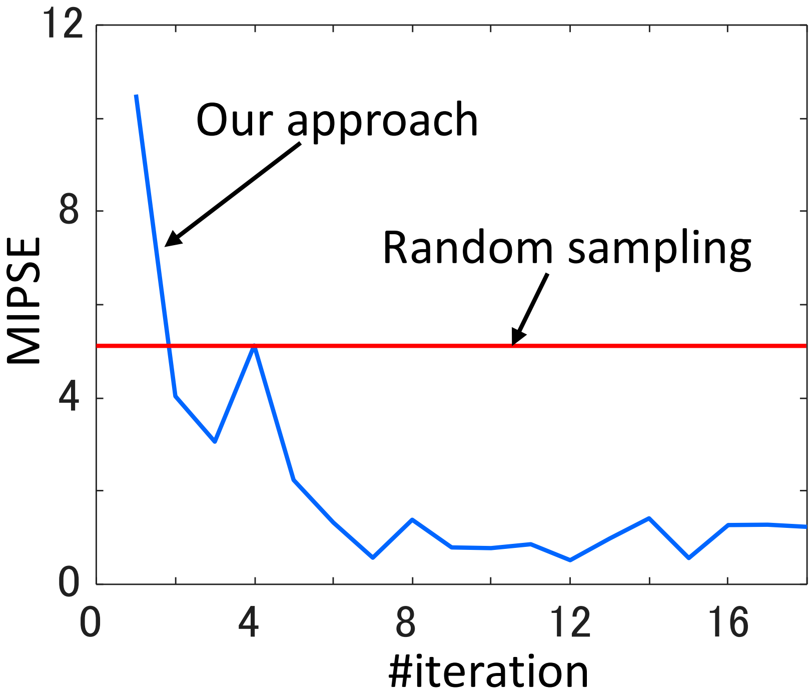

Figure 2. Change of MIPSE over the iterations of the factor updates and the histogram reconstructions in our approach and comparison with MIPSE provided by the standard random sampling.

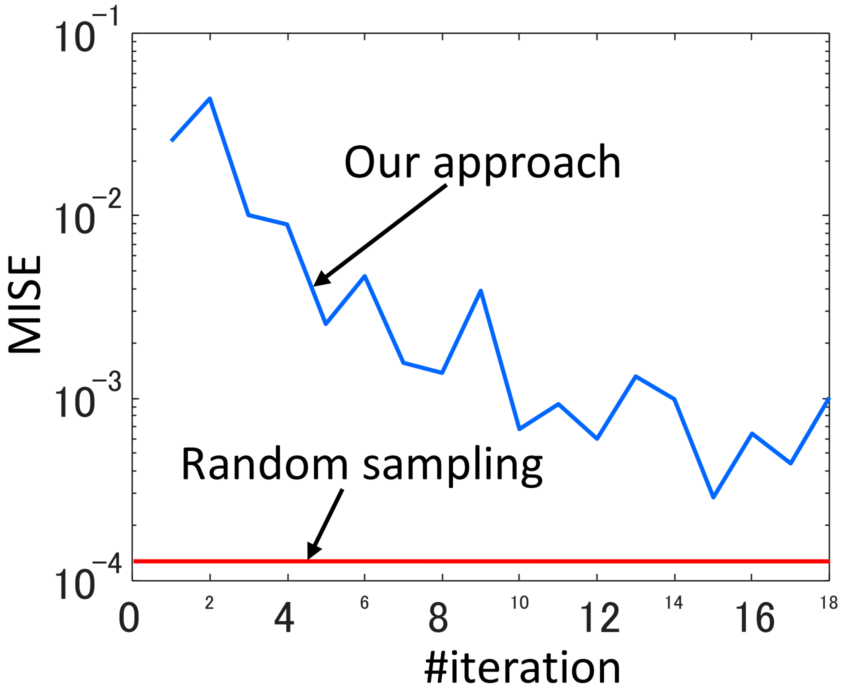

Figure 3. Change of MISE over the iterations of the factor updates and the histogram reconstructions in our approach and comparison with MISE provided by the standard random sampling.

We further investigated an instance weighing and sub-sampling algorithm named “Wang-Landau Algorithm” for the prototype selection preserving the variety of the instance distribution in a large data set. This enables highly tractable probabilistic modeling of the objective system over its multiple scopes by efficiently extracting instances representing each scope of the system from the big data.

Finally, we evaluated the performance of the developed method for the instance weighing and the prototype selection through some simulation examples. The achievement of the prototype selection preserving the variety of the instance distribution in a large data set has been demonstrated with its efficiency. The approach can be used for the efficient prototype selection from the big data set. Moreover, the superior accuracy of the proposed prototype selection method in terms of the MIPSE measure has been confirmed.

An issue remained for future work is the implementation of the prototype selection algorithm using the alpha-divergence measure. Since the distributions of the prototypes meeting with this measure have not been solved in analytical manner yet, more extensive studies to clarify the characteritics of the solutions and to develop some numerical method for their implemenation to the computational algorithm are needed.

Appendix A: Analysis on Optimal g(x) for MISPE

We design $g(x)$ to minimize the error measure $MISPE\left[\hat{f}(x|g(x),S,D)\right]$ between $f(x)$ and $\hat{f}(x|g(x),S,D)$. Assume the support of $f(x)$ is $[x\_\min, x\_\max]$ where $f(x)=0$ in its outside. Then $MISPE\left[\hat{f}(x|g(x),S,D)\right]$ is expressed by

$$\begin{aligned}
&MISPE\left[\hat{f}(x|g(x),S,D)\right] \\
&=E_g\left[\int_{x\_\min}^{x\_\max}\left(\frac{f(x)-\hat{f}(x|g(x),S,D)}{f(x)}\right)^2 \mathrm{d}x\right] \\
&=\int_{x\_\min}^{x\_\max} E_g\left[\left(\frac{f(x)-\hat{f}(x|g(x),S,D)}{f(x)}\right)^2\right]\mathrm{d}x \\
&=\int_{x\_\min}^{x\_\max}\frac{1}{f^2(x)}E_g\left[\left(f(x)-\hat{f}(x|g(x),S,D)\right)^2\right]\mathrm{d}x \\
&=\int_{x\_\min}^{x\_\max} L\,(x,g(x))\mathrm{d}x,
\end{aligned} \tag{20}$$

where $L\,(x,g(x))=\frac{1}{f^2(x)}E_g\left[\left(f(x)-\hat{f}(x|g(x),S,D)\right)^2\right]$. According to Eq. (9), $L^{\sim}(x,g(x))$ is set as $L^{\sim}(x,g(x))=L\,(x,g(x))+\lambda g(x)$. By the constraint $\int_{x\_\min}^{x\_\max} g(x)dx=1$, $MISPE\left[\hat{f}(x|g(x),S,D)\right]$ is written by

$$MISPE\left[\hat{f}(x|g(x),S,D)\right]=\int_{x\_\min}^{x\_\max} L^{\sim}(x,g(x))\mathrm{d}x, \quad s.t. \quad \int_{x\_\min}^{x\_\max} g(x)dx=1. \tag{21}$$

Our target is to obtain the optimal $g(x)$ by minimizing $MISPE\left[\hat{f}(x|g(x),S,D)\right]$ as

$$g_{opt}(x)=\arg\min_{g(x)} MISPE\left[\hat{f}(x|g(x),S,D)\right]=\arg\min_{g(x)}\int_{x\_\min}^{x\_\max} L^{\sim}(x,g(x))\mathrm{d}x, \quad s.t. \quad \int_{x\_\min}^{x\_\max} g(x)dx=1, \tag{22}$$

where $g_{opt}(x)$ is the optimal solution. To achieve this goal, the calculus of variation is applied. If we want to obtain the optimal $g(x)$ by minimizing $MISPE\left[\hat{f}(x|g(x),S,D)\right]$, then this problem is called calculus of variation with integral constraint. In our problem, y(x)=g(x) and there is no g'(x). Thus Eq. (11) is reduced to $\partial\tilde{L}/\partial y(x)=0$ which is Eq. (23). The optimal g(x) should be obtained by solving the following equations by

$$\begin{cases}
\dfrac{\partial L^{\sim}(x,g(x))}{\partial g(x)}=\dfrac{\partial L\,(x,g(x))}{\partial g(x)}+\lambda=\dfrac{1}{f^2(x)}\dfrac{\partial E_g\left[\left(f(x)-\hat{f}(x|g(x),S,D)\right)^2\right]}{\partial g(x)}+\lambda=0 \\
\displaystyle\int_{x\_\min}^{x\_\max} g(x)dx=1.
\end{cases} \tag{23}$$

Eq. (23) is written in another form by

$$\begin{cases} \dfrac{\partial E_g\left[\left(f(x)-\hat{f}(x\,|\,g(x),S,D)\right)^2\right]}{\partial g(x)} = -\lambda f^2(x) \\ \int_{x\_\min}^{x\_\max} g(x)dx = 1. \end{cases} \tag{24}$$

$\dfrac{\partial E_g\left[\left(f(x)-\hat{f}(x\,|\,g(x),S,D)\right)^2\right]}{\partial g(x)}$ is written by

$$\begin{aligned} &\frac{\partial E_g\left[\left(f(x)-\hat{f}(x\,|\,g(x),S,D)\right)^2\right]}{\partial g(x)} \\ &= \frac{\partial E_g\left[\hat{f}^2(x\,|\,g(x),S,D)\right]}{\partial g(x)} - 2f(x)\frac{\partial E_g\left[\hat{f}(x\,|\,g(x),S,D)\right]}{\partial g(x)}. \end{aligned} \tag{25}$$

By substituting Eq. (25) to Eq. (24), we obtain

$$\begin{cases} \dfrac{\partial E_g\left[\hat{f}^2(x\,|\,g(x),S,D)\right]}{\partial g(x)} - 2f(x)\dfrac{\partial E_g\left[\hat{f}(x\,|\,g(x),S,D)\right]}{\partial g(x)} = -\lambda f^2(x) \\ \int_{x\_\min}^{x\_\max} g(x)dx = 1. \end{cases} \tag{26}$$

Here, kernel estimator is employed with the points in the weighted sample set to estimate $\hat{f}(x\,|\,g(x),S,D)$ as

$$f(x\,|\,g(x),S,D) = \frac{1}{n}\sum_{i=1}^{n}\frac{f(\tilde{X}_i)}{g(\tilde{X}_i)}K\left(\frac{x-\tilde{X}_i}{h}\right), \quad \tilde{X}_i \leftarrow g(x) \tag{27}$$

where $\tilde{X}_i, i=1,\cdots,n$ are samples from the weighted sample set D . The expectation of $\hat{f}(x\,|\,g(x),S,D)$ is

$$\begin{aligned} E_g\left[\hat{f}(x\,|\,g(x),S,D)\right] &= \frac{1}{h}\sum_{i=1}^{n}E_g\left[\frac{f(\tilde{X}_i)}{g(\tilde{X}_i)}K\left(\frac{x-\tilde{X}_i}{h}\right)\right] = \frac{n}{h}\int_{x\_\min}^{x\_\max}\frac{f(y)}{g(y)}K\left(\frac{x-y}{h}\right)g(y)dy \\ &= \frac{n}{h}\int_{x\_\min}^{x\_\max} f(y)K\left(\frac{x-y}{h}\right)dy. \end{aligned} \tag{28}$$

So we obtain $\dfrac{\partial E_g\left[\hat{f}(x\,|\,g(x),S,D)\right]}{\partial g(x)} = 0$, because the expectation of $\hat{f}(x\,|\,g(x),S,D)$ does not contain $g(x)$ term.

$$
\begin{aligned}
&E_g\left[\hat{f}^2(x\,|\,g(x),S,D)\right]=\frac{1}{h^2}E_g\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{f(\tilde{X}_i)}{g(\tilde{X}_i)}K(\frac{x-\tilde{X}_i}{h})\frac{f(\tilde{X}_j)}{g(\tilde{X}_j)}K(\frac{x-\tilde{X}_j}{h})\right)\right]\\
&=\frac{n}{h^2}E_g\left[\left(\frac{f(\tilde{X}_i)}{g(\tilde{X}_i)}K(\frac{x-\tilde{X}_i}{h})\right)^2\right]+\frac{n(n-1)}{h^2}\left(E_g\left[\frac{f(\tilde{X}_i)}{g(\tilde{X}_i)}K(\frac{x-\tilde{X}_i}{h})\right]\right)^2\\
&=\frac{n}{h^2}\int_{x\_\min}^{x\_\max}\left(\frac{f(y)}{g(y)}\right)^2K^2\left(\frac{x-y}{h}\right)g(y)dy+\frac{n(n-1)}{h^2}\left(\int_{x\_\min}^{x\_\max}\frac{f(y)}{g(y)}K\left(\frac{x-y}{h}\right)g(y)dy\right)^2\\
&=\frac{n}{h^2}\int_{x\_\min}^{x\_\max}\frac{f^2(y)}{g(y)}K^2\left(\frac{x-y}{h}\right)dy+\frac{n(n-1)}{h^2}\left(\int_{x\_\min}^{x\_\max}f(y)K\left(\frac{x-y}{h}\right)dy\right)^2\\
&\overset{t=\frac{x-y}{h}}{=}\frac{n}{h}\int_{x\_\min}^{x\_\max}\frac{f^2(x-ht)}{g(x-ht)}K^2(t)dt+\frac{n(n-1)}{h^2}\left\{\int_{x\_\min}^{x\_\max}f(y)K\left(\frac{x-y}{h}\right)dy\right\}^2\\
&\approx\frac{n}{h}\int_{x\_\min}^{x\_\max}K^2(t)\left[\frac{f^2(x)}{g(x)}-ht\left(\frac{f^2(x)}{g(x)}\right)'\right]dt+\frac{n(n-1)}{h^2}\left\{\int_{x\_\min}^{x\_\max}f(y)K\left(\frac{x-y}{h}\right)dy\right\}^2\\
&=\frac{n}{h}\left[\frac{f^2(x)}{g(x)}\int_{x\_\min}^{x\_\max}K^2(t)dt\right]+\frac{n(n-1)}{h^2}\left\{\int_{x\_\min}^{x\_\max}f(y)K\left(\frac{x-y}{h}\right)dy\right\}^2\\
&=\frac{n\sigma_t}{h}\frac{f^2(x)}{g(x)}+\frac{n(n-1)}{h^2}\left\{\int_{x\_\min}^{x\_\max}f(y)K\left(\frac{x-y}{h}\right)dy\right\}^2, \qquad (29)
\end{aligned}
$$

where $K(t)$ is assumed to be symmetric and $\sigma_t=\int_{x\_\min}^{x\_\max}K^2(t)dt$. Then we obtain $\frac{\partial E_g\left[\hat{f}^2(x\,|\,g(x),S,D)\right]}{\partial g(x)}$ by

$$\frac{\partial E_g\left[\hat{f}^2(x\,|\,g(x),S,D)\right]}{\partial g(x)}=-\frac{n\sigma_t}{h}\frac{f^2(x)}{g^2(x)}. \qquad (30)$$

Then $\frac{\partial E_g\left[\left(f(x)-\hat{f}(x\,|\,g(x),S,D)\right)^2\right]}{\partial g(x)}$ in Eq. (25) is written by

$$
\begin{aligned}
&\frac{\partial E_g\left[\left(f(x)-\hat{f}(x\,|\,g(x),S,D)\right)^2\right]}{\partial g(x)}\\
&=\frac{\partial E_g\left[\hat{f}^2(x\,|\,g(x),S,D)\right]}{\partial g(x)}-2f(x)\frac{\partial E_g\left[\hat{f}(x\,|\,g(x),S,D)\right]}{\partial g(x)} \qquad (31)\\
&=-\frac{n\sigma_t}{h}\frac{f^2(x)}{g^2(x)}=-\lambda f^2(x).
\end{aligned}
$$

The expression of $g(x)$ is obtained by

$$g(x)=\sqrt{\frac{n\sigma_t}{\lambda h}}. \qquad (32)$$

The following the constraint should be used to fix the constant λ in Eq. (32)

$$\int_{x\_\min}^{x\_\max} g(x)dx = 1. \tag{33}$$

By substituting Eq. (32) into $\int_{x\_\min}^{x\_\max} g(x)dx = 1$, we obtain

$$\int_{x\_\min}^{x\_\max} g(x)dx = \sqrt{\frac{n\sigma_t}{\lambda h}} \int_{x\_\min}^{x\_\max} dx = 1. \tag{34}$$

Since we obtain λ by

$$\begin{aligned} \lambda &= \frac{n\sigma_t}{h}\left(\int_{x\_\min}^{x\_\max} dx\right)^2 \\ &= \frac{n\sigma_t}{h}\left(x\_\max - x\_\min\right)^2. \end{aligned} \tag{35}$$

By substituting Eq. (35) into Eq. (33), we obtain $g(x)$ by

$$g(x) = \sqrt{\frac{n\sigma_t}{\lambda h}} = \frac{1}{x\_\max - x\_\min}. \tag{36}$$

Appendix B: Analysis on Optimal g(x) for α -divergence

Let alpha-divergence $D^{(\alpha)}(f(x) \mid \hat{f}(x \mid g(x), S, D))$ be an error measure M (g, S, D) between $f(x)$ and $\hat{f}(x|g(x),S,D)$. In our analysis, we want to design $g(x)$ to minimize this divergence measure. The original definition of the alpha-divergence is

$$D^{(\alpha)}(f(x) \mid \hat{f}(x \mid g(x), X, Y)) = \frac{\int_{-\infty}^{\infty}(f^{\alpha}(x)\hat{f}^{1-\alpha}(x \mid g(x), X, Y) - \alpha f(x) + (\alpha - 1)\hat{f}(x \mid g(x), X, Y))dx}{\alpha(\alpha - 1)}, \alpha \neq 0,1. \tag{37}$$

The case with $\alpha \to 0$ or 1 is defined as KL-divergence, which is given by

$$\lim_{\alpha \to 0} D^{(\alpha)}(f(x) \mid \hat{f}(x \mid g(x), X, Y)) = KL(\hat{f}(x \mid g(x), X, Y) \mid f(x)), \tag{38}$$

$$\lim_{\alpha \to 1} D^{(\alpha)}(f(x) \mid \hat{f}(x \mid g(x), X, Y)) = KL(f(x) \mid \hat{f}(x \mid g(x), X, Y)). \tag{39}$$

Eq. (37) can be reformulated to a simple expression by

$$D^{(\alpha)}(f(x) \mid \hat{f}(x \mid g(x), X, Y)) = \frac{1 + \int_{-\infty}^{\infty} f^{\alpha}(x)\hat{f}^{1-\alpha}(x \mid g(x), X, Y)dx}{\alpha(\alpha - 1)}, \alpha \neq 0,1,\ f(x) \neq 0. \tag{40}$$

In Eq. (40), $f(x)$ causes singularity of alpha divergence when $f(x) = 0$. Therefore, without loss of generality, we exclude the area of $f(x) = 0$ from the integral and assume $f(x) \neq 0$ in the following analysis. $\hat{f}(x)$ has the similar effect on the alpha divergence, and we assume that $\hat{f}(x) \neq 0$ whenever $f(x) \neq 0$.

1. **For the case $\alpha \neq 0,1$**

Since $\hat{f}(x|g(x),S,D)$ is a function having a probability distribution according to the statistics of D sampled from S and D follows g(x), we take the expectation of alpha-divergence to measure the difference between $f(x)$ and $\hat{f}(x|g(x),S,D)$ over $g(x)$ as

$$\begin{aligned} &E_g[D^{(\alpha)}(f(x)\,|\,\hat{f}(x\,|\,g(x),S,D))] \\ &=\frac{1}{\alpha(1-\beta)}\left(1-E_g\left[\int_{-\infty}^{\infty}\left(f^{\alpha}(x)\hat{f}^{1-\alpha}(x\,|\,g(x),S,D)\right)dx\right]\right) \\ &=\frac{1}{\alpha(1-\alpha)}\left(1-\int_{-\infty}^{\infty}\left(f^{\alpha}(x)\mathrm{E}_{\mathrm{g}}\left[\hat{f}^{1-\alpha}(x\,|\,g(x),S,D)\right]\right)dx\right). \end{aligned} \tag{41}$$

Our research target is to obtain the optimal $g(x)$ to minimize $E_g[D^{(\alpha)}(f(x)\,|\,\hat{f}(x\,|\,g(x),X,Y))]$ by

$$g_{opt}(x)=\arg\min_{g(x)} E_g[D^{(\alpha)}(f(x)\,|\,\hat{f}(x\,|\,g(x),S,\mathrm{D}))] \quad s.t.\int_{-\infty}^{\infty}g(x)dx=1 \tag{42}$$

By substituting Eq. (41) into Eq. (42), the following expression is derived

$$g_{opt}(x)=\arg\min_{g(x)}\int_{-\infty}^{\infty}\left(L\ (x,g(x))\right)dx \quad s.t.\int_{-\infty}^{\infty}g(x)dx=1, \tag{43}$$

where $L\ (x,g(x))=-\frac{1}{\alpha(1-\alpha)}f^{\alpha}(x)E_g\left[\hat{f}^{1-\alpha}(x|g(x),S,D)\right]$. According to Eq. (9), $\tilde{L}(x,g(x))$ is set as $\tilde{L}(x,g(x))=L\ (x,g(x))+\lambda g(x)$. Correspondingly, Eq. (10) is written by

$$\begin{aligned} &\frac{\partial \tilde{L}(x,g(x))}{\partial g(x)}=\frac{\partial L\ (x,g(x))}{\partial g(x)}+\lambda \\ &=-\frac{1}{\alpha(1-\alpha)}f^{\alpha}(x)\frac{\partial E_g\left[\hat{f}^{1-\alpha}(x|g(x),S,D)\right]}{\partial g(x)}+\lambda=0 \quad s.t.\int_{-\infty}^{\infty}g(x)dx=1. \end{aligned} \tag{44}$$

Then $\frac{\partial E_g\left[\hat{f}^{1-\alpha}(x|g(x),S,D)\right]}{\partial g(x)}$ is written by

$$\frac{\partial E_g\left[\hat{f}^{1-\alpha}(x|g(x),S,D)\right]}{\partial g(x)}=\alpha(1-\alpha)\lambda f^{-\alpha}(x). \tag{45}$$

Here, we consider to employ kernel estimator to estimate $\hat{f}(x|g(x),S,D)$ by

$$\hat{f}(x|g(x),S,D)=\frac{1}{h}\sum_{i=1}^{n}w(\tilde{X}_i)K(\frac{x-\tilde{X}_i}{h}), \tag{46}$$

where $\tilde{X}_i, i=1,\cdots,n$ are samples in the weighted sample set D and $g(x)$ should be $g(x)\neq 0$ whenever $f(x)\neq 0$. w(x) is unknown since f(x) is not given in practice. However, we assume that w(x) is given by Eq.(11) from f(x) and g(x) in our current problem setting as noted earlier in the general problem description. Now, we want to calculate $E_g\left[\hat{f}^{1-\alpha}(x|g(x),S,D)\right]$, which is written by

$$\begin{aligned}
&E_g\left[\hat{f}^{1-\alpha}(x\,|\,g(x),S,D)\right]\\
&=E_g\left[\left(\frac{1}{h}\sum_{i=1}^{n}w(\tilde{X}_i)K(\frac{x-\tilde{X}_i}{h})\right)^{1-\alpha}\right]\\
&=E_g\left[\left(\frac{1}{h}\sum_{i=1}^{n}\frac{f(\tilde{X}_i)}{g(\tilde{X}_i)}K(\frac{x-\tilde{X}_i}{h})\right)^{1-\alpha}\right]\\
&=\frac{1}{h^{1-\alpha}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}\frac{f(y_i)}{g(y_i)}K(\frac{x-y_i}{h})\right)^{1-\alpha} g(y_1,\cdots,y_n)dy_1\cdots dy_n.
\end{aligned}\tag{47}$$

Since $X_1,\cdots,X_n$ are i.i.d. sampled from $g(x)$, we take the form $g(y_1,\cdots,y_n)=g(y_1)\cdots g(y_n)$. Eq. (47) is written by

$$\begin{aligned}
&E_g\left[\hat{f}^{1-\alpha}(x\,|\,g(x),S,D)\right]\\
&=\frac{1}{h^{1-\alpha}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}\frac{f(y_i)}{g(y_i)}K(\frac{x-y_i}{h})\right)^{1-\alpha}\prod_{j=1}^{n}g(y_j)dy_1\cdots dy_n\\
&\overset{t_i=\frac{x-y_i}{h}}{=}\frac{h^n}{h^{1-\alpha}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}\frac{f(x-ht_i)}{g(x-ht_i)}K(t_i)\right)^{1-\alpha}\prod_{j=1}^{n}g(x-ht_j)dt_1\cdots dt_n.
\end{aligned}\tag{48}$$

Set $s(x,\boldsymbol{t})=s(x,\left[t_1,\cdots,t_n\right]^T)=\left(\sum_{i=1}^{n}\frac{f(x-ht_i)}{g(x-ht_i)}K(t_i)\right)^{1-\alpha}$. By applying Taylor's expansion for multivariate $t_1,\cdots,t_n$ to $s(x,\boldsymbol{t})$. Then, $s(x,\boldsymbol{t})$ is written by

$$\begin{aligned}
s(x,\boldsymbol{t})&=\sum_{j=0}^{\infty}\sum_{|\beta|=j}\frac{\nabla^{\beta}s(x,\boldsymbol{0})}{\beta!}(\boldsymbol{t}-\boldsymbol{0})^{\beta}\\
&=\sum_{j=0}^{\infty}\sum_{|\beta|=j}\frac{1}{\beta!}\left.\frac{\partial^{|\beta|}s(x,\boldsymbol{t})}{\partial t_1^{\beta_1}\cdots\partial t_n^{\beta_n}}\right|_{\boldsymbol{t}=\boldsymbol{0}}(\boldsymbol{t}-\boldsymbol{0})^{\beta}\\
&\approx\sum_{j=0}^{1}\sum_{|\beta|=j}\frac{1}{\beta!}\left.\frac{\partial^{|\beta|}s(x,\boldsymbol{t})}{\partial t_1^{\beta_1}\cdots\partial t_n^{\beta_n}}\right|_{\boldsymbol{t}=\boldsymbol{0}}(\boldsymbol{t}-\boldsymbol{0})^{\beta}\\
&=s(x,\boldsymbol{0})+\sum_{i=1}^{n}\left.\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}\right|_{\boldsymbol{t}=\boldsymbol{0}}t_i,
\end{aligned}\tag{49}$$

for the first order approximation, where $\beta=(\beta_1,\cdots,\beta_n)$, $|\beta|=\beta_1+\cdots+\beta_n$, $\beta!=\beta_1!\cdots\beta_n!$, and $\left(\boldsymbol{t}\right)^{\beta}=\left(t_1\right)^{\beta_1}\cdots\left(t_n\right)^{\beta_n}$, in which the term $s(x,\boldsymbol{t})$ at the point $\boldsymbol{t}=0$ is expressed by

$$s(x,\boldsymbol{t})\big|_{\boldsymbol{t}=0}=\left.\left(\sum_{i=1}^{n}\frac{f(x-ht_i)}{g(x-ht_i)}K(t_i)\right)^{1-\alpha}\right|_{\boldsymbol{t}=0}=\left(n\frac{f(x)}{g(x)}K(0)\right)^{1-\alpha}.\tag{50}$$

Also, the term $\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}$ is written by

$$\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}=(1-\alpha)\left(\sum_{i=1}^{n}\frac{f(x-ht_i)}{g(x-ht_i)}K(t_i)\right)^{-\alpha}\left[-h\left(\frac{f(x-ht_i)}{g(x-ht_i)}\right)'K(t_i)+\frac{f(x-ht_i)}{g(x-ht_i)}K'(t_i)\right].\tag{51}$$

Then at the point $\boldsymbol{t}=0$, Eq. (27) is expressed by

$$\left.\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}\right|_{\boldsymbol{t}=\boldsymbol{0}}=(1-\alpha)\left(n\frac{f(x)}{g(x)}K(0)\right)^{-\alpha}\left[-h\left(\frac{f(x)}{g(x)}\right)'K(0)+\frac{f(x)}{g(x)}K'(0)\right]. \tag{52}$$

By substituting Eq. (49), Eq. (50) and Eq. (52) into Eq.(48), we obtain

$$\begin{aligned}
&E_g\left[\hat{f}^{1-\alpha}(x\,|\,g(x),S,D)\right]\\
&\approx\frac{h^n}{h^{1-\alpha}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(s(x,\boldsymbol{0})+\sum_{i=1}^{n}\left.\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}\right|_{\boldsymbol{t}=\boldsymbol{0}}t_i\right)\prod_{j=1}^{n}g(x-ht_j)dt_1\cdots dt_n\\
&=\frac{h^n}{h^{1-\alpha}}s(x,\boldsymbol{0})\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{j=1}^{n}g(x-ht_j)dt_1\cdots dt_n+\frac{h^n}{h^{1-\alpha}}\sum_{i=1}^{n}\left.\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}\right|_{\boldsymbol{t}=\boldsymbol{0}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}t_i\prod_{j=1}^{n}g(x-ht_j)dt_1\cdots dt_n\\
&=\frac{h^n}{h^{1-\alpha}}\left(n\frac{f(x)}{g(x)}K(0)\right)^{-\alpha}\frac{1}{h^n}+\frac{h^n}{h^{1-\alpha}}\left((1-\alpha)\left(n\frac{f(x)}{g(x)}K(0)\right)^{-\alpha-1}\left[-h\left(\frac{f(x)}{g(x)}\right)'K(0)+\frac{f(x)}{g(x)}K'(0)\right]\right)\frac{n}{h^{n+1}}\left(x-E_g[\tilde{X}]\right)\\
&=\frac{1}{h}\left(\frac{n}{h}K(0)\right)^{-\alpha}\left(\frac{f(x)}{g(x)}\right)^{-\alpha}-\frac{1-\alpha}{h}\left(x-E_g[\tilde{X}]\right)\left(\frac{n}{h}K(0)\right)^{-\alpha}\left[\frac{f(x)}{g(x)}\right]^{-\alpha-1}\left[\left(\frac{f(x)}{g(x)}\right)'-\frac{K'(0)}{hK(0)}\frac{f(x)}{g(x)}\right].
\end{aligned} \tag{53}$$

From Eq. (53), the partial derivative of $E_g\left[\hat{f}^{1-\alpha}(x\,|\,g(x),S,D)\right]$ with respect of $g(x)$ is given by

$$\begin{aligned}
&\frac{\partial E_g\left[\hat{f}^{1-\alpha}(x\,|\,g(x),S,D)\right]}{\partial g(x)}\\
&=\frac{\alpha}{h}\left(\frac{n}{h}K(0)\right)^{-\alpha}\frac{f^{-\alpha}(x)}{g^{-\alpha+1}(x)}-\frac{1-\alpha}{h}\left(x-E_g[\tilde{X}]\right)\left(\frac{n}{h}K(0)\right)^{-\alpha}\left\{(\alpha+1)\frac{f^{-\alpha-1}(x)}{g^{-\alpha}(x)}\left[\left(\frac{f(x)}{g(x)}\right)'-\frac{K'(0)}{hK(0)}\frac{f(x)}{g(x)}\right]\right.\\
&\qquad\left.+\left(\frac{f(x)}{g(x)}\right)^{-\alpha-1}\left(-\frac{f'(x)}{g^2(x)}+2\frac{f(x)}{g(x)}\frac{g'(x)}{g^2(x)}+\frac{K'(0)}{hK(0)}\frac{f(x)}{g^2(x)}\right)\right\}\\
&=\frac{\alpha}{h}\left(\frac{n}{h}K(0)\right)^{-\alpha}\left(\frac{f(x)}{g(x)}\right)^{-\alpha}\left[\frac{1}{g(x)}-\frac{1-\alpha}{\alpha}\left(x-E_g[\tilde{X}]\right)\left\{\frac{\alpha+1}{f(x)}\left(\frac{f(x)}{g(x)}\right)'\underbrace{-\frac{f'(x)}{g(x)}\frac{1}{f(x)}+\frac{f(x)}{g(x)}\frac{g'(x)}{g(x)}\frac{1}{f(x)}}_{\frac{1}{f(x)}\left(\frac{f(x)}{g(x)}\right)'}+\frac{g'(x)}{g^2(x)}-\frac{\alpha}{h}\frac{K'(0)}{K(0)}\frac{1}{g(x)}\right\}\right]\\
&=\frac{\alpha}{h}\left(\frac{n}{h}K(0)\right)^{-\alpha}\left(\frac{f(x)}{g(x)}\right)^{-\alpha}\left[\frac{1}{g(x)}-\frac{1-\alpha}{\alpha}\left(x-E_g[\tilde{X}]\right)\left\{\frac{\alpha}{f(x)}\left(\frac{f(x)}{g(x)}\right)'+\frac{g'(x)}{g^2(x)}-\frac{\alpha}{h}\frac{K'(0)}{K(0)}\frac{1}{g(x)}\right\}\right].
\end{aligned} \tag{54}$$

By substituting Eq. (30) into Eq. (21), the following equation is derived by

$$\frac{\alpha}{h}\left(\frac{n}{h}K(0)\right)^{-\alpha}\left(\frac{f(x)}{g(x)}\right)^{-\alpha}\left[\frac{1}{g(x)}-\frac{1-\alpha}{\alpha}\left(x-E_g[\tilde{X}]\right)\left\{\frac{\alpha}{f(x)}\left(\frac{f(x)}{g(x)}\right)'+\frac{g'(x)}{g^2(x)}-\frac{\alpha}{h}\frac{K'(0)}{K(0)}\frac{1}{g(x)}\right\}\right]=\alpha(1-\alpha)\lambda f^{-\alpha}(x). \tag{55}$$

Eq. (55) is rewritten by

$$g'(x)=\frac{\alpha}{\alpha-1}\left[\frac{1}{\alpha-1}\frac{1}{\left(x-E_g[\tilde{X}]\right)}+\frac{f'(x)}{f(x)}-\frac{1}{h}\frac{K'(0)}{K(0)}\right]g(x)+\frac{\alpha}{\alpha-1}\frac{\lambda h}{\left(x-E_g[\tilde{X}]\right)}\left(\frac{n}{h}K(0)\right)^{\alpha}g^{-\alpha+2}(x). \tag{56}$$

where $f(x)$ is assumed to satisfy $f(x)\neq 0$ and Eq. (56) doesn't involve any singularity by this assumption. Since $\dfrac{f'(x)}{f(x)}$ is included as the coefficient of $g(x)$, Eq. (56) is referred as ordinary differential equation(ODE) with variable coefficient. When $\alpha\neq 2$, this special form of Eq. (32) is called as Bernoulli equation and that has general solution as shown later in Theorem 1. Besides, when $\alpha=2$, Bernoulli equation in Eq. (56) is changed to a first order differential equation as shown later in Theorem 2. Thus, the optimal solution of $g(x)$ is discussed in the following selection of α :

(1) When $\alpha \neq 2$, a Bernoulli equation is analyzed.

(2) When $\alpha = 2$, a first order differential equation is analyzed.

(1) For the case $\alpha \neq 2$

Theorem 1. The Bernoulli equation is given by

$$p'(x) = a_1(x)p(x) + a_2(x)p^{\beta}(x), \quad \beta \neq 0,1, \tag{57}$$

where β can be any real number other than 0 or 1. The general solution in [4] is given by

$$p^{1-\beta}(x) = (1-\beta)e^{u(x)}\int e^{-u(x)}a_2(x)dx, \quad where\ u(x) = (1-\beta)\int a_1(x)dx, \tag{58}$$

where the function $\exp\left[u(x)\right]$ is called as an integrating factor.

Proof.

Set $\omega(x) = p^{1-\beta}(x)$ for changing the Eq. (57) to a general form, the derivative of $\omega(x)$ is

$$\begin{aligned} \omega'(x) &= (1-\beta)p^{-\beta}(x)p'(x) \\ &= (1-\beta)p^{-\beta}(x)\left[a_1(x)p(x) + a_2(x)p^{\beta}(x)\right] \\ &= (1-\beta)a_1(x)\omega(x) + (1-\beta)a_2(x), \end{aligned} \tag{59}$$

which leads to nonhomogeneous first order differential equation. Eq. (59) can be written in another form by

$$\omega'(x) - (1-\beta)a_1(x)\omega(x) = (1-\beta)a_2(x). \tag{60}$$

A new function $q(x)$ is introduced to Eq. (60) to make the left hand side of Eq. (60) have the form like $\left(\dfrac{\omega(x)}{q(x)}\right)' q(x)$. Eq. (60) is rewritten by

$$\left(\frac{\omega(x)}{q(x)}\right)' = \frac{\omega'(x)q(x) - (1-\beta)a_1(x)\omega(x)q(x)}{q^2(x)} = (1-\beta)\frac{a_2(x)}{q(x)}. \tag{61}$$

By comparing with $\left(\dfrac{\omega(x)}{q(x)}\right)' = \dfrac{\omega'(x)q(x) - \omega(x)q'(x)}{q^2(x)}$, the following equation is derived

$$q'(x) = (1-\beta)a_1(x)q(x), \tag{62}$$

where $q(x)$ is solved by

$$q(x) = \exp\left[(1-\beta)\int a_1(x)dx\right]. \tag{63}$$

Moreover, Eq. (61) provides another form as follows by $q(x)$ of Eq. (63)

$$\left(\frac{\omega(x)}{q(x)}\right)' = (1-\beta)\frac{a_2(x)}{q(x)}. \tag{64}$$

Then, $\omega(x)$ can be solved by

$$\omega(x) = (1-\beta)q(x)\int \frac{a_2(x)}{q(x)}dx. \tag{65}$$

By substituting Eq. (63) into Eq.(65), the following equation is derived

$$\omega(x)=(1-\beta)\exp\left(u(x)\right)\int a_2(x)\exp\left(-u(x)\right)dx, \beta\neq 0,1 \quad where\ \ u(x)=(1-\beta)\int a_1(x)dx, \tag{66}$$

which is the general solution of Eq.(57). □

We employ the Theorem 1 to obtain the solution Eq. (32). Corresponding to the definition formulas of Bernoulli equation in Theorem 1, we have $a_1(x)=\frac{\alpha}{\alpha-1}\left[\frac{1}{\alpha-1}\frac{1}{\left(x-E_g[\tilde{X}]\right)}+\frac{f'(x)}{f(x)}-\frac{1}{h}\frac{K'(0)}{K(0)}\right]$, $a_2(x)=\frac{\alpha}{\alpha-1}\frac{\lambda h}{\left(x-E_g[\tilde{X}]\right)}\left(\frac{n}{h}K(0)\right)^{\alpha}$ and $\beta=-\alpha+2$ where $\alpha\neq 2$ in Eq.(55). Note that $\beta=1$ is automatically excluded, since $\alpha\neq 0,1$ originally hold. The expression of $u(x)$ is obtained by

$$\begin{aligned}u(x)&=(1-\beta)\int a_1(x)dx\\&=(1+\alpha-2)\frac{\alpha}{\alpha-1}\int\left[\frac{1}{\alpha-1}\frac{1}{\left(x-E_g[\tilde{X}]\right)}+\frac{f'(x)}{f(x)}-\frac{1}{h}\frac{K'(0)}{K(0)}\right]dx\\&=\alpha\ln\left(f(x)\right)+\frac{\alpha}{\alpha-1}\ln\left(x-E_g[\tilde{X}]\right)-\frac{\alpha}{h}\frac{K'(0)}{K(0)}\int dx.\end{aligned}\tag{67}$$

The general solution of Eq. (56) is given by

$$g^{\alpha-1}(x)=\alpha\lambda h\left(\frac{n}{h}K(0)\right)^{\alpha}\exp\left(u(x)\right)\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx, \tag{68}$$

where $u(x)=\alpha\ln\left(f(x)\right)+\frac{\alpha}{\alpha-1}\ln\left(x-E_g[\tilde{X}]\right)-\frac{\alpha}{h}\frac{K'(0)}{K(0)}\int dx$ and $\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx$ is an indefinite integral, which is a function of *x*. λ is Lagrange multiplier. To express $g(x)$, Eq. (68) is rewritten by

$$\int_{-\infty}^{\infty}g(x)dx=\left[\alpha\lambda h\left(\frac{n}{h}K(0)\right)^{\alpha}\right]^{\frac{1}{\alpha-1}}\int_{-\infty}^{\infty}\left[\exp\left(u(x)\right)\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx\right]^{\frac{1}{\alpha-1}}dx=1. \tag{69}$$

Then, $\alpha\lambda h\left(\frac{n}{h}K(0)\right)^{\alpha}$ is expressed by

$$\alpha\lambda h\left(\frac{n}{h}K(0)\right)^{\alpha}=\frac{1}{\left\{\int_{-\infty}^{\infty}\left[\exp\left(u(x)\right)\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx\right]^{\frac{1}{\alpha-1}}dx\right\}^{\alpha-1}}. \tag{70}$$

By substituting Eq.(70) into Eq.(68), $g^{\alpha-1}(x)$ is given by

$$g^{\alpha-1}(x)=\frac{\exp\left(u(x)\right)\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx}{\left\{\int_{-\infty}^{\infty}\left[\exp\left(u(x)\right)\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx\right]^{\frac{1}{\alpha-1}}dx\right\}^{\alpha-1}},\quad \alpha\neq 2. \tag{71}$$

Eq.(71) is written in another form as

$$g(x)=\frac{\left[\exp\left(u(x)\right)\delta(x)\right]^{\frac{1}{\alpha-1}}}{\int_{-\infty}^{\infty}\left[\exp\left(u(x)\right)\delta(x)\right]^{\frac{1}{\alpha-1}}dx},\quad \alpha\neq 2, \tag{72}$$

where $u(x)=\alpha\ln\left(f(x)\right)+\frac{\alpha}{\alpha-1}\ln\left(x-E_g[\tilde{X}]\right)-\frac{\alpha}{h}\frac{K'(0)}{K(0)}\int dx$ and $\delta(x)=\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx$.

(2) For the case $\alpha=2$

Theorem 2. If the equations have the form

$$p'(x)=a_1(x)p(x)+a_2(x), \tag{73}$$

which is called as first order differential equation, the general solution is given by

$$p(x)=\exp\left(u(x)\right)\int\exp\left(-u(x)\right)a_2(x)dx, \tag{74}$$

where $u(x)=\int a_1(x)dx$ and the function $\exp\left[u(x)\right]$ is called as an integrating factor.

Proof.

Eq. (73) is rewritten as

$$p'(x)-a_1(x)p(x)=a_2(x). \tag{75}$$

We assume the existence of a new function $q(x)$. Both sides of Eq.(75) are multiplied by $q(x)$, which is given by

$$p'(x)q(x)-a_1(x)q(x)p(x)=a_2(x)q(x). \tag{76}$$

We assume $q(x)$ satisfy the following

$$-a_1(x)q(x)=q'(x). \tag{77}$$

So Eq.(76) is rewritten by

$$\left(p(x)q(x)\right)'=a_2(x)q(x). \tag{78}$$

From Eq.(78), $p(x)$ is obtained by

$$p(x)=\frac{1}{q(x)}\int a_2(x)q(x)dx. \tag{79}$$

Since $q(x)$ should satisfy Eq.(77), $q(x)$ can be solved by

$$q(x)=\exp\left(-\int a_1(x)dx\right). \tag{80}$$

By substituting Eq.(80) into Eq.(79), $p(x)$ is obtained by

$$p(x)=\exp\left(\int a_1(x)dx\right)\int \exp\left(-\int a_1(x)dx\right)a_2(x)dx, \tag{81}$$

which can be written in another form by

$$p(x)=\exp\left(u(x)\right)\int \exp\left(-u(x)\right)a_2(x)dx, \tag{82}$$

where $u(x)=\int a_1(x)dx$. □

When $\alpha=2$, the expression of Eq. (56) is changed to

$$g'(x)=2\left[\frac{1}{\left(x-E_g[\tilde{X}]\right)}+\frac{f'(x)}{f(x)}-\frac{1}{h}\frac{K'(0)}{K(0)}\right]g(x)+\frac{2\lambda h}{\left(x-E_g[\tilde{X}]\right)}\left(\frac{n}{h}K(0)\right)^2. \tag{83}$$

Corresponding to the definition formulas in Theorem 2, we have $a_1(x)=2\left[\frac{1}{\left(x-E_g[\tilde{X}]\right)}+\frac{f'(x)}{f(x)}-\frac{1}{h}\frac{K'(0)}{K(0)}\right]$, $a_2(x)=\frac{2\lambda h}{\left(x-E_g[\tilde{X}]\right)}\left(\frac{n}{h}K(0)\right)^2$. The expression of $u(x)$ is obtained by substituting $a_1(x)$ into $\int a_1(x)dx$ by

$$\begin{aligned}u(x)&=\int a_1(x)dx\\&=2\int\left[\frac{1}{\left(x-E_g[\tilde{X}]\right)}+\frac{f'(x)}{f(x)}-\frac{1}{h}\frac{K'(0)}{K(0)}\right]dx\\&=2\ln\left(f(x)\right)+2\ln\left(x-E_g[\tilde{X}]\right)-\frac{2}{h}\frac{K'(0)}{K(0)}\int dx.\end{aligned} \tag{84}$$

By applying Theorem 2, the solution of $g(x)$ is given by

$$g(x)=2\lambda h\left(\frac{n}{h}K(0)\right)^2\exp\left(\mathbf{u}(x)\right)\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx, \tag{85}$$

where $u(x)=2\ln\left(x-E_g[\tilde{X}]\right)+2\ln\left(f(x)\right)-\frac{2}{h}\frac{K'(0)}{K(0)}\int dx$. By the constraint $\int_{-\infty}^{\infty}g(x)dx=1$ and Eq.(84), $\int_{-\infty}^{\infty}g(x)dx$ is written by

$$\int_{-\infty}^{\infty}g(x)dx=2\lambda h\left(\frac{n}{h}K(0)\right)^2\int_{-\infty}^{\infty}\left[\exp\left(\mathbf{u}(x)\right)\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx\right]dx=1, \tag{86}$$

From Eq.(86), $2\lambda h\left(\frac{n}{h}K(0)\right)^2$ is expressed by

$$2\lambda h\left(\frac{n}{h}K(0)\right)^2=\frac{1}{\int_{-\infty}^{\infty}\left[\exp\left(\mathbf{u}(x)\right)\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx\right]dx}, \tag{87}$$

By substituting $2\lambda h\left(\frac{n}{h}K(0)\right)^2$ into (85), $g(x)$ is written by

$$g(x)=\frac{\exp\left(\mathbf{u}(x)\right)\delta(x)}{\int_{-\infty}^{\infty}\left[\exp\left(\mathbf{u}(x)\right)\delta(x)\right]dx}, \tag{88}$$

where $u(x)=2\ln\left(f(x)\right)+2\ln\left(x-E_g[\tilde{X}]\right)-\frac{2}{h}\frac{K'(0)}{K(0)}\int dx$ and $\delta(x)=\int\frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}dx$.

2. For the case $\alpha\to 0$ or 1

If $\alpha\to 0$ or 1, Eq. (4) is known to converge to KL-divergence given by

$$\begin{aligned}KL\left(f(x)\Box\hat{f}(x\,|\,g(x),S,D)\right)&=\int_{-\infty}^{\infty}\left(f(x)\ln\frac{f(x)}{\hat{f}(x\,|\,g(x),S,D)}\right)dx\\&=\int_{-\infty}^{\infty}f(x)\ln f(x)dx-\int_{-\infty}^{\infty}f(x)\ln\hat{f}(x\,|\,g(x),S,D)dx.\end{aligned} \tag{89}$$

The expectation of $KL\left(f(x)\Box\hat{f}(x\,|\,g(x),S,D)\right)$ is applied to measure the difference between $f(x)$ and $\hat{f}(x\,|\,g(x),S,D)$ over $g(x)$ as

$$E_g\left[KL\left(f(x)\Box\hat{f}(x\,|\,g(x),S,D)\right)\right]=\int_{-\infty}^{\infty}f(x)\ln f(x)dx-\int_{-\infty}^{\infty}f(x)E_g\left[\ln\hat{f}(x\,|\,g(x),S,D)\right]dx. \tag{90}$$

Our research target is to obtain the optimal $g(x)$ to minimize $E_g\left[KL\left(f(x)\Box\hat{f}(x\,|\,g(x),S,D)\right)\right]$ as

$$g_{opt}(x)=\arg\min_{g(x)}E_g\left[KL\left(f(x)\Box\hat{f}(x\,|\,g(x),S,D)\right)\right]\quad s.t.\int_{-\infty}^{\infty}g(x)dx=1. \tag{91}$$

By substituting Eq.(89) into Eq.(90), the following expression is derived

$$g_{opt}(x)=\arg\min_{g(x)}\int_{-\infty}^{\infty}\left(L\ (x,g(x))\right)dx\quad s.t.\int_{-\infty}^{\infty}g(x)dx=1, \tag{92}$$

where $L\ (x,g(x))=-f(x)E_g\left[\ln\hat{f}(x\,|\,g(x),S,D)\right]$. According to Eq. (9), $\tilde{L}(x,g(x))$ is set as $\tilde{L}(x,g(x))=L\ (x,g(x))+\lambda g(x)$. Correspondingly, Eq. (11) is written by

$$\begin{aligned}\frac{\partial\tilde{L}(x,g(x))}{\partial g(x)}&=\frac{\partial L\ (x,g(x))}{\partial g(x)}+\lambda\\&=-f(x)\frac{\partial E_g\left[\ln\hat{f}(x\,|\,g(x),S,D)\right]}{\partial g(x)}+\lambda=0\quad s.t.\int_{-\infty}^{\infty}g(x)dx=1.\end{aligned} \tag{93}$$

Then $\frac{\partial E_g\left[\ln\hat{f}(x\,|\,g(x),S,D)\right]}{\partial g(x)}$ is written by

$$\frac{\partial E_g\left[\ln\hat{f}(x\,|\,g(x),S,D)\right]}{\partial g(x)}=\frac{\lambda}{f(x)}. \tag{94}$$

Here, we consider to employ kernel estimator to estimate $\hat{f}(x\,|\,g(x),S,D)$ by

$$\hat{f}(x\,|\,g(x),S,D)=\frac{1}{h}\sum_{i=1}^{n}w(\tilde{X}_i)K(\frac{x-\tilde{X}_i}{h}), \tag{95}$$

where $\tilde{X}_i, i=1,\cdots,n$ are samples in the weighted sample set D and $g(x)$ should be $g(x)\neq 0$ whenever $f(x)\neq 0$. *w*(x) is unknown since *f*(x) is not given in practice. However, we assume that *w*(x) is given by Eq.(11) from *f*(x) and *g*(x) in our current problem setting as noted earlier in the general problem description. Now, we want to calculate $E_g\left[\ln\hat{f}(x\,|\,g(x),S,D)\right]$, which is written by

$$\begin{aligned}
&E_g\left[\ln\hat{f}(x\,|\,g(x),S,D)\right]\\
&=E_g\left[\ln\left(\frac{1}{h}\sum_{i=1}^{n}w(\tilde{X}_i)K(\frac{x-\tilde{X}_i}{h})\right)\right]\\
&=E_g\left[\ln\left(\frac{1}{h}\sum_{i=1}^{n}\frac{f(\tilde{X}_i)}{g(\tilde{X}_i)}K(\frac{x-\tilde{X}_i}{h})\right)\right]\\
&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\ln\left(\frac{1}{h}\sum_{i=1}^{n}\frac{f(y_i)}{g(y_i)}K(\frac{x-y_i}{h})\right)g(y_1,\cdots,y_n)dy_1\cdots dy_n.
\end{aligned} \tag{96}$$

Since $X_1,\cdots,X_n$ are i.i.d. sampled from $g(\mathrm{x})$, we take the form $g(y_1,\cdots,y_n)=g(y_1)\cdots g(y_n)$. Eq.(96) is written by

$$\begin{aligned}
&E_g\left[\ln\hat{f}(x\,|\,g(x),S,D)\right]\\
&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\ln\left(\frac{1}{h}\sum_{i=1}^{n}\frac{f(y_i)}{g(y_i)}K(\frac{x-y_i}{h})\right)\prod_{j=1}^{n}g(y_j)dy_1\cdots dy_n\\
&\overset{t_i=\frac{x-y_i}{h}}{=}\ h^n\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\ln\left(\frac{1}{h}\sum_{i=1}^{n}\frac{f(x-ht_i)}{g(x-ht_i)}K(t_i)\right)\prod_{j=1}^{n}g(x-ht_j)dt_1\cdots dt_n.
\end{aligned} \tag{97}$$

Set $s(x,\boldsymbol{t})=s(x,\left[t_1,\cdots,t_n\right]^T)=\ln\left(\frac{1}{h}\sum_{i=1}^{n}\frac{f(x-ht_i)}{g(x-ht_i)}K(t_i)\right)$. By applying Taylor's expansion for multivariate $t_1,\cdots,t_n$ to $s(x,\boldsymbol{t})$. Then, $s(x,\boldsymbol{t})$ is written by

$$\begin{aligned}
s(x,\boldsymbol{t})&=\sum_{j=0}^{\infty}\sum_{|\beta|=j}\frac{\nabla^{\beta}s(x,\mathbf{0})}{\beta!}(\boldsymbol{t}-\mathbf{0})^{\beta}\\
&=\sum_{j=0}^{\infty}\sum_{|\beta|=j}\frac{1}{\beta!}\frac{\partial^{|\beta|}s(x,\boldsymbol{t})}{\partial t_1^{\beta_1}\cdots\partial t_n^{\beta_n}}\bigg|_{\boldsymbol{t}=\mathbf{0}}(\boldsymbol{t}-\mathbf{0})^{\beta}\\
&\approx\sum_{j=0}^{1}\sum_{|\beta|=j}\frac{1}{\beta!}\frac{\partial^{|\beta|}s(x,\boldsymbol{t})}{\partial t_1^{\beta_1}\cdots\partial t_n^{\beta_n}}\bigg|_{\boldsymbol{t}=\mathbf{0}}(\boldsymbol{t}-\mathbf{0})^{\beta}\\
&=s(x,\mathbf{0})+\sum_{i=1}^{n}\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}\bigg|_{\boldsymbol{t}=\mathbf{0}}t_i,
\end{aligned} \tag{98}$$

for the first order approximation, where $\beta=(\beta_1,\cdots,\beta_n)$, $|\beta|=\beta_1+\cdots+\beta_n$, $\beta!=\beta_1!\cdots\beta_n!$, and $(\boldsymbol{t})^{\beta}=(t_1)^{\beta_1}\cdots(t_n)^{\beta_n}$, in which the term $s(x,\boldsymbol{t})$ at the point $\boldsymbol{t}=0$ is expressed by

$$s(x,\boldsymbol{t})\Big|_{\boldsymbol{t}=0}=\ln\left(\frac{1}{h}\sum_{i=1}^{n}\frac{f(x-ht_i)}{g(x-ht_i)}K(t_i)\right)\Bigg|_{\boldsymbol{t}=0}=\ln\left(\frac{n}{h}\frac{f(x)}{g(x)}K(0)\right). \tag{99}$$

Also, the term $\dfrac{\partial s(x,\boldsymbol{t})}{\partial t_i}$ is written by

$$\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}=\left(\sum_{i=1}^{n}\frac{f(x-ht_i)}{g(x-ht_i)}K(t_i)\right)^{-1}\left[-h\left(\frac{f(x-ht_i)}{g(x-ht_i)}\right)'K(t_i)+\frac{f(x-ht_i)}{g(x-ht_i)}K'(t_i)\right]. \tag{100}$$

Then at the point $\boldsymbol{t}=0$, Eq.(100) is expressed by

$$\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}\Bigg|_{\boldsymbol{t}=\boldsymbol{0}}=\left(n\frac{f(x)}{g(x)}K(0)\right)^{-1}\left[-h\left(\frac{f(x)}{g(x)}\right)'K(0)+\frac{f(x)}{g(x)}K'(0)\right]. \tag{101}$$

By substituting Eq.(98), Eq.(99), Eq.(100) into Eq.(101), we obtain

$$\begin{aligned}
&E_g\left[\ln\hat{f}(x\,|\,g(x),S,D)\right]\\
&\approx h^n\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(s(x,\boldsymbol{0})+\sum_{i=1}^{n}\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}\Bigg|_{\boldsymbol{t}=\boldsymbol{0}}t_i\right)\prod_{j=1}^{n}g(x-ht_j)dt_1\cdots dt_n.\\
&=h^n s(x,\boldsymbol{0})\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{j=1}^{n}g(x-ht_j)dt_1\cdots dt_n+h^n\sum_{i=1}^{n}\frac{\partial s(x,\boldsymbol{t})}{\partial t_i}\Bigg|_{\boldsymbol{t}=\boldsymbol{0}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}t_i\prod_{j=1}^{n}g(x-ht_j)dt_1\cdots dt_n\\
&=\ln\left(\frac{n}{h}\frac{f(x)}{g(x)}K(0)\right)+h^n\left(\left(n\frac{f(x)}{g(x)}K(0)\right)^{-1}\left[-h\left(\frac{f(x)}{g(x)}\right)'K(0)+\frac{f(x)}{g(x)}K'(0)\right]\right)\frac{n}{h^{n+1}}\left(x-E_g[\tilde{X}]\right)\\
&=\ln\left(\frac{n}{h}\frac{f(x)}{g(x)}K(0)\right)-\left(x-E_g[\tilde{X}]\right)\left[\frac{f'(x)}{f(x)}-\frac{g'(x)}{g(x)}-\frac{K'(0)}{hK(0)}\right].
\end{aligned} \tag{102}$$

Then, by taking the derivative of Eq.(102), we obtain

$$\frac{\partial E_g\left[\ln\hat{f}(x\,|\,g(x),S,D)\right]}{\partial g(x)}=-\frac{1}{g(x)}-\left(x-E_g[\tilde{X}]\right)\frac{g'(x)}{g^2(x)} \tag{103}$$

By substituting Eq.(103) into Eq.(93), the following equation is derived by

$$-\frac{1}{g(x)}-\left(x-E_g[\tilde{X}]\right)\frac{g'(x)}{g^2(x)}=\frac{\lambda}{f(x)}. \tag{104}$$

Eq.(102) is transformed by

$$g'(x)=-\frac{g(x)}{\left(x-E_g[\tilde{X}]\right)}-\frac{\lambda}{\left(x-E_g[\tilde{X}]\right)f(x)}g^2(x). \tag{105}$$

Eq.(103) can be solved by Theorem 1. We have $a_1(x)=-\dfrac{1}{\left(x-E_g[\tilde{X}]\right)}$, $a_2(x)=-\dfrac{\lambda}{\left(x-E_g[\tilde{X}]\right)f(x)}$ and $\beta=2$. The expression of $u(x)$ is obtained by

$$\begin{aligned} u(x) &= (1-\beta)\int a_1(x)dx \\ &= \int \frac{1}{\left(x-E_g[\tilde{X}]\right)}dx \\ &= \ln\left(x-E_g[\tilde{X}]\right)+C, \end{aligned} \tag{106}$$

where C is a constant. The general solution of Eq.(105) is

$$g^{-1}(x) = \lambda \exp\left(\mathbf{u}(x)\right)\int \frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}\frac{1}{f(x)}dx, \tag{107}$$

where $u(x)=\ln\left(x-E_g[\tilde{X}]\right)+C$. To express $g(x)$, Eq.(107) is rewritten by

$$g(x) = \lambda^{-1}\left[\exp\left(\mathbf{u}(x)\right)\int \frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}\frac{1}{f(x)}dx\right]^{-1}. \tag{108}$$

By the constraint $\int_{-\infty}^{\infty} g(x)dx = 1$ and Eq.(108), $g(x)$ must satisfy the following condition

$$\int_{-\infty}^{\infty} g(x)dx = \lambda^{-1}\int_{-\infty}^{\infty}\left[\exp\left(\mathbf{u}(x)\right)\int \frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}\frac{1}{f(x)}dx\right]^{-1} dx = 1. \tag{109}$$

Then λ^{-1} is expressed by

$$\lambda^{-1} = \frac{1}{\int_{-\infty}^{\infty}\left[\exp\left(\mathbf{u}(x)\right)\int \frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}\frac{1}{f(x)}dx\right]^{-1} dx}. \tag{110}$$

By substituting Eq.(110) into Eq.(108), $g(x)$ is written by

$$g(x) = \frac{\left[\exp\left(\mathbf{u}(x)\right)\delta(x)\right]^{-1}}{\int_{-\infty}^{\infty}\left[\exp\left(\mathbf{u}(x)\right)\delta(x)\right]^{-1}dx}, \tag{111}$$

where $u(x)=\ln\left(x-E_g[\tilde{X}]\right)+C$ and $\delta(x) = \int \frac{\exp\left(-u(x)\right)}{\left(x-E_g[\tilde{X}]\right)}\frac{1}{f(x)}dx$.

Reference

[1] Silverman, B.W.: Density Estimation for Statistics and Data Analysis, Chaoman & Hall/CRC (1986)

[2] Khan, A.U., Hildreth, W.B.: Case studies in public budgeting and financial management. New York, N.Y: Marcel Dekker (2003).

[3] Waters, A.: Alpha divergence, Lecture note of Electrical and Computer Engineering,Rice University, Available at: http://www.ece.rice.edu/~vc3/elec633/AlphaDivergence.pdf (2008).

[4] Minka, T.: Divergence measures and message passing, Tech. Rep. No. MSR-TR-2005-

173, Microsoft Research Ltd., Cambridge, UK (2005)

[5] Hunter, J. K.: Lecture notes on applied mathematics. Available at: https://www.math.ucdavis.edu/ hunter/m280_09/ch3.pdf (2009).

[6] Jost, J. and Li-Jost, X., Hunter, J. K.: Calculus of variations, Cambridge University Press (1998).

[7] Kil, R.M.: The euler-lagrange equation. Available at: http://mathsci.kaist.ac.kr/ nipl/am621/lecturenotes/Euler-Lagrange_equation.pdf (2006).

[8] Iba, Y., Saito N., Kitajima A.: Multicanonical MCMC for sampling rare events: an illustrative review, Ann. Inst. Stat. Math., Vol.66, pp.611–645 (2014).

List of Publications and Significant Collaborations that resulted from your AOARD supported project: In standard format showing authors, title, journal, issue, pages, and date, for each category list the following:

a) papers published in peer-reviewed journals,

a-1

Kai Ming Ting, Takashi Washio, Jonathan R. Wells and Sunil Aryal, Defying the gravity of learning curve: a characteristic of nearest neighbour anomaly detectors, Machine Learning, Vol.106, No.1, pp.55–91, Aug., 29, 2016.

a-2

Bo Chen, Kai Ming Ting, Takashi Washio, Gholamreza Haffari, Half-space Mass: A maximally robust and efficient data depth method, Machine Learning, Vol.100, pp.677–699, 2015, Aug., 5, 2015.

b) papers published in peer-reviewed conference proceedings,

Sunil Aryal, Kai Ming Ting, Gholamreza Haffari, Takashi Washio, Beyond tf-idf and cosine distance in documents dissimilarity measure, In Information Retrieval Technology Vol. 9460 of the series Lecture Notes in Computer Science: Proc. the 11th Asia Information Retrieval Societies Conference (AIRS 2015), Springer International Publishing, pp.400-406, Dec., 2. 2015.

d) conference presentations without papers,

Takashi Washio, Defying the Gravity of Learning Curves: Are More Samples Better for Nearest Neighbor Anomaly Detectors ?, SISAP 2016: the 9th International Conference on Similarity Search and Applications, Invited Talk, Oct., 24, 2016.