SUPPLEMENTARY MATERIAL

Numerical Simulations

This section implements Monte Carlo simulation on different data generation processes to evaluate the performance of our proposed method in comparison with other state-of-the-art approaches. We adopt the Gaussian kernel function in our KTCS method and set the hyper-parameters according to the suggestions in[1]. The first competitive method is the most intuitive uniform subsampling method (Uniform), where all the selected data points are uniformly selected. We also consider Support Point (SP)[2], and SPARTAN[3], which primarily aim to achieve *representativeness*. Additionally, we evaluate two competitive methods that focus on *coverage*: the MMD-Critic[4] and MED[5]. The simulation is conducted through Python. For each simulation setup, we repeat 200 times.

We first generate the the synthetic dataset $\boldsymbol{x}_1,...,\boldsymbol{x}_N$ and each of the $\boldsymbol{x}_i \in \mathbb{R}^d$ is a $1 \times d$ vector. In order to introduce non-linearity into the data, we consider setting it under a generative model setup. That is, we firstly generate $\boldsymbol{x}_1^*,...,\boldsymbol{x}_N^*$ from a simple distribution, for example, a mixture of Gaussian distribution. We name these $\boldsymbol{x}_1^*,...,\boldsymbol{x}_N^*$ as *latents*. Here, each of the $\boldsymbol{x}_i^* \in \mathbb{R}^{d_*}$ is a $1 \times d_*$ vector, and the dimension of *latent variables* is much smaller than the dimension of dataset, i.e., $d_* < d$. Based on the generated latent variables, we then apply a nonlinear transformation to generate a synthetic dataset $\boldsymbol{x}_1,...,\boldsymbol{x}_N$. We set $\boldsymbol{x}_i^\top = A(\boldsymbol{W}\boldsymbol{x}_i^{*\top})$, where $\boldsymbol{W}$ is a $d \times d_*$ matrix and $A(\cdot)$ is a nonlinear activation function, for example a Logit function. Such a setup is able to generate a very complicated nonlinear data structure, for example[6] shows that sampling from a Gaussian distribution with a similar generation setup can generate highly complicated images.

In application, researchers have observed that safety-critical events are not only rarely observed, but also largely deviate from the center and show great diversity[7]. To mimic this, we separate the generations for the latent into two parts, *major* and *minor*. For the *major* part, we set the number of cases to be larger and the variance to be small. For the *minor* part, we set the number of cases to be smaller and the variance to be larger. The center of latents for the *major* part and *minor* part are far away from each other.

For each of the generated synthetic datasets, we use our proposed methods and the five other competitive methods to select $0.5\sqrt{N}, \sqrt{N}, 2\sqrt{N}$ points individually. The performance of the different methods is evaluated based on IP and MMD. We also calculate the proportion of the minor modes covered through different methods—namely, as PCT in the proceeding analysis. Intuitively, larger values in the PCT indicate these sub-selected points are more likely to cover corner cases, meaning a larger coverage. For example, if we set the number of minor modes to be 10, and 8 of these minor modes can exhibit at least one point selected through the sub-sampling method, then here PCT = 0.8.

Example 1: We consider the generation for the latents to come from a two-dimensional ring-shaped Gaussian mixture

Table 1. Performance comparison under simulation example 1.

			IP			MMD			PCT		
N	d	Method	$0.5\sqrt{N}$	$\sqrt{N}$	$2\sqrt{N}$	$0.5\sqrt{N}$	$\sqrt{N}$	$2\sqrt{N}$	$0.5\sqrt{N}$	$\sqrt{N}$	$2\sqrt{N}$
1100	20	**KTCS**	0.184	0.153	0.129	0.038	0.011	0.004	0.566	0.722	0.905
		MED	0.121	0.120	0.120	0.036	0.019	0.014	0.987	1.000	1.000
		MMD-Critic	0.143	0.135	0.146	0.026	0.012	0.014	0.727	0.997	1.000
		SP	0.169	0.148	0.137	0.011	0.006	0.004	0.107	0.271	0.520
		SPARTAN	0.221	0.188	0.168	0.052	0.026	0.012	0.134	0.251	0.442
		Uniform	0.223	0.187	0.168	0.054	0.025	0.012	0.137	0.261	0.467
1100	50	**KTCS**	0.130	0.100	0.085	0.056	0.023	0.008	0.688	0.899	0.982
		MED	0.079	0.070	0.073	0.051	0.025	0.016	0.998	1.000	1.000
		MMD-Critic	0.096	0.084	0.089	0.039	0.019	0.014	0.723	0.999	1.000
		SP	0.120	0.103	0.095	0.012	0.005	0.003	0.080	0.223	0.444
		SPARTAN	0.165	0.136	0.121	0.055	0.029	0.014	0.123	0.238	0.400
		Uniform	0.163	0.132	0.118	0.055	0.026	0.013	0.146	0.269	0.479
5500	20	**KTCS**	0.146	0.129	0.118	0.010	0.002	0.000	0.837	0.949	0.984
		MED	0.113	0.113	0.111	0.024	0.020	0.017	1.000	1.000	1.000
		MMD-Critic	0.130	0.146	0.172	0.013	0.018	0.037	0.998	0.998	1.000
		SP	0.151	0.145	0.138	0.004	0.002	0.001	0.265	0.482	0.760
		SPARTAN	0.184	0.166	0.154	0.022	0.011	0.006	0.270	0.493	0.748
		Uniform	0.184	0.166	0.154	0.023	0.011	0.006	0.284	0.498	0.741
5500	50	**KTCS**	0.103	0.087	0.080	0.016	0.004	0.000	0.850	0.988	0.995
		MED	0.065	0.067	0.070	0.031	0.024	0.020	1.000	1.000	1.000
		MMD-Critic	0.079	0.088	0.114	0.021	0.016	0.029	1.000	1.000	1.000
		SP	0.104	0.101	0.098	0.004	0.002	0.001	0.244	0.449	0.767
		SPARTAN	0.129	0.118	0.110	0.023	0.012	0.006	0.293	0.479	0.734
		Uniform	0.131	0.117	0.110	0.024	0.012	0.006	0.282	0.488	0.736

distribution. We set the generation of latent as: $\boldsymbol{x}^*_{i,k} \sim \sum_{j=1}^{10} 0.1 \times \mathcal{N}\left(\mu_{j,k}, \Sigma_k\right)$ where $k \in [\text{Major}, \text{Minor}]$. This implies that both Major and Minor parts each have 10 modes. For the vector $\boldsymbol{\mu}_{j.k}$, we set the parameters to $j = 1, ..., 10$, and $\boldsymbol{\mu}_{j,\text{Major}} = \left(2\cos(\frac{j\pi}{10}), 2\sin(\frac{j\pi}{10})\right)$ and $\boldsymbol{\mu}_{j,\text{Minor}} = \left(4\cos(\frac{j\pi}{10}), 4\sin(\frac{j\pi}{10})\right)$. We set the covariance matrix as follows, where ρ is used to control the correlations:

$$\Sigma_{\text{Major}} = \begin{pmatrix} 0.02 & 0.02\rho \\ 0.02\rho & 0.02 \end{pmatrix} \quad \text{and} \quad \Sigma_{\text{Minor}} = \begin{pmatrix} 0.1 & 0.1\rho \\ 0.1\rho & 0.1 \end{pmatrix} \tag{1}$$

We set the ratio between Major and Minor as 10:1. We consider the total number of points to be $N = 5500$ and $N = 1100$. For example, if the total number of points $N = 5500$, then 5000 points come from the Major part and each mode has 500 points. The remaining 500 come from the Minor part and each mode has 50 points. In the simulations, we consider $\rho = 0.5$ and the dimension $d = 20, 50$. The entries for the transformation matrix come from a standard Normal distribution $\boldsymbol{W}_{i,j} \sim \mathcal{N}(0,1)$, where $\boldsymbol{W}_{i,j}$ denotes the row i and column j for matrix $\boldsymbol{W}$. The results are reported in Table 1.

Example 2: We consider the generation for the latent to come from a factorial design-based process. We first consider a 2^4 full factorial design for generating the synthetic data, so as the number of modes here to be $2^4 = 16$. Each of the modes k and $k = 1, ...16$, is generated from a multivariate Gaussian distribution as $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma)$, the covariance matrix is as $\Sigma_{i,j} = 0.5^{|i-j|}$, and here we set the parameter $\rho = 0.5$.

Table 2. Performance comparison under simulation example 2.

			IP			MMD			PCT		
N	d	Method	$0.5\sqrt{N}$	$\sqrt{N}$	$2\sqrt{N}$	$0.5\sqrt{N}$	$\sqrt{N}$	$2\sqrt{N}$	$0.5\sqrt{N}$	$\sqrt{N}$	$2\sqrt{N}$
840	20	**KTCS**	0.160	0.118	0.097	0.074	0.032	0.008	0.451	0.683	0.830
		MED	0.095	0.077	0.075	0.070	0.052	0.038	0.566	0.813	0.951
		MMD-Critic	0.192	0.117	0.095	0.077	0.031	0.021	0.375	0.685	0.902
		SP	0.158	0.136	0.123	0.014	0.008	0.005	0.063	0.128	0.276
		SPARTAN	0.211	0.171	0.150	0.064	0.029	0.015	0.084	0.153	0.304
		Uniform	0.209	0.172	0.150	0.062	0.029	0.015	0.087	0.151	0.296
840	50	**KTCS**	0.116	0.078	0.061	0.080	0.037	0.013	0.517	0.778	0.899
		MED	0.076	0.048	0.040	0.072	0.052	0.038	0.558	0.803	0.946
		MMD-Critic	0.143	0.074	0.054	0.100	0.041	0.024	0.363	0.685	0.902
		SP	0.118	0.095	0.084	0.014	0.008	0.004	0.042	0.121	0.265
		SPARTAN	0.161	0.126	0.105	0.068	0.035	0.017	0.074	0.151	0.286
		Uniform	0.157	0.120	0.103	0.067	0.032	0.015	0.094	0.177	0.308
4200	20	**KTCS**	0.120	0.101	0.090	0.020	0.005	0.001	0.674	0.789	0.870
		MED	0.071	0.067	0.067	0.058	0.048	0.042	0.880	0.958	0.987
		MMD-Critic	0.111	0.090	0.104	0.033	0.024	0.020	0.820	0.945	0.977
		SP	0.137	0.124	0.115	0.006	0.004	0.003	0.165	0.345	0.566
		SPARTAN	0.174	0.152	0.138	0.028	0.013	0.007	0.182	0.323	0.553
		Uniform	0.171	0.151	0.139	0.027	0.014	0.007	0.177	0.330	0.554
4200	50	**KTCS**	0.081	0.066	0.058	0.028	0.009	0.003	0.773	0.868	0.925
		MED	0.043	0.035	0.034	0.057	0.047	0.041	0.870	0.957	0.987
		MMD-Critic	0.071	0.051	0.055	0.045	0.030	0.021	0.831	0.945	0.983
		SP	0.096	0.086	0.081	0.006	0.004	0.002	0.164	0.374	0.612
		SPARTAN	0.125	0.106	0.099	0.030	0.015	0.008	0.167	0.331	0.571
		Uniform	0.121	0.106	0.098	0.028	0.014	0.007	0.181	0.337	0.551

The vector $\boldsymbol{\mu}_k$ is generated based on a 2^4 full factorial design process. The matrix Π is defined to take the form of 2^4 full factorial design as follows:

$$\Pi^\top = \begin{pmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\ +1 & +1 & -1 & -1 & +1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & +1 & -1 & +1 & -1 \\ +1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & +1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 \end{pmatrix}$$

The matrix Π is used to control the sign for the vector of $\boldsymbol{\mu}_k$. $\boldsymbol{U}$ is used to control the coefficient for the vector $\boldsymbol{\mu}_k$, and it is defined to be a 16×4 matrix and each of the entries for matrix $\boldsymbol{U}$ comes from a Uniform distribution as $\boldsymbol{U}_{i,j} \sim \text{Unif}(0,1)$. We define the $\odot$ as the Hadmard matrix product, meaning that for these two matrices A and B, it will have the following: $(\Pi \odot \boldsymbol{U})_{i,j} = \Pi_{i,j} \times \boldsymbol{U}_{i,j}$. Finally, the vector $\boldsymbol{\mu}_k = (\Pi \odot \boldsymbol{U})_{i,\cdot}$.

We randomly split the 16 modes into two parts, half of them labeled as Major, and the remaining half of them labeled as Minor. The ratio of the number of points between these two parts is 20:1. We consider two situations. In the first situation, the total number of points is $N = 840$, meaning that each of the Major modes has 100 points, and each of the Minor modes has 5 points. In the second situation, the total number of points $N = 4200$, meaning that each of the Major modes has 500 points, and each of the Minor modes has 25 points. The entries for the transformation matrix come from a standard Normal distribution

$\boldsymbol{W}_{i,j} \sim \mathcal{N}(0,1)$, where $\boldsymbol{W}_{i,j}$ denotes the row i and column j for matrix $\boldsymbol{W}$. We consider the dimension for the final generated data to be either 20 or 50. The results are reported in Table 2. Our methods achieve great balance between IP and MMD.

Example 3: We further extend the 2^4 full factorial design into a more complicated 2^5 full factorial design. Here, a total of 32 modes are considered, and 16 of them are labeled as Major while the remaining 16 are labeled as Minor. The ratio of the number of points between these two parts is 20:1. We also consider two situations. In the first situation, the total number of points $N = 1680$, meaning that each of the Major modes has 100 points, and each of the Minor modes has 5 points. In the second situation, the total number of points $N = 8400$, meaning that each of the Major modes has 500 points, and each of the Minor modes has 25 points. The results are reported in the supplementary material Section . The simulation results also confirm our KTCS method can simultaneously achieve good performance in representativeness and coverage.

Table 3. Performance comparison under simulation example 3.

			IP			MMD			PCT		
N	d	Method	$0.5\sqrt{N}$	$\sqrt{N}$	$2\sqrt{N}$	$0.5\sqrt{N}$	$\sqrt{N}$	$2\sqrt{N}$	$0.5\sqrt{N}$	$\sqrt{N}$	$2\sqrt{N}$
1680	20	**KTCS**	0.103	0.075	0.059	0.049	0.021	0.008	0.311	0.510	0.696
		MED	0.061	0.046	0.042	0.054	0.037	0.026	0.502	0.708	0.857
		MMD-Critic	0.126	0.072	0.056	0.047	0.021	0.013	0.299	0.608	0.817
		SP	0.101	0.086	0.076	0.010	0.006	0.004	0.043	0.101	0.209
		SPARTAN	0.138	0.110	0.093	0.046	0.024	0.011	0.056	0.121	0.220
		Uniform	0.137	0.108	0.092	0.045	0.022	0.011	0.063	0.128	0.236
1680	50	**KTCS**	0.073	0.047	0.035	0.053	0.025	0.011	0.376	0.614	0.797
		MED	0.051	0.030	0.023	0.048	0.031	0.021	0.507	0.716	0.862
		MMD-Critic	0.092	0.045	0.031	0.060	0.025	0.014	0.296	0.611	0.817
		SP	0.064	0.051	0.043	0.012	0.006	0.004	0.039	0.100	0.207
		SPARTAN	0.094	0.069	0.056	0.048	0.024	0.012	0.062	0.119	0.219
		Uniform	0.094	0.069	0.055	0.047	0.023	0.011	0.057	0.118	0.218
8400	20	**KTCS**	0.075	0.061	0.052	0.017	0.006	0.002	0.514	0.665	0.770
		MED	0.041	0.036	0.034	0.043	0.034	0.029	0.749	0.852	0.907
		MMD-Critic	0.068	0.052	0.058	0.022	0.016	0.011	0.707	0.871	0.943
		SP	0.086	0.076	0.069	0.005	0.003	0.002	0.114	0.237	0.424
		SPARTAN	0.109	0.091	0.081	0.020	0.010	0.005	0.128	0.246	0.430
		Uniform	0.108	0.092	0.082	0.020	0.010	0.005	0.131	0.242	0.424
8400	50	**KTCS**	0.048	0.037	0.031	0.022	0.009	0.003	0.609	0.763	0.847
		MED	0.027	0.020	0.017	0.034	0.027	0.023	0.751	0.847	0.899
		MMD-Critic	0.042	0.028	0.025	0.026	0.017	0.012	0.707	0.864	0.946
		SP	0.051	0.044	0.041	0.005	0.003	0.002	0.112	0.245	0.442
		SPARTAN	0.068	0.057	0.050	0.022	0.011	0.005	0.127	0.243	0.410
		Uniform	0.069	0.057	0.050	0.021	0.010	0.005	0.124	0.245	0.442

In conclusion, our simulation results for multiple data generation scenarios consistently demonstrate that KTCS strikes a balance between representativeness and coverage. Compared with existing state-of-the-art methods, our proposed KTCS method achieves superior IP values and remarkably low MMD scores while also capturing an impressively high proportion of minor modes. These findings confirm that KTCS not only selects representative samples from complex data but also successfully covers safety-critical and corner cases that lie far from the center of the distribution. Moreover, performance advantages persist as we vary both the sample size and the dimension, reflecting the robustness and generalizability of our approach.

All Selected Test Cases

The 48 extracted features with their corresponding calculation formula and engineering interpretations are presented in Table 4. Figures 1 and Figure 2 present the 118 cases using radar plots. Each case is accompanied by its corresponding weights generated by the attention mechanism, case ID, and descriptive metrics for the selected features.

Technical Lemmas

Lemma 1. *Let $X_1,\ldots,X_n$ be i.i.d. random variables on Hilbert space with the condition that $\sup_{i=1,\ldots,n}\|X_i\|_{\mathscr{H}} \leq A$ for some real number $A > 0$. Then, for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$ that*

$$\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\|_{\mathscr{H}} \leq A\frac{\sqrt{2\log(2/\delta)}}{\sqrt{n}}\right) \geq 1-\delta \tag{2}$$

Lemma 2. *Let $X_1,\ldots,X_n$ be i.i.d. random variables in a Hilbert space $(\mathscr{H},\|\cdot\|)$ such that - $\forall i \in [n], \mathbb{E}X_i = \mu$, - $\exists \sigma > 0, \exists H > 0, \forall i \in [n], \forall p \geq 2, \mathbb{E}\|X_i-\mu\|^p \leq \frac{1}{2}p!\sigma^2 H^{p-2}$. Then, for any $\delta \in [0,1]$, we have with probability at least $1-\delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right\|_{\mathscr{H}} \leq \frac{2H\log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}} \tag{3}$$

Lemma 1 rely on the Theorem 3.5 in[8], and Lemma 2 rely on Theorem 3.3.4 of[9].[10] and[11] use similar approaches.

Lemma 3 (McDiamrd's inequality)**.** *Suppose $X_1,...,X_N$ in the domain of $\mathscr{X}$ are independent random variables. For any function $f : \mathscr{X}^N \to \mathbb{R}^N$, there exists a constant $c_i > 0$, such that $\left|f(X_1,...,X_i,...,X_N) - f(X_1,...,X_i',...,X_N\right| \leq c_i$, then the following holds:*

$$\Pr\left(\left|f(X_1,\ldots,X_N) - \mathbb{E}[f(X_1,\ldots,X_N)]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{N} c_i^2}\right) \tag{4}$$

where ε is some constant here.

Lemma 4 (Hoeffding inequality)**.** *Suppose $X_1,\ldots,X_n$ are independent random variables satisfying that $\mathbb{E}[X_i] = 0$ and $a_i \leq X_i \leq b_i$ for all i. If $\varepsilon > 0$ and for all $t > 0$, then:*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq \varepsilon\right] \leq e^{-nt\varepsilon}\prod_{i=1}^{n} e^{t^2(b_i-a_i)^2/8} \tag{5}$$

where ε is some constant here.

Table 4. All extracted features and their associated calculation formula and descriptions

Driving Data	Feature	Formula	Engineering Interpretation
X-Axis Acceleration	Initiation	$accx\text{-}init = a_{x,1}$	Longitudinal acceleration at initial time
	Mean	$accx\text{-}mean = \frac{1}{T}\sum_{t=1}^{T} a_{x,t}$	Average of the longitudinal acceleration
	Maximum	$accx\text{-}max = \max[a_{x,1}, ...a_{x,T}]$	Maximum of longitudinal acceleration
	Minimum	$accx\text{-}min = \min[a_{x,1}, ...a_{x,T}]$	Minimum of longitudinal acceleration
	Std Dev	$accx\text{-}std = \mathrm{SD}[a_{x,1}, ...a_{x,T}]$	Level of volatility in longitudinal acceleration
	Skewness	$accx\text{-}skew = \mathrm{Skew}[a_{x,1}, ...a_{x,T}]$	Skewness of the longitudinal acceleration
	Kurtosis	$accx\text{-}kurt = \mathrm{Kurt}[a_{x,1}, ...a_{x,T}]$	Kurtosis of the longitudinal acceleration
Y-Axis Acceleration	Initiation	$accy\text{-}init = a_{y,1}$	Lateral acceleration at initial time
	Mean	$accy\text{-}mean = \frac{1}{T}\sum_{t=1}^{T} a_{y,t}$	Average of the lateral acceleration
	Maximum	$accy\text{-}max = \max[a_{y,1}, ...a_{y,T}]$	Maximum of lateral acceleration
	Minimum	$accy\text{-}min = \min[a_{y,1}, ...a_{y,T}]$	Minimum of lateral acceleration
	Std Dev	$accy\text{-}std = \mathrm{SD}[a_{y,1}, ...a_{y,T}]$	Standard deviation of lateral acceleration
	Skewness	$accy\text{-}skew = \mathrm{Skew}[a_{y,1}, ...a_{y,T}]$	Skewness of the lateral acceleration
	Kurtosis	$accy\text{-}kurt = \mathrm{Kurt}[a_{y,1}, ...a_{y,T}]$	Kurtosis of the lateral acceleration
Z-Axis Acceleration	Initiation	$accz\text{-}init = a_{z,1}$	Horizontal acceleration at initial time
	Mean	$accz\text{-}mean = \frac{1}{T}\sum_{t=1}^{T} a_{z,t}$	Average of the horizontal acceleration
	Maximum	$accz\text{-}max = \max[a_{z,1}, ...a_{z,T}]$	Maximum of horizontal acceleration
	Minimum	$accz\text{-}min = \min[a_{z,1}, ...a_{z,T}]$	Minimum of horizontal acceleration
	Std Dev	$accz\text{-}std = \mathrm{SD}[a_{z,1}, ...a_{z,T}]$	Standard deviation of horizontal acceleration
	Skewness	$accz\text{-}skew = \mathrm{Skew}[a_{z,1}, ...a_{z,T}]$	Skewness of the horizontal acceleration
	Kurtosis	$accz\text{-}kurt = \mathrm{Kurt}[a_{z,1}, ...a_{z,T}]$	Kurtosis of the horizontal acceleration
Driving Speed	Initiation	$speed\text{-}init = v_1$	Driving speed at initial time of the case
	Mean	$speed\text{-}mean = \frac{1}{T}\sum_{t=1}^{T} v_t$	Average of speed within the case
	Maximum	$speed\text{-}max = \max[v_1, ...v_T]$	Maximum of speed within the case
	Minimum	$speed\text{-}min = \min[v_1, ...v_T]$	Minimum of speed within the case
	Std Dev	$speed\text{-}std = \mathrm{SD}[v_1, ...v_T]$	Standard deviation of speed within the case
	Skewness	$speed\text{-}skew = \mathrm{Skew}[v_1, ...v_T]$	Skewness of speed within the case
	Kurtosis	$speed\text{-}kurt = \mathrm{Kurt}[v_1, ...v_T]$	Kurtosis of speed within the case
Relative X-Axis Position	Initiation	$xpos\text{-}init = \Delta s_{x,1}$	Relative longitudinal distance at initial time
	Mean	$xpos\text{-}mean = \frac{1}{T}\sum_{t=1}^{T} \Delta s_{x,t}$	Average of relative longitudinal distance
	Maximum	$xpos\text{-}max = \max[\Delta s_{x,1}, ...\Delta s_{x,T}]$	Largest relative longitudinal distance
	Minimum	$xpos\text{-}min = \min[\Delta s_{x,1}, ...\Delta s_{x,T}]$	Closest relative longitudinal distance
	Std Dev	$xpos\text{-}std = \mathrm{SD}[\Delta s_{x,1}, ...\Delta s_{x,T}]$	Level of interactions in relative longitudinal position
Relative Y-Axis Position	Initiation	$ypos\text{-}init = \Delta s_{y,1}$	Relative lateral distance at initial time
	Mean	$ypos\text{-}mean = \frac{1}{T}\sum_{t=1}^{T} \Delta s_{y,t}$	Average of relative lateral distance
	Maximum	$ypos\text{-}max = \max[\Delta s_{y,1}, ...\Delta s_{y,t}]$	Maximum of relative lateral distance
	Minimum	$ypos\text{-}min = \min[\Delta s_{y,1}, ...\Delta s_{y,T}]$	Minimum of relative lateral distance
	Std Dev	$ypos\text{-}std = \mathrm{SD}[\Delta s_{y,1}, ...\Delta s_{y,T}]$	Level of interactions in relative lateral position
Relative X-Axis Velocity	Initiation	$xvel\text{-}init = \Delta v_{x,1}$	Relative longitudinal velocity at initial time
	Mean	$xvel\text{-}mean = \frac{1}{T}\sum_{t=1}^{T} \Delta v_{x,t}$	Average of longitudinal velocity
	Maximum	$xvel\text{-}max = \max[\Delta v_{x,1}, ...\Delta v_{x,T}]$	Maximum of longitudinal velocity
	Minimum	$xvel\text{-}min = \min[\Delta v_{x,1}, ...\Delta v_{x,T}]$	Minimum of longitudinal velocity
	Std Dev	$xvel\text{-}std = \mathrm{SD}[\Delta v_{x,1}, ...\Delta v_{x,T}]$	Level of interactions in relative longitudinal velocity
Relative Y-Axis Velocity	Initiation	$yvel\text{-}init = \Delta v_{y,1}$	Relative lateral velocity at initial time
	Mean	$yvel\text{-}mean = \frac{1}{T}\sum_{t=1}^{T} \Delta v_{y,t}$	Average of relative lateral velocity
	Maximum	$yvel\text{-}max = \max[\Delta v_{y,1}, ...\Delta v_{y,T}]$	Maximum of relative lateral velocity
	Minimum	$yvel\text{-}min = \min[\Delta v_{y,1}, ...\Delta v_{y,T}]$	Minimum of relative lateral velocity
	Std Dev	$yvel\text{-}std = \mathrm{SD}[\Delta v_{y,1}, ...\Delta v_{y,T}]$	Level of interactions in relative lateral velocity

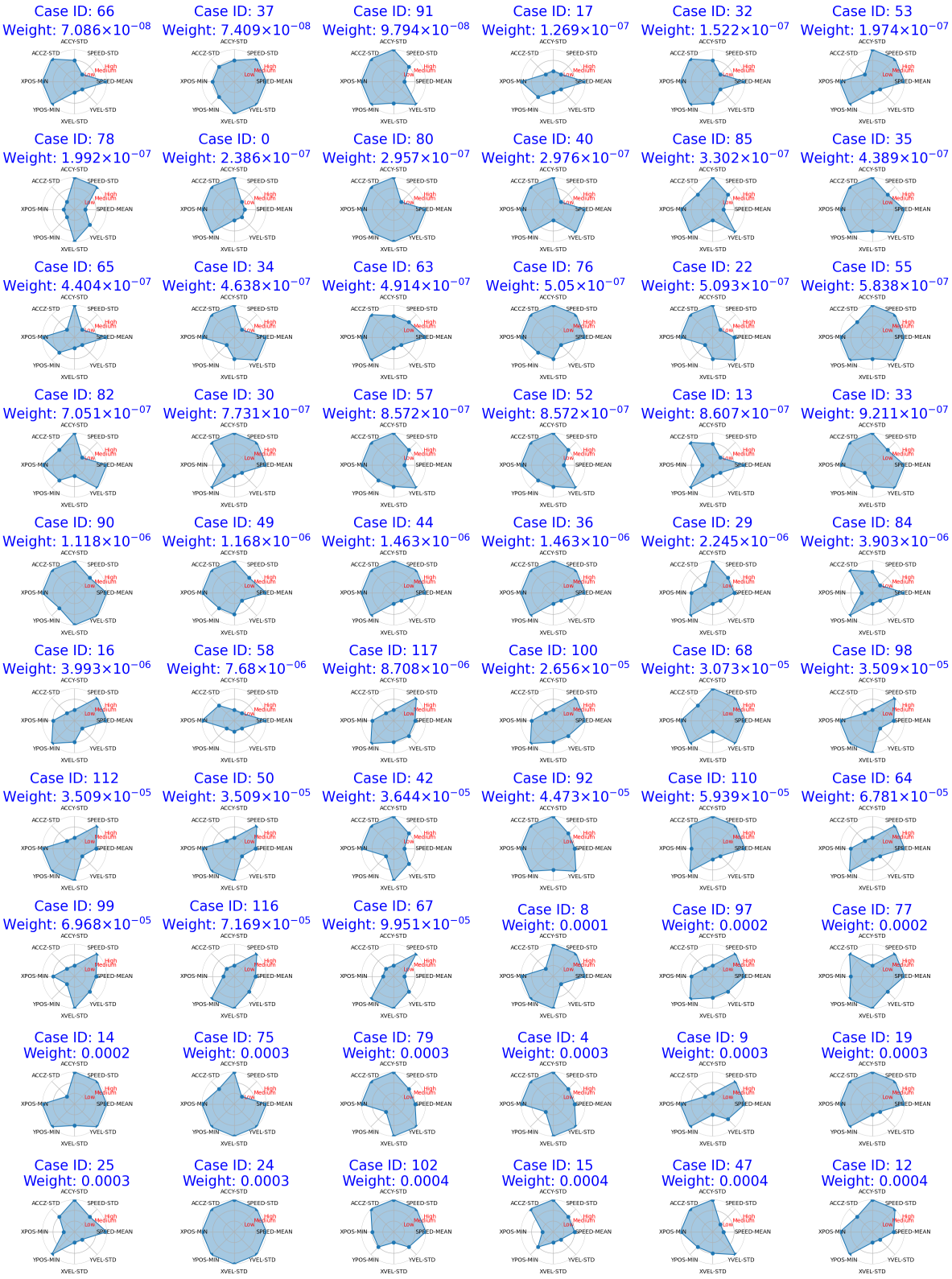

Figure 1. Part 1 of the 118 selected cases with corresponding weight and selected features.

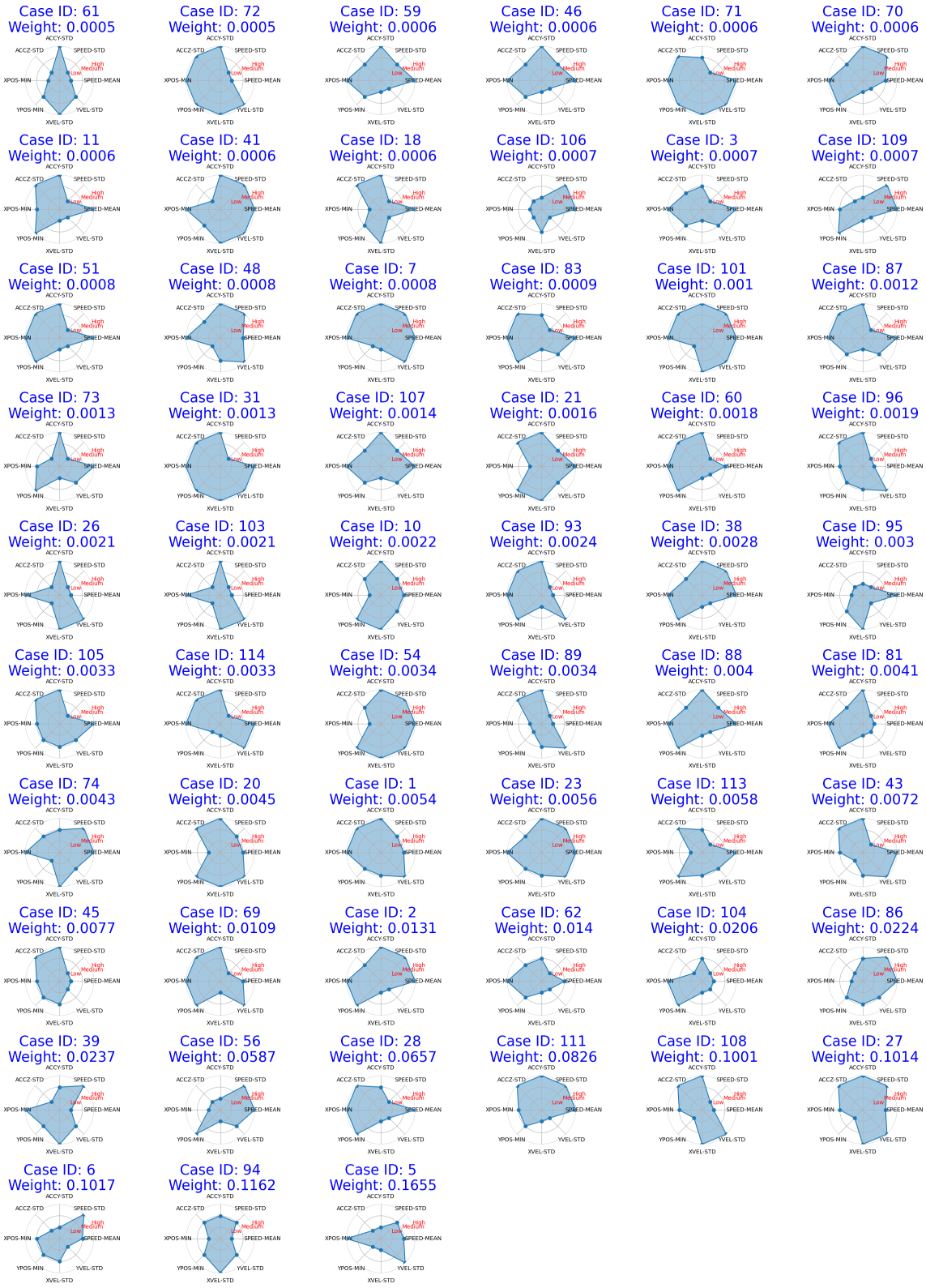

Figure 2. Part 2 of the 118 selected cases with corresponding weight and selected features.

Proof for Theorem 1

Proof. For simplicity, we define $w_{\theta_{\text{opt}}}(\boldsymbol{x}_i) = \psi_i$. Based on this, we can have the follows:

$$\mathscr{E}_X(\theta_{\text{opt}}) = \sum_{i=1}^{N}\sum_{j\neq i}^{N} w_{\theta_{\text{opt}}}(\boldsymbol{x}_i) w_{\theta_{\text{opt}}}(\boldsymbol{x}_j)\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \sum_{i=1}^{N}\sum_{j\neq i}^{N} \psi_i\psi_j\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{6}$$

Similar to the definition in[12], we define π_i and as the factual importance measure approximated through Pareto order sampling and ψ_i as the target importance measure. In order to investigate the asymptotic behaviors of $\widehat{\text{IP}}_{\boldsymbol{z}}$, we can decompose the difference between $\widehat{\text{IP}}_{\boldsymbol{z}}$ and $\mathscr{E}_X(\theta_{\text{opt}})$ as the follows:

$$\widehat{\text{IP}}_{\boldsymbol{z}} - \mathscr{E}_X(\theta_{\text{opt}}) = \underbrace{\widehat{\text{IP}}_{\boldsymbol{z}} - \sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_i\pi_j\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)}_{T_1\text{ Term}} + \underbrace{\sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_i\pi_j\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) - \mathscr{E}_X(\theta)}_{T_2\text{ Term}} \tag{7}$$

So as investigating the asymptotic behaviors of $\widehat{\text{IP}}_{\boldsymbol{z}} - \mathscr{E}_X(\theta_{\text{opt}})$ is equivalent with verifying T_1 and T_2 term individually. First we look at the T_1 term. We use $I(\boldsymbol{x}_i)$ to define whether $\boldsymbol{x}_i$ is included in the Pareto order sampling. With this, T_1 term can be decomposed as:

$$\begin{aligned}
T_1 =& \frac{1}{M(M-1)}\sum_{t=1}^{M}\sum_{s=1,s\neq t}^{M}\mathscr{K}(\boldsymbol{z}_t,\boldsymbol{z}_s) - \sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_i\pi_j\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)\\
=& \sum_{t=1}^{M}\sum_{s=1,s\neq t}^{M}\frac{1}{M(M-1)}\mathscr{K}(\boldsymbol{z}_t,\boldsymbol{z}_s) - \sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_i\pi_j\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)\\
=& \sum_{t=1}^{N}\sum_{s=1,s\neq t}^{N}\frac{I(\boldsymbol{x}_i)I(\boldsymbol{x}_j)}{M(M-1)}\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) - \sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_i\pi_j\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)\\
=& \sum_{t=1}^{N}\sum_{s=1,s\neq t}^{N}\left[\frac{I(\boldsymbol{x}_i)I(\boldsymbol{x}_j)}{M(M-1)} - \pi_i\pi_j\right]\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)
\end{aligned} \tag{8}$$

Here, based on the fact $\mathbb{E}[I(\boldsymbol{x}_i)] = M\pi_i$ that we can know that:

$$\mathbb{E}[I(\boldsymbol{x}_i)I(\boldsymbol{x}_j)] = \frac{M(M-1)}{N(N-1)} \times N(N-1)\pi_i\pi_j = \frac{M(M-1)}{N(N-1)}\pi_i'\pi_j' \tag{9}$$

So as we can have the following equations:

$$\mathbb{E}\left[\frac{N(N-1)}{M(M-1)}I(\boldsymbol{x}_i)I(\boldsymbol{x}_j) - \pi_i'\pi_j'\right] = 0 \tag{10}$$

Based on this equation, we can have the follows:

$$\frac{I(\boldsymbol{x}_i)I(\boldsymbol{x}_j)}{M(M-1)} - \pi_i\pi_j = \frac{1}{N(N-1)}\left[\frac{N(N-1)}{M(M-1)}I(\boldsymbol{x}_i)I(\boldsymbol{x}_j) - \pi_i'\pi_j'\right] \tag{11}$$

Then, based on Equation 11, the T_1 term can be transformed into the following:

$$\begin{aligned} T_1 &= \frac{1}{N(N-1)}\sum_{t=1}^{N}\sum_{s=1,s\neq t}^{N}\left[\frac{N(N-1)}{M(M-1)}I(\boldsymbol{x}_i)I(\boldsymbol{x}_j) - \pi_i'\pi_j'\right] \\ &= \frac{1}{N(N-1)}\sum_{i\neq j}\left[\frac{N(N-1)}{M(M-1)}I(\boldsymbol{x}_i)I(\boldsymbol{x}_j) - \pi_i'\pi_j'\right] \end{aligned} \tag{12}$$

Here we set that $\left[\frac{N(N-1)}{M(M-1)}I(\boldsymbol{x}_i)I(\boldsymbol{x}_j) - \pi_i'\pi_j'\right] \in [-u,u]$. For simplicity, we can define $N' = N(N-1)$ we apply the Lemma 4, we can have the following inequality:

$$\Pr\left(\frac{1}{N(N-1)}\sum_{i\neq j}\left[\frac{N(N-1)}{M(M-1)}I(\boldsymbol{x}_i)I(\boldsymbol{x}_j) - \pi_i'\pi_j'\right] \geq \varepsilon\right) \leq \exp\left(-N't\varepsilon + \frac{u^2t^2N'}{2}\right) \tag{13}$$

If we let $t = \varepsilon/u^2$ and $\varepsilon = \sqrt{\frac{-2u^2\log\delta}{N'}}$ where $\delta \in (0,1)$. The with at least a probability of $1-\delta$, we can have the following bound:

$$\frac{1}{N(N-1)}\sum_{i\neq j}\left[\frac{N(N-1)}{M(M-1)}I(\boldsymbol{x}_i)I(\boldsymbol{x}_j) - \pi_i'\pi_j'\right] \leq \sqrt{\frac{-2u^2\log\delta}{N'}} \tag{14}$$

As long as $u < N$, then $\sqrt{\frac{-2u^2\log\delta}{N'}} \to 0$ when $N \to 0$. Here, as the π_i denotes the factual importance measure through Pareto order sampling with the fact that $\pi > 0$ and $\sum_{i=1}^{N}\pi_i = 1$, so as $\pi_i = o(1/N)$, so as $\pi_i'\pi_j' = N(N-1)\pi_i\pi_j < N$. So as the condition that $u < N$ can be satisfied. Based on this, when N goes sufficiently large, we have $T_1 \to 0$

As for the T_2 term, from the equation (1.25) in[12], we can have:

$$\pi_i/\psi_i \to 1 \quad \text{as} \quad N \to \infty, \quad i = 1,...N \tag{15}$$

According to the continuous mapping theorem, we can have that:

$$\frac{\sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_i\pi_j\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sum_{i=1}^{N}\sum_{j\neq i}^{N}\psi_i\psi_j\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)} \to 1 \quad \text{as} \quad N \to \infty \tag{16}$$

which is equivalent as $T_2 \to 0$ when $N \to \infty$. Combining the previous results in the asymptotic behaviors in the T_1 term, we can know that:

$$\widehat{\mathrm{IP}}_{\boldsymbol{z}} / \mathscr{E}_X(\theta) \to 1 \quad \text{as} \quad N \to \infty \tag{17}$$

This completes the proof. □

0.1 Proof of Theorem 2

Proof. We first define $\mathscr{E}(\theta)$ as the population version of $\mathscr{E}_X\theta$ as follows:

$$\mathscr{E}(\theta) = \int_{\boldsymbol{x}\in\mathscr{X}} \int_{\boldsymbol{x}'\in\mathscr{X}} \mathscr{K}(\boldsymbol{x},\boldsymbol{x}') w_\theta(\boldsymbol{x}) w_\theta(\boldsymbol{x}') d\mathbb{P}(\boldsymbol{x},\boldsymbol{x}') \tag{18}$$

and $\theta^* = \underset{\theta\in\mathscr{B}^p}{\operatorname{argmin}} \mathscr{E}(\theta)$. We can take the differences between these two functions between $\mathscr{E}(\theta)$, which is the population version of $\mathscr{E}_X(\theta)$. The difference can be bounded as following:

$$\mathscr{E}(\theta) - \mathscr{E}_X(\theta) \leq \sup_{\theta\in\mathscr{B}^p} \left\{ \mathscr{E}(\theta) - \mathscr{E}_X(\theta) \right\} \tag{19}$$

And $\mathscr{E}_X(\theta)$ is calculated based on $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$, and here we replace $\boldsymbol{x}_i \in \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ with $\boldsymbol{x}_i'$. Based on these new sequences $(\boldsymbol{x}_1, ...\boldsymbol{x}_i', ..., \boldsymbol{x}_N)$ we can construct the calculations for $\mathscr{E}_{X'}(\theta)$. For simplicity, we can define $\Lambda(X) = \mathscr{E}(\theta) - \mathscr{E}_X(\theta)$ and similarly $\Lambda(X') = \mathscr{E}(\theta) - \mathscr{E}_{X'}(\theta)$. We further define $\eta_\theta(\boldsymbol{x}_i,\boldsymbol{x}_j) = w_\theta(\boldsymbol{x}_i) w_\theta(\boldsymbol{x}_j) \mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)$

Such that we can calculate the differences between $\Lambda(X)$ and the $\Lambda(X')$:

$$\begin{aligned} \left|\Lambda(X) - \Lambda(X')\right| &= \sup_{\theta\in\mathscr{B}^p} \left| \mathscr{E}_X(\theta) - \mathscr{E}_{X'}(\theta) \right| \\ &= \sup_{\theta\in\mathscr{B}^p} \sum_{j\neq k} \left| \eta_\theta(\boldsymbol{x}_i,\boldsymbol{x}_j) - \eta_\theta(\boldsymbol{x}_i',\boldsymbol{x}_j) \right| \\ &\leq \sup_{\theta\in\mathscr{B}^p} \sum_{j\neq k} \Big(\left|\eta_\theta(\boldsymbol{x}_i,\boldsymbol{x}_j)\right| + \left|\eta_\theta(\boldsymbol{x}_i',\boldsymbol{x}_j)\right| \Big) \end{aligned} \tag{20}$$

Here, we assume that $w_\theta(\boldsymbol{x}) \leq s/N$ for any $\boldsymbol{x} \in \mathscr{X}$. This can simply be achieved by doing the threshold for the output of the approximated function $w_\theta(\cdot)$. Now the key question turns to investigating the upper bound for $\eta_\theta(\boldsymbol{x}_i,\boldsymbol{x}_j)$ and $\eta_\theta(\boldsymbol{x}_i',\boldsymbol{x}_j)$:

$$\eta_\theta(\boldsymbol{x}_i,\boldsymbol{x}_j) = w_\theta(\boldsymbol{x}_i) w_\theta(\boldsymbol{x}_j) \mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) \leq \frac{s^2}{N^2} \sup_{\boldsymbol{x},\boldsymbol{x}'} \mathscr{K}(\boldsymbol{x},\boldsymbol{x}') = \frac{s^2 K}{N^2} \tag{21}$$

Based on this and the results from Equation 21, we can have the following:

$$\left|\Lambda(X)-\Lambda(X')\right| \leq \frac{2s^2K}{N} \tag{22}$$

Applying the McDiamard inequality in Lemma 3, we can have the follows:

$$\Pr\left(|\Lambda(X)-\mathbb{E}(\Lambda(X)|>\Delta\right) \leq 2\exp\left(\frac{-N\Delta^2}{2s^4K^2}\right) = \frac{\delta}{(1+2/\varepsilon)^p} \tag{23}$$

From equation 23, we can know the follows toward Δ:

$$\Delta = \frac{s^2K}{\sqrt{N}}\sqrt{2\log(2/\delta)+2p\log\left(1+\frac{2}{\varepsilon}\right)} \tag{24}$$

Let Θ be an ε-cover of $\mathbb{B}^p$. Here, this means that $\forall\theta\in\mathbb{B}^p\ \exists\ \theta'\in\Theta$ such that $\|\theta-\theta'\|\leq\varepsilon$. By Lemma 5.2 of[13] we have $|\Theta|\leq(1+2/\varepsilon)^d$. Using the above and union bounding over all elements of $\mathscr{A}$ and all k, we have:

$$\Pr\left\{\forall\theta\in\Theta:\left|\Lambda(X)-\mathbb{E}(\Lambda(X))\right|\leq\Delta\right\}\geq 1-\delta \tag{25}$$

Combining equation 24 and 25, with at least probability of $1-\delta$ and for any $\theta\in\Theta$:

$$\mathscr{E}(\theta)-\mathscr{E}_X(\theta) \leq \mathbb{E}\Big\{\sup_{\theta\in\mathscr{B}^p}\left(\mathscr{E}(\theta)-\mathscr{E}_X(\theta)\right)\Big\}+\frac{s^2K}{\sqrt{N}}\sqrt{2\log(2/\delta)+2p\log\left(1+\frac{2}{\varepsilon}\right)} \tag{26}$$

Next, we can investigate the upper bound for the $\mathbb{E}\Big\{\sup_{\theta\in\mathscr{B}^p}\left(\mathscr{E}(\theta)-\mathscr{E}_X(\theta)\right)\Big\}$. Through the symmetrization technical in[14], we can have the following:

$$\begin{aligned}\mathbb{E}\Big\{\sup_{\theta\in\mathscr{B}^p}\left(\mathscr{E}(\theta)-\mathscr{E}_{\boldsymbol{x}}(\theta)\right)\Big\} &\leq \mathbb{E}_{\boldsymbol{x},\sigma}\frac{1}{\lfloor N/2\rfloor}\sup_{\theta\in\mathscr{B}^p}\left|\sum_{i=1}^{\lfloor N/2\rfloor}\sigma_i\eta_\theta\left(\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{x}}_{\lfloor\frac{N}{2}\rfloor+i}\right)\right| \\ &\leq \mathbb{E}_{\boldsymbol{x},\sigma}\frac{1}{\lfloor N/2\rfloor}\sup_{\theta\in\mathscr{B}^p}\left|\sum_{i=1}^{\lfloor N/2\rfloor}\sigma_i\xi_i\right| = R_N\end{aligned} \tag{27}$$

Here, we define $\sigma_1,...\sigma_N$ to be a sequence of Rademacher variables and for each σ_i. we have $\Pr\{\sigma_i=1\}=\Pr\{\sigma_i=-1\}=0.5$. We further define $\xi_i=\mathscr{K}(\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{x}}_{\lfloor N/2\rfloor+i})$ and $\bar{\boldsymbol{x}}_1,...,\bar{\boldsymbol{x}}_N$ represents an i.i.d independent copy of $\boldsymbol{x}_1,...,\boldsymbol{x}_N$. The notation $\lfloor N/2\rfloor$ denotes the largest integer less than N and the notation R_N defines the empirical Rademacher complexity.

With this, we can have the following bound with at least a probability of $1-\delta$:

$$\mathscr{E}(\theta)-\mathscr{E}_X(\theta)\leq R_N+\frac{s^2K}{\sqrt{N}}\sqrt{2\log(2/\delta)+2p\log\left(1+\frac{2}{\varepsilon}\right)} \tag{28}$$

Combining the bound in equation 19, we can have with at least probability of $1-\delta$:

$$\mathscr{E}(\theta^*)-\mathscr{E}_X(\theta_{\text{opt}})\leq R_N+\frac{s^2K}{\sqrt{N}}\sqrt{2\log(2/\delta)+2p\log\left(1+\frac{2}{\varepsilon}\right)} \tag{29}$$

This completes the proof. □

Proof for Theorem 3

Proof. Here, this empirical density estimation can be constructed as $\widehat{\mu}_N=\frac{1}{N}\sum_{i=1}^{N}\phi(\boldsymbol{x}_i)$ and $\mu=\int\phi(\boldsymbol{x})\,\mathrm{d}\,\mathbb{P}(\boldsymbol{x})$. The estimator based on the proposed KTCS is defined $\|\widehat{\mu}_M-\mu\|_{\mathscr{H}}$, and we define the constructed kernel mean embedding based on the ground truth weights $\lambda_1^*,...\lambda_M^*$ is as $\widehat{\mu}_{M^*}=\sum_{j=1}^{M}\lambda_j^*\phi(\boldsymbol{z}_j)$. In order to generate an upper bound for $\|\widehat{\mu}_M-\mu\|_{\mathscr{H}}$, we can constitute the error decomposition as follows:

$$\|\widehat{\mu}_M-\mu\|_{\mathscr{H}}\leq\|\widehat{\mu}_{M^*}-\widehat{\mu}_M\|_{\mathscr{H}}+\|\widehat{\mu}_{M^*}-\widehat{\mu}_N\|_{\mathscr{H}}+\|\widehat{\mu}_N-\mu\|_{\mathscr{H}} \tag{30}$$

First, let's look at the third component $\|\widehat{\mu}_N-\mu\|_{\mathscr{H}}$, which investigates the concentration ability of $\widehat{\mu}_N$ around its mean μ on a Hilbert space:

$$\|\widehat{\mu}_N-\mu\|_{\mathscr{H}}=\left\|\frac{1}{N}\sum_{i=1}^{N}\left[\phi(\boldsymbol{x}_i)-\mu\right]\right\|_{\mathscr{H}} \tag{31}$$

Furthermore, we can have $\|\phi(\boldsymbol{x}_i)-\mu\|\leq 2\sup_{\boldsymbol{x}\in\mathscr{X}}\|\phi(\boldsymbol{x})\|=2\sup_{\boldsymbol{x}\in\mathscr{X}}\|\mathscr{K}(\boldsymbol{x},\cdot)\|=2K$, indicating that each individual term is bounded by $2K$. With this bounded term, we can apply the the Lemma 1 with at least $1-\delta/3$ probability to have the following bound

$$\|\widehat{\mu}_N-\mu\|_{\mathscr{H}}\leq\frac{2K\sqrt{2\log(6/\delta)}}{\sqrt{N}} \tag{32}$$

Next, we move to investigate the bound toward the second term. Considering the underlying covariance structure among the selected points $\boldsymbol{z}_1,...,\boldsymbol{z}_M$, we define $C:=\int\phi(\boldsymbol{x})\otimes_{\mathscr{H}}\phi(\boldsymbol{z})d\mathbb{P}(\boldsymbol{z})$. To avoid singularity in C and C^{-1}, we introduce $C_\lambda=C+\lambda I$

where I is an identity matrix. And for the first component, we can define:

$$\begin{aligned} \forall \boldsymbol{z} \in \mathscr{Z}, \mathscr{N}_{\boldsymbol{z}}(\lambda) &:= \left\langle \phi(\boldsymbol{z}), C_{\lambda}^{-1}\phi(\boldsymbol{z}) \right\rangle_{\mathscr{H}} \\ \mathscr{N}_{\infty}(\lambda) &:= \sup_{\boldsymbol{z}\in\mathscr{Z}} \mathscr{N}_{\boldsymbol{z}}(\lambda) \end{aligned} \tag{33}$$

Based on these, we extend the second component as:

$$\|\widehat{\mu}_N - \widehat{\mu}_{M^*}\|_{\mathscr{H}} = \|(I-P_M)\widehat{\mu}_N\|_{\mathscr{H}} = \|(I-P_M)(\widehat{\mu}_N - \widetilde{\mu}_M)\|_{\mathscr{H}} \tag{34}$$

where here $\widetilde{\mu}_M = \frac{1}{M}\sum_{j=1}^{M}\phi(\boldsymbol{z}_j)$ is defined to be the average of feature map aggregations.

Based on this C_λ matrix, we can generate the following bound:

$$\|(I-P_M)(\widehat{\mu}_N - \widetilde{\mu}_M)\|_{\mathscr{H}} \leq \left\|(I-P_M)C_{\lambda}^{1/2}\right\|_{\mathscr{H}} \left\|C_{\lambda}^{-1/2}(\widehat{\mu}_N - \widetilde{\mu}_M)\right\|_{\mathscr{H}} \tag{35}$$

In the proceeding part, we will derive an upper bound for the $\left\|C_{\lambda}^{-1/2}(\widehat{\mu}_N - \widetilde{\mu}_M)\right\|_{\mathscr{H}}$ part. As the selected points $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M$ are independent from the original dataset $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$. For the random variable $\boldsymbol{z}$, its expectation is calculated as follows:

$$\mathbb{E}\big[\phi(\boldsymbol{z})\big] = \int \phi(\boldsymbol{z}) d\mathbb{P}(\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^{N}\phi(\boldsymbol{x}_i) = \widehat{\mu}_N \tag{36}$$

Based on the calculation of the $\mathbb{E}\big[\phi(\boldsymbol{z})\big]$, we can decompose the $\left\|C_{\lambda}^{-1/2}(\widehat{\mu}_N - \widetilde{\mu}_M)\right\|_{\mathscr{H}}$ as:

$$\left\|C_{\lambda}^{-1/2}(\widehat{\mu}_N - \widetilde{\mu}_M)\right\|_{\mathscr{H}} = \left\|\frac{1}{M}\sum_{j=1}^{M}\left(C_{\lambda}^{-1/2}\phi(\boldsymbol{z}_j) - C_{\lambda}^{-1/2}\widehat{\mu}_N\right)\right\|_{\mathscr{H}} \tag{37}$$

For simplicity, we define $\xi_j = C_{\lambda}^{-1/2}\phi(\boldsymbol{z}_j)$, and it is easy to verify that $\mathrm{E}(\xi_j) = C_{\lambda}^{-1/2}\widehat{\mu}_N$ for any $j = [1, ..., M]$. With the application of the Cauthy-Schwartz inequality, we can derive the upper bound for the second order moment condition for random variable ξ_i

$$\begin{aligned} \mathbb{E}\|\zeta_i - \mathbb{E}[\zeta_i]\|_{\mathscr{H}}^2 &= \mathbb{E}\left\|C_{\lambda}^{-1/2}(\phi(\boldsymbol{z}_j) - \widehat{\mu}_N)\right\|_{\mathscr{H}}^2 \\ &\leq \mathbb{E}(\|C_{\lambda}^{-1/2}\|_{\mathscr{H}})\mathbb{E}(\|\phi(\boldsymbol{z}_j) - \widehat{\mu}_N\|_{\mathscr{H}}) = \nu_{\boldsymbol{z}}\,\mathrm{tr}(C_{\lambda}^{-1/2}) \end{aligned} \tag{38}$$

Here, $\nu_{\boldsymbol{z}}^2$ is the variance of $\boldsymbol{z}$ in the RKHS as introduced in[15]. Here, $\nu_{\boldsymbol{z}}^2 = \mathbb{E}(\|\phi(\boldsymbol{z}_j) - \widehat{\mu}_N\|_{\mathscr{H}}^2)$. For simplicity, we define

$c = \mathrm{tr}(C_\lambda^{-1/2})$. Then:

$$\mathbb{E}\,\|\zeta_i - \mathbb{E}\,[\zeta_i]\|_{\mathscr{H}}^2 \leq c\nu_z \tag{39}$$

After considering the second moment condition, we investigate the upper bound for higher order moments $\mathbb{E}\,\|\zeta_i - \mathbb{E}\,[\zeta_i]\|_{\mathscr{H}}^p$, when $p \geq 3$:

$$\begin{aligned}\mathbb{E}\,\|\zeta_i - \mathbb{E}\,[\zeta_i]\|_{\mathscr{H}}^p &= \mathbb{E}\,\|\zeta_i - \mathbb{E}\,[\zeta_i]\|_{\mathscr{H}}^2 \times \mathbb{E}\,\|\zeta_i - \mathbb{E}\,[\zeta_i]\|_{\mathscr{H}}^{p-2} \\ &\leq \nu\,\mathrm{tr}(C_\lambda^{-1/2})\,(\mathrm{ess\,sup}\,\|\zeta_i - \mathbb{E}\,[\zeta_i]\|_{\mathscr{H}})^{p-2} \\ &\leq \frac{1}{2}p!c\nu_z\left(2N_{Q,\infty}(\lambda)^{1/2}\right)^{p-2}\end{aligned} \tag{40}$$

as $\frac{1}{2}p! \geq 1$ always holds whenever $p \geq 3$. Combining the equations 39 and 40, we have:

$$\mathbb{E}\,\|\zeta_i - \mathbb{E}\,[\zeta_i]\|_{\mathscr{H}}^p \leq \frac{1}{2}p!c\nu_z\left(2N_{Q,\infty}(\lambda)^{1/2}\right)^{p-2} \tag{41}$$

holds for all $p \geq 2$. With the application of Lemma 2, with at least a probability of $1 - \delta/6$, we can get the following upper bound:

$$\left\|C_\lambda^{-1/2}\left(\widetilde{\mu}_M - \widehat{\mu}_N\right)\right\|_{\mathscr{H}} \leq \frac{4\sqrt{\mathscr{N}_\infty(\lambda)}\log(12/\delta)}{M} + \sqrt{\frac{2c\nu_z\log(12/\delta)}{M}} \tag{42}$$

Adopting the results from[10], with at least probability of $1 - \delta/6$:

$$\left\|(I - P_M)C_\lambda^{1/2}\right\|_{\mathscr{H}} \leq \sqrt{3\lambda} \tag{43}$$

Combining equation 42 and 43, with at least probability of $1 - \delta/3$, the follow holds:

$$\|\widehat{\mu}_N - \widehat{\mu}_{M^*}\|_{\mathscr{H}} \leq \sqrt{3\lambda} \times \left(\frac{4\sqrt{\mathscr{N}_\infty(\lambda)}\log(12/\delta)}{M} + \sqrt{\frac{2c\nu_z\log(12/\delta)}{M}}\right) \tag{44}$$

In the last step, we investigate the upper bound for the first part. We define the parameters are trying to optimize the $\boldsymbol{\lambda}^* = \mathrm{argmin}_{\mathbf{w}}\,\|\Phi_M\mathbf{w} - \widehat{\mu}_N\|_{\mathscr{H}}$, this is equivalent with minimizing the associated MMD values. And $\boldsymbol{\lambda}^* = (\lambda_1^*, ..., \lambda_M^*)$. We

can achieve this $\boldsymbol{\lambda}^* = (\Phi_M^* \Phi_M)^+ \Phi_M^* \widehat{\mu}_N$. We work on the bound:

$$\|\widehat{\mu}_M - \widehat{\mu}_{M^*}\|_{\mathscr{H}} = \|\Phi_M \boldsymbol{\lambda} - \Phi_M \boldsymbol{\lambda}^*\|_{\mathscr{H}} = \Big\| \sum_{j=1}^{M} (\lambda_j - \lambda_j^*) \phi(\boldsymbol{z}_j) \Big\|_{\mathscr{H}} \tag{45}$$

We make the assumption that $|\lambda_j - \lambda_j^*| \leq c'/M$. The rationale behind this assumption is when $\|\boldsymbol{\lambda}\|_1 = 1$ and there are M sub-components within the vector $\boldsymbol{\lambda}$, we aim to limit the deviation of λ_j from λ_j^* to prevent excessively large disparities. Large deviations $|\lambda_j - \lambda_j^*|$ in could potentially lead to trivial solutions, where a few points dominate all the weights within $\boldsymbol{\lambda}$. With this, we have:

$$\Big\| \sum_{j=1}^{M} (\lambda_j - \lambda_j^*) \phi(\boldsymbol{z}_j) \Big\|_{\mathscr{H}} \leq \Big\| \frac{1}{M} \sum_{j=1}^{M} c' \phi(\boldsymbol{z}_j) \Big\|_{\mathscr{H}} \tag{46}$$

Meanwhile, according to the definitions of the feature map, the upper bound for the $\phi(\boldsymbol{z}_j)$ is as $\phi(\boldsymbol{z}_j) = \mathscr{K}(\boldsymbol{z}_j, \cdot) \leq K$. Based on the Lemma 1, with at least a probability of $1 - \delta/3$:

$$\Big\| \frac{1}{M} \sum_{j=1}^{M} c' \phi(\boldsymbol{z}_j) \Big\|_{\mathscr{H}} \leq c' K \frac{\sqrt{2\log(2/\delta)}}{\sqrt{M}} \tag{47}$$

Based on equation 46, this implies with at least a probability of $1 - \delta/3$:

$$\|\widehat{\mu}_M - \widehat{\mu}_{M^*}\|_{\mathscr{H}} \leq c' K \frac{\sqrt{2\log(6/\delta)}}{\sqrt{M}} \tag{48}$$

As we can have $\left(1 - \frac{\delta}{3}\right)^3 \geq (1 - \delta)$ so combining all the previous results in equation 32 44 and 48, we can have the follows with at least probability of $1 - \delta$:

$$\begin{aligned} \|\widehat{\mu}_M - \mu\|_{\mathscr{H}} \leq & \frac{2K\sqrt{2\log(6/\delta)}}{\sqrt{N}} + c' K \frac{\sqrt{2\log(6/\delta)}}{\sqrt{M}} \\ & + \sqrt{3\lambda} \times \left(\frac{4\sqrt{\mathscr{N}_\infty(\lambda)}\log(12/\delta)}{M} + \sqrt{\frac{2c\nu_{\boldsymbol{z}}\log(12/\delta)}{M}} \right) \end{aligned} \tag{49}$$

Proof of Remark 1: We can make several further assumptions toward the parameters investigated above, to discuss more on the order of M here. Given the fact that:

$$\mathscr{N}_\infty(\lambda) = \sup_{\boldsymbol{z} \in \mathscr{Z}} \mathscr{N}_{\boldsymbol{z}}(\lambda) = \sup_{\boldsymbol{z} \in \mathscr{Z}} \left\langle \phi(\boldsymbol{z}), C_\lambda^{-1} \phi(\boldsymbol{z}) \right\rangle \leq \frac{K^2}{\lambda} \tag{50}$$

So as for the third part of the equation 49, we can have the bound as:

$$\sqrt{3\lambda} \times \frac{4\sqrt{\mathcal{N}_\infty(\lambda)}\log(12/\delta)}{M} \leq \frac{4\sqrt{3}K\log(12/\delta)}{M} \tag{51}$$

And we make further assumptions toward $c = \mathrm{tr}(C_\lambda^{-1/2})$. As the parameter c is a constant determined by λ, here we assume the c satisfying a polynomial decaying toward *lambda* as $c = \mathrm{tr}(C_\lambda^{-1/2}) = c_\lambda \lambda^{1-\gamma}$, here c_λ is some constant and $\gamma \in (0,1)$. Similar assumptions have been adopted in[10]. As λ is some constant, here we set λ inversely proportion to the size M. The larger the size M is, the smaller λ will be. We assume $\lambda = (\frac{1}{M})^{1-\gamma}$, and with all of these, we can have the bound:

$$\sqrt{3\lambda} \times \sqrt{\frac{2c\nu_{\mathbf{z}}\log(12/\delta)}{M}} = \frac{\sqrt{6c_\gamma\log(12/\delta)}}{M} \tag{52}$$

Combining the equation 51 and equation 52, we can have the following:

$$\sqrt{3\lambda} \times \left(\frac{4\sqrt{\mathcal{N}_\infty(\lambda)}\log(12/\delta)}{M} + \sqrt{\frac{2c\nu_{\mathbf{z}}\log(12/\delta)}{M}} \right) = \mathcal{O}\left(\frac{1}{M}\right) \tag{53}$$

Furthermore, we assume the deviance between $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^*$ satisfying $M|\lambda_j - \lambda_j^*| \leq \frac{1}{M^{1/2}}$, indicating that $c' = \mathcal{O}(M^{-1/2})$ Similar conditions and theoretical results can be found in[16]. Then the following will satisfy:

$$c'K\frac{\sqrt{2\log(6/\delta)}}{\sqrt{M}} = \mathcal{O}\left(\frac{1}{M}\right) \tag{54}$$

If we set the the size for subselected samples $M = \mathcal{O}(N^{1/2})$, then $\|\widehat{\mu}_M - \mu\|_{\mathscr{H}} = \mathcal{O}(\frac{1}{\sqrt{N}})$ □

References

1. Li, T. & Yuan, M. On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302* (2019).

2. Mak, S. & Joseph, V. R. Support points. *The Annals Stat.* **46**, 2562–2592 (2018).

3. Zhang, J. *et al.* An optimal transport approach for selecting a representative subsample with application in efficient kernel density estimation. *J. Comput. Graph. Stat.* **32**, 329–339 (2023).

4. Kim, B., Khanna, R. & Koyejo, O. O. Examples are not enough, learn to criticize. criticism for interpretability. *Adv. Neural Inf. Process. Syst.* **29** (2016).

5. Joseph, V. R., Dasgupta, T., Tuo, R. & Wu, C. J. Sequential exploration of complex surfaces using minimum energy designs. *Technometrics* **57**, 64–74 (2015).

6. Chen, Y., Gao, Q. & Wang, X. Inferential wasserstein generative adversarial networks. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **84**, 83–113 (2022).

7. Guo, F. Statistical methods for naturalistic driving studies. *Annu. Rev. Stat. its Appl.* **6**, 309–328 (2019).

8. Pinelis, I. Optimum bounds for the distributions of martingales in banach spaces. *The Annals Probab.* 1679–1706 (1994).

9. Yurinsky, V. *Sums and Gaussian vectors* (Springer, 2006).

10. Chatalic, A., Schreuder, N., Rosasco, L. & Rudi, A. Nyström kernel mean embeddings. In *International Conference on Machine Learning*, 3006–3024 (PMLR, 2022).

11. Rudi, A., Camoriano, R. & Rosasco, L. Less is more: Nyström computational regularization. *Adv. Neural Inf. Process. Syst.* **28** (2015).

12. Rosén, B. On inclusion probabilities for order πps sampling. *J. Stat. Plan. Inference* **90**, 117–143 (2000).

13. Eldar, Y. C. & Kutyniok, G. *Compressed Sensing : Theory and Applications* (Cambridge University Press, Cambridge, United Kingdom, 2012).

14. Bartlett, P. L. & Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3**, 463–482 (2002).

15. Wolfer, G. & Alquier, P. Variance-aware estimation of kernel mean embedding. *arXiv preprint arXiv:2210.06672* (2022).

16. Liu, W., Yu, X., Zhong, W. & Li, R. Projection test for mean vector in high dimensions. *J. Am. Stat. Assoc.* 1–13 (2022).