6 Appendix

6.1 Proof of Lemma 1

Lemma 1 (Lipschitz parameter)**.** *Let $\tau(A\omega, Bx)$ be Lipschitz-continuous w.r.t. x with parameter $\beta(\omega)$. Then given $\|x_1 - x_2\| \leq \epsilon$ and $\|y_1 - y_2\| \leq \epsilon$, we have*

$$f(x, y) = s_R(x, y) - k(x, y)$$

satisfies

$$|f(x, y) - f(x', y')| \leq 2\gamma\sigma_R\epsilon t.$$

with probability at least $1 - 1/t^2$, where $\sigma_\tau^2 = Var[\beta(\omega)]/R$ is the variance of the Lipschitz parameter averaged over R samples.

Proof. Consider an arbitrary pair of series $(x_1, x_2) \in \mathcal{X}$ of the same length. From Lipschitz-continuity of $\tau(.)$, we have

$$\tau(A\omega, Bx_2) = \tau(A\omega, Bx_1) + \beta(\omega)\Delta$$

for some $|\Delta| \leq \|x_2 - x_1\|$. Then let (A_1, B_1) and (A_2, B_2) be the minimizers of $\tau(A\omega, Bx_1)$ and $\tau(A\omega, Bx_2)$ respectively, we have

$$\tau(A_2\omega, B_2x_2) \leq \tau(A_1\omega, B_1x_2) \leq \tau(A_1\omega, B_1x_1)+\beta(\omega)|\Delta|.$$

and

$$\tau(A_1\omega, B_1x_1) \leq \tau(A_2\omega, B_2x_1) \leq \tau(A_2\omega, B_2x_2)+\beta(\omega)|\Delta|.$$

Therefore, $\phi_\omega(x_2) = \phi_\omega(x_1) + \beta(\omega)\Delta$ and

$$\begin{aligned} s_R(x_2, y_2) &= \frac{1}{R}\sum_{i=1}^{R}\phi_{\omega_i}(x_2)\phi_{\omega_i}(y_2) \\ &\leq \frac{1}{R}\sum_{i=1}^{R}\phi_{\omega_i}(x_1)\phi_{\omega_i}(y_2) + r\tilde{\beta}\epsilon \leq s_R(x_1, y_1) + 2r\tilde{\beta}\epsilon. \end{aligned}$$

where $\tilde{\beta} = \frac{1}{R}\sum_{i=1}^{R}\beta(\omega_i)$. With the similar argument we have $|s_R(x_2, y_2) - s_R(x_1, y_1)| \leq 2\gamma\tilde{\beta}\epsilon$ and $|k(x_2, y_2) - k(x_1, y_1)| \leq 2\gamma\bar{\beta}\epsilon$. Then since $E[\tilde{\beta}] = \bar{\beta}$ and Let $Var[\tilde{\beta}] = \sigma_\tau^2$. By Chebyshev inequality, we have

$$P\left[|f(x_1, y_1) - f(x_2, y_2)| \geq 2\gamma\sigma_\tau\epsilon * t\right] \leq \frac{1}{t^2}$$

□

6.2 Experimental settings and parameters for RWS

As shown in Table 1, we choose 16 datasets that come from various applications, including ECG, sensor, image, spectro, simulated and device, and have various numbers of classes, varying numbers of time series, and a wide range of lengths of time series, as shown in Table 1. For all experiments, we generate random document from uniform distribution with mean centered in Word2Vec embedding space since we observe the best performance with this setting. We perform 10-fold cross-validation to search for best parameters for σ, and $DMax$ as well as parameter C for LIBLINEAR on training set for each dataset. We simply fix the $DMin = 1$, and vary $DMax$ in the range of [10 20 30 40 50 60 70 80 90 100], σ in the range of [1e-4 1e-3 3e-3 1e-2 3e-2 0.10 0.14 0.19 0.28 0.39 0.56 0.79 1.12 1.58 2.23 3.16 4.46 6.30 8.91 10 31.62 1e2 3e2 1e3 1e4], and C in the range of [1e-5 1e-4 1e-3 1e-2 1e-1 1 1e1 1e2 1e3 1e4 1e5] respectively in all experiments. All computations were carried out on a DELL dual socket system with Intel Xeon processors 272 at 2.93GHz for a total of 16 cores and 250 GB of memory, running the SUSE Linux operating system.

Table 5: Properties of the datasets: Beef, ChlorineConcentration (CHCO), DistalPhalanxTW (DPTW), ECG5000 (ECG5T), FordB, HandOutlines (HO), InsectWingbeatSound (IWBS), ItalyPowerDemand (IPD), LargeKitchenAppliances (LKA), MALLAT, MiddlePhalanxOutlineCorrect (MPOC), NonInvasiveFatalECG_Thorax2 (NIFECG), PhalangesOutlinesCorrect (POC), ProximalPhalanxOutlineAgeGroup (PPOAG), Two_Patterns (TWOP), and Wafer. We define C:Classes, N:Train, M:Test, and L:length.

Name	C	N	M	L	App
Beef	5	30	30	470	Spectro
DPTW	6	400	139	80	Image
IPD	2	67	1,029	24	Sensor
PPOAG	3	400	205	80	Image
MPOC	2	600	291	80	Image
POC	2	1,800	858	80	Image
LKA	3	375	375	720	Device
IWBS	11	220	1,980	256	Sensor
TWOP	4	1,000	4,000	128	Simulated
ECG5T	5	500	4,500	140	ECG
CHCO	3	467	3,840	166	Simulated
Wafer	2	1,000	6,174	152	Sensor
MALLAT	8	55	2,345	1,024	Simulated
FordB	2	3636	810	500	Sensor
NIFECG	42	1,800	1,965	750	ECG
HO	2	370	1,000	2,709	Image

6.3 More Results on Effects of σ, R and D on Random Features

To fully investigate the behavior of the WME method, we study the effect of the kernel parameter σ, the R number of random documents and the D length of random documents on training and testing accuracy for all 16 datasets. Clearly, the training and testing accuracy can converge rapidly to the exact kernels when

varying R from 4 to 512, which confirms our analysis in Theory 1. When varying D from 10 to 100, we can see that in the majority of cases $DMax = [10\ 40]$ generally yields a near-peak performance except FordB.

6.4 Parameters and Settings on Comparisons of Feature Representations

For TSEigen Hayashi et al. [2005], we implemented this method in Matlab where we apply SVD to compute R number of largest dominant components on the similar matrix computed using DTW. For TSMC Lei et al. [2017], we used their open source in code in Github: `https://github.com/cecilialeiqi/SPIRAL`. Since the default rank size of TSMC is 32, we keep all methods consistent with this setting to make a fair comparison. For all methods, we choose the parameter C by 10-fold cross validation on training data in LIBLINEAR on all 16 datasets.

6.5 Parameters and Settings on Comparisons for Large-Scale Classification

For 1NN-DTW and 1NN-DTW$^{\text{opt}}$, we implemented them using Matlab internal fitcknn with DTW using the same C Mex file [2] as our method RWS. Although our implementations may not be highly optimized, we believe the runtime comparisons among these methods are reasonably fair. For DTWF Kate [2016], we used their open source code [3]. To make a fair comparison with other methods, we set the window size as $min(L/10, 40)$. The feature representation generated by DTWF combines SAX, DTW, and DTW_R where we use recommended parameter ranges $n = [8\ 16\ 24\ 32\ 40\ 48\ 56\ 64\ 72\ 80\ 96\ 112\ 128\ 144\ 160]$, $w = [4\ 8]$, and $a = [3\ 4\ 5\ 6\ 7\ 8\ 9]$ for cross validation. For TGAK Cuturi [2011], we took their open source code [4] for the experiments. We choose recommended window size $T = 0.25$ due to a good trade off between testing accuracy and computational time. We also perform cross validation to search for good kernel parameter σ in the range of [0.01, 0.033, 0.066, 0.1, 0.33, 0.66, 1, 3.3, 6.6, 10] and the LIBLINEAR parameter C in the range of [1e-5 1e-4 1e-3 1e-2 1e-1 1 1e1 1e2 1e3 1e4 1e5 1e6].

6.6 Parameters and Settings on Comparisons for Large-Scale Clustering

For KMeans-DTW Petitjean et al. [2011], we used the public available python code [5], which also implements LB_Keogh lower bound with DTW. However, the efficiency of python code may be significantly worse than C mex file of DTW we used, which could be the reason we observed larger margin speedup compared to 1NN-DTW. Nevertheless, note that the computational complexity of RWS over Kmeans-DTW reduces from quadratic complexity to linear complexity. For CLDS Li and Prakash [2011], we used the open source code published by authors [6]. We choose the parameter C by cross validation while using recommended parameters for generating the representations on all datasets. For K-Shape Paparrizos and Gravano [2015], we used the public available python code [7]. Similarly, we choose the parameter C by cross validation while using recommended parameters for generating the representations on all datasets.

[2] `https://www.mathworks.com/matlabcentral/fileexchange/43156-dynamic-time-warping\T1\textendashdtw-`

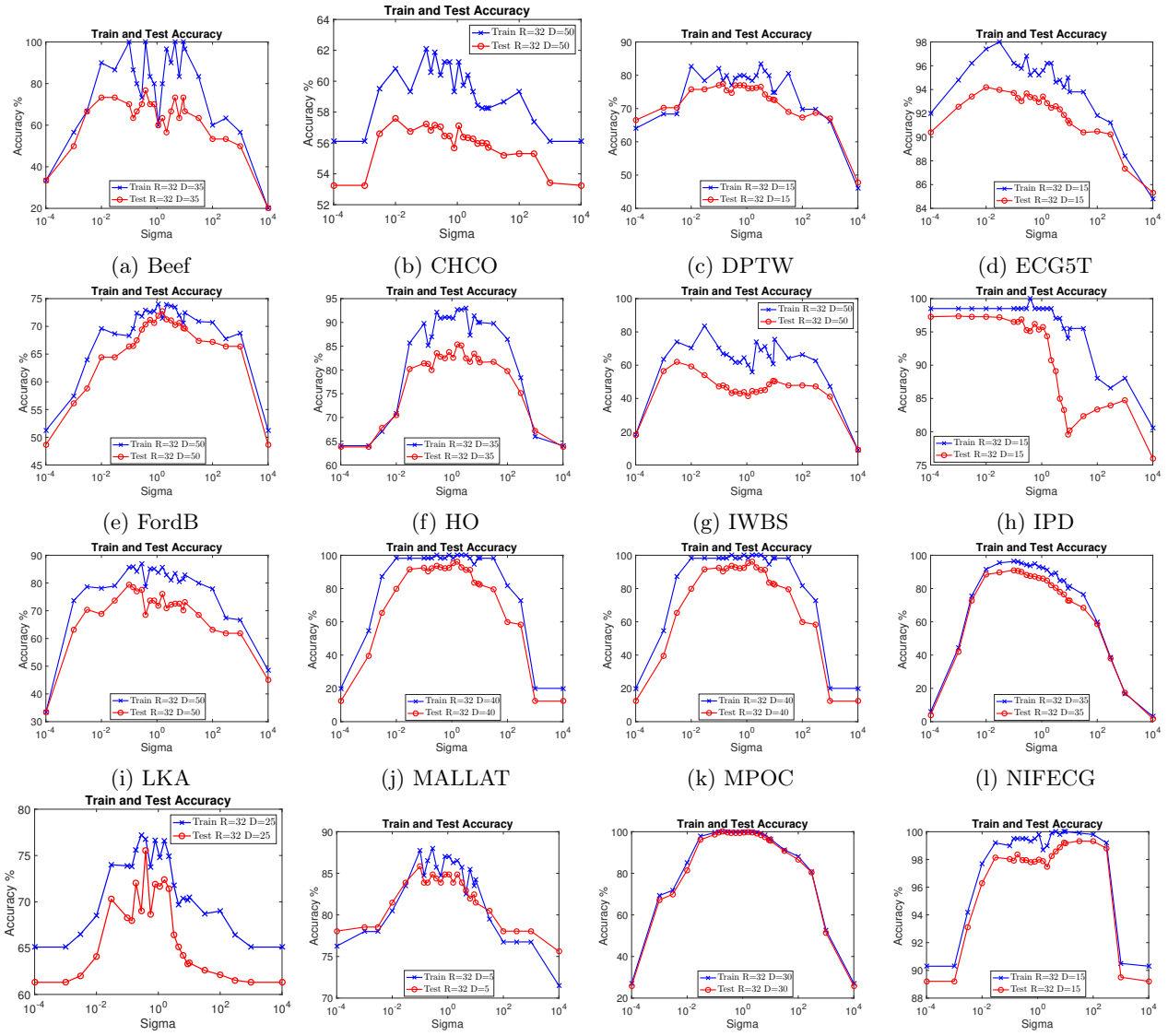
[3] `https://people.uwm.edu/katerj/timeseries/`

[4] `http://marcocuturi.net/GA.html`

[5] `https://github.com/alexminnaar/time-series-classification-and-clustering`

[6] `http://www.cs.cmu.edu/~./leili/software.html`

[7] `https://github.com/Mic92/kshape`



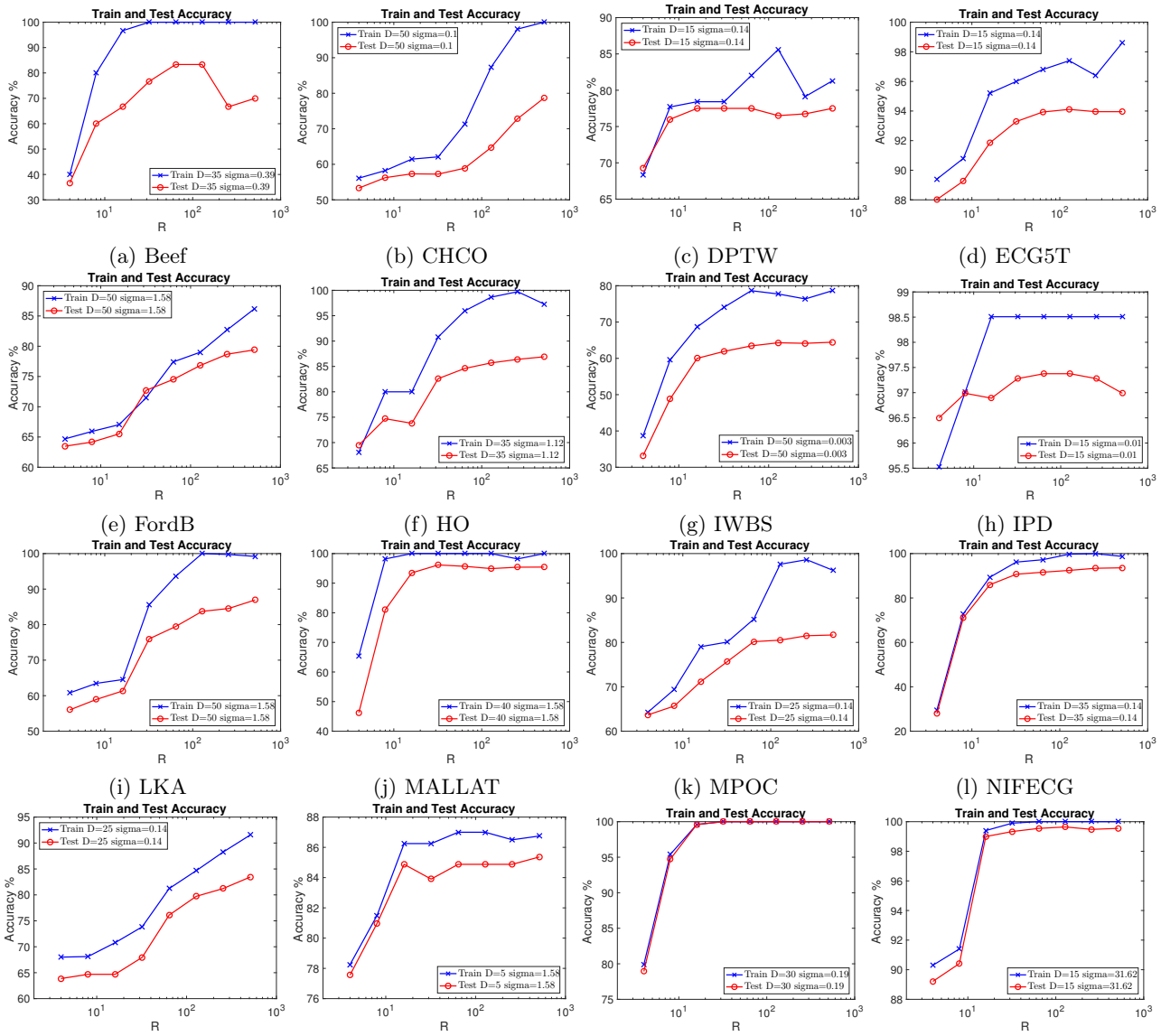
(a) Beef (b) CHCO (c) DPTW (d) ECG5T

(e) FordB (f) HO (g) IWBS (h) IPD

(i) LKA (j) MALLAT (k) MPOC (l) NIFECG

(m) POC (n) PPOAG (o) TWOP (p) Wafer

Figure 5: Train (Blue) and test (Red) accuracy when varying σ with fixed D and R. We denote $D = DMax/2$.



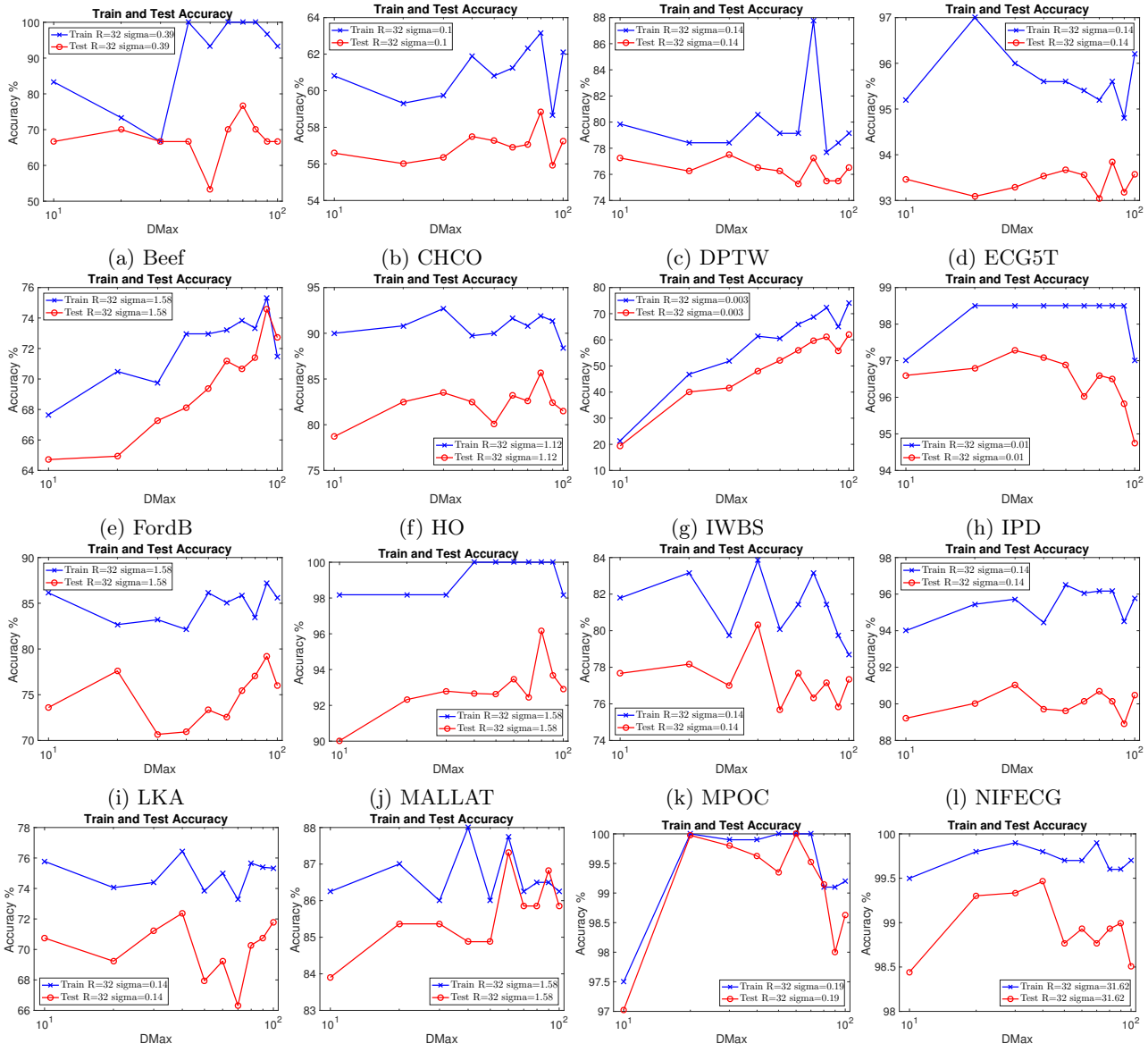
(a) Beef (b) CHCO (c) DPTW (d) ECG5T

(e) FordB (f) HO (g) IWBS (h) IPD

(i) LKA (j) MALLAT (k) MPOC (l) NIFECG

(m) POC (n) PPOAG (o) TWOP (p) Wafer

Figure 6: Train (Blue) and test (Red) accuracy when varying R with fixed σ and D. We denote $D = DMax/2$.



(a) Beef (b) CHCO (c) DPTW (d) ECG5T

(e) FordB (f) HO (g) IWBS (h) IPD

(i) LKA (j) MALLAT (k) MPOC (l) NIFECG

(m) POC (n) PPOAG (o) TWOP (p) Wafer

Figure 7: Train (Blue) and test (Red) accuracy when varying D with fixed σ and R. We denote $D = DMax/2$.