A. Bounds that hold with High Probability

To show high probability bounds we employ concentration results on homogeneous strongly Rayleigh measures. Specifically, we use the following theorem.

Theorem 6 (Pemantle & Peres (2014))**.** *Let $\mathbb{P}$ be a k-homogeneous strongly Rayleigh probability measure on $\{0,1\}^N$ and f an ℓ-Lipschitz function on $\{0,1\}^N$, then*

$$\mathbb{P}(f - \mathbb{E}[f] \geq a\ell) \leq \exp\{-a^2/8k\}.$$

It has been known that a k-DPP is a homogeneous strongly Rayleigh measure on $\{0,1\}^N$ (Borcea et al., 2009; Anari et al., 2016), thus Theorem 6 applies to results obtained with k-DPP. Concretely, for the bound in Theorem 1 that holds in expectation, we have the following bound that holds with high probability:

Corollary 7. *When sampling $C \sim k$-DPP(K), for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\frac{\|K - K_{\cdot C}(K_{C,C})^\dagger K_{C\cdot}\|_F}{\|K - K_k\|_F} \leq \left(\frac{c+1}{c+1-k}\right)\sqrt{N-k} + \sqrt{8c\log(1/\delta)}\sqrt{\frac{\sum_{i=1}^N \lambda_i^2}{\sum_{i=k+1}^N \lambda_i^2}},$$
$$\frac{\|K - K_{\cdot C}(K_{C,C})^\dagger K_{C\cdot}\|_2}{\|K - K_k\|_2} \leq \left(\frac{c+1}{c+1-k}\right)(N-k) + \sqrt{8c\log(1/\delta)}\frac{\lambda_1}{\lambda_{k+1}},$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$ are the eigenvalues of K.

Proof. We let $f_F(C) = \frac{\|K - K_{\cdot C}(K_{C,C})^\dagger K_{C\cdot}\|_F}{\|K-K_k\|_F}$ and $f_2(C) = \frac{\|K - K_{\cdot C}(K_{C,C})^\dagger K_{C\cdot}\|_2}{\|K-K_k\|_2}$. Then we have

$$0 \leq f_F(C) \leq \sqrt{\frac{\sum_{i=1}^N \lambda_i^2}{\sum_{i=k+1}^N \lambda_i^2}},$$
$$0 \leq f_2(C) \leq \frac{\lambda_1}{\lambda_{k+1}},$$

which indicates that the Lipschitz constants of $f_F(C)$ and $f_2(C)$ is upper-bounded by $\sqrt{\frac{\sum_{i=1}^N \lambda_i^2}{\sum_{i=k+1}^N \lambda_i^2}}$ and $\frac{\lambda_1}{\lambda_{k+1}}$ respectively. Applying Theorem 6 will give the results in Corollary 7. □

For the bound in Theorem 3 that holds in expectation, we have the following bound that holds with high probability:

Corollary 8. *If $\tilde{K}$ is constructed via* DPP*-Nyström, then with probability at least $1-\delta$, $\sqrt{\frac{\mathrm{bias}(\tilde{K})}{\mathrm{bias}(K)}}$ is upper-bounded by*

$$1 + \frac{1}{N\gamma}\left(\frac{(c+1)e_{c+1}(K)}{e_c(K)} + \sqrt{8c\log(1/\delta)}tr(K)\right).$$

Proof. Consider the function $f_C(K) = \nu_C = \sum_i \|b_i^\top (U^C)^\perp\|_2^2 \leq \sum_i \|b_i^\top\|_2^2 = \mathrm{tr}(K)$. Since $0 \leq f_C(K) \leq \mathrm{tr}(K)$, it follows that the Lipschitz constant for f_C is at most $\mathrm{tr}(K)$. Thus when $C \sim k$-DPP and $\delta \in (0,1)$, by applying Theorem 6 we see that the inequality $\nu_C \leq \mathbb{E}[\nu_C] + \sqrt{8c\log(1/\delta)}\mathrm{tr}(K)$ holds with probability at least $1-\delta$. Hence

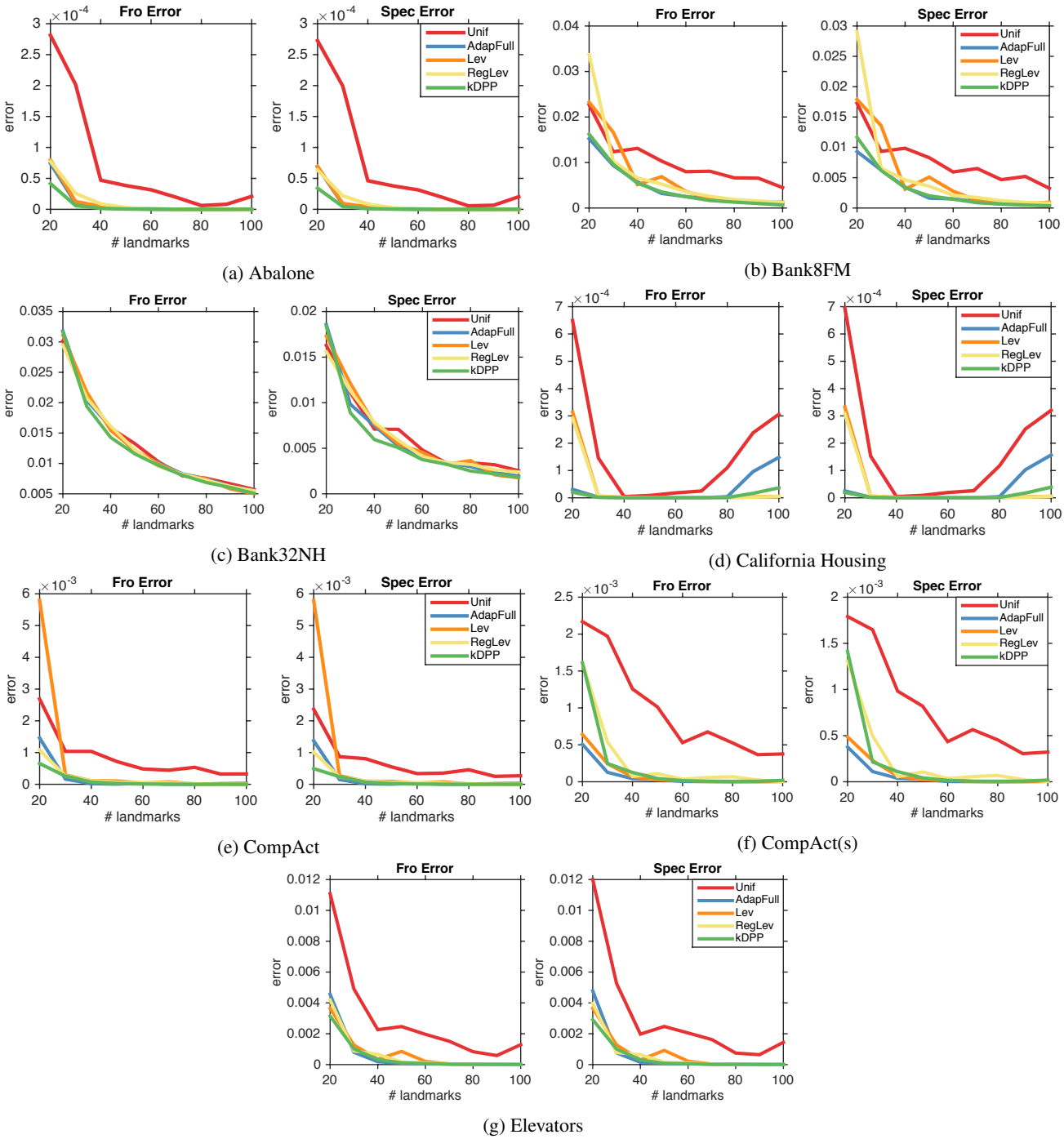
$$\mathbb{E}_C\left[\sqrt{\frac{\mathrm{bias}(\tilde{K})}{\mathrm{bias}(K)}}\right] \leq 1 + \mathbb{E}\left[\frac{\nu_C}{N\gamma}\right] + \sqrt{8c\log(1/\delta)}\frac{\mathrm{tr}(K)}{N\gamma}$$
$$= 1 + \frac{1}{N\gamma}\left(\frac{(c+1)e_{c+1}(K)}{e_c(K)} + \sqrt{8c\log(1/\delta)}\mathrm{tr}(K)\right)$$

holds with probability at least $1-\delta$. □

B. Supplementary Experiments

B.1. Kernel Approximation

Fig. 7 shows the matrix norm relative error of various methods in kernel approximation on the remaining 7 datasets mentioned in the main text.



(a) Abalone

(b) Bank8FM

(c) Bank32NH

(d) California Housing

(e) CompAct

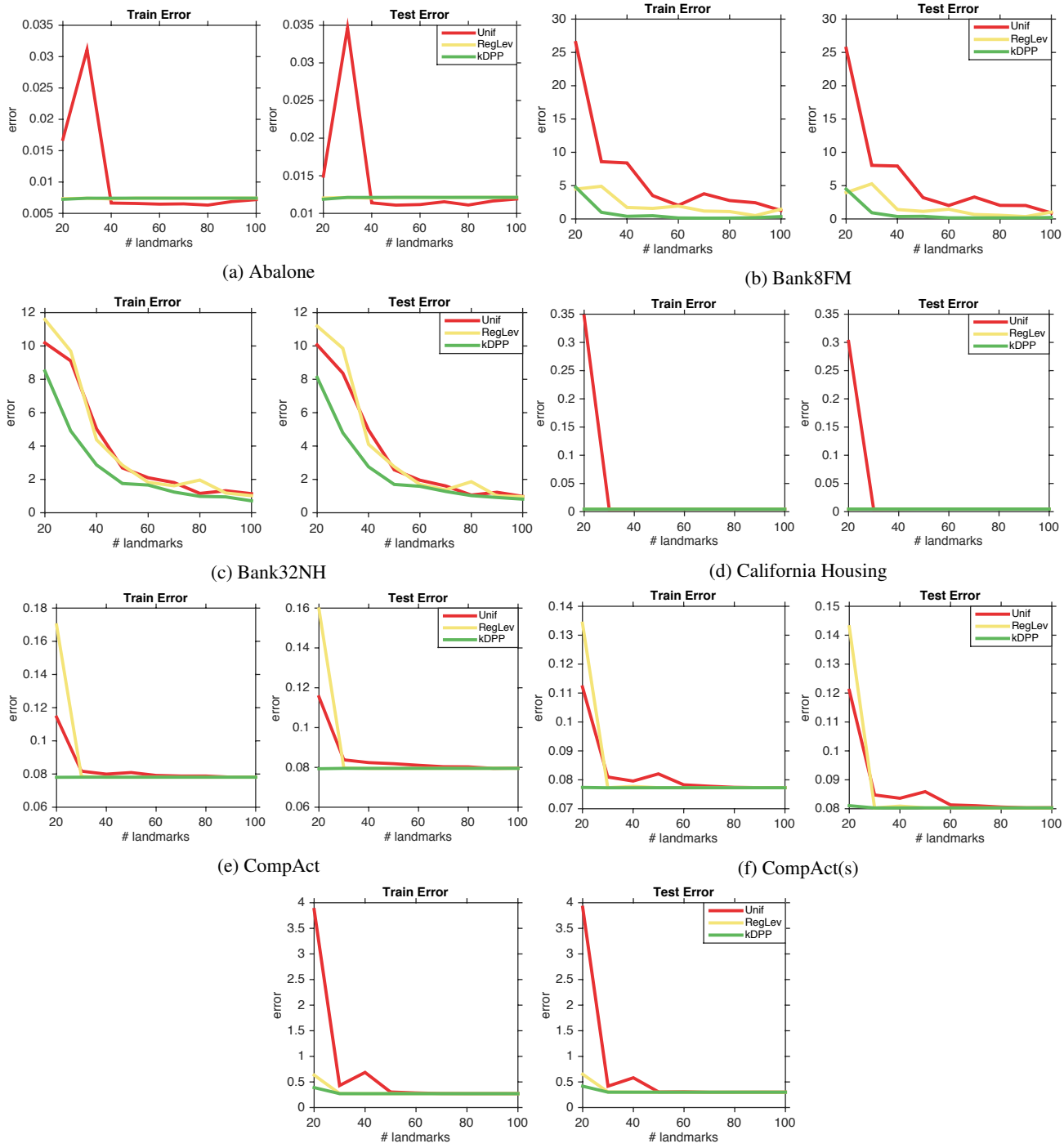
(f) CompAct(s)

(g) Elevators

Figure 7: Relative Frobenius norm and spectral norm error achieved by different kernel approximation algorithms on the remaining 7 data sets.

B.2. Approximated Kernel Ridge Regression

Fig. 8 shows the training and test error of various methods for kernel ridge regression on the remaining 7 datasets.



(a) Abalone

(b) Bank8FM

(c) Bank32NH

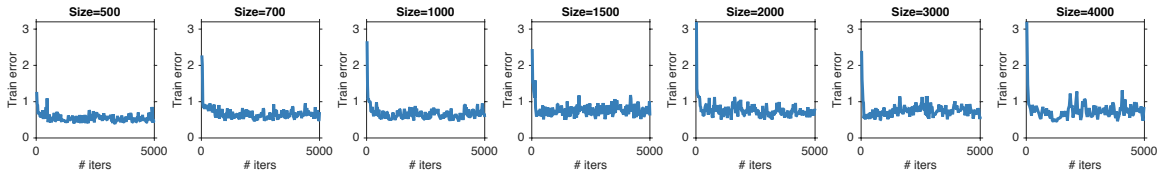
(d) California Housing

(e) CompAct

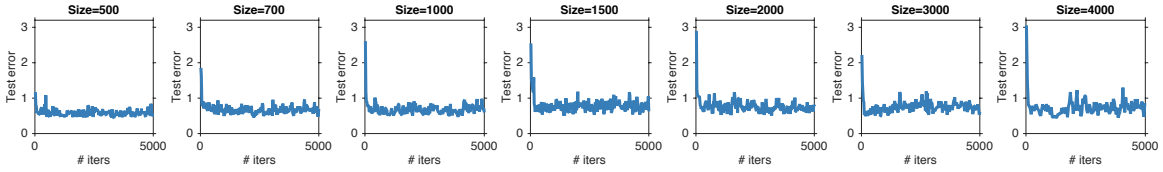
(f) CompAct(s)

(g) Elevators

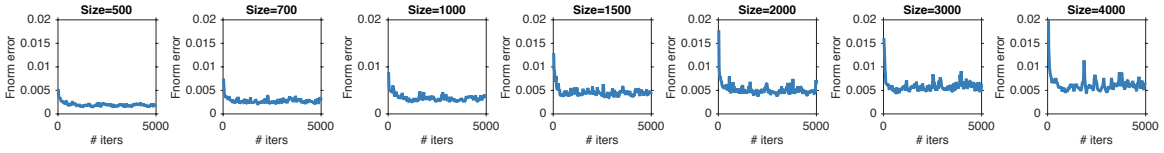
Figure 8: Training and test error achieved by different Nyström kernel ridge regression algorithms on the remaining 7 regression datasets.



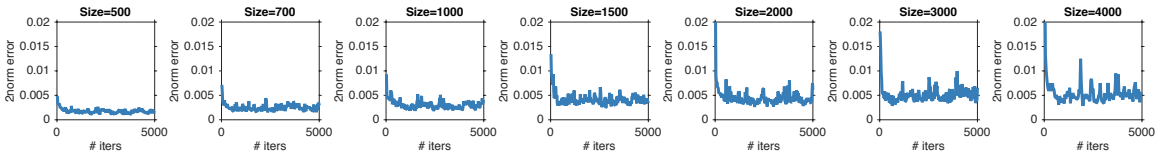
(a) Training error



(b) Test error



(c) Relative Frobenius norm error



(d) Relative Spectral norm error

Figure 9: Performance of Markov chain DPP-Nyström with 50 landmarks on Ailerons. Runs for 5,000 iterations.

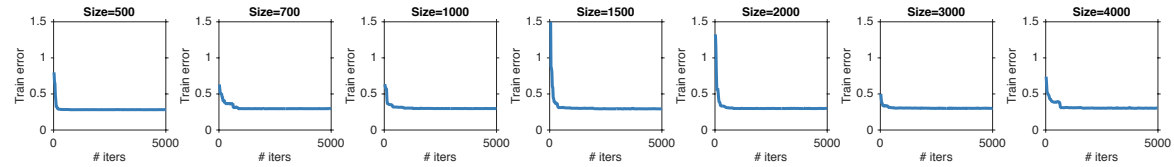
B.3. Mixing of Markov Chain k-DPP

We first show the mixing of the Gibbs DPP-Nyström with 50 landmarks with different performance measures: relative spectral norm error, training error and test error of kernel ridge regression in Fig. 9.

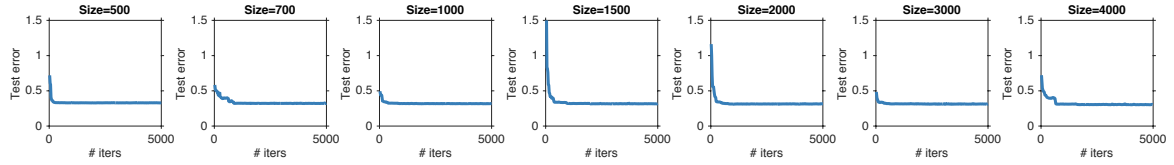
We also show corresponding results with respect to 100 and 200 landmarks in Fig. 10 and Fig. 11, so as to illustrate that for varying number of landmarks the chain is indeed fast mixing and will give reasonably good result within a small number of iterations.

B.4. Running Time Analysis

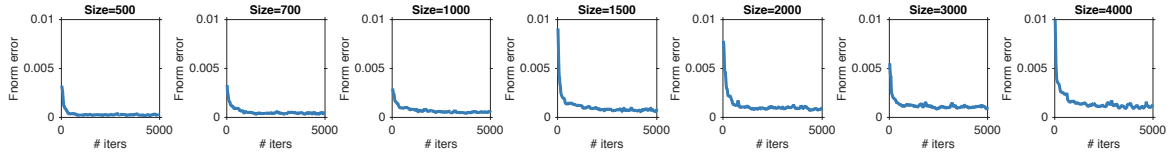
We next show time-error trade-offs for various sampling methods on small and larger datasets with respect to Fnorm and 2norm errors. We sample 20 landmarks from Ailerons dataset of size 4,000 and California Housing of size 12,000. The result is shown in Figure 12 and Figure 13 and similar trends as the example results in the main text could be spotted: on small scale dataset (size 4,000) `kDPP` get very good time-error trade-off. It is more efficient than `Kmeans`, though the error is a bit larger. While on larger dataset (size 12,000) the efficiency is further enhanced while the error is even lower than `Kmeans`. It also have lower variances in both cases compared to `AppLev` and `AppRegLev`. Overall, on larger dataset we obtain the best time-error trade-off with `kDPP`.



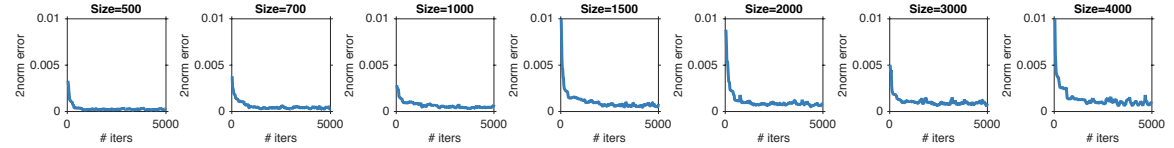
(a) Training error



(b) Test error

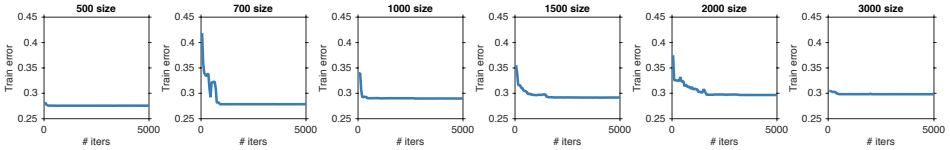


(c) Relative Frobenius norm error

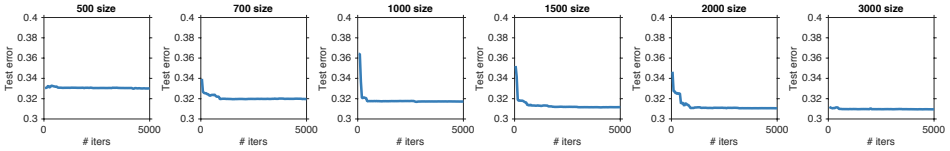


(d) Relative Spectral norm error

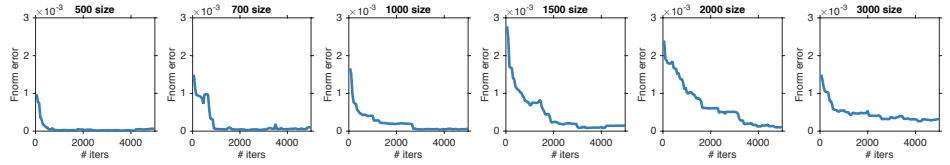
Figure 10: Performance of Markov chain DPP-Nyström with 100 landmarks on Ailerons. Runs for 5,000 iterations.



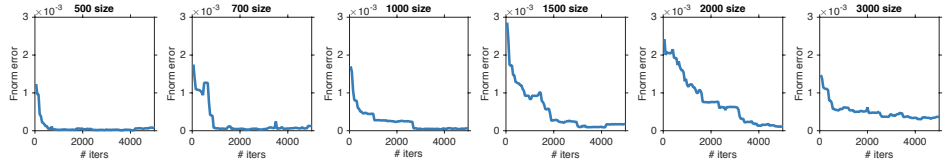
(a) Training error



(b) Test error

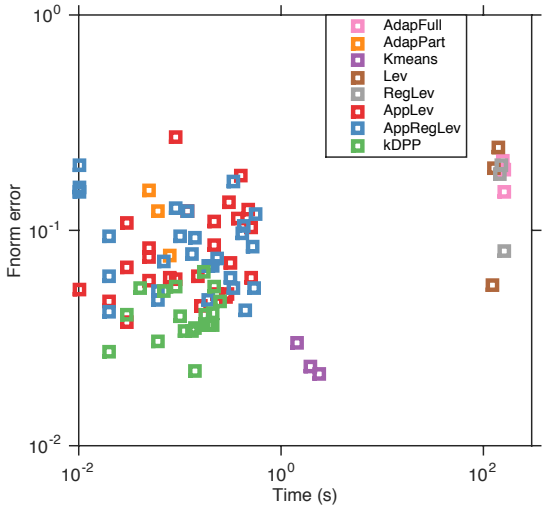


(c) Relative Frobenius norm error

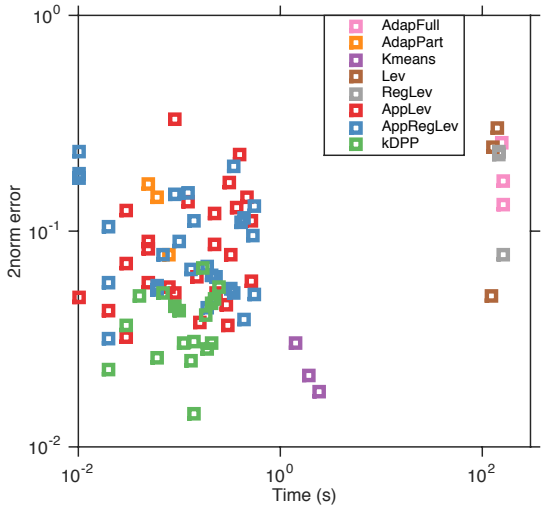


(d) Relative Spectral norm error

Figure 11: Performance of Markov chain DPP-Nyström with 200 landmarks on Ailerons. Runs for 5,000 iterations.

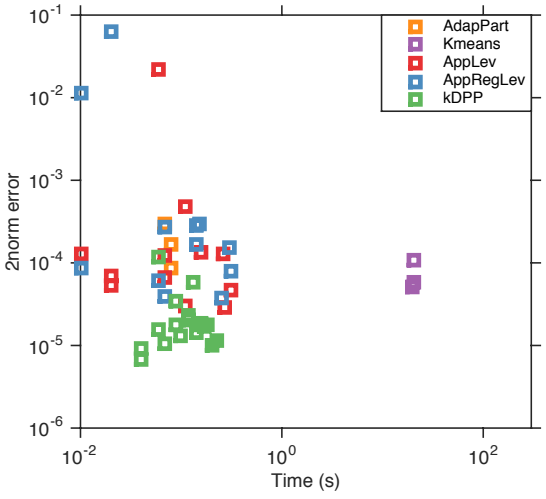


(a) Fnorm Error vs. Time

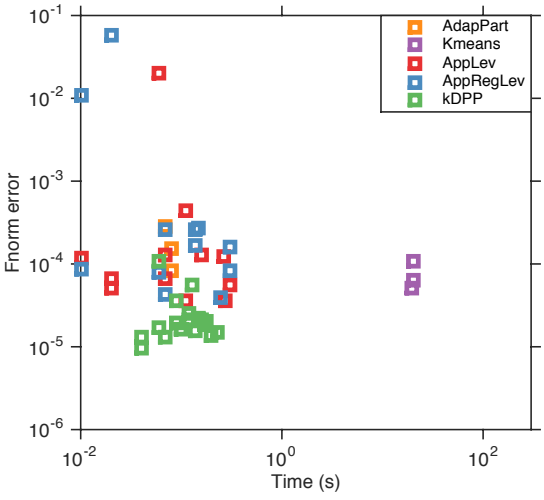


(b) 2norm Error vs. Time

Figure 12: Time-Error tradeoff with 20 landmarks on Ailerons of size 4,000. Time and Errors shown in log-scale.



(a) 2norm Error vs. Time



(b) Training Error vs. Time

Figure 13: Time-Error tradeoff with 20 landmarks on California Housing of size 12,000. Time and Errors shown in log-scale. We didn't include `AdapFull`, `Lev` and `RegLev` due to their inefficiency on larger datasets.