A Proof of Theorem 3.4

The proofs of Theorem 3.2 and Theorem 3.4 rely on the main error bound for the *Hilbert coreset construction problem* given in Eq. (9) (Campbell and Broderick, 2019). We restate this error bound in Lemma A.2, which depends on several key quantities given below:

- $c_{ls} := \frac{1}{J_+}\cos(\omega_l^T x_{i_s} + b_l)\cos(\omega_l^T x_{j_s} + b_l)$, such that $1 \le s \le S$ and $1 \le l \le J_+$
- $\hat{\sigma}_j^2 := \frac{1}{S}\sum_{s=1}^{S} c_{js}^2 = \frac{1}{S}\|R_j\|_2^2$
- $\hat{\sigma}^2 := \left(\sum_{j=1}^{J_+}\hat{\sigma}_j\right)^2$

Definition A.1. (Campbell and Broderick, 2019) The *Hilbert construction problem* is based on solving the quadratic program,

$$\underset{w\in\mathbb{R}_+^{J_+}}{\operatorname{argmin}} \quad \frac{1}{S}\|r - r(w)\|_2^2 \quad \text{s.t.} \sum_{j=1}^{J_+} w_j\hat{\sigma}_j = \hat{\sigma}. \tag{9}$$

Remark. The minimizer of Eq. (9) is $w^* = (1, \cdots, 1)$ since $r(w^*) = r$. However, the goal is to find a sparse w. Instead of adding sparsity-inducing constraints (such as L_1 penalties), which would lead to computational difficulties for large-scale problems, Campbell and Broderick (2019) minimize Eq. (9) greedily through the Frank-Wolfe algorithm. Frank-Wolfe outputs a sparse w since the sparsity of w is bounded by the number of iterations Frank-Wolfe is run for.

Lemma A.2. *(Campbell and Broderick, 2019, Theorem 4.4) Solving Eq. (9) with J iterations of Frank-Wolfe satisfies*

$$\begin{aligned}\frac{1}{S}\|r - r(w)\|_2^2 &\le \frac{\hat{\sigma}^2\eta^2\bar{\eta}^2\nu_J^2}{\bar{\eta}^2\nu^{-2(J-2)} + \eta^2(J-1)} \\ &\le \nu_J^{2J-2},\end{aligned} \tag{10}$$

where $0 \le \nu_J < 1$. Furthermore, $\nu_J^2 = 1 - \frac{d^2}{\sigma^2\bar{\eta}^2}$ where d is the distance from r to the nearest boundary of the convex hull of $\left\{\frac{\hat{\sigma}}{\hat{\sigma}_j}R_j\right\}_{j=1}^{J_+}$ and $\bar{\eta}^2 := \frac{1}{S}\max_{i,j\in[J_+]}\left\|\frac{R_i}{\hat{\sigma}_i} - \frac{R_j}{\hat{\sigma}_j}\right\|^2, 0 \le \bar{\eta} \le 2$.

We prove Theorem 3.4 first since the main idea is captured in this proof. The proof of Theorem 3.2 is more involved since we must use a number of concentration bounds to justify subsampling only S datapoint pairs instead of all $\frac{N(N-1)}{2}$ possible datapoint pairs. Both proofs will also depend on the following constants.

- $\sigma_j^2 := \frac{1}{V^*}\sum_{s=1}^{V^*} c_{js}^2 = \frac{1}{V^*}\|R_j\|_2^2$
- $\sigma^2 := \left(\sum_{j=1}^{J_+}\sigma_j\right)^2$

Here, $V^* = \frac{N(N-1)}{2}$, that is when all datapoint pairs above the diagonal are included. $\hat{\sigma}_j^2$ and $\hat{\sigma}^2$ are simply unbiased estimates of σ_j^2 and σ^2 based on sampling only S instead of all V^* datapoint pairs.

While Lemma A.2 guarantees $0 < \nu_{J_+} < 1$, it does not guarantee that $\nu_{J_+} \to 1$ as the number of random features $J_+ \to \infty$. The following Lemma is critical in showing that ν_{J_+} does not approach 1, which would result in no compression.

Lemma A.3. *Let $\{x_i\}_{i=1}^K$ be a set of points in $\mathbb{R}^p$ that satisfies Assumption 3.1(a). Consider the vector $v_{\omega,b} = (\cos(\omega^T x_i + b)\cos(\omega^T x_j + b))_{i<j, i\in[K-1]} \in \mathbb{R}^{\frac{K(K-1)}{2}}$. Let the unit vector $u_{\omega,b} := \frac{v_{\omega,b}}{\|v_{\omega,b}\|}$. If $\omega_j \overset{i.i.d.}{\sim} F$ and $b_j \overset{i.i.d.}{\sim} G$, where F has positive density on all of $\mathbb{R}^p$ and G has positive density on $[0, 2\pi]$, then*

$$\begin{aligned}&\mathrm{d}\left(\mathrm{ConvexHull}\{u_{\omega_j,b_j}\}_{j=1}^J, \mathcal{S}^{\frac{K(K-1)}{2}-1}\right) \to 0 \quad \textit{for} \quad J \to \infty \\ &\textit{s.t.} \quad \mathrm{d}(A,B) := \max_{a\in A, b\in B}||a-b||_2.\end{aligned} \tag{11}$$

Here, $\mathcal{S}^{\frac{K(K-1)}{2}-1}$ denotes the surface of the unit sphere in $\mathbb{R}^{\frac{K(K-1)}{2}}$.

Proof. By construction, each unit vector $u_i \coloneqq u_{\omega_i, b_i}$ lies on the boundary of the unit sphere in $\mathbb{R}^{\frac{K(K-1)}{2}}$. Hence, F, G induce a distribution on $\mathcal{S}^{\frac{K(K-1)}{2}-1}$. It suffices to show $\mathcal{S}^{\frac{K(K-1)}{2}-1}$ has strictly positive density everywhere since, as $J \to \infty$, any arbitrarily small neighborhood around a collection of points that cover $\mathcal{S}^{\frac{K(K-1)}{2}-1}$ will be hit by some u_i with probability 1. By standard convexity arguments, the convex hull of the u_i will arbitrarily approach $\mathcal{S}^{\frac{K(K-1)}{2}-1}$ by taking the radius of the neighborhoods to zero. We now show $\mathcal{S}^{\frac{K(K-1)}{2}-1}$ has strictly positive density everywhere. Since u_i is the normalized vector of $v_i \coloneqq v_{\omega_i, b_i}$ and each component of v_i is between -1 and 1, it suffices to show, by the continuity of the cosine function, that for any $a \in \{-1, 1\}^{\frac{K(K-1)}{2}}$ there exist some ω_i, b_i such that $\text{sign}(v_i) \coloneqq (\text{sign}(v_{il}))_{l \in \frac{K(K-1)}{2}}$ equals a. Recall that

$$\cos(a)\cos(b) = \frac{1}{2}(\cos(a+b) + \cos(a-b)). \tag{12}$$

Take $b_i = 0$. Then, Equation (12) implies $v_{il} = \frac{1}{2}(\cos(\omega_i^T(x_{i_l} + x_{j_l}) + \cos(\omega^T(x_{i_l} - x_{j_l}))$. Consider the vector $\tilde{v}_i = (\cos(\omega_i^T(x_{i_l} + x_{j_l}), \cos(\omega_i^T(x_{i_l} - x_{j_l}))_{l \in \frac{K(K-1)}{2}} \in \mathbb{R}^{K(K-1)}$. It suffices to show that for any $\tilde{a} \in \{-1, 1\}^{K(K-1)}$, there exists an ω_i such that $\text{sign}(\tilde{v}_i) = \tilde{a}$. Recall that the cosine function has infinite *VC dimension*, namely that for any labeling $y_1, \cdots, y_M \in \{-1, 1\}$ of distinct points $x_1, \cdots x_M \in \mathbb{R}^p$, there exists an ω^* such that $\text{sign}(\cos((\omega^*)^T x_m)) = y_m$. Take $M = K(K-1)$, $y_m = \tilde{a}_m$, $x_m = x_{i_m} + x_{j_m}$, and $x_{m+1} = x_{i_m} - x_{j_m}$. Since all the x_m are distinct by Assumption 3.1(a), we can find an ω_i such that $\text{sign}(\tilde{v}_i) = \tilde{a}$ as desired. □

We now prove Theorem 3.4.

Proof. Each $R_j \in \mathbb{R}^{\frac{N(N-1)}{2}}$ and the R_j's are i.i.d. since each ω_j is drawn i.i.d. from Q. The induced Hilbert norm $\|\cdot\|_H$ of each R_j is given by $\|R_j\|_H^2 = \frac{2}{N(N-1)}\|R_j\|_2^2$ (Campbell and Broderick, 2019). Hence, $\tilde{R}_j \coloneqq \frac{R_j}{\sigma_j}$ is a unit vector in the vector space with norm $\|\cdot\|_H$. By Lemma A.3,

$$\text{d}\left(\text{ConvexHull}\{\tilde{R}_j\}_{j=1}^{J_+}, \mathcal{S}^{\frac{N(N-1)}{2}-1}\right) \to 0 \tag{13}$$

Let $\tilde{r} \coloneqq \frac{1}{\sigma}\sum_{j=1}^{J_+}\sigma_j\tilde{R}_j \in \text{ConvexHull}\{\tilde{R}_j\}_{j=1}^{J_+}$ and observe that $\tilde{r} = \frac{r}{\sigma}$. The distance, which we denote as d_{J_+}, between $\tilde{r}$ and the $\text{ConvexHull}\{\tilde{R}_j\}_{j=1}^{J_+}$ approaches $1 - \|\tilde{r}\|_H$ since the $\text{ConvexHull}\{\tilde{R}_j\}_{j=1}^{J_+}$ approaches $\mathcal{S}^{\frac{N(N-1)}{2}-1}$. Hence,

$$\lim_{J_+ \to \infty} d_{J_+} = 1 - \lim_{J_+ \to \infty}\|\tilde{r}\|_H = 1 - \frac{\lim_{J_+ \to \infty}\|r\|_H}{\lim_{J_+ \to \infty}\sigma}. \tag{14}$$

Now,

$$r_s = \frac{1}{J_+}\sum_{j=1}^{J_+} c_{js} \overset{J_+ \to \infty}{\longrightarrow} k(x_{i_s}, x_{j_s}). \tag{15}$$

Hence, as $J_+ \to \infty$,

$$\|r\|_H \to \sqrt{\frac{2}{N(N-1)}\sum_{i<j}(k(x_i, x_j))^2}. \tag{16}$$

Now,

$$\begin{aligned}
\sigma &= \sum_{j=1}^{J_+} \sigma_j \\
&= \sum_{j=1}^{J_+} \sqrt{\frac{1}{V^*} \sum_{s=1}^{V^*} c_{js}^2} \\
&= \sum_{j=1}^{J_+} \sqrt{\frac{1}{V^*} \sum_{s=1}^{V^*} \frac{1}{J_+^2} \cos^2(\omega_j^T x_{i_s} + b_j) \cos^2(\omega_j^T x_{j_s} + b_j)} \\
&= \frac{1}{J_+} \sum_{j=1}^{J_+} \sqrt{\frac{1}{V^*} \sum_{s=1}^{V^*} \cos^2(\omega_j^T x_{i_s} + b_j) \cos^2(\omega_j^T x_{j_s} + b_j)} \\
&= \sqrt{\frac{2}{N(N-1)}} \frac{1}{J_+} \sum_{j=1}^{J_+} \|(\cos(\omega_j^T x_m + b_j) \cos(\omega_j^T x_n + b_j))_{m<n}\|_2 \\
&\to \sqrt{\frac{2}{N(N-1)}} \mathbb{E}_{\omega,b} \|(\cos(w^T x_m + b) \cos(w^T x_n + b))_{m<n}\|_2
\end{aligned} \tag{17}$$

If $x \neq y$ and $w \neq 0$, then

$$\begin{aligned}
k(x,y) &= \mathbb{E}_{\omega,b} \cos(w^T x + b) \cos(w^T y + b) \\
&< \mathbb{E}_{\omega_i,b} |\cos(w^T x + b) \cos(w^T y + b)|.
\end{aligned} \tag{18}$$

by Jensen's inequality. Hence, Eq. (18) and Assumption 3.1(a-b) together imply

$$\frac{\lim_{J_+ \to \infty} \|r\|_2}{\lim_{J_+ \to \infty} \sigma} < 1. \tag{19}$$

By Eq. (16) and Eq. (17),

$$\frac{\lim_{J_+ \to \infty} \|r\|_H}{\lim_{J_+ \to \infty} \sigma} \leq \frac{\|K\|_F}{\mathbb{E}_{\omega,b} \|u(\omega,b)\|_2}, \tag{20}$$

where $u(\omega, b)$ is defined in Theorem 3.4. Lemma A.2 says that $\nu_{J_+}^2 = 1 - \frac{d^2}{\sigma^2 \bar{\eta}^2}$, where d is the distance from r to the nearest boundary of the convex hull of $\left\{\frac{\sigma}{\sigma_j} R_j\right\}_{j=1}^{J_+}$. Hence, $d = \sigma d_{J_+}$ and $\nu_{J_+}^2 = 1 - \frac{d_{J_+}^2}{\bar{\eta}^2}$. Eq. (14) and Eq. (20) together imply,

$$\liminf_{J_+ \to \infty} d_{J_+} \leq 1 - \frac{\|K\|_F}{\mathbb{E}_{\omega,b} \|u(\omega,b)\|_2}. \tag{21}$$

Therefore, since $0 \leq \bar{\eta}^2 \leq 2$ by Lemma A.2,

$$\begin{aligned}
\limsup_{J_+ \to \infty} \nu_{J_+}^2 &\leq \limsup_{J_+ \to \infty} 1 - \frac{d_{J_+}^2}{2} \\
&= 1 - \liminf_{J_+ \to \infty} \frac{d_{J_+}^2}{2} \\
&\leq 1 - \frac{\left(1 - \frac{\|K\|_F}{\mathbb{E}_{\omega,b} \|u(\omega,b)\|_2}\right)^2}{2}.
\end{aligned} \tag{22}$$

□

B Proof of Theorem 3.2

The following technical lemma is needed to derive the probability bound in Theorem 3.2.

Lemma B.1. *Suppose* $\frac{\sigma^2}{J_+^2\sigma_i^2} \le M$ *for some* $1 \le M < \infty$ *for all* $i \in [J_+]$. *For* $S \ge 8\frac{M^2}{\sigma^4}\log\left(\frac{2J_+}{\delta^2}\right)$

$$\mathbb{P}\left(\frac{\hat{\sigma}^2}{J_+^2 {\hat{\sigma}_i}^2} \ge 5M\right) \le \delta \tag{23}$$

for all $i \in [J_+]$.

Proof. Notice that

$$\begin{aligned}\mathbb{E}_{i_s,j_s}{\hat{\sigma}_l}^2 &= \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{i_s,j_s}c_{ls}^2 \\ &= \frac{1}{N^2}\sum_{s=1}^{N^2}c_{ls}^2 \\ &= \sigma_l^2.\end{aligned}$$

Hence, ${\hat{\sigma}_l}^2$ is an unbiased estimator of σ_l^2. Each $c_{ls}^2 \le \frac{1}{{J_+}^2}$ is a bounded random variable, and the collection of random variables $\{c_{ls}^2\}_{s=1}^S$ are i.i.d. since $i_s, j_s \overset{\text{i.i.d.}}{\sim} \pi$. Hence, by Hoeffding's inequality,

$$\mathbb{P}\left(|\hat{\sigma}_l^2 - \sigma_l^2| \ge t\right) \le 2\exp\left(-2SJ_+^4t^2\right). \tag{24}$$

Define the event $A_t := \cup_{i=1}^{J_+}\{|\hat{\sigma}_i^2 - \sigma_i^2| < t\}$ and pick t such that $t \le \min_{i\in[J_+]}\sigma_i^2$. Since $\sigma_i^2 \ge \frac{\sigma^2}{M}$ by assumption, it suffices to pick $0 < t \le \frac{\sigma^2}{M}$. Conditioned on A_t, $\hat{\sigma}_i \le \sqrt{\sigma_i^2 + t} \le \sigma_i + \sqrt{t}$, which implies $\hat{\sigma}^2 \le (\sigma + J_+\sqrt{t})^2$. Therefore,

$$\begin{aligned}\mathbb{P}\left(\frac{\hat{\sigma}^2}{J_+^2{\hat{\sigma}_i}^2} \ge cM\right) &= \mathbb{P}\left(A_t^c \cup \left\{\frac{\hat{\sigma}^2}{J_+^2{\hat{\sigma}_i}^2} \ge cM\right\}\right) + \mathbb{P}\left(A_t \cup \left\{\frac{\hat{\sigma}^2}{J_+^2{\hat{\sigma}_i}^2} \ge cM\right\}\right) \\ &\le \mathbb{P}\left(A_t^c\right) + \mathbb{P}\left(A_t, \left\{\frac{\hat{\sigma}^2}{J_+^2{\hat{\sigma}_i}^2} \ge cM\right\}\right) \\ &\le \mathbb{P}\left(A_t^c\right) + \mathbb{P}\left(\frac{\hat{\sigma}^2}{J_+^2{\hat{\sigma}_i}^2} \ge cM \mid A_t\right) \\ &\le \mathbb{P}\left(A_t^c\right) + \mathbb{P}\left(\frac{(\sigma+\sqrt{t}J_+)^2}{J_+^2(\sigma_i^2 - t)} \ge cM \mid A_t\right).\end{aligned} \tag{25}$$

Notice that $\mathbb{P}\left(\frac{(\sigma+\sqrt{t}J_+)^2}{\sigma_i^2 - t} \ge cM^2 \mid A_t\right)$ is either 0 or 1 since σ_i and σ are constants. We pick t so that this probability is 0. To pick t, notice that,

$$\begin{aligned}\frac{(\sigma+\sqrt{t}J_+)^2}{J_+^2(\sigma_i^2 - t)} &= \frac{\left(\frac{\sigma}{\sigma_i} + \frac{\sqrt{t}J_+}{\sigma_i}\right)^2}{J_+^2(1 - \frac{t}{\sigma_i^2})} \\ &\le \frac{\left(J_+\sqrt{M} + \frac{J_+\sqrt{tM}J_+}{\sigma}\right)^2}{J_+^2(1 - \frac{t}{\sigma_i^2})} \\ &\le \frac{M\left(1 + \frac{\sqrt{t}J_+}{\sigma}\right)^2}{1 - \frac{MJ_+^2t}{\sigma^2}},\end{aligned} \tag{26}$$

where the last inequality holds as long as $0 < t < \frac{\sigma^2}{MJ_+^2}$ and follows by noting that $\frac{1}{\sigma_i^2} \leq \frac{MJ_+^2}{\sigma^2}$ by assumption. Pick $t = \frac{\sigma^2}{4J_+^2 M}$. Since $0 \leq \sigma \leq 1$, this choice of t implies $\frac{M\left(1+\frac{\sqrt{t}J_+}{\sigma}\right)^2}{1-\frac{MJ_+^2 t}{\sigma^2}} \leq 5M$. Hence, for $c = 5$ and this choice of t, $\mathbb{P}\left(\frac{(\sigma+\sqrt{t}J_+)^2}{J_+^2(\sigma_i^2-t)} \geq 5M \mid A_t\right) = 0$. Combining Eq. (25) and Eq. (24), we have by a union bound that,

$$\mathbb{P}\left(\frac{\hat{\sigma}^2}{J_+^2 \hat{\sigma_i}^2} \geq 5M\right) \leq 2J_+ \exp\left(-\frac{1}{8} S \frac{\sigma^4}{M^2}\right), \tag{27}$$

for all $i \in [J_+]$. Solving for S by setting the right hand side above to δ yields the claim. □

We have all the pieces to prove Theorem 3.2. We follow the proof strategy in (Campbell and Broderick, 2019, Theorem 5.2).

Proof. Let $R^* = \left[z_{+1}^T \circ z_{+1}^T, \cdots z_{+N-1}^T \circ z_{+N}^T, z_{+N}^T \circ z_{+N}^T\right] \in \mathbb{R}^{J_+ \times N^2}$. Notice,

$$\frac{1}{N^2}\|Z_+ {Z_+}^T - Z(w)Z(w)^T\|_F^2 = (1-w)^T \frac{R^*}{N}\frac{R^{*T}}{N}(1-w). \tag{28}$$

We approximate Eq. (28) with $(1-w)^T \frac{R}{\sqrt{S}}\frac{R^T}{\sqrt{S}}(1-w)$ and bound the error. Suppose

$$D^* := \max_{i,j \in [J_+]} \left| \left(\frac{R^*}{N}\frac{R^{*T}}{N}\right)_{ij} - \left(\frac{R}{\sqrt{S}}\frac{R^T}{\sqrt{S}}\right)_{ij} \right| \leq \frac{\epsilon}{2}.$$

Then,

$$\begin{aligned}(1-w)^T \frac{R^*}{N}\frac{R^{*T}}{N}(1-w) - (1-w)^T \frac{R}{\sqrt{S}}\frac{R^T}{\sqrt{S}}(1-w) &\leq \sum_{i,j\in[J_+]} |w_i - 1||w_j - 1| D^* \\ &\leq \|w-1\|_1^2 \frac{\epsilon}{2}.\end{aligned} \tag{29}$$

Notice,

$$\begin{aligned}\mathbb{E}_{i_s,j_s}\left[\left(\frac{R}{\sqrt{S}}\frac{R^T}{\sqrt{S}}\right)_{ij}\right] &= \mathbb{E}_{i_s,j_s}\left[\frac{1}{S}\sum_{s=1}^{S} c_{is}c_{js}\right] \\ &= \frac{1}{S}\sum_{s=1}^{S} E_{i_s,j_s}\left[c_{is}c_{js}\right] \\ &= E_{i_s,j_s}\left[c_{is}c_{js}\right] \\ &= \frac{1}{N^2}\sum_{s=1}^{N^2} c_{is}c_{js} \\ &= \left(\frac{R^*}{N}\frac{R^{*T}}{N}\right)_{ij}.\end{aligned} \tag{30}$$

Hence, the i.i.d. collection of random variables $\{c_{is}c_{js}\}_{s=1}^S$ yields an unbiased estimate of $\left(\frac{R^*}{N}\frac{R^{*T}}{N}\right)_{ij}$. Each $c_{is}c_{js}$ is bounded by $\frac{1}{J_+^2}$. Therefore, by Hoeffding's inequality and a simple union bound,

$$\mathbb{P}\left(D^* \geq \frac{\epsilon}{2}\right) \leq 2J_+^2 \exp\left(-2SJ_+^4\epsilon^2\right). \tag{31}$$

Setting the right-hand side to $\frac{\delta^*}{2}$ and solving for $\frac{\epsilon}{2}$ implies with probability at least $1 - \frac{\delta^*}{2}$,

$$\frac{\epsilon}{2} \leq \frac{1}{\sqrt{S}{J_+}^2}\log\left[\frac{4J_+^2}{\delta^*}\right]^{\frac{1}{2}}. \tag{32}$$

Hence, with probability at least $1-\frac{\delta^*}{2}$,

$$\begin{aligned}\frac{1}{N^2}\|Z_+Z_+{}^T - Z(w)Z(w)^T\|_F^2 &\leq (1-w)^T\frac{R}{\sqrt{S}}\frac{R^T}{\sqrt{S}}(1-w) + \|1-w\|_1^2\frac{1}{\sqrt{S}J_+^2}\log\left[\frac{4J_+^2}{\delta^*}\right]^{\frac{1}{2}}\\ &= \frac{1}{S}\|r-r(w)\|_2^2 + \|1-w\|_1^2\frac{1}{\sqrt{S}J_+^2}\log\left[\frac{4J_+^2}{\delta^*}\right]^{\frac{1}{2}}\end{aligned}$$

Lemma A.2 implies that there exists a $0\leq\nu<1$ such that $\frac{1}{S}\|r-r(w)\|_2^2\leq\nu^{2J-2}$. Since ν depends on the pairs i_l,j_l picked, we can take ν^* to be the largest ν possible. Since the set of all possible S pairs is finite, that implies $0\leq\nu^*<1$. Hence, setting $J=\frac{1}{2}\log_{\nu^*}\left(\frac{\epsilon}{2}\right)+2$ guarantees that $\frac{1}{S}\|r-r(w)\|_2^2\leq\frac{\epsilon}{2}$ for any collection of drawn $i_l,j_l, 1\leq l\leq S$. Assume for any $a\in(0,1]$ and $\delta>0$, we can find an M such that

$$\mathbb{P}\left(\max_j \sigma^2/(J_+^2\sigma_j^2) > M\right) < a\delta. \tag{33}$$

If Eq. (33) holds, we may assume $\max_j \sigma^2/(J_+^2\sigma_j^2) < M$ by setting M large enough since we just need a $1-\delta$ probabilistic guarantee. By the polytope constraint in Eq. (9), $w_i^*\leq\frac{\hat{\sigma}}{\hat{\sigma}_i}$ for all $i\in[J_+]$. Without loss of generality, assume the first J components of w^* can be the only non-zero values since w^* is at least J sparse. For $S\geq 8\frac{M^4}{\sigma^4}\log\left(\frac{2J_+}{\delta^2}\right)$, Lemma B.1 implies with probability at least $1-\frac{\delta^*}{2}$,

$$\begin{aligned}\|1-w^*\|_1^2 &\leq \left(\frac{\hat{\sigma}}{\hat{\sigma}_i}J + (J_+-J)\right)^2\\ &\leq (JMJ_+ + J_+))^2\\ &\leq (2JM\sqrt{5}J_+)^2\\ &\leq 10J_+^2M^2J^2\\ &\leq 10J_+^2M^2\frac{(\log\frac{2}{\epsilon})^2}{(\log\nu)^2}\end{aligned} \tag{34}$$

Therefore, with probability at least $1-\delta^*$,

$$\frac{1}{N^2}\|Z_+Z_+{}^T - Z(w)Z(w)^T\|_F^2 \leq \frac{\epsilon}{2} + \frac{10M^2(\log\frac{2}{\epsilon})^2}{\sqrt{S}(\log\nu)^2}\log\left[\frac{4J_+^2}{\delta^*}\right]^{\frac{1}{2}}. \tag{35}$$

Finally, setting $S\geq\max\left(\frac{100}{\epsilon^2}\left[M\frac{(\log\frac{2}{\epsilon})}{(\log\nu)}\right]^4\log\left[\frac{4J_+{}^2}{\delta^*}\right], 8\frac{M^4}{\sigma^4}\log\left(\frac{2J_+}{\delta^2}\right)\right)$ implies $\frac{1}{N^2}\|Z_+Z_+{}^T - Z(w)Z(w)^T\|_F^2\leq\epsilon$ with probability at least $1-\delta^*$ which matches the rate provided in Theorem 3.2. It remains to show Eq. (33). Notice that

$$\frac{\sigma}{J_+\sigma_j} = \frac{1}{J_+} + \frac{1}{J_+}\sum_{i\neq j}\tilde{\sigma_{ij}}, \tag{36}$$

where $\sigma_{ij}\coloneqq\frac{\sigma_i}{\sigma_j}$. Notice that each σ_{ij} are i.i.d. for $i\neq j$. Let the $\mu_j=\mathbb{E}\sigma_{ij}$ and s_j be the standard deviation of σ_{ij}. Since each σ_j is i.i.d. that implies μ_j and s_j are both constant across j so we drop the subscript. By a union bound, it suffices to show for any $\tau>0$ we can find an M such that

$$\mathbb{P}\left(\max_{1\leq j\leq J_+}\frac{1}{J_+}\sum_{i\neq j}\tilde{\sigma_{ij}} > M\right) < \tau. \tag{37}$$

By Chebyshev's inequality,

$$\mathbb{P}\left(\frac{1}{J_+}\sum_{i\neq j}\tilde{\sigma_{ij}} - \mu > \frac{cs}{J_+}\right) \leq \frac{1}{c^2}. \tag{38}$$

Take $c = J_+\tau$. Then,

$$\mathbb{P}\left(\frac{1}{J_+}\sum_{i\neq j}\tilde{\sigma_{ij}} - \mu > \frac{cs}{J_+}\right) \leq \frac{1}{J_+^2\tau} < \tau. \tag{39}$$

By a union bound, Eq. (38) implies

$$\mathbb{P}\left(\max_{1\leq j\leq J_+}\frac{1}{J_+}\sum_{i\neq j}\tilde{\sigma_{ij}} > M\right) < \frac{1}{\tau J_+} < \tau$$

for $M = \mu + s\tau$ as desired.

The proof showing that $\limsup_{J_+\to\infty}\nu_{J_+} < 1$ is the same as the proof Theorem 3.4.

□

C Runtime analysis of methods

The ridge regression and PCA runtimes depend on the number of features used, as specified in Table 1, and therefore follow from the first column of the table.

First, we show that using RFM with $J_+ = O\left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)$ number of random features ensures that $\frac{1}{N^2}\|K-\hat{K}\|_F^2 = O(\epsilon)$ with high probability. By a union bound, $\mathbb{P}\left(\frac{1}{N^2}\|K-\hat{K}\|_F^2 \leq \epsilon\right) \geq \mathbb{P}\left(\max_{i,j\in[N]}|K_{ij}-\hat{K}_{ij}| \leq \sqrt{\epsilon}\right)$. Now, Claim 1 of Rahimi and Recht (2007) implies

$$\mathbb{P}\left(\max_{i,j\in[N]}|K_{ij}-\hat{K}_{ij}| \geq \sqrt{\epsilon}\right) = O\left(\frac{1}{\epsilon}e^{-J_+\epsilon}\right). \tag{40}$$

Setting the right-hand side of Eq. (40) to some fixed probability threshold δ^* implies $J_+ = O\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon\delta^*}\right)\right)$. Since δ^* is some fixed constant, $J_+ = O\left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)$ number of random features suffices for an $O(\epsilon)$ error guarantee. Hence, it suffices to use $J_+ = O\left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)$ as the up-projection dimension for both RFM-FW and RFM-JL.

To prove the bounds for RFM-FW, take $S = \Omega(J_+^2(\log J_+)^2)$. It is straightforward to check that this choice of S satisfies the requirements of Theorem 3.2. By Theorem 3.2, it suffices to set $J = O(\log J_+)$ for an $O(\epsilon)$ error guarantee. Hence, Algorithm 1 takes $O(SJ_+\log J_+)$ time to compute the random feature weights w since Frank-Wolfe has to be run for a total of $O(\log J_+)$ iterations. Finally, it takes $O(N\log J_+)$ to apply these $O(\log J_+)$ weighted random features to the N datapoints. We conclude by proving the time complexity of RFM-JL.

Denote $\tilde{x_i} \coloneqq (Z_+)_i \in \mathbb{R}^{J_+}$ as the mapped datapoints from RFM. Let $A \in \mathbb{R}^{J\times J_+}$ for $J \leq J_+$ be a matrix filled with i.i.d. $N(0,\frac{1}{J})$ random variables for the JL compression step. Let $f(x) := Ax$. It suffices to pick a J such that,

$$\mathbb{P}\left(\max_{i,j\in[N]}\left|\tilde{x_i}^T\tilde{x_j} - f(\tilde{x_i})^T f(\tilde{x_j})\right| \geq \sqrt{\epsilon}\right) \leq \delta^* \tag{41}$$

for RFM-JL. We use the following corollary from Kakade and Shakhnarovich (2009, Corollary 2.1) to bound the above probability.

Lemma C.1. *Let $u, v \in \mathbb{R}^d$ and such that $\|u\| \leq 1$ and $\|v\| \leq 1$. Let $f(x) = Ax$, where A is a $k \times d, k \leq d$ matrix of i.i.d. $N(0,\frac{1}{k})$ random variables. Then,*

$$\mathbb{P}\left(\mid u^Tv - f(u)^Tf(v) \mid\right) \leq 4e^{-\frac{1}{4}(\epsilon^2-\epsilon^3)k}. \tag{42}$$

$\|\tilde{x_i}\|_2 = 1$ since $\tilde{x_i} = \frac{1}{\sqrt{J_+}}\left(\cos(\omega_1^Tx_i+b),\cdots,\cos(\omega_{J_+}^Tx_i+b)\right)$. Hence, we may apply Lemma C.1 to $\tilde{x_i}$. By a union bound and an application of Lemma C.1, Eq. (41) is bounded by $O\left(N^2e^{-J\epsilon}\right)$. Setting $N^2e^{-J\epsilon}$ equal to δ^* and solving for J implies that $J = \Omega\left(\frac{1}{\epsilon}\log\left(\frac{N^2}{\delta^*}\right)\right)$. Hence, $J = O\left(\frac{1}{\epsilon}\log N\right)$. Now, $O\left(\frac{1}{\epsilon}\right) = O\left(\frac{J_+}{\log\frac{1}{\epsilon}}\right)$ which implies $J = O\left(\frac{J_+\log N}{\log\frac{1}{\epsilon}}\right)$. Since $N > J_+ > O\left(\frac{1}{\epsilon}\right)$, $J = \Omega(J_+)$ suffices for an for an $O(\epsilon)$ error guarantee. While the JL algorithm typically takes $O(NJ_+k)$ time to map a $N \times J_+$ matrix to a $N \times k$ matrix, the techniques in Hamid et al. (2014, Section 3.5) show that only $O(NJ_+\log J)$ time is required by using the Fast-JL algorithm.

D Impact of kernel approximation

Here we provide the precise error bound and runtimes for kernel ridge regression, kernel SVM, and kernel PCA when using a low-rank factorization ZZ^T of K. We denote $X \subset \mathbb{R}^p$ as the input space and define $c > 0$ such that $K(x,x) \leq c$ and $\hat{K}(x,x) \leq c$ for all $x \in X$. This condition is verified with $c = 1$ for Gaussian kernels for example. All the bounds provided follow from Cortes et al. (2010); Talwalkar (2010), where we simply replace the spectral norm with the Frobenius norm since the Frobenius norm upper bounds the spectral norm.

D.1 Kernel ridge regression

Exact kernel ridge regression takes $O(N^3)$ since K must be inverted. Suppose $K \approx ZZ^T := \hat{K}$, where Z could be found using RFM for example. Running ridge regression with the feature matrix Z just requires computing and inverting the covariance matrix $Z^TZ \in R^{J\times J}$ which takes $\Theta(\max(J^3, NJ^2))$ time. Proposition D.1 quantifies the error between the regressor obtained from K and the one from $\hat{K}$.

Proposition D.1. *(Proposition 1 of Cortes et al. (2010)) Let $\hat{f}$ denote the regression function returned by kernel ridge regression when using the approximate kernel matrix $\hat{K} \in \mathbb{R}^{N\times M}$, and f^* the function returned when using the exact kernel matrix K. Assume that every response y is bounded in absolute value by M for some $0 < M < \infty$. Let $\lambda := N\lambda_0 > 0$ be the ridge parameter. Then, the following inequality holds for all $x \in X$:*

$$\begin{aligned}
|\hat{f}(x) - f^*(x)| &\leq \frac{cM}{\lambda_0^2 N}\|\hat{K} - K\|_2 \\
&\leq \frac{cM}{\lambda_0^2 N}\|\hat{K} - K\|_F \\
&= O\left(\frac{1}{N}\|\hat{K} - K\|_F\right)
\end{aligned}$$

D.2 Kernel SVM

Kernel SVM regression takes $O(N^3)$ using K since K must be inverted. Again suppose $K \approx ZZ^T := \hat{K}$. Then, training a linear SVM via dual-coordinate decent on Z has time complexity $O\left(NJ\log\rho\right)$, where ρ is the optimization tolerance Hsieh et al. (2008).

Proposition D.2. *(Proposition 2 of Cortes et al. (2010)) Let $\hat{f}$ denote the hypothesis returned by SVM when using the approximate kernel matrix $\hat{K}$, f^* the hypothesis returned when using the exact kernel matrix K, and C_0 be the penalty for SVM. Then, the following inequality holds for all $x \in X$:*

$$\begin{aligned}
|\hat{f}(x) - f^*(x)| &\leq \sqrt{2}c^{\frac{3}{4}}C_0\|\hat{K} - K\|_2^{\frac{1}{4}}\left[1 + \frac{\|\hat{K} - K\|_2^{\frac{1}{4}}}{4c}\right] \\
&\leq \sqrt{2}c^{\frac{3}{4}}C_0\|\hat{K} - K\|_F^{\frac{1}{4}}\left[1 + \frac{\|\hat{K} - K\|_F^{\frac{1}{4}}}{4c}\right]. \\
&= O\left(\|\hat{K} - K\|_F^{\frac{1}{2}}\right).
\end{aligned}$$

D.3 Kernel PCA

We follow Talwalkar (2010) to understand the effect matrix approximation has on kernel PCA. For a more in-depth analysis, see pg. 92-98 of Talwalkar (2010). Without loss of generality, we assume the data are mean zero.

Let $\Phi(\cdot)$ be the unique feature map such that $k(x,y) = \langle\Phi(x), \Phi(y)\rangle$. Let the feature covariance matrix be denoted as $\Sigma_\Phi := \Phi(X_N)\Phi(X_N)^T$, where $\Phi(X_N) := [\Phi(x_1)\cdots\Phi(x_n)]$. Since the rank of Σ_Φ is at most N, let v_i $1 \leq i \leq N$ be the N singular vectors of Σ_Φ. For certain kernels, e.g., the RBF kernel, the v_i are infinite dimensional. However, the projection of $\Phi(x)$ onto each v_i is tractable to compute via the kernel trick:

$$\Phi(x)^T v_i = \Phi(x)\frac{\Phi(X_N)u_i}{\sqrt{\sigma_i}} = \frac{k_x^T u_i}{\sqrt{\sigma_i}}, \tag{43}$$

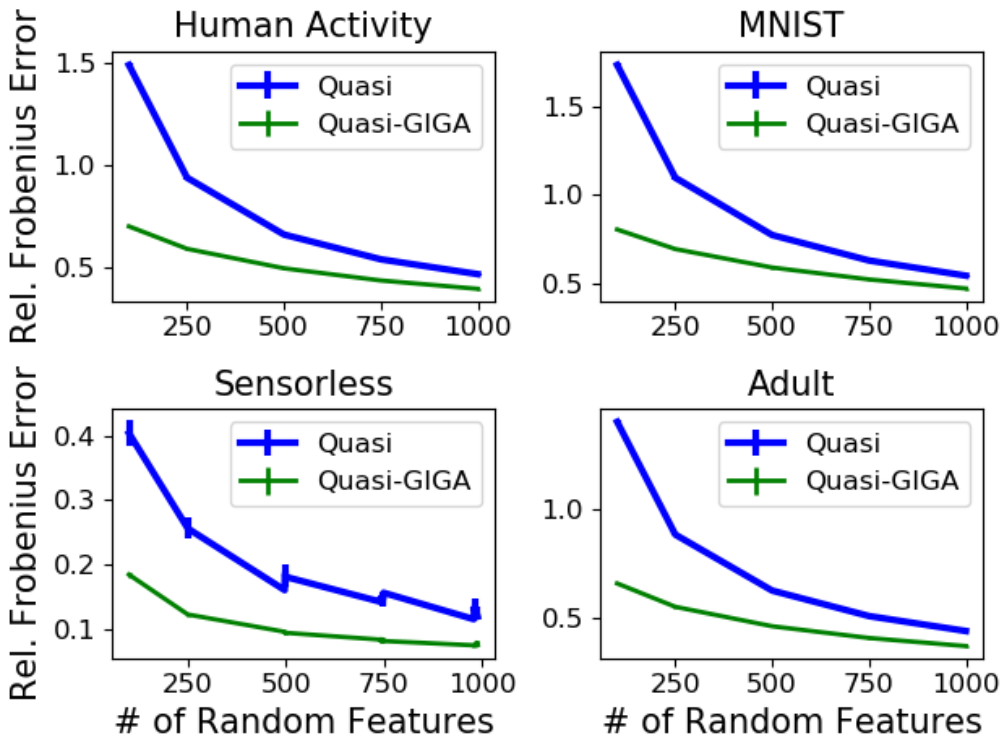

Figure 7: Kernel matrix approximation errors. Lower is better. Each point denotes the average over 20 simulations and the error bars represent one standard deviation. The HALTON sequence was used to generate the quasi random features.

where $k_x := (K(x_1, x), \cdots, K(x_N, x))$ and u_i is the ith singular vector of K with associated eigenvalue σ_i. Often, the goal is to project $\Phi(x)$ onto the first l eigenvectors of Σ_Φ for dimensionality reduction. To analyze the error of the projection, let P_{V_l} be defined as the subspace V_l spanned by the top l eigenvectors of Σ_Φ. Then, the *average empirical residual* $R_l(K)$ of a kernel matrix K is defined as,

$$\begin{aligned} R_l(K) &:= \frac{1}{N}\sum_{n=1}^{N} \|\Phi(x_n)\|^2 - \frac{1}{N}\sum_{n=1}^{N} \|P_{V_l}(\Phi(x_n))\|^2 \\ &= \sum_{i>l} \sigma_i \end{aligned} \tag{44}$$

$R_l(K)$ is simply the spectral error of a low-rank decomposition of Σ_Φ using the SVD. If we instead use $\hat{K}$ for the eigendecomposition, the following proposition bounds the difference between $R_l(K)$ and $R_l(\hat{K})$.

Proposition D.3. *(Proposition 5.4 of Talwalkar (2010)) For $R_l(K)$ and $R_l(\hat{K})$ defined as above,*

$$\begin{aligned} |R_l(K) - R_l(\hat{K})| &\leq \left(1 - \frac{l}{N}\right) \|K - \hat{K}\|_2 \\ &\leq \left(1 - \frac{l}{N}\right) \|K - \hat{K}\|_F. \end{aligned}$$

E Additional Experiments

As stated in Section 4, our method may be applied on top of other random feature methods. In particular, many previous works have reduced the number of random features needed for a given level of approximation by sampling them from a different distribution (e.g., through importance sampling or Quasi-Monte-Carlo techniques). Regardless of the way the random features are sampled, our method can still be used for compression.

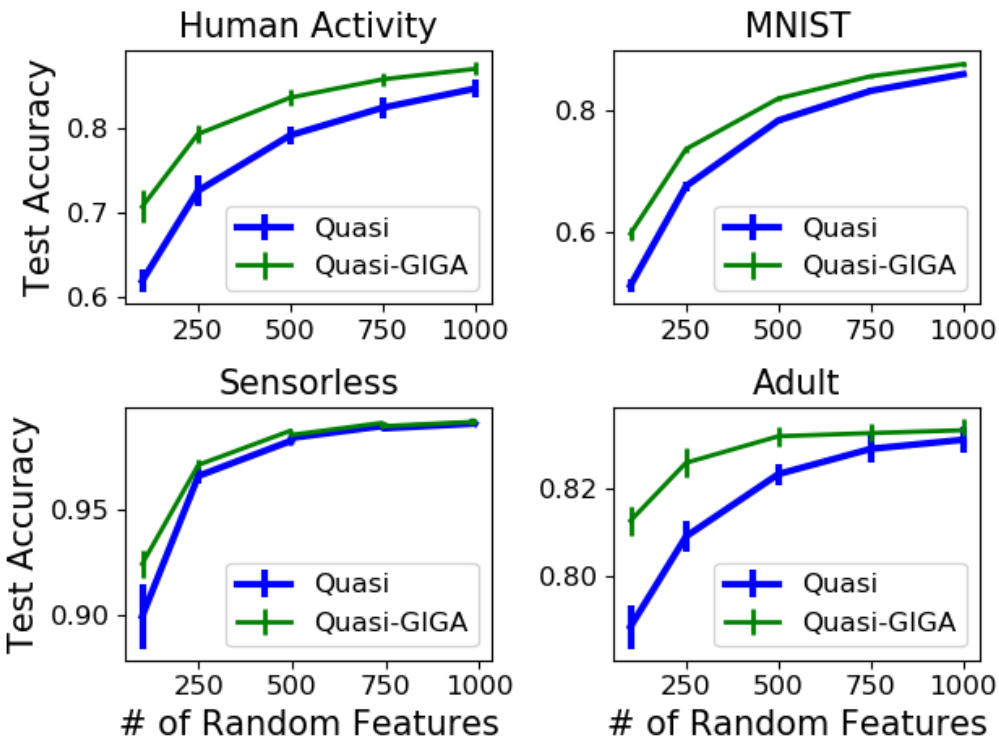

Figure 8: Classification accuracy. Higher is better. Each point denotes the average over 20 simulations and the error bars represent one standard deviation. The HALTON sequence was used to generate the Quasi random features.

To demonstrate this point further, we consider generating random features using Quasi-Monte-Carlo (Avron et al., 2016). Quasi random features work by generating a sequence of points from a (low-discrepancy) grid of points in $[0, 1]^p$. Points are sampled from the target random-features distribution Q by applying the inverse CDF of Q on each of these points in the sequence. In Avron et al. (2016), the authors showed that generating random features in this way improved performance over the classical random features method provided in Rahimi and Recht (2007). In Fig. 7 and Fig. 8, we see that our method is able to compress the number of quasi random features, which is similar to the behavior in Fig. 1 and Fig. 2. Note that the experimental setup is exactly the same as in Section 4 except that the random features are now generated using Quasi-Monte-Carlo.