Recycling Randomness with Structure for Sublinear time Kernel Expansions

Krzysztof Choromanski KCHORO@GOOGLE.COM
Vikas Sindhwani SINDHWANI@GOOGLE.COM

Abstract

We propose a scheme for recycling Gaussian random vectors into structured matrices to approximate various kernel functions in sublinear time via random embeddings. Our framework includes the Fastfood construction of Le et al. (2013) as a special case, but also extends to Circulant, Toeplitz and Hankel matrices, and the broader family of structured matrices that are characterized by the concept of low-displacement rank. We introduce notions of coherence and graph-theoretic structural constants that control the approximation quality, and prove unbiasedness and low-variance properties of random feature maps that arise within our framework. For the case of low-displacement matrices, we show how the degree of structure and randomness can be controlled to reduce statistical variance at the cost of increased computation and storage requirements. Empirical results strongly support our theory and justify the use of a broader family of structured matrices for scaling up kernel methods using random features.

1. Introduction

Consider a k-dimensional feature map of the form,

$$\Psi(\mathbf{x}) = \frac{1}{\sqrt{k}} s(\mathbf{M}\mathbf{x}) \tag{1}$$

where the input data vector $\mathbf{x}$ is drawn from $\mathbb{R}^n$, $s(\cdot)$ denotes a real-valued or complex-valued pointwise nonlinearity (activation function), and $\mathbf{M}$ is a $k \times n$ Gaussian random matrix. It is well known that as a function of a pair of data vectors, the Euclidean inner product $\Psi(\mathbf{x})^T \Psi(\mathbf{z})$, converges to a positive definite kernel function $\mathcal{K}(\mathbf{x}, \mathbf{z})$ depending on the choice of the scalar nonlinearity, as $k \to \infty$. For example, the complex exponential nonlinearity $s(x) = e^{-i\frac{x}{\sigma}}$ corresponds to the Gaussian kernel (Rahimi & Recht, 2007), while the rectified linear function (ReLU), $s(x) = \max(x, 0)$, leads to the Arc-cosine kernel (Cho & Saul, 2009).

In recent years, such random feature maps have been used to dramatically accelerate the training time and inference speed of kernel methods (Schölkopf & Smola, 2002) across a variety of statistical modeling problems (Rahimi & Recht, 2007; Xie et al., 2015) and applications (Huang et al., 2014; Vedaldi & Zisserman, 2012). Standard linear techniques applied to random nonlinear embeddings of data are equivalent to learning with approximate kernels. To quantify the benefits, consider solving a kernel ridge regression task given l training examples. With traditional kernel methods, dense linear algebra operations on the Gram matrix associated with the exact kernel function imply that the training complexity grows as $O(l^3 + l^2 n)$ and the time to make a prediction on a test sample grows as $O(ln)$. By contrast, random feature approximations reduce training complexity to $O(lk^2 + lkn)$ and test speed to $O(kn)$. This is a major win on big datasets where l is very large, provided that a small value of k can provide a good approximation to the kernel function.

In practice, though, the optimal value of k is often large, albeit still much smaller than l. For example, in a speech recognition application (Huang et al., 2014) involving around two million training examples, about hundred thousand random features are required to achieve state of the art results. In such settings, the time to construct the random feature map is dominated by matrix multiplication against the dense Gaussian random matrix, which becomes the new computational bottleneck. To alleviate this bottleneck, (Le et al., 2013) introduce the "Fastfood" approach where Gaussian random matrices are replaced by Hadamard matrices combined with diagonal matrices with Gaussian distributed diagonal entries. It was shown in (Le et al., 2013) that for the specific case of the complex exponential nonlinearity, the Fastfood feature maps provide unbiased estimates for the Gaussian kernel function, at the expense of additional statistical variance, but with the computational benefit of reducing the feature map construction time from $O(kn)$ to $O(k \log\ n)$ by using the Fast Walsh-Hadamard transform for matrix multiplication. The Fastfood construction for kernel approximations is akin to

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

the use of structured matrices - in lieu of Gaussian random matrices - in Fast Johnson-Lindenstrauss transform (FJLT) (Alon & Chazelle, 2009) for dimensionality reduction, fast compressed sensing (Bajwa et al., 2007; Rauhut et al., 2012), and randomized numerical linear algebra techniques (Alon & Chazelle, 2011; Mahoney, 2011) Specific structured matrices were recently applied for approximating angular kernels (Choromanska et al., 2016). Some heuristic results for approximating kernels with circulant matrices were given in (Yu et al., 2015).

Our contributions in this paper are as follows:

- We study a general family of structured random matrices that can be constructed by recycling a Gaussian random vector using a sequence of elementary generator matrices (introduced in Section 3). This family includes Circulant, Toeplitz and Hankel matrices. It also includes the Fastfood construction of (Le et al., 2013) as a special case. We show that fast sublinear time random feature maps obtained from these matrices provide unbiased estimates of the exact kernel, with variance comparable to the fully unstructured Gaussian case (Section 4). We introduce various structural coherence and graph-theoretic constants that control the quality of randomness we get from our model. Our approach generalizes across various choices of nonlinearities and kernel functions.
- Of particular interest for us is the class of generalized structured matrices that have low-displacement rank (Pan, 2001; Sindhwani et al., 2015). Such matrices span an increasingly rich class of structures as the displacement rank is increased: from Circulant and Toeplitz matrices, to inverses and products of Toeplitz matrices, and more. The displacement rank provides a knob with which the degree of structure and randomness can be controlled to tradeoff computational and storage requirements against statistical variance.
- We provide empirical support for our theoretical results (Section 5). In particular, we show that Circulant, Fastfood and low-displacement Toeplitz-like matrices provide high quality sublinear-time feature maps for approximating various kernels. With increasing displacement rank, the quality of the approximation approaches that of the fully Gaussian random matrix.

2. Background and Preliminaries

We start by giving a brisk background on random feature maps and structured matrices.

2.1. Random Embeddings, Nonlinearities and Kernels

Random feature maps may be viewed as arising from Monte-Carlo approximations to integral representations of kernel functions. The original construction by Rahimi & Recht (2007) was motivated by a classical result that characterizes the class of shift-invariant positive definite functions.

Theorem 2.1 (Bochner's Theorem (Bochner, 1933))**.** *A continuous shift-invariant scaled kernel function $\mathcal{K}(\mathbf{x},\mathbf{z}) \equiv \phi(\mathbf{x}-\mathbf{z})$ on $\mathbb{R}^n$ is positive definite if and only if it is the Fourier transform of a unique finite probability measure p on $\mathbb{R}^n$. That is, for any $\mathbf{x},\mathbf{z} \in \mathbb{R}^d$,*

$$\mathcal{K}(\mathbf{x},\mathbf{z}) = \int_{\mathbb{R}^n} e^{-i(\mathbf{x}-\mathbf{z})^T\mathbf{w}} p(\mathbf{w}) d\mathbf{w} = \mathbb{E}_{\mathbf{w}\sim p}[e^{-i(\mathbf{x}-\mathbf{z})^T\mathbf{w}}] \,.$$

Bochner's theorem stablishes one-to-one correspondence between shift-invariant kernel functions and probability densities on $\mathbb{R}^n$, via the Fourier transform. In the case of the Gaussian kernel with bandwidth σ, the associated density is also Gaussian with covariance matrix σ^{-2} times the identity.

While studying synergies between kernel methods and deep learning, (Cho & Saul, 2009) introduce b^{th}-order *arc-cosine* kernels via the following integral representation:

$$\mathcal{K}_b(\mathbf{x},\mathbf{z}) = \int_{\mathbb{R}^d} i(\mathbf{w}^T\mathbf{x}) i(\mathbf{w}^T\mathbf{z}) (\mathbf{w}^T\mathbf{x})^b (\mathbf{w}^T\mathbf{z})^b \; p(\mathbf{w}) d\mathbf{w}$$

where $i(\cdot)$ is the step function, i.e. $i(x) = 1$ if $x > 0$ and 0 otherwise; and the density p is chosen to be standard Gaussian. These kernels evaluate inner products in the representation induced by an infinitely wide single hidden layer neural network with random Gaussian weights, and admit closed form expressions in terms of the angle $\theta = cos^{-1}(\frac{\mathbf{x}^T\mathbf{z}}{\|\mathbf{x}\|_2\|\mathbf{z}\|_2})$ between $\mathbf{x}$ and $\mathbf{z}$:

$$\mathcal{K}_0(\mathbf{x},\mathbf{z}) = 1 - \frac{\theta}{\pi} \quad (2)$$

$$\mathcal{K}_1(\mathbf{x},\mathbf{z}) = \frac{\|\mathbf{x}\|_2\|\mathbf{z}\|_2}{\pi}[sin(\theta) + (\pi-\theta)cos(\theta)] \quad (3)$$

where $\|\cdot\|_2$ denotes l_2 norm.

Monte Carlo approximations to the integral representations above lead to the following,

$$\mathcal{K}(\mathbf{x},\mathbf{z}) \approx \frac{1}{k}\sum_{j=1}^{k} s(\mathbf{x}^T\mathbf{w}_j) s(\mathbf{z}^T\mathbf{w}_j) = \Psi(\mathbf{x})^T\Psi(\mathbf{z}) \quad (4)$$

where the feature map $\Psi(\mathbf{x})$ has the form given in Eqn. 1, with rows of $\mathbf{M}$, i.e. $\mathbf{w}_j$ vectors, drawn from the Gaussian density, and the nonlinearity s set to the following: complex exponential, $s(x) = e^{i\frac{x}{\sigma}}$, for the Gaussian kernel with bandwidth σ; hard-thresholding, $s(x) = i(x)$, for the angular similarity kernel in Eqn. 2; and ReLU activation, $s(x) = \max(x, 0)$, for the first order arc-cosine kernel in Eqn. 3.

2.2. Structured Matrices

A $m \times n$ matrix is called a structured matrix if it satisfies the following two properties: (1) it has much fewer degrees of freedom than mn independent entries, and hence can be implicitly stored more efficiently than general matrices, and (2) the structure in the matrix can be exploited for fast linear algebra operations such as fast matrix-vector multiplication. Examples include the Discrete Fourier Transform (DFT), the Discrete Cosine Transform (DCT) and the Walsh-Hadamard Transform (WHT) matrices. Here, we give other examples particularly relevant to this paper. The matrices described below are square. Rectangular matrices can be obtained by appropriately selecting rows or columns.

Circulant Matrices: These matrices are intimately associated with circular convolutions and have been used for fast compressed sensing in (Rauhut et al., 2012). A $n \times n$ Circulant matrix is completely determined by its first column/row, i.e., n parameters. Each column/row of a Circulant matrix is generated by cyclically down/right-shifting the previous column/row. A *skew-Circulant* matrix has identical structure to Circulant, except that the upper triangular part of the matrix is negated. This general structure looks like,

$$\begin{bmatrix} \mathbf{g_0} & f\mathbf{g_{n-1}} & \dots & f\mathbf{g_1} \\ \mathbf{g_1} & \mathbf{g_0} & \dots & \vdots \\ \vdots & \vdots & \vdots & f\mathbf{g_{n-1}} \\ \mathbf{g_{n-1}} & \dots & \mathbf{g_1} & \mathbf{g_0} \end{bmatrix}$$

with $f = 1$ for Circulant and $f = -1$ for skew-Circulant matrix. Both these matrices admit $O(n \log n)$ matrix-vector multiplication as they are diagonalized by the DFT matrix (Pan, 2001). We will use the notation $\texttt{circ}[\mathbf{g}]$ and $\texttt{scirc}[\mathbf{g}]$ for Circulant and skew-Circulant matrices respectively.

Toeplitz and Hankel Matrices: These matrices implement discrete linear convolution and arise naturally in dynamical systems and time series analysis. Toeplitz matrices are characterized by constant diagonals as follows,

$$\begin{bmatrix} \mathbf{t_0} & \mathbf{t_{-1}} & \dots & \mathbf{t_{-(n-1)}} \\ \mathbf{t_1} & \mathbf{t_0} & \dots & \vdots \\ \vdots & \vdots & \vdots & \mathbf{t_{-1}} \\ \mathbf{t_{n-1}} & \dots & \mathbf{t_1} & \mathbf{t_0} \end{bmatrix}$$

Closely related Hankel matrices have constant anti-diagonals. Toeplitz-vector multiplication can be reduced to $O(n \log n)$ Circulant-vector multiplication. For detailed properties of Circulant and Toeplitz matrices, we point the reader to (Gray, 2006)

Structured Matrices with Low-displacement Rank: The notion of displacement operators and displacement rank (Golub & Loan, 2012; Pan, 2001; Kailath et al., 1979) can be used to broadly generalize various classes of structured matrices. For example, under the action of the *Sylvester displacement operator* defined as $L[\mathbf{T}] = \mathbf{Z}_1\mathbf{T} - \mathbf{T}\mathbf{Z}_{-1}$, every Toeplitz matrix can be transformed into a matrix of rank at most 2 using elementary shift and scale operations implemented by matrices of the form $\mathbf{Z}_f = [\mathbf{e}_2\mathbf{e}_3 \dots \mathbf{e}_n \; f\mathbf{e}_1]$ for $f = \pm 1$ where $\mathbf{e}_1 \dots \mathbf{e}_n$ are column vectors representing the standard basis of $\mathbb{R}^n$.

For a given displacement rank parameter r, the class of matrices for which the rank of $L[\mathbf{T}]$ is at most r is called *Toeplitz-like*. Remarkably, this class of matrices admits a closed-form parameterization in terms of the low-rank factorization of $L[\mathbf{T}]$:

Theorem 2.2 (Parameterization of Toeplitz-like matrices with displacement rank r (Pan, 2001))**.** *: If an $n \times n$ matrix $\mathbf{T}$ satisfies $rank(\mathbf{Z}_1\mathbf{T} - \mathbf{T}\mathbf{Z}_{-1}) \leq r$, then it can be written as,*

$$\mathbf{T} = \sum_{i=1}^{r} \texttt{circ}[\mathbf{g}^i] \; \texttt{scirc}[\mathbf{h}^i] \tag{5}$$

for some choice of vectors $\{\mathbf{g}^i, \mathbf{h}^i\}_{i=1}^r \in \mathbb{R}^n$.

The family of matrices expressible by Eqn. 5 is very rich (Pan, 2001), i.e., it covers (i) all Circulant and Skew-circulant matrices for $r = 1$, (ii) all Toeplitz matrices and their inverses for $r = 2$, (iii) Products, inverses, linear combinations of distinct Toeplitz matrices with increasing r, and (iv) all $n \times n$ matrices for $r = n$. Since Toeplitz-like matrices under the parameterization of Eqn. 5 are a sum of products between Circulant and Skew-circulant matrices, they inherit fast FFT based matrix-vector multiplication with cost $O(nrlog\ n)$, where r is the displacement rank. Hence, r provides a knob on the degree of structure imposed on the matrix with which storage requirements, computational constraints and statistical capacity can be explicitly controlled. Recently such matrices were used in the context of learning mobile-friendly neural networks in (Sindhwani et al., 2015). We note in passing that the displacement rank framework generalizes to other types of base structures (e.g. Vandermonde); see (Pan, 2001).

2.3. FastFood

In the context of fast kernel approximations, (Le et al., 2013) introduce the Fastfood technique where the matrix $\mathbf{M}$ in Eqn. 1 is parameterized by a product of diagonal and simple matrices as follows:

$$\mathbf{F} = \frac{1}{\sqrt{n}}\mathbf{SHGPHB}. \tag{6}$$

Here, $\mathbf{S}, \mathbf{G}, \mathbf{B}$ are diagonal random matrices, $\mathbf{P}$ is a permutation matrix and $\mathbf{H}$ is the Walsh-Hadamard matrix. The $k \times n$ matrix $\mathbf{M}$ is obtained by vertically stacking k/n independent copies of the $n \times n$ matrix $\mathbf{F}$. Multiplication

against such a matrix can be performed in time $O(k \log\ n)$. The authors prove that (1) the Fastfood approximation is unbiased, (2) its variance is at most the variance of standard Gaussian random features with an additional $O(\frac{1}{k})$ term, and (3) for a given error probability δ, the pointwise approximation error of a $n \times n$ block of Fastfood is at most $O(\sqrt{\log(n/\delta)})$ larger than that of standard Gaussian random features. However, note that the Fastfood analysis is limited to the Gaussian kernel and their variance bound uses properties of the complex exponential. The authors also conjecture that the Hadamard matrix $\mathbf{H}$ above, can be replaced by any matrix $\mathbf{T}$ such that $\mathbf{T}/\sqrt{n}$ is orthonormal, the maximum entry in $\mathbf{T}$ is small, and matrix-vector product against $\mathbf{T}$ can be computed in $O(n \log\ n)$ time.

3. Structured Matrices from Gaussian Vectors

In this section, we present a general structured matrix model that allows a small Gaussian vector to be recycled in order to mimic the properties of a Gaussian random matrix suitable for generating random features. We first introduce some basic concepts in our construction. Note that we emphasize intuitions in our exposition - formal proofs are provided in our supplementary material.

3.1. The $\mathcal{P}$-model

Budget of Randomness: Let t be some given parameter. Consider the column vector $\mathbf{g} = (g_1, ..., g_t)^T$, where each entry is an independent Gaussian taken from $\mathcal{N}(0, 1)$. This vector stands for the "budget of randomness" used in our structured matrix construction scheme.

Our goal is to recycle the Gaussian vector $\mathbf{g}$ to construct random matrices with desirable properties. This is accomplished using a sequence of matrices which we call the $\mathcal{P}$-model.

Definition 3.1 ($\mathcal{P}$-model)**.** *Given the budget of uncertainty parameter t, a sequence of m matrices with unit l_2 norm columns, denoted as $\mathcal{P} = \{\mathbf{P}_i\}_{i=1}^m$, where $\mathbf{P}_i \in \mathbb{R}^{t \times n}$, specifies a $\mathcal{P}$-model. Such a sequence defines an $m \times n$ random matrix of the form:*

$$\mathbf{S}[\mathcal{P}] = \begin{pmatrix} \mathbf{g}^T\mathbf{P}_1 \\ \mathbf{g}^T\mathbf{P}_2 \\ \vdots \\ \mathbf{g}^T\mathbf{P}_m \end{pmatrix} \tag{7}$$

where $\mathbf{g}$ is a Gaussian random vector of length t.

In the constructions of interest to us, the sequence $\mathcal{P}$ is designed to separate structure from Gaussian randomness; though elements of $\mathcal{P}$ can be deterministic or itself random, Gaussianity is restricted to the vector $\mathbf{g}$. The ability of $\mathcal{P}$ to recycle a Gaussian vector effectively depends on certain structural constants that we now define.

Definition 3.2 (Coherence of a $\mathcal{P}$-model)**.** *For $\mathcal{P} = \{\mathbf{P}_i\}_{i=1}^m$, let $\mathbf{P}_{ij}$ denote the j^{th} column of the i^{th} matrix. The coherence of a $\mathcal{P}$-model is defined as,*

$$\mu[\mathcal{P}] = \max_{1 \leq i \leq j \leq m} \sqrt{\frac{\sum_{1 \leq n_1 < n_2 \leq n} (\mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_2})^2}{n}} \tag{8}$$

Note that $\mu[\mathcal{P}]$ is a maximum over all pairs of rows $1 \leq i \leq j \leq m$ of the rescaled sums of cross-correlations $\mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_2}$ for all pairs of different column indices n_1, n_2. Lower values of $\mu[\mathcal{P}]$ will lead to better quality models. In practice, as we will see in subsequent analysis, it suffices if $\mu[\mathcal{P}] = O(poly(\log(n)))$ which is the case for instance for Toeplitz and Circulant matrices.

The coherence of the $\mathcal{P}$-model is an extremal statistic of pairwise correlations. We couple it with another set of objects describing global structural properties of the model, namely the *coherence graphs*.

Definition 3.3 (Coherence Graphs for $\mathcal{P}$-model and their Chromatic Numbers)**.** *Let $1 \leq i, j \leq m$. We define by $\mathcal{G}_{i,j}$ an undirected graph with the set of vertices $V(\mathcal{G}_{i,j}) = \{\{n_1, n_2\} : 1 \leq n_1 \neq n_2 \leq n$ and $\mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_2} \neq 0\}$ and the set of edges $E(\mathcal{G}_{i,j}) = \{\{\{n_1, n_2\}, \{n_2, n_3\}\} : \{n_1, n_2\}, \{n_2, n_3\} \in V(\mathcal{G}_{i,j})\}$. In other words, edges are between these vertices such that their corresponding 2-element subsets intersect. The chromatic number $\chi(i, j)$ of a graph $\mathcal{G}_{i,j}$ is the smallest number of colors that can be used to color all vertices of $\mathcal{G}_{i,j}$ in such a way that no two adjacent vertices share the same color.*

The chromatic number of a $\mathcal{P}$-model is defined as follows:

Definition 3.4 (Chromatic number of a $\mathcal{P}$-model)**.** *The chromatic number $\chi[\mathcal{P}]$ of a $\mathcal{P}$-model is given as:*

$$\chi[\mathcal{P}] = \max_{1 \leq i \leq j \leq m} \chi(i, j),$$

where $\mathcal{G}_{i,j}$ are associated coherence graphs.

As it was the case for the coherence $\mu[\mathcal{P}]$, smaller values of the chromatic number $\chi[\mathcal{P}]$ lead to better theoretical results regarding the quality of the model. Intuitively speaking, coherence graphs encode in a compact combinatorial way correlations between different rows of the structured matrix produced by the $\mathcal{P}$-model. The chromatic number $\chi[\mathcal{P}]$ is a single combinatorial parameter measuring quantitatively these dependencies. It can be easily computed or at least upper-bounded (which is enough for us) for $\mathcal{P}$-models related to all structured matrices considered in this paper. The following is a well-known fact from graph theory:

Lemma 3.1. *The chromatic number $\chi(G)$ of an undirected graph G with maximum degree d_{max} satisfies: $\chi(G) \leq d_{max} + 1$.*

For all instantiations of $\mathcal{P}$-models considered in this paper leading to various structured matrices, the vertices of associated coherence graphs will turn out to have small degrees and hence, by Lemma 3.1, small chromatic numbers.

We will introduce one more structural parameter of the $\mathcal{P}$-model, depending on whether it is specified deterministically or randomly.

Definition 3.5. *The uni-coherence $\tilde{\mu}[\mathcal{P}]$ of the $\mathcal{P}$-model is defined as follows. If matrices $\mathbf{P}_i$ are constructed deterministically then $\tilde{\mu}[\mathcal{P}] = \max_{1 \leq i < j \leq m} \sum_{n_1=1}^{n} |\mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_1}|$. If the matrices that specify $\mathcal{P}$ are constructed randomly, then we take $\tilde{\mu}[\mathcal{P}] = \max_{1 \leq i < j \leq m} \mathbb{E}[|\sum_{n_1=1}^{n} \mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_1}|]$.*

It turns out that the sublinearity in n of uni-coherence $\tilde{\mu}[\mathcal{P}]$ helps to establish strong theoretical results regarding the quality of the $\mathcal{P}$-model.

3.2. Examples of $\mathcal{P}$-model structured matrices

Below we observe that various structured random matrices can be constructed according to the $\mathcal{P}$-model, i.e. by specifying a sequence of matrices $\mathbf{P}_i$ in Eqn. 7. We note that chromatic numbers and coherence values of these $\mathcal{P}$-models are low. In the next section, we show that this implies that we can get unbiased, low-variance kernel approximations from these matrices, for various choices of nonlinearities. Here we consider square structured matrices for which $m = n$, or rectangular matrices with $m < n$ obtained by selecting first m rows of a structured matrix.

3.2.1. Circulant matrices

Circulant matrices can be constructed via the $\mathcal{P}$-model with budget of randomness $t = n$ and matrices $\{\mathbf{P}_i\}_{i=1}^m$ of entries in $\{0, 1\}$. See Fig. 1 for an illustrative construction. The coherence of the related $\mathcal{P}$-model trivially satisfies: $\mu[\mathcal{P}] = O(1)$ and $\tilde{\mu}[\mathcal{P}] = 0$. The coherence graphs are vertex disjoint cycles. Since each cycle can be colored with at most 3 colors, the chromatic number of the $\mathcal{P}$-model satisfies: $\chi[\mathcal{P}] \leq 3$.

3.2.2. Toeplitz and Hankel matrices

The associated $\mathcal{P}$-models are obtained in a similar way as for circulant matrices, in particular each column of each $\mathbf{P}_i$ is a binary vector. The corresponding coherence graphs have vertices of degrees at most 2 and thus the chromatic number $\chi[\mathcal{P}]$ is at most 3. As for the previous case, coherence $\mu[\mathcal{P}]$ is of the order $O(1)$ and $\tilde{\mu}[\mathcal{P}] = 0$.

3.2.3. Fastfood matrices

The Fastfood (Le et al., 2013) approach is a very special case of the $\mathcal{P}$-model. Note that the core term in the Fastfood transform, Eqn. 6, is the structured matrix $\mathbf{HG}$, where $\mathbf{H} = \{h_{i,j}\}$ is Hadamard and $\mathbf{G}$ is a random diagonal gaussian matrix (the rightmost terms $\mathbf{HB}$ in Eqn. 6 implement data preprocessing to make all datapoints dense, and normalization is implemented by the leftmost scaling matrix $\mathbf{S}$). The matrix $\mathbf{HG}$ can be constructed via the $\mathcal{P}$-model with the fixed budget of randomness $\mathbf{g} = (g_1, ..., g_n)$ and using the sequence of matrices $\mathcal{P} = (\mathbf{P}_1, ..., \mathbf{P}_n)$, where each $\mathbf{P}_i$ is a random diagonal matrix with entries on the diagonal of the form: $h_{i,1}, ..., h_{i,n}$. The quality of the FastFood approach can be now explained in the general $\mathcal{P}$-model method framework. One can easily see that the graphs related to the model are empty (since $\mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_2} = 0$ for $n_1 \neq n_2$). The sublinearity of $\tilde{\mu}[\mathcal{P}]$ comes from the fact that with high probability any two rows of $\mathbf{HG}$ are close to be orthogonal.

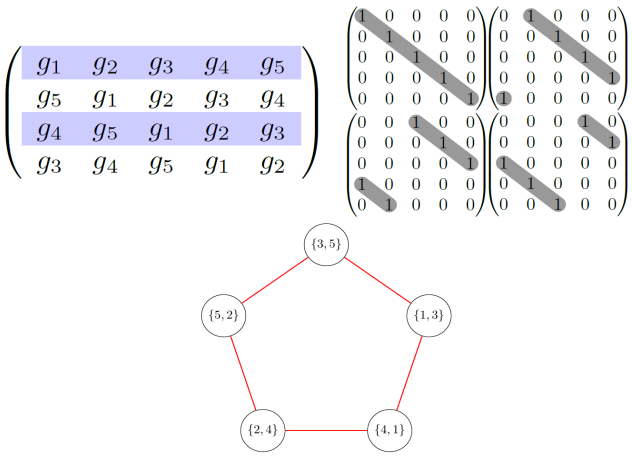


Figure 1. Top left: Circulant gaussian matrix $\mathcal{C}$. Top right: matrices $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4$ from the $\mathcal{P}$-model generating $\mathcal{C}$ from the "budget of randomness" $(g_1, ..., g_5)$. Bottom: Graph $\mathcal{G}_{i_1,i_2}$ corresponding to two highlighted rows of $\mathcal{C}$. Graphs obtained from circulant matrices are collections of cycles thus their chromatic number is at most 3.

3.2.4. Toeplitz-like semi-Gaussian matrices

Consider Toeplitz-like matrices expressible by Eqn. 5 with displacement rank r. We will assume that $\mathbf{g}^1, ..., \mathbf{g}^r \in \mathbb{R}^n$ defining the Circulant-components in Eqn. 5 are independent Gaussian vectors. They will serve as a "budget of randomness" in the related $\mathcal{P}$-model that we are about to describe, with r allowing a tunable tradeoff between structure and randomness. The vectors $\mathbf{h}^1, ..., \mathbf{h}^r$defining the skew-Circulant components in Eqn. 5 can be defined in different ways. Below we present two general schemes:

Random discretized vectors $\mathbf{h}^i$: Each dimension of each $\mathbf{h}^i$ is chosen independently at random from the binary set $\{-\frac{1}{\sqrt{nr}}, \frac{1}{\sqrt{nr}}\}$.

Sparse setting: Each $\mathbf{h}^i$ is sparse (but nonzero), i.e. has

only few nonzero entries. Furthermore, the sign of each $\mathbf{h}^i_j$ is chosen independently at random and the following holds: $\|\mathbf{h}^1\|^2 + ... + \|\mathbf{h}^r\|^2 = 1$. This setting is characterized by a parameter κ defining the size of the set of dimensions that are nonzero for at least one $\mathbf{h}^i$.

We refer to such matrices as Toeplitz-like semi-Gaussian matrices. We now sketch how they can be obtained from the $\mathcal{P}$-model. We take $t = nr$ and $\mathbf{g} = (g^1_1, ..., g^1_n, ..., g^r_1, ..., g^r_n)^T$. The matrix $\mathbf{P}_1$ is constructed by vertically stacking r matrices $\mathbf{S}_j$ for $j = 1, ..., r$, where each $\mathbf{S}_j$ is constructed as follows. The first column of $\mathbf{S}_j$ is $\mathbf{h}^j$ and the subsequent columns are obtained from previous by skew-Circulant downward shifts. Matrix $\mathbf{P}_i$ for $i > 1$ is obtained from $\mathbf{P}_{i-1}$ by upward Circulant shifts, independently for each column at each block $\mathbf{S}_j$.

Matrices constructed according to this procedure satisfy conditions regarding certain structural parameters of the $\mathcal{P}$-model (see: Theorem 4.4). In particular, in the sparse semi-Gaussian setting the corresponding coherence graphs have vertices of degrees bounded by a constant; thus, by Lemma 3.1 the $\mathcal{P}$-models associated with them have low chromatic numbers.

3.3. Construction of Random Feature Maps

Given $S[\mathcal{P}]$, the $m \times n$ structured random matrix defined by a $\mathcal{P}$-model, in lieu of using the $k \times n$ Gaussian random matrix $\mathbf{M}$ in Eqn. 1, the feature map for a data vector $\mathbf{x}$ is constructed as follows.

- Preprocessing phase: Compute $\mathbf{x}' = D_1 H D_0 \mathbf{x}$, where $H \in \mathbb{R}^{n \times n}$ is a l_2-normalized Hadamard matrix and $D_0, D_1 \in \{-1, +1\}^{n \times n}$ are independent random diagonal matrices. Note that this transformation does not change the values of Gaussian or Arc-cosine kernels, since they are spherically-invariant. This preprocessing densifies the input data vector.
- Compute $\mathbf{x}'' = S[\mathcal{P}]\mathbf{x} \in \mathbb{R}^m$.
- Compute $\bar{\mathbf{x}} \in \mathbb{R}^k$ by concatenating random instantiations of the vector $\mathbf{x}''$ above obtained from k/m independent constructions of $S[\mathcal{P}]$.
- Return $\Psi(\mathbf{x}) = \frac{1}{\sqrt{k}} s(\bar{\mathbf{x}})$

Note that the displacement rank r for low displacement rank matrices and the number of rows m of a single structured block can be used to control the "budget of randomness"; $m = 1$ reduces to a completely unstructured matrix.

4. Theoretical results

In this section we provide concentration results regarding $\mathcal{P}$-model for Gaussian and arc-cosine kernels, showing in particular that the variance of the computed structured approximation of the kernel is close to the unstructured one. We also present results targeting specifically low displacement rank structured matrices, and show how the displacement rank knob can be used to increase the budget of randomness and reduce the variance.

Let us denote by $\tilde{\mathcal{K}}_{\mathcal{P}}(\mathbf{x}, \mathbf{z})$ the approximation of the kernel for two vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$ if the $\mathcal{P}$-model is used. By $\tilde{\mathcal{K}}_{\mathbf{G}}(\mathbf{x}, \mathbf{z})$ we denote the approximation of the kernel for two vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$ if the fully unstructured setting with truly random Gaussian matrix $\mathbf{G}$ is applied. All the proofs are in the Appendix. We start with the following result.

Lemma 4.1 (Unbiasedness of the $\mathcal{P}$-model)**.** *Presented $\mathcal{P}$-model mechanism gives an unbiased estimation of the Gaussian and b^{th}-order arc-cosine kernels for $b \in \{0, 1\}$ if for every $\mathbf{P}_i$ any two different columns $\mathbf{P}_{i,j}$,$\mathbf{P}_{i,k}$ of $\mathbf{P}_i$ satisfy $\mathbf{P}^T_{i,j}\mathbf{P}_{i,k} = 0$. Thus, $\mathbb{E}[\tilde{\mathcal{K}}_{\mathcal{P}}(\boldsymbol{x}, \boldsymbol{z})] = \mathcal{K}(\mathbf{x}, \mathbf{z})$.*

The orthogonality condition $\mathbf{P}^T_{i,j}\mathbf{P}_{i,k} = 0$ is trivially satisfied by Hankel, circulant or Toeplitz structured matrices produced by the $\mathcal{P}$-model as well as Toeplitz-like semi-Gaussian matrices, where each $\mathbf{h}^i$ has one nonzero entry. It is also satisfied in expectation (which in practice suffices) for all presented Toeplitz-like semi-Gaussian matrices.

For a $\mathcal{P}$-model, where matrices $\mathbf{P}_i$ were chosen randomly we denote as $\eta[\mathcal{P}]$ the maximum possible value that a random variable $(\mathbf{P}^T_{i,n_1}\mathbf{P}_{j,n_1})^2$ can take for $1 \leq i < j \leq m, 1 \leq n_1 \leq n$. Without loss of generality we will assume that data vectors are drawn from the ball $\mathcal{B}(0, 1)$ centered at 0 of unit l_2 norm. Below we state results regarding d^{th} moments of the obtained kernel's approximation via the $\mathcal{P}$-model that lead to the concentration results.

Theorem 4.1. *Let $\mathbf{x}, \mathbf{z} \in \mathcal{B}(0, 1)$ and let $d \in \mathbb{N}$. Assume that each structured block of a matrix $\mathbf{A}$ (see: Section 3.3) produced according to the $\mathcal{P}$-model has m rows and $\tilde{\mu}[\mathcal{P}] = o(\frac{n}{\log^2(n)})$. If matrices $\mathbf{P}_i$ of the $\mathcal{P}$-model are chosen randomly then assume furthermore that for any $1 \leq i < j \leq m$ and $1 \leq n_1 < n_2 \leq n$ the n_1^{th} column of $\mathbf{P}_i$ is chosen independently from the n_2^{th} column of $\mathbf{P}_j$. If matrices $\mathbf{P}_i$ are chosen deterministically then for any $T, \epsilon > 0$ the following is true for n large enough:*

$$|\mathbb{E}[\tilde{\mathcal{K}}^d_{\mathcal{P}}(\boldsymbol{x}, \boldsymbol{z})] - \mathbb{E}[\tilde{\mathcal{K}}^d_{\mathbf{G}}(\boldsymbol{x}, \boldsymbol{z})]| \leq O(p_{gen}(T) + p_{struct}(T) + d\epsilon),$$

where:

$$p_{gen}(T) = \frac{4d}{\sqrt{2\pi T}} e^{-\frac{T}{2}} + 4ne^{-\frac{\log^2(n)}{8}}, \tag{9}$$

$$\begin{aligned} p_{struct}(T) = 4 \sum_{i=1}^{m} \chi(i, i) e^{-\frac{1}{8\mu^2[\mathcal{P}]\chi^2[\mathcal{P}]} \frac{n}{\log^6(n)}} \\ +2 \sum_{1 \leq i \leq j \leq m} \chi(i, j) e^{-\frac{\epsilon^2\sqrt{n}}{8\mu^2[\mathcal{P}]\chi^2[\mathcal{P}]T \log^4(n)}} \end{aligned} \tag{10}$$

and expectations are taken in respect to random choice for a Gaussian vector $\mathbf{g}$. *If* $\mathbf{P}_i s$ *are chosen from the probabilistic model then the above holds with probability at least* $1 - p_{wrong}$ *in respect to random choices of* $\mathbf{P}_i s$, *where*

$$p_{wrong} = 2 \sum_{i \leq i < j \leq m} e^{-\frac{n}{8 \log^6(n) \eta[\mathcal{P}]}}.$$

Let us comment on the result above. The upper bound is built from two main components: p_{gen} and p_{struct}. The first one depends on the general parameters of the setting: dimensionality of the data n and order of the computed moment d. The second one is crucial to understand how the structure of the matrix influences the quality of the model. We can immediately see that low chromatic numbers $\chi(i,j)$ (see: Section 3.1) improve quality since they decrease computed upper bound. Furthermore, low values of the coherence $\mu[\mathcal{P}]$ and chromatic number $\chi[\mathcal{P}]$ also lead to stronger concentration results. Both observations were noticed by us before, but now we see how they are implied by general theoretical results. Finally, for all considered settings, where matrices $\mathbf{P}_i$ are constructed randomly parameter $\eta[\mathcal{P}]$ is of order $O(1)$ thus p_{wrong} in negligibly small.

In particular, if both the chromatic number $\chi[\mathcal{P}]$ and the coherence $\mu[\mathcal{P}]$ are of the order $O(poly(\log(n)))$ then p_{struct} if inversely proportional to the superpolynomial function of n thus is negligible in practice. That, as we will see soon, will be the case for proposed Toeplitz-like semi-Gaussian matrices with sparse vectors h^i.

Let us also note that Theorem 4.1 can be straightforwardly applied to the structured matrix from the Fastfood model since the condition regarding $\tilde{\mu}[\mathcal{P}]$ is satisfied and so is the independence condition. Since all the chromatic numbers are equal to zero (because corresponding graphs are empty), $p_{struct} = 0$ and thus the theorem holds.

Theorem 4.1 implies also that variances of the kernel approximation for the structured $\mathcal{P}$-model case and unstructured setting are very similar (we borrow denotation from Theorem 4.1).

Theorem 4.2. *Consider the setting as in Theorem 4.1. If matrices* $\mathbf{P}_i$ *are chosen deterministically then for any* $T, \epsilon > 0$ *the following is true for* n *large enough:*

$$|Var(\tilde{\mathcal{K}}_{\mathcal{P}}(\boldsymbol{x},\boldsymbol{z})) - Var(\tilde{\mathcal{K}}_{\mathbf{G}}(\boldsymbol{x},\boldsymbol{z}))| = O(\frac{m-1}{2k}\Delta), \quad (11)$$

where Var *stands for the variance and* $\Delta = p_{gen}(T) + p_{struct} + \epsilon$. *If* $\mathbf{P}_i s$ *are chosen from the probabilistic model then the above holds with probability at least* $1 - p_{wrong}$, *where* p_{wrong} *is as in Theorem 4.1.*

Note that in practice it means that the variance in the structured and unstructured setting is similar. In particular, choosing $\epsilon = O(\frac{1}{m^2})$, $T > 7\log(m)$, one can deduce that the variance in the structured setting is of the order $O(\frac{1}{m})$ for n large enough (the well known fact is that the unstructured variance is of the order $O(\frac{1}{m})$). Note also that as expected, for $m = 1$ the structured setting becomes an unstructured one, since each structured block consists of just one row and different blocks are constructed independently.

Toeplitz-like semi-Gaussian Low-displacement rank matrices: Note that the structure of a matrix affects only the p_{struct} factor in the statements above. Thus, we will focus on the structured parameters of the $\mathcal{P}$-model. We will show that Toeplitz-like semi-Gaussian matrices can be set up so that the above parameters are of required order.

Theorem 4.3. *Consider Toeplitz-like semi-Gaussian matrices with sparse skew-Circulant factors (as in Subsection 3.2.4). Let* κ *denote the number of dimensions that are nonzero for at least one* $\mathbf{h}^i$. *Then for* $1 \leq i \leq j \leq m$ *we have:* $\chi(i,j) \leq \kappa^2 + 1$. *Furthermore,* $\mu[\mathcal{P}] \leq \kappa$ *and the bound on* $|\mathbb{E}[\tilde{\mathcal{K}}^d_{\mathcal{P}}(\boldsymbol{x},\boldsymbol{z})] - \mathbb{E}[\tilde{\mathcal{K}}^d_{\mathbf{G}}(\boldsymbol{x},\boldsymbol{z})]|$ *derived in Theorem 4.1 is valid also here if* $r \geq 3log^5(n)$ *and for* p_{wrong} *of the order* $o(\frac{1}{n})$.

The richness of the low displacement rank mechanism comes from the fact that the budget of randomness can be controlled by the rank parameter r and increasing r leads to better quality approximations. In particular, we have:

Theorem 4.4. *Consider Toeplitz-like semi-Gaussian matrices with sparse skew-Circulant factors and parameter* κ. *Assume that each* $\mathbf{h}^i$ *has exactly* α *nonzero dimensions, each nonzero dimensions taken independently at random from* $\{-\frac{1}{\alpha r}, \frac{1}{\alpha r}\}$. *Then,* $\mathbb{P}[|\mu[\mathcal{P}]| > \tau] \leq 4n^2 e^{-\frac{\tau^2 \alpha r}{O(\kappa^2)}}$.

Note that increasing rank r leads to sharper upper bounds on the coherence $\mu[\mathcal{P}]$ (in practice r polynomial in $log(n)$ suffices) and thus, from what we have said so far, to better concentration results for the entire structured scheme. Analogous variance bounds can also be derived for Toeplitz-like semi-Gaussian matrices where the $\mathbf{h}^i$ vectors are chosen to be dense. But due to lack of space, these results are included in our supplementary material.

5. Empirical Support

In this section, we compare feature maps obtained with fully Gaussian, Fastfood, Circulant, and Toeplitz-like matrices with increasing displacement rank. Our goal is to lend support to the theoretical contributions of this paper by showing that high-quality feature maps can be constructed from a broad class of structured matrices as instantiations of the proposed $\mathcal{P}$-model.

Kernel Approximation Quality: In Figure 5, we report relative Frobenius error in reconstructing the Gram matrix, i.e. $\frac{\|\mathbf{K} - \tilde{\mathbf{K}}\|_{fro}}{\|\mathbf{K}\|_{fro}}$ where $\mathbf{K}, \tilde{\mathbf{K}}$ denote the exact and ap-

Table 1. Kernel approximation (first row) and classification error (second row) in percentage for Complex Exponential (Gaussian Kernel).

	Gaussian	QMC (Halton)	Fastfood	Circulant	ToeplitzLike(1)	ToeplitzLike(5)	ToeplitzLike(10)	ToeplitzLike(20)
USPS (k=256)	5.06	5.05	6.76	7.61	9.66	7.55	6.86	**6.68**
	7.12	6.90	7.37	7.54	7.72	7.44	7.46	**7.29**
USPS (k=1280)	2.32	2.15	3.06	3.32	4.41	3.35	3.16	**3.00**
	4.52	4.73	4.62	4.53	4.62	4.58	**4.53**	4.65
DNA (k=80)	3.6	3.51	5.01	4.62	6.26	4.65	4.40	**4.10**
	31.04	30.94	31.04	30.94	31.35	30.82	**30.29**	30.70
DNA ($k = 900$)	1.61	1.59	2.23	2.06	2.88	2.09	1.93	**1.83**
	16.5	15.01	16.94	16.63	16.82	**16.34**	16.57	16.57
COIL ($k = 1024$)	2.74	2.41	**3.67**	4.45	5.60	4.47	4.09	3.79
	0.52	1.11	**0.49**	0.62	0.62	0.48	0.57	0.52
COIL ($k = 2048$)	1.92	1.87	**2.64**	3.14	4.18	3.04	2.87	2.76
	0.17	0.28	**0.15**	0.19	0.19	0.20	0.19	0.19

proximate Gram matrices, as a function of the number of random features. We use the g50c dataset which comprises of 550 examples drawn from multivariate Gaussians in 50-dimensional space with means separated such that the Bayes error is 5%. We see that Circulant matrices and Toeplitz-like matrices with very low displacement rank (1 or 2) perform as well as Fastfood feature maps. In all experiments, for Toeplitz-like matrices, we used skew-Circulant parameters (the $\mathbf{h}$ vectors in Eqn. 5) with average sparsity of 5. As the displacement rank is increased, the budget of randomness increases and the reconstruction error approaches that of Gaussian Random features, as expected based on our theoretical results. Results on publicly

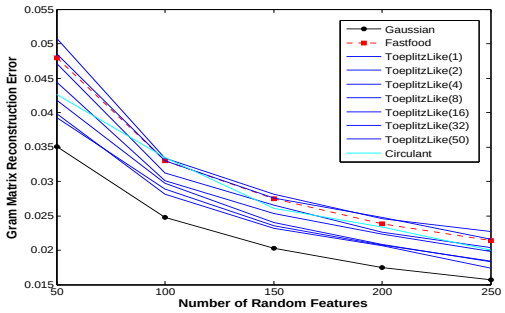
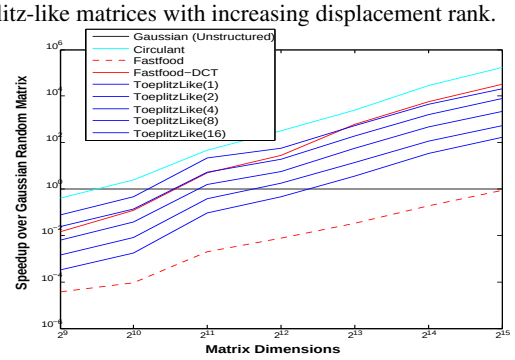


Figure 2. Lower blue curves (better reconstruction) correspond to Toeplitz-like matrices with increasing displacement rank.

available real-world classification datasets, averaged over 100 runs, are reported in Table 1 for complex exponential nonlinearity (Gaussian kernel). Results with ReLU (arccosine) are similar but not shown for lack of space. As observed in previous papers, better Gram matrix approximation is not often correlated with higher classification accuracy. Nonetheless, it is clear that the design of space of valid feature map constructions based on structured matrices is much larger than what has so far been explored in the literature: Circulant and Toeplitz-like matrices are very competitive with Fastfood, and sometimes give better results particularly with increasing displacement rank. The effectiveness of such feature maps for nonlinearities other than the complex exponential also validates our theoretical contributions. Among the unstructured baselines, we also include Quasi-Monte Carlo (QMC) feature maps of (Yang et al., 2014) using Halton low-discrepancy sequences. The use of structured matrices to accelerate QMC techniques building on (Dick et al., 2015) is of interest for future work.

Figure 3. Lower blue curves (smaller speedup) correspond to Toeplitz-like matrices with increasing displacement rank.



Speedups: Figure 3 shows the speedup obtained in featuremap construction time using structured matrices relative to using unstructured Gaussian random matrices (on a 6-core 32-GB Intel(R) Xeon(R) machine running Matlab R2014a). The benefits of sub-quadratic matrix-vector multiplication with FFT-variations tend to show up beyond 1024 dimensions. Circulant-based feature maps are the fastest to compute. Fastfood (with DCT instead of Hadamard matrices) is about as fast as Toeplitz-like matrices with displacement rank 1 or 2. Higher displacement rank matrices show speedups at higher dimensions as expected. Fastfood with inbuilt `fwht` routine in Matlab performed poorly in our experiments.

6. Conclusions

We have theoretically justified and empirically validated the use of a broad family of structured matrices for accelerating the construction of random embeddings for approximating various kernel functions. In particular, the class of Toeplitz-like semi-Gaussian matrices allows our construction to span highly compact to fully random matrices.

References

Alon, N. and Chazelle, B. The fast johnson lindenstrauss transform and approximate nearest neighbors. In *SIAM J. COMPUT.*, 2009.

Alon, N. and Chazelle, B. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. In *SIAM Review*, 2011.

Bajwa, W., Haupt, J., Raz, G., Wright, S., and Nowak, R. Toeplitz structured compressed sensing matrices. In *IEEE/SP Workshop on Statistical Signal Processing-SSP*, 2007.

Bochner, S. Monotone funktionen, Stieltjes integrale und harmonische analyse. *Math. Ann.*, 108, 1933.

Cho, Youngmin and Saul, Lawrence K. Kernel methods for deep learning. In *Neural Information Processing Systems*, 2009.

Choromanska, Anna, Choromanski, Krzysztof, Bojarski, Mariusz, Jebara, Tony, Kumar, Sanjiv, and LeCun, Yann. Binary embeddings with structured hashed projections. *ICML*, 2016.

Dick, HJ., Gia, Q.T. Le, Kuo, F. Y., and Schwab, Ch. Fast qmc matrix-vector multiplication. *SIAM J. Sci. Comput.*, 37, 2015.

Golub, G. and Loan, C. V. *Matrix Computations*. Johns Hopkins University Press, 4rth edition, 2012.

Gray, R. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory ¿*, 2, 2006.

Huang, P., Avron, H., Sainath, T., Sindhwani, V., and Ramabhadran, B. Kernel methods match deep neural networks on timit. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014.

Kailath, T., Kung, S. Y., and Morf, M. Displacement ranks of matrices and linear equations. *Journal of Mathematical Analysis and Applications*, pp. 395–407, 1979.

Le, Q., Sarlós, T., and Smola, A. Fastfood – Approximating kernel expansions in loglinear time. In *Proc. of the 30th International Conference on Machine Learning (ICML)*, 2013.

Mahoney, M. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3, 2011.

Pan, V. *Structured Matrices and Polynomials: Unified Superfast Algorithms*. Springer, 2001.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *NIPS*, 2007.

Rauhut, H., Romberg, J., and Tropp, J. Restricted isometries for partial random circulant matrices. *Appl. Comput. Harmonic Anal*, 32(2), 2012.

Schölkopf, B. and Smola, A. (eds.). *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

Sindhwani, V., Sainath, T., and Kumar, S. Structured transforms for small footprint deep learning. In *NIPS*, 2015.

Vedaldi, A. and Zisserman, A. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(34):480–492, 2012.

Xie, B., Liang, Y., and Song, L. Scale up nonlinear component analysis with doubly stochastic gradients. In *NIPS*, 2015.

Yang, J., Sindhwani, V., Avron, H., and Mahoney, M. Qmc feature maps for shift-invariant kernels. In *ICML*, 2014.

Yu, Felix X., Kumar, Sanjiv, Rowley, Henry A., and Chang, Shih-Fu. Compact nonlinear maps and circulant extensions. *CoRR*, abs/1503.03893, 2015. URL http://arxiv.org/abs/1503.03893.