Diagonal Estimation with Probing Methods

Bryan J. Kaperick

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Mathematics

Matthias C. Chung, Chair
Julianne M. Chung
Serkan Gugercin
Jayanth Jagalur-Mohan

May 3, 2019
Blacksburg, Virginia

Keywords: Probing Methods, Numerical Linear Algebra, Computational Inverse Problems

Diagonal Estimation with Probing Methods

Bryan J. Kaperick

(ABSTRACT)

Probing methods for trace estimation of large, sparse matrices has been studied for several decades. In recent years, there has been some work to extend these techniques to instead estimate the diagonal entries of these systems directly. We extend some analysis of trace estimators to their corresponding diagonal estimators, propose a new class of deterministic diagonal estimators which are well-suited to parallel architectures along with heuristic arguments for the design choices in their construction, and conclude with numerical results on diagonal estimation and ordering problems, demonstrating the strengths of our newly-developed methods alongside existing methods.

Diagonal Estimation with Probing Methods

Bryan J. Kaperick

(GENERAL AUDIENCE ABSTRACT)

In the past several decades, as computational resources increase, a recurring problem is that of estimating certain properties very large linear systems (matrices containing real or complex entries). One particularly important quantity is the trace of a matrix, defined as the sum of the entries along its diagonal. In this thesis, we explore a problem that has only recently been studied, in estimating the diagonal entries of a particular matrix explicitly. For these methods to be computationally more efficient than existing methods, and with favorable convergence properties, we require the matrix in question to have a majority of its entries be zero (the matrix is sparse), with the largest-magnitude entries clustered near and on its diagonal, and very large in size. In fact, this thesis focuses on a class of methods called probing methods, which are of particular efficiency when the matrix is not known explicitly, but rather can only be accessed through matrix vector multiplications with arbitrary vectors. Our contribution is new analysis of these diagonal probing methods which extends the heavily-studied trace estimation problem, new applications for which probing methods are a natural choice for diagonal estimation, and a new class of deterministic probing methods which have favorable properties for large parallel computing architectures which are becoming ever-more-necessary as problem sizes continue to increase beyond the scope of single processor architectures.

Dedication

To Mom, Dad, Kevin, and Molly.

Acknowledgments

I never would've gotten this done without lots of persistence and lots of patience from Tia, and the warm research environment fostered by Tia, Julianne, and the rest of the Chung Lab throughout the years.

Contents

0.1 Introduction

The diagonal entries of a matrix reveal useful information regarding the structure and spectral properties of the underlying operator. We have classical results such as Geršgorin's Theorem [44], restricting the eigenvalues of a matrix to lie on disks in the complex plane centered at its diagonal entries. More generally, in the forward problem of a matrix vector multiplication, the i^{th} diagonal entry corresponds to how much the i^{th} entry of $\mathbf{Ax}$ is impacted by the i^{th} entry of $\mathbf{x}$. The trace of a matrix is defined as the sum of its diagonal entries,

$$\operatorname{trace}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

This quantity is also equal to the sum of the eigenvalues (counting multiplicities), which unifies the diagonal entries to the spectrum. This makes a connection between the spectrum of the underlying linear operator and the representation of that operator in a particular basis. This invariant arises in numerous applications from statistical learning to lattice quantum chromodynamics [6, 8, 12, 27, 32, 34]. In particular, the estimation of the trace for matrices which are not explicitly-given — but instead through matrix-vector multiplications — has fueled much research over the past several decades [3, 11, 18, 28].

In this thesis, we explore the more fundamental problem of estimating the diagonal entries directly

$$\operatorname{diag}(\mathbf{A}) = \begin{bmatrix} a_{11} & a_{22} & \dots & a_{nn} \end{bmatrix}^{\top}.$$

We use techniques that have been developed in recent years which have been adapted from efficient methods for trace estimation. Our contributions include new analysis of these diagonal estimators, novel blocked methods adapted to resolve certain problems arising in the use of Hadamard vectors as probing vectors and which exploit large-scale parallelized computational frameworks, which are more necessary than ever before for solving modern scale

computational problems. Additionally, we discuss the more general examples of estimating the diagonals for generalized cross-validation for ill-posed linear inverse problems, of matrix functions (defined in the spectral sense) and applications where other information about the diagonal is desired such as the ranking of the diagonal entries.

We make a few comments on notation before proceeding.

- **Bold-face letters** indicate a vector (with lowercase) or matrix (with uppercase) quantity. Scalars are written in standard math print. An implicit relationship will be assumed when the same letter is used for multiple variables. For example $\mathbf{A}$ is a matrix, $\mathbf{a}_i$ is its i^{th} column, $\hat{\mathbf{a}}_i$ is its i^{th} row (but still interpreted as a column vector) and lastly a_{ij} is its ij^{th} entry.
- $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ denote the canonical basis vectors in $\mathbb{R}^n$, that is $\mathbf{e}_1 = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^\top$.
- Structured matrices which are explicitly given will have blank entries to indicate zeros outside the main entries of the matrix.

0.1.1 Characterizations of the Diagonal Entries

We begin with a brief overview of some different representations of the diagonal entries of a matrix.

1. Let $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{n \times n}$ be arbitrary. Then, we can express its i^{th} diagonal entries as a bilinear form
$$a_{ii} = \mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i. \tag{1}$$
2. Further, if $\mathbf{A}$ is diagonalizable, then $\mathbf{A}$ can be expressed as $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ where λ_i are the diagonals of $\mathbf{\Lambda}$ as well as eigenvalues of $\mathbf{A}$, ordered in monotone decreasing mag-

nitude. The columns of $\mathbf{V}$ are associated normalized eigenvectors $\mathbf{v}_i$. Then, denoting the columns of $\mathbf{V}^{-1}$ as $\mathbf{w}_i$, this diagonalization can be expressed in dyadic form as the following sum of outer products.

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{w}_i^\top.$$

Now, denoting the ij^{th} entry of $\mathbf{V}$ as v_{ij} and the ij^{th} entry of $\mathbf{V}^{-1}$ as w_{ij}, the i^{th} diagonal entry of $\mathbf{A}$ can be expressed as

$$a_{ii} = \sum_{k=1}^{n} \lambda_k v_{ik} w_{ki} = \hat{\mathbf{v}}_i^\top \mathbf{\Lambda} \hat{\mathbf{w}}_i$$

where the $\hat{\mathbf{v}}_i$ denotes the i^{th} row of $\mathbf{V}$ as a column vector, and similarly for $\hat{\mathbf{w}}$.

3. If $\mathbf{A}$ is symmetric, and hence real orthogonally diagonalizable, $\mathbf{V}^{-1} = \mathbf{V}^\top$ and this becomes

$$a_{ii} = \hat{\mathbf{v}}_i^\top \mathbf{\Lambda} \mathbf{v}_i.$$

4. If $\mathbf{A}$ is symmetric positive semi-definite with rank r, then $\mathbf{\Lambda}_r$ – the $r \times r$ block of the first r rows and columns of $\mathbf{\Lambda}$ – is symmetric positive definite, so this can be expressed as an inner product

$$a_{ii} = \left\langle \hat{\mathbf{v}}_i^{(r)}, \hat{\mathbf{v}}_i^{(r)} \right\rangle_{\mathbf{\Lambda}_r} = \|\hat{\mathbf{v}}_i^{(r)}\|_{\mathbf{\Lambda}_r}^2$$

where $\hat{\mathbf{v}}_i^{(r)}$ is the i^{th} row of $\mathbf{V}_r$, which is comprised of the r leading columns of $\mathbf{V}$.

5. For general $\mathbf{A}$, it can be expressed in it singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Then, again using the dyadic form

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

we achieve a similar result

$$a_{ii} = \sum_{k=1}^{r} \sigma_k \mathbf{u}_{ik} \mathbf{v}_{ki} = \hat{\mathbf{u}}_i^\top \boldsymbol{\Sigma} \hat{\mathbf{v}}_i = \left\langle \hat{\mathbf{u}}_i^{(r)}, \hat{\mathbf{v}}_i^{(r)} \right\rangle_{\boldsymbol{\Sigma}_r},$$

where the final equality is an inner product with respect to the reduced $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$ and $\hat{\mathbf{u}^{(r)}}_i$ is the i^{th} row of the reduced left singular matrix $\mathbf{U}_r \in \mathbb{R}^{n \times r}$, and similar for $\hat{\mathbf{v}^{(r)}}_i$ with the reduced right singular matrix $\mathbf{V}_r$.

Chapter 1

Probing Methods

Probing methods are a class of algorithms which seek to extract information from a matrix by multiplying it against a sequence of vectors. This sequence may be stochastic, in which each vector is an independent, identically-distributed (i.i.d.) realization of some predetermined probability distribution chosen for favorable properties suited to this problem. Alternatively, deterministic sequences can be used which have some favorable structure and, as we will show, some favorable convergence properties. Probing methods are most applicable when the cost of matrix-vector multiplication is tolerable, but more sophisticated approaches involving matrix decompositions are infeasible. In general, this is the cusp at which $\mathcal{O}(n^3)$ algorithms are largely intractable, while $\mathcal{O}(sn^2)$ algorithms are likely feasible for $s \ll n$. An important factor in designing a probing method is choosing how to generate the sequence of vectors $\{\mathbf{v}_k\}$. Assuming it is cheap to generate the probing vectors (the primary two choices discussed here will require $\mathcal{O}(n)$ and $\mathcal{O}(n \log n)$ work, both with a small constant factor) then probing methods require $\mathcal{O}(n^2)$ work for each matrix vector multiplication. Using $s \ll n$ probing vectors leads to the desired complexity $\mathcal{O}(sn^2)$.

A general weakness of probing methods using stochastic probing vectors is their slow convergence behavior, which is typically on the order of $\mathcal{O}\left(s^{-1/2}\right)$ with s probing vectors. Fortunately, if some structure about the matrix $\mathbf{A}$ can be assumed, the rate of convergence can be improved. In this thesis, we discuss two stochastic choices of probing vectors, and one deterministic choice.

- Let $\mathbf{v} \sim \textit{Rademacher}\,(n)$ denote a vector whose components are random variables independently distributed according to the Rademacher distribution. That is, realizations of $\mathbf{v}$ are vectors in $\mathbb{R}^n$ with i.i.d. entries taking -1 or 1 each with probability $1/2$. This distribution is also referred to as the vectorized version of the symmetric or signed Bernoulli distribution in the literature.

- Let $\mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$ denote a vector whose components are random variables independently distributed by $v_i \sim \mathcal{N}(\mu, \sigma^2)$.

- Let $\mathbf{v} \in \mathbb{R}^n$ be called a Hadamard vector if it is a column of a Hadamard matrix. A Hadamard matrix is any $\mathbf{H} \in \mathbb{R}^{n\times n}$ taking only entries ± 1 such that $\mathbf{H}\mathbf{H}^\top = n\mathbf{I}_n$. These will be discussed in detail in Chapter 2.

1.1 Hutchinson's Trace Estimator

To motivate the discussion of diagonal estimation, we begin with a probing method for the closely-related problem of trace estimation, which has historically received more attention. In 1987, D. Girard [18] first introduced the first probing method for estimating the trace, and in 1990, M.F. Hutchinson [28] provided analysis and introduced the use of Rademacher variables in Girard's probing method. Since then, the method is known as *Hutchinson's trace estimator* for estimating the trace of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and takes the form[1]

$$\operatorname{trace}(\mathbf{A}) \approx t_s^R = \frac{1}{s}\sum_{k=1}^{s} \mathbf{v}^{(k)\top}\mathbf{A}\mathbf{v}^{(k)}, \quad \mathbf{v}^{(k)} \sim \textit{Rademacher}\,(n). \tag{1.1}$$

The validity of this estimator is seen by computing the expectation of an arbitrary term

[1]The superscript will be used throughout the thesis to denote the choice of probing vectors, and may be omitted where the context has already made this clear.

$f(\mathbf{v}) := \mathbf{v}^\top \mathbf{A}\mathbf{v}$. Note that Rademacher variables clearly have mean 0 and variance 1, which will be exploited along with the pairwise independence of components of each $\mathbf{v}_k$ realization

$$\mathbb{E}\left[f(\mathbf{v})\right] = \mathbb{E}\left[\mathbf{v}^\top \mathbf{A}\mathbf{v}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\,\mathbb{E}\left[v_i v_j\right] = \sum_{i=1}^{n} a_{ii} = \operatorname{trace}\left(\mathbf{A}\right).$$

Hutchinson's trace estimator t_s^R is nothing more than a standard Monte Carlo estimator, being of the form

$$\operatorname{trace}\left(\mathbf{A}\right) = \lim_{s\to\infty} t_s^R = \lim_{s\to\infty} \frac{1}{s}\sum_{k=1}^{s} f(\mathbf{v}^{(k)}).$$

1.2 Convergence Behavior of Trace and Diagonal Estimators

1.2.1 Hutchinson's Analysis of the Trace Estimator

We now further analyze Hutchinson's trace estimator for the purpose of translating these results to the diagonal estimator. We have already seen t_s is a Monte Carlo estimator, so it is sufficient to analyze $f(\mathbf{v}) = \mathbf{v}^\top \mathbf{A}\mathbf{v}$ to understand its behavior. Indeed, the expectation of t_s reduces to the expectation of $f(\mathbf{v}_k)$ by simple application of the linearity of $\mathbb{E}\left[\cdot\right]$. Let the probing vectors be distributed according to $\mathcal{V}$.

$$\mathbb{E}\left[t_s\right] = \frac{1}{s}\sum_{k=1}^{s}\mathbb{E}\left[f(\mathbf{v}^{(k)})\right] = \frac{1}{s}\cdot s\mathbb{E}\left[f(\mathbf{v})\right] = \mathbb{E}\left[f(\mathbf{v})\right], \quad \mathbf{v}\sim\mathcal{V}.$$

Thus, we need to choose the distribution of probing vectors, $\mathcal{V}$, such that if $\mathbf{v}\sim\mathcal{V}$, then $\mathbb{E}\left[f(\mathbf{v})\right] = \operatorname{trace}\left(\mathbf{A}\right)$. We have already seen that this holds for $\mathcal{V}\sim \textit{Rademacher}\left(n\right)$, but in fact this choice is not necessary. We will now demonstrate that unbiasedness can be achieved

with only mild assumptions on $\mathcal{V}$. Let $\mathbf{v} \sim \mathcal{V}$ with each component v_i having $\mathbb{E}[v_i] = \mu$ and $\mathrm{Var}[v_i] = \sigma^2$. Then, $\mathbb{E}[v_i^2] = \sigma^2 - \mu^2$.

$$\begin{aligned}\mathbb{E}\left[\mathbf{v}^\top \mathbf{A}\mathbf{v}\right] &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\,\mathbb{E}[v_i v_j] = \sum_{i=1}^{n} a_{ii}\mathbb{E}\left[v_i^2\right] + \sum_{\substack{i,j=1\\ j\neq i}}^{n} a_{ij}\mathbb{E}[v_i]\,\mathbb{E}[v_j] \\ &= (\sigma^2-\mu^2)\sum_{i=1}^{n} a_{ii} + \mu^2 \sum_{\substack{i,j=1\\ j\neq i}}^{n} a_{ij} = (\sigma^2-\mu^2)\,\mathrm{trace}(\mathbf{A}) + \mu^2 \sum_{\substack{i,j=1\\ j\neq i}}^{n} a_{ij}.\end{aligned}$$

Hence, without any structural assumptions on $\mathbf{A}$, we require $\sigma^2 - \mu^2 = 1$ and $\mu^2 = 0$. That is,

$$\mu = \mathbb{E}[v_i] = 0, \quad \sigma^2 = \mathrm{Var}\left[v_i^2\right] = 1$$

is a necessary condition for t_s to be an unbiased estimator of $\mathbf{A}$ in general[2].

As a Monte Carlo estimator, choosing $\mathcal{V}$ such that the variance of $f(\mathbf{v}_k)$ is minimized results in optimal convergence behavior.

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{v})^2\right] &= \mathbb{E}\left[\mathbf{v}^\top \mathbf{A}\mathbf{v}\mathbf{v}^\top \mathbf{A}\mathbf{v}\right] = \mathbb{E}\left[\left(\sum_{i,j} a_{ij} v_i v_j\right)\left(\sum_{k,\ell} a_{k\ell} v_k v_\ell\right)\right] \\ &= \sum_{i,j,k,\ell} a_{ij}a_{k\ell}\mathbb{E}[v_i v_j v_k v_\ell] = \sum_i a_{ii}^2 \mathbb{E}\left[v_i^4\right] + \mathbb{E}\left[v_i^2\right]\mathbb{E}\left[v_j^2\right]\sum_{\substack{i,j\\ i\neq j}}\left(a_{ii}a_{jj} + a_{ij}^2 + a_{ij}a_{ji}\right) \\ &= \sum_i a_{ii}^2 \mathbb{E}\left[v_i^4\right] + \sum_{\substack{i,j\\ i\neq j}}\left(a_{ii}a_{jj} + a_{ij}^2 + a_{ij}a_{ji}\right). \qquad (1.2)\end{aligned}$$

Expanding $\mathbb{E}[f(\mathbf{v})]^2$,

$$\mathbb{E}[f(\mathbf{v})]^2 = \mathrm{trace}(\mathbf{A})^2 = \sum_{i,j} a_{ii}a_{jj} = \sum_i a_{ii}^2 + \sum_{\substack{i,j\\ i\neq j}} a_{ii}a_{jj}. \qquad (1.3)$$

[2]One may observe that a rescaling of t_s to t_s/σ^2 would remove the condition of unit variance. Since this normalization will not artificially restrict any of our analysis, we keep it for simplicity of discussion.

Then, $\mathrm{Var}\left[f(\mathbf{v})\right]$ is found by subtracting (1.3) from (1.2), yielding

$$\mathrm{Var}\left[f(\mathbf{v})\right] = \mathbb{E}\left[f(\mathbf{v})^2\right] - \mathbb{E}\left[f(\mathbf{v})\right]^2 = \left(\mathbb{E}\left[v_1^4\right] - 1\right)\sum_i a_{ii}^2 + \sum_{\substack{i,j \\ i\neq j}} \left(a_{ij}^2 + a_{ij}a_{ji}\right).$$

Following Hutchinson's analysis, we see that in order to minimize this variance, we need to minimize $\mathbb{E}\left[v_i^4\right] - 1$ for each component v_i. Then,

$$0 \leq \mathrm{Var}\left[v_i^2\right] = \mathbb{E}\left[(v_i^2 - 1)^2\right] = \mathbb{E}\left[v_i^4\right] - 2\mathbb{E}\left[v_i^2\right] + 1 = \mathbb{E}\left[v_i^4\right] - 1.$$

That is, $\mathbb{E}\left[v_i^4\right]$ has a lower bound of 1, which occurs precisely when $\mathrm{Var}\left[v_i^2\right] = 0$, in which case each component v_i of $\mathbf{v}$ satisfies $v_i^2 = 1$ almost surely. In summary, to achieve unbiasedness and minimal variance, the probing vector entries must satisfy the three conditions

$$\mathbb{E}\left[v_i^2\right] = 1, \quad \mathbb{E}\left[v_i\right] = 0, \quad \mathrm{Var}\left[v_i^2\right] = 0.$$

With the additional assumption that $v_i \in \mathbb{R}$, these conditions uniquely constrain $\mathcal{V}$ to equal the Rademacher distribution. Without the real-valued condition, probing vector entries uniformly distributed on the ring of magnitude one in the complex plane would also satisfy these conditions. This idea is explored by Iitaka and Ebisuzaki [29] and the method achieves a lower variance (one half that of real-valued Rademacher probing vectors). Due to the computational overhead necessary for complex arithmetic, we only analyze real-valued probing variables in this paper, as all the matrices we consider are real-valued.

To conclude, Rademacher probing vectors produce an estimator t_s^R with variance

$$\mathrm{Var}\left[t_s^R\right] = \frac{1}{s}\mathrm{Var}\left[f(\mathbf{v})\right] = \frac{1}{s}\sum_{\substack{i,j \\ i\neq j}}\left(a_{ij}^2 + a_{ij}a_{ji}\right) = \frac{1}{s}\left(\sum_{\substack{i,j \\ i\neq j}} a_{ij}a_{ji} + \|\mathbf{A}\|_F^2 - \sum_i a_{ii}^2\right).$$

When $\mathbf{A}$ is symmetric, we can write this without the $a_{ij}a_{ji}$ terms as

$$\mathrm{Var}\left[t_s^R\right] = \frac{2}{s}\left(\|A\|_{\mathrm{F}}^2 - \sum_i a_{ii}^2\right). \tag{1.4}$$

Hence, the variance is controlled by the squared magnitudes of the off-diagonal entries, as one would expect, seeing as these are the sources of error in the approximator. For a diagonal matrix, this variance vanishes completely.

1.2.2 Extensions of Hutchinson's Analysis to the Diagonal Estimator

The motivation for presenting Hutchinson's analysis in the above-detail has been to give insight into how these techniques extend to the diagonal estimation case. Clearly, trace estimation and diagonal estimation are closely related problems, so one would expect only a small change to the Hutchinson estimator is necessary to yield a diagonal estimator. Bekas et al. [9] propose such a modification Hutchinson's estimator, Equation (1.1) for this purpose in the natural way.

$$\mathbf{d}_s = \left(\sum_{k=1}^{s}\mathbf{v}^{(k)} \odot \mathbf{A}\mathbf{v}^{(k)}\right) \oslash \left(\sum_{k=1}^{s}\mathbf{v}^{(k)} \odot \mathbf{v}^{(k)}\right).$$

Here, $\odot$ denotes element-wise multiplication of vectors and $\oslash$ denotes element-wise division of vectors.

The algorithm to compute the diagonal is presented here. One important observation is the

input : *MatVec* Routine satisfying $MatVec(\mathbf{x}) = \mathbf{A}\mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^n$
input : *Generate* Routine which produces an i.i.d. random probing vector or a deterministic process
input : s Number of probes

$\mathbf{t}^{(0)} \leftarrow \mathbf{0}$;
$\mathbf{q}^{(0)} \leftarrow \mathbf{0}$;
for $\mathbf{k} \leftarrow 1$**to** s **do**
  $\mathbf{v}^{(k)} \leftarrow$ *Generate()*;
  $\mathbf{t}^{(k)} = \mathbf{t}^{(k-1)} + \mathbf{v}^{(k)} \odot \mathbf{f}(\mathbf{v}^{(k)})$;
  $\mathbf{q}^{(k)} = \mathbf{v}^{(k)} \odot \mathbf{v}^{(k)}$;
  $\mathbf{d}^{(k)} = \mathbf{t}^{(k)} \oslash \mathbf{q}^{(k)}$;
end
output: $\mathbf{d}^{(s)}$ The diagonal estimate of $\mathbf{A}$

Algorithm 1: Diagonal Probing

flexibility of this framework for implicitly-defined $\mathbf{A}$. The only information required about the matrix is a procedure for (approximately) multiplying the desired probing vectors against it. There has been some work on understanding convergence behavior of the trace estimator with different assumptions on the uncertainty in the matrix vector multiplications [11], and we expect that some of this work could be extended in the future to the problem of diagonal estimation.

The diagonal estimator differs from the trace estimator in that it no longer takes the Monte Carlo estimator form, due to an additional normalization term. Let $d_s^{\mathcal{V},i}$ be the estimator for the i^{th} diagonal entry, a_{ii}, using s vectors i.i.d. by $\mathcal{V}$. Let $\mathbf{V} = [v_{ij}] \in \mathbb{R}^{n \times s}$ be the matrix containing these realizations $\mathbf{v}^{(k)}$ as its columns. [3]

$$d_s^{\mathbf{V},i} = a_{ii} + \sum_{j,j \neq i} a_{ij} \frac{\sum_{k=1}^{s} v_{ik} v_{jk}}{\sum_{k=1}^{s} v_{ik}^2}.$$

[3] We will also use the notation $d_s^{\mathbf{V},i}$ later on, since the probing vectors may be generated by some deterministic process rather than a distribution $\mathcal{V}$. For now, we discuss only the stochastic case.

First, we analyze its expectation to understand the necessary conditions on $\mathcal{V}$ for $d_s^{\mathbf{V},i}$ to be unbiased.

$$\mathbb{E}\left[d_s^{\mathbf{V},i}\right] = a_{ii} + \sum_{j,j\neq i} a_{ij} \sum_{k=1}^{s} \mathbb{E}\left[\frac{v_{ik}v_{jk}}{\sum_m v_{im}^2}\right] = a_{ii} + s\mathbb{E}\left[v_{11}\right]\mathbb{E}\left[\frac{v_1}{v_{11}^2 + \cdots + v_{1s}^2}\right] \sum_{j,j\neq i} a_{ij} \tag{1.5}$$

$$\tag{1.6}$$

In the last line, we have used symmetry due to the i.i.d. entries to express the expectation as a single expectation of a function of s realizations of $\mathcal{V}$. Without loss of generality, we have used entries in the first row of $\mathbf{V}$, $v_{11}, \ldots, v_{1s}$ as representatives.

Then, a sufficient condition for unbiasedness, with no consideration of the structure of $\mathbf{A}$ is that $\mathbb{E}\left[\mathbf{v}\right] = 0$. A similar computation leads to variance of the estimator, assuming $\mathbb{E}\left[\mathbf{v}\right] = \mathbf{0}$,

$$\mathrm{Var}\left[d_s^{\mathbf{V},i}\right] = s\,\mathrm{Var}\left[v^2\right]\mathbb{E}\left[\frac{v_1^2}{(v_1^2 + \cdots + v_s^2)^2}\right] \sum_{j,j\neq i} a_{ij}^2.$$

Observe if $\mathcal{V}$ is the Rademacher distribution, then each v_i^2 term above is 1 and

$$\mathrm{Var}\left[d_s^{\mathbf{V},i}\right] = s \cdot 1 \cdot \frac{1}{s^2} \sum_{j,j\neq i} a_{ij}^2 = \frac{1}{s} \sum_{j,j\neq i} a_{ij}^2.$$

1.2.3 Equivalence of $d_s^{R,i}$ to a Particular Trace Estimator

An additional observation about the Rademacher case is that the normalization term $(\cdot) \oslash (\sum_{k=1}^{s} \mathbf{v}_k \odot \mathbf{v}_k)$ reduces to simply $\begin{bmatrix} 1/s & \ldots & 1/s \end{bmatrix}^\top$ so $d_s^{\mathcal{V},i}$ is again a Monte Carlo estimator as in the trace estimation case. In fact, the diagonal estimator $d_s^{R,i}$ is precisely the Hutchinson trace estimator applied to $\mathbf{A}_i$, where $\mathbf{A}_i$ is the $n \times n$ matrix with the i^{th} row of $\mathbf{A}$ in its i^{th}

row, and zero vectors for every other row

$$\mathbf{A}_i = \left[\begin{array}{c} \mathbf{0} \\ \hline \hat{\mathbf{a}}_i^\top \\ \hline \mathbf{0} \end{array}\right], \quad \operatorname{trace}(\mathbf{A}_i) = a_{ii}.$$

Thus, we immediately see that for any $\mathbf{x} \in \mathbb{R}^n$,

$$[\mathbf{x} \odot \mathbf{A}_i\mathbf{x}]_i = \mathbf{x}^\top \mathbf{A}_i \mathbf{x}.$$

$\mathbf{d}_s^R$ can be interpreted as a stacking of n Hutchinson trace estimators for $\mathbf{A}_1, \ldots, \mathbf{A}_n$. More generally, any sequence of probing vectors for which the normalization term collapses into a scalar fixed for each iteration reduce equivalently to a trace estimator with that same choice of probing vectors in this way. In particular, this property holds for Hadamard vectors discussed in Chapter 2. However, for normally-distributed vectors, this normalization is still a random variable, so we cannot make this equivalence, and the computation of moments of the estimator $d_s^{G,i}$ remains more complex than in the trace case.

1.2.4 Derivation of Variance for $d_s^{G,i}$

While the superscript R in the trace and diagonal estimators reflects the choice of Rademacher probing vectors, the superscript G reflects the choice of Gaussian probing vectors. We can resolve the variance of the diagonal estimator $d_s^{G,i}$ be evaluating the expectation term in

Equation (1.5), where $p(\cdot)$ is the probability density function for $\mathcal{N}(\mu, \sigma^2)$. We have

$$\mathbb{E}\left[\frac{v_1^2}{(v_1^2+\cdots+v_s^2)^2}\right] = \int_{\mathbb{R}^s} \frac{x_1^2}{\|\mathbf{x}\|_2^4} p(x_1)\dots p(x_s)\,\mathrm{d}x_1,\dots\mathrm{d}x_s$$
$$= \int_{\mathbb{R}^s} \frac{x_1^2}{\|\mathbf{x}\|_2^4} \prod_{i=1}^{s}\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x_i^2}{2\sigma^2}}\right)\mathrm{d}x_1,\dots\mathrm{d}x_s = (2\pi\sigma^2)^{-s/2}\int_{\mathbb{R}^s} \frac{x_1^2}{\|\mathbf{x}\|_2^4} e^{\frac{-\|\mathbf{x}\|^2}{2\sigma^2}}\,\mathrm{d}x_1,\dots\mathrm{d}x_s.$$

First, we transform to s-dimensional spherical coordinates [42] $(r, \theta_1, \dots, \theta_{n-1})$ given by

$$\begin{aligned} x_1 &= r\cos\theta_1 \\ x_2 &= r\sin\theta_1\cos\theta_2 \\ &\vdots \\ x_s &= r\sin\theta_1\dots\sin\theta_{s-2}\cos\theta_{s-1} \end{aligned}$$

with $0 \le \theta_i \le \pi$ for $i = 1, \dots, s-2$ and $0 \le \theta_{s-1} \le 2\pi$. Transformation of differentials is scaled by the Jacobian of this transformation by

$$\mathrm{d}x_1\,\mathrm{d}x_2\,\dots\,\mathrm{d}x_s = r^{s-1}\sin^{s-2}\theta_1\,\sin^{s-3}\theta_2\,\dots\,\sin\theta_{s-2}\mathrm{d}r\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2\,\dots\,\mathrm{d}\theta_{s-1}.$$

Proceeding by letting $c_s = (2\pi\sigma^2)^{-s/2}$,

$$\mathbb{E}\left[\frac{v_{11}^2}{(v_{11}^2+\cdots+v_{1s}^2)^2}\right]$$
$$= c_s\int_0^{2\pi}\cdots\int_0^{\pi}\int_0^{\infty} \frac{r^2\cos^2\theta_1}{r^4} e^{\frac{-r^2}{2\sigma^2}} r^{s-1}\sin^{s-2}\theta_1\,\dots\,\sin\theta_{s-2}\mathrm{d}r\,\mathrm{d}\theta_1\,\dots\,\mathrm{d}\theta_{s-1}$$
$$= c_s\int_0^{\infty} r^{s-3}e^{\frac{-r^2}{2\sigma^2}}\,\mathrm{d}r\int_0^{\pi}\cos^2\theta_1\sin^{s-2}\theta_1\,\mathrm{d}\theta_1\int_0^{2\pi}\mathrm{d}\theta_{s-1}\prod_{i=2}^{s-2}\int_0^{\pi}\sin^{s-i-1}\theta_i\,\mathrm{d}\theta_i.$$

For each of these component integrals, there are well-known solutions [37] of the forms

$$\int_0^\infty r^{s-3} e^{\frac{-r^2}{2\sigma^2}}\, dr = 2^{s/2-2}\sigma^{s-2}\Gamma(s/2-1)$$
$$\int_0^\pi \cos^2\theta_1 \sin^{s-2}\theta_1 = 2^{-1}\pi^{1/2}\frac{\Gamma\left(\frac{s-1}{2}\right)}{\Gamma(s/2+1)}$$
$$\prod_{i=2}^{s-2}\int_0^\pi \sin^{s-i-1}\theta_i\, d\theta_i = \prod_{i=2}^{s-2}\pi^{1/2}\frac{\Gamma(i/2)}{\Gamma\left(\frac{i+1}{2}\right)} = \frac{\pi^{\frac{s-3}{2}}}{\Gamma\left(\frac{s-1}{2}\right)} = \int_0^{2\pi} d\theta_{n-1} = 2\pi,$$

where the final steps follow from the telescoping form of this product. Combining all of these yields

$$c_s\left[2^{s/2-2}\sigma^{s-2}\Gamma(s/2-1)\right]\left[2^{-1}\pi^{1/2}\frac{\Gamma\left(\frac{s-1}{2}\right)}{\Gamma(s/2+1)}\right]\left[\frac{\pi^{\frac{s-3}{2}}}{\Gamma\left(\frac{s-1}{2}\right)}\right][2\pi] = \frac{\Gamma(s/2-1)}{4\sigma^2\Gamma(s/2+1)}.$$

Exploiting the well-known recurrence of the Gamma function $z\Gamma(z) = \Gamma(z+1)$ for $z \in \mathbb{C}$ such that $\Re z > 0$, with $z = s/2-1$ and then again with $z = s/2$, we have

$$\frac{\Gamma(s/2-1)}{4\sigma^2\Gamma(s/2+1)} = \frac{1}{4\sigma^2(s/2)(s/2-1)} = \frac{1}{\sigma^2 s(s-2)}$$

and so

$$\mathbb{E}\left[\frac{v_{11}^2}{(v_{11}^2+\cdots+v_{1s}^2)^2}\right] = \frac{1}{\sigma^2 s(s-2)}$$

Finally, we substitute this into the formula for variance above,

$$\mathrm{Var}\left[d_s^{G,i}\right] = \frac{1}{s-2}\sum_{j,j\neq i} a_{ij}^2.$$

We note that both Rademacher and Normally distributed estimators exhibit decay of variance in $\mathcal{O}(1/s)$, yet with the choice of Rademacher we see marginally lower variance for any finite s. By choosing $\{\mathbf{v}_k\}$ as Rademacher realizations, we see that the diagonal estimator

simplifies to a Monte Carlo estimator like Hutchinson's trace estimator

$$\mathbf{d}_s^R = \left(\sum_{k=1}^{s} \mathbf{v}_k \odot \mathbf{A}\mathbf{v}_k\right) \oslash \left(\sum_{k=1}^{s} \mathbf{v}_k \odot \mathbf{v}_k\right) = \frac{1}{s}\sum_{k=1}^{s} \mathbf{v}_k \odot \mathbf{A}\mathbf{v}_k.$$

As we have already proved the superiority of the diagonal estimator using Rademacher vectors compared to the estimator with Gaussian vectors with respect to variance, we can analyze the diagonal estimator treating it as a Monte Carlo estimator.

1.3 Classical Monte Carlo Convergence Theory

We derive some classical results on Monte Carlo estimators and their rates of convergence to aid with analyzing the diagonal estimator. Each of these results will be presented not in their most general form, but with only the scope of generality most useful for understanding the trace and diagonal estimator. Throughout this section, let $\{X_k\}_{k=1,\dots}$ denote a sequence of i.i.d. random variables with $\mathbb{E}[X_k] = \mu$ and $\mathrm{Var}[X_k] = \sigma^2 < \infty$. Additionally, the n^{th} sample mean is given $S_n := \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. While the strong law of large numbers is sufficient to conclude that $S_n \to \mu$ almost surely as $n \to \infty$, it gives no insight about the rate of convergence, which is of important practical interest.

The central limit theorem says that $\sqrt{n}(S_n - \mu)$ converges in distribution (but not probability or almost surely) to a Gaussian distribution of mean 0 and variance σ^2. The form $\sqrt{n}(S_n - \mu) = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) - \sqrt{n}$ suggests that for sufficiently large n, S_n *acts* like a Gaussian of mean μ and variance $\sigma^2/\sqrt{n}$. From this perspective, with variance of the convergent distribution decaying like $\mathcal{O}\left(n^{-1/2}\right)$, this can be naturally considered a rate of convergence. This is made precise with the following well-known result.

Theorem 1.1 Berry-Esséen Theorem [14]. *Let $\{X_k\}_{k=1,\dots}$ and S_n have the same as-*

sumptions as above. Additionally, assume a finite third moment, $\mathbb{E}\left[|X_k|^3\right] = \rho < \infty$. *Let* F_{Z_n} *be the (cumulative) distribution function of the scaled sum* $Z_n := \sqrt{n}(S_n - \mu)/\sigma$ *and* F_G *be the distribution function of the standard normal distribution. Then, for all* x *and* $n > 0$

$$|F_{Z_n}(x) - F_G(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}},$$

where $C < 0.4748$ *[41].*

Notice we have arrived at a uniform bound, so we scale the classical statement to better suit our discussion. Let F_{S_n} be the density function for S_n and F_{G_n} be the density function for G_n, a random variable normally-distributed with mean μ and variance σ^2/n.

$$F_{Z_n}(x) = P\left(\frac{\sqrt{n}(S_n - \mu)}{\sigma} \leq x\right) = P\left(S_n \leq \frac{\sigma}{\sqrt{n}x + \mu}\right) = F_{S_n}(\frac{\sigma}{\sqrt{n}}x + \mu)$$
$$F_G(x) = F_{G_n}(\sigma x + \mu).$$

Let $x \in \mathbb{R}$ be arbitrary, and apply the bound in Theorem 1.1 for point $\frac{\sqrt{n}}{\sigma}(x - \mu)$, and we have

$$|F_{S_n}(x) - F_{G_n}(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}.$$

More succinctly,

$$F_{S_n}(x) = F_{G_n}(x) + \mathcal{O}\left(n^{-1/2}\right) \tag{1.7}$$

and we have now arrived at precisely what is meant in saying the Central Limit Theorem implies S_n *acts* like a Gaussian with variance σ^2/n as n grows. Theorem 1.1 implies that the density functions vary by a factor which decays with $n^{-1/2}$. In the end, we will want quantitative statements about $P\left(|S_n - \mu| < \varepsilon|\mu|\right)$ for small $\varepsilon > 0$. This particular probability is useful in the context of estimation, because it is the probability that the relative error of the estimation S_n of μ is at most ε. For example, setting $\varepsilon = 10^{-k}$ provides a probability that

the estimator will be within roughly k decimal digits of accuracy [22].

A useful framework for the ensuing discussion will be the following definition of an $(\varepsilon, \delta)-$estimator.

Definition 1.2. A random variables S is called an $(\varepsilon, \delta)-$estimator of μ if it satisfies

$$P\left(|S-\mu| \leq |\mu|\varepsilon\right) \geq 1-\delta.$$

Hence, we can restate Theorem 1.1 in this notation by noting that (1.7) implies that for any $\varepsilon > 0$

$$P\left(|S_n-\mu| \leq \varepsilon\right) = P\left(|G_n-\mu| \leq \varepsilon\right) + \mathcal{O}\left(n^{-1/2}\right) \tag{1.8}$$

For a general Gaussian variable with variance σ^2 a simple rescaling yields that the probability it falls within x of its mean is given by $\text{Erf}\left(x/\sqrt{2\sigma^2}\right)$, where the error function, $\text{Erf}\left(\cdot\right)$, is defined $\text{Erf}\left(x\right) = P\left(|X| < x\right)$ where X is a normal random variable with mean 0 and variance $1/2$. Hence,

$$P\left(|G_n-\mu| \leq \varepsilon\right) = \text{Erf}\left(\varepsilon\sqrt{\frac{n}{2\sigma^2}}\right). \tag{1.9}$$

Then, combining the constant from Theorem 1.1, (1.8), and (1.9) we see that S_n as defined above, with the additional constraint of finite third moment ρ on the X_k's is an $\left(\varepsilon, 1-\text{Erf}\left(\varepsilon|\mu|\sqrt{\frac{n}{2\sigma^2}}\right)+\frac{2\rho C}{\sigma^3\sqrt{n}}\right)-$estimator of μ. With this notation, we discuss Chebychev and Chernoff equalities. We next make precise our notions of convergence and rates of convergence.

A classical result is Chebychev's inequality, a simple consequence of Markov's inequality [33].

Theorem 1.3 Chebychev's Inequality. *Let S be a random variable with expectation*

$\mu < \infty$ and variance $0 < \sigma^2 < \infty$. For any $k > 0$,

$$P\left(|S - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}.$$

Since S_n has variance σ^2/n, by choosing $k = \frac{|\mu|\varepsilon\sqrt{n}}{\sigma}$, Theorem 1.3 yields

$$P\left(|S_n - \mu| \leq \varepsilon|\mu|\right) \geq 1 - \frac{\sigma^2}{\mu^2\varepsilon^2 n} = 1 - \mathcal{O}\left(n^{-1}\right). \tag{1.10}$$

Adding some additional regularity by assuming boundedness of each of the X_k's, we have a Chernoff-style bound for S_n which is in general much sharper than Chebychev's inequality.

Theorem 1.4 Hoeffding's Inequality [26]. *Given the assumptions for $\{X_k\}$ and S_n defined above, and also assuming that each X_k is supported on a bounded interval $[a, b] \subset \mathbb{R}$*

$$P\left(|S_n - \mu| \leq \varepsilon|\mu|\right) \geq 1 - 2\exp\left(\frac{-2n\varepsilon^2\mu^2}{(b-a)^2}\right).$$

1.4 Rewriting Classical Bounds as (ε, δ) Statements

We have discussed different bounds for sample means S_n of i.i.d. random variables. For a fixed $\varepsilon > 0$, we arrive at statements of the form "S_n is an $(\varepsilon, \delta_{n,\varepsilon})-$estimator of μ". Here, we have written $\delta_{n,\varepsilon}$ to enforce the point that δ depends on both n and ε. This is only a useful statement if $\delta_{n,\varepsilon} \to 0$ by taking either $n \to \infty$ or $\varepsilon \to 0^+$. To use these results practically, it is instead useful to think of n as a function of ε and δ. Recalling our above interpretation of Definition 1.2, we want the following:

Problem 1 (Minimum Samples For (ε, δ)-Estimate). Given a sample mean S_n as

defined above, a fixed $\varepsilon > 0$, and $\delta > 0$, determine the minimum n such that S_n is an $(\varepsilon, \delta)-$estimator for μ.

Theorem 1.1 does not yield any closed-form statement for Problem 1, but the two ensuing inequalities conclude

Result	n satisfying Problem 1	Conditions on $\{X_k\}$
Chebychev's Inequality	$\sigma^2\mu^{-2}\varepsilon^{-2}\delta$	non-zero, finite variance
Hoeffding's Inequality	$(b-a)^2 2^{-1}\mu^{-2}\varepsilon^{-2}\log(2\delta^{-1})$	$X_k \in [a,b]$ for each k

1.5 Application Stochastic Analysis to Trace and Diagonal Estimators

We can place this in the context of trace estimation naturally. With Rademacher probing vectors, we can establish a bound on each term

$$\mathbf{v}^\top \mathbf{A}\mathbf{v} = \sum_{i,j} a_{ij} v_i v_j = \sum_i a_{ii} + \sum_{i<j} (a_{ij} + a_{ji}) v_i v_j.$$

Since $v_i v_j$ is again a Rademacher variable, we see the total range of values is controlled by the sum of terms of the form $\pm(a_{ij} + a_{ji})$ and we have that $\mathbf{v}^\top \mathbf{A}\mathbf{v}$ is contained on an interval of length $2\sum_{i<j} |a_{ij} + a_{ji}|$.

These results are consistent with the variance of t_s^R since this interval is length zero if and only if each of the off-diagonal entries satisfy $a_{ij} = -a_{ji}$ which is precisely the case for which the variance vanishes.

Thus, for Rademacher trace probing, we can translate the results of the previous table immediately,

Result	s for Problem 1 with t_s^R	s for Problem 1 with $d_s^{R,i}$
Chebychev's Inequality	$\frac{1}{\varepsilon^2 \operatorname{trace}(\mathbf{A})^2} \frac{\sum_{\substack{i,j \\ i \neq j}} a_{ij}^2 + a_{ij}a_{ji}}{\delta}$	$\frac{1}{\varepsilon^2 a_{ii}^2} \frac{\sum_{j \neq i} a_{ij}^2}{\delta}$
Hoeffding's Inequality	$\frac{\log(2/\delta)\left(\sum_{\substack{i,j \\ i<j}} \lvert a_{ij}+a_{ji} \rvert\right)^2}{2\varepsilon^2 \operatorname{trace}(\mathbf{A})^2}$	$\frac{\log(2/\delta)\sum_{j \neq i} a_{ij}^2}{2\varepsilon^2 a_{ii}^2}$

There has been work to sharpen these bounds for trace estimation in the case that the underlying matrix $\mathbf{A}$ is symmetric positive semi-definite [4]. We recall that symmetry gives the variance of t_s^R as in Equation (1.4) which avoids the somewhat opaque $a_{ij}a_{ji}$ terms.

Theorem 1.5 Theorem (Hutchinson Trace Estimator Convergence Rate [4]). *The Hutchinson estimator t_s^R is an (ε, δ)-estimator of* $\operatorname{trace}(\mathbf{A})$ *for* $s \geq 6\varepsilon^{-2}\log(2\operatorname{rank}(\mathbf{A})/\delta)$.

The proof of Theorem 1.5 relies on the orthogonal diagonalization of $\mathbf{A}$, and relies on Lemma 5 from [1], which gives an exponentially decaying bound on how far the average inner product between a fixed orthogonal vector and i.i.d. Rademacher vectors will be from magnitude 1. Given that $\mathbf{A}_i$ is no longer symmetric, $\mathbf{A}_i$ is diagonalizable (all rank-1 matrices are diagonalizable), but not orthogonally so. Avron and Toledo provide another (ε, δ)-estimate result which is weaker, yet applicable to a wider class of probing bases.

Theorem 1.6 Theorem (Weaker Hutchinson Trace Estimator Convergence Rate [4]). *The Hutchinson estimator t_s^R is an (ε, δ)-estimator of* $\operatorname{trace}(\mathbf{A})$ *for* $s \geq \frac{1}{2}\varepsilon^{-2}n^{-2}\operatorname{rank}(\mathbf{A})^2\log(2/\delta)\kappa_f($ *where $\kappa_f(\mathbf{A})$ is the ratio between the largest and smallest nonzero eigenvalue of $\mathbf{A}$.*

This argument can be modified for the Rademacher diagonal estimator by noting the equivalencebetween the Rademacher diagonal estimator of $\mathbf{A}$, and the Hutchinson trace estimator of $\mathbf{A}_i$ as described in Section 1.2.3.

This yields our result.

Theorem 1.7 Theorem (Rademacher Diagonal Estimator Convergence Rate). *Let* $\mathbf{A}$ *be semi symmetric positive definite, with non-zero* i^{th} *diagonal entry* a_{ii}*. The diagonal estimator* $d_s^{R,i}$ *for* a_{ii} *is an* (ε, δ)*-estimator for*

$$s \geq \frac{2n^2}{\varepsilon^2} \operatorname{rank}(\mathbf{A}) \log(2/\delta) \left(\frac{\operatorname{trace}(\mathbf{A})^2}{a_{ii}} \right).$$

Proof. Let $\mathbf{E}_i$ be the matrix of all zeros except a 1 in its i^{th} diagonal entry. Note $\mathbf{A}_i = \mathbf{E}_i\mathbf{A}$. Then, since $\mathbf{A}$ is symmetric semi-positive definite, it has an orthogonal diagonalization $\mathbf{A} =$

$\mathbf{QDQ}^\top$. Then,

$$\begin{aligned}
\left|\mathbf{v}^\top \mathbf{A}_i \mathbf{v}\right| &= \left|\mathbf{v}^\top \mathbf{E}_i \mathbf{Q}\mathbf{D}\mathbf{Q}^\top \mathbf{v}\right| \\
&= \left|\mathbf{v}^\top \mathbf{E}_i \sum_{j=1}^{\operatorname{rank}(\mathbf{A})} \lambda_j \mathbf{q}_j \mathbf{q}_j^\top \mathbf{v}\right| \\
&= \left|\mathbf{v}^\top \sum_{j=1}^{\operatorname{rank}(\mathbf{A})} (\lambda_j \mathbf{q}_j^\top \mathbf{v}) \mathbf{E}_i \mathbf{q}_j\right| \\
&= \left|\mathbf{v}^\top \begin{bmatrix} \mathbf{0} \\ \sum_{j=1}^{\operatorname{rank}(\mathbf{A})} (\lambda_j \mathbf{q}_j^\top \mathbf{v}) q_{ij} \\ \mathbf{0} \end{bmatrix}\right| \\
&= \left|v_i \sum_{j=1}^{\operatorname{rank}(\mathbf{A})} (\lambda_j \mathbf{q}_j^\top \mathbf{v}) q_{ij}\right| \\
&= \left|\sum_{j=1}^{\operatorname{rank}(\mathbf{A})} (\lambda_j \mathbf{q}_j^\top \mathbf{v}) q_{ij}\right| \\
&\leq \left|\sum_{j=1}^{\operatorname{rank}(\mathbf{A})} \lambda_j \|\mathbf{q}_j\|_2^2 \|\mathbf{v}\|^2 q_{ij}\right| \\
&= n \left|\sum_{j=1}^{\operatorname{rank}(\mathbf{A})} \lambda_j q_{ij}\right| \\
&\leq n \|\mathbf{D}\hat{\mathbf{q}}_i\|_1 \leq n\sqrt{\operatorname{rank}(\mathbf{A})}\operatorname{trace}(\mathbf{A}).
\end{aligned}$$

Then, $\mathbf{v}^\top \mathbf{A}_i \mathbf{v}$ is supported on the interval centered at zero of length $2n\sqrt{\operatorname{rank}(\mathbf{A})}\operatorname{trace}(\mathbf{A})$. Applying Hoeffding's inequality (1.4), we have for any $\varepsilon > 0$,

$$P\left(|d_s^{R,i} - a_{ii}| \geq \varepsilon |a_{ii}|\right) \leq 2\exp\left(\frac{-s\varepsilon a_{ii}}{2n^2 \operatorname{rank}(\mathbf{A}) \operatorname{trace}(\mathbf{A})^2}\right).$$

Bounding the right-hand side by a fixed $\delta > 0$ and rearranging yields the lower bound for

s. □

We make some comments about this result. Firstly, it is an incredibly weak statement, given the scaling with rank $(\mathbf{A})\, n^2$, so that if $\mathbf{A}$ is SPD, this bound scales with n^3, which quickly becomes well beyond a reasonable value for s. Additionally, the bound improves if $|a_{ii}|$ is large relative to the other diagonal entries. So we expect the largest diagonals to converge most quickly.

It is expected that this result could be drastically improved to be closer to a bound resembling Theorem 1.5.

1.6 A Unified Framework for Analyzing $\mathbf{d}_s$

Thus far, only stochastic choices of probing vectors have been discussed. We have discussed issues of variance and bias in an estimator comprised of i.i.d. samples from a random distribution. We will now discuss the estimator within a unified perspective. For example, with the choice of Hadamard vectors, the probing vectors are successive columns of a fixed Hadamard matrix, and so they are not independent of each other. Since all stochastic choices have an assumption of inependence between probing vectors, this demonstrates the need for a new framework. It is of interest to analyze what properties this sequence of probing vectors must satisfy in order to guarantee accurate results.

Recall $\mathbf{d}_s$ is the generic diagonal estimator, and d_s^i is the i^{th} entry. First, we rearrange the

form of the unnormalized estimator, $\mathbf{d}'_s$

$$\mathbf{d}'_s := \sum_{k=1}^{s} \mathbf{v}_k \odot \mathbf{A}\mathbf{v}_k = \sum_{k=1}^{s} \begin{bmatrix} v_{1k} & & \\ & \ddots & \\ & & v_{nk} \end{bmatrix} \mathbf{A}\mathbf{v}_k.$$

$$d_s^{i'} = \sum_{k=1}^{s} v_{ik} \sum_{j=1}^{n} a_{ij} v_{jk} = \sum_{k=1}^{s} \sum_{j=1}^{n} a_{ij} v_{ik} v_{jk} = \sum_{j=1}^{n} a_{ij} \sum_{k=1}^{s} v_{ik} v_{jk}$$

Now, including the normalization term,

$$\begin{aligned} d_s^i = \frac{\sum_{j=1}^{n} a_{ij} \sum_{k=1}^{s} v_{ik} v_{jk}}{\sum_{k=1}^{s} v_{ik}^2} &= a_{ii} \left(\frac{\sum_{k=1}^{s} v_{ik}^2}{\sum_{k=1}^{s} v_{ik}^2} \right) + \sum_{j=1, j\neq i}^{n} a_{ij} \left(\frac{\sum_{k=1}^{s} v_{ik} v_{jk}}{\sum_{k=1}^{s} v_{ik}^2} \right) \\ &= a_{ii} + \sum_{j\neq i}^{n} a_{ij} \left(\frac{\sum_{k=1}^{s} v_{ik} v_{jk}}{\sum_{k=1}^{s} v_{ik}^2} \right) \end{aligned}$$

To enforce the desired result, $d_s^i = a_{ii}$, the following constraint is imposed.

$$\sum_{k=1}^{s} v_{ik} v_{jk} = 0, \quad \text{when } i \neq j.$$

By building the sequence $\{\mathbf{v}_k\}$ into a matrix $\mathbf{V} \in \mathbb{R}^{n\times s}$ whose columns are $\mathbf{v}_1, \ldots, \mathbf{v}_s$, this constraint can be viewed as an orthogonality condition on the rows of $\mathbf{V}$.

Proposition 1.8 (Exactness condition, [9]). *For $\mathbf{V} = [\mathbf{v}_1|\mathbf{v}_2|\ldots|\mathbf{v}_s]$, if the i^{th} row of $\mathbf{V}$ is orthogonal to all those rows j of $\mathbf{V}$ for which $a_{ij} \neq 0$, then the diagonal estimator will yield an exact result for a_{ii}, the i^{th} diagonal entry of $\mathbf{A}$.*

If $\mathbf{V}\mathbf{V}^\top$ is diagonal, then this proposition implies the diagonal estimator is equal to the

diagonal of $\mathbf{A}$ if $s = n$. However, the utility of probing methods is realized when $s \ll n$. That is, $\mathbf{V}$ will have more rows than columns, and since it is impossible to find more than s mutually orthogonal vectors in an s dimensional vector space, $\mathbf{V}\mathbf{V}^\top$ cannot be diagonal.

To emphasize the importance that $s \ll n$, we observe that the standard basis vectors as probing vectors as in Equation (1) would set $\mathbf{V} = \mathbf{I}_n$ which produces an exact estimator. Obviously, using fewer than n standard basis vectors will result in an estimator that is exact for a subset of diagonal entries, but will produce no information about the $n - s$ remaining entries. This approach has been used for trace estimation [4], however for diagonal estimation this approach will not be sufficient, as the applications of interest in this paper prioritize the learning of incomplete information about the entire diagonal as opposed to complete information about any subset. Moreover, Proposition 1.8 suggests a deterministic choice of probing vectors may be a viable alternative to stochastic choices. There is no guarantee that an exact estimate will ever be reached within n steps for the stochastic estimators previously discussed. Meanwhile, it is certainly possible with a deterministic selection of $\mathbf{V}$ to produce an exact estimator with no more than n probing vectors. The natural question is whether there is a better deterministic choice of probing vectors than the standard basis vectors. An important question is of how to quantify the performance of a particular choice of probing vectors.

1.7 Off-Diagonal Interaction and The Welch Bound

It has been established that the estimator cannot be exact with fewer than n probing vectors if no structural information about the system. We present some well-known general bounds on how accurate the estimator can be expected to perform for a given (n, s) pair.

These bounds are framed in the minimization of two error functions motivated by Propo-

sition 1.8. Specifically, these two functions indicate how far $\mathbf{V}\mathbf{V}^\top$ differs from a diagonal matrix. Let $\widehat{\mathbf{v}}_i^\top$ is the i^{th} row vector of $\mathbf{V}$. First, the off-diagonal interactions can be quantified in a least square sense,

$$E_{rms} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} |\widehat{\mathbf{v}}_i^\top \widehat{\mathbf{v}}_j|^2}$$

as well as in a maximal magnitude sense.

$$E_{max} = \max_{1 \leq i < j < n} |\widehat{\mathbf{v}}_i^\top \widehat{\mathbf{v}}_j|.$$

E_{rms} quantifies error as the average magnitude of off-diagonal entries of $\mathbf{V}\mathbf{V}^\top$, while E_{max} is the largest magnitude off-diagonal entry. Clearly, $E_{max} \geq E_{rms}$. Bounds for these quantities are given in the theory of binary codewords. Minimizing E_{rms} is equivalent to minimizing the maximum cross-correlation amplitude between code words (rows of $\mathbf{V}$) [40]. The so-called *Welch Bounds* for these values are determined to be

$$E_{max} \geq E_{rms} \geq \sqrt{\frac{n-s}{(n-1)s}}.$$

Hence, the sequence of probing vectors which are optimal in an E_{rms} sense will satisfy the Welch bound with equality. We discuss one such class of matrices in the following section.

Chapter 2

Hadamard Matrices

Definition 2.1. A matrix $\mathbf{H} \in \{\pm 1\}^{n \times n}$ is a *Hadamard matrix* if $\mathbf{H}\mathbf{H}^\top = n\mathbf{I}$. We denote the set of Hadamard matrices of order n as $\mathrm{Had}(n)$.

By definition, $\mathbf{H}\mathbf{H}^\top$ is a diagonal matrix, so n columns of a Hadamard matrix as probing vectors form an exact diagonal estimator by Proposition 1.8. Even better, Hadamard matrices are a class of matrices with entries ± 1 which satisfy the Welch bounds. This suggests that the average magnitude of off-diagonal interactions in $\mathbf{H}\mathbf{H}^\top$ is minimized for any $1 \leq s < n$, and that error is 0 when $s = n$.

It is widely conjectured that there exists a Hadamard matrix of order n if and only if $n = 1$, $n = 2$ or $n = 4 \mod 4$. It is a straightforward combinatorial proof to show that this is a necessary condition for existence. As to whether it is also sufficient remains an open question [38]. Unfortunately, this is not a problem which can be ignored for practical applications. In fact, as of 2005 [31], the smallest n divisible by four which has no known Hadamard matrix is 668, which is well within the scale of problems of interest for probing methods. Moreover, the problem of existence is paired with the problem of construction. Even in cases for which existence can be proved, we will need to even further refine the set of Hadamard matrices to those which can be constructed efficiently.

As is clear from Algorithm 1, the cost of a specific probing method is controlled by the number of probing vectors, the cost of evaluating a matrix-vector multiplication with $\mathbf{A}$,

and the generation of each probing vector $\mathbf{v}_k$. Stochastic methods generally have a cheap cost associated with constructing probing vectors, and because of independence of entries, can be done entirely independently of each other. Rademacher variables in particular require only n random bits to generate a length n vector. Gaussian random vectors in floating point arithmetic also require $\mathcal{O}(n)$ random bits and time, although with a larger constant. With this in mind, we seek to understand the complexity of constructing Hadamard columns, and design choices that should be considered in constructing them. We observe immediately from the definition that if $\mathbf{H}$ is a Hadamard matrix, then so too is any transformation of $\mathbf{H}$ through row and column reordering. In the context of probing, we need only a small subset of columns of $\mathbf{H}$, so naturally we want to understand whether there is different convergence behavior of $\mathbf{d}_s^H$ by choosing different columns of $\mathbf{H}$, or different matrices from $\mathrm{Had}(n)$ altogether.

In order to adaquetely adress these the two concerns of construction and existence, we divide the discussion into two sections. For construction, we apply known theory of Hadamard matrices to the context of probing in order to motivate the design decisions in the construction of Hadamard columns with arguments grounded in convergence behavior and computational efficiency. As for the question of existence, we address the limitations of current methods, and propose a new block matrix construction which can serve to ameliorate some of these concerns, particularly in the case of diagonal estimation moreso than trace estimation.

2.1 The Problem of Construction

We proceed by making the following observation which suggests the construction can be quite efficient. If $\mathbf{H_1}$ and $\mathbf{H_2}$ are Hadamard matrices of sizes n and m, then $\mathbf{H_1} \otimes \mathbf{H_2}$, the Kronecker product, is a Hadamard matrix of size nm.

For example, let $\mathbf{H}$ be an arbitrary Hadamard matrix. Then

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \mathbf{H} = \left[\begin{array}{c|c} \mathbf{H} & \mathbf{H} \\ \hline \mathbf{H} & -\mathbf{H} \end{array}\right]$$

is also a Hadamard matrix. This fact will allow feasibly the construction of a single column of large Hadamard matrice. [1] Using this Kronecker product trick, we define an important class of Hadamard matrices called the Walsh (Hadamard) matrices. By applying this rule multiple times, we have a definition for the Walsh matrices.

Definition 2.2 Walsh Matrices. The Walsh matrices consist of the Hadamard matrices of order N^k which can be constructed by the Kronecker product of k order N Hadamard matrices for some $N, k \in \mathbb{N}$. That is,

$$\text{Walsh}(N,\, k) = \{\mathbf{H}^{(1)} \otimes \ldots \otimes \mathbf{H}^{(k)} \,|\, \mathbf{H}^{(i)} \in \text{Had}\,(N)\,,\, i = 1, \ldots, k\}.$$

Definition 2.3 Bandwidth of a matrix. Let $\mathbf{A} \in \mathbb{R}^{n \times n}$. Let a_{ij} denote the ij^{th} entry of $\mathbf{A}$ for $1 \leq i, j \leq n$. Let $b \in \mathbb{N}$ value such that $|i - j| > b$ implies $a_{ij} = 0$. Then we say $\mathbf{A}$ has bandwidth b.

We can say the following about Walsh(N, k).

Theorem 2.4. *Let $N, k \in \mathbb{N}$ and $\mathbf{H} \in \text{Walsh}(N, k)$. Suppose $\mathbf{A}$ of size $N_1 N_2 \ldots N_k \times N_1 N_2 \ldots N_k$ has bandwidth strictly less than N^ℓ for some $\ell \in \{0, \ldots, k-1\}$. If the diagonal of $\mathbf{A}$ is estimated with the first jN^ℓ columns of $\mathbf{H}$ for some $j \in \{1, \ldots, N^{k-\ell}\}$, the estimate will be exact.*

[1] For a very rough estimate of the efficiency of constructing one Hadamard column compared to the full matrix, for $n = 2^25$ in MatLab 2016a, the built-in hadamard command takes over a minute to complete, while accessing all of the columns independently exploiting an efficient Kronecker product construction takes approximately 0.5 seconds.

We give the proof of this statement in the appendix.

In practice, the bandwidth of $\mathbf{A}$ is not known a priori. If the sparsity pattern of $\mathbf{A}$ is already given, then the choice of optimal probing vectors can be made by solving an optimization problem [43] which is outside of the scope of the assumptions made in this thesis. However, if instead we make a more reasonable assumption that the entries of $\mathbf{A}$ decay away from the diagonal, then Theorem 2.4 can be instead interpreted as a claim that the diagonal estimate incurs no error from the nonzero entries within the primary $N^\ell - 1$ off-diagonal bands.

We can generalize this result further by introducing a more general set of Hadamard matrices.

Definition 2.5 Generalized Walsh Matrices. The Generalized Walsh matrices consist of the Hadamard matrices of order $N_1 N_2 \dots N_k$ which can be constructed by the Kronecker product of k Hadamard matrices of order $N_1, \dots, N_k$, respecively. That is,

$$\mathcal{G}(N_1, \dots, N_k) = \{\mathbf{H}^{(1)} \otimes \cdots \otimes \mathbf{H}^{(k)} \,|\, \mathbf{H}^{(i)} \in \mathrm{Had}\,(N_i)\,,\ i = 1, \dots, k\}.$$

Observe $\mathcal{G}(N_1, \dots, N_k) \subseteq \mathrm{Had}\,(N_1 N_2 \cdots N_k)$ and if $N_1 = \cdots = N_k$ then $\mathcal{G}(N_1, \dots, N_k) = \mathrm{Walsh}(N,\, k)$.

Then, the more general form of Theorem 2.4 is the following.

Theorem 2.6. *Let $k, N_1, \dots, N_k \in \mathbb{N}$ and $\mathbf{H} \in \mathcal{G}(N_1, \dots, N_k)$. Suppose $\mathbf{A}$ of size $N_1 N_2 \dots N_k \times N_1 N_2 \dots N_k$ has bandwidth strictly less than $N_\ell N_{\ell+1} \cdots N_k$ for some $\ell \in \{1, \dots, k\}$. If the diagonal of $\mathbf{A}$ is estimated with the first $j \cdot N_\ell N_{\ell+1} \cdots N_k$ columns of $\mathbf{H}$ for some $j \in \{1, \dots, N_1 N_2 \cdots N_{\ell-1}\}$, the estimate will be exact.*

2.1.1 Computational Discusion

In diagonal probing applications using block Hadamard matrices, we are given a fixed N and need to construct a small subset of the columns of some element of $\mathrm{Had}(N)$ efficiently. The Generalized Walsh matrices are exactly the set of Hadamard matrices which can be efficiently generated column-wise for large N. However, as N grows even moderately large, the set of Generalized Walsh matrices grows very large. This raises the question of whether the choice of Hadamard matrix is important. Theorem 2.6 suggests that the choice of matrix can significantly affect the amount of probing vectors needed for certain convergence guarantees.

For example, suppose we have $\mathbf{A} \in \mathbb{R}^{288\times 288}$ with bandwidth 7. Consider $\mathbf{H} \in \mathcal{G}(36,2,2,2)$ and $K \in \mathcal{G}(2,2,2,36)$. By Proposition 2.6, the diagonal of $\mathbf{A}$ is recovered exactly by probing with 8 columns of $\mathbf{H}$. However, we have no guarantees about probing with any fewer than 36 columns of $\mathbf{K}$.

In general, Proposition 2.6 tells us that if we are given a fixed N and seek an $\mathbf{H} \in \mathcal{G}(N_1,\ldots,N_k)$ for some $N_1,\ldots,N_k$ such that $N_1N_2\cdots N_k = N$, then we should order the $N_1,\ldots,N_k$ such that $N_1 \geq N_2 \geq \cdots \geq N_k$ in order to have the most flexibility in choosing the number of probing vectors to take from $\mathbf{H}$ while maintaining theoretical convergence guarantees.

2.1.2 The Fast Walsh-Hadamard Transform

We use a variation on the Fast Walsh-Hadamard Transform (FWHT) [5] in order to efficiently compute columns of Hadamard matrices of a desired order. For a fixed $n = 2^m$ for some $m \in \mathbb{N}$, this algorithm allows for the construction of an element of $\mathrm{Had}(n)$ with computational complexity $\mathcal{O}(n \log n)$ (as opposed to the $\mathcal{O}(n^2)$ complexity of a naive algorithm). We have modified this approach to compute, for fixed $n = 2^m x$ for some $m \in \mathbb{N}$ and $x \in$

$2\mathbb{N}-1$, a single specified column of an element of $\mathrm{Had}(n)$ with complexity $\mathcal{O}((\log n))$.
input : N the order of the system, j the index of probing vector requested

Find m,k such that $N = 2^m k$;

if $m < 2$ **then**
| **output:** Failure – N is not a Hadamard dimension

end

if $m = 2$ **then**

| **if** $\mathbf{H} \in \mathrm{Had}(N)$ *in memory* **then**
| | **output:** Retrieve j^{th} column of $\mathbf{H}$
| **end**

end

else
| **output:** Failure – No stored Hadamard matrix of order N

end

else

| **for** $\ell = m, m-1, \ldots, 2$ **do**
| | **if** $\mathbf{H} \in \mathrm{Had}\left(2^\ell k\right)$ *in memory* **then**
| | | Retrieve $\mathbf{h}$, the $\lceil j/2^\ell \rceil^{th}$ column of $\mathbf{H}$;
| | | **output:** $\mathbf{h} \otimes \mathrm{FWHT}(2^{m-\ell}, \mathbf{e}_j)$
| | **end**
| **end**

end

output: Failure – Insufficient stored Hadamard matrices

2.2 The Problem of Existence

One of the most limiting aspects of using Hadamard matrices for probing is the restriction on the problem sizes for which there corresponds a Hadamard matrix. At the onset of Chapter 2, we stated the conjecture that Hadamard matrices exist at each size which is a multiple of 4. Restricting ourselves further to *easily computable* Hadamard matrices, i.e. Generalized Walsh matrices (Definition 2.5), reveals an even more pessimistic situation. Consider an $\mathbf{H} \in \mathcal{G}(N_1, \ldots, N_k)$. Assume that each $N_k > 2$. It is comprised of kronecker products between k Hadamard matrices, each of which have order divisible by 4 (recall this direction of implication is proven true– only the converse remains to be proven). Thus, $\mathbf{H}$ is divisible by 4^k. This suggests the Generalized Walsh matrices are much more sparsely available than at every multiple of 4. Given an arbitrary dimension for $\mathbf{A}$, our options from Generalized Walsh are indeed quite limited. Our only hope is to precompute higher orders of seed matrices (which can eventually met by limitations of existence discussed earlier, such as $n = 668$) or by constructing these

One possible remedy is to probe with the smallest valid Hadamard size at least as large as n, truncating rows to equal n. However, this new matrix is certainly not guaranteed to be a Hadamard matrix, and hence loses its properties of exactness when $s = n$, and no longer is guaranteed to match the Welch error bound.

2.2.1 Block Hadamard Matrices

We propose a modification to Hadamard probing methods which can resolve some fundamental constraints of these methods resulting from the problem of existence.

We define a new class of matrices which is intrinsically related to Hadamard matrices.

Definition 2.7. A matrix $\mathbf{H} \in \mathbb{R}^{n\times n}$ is a *block Hadamard Matrix* if it is of the form

$$\mathbf{H} = \begin{bmatrix} \mathbf{H_1} & & \\ & \ddots & \\ & & \mathbf{H_p} \end{bmatrix}$$

where $\mathbf{H}_1, \ldots, \mathbf{H}_p$ are Hadamard matrices and all off-diagonal blocks are $\mathbf{0}$.

We consider the block Hadamard matrix

$$\mathbf{H} = \begin{bmatrix} \mathbf{H_1} & & \\ & \ddots & \\ & & \mathbf{H_p} \end{bmatrix}.$$

In which each block $\mathbf{H}_i$ is of size $s_i \times s_i$. It is observed that $\mathbf{H}$ satisfies

$$\mathbf{H}\mathbf{H}^\top = \begin{bmatrix} \mathbf{H_1}\mathbf{H_1}^\top & & \\ & \ddots & \\ & & \mathbf{H_p}\mathbf{H_p}^\top \end{bmatrix} = \begin{bmatrix} s_1\mathbf{I}_{s_1} & & \\ & \ddots & \\ & & s_p\mathbf{I}_{s_p} \end{bmatrix}$$

which is diagonal. We are motivated to use these columns as probing vectors in diagonal estimation. There are two reasons for using this Block Hadamard scheme.

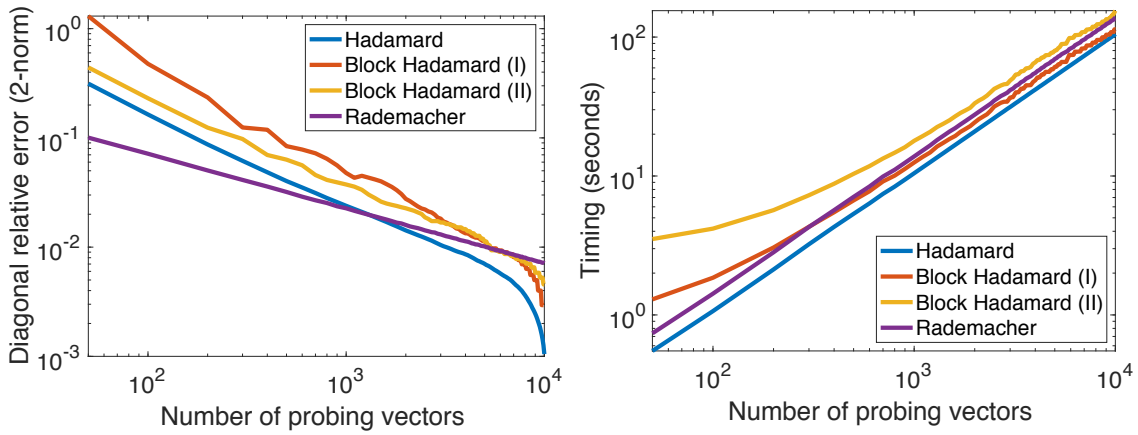
Firstly, this scheme lends itself to parallelization. Each set of s_i probing vectors reveals information only about a subset of s_i diagonal entries of $\mathbf{A}$. Thus, by assigning a set of probing vectors corresponding to one block, $\mathbf{H}_i$ of $\mathbf{H}$, to one processor which can work independently of the others, p processors can collect information about disjoint sets of entries of the diagonal and reconstruct the solution at the end of the process.

Secondly, the block method can be exact for any problem size n, since any natural number

can be decomposed into valid Hadamard sizes which sum up to it. (This is clearly true since 1 is a valid Hadamard size). There arises the question of how to distribute probing vectors within the blocks. We can return to Theorem 2.4 for guidance. Block $\mathbf{H}_i$ is size $s_i \times s_i$, which must be a Hadamard dimension. Specifically, in our implementation $\mathbf{H}_i \in \mathcal{G}(N_1, \ldots, N_k)$ such that $s_i = N_1 N_2 \ldots N_k$. Using Algorithm 2, the final $N_\ell, \ldots, N_k$ will all equal 2, and so probing each block with small powers of 2 will yield the best convergence properties under the above assumptions. There are certainly further optimizations to be done in construction and block size allocation. Our numerical experiments demonstrate that the measures taken thus far reliably outperform a blocking scheme with no optimizations. Namely, with no optimizations, we split $\mathbf{H}$ into p blocks, where p is the number of processors available, and the blocks are of the largest valid Hadamard dimension n_0 such that $n_0 p \geq n$ and the final block is truncated down to n. The number of probing vectors are split evenly across the p blocks. We denote this naive strategy as "Hadamard (I)" and the strategy using Generalized Walsh heuristics to choose better sizes s_i and allocations of probing vectors in accordance with Theorem 2.6 as "Hadamard (II)".

2.3 Numerical Results for Diagonal Probing

Consider an example $\mathbf{A} \in \mathbb{R}^{10^4 \times 10^4}$ with ones on diagonal, and off diagonal are i.i.d. normal variables distributed by $\mathcal{N}(0, .01)$ 5% are nonzero.



(a) Error comparison

(b) Timing comparison

Figure 2.1: Performance comparison of the three Hadamard variations as well as Rademacher probing displaying expected convergence behavior on a 12 processor system.

We make a few comments here. In comparing the Hadamard and Rademacher probing, we see the Rademacher performing better in the first thousand probing vectors, before being overtaken by the Hadamard approach. In particular, we see rapid decay in error of the Hadamard probing near the complete probing.

As expected, we see some additional computational overhead in computing the Block Hadamard (II) estimator compared to Block Hadamard (I), but in exchange this effort results in better convergence properties. We expect that for future research there is the potential for more refinement that can be done to both reduce the overheard in Block Hadamard (II) as well as making convergence even faster.

Chapter 3

Applications and Numerical Experiments

3.1 Generalized Cross-Validation

A diagonal estimation problem arises naturally in the problem of Tikhonov parameter selection Consider the linear inverse problem of the form $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$ is symmetric positive definite, $\mathbf{A} \in \mathbb{R}^{m \times n}$, and $\mathbf{b} \in \mathbb{R}^m$. From this, the goal is to reconstruct $\mathbf{x}$. Specifically, we discuss the ill-posed problem for which $\mathbf{A}$ is highly sensitive to error, and so with $\boldsymbol{\varepsilon}$ present, we need to employ regularization to recover $\mathbf{x}$. By imposing smoothness on the solution, the Tikhonov regularized solution, $\mathbf{x}_\lambda$, takes the form

$$\mathbf{x}_\lambda = \text{argmin}_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda \|\mathbf{x}\|^2 \tag{3.1}$$

Suppose also that $\mathbf{A}$ has rank r. Defining monotonically decreasing filter factors $1 \geq \phi_1, \ldots, \phi_r \geq 0$, then define the filterd $\mathbf{A}$ as

$$\mathbf{A}_\lambda = \mathbf{U}\mathbf{\Phi}\mathbf{\Sigma}\mathbf{V}^\top, \quad \mathbf{\Phi} = \begin{bmatrix} \phi_1 & & & & & \\ & \ddots & & & & \\ & & \phi_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}.$$

Then, one can derive the regularized solution $\mathbf{x}_\lambda$ as a linear operation of the regularization matrix $\mathbf{A}_\lambda$ applied to the observed data $\mathbf{b}$.

$$\mathbf{x}_\lambda = \mathbf{A}_\lambda \mathbf{b}, \quad \text{where } \mathbf{A}_\lambda = \mathbf{V}\mathbf{\Phi}_\lambda \mathbf{\Sigma}^\dagger \mathbf{U}^\top \mathbf{b}. \tag{3.2}$$

The natural question arises as to how to choose $\lambda \geq 0$ to yield the best regularized solution, $\mathbf{x}_\lambda$. One idea is through *leave one out cross-validation* (LOOCV). The general idea of cross validation is that the optimal statistical model is one which, when derived in absence of certain data, can accurately recreate that missing data. LOOCV specifically derives the model for a set of data based on $m-1$ observations. Then, $\lambda \geq 0$ is chosen to minimize an average of a least-squares loss function related to the model derived from $m-1$ observations. That is, we examine instead the Tikhonov regularization problem

$$\min \sum_{k \neq i} \left((\mathbf{A}\mathbf{x})_k - b_k \right)^2 + \lambda^2 \|\mathbf{x}\|_2^2. \tag{3.3}$$

To reduce confusion, we briefly comment on the overloaded use of the subscript $[i]$ in the following discussion. Attached to $\mathbf{b}$ or $\mathbf{A}$, the subscript $[i]$ is used to denote the quantity

with its i^{th} row removed. However, $\mathbf{A}_\lambda^{[i]}$ is the regularization matrix corresponding to $\mathbf{A}^{[i]}$ by way of Equation (3.2) using the filtered $\mathbf{A}^{[i]}$ instead of the filtered $\mathbf{A}$. From this, we define $\mathbf{x}_\lambda^{[i]} := \mathbf{A}_\lambda^{[i]}\mathbf{b}^{[i]}$. In the case of $\mathbf{x}_\lambda$, however, we set $\mathbf{x}_\lambda := \mathbf{A}_\lambda^{[i]}\mathbf{b}^{[i]}$. Define the generalized cross-validation (GCV) function $G(\lambda)$ as the average least squares error for each of the m LOOCV models for the given λ [19]. That is,

$$G(\lambda) = \frac{1}{m}\sum_{i=1}^{m}\left\|\mathbf{b}^{[i]} - \mathbf{A}^{[i]}\mathbf{x}_\lambda^{[i]}\right\|_2^2. \tag{3.4}$$

The regularization parameter is chosen to minimize this loss function.

$$\lambda = \text{argmin}_{\lambda\geq 0}\, G(\lambda).$$

3.1.1 Deriving Generalized Cross-Validation from LOOCV

Equation (3.4) for $G(\lambda)$ is problematic since it requires m system solves to produce the m $\mathbf{A}^{[i]}\mathbf{x}_\lambda^{[i]}$ terms. In order to reduce this complexity, we seek to express $G(\lambda)$ in a more convenient form, following the derivation presented in **??**. First, we fix and i and define $\tilde{\mathbf{b}}$ entrywise as follows:

$$\tilde{b}_k = \begin{cases} [\mathbf{A}\mathbf{x}_\lambda^{[i]}]_k & \text{if } k = i \\ b_i & \text{if } k \neq i. \end{cases}$$

Then, $\mathbf{x}_\lambda^{[i]}$ is also a solution of the LOOCV Tikhonov Problem 3.3 with $\mathbf{b}$ replaced by $\tilde{\mathbf{b}}$, since the two vectors differ only in the i^{th} entry, which is not present in the summation in Equation (3.3).

Since $([\mathbf{A}\mathbf{x}]_i - \tilde{\mathbf{b}}_i)^2$ vanishes when $\mathbf{x} = \mathbf{x}_\lambda^{[i]}$ by the construction of $\tilde{\mathbf{b}}$, we add this term to

Equation (3.3) and note $\mathbf{x}_\lambda^{[i]}$ is then a solution to the modified minimization,

$$\min_{\mathbf{x}} \left([\mathbf{Ax}]_i - \tilde{b}_i \right)^2 + \sum_{k \neq i} \left([\mathbf{Ax}]_k - \tilde{b}_i \right)^2 + \lambda^2 \|\mathbf{x}\|_2^2.$$

By grouping the first two terms together, this is equivalent to

$$\min_{\mathbf{x}} \|\mathbf{Ax} - \tilde{\mathbf{b}}\|_2^2 + \lambda^2 \|\mathbf{x}\|_2^2.$$

This is a Tikhonov minimization problem for the linear model $\mathbf{Ax} = \tilde{\mathbf{b}}$ satisfied by $\mathbf{x}_\lambda^{[k]}$, so by Equation (3.2) the following relationship holds:

$$\mathbf{x}_\lambda^{[k]} = \mathbf{A}_\lambda \tilde{\mathbf{b}}.$$

Then, noticing

$$\mathbf{Ax}_\lambda^{[i]} - \mathbf{Ax}_\lambda = \mathbf{A}(\mathbf{A}_\lambda \tilde{\mathbf{b}}) - \mathbf{A}(\mathbf{A}_\lambda \mathbf{b}) = \mathbf{AA}_\lambda(\tilde{\mathbf{b}} - \mathbf{b})$$

and considering the i^{th} row of this value,

$$\left[\mathbf{Ax}_\lambda^{[i]} - \mathbf{Ax}_\lambda \right]_i = [\mathbf{AA}_\lambda]_i \, (\tilde{\mathbf{b}} - \mathbf{b})$$

and since $\tilde{\mathbf{b}} - \mathbf{b}$ is nonzero only in the i^{th} entry,

$$\left(\mathbf{Ax}_\lambda^{[i]} - \mathbf{Ax}_\lambda \right)_i = [\mathbf{AA}_\lambda]_{ii} \, (\tilde{b}_i - b_i).$$

Next, we subtract both sides of the equation from $\tilde{b}_i - b_i$ producing

$$\begin{aligned}
\tilde{b}_i - b_i - \left[\mathbf{A}\mathbf{x}_\lambda^{[i]} - \mathbf{A}\mathbf{x}_\lambda\right]_i &= \tilde{b}_i - b_i - [\mathbf{A}\mathbf{A}_\lambda]_{ii}\,(\tilde{b}_i - b_i) \\
\tilde{b}_i - b_i - \left[\mathbf{A}\mathbf{x}_\lambda^{[i]}\right]_i + [\mathbf{A}\mathbf{x}_\lambda]_i &= \tilde{b}_i - b_i - [\mathbf{A}\mathbf{A}_\lambda]_{ii}\,(\tilde{b}_i - b_i) \\
\text{Substituting } \tilde{\mathbf{b}}_i &= \mathbf{A}\mathbf{x}_\lambda^{[i]}, \\
\left[\mathbf{A}\mathbf{x}_\lambda^{[i]}\right]_i - \left[\mathbf{A}\mathbf{x}_\lambda^{[i]}\right]_i + [\mathbf{A}\mathbf{x}_\lambda]_i - b_i &= [\mathbf{A}\mathbf{x}_\lambda^{[i]}]_i - b_i - [\mathbf{A}\mathbf{A}_\lambda]_{ii}\left(\left[\mathbf{A}\mathbf{x}_\lambda^{[i]}\right]_i - b_i\right) \\
(\mathbf{A}\mathbf{x}_\lambda)_i - b_i &= \left([\mathbf{A}\mathbf{x}_\lambda^{[i]}]_i - b_i\right)(1 - [\mathbf{A}\mathbf{A}_\lambda]_{ii}) \\
\frac{(\mathbf{A}\mathbf{x}_\lambda)_i - b_i}{1 - [\mathbf{A}\mathbf{A}_\lambda]_{ii}} &= [\mathbf{A}\mathbf{x}_\lambda^{[i]}]_i - b_i.
\end{aligned}$$

Then, substituting this into the formula for $G(\lambda)$ yields

$$\begin{aligned}
G(\lambda) &= \frac{1}{m}\sum_{i=1}^{m}\left(\left[\mathbf{A}^{[i]}\mathbf{x}_\lambda^{[i]}\right]_i - b_i\right)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{[\mathbf{A}\mathbf{x}_\lambda]_i - b_i}{1 - [\mathbf{A}\mathbf{A}_\lambda]_{ii}}\right)^2 \\
G(\lambda) &= \frac{1}{m}\left\|(\mathbf{A}\mathbf{x}_\lambda - \mathbf{b}) \oslash (1 - \operatorname{diag}(\mathbf{A}\mathbf{A}_\lambda))\right\|^2
\end{aligned}$$

Lastly, since $\mathbf{x}_\lambda$ is entirely dependent on $\mathbf{A}$, λ and $\mathbf{b}$ by $\mathbf{x}_\lambda = \mathbf{A}_\lambda\mathbf{b}$, we can write $G(\lambda)$ instead as

$$G(\lambda) = \frac{1}{m}\left\|((\mathbf{I} - \mathbf{A}\mathbf{A}_\lambda)\mathbf{b}) \oslash \operatorname{diag}(\mathbf{I} - \mathbf{A}\mathbf{A}_\lambda)\right\|^2.$$

In general, one assumes it is not feasible to approximate the diagonal entries $[\mathbf{A}\mathbf{A}_\lambda]_{ii}$. The standard generalized cross-validation approach to this problem is to make the approximation

$$[\mathbf{A}\mathbf{A}_\lambda]_{ii} \approx \frac{1}{m}\operatorname{trace}(\mathbf{A}\mathbf{A}_\lambda).$$

That is, substituting the i^{th} diagonal entry with the average diagonal entry of $\mathbf{A}\mathbf{A}_\lambda$. Then,

$$\begin{aligned}
G(\lambda) &= \frac{1}{m} \left\| ((\mathbf{I} - \mathbf{A}\mathbf{A}_\lambda)\mathbf{b}) \oslash \operatorname{diag}(\mathbf{I} - \mathbf{A}\mathbf{A}_\lambda) \right\|^2 \\
&= \frac{1}{m} \sum_{i=1}^{m} \left(\frac{[\mathbf{A}\mathbf{x}]_i - \mathbf{b}}{1 - [\mathbf{A}\mathbf{A}_\lambda]_{ii}} \right)^2 \\
&\approx \frac{1}{m} \sum_{i=1}^{m} \left(\frac{[\mathbf{A}\mathbf{x}]_i - \mathbf{b}}{1 - \operatorname{trace}(\mathbf{A}\mathbf{A}_\lambda)/m} \right)^2 \\
&= \frac{1}{m} \frac{1}{(1 - \operatorname{trace}(\mathbf{A}\mathbf{A}_\lambda)/m)^2} \sum_{i=1}^{m} ([\mathbf{A}\mathbf{x}]_i - \mathbf{b})^2 \\
&= \frac{1}{m} \frac{1}{(1 - \operatorname{trace}(\mathbf{A}\mathbf{A}_\lambda)/m)^2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \\
&= \frac{1}{m} \frac{1}{1/m^2 (m - \operatorname{trace}(\mathbf{A}\mathbf{A}_\lambda))^2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \\
&= \frac{1}{m} \frac{1}{1/m^2 [\operatorname{trace}(\mathbf{I} - \mathbf{A}\mathbf{A}_\lambda)]^2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \\
&= \frac{m \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2}{[\operatorname{trace}(\mathbf{I} - \mathbf{A}\mathbf{A}_\lambda)]^2} \\
&=: H(\lambda).
\end{aligned}$$

In fact, the approximation of $\operatorname{trace}(\mathbf{A}\mathbf{A}_\lambda)$ for approximating the GCV function of a particular linear model was Hutchinson's original motivation for developing t_s^R [28]. Now equipped with the diagonal estimator, we are instead interested in approximating the diagonal entries of $\mathbf{A}\mathbf{A}_\lambda$ in order to estimate $G(\lambda)$ exactly, rather than its approximation, which we denote by $H(\lambda)$.

$$\begin{aligned}
G(\lambda) &= \frac{1}{m} \left\| ((\mathbf{I} - \mathbf{A}\mathbf{A}_\lambda)\mathbf{b}) \oslash \operatorname{diag}(\mathbf{I} - \mathbf{A}\mathbf{A}_\lambda) \right\|^2 \\
H(\lambda) &= m \left\| ((\mathbf{I} - \mathbf{A}\mathbf{A}_\lambda)\mathbf{b}) / \operatorname{trace}(\mathbf{I} - \mathbf{A}\mathbf{A}_\lambda) \right\|^2
\end{aligned}$$

Equivalently, these can be expressed as $G(\lambda) = \|\mathbf{T}\,\mathbf{r}_\lambda\|_2^2$ and $H(\lambda) = \|\mathbf{D}\,\mathbf{r}_\lambda\|_2^2$ using

$$\mathbf{r}_\lambda = \mathbf{A}\mathbf{x}_\lambda - \mathbf{b}, \quad \mathbf{T} = \frac{\sqrt{m}}{m - \operatorname{trace}(\mathbf{A}\mathbf{A}_\lambda)}\mathbf{I}_m, \quad \mathbf{D} = \frac{1}{\sqrt{m}} \begin{bmatrix} (1 - [\mathbf{A}\mathbf{A}_\lambda]_{11})^{-1} & & \\ & \ddots & \\ & & (1 - [\mathbf{A}\mathbf{A}_\lambda]_{mm})^{-1} \end{bmatrix}.$$

We note that while $\mathbf{A}$ is of size $m \times n$ and the matrix whose diagonal must be approximated, $\mathbf{A}\mathbf{A}_\lambda$, is of size $m \times m$. The probing of $\mathbf{A}\mathbf{A}_\lambda$ requires the ability to produce $\mathbf{A}\mathbf{A}_\lambda\mathbf{v}$ for arbitrary $\mathbf{v} \in \mathbb{R}^m$. In real-sized Tikhonov regularization problems, it is assumed to be infeasible to compute the SVD of $\mathbf{A}$, and so we are left with solving

$$\hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left\| \begin{bmatrix} \mathbf{A} \\ \lambda\mathbf{I}_n \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{v} \\ \mathbf{0}_n \end{bmatrix} \right\|_2$$

using, for example, a conjugate-gradient scheme such as Conjugate Gradient Least Squares [36]. This yields $\hat{\mathbf{x}} = \mathbf{A}_\lambda\mathbf{v}$ and then obtaining the desired result through a final matrix vector multiplication by $\mathbf{A}$, $\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}\mathbf{A}_\lambda\mathbf{v}$.

The fundamental assumption made in using H as an approximation of G is that $\operatorname{argmin} H(\lambda) \approx \operatorname{argmin} G(\lambda)$, since this minimizer is the quantity of importance in following the induced Tikhonov regularized system. However, in studying examples from the Regularization Tools [21] toolbox, we see that this is not the case even on some small, simple systems. Test problems for which $\mathbf{A}\mathbf{A}_\lambda$ has high variance along its diagonal should yield large differences between $G(\lambda)$ and $H(\lambda)$.

One particularly extreme example is that of a Hilbert Matrix, described by its ij^{th} entries

$h_{ij} = \frac{1}{i+j-1}$. That is, reciprocals of the integers on the antidiagonals.

$$\mathbf{H}_3 = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix}.$$

David Hilbert first introduced matrices of this form [25] to answer the question of making $\int_0^1 P(x)^2\,dx$ smaller than any fixed $\varepsilon > 0$ for a polynomial P of degree n with integral coefficients. The result is the determinant of the Hilbert matrix of order n.

$\mathbf{H}_n$ is the Gram matrix for the monomial basis of polynomials $\{1, x, \ldots, x^{n-1}\}$ on $L^2([0,1])$ and thus is extremely ill-conditioned, with growth of its condition number according to

$$\mathcal{O}\left(\frac{(1+\sqrt{2})^{4n}}{\sqrt{n}}\right).$$

For example, $\mathbf{H}_3$ above already has $\kappa(\mathbf{H}_3) \approx 524$ and $\kappa(\mathbf{H}_4) \approx 15514$.

We observe that $G(\lambda)$ and $H(\lambda)$ behave remarkably differently for Hilbert matrices. Consider the case $\mathbf{H}_{400}$, as follows.

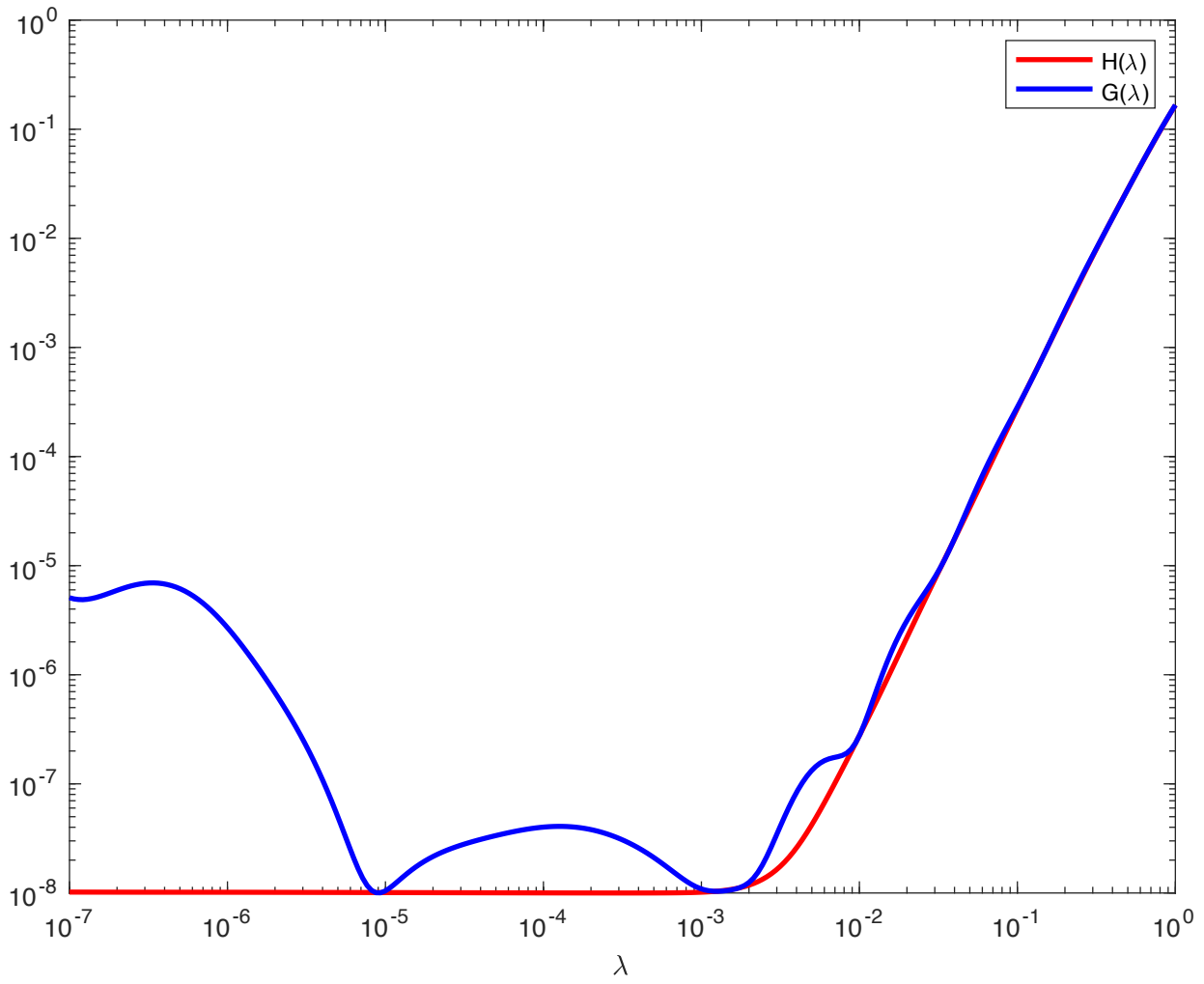


Figure 3.1: Comparing the cross-validation function and the generalized cross validation for the Hilbert matrix of size $n = 400$.

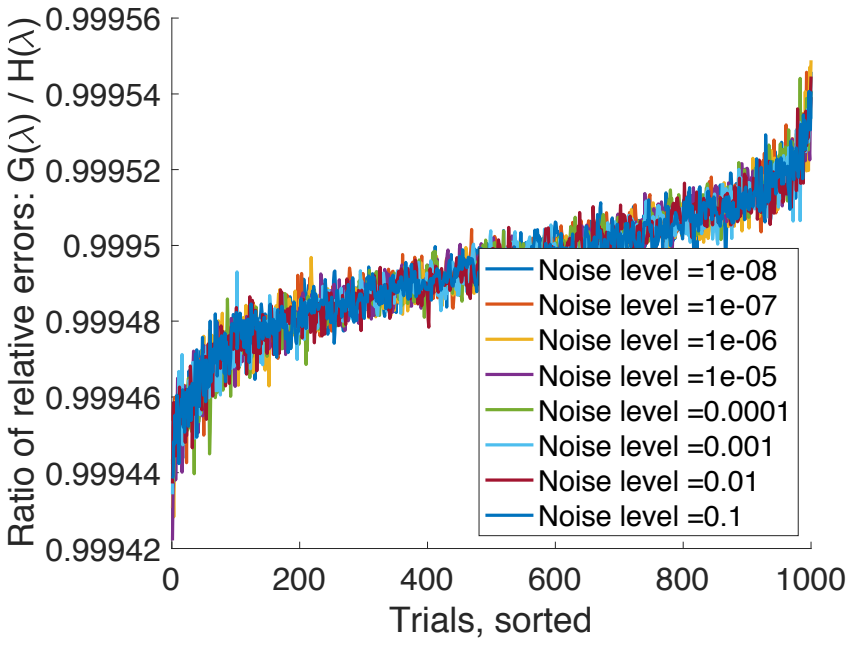
3.1.2 GCV Experiments

Our claim is that $G(\lambda)$ can give more information about the *true* Tikhonov regularization parameter than $H(\lambda)$. $H(\lambda)$ is a function which was born out of necessity, at a time when diagonal approximation was not considered. With a diagonal estimator which requires the same computational effort as the trace estimator, there is no reason to use $H(\lambda)$ over $G(\lambda)$.

Minimizing $G(\lambda)$ should be closer to the optimal value, within the LOOCV philosophy. We seek to uncover examples for which $G(\lambda)$ and $H(\lambda)$ have significantly different minimizers. $G(\lambda)$ is the true GCV function, which exactly is equal to the LOOCV least squares error. $H(\lambda)$ approximates $G(\lambda)$ by replacing the diagonal entries of $\mathbf{A}\mathbf{A}_\lambda$ with their average value, $\text{trace}\left(\mathbf{A}\mathbf{A}_\lambda\right)/m$.

On a standard image deblurring problem, we generate random right-hand sides $\mathbf{b}$ of the form $\mathbf{A}\mathbf{x}+\boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, testing with the noise level σ^2 at various intensities.

We run $N = 1000$ trials at each noise level. First, we construct the right hand side, and then do a brute-force optimization to determine the λ which minimizes least-square reconstruction error in the Tikhonov regularized problem. Then, we compare the relative error between the parameter found by minimizing G and H, respectively.



We observe that G performs slightly yet consistently better than H in selecting a λ parameter closer to optimal. This aligns with our expectations, and justifies the use of $G(\lambda)$ and the induced diagonal estimation problem rather than $H(\lambda)$ and its associated trace estimation

problem.

3.2 Diagonal Ordering Experiments

Up until now, we have only considered the problem of using the diagonal estimator to approximate the diagonal of $\mathbf{A}$ in the 2-norm sense. That is, we have sought to minimize

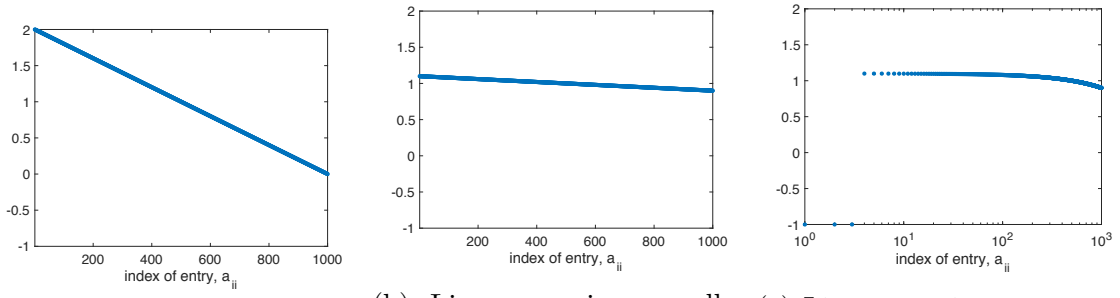
$$error_{rel}(\mathbf{d}_s) = \frac{\|\mathbf{d}_s - \operatorname{diag}(\mathbf{A})\|_2}{\|\operatorname{diag}(\mathbf{A})\|_2}.$$

However, there exist other applications in which only an estimate on information about the ordering of the diagonals of $\mathbf{A}$ is required. One may expect this should be an easier problem, as the difficulty of achieving high precision estimates has already been discussed as one of the primary weaknesses of stochastic probing vectors. There are many choices for an error function which compares the rankings of two arrays [17]. Ultimately, for the scope of this thesis, there is not a significant difference between these choices, as we do not present any detailed convergence analysis which would depend on these details. For our comparison tests, we choose the Kendall tau distance, which counts the number of "disagreements" in pair orderings between two arrays. For two arrays of length n, each of the $n(n-1)/2$ pairs of each array are identified as in agreement, or in disagreement between the two orderings. The sum of the number of disagreements, normalized by $n(n-1)/2$ results in the Kendall tau distance. Hence, it lies on $[0, 1]$, equal to zero if and only if the two arrays are reverse orderings of each other, and equal to 1 if and only if the two orderings are the same. In additional defense of its suitability, it is a metric if we assume no repeated entries in the arrays, but this condition is not necessary for our purposes.

Consider $\mathbf{A} \in \mathbb{R}^{10^3 \times 10^3}$ with off-diagonal entries having the form $a_{ij} = \exp(-0.1|i-j|)$ with

90% of them chosen at random and set to 0.

We conduct three experiments to compare the efficacy of each of the probing methods for identifying the proper ranking of the diagonal entries, as well as the easier problem of identifying only the minimum diagonal entry.

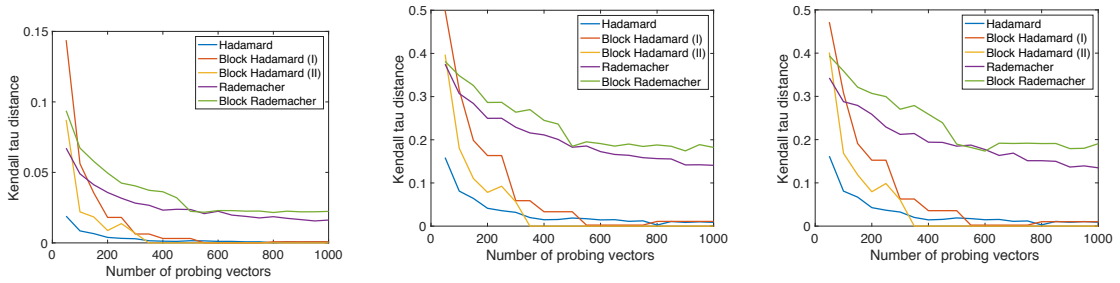


(a) Linear spacing

(b) Linear spacing, smaller step

(c) Linear spacing, separated cluster

Figure 3.2: Visual representation of the diagonal entries of the three test matrices.



(a) Linear spacing

(b) Linear spacing, smaller step

(c) Linear spacing, separated cluster

Figure 3.3: Visual representation of the diagonal entries of the three test matrices.

From these experiments, we see that the Hadamard probing strategies all outperform the Rademacher strategies. In particular the Block Hadamard estimators using a smarter blocking and probing vector distribution consistently outperforms the Block Hadamard estimator using a naive strategy.

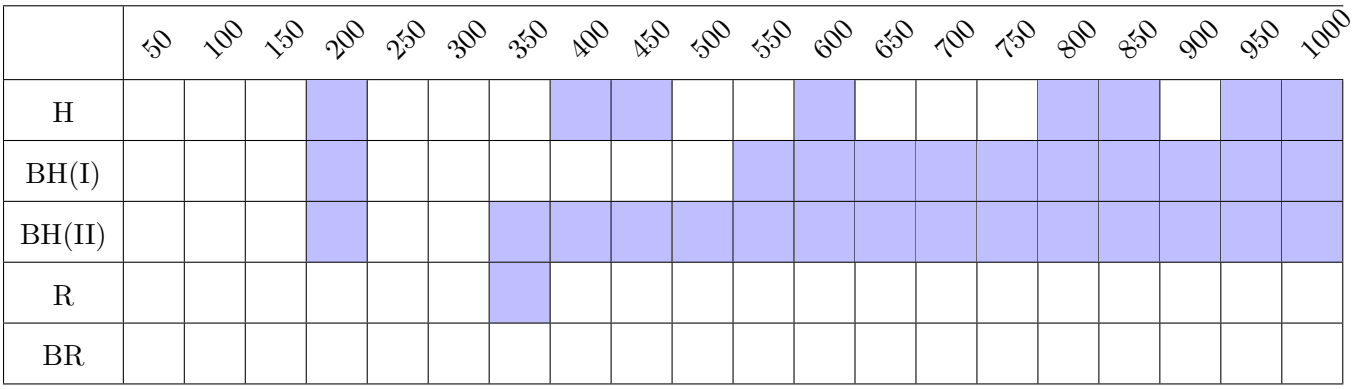


Figure 3.4: Colored cells flag which estimators were able to accurately identify the smallest-magnitude diagonal entry in the linear spacing test matrix.

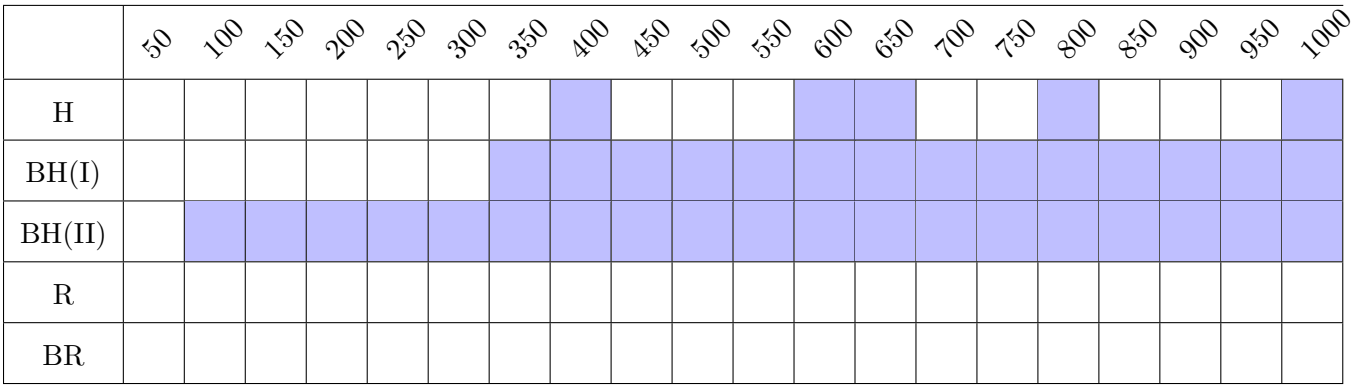


Figure 3.5: Colored cells flag which estimators were able to accurately identify the smallest-magnitude diagonal entry in the small increment linear spacing test matrix.

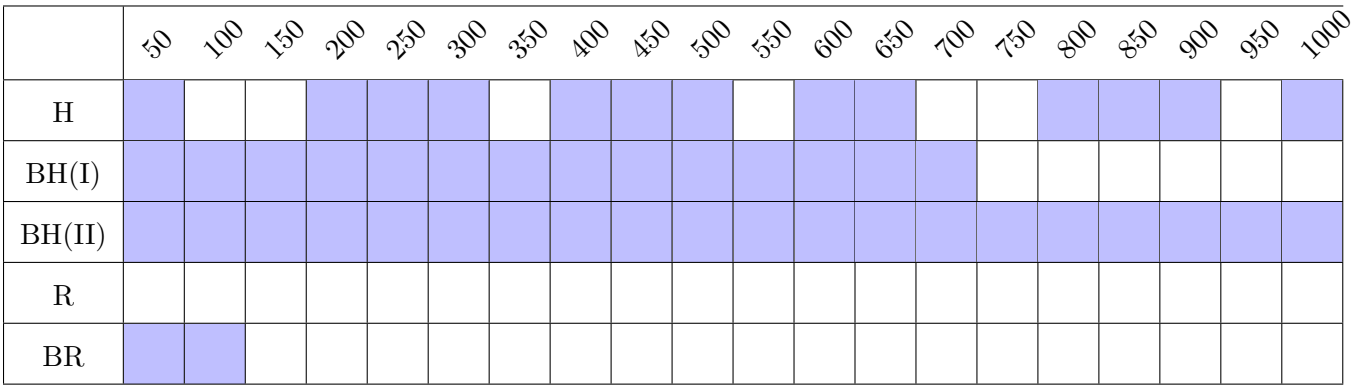


Figure 3.6: Colored cells flag which estimators were able to accurately identify the smallest-magnitude diagonal entry in the test matrix with a separated cluster of 3 diagonal entries

We see a slightly different result when examining only the convergence of the minimum diagonal entry to the correct value. We see the smart block procedure performs even better than the full Hadamard estimator. The two Rademacher estimators perform extremely poorly, with only occasional hits on the correct minimum, before losing it again. More analysis could be fruitful in shedding further light on the performance of the Block Hadamard estimator for these kinds of estimation problems.

3.3 Future Work

3.3.1 Matrix Function Approximation

Matrix functions are an active area of research for many problems in statistical inference, differential equations, and approximation theory [13, 20, 23, 39]. We present some basic context for this problem, and mention some specific application areas in which trace and diagonal estimation arise. First we provide some explicit representations for $f(\mathbf{A})$. For a function f analytic on Σ containing the spectrum of $\mathbf{A}$, we can defined the matrix function $f(\mathbf{A})$ uniquely using Hermite interpolation.

Definition 3.1. (Matrix function via Hermite interpolation, [24]) Let f be defined on the spectrum of $\mathbf{A} \in \mathbb{C}^{n\times n}$ and let ϕ be the minimal polynomial of $\mathbf{A}$. Then, $f(\mathbf{A}) := p(\mathbf{A})$ where p is the polynomial of degree less than

$$\sum_{i=1}^{s} n_i = \deg \phi$$

(where n_i is the index of λ_i, the size of the largest Jordan block containing λ_i) that satisfies

the interpolation conditions

$$p^{(j)}(\lambda_i) = f^{(j)}(\lambda_i), \quad j = 0, \ldots, n_i - 1, \ i = 1, \ldots, s.$$

There is a unique p and it is known as the Hermite interpolating polynomial.

If $\mathbf{A}$ has an eigendecomposition $\mathbf{PDP}^{-1}$, then $f(\mathbf{A})$ is simply expressed

$$f(\mathbf{A}) = \mathbf{P} \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_n) \end{bmatrix} \mathbf{P}^{-1}.$$

We note that the function acts on the eigenvalues of $\mathbf{A}$, but not its eigenvectors. Another representation which proves useful particularly in error analysis is the Cauchy integral form of $f(\mathbf{A})\mathbf{b}$ for $\mathbf{b} \in \mathbb{R}^n$.

Definition 3.2. (matrix function via Cauchy integral, [24]). For $\mathbf{A} \in \mathbb{C}^{n \times n}$,

$$f(\mathbf{A}) := \frac{1}{2\pi i} \int_\Gamma f(z)(z\mathbf{I} - \mathbf{A})^{-1} \, dz,$$

where f is analytic on and inside a closed contour Γ that encloses the spectrum of $\mathbf{A}$.

These different representations suggest a rich variety of strategies for approximating $f(\mathbf{A})$ and $f(\mathbf{A})\mathbf{b}$. The Hermite interpolation form suggests Krylov subspace methods can prove helpful. The Cauchy integral form motivates quadrature-based methods.

For large n, the eigendecomposition is too expensive to compute. For diagonal and trace estimation of $f(\mathbf{A})$, we need only compute $s \ll n$ matrix-vector multiplications rather than the full matrix $f(\mathbf{A})$. That is, we need only estimate $f(\mathbf{A})$ accurately in one specified

direction. In the context of diagonal probing, this is precisely what is desired, in which $\mathbf{b}$ is a specified probing vector.

3.3.2 Log-determinant Computation

One example where trace estimation has played a crucial role is in the computation of the log-determinant of a covariance matrix [35]. Given a data vector $\mathbf{y} \in \mathbb{R}^n$ and Gaussian process regression requires the training of a Gaussian process characterized by the covariance matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$ depending on parameters θ. One way to train the Gaussian process is through maximum likelihood estimation, which involves minimizing the negative log-likelihood function

$$\mathcal{L}(\theta) = \frac{1}{2} \log \det \mathbf{C} + \frac{1}{2} \mathbf{y}^\top \mathbf{C}^{-1} \mathbf{y}.$$

Being a covariance matrix, $\mathbf{C}$ is symmetric positive-definite and so its eigenvalues are all positive. Then, we can write the $\log \det(\mathbf{A})$ as follows.

$$\log \det(\mathbf{A}) = \log(\det(\mathbf{A})) = \log(\prod_{i=1}^{n} \lambda_i) = \sum_{i=1}^{n} \log(\lambda_i) = \operatorname{trace}(\log \mathbf{A}).$$

Note, the final log is a matrix function. The problem then of estimating $\log \det(\mathbf{A})$ can be reduced a trace estimation. $\log(\mathbf{A})$ has a convergent series representation, provided $\|\mathbf{A}\|_2 < 1$ given by

$$\log(\mathbf{I} - \mathbf{A}) = -\sum_{k=1}^{\infty} \frac{1}{k} \mathbf{A}^k.$$

Then, for general symmetric positive-definite matrices $\mathbf{A}$, one can estimate an upper bound α on the spectrum of $\mathbf{A}$ and then one has

$$\log \det(\mathbf{A}) = n \log(\alpha) - \sum_{k=1}^{\infty} \frac{1}{k} \operatorname{trace}\left((\mathbf{I} - \mathbf{A}/\alpha)^k\right),$$

requiring only the estimation of the trace of $\mathbf{A}^k$ for k up to some predefined number of terms proportional to the accuracy of the approximation [10].

3.3.3 Subset Selection Inequality

Corollary 3.3 [30]. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric positive definite with eigenvalues contained within the interval $[a, b]$. Define $\varrho = \frac{4ab}{(a+b)^2}$. If $\mathbf{K} \in \mathbb{R}^{n\times m}$, $m < n$ is an isometry, $\mathbf{K}^\top\mathbf{K} = \mathbf{I}_k$, then:*

$$\operatorname{trace}\left(\mathbf{K}^\top \log \mathbf{AK}\right) \leq \log\det(\mathbf{K}^\top\mathbf{AK}).$$

When $\mathbf{K}$ is a selection operator, so that $\mathbf{K}^\top \log \mathbf{AK}$ is just a principal submatrix of $\log \mathbf{A}$, then estimating information about the subset of diagonals selected by a particular $\mathbf{K}$ gives estimation information about the trace of $\mathbf{K}^\top \log \mathbf{AK}$.

In this context, ranking the diagonal entries as was done in Section 3.2 can be useful in selecting a particular submatrix of $\mathbf{A}$ in the context of Bayesian experimental design.

3.3.4 Matrix Updating and Network Analysis

Another important problem in matrix function theory is that of approximating information about a matrix function update [7]. That is, given f and $\mathbf{A}$ as before, and a (typically low-rank) matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, we seek to approximate information about

$$f(\mathbf{A} + \mathbf{B}) - f(\mathbf{A}). \tag{3.5}$$

Additionally, for this to be well-defined, we need f to be analytic not just on the spectrum of $\mathbf{A}$, but also on the spectrum of $\mathbf{A} + \mathbf{B}$ as well.

One application which requires the diagonal entries of (3.5) is in subgraph centrality measurement [15, 16]. Let $G = (V, E)$ with $V = \{1, \ldots, n\}$ and $E \subseteq V \times V$ be an undirected graph, and $\mathbf{A}$ its adjacency matrix. That is, $a_{ij} = 1$ if $(i, j) \in E$ and 0 otherwise. *Subgraph centrality* of the i^{th} node is given

$$f_i(\mathbf{A}) = \frac{[\exp(\mathbf{A})]_{ii}}{\operatorname{trace}(\exp(\mathbf{A}))}.$$

In particular, in studying the problem of *total communication* in the network, one studies these subgraph centralities under an updating or downdating of the network [2]. That is, as an edge is added/removed between the i^{th} and j^{th} nodes, corresponding to a rank two update to $\mathbf{A}$.

Chapter 4

Conclusion

We have addressed the problem of diagonal estimation with probing methods by analyzing convergence behavior with both stochastic and deterministic choices of probing vectors. We have extended analysis for the trace estimator to the corresponding results for the diagonal estimator for Rademacher and Gaussian probing vectors. We have analyzed thoroughly Hadamard matrices and design decisions that must be considered when using their columns as a probing basis. We provide context for classical combinatorial results for Hadamard matrices, and present some of these results for the first time as results on error accumulation in diagonal estimators. As a response to some of the inherent difficulties of Hadamard matrix columns as probing vectors, we have proposed a new class of probing methods using a blocked structure, and analyzed the convergence behavior of these methods. In particular, we provide heuristic as well as combinatorial arguments for the best way to construct these vectors. We proceed by discussing different contexts where diagonal information about large, implicitly-defined matrices is required, and these methods can be applied.

To accompany this analysis, we provide numerical examples to support the claims with numerical results on real-world diagonal estimation problems, and demonstrate circumstances where our analysis and new methods can be of use. Our results show that the Block Hadamard estimators come close to the performance of the standard Hadamard estimator, with the potential for further improvement. Somewhat surprisingly, on our tests thus far, we observe the Block Hadamard estimator outperforming all other methods in the task

of identifying extremal diagonal entries, which suggests further analysis may support this finding.

Bibliography

[1] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281. ACM, 2001.

[2] Francesca Arrigo and Michele Benzi. Updating and downdating techniques for optimizing network communicability. *SIAM Journal on Scientific Computing*, 38(1):B25–B49, 2016.

[3] Haim Avron. Counting triangles in large graphs using randomized matrix trace estimation. In *Workshop on Large-scale Data Mining: Theory and Applications*, volume 10, pages 10–9, 2010.

[4] Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2): 8, 2011.

[5] V. R. Algazi B. J. Fino. Unified matrix treatment of the fast walsh-hadamard transform. *IEEE Transactions on Computers*, 25(11):1142–1146, 11 1976.

[6] Roi Baer and Martin Head-Gordon. Chebyshev expansion methods for electronic structure calculations on large molecular systems. *The Journal of Chemical Physics*, 107 (23):10003–10013, 1997. doi: 10.1063/1.474158.

[7] Bernhard Beckermann, Daniel Kressner, and Marcel Schweitzer. Low-rank updates of matrix functions. *SIAM Journal on Matrix Analysis and Applications*, 39(1):539–565, 2018.

[8] C. Bekas, A. Curioni, and I. Fedulova. Low cost high performance uncertainty quantification. In *Proceedings of the 2nd Workshop on High Performance Computational Finance*, WHPCF '09, pages 8:1–8:8, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-716-5. doi: 10.1145/1645413.1645421. URL http://doi.acm.org/10.1145/1645413.1645421.

[9] Costas Bekas, Effrosyni Kokiopoulou, and Yousef Saad. An estimator for the diagonal of a matrix. *Applied numerical mathematics*, 57(11):1214–1229, 2007.

[10] Christos Boutsidis, Petros Drineas, Prabhanjan Kambadur, Eugenia-Maria Kontopoulou, and Anastasios Zouzias. A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix. *Linear Algebra and its Applications*, 533:95–117, 2017.

[11] Jie Chen. How accurately should i compute implicit matrix-vector products when applying the hutchinson trace estimator? *SIAM Journal on Scientific Computing*, 38 (6):A3515–A3539, 2016. doi: 10.1137/15M1051506. URL https://doi.org/10.1137/15M1051506.

[12] Lingji Chen, Pablo O Arambel, and Raman K Mehra. Estimation under unknown correlation: Covariance intersection revisited. *IEEE Transactions on Automatic Control*, 47(11):1879–1882, 2002.

[13] Eugene D Denman and Alex N Beavers Jr. The matrix sign function and computations in systems. *Applied mathematics and Computation*, 2(1):63–94, 1976.

[14] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[15] Ernesto Estrada and Desmond J Higham. Network properties revealed through matrix functions. *SIAM review*, 52(4):696–714, 2010.

[16] Ernesto Estrada and Juan A Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005.

[17] Ronald Fagin, Ravi Kumar, and Dakshinamurthi Sivakumar. Comparing top k lists. *SIAM Journal on discrete mathematics*, 17(1):134–160, 2003.

[18] D. Girard. Un algorithme simple et rapide pour la validation croisée généralisée sur des problèmes de grande taille. *Inf. et Math.Appl. de Grenoble*, 669-M, 1987.

[19] Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[20] Nicholas Hale, Nicholas J Higham, and Lloyd N Trefethen. Computing a^α,\log(a), and related matrix functions by contour integrals. *SIAM Journal on Numerical Analysis*, 46(5):2505–2523, 2008.

[21] Per Christian Hansen. Regularization tools version 4.0 for matlab 7.3. *Numerical algorithms*, 46(2):189–194, 2007.

[22] Nicholas J Higham. *1.2 Relative Error and Significant Digits*, volume 80. Siam, 2002.

[23] Nicholas J. Higham. 2013. doi: 10.1137/1.9780898717778.ch1. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898717778.ch1.

[24] Nicholas J. Higham. *1. Theory of Matrix Functions*, pages 1–34. 2013. doi: 10.1137/1.9780898717778.ch1. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898717778.ch1.

[25] David Hilbert. Ein beitrag zur theorie des legendre'schen polynoms. *Acta mathematica*, 18(1):155–159, 1894.

[26] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[27] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136 (3B):B864, 1964.

[28] MF Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *communications in statistics-simulation and computation*, 19(2):433–450, 1990.

[29] Toshiaki Iitaka and Toshikazu Ebisuzaki. Random phase vector for calculating the trace of a large matrix. *Physical Review E*, 69(5):057701, 2004.

[30] Jayanth Jagular-Mohan and Youssef Marzouk. Bayesian optimal experimental design in non-submodular settings: batch algorithms and guarantees. private communication.

[31] Hadi Kharaghani and Behruz Tayfeh-Rezaie. A hadamard matrix of order 428. *Journal of Combinatorial Designs*, 13(6):435–440, 2005.

[32] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

[33] A.N. Kolmogorov and Albert T Bharucha-Reid. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications, 2018.

[34] Donald J. Kouri, Youhong Huang, and David K. Hoffman. Direct approach to density functional theory: Heaviside-fermi level operator using a pseudopotential treatment. *The Journal of Physical Chemistry*, 100(19):7903–7910, 1996.

[35] WE Leithead, Yunong Zhang, and DJ Leith. Efficient gaussian process based on bfgs updating and logdet approximation. *IFAC Proceedings Volumes*, 38(1):1305–1310, 2005.

[36] Michael Saunders, Per Christian Hansen, Folkert Bleichrodt, Christopher Fougner. Cgls: Cg method for $ax = b$ and least squares. URL http://stanford.edu/group/SOL/software/cgls/.

[37] Victor H Moll. Special integrals of gradshteyn and ryzhik: the proofs. 2014.

[38] Chan-Hyoung Park, Hong-Yeop Song, and Kyu Tae Park. Existence and classification of hadamard matrices. In *Signal Processing Proceedings, 1998. ICSP '98. 1998 Fourth International Conference on*, pages 117–121 vol.1, 1998. doi: 10.1109/ICOSP.1998.770165.

[39] Y. Saad. Analysis of some krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis*, 29(1):209–228, 1992. ISSN 00361429. URL http://www.jstor.org/stable/2158085.

[40] Dilip V. Sarwate. *Meeting the Welch Bound with Equality*, pages 79–102. Springer London, London, 1999. ISBN 978-1-4471-0551-0. doi: 10.1007/978-1-4471-0551-0_6. URL http://dx.doi.org/10.1007/978-1-4471-0551-0_6.

[41] Irina Shevtsova. On the absolute constants in the berry-esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*, 2011.

[42] Wen Shih. n-dimension spherical coordinates and the volumes of the n-ball in n, 1999. URL http://www.ams.sunysb.edu/~wshih/mathnotes/n-D_Spherical_coordinates.pdf.

[43] Jok M. Tang and Yousef Saad. A probing method for computing the diagonal of a matrix inverse. *Numerical Linear Algebra with Applications*, 2012. ISSN 1099-1506.

[44] Richard S Varga. *Geršgorin and his circles*, volume 36. Springer Science & Business Media, 2010.

.1 Proofs for Section 2.1 (The Problem of Construction)

Proof (Theorem 2.4). Let $N, k \in \mathbb{N}$ and $\mathbf{H} \in \text{Walsh}(N, k)$. By definition of Walsh$(N, k)$, there exist $\mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(k)}$ all elements of Had (N) such that

$$\mathbf{H} = \mathbf{H}^{(1)} \otimes \mathbf{H}^{(2)} \otimes \cdots \otimes \mathbf{H}^{(k)}.$$

Fix $\ell \in \{0, \ldots, k-1\}$. Set

$$\mathbf{W}_0 = \mathbf{H}^{(1)} \otimes \mathbf{H}^{(2)} \otimes \cdots \otimes \mathbf{H}^{(k-\ell)}$$
$$\mathbf{W}_1 = \mathbf{H}^{(k-\ell+1)} \otimes \mathbf{H}^{(k-\ell+2)} \otimes \cdots \otimes \mathbf{H}^{(k)}.$$

Observe $\mathbf{H} = \mathbf{W}_0 \otimes \mathbf{W}_1$ with $\mathbf{W}_0 \in \text{Walsh}(N, k-\ell)$ and $\mathbf{W}_1 \in \text{Walsh}(N, \ell)$, so both $\mathbf{W}_0$ and $\mathbf{W}_1$ are Hadamard matrices. Denote the columns of $\mathbf{W}_0$ as $\mathbf{W}_0 = \begin{bmatrix} \mathbf{h}_1, & \mathbf{h}_2, & \cdots, & \mathbf{h}_{N^{k-\ell}} \end{bmatrix}$. Then,

$$\mathbf{H} = \mathbf{W}_0 \otimes \mathbf{W}_1 = \begin{bmatrix} \mathbf{h}_1 \otimes \mathbf{W}_1, & \mathbf{h}_2 \otimes \mathbf{W}_1, & \cdots, & \mathbf{h}_{N^{k-\ell}} \otimes \mathbf{W}_1 \end{bmatrix}.$$

Each of these blocks $\mathbf{h}_i \otimes \mathbf{W}_1$ is of dimension $N^k \times N^\ell$. Let $j \in \{1, \ldots, N^{k-\ell}\}$ be arbitrary and denote $\tilde{\mathbf{H}}$ to be the first jN^ℓ columns of $\mathbf{H}$. $\tilde{\mathbf{H}}$ consists precisely of the first j blocks of

the form $\mathbf{h}_i \otimes \mathbf{W}_1$. Consider $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^\top$.

$$\begin{aligned}
\tilde{\mathbf{H}}\tilde{\mathbf{H}}^\top &= \begin{bmatrix} \mathbf{h}_1 \otimes \mathbf{W}_1, & \mathbf{h}_2 \otimes \mathbf{W}_1, & \cdots, & \mathbf{h}_j \otimes \mathbf{W}_1 \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \otimes \mathbf{W}_1, & \mathbf{h}_2 \otimes \mathbf{W}_1, & \cdots, & \mathbf{h}_j \otimes \mathbf{W}_1 \end{bmatrix}^\top \\
&= \begin{bmatrix} \mathbf{h}_1 \otimes \mathbf{W}_1, & \mathbf{h}_2 \otimes \mathbf{W}_1, & \cdots, & \mathbf{h}_j \otimes \mathbf{W}_1 \end{bmatrix} \begin{bmatrix} (\mathbf{h}_1 \otimes \mathbf{W}_1)^\top \\ (\mathbf{h}_2 \otimes \mathbf{W}_1)^\top \\ \vdots \\ (\mathbf{h}_j \otimes \mathbf{W}_1)^\top \end{bmatrix} \\
&= \sum_{i=1}^{j} (\mathbf{h}_i \otimes \mathbf{W}_1)(\mathbf{h}_i \otimes \mathbf{W}_1)^\top \\
&= \sum_{i=1}^{j} (\mathbf{h}_i \otimes \mathbf{W}_1)(\mathbf{h}_i^\top \otimes \mathbf{W}_1^\top) \\
&= \sum_{i=1}^{j} (\mathbf{h}_i \mathbf{h}_i^\top) \otimes (\mathbf{W}_1 \mathbf{W}_1^\top) \\
&= \sum_{i=1}^{j} (\mathbf{h}_i \mathbf{h}_i^\top) \otimes (N^\ell \mathbf{I}_{N^\ell}).
\end{aligned}$$

Since $N^\ell \mathbf{I}_{N^\ell}$ is diagonal, each term in this summation is the Kronecker product of a matrix with a diagonal matrix of size $N^\ell \times N^\ell$ which implies they will have all zero entries on sub and super diagonals $1, \ldots, N^\ell - 1$. This structure is preserved under summation, so $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^\top$ too will have all zero entries on its first through $N^\ell - 1^{th}$ sub and super diagonals. The ij^{th} entry of $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^\top$ is 0 if and only if the i^{th} and j^{th} rows of $\tilde{\mathbf{H}}$ are orthogonal. Then, using Proposition 1.8, we conclude the diagonal estimation of a matrix $\mathbf{A} \in \mathbb{R}^{N^k \times N^k}$ performed with the columns of $\tilde{\mathbf{H}}$ will be degraded only by the nonzero entries of $\mathbf{A}$ outside the $N^\ell - 1$ band around the diagonal of $\mathbf{A}$. Thus, if $\mathbf{A}$ has bandwidth at most $N^\ell - 1$ (i.e., strictly less than N^ℓ), the estimate will be exact, as desired. □

The proof for Theorem 2.6 follows closely to that of Theorem 2.4.

Proof (Theorem 2.6). Let $k, N_1, \ldots, N_k \in \mathbb{N}$ and $\mathbf{H} \in \mathcal{G}(N_1, \ldots, N_k)$. By definition of $\mathcal{G}(N_1, \ldots, N_k)$, there exist $\mathbf{H}^{(1)} \in \mathrm{Had}\,(N_1), \ldots, \mathbf{H}^{(k)} \in \mathrm{Had}\,(N_k)$ such that

$$\mathbf{H} = \mathbf{H}^{(1)} \otimes \mathbf{H}^{(2)} \otimes \cdots \otimes \mathbf{H}^{(k)}.$$

Fix $\ell \in \{1, \ldots, k\}$. Set

$$\mathbf{W}_0 = \mathbf{H}^{(1)} \otimes \mathbf{H}^{(2)} \otimes \cdots \otimes \mathbf{H}^{(\ell-1)}$$
$$\mathbf{W}_1 = \mathbf{H}^{(\ell)} \otimes \mathbf{H}^{(\ell+1)} \otimes \cdots \otimes \mathbf{H}^{(k)}.$$

Observe $\mathbf{H} = \mathbf{W}_0 \otimes \mathbf{W}_1$ with $\mathbf{W}_0 \in \mathcal{G}(N_1, \ldots, N_{\ell-1})$ and $\mathbf{W}_1 \in \mathcal{G}(N_\ell, \ldots, N_k)$, so both $\mathbf{W}_0$ and $\mathbf{W}_1$ are Hadamard matrices. Denote the columns of $\mathbf{W}_0$ as $\mathbf{W}_0 = \begin{bmatrix} \mathbf{h}_1, & \mathbf{h}_2, & \cdots, & \mathbf{h}_{N^{\ell-1}} \end{bmatrix}$. Then,

$$\mathbf{H} = \mathbf{W}_0 \otimes \mathbf{W}_1 = \begin{bmatrix} \mathbf{h}_1 \otimes \mathbf{W}_1, & \mathbf{h}_2 \otimes \mathbf{W}_1, & \cdots, & \mathbf{h}_{N^{\ell-1}} \otimes \mathbf{W}_1 \end{bmatrix}.$$

Each of these blocks $\mathbf{h}_i \otimes \mathbf{W}_1$ is of dimension $N_1 N_2 \cdots N_k \times N_\ell N_{\ell+1} \cdots N_k$. Let $j \in \{1, \ldots, N_1 N_2 \cdots N_{\ell-1}\}$ be arbitrary and denote $\tilde{\mathbf{H}}$ to be the first $j N_1 N_2 \cdots N_{\ell-1}$ columns of $\mathbf{H}$. $\tilde{\mathbf{H}}$ consists precisely of the first j blocks of the form $\mathbf{h}_i \otimes \mathbf{W}_1$.

Consider $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^\top$.

$$\begin{aligned}
\tilde{\mathbf{H}}\tilde{\mathbf{H}}^\top &= \begin{bmatrix} \mathbf{h}_1 \otimes \mathbf{W}_1, & \mathbf{h}_2 \otimes \mathbf{W}_1, & \cdots, & \mathbf{h}_j \otimes \mathbf{W}_1 \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \otimes \mathbf{W}_1, & \mathbf{h}_2 \otimes \mathbf{W}_1, & \cdots, & \mathbf{h}_j \otimes \mathbf{W}_1 \end{bmatrix}^\top \\
&= \begin{bmatrix} \mathbf{h}_1 \otimes \mathbf{W}_1, & \mathbf{h}_2 \otimes \mathbf{W}_1, & \cdots, & \mathbf{h}_j \otimes \mathbf{W}_1 \end{bmatrix} \begin{bmatrix} (\mathbf{h}_1 \otimes \mathbf{W}_1)^\top \\ (\mathbf{h}_2 \otimes \mathbf{W}_1)^\top \\ \vdots \\ (\mathbf{h}_j \otimes \mathbf{W}_1)^\top \end{bmatrix} \\
&= \sum_{i=1}^{j} (\mathbf{h}_i \otimes \mathbf{W}_1)(\mathbf{h}_i \otimes \mathbf{W}_1)^\top \\
&= \sum_{i=1}^{j} (\mathbf{h}_i \otimes \mathbf{W}_1)(\mathbf{h}_i^\top \otimes \mathbf{W}_1^\top) \\
&= \sum_{i=1}^{j} (\mathbf{h}_i \mathbf{h}_i^\top) \otimes (\mathbf{W}_1 \mathbf{W}_1^\top) \\
&= \sum_{i=1}^{j} (\mathbf{h}_i \mathbf{h}_i^\top) \otimes (N_\ell N_{\ell+1} \cdots N_k \mathbf{I}_{N_\ell N_{\ell+1} \cdots N_k}).
\end{aligned}$$

Since $N_\ell N_{\ell+1} \cdots N_k \mathbf{I}_{N_\ell N_{\ell+1} \cdots N_k}$ is diagonal, each term in this summation is the Kronecker product of a matrix with a diagonal matrix of size $N_\ell N_{\ell+1} \cdots N_k \times N_\ell N_{\ell+1} \cdots N_k$ which implies they will have all zero entries on sub and super diagonals $1, \ldots, N_\ell N_{\ell+1} \cdots N_k - 1$. This structure is preserved under summation, so $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^\top$ too will have all zero entries on its first through $N_\ell N_{\ell+1} \cdots N_k - 1^{th}$ sub and super diagonals. The ij^{th} entry of $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^\top$ is 0 if and only if the i^{th} and j^{th} rows of $\tilde{\mathbf{H}}$ are orthogonal. Then, using Proposition 1.8, we conclude the diagonal estimation of a matrix $\mathbf{A} \in \mathbb{R}^{N_1 N_2 \cdots N_k \times N_1 N_2 \cdots N_k}$ performed with the columns of $\tilde{\mathbf{H}}$ will be degraded only by the nonzero entries of $\mathbf{A}$ outside the $N_\ell N_{\ell+1} \cdots N_k - 1$ band around the diagonal of $\mathbf{A}$. Thus, if $\mathbf{A}$ has bandwidth at most $N_\ell N_{\ell+1} \cdots N_k - 1$ (i.e., strictly

less than $N_\ell N_{\ell+1} \cdots N_k$), the estimate will be exact, as desired. □