4 Information-Theoretic Bounds on Sketching

Mert Pilanci

Summary

Approximate computation methods with provable performance guarantees are becoming important and relevant tools in practice. In this chapter, we focus on sketching methods designed to reduce data dimensionality in computationally intensive tasks. Sketching can often provide better space, time, and communication complexity trade-offs by sacrificing minimal accuracy. This chapter discusses the role of information theory in sketching methods for solving large-scale statistical estimation and optimization problems. We investigate fundamental lower bounds on the performance of sketching. By exploring these lower bounds, we obtain interesting trade-offs in computation and accuracy. We employ Fano's inequality and metric entropy to understand fundamental lower bounds on the accuracy of sketching, which is parallel to the information-theoretic techniques used in statistical minimax theory.

4.1 Introduction

In recent years we have witnessed an unprecedented increase in the amount of available data in a wide variety of fields. Approximate computation methods with provable performance guarantees are becoming important and relevant tools in practice to attack larger-scale problems. The term *sketching* is used for randomized algorithms designed to reduce data dimensionality in computationally intensive tasks. In large-scale problems, sketching allows us to leverage limited computational resources such as memory, time, and bandwidth, and also explore favorable trade-offs between accuracy and computational complexity.

Random projections are widely used instances of sketching, and have attracted substantial attention in the literature, especially very recently in the machine-learning, signal processing, and theoretical computer science communities [1–6]. Other popular sketching techniques include leverage score sampling, graph sparsification, core sets, and randomized matrix factorizations. In this chapter we overview sketching methods, develop lower bounds using information-theoretic techniques, and present upper bounds on their performance. In the next section we begin by introducing commonly used sketching methods.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

This chapter focuses on the role of information theory in sketching methods for solving large-scale statistical estimation and optimization problems, and investigates fundamental lower bounds on their performance. By exploring these lower bounds, we obtain interesting trade-offs in computation and accuracy. Moreover, we may hope to obtain improved sketching constructions by understanding their information-theoretic properties. The lower-bounding techniques employed here parallel the information-theoretic techniques used in statistical minimax theory [7, 8]. We apply Fano's inequality and packing constructions to understand fundamental lower bounds on the accuracy of sketching.

Randomness and sketching also have applications in privacy-preserving queries [9, 10]. Privacy has become an important concern in the age of information where breaches of sensitive data are frequent. We will illustrate that randomized sketching offers a computationally simple and effective mechanism to preserve privacy in optimization and machine learning.

We start with an overview of different constructions of sketching matrices in Section 4.2. In Section 4.3, we briefly review some background on convex analysis and optimization. Then we present upper bounds on the performance of sketching from an optimization viewpoint in Section 4.4. To be able to analyze upper bounds, we introduce the notion of *localized Gaussian complexity*, which also plays an important role in the characterization of minimax statistical bounds. In Section 4.5, we discuss information-theoretic lower bounds on the statistical performance of sketching. In Section 4.6, we turn to non-parametric problems and information-theoretic lower bounds. Finally, in Section 4.7 we discuss privacy-preserving properties of sketching using a mutual information characterization, and communication-complexity lower bounds.

4.2 Types of Randomized Sketches

In this section we describe popular constructions of sketching matrices. Given a sketching matrix $\mathbf{S}$, we use $\{\mathbf{s}_i\}_{i=1}^m$ to denote the collection of its n-dimensional rows. Here we consider sketches which are zero mean, and are normalized, i.e., they satisfy the following two conditions:

$$\text{(a)} \quad \mathbb{E}\mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{I}_{d\times d}, \tag{4.1}$$

$$\text{(b)} \quad \mathbb{E}\mathbf{S} = \mathbf{0}_{n\times d}. \tag{4.2}$$

The reasoning behind the above conditions will become clearer when they are applied to sketching optimization problems involving data matrices.

A very typical use of sketching is to obtain compressed versions of a large data matrix $\mathbf{A}$. We obtain the matrix $\mathbf{SA} \in \mathbb{R}^{m\times d}$ using simple matrix multiplication. See Fig. 4.1 for an illustration. As we will see in a variety of examples, random matrices preserve most of the information in the matrix $\mathbf{A}$.

4.2.1 Gaussian Sketches

The most classical sketch is based on a random matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ with i.i.d. standard Gaussian entries. Suppose that we generate a random matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ with entries drawn

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

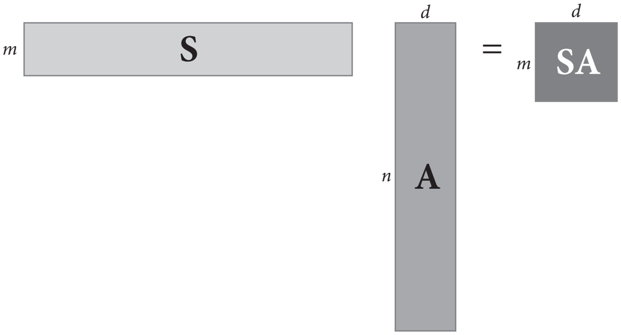

Figure 4.1 Sketching a tall matrix $\mathbf{A}$. The smaller matrix $\mathbf{SA} \in \mathbb{R}^{m \times d}$ is a compressed version of the original data $\mathbf{A} \in \mathbb{R}^{n \times d}$.

from i.i.d. zero-mean Gaussian random variables with variance $1/m$. Note that we have $\mathbb{E}\mathbf{S} = \mathbf{0}_{m \times d}$ and also $\mathbb{E}\mathbf{S}^{\mathrm{T}}\mathbf{S} = \sum_{i=1}^{m} \mathbb{E}\mathbf{s}_i\mathbf{s}_i^{\mathrm{T}} = \sum_{i=1}^{m} \mathbf{I}_d(1/m) = \mathbf{I}_d$. Analyzing the Gaussian sketches is considerably easier than analyzing sketches of other types, because of the special properties of the Gaussian distribution such as rotation invariance. However, Gaussian sketches may not be the most computationally efficient choice for many data matrices, as we will discuss in the following sections.

4.2.2 Sub-Gaussian Sketches

A generalization of the previous construction is a random sketch with rows, drawn from i.i.d. sub-Gaussian random variables. In particular, a zero-mean random vector $\mathbf{s} \in \mathbb{R}^n$ is 1-sub-Gaussian if, for any $\mathbf{u} \in \mathbb{R}^n$, we have

$$\mathbb{P}[\langle \mathbf{s}, \mathbf{u} \rangle \geq \varepsilon \|\mathbf{u}\|_2] \leq e^{-\varepsilon^2/2} \qquad \text{for all } \varepsilon \geq 0. \tag{4.3}$$

For instance, a vector with i.i.d. $\mathcal{N}(0,1)$ entries is 1-sub-Gaussian, as is a vector with i.i.d. Rademacher entries (uniformly distributed over $\{-1, +1\}$). In many models of computation, multiplying numbers by random signs is simpler than multiplying by Gaussian variables, and only costs an addition operation. Note that multiplying by -1 only amounts to flipping the sign bit in the signed number representation of the number in the binary system. In modern computers, the difference between addition and multiplication is often not appreciable. However, the real disadvantage of sub-Gaussian and Gaussian sketches is that they require matrix–vector multiplications with unstructured and dense random matrices. In particular, given a data matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, computing its sketched version $\mathbf{SA}$ requires $O(mnd)$ basic operations using classical matrix multiplication algorithms, in general.

4.2.3 Randomized Orthonormal Systems

The second type of randomized sketch we consider is the *randomized orthonormal system* (ROS), for which matrix multiplication can be performed much more efficiently.

In order to define an ROS sketch, we first let $\mathbf{H} \in \mathbb{R}^{n \times n}$ be an orthonormal matrix with entries $\mathbf{H}_{ij} \in [(-1/\sqrt{n}), (1/\sqrt{n})]$. Standard classes of such matrices are the Hadamard

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

or Fourier bases, for which matrix–vector multiplication can be performed in $O(n\log n)$ time via the fast Hadamard or Fourier transforms, respectively. For example, an $n \times n$ Hadamard matrix $\mathbf{H} = \mathbf{H}_n$ can be recursively constructed as follows:

$$\mathbf{H}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{H}_4 = \frac{1}{\sqrt{2}}\begin{bmatrix} \mathbf{H}_2 & \mathbf{H}_2 \\ \mathbf{H}_2 & -\mathbf{H}_2 \end{bmatrix}, \quad \mathbf{H}_{2^t} = \underbrace{\mathbf{H}_2 \otimes \mathbf{H}_2 \otimes \cdots \otimes \mathbf{H}_2}_{\text{Kronecker product } t \text{ times}}.$$

From any such matrix, a sketching matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ from a ROS ensemble can be obtained by sampling i.i.d. rows of the form

$$\mathbf{s}^{\mathrm{T}} = \sqrt{n}\mathbf{e}_j^{\mathrm{T}}\mathbf{H}\mathbf{D} \qquad \text{with probability } 1/n \text{ for } j = 1,\ldots,n,$$

where the random vector $\mathbf{e}_j \in \mathbb{R}^n$ is chosen uniformly at random from the set of all n canonical basis vectors, and $\mathbf{D} = \mathrm{diag}(\mathbf{r})$ is a diagonal matrix of i.i.d. Rademacher variables $\mathbf{r} \in \{-1,+1\}^n$, where $\mathbb{P}[\mathbf{r}_i = +1] = \mathbb{P}[\mathbf{r}_i = -1] = \frac{1}{2}\,\forall i$. Alternatively, the rows of the ROS sketch can be sampled without replacement and one can obtain similar guarantees to sampling with replacement. Given a fast routine for matrix–vector multiplication, ROS sketch $\mathbf{SA}$ of the data $\mathbf{A} \in \mathbb{R}^{n\times d}$ can be formed in $O(nd\log m)$ time (for instance, see [11]).

4.2.4 Sketches Based on Random Row Sampling

Given a probability distribution $\{\mathbf{p}_j\}_{j=1}^n$ over $[n] = \{1,\ldots,n\}$, another choice of sketch is to randomly sample the rows of the extended data matrix $\mathbf{A}$ a total of m times with replacement from the given probability distribution. Thus, the rows of S are independent and take on the values

$$\mathbf{s}^{\mathrm{T}} = \frac{\mathbf{e}_j^{\mathrm{T}}}{\sqrt{\mathbf{p}_j}} \qquad \text{with probability } \mathbf{p}_j \text{ for } j = 1,\ldots,n,$$

where $\mathbf{e}_j \in \mathbb{R}^n$ is the jth canonical basis vector. Different choices of the weights $\{\mathbf{p}_j\}_{j=1}^n$ are possible, including those based on the *leverage scores* of $\mathbf{A}$. Leverage scores are defined as

$$\mathbf{p}_j := \frac{\|\mathbf{u}_j\|_2^2}{\sum_{i=1}^n \|\mathbf{u}_i\|_2^2},$$

where $u_1, u_2, \ldots, u_n$ are the rows of $\mathbf{U} \in \mathbb{R}^{n\times d}$, which is the matrix of left singular vectors of $\mathbf{A}$. Leverage scores can be obtained using a singular value decomposition $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$. Moreover, there also exist faster randomized algorithms to approximate the leverage scores (e.g., see [12]). In our analysis of lower bounds to follow, we assume that the weights are α-balanced, meaning that

$$\max_{j=1,\ldots,n} \mathbf{p}_j \le \frac{\alpha}{n} \tag{4.4}$$

for some constant α that is independent of n.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

4.2.5 Graph Sparsification via Sub-Sampling

Let $G = (V, E)$ be a weighted, undirected graph with d nodes and n edges, where V and E are the set of nodes and the set of edges, respectively. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be the node–edge incidence matrix of the graph G. Suppose we randomly sample the edges in the graph a total of m times with replacement from a given probability distribution over the edges. The obtained graph is a weighted subgraph of the original, whose incidence matrix is $\mathbf{SA}$. Similarly to the row sampling sketch, the sketch can be written as

$$\mathbf{s}^{\mathrm{T}} = \frac{\mathbf{e}_j^{\mathrm{T}}}{\sqrt{\mathbf{p}_j}} \qquad \text{with probability } \mathbf{p}_j \text{ for } j = 1, \dots, n.$$

We note that row and graph sub-sampling sketches satisfy the condition (4.1). However, they do not satisfy the condition (4.2). In many computational problems on graphs, sparsifying the graph has computational advantages. Notable examples of such problems are solving Laplacian linear systems and graph partitioning, where sparsification can be used. We refer the reader to Spielman and Srivastava [13] for details.

4.2.6 Sparse Sketches Based on Hashing

In many applications, the data matrices contain very few non-zero entries. For sparse data matrices, special constructions of the sketching matrices yield greatly improved performance. Here we describe the count-sketch construction from [14, 15]. Let $h : [n] \to [m]$ be a hash functions from a pair-wise independent family.[1] The entry $\mathbf{S}_{ij}$ of the sketch matrix is given by σ_j if $i = h(j)$, and otherwise it is zero, where $\sigma \in \{-1, +1\}^n$ is a random vector containing 4-wise independent variables. Therefore, the jth column of $\mathbf{S}$ is non-zero only in the row indexed by $h(j)$. We refer the reader to [14, 15] for the details. An example realization of the sparse sketch is given below, where each column contains a single non-zero entry which is uniformly random sign ±1 at a uniformly random index:

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

Figure 4.2 shows examples of different randomized sketching matrices $S \in \mathbb{R}^{m\times n}$, where $m = 64, n = 1024$, which are drawn randomly. We refer readers to Nelson and Nguyên [16] for details on sparse sketches.

4.3 Background on Convex Analysis and Optimization

In this section, we first briefly review relevant concepts from convex analysis and optimization. A set $C \subseteq \mathbb{R}^d$ is convex if, for any $\mathbf{x}, \mathbf{y} \in C$,

$$t\mathbf{x} + (1 - t)\mathbf{y} \in C \text{ for all } t \in [0, 1].$$

[1] A hash function is from a pair-wise independent family if $\mathbb{P}[h(j) = i, h(k) = l] = 1/m^2$ and $\mathbb{P}[h(j) = i] = 1/m$ for all i, j, k, l.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

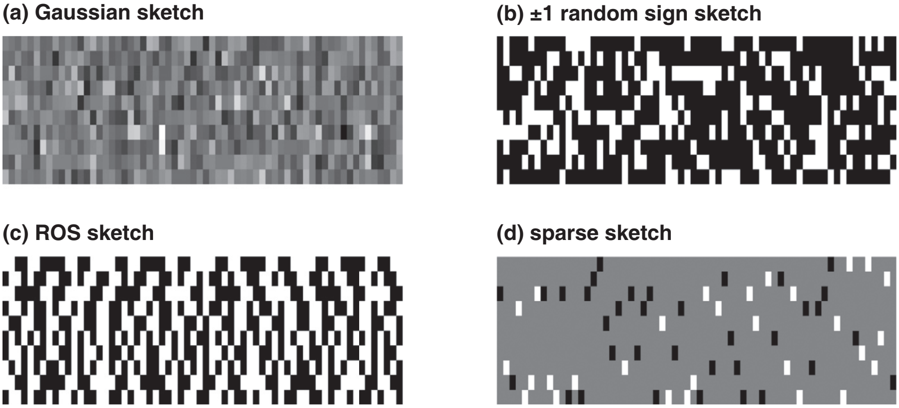

Figure 4.2 Different types of sketching matrices: (a) Gaussian sketch, (b) ±1 random sign sketch, (c) randomized orthogonal system sketch, and (d) sparse sketch.

Let $\mathcal{X}$ be a convex set. A function $f : \mathcal{X} \to \mathbb{R}$ is convex if, for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq t f(\mathbf{x}) + (1-t) f(\mathbf{y}) \text{ for all } t \in [0,1].$$

Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, we define the linear transform of the convex set C as $\mathbf{A}C = \{\mathbf{A}\mathbf{x} | \mathbf{x} \in C\}$. It can be shown that $\mathbf{A}C$ is convex if C is convex.

A convex optimization problem is a minimization problem of the form

$$\min_{\mathbf{x} \in C} f(\mathbf{x}), \tag{4.5}$$

where $f(\mathbf{x})$ is a convex function and C is a convex set. In order to characterize optimality of solutions, we will define the tangent cone of C at a fixed vector $\mathbf{x}^*$ as follows:

$$\mathcal{T}_C(\mathbf{x}^*) = \{t(\mathbf{x} - \mathbf{x}^*) \quad | \quad t \geq 0 \text{ and } \mathbf{x} \in C\}. \tag{4.6}$$

Figures 4.3 and 4.4 illustrate[2] examples of tangent cones of a polyhedral convex set in $\mathbb{R}^2$. A first-order characterization of optimality in the convex optimization problem (4.5) is given by the tangent cone. If a vector $\mathbf{x}^*$ is optimal in (4.5), it holds that

$$\mathbf{z}^{\mathrm{T}} \, \nabla f(\mathbf{x}^*) \geq 0, \forall \mathbf{z} \in \mathcal{T}_C(\mathbf{x}^*). \tag{4.7}$$

We refer the reader to Hiriart-Urruty and Lemaréchal [17] for details on convex analysis, and Boyd and Vandenberghe [18] for an in-depth discussion of convex optimization problems and applications.

[2] Note that the tangent cones extend toward infinity in certain directions, whereas the shaded regions in Figs. 4.3 and 4.4 are compact for illustration.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

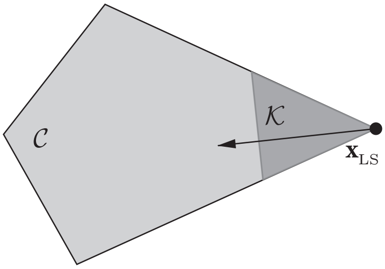

Figure 4.3 A narrow tangent cone where the Gaussian complexity is small.

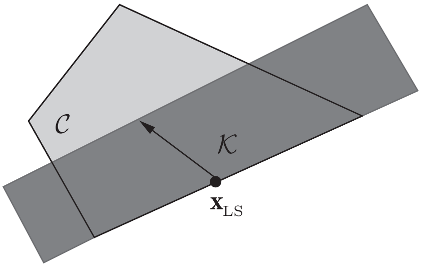

Figure 4.4 A wide tangent cone where the Gaussian complexity is large.

4.4 Sketching Upper Bounds for Regression Problems

Now we consider an instance of a convex optimization problem. Consider the least-squares optimization

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in C} \underbrace{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2}_{f(\mathbf{x})}, \tag{4.8}$$

where $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$ are the input data and $C \subseteq \mathbb{R}^d$ is a closed and convex constraint set. In statistical and signal-processing applications, it is typical to use the constraint set to impose structure on the obtained solution $\mathbf{x}$. Important examples of the convex constraint C include the non-negative orthant, ℓ_1-ball for promoting sparsity, and the ℓ_∞-ball as a relaxation to the combinatorial set $\{0,1\}^d$.

In the unconstrained case when $C = \mathbb{R}^d$, a closed-form solution exists for the solution of (4.8), which is given by $\mathbf{x}^* = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b}$. However, forming the Gram matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and inverting using direct methods such as QR decomposition, or the singular value decomposition, typically requires $O(nd^2) + O(nd\min(n,d))$ operations. Faster iterative algorithms such as the conjugate gradient (CG) method can be used to obtain an approximate solution in $O(nd\kappa(\mathbf{A}))$ time, where $\kappa(\mathbf{A})$ is the condition number of the data matrix $\mathbf{A}$. Using sketching methods, it is possible to obtain even faster approximate solutions, as we will discuss in what follows.

In the constrained case, a variety of efficient iterative algorithms have been developed in the last couple of decades to obtain the solution, such as proximal and projected gradient methods, their accelerated variants, and barrier-based second-order methods. Sketching can also be used to improve the run-time of these methods.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

4.4.1 Over-Determined Case ($n > d$)

In many applications, the number of observations, n, exceeds the number of unknowns, d, which gives rise to the tall $n \times d$ matrix A. In machine learning, it is very common to encounter datasets where n is very large and d is of moderate size. Suppose that we first compute the sketched data matrices $\mathbf{SA}$ and $\mathbf{Sb}$ from the original data $\mathbf{A}$ and $\mathbf{b}$, then consider the following approximation to the above optimization problem:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x} \in C} \|\mathbf{SAx} - \mathbf{Sb}\|_2^2 . \tag{4.9}$$

After applying the sketch to the data matrices, the sketched problem has dimensions $m \times d$, which is lower than the original dimensions when $m < n$. Note that the objective in the above problem (4.9) can be seen as an unbiased approximation of the original objective function (4.8), since it holds that

$$\mathbb{E}\,\|\mathbf{SAx} - \mathbf{Sb}\|_2^2 = \|\mathbf{Ax} - \mathbf{b}\|_2^2$$

for any fixed choice of $\mathbf{A}$, $\mathbf{x}$, and $\mathbf{b}$. This is a consequence of the condition (4.1), which is satisfied by all of the sketching matrices considered in Section 4.2.

4.4.2 Gaussian Complexity

Gaussian complexity plays an important role in statistics, empirical process theory, compressed sensing, and the theory of Banach spaces [19–21]. Here we consider a localized version of the Gaussian complexity, which is defined as follows:

$$\mathcal{W}_t(C) := \mathbb{E}_g\Big[\sup_{\substack{\mathbf{z} \in C \\ \|\mathbf{z}\|_2 \le t}} |\langle \mathbf{g}, \mathbf{z} \rangle|\Big], \tag{4.10}$$

where $\mathbf{g}$ is a random vector with i.i.d. standard Gaussian entries, i.e., $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. The parameter $t > 0$ controls the radius at which the random deviations are localized. For a finite value of t, the supremum in (4.10) is always achieved since the constraint set is compact.

Analyzing the sketched optimization problem requires us to control the random deviations constrained to the set of possible descent directions $\{\mathbf{x} - \mathbf{x}^* \mid \mathbf{x} \in C\}$. We now define a transformed tangent cone at x^* as follows:

$$\mathcal{K} = \{t\mathbf{A}(\mathbf{x} - \mathbf{x}^*) \quad | \quad t \ge 0 \text{ and } \mathbf{x} \in C\},$$

which can be alternatively defined as $A\mathcal{T}_C(\mathbf{x}^*)$ using the definition given in (4.6). The next theorem provides an upper bound on the performance of the sketching method for constrained optimization based on localized Gaussian complexity.

THEOREM 4.1 *Let S be a Gaussian sketch, and let $\widehat{\mathbf{x}}$ be the solution of* (4.9). *Suppose that $m \ge c_0 \mathcal{W}_1(\mathcal{K})^2/\epsilon^2$, where c_0 is a universal constant, then it holds that*

$$\frac{\|A(\widehat{\mathbf{x}} - \mathbf{x}^*)\|_2}{f(\mathbf{x}^*)} \le \epsilon ,$$

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

and consequently we have

$$f(\mathbf{x}^*) \le f(\widehat{\mathbf{x}}) \le f(\mathbf{x}^*)(1+\epsilon)\,. \tag{4.11}$$

As predicted by the theorem, the approximation ratio improves as the sketch dimension m increases, and converges to one as $m \to \infty$. However, we are often interested in the rate of convergence of the approximation ratio. Theorem 4.1 characterizes this rate by relating the geometry of the constraint set to the accuracy of the sketching method (4.9). As an illustration, Figs. 4.3 and 4.4 show narrow and wide tangent cones in $\mathbb{R}^2$, respectively. The proof of Theorem 4.1 combines the convex optimality condition involving the tangent cone in (4.7) with results on empirical processes, and can be found in Pilanci and Wainwright [22]. An important feature of Theorem 4.1 is that the approximation quality is relative to the optimal value $f(\mathbf{x}^*)$. This is advantageous when $f(\mathbf{x}^*)$ is small, e.g., the optimal value can be zero in noiseless signal recovery problems. However, in problems where the signal-to-noise ratio is low, $f(\mathbf{x}^*)$ can be large, and hence negatively affects the approximation quality. We illustrate the implications of Theorem 4.1 on some concrete examples in what follows.

Example 4.1 Unconstrained Least Squares For unconstrained problems, we have $\mathcal{K} = \mathrm{range}(\mathbf{A})$, i.e., the tangent cone is equal to the range space of the data matrix $\mathbf{A}$. In order to apply Theorem 4.1, we need the following lemma about the Gaussian complexity of a subspace.

LEMMA 4.1 *Let Q be a subspace of dimension q. The Gaussian complexity of Q satisfies*

$$\mathcal{W}_t(Q) \le t\sqrt{q}\,.$$

Proof Let $\mathbf{U}$ be an orthonormal basis for the subspace Q. We have the following representation: $\mathbf{L} = \{\mathbf{U}\mathbf{x} \,|\, \mathbf{x} \in \mathbb{R}^q\}$. Consequently the Gaussian complexity $\mathcal{W}_1(Q)$ can be written as

$$\begin{aligned}\mathbb{E}_{\mathbf{g}}\Big[\sup_{\substack{\mathbf{x}\\ \|\mathbf{U}\mathbf{x}\|_2\le \mathbf{t}}} \langle \mathbf{g}, \mathbf{U}\mathbf{x}\rangle\Big] = \mathbb{E}_{\mathbf{g}}\Big[\sup_{\substack{\mathbf{x}\\ \|\mathbf{x}\|_2\le t}} \langle \mathbf{U}^{\mathrm{T}}\mathbf{g}, \mathbf{x}\rangle\Big] = t\,\mathbb{E}_{\mathbf{g}}\|\mathbf{U}^{\mathrm{T}}\mathbf{g}\|_2 &\le t\sqrt{\mathbb{E}\ \mathrm{tr}\ \mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{g}\mathbf{g}^{\mathrm{T}}}\\ &= t\sqrt{\mathrm{tr}\ \mathbf{U}^{\mathrm{TU}}}\\ &= t\sqrt{q}.\end{aligned}$$

Where the inequality follows from Jensen's inequality and concavity of the square root, and first and fifth equality follow since $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_q$. Therefore, the Gaussian complexity of the range of $\mathbf{A}$ for $t = 1$ satisfies

$$\mathcal{W}_1(\mathrm{range}(\mathbf{A})) \le \sqrt{\mathrm{rank}(\mathbf{A})}\,.$$

Setting the dimension of the sketch $m \ge c_0\,\mathrm{rank}(\mathbf{A})/\epsilon^2$ suffices to obtain an ϵ approximate solution in the sense of (4.11). We note that $\mathrm{rank}(\mathbf{A})$ might not be known *a priori*, but the upper bound $\mathrm{rank}(\mathbf{A}) \le d$ may be useful when $n \gg d$.

Example 4.2 ℓ_1 Constrained Least Squares For ℓ_1-norm-constrained problems we have $C = \{\mathbf{x} \mid \|\mathbf{x}\|_1 \leq r\}$ for some radius parameter r. The tangent cone $\mathcal{K}$ at the optimal point $\mathbf{x}^*$ depends on the support[3] of $\mathbf{x}^*$, and hence on its cardinality $\|\mathbf{x}^*\|_0$. In [23], it is shown that the localized Gaussian complexity satisfies

$$\mathcal{W}_1(\mathcal{K}) \leq c_1 \frac{\gamma_{k_+}(\mathbf{A})}{\beta_{k_-}(\mathbf{A})} \sqrt{\|\mathbf{x}^*\|_0 \log d},$$

where c_1 is a universal constant and γ_k are β_k are ℓ_1-restricted maximum and minimum eigenvalues defined as follows:

$$\gamma_k := \max_{\substack{\|\mathbf{z}\|=1 \\ \|\mathbf{z}\|_1 \leq \sqrt{k}}} \|\mathbf{A}\mathbf{z}\|_2^2 \quad \text{and} \quad \beta_k := \min_{\substack{\|\mathbf{z}\|=1 \\ \|\mathbf{z}\|_1 \leq \sqrt{k}}} \|\mathbf{A}\mathbf{z}\|_2^2 .$$

As a result, we conclude that for ℓ_1-constrained problems, the sketch dimension can be substantially smaller when ℓ_1-constrained eigenvalues are well behaved.

4.4.3 Under-Determined Case ($n \leq d$)

In many applications the dimension of the data vectors may be larger than the sample size. In these situations, it makes sense to reduce the dimensionality by applying the sketch on the right, i.e., $\mathbf{A}\mathbf{S}^{\mathrm{T}}$, and solve

$$\arg \min_{\substack{\mathbf{z} \in \mathbb{R}^m \\ \mathbf{S}^{\mathrm{T}}\mathbf{z} \in C}} \|(\mathbf{A}\mathbf{S}^{\mathrm{T}}\mathbf{z} - \mathbf{b})\|_2 . \tag{4.12}$$

Note that the vector $\mathbf{z} \in \mathbb{R}^m$ is of smaller dimension than the original variable $\mathbf{x} \in \mathbb{R}^d$. After solving the reduced-dimensional problem and obtaining its optimal solution $\mathbf{z}^*$, the final estimate for the original variable $\mathbf{x}$ can be taken as $\widehat{\mathbf{x}} = \mathbf{S}^{\mathrm{T}}\mathbf{z}^*$. We will investigate this approach in Section 4.5 in non-parametric statistical estimation problems and present concrete theoretical guarantees.

It is instructive to note that, in the special case where we have ℓ_2 regularization and $C = \mathbb{R}^d$, we can easily transform the under-determined least-squares problem into an over-determined one using convex duality, or the matrix-inversion lemma. We first write the sketched problem (4.12) as the constrained convex program

$$\min_{\substack{\mathbf{z} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n \\ \mathbf{y} = \mathbf{A}\mathbf{S}^{\mathrm{T}}\mathbf{z}}} \frac{1}{2}\|\mathbf{y} - \mathbf{b}\|_2^2 + \rho\|\mathbf{z}\|_2^2 ,$$

and form the convex dual. It can be shown that strong duality holds, and consequently primal and dual programs can be stated as follows:

$$\min_{\mathbf{z} \in \mathbb{R}^m} \frac{1}{2}\|\mathbf{A}\mathbf{S}^{\mathrm{T}}z - \mathbf{b}\|_2^2 + \rho\|\mathbf{z}\|_2^2 = \max_{\mathbf{x} \in \mathbb{R}^d} -\frac{1}{4\rho}\|\mathbf{S}\mathbf{A}^{\mathrm{T}}\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{x}\|_2^2 + \mathbf{x}^{\mathrm{T}}\mathbf{b} ,$$

[3] The term support refers to the set of indices where the solution has a non-zero value.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

where the primal and dual solutions satisfy $\mathbf{z}^* = (1/2\rho)\mathbf{S}\mathbf{A}^{\mathrm{T}}\mathbf{x}^*$ at the optimum [18]. Therefore the sketching matrix applied from the right, $\mathbf{A}\mathbf{S}^{\mathrm{T}}$, corresponds to a sketch applied on the left, $\mathbf{S}\mathbf{A}^{\mathrm{T}}$, in the dual problem which parallels (4.9). This observation can be used to derive approximation results on the dual program. We refer the reader to [22] for an application in support vector machine classification where $\mathbf{b} = \mathbf{0}_n$.

4.5 Information-Theoretic Lower Bounds

4.5.1 Statistical Upper and Lower Bounds

In order to develop information-theoretic lower bounds, we consider a statistical observation model for the constrained regression problem. Consider the following model:

$$\mathbf{b} = \mathbf{A}\mathbf{x}^\dagger + \mathbf{w}, \quad \text{where } \mathbf{w} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n) \text{ and } \mathbf{x}^\dagger \in \mathcal{C}_0, \tag{4.13}$$

where $\mathbf{x}^\dagger$ is the unknown vector to be estimated and $\mathbf{w}$ is an i.i.d. noise vector whose entries are distributed as $\mathcal{N}(0,\sigma^2)$. In this section we will focus on the observation model (4.13) and present a lower bound on all estimators which use the sketched data $(\mathbf{S}\mathbf{A}, \mathbf{S}\mathbf{b})$ to form an estimate $\widehat{\mathbf{x}}$.

We assume that the unknown vector $\mathbf{x}^\dagger$ belongs to some set $\mathcal{C}_0 \subseteq \mathcal{C}$ that is star-shaped around zero.[4] In many cases of interest we have $\mathcal{C} = \mathcal{C}_0$, i.e., when the set $\mathcal{C}$ is convex and simple to describe. In this case, the constrained least-squares estimate $\mathbf{x}^*$ from equation (4.8) corresponds to the constrained maximum-likelihood estimator for estimating the unknown regression vector $\mathbf{x}^\dagger$ under the Gaussian observation model (4.13). However $\mathcal{C}_0$ may not be computationally tractable as an optimization constraint set, such as a non-convex set, and we can consider a set $\mathcal{C}$ which is a convex relaxation[5] of this set, such that $\mathcal{C} \subset \mathcal{C}_0$. An important example is the set of s sparse and bounded vectors given by $\mathcal{C}_0 = \{\mathbf{x} \,:\, \|\mathbf{x}\|_0 \le s,\ \|\mathbf{x}\|_\infty \le 1\}$, which has combinatorially many elements. The well-known ℓ_1 relaxation given by $\mathcal{C} = \{\mathbf{x} \,:\, \|\mathbf{x}\|_1 \le \sqrt{s},\ \|\mathbf{x}\|_\infty \le 1\}$ satisfies $\mathcal{C} \subset \mathcal{C}_0$, which follows from the Cauchy–Schwartz inequality, and is widely used [24, 25] to find sparse solutions.

We now present a theoretical result on the statistical performance of the original constrained least-squares estimator in (4.8)

THEOREM 4.2 *Let $\mathcal{C}$ be any set that contains the true parameter $\mathbf{x}^\dagger$. Then the constrained estimator $\mathbf{x}^*$ in* (4.8) *under the observation model* (4.13) *has mean-squared error upper-bounded as*

$$\mathbb{E}_w\left[\frac{1}{n}\|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^\dagger)\|_2^2\right] \le c_2\Big(\delta^*(n)^2 + \frac{\sigma^2}{n}\Big),$$

[4] This assumption means that, for any $x \in \mathcal{C}_0$ and scalar $t \in [0,1]$, the point tx also belongs to $\mathcal{C}_0$.

[5] We may also consider an approximation of $\mathcal{C}_0$ which doesn't necessarily satisfy $\mathcal{C} \subset \mathcal{C}_0$, for example, the ℓ_1 and ℓ_0 unit balls.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

where $\delta^*(n)$ is the critical radius, equal to the smallest positive solution $\delta > 0$ to the inequality

$$\frac{\mathcal{W}_\delta(C)}{\delta\sqrt{n}} \leq \frac{\delta}{\sigma}. \tag{4.14}$$

We refer the reader to [20, 23] for a proof of this theorem. This result provides a baseline against which to compare the statistical recovery performance of the randomized sketching method. In particular, an important goal is characterizing the minimal projection dimension m that will enable us to find an estimate $\widehat{\mathbf{x}}$ with the error guarantee

$$(1/n)\|\mathbf{A}(\widehat{\mathbf{x}} - \mathbf{x}^\dagger)\|_2^2 \approx (1/n)\|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^\dagger)\|_2^2,$$

in a computationally simpler manner using the compressed data $\mathbf{SA}, \mathbf{Sb}$.

An application of Theorem 4.1 will yield that the sketched solution $\widehat{\mathbf{x}}$ in (4.9), using the choice of sketch dimension $m = c_0\mathcal{W}_1(\mathcal{K})^2/\epsilon^2$, satisfies the bound

$$\|\mathbf{A}(\widehat{\mathbf{x}} - \mathbf{x}^*)\| \leq \epsilon\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2,$$

where $\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2 = f(\mathbf{x}^*)$ is the optimal value of the optimization problem (4.8). However, under the model (4.13) we have

$$\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2 = \|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^\dagger) - \mathbf{w}\|_2 \leq \|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^\dagger)\|_2 + \|\mathbf{w}\|_2,$$

which is at least $O(\sigma\sqrt{n})$ because of the term $\|\mathbf{w}\|_2$. This upper bound suggests that $(1/n)\|\mathbf{A}(\widehat{\mathbf{x}} - \mathbf{x}^\dagger)\|_2^2$ is bounded by $O(\epsilon^2\sigma^2) = O(\sigma^2\mathcal{W}_1(\mathcal{K})^2/m)$. This can be considered as a negative result for the sketching method, since the error scales as $O(1/m)$ instead of $O(1/n)$. We will show that this upper bound is tight, and the $O(1/m)$ scaling is unavoidable for all methods that sketch the data once. In contrast, as we will discuss in Section 4.5.7, an iterative sketching method can achieve optimal prediction error using sketches of comparable dimension.

We will in fact show that, unless $m \geq n$, *any method* based on observing *only* the pair $(\mathbf{SA}, \mathbf{Sb})$ necessarily has a substantially larger error than the least-squares estimate. In particular, our result applies to an arbitrary measurable function $(\mathbf{SA}, \mathbf{Sb}) \mapsto \widehat{\mathbf{x}}$, which we refer to as an *estimator*.

More precisely, our lower bound applies to any random matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ for which

$$|||\mathbb{E}[\mathbf{S}^{\mathrm{T}}(\mathbf{S}\mathbf{S}^{\mathrm{T}})^{-1}\mathbf{S}]|||_{\mathrm{op}} \leq \eta\,\frac{m}{n}, \tag{4.15}$$

where η is a constant that is independent of n and m, and $|||\mathbf{A}|||_{\mathrm{op}}$ denotes the ℓ_2-operator norm, which reduces to the maximum eigenvalue for a symmetric matrix. These conditions hold for various standard choices of the sketching matrix, including most of those discussed in the Section 4.2: the Gaussian sketch, the ROS sketch,[6] the sparse sketch, and the α-balanced leverage sampling sketch. The following lemma shows that the condition (4.15) is satisfied for Gaussian sketches with equality and $\eta = 1$.

[6] See [23] for a proof of this fact for Gaussian and ROS sketches. To be more precise, for ROS sketches, the condition (4.15) holds when rows are sampled without replacement.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

LEMMA 4.2 *Let* $\mathbf{S} \in \mathbb{R}^{m\times n}$ *be a random matrix with i.i.d. Gaussian entries. We have*

$$\left\|\mathbb{E}\left[\mathbf{S}^{\mathrm{T}}(\mathbf{S}\mathbf{S}^{\mathrm{T}})^{-1}\mathbf{S}\right]\right\|_{\mathrm{op}} = \frac{m}{n}.$$

Proof Let $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ denote the singular value decomposition of the random matrix $\mathbf{S}$. Note that we have $\mathbf{S}^{\mathrm{T}}(\mathbf{S}\mathbf{S}^{\mathrm{T}})^{-1}\mathbf{S} = \mathbf{V}\mathbf{V}^{\mathrm{T}}$. By virtue of the rotation invariance of the Gaussian distribution, columns of $\mathbf{V}$ denoted by $\{\mathbf{v}_i\}_{i=1}^m$ are uniformly distributed over the n-dimensional unit sphere, and it holds that $\mathbb{E}\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}} = (1/n)\mathbf{I}_n$ for $i = 1,\ldots,m$. Consequently, we obtain

$$\mathbb{E}\left[\mathbf{S}^{\mathrm{T}}(\mathbf{S}\mathbf{S}^{\mathrm{T}})^{-1}\mathbf{S}\right] = \mathbb{E}\sum_{i=1}^{m}\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}} = m\mathbb{E}\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}} = \frac{m}{n}\mathbf{I}_n,$$

and the bound on the operator norm follows.

4.5.2 Fano's Inequality

Let X and Y represent two random variables with a joint probability distribution $P_{x,y}$, where X is discrete and takes values from a finite set $\mathcal{X}$. Let $\widehat{X} = g(Y)$ be the predicted value of X for some deterministic function g which also takes values in $\mathcal{X}$. Then Fano's inequality states that

$$P\left[X \neq \widehat{X}\right] \geq \frac{H(X|Y) - 1}{\log_2(|\mathcal{X}| - 1)}.$$

Fano's inequality follows as a simple consequence of the chain rule for entropy. However, it is very powerful for deriving lower bounds on the error probabilities in coding theory, statistics, and machine learning [7, 26–30].

4.5.3 Metric Entropy

For a given positive tolerance value $\delta > 0$, we define the δ-packing number $M_{\delta,\|\cdot\|}$ of a set $C \subseteq \mathbb{R}^d$ with respect to a norm $\|\cdot\|$ as the largest number of vectors $\{\mathbf{x}^j\}_{j=1}^M \subseteq C$ which are elements of C and satisfy

$$\|\mathbf{x}^k - \mathbf{x}^l\| > \delta \;\; \forall k \neq l.$$

We define the *metric entropy* of the set C with respect to a norm $\|\cdot\|$ as the logarithm of the corresponding packing number

$$N_{\delta,\|\cdot\|}(C) = \log_2 M_{\delta,\|\cdot\|}.$$

The concept of metric entropy provides a way to measure the complexity, or effective size, of a set with infinitely many elements and dates back to the seminal work of Kolmogorov and Tikhomirov [31].

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

4.5.4 Minimax Risk

In this chapter, we will take a frequentist approach in modeling the unknown vector $\mathbf{x}^\dagger$ we are trying to estimate from the data. In order to assess the quality of estimation, we will consider a risk function associated with our estimation method. Note that, for a fixed value of the unknown vector $\mathbf{x}^\dagger$, there exist estimators which make no error for that particular vector $\mathbf{x}^\dagger$, such as the estimator which always returns $\mathbf{x}^\dagger$ regardless of the observation. We will take the worst-case risk approach considered in the statistical estimation literature, which focuses on the *minimax* risk. More precisely, we define the minimax risk as follows:

$$\mathcal{M}(Q) = \inf_{\widehat{\mathbf{x}}\in Q} \sup_{\mathbf{x}^\dagger\in\mathcal{X}} \mathbb{E}\left[\frac{1}{n}\|\mathbf{A}(\widehat{\mathbf{x}} - \mathbf{x}^\dagger)\|_2^2\right], \tag{4.16}$$

where the infimum ranges over all estimators that use the input data $\mathbf{A}$ and $\mathbf{b}$ to estimate $\mathbf{x}^\dagger$.

4.5.5 Reduction to Hypothesis Testing

In this section we present a reduction of the minimax estimation risk to hypothesis testing. Suppose that we have a packing of the constraint set C given by the collection $\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(M)}$ with radius 2δ. More precisely, we have

$$\|A(\mathbf{z}^{(i)} - \mathbf{z}^{(j)})\|_2 \geq 2\delta \quad \forall i \neq j,$$

where $\mathbf{z}^{(i)} \in C$ for all $i = 1,\ldots,M$. Next, consider a set of probability distributions $\{P_{\mathbf{z}^{(j)}}\}_{j=1}^M$ corresponding to the distribution of the observation when the unknown vector is $\mathbf{x}^\dagger = \mathbf{z}^j$. Suppose that we have an M-ary hypothesis-testing problem constructed as follows. Let J_δ denote a random variable with uniform distribution over the index set $\{1,\ldots,M\}$ that allows us to pick an element of the packing set at random. Note that M is a function of δ, hence we keep the dependence of J_δ on δ explicit in our notation. Let us set the random variable Z according to the probability distribution $P_{\mathbf{z}^{(j)}}$ in the event that $J_\delta = j$, i.e.,

$$Z \sim \mathbb{P}_{\mathbf{x}^{(j)}} \quad \text{whenever} \quad J_\delta = j.$$

Now we will consider the problem of detecting the index set given the value of Z. The next lemma is a standard reduction in minimax theory, and relates the minimax estimation risk to the M-ary hypothesis-testing error (see Birgé [30] and Yu [7]).

LEMMA 4.3 *The minimax risk Q is lower-bounded by*

$$\mathcal{M}(Q) \geq \delta^2 \inf_{\psi} \mathbb{P}[\psi(Z) \neq J_\delta]. \tag{4.17}$$

A proof of this lemma can be found in Section A4.2. Lemma 4.3 allows us to apply Fano's method after transforming the estimation problem into a hypothesis-testing problem based on sketched data. Let us recall the condition on sketching matrices stated earlier,

$$\|\mathbb{E}[\mathbf{S}^{\mathrm{T}}(\mathbf{S}\mathbf{S}^{\mathrm{T}})^{-1}\mathbf{S}]\|_{\mathrm{op}} \leq \eta\,\frac{m}{n}, \tag{4.18}$$

where η is a constant that is independent of n and m. Now we are ready to present the lower bound on the statistical performance of sketching.

THEOREM 4.3 *For any random sketching matrix* $\mathbf{S} \in \mathbb{R}^{m\times n}$ *satisfying condition* (4.18), *any estimator* $(\mathbf{SA}, \mathbf{Sb}) \mapsto \mathbf{x}^\dagger$ *has MSE lower-bounded as*

$$\sup_{\mathbf{x}^\dagger \in C_0} \mathbb{E}_{\mathbf{S},\mathbf{w}}\left[\frac{1}{n}\|\mathbf{A}(\mathbf{x}^\dagger - \mathbf{x}^*)\|_2^2\right] \geq \frac{\sigma^2}{128\eta}\,\frac{\log_2(\frac{1}{2}M_{1/2})}{\min\{m,n\}}, \tag{4.19}$$

where $M_{1/2}$ *is the* $1/2$*-packing number of* $C_0 \cap \mathbb{B}_A(1)$ *in the semi-norm* $(1/\sqrt{n})\|\mathbf{A}(\,\cdot\,)\|_2$.

We defer the proof to Section 4.8, and investigate the implications of the lower bound in the next section. It can be shown that Theorem 4.3 is tight, since Theorem 4.1 provides a matching upper bound.

4.5.6 Implications of the Information-Theoretic Lower Bound

We now investigate some consequences of the lower bound given in Theorem 4.3. We will focus on concrete examples of popular statistical estimation and optimization problems to illustrate its applicability.

Example 4.3 Unconstrained Least Squares We first consider the simple unconstrained case, where the constraint is the entire d-dimensional space, i.e., $C = \mathbb{R}^d$. With this choice, it is well known that, under the observation model (4.13), the least-squares solution $\mathbf{x}^*$ has prediction mean-squared error upper-bounded as follows:[7]

$$\mathbb{E}\left[\frac{1}{n}\|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^\dagger)\|_2^2\right] \lesssim \frac{\sigma^2 \operatorname{rank}(\mathbf{A})}{n} \tag{4.20a}$$

$$\leq \frac{\sigma^2 d}{n}, \tag{4.20b}$$

where the expectation is over the noise variable w in (4.13). On the other hand, with the choice $C_0 = \mathbb{B}_2(1)$, it is well known that we can construct a $1/2$-packing with $M = 2^d$ elements, so that Theorem 4.3 implies that any estimator $\mathbf{x}^\dagger$ based on $(\mathbf{SA}, \mathbf{Sb})$ has prediction MSE lower-bounded as

$$\mathbb{E}_{\mathbf{S},\mathbf{w}}\left[\frac{1}{n}\|\mathbf{A}(\widehat{\mathbf{x}} - \mathbf{x}^\dagger)\|_2^2\right] \gtrsim \frac{\sigma^2 d}{\min\{m,n\}}. \tag{4.20c}$$

Consequently, the sketch dimension m must grow proportionally to n in order for the sketched solution to have a mean-squared error comparable to the original least-squares estimate. This may not be desirable for least-squares problems in which $n \gg d$, since it should be possible to sketch down to a dimension proportional to $\operatorname{rank}(\mathbf{A})$ which is always upper-bounded by d. Thus, Theorem 4.3 reveals a surprising gap between the classical least-squares sketch (4.9) and the accuracy of the original least-squares estimate. In the regime $n \gg m$, the prediction MSE of the sketched solution is $O(\sigma^2(d/m))$

[7] In fact, a closed-form solution exists for the prediction error, which it is straightforward to obtain from the closed-form solution of the least-squares estimator. However, this simple form is sufficient to illustrate information-theoretic lower bounds.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

which is a factor of n/m larger than the optimal prediction MSE in (4.20b). In Section 4.5.7, we will see that this gap can be removed by iterative sketching algorithms which don't obey the information-theoretic lower bound (4.20c).

Example 4.4 ℓ_1 Constrained Least Squares We can consider other forms of constrained least-squares estimates as well, such as those involving an ℓ_1-norm constraint to encourage sparsity in the solution. We now consider the sparse variant of the linear regression problem, which involves the ℓ_0 "ball"

$$\mathbb{B}_0(s) := \left\{ \mathbf{x} \in \mathbb{R}^d \mid \sum_{j=1}^{d} \mathbb{I}[\mathbf{x}_j \neq 0] \leq s \right\},$$

corresponding to the set of all vectors with at most s non-zero entries. Fixing some radius $R \geq \sqrt{s}$, consider a vector $\mathbf{x}^\dagger \in C_0 := \mathbb{B}_0(s) \cap \{\|\mathbf{x}\|_1 = R\}$, and suppose that we have noisy observations of the form $\mathbf{b} = \mathbf{A}\mathbf{x}^\dagger + \mathbf{w}$.

Given this setup, one way in which to estimate $\mathbf{x}^\dagger$ is by computing the least-squares estimate $\mathbf{x}^*$ constrained to the ℓ_1-ball $C = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_1 \leq R\}$.[8] This estimator is a form of the Lasso [2, 32] which has been studied extensively in the context of statistical estimation and signal reconstruction.

On the other hand, the 1/2-packing number M of the set C_0 can be lower-bounded as $\log_2 M \gtrsim s \log_2(ed/s)$. We refer the reader to [33] for a proof. Consequently, in application to this particular problem, Theorem 4.3 implies that any estimator $\widehat{\mathbf{x}}$ based on the pair $(\mathbf{SA}, \mathbf{Sb})$ has mean-squared error lower-bounded as

$$\mathbb{E}_{\mathbf{w},\mathbf{S}}\left[\frac{1}{n}\|\mathbf{A}(\widehat{\mathbf{x}} - \mathbf{x}^\dagger)\|_2^2\right] \gtrsim \frac{\sigma^2 s \log_2(ed/s)}{\min\{m,n\}}. \tag{4.21}$$

Again, we see that the projection dimension m must be of the order of n in order to match the mean-squared error of the constrained least-squares estimate $\mathbf{x}^*$ up to constant factors.

Example 4.5 Low-Rank Matrix Estimation In the problem of multivariate regression, the goal is to estimate a matrix $\mathbf{X}^\dagger \in \mathbb{R}^{d_1 \times d_2}$ model based on observations of the form

$$\mathbf{Y} = \mathbf{A}\mathbf{X}^\dagger + \mathbf{W}, \tag{4.22}$$

where $\mathbf{Y} \in \mathbb{R}^{n \times d_2}$ is a matrix of observed responses, $\mathbf{A} \in \mathbb{R}^{n \times d_1}$ is a data matrix, and $\mathbf{W} \in \mathbb{R}^{n \times d_2}$ is a matrix of noise variables. A typical interpretation of this model is a collection of d_2 regression problems, where each one involves a d_1-dimensional regression vector, namely a particular column of the matrix $\mathbf{X}^\dagger$. In many applications, including reduced-rank regression, multi-task learning, and recommender systems

[8] This setup is slightly unrealistic, since the estimator is assumed to know the radius $R = \|\mathbf{x}^\dagger\|_1$. In practice, one solves the least-squares problem with a Lagrangian constraint, but the underlying arguments are essentially the same.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

(e.g., [34–37]), it is reasonable to model the matrix $\mathbf{X}^{\dagger}$ as being a low-rank matrix. Note that a rank constraint on the matrix $\mathbf{X}$ can be written as an ℓ_0-"norm" sparsity constraint on its singular values. In particular, we have

$$\operatorname{rank}(\mathbf{X}) \le r \quad \text{if and only if} \quad \sum_{j=1}^{\min\{d_1,d_2\}} \mathbb{I}[\gamma_j(\mathbf{X}) > 0] \le r,$$

where $\gamma_j(\mathbf{X})$ denotes the jth singular value of $\mathbf{X}$. This observation motivates a standard relaxation of the rank constraint using the nuclear norm $\|\mathbf{X}\|_{\text{nuc}} := \sum_{j=1}^{\min\{d_1,d_2\}} \gamma_j(\mathbf{X})$.

Accordingly, let us consider the constrained least-squares problem

$$\mathbf{X}^* = \arg \min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \left\{ \frac{1}{2}\|\mathbf{Y}-\mathbf{A}\mathbf{X}\|_{\text{fro}}^2 \right\} \qquad \text{such that } \|\mathbf{X}\|_{\text{nuc}} \le R, \tag{4.23}$$

where $\|\cdot\|_{\text{fro}}$ denotes the Frobenius norm on matrices, or equivalently the Euclidean norm on its vectorized version. Let C_0 denote the set of matrices with rank $r < \frac{1}{2}\min\{d_1,d_2\}$, and Frobenius norm at most one. In this case the constrained least-squares solution $\mathbf{X}^*$ satisfies the bound

$$\mathbb{E}\left[\frac{1}{n}\|\mathbf{A}(\mathbf{X}^* - \mathbf{X}^{\dagger})\|_2^2\right] \lesssim \frac{\sigma^2 r(d_1+d_2)}{n}. \tag{4.24a}$$

On the other hand, the $1/2$-packing number of the set C_0 is lower-bounded as $\log_2 M \gtrsim r(d_1+d_2)$ (see [36] for a proof), so that Theorem 4.3 implies that any estimator $\widehat{\mathbf{X}}$ based on the pair $(\mathbf{SA},\mathbf{SY})$ has MSE lower-bounded as

$$\mathbb{E}_{\mathbf{w},\mathbf{S}}\left[\frac{1}{n}\|\mathbf{A}(\widehat{\mathbf{X}} - \mathbf{X}^{\dagger})\|_2^2\right] \gtrsim \frac{\sigma^2 r(d_1+d_2)}{\min\{m,n\}}. \tag{4.24b}$$

As with the previous examples, we see the sub-optimality of the sketched approach in the regime $m < n$.

4.5.7 Iterative Sketching

It is possible to improve the basic sketching estimator using adaptive measurements. Consider the constrained least-squares problem in (4.8):

$$x^* = \arg\min_{\mathbf{x}\in C} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \tag{4.25}$$

$$= \arg\min_{\mathbf{x}\in C} \underbrace{\frac{1}{2}\|\mathbf{A}\mathbf{x}\|_2^2 - \mathbf{b}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \frac{1}{2}\|\mathbf{b}\|_2^2}_{f(\mathbf{x})}. \tag{4.26}$$

We may use an iterative method to obtain $\mathbf{x}^*$ which uses the gradient $\nabla f(\mathbf{x}) = \mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{x}-\mathbf{b})$ and Hessian $\nabla^2 f(\mathbf{x}) = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ to minimize the second-order Taylor expansion of $f(\mathbf{x})$ at a current iterate $\mathbf{x}_t$ using $\nabla f(\mathbf{x}_t)$ and $\nabla^2 f(\mathbf{x}_t)$ as follows:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \arg\min_{\mathbf{x}\in C} \left\|\left[\nabla^2 f(\mathbf{x})\right]^{1/2}\mathbf{x}\right\|_2^2 + \mathbf{x}^{\mathrm{T}}\,\nabla f(\mathbf{x}_t) \tag{4.27}$$

$$= \mathbf{x}_t + \arg\min_{\mathbf{x}\in C} \|\mathbf{A}\mathbf{x}\|_2^2 - \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{b}-\mathbf{A}\mathbf{x}_t). \tag{4.28}$$

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

We apply a sketching matrix **S** to the data **A** on the formulation (4.28) and define this procedure as an *iterative sketch*

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \arg\min_{\mathbf{x}\in C} \|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 - 2\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{b} - \mathbf{A}\mathbf{x}_t)\,. \tag{4.29}$$

Note that this procedure uses more information then the classical sketch (4.9), in particular it calculates the left matrix–vector multiplications with the data A in the following order:

$$\begin{gathered}\mathbf{s}_1^{\mathrm{T}}\mathbf{A}\\ \mathbf{s}_2^{\mathrm{T}}\mathbf{A}\\ \vdots\\ \mathbf{s}_m^{\mathrm{T}}\mathbf{A}\\ \vdots\\ (\mathbf{b} - \mathbf{A}\mathbf{x}_1)^{\mathrm{T}}\mathbf{A}\\ \vdots\\ (\mathbf{b} - \mathbf{A}\mathbf{x}_t)^{\mathrm{T}}\mathbf{A}\,,\end{gathered}$$

where $\mathbf{s}_1^{\mathrm{T}},\ldots,\mathbf{s}_m^{\mathrm{T}}$ are the rows of the sketching matrix **S**. This can be considered as an adaptive form of sketching where the residual directions $(\mathbf{b} - \mathbf{A}\mathbf{x}_t)$ are used after the random directions $\mathbf{s}_1,\ldots,\mathbf{s}_m$. As a consequence, the information-theoretic bounds we considered in Section 4.4.6 do not apply to iterative sketching. In Pilanci and Wainwright [23], it is shown that this algorithm achieves the minimax statistical risk given in (4.16) using at most $O(\log_2 n)$ iterations while obtaining equivalent speedups from sketching. We also note that the iterative sketching method can also be applied to more general convex optimization problems other than the least-squares objective. We refer the reader to Pilanci and Wainwright [38] for the application of sketching in solving general convex optimization problems.

4.6 Non-Parametric Problems

4.6.1 Non-Parametric Regression

In this section we discuss an extension of the sketching method to non-parametric regression problems over Hilbert spaces. The goal of non-parametric regression is making predictions of a continuous response after observing a covariate, where they are related via

$$y_i = f^*(x_i) + v_i, \tag{4.30}$$

where $\mathbf{v} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$, and the function $f^*(\mathbf{x})$ needs to be estimated from $\{x_i, y_i\}_{i=1}^n$. We will consider the well-studied case where the function f^* is assumed to belong to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, and has a bounded Hilbert norm $\|f\|_{\mathcal{H}}$

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

[39, 40]. For these regression problems it is customary to consider the kernel ridge regression (KRR) problem based on convex optimization

$$\widehat{f} = \arg\min_{f\in\mathcal{H}}\left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|_{\mathcal{H}}^2\right\}. \tag{4.31}$$

An RKHS is generated by a kernel function which is positive semidefinite (PSD). A PSD kernel is a symmetric function $K : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ that satisfies

$$\sum_{i,j=1}^{r} y_i y_j K(x_i, x_j) \geq 0$$

for all collections of points $\{x_1,\ldots,x_n\}$, $\{y_1,\ldots,y_n\}$ and $\forall r \in \mathbb{Z}_+$. The vector space of all functions of the form

$$f(\cdot) = \sum_{i}^{r} y_i K(\cdot, x_i)$$

generates an RKHS by taking closure of all such linear combinations. It can be shown that this RKHS is uniquely associated with the kernel function K (see Aronszajn [41] for details). Let us define a finite-dimensional kernel matrix $\mathbf{K}$ using n covariates as follows

$$\mathbf{K}_{ij} = \frac{1}{n}K(x_i, x_j),$$

which is a positive semidefinite matrix. In the linear least-squares regression the kernel matrix reduces to the Gram matrix given by $\mathbf{K} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$. It is also known that the above infinite-dimensional program can be recast as a finite-dimensional quadratic optimization problem involving the kernel matrix

$$\widehat{w} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{K}\mathbf{w} - (1/\sqrt{n})\mathbf{y}\|_2^2 + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{K}\mathbf{w} \tag{4.32}$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^n} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{K}^2\mathbf{w} - \mathbf{w}^{\mathrm{T}}\frac{\mathbf{K}\mathbf{y}}{\sqrt{n}} + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{K}\mathbf{w}, \tag{4.33}$$

and we can find the optimal solution to the infinite-dimensional problem (4.31) via the following relation:[9]

$$\widehat{f}(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\widehat{w}_i K(\cdot, x_i). \tag{4.34}$$

We now define a kernel complexity measure that is based on the eigenvalues of the kernel matrix $\mathbf{K}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ correspond to the real eigenvalues of the symmetric positive-definite kernel matrix $\mathbf{K}$. The kernel complexity is defined as follows.

[9] Our definition of the kernel optimization problem slightly differs from the literature. The classical kernel problem can be recovered by a variable change $\mathbf{w}' = \mathbf{K}^{1/2}w$, where $\mathbf{K}^{1/2}$ is the matrix square root. We refer the reader to [40] for more details on kernel-based methods.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

DEFINITION 4.1 (Kernel complexity)

$$\mathcal{R}(\delta) = \sqrt{\sum_{i=1}^{n} \min\{\delta^2, \lambda_i\}},$$

which is the sum of eigenvalues truncated at level δ. As in (4.14)*, we define a critical radius $\delta^*(n)$ as the smallest positive solution $\delta^*(n) > 0$ to the following inequality:*

$$\frac{R(\delta)}{\delta\sqrt{n}} \leq \frac{\delta}{\sigma}, \tag{4.35}$$

where σ is the noise standard deviation in the statistical model (4.30)*. The existence of a unique solution is guaranteed for all kernel classes (see Bartlett et al. [20]). The critical radius plays an important role in the minimax risk through an information-theoretic argument. The next theorem provides a lower bound on the statistical risk of any estimator applied to the observation model* (4.30).

THEOREM 4.4 *Given n i.i.d. samples from the model* (4.30)*, any estimator $\widehat{f}$ has prediction error lower-bounded as*

$$\sup_{\|f^*\|_{\mathcal{H}\leq 1}} \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(x_i) - f^*(x_i))^2 \geq c_0\delta^*(n)^2, \tag{4.36}$$

where c_0 is a numerical constant and $\delta^*(n)$ is the critical radius defined in (4.35).

The lower bound given by Theorem 4.4 can be shown to be tight, and is achieved by the kernel-based optimization procedure (4.33) and (4.34) (see Bartlett *et al*. [20]). The proof of Theorem 4.4 can be found in Yang *et al*. [42]. We may define the *effective dimension* $d^*(n)$ of the kernel via the relation

$$d^*(n) := n\delta^*(n)^2.$$

This definition allows us to interpret the convergence rate in (4.36) as $d^*(n)/n$, which resembles the classical parametric convergence rate where the number of variables is $d^*(n)$.

4.6.2 Sketching Kernels

Solving the optimization problem (4.33) becomes a computational challenge when the sample size n is large, since it involves linear algebraic operations on an $n \times n$ matrix $\mathbf{K}$. There is a large body of literature on approximating kernel matrices using randomized methods [43–46]. Here we assume that the matrix $\mathbf{K}$ is available, and a sketching matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ can be applied to form a randomized approximation of the kernel matrix. We will present an extension of (4.9), which achieves optimal statistical accuracy. Specifically, the sketching method we consider solves

$$\widehat{\mathbf{v}} = \arg\min_{\mathbf{v}\in\mathbb{R}^m} \frac{1}{2}\mathbf{v}^{\mathrm{T}}(\mathbf{SK})(\mathbf{KS}^{\mathrm{T}})\mathbf{v} - \mathbf{v}^{\mathrm{T}}\frac{\mathbf{SKy}}{\sqrt{n}} + \lambda\mathbf{v}^{\mathrm{T}}\mathbf{SKS}^{\mathrm{T}}\mathbf{v}, \tag{4.37}$$

which involves smaller-dimensional sketched kernel matrices $\mathbf{SK}$, $\mathbf{SKS}^{\mathrm{T}}$ and a lower-dimensional decision variable $\mathbf{v} \in \mathbb{R}^m$. Then we can recover the original variable via

$\mathbf{w} = \mathbf{S}^{\mathrm{T}}\mathbf{v}$. The next theorem shows that the sketched kernel-based optimization method achieves the optimal prediction error.

THEOREM 4.5 *Let $\mathbf{S} \in \mathbb{R}^{m\times n}$ be a Gaussian sketching matrix where $m \geq c_3 d_n$, and choose $\lambda = 3\delta^*(n)$. Given n i.i.d. samples from the model* (4.30)*, the sketching procedure* (4.42) *produces a regression estimate $\widehat{f}$ which satisfies the bound*

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(x_i) - f^*(x_i))^2 \leq c_2\delta^*(n)^2\,,$$

where $\delta^*(n)$ is the critical radius defined in (4.35).

A proof of this theorem can be found in Yang *et al*. [42]. We note that a similar result holds for the ROS sketch matrices with extra logarithmic terms in the dimension of the sketch, i.e., when $m \geq c_4 d_n \log^4(n)$ holds. Notably, Theorem 4.5 guarantees that the sketched estimator achieves the optimal error. This is in contrast to the lower-bound case in Section 4.4.6, where the sketching method does not achieve a minimax optimal error. This is due to the fact that the sketched problem in (4.37) is using the observation $\mathbf{SKy}$ instead of $\mathbf{Sy}$. Therefore, the lower bound in Section 4.4.6 does not apply for this construction. It is worth noting that one can formulate the ordinary least-squares case as a kernel regression problem with kernel $K = \mathbf{AA}^{\mathrm{T}}$, and then apply the sketching method (4.37), which is guaranteed to achieve the minimax optimal risk. However, computing the kernel matrix $\mathbf{AA}^{\mathrm{T}}$ would cost $O(nd^2)$ operations, which is more than would be required for solving the original least-squares problem.

We note that some kernel approximation methods avoid computing the kernel matrix $\mathbf{K}$ and directly form low-rank approximations. We refer the reader to [43] for an example, which also provides an error guarantee for the approximate kernel.

4.7 Extensions: Privacy and Communication Complexity

4.7.1 Privacy and Information-Theoretic Bounds

Another interesting property of randomized sketching is privacy preservation in the context of optimization and learning. Privacy properties of random projections for various statistical tasks have been studied in the recent literature [10, 11, 47]. It is of great theoretical and practical interest to characterize fundamental privacy and optimization trade-offs of randomized algorithms. We first show the relation between sketching and a mutual information-based privacy measure.

4.7.2 Mutual Information Privacy

Suppose we model the data matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ as stochastic, where each entry is drawn randomly. One way we can assess the information revealed to the server is considering the mutual information per symbol, which is given by the formula

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

$$\frac{I(\mathbf{SA};\mathbf{A})}{nd} = \frac{1}{nd}\{H(\mathbf{A}) - H(\mathbf{A}|\mathbf{SA})\}$$
$$= \frac{1}{nd}D(\mathbb{P}_{\mathbf{SA},\mathbf{A}}||\mathbb{P}_{\mathbf{SA}}\mathbb{P}_{\mathbf{A}}),$$

where we normalize by nd since the data matrix $\mathbf{A}$ has nd entries in total. The following corollary is a direct application of Theorem 4.1.

COROLLARY 4.1 *Let the entries of the matrix* $\mathbf{A}$ *be i.i.d from an arbitrary distribution with finite variance* σ^2*. Using sketched data, we can obtain an* ϵ*-approximate*[10] *solution to the optimization problem while ensuring that the revealed mutual information satisfies*

$$\frac{I(\mathbf{SA};\mathbf{A})}{nd} \leq \frac{c_0}{\epsilon^2}\frac{\mathcal{W}^2(\mathbf{A}\mathcal{K})}{n}\log_2(2\pi e\sigma^2).$$

Therefore, we can guarantee the mutual information privacy of the sketching-based methods, whenever the term $\mathcal{W}(\mathbf{A}\mathcal{K})$ is small.

An alternative and popular characterization of privacy is referred to as the differential privacy (see Dwork *et al.* [9]), where other randomized methods, such as additive noise for preserving privacy, were studied. It is also possible to directly analyze differential privacy-preserving aspects of random projections as considered in Blocki *et al.* [10].

4.7.3 Optimization-Based Privacy Attacks

We briefly discuss a possible approach an adversary might take to circumvent the privacy provided by sketching. If the data matrix is sparse, then one might consider optimization-based recovery techniques borrowed from compressed sensing to recover the data $\mathbf{A}$ given the sketched data $\widetilde{\mathbf{A}} = \mathbf{SA}$,

$$\min_{\mathbf{A}} ||\mathbf{A}||_1$$
$$\text{s.t. } \mathbf{SA} = \widetilde{\mathbf{A}},$$

where we have used the matrix ℓ_1 norm $||\mathbf{A}||_1 := \sum_{i=1}^{n}\sum_{j=1}^{d}|\mathbf{A}_{ij}|$. The success of the above optimization method will critically depend on the sparsity level of the original data $\mathbf{A}$. Most of the randomized sketching constructions shown in Section 4.2 can be shown to be susceptible to data recovery via optimization (see Candès and Tao [25] and Candès *et al.* [48]). However, this method assumes that the sketching matrix $\mathbf{S}$ is available to the attacker. If $\mathbf{S}$ is not available to the adversary, then the above method cannot be used and the recovery is not straightforward.

4.7.4 Communication Complexity-Space Lower Bounds

In this section we consider a streaming model of computation, where the algorithm is allowed to make only one pass over the data. In this model, an algorithm receives updates to the entries of the data matrix $\mathbf{A}$ in the form "add a to $\mathbf{A}_{ij}$." An entry can

[10] Here ϵ-approximate solution refers to the approximation defined in Theorem 4.1, relative to the optimal value.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

be updated more than once, and the value a is any arbitrary real number. The sketches introduced in this chapter provide a valuable data structure when the matrix is very large in size, and storing and updating the matrix directly can be impractical. Owing to the linearity of sketches, we can update the sketch $\mathbf{SA}$ by adding $a\mathbf{S}\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}$ to $\mathbf{SA}$, and maintain an approximation with limited memory.

The following theorem due to Clarkson and Woodruff [49] provides a lower bound of the space used by any algorithm for least-squares regression which performs a single pass over the data.

THEOREM 4.6 *Any randomized 1-pass algorithm which returns an ϵ-approximate solution to the unconstrained least-squares problem with probability at least $7/9$ needs $\Omega(d^2(1/\epsilon + \log(nd)))$ bits of space.*

This theorem confirms that the space complexity of sketching for unconstrained least-squares regression is near optimal. Because of the choice of the sketching dimension $m = O(d)$, the space used by the sketch $\mathbf{SA}$ is $O(d^2)$, which is optimal up to constants according to the theorem.

4.8 Numerical Experiments

In this section, we illustrate the sketching method numerically and confirm the theoretical predictions of Theorems 4.1 and 4.3. We consider both the classical low-dimensional statistical regime where $n > d$, and the ℓ_1-constrained least-squares minimization known as LASSO (see Tibshirani [51]):

$$\mathbf{x}^* = \arg \min_{\mathbf{x}\, s.t.\, \|\mathbf{x}\|_1 \le \lambda} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 .$$

We generate a random i.i.d. data matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, where $n = 10\,000$ and $d = 1000$, and set the observation vector $\mathbf{b} = \mathbf{A}\mathbf{x}^{\dagger} + \sigma\mathbf{w}$, where $\mathbf{x}^{\dagger} \in \{-1, 0, 1\}^d$ is a random s-sparse vector and $\mathbf{w}$ has i.i.d. $\mathcal{N}(0, 10^{-4})$ components. For the sketching matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$, we consider the Gaussian and Rademacher (±1 i.i.d.-valued) random matrices, where m ranges between 10 and 400. Consequently, we solve the sketched program

$$\widehat{\mathbf{x}} = \arg \min_{\mathbf{x}\, s.t.\, \|\mathbf{x}\|_1 \le \lambda} \|\mathbf{S}\mathbf{A}\mathbf{x} - \mathbf{S}\mathbf{b}\|_2 .$$

Figures 4.5 and 4.6 show the relative prediction mean-squared error given by the ratio

$$\frac{(1/n)\|\mathbf{A}(\widehat{\mathbf{x}} - \mathbf{x}^{\dagger})\|_2^2}{(1/n)\|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^{\dagger})\|_2^2} ,$$

where it is averaged over 20 realizations of the sketching matrix, and $\widehat{\mathbf{x}}$ and $\mathbf{x}^*$ are the sketched and the original solutions, respectively. As predicted by the upper and lower bounds given in Theorems 4.1 and 4.3, the prediction mean-squared error of the sketched estimator scales as $O((s \log d)/m)$, since the corresponding Gaussian complexity $\mathcal{W}_1(\mathcal{K})^2$ is $O(s \log d)$. These plots reveal that the prediction mean-squared error of the sketched estimators for both Gaussian and Rademacher sketches are in agreement with the theory.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

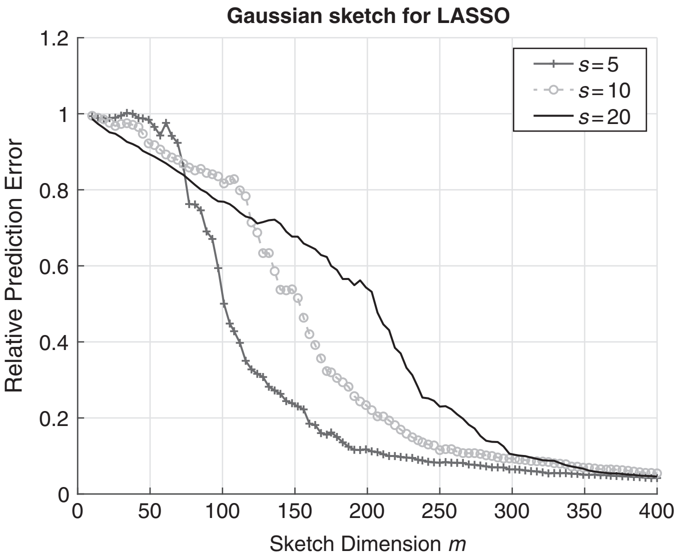

Figure 4.5 Sketching LASSO using Gaussian random projections.

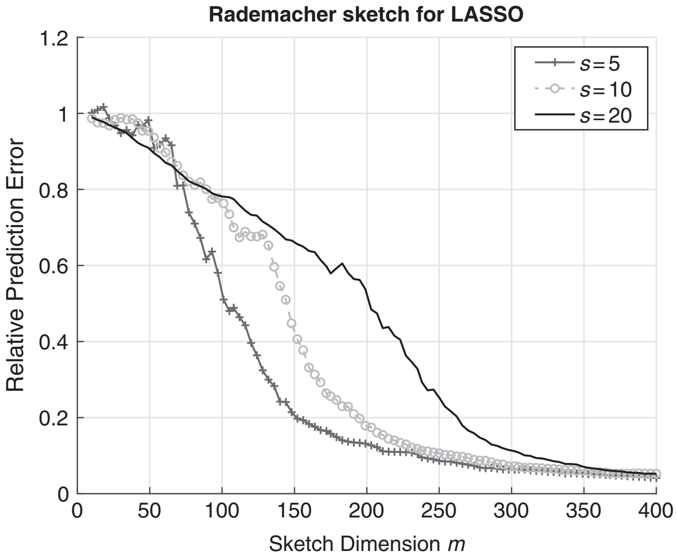

Figure 4.6 Sketching LASSO using Rademacher random projections.

4.9 Conclusion

This chapter presented an overview of random projection-based methods for solving large-scale statistical estimation and constrained optimization problems. We investigated fundamental lower bounds on the performance of sketching using information-theoretic tools. Randomized sketching has interesting theoretical properties, and also has numerous practical advantages in machine-learning and optimization

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

problems. Sketching yields faster algorithms with lower space complexity while maintaining strong approximation guarantees.

For the upper bound on the approximation accuracy in Theorem 4.2, Gaussian complexity plays an important role, and also provides a geometric characterization of the dimension of the sketch. The lower bounds given in Theorem 4.3 are statistical in nature, and involve packing numbers, and consequently metric entropy, which measures the complexity of the sets. The upper bounds on the Gaussian sketch can be extended to Rademacher sketches, sub-Gaussian sketches, and randomized orthogonal system sketches (see Pilanci and Wainwright [22] and also Yun *et al*. [42] for the proofs). However, the results for non-Gaussian sketches often involve superfluous logarithmic factors and large constants as artifacts of the analysis. As can be observed in Figs. 4.5 and 4.6, the mean-squared error curves for Gaussian and Rademacher sketches are in agreement with each other. It can be conjectured that the approximation ratio of sketching is universal for random matrices with entries sampled from well-behaved distributions. This is an important theoretical question for future research. We refer the reader to the work of Donoho and Tanner [51] for observations of the universality in compressed sensing.

Finally, a number of important limitations of the analysis techniques need to be considered. The minimax criterion (4.16) is a worst-case criterion in nature by virtue of its definition, and may not correctly reflect the average error of sketching when the unknown vector $\mathbf{x}^{\dagger}$ is randomly distributed. Furthermore, in some applications, it might be suitable to consider prior information on the unknown vector. As an interesting direction of future research, it would be interesting to study lower bounds for sketching in a Bayesian setting.

A4.1 Proof of Theorem 4.3

Let us define the shorthand notation $\|\cdot\|_{\mathbf{A}} := (1/\sqrt{n})\|\mathbf{A}(\cdot)\|_2$. Let $\{\mathbf{z}^j\}_{j=1}^{M}$ be a 1/2-packing of $C_0 \cap \mathbb{B}_A(1)$ in the semi-norm defined by $\|\cdot\|_{\mathbf{A}}$, and, for a fixed $\delta \in (0, 1/4)$, define $\mathbf{x}^j = 4\delta\mathbf{z}^j$. Since $4\delta \in (0,1)$, the star-shaped assumption guarantees that each $\mathbf{x}^j$ belongs to C_0. We thus obtain a collection of M vectors in C_0 such that

$$2\delta \leq \|\mathbf{x}^j - \mathbf{x}^k\|_{\mathbf{A}} \leq 8\delta \qquad \text{for all } j \neq k.$$

Letting J be a random index uniformly distributed over $\{1, \ldots, M\}$, suppose that, conditionally on $J = j$, we observe the sketched observation vector $\mathbf{Sb} = \mathbf{SAx}^j + \mathbf{Sw}$, as well as the sketched matrix $\mathbf{SA}$. Conditioned on $J = j$, the random vector $\mathbf{Sb}$ follows an $\mathcal{N}(\mathbf{SAx}^j, \sigma^2\mathbf{SS}^{\mathrm{T}})$ distribution, denoted by $\mathbb{P}_{\mathbf{x}^j}$. We let $\bar{Y}$ denote the resulting mixture variable, with distribution $(1/M)\sum_{j=1}^{M} \mathbb{P}_{\mathbf{x}^j}$.

Consider the multi-way testing problem of determining the index J by observing $\bar{Y}$. With this setup, we may apply Lemma 4.3 (see, e.g., [30, 46]), which implies that, for any estimator $\mathbf{x}^{\dagger}$, the worst-case mean-squared error is lower-bounded as

$$\sup_{\mathbf{x}^* \in C} \mathbb{E}_{\mathbf{S},\mathbf{w}} \|\mathbf{x}^{\dagger} - \mathbf{x}^*\|_{\mathbf{A}}^2 \geq \delta^2 \inf_{\psi} \mathbb{P}[\psi(\bar{Y}) \neq J], \tag{4.38}$$

where the infimum ranges over all testing functions ψ. Consequently, it suffices to show that the testing error is lower-bounded by 1/2.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

In order to do so, we first apply Fano's inequality [27] conditionally on the sketching matrix $\mathbf{S}$ and get

$$\mathbb{P}[\psi(\bar{Y}) \neq J] \; = \; \mathbb{E}_{\mathbf{S}}\Big\{\mathbb{P}[\psi(\bar{Y}) \neq J \mid \mathbf{S}]\Big\} \geq 1 - \frac{\mathbb{E}_{\mathbf{S}}[I_{\mathbf{S}}(\bar{Y}; J)] + 1}{\log_2 M}, \tag{4.39}$$

where $I_{\mathbf{S}}(\bar{Y}; J)$ denotes the mutual information between $\bar{Y}$ and J with $\mathbf{S}$ fixed. Our next step is to upper-bound the expectation $\mathbb{E}_{\mathbf{S}}[I(\bar{Y}; J)]$.

Letting $D(\mathbb{P}_{\mathbf{x}^j} \,||\, \mathbb{P}_{\mathbf{x}^k})$ denote the Kullback–Leibler (KL) divergence between the distributions $\mathbb{P}_{\mathbf{x}^j}$ and $\mathbb{P}_{\mathbf{x}^k}$, the convexity of KL divergence implies that

$$\begin{aligned} I_{\mathbf{S}}(\bar{Y}; J) &= \frac{1}{M}\sum_{j=1}^{M} D\left(\mathbb{P}_{\mathbf{x}^j} \,||\, \frac{1}{M}\sum_{k=1}^{M}\mathbb{P}_{\mathbf{x}^k}\right) \\ &\leq \frac{1}{M^2}\sum_{j,k=1}^{M} D(\mathbb{P}_{\mathbf{x}^j} \,||\, \mathbb{P}_{\mathbf{x}^k}). \end{aligned}$$

Computing the KL divergence for Gaussian vectors yields

$$I_{\mathbf{S}}(\bar{Y}; J) \leq \frac{1}{M^2}\sum_{j,k=1}^{M} \frac{1}{2\sigma^2}(\mathbf{x}^j - \mathbf{x}^k)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\left[\mathbf{S}^{\mathrm{T}}(\mathbf{S}\mathbf{S}^{\mathrm{T}})^{-1}\mathbf{S}\right]\mathbf{A}(\mathbf{x}^j - \mathbf{x}^k).$$

Thus, using condition (4.15), we have

$$\mathbb{E}_{\mathbf{S}}[I(\bar{Y}; J)] \leq \frac{1}{M^2}\sum_{j,k=1}^{M}\frac{m\,\eta}{2\,n\sigma^2}||\mathbf{A}(\mathbf{x}^j - \mathbf{x}^k)||_2^2 \;\leq\; \frac{32\,m\eta}{\sigma^2}\,\delta^2,$$

where the final inequality uses the fact that $||\mathbf{x}^j - \mathbf{x}^k||_{\mathbf{A}} = 1/\sqrt{n}||\mathbf{A}(\mathbf{x}^j - \mathbf{x}^k)||_2 \leq 8\delta$ for all pairs.

Combined with our previous bounds (4.38) and (4.39), we find that

$$\sup_{\mathbf{x}^* \in C} \mathbb{E}||\widehat{\mathbf{x}} - \mathbf{x}^*||_2^2 \geq \delta^2\left\{1 - \frac{32(m\eta\delta^2/\sigma^2) + 1}{\log_2 M}\right\}.$$

Setting $\delta = \sigma^2\log_2(M/2)/64\,\eta\, m$ yields the lower bound (4.19).

A4.2 Proof of Lemma 4.3

By Markov's inequality applied on the random variable $||\widehat{\mathbf{x}} - \mathbf{x}^\dagger||_{\mathbf{A}}^2$ we have

$$\mathbb{E}\,||\widehat{\mathbf{x}} - \mathbf{x}^\dagger||_{\mathbf{A}}^2 \geq \delta^2\,\mathbb{P}[||\widehat{\mathbf{x}} - \mathbf{x}^\dagger||_{\mathbf{A}}^2 \geq \delta^2]. \tag{4.40}$$

Now note that

$$\begin{aligned} \sup_{\mathbf{x}^* \in C}\mathbb{P}[||\widehat{\mathbf{x}} - \mathbf{x}^\dagger||_{\mathbf{A}} \geq \delta] &\geq \max_{j \in \{1,\ldots,M\}} \mathbb{P}[\;||\widehat{\mathbf{x}} - \mathbf{x}^{(j)}||_{\mathbf{A}} \geq \delta \mid J_\delta = j] \\ &\geq \frac{1}{M}\sum_{j=1}^{M}\mathbb{P}[||\widehat{\mathbf{x}} - \mathbf{x}^{(j)}||_{\mathbf{A}} \geq \delta \mid J_\delta = j], \end{aligned} \tag{4.41}$$

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

since every element of the packing set satisfies $\mathbf{x}^{(j)} \in C$ and the discrete maximum is upper-bounded by the average over $\{1,\ldots,M\}$. Since we have $\mathbb{P}[J_\delta = j] = 1/M$, we equivalently have

$$\frac{1}{M}\sum_{j=1}^{M}\mathbb{P}[\|\widehat{\mathbf{x}}-\mathbf{x}^{(j)}\|_{\mathbf{A}}] \geq \delta \mid J_\delta = j] = \sum_{j=1}^{M}\mathbb{P}\Big[\|\widehat{\mathbf{x}}-\mathbf{x}^{(j)}\|_{\mathbf{A}} \geq \delta \Big| J_\delta = j\Big]\mathbb{P}[J_\delta = j]$$
$$= \mathbb{P}\Big[\|\widehat{\mathbf{x}}-\mathbf{x}^{(J_\delta)}\|_{\mathbf{A}} \geq \delta\Big]. \tag{4.42}$$

Now we will argue that, whenever the true index is $J_\delta = j$ and if $\|\widehat{x}-x^{(j)}\|_{\mathbf{A}} < \delta$, then we can form a hypothesis tester $\psi(Z)$ identifying the true index j. Consider the test

$$\psi(Z) := \arg\min_{j\in[M]} \|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|_{\mathbf{A}}.$$

Now note that $\|\mathbf{x}^j - \widehat{\mathbf{x}}\|_{\mathbf{A}} < \delta$ ensures that

$$\|\mathbf{x}^{(i)} - \widehat{\mathbf{x}}\|_{\mathbf{A}} \geq \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|_{\mathbf{A}} - \|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|_{\mathbf{A}} \geq 2\delta - \delta = \delta,$$

where the second inequality follows from the 2δ-packing construction of our collection $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(M)}$. Consequently $\|\mathbf{x}^{(i)} - \widehat{\mathbf{x}}\|_{\mathbf{A}} > \delta$ for all $i \in \{1,\ldots,N\} - \{j\}$, and the test $\psi(Z)$ identifies the true index $J = j$. Therefore we obtain

$$\Big\{\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|_{\mathbf{A}} < \delta\Big\} \quad \Rightarrow \quad \{\phi(Z) = j\},$$

and conclude that the complements of these events obey

$$\mathbb{P}\Big[\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|_{\mathbf{A}} \geq \delta \mid J_\delta = j\Big] \geq \mathbb{P}[\phi(Z) \neq j \mid J_\delta = j].$$

Taking averages over the indices $1,\ldots,M$, we obtain

$$\mathbb{P}\Big[\|\mathbf{x}^{(J_\delta)} - \widehat{\mathbf{x}}\|_{\mathbf{A}} \geq \delta\Big] = \frac{1}{M}\sum_{j=1}^{M}\mathbb{P}\Big[\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|_{\mathbf{A}} \geq \delta \mid J_\delta = j\Big] \geq \mathbb{P}[\phi(Z) \neq J_\delta].$$

Combining the above with the earlier lower bound (4.41) and the identity (4.42), we obtain

$$\sup_{\mathbf{x}^*\in C}\mathbb{P}[\|\widehat{\mathbf{x}} - \mathbf{x}^*\|_{\mathbf{A}} \geq \delta] \geq \mathbb{P}[\phi(Z) \neq J_\delta] \geq \inf_{\phi}\mathbb{P}[\phi(Z) \neq J_\delta],$$

where the second inequality follows by taking the infimum over all tests, which can only make the probability smaller. Plugging in the above lower bound in (4.40) completes the proof of the lemma.

References

[1] S. Vempala, *The random projection method*. American Mathematical Society, 2004.

[2] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

[3] N. Halko, P. Martinsson, and J. A. Tropp, “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions,” *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, 2011.

[4] M. W. Mahoney, *Randomized algorithms for matrices and data*. Now Publishers, 2011.

[5] D. P. Woodruff, “Sketching as a tool for numerical linear algebra,” *Foundations and Trends Theoretical Computer Sci.*, vol. 10, nos. 1–2, pp. 1–157, 2014.

[6] S. Muthukrishnan, “Data streams: Algorithms and applications,” *Foundations and Trends Theoretical Computer Sci.*, vol. 1, no. 2, pp. 117–236, 2005.

[7] B. Yu, “Assouad, Fano, and Le Cam,” in *Festschrift in Honor of Lucien Le Cam.* Springer, 1997, pp. 423–435.

[8] Y. Yang and A. Barron, “Information-theoretic determination of minimax rates of convergence,” *Annals Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Proc. Theory of Cryptography Conference*, 2006, pp. 265–284.

[10] J. Blocki, A. Blum, A. Datta, and O. Sheffet, “The Johnson–Lindenstrauss transform itself preserves differential privacy,” in *Proc. 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, 2012, pp. 410–419.

[11] N. Ailon and B. Chazelle, “Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform,” in *Proc. 38th Annual ACM Symposium on Theory of Computing*, 2006, pp. 557–563.

[12] P. Drineas and M. W. Mahoney, “Effective resistances, statistical leverage, and applications to linear equation solving,” *arXiv:1005.3097*, 2010.

[13] D. A. Spielman and N. Srivastava, “Graph sparsification by effective resistances,” *SIAM J. Computing*, vol. 40, no. 6, pp. 1913–1926, 2011.

[14] M. Charikar, K. Chen, and M. Farach-Colton, “Finding frequent items in data streams,” in *International Colloquium on Automata, Languages, and Programming*, 2002, pp. 693–703.

[15] D. M. Kane and J. Nelson, “Sparser Johnson–Lindenstrauss transforms,” *J. ACM*, vol. 61, no. 1, article no. 4, 2014.

[16] J. Nelson and H. L. Nguyên, “Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings,” in *Proc. 2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, 2013, pp. 117–126.

[17] J. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms*. Springer, 1993, vol. 1.

[18] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[19] M. Ledoux and M. Talagrand, *Probability in Banach spaces: Isoperimetry and processes*. Springer, 1991.

[20] P. L. Bartlett, O. Bousquet, and S. Mendelson, “Local Rademacher complexities,” *Annals Statist.*, vol. 33, no. 4, pp. 1497–1537, 2005.

[21] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, “The convex geometry of linear inverse problems,” *Foundations Computational Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[22] M. Pilanci and M. J. Wainwright, “Randomized sketches of convex programs with sharp guarantees,” UC Berkeley, Technical Report, 2014, full-length version at *arXiv:1404.7203*; Presented in part at ISIT 2014.

[23] M. Pilanci and M. J. Wainwright, “Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares,” *J. Machine Learning Res.*, vol. 17, no. 1, pp. 1842–1879, 2016.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

[24] S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. Sci. Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[25] E. J. Candès and T. Tao, “Decoding by linear programming,” *IEEE Trans. Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[26] R. M. Fano and W. Wintringham, “Transmission of information,” *Phys. Today*, vol. 14, p. 56, 1961.

[27] T. Cover and J. Thomas, *Elements of information theory*. John Wiley & Sons, 1991.

[28] P. Assouad, “Deux remarques sur l’estimation,” *Comptes Rendus Acad. Sci. Paris*, vol. 296, pp. 1021–1024, 1983.

[29] I. A. Ibragimov and R. Z. Has’minskii, *Statistical estimation: Asymptotic theory*. Springer, 1981.

[30] L. Birgé, “Estimating a density under order restrictions: Non-asymptotic minimax risk,” *Annals Statist.*, vol. 15, no. 3, pp. 995–1012, 1987.

[31] A. Kolmogorov and B. Tikhomirov, “ϵ-entropy and ϵ-capacity of sets in functional spaces,” *Uspekhi Mat. Nauk*, vol. 86, pp. 3–86, 1959, English transl. *Amer. Math. Soc. Translations*, vol. 17, pp. 277–364, 1961.

[32] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[33] G. Raskutti, M. J. Wainwright, and B. Yu, “Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls,” *IEEE Trans. Information Theory*, vol. 57, no. 10, pp. 6976–6994, 2011.

[34] N. Srebro, N. Alon, and T. S. Jaakkola, “Generalization error bounds for collaborative prediction with low-rank matrices,” in *Proc. Advances in Neural Information Processing Systems*, 2005, pp. 1321–1328.

[35] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *J. Roy. Statist. Soc. B*, vol. 1, no. 68, p. 49, 2006.

[36] S. Negahban and M. J. Wainwright, “Estimation of (near) low-rank matrices with noise and high-dimensional scaling,” *Annals Statist.*, vol. 39, no. 2, pp. 1069–1097, 2011.

[37] F. Bunea, Y. She, and M. Wegkamp, “Optimal selection of reduced rank estimators of high-dimensional matrices,” *Annals Statist.*, vol. 39, no. 2, pp. 1282–1309, 2011.

[38] M. Pilanci and M. J. Wainwright, “Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence,” *SIAM J. Optimization*, vol. 27, no. 1, pp. 205–245, 2017.

[39] H. L. Weinert, (ed.), *Reproducing kernel hilbert spaces: Applications in statistical signal processing*. Hutchinson Ross Publishing Co., 1982.

[40] B. Schölkopf and A. Smola, *Learning with kernels*. MIT Press, 2002.

[41] N. Aronszajn, “Theory of reproducing kernels,” *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[42] Y. Yang, M. Pilanci, and M. J. Wainwright, “Randomized sketches for kernels: Fast and optimal nonparametric regression,” *Annals Statist.*, vol. 45, no. 3, pp. 991–1023, 2017.

[43] A. Rahimi and B. Recht, “Random features for large-scale kernel machines,” in *Proc. Advances in Neural Information Processing Systems*, 2008, pp. 1177–1184.

[44] A. Rahimi and B. Recht, “Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning,” in *Proc. Advances in Neural Information Processing Systems*, 2009, pp. 1313–1320.

[45] P. Drineas and M. W. Mahoney, “On the Nyström method for approximating a Gram matrix for improved kernel-based learning,” *J. Machine Learning Res.*, vol. 6, no. 12, pp. 2153–2175, 2005.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press

[46] Q. Le, T. Sarlós, and A. Smola, "Fastfood – approximating kernel expansions in loglinear time," in *Proc. 30th International Conference on Machine Learning*, 2013, 9 unnumbered pages.

[47] S. Zhou, J. Lafferty, and L. Wasserman, "Compressed regression," *IEEE Trans. Information Theory*, vol. 55, no. 2, pp. 846–866, 2009.

[48] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 489–509, 2004.

[49] K. L. Clarkson and D. P. Woodruff, "Numerical linear algebra in the streaming model," in *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, 2009, pp. 205–214.

[50] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[51] D. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Phil. Trans. Roy. Soc. London A: Math., Phys. Engineering Sci.*, vol. 367, no. 1906, pp. 4273–4293, 2009.

https://doi.org/10.1017/9781108616799.005 Published online by Cambridge University Press