CS395T: Continuous Algorithms, Part IX
Sparse recovery

Kevin Tian

1 Basis pursuit

In Part VIII, we developed efficient algorithms for *overconstrained* linear systems, i.e. solving a regression problem $\mathbf{A}x \approx b$ where $\mathbf{A}$ has more rows (constraints) than columns (features). This lecture considers the *underconstrained* counterpart of this problem, where we model observations $b \in \mathbb{R}^n$ as being the result of noisy linear measurements of a hidden vector $x^\star \in \mathbb{R}^d$, i.e. for $n < d$,[1]

$$b = \mathbf{A}x^\star + \xi,\ \mathbf{A} \in \mathbb{R}^{n \times d},\ b, \xi \in \mathbb{R}^n,\ x^\star \in \mathbb{R}^d \text{ with } \text{nnz}(x^\star) \leq k. \tag{1}$$

In the noiseless case $\xi = \mathbb{0}_n$, the problem (1) has infinitely-many solutions, because $\mathbf{A}$ has a kernel, and moving arbitrarily within this kernel induces new solutions. One may posit that this means the underconstrained linear regression problem is meaningless in general, because recovering $x^\star$ from $\mathbf{A}x^\star$ is ill-posed when $n < d$. However, a sequence of striking breakthroughs by [CR05, CT05, Don06, CT06, CRT06] pushed back on this conventional wisdom, by showing that the problem of recovering $x^\star$ from underconstrained measurements is both well-defined and algorithmically tractable, under additional structural assumptions on $\mathbf{A}$ and $x^\star$. These works and follow-ups led to the development of the field of *sparse recovery*, also known as *compressed sensing* because of the ability to detect or approximate $x^\star$ from a compressed number of $\ll d$ measurements.

As implied by the name sparse recovery, in this lecture we consider the case where $x^\star$ is k-sparse, i.e. $\text{nnz}(x^\star) \leq k$, and we wish to approximately recover $x^\star$ from noisy measurements. This modeling assumption, subject to accounting for noise, is well-motivated in practice. To see why, the noise-tolerant model (1) extends to the case where a parameter vector of interest x is *heavy*, i.e. there is a subset $S \subseteq [d]$ with $|S| \leq k$ such that $\|x_S\|_2$ is responsible for a large portion of $\|x\|_2$. In this case, we can treat the heavy coordinates x_S as our hidden parameter vector $x^\star$, and lump $\mathbf{A}x_{S^c}$ into the noise component ξ where $S^c := [d] \setminus S$, with the hope that $\|\mathbf{A}x_{S^c}\|_2 \ll \|\mathbf{A}x^\star_S\|_2$. Indeed, recovery of heavy parameter vectors is a ubiquitous problem in many application domains. For example, in signal processing (e.g. for recording music), only a few coordinates in the Fourier spectrum are large; similarly, images are often sparse in the Haar wavelet basis, which captures locality (e.g. features localized to a few adjacent pixels). Moreover, various real-world data, such as the number of inlinks per page in internet networks, follow power-law distributions which are naturally heavy. We defer additional discussion of this phenomenon to the excellent reference [Pri20].

In this section, to introduce the algorithmic tractability of (1), we focus on the noiseless problem

$$b = \mathbf{A}x^\star,\ \mathbf{A} \in \mathbb{R}^{n \times d},\ b \in \mathbb{R}^n,\ x^\star \in \mathbb{R}^d \text{ with } \text{nnz}(x^\star) \leq k, \tag{2}$$

i.e. (1) where $\xi = \mathbb{0}_n$, handling more general noisy cases in Sections 2 and 3. As a starting point, observe that it is ideal that $\mathbf{A}$ has no $O(k)$-sparse vectors in its kernel. For example, suppose our problem is mildly-misspecified and $x^\star$ is actually $\frac{k}{2}$-sparse,[2] but $\mathbf{A}$ has a $\frac{k}{2}$-sparse vector v in its kernel. Then, $x^\star + tv$ for any $t \in \mathbb{R}$ is an equally-plausible solution to (2), and hence our sparse recovery problem is again ill-posed. The following robust variant of this "no sparse vectors in $\mathbf{A}$'s kernel" assumption was formalized by [CDD09], where it was termed the nullspace property.

[1] We mention that there is a design decision in how to model the noise ξ. The definition (1) considers "output noise," i.e. ξ is added to $\mathbf{A}x^\star$, which captures applications where $x^\star$ is truly-sparse and we can only access noisy linear measurements of it. Another natural assumption is "input noise" where we let $b = \mathbf{A}(x^\star + \xi)$. As we explore later in this lecture, under structural assumptions on $\mathbf{A}$ these notions are comparable.

[2] It is more realistic to assume that k is an upper bound on the sparsity of $x^\star$ (than $\text{nnz}(x^\star) = k$, for example), since it is often an unknown parameter which must be estimated using domain knowledge.

Definition 1 (Nullspace property)**.** *We say* $\mathbf{A} \in \mathbb{R}^{n \times d}$ *satisfies the* k-nullspace property*, or* k*-NSP, if for all* $S \subseteq [d]$ *with* $|S| \leq k$*, all vectors* $v \in \mathbb{R}^d$ *with* $\mathbf{A}v = \mathbb{0}_n$ *have*

$$\|v_S\|_1 < \|v_{S^c}\|_1 .$$

Ignoring computational issues of how to determine whether $\mathbf{A}$ satisfies k-NSP until later in the lecture, we demonstrate how Definition 1 leads to polynomial-time algorithms for sparse recovery in the noiseless setting (2). Intuitively, Definition 1 prevents the type of sparse movement in the kernel of $\mathbf{A}$ described earlier, because it enforces that all v with $\mathbf{A}v = 0$ cannot have a k-sized set S which accounts for most of v's ℓ_1 mass. We now formalize this intuition.

Lemma 1. *Let* $x^\star \in \mathbb{R}^d$ *have* $x_i^\star = 0$ *for all* $i \notin S$*, for* $S \subseteq [d]$*. Then for all* $x \in \mathbb{R}^d$ *with* $\|x\|_1 \leq \|x^\star\|_1$*, we have for* $v := x - x^\star$ *that* $\|v_S\|_1 \geq \|v_{S^c}\|_1$*.*

Proof. We apply the triangle inequality and the support assumption on $x^\star$:

$$\begin{aligned} \|x\|_1 &= \|x^\star + v_S + v_{S^c}\|_1 = \|x^\star + v_S\|_1 + \|v_{S^c}\|_1 \\ &\geq \|x^\star\|_1 - \|v_S\|_1 + \|v_{S^c}\|_1 \geq \|x\|_1 - \|v_S\|_1 + \|v_{S^c}\|_1 . \end{aligned}$$

In particular, the second equality used coordinatewise-separability of ℓ_1, and the last inequality applied the assumption $\|x\|_1 \leq \|x^\star\|_1$. The conclusion follows by rearranging. □

Lemma 1 motivates a simple algorithm for recovering $x^\star$ in (2), when $\mathbf{A}$ satisfies k-NSP.

Theorem 1 (Basis pursuit)**.** *Consider an instance of the problem* (2)*. If* $\mathbf{A}$ *satisfies* k*-NSP, then* $x^\star$ *is the unique solution to*

$$\min_{\substack{x \in \mathbb{R}^d \\ \mathbf{A}x = b}} \|x\|_1 . \tag{3}$$

Proof. Suppose for contradiction that there is $x \neq x^\star$ satisfying $\mathbf{A}x = b$ and $\|x\|_1 \leq \|x^\star\|_1$. By definition, this means $v := x - x^\star$ has $\mathbf{A}v = \mathbb{0}_n$, but additionally Lemma 1 implies $\|v_S\|_1 \geq \|v_{S^c}\|_1$, where $S \subseteq [d]$ is the support of $x^\star$. Since $|S| \leq k$ by assumption, this contradicts k-NSP. □

We note that the problem (3) is equivalent to a linear program, by introducing d auxiliary variables $\{t_i\}_{i \in [d]} \in \mathbb{R}_{\geq 0}$, adding the linear constraints $-t_i \leq x_i \leq t_i$ for all $i \in [d]$, and minimizing $\langle t, \mathbb{1}_d \rangle$. Thus, there are polynomial-time algorithms for solving (3) to high accuracy, via Theorem 1, Part I (or alternatively, the more specialized algorithms developed in the next lecture).

Remark 1. *The linear programming-based solution* (3) *to underconstrained linear systems has a longer history than discussed here, surveyed in e.g. [CDS01]. The intuition for this* ℓ_1*-based minimization problem is that the* ℓ_1 *norm serves as a proxy for the* ℓ_0 *"norm," another name for number of nonzero elements* $\|x\|_0 := \mathrm{nnz}(x)$*. We would like to find sparse solutions to* $\mathbf{A}x = b$*, but* $\|\cdot\|_0$ *is nonconvex (so it is not an actual norm), and in fact it is discontinuous. Instead, guided by the intuition that for a fixed* ℓ_2 *budget, the* ℓ_1 *norm is smaller for sparse vectors, since*

$$\|x\|_1 \leq \sqrt{\mathrm{nnz}(x)}\, \|x\|_2 \tag{4}$$

by the Cauchy-Schwarz inequality, (3) *has long been used as a heuristic to choose sparse solutions. Indeed,* (3) *is often described as a convex relaxation of the nonconvex problem*

$$\min_{\substack{x \in \mathbb{R}^d \\ \mathbf{A}x = b}} \mathrm{nnz}(x),$$

which models the actual goal of sparse recovery when there is a uniquely-sparsest solution. The name basis pursuit arises from a signal processing viewpoint which casts $\mathbf{A}$ *as a dictionary of columns, and hence the goal of sparse recovery in this language is to find a linear combination of a small subset of these columns (i.e. a basis) to reconstruct* b*. It was finally proven by [CRT06] that choosing* $\mathbf{A}$ *to be a subsampled Fourier matrix, solving* (4) *succeeds in recovering sparse signal* $x^\star$ *with few samples (i.e. rows of* $\mathbf{A}$*). This result inspired a flurry of work which abstracted properties of* $\mathbf{A}$ *needed for basis pursuit to succeed in sparse recovery, see e.g. the survey [Rau10].*

2 Restricted isometry property

In this section, we consider an alternative structural property of a matrix $\mathbf{A}$ which makes recovering $x^\star$ in (1) tractable. This structural property, known as the restricted isometry property, can be thought of as a quantitative variant of Definition 1 from earlier.

Definition 2 (Restricted isometry property)**.** *We say* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *satisfies the* (ϵ, k)-restricted isometry property*, or* (ϵ, k)*-RIP, if for all* $x \in \mathbb{R}^d$ *with* $\mathrm{nnz}(x) \le k$*,*

$$(1-\epsilon)\|x\|_2^2 \le \|\mathbf{A}x\|_2^2 \le (1+\epsilon)\|x\|_2^2.$$

Definition 2 requires that beyond having no sparse vectors in its kernel, $\mathbf{A}$ should also act as an approximate isometry when restricted to sparse vectors. That is, while NSP is closer to a full-rank condition on sparse vectors, RIP posits that $\mathbf{A}^\top\mathbf{A}$ is well-conditioned when restricted to this set. In this lecture, we primarily focus on the case when $\epsilon \in (0,1)$ is a fixed constant for simplicity. For various random models of $\mathbf{A}$, e.g. when $\mathbf{A}$ has rows sampled from an appropriately-rescaled isotropic Gaussian $\mathcal{N}(\mathbb{0}_d, \sigma^2\mathbf{I}_d)$, we can establish Definition 2. We remark that preconditioned variants of the methods we discuss are an active research area, see e.g. [KKMR21] which established lower bounds on the sample complexity required by these preconditioned methods when $\mathbf{A}$ has rows $\sim \mathcal{N}(\mathbb{0}_d, \mathbf{\Sigma})$ for ill-conditioned $\mathbf{\Sigma}$ with certain graph structure.

Unfortunately, in general both RIP and NSP are NP-hard to certify [BDMS13, TP14]. However, the sparse recovery community has developed various arguments for showing that certain random matrix designs satisfy RIP with high probability. To give a simple example, we prove that the isotropic Gaussian model satisfies Definition 2 using few samples.

Lemma 2. *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *have i.i.d. rows* $\sim \mathcal{N}(\mathbb{0}_d, \frac{1}{n}\mathbf{I}_d)$*. For* $\epsilon, \delta \in (0,1)$*, if*

$$n \ge C \cdot \frac{k\log(\frac{d}{k}) + \log(\frac{1}{\delta})}{\epsilon^2},$$

for an appropriate constant C*, then* $\mathbf{A}$ *satisfies* (ϵ, k)*-RIP with probability* $\ge 1-\delta$*.*

Proof. To prove Definition 2, it suffices to show that for each $T \subseteq [d]$ with $|T| \le k$, we have $\|[\mathbf{A}^\top\mathbf{A}]_{T\times T} - \mathbf{I}_T\|_{\mathrm{op}} \le \epsilon$, where $\mathbf{I}_T$ is shorthand for the identity matrix restricted to coordinates in T, and $\mathbf{M}_{T\times T}$ denotes the submatrix of $\mathbf{M}$ with rows and columns T. Note that there are $\binom{d}{k}$ such matrices, since it is enough to consider T with $|T| = k$, as this bounds the operator norm of all smaller submatrices as well. For each such T, let

$$S_T := \left\{x \in \mathbb{R}^d \mid \|x\|_2 = 1,\ \mathrm{supp}(x) \subseteq T\right\}$$

be the surface of the unit ball in $\mathbb{R}^T$. Moreover, let $N_T \subset S_T$ be a set such that

$$\max_{x\in S_T}\min_{v\in N_T}\|v-x\|_2 \le \frac{1}{4}.$$

Lemma 4, Part VIII shows that to prove $\|[\mathbf{A}^\top\mathbf{A}]_{T\times T} - \mathbf{I}_T\|_{\mathrm{op}} \le \epsilon$, it suffices to show that for all $v \in N_T$, we have $|v^\top(\mathbf{A}^\top\mathbf{A} - \mathbf{I}_d)v| \le \frac{\epsilon}{3}$. Moreover, Lemma 3, Part VIII proves that N_T exists satisfying the above display with $|N_T| \le 9^k$. Therefore, it suffices to prove that for the $\binom{dk}{\cdot}9^k$ different vectors $v \in N_T$ for some $T \subseteq [d]$, $|T| = k$, we have

$$1-\epsilon \le \|\mathbf{A}v\|_2^2 \le 1+\epsilon.$$

By taking n as specified, the result follows from the Johnson-Lindenstrauss lemma (Corollary 1, Part V), with failure probability set to $\frac{\delta}{9^k}\cdot\binom{d}{k}^{-1}$, where we use $\log\binom{d}{k} = O(k\log\frac{d}{k})$. □

Beyond Gaussian models, a variety of random design matrices $\mathbf{A}$ have been shown to satisfy RIP using $n \approx k$ rows, including Rademacher matrices and subsampled Fourier matrices. These design matrices are reasonable in settings such as signal processing, where we have a choice in what linear measurements we apply to the sparse signal of interest, e.g. Fourier measurements of images in MRI. Intuitively, all of these RIP design matrices have rows which are spanned by a few elements of

an orthonormal basis that is entirely non-sparse in the standard coordinate system, so their kernels also do not contain sparse vectors. There is a perspective which views these opposing orthonormal bases (i.e. the standard basis and a basis spanning $\mathbf{A}$'s rows) as inducing an *uncertainty principle*, which states that no vector is sparse in both bases. For a formalization of why uncertainty principles lead to tractable sparse recovery, we refer the reader to [DS89, Moi18].

In the remainder of the section, we develop a more continuous variant of the RIP definition, which is in some sense discrete because of its imposition of exact sparsity. We instead use the following definition as a continuous proxy for sparsity, motivated by the observation (4).

Definition 3 (Numerical sparsity)**.** *We say $x \in \mathbb{R}^d$ is* k-numerically sparse *if*

$$\mathrm{NS}(x) := \frac{\|x\|_1^2}{\|x\|_2^2} \le k.$$

Because of (4), all truly k-sparse vectors are also k-numerically sparse. However, Definition 3 allows for some flexibility, e.g. x may have many nonzero coordinates, as long as they are sufficiently small. To relate Definition 3 to the standard notion of sparsity, the following decomposition is helpful.

Lemma 3. *Let $x \in \mathbb{R}^d$ have $\mathrm{NS}(x) \le k$. For any $C > 1$, we can write $x = \sum_{i\in[m]} v_i$ for $\{v_i\}_{i\in[m]} \subset \mathbb{R}^d$ with disjoint supports, such that $\mathrm{nnz}(v_i) \le Ck$ for all $i \in [k]$, and*

$$\sum_{i=2}^{m} \|v_i\|_2 \le \frac{1}{\sqrt{C}} \|x\|_2 .$$

Proof. Consider a greedy decomposition of x which, starting from $i \leftarrow 1$, lets v_i be x restricted to its Ck largest remaining coordinates by magnitude, and updates $x \leftarrow x - v_i$, until $x = \mathbb{0}_d$. Because every coordinate of v_i is larger than every coordinate of v_{i+1} by magnitude, we have

$$\|v_{i+1}\|_2 \le \sqrt{Ck}\, \|v_{i+1}\|_\infty \le \frac{1}{\sqrt{Ck}} \|v_i\|_1 \text{ for all } i \in [m-1].$$

Summing this equation for all $i \in [m-1]$, and using that the $\{\mathrm{supp}(v_i)\}_{i\in[m]}$ are disjoint,

$$\sum_{i=2}^{m} \|v_i\|_2 \le \frac{1}{\sqrt{Ck}} \|v\|_1 \le \frac{1}{\sqrt{C}} \|v\|_2 ,$$

where the last inequality used that $\mathrm{NS}(x) \le k$ by assumption. □

The decomposition in Lemma 3 is sometimes referred to as a shelling decomposition, because it recursively partitions a vector into sparse shells. By using this decomposition, we can show that RIP matrices satisfy a related definition we call restricted well-conditioning.[3]

Definition 4 (Restricted well-conditioned)**.** *We say $\mathbf{A} \in \mathbb{R}^{n\times d}$ is* (ϵ, k)-restricted well-conditioned*, or* (ϵ, k)*-RWC, if for all $x \in \mathbb{R}^d$ with $\mathrm{NS}(x) \le k$,*

$$(1-\epsilon)\|x\|_2^2 \le \|\mathbf{A}x\|_2^2 \le (1+\epsilon)\|x\|_2^2 .$$

Clearly, Definition 4 implies Definition 2, because all sparse vectors are numerically sparse (see (4)). We briefly formalize the fact that RWC is also a quantitative strengthening of NSP.

Lemma 4. *For $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\epsilon \in (0,1)$, if $\mathbf{A}$ is $(\epsilon, 4k)$-RWC, it also satisfies k-NSP.*

Proof. Suppose for contradiction that $\mathbf{A}v = \mathbb{0}_n$ for $v \ne \mathbb{0}_d$, yet $\|v_S\|_1 \ge \|v_{S^c}\|_1$ for some $S \subseteq [d]$ with $|S| \le k$. We observe that this implies $\mathrm{NS}(v) \le 4k$:

$$\|v\|_1 \le 2\|v_S\|_1 \le 2\sqrt{k}\,\|v_S\|_2 \le 2\sqrt{k}\,\|v\|_2 . \tag{5}$$

However, by RWC this also means $\|\mathbf{A}v\|_2^2 \ge (1-\epsilon)\|v\|_2^2 > 0$, a contradiction. □

By applying Lemma 3, we now relate Definitions 2 and 4.

Lemma 5. *For $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\epsilon \in (0,1)$, if $\mathbf{A}$ satisfies $(\frac{\epsilon}{5}, \frac{25k}{\epsilon^2})$-RIP,[4] it is also (ϵ, k)-RWC.*

[3] Our Definition 4 is related to the restricted well-conditioning definition in [ANW10], which implies ours.

[4] These constants are not optimized for ease of exposition, and can be sharpened.

Proof. Let $\mathrm{NS}(x) \le k$, and let $\{v_i\}_{i\in[m]}$ all be $\frac{25k}{\epsilon^2}$-sparse with disjoint supports, so that $\sum_{i\in[m]} v_i = x$ and $\sum_{i=2}^m \|v_i\|_2 \le \frac{\epsilon}{5}\|x\|_2$, from Lemma 3. For the upper bound, we have by applying RIP,

$$\begin{aligned}\|\mathbf{A}x\|_2^2 &\le \left(1+\frac{\epsilon}{5}\right)\|\mathbf{A}v_1\|_2^2 + \left(1+\frac{5}{\epsilon}\right)\left\|\sum_{i=2}^m \mathbf{A}v_i\right\|_2^2 \\ &\le \left(1+\frac{\epsilon}{5}\right)^2\|v_1\|_2^2 + \frac{6}{\epsilon}\left(\sum_{i=2}^m \|\mathbf{A}v_i\|_2\right)^2 \\ &\le \left(1+\frac{\epsilon}{5}\right)^2\|x\|_2^2 + \frac{6}{\epsilon}\cdot\frac{\epsilon^2}{25}\sqrt{1+\epsilon}\,\|x\|_2^2 \le (1+\epsilon)\|x\|_2^2.\end{aligned}$$

Similarly, the lower bound follows from

$$\begin{aligned}\|\mathbf{A}x\|_2^2 &\ge \left(1-\frac{\epsilon}{5}\right)\|\mathbf{A}v_1\|_2^2 + \left(1-\frac{5}{\epsilon}\right)\left\|\sum_{i=2}^m \mathbf{A}v_i\right\|_2^2 \\ &\ge \left(1-\frac{\epsilon}{5}\right)^2\|v_1\|_2^2 - \frac{5}{\epsilon}\left(\sum_{i=2}^m \|\mathbf{A}v_i\|_2\right)^2 \\ &\ge \left(1-\frac{\epsilon}{5}\right)^4\|x\|_2^2 - \frac{5}{\epsilon}\cdot\frac{\epsilon^2}{25}\sqrt{1+\epsilon}\,\|x\|_2^2 \ge (1-\epsilon)\|x\|_2^2.\end{aligned}$$

□

Lemma 5 states that for constant ϵ, we can transfer RIP (approximate isometry on sparse vectors) to RWC (approximate isometry on numerically sparse vectors) with constant overhead in the parameters. This also yields as a corollary of Lemma 4 the fact that $(\epsilon, O(k))$-RIP implies k-NSP, for any sufficiently small constant ϵ. In summary, RIP and RWC imply each other, and both imply NSP, where all statements hold with potentially a constant factor loss in parameters.

Finally, we generalize Theorem 1 to the noisy setting (1), under the RWC assumption.

Theorem 2 (Noise-tolerant basis pursuit)**.** *Consider an instance of the problem (1), where $\|\xi\|_2 \le \Delta$. If $\mathbf{A}$ is $(\epsilon, 4k)$-RWC for any $\epsilon \in (0, \frac{1}{2}]$, then $\|x - x^\star\|_2 \le 4\Delta$, where*

$$x := \operatorname{argmin}_{\substack{x\in\mathbb{R}^d \\ \|\mathbf{A}x-b\|_2\le\Delta}} \|x\|_1.$$

Proof. Because $x^\star$ is feasible, we have $\|x\|_1 \le \|x^\star\|_1$, so Lemma 1 and (5) imply that $v := x - x^\star$ has $\mathrm{NS}(v) \le 4k$. Therefore, by RWC and the assumption $\|\mathbf{A}x - b\|_2$, we have the desired

$$\|x - x^\star\|_2^2 \le \frac{1}{1-\epsilon}\|\mathbf{A}(x-x^\star)\|_2^2 \le \frac{4}{1-\epsilon}\left(\|\mathbf{A}x-b\|_2^2 + \|\mathbf{A}x^\star - b\|_2^2\right) \le \frac{8\Delta^2}{1-\epsilon} \le 16\Delta^2.$$

□

Theorem 2 shows we can extend basis pursuit to solve noisy sparse recovery in polynomial time, recovering the true parameter vector up to a distance scaling linearly in the amount of noise added.

Remark 2. *In situations where it is important to output proper hypotheses, i.e. estimates x which are truly k-sparse, we can take any hypothesis $\hat{x}$ and truncate it to its largest k coordinates. This will at most double the distance to $x^\star$, due to the observation*

$$\begin{aligned}\hat{x}_{S_k(\hat{x})} &= \operatorname{argmin}_{\substack{x\in\mathbb{R}^d \\ \mathrm{nnz}(x)\le k}} \|x - \hat{x}\|_2 \\ \implies \left\|\hat{x}_{S_k(\hat{x})} - x^\star\right\|_2 &\le \left\|\hat{x} - \hat{x}_{S_k(\hat{x})}\right\|_2 + \|\hat{x} - x^\star\|_2 \le 2\|\hat{x} - x^\star\|_2,\end{aligned}$$

where we let $S_k(\hat{x}) \subseteq [d]$ denote the indices of $\hat{x}$'s k largest coordinates by magnitude.

In the remainder of the lecture, to be consistent with Remark 2, we let $S_k(v)$ return the indices of the k largest coordinates of a vector v by magnitude, breaking ties arbitrarily.

3 Projected gradient descent

In this section, we develop a simple first-order method which qualitatively matches the guarantees of Theorem 2, but which can be implemented to run in nearly-linear time. We mention that under the RIP (or RWC) assumption, there are a host of efficient first-order methods which achieve this type of guarantee, including greedy combinatorial methods such as matching pursuit and its variants [MZ93, PRK93, NV10], and iterative methods which directly analyze progress on a nonconvex objective (i.e. over the set of sparse vectors) [NT09, BD09, MD10, Fou11].

The algorithm we present in this section is inspired by several of these first-order algorithms, but its analysis is closer in spirit to the projected gradient descent methods we have seen in convex settings throughout earlier lectures. It also was extended recently by [KLL+23b] to robustly handle a certain type of semi-random adversarial noise in the observations, giving some credence to its general flexibility as a framework for sparse recovery. To motivate it, note that half the gradient at x of the least-squares objective in $\mathbf{A}$ (i.e. Eq. (1), Part VIII) is $\mathbf{A}^\top(\mathbf{A}x - b)$. Now, let us consider the noiseless setting $b = \mathbf{A}x^\star$ for simplicity, and suppose $\mathbf{A} = \frac{1}{\sqrt{n}}\mathbf{G}$ where $\mathbf{G} \in \mathbb{R}^{n\times d}$ is entrywise i.i.d. $\sim \mathcal{N}(0,1)$. In this case, if $x, x^\star$ is independent of $\mathbf{G}$, letting $\{g_i\}_{i\in[n]}$ be the rows of $\mathbf{G}$,

$$\begin{aligned}\mathbf{A}^\top(\mathbf{A}x - b) &= \mathbf{A}^\top\mathbf{A}(x - x^\star)\\ &= \frac{1}{n}\sum_{i\in[n]} \langle g_i, x - x^\star\rangle\, g_i.\end{aligned}$$

Next, consider a single one of these summands. Writing $v := x - x^\star$ and supposing $\|v\|_2 = 1$ for simplicity, we can decompose $g_i = \xi_i v + h_i$, where $\xi \sim \mathcal{N}(0,1)$ and $h_i \sim \mathcal{N}(\mathbb{0}_d, \mathbf{I}_d - vv^\top)$, i.e. we separate out the components of g_i in the v direction and in the orthogonal subspace. Then,

$$\frac{1}{n}\sum_{i\in[n]} \langle g_i, v\rangle\, g_i = \left(\frac{1}{n}\sum_{i\in[n]} \xi_i^2\right) v + \frac{1}{n}\sum_{i\in[n]} \xi_i h_i.$$

Viewing each $\xi_i h_i$ as essentially an independent Gaussian vector, we expect the maximum coordinate of $\frac{1}{n}\sum_{i\in[n]}\xi_i h_i$ to scale as $\frac{1}{\sqrt{n}}$, and moreover $\frac{1}{n}\sum_{i\in[n]}\xi_i^2$ is very tightly concentrated around 1. The point of this digression is that we showed that in this special case, the gradient $\mathbf{A}^\top(\mathbf{A}x - b)$ is highly-correlated with v, the desired descent direction, up to a "flat" additive term, i.e. a noise component which is small coordinatewise. We view the role of ℓ_1 projection in gradient descent as denoising this flat term. We begin by making this decomposition of the gradient rigorous.

Lemma 6. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $v \in \mathbb{R}^d$. If $\mathbf{A}$ is $(\frac{1}{2}, k)$-RWC and $\mathrm{NS}(v) \le k$,*

$$\left\|g_{S_k(g)}\right\|_2 \le \frac{3}{2}\|v\|_2, \text{ for } g := \mathbf{A}^\top\mathbf{A}v.$$

Proof. Let $h = g_{S_k(g)}$ be k-sparse, so our goal is to show $\|h\|_2 \le \frac{3}{2}\|v\|_2$. We have

$$\begin{aligned}\|h\|_2^2 = \langle h, g\rangle &= \left\langle h, \mathbf{A}^\top\mathbf{A}v\right\rangle\\ &\le \|\mathbf{A}h\|_2\,\|\mathbf{A}v\|_2 \le \frac{3}{2}\|h\|_2\,\|v\|_2,\end{aligned}$$

where we used the RWC assumption and $\mathrm{NS}(h), \mathrm{NS}(v) \le k$. □

Lemma 6 shows that if $x - x^\star$ is numerically sparse, then the gradient step $\mathbf{A}^\top(\mathbf{A}x - b) = \mathbf{A}^\top\mathbf{A}(x - x^\star)$ can be decomposed into a "short" component with ℓ_2 norm $O(\|v\|_2)$, and a "flat" component with ℓ_∞ norm $O(\frac{1}{\sqrt{k}}\|v\|_2)$. This latter claim is because every coordinate outside the top k coordinates of g by magnitude is smaller than $\frac{1}{\sqrt{k}}$ times the ℓ_2 norm of the top k coordinates. We now show how to turn this observation into an efficient algorithm based on projected gradient descent.

We begin by describing one phase of the algorithm. Let $x_0 \in \mathbb{R}^d$ be k-sparse, and assume we know an upper bound $R \ge \|x_0 - x^\star\|_2$. We show how to produce a k-sparse point $\hat{x} \in \mathbb{R}^d$ such that $\|\hat{x} - x^\star\|_2 \le \frac{R}{2}$ in $O(nd)$ time.[5] Our main helper tool is the following one-step analysis.

[5] The algorithm technically runs in $O(\mathrm{nnz}(\mathbf{A}))$ time, but RIP (and hence RWC) matrices are typically dense.

Proposition 1. *In the setting of (2), let $x_0 \in \mathbb{R}^d$ have $\mathrm{nnz}(x_0) \le k$ and $\|x_0 - x^\star\|_2 \le R$. Suppose $\mathbf{A}$ is $(\frac{1}{2}, 2^{16}k)$-RWC, and $x \in \mathbb{R}^d$ satisfies*

$$\|x - x^\star\|_2 \le R,\ x \in \mathcal{X} := \left\{x \in \mathbb{R}^d \mid \|x - x_0\|_1 \le \sqrt{2k}R\right\}.$$

Then, one of the following two situations must occur, where we denote $v := x - x^\star$.

1. *$\|\mathbf{A}v\|_2^2 \ge \frac{R^2}{32}$ and $\mathbf{A}^\top \mathbf{A}v = s + f$ for some $s, f \in \mathbb{R}^d$ with*

$$\|s\|_2 \le \frac{3R}{2},\ \|f\|_\infty \le \frac{3}{2^9\sqrt{k}}R.$$

2. *$\|v\|_2 \le \frac{R}{4}$.*

In the former case, letting $\eta := \frac{1}{180}$ and $x' \leftarrow \mathrm{argmin}_{x' \in \mathcal{X}}\{\langle \eta g, x' - x\rangle + \frac{1}{2}\|x' - x\|_2^2\}$, we have

$$\|x' - x^\star\|_2^2 \le \|x - x^\star\|_2^2 - \frac{R^2}{14400}.$$

Proof. Note that sparsity of x_0 and $x^\star$ implies that $x^\star \in \mathcal{X}$, by using (4). Next, if it is not the case that $\|v\|_2 \le \frac{R}{4}$, then since $\|v\|_1 \le \|x_0 - x^\star\|_1 + \|x_0 - x\|_1 \le \sqrt{32k}R$, it is the case that $\mathrm{NS}(v) \le 32k$. Therefore, the fact that $\mathbf{A}$ is RWC with the given parameters implies $\|\mathbf{A}v\|_2^2 \ge \frac{1}{2}\|v\|_2^2 \ge \frac{R^2}{32}$, as claimed. Further, applying Lemma 6 with $k \leftarrow 2^{16}k$, and letting s be the $2^{16}k$ largest coordinates of g by magnitude and $f := g - s$, we have

$$\|f\|_\infty \le \min_{i \in \mathrm{supp}(s)} |s_i| \le \frac{1}{2^8\sqrt{k}}\|s\|_2 \le \frac{3}{2^9\sqrt{k}}R.$$

This proves our first claim. The first-order optimality condition on x' (see e.g. the proof of Theorem 2, Part II) then implies, letting $v' := x' - x^\star$, that

$$\begin{aligned}
\|v\|_2^2 - \|v'\|_2^2 &\ge \langle 2\eta g, x - x^\star\rangle + \langle 2\eta g, x' - x\rangle + \|x - x'\|_2^2 \\
&\ge 2\eta\|\mathbf{A}v\|_2^2 + \langle 2\eta f, x' - x\rangle + \langle 2\eta s, x' - x\rangle + \|x - x'\|_2^2 \\
&\ge \frac{\eta R^2}{16} - 2\eta\|f\|_\infty\|x' - x\|_1 - \eta^2\|s\|_2^2 \\
&\ge \frac{\eta R^2}{40} - \frac{9\eta^2 R^2}{4} \ge \frac{R^2}{14400}.
\end{aligned}$$

where we used $\|x' - x\|_1 \le \sqrt{8k}R$ since $x, x' \in \mathcal{X}$, and plugged in our choice of η. □

By iterating on Proposition 1, we can implement the phase described earlier.

Corollary 1. *In the setting of (2), let $x_0 \in \mathbb{R}^d$ have $\mathrm{nnz}(x_0) \le k$ and $\|x_0 - x^\star\|_2 \le R$, and suppose $\mathbf{A}$ is $(\frac{1}{2}, 2^{16}k)$-RWC. There is an algorithm which produces $\hat{x} \in \mathbb{R}^d$ satisfying $\mathrm{nnz}(\hat{x}) \le k$ and $\|\hat{x} - x^\star\|_2 \le \frac{R}{2}$, in time $O(nd)$.*

Proof. We simply iterate the procedure in Proposition 1, until the condition in Item 1 fails and hence we can conclude we are in the case of Item 2. Note that we can verify if Item 1 holds in every iteration, since $\mathbf{A}v = \mathbf{A}x - b$, which we can compute in $O(\mathrm{nnz}(\mathbf{A}))$ time. Moreover, we will reach the case of Item 2 in $O(1)$ iterations, since each iteration makes $\Omega(R^2)$ progress on the squared distance to $x^\star$. Finally, once we are in the latter case, we can truncate our final iterate to its top k coordinates, which at most doubles the distance to $x^\star$ by applying Remark 2. □

Finally, by iterating on Corollary 1, we obtain a high-precision solver for (2).

Theorem 3 (Sparse recovery via projected gradient descent)**.** *In the setting of (2), let $\|x^\star\|_2 \le R$, and suppose $\mathbf{A}$ is $(\frac{1}{2}, 2^{16}k)$-RWC. There is an algorithm which produces $\hat{x} \in \mathbb{R}^d$ satisfying $\mathrm{nnz}(\hat{x}) \le k$ and $\|\hat{x} - x^\star\|_2 \le r$, in time*

$$O\left(nd\log\left(\frac{R}{r}\right)\right).$$

Proof. It suffices to recursively apply the algorithm in Corollary 1, $\log_2(\frac{R}{r})$ times. □

We note that the strategy in Theorem 3 generalizes to handle the noisy setting (1), recovering $x^\star$ up to distance $O(\Delta)$, similarly to Theorem 2. This was shown in [KLL+23b], which, as mentioned previously, extended this analysis to handle a wider family of matrices $\mathbf{A}$ perturbed by an adversary which can augment $\mathbf{A}$ with additional rows, potentially destroying the restricted isometry property. In this endeavor, the short-flat decomposition strategy of Proposition 1 was crucial, because it gives verifiable conditions (the size of residuals and the existence of a decomposition) which allow one to provably make progress, whereas checking for RIP is computationally hard.

Remark 3. *In addition to sparse recovery (i.e. underconstrained linear regression), the frameworks described in this lecture extend to varying degrees to solve a variety of other linear inverse problems with structural assumptions on the solution, including low-rank matrix generalizations of compressed sensing, and recovering permutation or orthogonal matrices [ANW10, CRPW12].*

One particularly challenging variation of linear inverse problems is the matrix completion problem, where we are given random observations of a low-rank matrix, and wish to recover it. For example, the Netflix Prize was a famous competition for designing collaborative filtering algorithms [SN07], an instance of matrix completion [RS05]. The key challenge in this setting is that a "for all low-rank matrices" type of statement such as RIP (which preserves the norm for all sparse vectors) cannot hold for the linear measurements taken by matrix completion, as we can always plant a 1*-sparse, rank-*1 *matrix which entirely dodges the measurements, and hence its norm will not be preserved. Under additional structural assumptions on the target matrix, such as the popular notion of incoherence, a sequence of works [CT10, KMO10, CR12] ultimately showed that rank-k, $d \times d$ matrices can be recovered using $\approx dk$ observations [Rec11], which can be far fewer than d^2 when k is small. We mention that the short-flat decomposition strategy of this section was recently extended to the matrix completion case [KLL+23a], giving a nearly-linear time algorithm with the state-of-the-art sample complexity and recovery rate among such algorithms in theory. However, it remains to be seen whether such frameworks lead to improved performance in practice.*

Source material

Portions of this lecture are based on reference material in [Moi18, Pri20], as well as the author's own experience working in the field.

References

[ANW10] Alekh Agarwal, Sahand N. Negahban, and Martin J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010*, pages 37–45. Curran Associates, Inc., 2010.

[BD09] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

[BDMS13] Afonso S. Bandeira, Edgar Dobriban, Dustin G. Mixon, and William F. Sawin. Certifying the restricted isometry property is hard. *IEEE Trans. Inf. Theory*, 59(6):3448–3450, 2013.

[CDD09] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22:211–231, 2009.

[CDS01] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[CR05] Emmanuel J. Candès and Justin K. Romberg. Signal recovery from random projections. In *Computational Imaging III, 2005*, volume 5674 of *SPIE Proceedings*, pages 76–86. SPIE, 2005.

[CR12] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, 2012.

[CRPW12] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[CRT06] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

[CT05] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, 2005.

[CT06] Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, 2006.

[CT10] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[Don06] David L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[DS89] David L. Donoho and Philip B. Stark. Uncertainty principles and signal recovery. *SIAM J. Appl. Math*, 49(3):906–931, 1989.

[Fou11] Simon Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.

[KKMR21] Jonathan A. Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. On the power of preconditioning in sparse linear regression. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021*, pages 550–561. IEEE, 2021.

[KLL+23a] Jonathan A. Kelner, Jerry Li, Allen Liu, Aaron Sidford, and Kevin Tian. Matrix completion in almost-verification time. In *64th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2023*, pages 2102–2128. IEEE, 2023.

[KLL+23b] Jonathan A. Kelner, Jerry Li, Allen Liu, Aaron Sidford, and Kevin Tian. Semi-random sparse recovery in nearly-linear time. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 2352–2398. PMLR, 2023.

[KMO10] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[MD10] Arian Maleki and David L Donoho. Optimally tuned iterative reconstruction algorithms for compressed sensing. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):330–341, 2010.

[Moi18] Ankur Moitra. *Algorithmic Aspects of Machine Learning*. Cambridge University Press, 2018.

[MZ93] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

[NT09] Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and computational harmonic analysis*, 26(3):301–321, 2009.

[NV10] Deanna Needell and Roman Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE Journal of selected topics in signal processing*, 4(2):310–316, 2010.

[Pri20] Eric Price. *Sparse Recovery*. 2020.

[PRK93] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993.

[Rau10] Holger Rauhut. Compressive sensing and structured random matrices. *Radon Series Comp. Appl. Math*, 9:1–92, 2010.

[Rec11] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

[RS05] Jason D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005)*, volume 119 of *ACM International Conference Proceeding Series*, pages 713–719. ACM, 2005.

[SN07] ACM SIGKDD and Netflix. Proceedings of kdd cup and workshop. https://www.cs.uic.edu/ liub/KDD-cup-2007/proceedings.html, 2007. Accessed: 2023-04-01.

[TP14] Andreas M. Tillmann and Marc E. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inf. Theory*, 60(2):1248–1259, 2014.