Sparse & Redundant Representations in Computer Vision and Machine Learning.

Course Notes - Fall '19. - Johns Hopkins University.

Jeremias Sulam.

Chapter 1: Underdetermined Systems of Equations.

Linear Algebra provides a series of results that are profound, timeless, precise - and they set the basis for many engineering solutions and scientific advancement. Thus, it might appear surprising that it contains an elementary problem that only recently has been well understood, and that it continues to drive much of current research.

Consider a matrix $A \in \mathbb{R}^{n \times m}$, with $n < m$, and the system $b = Ax$. What is the solution (x) to this linear system?

a) Option 1: If b is not in the span of the columns of A, then there's no solution

b) If $b \in \text{span}(A)$, then $\exists$ infinite many $x : Ax = b$.

To avoid the issue of no solution, we'll assume hereafter that A is full rank: What is rank?
rank $(A) = n \Rightarrow$ so that span $(A) = \mathbb{R}^n$, and so $b \in$ span (A).

This underdetermined problem is very common - consider the observation of a subsampled signal; ie $A = \begin{bmatrix} 1 0 0 \cdots \\ 0 0 1 0 \cdots \\ 0 0 0 0 1 \cdots \end{bmatrix}$. There $\exists$ infinitely many ways ("x") to "fill-in the blanks".

These are "ill-posed" inverse problems, and to address them, it is typical to introduce Regularization - regularizing the solution. - or "narrowing down" the space of possible solutions. A common way to do this is through a function, that evaluates how "desirable" or useful a candidate x is. Thus, define $J(x)$.

$$(P_J) : \min_x J(x) \text{ s.t. } b = Ax.$$

Great - but how do we choose $J(x)$? A traditional route is to prefer solutions with small l_2-norm. What is an lp-norm?

$$(P_{l_2}) : \min_x \|x\|_2^2 \text{ s.t. } b = Ax.$$

Let's solve this through Lagrange Multipliers - What are Lagrange Multipliers?

Define $\mathcal{L}(x) = \|x\|_2^2 + \lambda^T(Ax-b)$.

then: $\frac{\partial \mathcal{L}}{\partial \lambda} = Ax - b = 0$ — of course. — (1)

$$\frac{\partial \mathcal{L}}{\partial x} = 2x + A^T\lambda = 0. \Rightarrow x^* = -\frac{1}{2} A^T \lambda \qquad (2)$$

from (1), $Ax^* = -\frac{1}{2} AA^T\lambda = b. \Rightarrow \lambda^* = -2\underbrace{(AA^T)}^{-1} b.$

is AA^T invertible?

Thus, x^*, from (2), $\boxed{x^* = \underbrace{A^T(AA^T)^{-1}} b}$

pseudo-inverse or "Moore-Penrose" inverse, A^+

Here, A^+ is a "right-pseudo" inverse, as $AA^+ = AA^T(AA^T)^{-1} = I$.

When could I compute a left pseudo-inverse?

The $\|\cdot\|_2^2$ is widely spread precisely because of this: it often leads to closed-form solutions (simple).

Is the solution of (P_2) any good then? It gives us a unique solution, so is this the end of the course? In fact, this is also true for any strictly convex function $J(x)$. What's convexity?

Convex Set: A set Ω is convex if $\forall\, x_1, x_2 \in \Omega$, $x_1 \cdot t + (1-t)\, x_2 \in \Omega$

$\forall\, t \in [0,1]$. $\Rightarrow$ (Q.) is the set $\{x : Ax = b\}$ convex?

Convex Function: A function $J(x): \Omega \to \mathbb{R}$ is convex is $\forall x_1, x_2 \in \Omega$ and $t \in [0,1]$,

$$J(t x_1 + (1-t)\, x_2) \leq t\, J(x_1) + (1-t)\, J(x_2).$$

(Q): show that l_2 is convex.

$J(x)$ is strictly convex if the inequality is strictly $<$.

All strictly convex functions have a unique minimizer. – and the constraint set was convex – guaranteeing a unique solution for x^*.

So why use l_2? All p-norm (l_p) are convex, so how about others?

→ Consider adding an example of getting observations $Ax^* = b$, and trying to recover x^* with $x_{l.s.}$, and showing that they are different. "So what can we do"?

Looking at the l_1 case: Consider $J(x) = \|x\|_1$ then,

$$(P_1): \min_x \|x\|_1 \ \text{s.t.}\ Ax = b.$$

and bounded! if $\hat{x}_1$ solution $\Rightarrow$ $\upsilon = \|x_1\|_1$
other $\|x_2\|_1$: $\|x_1 - x_2\|_1 \leq 2\upsilon$

(P_1) is convex, alas not strictly so. So the solution might not be unique.

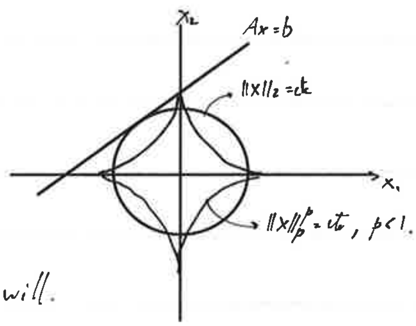
(Q): Is the set of solutions to (P_1) convex? Show.

Nonetheless, even if there might be possibly infinitely many solutions, all these live in a set that is convex and bounded - i.e: they are all "nearby".

Importantly, using $J = \|\cdot\|_1$ promotes sparse solutions. Why does this happen? While no precise relationship between the ℓ_1 and level of sparsity of solutions can be expressed in general, we can gain intuition from a geometric and analytical perspective. (maybe leave the analytical one as optional).

Geometrically, the linear system $Ax=b$ defines an affine subspace. (i.e: the solutions live in an affine subspace: a solution x_0 plus any vector $h \in Null(A)$ provides a feasable solution, too. This is a hyperplane in $\mathbb{R}^{m-n}$ embedded in $\mathbb{R}^m$.

The solution to $\left[\min_x \|x\|_p^p \text{ s.t. } Ax=b\right]$ is then obtained by "inflating" the ℓ_p-ball until it intersects the constraint set - the hyperplane. As can be seen, the ℓ_2 will not provide sparse solution, whereas ℓ_p with $0 \leq p \leq 1$, will.



For an analytical motivation, look at pg. 8-9 in Miki's book, showing that "for a pair of ℓ_p-ℓ_q norms with $q<p$, a unit length ℓ_p-norm vector becomes the shortest in ℓ_q when it is the sparsest possible." For an optimality perspective, look at Mairal's monograph - pg. 16.

$\Rightarrow$ To complement the previous example, run now the solution to (P_1) to demonstrate recovery of the original x^* $\Rightarrow$ show how to turn this to linear programming.

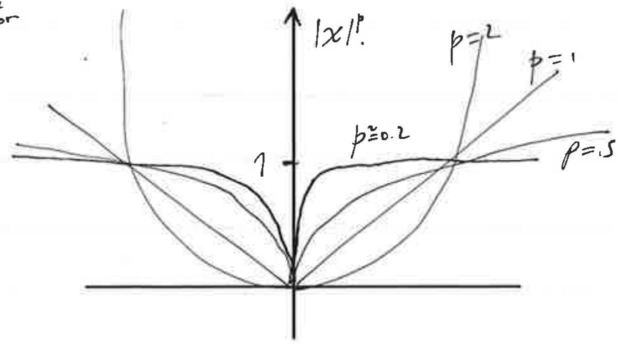
So, it seems that if we are after sparsity, we should use ℓ_p with $p<1$. The problem, however, is that this now becomes a non-convex problem - very hard to solve! However, we should not retreat to the "comfortable" ℓ_2, but rather find way of still producing solutions to this tough problem.

The l_0-norm: So it seems that we should consider, in an ideal case, the l_0-pseudo norm:

$$\|x\|_0 = \lim_{p \to 0} \|x\|_p^p = \lim_{p \to 0} \sum_{i=1}^{m} |x_i|^p = \# \{ i : x_i \neq 0 \}.$$

Note that $\|x\|_p^p$ is not a norm for $p < 1$. Why is l_p not a norm for $p < 1$?

Indeed, as $p \to 0$, $|x|^p$ becomes an indicator function, and this represents some sort of ideal interest; "searching for the sparsest solution":



(P_0): $\min_x \|x\|_0$ s.t. $Ax=b$.

However simple to understand, this problem is hard — and often disregarded as too challenging for practitioners and thus replaced by convex alternatives (such as the l_1). Note that the non-convexity and non-smoothness of the l_0 make (P_0) particularly problematic. Consider even the simple question:

- Is there a unique solution to (P_0)?
- Say I give you a candidate solution — can you check if this is the sparsest one?

(P_0) is basically a combinatorial problem: to solve it, one should explore all possible solutions with $\|x\|_0 = 1$. If none satisfy the system $Ax=b$, one should move on to all combinations of 2 columns for A, i.e: $\|x\|_0 = 2$, etc.

This is intractable. Say $n = 500$, $m = 2000$, and say the solution x^*: $\|x^*\|_0 = 20$. One thus have to sweep through all $\binom{m}{20} \simeq 3.9 \cdot 10^{47}$ options. If each test takes 1 nanosecond ($1e^{-9}$), this would still take $1.2\ 10^{31}$ years!

These examstive-search approaches have a complexity which is exponential in m. Indeed, (P_0) has been proven to be NP-hard. (non-deterministic polynomial time)
↳ "as hard as the hardest problem in NP" ↳ non-deterministic polynomial time.

Despite these difficulties, could we maybe find an approximate solution? Under what conditions? and how accurately and efficiently? These are the questions explored in the first part of the course.

$\Rightarrow$ To conclude and contrast with the previous cases (l_2 and l_1), run the same case example with l_0 (IHT or OMP) to show recovery.

L_1 via LP:

Given $b = Ax$, let $x = u - v$, $u_i, v_i \geq 0 \ \forall i$.

$$\Rightarrow \quad z = \begin{bmatrix} u \\ v \end{bmatrix}, \quad \tilde{A} = [A, -A] \quad \Rightarrow \quad Ax = \tilde{A}z = b.$$

Now,
$$\min_x \|x\|_1 = \min_z \|z\|_1 = \min_z \mathbb{1}^T z.$$

$$\Rightarrow \boxed{\min \ \mathbb{1}^T z \quad \text{s.t.} \quad \begin{aligned} z_i &\geq 0 \\ \tilde{A}z &= b. \end{aligned}}$$

solution $\hat{x} = \hat{u} - \hat{v}$, $\hat{z} = \begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix}$.