Class 3: Greedy Algorithms. – "Greed is Good": J.T.

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

or (P_0): $\min \|Ax-b\|_2^2$ s.t. $\|x\|_0 \leq k$.

One can decompose this problem in two parts:

a) Identify the support

b) Find coefficients.

Note that b is simple! indeed, if we know $S^* = supp(x^*)$ then

$$x^* = \arg\min_x \|Ax - b\|_2^2 \text{ s.t. } supp(x) = S^*, \ |S| = k.$$

$$= \arg\min_{x_{S^*}} \| A_{S^*} x_{S^*} - b\|_2^2$$

$$\Rightarrow A_S^T (A_S x_S - b) = 0$$

$$\Rightarrow \hat{x}_S = (A_S^T A_S)^{-1} A_S^T b.$$

Least Squares Solution.

So often the main problem is $\boxed{\text{finding the correct support.}}$

Suppose someone gave us $b = \underset{\text{m×n}}{A} x^*$ and told us $\|x^*\|_0 = 1$, and unique. What would we do?

We could perform 'm' tests: $\min_z \underbrace{\|a_j z - b\|_2^2}$ $= \epsilon(j)$.

which is minimized by $\hat{z}_j = \dfrac{a_j^T b}{\|a_j\|_2^2}$.

The error of approximating b with the j'th atom is then

$$\|a_j z_j - b\|_2^2 = \|a_j \frac{a_j^T b}{\|a_j\|_2^2} - b\|_2^2 = \|b\|_2^2 - \frac{(a_j^T b)^2}{\|a_j\|_2^2}.$$

If one of these $\epsilon(j)$ terms become $\epsilon(j^*) = 0 \Rightarrow$ found solution!

This costs $O(mn)$ flops: m 'checks' of $\langle a_j, b\rangle$.

If $\|X^*\|_0 > 1$, however, and no $\epsilon(j)$ is found to be zero, then we should move to all $\binom{m}{k}$ possibilities, which is $O(m^k)$ prohibitive!

Greedy Approaches attemp to (hopefully) find the correct support by building it more or less sequentially.

Orthogonal Matching Pursuit:

→ Choose the next atom that best approximates a residual.

Init: $x^{(0)} = 0$, $r^{(0)} = b - Ax^{(0)} = b$, $S = \text{supp}\{x^{(0)}\} = \emptyset$.

for $k = 1, \ldots, k_0$:

- Find "best" next atom:

$$j_0 = \underset{j}{\arg\min} \| a_j z_j - r^{k-1} \|_2^2 = \underset{j}{\arg\max} \left| \frac{a_j^T r^{k-1}}{\|a_j\|_2^2} \right|$$

- Update support: $S \leftarrow S \cup \{j_0\}$.

- Update solution: $x_S^{(k)} = \underset{x}{\arg\min} \|A_S x - b\|_2^2 = A_S^+ b$.

- Update residual: $r^{(k)} = b - Ax^{(k)}$.

Some observations:

1) Why orthogonal?

Note that by computing $x_s^{(k)} = \underset{x}{\operatorname{argmin}} \|b - A_s x\|_2^2$

$$\Rightarrow A_s^T (A_s x - b) = 0.$$

$$A_s^T r^{(k)} = 0.$$

$\Rightarrow$ residual is orthogonal to selected atoms.

Q: Can an atom be selected twice?

2) Matching Pursuit: (Mallat '93).

solution is updates by: $x_{(j_0)}^{k} = x_{(j_0)}^{k-1} + \underbrace{\frac{a_{j_0}^T r^{k-1}}{\|a_{j_0}\|_2^2}}_{z_{j_0}}$.

which replaces the Least Squares step. Thus cheaper.

Likewise, there are different variations of these methods with different level of complexity. Eg: Least-Squares OMP, where $m - |S^k|$ Least-Squares steps are evaluated to select the next atom - replacing the cheaper inner product in OMP. This, though, is typically much more expensive.

3) Note that normalization does not influence which atoms are selected. Thus, easier to work with normalized atoms.

$\tilde{A} = AW \Rightarrow \tilde{x}^{OMP}, \Rightarrow x = W\tilde{x}$ since

$b = \tilde{A}\tilde{x} = A\underbrace{W\tilde{x}}_{x}$.

How fast does the residual decay? i.e: how fast is an approximation constructed?

We'll look at MP for simplicity, and $\|a_i\|_2 = 1 \ \forall i$.

In MP, $X^k = X^{k-1}$ where $j_0 = \arg\max_j \langle a_j, r^{k-1} \rangle$

$$X^k_{(j_0)} = X^k_{(j_0)} + \underbrace{a_{j_0}^T b}_{z_{j_0}}$$

Then, $r^k = b - AX^k = r^{k-1} - (a_{j_0}^T r^{k-1})\, a_{j_0}$.

Thus,

$$\epsilon(j_0) = \|r^k\|_2^2 = \|r^{k-1}\|_2^2 + (a_{j_0}^T r^{k-1})^2 - 2(a_{j_0}^T r^{k-1})^2.$$

$$= \|r^{k-1}\|_2^2 - (a_{j_0}^T r^{k-1})^2$$

$$= \|r^{k-1}\|_2^2 - \max_{1 \le j \le m} (a_j^T r^{k-1})^2.$$

Define the decaying factor $\delta(A, v) = \max_{1 \le j \le m} \frac{|a_j^T v|^2}{\|v\|_2^2}$.

and the "universal" decaying factor: $\boxed{\delta(A) = \inf_v \delta(A, v)}$

This is the "worst" vector: the one that leads to a smallest decrease in energy of the residual. Thus,

$$\|r^k\|_2^2 = \|r^{k-1}\|_2^2 - \delta(A, r^{k-1}) \cdot \|r^{k-1}\|_2^2.$$

$$\|r^k\|_2^2 \le \|r^{k-1}\|_2^2 (1 - \delta(A)).$$

This leads to $\|r^k\|_2^2 \le (1 - \delta(A))^k \|b\|_2^2 \Rightarrow$ exponential Decay.

...ut now: · Is $\delta(A) > 0$? Yes, because A: full rank.
so no nullspace : $\nexists v \neq 0$: $Av = 0$.

· Best $\delta(A)$? $\delta(A) \leq 1/\sqrt{n}$. (Say, for orthonormal A).

—o—.

So, OMP approximates exponentially fast. But to what?
Do it succeed in recovering the sparsest vector?

· Let $b = Ax$, as input. Assume without loss of generality that the k non-zero are the first k entries, and ordered:

$$\Rightarrow \quad b = \sum_{j=1}^{k} a_j x_j . \qquad x_i \geq x_j \text{ if } i < j .$$

x } k.

In the begining, $r^{(0)} = b$, and $j_0 = \arg\max_j |a_j^T b|$.
For this first stage to succeed, $j_0 \in [1, k]$ - one of the correct atoms -. Thus, we require: $\max_{i \in S} |a_i^T b| > \max_{j \notin S} |a_j^T b|$: necessary and sufficient.

Thus, a sufficient condition is: $|a_1^T b| > \max_{j > k} |a_j^T b|$ ~~...~~.
where X_1 is the largest coeff.

$$\Rightarrow \quad \underbrace{\left| \sum_{t=1}^{k} x_t a_t^T a_1 \right|}_{a} > \max_{j > k} \underbrace{\left| \sum_{t=1}^{k} x_t a_t^T a_j \right|}_{b}$$

We will construct a lower bound for left, upper for right, and enforce it.

a): $\left| \sum_{t=1}^{k} x_t a_t^T a_1 \right| \geq |x_1| - \sum_{t=2}^{k} |x_t a_t^T a_1|$ -reverse tri. inequality-.

$$\geq |x_1| - |x_1| \sum_{t=2}^{k} |a_t^T a_1|$$

$$\geq |x_1| - |x_1| (k-1) \mu(A)$$

$$= |x_1| \left(1 - (k-1) \mu(A)\right).$$

$j > k$:

b): $\left|\sum_{t=1}^{k} x_t a_t^T a_j\right| \le \sum_{t=1}^{k} |x_t||a_t^T a_j| \le |x_1| k \cdot \mu(A)$.

Thus, combining (a), (b); we require that

$$1-(k-1)\mu(A) > k\cdot\mu(A) \Rightarrow k < \frac{1}{2}\left(1+\frac{1}{\mu(A)}\right).$$

This condition guarantees that a correct atom is chosen in the first step.

The residual is updated by subtracting from b (or $r^{(k)}$) a term that is proportional to the chosen atom (or atoms, in general). Thus, r^k is still a linear combination of the same k atoms in b, at most.

Repeating the same steps with $r^{(1)}$ for b guarantees the recovery of the second atom, and so forth. Further, due to orthogonality no atom is chosen twice, and the algorithm terminates after k iterations, retrieving all k non-zeros.

Thus, we have the following result:

For $b = Ax$, A: full rank, if a solution exist: $\|x\|_0 \le \frac{1}{2}\left(1+\frac{1}{\mu}\right)$, OMP finds it in $\|x\|_0$ steps.

Some computational ~~Constraints~~ Considerations: (4)

- The bottleneck is the L.S. But this can be ~~bypass~~ alleviated by using a Choleski decomposition progressively. All in all, OMP is $O(k^3 + kmn)$.
- Futher improvemenst can be obtained if one uses many signals.

→ Demo of Recovery. OMP (and maybe MP). observation: they work way beyond the bound. too pessimistic.

⇒ Better high-probability bounds exist. For example, while this holds when $k \leq \frac{1}{2\mu}$, Schnass (19'!) showed that

$$s \leq \frac{1-\alpha}{\mu^2 \log(m)}$$ suffices with h.p.

What about noisy/real data? Then we would rather want to solve, given $b = Ax^* + v$, $\|v\|_2 \leq \epsilon$, $\|x^*\|_0 \leq k$.

$$(P_0^\epsilon): \min_x \|x\|_0 \text{ s.t. } \|Ax - y\|_2^2 \leq \epsilon^2$$

<u>OMP still works</u>: if $\exists$ solution x: $\|x\|_0 \leq \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right) - \frac{1}{\mu(A)} \cdot \frac{\epsilon}{|x_{min}|}$,

then a) OMP recovers the true support of x^*,

b) $$\|\hat{x}_{OMP} - x^*\|_2^2 \leq \frac{\epsilon^2}{1 - \mu(A)(k-1)}.$$

proof sketch:

$b = \sum_{i=1}^{k} x_i a_i + v$

As before, ~~we~~ we require the sufficient condition

$$|a_1^T b| > \max_{j > k} |a_j^T b| \quad \text{- recall ordering.}$$

$$\left|\sum_{t=1}^{k} x_t a_t^T a_1 + a_1^T v\right| > \max_{j > k} \left|\sum_{t=1}^{k} a_j^T a_t x_t + a_j^T v\right|$$

LHS:
$$\left|\sum_{j=1}^{k} x_t a_t^T a_1 + a_1^T v\right| \geq \left|\sum_{t=1}^{k} x_t a_t^T a_1\right| - |a_1^T v| \qquad (\|a_1\|_2 = 1.)$$

$$\geq |x_1|(1 - (k-1)\mu(A)) - \epsilon$$

RHS:
$$\left|\sum_{t=1}^{k} x_t a_t^T a_j + a_j^T v\right| \geq \left|\sum_{t=1}^{k} x_t a_t^T a_j\right| + |a_j^T v|$$

$$\geq |x_1| \mu(A) \cdot k + \epsilon.$$

$\Rightarrow$ Enforcing.

$$|x_1|(1 - (k-1)\mu(A)) - \epsilon \geq |x_1| \mu(A) k + \epsilon.$$

$$\Rightarrow k \leq \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right) - \frac{1}{\mu(A)} \cdot \frac{\epsilon}{|x_1|}.$$

recall that $|x_1| \geq |x_i|$ $i \in S$, and we have assumed $k < \frac{1}{2}\left(1 + \frac{1}{\mu}\right) - \frac{1}{\mu} \frac{\epsilon}{|x_{min}|}$.
Thus, the above holds for every $i \in S$.
One can then reiterate the argument, except one has to make sure that the sparsity of the representation of the residual and the amount of noise is preserved, until all atoms are recovered.

To see the second claim, note that

$$\hat{x}_{s,omp} = \underset{x}{\operatorname{argmin}} \|A_s x - b\|_2^2 = A_s^+ b = A_s^+(A_s x_s^* + v)$$
$$= x_s^* + A_s^+ v.$$

$$\Rightarrow \|\hat{x}_{omp} - x^*\|_2^2 \leq \|A_s^+\|_2^2 \underbrace{\|v\|_2^2}_{\leq \epsilon^2}$$

$$\leq \frac{1}{\underbrace{\lambda_{min}(A_s^T A_s)}_{(1-\delta_s)}} \leq \frac{1}{1-\mu(A)(k-1)}$$

Gershgorin Disk Theorem.

$A_s = U\Sigma V^T$

$A_s^+ = (V\Sigma U^T U \Sigma V^T)^{-1} V \Sigma U^T$

$\vdots$

$= V\Sigma^{-1} U^T$

$\Rightarrow \|A_s^+\|_2 = \|\Sigma^{-1}\|_2 = \frac{1}{\sigma_{min}(A_s)}$

$\Rightarrow \|A_s^+\|_2^2 = \frac{1}{\lambda_{min}(A_s^T A_s)}$

$\boxed{\text{Denoising}}$

From this we see that if we seek to "denoise" with omp:

$$\|\hat{x} - x\|_2^2 \leq \frac{\epsilon^2}{1-\mu(A)(k-1)} > \epsilon^2 \text{ that we started with!}$$

This is because of the worst-case assumption. In fact, one can obtain great denoising!

Consider the "oracle" estimator, having retrieved the correct support:

$$\hat{x}_{omp} = A_s^+ b = A_s^+(A_s x + v).$$

But now assume $v \sim \mathcal{N}(0, I\sigma^2)$.

$tr(A^{-1}) = \sum_i \lambda_i(A^{-1}) \leq$

$$\Rightarrow \mathbb{E}[\|x - \hat{x}_{omp}\|_2^2] = \mathbb{E}[\|A_s^+ v\|_2^2]$$
$$= \sigma^2 tr((A_s^T A_s)^{-1}) \leq \frac{k \cdot \sigma^2}{1-\delta_s}$$

and likewise $\mathbb{E}[\|x - \hat{x}_{oracle}\|^2] \leq \frac{k\sigma^2}{1-\delta_s}$ which $\ll n\sigma^2$!

When considering Gaussian Noise v, it can be shown that OMP has a nearly-oracle performance, in the sense that achieves an MSE of

$$\|\hat{X}_{OMP} - X_0\|_2^2 \leq 2(1+\alpha) k \sigma^2 \log(m) = C\log(m) \cdot \underbrace{k\sigma^2}_{\text{oracle}}$$

This holds for $v \sim \mathcal{N}(0, \sigma^2 I)$ and if w.h.p.

$$|x_{min}| - (2k-1)\mu |x_{min}| \geq 2\sigma \sqrt{2\underbrace{(1+\alpha)}_{(1+\alpha)} \log m} \quad \text{for some } \alpha > 0,$$

with Probability at least $1 - \dfrac{1}{m^\alpha \sqrt{\pi(1+\alpha)\log m}}$.

[Ben-Haim, Eldar, Elad].

Denoising Demo: ..

IHT: What if signals are very large? Recall that OMP is $O(k^3 + 2kmn)$

Cheaper alternatives are preferred. One of them is IHT, which can be thought of Projected Gradient Descent:

Given $\min_x \|b - Ax\|_2^2$ s.t. $\|x\|_0 \leq k$,

do $$X^{k+1} = H_k(X^k - \eta A^T(A^T x^k - b))$$

$\downarrow$ Projection onto set of k-sparse vector:

$$H_k(a) = \arg\min_x \tfrac{1}{2}\|x - a\|_2^2 \text{ s.t. } \|x\|_0 \leq k.$$

H_k: keeps k-largest entries.

This is cheap: $O(\underset{\text{iterations}}{T}mn + \underbrace{Tn\log(n)}_{\text{sorting}})$.

Theorem (Blumensath & Davies) '09.

If $b = Ax^* + v$, $\|x\|_0 = s$, A has RIP : $\delta_{3s} < 1/\sqrt{32}$, then

$$\|x^k - x^*\|_2 \le 2^{-k}\|x^*\|_2 + 5\|v\|_2.$$

There $\exists$ also accelerated versions of this, AIHT, which perform quite well.

Derivation of IHT: given $\min_x \overbrace{\|b-Ax\|_2^2}^{\ell(x)}$ s.t. $\|x\|_0 \le k$ — hard —

Let's consider a Majorization-Minimization approach:

For $\ell(x)$, hard to minimize; construct $g(x,z) \ge \ell(x)\ \forall z$, and $g(x,x) = \ell(x)$.

Then, we can do

$$x^{k+1} = \arg\min_x g(x, x^k)$$

Consider $g(x,z) = \|b - Ax\|_2^2 - \|Ax - Az\|_2^2 + \|x - z\|_2^2$

$g(x,z) \ge \ell(x)$ if $\|A\|_2 \le 1$, so consider normalizing it.

→ because then $\|A(x-z)\|_2^2 \le \|x - z\|_2^2$

Now, $\min_x g(x,z)$: but $g(x,z) = \|x\|_2^2 - 2x^T(z - A^T(Az-b)) + \underbrace{\|Az\|_2^2 + \|z\|_2^2 + \|b\|_2^2}_{cte}$

So $g(x,z) = \|x - (z - A^T(Az-b))\|_2^2 + cte_2$

So $\min_x g(x,z)$ s.t. $\|x\|_0 \le k$

$\min_x \|x - (z - A^T(Az-b))\|_2^2$ s.t. $\|x\|_0 \le k$ $\Rightarrow$ $x = \arg\min_x(g(x,z)) =$

$x = H_k[z - A^T(Az-b)]$.

So: let $z = x^k$.

$x^{k+1} = H_k[z - A^T(Az-b)]$.