"Just Relax"

We have been studying the (P_0^ϵ), which are still combinatorial.
We have seen that approximating algorithm – greedy ones – often do succeed in finding the sparsest solution.

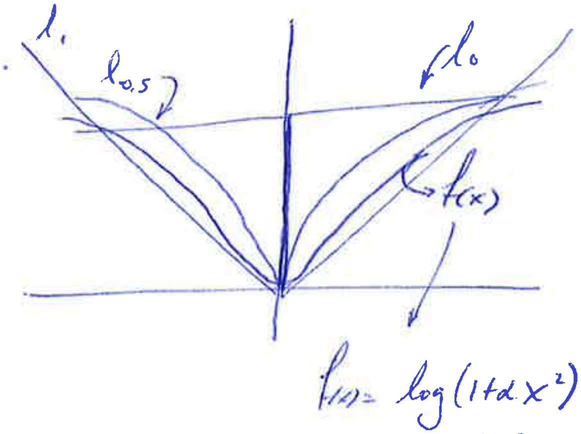
Now we attempt a different strategy: changing the problem objective.

So instead of

(P_0^ϵ): $\min_x \|b - Ax\|_2^2$ s.t. $\|x\|_0 \leq K$

which is equivalent to $\min_x \|b - Ax\|_2^2 + \lambda \|x\|_0$ for some λ,

Consider $\min_x \|b - Ax\|_2^2 + \lambda \|x\|_p^p$, ~~p~~ $0 < p \leq 1$



In fact, any "sensible" options could work.

The l_1 is the "closest" norm to the l_0 problem, and we already saw in the first class that it promotes sparse solutions.

So we'll first study (P_1): $\min_x \|x\|_1$ s.t. $Ax = b$.

How ~~the~~ good is this? How and when can we hope it'll work?

→ We'll follow Tropp's analysis → simpler. (

(Nope, let's do Bickel's).

A: $n \times m$, (full rank).

S: subset of columns $S \subseteq [m]$.

Jump to Bickel's Analysis ↗

Exact Recovery Condition (ERC).

For a given support S, A satisfies the ERC if

$$\max_{i \notin S} \|A_S^{+} a_i\|_1 < 1. \qquad ERC(A,S).$$

Intuitive, this requires ~~every solution~~ the least-squares (ℓ_2) solution to ~~the system~~

$$\min_x \|A_S x - a_i\|_2^2 \quad i \notin S$$

to have ℓ_1 norm less than 1.

Theorem: For x : $S = \text{supp}\{x\}$, $Ax = b$, If ERC is met, BP recovers x.

proof: ~~(We know x is feasible in $Ax=b$.)~~

Consider a candidate solution $y \neq x$, $Ax = Ay$.

and $e = x - y \in N(A)$.

$$Ae = A_S e_S + A_{\bar{S}} e_{\bar{S}} = 0 \quad \Rightarrow \quad e_S = -A_S^{+} A_{\bar{S}} e_{\bar{S}}.$$

$$\Rightarrow \|e_S\|_1 = \|-A_S^{+} A_{\bar{S}} e_{\bar{S}}\|_1 \leq \|A_S^{+} A_{\bar{S}}\|_1 . \|e_{\bar{S}}\|_1$$

The ℓ_1-operator norm $\|B\|_1 = \max_v \frac{\|Bv\|_1}{\|v\|_1}$: maximum ℓ_1 of a column.

Analysis from Bickel in Hastie. (1)

...alysis of Lasso

Observation: The Lasso error satisfies a cone constraint:

$$\|e_{S^c}\|_1 \leq \alpha \|e_S\|_1 , \quad \text{where } e = x_0 - \hat{x},$$

$$\hat{x} = \arg\min \|y - Ax\|_2^2 + \lambda \|x\|_1$$

for a proper value of λ.

Proof: ~~sketch~~ argument.
For the constraint version, i.e: s.t. $\|\hat{x}\|_1 \leq \|x_0\|_1 = R$, since $\hat{x}$ is feasible:

$$e = \hat{x} - x^\circ = x^\circ - \hat{x}$$

$$\|x_0\|_1 = \|x^\circ_S\|_1 \geq \|\hat{x}\|_1 = \|x^\circ + e\|_1 = \|x^\circ_S + e_S\|_1 + \|e_{S^c}\|_1$$

$$\geq \|x^\circ_S\|_1 - \|e_S\|_1 + \|e_{S^c}\|_1$$

$$\Rightarrow \|e_{\bar{S}}\|_1 \leq \|e_S\|_1 \quad (\alpha = 1).$$

Restricted Convexity

Note that for the function $f(x) = \|y - Ax\|_2^2$ to be strongly convex, we need $A^TA \succ 0$. Why strongly convex? Because we want a unique solution.

But since A^TA is rank deficient (n), $A^TA \succeq 0$. i.e: many eigenvalues are zero.

Thus, similarly to what was done for RIP, we can define

$$\min_v \frac{v^T X^T X v}{\|v\|_2^2} \geq \gamma \quad \forall\, v \in C \qquad \|Xv\|_2^2 \geq \gamma \|v\|_2^2$$

In particular, this will be usefull for $C(S, \alpha) = \{ v : \|v_{S^c}\|_1 < \|v_S\|_1 \}$

We are now going to prove a very general result:

Theorem:

Let $y = Ax^\circ + w$, $\|x^\circ\|_0 = k$, and X satisfies the REP with $\gamma > 0$.
Then any $\hat{x} \in \underset{x}{\operatorname{argmin}} \|y - Ax\|_2^2$ s.t. $\|x\|_1 \leq R = \|x^\circ\|_1$ satisfies

$$\boxed{\|\hat{x} - x^*\|_2 \leq \frac{4}{\gamma}\sqrt{k}\,\|A^T w\|_\infty.}$$

- Note that these results are deterministic.
- Different probabilistic results can be obtained imposing distributions over A and/or w.

proof:
Recall that x° is feasible, and $\hat{x}$ is optimal. Thus,

$$\|y - A\hat{x}\|_2^2 \leq \|y - Ax^\circ\|_2^2 = \|w\|_2^2$$

$\langle f, g\rangle = |\sum f_i g_i| \leq \sum_i |f_i g_i|$

$\|fg\|_1 \leq \|f\|_p \|g\|_q$ for $\frac{1}{p} + \frac{1}{q} = 1$

Holder's Ineq.

$$\|\underbrace{Ax^\circ - A\hat{x}}_{Ae} + w\|_2^2 = \|Ae\|_2^2 + \|w\|_2^2 - 2w^T A e \leq \|w\|_2^2.$$

$$\Rightarrow \frac{\|Ae\|_2^2}{2} \leq |w^T A e| \leq \|A^T w\|_\infty \|e\|_1$$

But $\|e\|_1 = \|e_S\|_1 + \|e_{\bar{S}}\|_1 \leq 2\|e_S\|_1 \leq 2\sqrt{k}\,\|e\|_2.$

On the other hand, we know that $e \in \mathcal{C}(S, \alpha)$

$$\Rightarrow \|Ae\|_2^2 \geq \|e\|_2^2\,\gamma$$

$$\Rightarrow \|e\|_2 = \|\hat{x} - x^\circ\|_2 \leq \frac{4}{\gamma}\sqrt{k}\,\|A^T w\|_\infty.$$

//

e relevant cases:

a) if $w = 0$, then $\hat{x} = x^\circ$: exact recovery.

if $\mathbb{E}[b_i]=0$

$Var(\sum_i b_i) = \sum_{i=1}^{n} Var(b_i) + \underbrace{\sum_{i \neq j} Cov(b_i, b_j)}_{0} = \sum E(b_i^2)$

c) Say $\|a_i\|_2 = 1$, and $w \sim \mathcal{N}(0, \sigma^2)$.

Then $a_i^T w \sim \mathcal{N}(0, \sigma^2)$, and so we can use the Gaussian tail bound:

$$P[|a_i^T w| \geq t] \leq 2e^{-t^2/2\sigma^2} \rightarrow \text{(Chernoff bound). for Gaussian distribution}$$

Now, since

$$Pr[\|A^T w\|_\infty \geq t] = Pr[\max_i |a_i^T w| \geq t] \leq 2e^{-t^2/2\sigma^2 + \log(m)}$$

$$= Pr[\bigcup_i \{|a_i^T w| \geq t\}] \leq \sum_{i=1}^{p} Pr[|a_i^T w| \geq t]$$

union bound

$$= 2p e^{-t^2/2\sigma^2}$$

$$= 2e^{-t^2/2\sigma^2 + \log(p)}.$$

Now, letting $t = \sigma\sqrt{\tau \log m}$

$\Rightarrow$ $Pr[\|A^T w\|_\infty \geq t] \leq 2e^{-\frac{1}{2}(\tau - 2)\log(m)}$. with $(\tau > 2)$

Thus, $\boxed{\|\hat{x} - x^\circ\|_2 \leq \frac{4\sigma}{\gamma}\sqrt{k\tau \log m}}$ with $Pr > 1 - 2e^{-\frac{1}{2}(\tau - 2)\log m}$.

b) Say, $\|a_i\|_2 = 1$, $\|w\|_2 \leq \epsilon$. - adversarial.

$$\Rightarrow \|\hat{x} - x^\circ\|_2 \leq \frac{4}{\gamma}\sqrt{k} \cdot \|A^T w\|_\infty \leq \frac{4\sqrt{k}}{\gamma}\|w\|_2 = \frac{4\sqrt{k}}{\gamma} \cdot \epsilon.$$

What we are now missing is how to know/bound γ, or how to know if $\gamma > 0$ – which is what really matters.

Recall that REP(s,α)

$$\gamma = \min_{e \neq 0} \frac{\|Ae\|_2^2}{\|e\|_2^2} > 0, \qquad \forall e \in \mathcal{C} = \{e : \|e_{\bar{S}}\|_1 \leq \alpha \|e_S\|_1\}$$

Note that, considering a support S, and Lasso solution $\hat{x}$,

$$\|Ae\|_2^2 = \|A(e_S + e_{\bar{S}})\|_2^2 = \|Ae_S\|_2^2 + \|Ae_{\bar{S}}\|_2^2 + 2\, e_{\bar{S}}^T \underbrace{G}_{A^TA} e_S$$

$$\geq \|Ae_S\|_2^2 - 2\,|e_{\bar{S}}^T G e_S|.$$

Now,

$$|e_{\bar{S}}^T G e_S| = |\tilde{e}_{\bar{S}}^T G_{\bar{S},S} \tilde{e}_S| \leq \|\tilde{e}_{\bar{S}}\|_1 \|G_{\bar{S},S} \tilde{e}_S\|_\infty$$

(maybe show?)

$$\leq \|\tilde{e}_{\bar{S}}\|_1 \,\mu(A) \cdot \|\tilde{e}_S\|_1.$$

$$\leq \|e_S\|_1^2 \mu(A)\, \alpha.$$

$$\leq k \cdot \|e_S\|_2^2 \,\mu(A) \cdot \alpha.$$

On the other hand, $\|e_S\|_0 = k$, so

$$\|Ae_S\|_2^2 \geq (1-\delta_k)\|e_S\|_2^2 \underset{\delta_k \leq (k-1)\mu(A).}{\geq} (1-(k-1)\mu(A))\,\|e_S\|_2^2.$$

Putting everything together:

if $e \in \mathcal{C}(s,\alpha)$,
$$\frac{\|Ae\|_2^2}{\|e\|_2^2} \geq 1-(k-1)\mu(A) - 2k\mu(A)\alpha.$$

$$> 1 - k\mu(A) - 2k\mu(A)\alpha = 1 - k\mu(A)(1+2\alpha).$$

Thus, $\gamma > 1 - k\cdot\mu(A)(1+2\alpha)$

and while $k < \dfrac{1}{\mu(A)(1+2\alpha)} \rightarrow \gamma > 0$ ✓

//

ompressed Sensing

Motivation: if signals / images are "compressible" - why can't we just acquire them in this "compressed" representation?

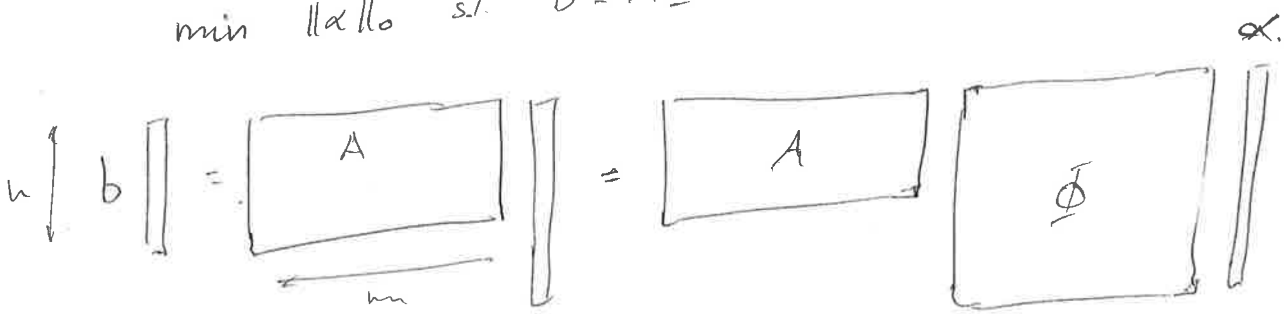
Say now that signal $x \in \mathbb{R}^m$. We want to take measurements $b_i = \langle a_i, x \rangle$, $i = 1 \ldots n$.

How small can 'n' be while still be able to recover x?

Note that if $n < m$ → undetermined. But if x is a "natural" signal, then is it "sparsifiable", say, under a unitary transformation Φ: $x = \Phi \alpha$, $\|\alpha\|_0 \ll m$.

Thus, we are trying to solve

$$\min \|\alpha\|_0 \quad \text{s.t.} \quad b = A \Phi \alpha.$$



How small can we make n?

Before we saw that BP recovers α if

$$\|\alpha\|_0 < \frac{1}{3\mu(A)} \qquad \left(\text{or } \|\alpha\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right)\right).$$

How small can $\mu(A)$ be?

From the Welch Bound, $\mu(A) \geq \sqrt{\frac{m-n}{n(m-1)}} = \Omega(1/\sqrt{n})$.

$\Rightarrow \quad K \lesssim \frac{1}{\mu(A)} \quad \Rightarrow \quad K \lesssim \sqrt{n} \quad \Rightarrow \quad n \gtrsim k^2$

this is not really good: Say $m = 100^2$, and say $K = 0.1m$

then $n \gtrsim 1000$

too many samples!

This is the "squared bottleneck", resulting from the pessimistic analysis relying on $\mu(A)$.

The alternative is to go back to other matrix ~~quanta~~ characterizations (e.g. RIP), even if they are not computable.

In fact, we've seen before that BP recovers x (or α) is A satisfies the restricted eigenvalue property. So which A satisfy this, and for what n?

Turns out (Raskutti, Wainwright, Yu) that for Gaussian matrices $n \times m$ with iid rows sampled from $\mathcal{N}(0, \Sigma_i)$,

and Σ_i satisfies the $REP(\gamma)$ $\left(\|\Sigma_i^{1/2} e\|_2^2 \geq \gamma \|e\|_2^2 \right)$ $e \in \mathcal{C}$.

if $\quad n > c \, \frac{\sigma_{max}^2 (1+\alpha)^2}{\gamma^2} \cdot k \cdot \log(m)$

then A satisfies the R.E. property with parameters $(\alpha, \gamma/8)$

with Probability $> 1 - c' e^{-c'' n}$.

c', c'', c''' universal positive constants.

historical reasons (and elegance), let's show a result

based on the Nullspace Property:

<u>NSP:</u> A matrix A satisfies the NSP for support S if

$$\|V_S\|_1 < \|V_{\bar{S}}\|_1 \quad \forall \quad v \in \mathcal{N}(A) \setminus \{0\}.$$

<u>Necessary and sufficient condition:</u>

Given $A: n \times m$, every vector $x \in \mathbb{R}^m$ supported on set S is the unique solution to $\min_z \|z\|_1$ s.t. $Az = Ax$ if and only if A satisfies the NSP relative to S.

<u>proof:</u> first implication:

Assume every vector x is the unique minim. of $\|z\|_1$ s.t. $Az = Ax$.

In particular, consider $v \in \mathcal{N}(A) \setminus \{0\}$, and $V_S = \arg\min_z \|z\|$ s.t. $Az = AV_S$.

Now, since $Av = 0 = AV_S + AV_{\bar{S}} \Rightarrow (-V_{\bar{S}})$ is also feasible since

$AV_S = A(-V_{\bar{S}})$. But V_S is the unique minimizer $\Rightarrow \|V_{\bar{S}}\|_1 > \|V_S\|_1$.

Second implic: assume NS holds for S.

Consider x supported on S, and a different vector $z \neq x$. ~~but~~ ~~feasible~~,

such that $Ax = Az$. Further, $V = x - z \in \mathcal{N}(A) \setminus \{0\}$.

Then:

$$\|x\|_1 = \|x - z_S + z_S\|_1 \leq \|x - z_S\|_1 + \|z_S\|_1 = \|V_S\|_1 + \|z_S\|_1$$

$$\xrightarrow{NSP} < \|V_{\bar{S}}\|_1 + \|z_S\|_1$$

$$= \|z_{\bar{S}}\|_1 + \|z_S\|_1 = \|z\|_1$$

$\Rightarrow \|x\|_1$ is minimal, and any other feasible solution has $\geq \ell_1$. //

Thus, we have the uniform recovery guarantee:

Given A, every k-sparse $x \in \mathbb{R}^m$ is the unique solution to BP iff A satisfies NUS of order k (for all S).

A classic Result is that RIP is sufficient for recovery, because it implies the NSP:

If A has the RIP with $\delta_{2k} < \frac{1}{3}$, then A has the NSP of order k. $\Rightarrow$ it recovers all k-sparse vectors.

What matrices satisfy the RIP? many random ones do:

Let $A: n \times m$ be a subGaussian matrix: all entries are ~~iid~~ independent zero-mean subgaussian random variables with variance 1 and subgaussian parameters β, k:

$$\mathbb{P}(|A_{ij}| \geq t) \leq \beta e^{-kt^2} \quad \forall\, t > 0.$$

Then, $\exists\, c > 0$ such that $\left(\frac{1}{\sqrt{n}} A\right)$ has RIP with $\delta_k < \delta$ if

$$n \geq \frac{2c}{\delta^2} \cdot k \log\left(\frac{em}{k}\right)$$

with $\mathbb{P} \geq 1 - 2e^{-\frac{\delta^2 n}{2c}}$.

is results then guarantees recovery of all k-sparse vectors w.h.p. by BP, for such random matrices.

This almost answers all questions for compressed sensing. Recall however that x is typically not sparse, but rather $x = \Phi\alpha$, α is.

This is ok, because if A is RIP, Φ unitary, $(A\Phi)$ also is RIP with different constants.

Note that Φ is irrelevant for sensing: $b = Ax$.

It should only be used for recovery/decoding:

$$\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad b = A\Phi\alpha.$$

Optimization for Lasso

Remember Gradient Descent:

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{for } f: \text{convex and differentiable.}$$

G.D: $x^{k+1} = x^k - t_k \nabla f(x^k)$.

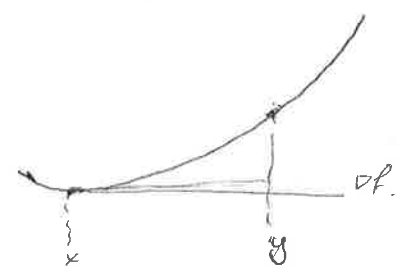
If ∇f is Lipschitz (i.e; f is smooth), then G.D converges with rate of $O(1/k)$.

But what if f is non-smooth?

Subgradients

for a convex $f: \mathbb{R}^n \to \mathbb{R}$.

$$f(y) \geq f(x) + \underbrace{\nabla f(x)}_{\text{gradient}}{}^T (y-x), \ \forall x, y.$$

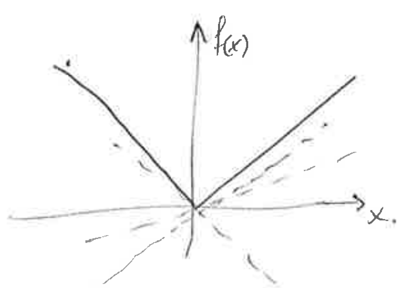


i.e: a linear approximation underestimates $f(x)$.

A subgradient of a convex f at x is any $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T (y-x). \ \forall y.$$

- always exists. (unlike the gradient).
- If f is differentiable at x, then $g = \nabla f(x)$. uniquely.



for $x \neq 0$, $g = \nabla f(x) = \text{sign}(x)$.

for $x = 0$, $g \in [-1, 1]$.

The set of all subgradients of a convex f is called the subdifferential $\partial f(x) = \{ g \in \mathbb{R}^n : g$ is a subgradient of f at $x \}$.

Note: - always non empty (for convex f).

- If f differentiable at x, $\partial f = \{ \nabla f(x) \}$.

Thus, they give optimality conditions: for convex f:

$$\boxed{f(x^*) = \min_x f(x) \iff 0 \in \partial f(x^*)}$$

proof: $f(y) \geq f(x^*) + g^T(y - x^*)$

if $g = 0 \Rightarrow f(y) \geq f(x^*) \;\; \forall y$.

Subgradient Method:

$$x^{k+1} = x^k - t_k g^k, \quad \text{where } g^k \in \partial f(x^k)$$

To get convergence, we need the step size to go to zero (but not too fast!)

Subgradient descent has ~~an~~ ~~of~~ a convergence rate of $O(1/\sqrt{k})$.

$\Rightarrow$ slow.

Algorithms for Lasso/BP

While any convex optimization algorithm would suffice, we'll pay close attention to first order methods.

Getting some intuition: Unitary Case.

Assume A: unitary/orthonormal.

then $\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda \|x\|_1$

$$\underbrace{\min_x \frac{1}{2}\|b - x\|_2^2 + \lambda \|x\|_1}, \quad \text{where } b = A^T y.$$

$$h(x) = \sum_{i=1}^{m} \frac{1}{2}(b_i - x_i) + \lambda |x_i|. \quad : \text{separable!}$$

$$h(x_i) = \begin{cases} \frac{1}{2}(b_i - x_i)^2 + \lambda & \text{if } x_i > 0. \\ \frac{1}{2}(b_i - x_i)^2 - \lambda & \text{if } x_i < 0 \end{cases}$$

Note that $h(x_i)$ is differentiable if $x_i > 0$ or $x_i < 0$, but not at $x_i = 0$.

recall that $h'(x_i) = \lim_{\epsilon \to 0} \frac{h(x_i+\epsilon) - h(x_i)}{\epsilon}$

must be unique!

thus, if $x_i > 0$:

$$x_i - b + \lambda = 0 \Rightarrow x_i = \max(b - \lambda, 0) \quad \text{if } b > \lambda$$

if $x_i < 0$: $x_i = \min(b + \lambda, 0)$ if $b < -\lambda$

if $|b| < \lambda$, then the min $h(x_i)$ must be attained at the only point it's not differentiable: $x = 0$.

Thus, for A: orthonorm, $x = \arg\min_x h(x) = S_\lambda(A^T y)$

$\hookrightarrow S_\lambda(b) = \text{sign}(b)(|b| - \lambda)_+$

this is $\text{prox}_g(b)$ / show.

What if $A : n \times m$?

$$F(x) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1$$

Following a Majorization - Minimization approach:

Consider $g(x,z) = F(x) + d(x,z) \geq F(x)$. and $F(x) = g(x,x)$.

$$d(x,z) = \frac{c}{2}\|x - z\|_2^2 - \frac{1}{2}\|Ax - Az\|_2^2 \geq 0.$$

we want $d(x,z)$ to be strongly convex $\Rightarrow$ $H(x) = c \cdot I - A^TA > 0$.

$\Rightarrow$ $c > \|A\|_2 = \lambda_{max}(A^TA)$.

Then: $g(x,z) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1 + \frac{c}{2}\|x - z\|_2^2 - \frac{1}{2}\|Ax - Az\|_2^2$.

$$= \frac{1}{2}\|y\|_2^2 + \cancel{\frac{1}{2}\|Ax\|_2^2} - x^TA^Ty + \frac{c}{2}\|x\|_2^2 + \frac{c}{2}\|z\|_2^2 - cx^Tz - \cancel{\frac{1}{2}\|Ax\|_2^2} - \frac{1}{2}\|Az\|_2^2 + x^TA^TAz.$$

$$= \frac{c}{2}\|x\|_2^2 - x^T\underbrace{(A^Ty + cz - A^TAz)}_{cz - A^T(Az - y) \,\triangleq\, v} + \text{cte.} + f(z,y). + \|x\|_1\lambda$$

$$g(x,z) = \frac{1}{2}\|x\|_2^2 - \frac{1}{c}x^T \cdot v + \frac{\lambda}{c}\|x\|_1 + \tilde{f}(z,y).$$

$$g(x,z) = \frac{1}{2}\|x - \frac{1}{c} \cdot v\|_2^2 + \frac{\lambda}{c}\|x\|_1 + f'(z,y).$$

$$\Rightarrow \arg\min_x g(x,z) = S_{\lambda/c}(v) = \underline{S_{\lambda/c}\left(z - \frac{1}{c}A^T(Az - y)\right)}.$$

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

$$\Rightarrow \quad x^{k+1} = S_{\lambda/c}\left(x^k - \frac{1}{c}A^T(Ax^k - y)\right). \quad \text{ISTA}.$$

... is is a particular case of Proximal-Gradient Methods, (5).

which solve $\min_x f(x) + g(x)$, : convex.

where f: convex & smooth, g: convex but non-smooth (non-diff.).

$$x^{k+1} = \text{prox}_g\left(x^k - \tfrac{1}{c}\nabla f(x^k)\right) = T_\lambda^{f,g}(x^k)$$

Recall that subgradient descent achieves convergence of $O(1/\sqrt{k})$.
We'll see (in part) that ISTA (prox.grad) achieves $O(1/k)$
basically at same cost (comput.), and it can be further accelerated to $O(1/k^2)$.

$$\begin{cases} x^{k+1} = S_\lambda\left(y^k - \tfrac{1}{c}A^T(Ay^k - b)\right) \\ y^{k+1} = x^{k+1} + t^k(x^{k+1} - x^k) \end{cases}$$

To make analysis shorter, we'll need a couple of Lemmas:

L1: Monotonic decrease:

Let $F(x) = f(x) + g(x)$ as above, and let $x^{k+1} = \text{ISTA}(x^k)$. with stepsize $c = L = \lambda_{max}(A^TA)$. Then

$$F(x^k) - F(x^{k+1}) \geq \frac{L}{2}\|x^k - x^{k+1}\|.$$

Note that this implies that $F(x^{k+1}) \leq F(x^k)$.

Observe that $f(x)$ is smooth: $\|\nabla f(x) - \nabla f(y)\|_2^2 \leq L\|x-y\|_2^2$

for $L = \lambda_{max}(A^TA)$

$\underline{L2}$:

$\forall \quad \frac{2}{L}\left[F(x) - F(T_\lambda^{f,g}(y))\right] \geq \|x - T_A^{f,g}(y)\|_2^2 - \|x - y\|_2^2 \quad \forall\, x, y.$

Theorem: For $F(x) = f(x) + g(x)$ as before, let $\{x^k\}$ be the sequence generated by ISTA with $L \geq \lambda_{max}(A^TA)$. Then, for any $x^* \in \mathcal{X}_{opt}$ and $k \geq 0$,

$$F(x^k) - F_{opt} \leq \frac{L}{2k}\|x^{(0)} - x^*\|^2$$

proof: From Lemma 2, set $x = x^*$, and $y = x^k$ to obtain:

$$\frac{2}{L}\left(F(x^*) - F(x^{k+1})\right) \geq \|x^* - x^{k+1}\|_2^2 - \|x^* - x^k\|_2^2.$$

Summing for all $k \geq 0$:

$$\frac{2}{L}\sum_{n=0}^{k-1}\left(F(x^*) - F(x^{n+1})\right) \geq \|x^* - x^k\|_2^2 - \|x^* - x^0\|_2^2$$

$$\Rightarrow \quad \frac{2}{L}\sum_{n=0}^{k-1}\left(F(x^{n+1}) - F(x^*)\right) \leq \|x^* - x^0\|_2^2$$

$F(x^*) = F_{opt}$: by def. Recall that $F(x^{n+1}) \leq F(x^n)$.

Thus

$$k\left(F(x^k) - F(x^*)\right) \leq \sum_{n=0}^{k-1} F(x^{n+1}) - F_{opt} \leq \|x^* - x^0\|_2^2 \cdot \frac{L}{2}.$$

$$\Rightarrow \quad F(x^k) - F(x^*) \leq \frac{L}{2k}\cdot\|x^0 - x^*\|_2^2.$$

//