An example, with a variety of algorithms

$$\min_x \|y - Ax\| + \lambda |x|_1$$

$$\sum_i |x_i|$$

This problem turns up <u>often</u>

- <u>Sparse reconstruction</u>
    - $y$ is a <u>signal</u>
    - $A$ is a <u>dictionary</u>
    - $x$ is a <u>sparse encoding</u>

## L1 reg linear regression

model $y_i$ by $x_i^T x + x_b$

the bias

Solve

$$\sum_i \left(y_i - (u_i^T x + x_b)\right)^2 + \lambda |x|_1$$

(it is not usual to regularize the bias.
- you could estimate it separately)

$$\|y - (Ax + x_b)\|^2 + \lambda |x|_1$$

Now estimate $x_b$ separately, to get $\tilde{y}$

$$\|\tilde{y} - Ax\|^2 + \lambda |x|_1$$

In both cases, the attraction is that the $L_1$ norm encourages sparsity (many zeros in $x$). One way to see this is notice that the penalty for small $x_i$ is large compared to $L_2$ — its worth making small values zero.

## Compressed Sensing:

· assume we have $m$ linear measurements of an unknown signal $z$

$$y_i = \phi_i \cdot z + \text{noise}$$

Suppose we know that $z$ is compressible, or has a sparse repn in a transform domain ~~(with a dictionary)~~ $W$ then we can recover $z$ by solving

$$\| \underbrace{\Phi}_{\substack{\text{measurement} \\ \text{vectors, forming the} \\ \text{compressed sensing} \\ \text{matrix}}} z - \overbrace{y}^{\substack{\text{measurement} \\ \text{values}}} \|^2 + \lambda \| \underbrace{W}_{\substack{\text{sparsifying} \\ \text{transform.}}} z \|_1$$

Now assume that $W$ is invertible

we can solve

$$\min \|Ax - y\|^2 + \lambda \|x\|_1$$

where $Wx = z$

$AW = \Phi$

<u>$A$ is the dictionary</u>

This problem is VERY DIFFERENT from

$$\|y - Ax\|^2 + \lambda \|x\|^2$$

2 norm.

2-norm problem:

$$\underline{(A^TA + \lambda I)\,x = A^Ty}$$

linear!

1 norm:

not linear

2 norm:

$$\lim_{\lambda \to 0} x = \cancel{(A^TA)^{-1}A^Ty}$$

$$= \underbrace{(A^TA)^+}_{} A^Ty$$

$+ \rightarrow$ Moore-Penrose pseudo inverse

2-norm: $\lambda \to \infty$ gives $x \to 0$

1norm: ~~$\lambda \to \infty$ gives~~

$x = 0$ for values of $\lambda \geq \lambda_{max}$

where $\lambda_{max} = \| 2A^T y \|_\infty$

$\left( \|u\|_\infty = \max_i |u_i| \right.$, inf norm).

2-norm

$x(\lambda)$ is a curve

(rational, algebraic, some mildly interesting geometry)

1-norm

$x(\lambda)$ is piecewise linear!

Some transformations of this problem

$$\min \|Ax-y\|^2 + \lambda |x|_1$$

can be turned into a quadratic program.

$$\min \|Ax-y\|^2 + \lambda \sum_i t_i$$

s.t. $\quad -t_i \leqslant x_i \leqslant t_i$

$$\begin{gathered} ||| \\ -x_i - t_i \leqslant 0 \\ x_i - t_i \leqslant 0 \end{gathered}$$

Notice this might be worrying — the objective is $\quad x^T A^T A x - 2y^T A x + y^T y$

This could be dense

Consider

$$\min \left[ f(x) + \lambda g(x) \right] \qquad (1)$$

and

$$\min \; f(x) \qquad (2)$$

$$\text{st} \quad g(x) \leq \mu .$$

Lagrangian for (2):

$$f(x) + \lambda \left( g(x) - \mu \right)$$

KKT

$$\nabla f + \lambda \nabla g$$

$$\lambda \geq 0$$

$$g \lambda = 0$$

$$g(x) - \mu = 0$$

~~exclude the~~

different values of $\mu$ → values of $\lambda \geq 0$

So for any $\lambda \geq 0$ in ①

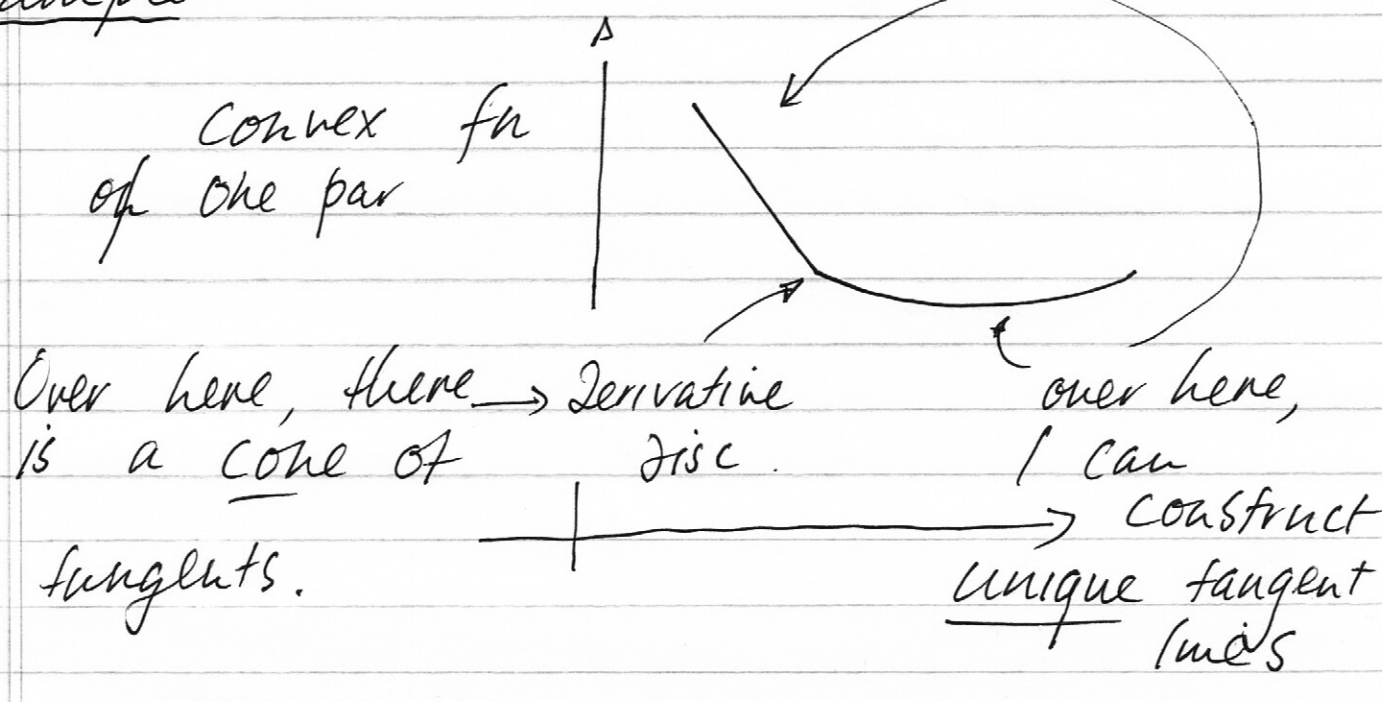
I can choose a $\mu \geq 0$ in ②

so that I get the <u>same soln</u>.

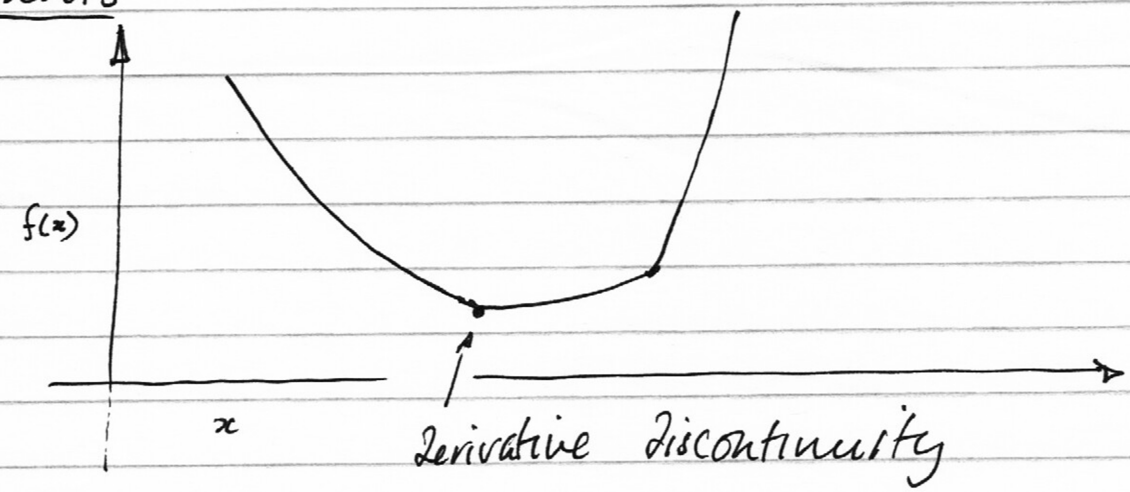
To proceed, we need a richer notion of gradient, to deal w/ the absolute value.

– the <u>subgradient</u>

<u>Example</u>



## Subgradients




consider the graph of a convex fn

$(\underline{x}, f(x))$.

- at a *differentiable* point, there is a *tangent plane*

*Equation* is easy.

*Surface* $g(x) = x_n - f(x_{1:n}) = 0$

*Normal* $= \nabla g = (-\nabla f, 1)$

(this *isn't* unit - doesn't matter

so the tangent plane at

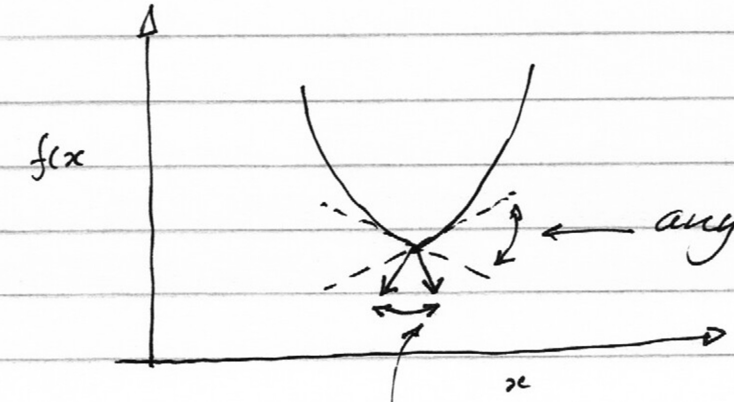
$(\underline{u}, f(\underline{u}))$

is $\quad -\nabla f\big|_{u} \cdot \underline{x}_{1:n} + x_n - \left(-\nabla f\big|_{u} \cdot \underline{u} + f(\underline{u})\right) = 0$

Now think about a non-differentiable point.

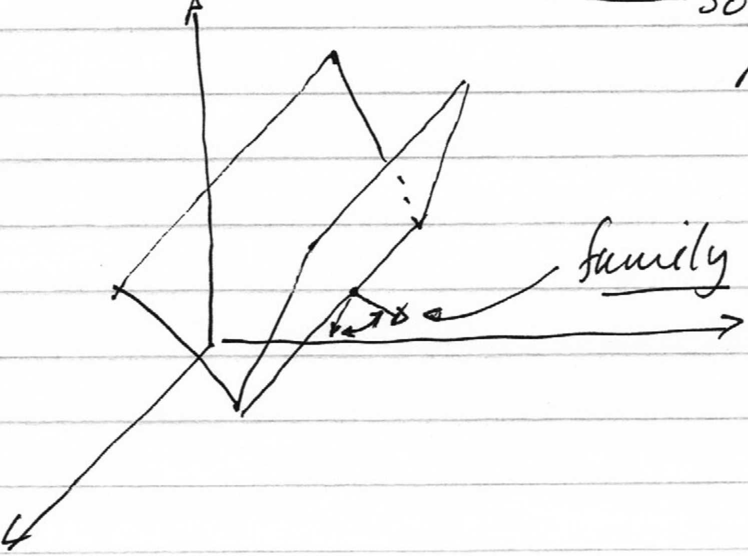
- there is a <u>cone</u> of normals

e.g.

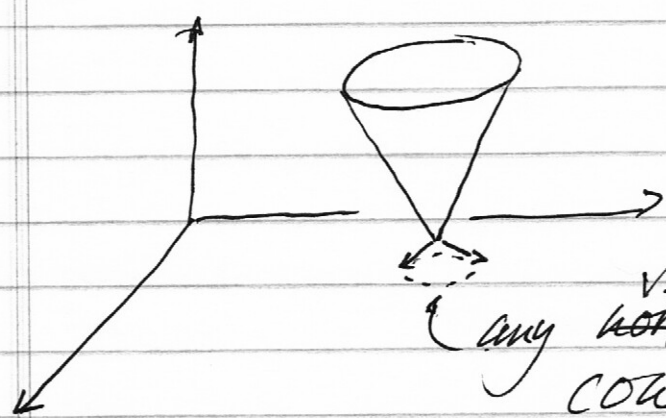



any <u>line</u> in this see-saw is tangent

so any vector in this range is tangent



<u>family</u> of normals
- 2D cone



any ~~normal~~ vector in this filled cone is normal.

All this means that at such points, our function has a family of derivatives — known as subgradients

Because a tangent plane (= normal) yields a derivative

| TP | Normal | SubGrad |
|---|---|---|
| $-p \cdot x_{1:n} + x_n - \alpha = 0$ | $(-p, 1)$ | $p.$ |

This yields an easy construction for subgradient of abs

$$\partial |x| = \begin{cases} -1 & x < 0 \\ 1 & x > 0 \\ [-1, 1] & x = 0 \end{cases}$$

subgradient

closed interval

Because my derivation of ~~Lagrangian~~ derivative condition was geometric, it works for subgradients

So for $f(x)$ to be a min, must have

$$0 \in \partial f$$

because subgradients produce intervals, or worse

So, for our problem

$$0 \in \partial\left[\|Ax-y\|^2 + \lambda |x|_1\right]$$

$$= 2A^T(Ax-y) + \lambda \partial |x|_1$$

$$= 2A^T(Ax-y) + \lambda s$$

where $s_i = \begin{cases} 1 & x_i > 0 \\ -1 & x_i < 0 \\ [-1,1] & x_i = 0 \end{cases}$

$s$ is the <u>sign vector</u>.

<u>Notice</u> if you <u>know</u> the sign vector for a soln, then the soln is easy to get.

write $\quad J(s) = \{ i \mid s_i \neq [-1;1] \}$

$$J_c(s) = \{\text{All indices}\} - J(s)$$

assume we <u>know</u> $s$;

then $\quad x_{J_c(s)} = 0$

wild index notation; sorry!

so we have

$$A_J^T (A_J x_J - y_J) + \lambda s_J = 0$$

which is a straightforward linear system (we assume that $A_J^T A_J$ has full rank for any $J$ we deal with — fairly reasonable)

Now, assume

$$\lambda_1 < \lambda_2 \quad , \quad S(\lambda_1) = S(\lambda_2)$$
$$= \sigma$$

1) $x_{J_c(\sigma)}(\lambda_1) = x_{J_c(\sigma)}(\lambda_2) = 0$ (obvious)

2) $S(t\lambda_1 + (1-t)\lambda_2) = \sigma$ (obvious)

3) $A^T_{J(\sigma)}\left(A_{J(\sigma)}\left[t\, x_{J(\sigma)}(\lambda_1) + (1-t)\, x_{J(\sigma)}(\lambda_2)\right] - y_{J(\sigma)}\right]$

$$+ \left(t\lambda_1 + (1-t)\lambda_2\right)\sigma = 0$$

(easy)

<u>But</u> this means

$x(\lambda)$ is <u>Piecewise linear</u>

So it is natural to try and <u>construct</u> this path. — but there is some bad news.

write $d$ for dimension of $x$.

then, clearly, # of verts in path is $\leq 3^d$

actually, the upper bound is $\frac{(3^d+1)}{2}$, and it can be <u>attained</u>

(see papers).

## Algorithms

### 1) Mathing pursuit

$r_0 = y$ ; ~~cho~~ $x_0 = 0$

- choose col j of A that has largest value of $r_i^T A_j$
- $x_{i+1} = x_i + (r_i^T A_j) \cdot \frac{1}{(A_j^T A_j)} \cdot \underline{e}_j$
- $r_{i+1} = r_i - \frac{(r_i^T A_j)}{(A_j^T A_j)} \cdot A_j$

Notice

- residual always gets smaller
- $|x|_1$ gets bigger

## Orthogonal matching pursuit

- like matching pursuit, BUT
  - readjust all ~~coeff~~ non-zero coeffs each time you insert a column to get best fit in that space.
  - Better estimate in $k$ steps, but each step takes more work.

## Homotopy algorithms

- a whole class of algorithm, constructing approx or exact

$x(\lambda)$.

<u>Notice</u>

1) we know $x(0)$

(because $A^T(Ax-y) = 0$, linear alg)

2) for sufficiently <u>large</u> $\lambda \geq \lambda_{max}$

$$x(\lambda_{max}) = 0$$

- and we can compute $\lambda_{max}$

$$0 \in A^T(Ax-y) + \lambda_{max}\begin{bmatrix} -1;1 \\ \vdots \\ -1;1 \end{bmatrix}$$

and $x = 0$

So we must have

$$0 \in -A^Ty + \lambda_{max}\begin{bmatrix} -1 & 1 \\ & \\ -1 & 1 \end{bmatrix}$$

$$\text{so } \lambda_{max} = \|A^Ty\|_\infty$$

largest abs. value.

## So we can do

$$\lambda_0 = 0 \;; \quad A^T A x_0 = A^T y$$

- Tangent to path:

$$(A^T A)_J \frac{dx_J}{d\lambda} + \sigma_J(\lambda) = 0$$

- search along tangent for $i$ that gives first sign change
- update $x$
- stop when $\lambda = \lambda_{max}$

## Advantage

- complete path
- Disadvantages:
  1) what if 2 knots coincide (rare)
  2) too many knots (approximations are available)

Another view of the PL: path.

- Problem

$$\min. \ \|Ax - y\|^2 + \lambda \sum_i t_i$$

$$\text{st} \quad -t_i \leq x_i \leq t_i$$

- Notice this is a *family* of q.p.'s
  - linear constraints, but the polytope *Doesn't* change when $\lambda$ changes.
- By inspection, at soln $x_i = \begin{cases} t_i \\ -t_i \end{cases}$
- *By inspection*, these are 1-faces or 0-faces

$\Rightarrow$ Soln is always on 1-face or 0-face

But look at SVM (linear)

$$\lambda \frac{w^T w}{2} + \frac{1}{N} \sum_i \xi_i$$

st

$$y_i (w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

1) as $\lambda$ changes, polytope does <u>not</u> change:

2) at soln, we always have

$$\xi_i = \max(0, 1 - y_i (w^T x_i + b))$$

so a soln is always on a

0-face or 1-face

3) $\Rightarrow$ homotopy path is P.L. !

(and fairly straightforward to construct)