This means we could think about

$$q(\lambda) = \inf_x \mathcal{L}(x, \lambda)$$

↑ recall: Lagrangian

dual

write x^c for a global min

$\operatorname{argmin}_x f(x)$ st $g(x) = 0$

Then we have

$$q(\lambda) \leq f(x^c)$$

(This is fairly easy. Think about a single constraint; $\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$

Now: either $q(\lambda) = -\infty$ or its finite;

if it's finite, then $\lambda g(x)$ must be bounded below; if the bound is 0, then $g(x) = 0$ or $\lambda = 0$

if $g(x) = 0$, then q can't be bigger than $f(x^c)$

if $\lambda = 0$, same

now if bound is bigger than zero,
we can't have a soln.
So $\lambda q(x) \neq 0$, so . . .

Example Duals:

$$\min_x \quad \frac{x^T x}{2} \qquad \text{s.t.} \quad Ax + b = 0$$

$$\mathcal{L}: \quad \frac{x^T x}{2} - \lambda^T (Ax + b).$$

now: $\inf_x$ occurs when

$$x - A^T\lambda = 0 \quad (\nabla_x \mathcal{L} = 0)$$

so:

$$q(\lambda) = -\frac{\lambda^T A A^T \lambda}{2} - b^T\lambda$$

Notice: $\max_\lambda \; q(\lambda)$ occurs when

$$AA^T\lambda - b = 0$$

(which we've seen before).

Ideas:

- form dual and solve

$$\max_{\lambda} \; q(\lambda)$$

- This isn't usually a good idea, cause we can't get the dual.

Idea:

- go down in x, then up in λ
- we've done something like that already! (ALM ;)
- Sometimes called "Dual Ascent"

recall:

start: x^0, $\lambda^0 = 0$, $c^0 > 0$

$$x^{k+1} = \arg\min_x A(x, \lambda^k, c^k) = f(x) - \lambda^{kT} g(x) + \frac{c^k}{2} g(x)^T g(x)$$

$$\lambda^{k+1} = \lambda^k - \frac{c^k}{2} g(x^{k+1})$$

$$c^{k+1} = r\, c^k .$$

Notice

$$A(x^{k+1}, \lambda^{k+1}, c^{k+1})$$

$$= A(x^{k+1}, \lambda^k, c^k) + r \frac{c^k}{2} g(x^k)^T g(x^k).$$

and this is +ve.

Another new idea: Splitting

- Example

want to $\min_x \|Ax - b\|^2$.

for very large A (big) ↓ A → tolerable.

- don't want to do the linear algebra A^TA might be tough to work with

- instead, write

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \qquad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

and consider

$$\|A_1 x_1 - b_1\|^2 + \|A_2 x_2 - b_2\|^2$$

$$\text{st} \quad x_1 = x_2$$

① these two problems have solutions in the same place.

Now look at ~~ALM~~ ALM

$$\text{Aug. Lag.} = \|A_1x_1 - b_1\|^2 + \|A_2x_2 - b_2\|^2 + \lambda^T(x_1 - x_2) + \frac{c}{2}(x_1 - x_2)^T(x_1 - x_2)$$

We could "split" the step obtaining a min in x

so

$$x_1^{k+1} = \underset{x_1}{\operatorname{argmin}}\, A(x_1, x_2^k, \lambda^k, c^k)$$

[and this is a linear system which is smaller than the $\|Ax-b\|^2$ system

$$x_2^{k+1} = \underset{x_2}{\operatorname{argmin}}\, A(x_1^{k+1}, x_2, \lambda^k, c^k)$$

λ^{k+1}, c^{k+1} } as usual.

An important working example

- Sparse regression.

recall linear regression.

- we have a dataset

$$(y_i, \underline{x}_i^T)$$

and we wish to predict best estimate of future y from new $\underline{x}$

one approach:

linear model: ← d-dimensional

$$y_{pred}(\underline{x}) = \underline{x}_i^T \theta$$

↑ parameters

How to choose θ?

min least-squares error

$$\underline{Y} = \begin{bmatrix} y_1 \\ \\ y_N \end{bmatrix}$$

$$\underline{\underline{X}} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \\ \underline{x}_N^T \end{bmatrix} \theta$$

and solve:

$$\min_{\theta} \left[\underline{Y} - \underline{\underline{X}}\theta \right]^T \left[\underline{Y} - \underline{\underline{X}}\theta \right]$$

which is solved by

$$X^T X \theta = X^T Y$$

But there are problems with this story

(I) imagine $d \gg N$ ↪ many θ!

(II) even if $d \ll N$, we could have

$X^T X$ has small rank, and this must lead to large prediction errors

(III) imagine some of the x-components are irrelevant — then we want the corresponding θ-components to be 0 — otherwise they contribute error.

Practical Q: can we force lots of θ-components to be 0?

Obvious strategy doesn't work

- Obvious strategy — penalize $\theta^T\theta$

so solve

$$\|\underline{Y} - \underline{\underline{X}}\theta\|^2 + \frac{\alpha}{2}\theta^T\theta.$$

Soln:

$$(X^TX + \alpha I)\theta = X^TY$$

- notice if α is big enough, $\underline{\underline{II}}$ ~~hat~~ is solved
- Also I, but not convincingly
- We DON'T GET 0's
- L^2 regularization

OR.

Very simple example

$y_i = c$

$\underline{x}_i^T = (1, \xi_i)$ ← a sample from noise, mean = 0, var = σ^2

then $E[X^T X] = \begin{pmatrix} N & 0 \\ 0 & N\sigma^2 \end{pmatrix}$ and $E[X^T Y] = \begin{pmatrix} Nc \\ 0 \end{pmatrix}$

~~approx...~~

so if we get $\begin{pmatrix} Nc \\ 0 \end{pmatrix}$ then $\theta = \begin{bmatrix} c \\ 0 \end{bmatrix}$

But imagine $X^T Y = \begin{pmatrix} Nc \\ \varepsilon \end{pmatrix}$

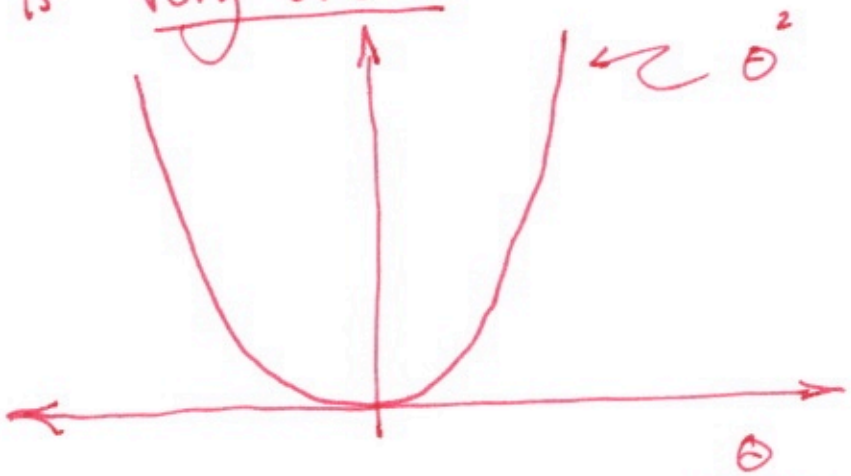
then $\theta = \begin{pmatrix} c \\ \frac{\varepsilon}{N\sigma^2} \end{pmatrix}$

if we regularize to h^2 we get

$\begin{pmatrix} \frac{c}{1+\alpha} \\ \frac{\varepsilon}{N\sigma^2+\alpha} \end{pmatrix}$ ← smaller, but not zero

Why:

- the penalty for θ_i small, but not 0 is very small



- we would like a small θ_i to be more expensive

→ rather than $\frac{1}{2}\theta^T\theta$

use $\sum_i |\theta_i| = \|\theta\|_1$

L^1 norm

absolute value.

So we could use

$$\min_{\theta} \|X\theta - Y\|^2 + \alpha |\theta|_1$$

(known as a lasso).

But how do we find a min?

Notice: $|\theta|_1$ isn't differentiable ~~at~~ (think $\theta_i = 0$)!

Notice: ignoring this will lead to problems → we won't get zero's!