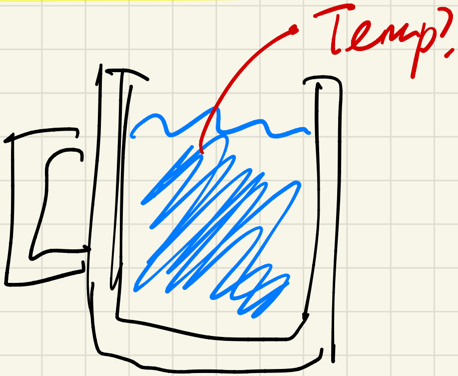
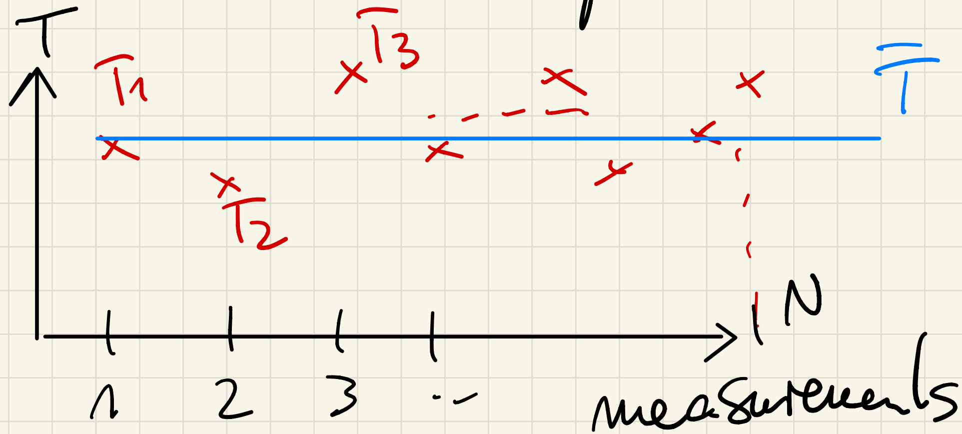
Lecture 5: Optimization meets linear algebra & calculus 2

Unit 2

Example: measure the temperature





$$\bar{T} = \frac{1}{N}\sum_{i=1}^{N} T_i \quad \leftarrow \text{input}$$

quantity to be estimated

Complexity?

$O(N)$

Where does the average come from?

Define cost function $J(T)$

$$J(T) = \sum_{i=1}^{N} (T_i - T)^2$$

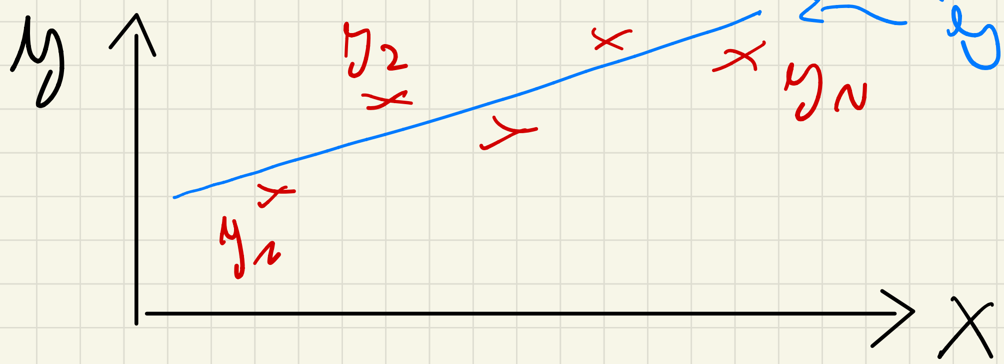
least squares

Then: $\bar{T} = \arg\min_{T} J(T)$

Indeed: $$\frac{dJ}{dT} = 0 \iff -2 \sum_{i=1}^{N} T_i - T = 0$$

$$\iff T = \frac{1}{N} \sum_{i=1}^{N} T_i \quad \blacksquare$$

Another example: linear regression



$\bar{y} = \bar{a}x + \bar{b}$

coefficients to be found.

Least squares:

$$\bar{a}, \bar{b} = \arg\min_{a,b} \sum_{i=1}^{N} \underbrace{(y_i - (ax_i + b))^2}_{J(a,b)}$$

$\bar{a}, \bar{b}$ must satisfy: $\frac{\partial J}{\partial a} \equiv 0$, $\frac{\partial J}{\partial b} \equiv 0$ $\Leftrightarrow$ $\bar{a}, \bar{b}$ as in lab 6.

What do these problems have in common?

General form: $\vec{y} = H\vec{\theta} + \vec{\varepsilon}$

"noise" (unknown)

model parameters (ex: T, $\{a, b\}$)

observation operator (model $\rightarrow$ measurements)

$H \in \mathbb{R}^{n \times p}$

$\vec{y} \in \mathbb{R}^{n}$

$\vec{\theta} \in \mathbb{R}^{p}$

Ex 1: $H = 1$

Ex 2: $H = \begin{bmatrix} 1 & x_1 \\ \vdots & \\ 1 & x_N \end{bmatrix}$

Linear least squares:

$$\hat{\vec{\theta}} = \arg\min_{\vec{\theta}} J(\vec{\theta}) \quad , \quad J(\vec{\theta}) = \|\vec{y} - H\vec{\theta}\|^2$$

2-norm

Question: Is $\hat{\vec{\theta}}$ unique?

Stationary points: $\nabla_{\vec{\theta}} J \equiv \vec{0}$

$$\partial_{\theta_k} J(\vec{\theta}) = \partial_{\theta_k} \sum_i \Big(y_i - \sum_{j=1} H_{ij}\theta_j\Big)^2$$

$$= \sum_i 2\Big(y_i - \sum_{j=1} H_{ij}\theta_j\Big) H_{ik}$$

Grouping all terms $\partial\theta_1 \cdots \partial\theta_p$:

$$\nabla_{\vec{\theta}} J = -2H^T\left(\vec{y} + H\vec{\theta}\right) \equiv \vec{0}$$

$$\Longleftrightarrow \left(H^T H\right)\vec{\theta} = H^T\vec{y}$$

$\vec{\theta}$ is unique when $H^T H$ is invertible, H is full rank

This occurs when $p \leq m$, and at least p rows of $H \in \mathbb{R}^{m \times p}$ are l.i.

$\Longleftrightarrow$ measurements need to be "different"

Is the stationary point a minimum?

In one variable, one checks if $\dfrac{d^2 J}{d\theta^2} > 0$

In more variable, the Hessian needs to be positive definite: OK if H is full-rank
$\Rightarrow H^T H$ is pos. def.

Complexity of $\underbrace{H^T H}_{O(p^2 n)} \vec{\theta} = \underbrace{H^T \vec{y}}_{O(pn)}$ ($p \times p$ lin system)

+ Solving linear system: $O(p^3)$

For a general matrix H, we can solve the optimization in $O(p^3) + O(p^2 n) + O(pn)$ ops.

What if p, n are large?..

or H is not sparse?

$\boxed{\text{BREAK}}$

Gradient descent

Given $\vec{\theta}^0$, for $K = 0, \dots, maxit$

$$\vec{\theta}^{K+1} = \vec{\theta}^K - \alpha \nabla_{\vec{\theta}} J(\vec{\theta}^K), \quad \alpha \in \mathbb{R}$$

Important! For using GD with several variables, all "$\vec{\theta}$" need to have same dimensions, s.t. $[\alpha] = [J]^1 [\vec{\theta}]^2$

Question: How to choose α to ensure convergence?

Answer: using Taylor (as in unit 1)

$$J(\vec{\theta}^{K+1}) \cong J(\vec{\theta}^{K}) + (\nabla_{\vec{\theta}} J(\vec{\theta}^{K})) \cdot (\vec{\theta}^{K+1} - \vec{\theta}^{K})$$

dot product

$$+ \frac{1}{2} (\vec{\theta}^{K+1} - \vec{\theta}^{K})^T S \, (\vec{\theta}^{K+1} - \vec{\theta}^{K})$$

Hessian (constant if J is quadratic)

Since in gradient descent

$$\vec{\theta}^{K+1} - \vec{\theta}^{K} = -\alpha \nabla_{\theta} J(\vec{\theta}^{K})$$

Then, $J(\vec{\theta}^{k+1}) = J(\vec{\theta}^{k}) \underbrace{- \alpha \|\nabla_\theta J\|^2 + \frac{\alpha^2}{2}\|\nabla_\theta J\|_S^2}$ ← S-norm

$\|x\|_S = x^T S x.$

The grad. descent decrease the cost function if < 0

$$\Longleftrightarrow (‡) \quad \frac{\alpha^{\cancel{2}}}{2}\|\nabla_\theta J\|_S^2 < \cancel{\alpha}\|\nabla_\theta J\|^2$$

So if: $\frac{1}{2}\max_x \frac{x^T S x}{x^T x} < \frac{1}{\alpha}, \ \forall x$

It holds for (‡)

But note that $\max\limits_{x} x^T S x / x^T x = \lambda_{max}$

with λ_{max} the largest eigenvalue of S.

$$\Longrightarrow \quad \boxed{\alpha < \frac{2}{\lambda_{max}}}$$

for Grad. descent to converge.

The closest you take $\alpha \to \frac{2}{\lambda_{max}}$

the fastest will be the convergence,

but this also depends on the initial guess

What about the complexity of gradient descent?

- Building $H^T H$: $O(p^2 n)$ as before

- Computation of λ_{max} with power iterations (lab 7)

$$O(n_p\, p^2)$$

n_p # of power iterations

- Grad. Descent iters

$$O(p^2 \text{ maxit})$$

p^2: matrix-vector multiplication

Total grad desc: $O(p^2 \cdot n) + O(n_p \cdot p^2) + O(pn \cdot maxit)$

v/s

Direct computation $O(p^3) + \dots$

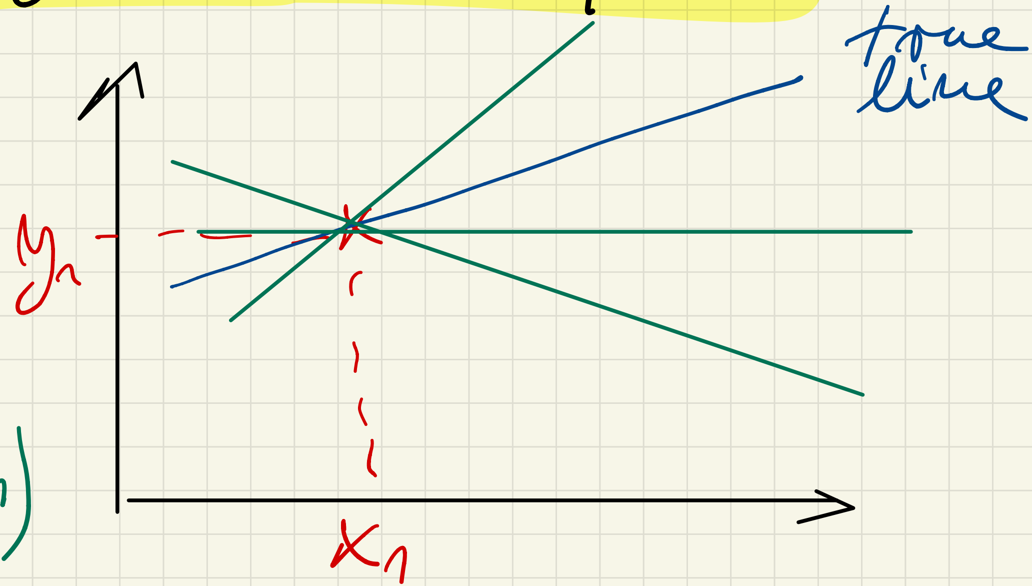
BUT: if H is sparse then complexity computations will change

$\longrightarrow$ project.

Regularized least squares

Example:

infinite lines passing through (x_1, y_1)



$\Longrightarrow$ non-unique solution

Why? 1 data point, but 2 params to estimate (slope + intersect).

Usually two
" fixes "

Assume something
on missing data

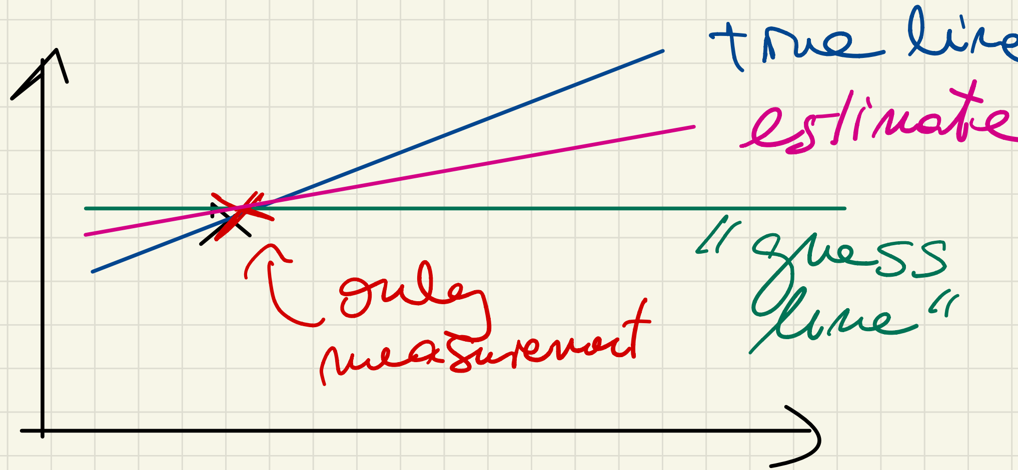
→ "create" more
measurements

→ often not good
to do.

Assume something
on parameters

we will do
this

Graphically:



Mathematically:

$$J(\vec{\theta}) = \| \vec{y} - H\vec{\theta} \|^2 + \beta \| \vec{\theta} - \vec{\theta}_0 \|_M^2$$

M pos def. given

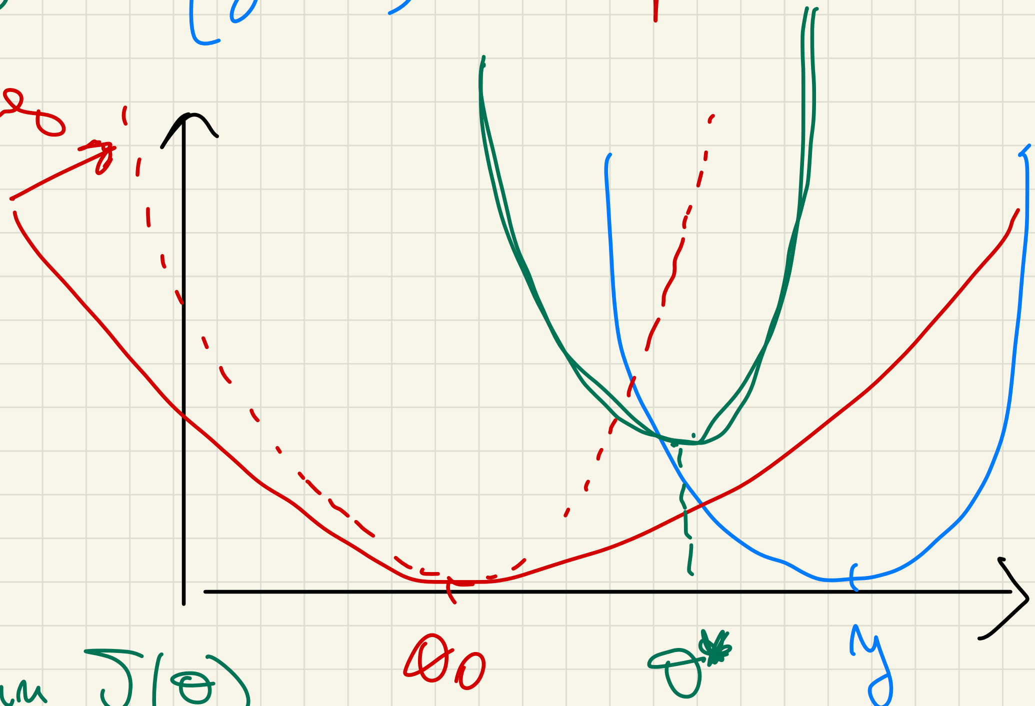
"prior" known value for the guess

$\beta > 0$. given

some convenient matrix norm

Example for 1 parameter

$$J(\theta) = (y - \theta)^2 + \beta(\theta - \theta_0)^2$$



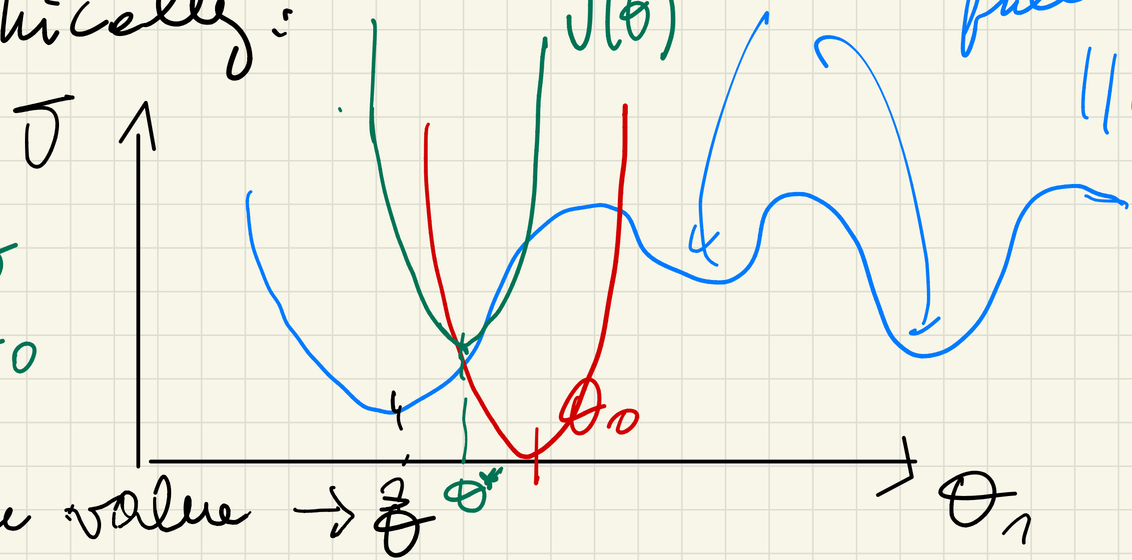
$$\theta^* = \operatorname{argmin} J(\theta)$$

Term $\beta\|\vec{\theta}-\vec{\theta}_0\|^2_M$ is called "regularization"

local minima since H is not full rank

Graphically:

$\theta^* \neq \hat{\theta}$
$\neq \theta_0$



Remark: you also can add regularizators when H is full rank. (project)

Last question: is $\vec{\theta}^*$ unique?

Yes: $\nabla_{\vec{\theta}} J = -2H^T\vec{y} + 2H^TH\vec{\theta} + 2\beta M(\vec{\theta} - \vec{\theta}_0)$

So, the stationary points satisfy

$$\underbrace{\left(\underbrace{H^TH}_{\substack{\text{semi-pos def!} \\ \text{(not invertible)}}} + \beta \underbrace{M}_{\substack{\text{sym pos def.} \\ \Rightarrow \text{invertible}}}\right)}_{\text{symm. pos. def} \longrightarrow \text{invertible!}} \vec{\theta}^* = H^T\vec{y} + \beta M\vec{\theta}_0$$

$\blacksquare$