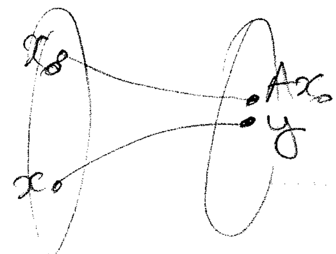
02/22/18. Regularization.

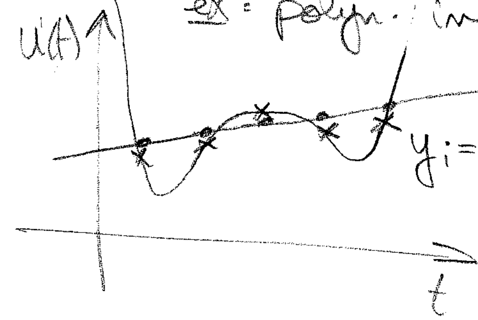
$y = Ax_0 + e$, $x = \operatorname{argmin} \|Ax - y\|$

$= (A^TA)^{-1}A^Ty$ (overdet)

$\frac{\|x - x_0\|}{\|x_0\|} \leq \kappa(A) \frac{\|e\|}{\|Ax_0\|}$ with $\kappa(A) = \|A^+\| \|A\|$



ex: polyn. interpolat$^{\circ}$



$y_i = \sum c_m t_i^m + e_i$

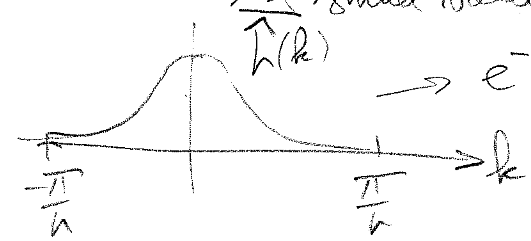
ex. image deblurring fig 4.1 p. 84 Bertero

$y = h * x_0 + e$ (h = blur kernel, PSF)

$\hat{y}(k) = \hat{h}(k)\, \hat{x}_0(k) + \hat{e}(k)$

$\hat{x}(k) = \hat{y}/\hat{h}(k) = \hat{x}_0(k) + \hat{e}(k)/\hat{h}(k)$

⚠ small values of $\hat{h}(k)$



$\to e^{-k^2\sigma^2/2}$ when spatial width σ.

like a heat equation backwards.

Tikhonov regularization: push $\|x\|$ down.

(T_λ) $\min_x \|Ax - y\|^2 + \lambda \|x\|^2$.

$\nabla = 0 \Rightarrow 2A^T(Ax - y) + 2\lambda x = 0$

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

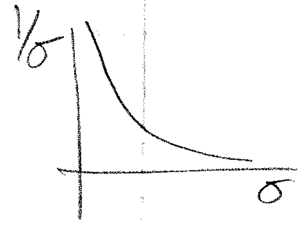
$x = (A^TA + \lambda I)^{-1} A^T y$ unreg. when $\lambda = 0$

In terms of the SVD:

$$A = U\Sigma V^* \qquad \text{with } U^*U = I \;, \quad V^*V = VV^* = I$$

$$\text{so } V^* = V^{-1}$$

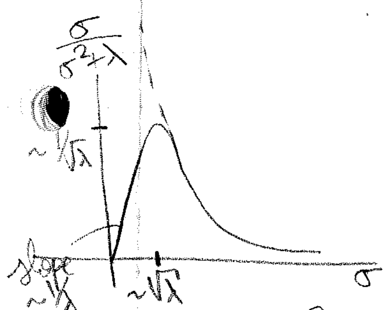
$$A^+ = (A^TA)^{-1}A^T = (V\Sigma U^* U\Sigma V^*)^{-1} V\Sigma U^*$$
$$= (V\Sigma^2 V^*)^{-1} V\Sigma U^*$$
$$= V\Sigma^{-2}V^* V\Sigma U^*$$
$$= V\Sigma^{-1}U^*$$



$$x_{LS} = A^+ y = \sum_i v_i \frac{1}{\sigma_i} u_i^* y$$

Fourier analysis if A is transl-inv.

$$(A^TA + \lambda I)^{-1}A^T = (V\Sigma U^* U\Sigma V^* + I)^{-1} V\Sigma U^*$$
$$= (V\Sigma^2 V^* + VV^*)^{-1} V\Sigma U^*$$
$$= (V(\Sigma^2 + \lambda I)V^*)^{-1} V\Sigma U^*$$
$$= V(\Sigma^2 + \lambda I)^{-1} V^* V\Sigma U^*$$
$$= V(\Sigma^2 + \lambda I)^{-1}\Sigma U^*$$



$$x_{T,\lambda} = A^{+,\lambda} y = \sum_i v_i \frac{\sigma_i}{\sigma_i^2 + \lambda} u_i^* y$$

(can do a truncated SVD instead)

Rmk: Tikhonov also proposed (early 1960s)

$$\min_x \|Ax - y\|^2 + \lambda \|Bx\|^2$$

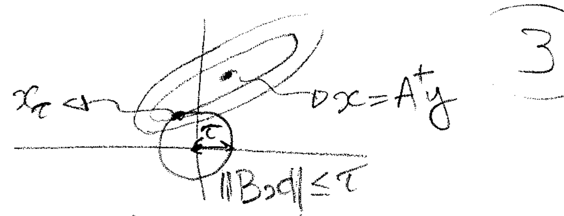
Prior: $\|Bx\|$ is small

$$\text{with } B = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & & \ddots & \\ & & & -1 & 1 \end{bmatrix}$$

$$\Rightarrow x = (A^TA + \lambda B^TB)^{-1}A^Ty$$

→ encodes smoothness of the object.
(Think $x_i = f(t_i)$ for some smooth function $f(t)$)

ex. Bertero p. 114. Intro bias-variance.
(Hubble space telescope, 1990)



Equivalent formulation (constrained).

(P_τ) $\quad \min \|Ax - y\| \quad$ s.t. $\|Bx\| \le \tau$

General: $\quad (T_\lambda) \ \min \boxed{f_0(x)} + \lambda \boxed{f_1(x)}$

$\qquad\qquad (P_\tau) \ \min f_0(x) : f_1(x) \le \tau$

Assume $\lambda, \tau > 0$

f_0, f_1 differentiable, convex.

Slater: $\forall \tau, \exists x : f_1(x) < \tau$.

<u>Prop</u>. • $\forall \tau, \exists \lambda(\tau)$: a solution x_τ of (P_τ) is also a solution of (T_λ)

• $\forall \lambda, \exists \tau(\lambda)$: a solution x_λ of (T_λ) is also a solution of (P_τ)

<u>Pf</u>. • Let x_τ be a solution to (P_τ).

$\mathcal{L}(x, \lambda) = f_0(x) + \lambda(f_1(x) - \tau)$.

Slater $\Rightarrow$ strong duality, KKT holds for some λ:

$$\begin{cases} (1)\ \nabla f_0(x_\tau) + \lambda \nabla f_1(x_\tau) = 0, & (2)\ \lambda \ge 0 \\ (3)\ f_1(x_\tau) - \tau \le 0, & (4)\ \lambda(f_1(x_\tau) - \tau) = 0. \end{cases}$$

$\mathcal{L}$ convex in x, (1) $\Rightarrow$ x_τ is a minimizer of $\mathcal{L}$, or $f_0(x) + \lambda f_1(x)$.

• Let x_λ be a solution to (T_λ)

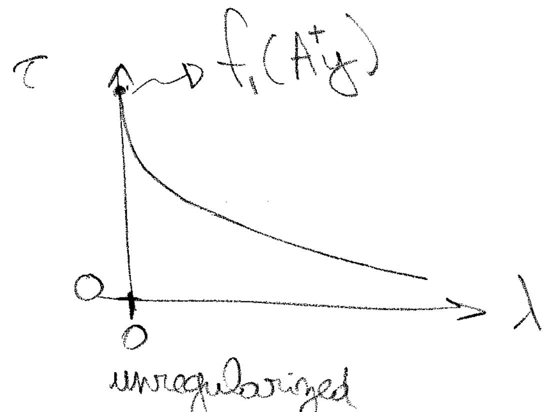
$f_0(x) + \lambda f_1(x)$ convex in x $\Rightarrow$

$\nabla f_0(x_\lambda) + \lambda \nabla f_1(x_\lambda) = 0 \qquad (1)$

Pick $\tau = f_1(x_\lambda)$, then (2), (3), (4) hold for $\mathcal{L}(x, \lambda) = f_0(x) + \lambda(f_1(x) - \tau)$

$\Rightarrow x_\lambda$ is optimal for (P_τ)

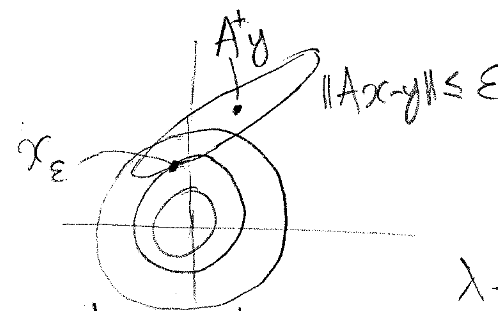
$\square$



$\lambda \to \tau = f_1(A^{+,\lambda} y)$
(other way is less direct).

Also equivalent to

$$(Q_\varepsilon) \quad \min \|Bx\| \quad \text{s.t.} \quad \|Ax - y\| \leq \varepsilon \, .$$



(similar argument)

$\lambda \to \varepsilon = f_0(A^{+,\lambda} y)$

In practice: how to choose $\tau / \lambda / \varepsilon$?

- provide all to user!
- estimate noise level ε such that $\|Ax - y\| \simeq \varepsilon$

  ex. median absolute deviation (MAD)

  $$\frac{\text{median}(X_1, \dots X_n)}{0.6745}$$ on a small patch where the mean is removed.

  (Then find λ / τ to match ε)
- cross-validation

  "leave one datum out":

  $$\min_{x_{\lambda,k}} \|A_{\sim k}\, x - y_{\sim k}\|^2 + \lambda \|Bx\|^2 \longrightarrow x_{\lambda,k}$$

  $A_{\sim k} =$ $\uparrow k$ removed $\qquad y_{\sim k} =$ $\leftarrow k$ removed

then measure prediction of remaining component: cross validation function to min over λ:

$$CV(\lambda) = \sum_k |(A x_{\lambda k})_k - y_k|^2$$

Hard to compute. But ($)

$$CV(\lambda) = \sum_k \frac{|(A x_\lambda)_k - y_k|^2}{|1 - P_{kk}(\lambda)|^2}$$

with $\begin{cases} P(\lambda) = A(A^*A + \lambda I)^{-1}A^* \\ \quad \text{(regularized projector onto Ran } A) \\ x_\lambda = \operatorname{argmin} \|Ax - y\|^2 + \lambda \|By\|^2 \end{cases}$

Issue: $V(\lambda)$ depends on the ordering of the rows of A / elements of y.

To fix this, consider the generalized CV function

$$GCV(\lambda) = \sum_k \frac{|(A x_\lambda)_k - y_k|^2}{|m - \operatorname{tr}(P(\lambda))|^2}. \qquad A \in \mathbb{R}^{m \times n}, \quad P(\lambda) \in \mathbb{R}^{m \times m}$$

$$= \frac{\|A x_\lambda - y\|^2}{(\operatorname{tr}(I - P(\lambda)))^2}$$

Mueller-Siltanen examples