Least squares regularization

Lecture 08

September 23, 2021

Beck, sections 3.3-3.4

Least squares problems

Recall: $\min_{\underline{x} \in \mathbb{R}^N} f(\underline{x})$, $\quad f(\underline{x}) = \underline{x}^T \underline{\underline{A}}^T \underline{\underline{A}}\, \underline{x} - 2 \underline{b}^T \underline{\underline{A}}\, \underline{x} + \|b\|_2^2$

$$= \|\underline{\underline{A}}\,\underline{x} - \underline{b}\|_2^2$$

If $\underline{\underline{A}}$ has full column rank ($\mathrm{rank}(\underline{\underline{A}}) = N$), then

$\underline{x}_* = (\underline{\underline{A}}^T \underline{\underline{A}})^{-1} \underline{\underline{A}}^T b$ is the (strict) global minimum

The problem: Small residuals don't translate into good models

or minimal

Idea: Penalize/Discourage "bad" behavior of models.

Quantitative ways to enforce this penalization are called regularization.

Almost all regularization strategies boil down to controlling or minimizing $\|\underline{\underline{R}}\,\underline{x}\|_2^2$, where $\underline{\underline{R}}$ is a regularizing matrix.

Overall: (for least squares)

$$\min_{\underline{x}\in\mathbb{R}^N} \text{"} \|\underline{\underline{A}}\,\underline{x}-b\|_2^2 \text{ and } \|\underline{\underline{R}}\,\underline{x}\|_2^2 \text{"}$$

Since we can't really do the above, we try to minimize a weighted sum of these:

Given $\lambda \in \mathbb{R}_{++}$, solve

$$\min_{\underline{x} \in \mathbb{R}^N} f(\underline{x}) + \lambda \|\underline{\underline{R}}\,\underline{x}\|_2^2$$

$$\shortparallel$$

$$\min_{\underline{x} \in \mathbb{R}^N} \|\underline{\underline{A}}\,\underline{x} - \underline{b}\|_2^2 + \lambda \|\underline{\underline{R}}\,\underline{x}\|_2^2$$

This type of regularization is called Tikhonov Regularization.

This regularization attempts to balance "data misfit" ($\|\underline{\underline{A}}\,\underline{x} - \underline{b}\|_2^2$) and regularization ($\|\underline{\underline{R}}\,\underline{x}\|_2^2$).

Since this is another quadratic function, we can globally minimize it:

$$\min_{\underline{x} \in \mathbb{R}^N} f(\underline{x}) \quad , \quad f(\underline{x}) = \|\underline{\underline{A}}\,\underline{x} - \underline{b}\|_2^2 + \lambda \|\underline{\underline{R}}\,\underline{x}\|_2^2$$

$$= \underline{x}^T (\underline{\underline{A}}^T \underline{\underline{A}} + \lambda \underline{\underline{R}}^T \underline{\underline{R}}) \underline{x} - 2 \underline{b}^T \underline{\underline{A}}\,\underline{x} + \|\underline{b}\|_2^2.$$

$$\nabla^2 f(\underline{x}) = 2 (\underline{\underline{A}}^T \underline{\underline{A}} + \lambda \underline{\underline{R}}^T \underline{\underline{R}})$$

Global optimality guaranteed if $\underline{\underline{A}}^T\underline{\underline{A}} + \lambda \underline{\underline{B}}^T\underline{\underline{B}}$ is of full column rank, i.e. $\operatorname{rank}(\underline{\underline{A}}^T\underline{\underline{A}} + \lambda \underline{\underline{B}}^T\underline{\underline{B}}) = N$.

$\Longleftrightarrow \underline{\underline{A}}^T\underline{\underline{A}} + \lambda \underline{\underline{B}}^T\underline{\underline{B}} \succ \underline{\underline{0}}$.

Under this assumption, $\nabla f = \underline{0}$ defines the unique global minimum.

$$\nabla f(\underline{x}) = 2(\underline{\underline{A}}^T\underline{\underline{A}} + \lambda \underline{\underline{B}}^T\underline{\underline{B}})\underline{x} - 2\underline{\underline{A}}^T\underline{b} = \underline{0}$$

The solution is

$$\underline{x}_* = (\underline{\underline{A}}^T\underline{\underline{A}} + \lambda \underline{\underline{B}}^T\underline{\underline{B}})^{-1}\underline{\underline{A}}^T\underline{b}$$

(normal equations)

Great, but how do we choose $\underline{\underline{B}}$? (and λ?)

Regularization: noisy data

In some applications, data ($\underline{b}$) is noisy.

I.e. we solve $\min_{\underline{x}} \|\underline{A}\,\underline{x} - \tilde{\underline{b}}\|_2^2$ instead of

$\min_{\underline{x}} \|\underline{A}\,\underline{x} - \underline{b}\|_2^2$, where $\tilde{b}$ is

a noisy/perturbed version of $\underline{b}$, e.g.

$\tilde{\underline{b}} = \underline{b} + \underline{\varepsilon}$ ← noise vector.

Frequently, the least squares solution is sensitive to data values, so $\underline{\tilde{x}}_* = (\underline{\underline{A}}^T \underline{\underline{A}})^{-1} \underline{\underline{A}}^T \tilde{b}$ can be very different from $\underline{x}_* = (\underline{\underline{A}}^T \underline{\underline{A}})^{-1} \underline{\underline{A}}^T \underline{b}$.

Typically, noise makes the norm of $\underline{\tilde{x}}$ very large compared to the norm of $\underline{x}$.

In this (particular) case, penalizing $\|\underline{x}\|_2^2$ <u>might</u> fix the problem.

E.g., $\underline{\underline{R}} = \underline{\underline{I}}$: $\|\underline{\underline{R}}\,\underline{x}\|_2^2 = \|\underline{x}\|_2^2$

I.e., $\underline{x}_* = (\underline{\underline{A}}^T \underline{\underline{A}} + \lambda \underline{\underline{I}})^{-1} \underline{\underline{A}}^T \underline{b}$.

Regularization: overfitting

In this case, poor model specification can be mitigated.

Challenge: $\underline{x}_* = (\underline{\underline{A}}^T \underline{\underline{A}})^{-1} \underline{\underline{A}}^T \underline{b}$ chooses a model that overfits the data: model "tries too hard" to fit data.

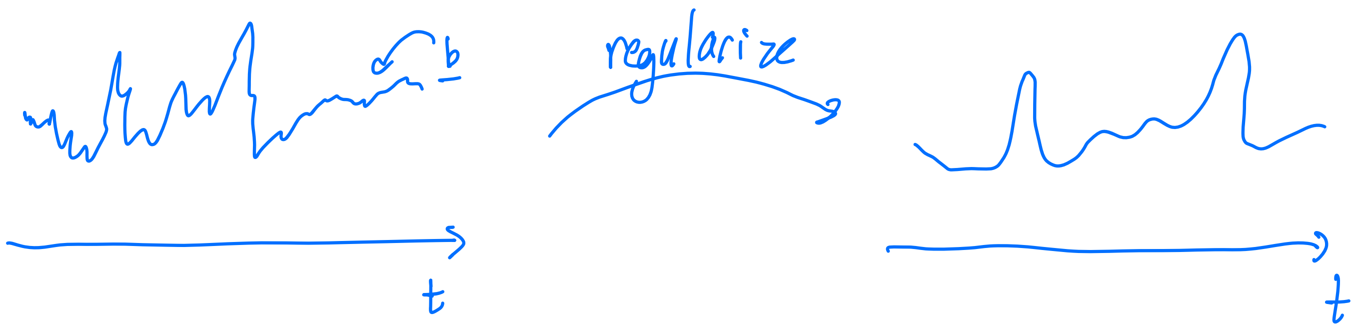
Goal: design $\underline{\underline{R}}$ to penalize models that are too "complex". This is, unfortunately, an art.

E.g., $\|\underline{\underline{B}}\,\underline{x}\|_2^2 = \|\underline{x}\|_2^2$ could work

Frequently: need to design $\underline{\underline{B}}$ to "capture" the "complex" that we want to penalize.

Regularization: denoising

Challenge: given signal data $\underline{b}$, "smooth" $\underline{b}$.



I.e.: construct a regularized signal $\underline{x}$ s.t.

(i) "$\underline{x} \approx \underline{b}$", i.e. $\|\underline{x}-\underline{b}\|_2^2$ is small.

(ii) $\underline{x}$ is "smooth", i.e. minimize $\|\underline{\underline{B}}\,\underline{x}\|_2^2$.

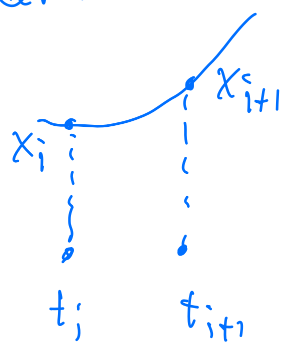
Once $\underline{\underline{R}}$ is chosen, this is computationally simple:

"$\underline{\underline{A}} = \underline{\underline{I}}$"

$$\min \|\underline{x} - \underline{b}\|_2^2 + \lambda \|\underline{\underline{R}}\,\underline{x}\|_2^2$$

$$\underline{x}_* = (\underline{\underline{I}} + \lambda \underline{\underline{R}}^T \underline{\underline{R}})^{-1} \underline{\underline{I}}^T \underline{b}.$$

To smooth out $\underline{x}$: enforce regularization on "derivative" of $\underline{x}$.



$$\left.\frac{dx}{dt}\right|_{t_i} \simeq \frac{x_{i+1} - x_i}{t_{i+1} - t_i}$$

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{pmatrix} \xrightarrow{\text{if points are equispaced}} \begin{pmatrix} x_2 - x_1 \\ x_3 - x_2 \\ \vdots \\ x_M - x_{M-1} \end{pmatrix}$$

measures derivatives.

Goal: make norm of ⤴ small.

$$\begin{pmatrix} x_2 - x_1 \\ \vdots \\ x_n - x_{n-1} \end{pmatrix} = \underbrace{\begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}}_{\underline{\underline{R}} \in \mathbb{R}^{(n-1)\times n}} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$$\min_{\underline{x} \in \mathbb{R}^n} \underbrace{\| \underline{x} - \underline{b} \|_2^2}_{\text{data misfit}} + \lambda \underbrace{\| \underline{\underline{R}}\, \underline{x} \|_2^2}_{\text{norm of derivative.}}$$

The choice of λ