REPRESENTER THEOREM

Dr. Matthieu R Bloch

Wednesday October 13, 2021

LOGISTICS

Assignment 4 due **October 14, 2021**

- Hard deadline on October 16
- Small update posted

Kayla's office hours tomorrow Thursday October 14, 2021: 11am

Assignment 2 grades released: (2.6/2.7/2.8 not graded)

- Mean: 22.64 - Median: 23.1 - Min: 7.5 - Max: 24.6 (clipped at 24)

Assignment 3: 45% graded

Midterm 1: 75% graded

WHAT'S ON THE AGENDA FOR TODAY?

Last time: solving least squares

- Minimum $||\cdot||_2$ solution
- Regularized least squares

Recall: $\min_{\theta} \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2$

hyperparameter

$\lambda > 0$

$$\hat{\theta} = (\underbrace{X^T X}_{d \times d} + \lambda I)^{-1} X^T y = X^T (\underbrace{X X^T}_{n \times n} + \lambda I)^{-1} y$$

$X \in \mathbb{R}^{n \times d}$

$$X = \begin{array}{c} \overset{d}{\longleftrightarrow} \\ \begin{pmatrix} - & x_1^T & - \\ & \vdots & \\ - & x_n^T & - \end{pmatrix} \end{array} \Big\updownarrow n$$

$$X X^T = \left[\; x_i^T x_j \; \right]_{i,j \in [\![1, n]\!]}$$

WHAT'S ON THE AGENDA FOR TODAY?

Last time: solving least squares

- Minimum $||\cdot||_2$ solution
- Regularized least squares

Today

- Extension to infinite dimension
- Representer theorem

Reading: Romberg, lecture notes 8/9

RIDGE REGRESSION

We can adapt the regularization approach to the situation of a $\boxed{\text{finite}}$ dimension Hilbert space $\mathcal{F}$

$\lambda > 0$ hyperparameter

$$\underbrace{\min_{f\in\mathcal{F}}}\sum_{i=1}^{n}(\underset{\in\mathbb{R}}{y_i} - f(\underset{\in\mathbb{R}^d}{\mathbf{x}_i}))^2 + \underbrace{\lambda\|f\|_{\mathcal{F}}^2}_{\text{controls complexity of solution}}$$

RIDGE REGRESSION

We can adapt the regularization approach to the situation of a finite dimension Hilbert space $\mathcal{F}$

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 + \lambda\|f\|_{\mathcal{F}}^2 \quad (*)$$

equivalent

- We are penalizing the norm of the entire function f

Using a basis for the space $\{\psi_i\}_{i=1}^{d}$, and constructing $\boldsymbol{\Psi}$ as earlier, we obtain

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\theta}\|_2^2 + \lambda\boldsymbol{\theta}^\intercal\mathbf{G}\boldsymbol{\theta} \quad (**)$$

depends on basis depends on basis

with $\mathbf{G}$ the Gram matrix for the basis.

Proof of equivalence : we introduce a basis $\{\psi_i\}_{i=1}^{d} \in \mathcal{F}$

$$\forall f \in \mathcal{F}, \text{ we can write } f = \sum_{i=1}^{d} \underset{\in \mathbb{R}}{\theta_i} \psi_i$$

$$\min_f \underbrace{\sum_{i=1}^{n} |y_i - f(x_i)|^2}_{1} + \lambda \underbrace{\|f\|_{\mathcal{F}}^2}_{2}$$

1 : equivalent to $\|\underset{\in \mathbb{R}^n}{y} - \Psi\theta\|_2^2$ $\qquad \Psi = \begin{bmatrix} \psi_1(x_1) & \cdots & \psi_d(x_1) \\ & \vdots & \\ & \psi_j(x_i) & \end{bmatrix}$

2 : $\|f\|_{\mathcal{F}}^2 = \left\|\sum_{i=1}^{d} \theta_i \psi_i\right\|_{\mathcal{F}}^2 = \left\langle \sum_{i=1}^{d} \theta_i \psi_i, \sum_{j=1}^{d} \theta_j \psi_j \right\rangle = \sum_{i=1}^{d}\sum_{j=1}^{d} \theta_i \theta_j \langle \psi_i, \psi_j \rangle \triangleq \theta^T G \theta$

where $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}$ $\qquad G = \left[\langle \psi_i, \psi_j \rangle \right]_{i,j \in [\![1;n]\!]}$

RIDGE REGRESSION

We can adapt the regularization approach to the situation of a finite dimension Hilbert space $\mathcal{F}$

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 + \lambda\|f\|_{\mathcal{F}}^2$$

- We are penalizing the norm of the entire function f

Using a basis for the space $\{\psi_i\}_{i=1}^{d}$, and constructing $\mathbf{\Psi}$ as earlier, we obtain

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{\Psi}\boldsymbol{\theta}\|_2^2 + \lambda\boldsymbol{\theta}^\intercal\mathbf{G}\boldsymbol{\theta}$$

with $\mathbf{G}$ the Gram matrix for the basis.

⚠ $G \neq I$ in general

If $\boxed{\mathbf{\Psi}^\intercal\mathbf{\Psi} + \lambda\mathbf{G}}$ is invertible, we find the solution as

$$\boldsymbol{\theta}^* = (\mathbf{\Psi}^\intercal\mathbf{\Psi} + \lambda\mathbf{G})^{-1}\mathbf{\Psi}^\intercal\mathbf{y} \quad ✓$$

and we can reconstruct the function as $f(\mathbf{x}) = \sum_{i=1}^{d} \theta_i^* \psi_i(\mathbf{x})$. ✗

$$M = X^T X + \lambda G$$

$M^T = X^T X + \lambda G^T = X^T X + \lambda G$ is symmetric semidefinite positive

For any $\theta \in \mathbb{R}^d$ $\quad \theta^T M \theta = \theta^T X^T X \theta + \lambda\, \theta^T G \theta = \|X\theta\|_2^2 + \lambda \left\| \sum_{i=1}^{d} \theta_i \psi_i \right\|_{\mathcal{F}}^2 \geq 0$

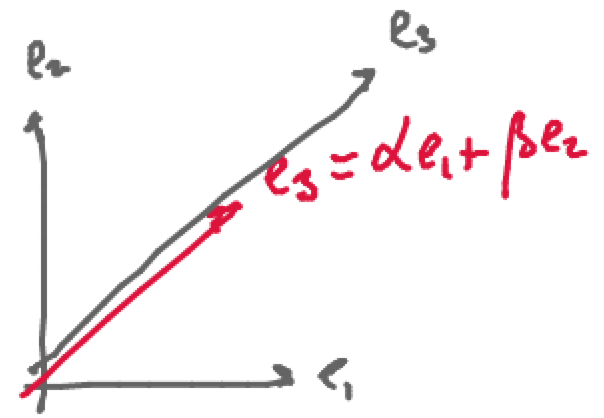
Because $\{\psi_i\}$ is a basis, we know that we can invert M

Assume θ is such that $\theta^T M \theta = 0$ Then $\|X\theta\|_2^2 + \lambda \left\| \sum_{i=1}^{d} \theta_i \psi_i \right\|_{\mathcal{F}}^2 = 0$ and $\begin{cases} \|X\theta\|_2^2 = 0 \\ \left\| \sum_{i=1}^{d} \theta_i \psi_i \right\|_{\mathcal{F}}^2 = 0 \end{cases}$

Hence $\theta \in \operatorname{Ker} X$ and $\sum_{i=1}^{d} \theta_i \psi_i = 0$ so that by linear independence $\forall i \; \theta_i = 0$

Hence $\theta = 0$

<u>Note</u>: $\mathbb{R}^2$



RIDGE REGRESSION

We can adapt the regularization approach to the situation of a finite dimension Hilbert space $\mathcal{F}$

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_{\mathcal{F}}^2$$

- We are penalizing the norm of the entire function f

Using a basis for the space $\{\psi_i\}_{i=1}^{d}$, and constructing $\boldsymbol{\Psi}$ as earlier, we obtain

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\theta}\|_2^2 + \lambda\boldsymbol{\theta}^\intercal\mathbf{G}\boldsymbol{\theta}$$

with $\mathbf{G}$ the Gram matrix for the basis.

If $\boldsymbol{\Psi}^\intercal\boldsymbol{\Psi} + \lambda\mathbf{G}$ is invertible, we find the solution as

$$\boldsymbol{\theta}^* = (\boldsymbol{\Psi}^\intercal\boldsymbol{\Psi} + \lambda\mathbf{G})^{-1}\boldsymbol{\Psi}^\intercal\mathbf{y} = \Psi^\intercal \underbrace{(\Psi\Psi^\intercal + \lambda G)^{-1} y}_{\alpha} = \text{linear comb o}$$

rows of Ψ

and we can reconstruct the function as $f(\mathbf{x}) = \sum_{i=1}^{d} \theta_i^*\psi_i(\mathbf{x})$.

$\Psi_1 = (\psi_1(x_1)\ \psi_2$

If $\mathbf{G}$ is well conditioned, the resulting function is not too sensitive to the choice of the basis

LEAST-SQUARES IN INFINITE DIMENSION HILBERT SPACES

In $\mathbb{R}^d$, the problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$ has a solution

$$\boldsymbol{\theta}^* = \mathbf{X}^\intercal\boldsymbol{\alpha} \text{ with } \boldsymbol{\alpha} = (\mathbf{X}\mathbf{X}^\intercal + \lambda\mathbf{I})^{-1}\mathbf{y}$$

$\mathbf{X}\mathbf{X}^\intercal \in \mathbb{R}^{n\times n}$ is dimension independent! We will be able to extend this to infinite dimensional Hilbert spaces!

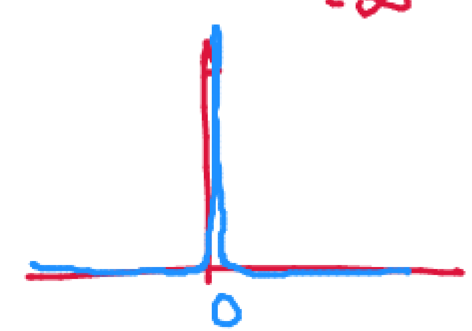
Let $\mathcal{F}$ be a Hilbert space and let $f \in \mathcal{F}$ be the function we are trying to estimate

- We will estimate $f \in \mathcal{F}$ using noisy observations $\langle f, x_i\rangle$ with $\{x_i\}_{i=1}^n$ elements of $\mathcal{F}$

fct of x_i

↳ for some Hilbert spaces $\langle f, x_i\rangle = f(t_i)$

⚠ $f(t) = \int_{-\infty}^{+\infty} f(u)\,\delta(t-u)\,du$ ←



$\int_{-\infty}^{+\infty} \delta(u)\,du = 1$

LEAST-SQUARES IN INFINITE DIMENSION HILBERT SPACES

In $\mathbb{R}^d$, the problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$ has a solution

$$\boldsymbol{\theta}^* = \mathbf{X}^\intercal\boldsymbol{\alpha} \text{ with } \boldsymbol{\alpha} = (\mathbf{X}\mathbf{X}^\intercal + \lambda\mathbf{I})^{-1}\mathbf{y}$$

$\mathbf{X}\mathbf{X}^\intercal \in \mathbb{R}^{n\times n}$ is dimension independent! We will be able to extend this to infinite dimensional Hilbert spaces!

Let $\mathcal{F}$ be a Hilbert space and let $f \in \mathcal{F}$ be the function we are trying to estimate

- We will estimate $f \in \mathcal{F}$ using noisy observations $\langle f, x_i\rangle$ with $\{x_i\}_{i=1}^n$ elements of $\mathcal{F}$
- This is the equivalent of saying $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$ in finite dimension

Proposition (Representer theorem)

$$\min_{f\in\mathcal{F}} \sum_{i=1}^n \left|y_i - \langle f, x_i\rangle_{\cancel{\mathcal{H}}}\right|^2 + \lambda\|f\|^{2}_{\cancel{\mathcal{H}}}$$

has solution

$\in \mathcal{F}$ $\quad\quad \in \mathbb{R}^{n\times n}$

$$f = \sum_{i=1}^n \alpha_i x_i \text{ with } \boldsymbol{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \qquad \mathbf{K} = [\langle x_i, x_j\rangle]_{1\le i,j\le n}$$

$\in \mathbb{R}$

Proof. $\min_f \sum_{i=1}^{n} |y_i - \langle f, x_i \rangle|^2 + \lambda \|f\|^2 \quad (*)$

$\forall f$ we can write $f = f_1 + f_2$ w/ $f_1 \in \mathrm{Span}(\{x_i\}_{i=1}^n)$ and $f_2 \in \mathrm{Span}(\{x_i\}_{i=1}^n)^\perp$

closest point to f in $\mathrm{Span}(\{x_i\}_{i=1}^n)$

Note that $\forall i \ \langle f_2, x_i \rangle = 0$

Then $\sum_{i=1}^{n} |y_i - \langle f, x_i \rangle|^2 = \sum_{i=1}^{n} |y_i - \langle f_1, x_i \rangle|^2$

$\|f\|^2 = \|f_1 + f_2\|^2 = \|f_1\|^2 + \underbrace{\|f_2\|^2}_{\geq 0}$ (Pythagorean theorem)

Hence any solution of $(*)$ must live in $\mathrm{Span}(\{x_i\}_{i=1}^n)$

LEAST-SQUARES IN INFINITE DIMENSION HILBERT SPACES

In $\mathbb{R}^d$, the problem $\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2$ has a solution

$$\boldsymbol{\theta}^* = \mathbf{X}^\intercal \boldsymbol{\alpha} \text{ with } \boldsymbol{\alpha} = (\mathbf{X}\mathbf{X}^\intercal + \lambda \mathbf{I})^{-1} \mathbf{y}$$

$\mathbf{X}\mathbf{X}^\intercal \in \mathbb{R}^{n \times n}$ is dimension independent! We will be able to extend this to infinite dimensional Hilbert spaces!

Let $\mathcal{F}$ be a Hilbert space and let $f \in \mathcal{F}$ be the function we are trying to estimate

- We will estimate $f \in \mathcal{F}$ using noisy observations $\langle f, x_i \rangle$ with $\{x_i\}_{i=1}^n$ elements of $\mathcal{F}$
- This is the equivalent of saying $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$ in finite dimension

Proposition (Representer theorem)

$$\min_{f \in \mathcal{F}} \sum_{i=1}^n \left| y_i - \langle f, x_i \rangle_{\mathcal{H}} \right|^2 + \lambda \|f\|_{\mathcal{H}}$$

has solution

$$f = \sum_{i=1}^n \alpha_i x_i \text{ with } \boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y} \qquad \mathbf{K} = \left[\langle x_i, x_j \rangle\right]_{1 \le i,j \le n}$$

We will see that the situation of the representer theorem happens in Reproducing Kernel Hilber Space (RKHS)