# REGRESSION

Dr. Matthieu R Bloch

Monday October 04, 2021

# LOGISTICS

**Assignment 4** assigned tonight

- Includes a programming component
- Due **October 13, 2021** (soft deadline, hard deadline on October 15)

# WHAT'S ON THE AGENDA FOR TODAY?

**Last time**: Non-Orthobases

- Dual basis

**Today**

- Wrap up non-orthobases in infinite dimension
- Least-square regression

**Reading:** Romberg, lecture notes 7/8

# NON-ORTHOGONAL BASES IN INFINITE DIMENSION

**Definition.**
$\{v_i\}_{i=1}^{\infty}$ is a **Riesz basis** for Hilbert space $\mathcal{H}$ if $\mathrm{cl}(\mathrm{span}(\{v_i\}_{i=1}^{\infty})) = \mathcal{H}$ and there exists $A, B > 0$ such that

$$A \sum_{i=1}^{\infty} \alpha_i^2 \leq \left\| \sum_{i=1}^{n} \alpha_i v_i \right\|_{\mathcal{H}}^2 \leq B \sum_{i=1}^{\infty} \alpha_i^2$$

**uniformly** for all sequences $\{\alpha_i\}_{i \geq 1}$ with $\sum_{i \geq 1} \alpha_i^2 < \infty$.

In infinite dimension, the existence of $A, B > 0$ is **not** automatic.

**Examples**

# NON-ORTHOGONAL BASES IN FINITE DIMENSION: DUAL BASIS

Computing expansion on Riesz basis not as simple in infinite dimension: Gram matrix is “infinite”

The Grammiam is a **linear operator**

$$\mathcal{G} : \ell_2(\mathbb{Z}) \to \ell_2(\mathbb{Z}) : \mathbf{x} \mapsto \mathbf{y} \text{ with } [\mathcal{G}(\mathbf{x})]_n \triangleq y_n = \sum_{\ell=-\infty^\infty} \langle v_\ell, v_n \rangle x_\ell$$

**Fact:** there exists another linear operator $\mathcal{H} : \ell_2(\mathbb{Z}) \to \ell_2(\mathbb{Z})$ such that

$$\mathcal{H}(\mathcal{G}(\mathbf{x})) = \mathbf{x}$$

We can replicate what we did in finite dimension!

# REGRESSION

A fundamental problem in unsupervised machine learning can be cast as follows

**Given a dataset $\mathcal{D} \triangleq \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, how do we find $f$ such that $f(\mathbf{x}_i) \approx y_i$ for all $i \in \{1, \cdots, n\}$?**

- Often $\mathbf{x}_i \in \mathbb{R}^d$, but sometimes $\mathbf{x}_i$ is a weirder object (think tRNA string)
- if $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ with $|\mathcal{Y}| < \infty$, the problem is called classification
- if $y_i \in \mathcal{Y} = \mathbb{R}$, the problem is called *regression*

We need to introduce several ingredients to make the question well defined

1. We need a class $\mathcal{F}$ to which $f$ should belong
2. We need a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ to measure the quality of our approximation

We can then formulate the question as

$$\min_{f \in \mathcal{F}} \sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i)$$

We will focus quite a bit on the *square loss* $\ell(u, v) \triangleq (u - v)^2$, called *least-square regression*

# LEAST SQUARE LINEAR REGRESSION

A classical choice of $\mathcal{F}$ is the set of continuous linear functions.

- $f : \mathbb{R}^d \to \mathbb{R}$ is *linear* iff

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \lambda, \mu \in \mathbb{R} \quad f(\lambda \mathbf{x} + \mu \mathbf{y}) = \lambda f(\mathbf{x}) + \mu f(\mathbf{y})$$

- We will see that every continuous linear function on $\mathbb{R}^d$ is actually an inner product, i.e.,

$$\exists \boldsymbol{\theta}_f \in \mathbb{R}^d \text{ s.t. } f(\mathbf{x}) = \boldsymbol{\theta}_f^\intercal \mathbf{x} \quad \forall \mathbf{x} \in \mathbb{R}^d$$

## Canonical form I

- Stack $\mathbf{x}_i$ as row vectors into a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, stack $y_i$ as elements of column vector $\mathbf{y} \in \mathbb{R}^n$

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \text{ with } \mathbf{X} \triangleq \begin{bmatrix} -\mathbf{x}_1^\intercal- \\ \vdots \\ -\mathbf{x}_n^\intercal- \end{bmatrix}$$

# LEAST SQUARE AFFINE REGRESSION

## Canonical form II

- Allow for *affine* functions (not just linear)
- Add a 1 to every $\mathbf{x}_i$

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{d+1}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \text{ with } \mathbf{X} \triangleq \begin{bmatrix} 1 - \mathbf{x}_1^\intercal - \\ \vdots \\ 1 - \mathbf{x}_n^\intercal - \end{bmatrix}$$

# NONLINEAR REGRESSION USING A BASIS

Let $\mathcal{F}$ be an $$d-dimensional subspace of a vector space with basis $\{\psi_i\}_{i=1}^d$

- We model $f(\mathbf{x}) = \sum_{i=1}^d \theta_i \psi_i(\mathbf{x})$

The problem becomes

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \|\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\theta}\|_2^2 \text{ with } \boldsymbol{\Psi} \triangleq \begin{bmatrix} -\psi(\mathbf{x}_1)^\intercal- \\ \vdots \\ -\psi(\mathbf{x}_n)^\intercal- \end{bmatrix} \triangleq \begin{bmatrix} \psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) & \cdots & \psi_d(\mathbf{x}_1) \\ \vdots & \vdots & \vdots & \vdots \\ \psi_1(\mathbf{x}_n) & \psi_2(\mathbf{x}_n) & \cdots & \psi_d(\mathbf{x}_n) \end{bmatrix}$$

We are recovering a nonlinear function of a continuous variable

- This is the exact same computational framework as linear regression.

# SOLVING THE LEAST-SQUARES PROBLEM

**Proposition.** Any solution $\boldsymbol{\theta}^*$ to the problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ must satisfy

$$\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\theta}^* = \mathbf{X}^\intercal\mathbf{y}$$

This system is called *normal equations*

**Facts:** for any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$

- $\ker \mathbf{A}^\intercal\mathbf{A} = \ker \mathbf{A}$
- $\mathrm{col}(\mathbf{A}^\intercal\mathbf{A}) = \mathrm{row}(\mathbf{A})$
- $\mathrm{row}(\mathbf{A})$ and $\ker \mathbf{A}$ are orthogonal complements

We can say a lot more about the normal equations

1. There is always a solution
2. If **rank**$(\mathbf{X}) = d$, there is a unique solution
3. if **rank**$(\mathbf{X}) < d$ there are infinitely many non-trivial solution
4. if **rank**$(\mathbf{X}) = n$, there exists a solution $\boldsymbol{\theta}^*$ for which $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^*$

In machine learning, there are often infinitely many solutions

# MINIMUM NORM 2 SOLUTIONS

One reasonable to choose a solution among infintely many is the *minimum energy* principle

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^\intercal \mathbf{y}$$

- We will see the solution is always unique

For now, assume that **rank**$(\mathbf{X}) = d$, so that the problem becomes

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \mathbf{X} \boldsymbol{\theta} = \mathbf{y}$$

**Proposition.** The solution is $\boldsymbol{\theta}^* = \mathbf{A}^\intercal (\mathbf{A}\mathbf{A}^\intercal)^{-1} \mathbf{y}$

# REGULARIZATION

Recall the problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \mathbf{X}^\top \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^\top \mathbf{y}$$

- There are infinitely many solution if $\ker \mathbf{X}$ is non trivial
- The space of solution is unbounded!
- Even if $\ker \mathbf{X} = \{0\}$, the system can be poorly conditioned

**Regularization** with $\lambda > 0$ consists in solving

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

- This problem *always* has a unique solution

# RIDGE REGRESSION

We can adapt the regularization approach to the situation of a 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

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

and we can reconstruct the function as

$$f(\mathbf{x}) = \sum_{i=1}^{d} \theta_i^* \psi_i(\mathbf{x})$$

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