Linear regression

CS 446 / ECE 449

2022-02-19 22:48:15 -0600 (383ae58)

Plan for today

- Linear regression setup revisited.
- Normal equations, SVD, and pseudoinverse.
- Example (if time).

“pytorch meta-algorithm”

1. Clean/augment data.
2. Pick model/architecture.
3. Pick a loss function measuring model fit to data.
4. Run a gradient descent variant to fit model to data.
5. Tweak 1-4 until training error is small.
6. Tweak 1-5, possibly reducing model complexity, until testing error is small.

Is that all of ML?
No, but these days it’s much of it!

Linear regression — basic setup

1. Start from training data $((\boldsymbol{x}_i, y_i))_{i=1}^n$, with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.
2. Model is a linear predictor: pick $\boldsymbol{w} \in \mathbb{R}^d$ with

$$\boldsymbol{x}_i \mapsto \boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i =: \hat{y}_i \approx y_i.$$

3. Loss function ℓ is squared loss ℓ_{sq} (standard regression loss):

$$\ell_{\mathrm{sq}}(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i, y_i) = \frac{1}{2}(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i - y_i)^2.$$

   We will minimize the empirical risk (average loss over training examples):

$$\widehat{\mathcal{R}}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\ell_{\mathrm{sq}}(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i, y_i) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 \qquad \text{where } \boldsymbol{X} := \begin{bmatrix} \leftarrow & \boldsymbol{x}_1^\mathsf{T} & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{x}_n^\mathsf{T} & \rightarrow \end{bmatrix}.$$

4. Basic method: gradient descent. Set $\boldsymbol{w}_0 = 0$, and thereafter

$$\boldsymbol{w}_{i+1} := \boldsymbol{w}_i - \eta\nabla\widehat{\mathcal{R}}(\boldsymbol{w}_i) = \boldsymbol{w}_i - \frac{\eta}{n}\boldsymbol{X}^\mathsf{T}\left(\boldsymbol{X}\boldsymbol{w}_i - \boldsymbol{y}\right),$$

   where η is a learning rate (step size).

2. Model is a linear predictor: pick $\boldsymbol{w} \in \mathbb{R}^d$ with

$$\boldsymbol{x}_i \mapsto \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_i \approx y_i.$$

- Our model/architecture/function class is $\{\boldsymbol{x} \mapsto \boldsymbol{w}^\mathsf{T}\boldsymbol{x} : \boldsymbol{w} \in \mathbb{R}^d\}$.
  For each $\boldsymbol{w} \in \mathbb{R}^d$, we have another predictor.
- This is a simple model; we'll build off of it to get more powerful ones!
- This model is insufficient for complicated tasks, but often does well, and forms a good baseline.

3. Loss function is squared loss (standard regression loss):

$$\ell(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i, y_i) = \frac{1}{2}(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i - y_i)^2.$$

We will minimize the empirical risk:

$$\widehat{\mathcal{R}}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i, y_i) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2.$$

- Regression towards the mean: if $\boldsymbol{x}_i = 1 \in \mathbb{R}^1$ for all i, then

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}} \|\boldsymbol{X}w - \boldsymbol{y}\|^2 = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

  Seems a reasonable notion of loss/error.
- There are many choices for ℓ. Next lecture we'll use logistic loss ℓ_{logistic}

$$\ell_{\text{logistic}}(\hat{y}, y) = \ln(1 + \exp(-\hat{y}y)).$$

  This and squared loss are the most common.

4. Basic method: gradient descent. Set $\boldsymbol{w}_0 = 0$, and thereafter

$$\boldsymbol{w}_{i+1} := \boldsymbol{w}_i - \eta \nabla \widehat{\mathcal{R}}(\boldsymbol{w}_i) = \boldsymbol{w}_i - \frac{\eta}{n} \boldsymbol{X}^{\mathsf{T}} \left(\boldsymbol{X} \boldsymbol{w}_i - \boldsymbol{y} \right),$$

where η is a learning rate (step size).

- In a few lectures, we'll see that this globally minimizes $\widehat{\mathcal{R}}$.
- We'll spend most of this lecture on other solutions via SVD.

Normal equations and SVD.

We want to find $\hat{\boldsymbol{w}}$ so that

$$2n\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \|\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y}\|^2 = \min_{\boldsymbol{w}\in\mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 .$$

Normal equations and SVD.

We want to find $\hat{\boldsymbol{w}}$ so that

$$2n\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \left\|\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y}\right\|^2 = \min_{\boldsymbol{w}\in\mathbb{R}^d} \left\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right\|^2 .$$

Idea from calculus: set gradient to zero and solve:

$$0 = \nabla_{\boldsymbol{w}} \left\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right\|^2 = 2\boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right),$$

meaning we want $\hat{\boldsymbol{w}}$ so that

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{y}.$$

These are called the normal equations.

The normal equations are the system of linear equalities

$$\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y}.$$

Proposition. $\hat{\boldsymbol{w}}$ satisfies $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$ iff $\hat{\boldsymbol{w}}$ satisfies the normal equations.

The normal equations are the system of linear equalities

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y}.$$

Proposition. $\hat{\boldsymbol{w}}$ satisfies $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$ iff $\hat{\boldsymbol{w}}$ satisfies the normal equations.

Proof (one direction). Consider $\boldsymbol{w}$ with $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y}$, and any $\boldsymbol{w}'$; then

$$\begin{aligned}\|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{y}\|^2 &= \|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w} + \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 \\ &= \|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w}\|^2 + 2(\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w})^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2.\end{aligned}$$

Since

$$(\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w})^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = (\boldsymbol{w}' - \boldsymbol{w})^\top(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^\top\boldsymbol{y}) = 0,$$

then

$$2n\widehat{\mathcal{R}}(\boldsymbol{w}') = \|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{y}\|^2 = \|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w}\|^2 + \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 \geq \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 = 2n\widehat{\mathcal{R}}(\boldsymbol{w}).$$

$\square$

Later we'll get a general version by convexity, but it's nice that we can check this directly so easily!

The normal equations are the system of linear equalities

$$\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^{\top}\boldsymbol{y}.$$

Proposition. $\hat{\boldsymbol{w}}$ satisfies $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$ iff $\hat{\boldsymbol{w}}$ satisfies the normal equations.

The normal equations are the system of linear equalities

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y}.$$

Proposition. $\hat{\boldsymbol{w}}$ satisfies $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$ iff $\hat{\boldsymbol{w}}$ satisfies the normal equations.

How do we solve for $\hat{\boldsymbol{w}}$?

- If $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, we can use $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$.
- In general, we will use the SVD.

The SVD (Singular Value Decomposition).

Let $\boldsymbol{M} \in \mathbb{R}^{n \times d}$ be given. $((s_i, \boldsymbol{u}_i, \boldsymbol{v}_i))_{i=1}^{r}$ is an SVD of $\boldsymbol{M}$ if:

- $\boldsymbol{M}$ has rank r;
- $s_1 \geq s_2 \cdots \geq s_r > 0$;
- $(\boldsymbol{u}_i)_{i=1}^{r}$ are orthonormal (orthogonal and unit length), and span the column space of $\boldsymbol{M}$;
- $(\boldsymbol{v}_i)_{i=1}^{r}$ are orthonormal, and span the row space of $\boldsymbol{M}$.
- $\boldsymbol{M} = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}$.

The SVD (Singular Value Decomposition).

Let $\boldsymbol{M} \in \mathbb{R}^{n \times d}$ be given. $((s_i, \boldsymbol{u}_i, \boldsymbol{v}_i))_{i=1}^r$ is an SVD of $\boldsymbol{M}$ if:

- $\boldsymbol{M}$ has rank r;
- $s_1 \geq s_2 \cdots \geq s_r > 0$;
- $(\boldsymbol{u}_i)_{i=1}^r$ are orthonormal (orthogonal and unit length), and span the column space of $\boldsymbol{M}$;
- $(\boldsymbol{v}_i)_{i=1}^r$ are orthonormal, and span the row space of $\boldsymbol{M}$.
- $\boldsymbol{M} = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$.

- The SVD always exists, and is real-valued.
  (When do real eigendecompositions not exist?)
- The ordered tuple $(s_1, \ldots, s_r)$ is unique,
  but the SVD is in general not unique (why not?).
- For $k < r$, the low rank approximation $\sum_{i=1}^k s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \approx \boldsymbol{M}$ has many applications (wait for the PCA lecture).

Pseudoinverse.

Given SVD $\boldsymbol{M} = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^\mathsf{T}$, the pseudoinverse is

$$\boldsymbol{M}^+ := \sum_{i=1}^{r} \frac{1}{s_i} \boldsymbol{v}_i \boldsymbol{u}_i^\mathsf{T}.$$

Pseudoinverse.

Given SVD $\boldsymbol{M} = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}$, the pseudoinverse is

$$\boldsymbol{M}^+ := \sum_{i=1}^{r} \frac{1}{s_i} \boldsymbol{v}_i \boldsymbol{u}_i^{\mathsf{T}}.$$

- The SVD may fail to be unique, but $\boldsymbol{M}^+$ is unique.
- $\boldsymbol{M}\boldsymbol{M}^+ = \sum_{i=1}^r \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}}$ and $\boldsymbol{M}^+\boldsymbol{M} = \sum_{i=1}^r \boldsymbol{v}_i \boldsymbol{v}_i^{\mathsf{T}}$; in general, neither is an identity matrix. (Consider the case $\boldsymbol{M} = \boldsymbol{e}_1 \boldsymbol{e}_1^{\mathsf{T}}$.)
- On the other hand,

$$\boldsymbol{M}\boldsymbol{M}^+\boldsymbol{M} =$$

$$\boldsymbol{M}^+\boldsymbol{M}\boldsymbol{M}^+ =$$

- If $\boldsymbol{M}^{-1}$ exists, then $\boldsymbol{M}^+ = \boldsymbol{M}^{-1}$.
- If $\boldsymbol{M} = 0$, then $r = 0$ and $\boldsymbol{M}^+ = 0$.

Pseudoinverse.

Given SVD $\boldsymbol{M} = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, the pseudoinverse is

$$\boldsymbol{M}^+ := \sum_{i=1}^{r} \frac{1}{s_i} \boldsymbol{v}_i \boldsymbol{u}_i^\top.$$

- The SVD may fail to be unique, but $\boldsymbol{M}^+$ is unique.
- $\boldsymbol{M}\boldsymbol{M}^+ = \sum_{i=1}^r \boldsymbol{u}_i \boldsymbol{u}_i^\top$ and $\boldsymbol{M}^+\boldsymbol{M} = \sum_{i=1}^r \boldsymbol{v}_i \boldsymbol{v}_i^\top$; in general, neither is an identity matrix. (Consider the case $\boldsymbol{M} = \boldsymbol{e}_1 \boldsymbol{e}_1^\top$.)
- On the other hand,

$$\boldsymbol{M}\boldsymbol{M}^+\boldsymbol{M} = \left(\sum_{i=1}^r s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top\right)\left(\sum_{j=1}^r \frac{1}{s_j} \boldsymbol{v}_j \boldsymbol{u}_j^\top\right)\left(\sum_{k=1}^r s_k \boldsymbol{u}_k \boldsymbol{v}_k^\top\right) = \boldsymbol{M},$$

$$\boldsymbol{M}^+\boldsymbol{M}\boldsymbol{M}^+ = \left(\sum_{i=1}^r \frac{1}{s_i} \boldsymbol{v}_i \boldsymbol{u}_i^\top\right)\left(\sum_{j=1}^r s_j \boldsymbol{u}_j \boldsymbol{v}_j^\top\right)\left(\sum_{k=1}^r \frac{1}{s_k} \boldsymbol{v}_k \boldsymbol{u}_k^\top\right) = \boldsymbol{M}^+.$$

- If $\boldsymbol{M}^{-1}$ exists, then $\boldsymbol{M}^+ = \boldsymbol{M}^{-1}$.
- If $\boldsymbol{M} = 0$, then $r = 0$ and $\boldsymbol{M}^+ = 0$.

OLS (Ordinary Least Squares) solution via SVD.

Given a least squares problem $\widehat{\mathcal{R}}(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2/(2n)$, the OLS solution

$$\hat{\boldsymbol{w}}_{\mathsf{ols}} = \boldsymbol{X}^{+}\boldsymbol{y}$$

satisfies the normal equations (whereby $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}_{\mathsf{ols}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$).

OLS (Ordinary Least Squares) solution via SVD.

Given a least squares problem $\widehat{\mathcal{R}}(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2/(2n)$, the OLS solution

$$\hat{\boldsymbol{w}}_{\mathsf{ols}} = \boldsymbol{X}^+\boldsymbol{y}$$

satisfies the normal equations (whereby $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}_{\mathsf{ols}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$).

Easy to check: writing $\boldsymbol{X} = \sum_{i=1}^r s_i \boldsymbol{u}_i \boldsymbol{v}_i^\mathsf{T}$,

$$\begin{aligned}
\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{ols}} &= \boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{y} \\
&= \left(\sum_{i=1}^r s_i \boldsymbol{v}_i \boldsymbol{u}_i^\mathsf{T}\right)\left(\sum_{j=1}^r s_j \boldsymbol{u}_j \boldsymbol{v}_j^\mathsf{T}\right)\left(\sum_{k=1}^r \frac{1}{s_k} \boldsymbol{v}_k \boldsymbol{u}_k^\mathsf{T}\right)\boldsymbol{y} \\
&= \boldsymbol{X}^\mathsf{T}\boldsymbol{y}.
\end{aligned}$$

SVD $\boldsymbol{M} = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ and orthonormal bases.

We can extend $(\boldsymbol{u}_i)_{i=1}^r$ and $(\boldsymbol{v}_i)_{i=1}^r$ to full orthonormal bases for $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively: write $\boldsymbol{M} \in \mathbb{R}^{n\times d}$ as

$$\left[\begin{array}{ccc|ccc} \uparrow & & \uparrow & \uparrow & & \uparrow \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_r & \boldsymbol{u}_{r+1} & \cdots & \boldsymbol{u}_n \\ \downarrow & & \downarrow & \downarrow & & \downarrow \end{array}\right] \cdot \left[\begin{array}{ccc|c} s_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & s_r & \\ \hline & 0 & & 0 \end{array}\right] \cdot \left[\begin{array}{ccc|ccc} \uparrow & & \uparrow & \uparrow & & \uparrow \\ \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_r & \boldsymbol{v}_{r+1} & \cdots & \boldsymbol{v}_d \\ \downarrow & & \downarrow & \downarrow & & \downarrow \end{array}\right]^\top .$$

The old parts span the column and row spaces of $\boldsymbol{M}$;
the new vectors span the left and right nullspaces.
Some call this a “full” SVD.

SVD and relationship to eigenvalues.

SVD and relationship to eigenvalues.

Note

$$\boldsymbol{M}\boldsymbol{M}^\top = \sum_{i=1}^{r} s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \sum_{j=1}^{r} s_j \boldsymbol{v}_j \boldsymbol{u}_j^\top = \sum_{i=1}^{r} s_i^2 \boldsymbol{u}_i \boldsymbol{u}_i^\top,$$

thus left singular vectors $(\boldsymbol{u})_{i=1}^{r}$ are top eigenvectors of $\boldsymbol{M}\boldsymbol{M}^\top$, with eigenvalues $s_1^2 \geq \cdots \geq s_r^2$.
Similarly,

$$\boldsymbol{M}^\top\boldsymbol{M} = \sum_{i=1}^{r} s_i \boldsymbol{v}_i \boldsymbol{u}_i^\top \sum_{j=1}^{r} s_j \boldsymbol{u}_j \boldsymbol{v}_j^\top = \sum_{i=1}^{r} s_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^\top,$$

obtaining right singular vectors from $\boldsymbol{M}^\top\boldsymbol{M}$.

Summary on least squares solutions

We want to approximately solve the empirical risk minimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \widehat{\mathcal{R}}(\boldsymbol{w}) = \min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|^2 .$$

Three approaches:

1. Gradient descent: $\boldsymbol{w}_0 := 0$, thereafter $\boldsymbol{w}_{i+1} := \boldsymbol{w} - \eta\nabla\widehat{\mathcal{R}}(\boldsymbol{w}_i)$.
2. Pick any $\hat{\boldsymbol{w}}$ satsifying the normal equations

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{y}.$$

3. Use the ordinary least squares (OLS) solution $\hat{\boldsymbol{w}}_{\mathsf{ols}} = \boldsymbol{X}^+\boldsymbol{y}$.

Summary on least squares solutions

We want to approximately solve the empirical risk minimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \widehat{\mathcal{R}}(\boldsymbol{w}) = \min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|^2 .$$

Three approaches:

1. Gradient descent: $\boldsymbol{w}_0 := 0$, thereafter $\boldsymbol{w}_{i+1} := \boldsymbol{w} - \eta\nabla\widehat{\mathcal{R}}(\boldsymbol{w}_i)$.
2. Pick any $\hat{\boldsymbol{w}}$ satsifying the normal equations

$$\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}.$$

3. Use the ordinary least squares (OLS) solution $\hat{\boldsymbol{w}}_{\mathsf{ols}} = \boldsymbol{X}^{+}\boldsymbol{y}$.

(Side note: are these different?. . .)

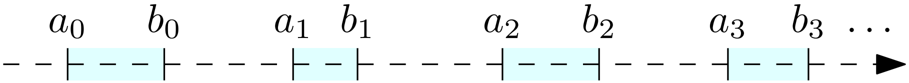
Example: Old Faithful geyser (Yellowstone)

Example: Old Faithful geyser (Yellowstone)

Task: Predict time of next eruption.

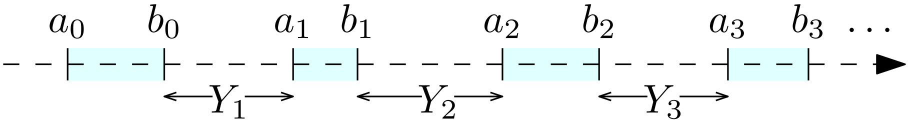
Time between eruptions

Source data: start and end times (a_i, b_i) of $n = 136$ eruptions.



Time between eruptions

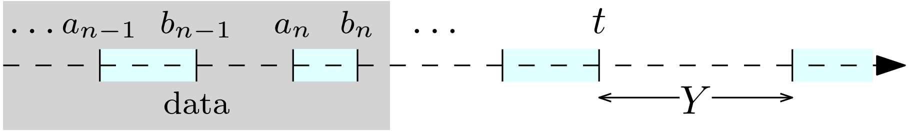
Source data: start and end times (a_i, b_i) of $n = 136$ eruptions.



Let's pre-process: form time between eruptions $y_i := a_{i+1} - b_i$.

Time between eruptions

Source data: start and end times (a_i, b_i) of $n = 136$ eruptions.

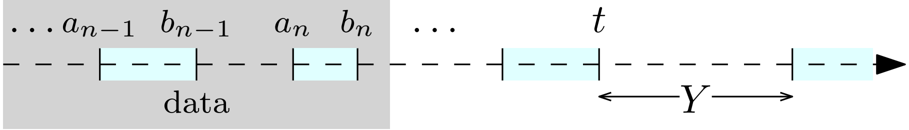


Let's pre-process: form time between eruptions $y_i := a_{i+1} - b_i$.

Reformulated task:
to estimate next eruption, find last end time t, compute $\hat{y}$, and output $t + \hat{y}$.

Time between eruptions

Source data: start and end times (a_i, b_i) of $n = 136$ eruptions.



Let's pre-process: form time between eruptions $y_i := a_{i+1} - b_i$.

Reformulated task:
to estimate next eruption, find last end time t, compute $\hat{y}$, and output $t + \hat{y}$.

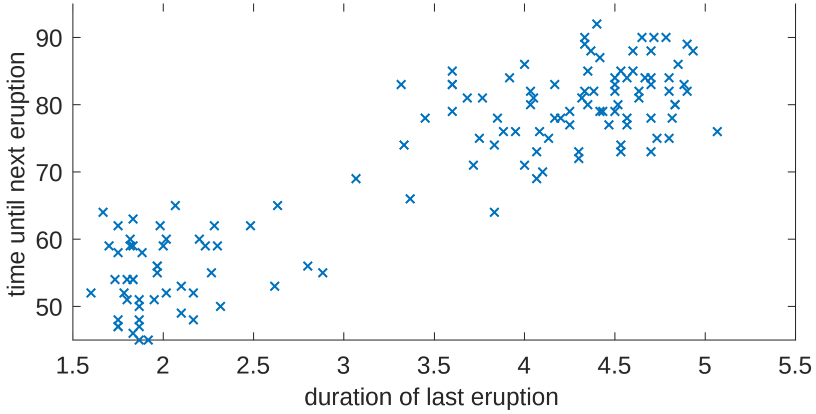
Let's use linear regression.

- Set $x_i = 1$ and the OLS solution is the mean:

$$\hat{y} = \frac{1}{136} \sum_{i=1}^{136} y_i = 70.7941.$$

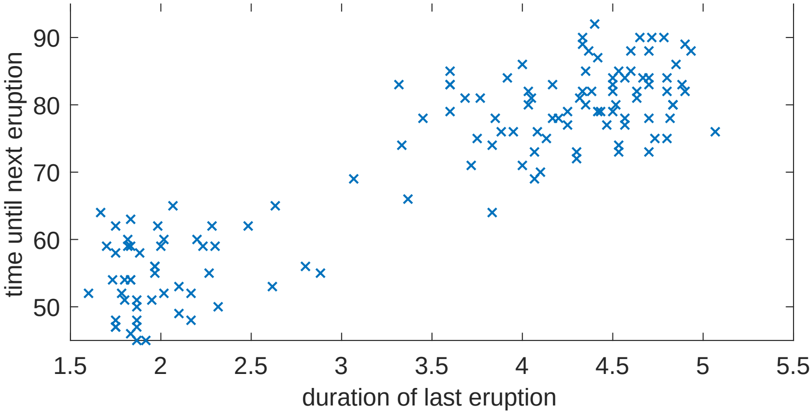
- Can we do better with another x_i?

Eruption length and time to eruption are correlated.



Let choose $\boldsymbol{x}_i := \left[\begin{smallmatrix} b_i - a_i \\ 1 \end{smallmatrix}\right]$.

Eruption length and time to eruption are correlated.



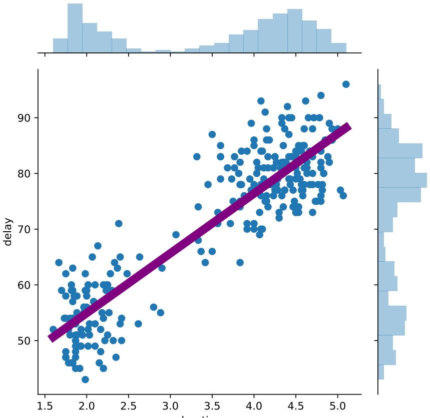
Let choose $\boldsymbol{x}_i := \begin{bmatrix} b_i - a_i \\ 1 \end{bmatrix}$.

(Side note: the extra "1" will be discussed extensively later.)

1. Form pairs $\boldsymbol{x}_i := \left[\begin{smallmatrix} b_i - a_i \\ 1 \end{smallmatrix}\right]$, and matrix

$$\boldsymbol{X} := \begin{bmatrix} \leftarrow & \boldsymbol{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{x}_n^\top & \rightarrow \end{bmatrix} = \begin{bmatrix} b_1 - a_1 & 1 \\ \vdots & \\ b_n - a_n & 1 \end{bmatrix} \in \mathbb{R}^{n \times 2}.$$

   Form labels $\boldsymbol{y} \in \mathbb{R}^n$, $y_i := a_{i+1} - b_i$.
2. Choose OLS solution $\hat{\boldsymbol{w}}_{\text{ols}} := \boldsymbol{X}^+ \boldsymbol{y}$.
3. Given a new eruption (a, b), estimate next eruption time $b + \boldsymbol{w}^\top \left[\begin{smallmatrix} b-a \\ 1 \end{smallmatrix}\right]$.



“pytorch meta-algorithm” on Old Faithful data

1. Clean/augment data.
   From (a_i, b_i), form $\boldsymbol{x}'_i = (1,)$ or $\boldsymbol{x}_i = (b_i - a_i, 1)$, and $y_i = a_i - b_{i-1}$.
2. Pick model/architecture (anything from lectures 2-13).
   Linear predictor.
3. Pick a loss function measuring model fit to data.
   Squared loss.
4. Run a gradient descent variant to fit model to data.
5. Tweak 1-4 until training error is small.
   $\boldsymbol{x}'_i$ was bad, so we added a feature and got $\boldsymbol{x}_i$.
6. Tweak 1-5, possibly reducing model complexity, until testing error is small.
   We didn’t try this!

Summary for today

- Linear regression setup revisited.
- Normal equations, SVD, and pseudoinverse.
- Example (if time).

(Appendix.)

The normal equations are the system of linear equalities

$$\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^{\top}\boldsymbol{y}.$$

Proposition. $\hat{\boldsymbol{w}}$ satisfies $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$ iff $\hat{\boldsymbol{w}}$ satisfies the normal equations.

The normal equations are the system of linear equalities

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y}.$$

Proposition. $\hat{\boldsymbol{w}}$ satisfies $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$ iff $\hat{\boldsymbol{w}}$ satisfies the normal equations.

Proof (other direction).
Suppose $\boldsymbol{w}$ is optimal; since $\hat{\boldsymbol{w}}_{\mathsf{ols}}$ satisfies the normal equations, then expanding the square as in the proof of the other direction gives

$$\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{ols}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{ols}} - \boldsymbol{y}\|^2.$$

Since $\boldsymbol{w}$ and $\hat{\boldsymbol{w}}_{\mathsf{ols}}$ are optimal, then $\widehat{\mathcal{R}}(\boldsymbol{w}) = \widehat{\mathcal{R}}(\hat{\boldsymbol{w}}_{\mathsf{ols}})$, so the preceding implies

The normal equations are the system of linear equalities

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y}.$$

Proposition. $\hat{\boldsymbol{w}}$ satisfies $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$ iff $\hat{\boldsymbol{w}}$ satisfies the normal equations.

Proof (other direction).
Suppose $\boldsymbol{w}$ is optimal; since $\hat{\boldsymbol{w}}_{\mathsf{ols}}$ satisfies the normal equations, then expanding the square as in the proof of the other direction gives

$$\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{ols}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{ols}} - \boldsymbol{y}\|^2.$$

Since $\boldsymbol{w}$ and $\hat{\boldsymbol{w}}_{\mathsf{ols}}$ are optimal, then $\widehat{\mathcal{R}}(\boldsymbol{w}) = \widehat{\mathcal{R}}(\hat{\boldsymbol{w}}_{\mathsf{ols}})$, so the preceding implies

$$0 = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{ols}}\|^2 = \|\boldsymbol{X}(\boldsymbol{w} - \hat{\boldsymbol{w}}_{\mathsf{ols}})\|^2,$$

therefore $\boldsymbol{X}(\boldsymbol{w} - \hat{\boldsymbol{w}}_{\mathsf{ols}}) = 0$ and $\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{ols}}$, which by the normal equations for $\hat{\boldsymbol{w}}_{\mathsf{ols}}$ means

$$\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{ols}} = \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w},$$

thus $\boldsymbol{w}$ satisfies the normal equations. $\square$

Geometric interpretation of least squares ERM

Let $\boldsymbol{a}_j \in \mathbb{R}^n$ be the j-th column (not row!) of matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, so

$$\boldsymbol{X} = \begin{bmatrix} \leftarrow & \boldsymbol{x}_1^{\mathsf{T}} & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{x}_n^{\mathsf{T}} & \rightarrow \end{bmatrix} = \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_d \\ \downarrow & & \downarrow \end{bmatrix}.$$

Geometric interpretation of least squares ERM

Let $\boldsymbol{a}_j \in \mathbb{R}^n$ be the j-th column (not row!) of matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, so

$$\boldsymbol{X} = \begin{bmatrix} \leftarrow & \boldsymbol{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{x}_n^\top & \rightarrow \end{bmatrix} = \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_d \\ \downarrow & & \downarrow \end{bmatrix}.$$

Minimizing $\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2$ means finding $\hat{\boldsymbol{b}} \in \operatorname{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d)$ closest to $\boldsymbol{b}$.

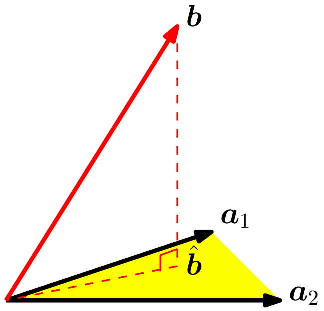
Geometric interpretation of least squares ERM

Let $\boldsymbol{a}_j \in \mathbb{R}^n$ be the j-th column (not row!) of matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, so

$$\boldsymbol{X} = \begin{bmatrix} \leftarrow & \boldsymbol{x}_1^\mathsf{T} & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{x}_n^\mathsf{T} & \rightarrow \end{bmatrix} = \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_d \\ \downarrow & & \downarrow \end{bmatrix}.$$

Minimizing $\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2$ means finding $\hat{\boldsymbol{b}} \in \operatorname{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d)$ closest to $\boldsymbol{b}$.

Solution $\hat{\boldsymbol{b}}$ is orthogonal projection of $\boldsymbol{b}$ onto $\operatorname{range}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^d\}$.



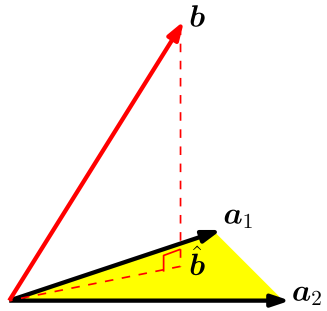
Geometric interpretation of least squares ERM

Let $\boldsymbol{a}_j \in \mathbb{R}^n$ be the j-th column (not row!) of matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, so

$$\boldsymbol{X} = \begin{bmatrix} \leftarrow & \boldsymbol{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{x}_n^\top & \rightarrow \end{bmatrix} = \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_d \\ \downarrow & & \downarrow \end{bmatrix}.$$

Minimizing $\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2$ means finding $\hat{\boldsymbol{b}} \in \operatorname{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d)$ closest to $\boldsymbol{b}$.

Solution $\hat{\boldsymbol{b}}$ is orthogonal projection of $\boldsymbol{b}$ onto $\operatorname{range}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^d\}$.



- $\hat{\boldsymbol{b}}$ is uniquely determined; indeed, $\hat{\boldsymbol{b}} = \boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{b} = \sum_{i=1}^{r} \boldsymbol{u}_i \boldsymbol{u}_i^\top \boldsymbol{b}$.

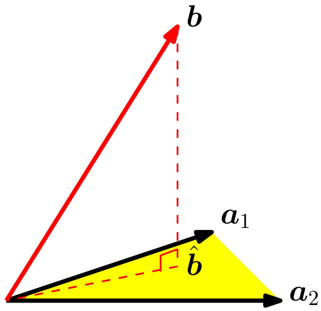
Geometric interpretation of least squares ERM

Let $\boldsymbol{a}_j \in \mathbb{R}^n$ be the j-th column (not row!) of matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, so

$$\boldsymbol{X} = \begin{bmatrix} \leftarrow & \boldsymbol{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{x}_n^\top & \rightarrow \end{bmatrix} = \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_d \\ \downarrow & & \downarrow \end{bmatrix}.$$

Minimizing $\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2$ means finding $\hat{\boldsymbol{b}} \in \operatorname{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d)$ closest to $\boldsymbol{b}$.

Solution $\hat{\boldsymbol{b}}$ is orthogonal projection of $\boldsymbol{b}$ onto $\operatorname{range}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^d\}$.



- $\hat{\boldsymbol{b}}$ is uniquely determined; indeed, $\hat{\boldsymbol{b}} = \boldsymbol{A}\boldsymbol{A}^+\boldsymbol{b} = \sum_{i=1}^r \boldsymbol{u}_i\boldsymbol{u}_i^\top \boldsymbol{b}$.
- If $r = \operatorname{rank}(\boldsymbol{A}) < d$, then >1 way to write $\hat{\boldsymbol{b}}$ as linear combination of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$.

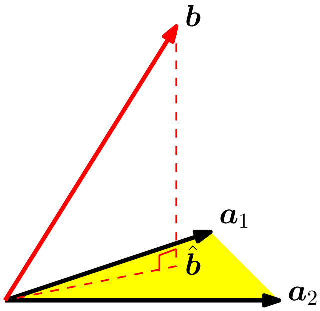
Geometric interpretation of least squares ERM

Let $\boldsymbol{a}_j \in \mathbb{R}^n$ be the j-th column (not row!) of matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, so

$$\boldsymbol{X} = \begin{bmatrix} \leftarrow & \boldsymbol{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{x}_n^\top & \rightarrow \end{bmatrix} = \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_d \\ \downarrow & & \downarrow \end{bmatrix}.$$

Minimizing $\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2$ means finding $\hat{\boldsymbol{b}} \in \operatorname{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d)$ closest to $\boldsymbol{b}$.

Solution $\hat{\boldsymbol{b}}$ is orthogonal projection of $\boldsymbol{b}$ onto $\operatorname{range}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^d\}$.



- $\hat{\boldsymbol{b}}$ is uniquely determined; indeed, $\hat{\boldsymbol{b}} = \boldsymbol{A}\boldsymbol{A}^+\boldsymbol{b} = \sum_{i=1}^r \boldsymbol{u}_i\boldsymbol{u}_i^\top \boldsymbol{b}$.
- If $r = \operatorname{rank}(\boldsymbol{A}) < d$, then >1 way to write $\hat{\boldsymbol{b}}$ as linear combination of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$.

If $\operatorname{rank}(\boldsymbol{A}) < d$, then **ERM solution is not unique**.

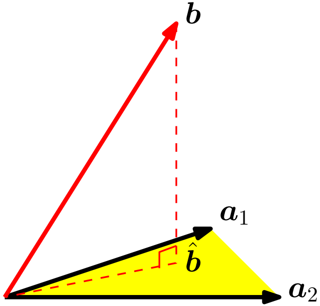
Geometric interpretation of least squares ERM

Let $\boldsymbol{a}_j \in \mathbb{R}^n$ be the j-th column (not row!) of matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, so

$$\boldsymbol{X} = \begin{bmatrix} \leftarrow & \boldsymbol{x}_1^\mathsf{T} & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{x}_n^\mathsf{T} & \rightarrow \end{bmatrix} = \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_d \\ \downarrow & & \downarrow \end{bmatrix}.$$

Minimizing $\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2$ means finding $\hat{\boldsymbol{b}} \in \operatorname{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d)$ closest to $\boldsymbol{b}$.

Solution $\hat{\boldsymbol{b}}$ is orthogonal projection of $\boldsymbol{b}$ onto $\operatorname{range}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^d\}$.



- $\hat{\boldsymbol{b}}$ is uniquely determined; indeed, $\hat{\boldsymbol{b}} = \boldsymbol{A}\boldsymbol{A}^+\boldsymbol{b} = \sum_{i=1}^r \boldsymbol{u}_i\boldsymbol{u}_i^\mathsf{T}\boldsymbol{b}$.
- If $r = \operatorname{rank}(\boldsymbol{A}) < d$, then >1 way to write $\hat{\boldsymbol{b}}$ as linear combination of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$.

If $\operatorname{rank}(\boldsymbol{A}) < d$, then **ERM solution is not unique**.

To get $\boldsymbol{w}$ from $\hat{\boldsymbol{b}}$:
solve system of linear equations $\boldsymbol{A}\boldsymbol{w} = \hat{\boldsymbol{b}}$.

Computing the SVD

Typical solver is an iterative, greedy method.
For more information, see the excellent data science book by Blum, Hopcroft, Kannan.

Why GD?

Why include GD, since pseudoinverse seems sufficient?

- GD is easy to implement, pseudoinverse more painful.
- Pseudoinverse after all implemented as an iterative solver.
- GD generalizes to other cases of squared loss (e.g., deep network training with squared loss).

Why $^1/_2$ in squared loss?

Note: this material is beyond the scope of this course.

1. (Sanity check: doesn't affect methods.) Note that multiplying the risk by a constant just scales the whole surface, and does not change the sign of $\widehat{\mathcal{R}}(\boldsymbol{v}) - \widehat{\mathcal{R}}(\boldsymbol{w})$ for any two choices $(\boldsymbol{v}, \boldsymbol{w})$, and therefore to some extent it doesn't matter (and can be treated as just rescaling the step size in gradient descent).
2. (Superficial reasons: many calculations seem nicer.) The usual answer is certain calculations are nicer: the gradient for linear regression becomes $\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$, the Hessian becomes $\boldsymbol{X}^\top\boldsymbol{X}$, and if $\|\boldsymbol{x}_i\| \leq 1$ for all examples then this has 1 as its largest eigenvalue and we can use a step size of 1 with gradient descent.
3. (Deeper, but related reason.) The squared loss is convex, and in fact the the related function $g(\boldsymbol{v}) = \|\boldsymbol{v}\|_2^2/2$ (in $\widehat{\mathcal{R}}$) satisfies a very interesting property. In convex analysis, a key concept for any convex function f is its <u>conjugate</u> f^*; interestingly, g is the <u>unique</u> convex function satisfying $g = g^*$. To some extent this is equivalent to the previous point, since conjugacy has connections to gradients, and moreover is also a bit arbitrary, since conjugacy relies upon a choice of primal and dual spaces and we could have obtained a different unique pair.

Overall, I think it is nice to use the $^1/_2$, and less controversial than the choice of placing 2π in the definition of the Fourier transform, which has become sort of standardized but does have various tradeoffs.

Supplemental reading

- Shalev-Shwartz/Ben-David: chapters 9, 24.
- SVD: see the related chapter in the "Foundations of Data Science" book by Blum, Hopcroft, Kannan; they use the same convention $\boldsymbol{M} = \sum_{i=1}^{r} s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ as we use here.