

# Linear Models

## Numerical Methods for Deep Learning

# Learning Problem

Given examples (inputs)

$$\mathbf{Y} = (\mathbf{y}_1 \quad \mathbf{y}_2 \quad \cdots \quad \mathbf{y}_n) \in \mathbb{R}^{n_f \times n}$$

and labels (outputs)

$$\mathbf{C} = (\mathbf{c}_1 \quad \mathbf{c}_2 \quad \cdots \quad \mathbf{c}_n) \in \mathbb{R}^{n_c \times n},$$

find a classification/prediction function  $f(\cdot, \boldsymbol{\theta})$ , i.e.,

$$f(\mathbf{y}_j, \boldsymbol{\theta}) \approx \mathbf{c}_j, \quad j = 1, \dots, n.$$

# Regression and Least-Squares

Simplest option, a linear model with  $\theta = (\mathbf{W}, \mathbf{b})$  and

$$f(\mathbf{Y}, \mathbf{W}, \mathbf{b}) = \mathbf{W}\mathbf{Y} + \mathbf{b}\mathbf{e}_n^\top \approx \mathbf{C}$$

- ▶  $\mathbf{W} \in \mathbb{R}^{n_c \times n_f}$  are *weights*
- ▶  $\mathbf{b} \in \mathbb{R}^{n_c}$  are *biases*
- ▶  $\mathbf{e}_n \in \mathbb{R}^n$  is a vector of ones

Equivalent notation:

$$f(\mathbf{Y}, \mathbf{W}, \mathbf{b}) = (\mathbf{W} \quad \mathbf{b}) \begin{pmatrix} \mathbf{Y} \\ \mathbf{e}_n^\top \end{pmatrix} \approx \mathbf{C}$$

Problem may not have a solution, or may have infinite solutions (when?). Solve through optimization

$$\min_{\mathbf{W}} \frac{1}{2} \|\mathbf{W}\mathbf{Y} - \mathbf{C}\|_F^2$$

(Frobenius norm:  $\|\mathbf{A}\|_F^2 = \text{trace}(\mathbf{A}^\top \mathbf{A}) = \sum_{i,j} \mathbf{A}_{i,j}^2$ .)

## Remark: Relation to Least-Squares

Consider the regression problem

$$\min_{\mathbf{W}} \frac{1}{2} \|\mathbf{W}\mathbf{Y} - \mathbf{C}\|_F^2.$$

It is easy to see that this is equivalent to

$$\min_{\mathbf{W}} \frac{1}{2} \|\mathbf{Y}^\top \mathbf{W}^\top - \mathbf{C}^\top\|_F^2,$$

which can be solved separately for each row in  $\mathbf{W}$

$$\mathbf{W}(j, :)^{\top} = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{Y}^{\top} \mathbf{w} - \mathbf{C}(j, :)^{\top}\|_F^2.$$

Notation: Let  $\mathbf{A} = \mathbf{Y}^\top$  and  $\mathbf{X} = \mathbf{W}^\top$  (easy to add bias here), we solve

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{A}\mathbf{X} - \mathbf{C}^\top\|_F^2$$

# Optimality Conditions for Least-Squares

To minimize a function need to differentiate and equate to 0

$$\frac{\partial \left( \frac{1}{2} \|\mathbf{AX} - \mathbf{C}^\top\|_F^2 \right)}{\partial \mathbf{X}} = 0$$

Compute the derivatives in three steps

1.

$$\frac{\partial \left( \frac{1}{2} \|\mathbf{R}\|_F^2 \right)}{\partial \mathbf{R}} = ???$$

2.

$$\frac{\partial (\mathbf{AX})}{\partial \mathbf{X}} = ???$$

3. Use chain rule

# Least-Squares: Normal Equations

The necessary and sufficient optimality conditions for the least-squares problem are

$$\frac{\partial \left( \frac{1}{2} \|\mathbf{A}\mathbf{X} - \mathbf{C}^\top\|_F^2 \right)}{\partial \mathbf{X}} = \mathbf{A}^\top (\mathbf{A}\mathbf{X} - \mathbf{C}^\top) = 0$$

Reorganize to obtain the **normal equations**

$$\mathbf{X} = (\mathbf{A}^\top \mathbf{A})^{-1} (\mathbf{A}^\top \mathbf{C}^\top).$$

Here,  $\mathbf{A}^\top \mathbf{A} \in \mathbb{R}^{n_f \times n_f}$  must be invertible, i.e.,

- ▶ sufficient amount of data ( $n > n_f$ )
- ▶ data is linearly independent

# Coding: Least-Squares Regression

1. Write a code for solving

$$\min_{\mathbf{W}, \mathbf{b}} \frac{1}{2} \|\mathbf{W}\mathbf{Y} + \mathbf{b}\mathbf{e}_n^\top - \mathbf{C}\|^2$$

and apply it to some of our test data (MNIST / CIFAR10)

2. Solve the problem using the normal equations derived above.
3. Use optimal weights to predict labels for test data. How well does your solution generalize?

## Ill-posedness and the SVD

If the data is linearly dependent or close to be linearly dependent, least-squares problem gives no good solution [2, 4, 3].

Understanding can be gained by the *Singular Value Decomposition* (SVD) (e.g., [1, Ch. 8])

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

where  $\mathbf{U} \in \mathbb{R}^{n_f \times n_f}$ ,  $\mathbf{V} \in \mathbb{R}^{n_f \times n}$  satisfy

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I}, \quad \text{and} \quad \mathbf{V}^\top \mathbf{V} = \mathbf{I}$$

Diagonal of  $\mathbf{\Sigma}$  contains the singular values  $\sigma_1 \geq \dots \sigma_{n_f} \geq 0$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_{n_f} \end{pmatrix}$$



## Ill-posedness and Regularization

Important is the *effective rank*: If  $\sigma_j \ll \sigma_1$  for all  $j \geq k$ , then the effective rank of the problem is  $k$ .

If  $k < n_f$ , the least squares problem is ill-posed, i.e., solution does not exist or is unstable.

Small perturbations in  $\mathbf{C}$  or  $\mathbf{A} = \mathbf{Y}^\top$  yield large perturbations in  $\mathbf{X} = \mathbf{W}^\top$

Solve regularized problem: For  $\lambda > 0$

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{AX} - \mathbf{C}^\top\|_F^2 + \frac{\lambda}{2} \|\mathbf{X}\|_F^2$$

*Exercise: solve the regularized least-squares problem*

## Ill-posedness and Regularization

Important is the *effective rank*: If  $\sigma_j \ll \sigma_1$  for all  $j \geq k$ , then the effective rank of the problem is  $k$ .

If  $k < n_f$ , the least squares problem is ill-posed, i.e., solution does not exist or is unstable.

Small perturbations in  $\mathbf{C}$  or  $\mathbf{A} = \mathbf{Y}^\top$  yield large perturbations in  $\mathbf{X} = \mathbf{W}^\top$

Solve regularized problem: For  $\lambda > 0$

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{A}\mathbf{X} - \mathbf{C}^\top\|_F^2 + \frac{\lambda}{2} \|\mathbf{X}\|_F^2$$

*Exercise: solve the regularized least-squares problem*

$$\mathbf{X} = (\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^\top \mathbf{C}^\top$$

# The Bias-Variance Decomposition

Assume  $\mathbf{C}^\top = \mathbf{A}\mathbf{X}_{\text{true}} + \epsilon$ ,  $\epsilon \sim \mathcal{N}(0, \sigma\mathbf{I})$ ,  $\lambda > 0$  fixed.

Then setting  $\mathbf{A}_\lambda^\dagger = (\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{I})^{-1}$

$$\begin{aligned}\mathbf{X} - \mathbf{X}_{\text{true}} &= \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \mathbf{C}^\top - \mathbf{X}_{\text{true}} \\ &= \left( \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \mathbf{A} - \mathbf{I} \right) \mathbf{X}_{\text{true}} + \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \epsilon \\ &= -\lambda \mathbf{A}_\lambda^\dagger \mathbf{X}_{\text{true}} + \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \epsilon\end{aligned}$$

# The Bias-Variance Decomposition

Assume  $\mathbf{C}^\top = \mathbf{A}\mathbf{X}_{\text{true}} + \epsilon$ ,  $\epsilon \sim \mathcal{N}(0, \sigma\mathbf{I})$ ,  $\lambda > 0$  fixed.

Then setting  $\mathbf{A}_\lambda^\dagger = (\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{I})^{-1}$

$$\begin{aligned}\mathbf{X} - \mathbf{X}_{\text{true}} &= \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \mathbf{C}^\top - \mathbf{X}_{\text{true}} \\ &= \left( \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \mathbf{A} - \mathbf{I} \right) \mathbf{X}_{\text{true}} + \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \epsilon \\ &= -\lambda \mathbf{A}_\lambda^\dagger \mathbf{X}_{\text{true}} + \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \epsilon\end{aligned}$$

Error depends on  $\epsilon \rightsquigarrow$  take expectation

$$\begin{aligned}\mathbb{E} \|\mathbf{X} - \mathbf{X}_{\text{true}}\|_F^2 &= \mathbb{E} \|\mathbf{A}_\lambda^\dagger \mathbf{A}^\top \epsilon - \lambda \mathbf{A}_\lambda^\dagger \mathbf{X}_{\text{true}}\|_F^2 \\ &= \underbrace{\|\text{bias}\|_F^2}_{\lambda^2 \|\mathbf{A}_\lambda^\dagger \mathbf{X}_{\text{true}}\|_F^2} + \underbrace{\text{variance}}_{\sigma^2 \text{trace} \left( \mathbf{A} \mathbf{A}_\lambda^{\dagger \top} \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \right)}\end{aligned}$$

# The Bias-Variance Decomposition

Assume  $\mathbf{C}^\top = \mathbf{A}\mathbf{X}_{\text{true}} + \epsilon$ ,  $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ ,  $\lambda > 0$  fixed.

Then setting  $\mathbf{A}_\lambda^\dagger = (\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{I})^{-1}$

$$\begin{aligned}\mathbf{X} - \mathbf{X}_{\text{true}} &= \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \mathbf{C}^\top - \mathbf{X}_{\text{true}} \\ &= \left( \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \mathbf{A} - \mathbf{I} \right) \mathbf{X}_{\text{true}} + \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \epsilon \\ &= -\lambda \mathbf{A}_\lambda^\dagger \mathbf{X}_{\text{true}} + \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \epsilon\end{aligned}$$

Error depends on  $\epsilon \rightsquigarrow$  take expectation

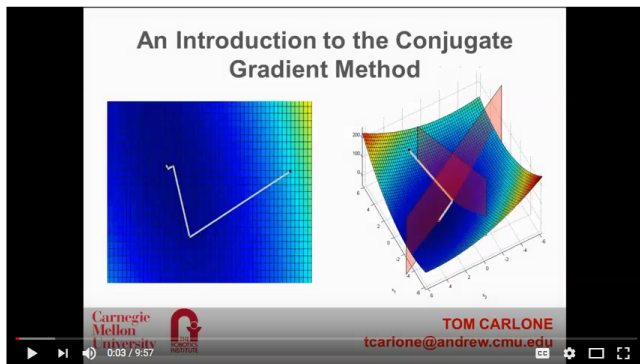
$$\begin{aligned}\mathbb{E} \|\mathbf{X} - \mathbf{X}_{\text{true}}\|_F^2 &= \mathbb{E} \|\mathbf{A}_\lambda^\dagger \mathbf{A}^\top \epsilon - \lambda \mathbf{A}_\lambda^\dagger \mathbf{X}_{\text{true}}\|_F^2 \\ &= \underbrace{\lambda^2 \|\mathbf{A}_\lambda^\dagger \mathbf{X}_{\text{true}}\|_F^2}_{\|\text{bias}\|_F^2} + \underbrace{\sigma^2 \text{trace} \left( \mathbf{A} \mathbf{A}_\lambda^{\dagger \top} \mathbf{A}_\lambda^\dagger \mathbf{A}^\top \right)}_{\text{variance}}\end{aligned}$$

Take home: No such thing as exact recovery!

# Next Time

Solving large-scale least-squares problems.

Watch: <https://www.youtube.com/watch?v=eAYohMUpPMA>



Overview of Conjugate Gradient Method

# References

- [1] U. M. Ascher and C. Greif. *A First Course on Numerical Methods*. SIAM, Philadelphia, 2011.
- [2] P. C. Hansen. *Rank-deficient and discrete ill-posed problems*. SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.
- [3] P. C. Hansen. *Discrete inverse problems*, volume 7 of *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2010.
- [4] C. R. Vogel. *Computational Methods for Inverse Problems*. SIAM, Philadelphia, 2002.