# 6.7900: Machine Learning
# Lecture 6

**Lecture start:** Tues/Thurs 2:35pm

**Who's speaking today?** Prof. Tamara Broderick

**Course website:** gradml.mit.edu

**Questions?** Ask here or on piazza.com/mit/fall2024/67900/

**Materials:** Slides, video, etc linked from gradml.mit.edu after the lecture (but there is no livestream)

## Last Time

I. Challenges with MLE/ERM for linear regression

II. Bayesian linear regression

   A. Bayes with Gaussians

   B. Posterior

## Today

I. Visualizations

II. Uncertainty

III. Ridge regression

IV. More flexible/complex features

# 6.7900: Machine Learning
# Lecture 6

**Lecture start:** Tues/Thurs 2:35pm

**Who's speaking today?** Prof. Tamara Broderick

**Course website:** gradml.mit.edu

**Questions?** Ask here or on piazza.com/mit/fall2024/67900/

**Materials:** Slides, video, etc linked from gradml.mit.edu after the lecture (but there is no livestream)

**Last Time**

I. Challenges with MLE/ERM for linear regression
II. Bayesian linear regression
   A. Bayes with Gaussians
   B. Posterior

**Today**

See board and demos

I. Visualizations
II. Uncertainty
III. Ridge regression
IV. More flexible/complex features

# Some points related to uncertainty

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities
- It's common to see publications discuss:

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities
- It's common to see publications discuss:
  - **Aleatoric uncertainty**: "intrinsic randomness"

[Aldous, "Is it practical and useful to distinguish aleatoric and epistemic uncertainty?"]

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities
- It's common to see publications discuss:
  - **Aleatoric uncertainty**: "intrinsic randomness"
  - **Epistemic uncertainty**: "lack of knowledge" about something that we could know for sure in principle

[Aldous, "Is it practical and useful to distinguish aleatoric and epistemic uncertainty?"]

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities
- It's common to see publications discuss:
  - **Aleatoric uncertainty**: "intrinsic randomness"
  - **Epistemic uncertainty**: "lack of knowledge" about something that we could know for sure in principle
  - Example: I plan to shuffle a deck of cards and then ask whether the top card is an ace.

[Aldous, "Is it practical and useful to distinguish aleatoric and epistemic uncertainty?"]

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities
- It's common to see publications discuss:
  - **Aleatoric uncertainty**: "intrinsic randomness"
  - **Epistemic uncertainty**: "lack of knowledge" about something that we could know for sure in principle
  - Example: I plan to shuffle a deck of cards and then ask whether the top card is an ace.
    - Before I shuffle, uncertainty is aleatoric. After shuffling but before looking, uncertainty is epistemic.

[Aldous, "Is it practical and useful to distinguish aleatoric and epistemic uncertainty?"]

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities
- It's common to see publications discuss:
  - **Aleatoric uncertainty**: "intrinsic randomness"
  - **Epistemic uncertainty**: "lack of knowledge" about something that we could know for sure in principle
  - Example: I plan to shuffle a deck of cards and then ask whether the top card is an ace.
    - Before I shuffle, uncertainty is aleatoric. After shuffling but before looking, uncertainty is epistemic.

[Aldous, "Is it practical and useful to distinguish aleatoric and epistemic uncertainty?"]

predictive vs posterior

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities
- It's common to see publications discuss:
  - **Aleatoric uncertainty**: "intrinsic randomness"
  - **Epistemic uncertainty**: "lack of knowledge" about something that we could know for sure in principle
  - Example: I plan to shuffle a deck of cards and then ask whether the top card is an ace.
    - Before I shuffle, uncertainty is aleatoric. After shuffling but before looking, uncertainty is epistemic.

[Aldous, "Is it practical and useful to distinguish aleatoric and epistemic uncertainty?"]

predictive vs posterior

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities
- It's common to see publications discuss:
  - **Aleatoric uncertainty**: "intrinsic randomness"
  - **Epistemic uncertainty**: "lack of knowledge" about something that we could know for sure in principle
  - Example: I plan to shuffle a deck of cards and then ask whether the top card is an ace.
    - Before I shuffle, uncertainty is aleatoric. After shuffling but before looking, uncertainty is epistemic.
- The Bayesian posterior uncertainty is conditional on our model. Goes to zero if enough data with enough $x$ coverage.

[Aldous, "Is it practical and useful to distinguish aleatoric and epistemic uncertainty?"]

predictive vs posterior

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities
- It's common to see publications discuss:
  - **Aleatoric uncertainty**: "intrinsic randomness"
  - **Epistemic uncertainty**: "lack of knowledge" about something that we could know for sure in principle
  - Example: I plan to shuffle a deck of cards and then ask whether the top card is an ace.
    - Before I shuffle, uncertainty is aleatoric. After shuffling but before looking, uncertainty is epistemic.
- The Bayesian posterior uncertainty is conditional on our model. Goes to zero if enough data with enough $x$ coverage.
  - "Uncertainty" as a term of art vs colloquial meaning

[Aldous, "Is it practical and useful to distinguish aleatoric and epistemic uncertainty?"]

predictive vs posterior

# Some points related to uncertainty

- Rather than return an error in the face of collinearity, the Bayesian approach expresses uncertainty over possibilities
- It's common to see publications discuss:
  - **Aleatoric uncertainty**: "intrinsic randomness"
  - **Epistemic uncertainty**: "lack of knowledge" about something that we could know for sure in principle
  - Example: I plan to shuffle a deck of cards and then ask whether the top card is an ace.
    - Before I shuffle, uncertainty is aleatoric. After shuffling but before looking, uncertainty is epistemic.

[Aldous, "Is it practical and useful to distinguish aleatoric and epistemic uncertainty?"]

predictive vs posterior

- The Bayesian posterior uncertainty is conditional on our model. Goes to zero if enough data with enough *x* coverage.
  - "Uncertainty" as a term of art vs colloquial meaning
  - Not just a Bayesian issue: (frequentist) standard errors go to 0 as data grows. So in real-life ("all models are wrong"), common p-values get arbitrarily small from enough data

# MAP and regularization

- What if I just want a single regression line and not a whole distribution?

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP.

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D})$$

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta}\ \text{-}\log p(\theta|\mathcal{D})$$

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{-}\log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \ \text{-}\log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

$$= \arg\min_{\theta} \left\{ (X\theta - Y)^\top (X\theta - Y) + \frac{\sigma^2}{\sigma_0^2}\theta^\top\theta \right\}$$

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{ -}\log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D \times D}$

$$= \arg\min_{\theta} \left\{ (X\theta - Y)^\top (X\theta - Y) + \frac{\sigma^2}{\sigma_0^2} \theta^\top \theta \right\}$$

the usual MLE/ERM
objective

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{-} \log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

$$= \arg\min_{\theta} \left\{ (X\theta - Y)^\top (X\theta - Y) + \frac{\sigma^2}{\sigma_0^2} \theta^\top \theta \right\}$$

the usual MLE/ERM objective

($\ell_2$/ridge) penalty/ regularizer

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{ -} \log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D \times D}$

**ridge regression**

$$= \arg\min_{\theta} \left\{ \underbrace{(X\theta - Y)^\top (X\theta - Y)}_{\text{the usual MLE/ERM objective}} + \underbrace{\frac{\sigma^2}{\sigma_0^2} \theta^\top \theta}_{(\ell_2\text{/ridge) penalty/regularizer}} \right\}$$

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{-}\log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

**ridge regression**

["ridge analysis" RW Hoerl 2020, Ridge Regression: A Historical Context; can think of "ridge" here due to collinearity in likelihood]

$$= \arg\min_{\theta} \left\{ (X\theta - Y)^\top (X\theta - Y) + \frac{\sigma^2}{\sigma_0^2}\theta^\top\theta \right\}$$

the usual MLE/ERM objective

($\ell_2$/ridge) penalty/ regularizer

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{-} \log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

**ridge regression**

["ridge analysis" RW Hoerl 2020, Ridge Regression: A Historical Context; can think of "ridge" here due to collinearity in likelihood]

$$= \arg\min_{\theta} \left\{ \underbrace{(X\theta - Y)^\top (X\theta - Y)}_{\text{the usual MLE/ERM objective}} + \underbrace{\frac{\sigma^2}{\sigma_0^2}\theta^\top\theta}_{(\ell_2\text{/ridge) penalty/regularizer}} \right\}$$

- Since the Gaussian posterior is symmetric, the MAP is equal to the mean here. So we already computed it in Lecture 5.

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{ -} \log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

**ridge regression**

["ridge analysis" RW Hoerl 2020, Ridge Regression: A Historical Context; can think of "ridge" here due to collinearity in likelihood]

$$= \arg\min_{\theta} \left\{ \underbrace{(X\theta - Y)^\top (X\theta - Y)}_{\text{the usual MLE/ERM objective}} + \underbrace{\frac{\sigma^2}{\sigma_0^2}\theta^\top\theta}_{(\ell_2\text{/ridge) penalty/ regularizer}} \right\}$$

- Since the Gaussian posterior is symmetric, the MAP is equal to the mean here. So we already computed it in Lecture 5.
  - Letting $\lambda = \sigma^2/\sigma_0^2$,

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{-} \log p(\theta|\mathcal{D})$$

- consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

**ridge regression**

["ridge analysis" RW Hoerl 2020, Ridge Regression: A Historical Context; can think of "ridge" here due to collinearity in likelihood]

$$= \arg\min_{\theta} \left\{ \underbrace{(X\theta - Y)^\top (X\theta - Y)}_{\text{the usual MLE/ERM objective}} + \underbrace{\frac{\sigma^2}{\sigma_0^2}\theta^\top\theta}_{(\ell_2\text{/ridge) penalty/regularizer}} \right\}$$

the usual MLE/ERM objective

($\ell_2$/ridge) penalty/regularizer

- Since the Gaussian posterior is symmetric, the MAP is equal to the mean here. So we already computed it in Lecture 5.
  - Letting $\lambda = \sigma^2/\sigma_0^2$, $\hat{\theta}_{\mathrm{MAP}} = (\lambda I_{D\times D} + X^\top X)^{-1}(X^\top Y)$

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{ -} \log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

**ridge regression**

$$= \arg\min_{\theta} \left\{ (X\theta - Y)^\top (X\theta - Y) + \frac{\sigma^2}{\sigma_0^2} \theta^\top \theta \right\}$$

["ridge analysis" RW Hoerl 2020, Ridge Regression: A Historical Context; can think of "ridge" here due to collinearity in likelihood]

the usual MLE/ERM objective

($\ell_2$/ridge) penalty/ regularizer

- Since the Gaussian posterior is symmetric, the MAP is equal to the mean here. So we already computed it in Lecture 5.
  - Letting $\lambda = \sigma^2/\sigma_0^2$, $\hat{\theta}_{\text{MAP}} = (\lambda I_{D\times D} + X^\top X)^{-1}(X^\top Y)$

What happens as $\lambda$ gets very small?

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{ -} \log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

**ridge regression**

["ridge analysis" RW Hoerl 2020, Ridge Regression: A Historical Context; can think of "ridge" here due to collinearity in likelihood]

$$= \arg\min_{\theta} \left\{ \underbrace{(X\theta - Y)^\top (X\theta - Y)}_{\text{the usual MLE/ERM objective}} + \underbrace{\frac{\sigma^2}{\sigma_0^2}\theta^\top\theta}_{(\ell_2\text{/ridge) penalty/regularizer}} \right\}$$

- Since the Gaussian posterior is symmetric, the MAP is equal to the mean here. So we already computed it in Lecture 5.
  - Letting $\lambda = \sigma^2/\sigma_0^2$, $\hat{\theta}_{\mathrm{MAP}} = (\lambda I_{D\times D} + X^\top X)^{-1}(X^\top Y)$

What happens as $\lambda$ gets very small? Very large?

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{ -}\log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

**ridge regression**

["ridge analysis" RW Hoerl 2020, Ridge Regression: A Historical Context; can think of "ridge" here due to collinearity in likelihood]

$$= \arg\min_{\theta} \left\{ \underbrace{(X\theta - Y)^\top (X\theta - Y)}_{\text{the usual MLE/ERM objective}} + \underbrace{\frac{\sigma^2}{\sigma_0^2}\theta^\top\theta}_{(\ell_2\text{/ridge) penalty/regularizer}} \right\}$$

- Since the Gaussian posterior is symmetric, the MAP is equal to the mean here. So we already computed it in Lecture 5.
  - Letting $\lambda = \sigma^2/\sigma_0^2$, $\hat{\theta}_{\mathrm{MAP}} = (\lambda I_{D\times D} + X^\top X)^{-1}(X^\top Y)$
- Avoids the inversion issue of the MLE by choosing a particular option among options "equally good" in MLE

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{-}\log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

**ridge regression**

["ridge analysis" RW Hoerl 2020, Ridge Regression: A Historical Context; can think of "ridge" here due to collinearity in likelihood]

$$= \arg\min_{\theta} \left\{ \underbrace{(X\theta - Y)^\top (X\theta - Y)}_{\text{the usual MLE/ERM objective}} + \underbrace{\frac{\sigma^2}{\sigma_0^2}\theta^\top\theta}_{(\ell_2\text{/ridge) penalty/ regularizer}} \right\}$$

- Since the Gaussian posterior is symmetric, the MAP is equal to the mean here. So we already computed it in Lecture 5.
  - Letting $\lambda = \sigma^2/\sigma_0^2$, $\hat{\theta}_{\text{MAP}} = (\lambda I_{D\times D} + X^\top X)^{-1}(X^\top Y)$
- Avoids the inversion issue of the MLE by choosing a particular option among options "equally good" in MLE

check: always invertible

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{-}\log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

**ridge regression**

["ridge analysis" RW Hoerl 2020, Ridge Regression: A Historical Context; can think of "ridge" here due to collinearity in likelihood]

$$= \arg\min_{\theta} \left\{ \underbrace{(X\theta - Y)^\top (X\theta - Y)}_{\text{the usual MLE/ERM objective}} + \underbrace{\frac{\sigma^2}{\sigma_0^2}\theta^\top\theta}_{(\ell_2\text{/ridge) penalty/regularizer}} \right\}$$

- Since the Gaussian posterior is symmetric, the MAP is equal to the mean here. So we already computed it in Lecture 5.
  - Letting $\lambda = \sigma^2/\sigma_0^2$, $\hat{\theta}_{\text{MAP}} = (\lambda I_{D\times D} + X^\top X)^{-1}(X^\top Y)$
- Avoids the inversion issue of the MLE by choosing a particular option among options "equally good" in MLE

check: always invertible

  - No uncertainty (e.g. when regression is underdetermined)

# MAP and regularization

- What if I just want a single regression line and not a whole distribution? Some options: posterior mean or MAP. Lecture 3:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\min_{\theta} \text{-}\log p(\theta|\mathcal{D})$$

  - consider the linear regression model we used, now with $\mu_0 = 0_D$ & $\Sigma_0 = \sigma_0^2 I_{D\times D}$

**ridge regression**

["ridge analysis" RW Hoerl 2020, Ridge Regression: A Historical Context; can think of "ridge" here due to collinearity in likelihood]

$$= \arg\min_{\theta} \left\{ \underbrace{(X\theta - Y)^\top (X\theta - Y)}_{\text{the usual MLE/ERM objective}} + \underbrace{\frac{\sigma^2}{\sigma_0^2}\theta^\top\theta}_{(\ell_2\text{/ridge) penalty/regularizer}} \right\}$$

- Since the Gaussian posterior is symmetric, the MAP is equal to the mean here. So we already computed it in Lecture 5.
  - Letting $\lambda = \sigma^2/\sigma_0^2$, $\hat{\theta}_{\text{MAP}} = (\lambda I_{D\times D} + X^\top X)^{-1}(X^\top Y)$
- Avoids the inversion issue of the MLE by choosing a particular option among options "equally good" in MLE

check: always invertible

  - No uncertainty (e.g. when regression is underdetermined)
  - May offer better generalization than vanilla MLE

# A note on features

# A note on features

- Linear models can be very flexible given non-trivial features

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$
- But we could take $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^\top \phi(x)$

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$
- But we could take $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^\top \phi(x)$
  - Now *x* can be any dimension (and need not be real-valued), and *D* is the dimension of the features $\phi(x)$

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$
- But we could take $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^\top \phi(x)$
  - Now *x* can be any dimension (and need not be real-valued), and *D* is the dimension of the features $\phi(x)$
  - E.g. for $x \in \mathbb{R}$, $\phi(x) = [1, x, x^2]^\top$

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$
- But we could take $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^\top \phi(x)$
  - Now *x* can be any dimension (and need not be real-valued), and *D* is the dimension of the features $\phi(x)$
  - E.g. for $x \in \mathbb{R}$, $\phi(x) = [1, x, x^2]^\top$ recall demo

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$
- But we could take $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^\top \phi(x)$
  - Now *x* can be any dimension (and need not be real-valued), and *D* is the dimension of the features $\phi(x)$
  - E.g. for $x \in \mathbb{R}$, $\phi(x) = [1, x, x^2]^\top$ recall demo
  - More generally, for $x \in \mathbb{R}^{D_x}$, $\phi(x)$ could collect all polynomials of some degree *r* or smaller

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$
- But we could take $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^\top \phi(x)$
  - Now *x* can be any dimension (and need not be real-valued), and *D* is the dimension of the features $\phi(x)$
  - E.g. for $x \in \mathbb{R}$, $\phi(x) = [1, x, x^2]^\top$ recall demo
  - More generally, for $x \in \mathbb{R}^{D_x}$, $\phi(x)$ could collect all polynomials of some degree *r* or smaller
  - Lots of other options: e.g. Fourier basis, logistic basis, wavelets, splines

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$
- But we could take $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^\top \phi(x)$
  - Now *x* can be any dimension (and need not be real-valued), and *D* is the dimension of the features $\phi(x)$
  - E.g. for $x \in \mathbb{R}$, $\phi(x) = [1, x, x^2]^\top$ recall demo
  - More generally, for $x \in \mathbb{R}^{D_x}$, $\phi(x)$ could collect all polynomials of some degree *r* or smaller
  - Lots of other options: e.g. Fourier basis, logistic basis, wavelets, splines
  - For any fixed $\phi(x)$, basically all the math we did goes through with $\phi(x)$ in place of *x*

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$
- But we could take $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^\top \phi(x)$
  - Now *x* can be any dimension (and need not be real-valued), and *D* is the dimension of the features $\phi(x)$
  - E.g. for $x \in \mathbb{R}$, $\phi(x) = [1, x, x^2]^\top$ recall demo
  - More generally, for $x \in \mathbb{R}^{D_x}$, $\phi(x)$ could collect all polynomials of some degree *r* or smaller
  - Lots of other options: e.g. Fourier basis, logistic basis, wavelets, splines
  - For any fixed $\phi(x)$, basically all the math we did goes through with $\phi(x)$ in place of *x*
    - E.g. for *N*x*D* matrix $\Phi$ with *n*th row $\phi(x^{(n)})^\top$

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$
- But we could take $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^\top \phi(x)$
  - Now *x* can be any dimension (and need not be real-valued), and *D* is the dimension of the features $\phi(x)$
  - E.g. for $x \in \mathbb{R}$, $\phi(x) = [1, x, x^2]^\top$ recall demo
  - More generally, for $x \in \mathbb{R}^{D_x}$, $\phi(x)$ could collect all polynomials of some degree *r* or smaller
  - Lots of other options: e.g. Fourier basis, logistic basis, wavelets, splines
  - For any fixed $\phi(x)$, basically all the math we did goes through with $\phi(x)$ in place of *x*
    - E.g. for *N*x*D* matrix $\Phi$ with *n*th row $\phi(x^{(n)})^\top$
      - Then OLS: $\hat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top Y$

# A note on features

- Linear models can be very flexible given non-trivial features
- We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^\top x$
- But we could take $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^\top \phi(x)$
  - Now *x* can be any dimension (and need not be real-valued), and *D* is the dimension of the features $\phi(x)$
  - E.g. for $x \in \mathbb{R}$, $\phi(x) = [1, x, x^2]^\top$ recall demo
  - More generally, for $x \in \mathbb{R}^{D_x}$, $\phi(x)$ could collect all polynomials of some degree *r* or smaller
  - Lots of other options: e.g. Fourier basis, logistic basis, wavelets, splines
  - For any fixed $\phi(x)$, basically all the math we did goes through with $\phi(x)$ in place of *x*
    - E.g. for *N*x*D* matrix $\Phi$ with *n*th row $\phi(x^{(n)})^\top$
      - Then OLS: $\hat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top Y$
- Deep neural networks perform nonlinear feature extraction/learning