

S&DS 265 / 565

Introductory Machine Learning

# Stochastic Gradient Descent and Bias-Variance Tradeoffs

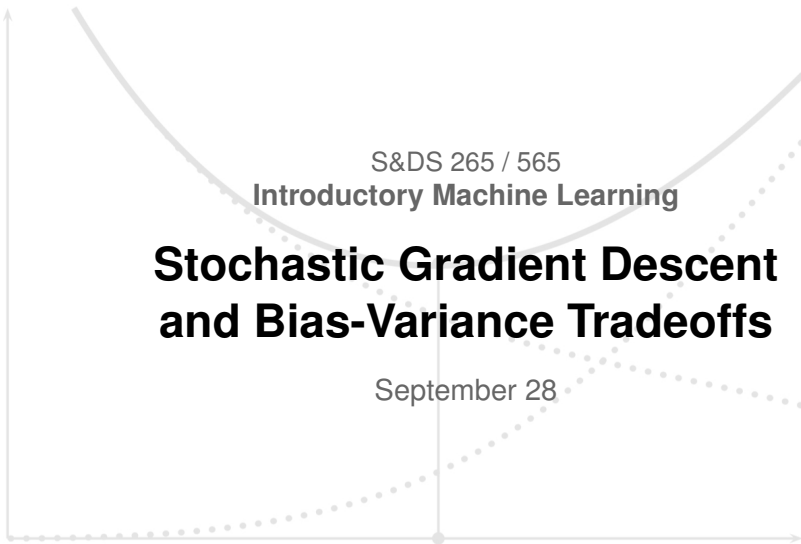
September 28

Risk

Bias squared

Variance

Yale










# Goings on

- Nothing due!
- Get started on Assn 2!
- Questions?

# Outline for today

- Stochastic gradient descent (redux)
- Regularization
- Jupyter notebook example
- Bias-variance tradeoffs

# You are here

					
5	Sept 28, 30	Bias and variance, cross-validation	 <a href="#">Bias-variance tradeoff</a>		
6	Oct 5, 7	Tree-based methods	 <a href="#">Trees and forests</a>	Tue: Assn 2 in; Assn 3 out	
7	Oct 12, 14	PCA and dimension reduction	 <a href="#">PCA examples</a>	Tue: Quiz 2 Thu: Assn 3 in; Assn 4 out	
8	Oct 19	Midterm exam (in class)			
9	Oct 26, 28	Language models, word embeddings	 <a href="#">Word embeddings</a>		
10	Nov 2, 4	Bayesian inference, topic models	 <a href="#">Bayesian inference</a>  <a href="#">Topic models</a>	Tue: Assn 4 in; Assn 5 out	

# SGD idea

Here's the idea:

- For each parameter  $\beta_j$ , see what happens to the loss if that parameter is increased a little bit.
- If the loss goes down (up), then increase (decrease)  $\beta_j$  proportionately
- Do this simultaneously for all of the parameters
- Rinse and repeat

# Stochastic gradient descent

Initialize all parameters to zero:  $\beta_j = 0, j = 1, \dots, p$ .

Read through the data one record at a time, and update the model.

- 1 Read data item  $x$
- 2 Make a prediction  $\hat{y}(x)$
- 3 Observe the true response/label  $y$
- 4 Update the parameters  $\beta$  so  $\hat{y}$  is closer to  $y$

# Stochastic gradient descent

Suppose we are doing *linear regression*. We initialize all parameters to zero:  $\beta_j = 0, j = 1, \dots, p$ .

We read through the data one record at a time, and update the model.

- 1 Read data item  $x$
- 2 Make a prediction  $\hat{y}(x) = \sum_{j=1}^p \beta_j x_j$
- 3 Observe the true response/label  $y$
- 4 Update the parameters  $\beta$  so  $\hat{y}$  is closer to  $y$

# SGD idea

Change  $\beta_j$  by a little bit:

$$\beta_j \rightarrow \beta_j + \varepsilon$$



# SGD idea

Change  $\beta_j$  by a little bit:

$$\beta_j \rightarrow \beta_j + \varepsilon$$

What happens to the squared error?

$$\begin{aligned}(y - \hat{y})^2 &\rightarrow (y - \hat{y} - \varepsilon x_j)^2 \\ &\approx (y - \hat{y})^2 + \underbrace{-2(y - \hat{y})x_j}_{\text{derivative of loss}} \varepsilon\end{aligned}$$

# SGD idea

Change  $\beta_j$  by a little bit:

$$\beta_j \rightarrow \beta_j + \varepsilon$$

What happens to the squared error?

$$\begin{aligned}(y - \hat{y})^2 &\rightarrow (y - \hat{y} - \varepsilon x_j)^2 \\ &\approx (y - \hat{y})^2 + \underbrace{-2(y - \hat{y})x_j}_{\text{derivative of loss}} \varepsilon\end{aligned}$$

Use adjustment

$$\begin{aligned}\beta_j &\rightarrow \beta_j + \overbrace{-\eta \cdot \text{derivative of loss}}^{\varepsilon} \\ &= \beta_j + \eta \cdot 2(y - \hat{y})x_j\end{aligned}$$

# SGD idea

Change  $\beta_j$  by a little bit:

$$\beta_j \rightarrow \beta_j + \varepsilon$$

What happens to the squared error?

$$\begin{aligned}(y - \hat{y})^2 &\rightarrow (y - \hat{y} - \varepsilon x_j)^2 \\ &\approx (y - \hat{y})^2 + \underbrace{-2(y - \hat{y})x_j}_{\text{derivative of loss}} \varepsilon\end{aligned}$$

Use adjustment

$$\begin{aligned}\beta_j &\rightarrow \beta_j + \overbrace{-\eta \cdot \text{derivative of loss}}^{\varepsilon} \\ &= \beta_j + \eta \cdot 2(y - \hat{y})x_j\end{aligned}$$

Squared error then decreases:

$$(y - \hat{y})^2 \approx (y - \hat{y})^2 - 4\eta(y - \hat{y})^2 x_j^2$$

# SGD for general loss

Suppose  $L(y, \beta^T x)$  is the loss for an input  $(x, y)$ , e.g.,  $(y - \beta^T x)^2$

SGD update:

$$\beta_j \leftarrow \beta_j - \eta \frac{\partial L(y, \beta^T x)}{\partial \beta_j}$$

$$\beta \leftarrow \beta - \eta \nabla_{\beta} L(y, \beta^T x) \quad (\text{vector notation})$$

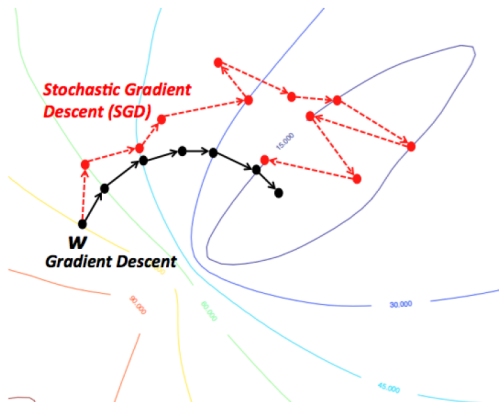
“Batch” gradient descent uses the entire training set in each step of gradient descent.

*Stochastic* gradient descent computes a quick approximation to this gradient, using only a single or a small “mini-batch” of data points

# Batch vs. stochastic gradient descent

- The average derivative over a mini-batch can be thought of as a noisy version of the average derivative over the entire data set
- (Which can in turn be thought of as a sample estimate of a population)
- The stochastic gradient is computed more cheaply, and updating the parameters makes progress more quickly

# Batch vs. stochastic gradient descent



<https://wikidocs.net/3413>

# SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose  $y = 1$  and probability  $p(x)$  is high?

# SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose  $y = 1$  and probability  $p(x)$  is high? *small change*
- Suppose  $y = 1$  and probability  $p(x)$  is small?



# SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose  $y = 1$  and probability  $p(x)$  is high? *small change*
- Suppose  $y = 1$  and probability  $p(x)$  is small? *big change*  $\uparrow$
- Suppose  $y = 0$  and probability  $p(x)$  is small?

# SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose  $y = 1$  and probability  $p(x)$  is high? *small change*
- Suppose  $y = 1$  and probability  $p(x)$  is small? *big change*  $\uparrow$
- Suppose  $y = 0$  and probability  $p(x)$  is small? *small change*
- Suppose  $y = 0$  and probability  $p(x)$  is big?

# SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose  $y = 1$  and probability  $p(x)$  is high? *small change*
- Suppose  $y = 1$  and probability  $p(x)$  is small? *big change*  $\uparrow$
- Suppose  $y = 0$  and probability  $p(x)$  is small? *small change*
- Suppose  $y = 0$  and probability  $p(x)$  is big? *big change*  $\downarrow$

# SGD: choice of learning rate

A conservative choice of learning rate is

$$\eta_t = \frac{1}{t}$$

A more aggressive choice is

$$\eta_t = \frac{1}{\sqrt{t}}$$

In practice: Try learning rates  $C/\sqrt{t}$  for different choices of  $C$ , and monitor the error

$$\frac{1}{T} \sum_{t=1}^T (Y_t - \hat{Y}_t)^2$$

# SGD: Regularization

A “ridge” penalty  $\frac{1}{2}\lambda \sum_{j=1}^d \beta_j^2$  is easily handled.

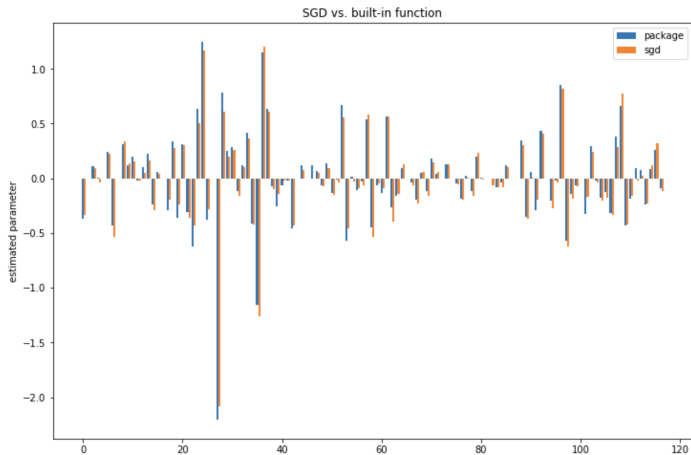
Gradient changes by an additive term  $2\lambda\beta_j$ . Update becomes

$$\begin{aligned}\beta_j &\longleftarrow \beta_j + \eta\{(y - p(x))x_j - \lambda\beta_j\} \\ &= (1 - \eta\lambda)\beta_j + \eta(y - p(x))x_j\end{aligned}$$

Check that this “does the right thing” whether  $\beta_j$  wants to be large positive or negative.

- *The penalty shrinks  $\beta_j$  toward zero*

# Recall from demo



# Bias and variance

Bias: How much are we off—on average?

Variance: How variable are we—on average?

# Bias and variance

$$\text{Bias: } \theta - \mathbb{E}\hat{\theta}$$

$$\text{Variance: } \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$



# Bias and variance

Examples of  $\theta, \hat{\theta}$ :

Estimating height, population, election outcome, ad click rate...

# Bias and variance

$$\text{Bias: } \theta - \mathbb{E}\hat{\theta}$$

$$\text{Variance: } \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$

# Bias and variance

$$\text{Bias: } \theta - \mathbb{E}\hat{\theta}$$

$$\text{Variance: } \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$

- $\hat{\theta}$  is an estimate from a sample
- $\mathbb{E}$  is the expectation (average) with respect to the sample
- So  $\mathbb{E}\hat{\theta}$  is the average estimate
- We can only directly compute  $\hat{\theta}$  for the sample we have
- *We don't know  $\theta$*

# Bias and variance

Bias and variance are two sides of the same coin: As squared bias goes up, variance goes down

# Bias-variance tradeoff

$$\text{Risk} = \text{Bias}^2 + \text{Variance}$$

# Bias-variance tradeoff

$$\mathbb{E}(\theta - \hat{\theta})^2 = \text{Bias}(\hat{\theta})^2 + \text{Variance}(\theta)$$

# Bias-variance tradeoff

$$\mathbb{E}(\theta - \hat{\theta})^2 = (\theta - \mathbb{E}\hat{\theta})^2 + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$

# Bias-variance tradeoff

Proof:

$$\mathbb{E}(\theta - \hat{\theta})^2 = \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2$$



# Bias-variance tradeoff

Proof:

$$\begin{aligned}\mathbb{E}(\theta - \hat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2 \\ &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})\right\} + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2\end{aligned}$$

# Bias-variance tradeoff

Proof:

$$\begin{aligned}\mathbb{E}(\theta - \hat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})\right\} + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2(\theta - \mathbb{E}\hat{\theta})\mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta}) + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2\end{aligned}$$

# Bias-variance tradeoff

Proof:

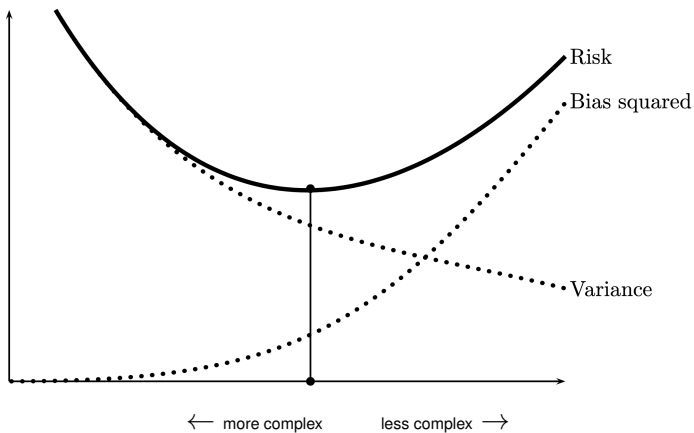
$$\begin{aligned}\mathbb{E}(\theta - \hat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})\right\} + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2(\theta - \mathbb{E}\hat{\theta})\mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta}) + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2\end{aligned}$$

# Bias-variance tradeoff

Proof:

$$\begin{aligned}\mathbb{E}(\theta - \hat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})\right\} + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2(\theta - \mathbb{E}\hat{\theta})\mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta}) + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \text{Bias}(\hat{\theta})^2 + \text{Variance}(\hat{\theta})\end{aligned}$$

# Bias-variance tradeoff



## Example: Regularization

Suppose that  $\mathbb{E}(Y) = \theta^*$  and we estimate

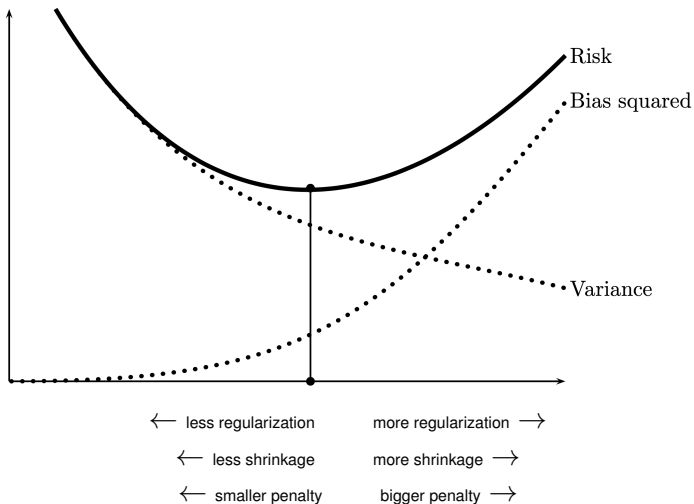
$$\hat{\theta} = \arg \min_{\theta} (Y - \theta)^2 + \lambda \theta^2$$

Then  $\hat{\theta} = \frac{Y}{1+\lambda}$ . What are the squared bias and variance?

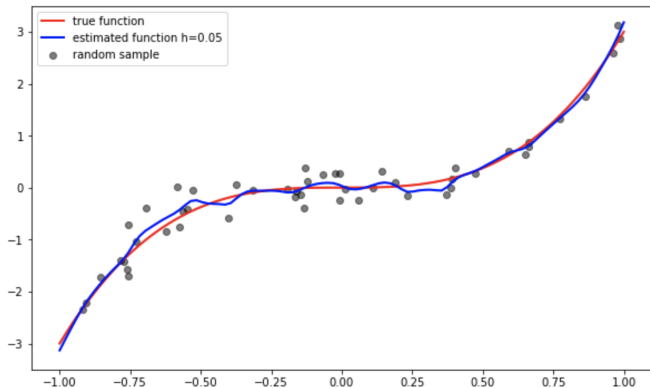
$$\text{Bias}^2 = \theta^{*2} \left( \frac{\lambda}{1+\lambda} \right)^2$$

$$\text{Variance} = \left( \frac{1}{1+\lambda} \right)^2 \text{Variance}(Y)$$

# Bias-variance tradeoff

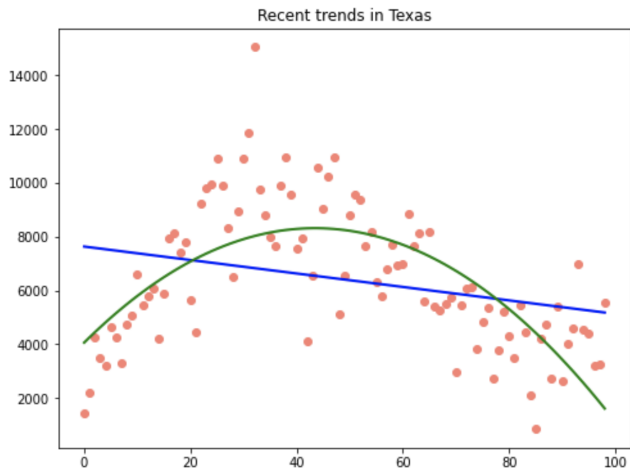


# Let's go to the first notebook





# Let's go to the second notebook



# What did we learn today?

- In SGD, a parameter is updated according to how much the loss changes when that parameter is changed by a little bit
- Mean squared error splits into squared bias plus variance
- As model complexity increases, squared bias decreases while variance increases