CS 4644 / 7643-A

DEEP LEARNING: LECTURE 3

DANFEI XU

(SLIDE ADAPTED FROM PROF. ZSOLT KIRA)

- Linear Classifier (cont.)
- SVM / Hinge Loss
- Softmax Classifier and Cross-Entropy Loss
- Gradient Descent

Recap:

Supervised Learning

- Train Input: $\{X, Y\}$
- Learning output: $f : X \rightarrow Y$, e.g. $P(y|x)$

Unsupervised Learning

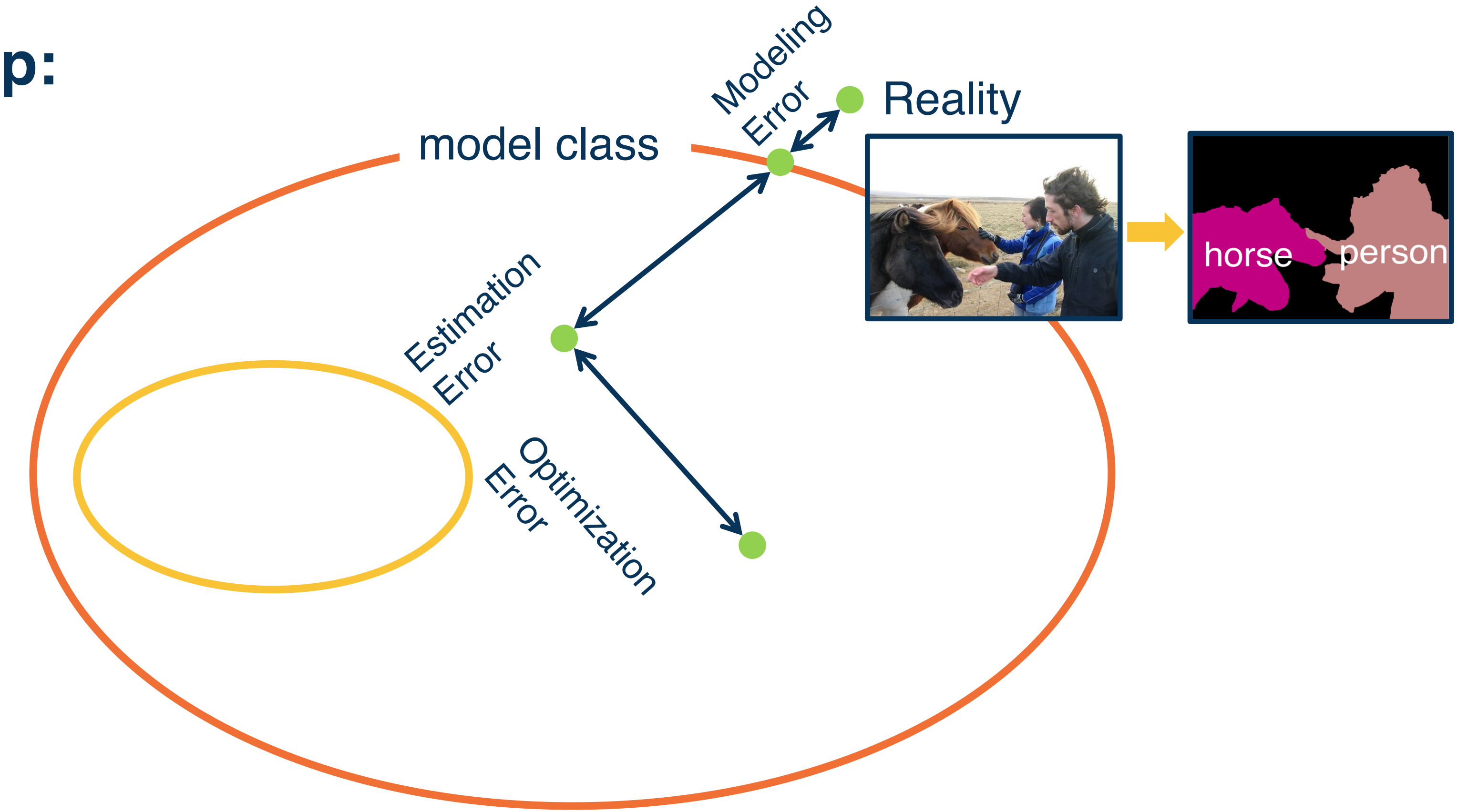
- Input: $\{X\}$
- Learning output: $P(x)$
- Example: Clustering, density estimation, etc.

Reinforcement Learning

- Supervision in form of **reward**
- No supervision on what action to take

Very often combined, sometimes within the same model!

Recap:



model class

Modeling Error

Reality

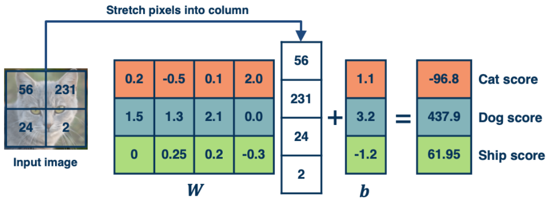
Estimation Error

Optimization Error

Recap:

Algebraic Viewpoint

$$f(x, W) = Wx$$



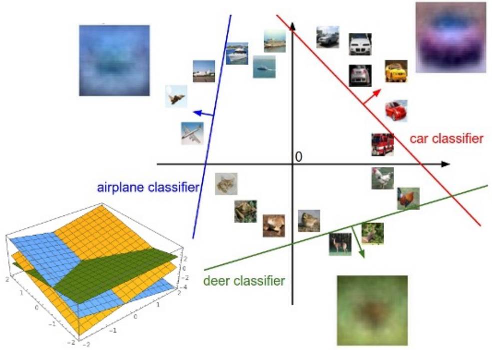
$$\begin{bmatrix} 0.2 & -0.5 & 0.1 & 2.0 \\ 1.5 & 1.3 & 2.1 & 0.0 \\ 0 & 0.25 & 0.2 & -0.3 \end{bmatrix} \begin{bmatrix} 56 \\ 231 \\ 24 \\ 2 \end{bmatrix} + \begin{bmatrix} 1.1 \\ 3.2 \\ -1.2 \end{bmatrix} = \begin{bmatrix} -96.8 \\ 437.9 \\ 61.95 \end{bmatrix}$$

Visual Viewpoint

One template per class

Geometric Viewpoint

Hyperplanes cutting up space



Adapted from from CS 231n slides

This time:

$$f(x, W) = Wx$$

airplane	-3.45	-0.51	3.42
automobile	-8.87	**6.04**	4.64
bird	0.09	5.31	2.65
cat	**2.9**	-4.22	5.1
deer	4.48	-4.19	2.64
dog	8.02	3.58	5.55
frog	3.78	4.49	**-4.34**
horse	1.06	-4.37	-1.5
ship	-0.36	-2.09	-4.79
truck	-0.72	-2.93	6.14

1. Define a **loss function** that quantifies our unhappiness with the scores across the training data.

2. Come up with a way of efficiently finding the parameters that minimize the loss function. **(optimization)**

SVM Loss Example

Suppose: 3 training examples, 3 classes.
With some W the scores $f(x,W)=Wx$ are:

cat	**3.2**	1.3	2.2
car	5.1	**4.9**	2.5
frog	-1.7	2.0	**-3.1**
	High Loss	Low Loss	High Loss

A **loss function** that tells how good the current classifier is

Given a dataset of examples:

$$\{(x_i, y_i)\}_{i=1}^{N}$$

Where x_i is image and
y_i is (integer) label

Loss over the dataset is a sum of loss over examples:

$$L = \frac{1}{N}\sum L(f(x_i, W), y_i)$$

Adapted from from CS 231n slides

Multiclass SVM loss:

Given an example $(\boldsymbol{x_{i,}} \boldsymbol{y_i})$
where $\boldsymbol{x_i}$ is the image and
where $\boldsymbol{y_i}$ is the (integer) label,

and using the shorthand for the
scores vector: $\boldsymbol{s = f(x_i, W)}$

Notation: $\boldsymbol{s_{y_i}}$ is the **score** given by the classifier for the correct label class of the i-th example (y_i)

the SVM loss has the form:

$$\boldsymbol{L_i = \sum_{j \neq y_i} \begin{cases} 0 & \text{if } s_{y_i} \geq s_j + 1 \\ s_j - s_{y_i} + 1 & \text{otherwise} \end{cases}}$$

$$\boldsymbol{= \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)}$$

Multiclass SVM loss:

Given an example (x_i, y_i) where x_i is the image and where y_i is the (integer) label,

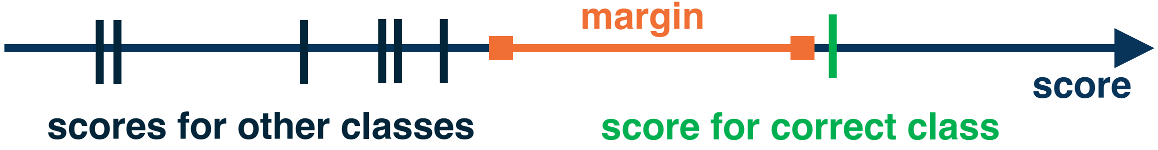
and using the shorthand for the scores vector: $s = f(x_i, W)$

the SVM loss has the form:

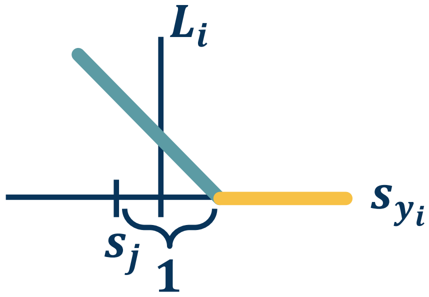
$$L_i = \sum_{j \neq y_i} \begin{cases} \boxed{0 \qquad \text{if } s_{y_i} \geq s_j + 1} \\ \boxed{s_j - s_{y_i} + 1 \quad \text{otherwise}} \end{cases}$$

$$= \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)$$

Loss = 0:



“Hinge Loss”



Multiclass SVM loss:

Given an example $(\boldsymbol{x_i}, \boldsymbol{y_i})$ where $\boldsymbol{x_i}$ is the image and where $\boldsymbol{y_i}$ is the (integer) label,

and using the shorthand for the scores vector: $\boldsymbol{s} = \boldsymbol{f}(\boldsymbol{x_i}, \boldsymbol{W})$

the SVM loss has the form:

$$\boxed{L_i = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)}$$

$= \max(0, 5.1 - 3.2 + 1) + \max(0, -1.7 - 3.2 + 1)$

$= \max(0, 2.9) + \max(0, -3.9)$

$= 2.9 + 0$

$= 2.9$

Suppose: 3 training examples, 3 classes.
With some $\boldsymbol{W}$ the scores $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{W}\boldsymbol{x}$ are:

cat	**3.2**	1.3	2.2
car	5.1	**4.9**	2.5
frog	-1.7	2.0	**-3.1**
Losses:	**2.9**		

Adapted from from CS 231n slides

SVM Loss Example

Multiclass SVM loss:

Given an example (x_i, y_i) where x_i is the image and where y_i is the (integer) label,

and using the shorthand for the scores vector: $s = f(x_i, W)$

the SVM loss has the form:

$$\boxed{L_i = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)}$$

= max(0, 1.3 - 4.9 + 1) + max(0, 2.0 - 4.9 + 1)
= max(0, -2.6) + max(0, -1.9)
= 0 + 0
= 0

Suppose: 3 training examples, 3 classes.
With some W the scores $f(x,W) = Wx$ are:

cat	**3.2**	1.3	2.2
car	5.1	**4.9**	2.5
frog	-1.7	2.0	**-3.1**
Losses:	2.9	**0.0**	

Adapted from from CS 231n slides

SVM Loss Example

Multiclass SVM loss:

Given an example (x_i, y_i)
where x_i is the image and
where y_i is the (integer) label,

and using the shorthand for the scores vector: $s = f(x_i, W)$

the SVM loss has the form:

$$L_i = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)$$

L = (2.9 + 0 + 12.9)/3
= **5.27**

Suppose: 3 training examples, 3 classes.
With some W the scores $f(x, W) = Wx$ are:

cat	**3.2**	1.3	2.2
car	5.1	**4.9**	2.5
frog	-1.7	2.0	**-3.1**
Losses:	2.9	0	12.9

Adapted from from CS 231n slides

SVM Loss Example

Multiclass SVM loss:

Suppose: 3 training examples, 3 classes.
With some W the scores $f(x,W)=Wx$ are:

$$L_i = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)$$

Q: What happens to loss if car image scores change a bit (e.g., ± 0.1)?

No change for small values

cat	**3.2**	1.3	2.2
car	5.1	**4.9**	2.5
frog	-1.7	2.0	**-3.1**

Adapted from from CS 231n slides

SVM Loss Example

SVM Loss Example

Multiclass SVM loss:

$$L_i = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)$$

Q: What is min/max of loss value?

[0,inf]

Suppose: 3 training examples, 3 classes.
With some W the scores $f(x,W)=Wx$ are:

cat	**3.2**	1.3	2.2
car	5.1	**4.9**	2.5
frog	-1.7	2.0	**-3.1**

Adapted from from CS 231n slides

Multiclass SVM loss:

$$L_i = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)$$

Q: At initialization W is close to 0 so all s ≈ 0. What is the loss?

num_class - 1

Suppose: 3 training examples, 3 classes.
With some $\boldsymbol{W}$ the scores $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{W})=\boldsymbol{W}\boldsymbol{x}$ are:

cat	**3.2**	1.3	2.2
car	5.1	**4.9**	2.5
frog	-1.7	2.0	**-3.1**

Adapted from from CS 231n slides

SVM Loss Example

Multiclass SVM loss:

Suppose: 3 training examples, 3 classes.
With some W the scores $f(x,W)=Wx$ are:

$$L_i = \frac{1}{C}\sum_{j\neq y_i} max(0, s_j - s_{y_i} + 1)$$

Q: What if we used mean instead of sum?

No difference
Scaling by constant

cat	**3.2**	1.3	2.2
car	5.1	**4.9**	2.5
frog	-1.7	2.0	**-3.1**

Adapted from from CS 231n slides

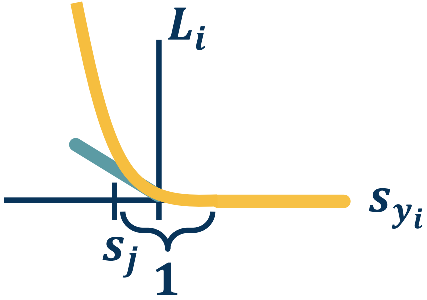
SVM Loss Example

SVM Loss Example

Multiclass SVM loss:

$$L_i = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)^2$$

Q: What if we used **squared** hinge loss?



- Smooth loss around hinge
- Sensitive to outliers (larger penalty)

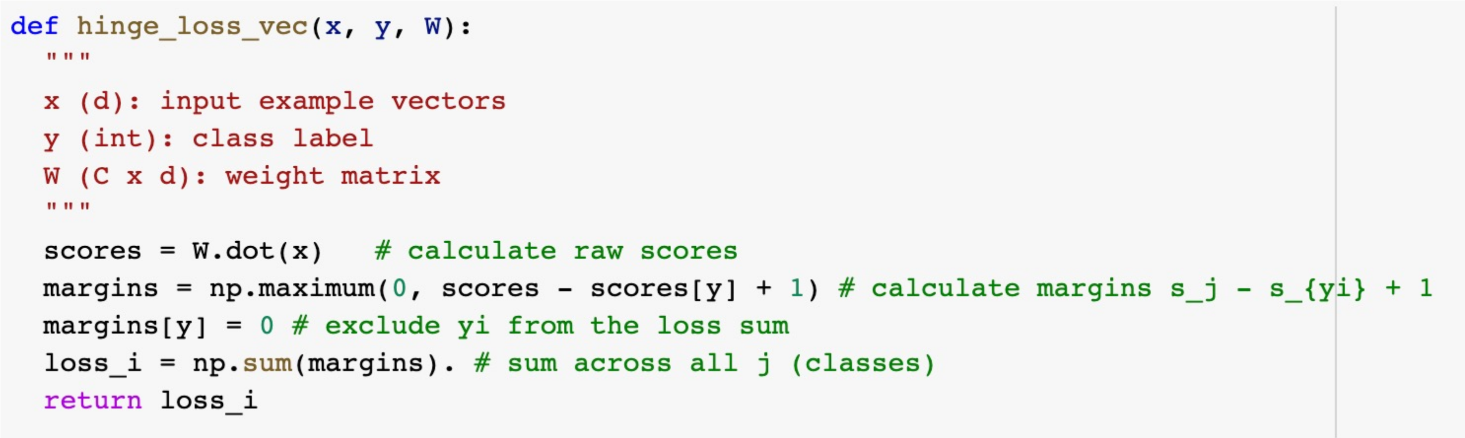
Suppose: 3 training examples, 3 classes.
With some W the scores $f(x,W)=Wx$ are:

cat	**3.2**	1.3	2.2
car	5.1	**4.9**	2.5
frog	-1.7	2.0	**-3.1**

Adapted from from CS 231n slides

Multiclass SVM loss:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$



```
def hinge_loss_vec(x, y, W):
  """
  x (d): input example vectors
  y (int): class label
  W (C x d): weight matrix
  """
  scores = W.dot(x)   # calculate raw scores
  margins = np.maximum(0, scores - scores[y] + 1) # calculate margins s_j - s_{yi} + 1
  margins[y] = 0 # exclude yi from the loss sum
  loss_i = np.sum(margins). # sum across all j (classes)
  return loss_i
```

Adapted from from CS 231n slides

$$f(x, W) = Wx$$

$$L = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq y_i} \max(0, f(x_i; W)_j - f(x_i; W)_{y_i} + 1)$$

E.g. Suppose that we found a W such that L = 0.

Q: Is this W unique?

Let’s look at an example

Adapted from from CS 231n slides

SVM Loss Example

Multiclass SVM loss:

Suppose: 3 training examples, 3 classes.
With some $\boldsymbol{W}$ the scores $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{W})=\boldsymbol{W}\boldsymbol{x}$ are:

cat	**3.2**	1.3	2.2
car	5.1	**4.9**	2.5
frog	-1.7	2.0	**-3.1**

Before:
= max(0, 1.3 - 4.9 + 1)
+max(0, 2.0 - 4.9 + 1)
= max(0, -2.6) + max(0, -1.9)
= 0 + 0
= 0

With W **twice as large:**
= max(0, 2.6 - 9.8 + 1)
+max(0, 4.0 - 9.8 + 1)
= max(0, -6.2) + max(0, -4.8)
= 0 + 0
= 0

Adapted from from CS 231n slides

$$f(x, W) = Wx$$

$$L = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq y_i} \max(0, f(x_i; W)_j - f(x_i; W)_{y_i} + 1)$$

E.g. Suppose that we found a W such that L = 0.

Q: Is this W unique?

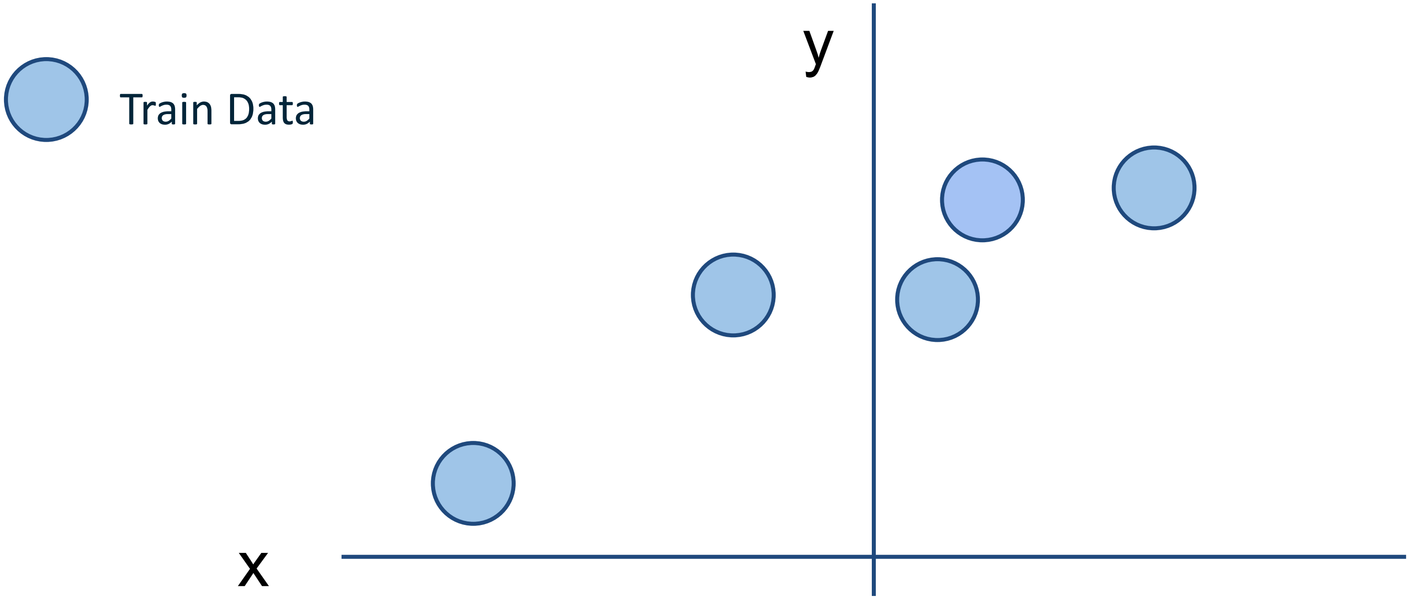
No, 2W also has L=0

How do we choose between W, 2W, and 1e+7W?

Adapted from from CS 231n slides

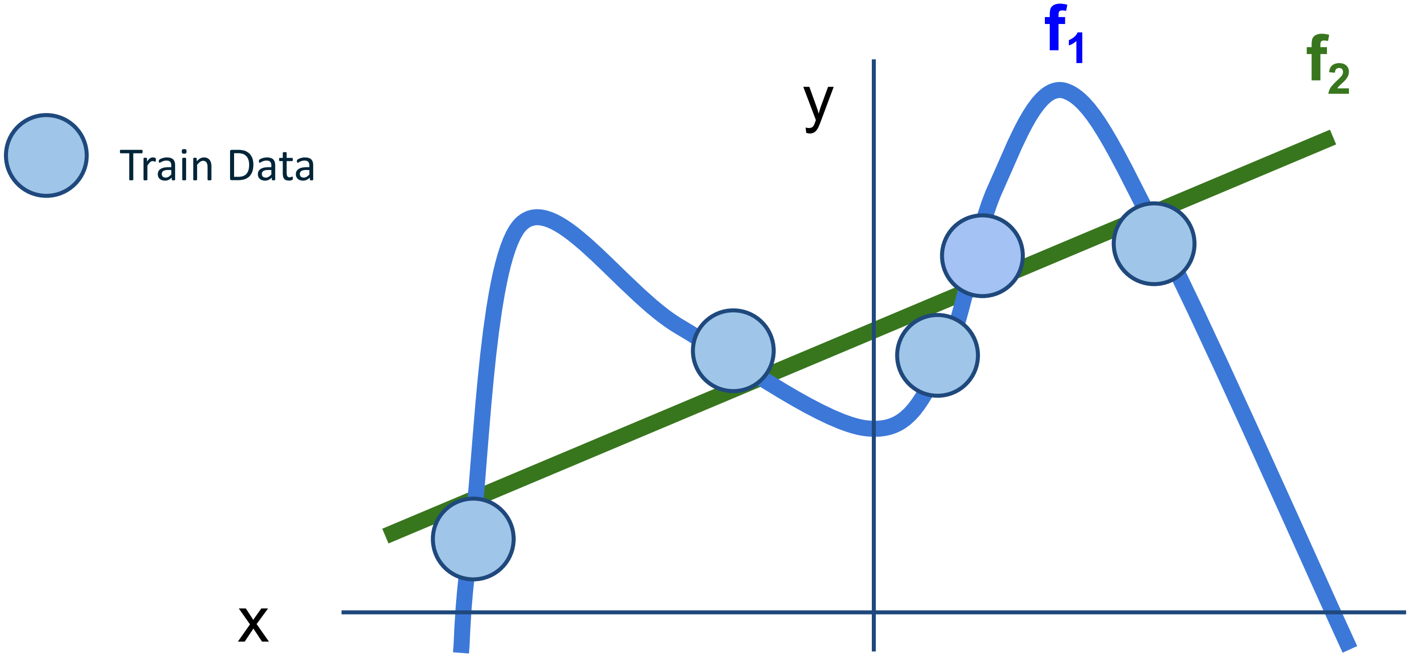
SVM Loss Example

Regularization intuition: fitting a polynomial function



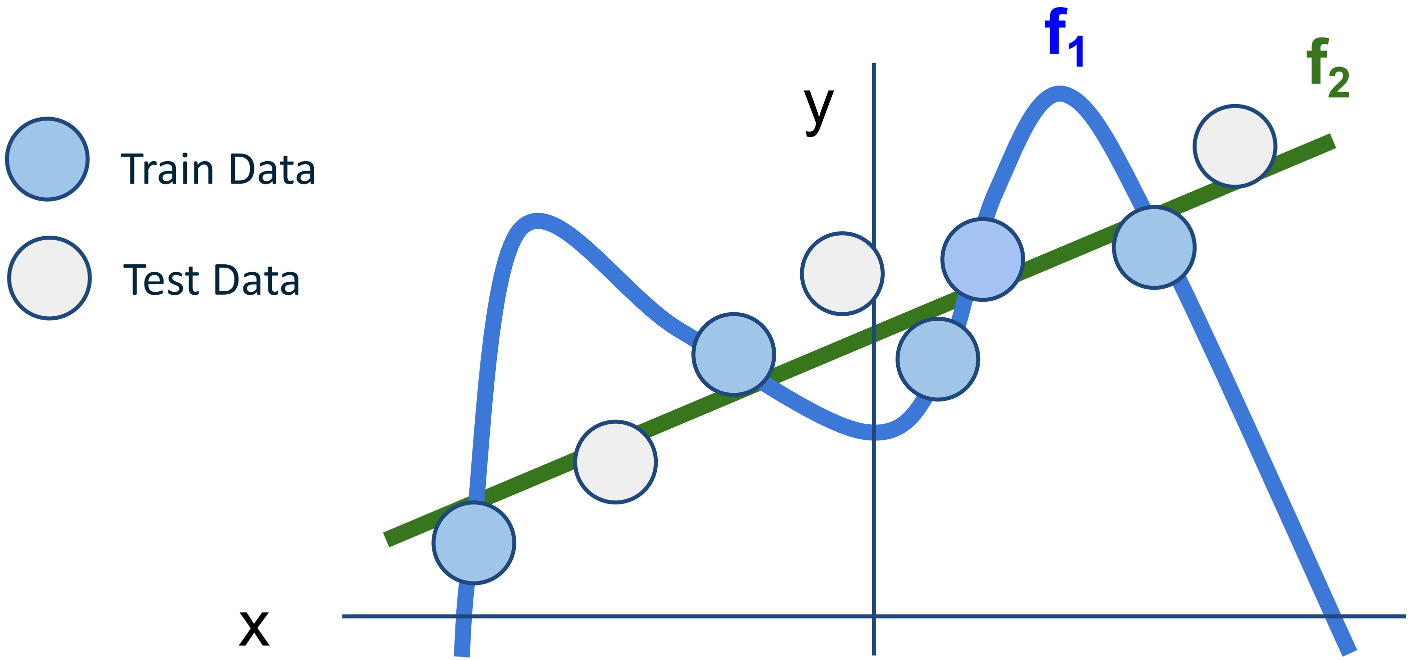
Adapted from from CS 231n slides

Regularization intuition: fitting a polynomial function



Adapted from from CS 231n slides

Regularization intuition: fitting a polynomial function



Regularization balances the simplicity of the function and loss, so we don't overfit to the noises in the data

Adapted from from CS 231n slides

Regularization

$$L(W) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} L_i(f(x_i, W), y_i)} + \underbrace{\lambda R(W)}$$

Data loss: Model predictions should match training data

Regularization: Prevent the model from doing *too* well on training data

Adapted from from CS 231n slides

Regularization

λ = regularization strength (hyperparameter)

$$L(W) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} L_i(f(x_i, W), y_i)} + \underbrace{\lambda R(W)}$$

Data loss: Model predictions should match training data

Regularization: Prevent the model from doing *too* well on training data

Simple examples

L2 regularization: $R(W) = \sum_k \sum_l W_{k,l}^2$

L1 regularization: $R(W) = \sum_k \sum_l |W_{k,l}|$

Elastic net (L1 + L2): $R(W) = \sum_k \sum_l \beta W_{k,l}^2 + |W_{k,l}|$

More complex (DNN-specific):

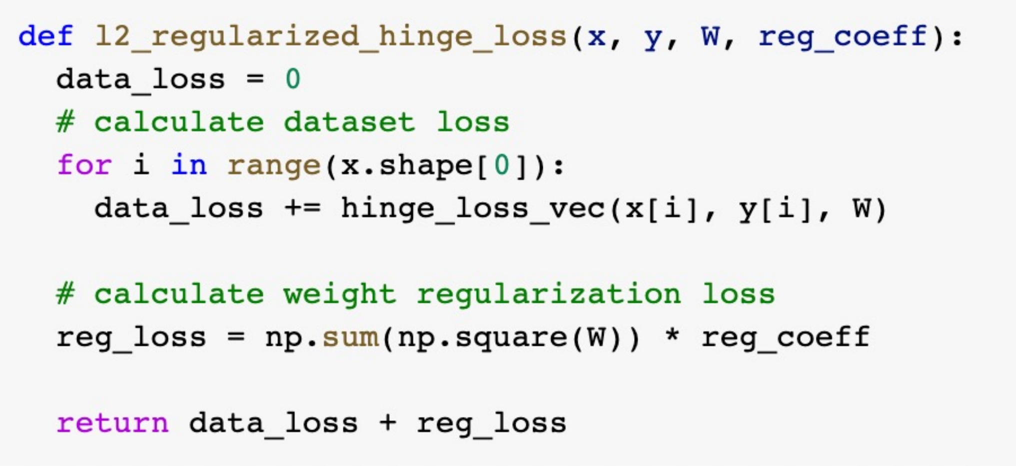
Dropout

Batch/layer normalization

Stochastic depth, fractional pooling, etc

Regularization: Implement a simple L2 regularizer

$$L(W) = \frac{1}{N} \sum_{i=1}^{N} L_i(f(x_i, W), y_i) + \lambda R(W)$$



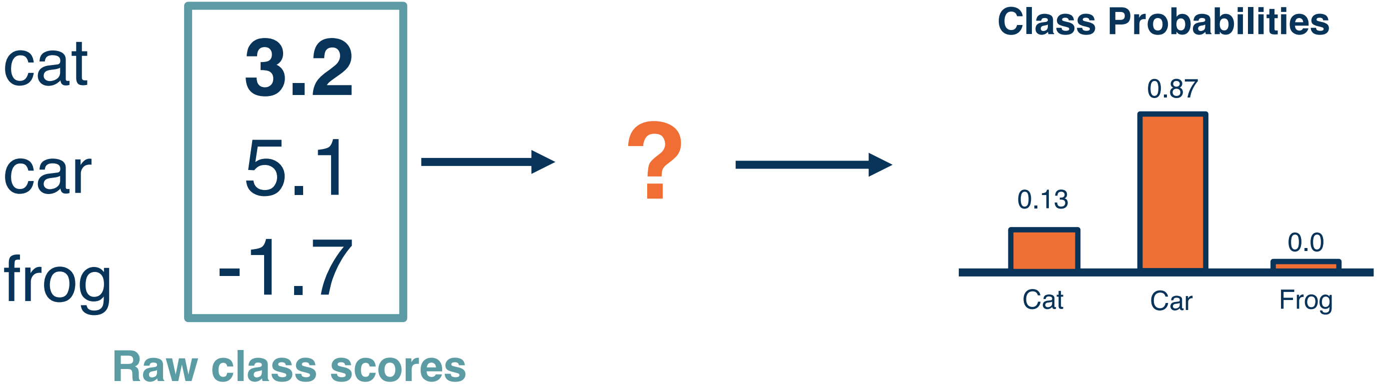
```
def l2_regularized_hinge_loss(x, y, W, reg_coeff):
  data_loss = 0
  # calculate dataset loss
  for i in range(x.shape[0]):
    data_loss += hinge_loss_vec(x[i], y[i], W)

  # calculate weight regularization loss
  reg_loss = np.sum(np.square(W)) * reg_coeff

  return data_loss + reg_loss
```

What if we want probabilities?

We need a different classifier!*



cat	**3.2**
car	5.1
frog	-1.7

Raw class scores

→ ? →

*Technically we can get probability from SVM classifiers too, see [Platt scaling]()

Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as **probabilities**

$$\boxed{s = f(x_i; \theta)}$$

Probabilities must be >= 0

$$\boxed{p_\theta(Y = y_i | X = x_i) = \frac{e^{s_{y_i}}}{\sum_j e^{s_j}}}$$

Softmax Function

Probabilities must sum to 1

	Unnormalized log-probabilities / logits	$\xrightarrow{\text{exp}}$	Unnormalized probabilities	$\xrightarrow{\text{normalize}}$	Probabilities
cat	**3.2**		**24.5**		**0.13**
car	5.1		164.0		0.87
frog	-1.7		0.18		0.00

How do we compute the loss?

Adapted from from CS 231n slides

Softmax Classifier (Multinomial Logistic Regression)

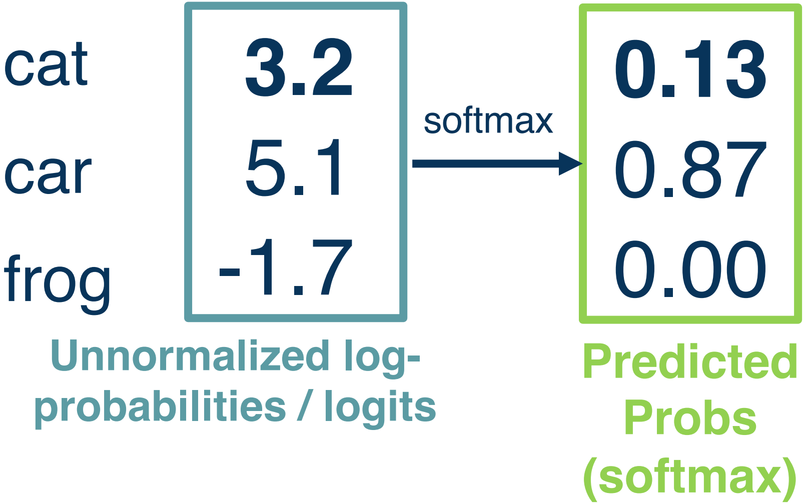
Want to interpret raw classifier scores as **probabilities**

$$\boxed{s = f(x_i; \theta)}$$

$$\boxed{p_\theta(Y = y_i | X = x_i) = \frac{e^{s_{y_i}}}{\sum_j e^{s_j}}}$$

Softmax Function

We maximize the probability of $p_\theta(y_i|x_i)$!



cat 3.2 → 0.13
car 5.1 → 0.87
frog -1.7 → 0.00

softmax

Unnormalized log-probabilities / logits

Predicted Probs (softmax)

Finding a set of weights θ that maximizes the probability of correct prediction: $\underset{\theta}{\operatorname{argmax}} \prod p_\theta(y_i|x_i)$

This is equivalent to:

$$\underset{\theta}{\operatorname{argmax}} \sum \ln p_\theta(y_i|x_i)$$

$$L_i = -\ln p_\theta(y_i|x_i) = -\ln\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right) = -\ln(0.13)$$

1. Maximum Likelihood Estimation (MLE): Choose weights to maximize the likelihood of observed data. In this case, the loss function is the **Negative Log-Likelihood (NLL)**.

Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as **probabilities**

$$\boxed{s = f(x_i;\theta)}$$

$$\boxed{p_\theta(Y = y_i | X = x_i) = \frac{e^{s_{y_i}}}{\sum_j e^{s_j}}}$$

Softmax Function

2. Information theory view

	Unnormalized log-probabilities / logits	softmax →	**Predicted Probs (softmax)**	→ maximize agreement ←	**Correct probs**
cat	**3.2**		**0.13**		**1.00**
car	5.1		0.87		0.00
frog	-1.7		0.00		0.00

Adapted from from CS 231n slides

Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as **probabilities**

$$\boxed{s = f(x_i;\theta)}$$

$$\boxed{p_\theta(Y = y_i|X = x_i) = \frac{e^{s_{y_i}}}{\sum_j e^{s_j}}}$$

Softmax Function

2. Information theory view

	Unnormalized log-probabilities / logits	softmax →	Predicted Probs (softmax)	→ maximize agreement ←	Correct probs
cat	**3.2**		**0.13**		**1.00**
car	5.1		0.87		0.00
frog	-1.7		0.00		0.00

Cross Entropy: $H(p,q) = -\sum p(x)\ln q(x)$

Cross Entropy Loss -> NLL

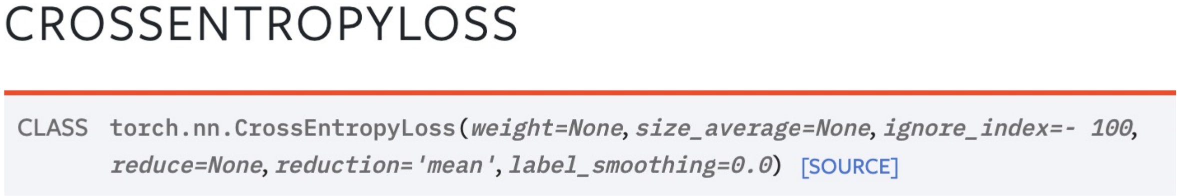
$$H_i(p,p_\theta) = -\sum_{y\in Y} p(y|x_i)\ln p_\theta(y|x_i)$$
$$= -\ln p_\theta(y_i|x_i)$$

$$L = \sum H_i(p,p_\theta) = -\sum \ln p_\theta(y_i|x_i) \equiv NLL$$

Adapted from from CS 231n slides

Softmax Classifier (Multinomial Logistic Regression)

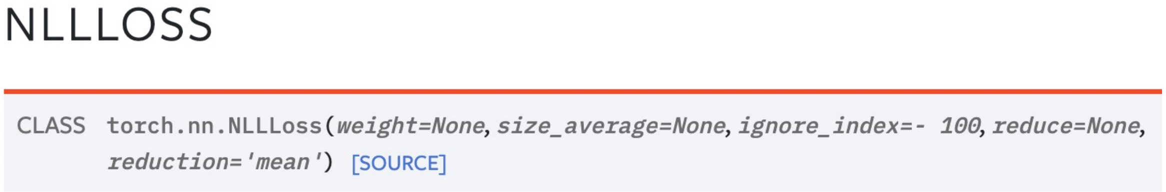
NLL and CrossEntropy are different loss functions in PyTorch!



CROSSENTROPYLOSS

CLASS `torch.nn.CrossEntropyLoss(`*`weight=None`*`,` *`size_average=None`*`,` *`ignore_index=- 100`*`,` *`reduce=None`*`,` *`reduction='mean'`*`,` *`label_smoothing=0.0`*`)` [SOURCE]

Expects unformalized logits as input (the function will apply softmax & log on top)



NLLLOSS

CLASS `torch.nn.NLLLoss(`*`weight=None`*`,` *`size_average=None`*`,` *`ignore_index=- 100`*`,` *`reduce=None`*`,` *`reduction='mean'`*`)` [SOURCE]

Expects log probabilities as input (do it yourself!)

Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as **probabilities**

$$\boxed{\boldsymbol{s = f(x_i; \theta)}}$$

$$\boxed{p_\theta(\boldsymbol{Y = y_i | X = x_i}) = \frac{\boldsymbol{e^{s_{y_i}}}}{\boldsymbol{\sum_j e^{s_j}}}}$$

Softmax Function

Cross-entropy loss:

$$L_i = -\log(p_\theta(y_i | x_i))$$

Q: What is the min/max of possible loss L_i?

Infimum is 0, max is unbounded (inf)

Adapted from from CS 231n slides

Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as **probabilities**

$$\boxed{\boldsymbol{s = f(x_i;\theta)}}$$

$$\boxed{p_\theta(\boldsymbol{Y = y_i | X = x_i}) = \frac{\boldsymbol{e^{s_{y_i}}}}{\boldsymbol{\sum_j e^{s_j}}}}$$

Softmax Function

Cross-entropy loss:

$$L_i = -\log(p_\theta(y_i | x_i))$$

Q: At initialization all s will be approximately equal; what is the loss?

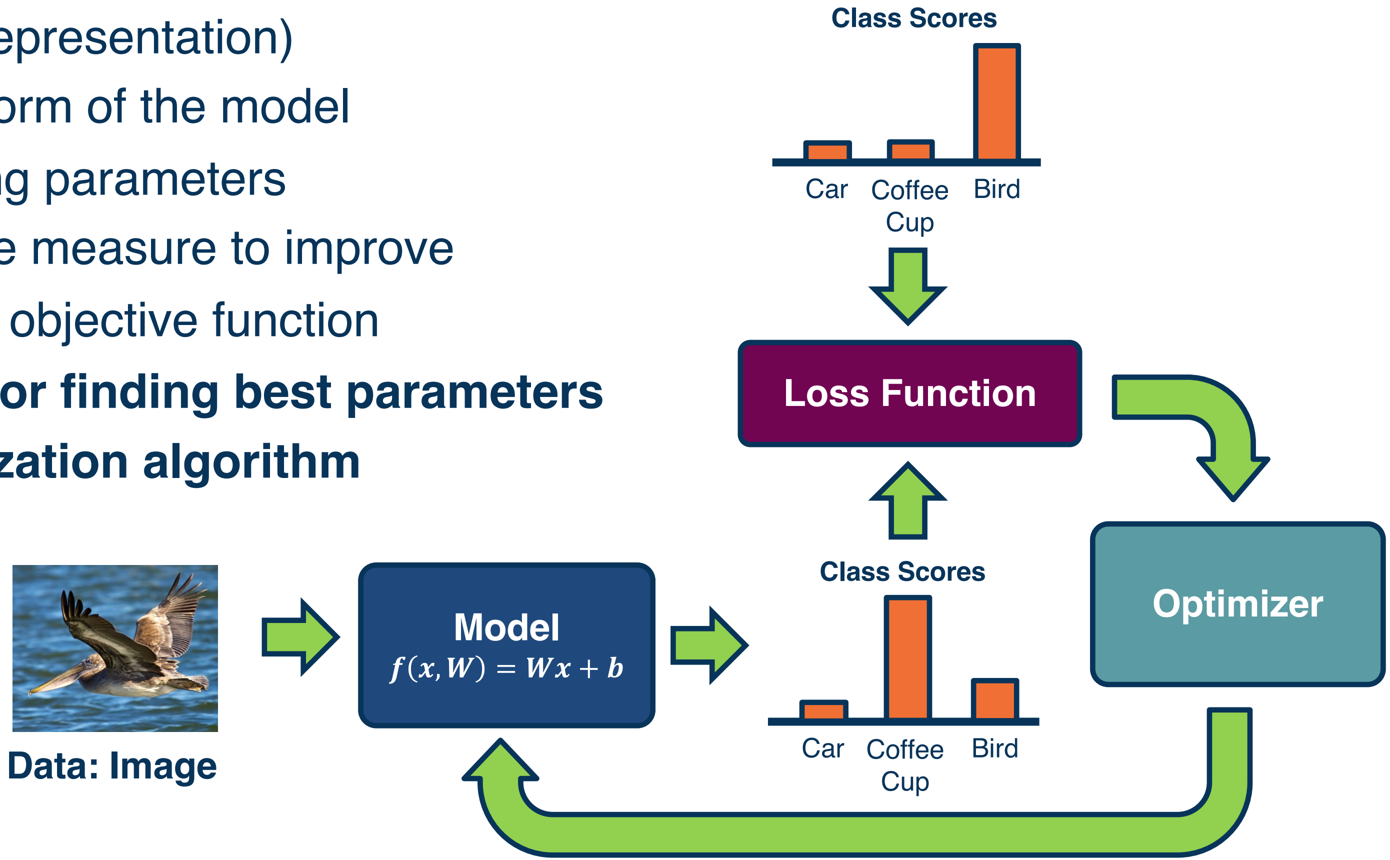
Log(C), e.g. log(3) ≈ 1.1

Adapted from from CS 231n slides

Loss functions: SVM and Softmax Classifier

- Loss function: performance measure to improve
  - Find weights that better satisfies the objective
- Multiclass SVM Classifier
  - Predicts class score
  - Hinge loss: “maximum margin” objective: $L_i = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)$
- Regularization
  - Prevent overly complex function that only works well on the training set
- Softmax Classifier
  - Predicts class probabilities
  - NLL and Cross Entropy Loss

- Input (and representation)
- Functional form of the model
  - Including parameters
- Performance measure to improve
  - Loss or objective function
- **Algorithm for finding best parameters**
  - **Optimization algorithm**



Class Scores

Car

Coffee Cup

Bird

Loss Function

Data: Image

Model

$f(x, W) = Wx + b$

Class Scores

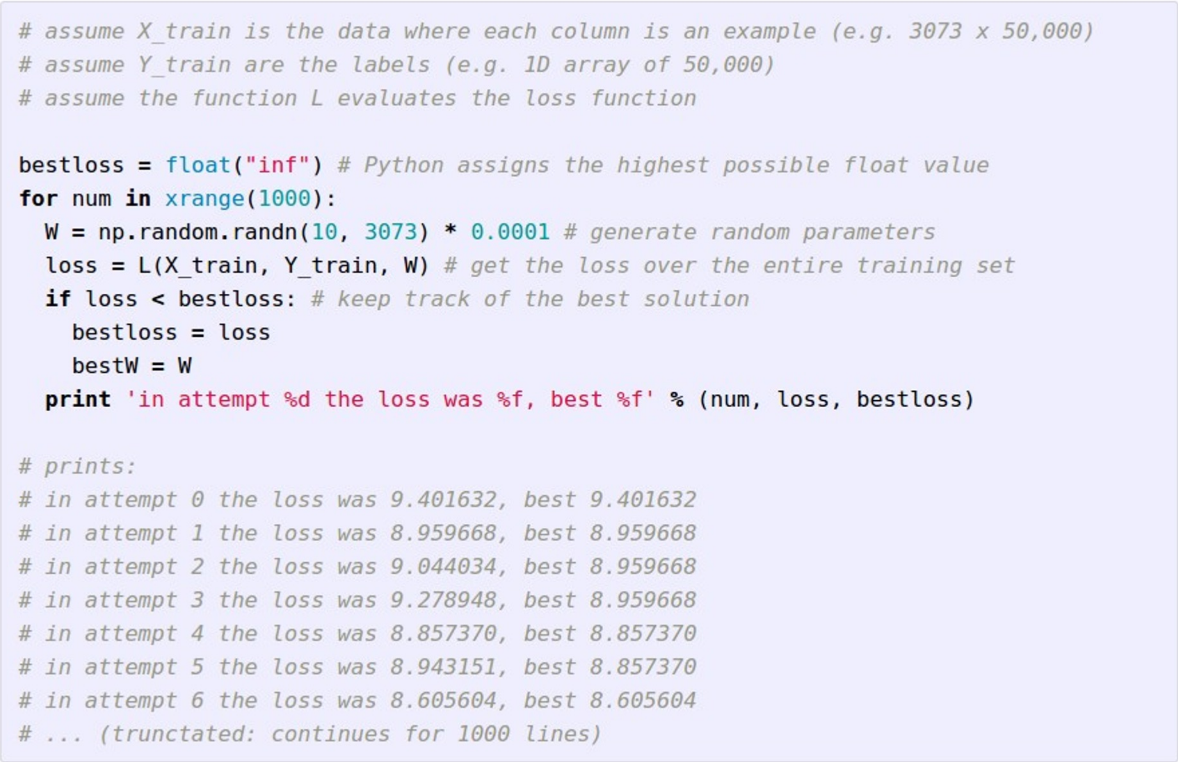
Car

Coffee Cup

Bird

Optimizer

Strategy #1: A first very bad idea solution: **Random search**

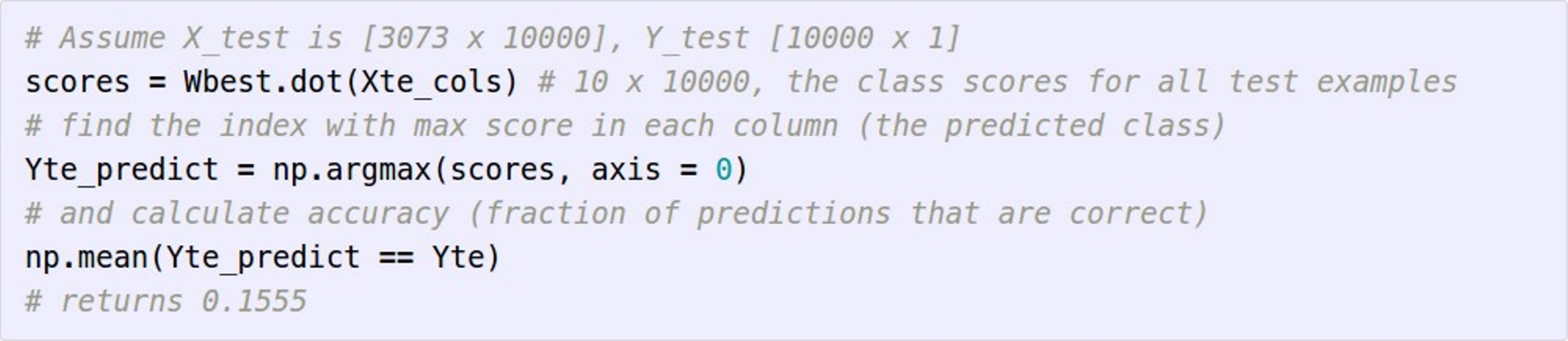


```
# assume X_train is the data where each column is an example (e.g. 3073 x 50,000)
# assume Y_train are the labels (e.g. 1D array of 50,000)
# assume the function L evaluates the loss function

bestloss = float("inf") # Python assigns the highest possible float value
for num in xrange(1000):
  W = np.random.randn(10, 3073) * 0.0001 # generate random parameters
  loss = L(X_train, Y_train, W) # get the loss over the entire training set
  if loss < bestloss: # keep track of the best solution
    bestloss = loss
    bestW = W
  print 'in attempt %d the loss was %f, best %f' % (num, loss, bestloss)

# prints:
# in attempt 0 the loss was 9.401632, best 9.401632
# in attempt 1 the loss was 8.959668, best 8.959668
# in attempt 2 the loss was 9.044034, best 8.959668
# in attempt 3 the loss was 9.278948, best 8.959668
# in attempt 4 the loss was 8.857370, best 8.857370
# in attempt 5 the loss was 8.943151, best 8.857370
# in attempt 6 the loss was 8.605604, best 8.605604
# ... (trunctated: continues for 1000 lines)
```

Lets see how well this works on the test set...



```
# Assume X_test is [3073 x 10000], Y_test [10000 x 1]
scores = Wbest.dot(Xte_cols) # 10 x 10000, the class scores for all test examples
# find the index with max score in each column (the predicted class)
Yte_predict = np.argmax(scores, axis = 0)
# and calculate accuracy (fraction of predictions that are correct)
np.mean(Yte_predict == Yte)
# returns 0.1555
```

15.5% accuracy! not bad!
(SOTA is ~99.7%)

Adapted from from CS 231n slides

Optimization

Given a model and loss function, finding the best set of weights is a **search problem**

- Find the best combination of weights that minimizes our loss function

Several classes of methods:

- Random search
- Genetic algorithms (population-based search)
- Gradient-based optimization

In deep learning, **gradient-based methods are dominant** although not the only approach possible

$$\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} & b1 \\ w_{21} & w_{22} & \cdots & w_{2m} & b2 \\ w_{21} & w_{22} & \cdots & w_{3m} & b3 \end{bmatrix}$$

Gradient

Loss

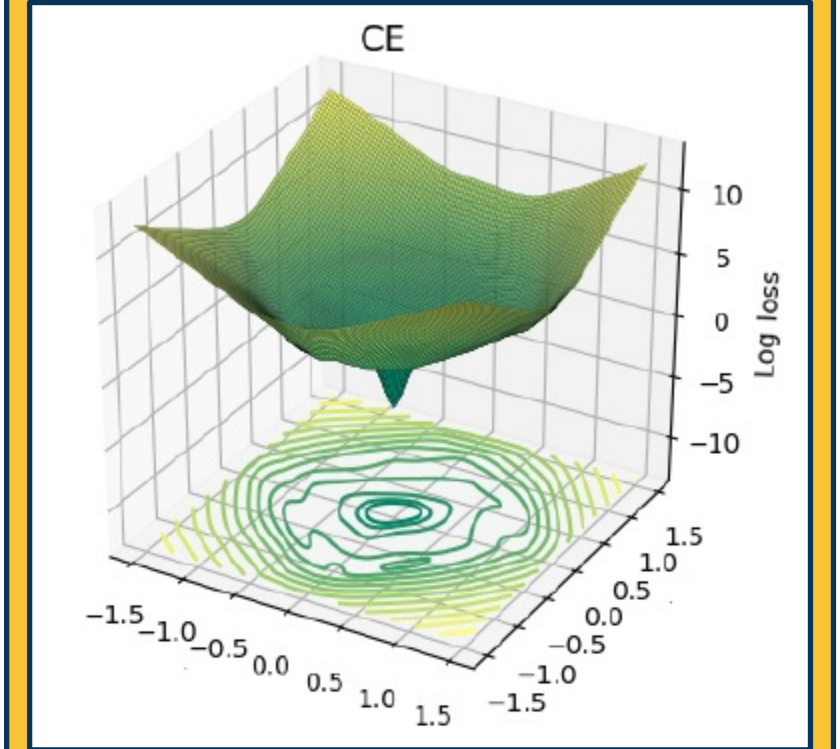
1. Calculate the gradients of a loss function with respect to a set of parameters (w's).
2. Update the parameters towards the gradient direction that minimizes the loss.

Gradient Descent: Follow the Slope!

As weights change, the gradients change as well

- This is often somewhat-smooth locally, so small changes in weights produce small changes in the loss

We can therefore think about **iterative algorithms** that take **current values of weights and modify them** a bit



- We can find the steepest descent direction by computing the **derivative:**

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

- **Gradient** is multi-dimensional derivatives
- Steepest descent direction is the **negative gradient**
- **Intuitively:** Measures how the function changes as the argument a changes by a small step size
- **In Machine Learning:** Want to know how to minimize loss by changing parameters
  - Can consider each parameter separately by taking **partial derivative** of loss function with respect to that parameter

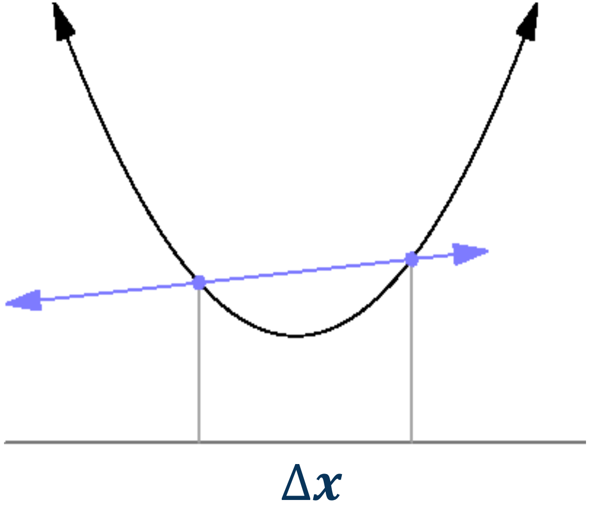


Image and equation from: https://en.wikipedia.org/wiki/Derivative #/media/File:Tangent_animation.gif

Derivatives

Calculate gradients: finite differences

current W:

[0.34,
-1.11,
0.78,
0.12,
0.55,
2.81,
-3.1,
-1.5,
0.33,…]
loss 1.25347

gradient dW:

[?,
?,
?,
?,
?,
?,
?,
?,
?,…]

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Calculate gradients: finite differences

current W:	**W + h** (first dim)**:**	**gradient dW:**
[0.34,	[0.34 + **0.0001**,	[?,
-1.11,	-1.11,	?,
0.78,	0.78,	?,
0.12,	0.12,	?,
0.55,	0.55,	?,
2.81,	2.81,	?,
-3.1,	-3.1,	?,
-1.5,	-1.5,	?,
0.33,…]	0.33,…]	?,…]
loss 1.25347	**loss 1.25322**	

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Calculate gradients: finite differences

current W:

[0.34,
-1.11,
0.78,
0.12,
0.55,
2.81,
-3.1,
-1.5,
0.33,…]
loss 1.25347

W + h (first dim)**:**

[0.34 + **0.0001**,
-1.11,
0.78,
0.12,
0.55,
2.81,
-3.1,
-1.5,
0.33,…]
loss 1.25322

gradient dW:

[**-2.5**,
?,
?,

(1.25322 - 1.25347)/0.0001
= -2.5

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

?,
?,…]

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Calculate gradients: finite differences

current W:	**W + h** (second dim)**:**	**gradient dW:**
[0.34,	[0.34,	[-2.5,
-1.11,	-1.11 + **0.0001**,	?,
0.78,	0.78,	?,
0.12,	0.12,	?,
0.55,	0.55,	?,
2.81,	2.81,	?,
-3.1,	-3.1,	?,
-1.5,	-1.5,	?,
0.33,…]	0.33,…]	?,…]
loss 1.25347	**loss 1.25353**	

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Calculate gradients: finite differences

current W:

[0.34,
-1.11,
0.78,
0.12,
0.55,
2.81,
-3.1,
-1.5,
0.33,…]
loss 1.25347

W + h (second dim):

[0.34,
-1.11 + **0.0001**,
0.78,
0.12,
0.55,
2.81,
-3.1,
-1.5,
0.33,…]
loss 1.25353

gradient dW:

[-2.5,
0.6,
?,
?,
?,…]

(1.25353 - 1.25347)/0.0001
= 0.6

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Calculate gradients: finite differences

current W:	**W + h** (third dim)**:**	**gradient dW:**
[0.34,	[0.34,	[-2.5,
-1.11,	-1.11,	0.6,
0.78,	0.78 + **0.0001**,	?,
0.12,	0.12,	?,
0.55,	0.55,	?,
2.81,	2.81,	?,
-3.1,	-3.1,	?,
-1.5,	-1.5,	?,
0.33,…]	0.33,…]	?,…]
loss 1.25347	**loss 1.25347**	

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Calculate gradients: finite differences

current W:

[0.34,
-1.11,
0.78,
0.12,
0.55,
2.81,
-3.1,
-1.5,
0.33,…]
loss 1.25347

W + h (third dim):

[0.34,
-1.11,
0.78 + **0.0001**,
0.12,
0.55,
2.81,
-3.1,
-1.5,
0.33,…]
loss 1.25347

gradient dW:

[-2.5,
0.6,
0,
?,
?,…]

(1.25347 - 1.25347)/0.0001
= 0

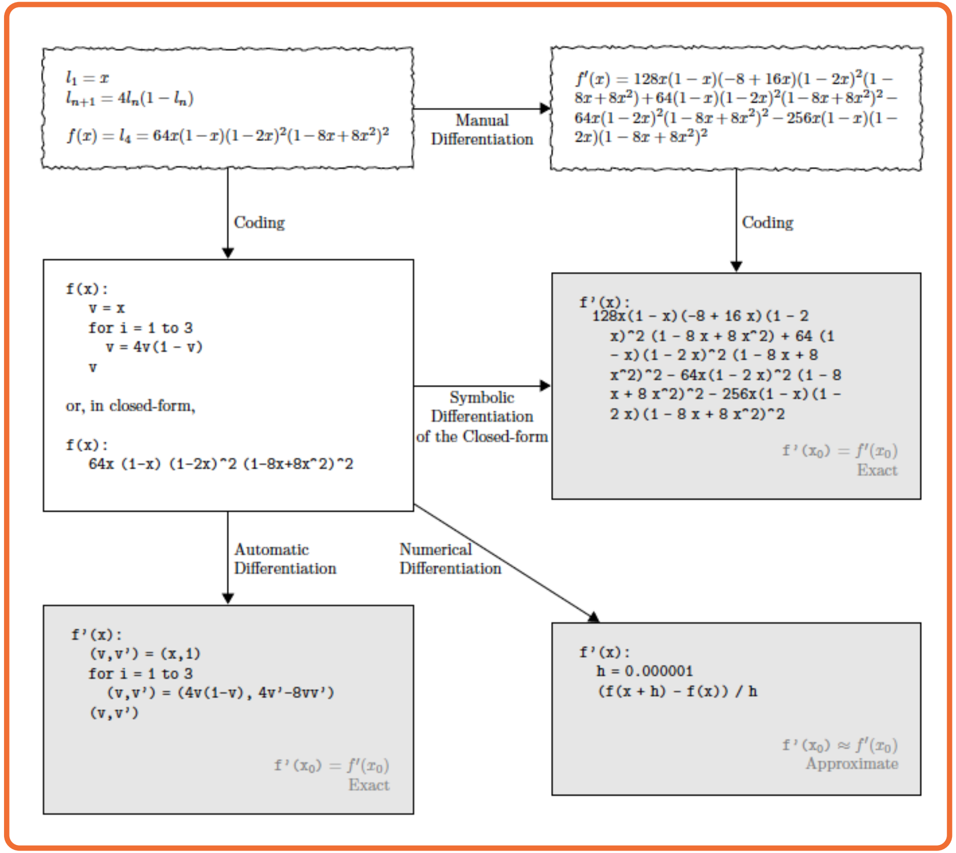
$$\boxed{\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}}$$

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Several ways to compute $\frac{\partial L}{\partial w_i}$

- Manual differentiation
- Symbolic differentiation
- Numerical differentiation
- **Automatic differentiation**

More on **autodiff**:
https://www.cs.toronto.edu/~rgrosse/courses/csc421_2019/readings/L06%20Automatic%20Differentiation.pdf



$l_1 = x$
$l_{n+1} = 4l_n(1 - l_n)$

$f(x) = l_4 = 64x(1-x)(1-2x)^2(1-8x+8x^2)^2$

$f'(x) = 128x(1-x)(-8+16x)(1-2x)^2(1-8x+8x^2) + 64(1-x)(1-2x)^2(1-8x+8x^2)^2 - 64x(1-2x)^2(1-8x+8x^2)^2 - 256x(1-x)(1-2x)(1-8x+8x^2)^2$

```
f(x):
  v = x
  for i = 1 to 3
    v = 4v(1 - v)
  v
```

or, in closed-form,

```
f(x):
  64x (1-x) (1-2x)^2 (1-8x+8x^2)^2
```

```
f'(x):
 128x(1 - x)(-8 + 16 x)(1 - 2
   x)^2 (1 - 8 x + 8 x^2) + 64 (1
   - x)(1 - 2 x)^2 (1 - 8 x + 8
   x^2)^2 - 64x(1 - 2 x)^2 (1 - 8
   x + 8 x^2)^2 - 256x(1 - x)(1 -
   2 x)(1 - 8 x + 8 x^2)^2
```

$f'(x_0) = f'(x_0)$
Exact

```
f'(x):
  (v,v') = (x,1)
  for i = 1 to 3
    (v,v') = (4v(1-v), 4v'-8vv')
  (v,v')
```

$f'(x_0) = f'(x_0)$
Exact

```
f'(x):
  h = 0.000001
  (f(x + h) - f(x)) / h
```

$f'(x_0) \approx f'(x_0)$
Approximate

Numerical vs Analytic Gradients

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

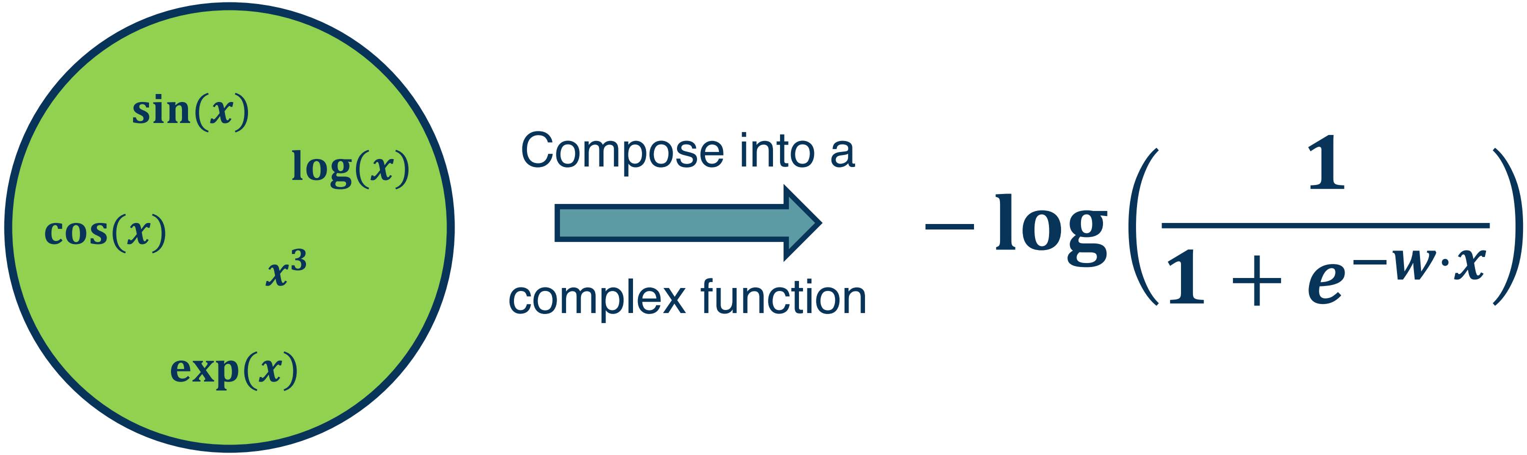
Numerical gradient: slow :(, approximate :(, easy to write :)
Analytic gradient: fast :), exact :), error-prone :(

Almost all differentiable functions that you can think of have analytical gradients implemented in popular libraries, e.g., PyTorch, TensorFlow.

If you want to derive your own gradients, check your implementation with numerical gradient.
This is called a **gradient check.**

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Composing simple functions creates complex analytical gradients



$\sin(x)$

$\log(x)$

$\cos(x)$

x^3

$\exp(x)$

Compose into a complex function

$$-\log\left(\frac{1}{1+e^{-w\cdot x}}\right)$$

$w \cdot x$

u

$\frac{1}{1+e^{-u}}$

p

$-\log(p)$

L

Adapted from slides by: Marc'Aurelio Ranzato, Yann LeCun

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial p}\frac{\partial p}{\partial u}\frac{\partial u}{\partial w}$$

Next time: Chain rule and Backpropagation!

Adapted from slides by: Marc'Aurelio Ranzato, Yann LeCun

Decomposing a Function

The gradient descent algorithm

- 1. Choose a model: $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{W}) = \mathrm{Wx}$
- 2. Choose loss function: $\boldsymbol{L_i} = |\boldsymbol{y} - \boldsymbol{W}\boldsymbol{x_i}|^2$
- 3. Calculate partial derivative for each parameter: $\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{w_i}}$
- 4. Update the parameters: $\boldsymbol{w_i} = \boldsymbol{w_i} - \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{w_i}}$
- 5. Add learning rate to prevent too big of a step: $\boldsymbol{w_i} = \boldsymbol{w_i} - \boldsymbol{\alpha}\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{w_i}}$
- **Repeat 3-5**

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial p}\frac{\partial p}{\partial u}\frac{\partial u}{\partial w}$$

Next time: Chain rule and Backpropagation!

Linear Algebra View: Vector and Matrix Sizes

Conventions:

- Size of derivatives for scalars, vectors, and matrices: Assume we have scalar $\boldsymbol{s} \in \mathbb{R}^{\mathbf{1}}$, vector $\boldsymbol{v} \in \mathbb{R}^{\boldsymbol{m}}$, i.e. $\boldsymbol{v} = [\boldsymbol{v}_{\mathbf{1}}, \boldsymbol{v}_{\mathbf{2}}, \dots, \boldsymbol{v}_{\boldsymbol{m}}]^{T}$ and matrix $\boldsymbol{M} \in \mathbb{R}^{\boldsymbol{k} \times \boldsymbol{\ell}}$
- What is the size of $\frac{\partial v}{\partial s}$? $\mathbb{R}^{m \times 1}$ (column vector of size m)
- What is the size of $\frac{\partial s}{\partial v}$? $\mathbb{R}^{1 \times m}$ (row vector of size m)

$$\begin{bmatrix} \frac{\partial v_1}{\partial s} \\ \frac{\partial v_2}{\partial s} \\ \vdots \\ \frac{\partial v_m}{\partial s} \end{bmatrix}$$

$$\begin{bmatrix} \frac{\partial s}{\partial v_1} & \frac{\partial s}{\partial v_1} & \dots & \frac{\partial s}{\partial v_m} \end{bmatrix}$$

Dimensionality of Derivatives

Georgia Tech

Conventions:

- What is the size of $\frac{\partial v^1}{\partial v^2}$? A matrix:

$$\begin{array}{rc}
 & \qquad\qquad\qquad \textbf{Col } j \\
\textbf{Row } i & \begin{bmatrix}
\frac{\partial v^1_1}{\partial v^2_1} & \cdots & \cdots & \cdots & \cdots & \\
\cdots & & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \frac{\partial v^1_i}{\partial v^2_j} & \cdots & \cdots & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \\
\cdots & \cdots & \cdots & \cdots & \cdots &
\end{bmatrix}
\end{array}$$

- This matrix of partial derivatives is called a **Jacobian**

(Note this is slightly different convention than on Wikipedia)

Conventions:

- What is the size of $\frac{\partial s}{\partial M}$? A matrix:

$$\begin{bmatrix} \frac{\partial s}{\partial m_{[1,1]}} & \cdots & \cdots & \cdots & \cdots & \\ \cdots & & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \frac{\partial s}{\partial m_{[i,j]}} & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \end{bmatrix}$$

- What is the size of $\frac{\partial L}{\partial W}$?
  - Remember that loss is a **scalar** and W is a matrix:

$$\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} & b1 \\ w_{21} & w_{22} & \cdots & w_{2m} & b2 \\ w_{31} & w_{32} & \cdots & w_{3m} & b3 \end{bmatrix}$$

$$W$$

Jacobian is also a matrix:

$$\begin{bmatrix} \frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{12}} & \cdots & \frac{\partial L}{\partial w_{1m}} & \frac{\partial L}{\partial b_1} \\ \frac{\partial L}{\partial w_{21}} & \cdots & \cdots & \frac{\partial L}{\partial w_{2m}} & \frac{\partial L}{\partial b_2} \\ \cdots & \cdots & \cdots & \frac{\partial L}{\partial w_{3m}} & \frac{\partial L}{\partial b_3} \end{bmatrix}$$

Dimensionality of Derivatives

Jacobians of Batches

Batches of data are **matrices** or **tensors** (multi-dimensional matrices)

Examples:

- Each instance is a vector of size $\boldsymbol{m}$, our batch is of size $[\boldsymbol{B} \times \boldsymbol{m}]$
- Each instance is a matrix (e.g. grayscale image) of size $\boldsymbol{W} \times \boldsymbol{H}$, our batch is $[\boldsymbol{B} \times \boldsymbol{W} \times \boldsymbol{H}]$
- Each instance is a multi-channel matrix (e.g. color image with R,B,G channels) of size $\boldsymbol{C} \times \boldsymbol{W} \times \boldsymbol{H}$, our batch is $[\boldsymbol{B} \times \boldsymbol{C} \times \boldsymbol{W} \times \boldsymbol{H}]$

Jacobians become tensors which is complicated

- Instead, flatten input to a vector and get a vector of derivatives!
- This can also be done for partial derivatives between two vectors, two matrices, or two tensors

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}$$

Flatten ⇩

$$\begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{21} \\ x_{22} \\ \vdots \\ x_{n1} \\ \vdots \\ x_{nn} \end{bmatrix}$$