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

Supervision in form of reward

No supervision on what action to take

Types of Machine Learning

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.

Recap:

Very often combined, sometimes within the same model!
Recap:

Types of Errors and Generalization

Model class

Estimation Error

Optimization Error

Modeling Error

Reality

horse person
Recap:

**Algebraic Viewpoint**

\[ f(x, W) = Wx \]

**Visual Viewpoint**

One template per class

**Geometric Viewpoint**

Hyperplanes cutting up space

Adapted from CS 231n slides
This time:

\[ f(x, W) = Wx \]

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)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>airplane</td>
<td>-3.45</td>
<td>-0.51</td>
<td>3.42</td>
</tr>
<tr>
<td>automobile</td>
<td>-8.87</td>
<td>6.04</td>
<td>4.64</td>
</tr>
<tr>
<td>bird</td>
<td>0.09</td>
<td>5.31</td>
<td>2.65</td>
</tr>
<tr>
<td>cat</td>
<td>2.9</td>
<td>-4.22</td>
<td>5.1</td>
</tr>
<tr>
<td>deer</td>
<td>4.48</td>
<td>-4.19</td>
<td>2.64</td>
</tr>
<tr>
<td>dog</td>
<td>8.02</td>
<td>3.58</td>
<td>5.55</td>
</tr>
<tr>
<td>frog</td>
<td>3.78</td>
<td>4.49</td>
<td>-4.34</td>
</tr>
<tr>
<td>horse</td>
<td>1.06</td>
<td>-4.37</td>
<td>-1.5</td>
</tr>
<tr>
<td>ship</td>
<td>-0.36</td>
<td>-2.09</td>
<td>-4.79</td>
</tr>
<tr>
<td>truck</td>
<td>-0.72</td>
<td>-2.93</td>
<td>6.14</td>
</tr>
</tbody>
</table>
Suppose: 3 training examples, 3 classes. With some $W$ the scores $f(x, W) = Wx$ are:

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th>car</th>
<th>frog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.2</td>
<td>5.1</td>
<td>-1.7</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>4.9</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>2.2</td>
<td>2.5</td>
<td>-3.1</td>
</tr>
</tbody>
</table>

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:

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

Adapted from from CS 231n slides
Multiclass SVM loss:

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

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

the SVM loss has the form:

\[
L_i = \sum_{j \neq y_i} \begin{cases} 
0 & \text{if } s_{y_i} \geq s_j + 1 \\
(s_j - s_{y_i} + 1) & \text{otherwise} 
\end{cases}
= \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)
\]

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

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

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

“Hinge Loss”
Multiclass SVM loss:

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>3.2</td>
</tr>
<tr>
<td>car</td>
<td>5.1</td>
</tr>
<tr>
<td>frog</td>
<td>-1.7</td>
</tr>
</tbody>
</table>

Losses: 2.9

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:

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

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

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

$$= 2.9 + 0$$

$$= 2.9$$

Adapted from from CS 231n slides
Suppose: 3 training examples, 3 classes. With some $W$ the scores $f(x, W) = Wx$ are:

<table>
<thead>
<tr>
<th>Class</th>
<th>Score 1</th>
<th>Score 2</th>
<th>Score 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>3.2</td>
<td>1.3</td>
<td>2.2</td>
</tr>
<tr>
<td>car</td>
<td>5.1</td>
<td>4.9</td>
<td>2.5</td>
</tr>
<tr>
<td>frog</td>
<td>-1.7</td>
<td>2.0</td>
<td>-3.1</td>
</tr>
</tbody>
</table>

Losses: 2.9

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:

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

$$= \max(0, 1.3 - 4.9 + 1) + \max(0, 2.0 - 4.9 + 1)$$

$$= \max(0, -2.6) + \max(0, -1.9)$$

$$= 0 + 0$$

$$= 0$$
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:

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

\[
L = \frac{(2.9 + 0 + 12.9)}{3} = 5.27
\]

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

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th>car</th>
<th>frog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.2</td>
<td>5.1</td>
<td>-1.7</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>4.9</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>2.2</td>
<td>2.5</td>
<td>-3.1</td>
</tr>
</tbody>
</table>

Losses: 2.9 0 12.9

Adapted from CS 231n slides
Multiclass SVM loss:

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

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

No change for small values

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

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th>frog</th>
<th>car</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores</td>
<td>3.2</td>
<td>-1.7</td>
<td>5.1</td>
</tr>
<tr>
<td>Correct</td>
<td>1.3</td>
<td>2.0</td>
<td>4.9</td>
</tr>
<tr>
<td>Loss</td>
<td>2.2</td>
<td>-3.1</td>
<td>2.5</td>
</tr>
</tbody>
</table>
Suppose: 3 training examples, 3 classes. With some $W$ the scores $f(x,W)=Wx$ are:

<table>
<thead>
<tr>
<th>Class</th>
<th>value1</th>
<th>value2</th>
<th>value3</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>3.2</td>
<td>1.3</td>
<td>2.2</td>
</tr>
<tr>
<td>car</td>
<td>5.1</td>
<td>4.9</td>
<td>2.5</td>
</tr>
<tr>
<td>frog</td>
<td>-1.7</td>
<td>2.0</td>
<td>-3.1</td>
</tr>
</tbody>
</table>

**Q: What is min/max of loss value?**

$[0, \infty]$
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

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

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

```
num_class - 1
```
Suppose: 3 training examples, 3 classes. With some $W$ the scores $f(x,W) = Wx$ are:

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th>car</th>
<th>frog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.2</td>
<td>1.3</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>5.1</td>
<td>4.9</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>-1.7</td>
<td>2.0</td>
<td>-3.1</td>
</tr>
</tbody>
</table>

Q: What if we used mean instead of sum?

No difference

Scaling by constant

Multiclass SVM loss:

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

Adapted from CS 231n slides
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:

<table>
<thead>
<tr>
<th>Class</th>
<th>SVM Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>3.2</td>
</tr>
<tr>
<td>car</td>
<td>5.1</td>
</tr>
<tr>
<td>frog</td>
<td>-1.7</td>
</tr>
</tbody>
</table>

SVM Loss Example

Adapted from from CS 231n slides
Multiclass SVM loss:

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

```python
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
Let's look at an example

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

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

E.g. Suppose that we found a $W$ such that $L = 0$.
**Q: Is this $W$ unique?**

Let’s look at an example
Multiclass SVM loss:

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

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th></th>
<th></th>
<th>frog</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>1.3</td>
<td>2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>4.9</td>
<td>2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1.7</td>
<td>2.0</td>
<td>-3.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 CS 231n slides
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 $1e+7W$?
Regularization intuition: fitting a polynomial function

Train Data

Adapted from CS 231n slides
Regularization intuition: fitting a polynomial function

Adapted 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 CS 231n slides
Regularization

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

**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

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

- **Data loss**: Model predictions should match training data
- **Regularization**: Prevent the model from doing too well on training data

\[ \lambda \text{ = regularization strength (hyperparameter)} \]

**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

[Georgia Tech logo]
Regularization: Implement a simple L2 regularizer

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

```python
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!*

*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**

\[
s = f(x_i; \theta)
\]

Probabilities must be \(\geq 0\)

\[
p_\theta(Y = y_i | X = x_i) = \frac{e^{s_{y_i}}}{\sum_{j} e^{s_{j}}}
\]

Probabilities must sum to 1

<table>
<thead>
<tr>
<th>Class</th>
<th>Unnormalized log-probabilities / logits</th>
<th>Unnormalized probabilities</th>
<th>Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>3.2</td>
<td>24.5</td>
<td>0.13</td>
</tr>
<tr>
<td>car</td>
<td>5.1</td>
<td>164.0</td>
<td>0.87</td>
</tr>
<tr>
<td>frog</td>
<td>-1.7</td>
<td>0.18</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Unnormalized log-probabilities / logits

**Cross-Entropy Loss Example**

Adapted from CS 231n slides

How do we compute the loss?

Probabilities must be \(\geq 0\)

Probabilities must sum to 1
Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as probabilities

\[ s = f(x_i; \theta) \]

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

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

<table>
<thead>
<tr>
<th></th>
<th>Cat</th>
<th>Frog</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unnormalized log-probabilities / logits</td>
<td>3.2</td>
<td>-1.7</td>
<td>5.1</td>
</tr>
<tr>
<td>Predicted Probs (softmax)</td>
<td>0.13</td>
<td>0.00</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Finding a set of weights \( \theta \) that maximizes the probability of correct prediction: \( \arg \max_\theta \prod p_\theta(y_i | x_i) \)

This is equivalent to:

\[ \arg \max_\theta \sum \ln p_\theta(y_i | x_i) \]

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

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 (Multinominal Logistic Regression)

Want to interpret raw classifier scores as probabilities

\[ s = f(x_i; \theta) \]

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

1. Information theory view

2. Information theory view

Unnormalized log-probabilities / logits

Predicted Probs (softmax)

Correct probs

Adapted from CS 231n slides
Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as probabilities

\[ s = f(x_i; \theta) \]

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

2. Information theory view

Cross Entropy:

\[ H(p, q) = - \sum p(x) \ln q(x) \]

Cross Entropy Loss \( \rightarrow \) NLL

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

\[ = - \ln p_\theta(y_i|x_i) \]

\[ L = \sum H_i(p, p_\theta) = - \sum \ln p_\theta(y_i|x_i) \equiv \text{NLL} \]

Adapted from CS 231n slides

Cross-Entropy Loss Example
Softmax Classifier (Multinomial Logistic Regression)

NLL and CrossEntropy are different loss functions in PyTorch!

CROSSENTROPYLOSS

```python
CLASS torch.nn.CrossEntropyLoss(weight=None, size_average=None, ignore_index=- 1, reduce=None, reduction='mean', label_smoothing=0.0) [SOURCE]
```

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

NLLLOSS

```python
CLASS torch.nn.NLLLoss(weight=None, size_average=None, ignore_index=- 1, 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

\[ s = f(x_i; \theta) \]

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

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 CS 231n slides
Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as probabilities

\[ s = f(x_i; \theta) \]

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

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. \( \log(3) \approx 1.1 \)

Adapted 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

**Components of a Parametric Model**

Data: Image

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

Class Scores

Loss Function

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

```python
# 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
# ... (truncated: 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
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

\[
\begin{bmatrix}
  w_{11} & w_{12} & \cdots & w_{1m} & b_1 \\
  w_{21} & w_{22} & \cdots & w_{2m} & b_2 \\
  w_{21} & w_{22} & \cdots & w_{3m} & b_3
\end{bmatrix}
\]

**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

Calculate gradients: finite differences

<table>
<thead>
<tr>
<th>current W:</th>
<th>gradient dW:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34,</td>
<td>[?,</td>
</tr>
<tr>
<td>-1.11,</td>
<td>?,</td>
</tr>
<tr>
<td>0.78,</td>
<td>?,</td>
</tr>
<tr>
<td>0.12,</td>
<td>?,</td>
</tr>
<tr>
<td>0.55,</td>
<td>?,</td>
</tr>
<tr>
<td>2.81,</td>
<td>?,</td>
</tr>
<tr>
<td>-3.1,</td>
<td>?,</td>
</tr>
<tr>
<td>-1.5,</td>
<td>?,</td>
</tr>
<tr>
<td>0.33,...]</td>
<td>?,...</td>
</tr>
<tr>
<td>loss 1.25347</td>
<td></td>
</tr>
</tbody>
</table>
## Calculate gradients: finite differences

<table>
<thead>
<tr>
<th>current $W$:</th>
<th>$W + h$ (first dim):</th>
<th>gradient $dW$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33, ...]</td>
<td>[0.34 + 0.0001, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33, ...]</td>
<td>[?, ?, ?, ?, ?, ?, ?, ?, ...]</td>
</tr>
</tbody>
</table>

**loss 1.25347**

**loss 1.25322**

---

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
Calculate gradients: finite differences

**current W:**

\[
[0.34, \\
-1.11, \\
0.78, \\
0.12, \\
0.55, \\
2.81, \\
-3.1, \\
-1.5, \\
0.33,...]
\]

**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 1.25347**

**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:**

\[
[-2.5, \\
?, \\
?, \\
?, \\
?, \\
?, \\
?, \\
?, \\
?, \\
?, \\
?, \\
?,...]
\]

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

\[
(1.25322 - 1.25347)/0.0001 = -2.5
\]
Calculate gradients: finite differences

<table>
<thead>
<tr>
<th>current $W$:</th>
<th>$W + h$ (second dim):</th>
<th>gradient $dW$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[0.34, -1.11 + 0.0001, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[-2.5, ?, ?, ?, ?, ?, ?, ?, ?,...]</td>
</tr>
</tbody>
</table>

loss 1.25347

loss 1.25353
Calculate gradients: finite differences

<table>
<thead>
<tr>
<th>current $W$:</th>
<th>$W + h$ (second dim):</th>
<th>gradient $dW$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[0.34, -1.11 + 0.0001, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[-2.5, 0.6, ?, ?, ?,...]</td>
</tr>
</tbody>
</table>
Calculate gradients: finite differences

<table>
<thead>
<tr>
<th>current $W$:</th>
<th>$W + h$ (third dim):</th>
<th>gradient $dW$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[0.34, -1.11, 0.78 + 0.0001, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[-2.5, 0.6, ?, ?, ?, ?, ?, ?, ?,...]</td>
</tr>
</tbody>
</table>

loss 1.25347  

loss 1.25347

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
Calculate gradients: finite differences

<table>
<thead>
<tr>
<th>current $W$:</th>
<th>$W + h$ (third dim):</th>
<th>gradient $dW$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[0.34, -1.11, 0.78 + 0.0001, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[-2.5, 0.6, 0, ?, ?, ?,...]</td>
</tr>
</tbody>
</table>
Several ways to compute $\frac{\partial L}{\partial w_i}$

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

More on autodiff:
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**.
Composing simple functions creates complex analytical gradients.

\[ \sin(x) \quad \log(x) \]
\[ \cos(x) \quad x^3 \]
\[ \exp(x) \]

Compose into a complex function:

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

Decomposing a Function

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

Next time: Chain rule and Backpropagation!

\[
\frac{\partial L}{\partial w} = \frac{\partial L}{\partial p} \frac{\partial p}{\partial u} \frac{\partial u}{\partial w}
\]

Adapted from slides by: Marc'Aurelio Ranzato, Yann LeCun
The gradient descent algorithm

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

Repeat 3-5
Decomposing a Function

\[ \frac{\partial L}{\partial w} = \frac{\partial L}{\partial p} \frac{\partial p}{\partial u} \frac{\partial u}{\partial w} \]

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 $s \in \mathbb{R}^1$, vector $v \in \mathbb{R}^m$, i.e. $v = [v_1, v_2, \ldots, v_m]^T$ and matrix $M \in \mathbb{R}^{k \times \ell}$.

- 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 $m$).

$$
\frac{\partial s}{\partial v} = \begin{bmatrix}
\frac{\partial s}{\partial v_1} & \frac{\partial s}{\partial v_2} & \cdots & \frac{\partial s}{\partial v_m}
\end{bmatrix}
$$
Conventions:

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

$$
\begin{bmatrix}
\frac{\partial v^1}{\partial v^1} & \cdots & \cdots & \cdots & \cdots \\
\frac{\partial v^2}{\partial v^1} & \cdots & \cdots & \cdots & \cdots \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
\vdots & \cdots & \cdots & \cdots & \cdots \\
\end{bmatrix}
$$

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

(Note this is slightly different convention than on [Wikipedia](https://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant))
Conventions:

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

$$
\begin{bmatrix}
\frac{\partial s}{\partial m_{[1,1]}} & \cdots & \cdots & \cdots & \cdots \\
\cdots & \frac{\partial s}{\partial m_{[i,j]}} & \cdots & \cdots & \cdots \\
\cdots & \cdots & \frac{\partial s}{\partial m_{[i,j]}} & \cdots & \cdots \\
\cdots & \cdots & \cdots & \frac{\partial s}{\partial m_{[i,j]}} & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots
\end{bmatrix}
$$
What is the size of $\frac{\partial L}{\partial W}$?

Remember that loss is a **scalar** and $W$ is a matrix:

$$
\begin{bmatrix}
w_{11} & \dots & w_{1m} & b_1 \\
w_{21} & \dots & w_{2m} & b_2 \\
w_{31} & \dots & w_{3m} & b_3
\end{bmatrix}
$$

Jacobian is also a matrix:

$$
\begin{bmatrix}
\frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{12}} & \dots & \frac{\partial L}{\partial w_{1m}} & \frac{\partial L}{\partial b_1} \\
\frac{\partial L}{\partial w_{21}} & \frac{\partial L}{\partial w_{22}} & \dots & \frac{\partial L}{\partial w_{2m}} & \frac{\partial L}{\partial b_2} \\
\frac{\partial L}{\partial w_{31}} & \frac{\partial L}{\partial w_{32}} & \dots & \frac{\partial L}{\partial w_{3m}} & \frac{\partial L}{\partial b_3}
\end{bmatrix}
$$
Batches of data are **matrices** or **tensors** (multi-dimensional matrices)

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

**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