LECTURE 7: NON-LINEAR OPTIMIZATION

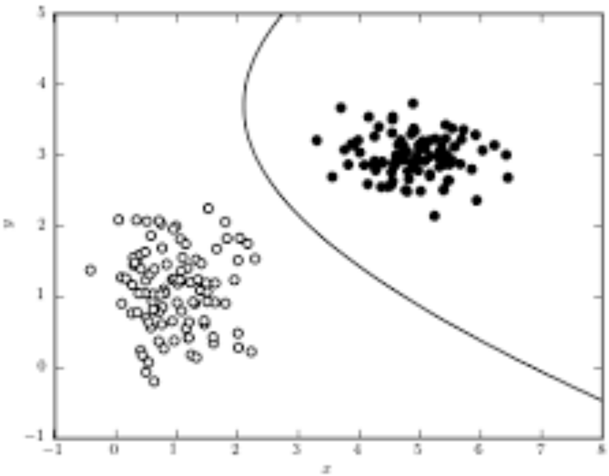
Ehsan Aryafar
earyafar@pdx.edu
http://web.cecs.pdx.edu/~aryafare/ML.html

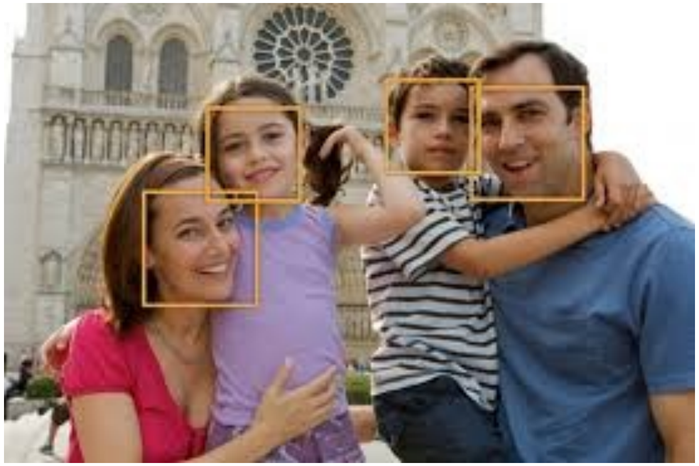
Recall: Classification

- Given features $\boldsymbol{x}$, determine its class label, $y = 1, \dots, K$
- Binary classification: $y = 0$ or 1
- Many applications:
  - Face detection: Is a face present or not?
  - Reading a digit: Is the digit 0,1,…,9?
  - Are the cells cancerous or not?
  - Is the email spam?
- Equivalently, determine classification function:

$$\hat{y} = f(\boldsymbol{x}) \in \{1, \dots, K\}$$

  - Like regression, but with a discrete response
  - May index $\{1, \dots, K\}$ or $\{0, \dots, K-1\}$





Recall: Linear Classifier

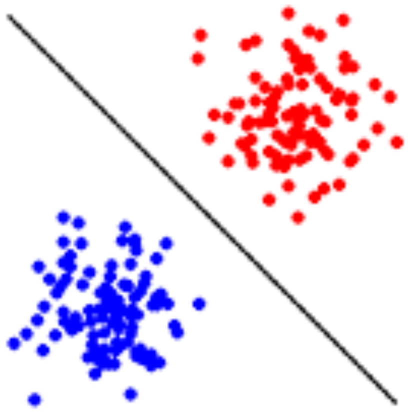
- General binary classification rule:

$$\hat{y} = f(x) = 0 \text{ or } 1$$

- Linear classification rule:
  - Take linear combination $z = w_0 + \sum_{j=1}^{d} w_d x_d$
  - Predict class from z

$$\hat{y} = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$$

- Decision regions described by a half-space.
- $\boldsymbol{w} = (w_0, \dots, w_d)$ is called the weight vector



Recall: Hard vs. Soft Decision Classifiers

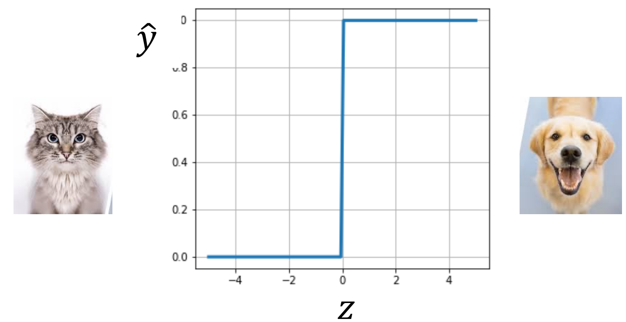
- Binary classification problem:
  - Given features x, estimate class label 0 or 1
  - Ex: cat vs. dog
- Hard decision classifier:
  - Output a class label: $\hat{y} = 0$ or 1
  - Ex: $\hat{y} = 1 \Rightarrow$ *Image is a dog!*
- Soft decision classifier:
  - Output a conditional probability $P(y = 1|x)$
  - $P(y = 1|x)$ is between 0 and 1
  - Ex: $P(y = 1|x) = 0.9 \Rightarrow$ *Given this image, there is a 90% chance it is a dog*

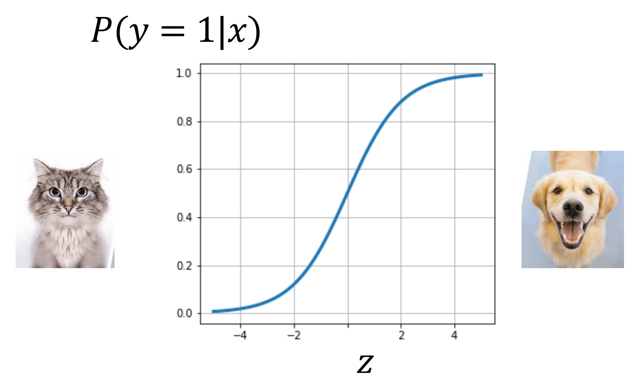
Cat
$y = 0$

Dog
$y = 1$

Recall: Logistic Model for Binary Classification

- Binary classification problem: $y = 0, 1$
- Hard decision linear classifier
  - Predict a class label $\hat{y} = 0$ or 1
  - $z = w_0 + \sum_j w_j x_j$
  - $\hat{y} = \begin{cases} 1 & z > 0 \\ 0 & z \leq 0 \end{cases}$

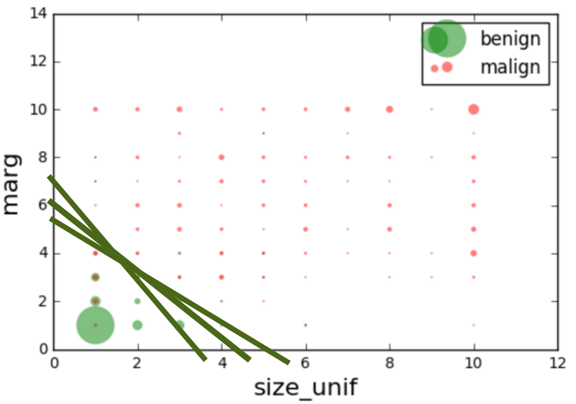




- Logistic soft decision classifier
  - Predict a probability $P(y = 1|x)$
  - $z = w_0 + \sum_j w_j x_j$
  - $P(y = 1|x) = \frac{1}{1+e^{-z}}$
    - Sometime called Sigmoid function

Recall: How To Fit Logistic Models?

- Given training data, $(\boldsymbol{x}_i, y_i), i = 1, \dots, N$
  - Binary labels $y_i \in \{0,1\}$



- Binary logistic model: Given *new* $\boldsymbol{x}$ predict class probability via:
  - Linear weights: $\boldsymbol{z} = \boldsymbol{w}_{1:p}^T \boldsymbol{x} + w_0$
  - Sigmoid: $P(y = 1|\boldsymbol{x}) = \frac{1}{1+e^{-z}}$

- Weight vector $\boldsymbol{w}$ represents unknown model parameters

- Learning problem: Learn weight vector $\boldsymbol{w}$ from the data

Recall: Maximum Likelihood Principle

- Likelihood function: From the logistic model, we can derive:

$P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$ = Probability of class labels given inputs $\boldsymbol{X}$ and weights $\boldsymbol{w}$

  - $(\boldsymbol{X}, \boldsymbol{y})$ are the data matrices for all n training samples
  - $\boldsymbol{w}$ is the vector of parameters

- Key idea: $P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$ is higher ⇒ data is a better match with the parameters

- Maximum Likelihood Principle: Given data $(\boldsymbol{X}, \boldsymbol{y})$:

Find parameters $\boldsymbol{W}$ to maximize $P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{W})$

Recall: Binary Cross Entropy

- Given data $(x_i, y_i), i = 1, \ldots, N$ with binary labels $y_i \in \{0,1\}$
- Theorem: MLE for logistic model is equivalent to minimizing the binary cross entropy:

$$J(\boldsymbol{w}) = \sum_{i=1}^{n} \left(\ln[1 + e^{z_i}] - y_i z_i\right), \qquad z_i = w_0 + \sum_{j=1}^{d} w_j x_{ij}$$

  - Find the weight vector $\boldsymbol{w}$ to minimize $J(\boldsymbol{w})$
  - Will prove below this is equivalent to maximizing $P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$
- Provides a simple cost function to minimize for fitting
- Note that z_i are implicitly function of weights $\boldsymbol{w}$

Learning Objectives

- Identify the objective function, parameters and constraints in an optimization problem
- Compute the gradient of a loss function for scalar, vector and matrix parameters
- Efficiently compute a gradient in python.
- Write the gradient descent update
- Describe the effect of the learning rate on convergence
- Determine if a loss function is convex

Outline

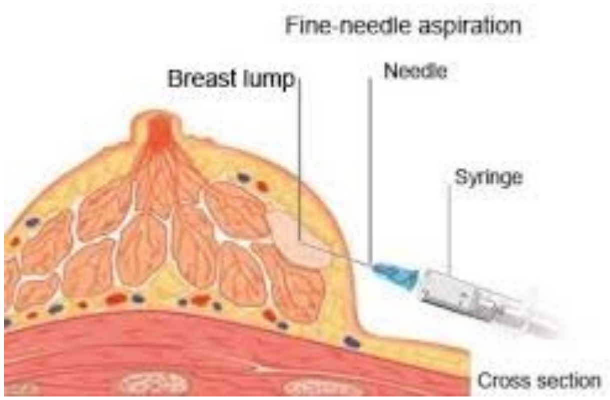
- Motivating example: Build an optimizer for logistic regression
- Gradients of multi-variable functions
- Gradient descent
- Adaptive step size
- Convexity

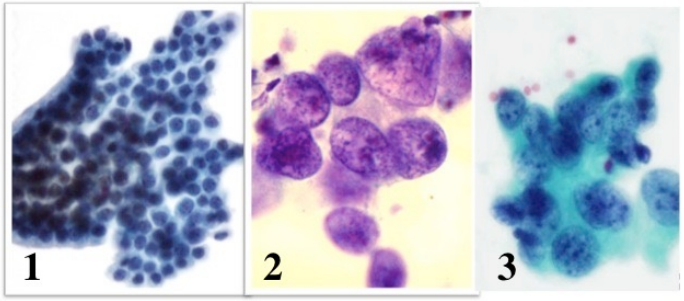
Recap: Breast Cancer Example

- Problem from lecture 6:
  Determine if sample indicates cancer
- Classification problem:
  - Input: $\boldsymbol{x} =$ 10 features of sample (size, cell mitosis, etc..)
  - Output: Is the sample benign or malignant?

$$\hat{y} = \begin{cases} 1 & \text{malignant (cancer)} \\ 0 & \text{benign (no cancer)} \end{cases}$$

- Training data $(\boldsymbol{x}_i, y_i), i = 1, \ldots, N$
  - Data from N =569 patients
- Learn a classification rule from $\boldsymbol{x}$ to y





Grades of carcinoma cells
http://breast-cancer.ca/5a-types/

Logistic Regression Maximum Likelihood

- Logistic model for the likelihood function:

$$P(y = 1|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1 + e^{-z}}, \qquad z = \boldsymbol{w}_{1:p}^{T}\boldsymbol{x} + w_0$$

  - $\boldsymbol{w}$ = unknown weights or parameters
- ML estimation : Minimize the negative log likelihood:

$$\widehat{w} = \arg\min_{w} f(w), \quad f(w) := -\sum_{i=1}^{N} \ln P(y_i|\boldsymbol{x}_i, \boldsymbol{w})$$

  - $f(\boldsymbol{w})$ = loss function = measure of goodness of fit of parameters
- Loss function: binary cross entropy (number of classes K=2)

$$f(\boldsymbol{w}) := \sum_{i=1}^{N} \{\ln[1 + e^{z_i}] - y_i\, z_i\}, \qquad z_i = \boldsymbol{w}_{1:p}^{T}\boldsymbol{x}_i + w_0$$

Minimizing the Loss Function

- No analytic solution to minimize loss
- Used sklearn LogisticRegression.fit method
  - Used built-in optimizer to minimize loss function
  - Very fast and achieves good results
- Questions for today:
  - How does this optimizer work?
  - How would we build one from scratch

```
# Fit on the scaled trained data
reg = linear_model.LogisticRegression(C=1e5)
reg.fit(Xtr1, ytr)
```

```
Accuracy on test data = 0.960976
```

Outline

- Motivating example: Build an optimizer for logistic regression
- Gradients of multi-variable functions
- Gradient descent
- Adaptive step size
- Convexity

Gradients and Optimization

- In machine learning, we often want to minimize a loss function $J(w)$
- Gradient $\nabla J(w)$: Key function
- Gradient has several important properties for optimization
  - Provides a simple linear approximation of a function
  - When at a local minima, $\nabla J(w) = 0$
  - At other points, $-\nabla J(w)$ provides a direction of maximum decrease

Gradient Defined

- Consider scalar-valued function $f(\boldsymbol{w})$
- Vector input $\boldsymbol{w}$. Then gradient is:

$$\nabla_w f(\boldsymbol{w}) = \begin{bmatrix} \partial f(\boldsymbol{w})/\partial w_1 \\ \vdots \\ \partial f(\boldsymbol{w})/\partial w_N \end{bmatrix}$$

- Matrix input $\boldsymbol{W}$, size $M \times N$. Then gradient is:

$$\nabla_w f(\boldsymbol{W}) = \begin{bmatrix} \partial f(\boldsymbol{W})/\partial W_{11} & \cdots & \partial f(\boldsymbol{W})/\partial W_{1N} \\ \vdots & \vdots & \vdots \\ \partial f(\boldsymbol{W})/\partial W_{M1} & \cdots & \partial f(\boldsymbol{W})/\partial W_{MN} \end{bmatrix}$$

- Gradient is same size as the argument!

Example 1

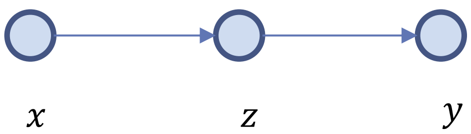
- $f(w_1, w_2) = w_1^2 + 2w_1w_2^3$
- Partial derivatives:
  - $\partial f / \partial w_1 = 2w_1 + 2w_2^3$
  - $\partial f / \partial w_2 = 6w_1w_2^2$
- Gradient: $\nabla f = \begin{bmatrix} 2w_1 + 2w_2^3 \\ 6w_1w_2^2 \end{bmatrix}$
- Example to right:
  - Computes gradient at $w = (2,4)$
  - Gradient is a numpy vector

```
def feval(w):

    # Function
    f = w[0]**2 + 2*w[0]*(w[1]**3)

    # Gradient
    df0 = 2*w[0]+2*(w[1]**3)
    df1 = 6*w[0]*(w[1]**2)
    fgrad = np.array([df0, df1])

    return f, fgrad

# Point to evaluate
w = np.array([2,4])
f, fgrad = feval(w)
```

```
f     = 260.000000
fgrad = [132 192]
```

Chain Rule

- We all know chain rule for scalar functions
- We have a composite function: $y = f(g(x))$
- This is the same as $y = f(z),\ z = g(x)$



- Chain rule says:

$$\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx} = f'(z)g'(x) = f'\big(g(x)\big)g'(x)$$

- Example: $y = \ln(z),\ z = \cos x$
  - Then $\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx} = \frac{1}{z}(-\sin x)$
  - We can leave it like this or substitute $z = \cos x \Rightarrow \frac{dy}{dx} = \frac{1}{\cos x}(-\sin x) = -\tan x$
- Excellent review at Khan Academy

Example 2: An Exponential Model

- Data fitting task:
  - Exponential model: $\hat{y}_i = ae^{-bx_i}$
  - Parameters $w = (a, b)$
  - MSE loss $J(w) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- Problem: Compute gradient ∇J
- Solution:
  - $\frac{\partial J}{\partial a} = \frac{1}{2}\sum_{i=1}^{N}\frac{\partial(y_i - \hat{y}_i)^2}{\partial a}$ [Linearity]

    $= \sum_{i=1}^{N}(\hat{y}_i - y_i)\,\frac{\partial \hat{y}_i}{\partial a}$ [Chain rule]

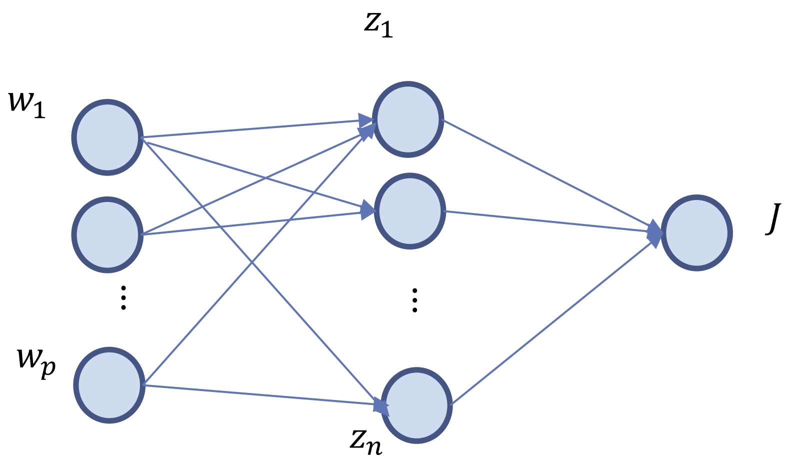
    $= \sum_{i=1}^{N}(\hat{y}_i - y_i)e^{-bx_i}$
  - $\frac{\partial J}{\partial b} = \sum_{i=1}^{N}(\hat{y}_i - y_i)(-ax_i e^{-bx_i})$
  - $\nabla J = \left[\frac{\partial J}{\partial a}, \frac{\partial J}{\partial b}\right]^T$

```
def Jeval(w):

    # Unpack vector
    a = w[0]
    b = w[1]

    # Compute the loss function
    yerr = y-a*np.exp(-b*x)
    J = 0.5*np.sum(yerr**2)

    # Compute the gradient
    dJ_da = -np.sum( yerr*np.exp(-b*x))
    dJ_db = np.sum( yerr*a*x*np.exp(-b*x))
    Jgrad = np.array([dJ_da, dJ_db])
    return J, Jgrad
```

Multi-Variable Chain Rule

- We have a multi-variable composite function:
  - $J = f(z_1, \dots, z_n)$
  - $z_i = g_i(w_1, \dots, w_p)$
- You can visualize the dependencies with a graph
- Multi-variable chain rule:

$$\boxed{\frac{\partial J}{\partial w_j} = \sum_{i=1}^{n} \frac{\partial J}{\partial z_i} \frac{\partial z_i}{\partial w_j}}$$



Example 3: Log-Linear Model

- Given:
  - Data $(\boldsymbol{x}_i, y_i), \quad i = 1, \ldots, N$
  - Model $\hat{y}_i = \log(z_i)\,, \quad z_i = w_0 + \sum_{j=1}^{d} X_{ij} w_j$
  - MSE loss function: $J = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$
- Problem: Find gradient component $\frac{\partial J}{\partial w_j}$
- Solution:
  - Define $A = [1 \;\; X]$, matrix with ones on the first column
  - Then, $z_i = w_0 + \sum_{j=1}^{d} X_{ij} w_j = \sum_{j=0}^{d} A_{ij} w_j$
  - Use multi-variable chain rule:

$$\frac{\partial J}{\partial w_j} = \sum_{i=1}^{N} \frac{\partial J}{\partial z_i} \frac{\partial z_i}{\partial w_j} = \sum_{i=1}^{N} \frac{\partial J}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_i} \frac{\partial z_i}{\partial w_j} = \sum_{i=1}^{N} 2(\hat{y}_i - y_i) \frac{1}{z_i} A_{ij}$$

Example 3: Matrix Version

- From previous slide:
  - $z_i = w_0 + \sum_{j=1}^{d} X_{ij} w_j = \sum_{j=0}^{d} A_{ij} w_j$
  - $\hat{y}_i = \log(z_i)$
  - $\frac{\partial J}{\partial w_j} = 2 \sum_{i=1}^{N} (\hat{y}_i - y_i) \frac{1}{z_i} A_{ij}$
- Can implement these with matrix operations:
  - Useful for efficient implementation in python
  - $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{w}$
  - $\hat{\boldsymbol{y}} = \log(\boldsymbol{z})$
  - $\frac{dJ}{dz} = 2(\hat{\boldsymbol{y}} - \boldsymbol{y}) \frac{1}{\boldsymbol{z}}$ [elementwise division]
  - $\frac{\partial J}{\partial \boldsymbol{w}} = \boldsymbol{A}^T \frac{dJ}{dz}$

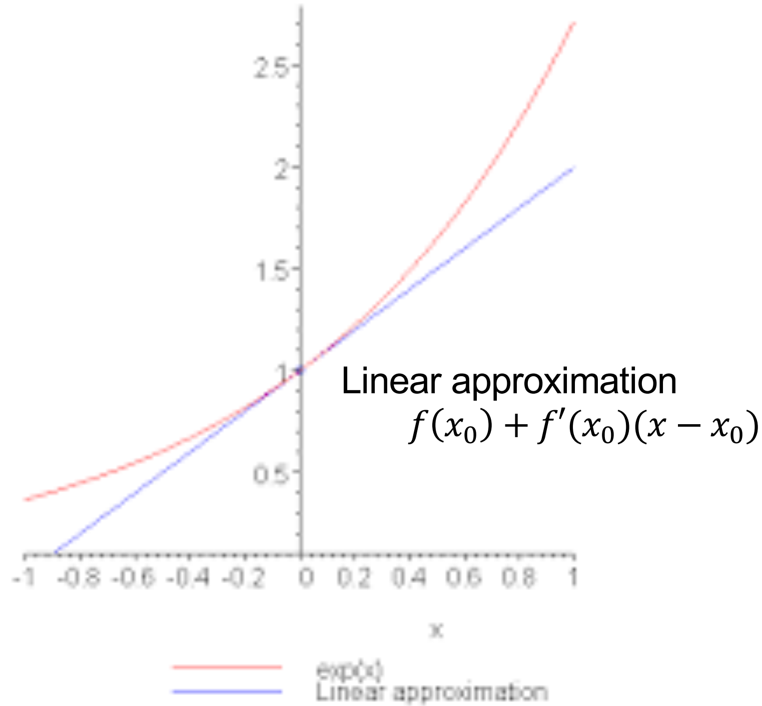
```
def Jeval(w,X,y):

    # Create matrix A=[1 X]
    n = X.shape[0]
    A = np.column_stack((np.ones(n), X))

    # Compute function
    z = A.dot(w)
    yhat = np.log(z)
    J = np.sum((y-yhat)**2)

    # Compute gradient
    dJ_dz = 2*(yhat-y)/z
    Jgrad = A.T.dot(dJ_dz)

    return J, Jgrad
```

First-Order Approximations
Scalar-Input Functions

- Consider function $f(x)$ with scalar input x
- First-order approximation for a scalar input function

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

- Approximates $f(x)$ by a linear function
  - Derivative = $f'(x_0)$ = slope
- What is the equivalent for vector-input functions?



First-Order Approximations

Vector Input Functions

- Suppose $f(\boldsymbol{x})$ takes a vector input $\boldsymbol{x} = (x_1, \dots, x_p)$
- Fix a point $\boldsymbol{x}_0 = (x_{01}, \dots, x_{0p})$
- Then for any other point $\boldsymbol{x} \approx \boldsymbol{x}_0$, gradients can be used for first order approximation

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x_0}) + \sum_{j=1}^{p} \frac{\partial f}{\partial x_j} \left(x_j - x_{0j}\right) = f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T (\boldsymbol{x} - \boldsymbol{x}_0)$$

- Linear function in $\boldsymbol{x}$
- Change in $f(\boldsymbol{x})$ given by inner product:

$$f(\boldsymbol{x}) - f(\boldsymbol{x}_0) \approx \nabla f(\boldsymbol{x}_0)^T (\boldsymbol{x} - \boldsymbol{x}_0) = \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 \rangle$$

Checking Gradients

- Always check gradients before using
  - Even good developers make mistakes!
- Simple check:
  - Take some point w_0
  - Evaluate $J(w_0)$ and $\nabla J(w_0)$
  - Take a second point w_1 close to w_0
  - Evaluate $J(w_1)$
  - Verify that:

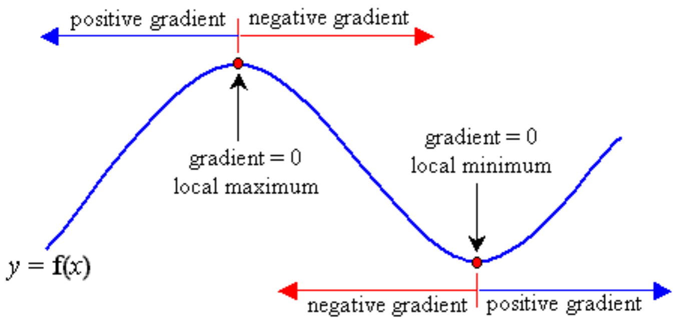
$$J(w_1) - J(w_0) \approx \nabla J(w_0)^T (w_1 - w_0)$$

```python
# Generate random positive data
n = 100
d = 5
X = np.random.uniform(0,1,(n,d))
w0 = np.random.uniform(0,1,(d+1,))
y = np.random.uniform(0,2,(n,))

# Compute function and gradient at point w0
J0, Jgrad0 = Jeval(w0,X,y)

# Take a small perturbation
step = 1e-4
w1 = w0 + step*np.random.normal(0,1,(d+1,))

# Evaluate the function at perturbed point
J1, Jgrad1 = Jeval(w1,X,y)

dJ = J1-J0
dJ_est = Jgrad0.dot(w1-w0)
print('Actual difference:    %12.4e' % dJ)
print('Estimated difference: %12.4e' % dJ_est)
```

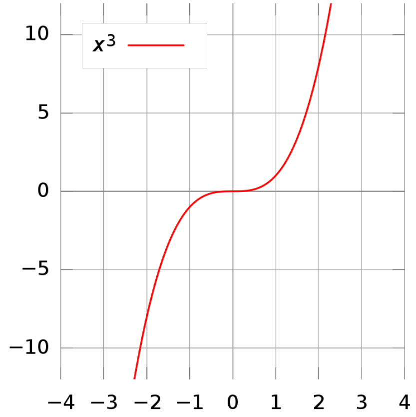
```
Actual difference:      -1.1895e-03
Estimated difference:   -1.1896e-03
```

Gradients and Stationary Points

- Stationary point: Any $\boldsymbol{w}$ where $\nabla f(\boldsymbol{w}) = 0$
- Occurs at any local maxima or minima
- Also, any saddle point
- In linear regression:
  - $f(\boldsymbol{w})$ = RSS loss function
  - Solved for $\boldsymbol{w}$ where $\nabla f(\boldsymbol{w}) = 0$
- But, often cannot explicitly solve for $\nabla f(\boldsymbol{w}) = 0$



Saddle point: a point on the surface of a function where the slopes are zero but is not a local extremum (e.g., x=0 on the right hand side).

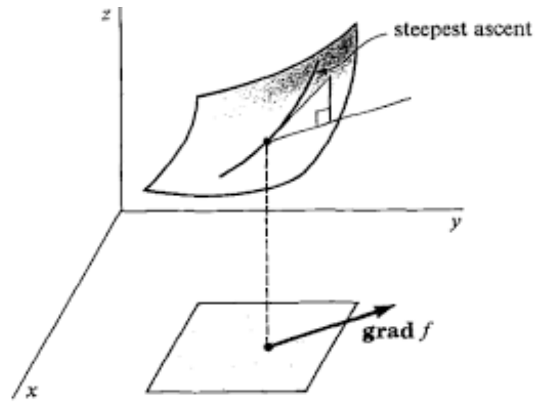


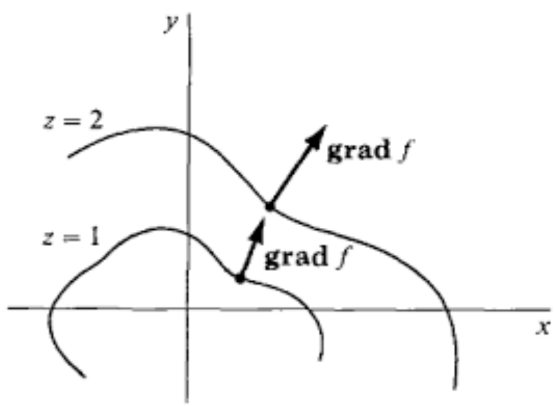
Direction of Maximum Increase

- Gradient indicates direction of maximum increase:
- Take a starting point x_0
- Change in $f(x)$ direction u

$$f(\boldsymbol{x}_0 + \boldsymbol{u}) - f(\boldsymbol{x}_0) \approx \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{u} \rangle$$
$$= \|\nabla f(\boldsymbol{x}_0)\| \|\boldsymbol{u}\| \cos\theta$$

  - Maximum increase when $\boldsymbol{u} = \alpha\, \nabla f(\boldsymbol{x}_0)$
  - Maximum decrease when $\boldsymbol{u} = -\alpha\, \nabla f(\boldsymbol{x}_0)$





First-Order Approximations
Matrix Input Functions (Advanced Concept)

- Suppose $f(\boldsymbol{W})$ takes a matrix input $\boldsymbol{W} = (W_{ij})$
- First order approximation formula:

$$f(\boldsymbol{W}) \approx f(\boldsymbol{W}_0) + \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{\partial f}{\partial W_{ij}} \left(W_{ij} - W_{0,ij}\right)$$

- Change in $f(\boldsymbol{W})$ given by matrix inner product:

$$f(\boldsymbol{W}) - f(\boldsymbol{W}_0) \approx \langle \nabla f(\boldsymbol{W}_0), \boldsymbol{W} - \boldsymbol{W}_0 \rangle, \qquad \langle \boldsymbol{A}, \boldsymbol{B} \rangle := \sum_{i=1}^{M}\sum_{j=1}^{N} A_{ij}B_{ij}$$

  - Similar to the vector formula

Example 4: Matrix-Input Function (Advanced Concept)

- Suppose

$$f(\boldsymbol{W}) = \boldsymbol{a}^T \boldsymbol{W} \boldsymbol{b}$$

  - Matrix input / scalar output
- Then, $f(\boldsymbol{W}) = \boldsymbol{a}^T \boldsymbol{W} \boldsymbol{b} = \sum_{ij} a_i b_j W_{ij}$
- Partial derivatives: $\dfrac{\partial f}{\partial W_{ij}} = a_i b_j$
- Gradient:

$$\nabla f(W) = \begin{bmatrix} a_1 b_1 & \cdots & a_1 b_N \\ \vdots & \vdots & \vdots \\ a_N b_1 & \cdots & a_N b_N \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} [b_1 \quad \cdots \quad b_N] = \boldsymbol{a}\boldsymbol{b}^T$$

  - $\boldsymbol{a}\boldsymbol{b}^T$ is called the outer product

Example 4 in Python (Advanced Concept)

- Function: $f(\boldsymbol{W}) = \boldsymbol{a}^T \boldsymbol{W} \boldsymbol{b}$
  - Use python `dot` for matrix-vector products
- Gradient: $\nabla f(\boldsymbol{W}) = \boldsymbol{a}\boldsymbol{b}^T$
  - Want fgrad[i,j] = a[i]b[j]
  - Avoid for-loops
  - Use python broadcasting
  - a[:,None] = m x 1
  - b[None,:] = 1 x n

```python
def feval(W,a,b):
    # Function
    f = a.dot(W.dot(b))

    # Gradient -- Use python broadcasting
    fgrad = a[:,None]*b[None,:]

    return f, fgrad

# Some random data
m = 4
n = 3
W = np.random.randn(m,n)
a = np.random.randn(m)
b = np.random.randn(n)

f, fgrad = feval(W,a,b)
```

Outline

- Motivating example: Build an optimizer for logistic regression
- Gradients of multi-variable functions
- Gradient descent
- Adaptive step size
- Convexity

Unconstrained Optimization

- Problem: Given $f(\boldsymbol{w})$ find the minimum:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} f(\boldsymbol{w})$$

  - $f(w)$ is called the objective function (in the optimization lingo)
    - Loss function in ML
  - $\boldsymbol{w} = (w_1, \cdots, w_M)$ is a vector of decision variables or parameters

- Called unconstrained since there are no constraints on $\boldsymbol{w}$
- Will discuss constrained optimization briefly later
  - Would require more math

Numerical Optimization

- We saw that we can find minima by setting $\nabla f(w) = 0$
  - Can also be a saddle point
  - M equations and M unknowns.
  - May not have closed-form solution

- Numerical methods: Finds a sequence of estimates w^k that (hopefully) converges to the true solution

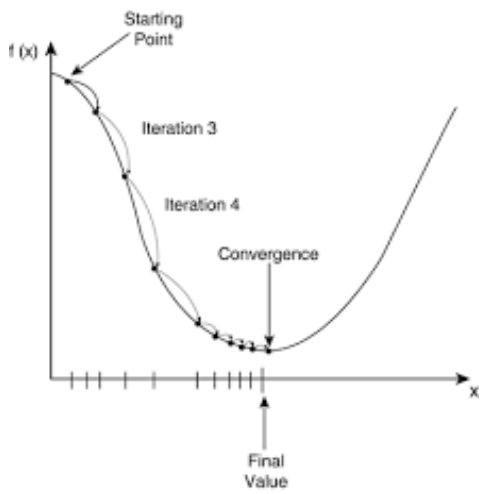
$$w^k \to w^*$$

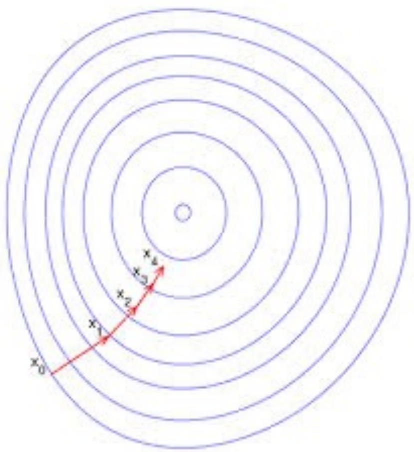
  - Or converges to some other “good” minima
  - Run on a computer program, like python

Gradient Descent

- Most simple method for **unconstrained optimization**
  - There are variants for constrained ones
- Key property of gradient, $\nabla_w f(\boldsymbol{w})$
  - $-\nabla_w f(\boldsymbol{w})$ = Points in the direction of steepest decrease
- Gradient descent algorithm:
  - Start with initial w^0 (which may not be very good ☹)
  - $w^{k+1} = w^k - \alpha_k \nabla f(w^k)$
  - Repeat until some stopping criteria
- α_k is called the step size
  - In machine learning, this is called the learning rate

Gradient Descent Illustrated





- $M = 1$
- $M = 2$

Gradient Descent Analysis (Advanced)

- Using gradient update rule

$$f\left(w^{k+1}\right) = f\left(w^{k}\right) + \nabla f\left(w^{k}\right) \cdot \left(w^{k+1} - w^{k}\right) + O\left\|w^{k+1} - w^{k}\right\|^{2}$$
$$= f\left(w^{k}\right) \; - \alpha \, \nabla f\left(w^{k}\right) \cdot \nabla f\left(w^{k}\right) + O(\alpha^{2})$$

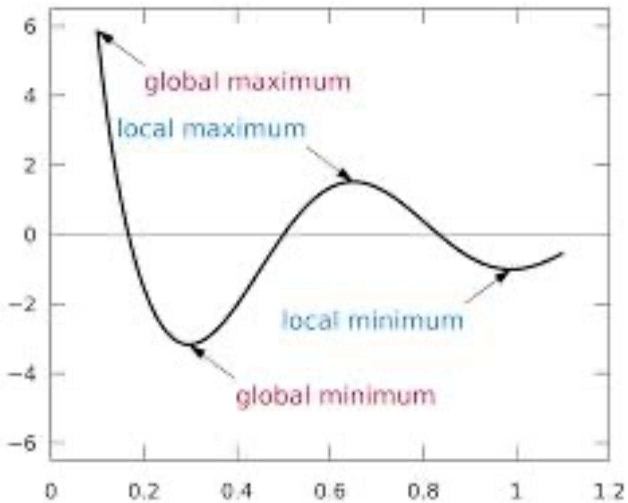
$$= f\left(w^{k}\right) \; - \alpha \left\|\nabla f\left(w^{k}\right)\right\|^{2} + O(\alpha^{2})$$

- Consequence: If step size α is small, then $f\left(w^{k}\right)$ decreases
- Theorem:

If $f''(w)$ is bounded above, $f(w)$ is bounded below, and α is chosen sufficiently small,

Then gradient descent converges to local minima

Local vs. Global Minima



- Definitions:
  - w^* is a global minima if $f(w) \geq f(w^*)$ for all w
  - w^* is a local minima if $f(w) \geq f(w^*)$ for all w in some open neighborhood of w^*
- Most numerical methods (including gradient descent):
  - Generally only guarantee convergence to local minima
- Convex functions: Have only global minima (more later)

Gradients for Logistic Regression

- Logistic regression
  - Linear function: $z_i = w_0 + \sum_{j=1}^{d} X_{ij} w_j$
  - Output probability: $P(y = 1|x) = \frac{1}{1+e^{-z_i}}$
  - Binary cross-entropy loss: $J(\boldsymbol{w}) = \sum_{i=1}^{n} \{\ln[1 + e^{z_i}] - y_i z_i\}$
- Compute gradients:
  - Define $A = [1 \;\; X]$, matrix with ones on the first column
  - Then, $z_i = w_0 + \sum_{j=1}^{d} X_{ij} w_j = \sum_{j=0}^{d} A_{ij} w_j$
  - Let $p_i = \frac{1}{1+e^{-z_i}}$
  - Observe $\frac{\partial J}{\partial z_i} = \frac{e^{z_i}}{1+e^{z_i}} - y_i = p_i - y_i$
  - Use multi−variable chain rule:

$$\frac{\partial J}{\partial w_j} = \sum_{i=1}^{N} \frac{\partial J}{\partial z_i} \frac{\partial z_i}{\partial w_j} = \sum_{i=1}^{N} (p_i - y_i)\, A_{ij}$$

Matrix Form

- Logistic regression
  - Linear function: $z_i = \sum_{j=0}^{d} A_{ij} w_j$
  - Output probability: $P(y=1|x) = \frac{1}{1+e^{-z_i}}$
  - BCE: $J = \sum_{i=1}^{n} \{\ln[1+e^{z_i}] - y_i z_i\}$
  - $\frac{\partial J}{\partial z_i} = p_i - y_i$
  - $\frac{\partial J}{\partial w_j} = \sum_{i=1}^{N} \frac{\partial J}{\partial z_i}\frac{\partial z_i}{\partial w_j} = \sum_{i=1}^{N} (p_i - y_i) A_{ij}$
- Matrix form:
  - $z = Aw$
  - Let $p = \frac{1}{1+e^{-z}}$
  - $\frac{\partial J}{\partial z} = p - y$
  - $\frac{\partial J}{\partial w} = A^T \frac{\partial J}{\partial z}$

```
def feval(w,X,y):
    """
    Compute the loss and gradient given w,X,y
    """
    # Construct transform matrix
    n = X.shape[0]
    A = np.column_stack((np.ones(n,), X))

    # The loss is the binary cross entropy
    z = A.dot(w)
    py = 1/(1+np.exp(-z))
    f = np.sum((1-y)*z - np.log(py))

    # Gradient
    df_dz = py-y
    fgrad = A.T.dot(df_dz)
    return f, fgrad
```

Implementation in Python

- Optimizer requires a python method to compute:
  - Objective function $f(\boldsymbol{w})$, and
  - Gradient $\nabla f(\boldsymbol{w})$
- For logistic loss:

$$f(\boldsymbol{w}) := \sum_{i=1}^{N} -y_i z_i + \ln[1 + e^{z_i}], \qquad z = Aw$$

- Thus, $f(\boldsymbol{w})$ and $\nabla f(\boldsymbol{w})$ depends on training data $(\boldsymbol{x}_i, y_i)$
  - How do we pass these?
- Two methods to pass data to the function:
  - Method 1: Use a class
  - Method 2: Use lambda calculus

Training data

```python
def feval(w,X,y):
    """
    Compute the loss and gradient given w,X,y
    """
    # Construct transform matrix
    n = X.shape[0]
    A = np.column_stack((np.ones(n,), X))

    # The loss is the binary cross entropy
    z = A.dot(w)
    py = 1/(1+np.exp(-z))
    f = np.sum((1-y)*z - np.log(py))

    # Gradient
    df_dz = py-y
    fgrad = A.T.dot(df_dz)
    return f, fgrad
```

Method 1: Create a Class

- Create a class for the objective function
- Pass data (x_i, y_i) in constructor
  - Also perform any pre-computations
- Pass argument $\boldsymbol{w}$ to method feval
  - Evaluates function and gradient
  - Can access the data as class members
- Instantiate the class with data (training data)

```
class LogisticFun(object):
    def __init__(self,X,y):
        """
        Class for computes the loss and gradient for a logistic regression problem.

        The constructor takes the data matrix `X` and response vector y for training.
        """
        self.X = X
        self.y = y
        n = X.shape[0]
        self.A = np.column_stack((np.ones(n,), X))

    def feval(self,w):
        """
        Compute the loss and gradient for a given weight vector
        """
        # The loss is the binary cross entropy
        z = self.A.dot(w)
        py = 1/(1+np.exp(-z))
        f = np.sum((1-self.y)*z - np.log(py))

        # Gradient
        df_dz = py-self.y
        fgrad = self.A.T.dot(df_dz)
        return f, fgrad
```

```
log_fun = LogisticFun(Xtr,ytr)
```

```
# Call the function
f, fgrad = log_fun.feval(w0)
```

Testing the Gradient

- Always test your implementation!
- Pick two points $\boldsymbol{w}_0, \boldsymbol{w}_1$ that are close
- Make sure: $f(\boldsymbol{w}_1) - f(\boldsymbol{w}_0) \approx \nabla f(\boldsymbol{w}_0)^T(\boldsymbol{w}_1 - \boldsymbol{w}_0)$

```
# Take a random initial point
p = X.shape[1]+1
w0 = np.random.randn(p)

# Perturb the point
step = 1e-6
w1 = w0 + step*np.random.randn(p)

# Measure the function and gradient at w0 and w1
f0, fgrad0 = log_fun.feval(w0)
f1, fgrad1 = log_fun.feval(w1)

# Predict the amount the function should have changed based on the gradient
df_est = fgrad0.dot(w1-w0)

# Print the two values to see if they are close
print("Actual f1-f0    = %12.4e" % (f1-f0))
print("Predicted f1-f0 = %12.4e" % df_est)
```

```
Actual f1-f0    =   3.3279e-04
Predicted f1-f0 =   3.3279e-04
```

Method 2: Lambda Calculus

- Create a function that take w, X, y
- Use lambda function to fix X, y

```
[6]: def feval(w,X,y):
         """
         Compute the loss and gradient given w,X,y
         """
         # Construct transform matrix
         n = X.shape[0]
         A = np.column_stack((np.ones(n,), X))

         # The loss is the binary cross entropy
         z = A.dot(w)
         py = 1/(1+np.exp(-z))
         f = np.sum((1-y)*z - np.log(py))

         # Gradient
         df_dz = py-y
         fgrad = A.T.dot(df_dz)
         return f, fgrad
```

```
[10]: # Create a function with X,y fixed
      feval_param = lambda w: feval(w,Xtr1,ytr)

      # You can now pass a parameter like w0
      f0, fgrad0 = feval_param(w0)
```

Gradient Descent

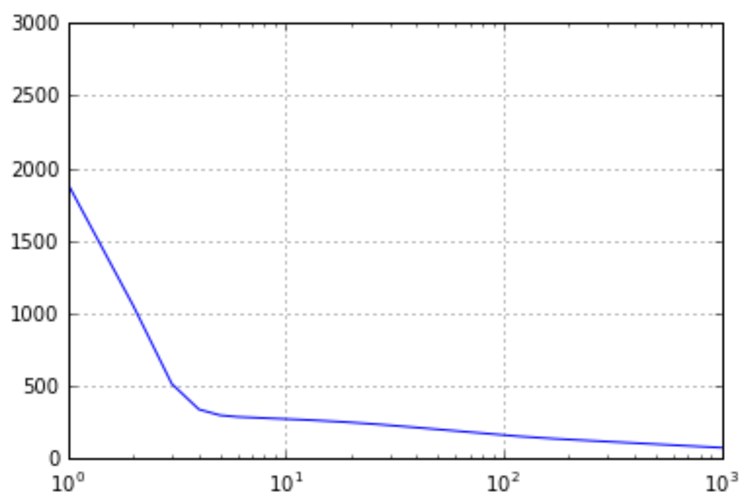
- Input parameters:
  - Function to return objective and gradient
  - Initial value w^0
  - Learning rate α
  - Number of iterations
- Code returns:
  - Final estimate w^k
  - Final function value $f(w^k)$
  - History (for debugging)

```
def grad_opt_simp(feval, winit, lr=1e-3,nit=1000):
    """
    Simple gradient descent optimization

    feval:  A function that returns f, fgrad, the objective
            function and its gradient
    winit:  Initial estimate
    lr:     learning rate
    nit:    Number of iterations
    """
    # Initialize
    w0 = winit

    # Create history dictionary for tracking progress per iteration.
    # This isn't necessary if you just want the final answer, but it
    # is useful for debugging
    hist = {'w': [], 'f': []}

    # Loop over iterations
    for it in range(nit):

        # Evaluate the function and gradient
        f0, fgrad0 = feval(w0)

        # Take a gradient step
        w0 = w0 - lr*fgrad0

         # Save history
        hist['f'].append(f0)
        hist['w'].append(w0)

    # Convert to numpy arrays
    for elem in ('f', 'w'):
        hist[elem] = np.array(hist[elem])
    return w0, hist
```

Gradient Descent on Logistic Regression

- Random initial condition
- 1000 iterations
- Convergence is slow.
- Final accuracy less than sk-learn optimizer!
  - estimate has not converged

```
# Initial condition
winit = np.random.randn(p)

# Parameters
feval = log_fun.feval
nit = 1000
lr = 1e-4

# Run the gradient descent
w, f0, hist = grad_opt_simp(feval, winit, lr=lr, nit=nit)

# Plot the training loss
t = np.arange(nit)
plt.semilogx(t, hist['f'])
plt.grid()
```



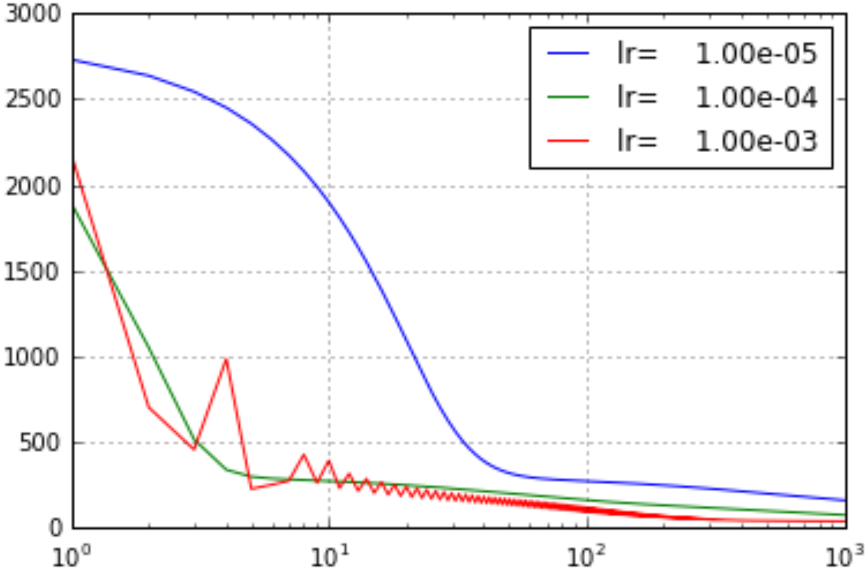

```
def predict(X,w):
    z = X.dot(w[1:]) + w[0]
    yhat = (z > 0)
    return yhat

yhat = predict(Xts,w)
acc = np.mean(yhat == yts)
print("Test accuracy = %f" % acc)
```

```
Test accuracy = 0.971731
```

Different Step Sizes

- Faster learning rate => Faster convergence
- But, may be unstable

```
lr=     1.00e-05  Test accuracy = 0.681979
lr=     1.00e-04  Test accuracy = 0.964664
lr=     1.00e-03  Test accuracy = 0.989399
```



Outline

- Motivating example: Build an optimizer for logistic regression
- Gradients of multi-variable functions
- Gradient descent
- Adaptive step size
- Convexity

Adaptive Step Size Selection

- Most practical algorithms change step size adaptively

$$w^{k+1} = w^k - \alpha_k \nabla f(w^k)$$

- Tradeoff: Selecting large α_k:
  - Larger steps, faster convergence
  - But, may overshoot

Armijo Rule

- Recall that we know if $w^{k+1} = w^k - \alpha \nabla f(w^k)$

$$f(w^{k+1}) = f\left(w^k\right) - \alpha \left\|\nabla f\left(w^k\right)\right\|^2 + O(\alpha^2)$$

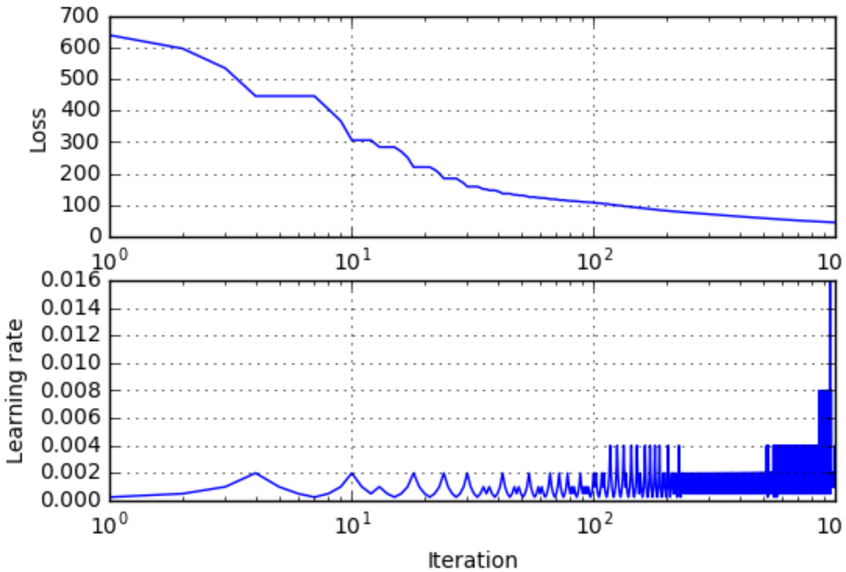
- Armijo Rule:
  - Select some $c \in (0,1)$. Usually $c = 1/2$
  - Select α such that

$$f\left(w^{k+1}\right) \leq f\left(w^k\right) - c\alpha \left\|\nabla f\left(w^k\right)\right\|^2$$

  - Decreases by at least at fraction c predicted by linear approx.
- Simple update:
  - If Armijo rule passes: Accept point and increase step size: $\alpha^{k+1} = \beta\alpha^k,\ \beta > 1$
  - If Armijo rule fails: Reject point and decrease step size: $\alpha^{k+1} = \beta^{-1}\alpha^k$
- Can also use a line search

Adaptive Gradient Descent in Python

- Simple modification of fixed step size case




```
for it in range(nit):

    # Take a gradient step
    w1 = w0 - lr*fgrad0

    # Evaluate the test point by computing the objective function, f1,
    # at the test point and the predicted decrease, df_est
    f1, fgrad1 = feval(w1)
    df_est = fgrad0.dot(w1-w0)

    # Check if test point passes the Armijo rule
    alpha = 0.5
    if (f1-f0 < alpha*df_est) and (f1 < f0):
        # If descent is sufficient, accept the point and increase the
        # learning rate
        lr = lr*2
        f0 = f1
        fgrad0 = fgrad1
        w0 = w1
    else:
        # Otherwise, decrease the learning rate
        lr = lr/2
```

What is β here?

Outline

- Motivating example: Build an optimizer for logistic regression
- Gradients of multi-variable functions
- Gradient descent
- Adaptive step size
- Convexity

Convex Sets

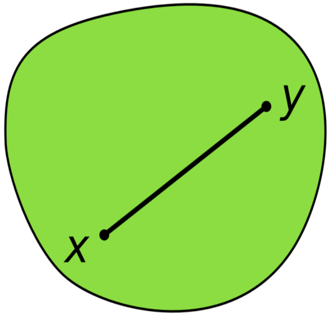
- Definition: A set X is convex if for any $x, y \in X$,

$$tx + (1-t)y \in X \text{ for all } t \in [0,1]$$

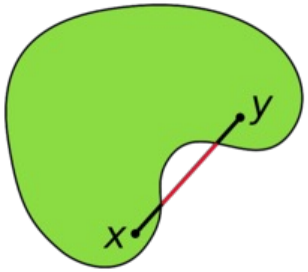
- Any line between two points remains in the set.
- Examples:
  - Square, circle, ellipse
  - $\{x \mid Ax \leq b\}$ for any matrix A and vector b

Convex Set Visualized

- Convex



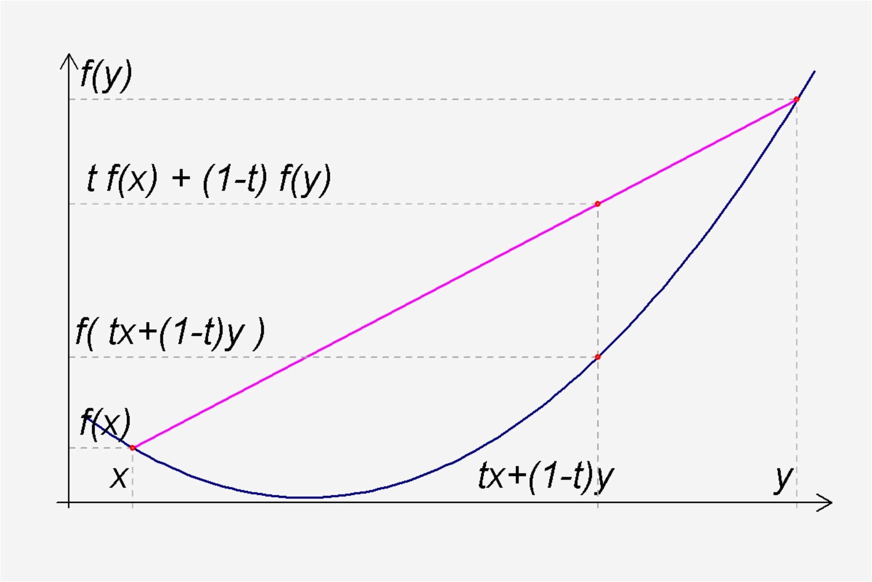
- Not convex



Convex Functions

- A real-valued function $f(x)$ is convex if:
  - Its domain is a convex set, and
  - For all x, y and $t \in [0,1]$:

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$



Convex Function Examples

- Linear function of a scalar $f(x) = ax + b$
- Linear function of a vector $f(x) = a^T x + b$
- Quadratic $f(x) = \frac{1}{2}ax^2 + bx + c$ is convex iff $a \geq 0$
- If $f''(x)$ exists everywhere, $f(x)$ is convex iff $f''(x) \geq 0$.
  - When x is a vector $f''(x) \geq 0$ means the Hessian must be positive semidefinite
- $f(x) = e^x$
- If $f(x)$ is convex, so is $f(Ax + b)$
- Logistic loss is convex!

Hessian Matrix (Advanced Concept)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a function taking as input a vector $\mathbf{x} \in \mathbb{R}^n$ and outputting a scalar $f(\mathbf{x}) \in \mathbb{R}$. If all second-order partial derivatives of f exist, then the Hessian matrix $\mathbf{H}$ of f is a square $n \times n$ matrix, usually defined and arranged as follows:

$$\mathbf{H}_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\[2.2ex] \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\[2.2ex] \vdots & \vdots & \ddots & \vdots \\[2.2ex] \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix},$$

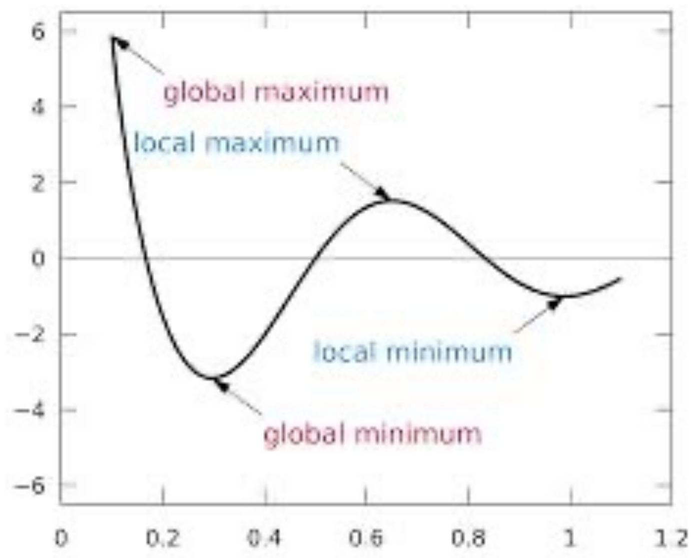
Positive-Definite and Positive Semi-Definite (Advanced Concept)

In mathematics, a symmetric matrix M with real entries is **positive-definite** if the real number $z^{\mathsf{T}} M z$ is positive for every nonzero real column vector z, where z^{T} is the transpose of z.[1] More generally, a Hermitian matrix (that is, a complex matrix equal to its conjugate transpose) is **positive-definite** if the real number $z^* M z$ is positive for every nonzero complex column vector z, where z^* denotes the conjugate transpose of z.

Positive semi-definite matrices are defined similarly, except that the scalars $z^{\mathsf{T}} M z$ and $z^* M z$ are required to be positive *or zero* (that is, nonnegative). **Negative-definite** and **negative semi-definite** matrices are defined analogously. A matrix that is not positive semi-definite and not negative semi-definite is sometimes called **indefinite**.

Global Minima and Convex Function

- Theorem: If $f(w)$ is convex and w is a local minima, then w is a global minima
- Implication for optimization:
  - Gradient descent only converges to local minima
  - In general, cannot guarantee optimality
  - Depends on initial condition
  - But, for convex functions can always obtain optimal



Other Topics We Did Not Cover

- Our optimizer is OK, but not nearly as fast as sklearn method
- Many techniques we did not cover
  - Newton’s method
  - Quasi-Newton’s method
  - Non-smooth optimization
  - Constrained optimization

- Take an optimization class and learn more.

What you should know

- Identify the objective function, parameters and constraints in an optimization problem
- Compute the gradient of a loss function for scalar, vector parameters
  - Matrix parameters are advanced (graduate students only)
- Efficiently compute a gradient in python.
- Write the gradient descent update
- Describe the effect of the learning rate on convergence
- Determine if a loss function is convex