45 Example (Inflection Point)
Draw the graph of a smooth function with an inflection point.

46 Example (Local vs Global Minimums)
Draw the graph of a function that has a two minimum values, a local minimum and a global minimum.

One strategy for minimizing a loss function is to perform a grid search. If the function is smooth, a gradient descent algorithm is likely to be much more efficient.[6]

47 Definition (Gradient Descent Algorithm)
The gradient descent algorithm for minimizing a smooth scalar function, $f(x)$, is outlined below.

```
Choose a step size, $h$.
Choose a starting point, $x_0$.
Define $f_{\text{min}} = f(x_0)$.
Initialize $k = 0$.
while $f(x_k) \leq f_{\text{min}}$:
    $f_{\text{min}} = f(x_k)$
    $x_{k+1} = x_k - f'(x_k)h$
    $k = k + 1$
return $x_k$
```

48 Definition (Gradient Vector)
The gradient vector, $\dfrac{dy}{d\mathbf{x}}$, of the function $y = f(\mathbf{x})$ is given by the vector

$$\frac{dy}{d\mathbf{x}} = \left(\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \ldots, \frac{\partial y}{\partial x_n}\right)^{\top}.$$

49 Theorem (Direction of Maximum Increase/Decrease)
The direction in the domain of the function $y = f(\mathbf{x})$ in which $f(\mathbf{x})$ increases the fastest is in the direction of the gradient vector $\dfrac{dy}{d\mathbf{x}}$. The direction in which $f(x)$ decreases the fastest is $-\dfrac{dy}{d\mathbf{x}}$.

50 Example (Gradient Vector)
Consider the function

$$y = f(x_1, x_2) = x_1 + x_1 x_2 + x_2^2.$$

(a) Compute the gradient vector of $f(x_1, x_2)$.

(b) In which direction does $f(x_1, x_2)$ increase the fastest at the point (2,1).

(c) In which direction does $f(x_1, x_2)$ decrease the fastest at the point (2,1).

51 Definition (Central Difference Approximation)
The partial derivative of the function $y = f(\mathbf{x})$ with respect to the variable x_i can be approximated by the central difference formula

$$\frac{\partial y}{\partial x_i} \approx \frac{f(x_1, \ldots, x_i + \epsilon, \ldots, x_n) - f(x_1, \ldots, x_i - \epsilon, \ldots, x_n)}{2\epsilon}$$

where ϵ is a small number (e.g. $\epsilon = 10^{-5}$).

52 Example (Numerical Approximation of the Gradient Vector)
Use a central difference approximation to numerically approximate the gradient vector $\dfrac{df(1,1)}{d\mathbf{x}}$ for the function given below using the values $\epsilon = 10^{-1}, 10^{-3}, 10^{-5}$. Compare your approximation to the exact answer.

$$f(x_1, x_2) = e^{1 - x_1 x_2}$$

[6] The gradient of a scalar function equals the ordinary derivative of the function.

The gradient descent algorithm can be easily adapted to functions with multiple input variables. Simply replace the scalar quantity $\frac{dy}{dx}$ with the vector quantity $\frac{dy}{d\mathbf{x}}$.

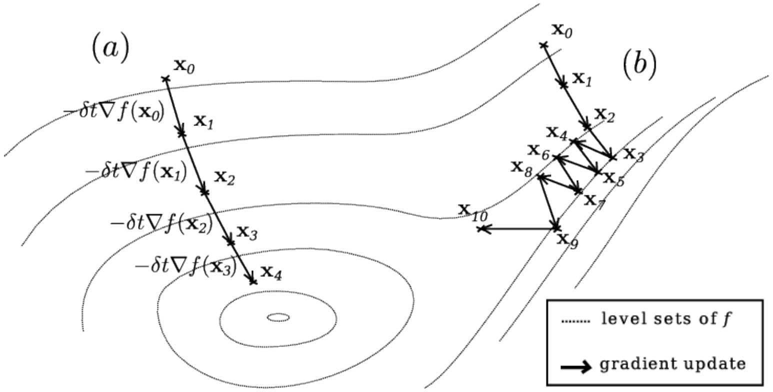
53 Example (Gradient Vector for Linear Regression)

Show that the gradient vector $\frac{d\ell}{d\mathbf{w}}$ for the MSE loss function (L^2 loss function) for linear regression $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$ is given by

$$\frac{d\ell}{d\mathbf{w}} = \frac{2}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y}).$$

In certain situations, the gradient descent algorithm for finding the minimum of a function can converge very slowly. Consider the following example.

54 Example (Slow Convergence of Gradient Descent)
Which initial condition, (a) or (b), in the diagram below is likely to result in slow convergence of the gradient descent algorithm?[7]



Standardizing input data to a neural network often accelerates convergence.

55 Definition (Standardizing Data)
A data set can be standardized by shifting and scaling the data set as follows:

$$z = \frac{x - \overline{x}}{s_x}$$

where $\overline{x}$ is the mean of x and s_x is the standard deviation of x and z is the standardized data.

56 Example (Standardize Linear Regression for Wine)
Which features are most important for determining the quality of (a) red wine (b) white wine? Standardize the features, perform linear regression and examine the feature weights.

Optimization problems can be divided into two types:
(i) convex optimization problems
(ii) non-convex optimization problems

Convex optimization problems are much easier than non-convex optimization problems. Much of classical machine learning deals with convex optimization problems. Training a neural network is a non-convex optimization problem. Until recently, many people believed that it was intractable to train large neural networks.

57 Definition (Convex Function)
A function is convex if

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2).$$

[7] http://ludovicarnold.altervista.org/teaching/