University of Stuttgart
Institute for Natural Language Processing

Thang Vu

Math for ML: Probabilities II & Optimization

Contents

Probabilities II

1

Summary Statistics and Independence

1.1

Expected Value

Definition (Expected Value)

The expected value of a function g of a univariate discrete random variable $X \sim p(x)$ is given by:

$$\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x)\, p(x)$$

- Here, $\mathcal{X}$ is the set of possible outcomes (the target space) of the random variable X.
- Intuition: If we perform the same experiment many times, what would be the average across all outcomes?

Expected Value: Example

- Given a fair, six-sided die:
  - Sample space: $\Omega = \{1, 2, 3, 4, 5, 6\}$ (analogously: event space A)
  - $P(X = 1) = P(X = 2) = ... = P(x = 6) = 1/6$

Expected Value: Example

- Given a fair, six-sided die:
  - Sample space: $\Omega = \{1, 2, 3, 4, 5, 6\}$ (analogously: event space A)
  - $P(X = 1) = P(X = 2) = ... = P(x = 6) = 1/6$

$$\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x)\, p(x) = \sum_{x \in \{1,2,\dots,6\}} g(x) \cdot \frac{1}{6} \quad (1)$$

$$= \frac{1}{6} \cdot (1 + 2 + 3 + 4 + 5 + 6) = \frac{1}{6} \cdot 21 = 3.5 \quad (2)$$

Expected Value

Definition (Expected Value)

The expected value of a function $g : \mathbb{R} \to \mathbb{R}$ of a univariate continuous random variable $X \sim p(x)$ is given by:

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)\, p(x)\, dx$$

- Again, $\mathcal{X}$ is the set of possible outcomes (the target space) of the random variable X.

Expected Value

Definition (Expected Value)

We can view multivariate random variables X as a finite vector of univariate random variables $X = [X_1, \ldots, X_D]^\top$. Then, the expected value of a multivariate random variable X is given by:

$$\mathbb{E}_X[g(x)] = \begin{pmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_D}[g(x_D)] \end{pmatrix}$$

- Here, the subscript $\mathbb{E}_{X_d}$ indicates that we are taking the expected value with respect to the d-th element of the vector x.

Mean

Definition (Mean)

The mean of a multivariate random variable X with states $x \in \mathbb{R}^D$ is an average and is defined as:

$$\mu = \mathbb{E}_X[x] = \begin{pmatrix} \mathbb{E}_{X_1}[x_1] \\ \vdots \\ \mathbb{E}_{X_D}[x_D] \end{pmatrix}$$

where

$$\mathbb{E}_{X_d}[x_d] = \begin{cases} \int_{\mathcal{X}} x_d \, p(x_d) \, dx_d, & \text{if } X \text{ is continuous} \\ \sum_{x_i \in \mathcal{X}} x_i \, p(x_d = x_i), & \text{if } X \text{ is discrete} \end{cases}$$

for $d = 1, \ldots, D$, where the subscript d indicates the corresponding dimension of x.

Covariance of Univariate Random Variables

Definition (Covariance)

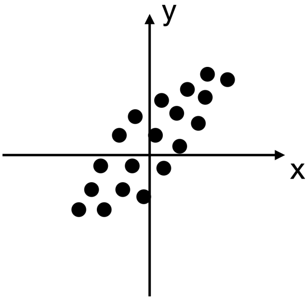
The covariance between two univariate random variables $X, Y \in \mathbb{R}$ is given by the expected product of their deviations from their respective means:

$$\mathrm{Cov}_{X,Y}[x, y] := \mathbb{E}_{X,Y}\left[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])\right]$$

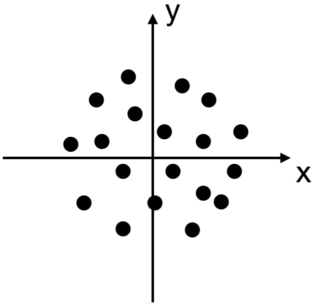
Remark:

- When the random variable associated with the expectation or covariance is clear from the context, the subscript is often suppressed. For example, $\mathbb{E}_X[x]$ is often written as $\mathbb{E}[x]$.
- Intuition: Is there a (linear) relationship between X and Y?

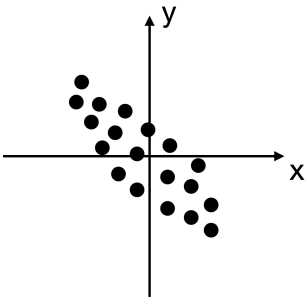
Example: Covariance



(a) $Cov(X, Y) > 0$



(b) $Cov(X, Y) \approx 0$



(c) $Cov(X, Y) < 0$

Covariance: Example

- Let X be the rainfall in cm
- Let Y be the amount of umrellas sold
- We observe 4 samples:
  $(0cm, 1), (2cm, 3), (4.5cm, 4umbrellas), (6.5cm, 6umbrellas)$
- Let's assume all events are equally likely, and so are the joint observations (i.e., $P(X = x) = \frac{1}{4}, P(Y = y) = \frac{1}{4}, P(X = x, Y = y) = \frac{1}{4}$)

Covariance: Example

- We observe:
  $(0cm, 0), (2cm, 3), (4.5cm, 4umbrellas), (6.5cm, 6umbrellas)$
- $\mathbb{E}_X[x] = \sum_{i=1}^4 x_i \cdot p(x_i) = \frac{1}{4}\sum_{i=1}^4 x_i = \frac{1}{4} \cdot (0 + 2 + 4.5 + 6.5) = 3.25$
- $\mathbb{E}_Y[y] = \sum_{i=1}^4 y_i \cdot p(y_i) = \frac{1}{4}\sum_{i=1}^4 y_i = \frac{1}{4} \cdot (1 + 3 + 4 + 6) = 3.5$
- $\mathrm{Cov}_{X,Y}[x, y] = \mathbb{E}_{X,Y}[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])] = \mathbb{E}_{X,Y}[(x - 3.25)(y - 3.5)]$
  $= \sum_{i=1}^4 (x - 3.25)(y - 3.5)p(x, y)$
- $P(X, Y) : 0$ except for our sample points, where it is $\frac{1}{4}$:
- $\frac{1}{4}[(0 - 3.25)(1 - 3.5) + (2 - 3.25)(3 - 3.5) + (4.5 - 3.25)(4 - 3.5) + (6.5 - 3.25)(6 - 3.5)]$
  $= \frac{1}{4}17.5 = 4.375$

Covariance: Example

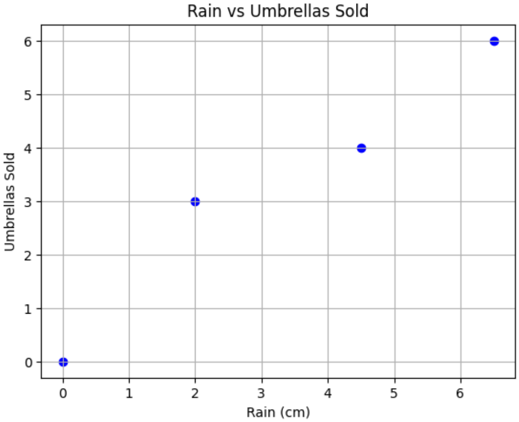


Figure: Plot for observed data points, positive covariance (4.375)

Covariance of Multivariate Random Variables

Definition (Covariance (Multivariate))

If we consider two multivariate random variables X and Y with states $x \in \mathbb{R}^D$ and $y \in \mathbb{R}^E$ respectively, the covariance between X and Y is defined as:

$$\mathrm{Cov}[x, y] = \mathbb{E}[xy^\top] - \mathbb{E}[x]\mathbb{E}[y]^\top = \mathrm{Cov}[y, x]^\top \in \mathbb{R}^{D \times E}$$

Note:

- When the same multivariate random variable is used in both arguments, i.e., $X = Y$, this definition results in the covariance matrix of X.
- The covariance matrix captures the relationship between individual dimensions of the random variable and intuitively describes its ”spread”.

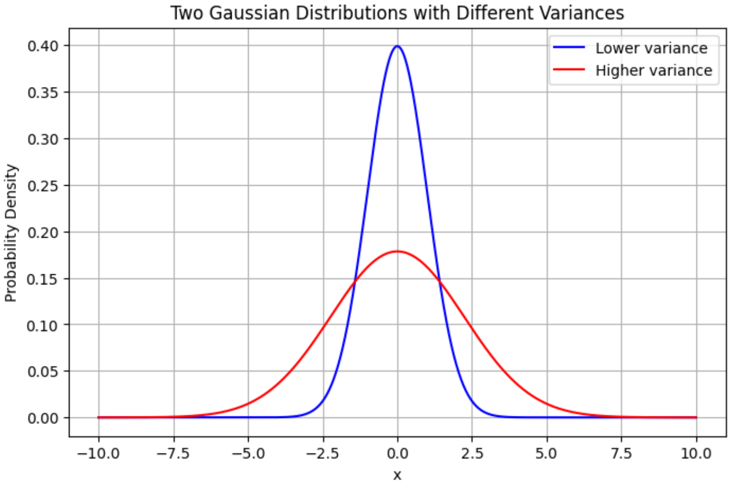
Variance of a Random Variable

Definition (Variance)

The variance of a random variable X with states $x \in \mathbb{R}^D$ and a mean vector $\mu \in \mathbb{R}^D$ is defined as:

$$
\begin{aligned}
V_X[x] &= \mathrm{Cov}_X[x, x] \\
&= \mathbb{E}_X\left[(x-\mu)(x-\mu)^\top\right] = \mathbb{E}_X[xx^\top] - \mathbb{E}_X[x]\mathbb{E}_X[x]^\top \\
&= \begin{pmatrix} \mathrm{Cov}[x_1, x_1] & \mathrm{Cov}[x_1, x_2] & \cdots & \mathrm{Cov}[x_1, x_D] \\ \mathrm{Cov}[x_2, x_1] & \mathrm{Cov}[x_2, x_2] & \cdots & \mathrm{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D, x_1] & \mathrm{Cov}[x_D, x_2] & \cdots & \mathrm{Cov}[x_D, x_D] \end{pmatrix}
\end{aligned}
$$

Variance: Example



Variance of a Random Variable

Properties of the Covariance Matrix:

- The $D \times D$ matrix is called the **covariance matrix** of the multivariate random variable X.
- The covariance matrix is **symmetric** and **positive semidefinite**.
- It describes the **spread** of the data.
- Diagonal elements contain the variances of the individual variables:

$$\mathrm{Var}[x_i] = \mathrm{Cov}[x_i, x_i]$$

- Off-diagonal elements are the **cross-covariance** terms:

$$\mathrm{Cov}[x_i, x_j], \quad \text{for } i \neq j$$

Correlation

Definition (Correlation)

The correlation between two random variables X and Y is given by:

$$\operatorname{corr}[x, y] = \frac{\operatorname{Cov}[x, y]}{\sqrt{V[x]\,V[y]}} \in [-1, 1]$$

Correlation Matrix:

- The correlation matrix is the covariance matrix of standardized random variables, where each variable is divided by its standard deviation (the square root of its variance).
- This standardization ensures that the variables are dimensionless, allowing for direct comparison of the strength of relationships between variables.

Empirical Mean and Covariance

Definition (Empirical Mean and Covariance)

The empirical mean vector is the arithmetic average of the observations for each variable and is defined as:

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

where $x_n \in \mathbb{R}^D$.

Definition (Empirical Covariance Matrix:)

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^\top$$

Independence of Random Variables

Definition (Independence)

Two random variables X and Y are *statistically independent* if and only if:

$$p(x, y) = p(x)\, p(y)$$

Intuition:

- X and Y are independent if knowing the value of Y does not provide any additional information about X, and vice versa.

Implications:

- **Conditional Probability:** $p(y \mid x) = p(y)$,
  $$p(x \mid y) = p(x)$$
- **Variance of Sum:** $V_{X,Y}[x + y] = V_X[x] + V_Y[y]$
- **Covariance:** $\mathrm{Cov}_{X,Y}[x, y] = 0$

Conditional Independence

Definition (Conditional Independence)

Two random variables X and Y are *conditionally independent* given a random variable Z if:

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z), \quad \forall z \in \mathcal{Z}$$

We denote this relationship as $X \perp Y \mid Z$.

Alternative Interpretation:

- Given Z, knowing Y provides no additional information about X:

$$p(x \mid y, z) = p(x \mid z)$$

Live Voting

Important Probability Distributions

1.2

Gaussian Distribution

Introduction:

- The **Gaussian distribution**, also known as the *normal distribution*, is the most well-studied probability distribution for continuous-valued random variables.
- Its importance originates from its many computationally convenient properties.
- Widely used in various areas of machine learning such as Gaussian processes, variational inference, and reinforcement learning.
- Also prevalent in other fields like signal processing (e.g., Kalman filter), control (e.g., linear quadratic regulator), and statistics (e.g., hypothesis testing).

Gaussian Distribution

Definition (Univariate Gaussian Distribution)

For a univariate random variable, the Gaussian distribution has a density given by:

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

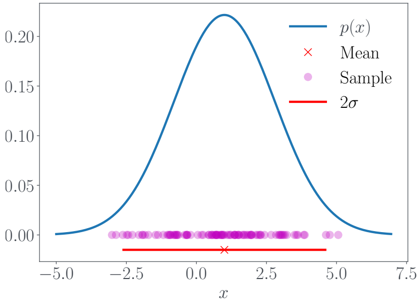
where:

- μ is the mean of the distribution.
- σ^2 is the variance of the distribution.

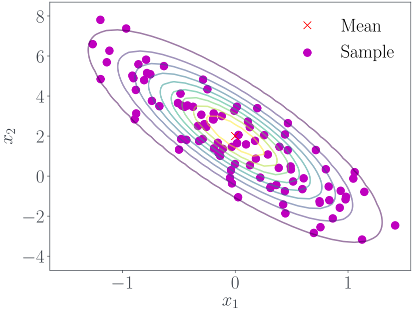
Properties:

- The Gaussian distribution is symmetric around the mean μ.
- Mean, median, and mode are all equal.
- Completely determined by its mean and variance.
- The area under the curve integrates to 1.

Gaussian Distribution – Plot



(a) Univariate (one-dimensional) Gaussian; The red cross shows the mean and the red line shows the extent of the variance.



(b) Multivariate (two-dimensional) Gaussian, viewed from top. The red cross shows the mean and the colored lines show the contour lines of the density.

Figure: Taken from Deisenroth, Faisal, and Ong (2020).

Multivariate Gaussian Distribution

Definition (Multivariate Gaussian Distribution)

The multivariate Gaussian distribution is fully characterized by a mean vector μ and a covariance matrix Σ, and is defined as:

$$p(x \mid \mu, \Sigma) = (2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

where $x \in \mathbb{R}^D$.

Notation:

- We write $p(x) = \mathcal{N}(x \mid \mu, \Sigma)$ or $X \sim \mathcal{N}(\mu, \Sigma)$.

Properties:

- $\mu \in \mathbb{R}^D$ is the mean vector, representing $\mathbb{E}[X] = \mu$.
- $\Sigma \in \mathbb{R}^{D \times D}$ is the covariance matrix, representing $\text{Cov}[X] = \Sigma$.
- The distribution is fully determined by its mean and covariance.

Bernoulli Distribution

Definition (Bernoulli Distribution)

The Bernoulli distribution is a distribution for a single binary random variable X with state $x \in \{0,1\}$. It is governed by a single continuous parameter $\mu \in [0,1]$ that represents the probability of $X = 1$. The Bernoulli distribution Ber(μ) is defined as:

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x}, \quad x \in \{0,1\}$$

Properties:

- **Mean (Expected Value):**

$$\mathbb{E}[X] = \mu$$

- **Variance:**

$$\text{Var}[X] = \mu(1-\mu)$$

Binomial Distribution

Definition (Binomial Distribution)

The Binomial distribution is a generalization of the Bernoulli distribution to a distribution over integers. It describes the probability of observing m occurrences of $X = 1$ in a set of N independent trials, where each trial is a Bernoulli experiment with success probability $\mu \in [0, 1]$. The Binomial distribution $\text{Bin}(N, \mu)$ is defined as:

$$p(m \mid N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

Properties:

- **Mean (Expected Value):** $\mathbb{E}[m] = N\mu$
- **Variance:** $\text{Var}[m] = N\mu(1 - \mu)$

Binomial Distribution – Plot

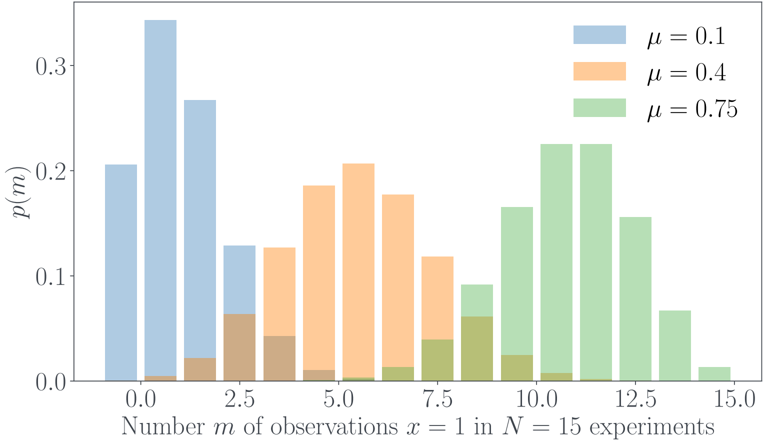


Figure: Taken from Deisenroth, Faisal, and Ong (2020).

Live Voting

Optimization

2

Loss Functions

2.1

Loss Functions

- Loss or Cost or Error functions measure how close an output or prediction is to the true, expected value, and therefore these functions map to a scalar
- Loss functions are designed to be minimized:
  Maximize likelihood $\rightarrow$ minimize negative likelihood
- In our function network, loss functions are usually at the top
- Loss functions are needed for optimization (as part of our objective) and should thus be differentiable

Loss Functions

Depending on the task, we either perform *classification* or *regression*.

Classification

- Prediction is a class, e.g., animal given image, emotion given speech, word given a vocabulary
- Binary classification: Two options
- Multi-class classification: Multiple (>2) options
- Multi-label classification: Classes are not exclusive, i.e., one can select one or more classes

Regression

- Prediction is a scalar, e.g., house price or similarity of two phrases

Example: Classification - Low loss

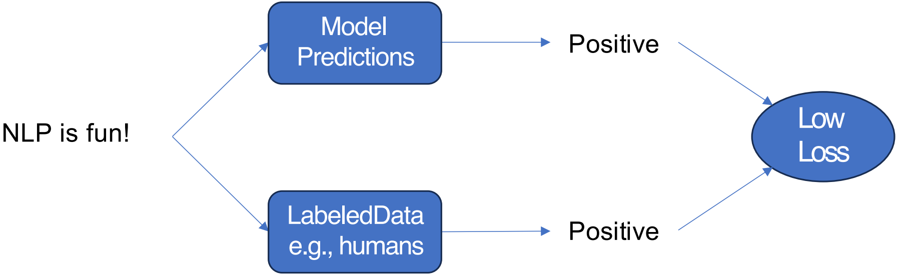


Figure: Loss of a correct prediction

Example: Classification - High loss

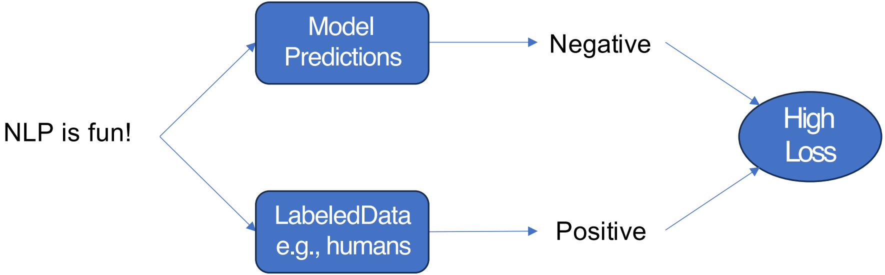


Figure: Loss of a wrong prediction

Common Loss Functions

Given predictions $y \in \mathbb{R}$ and expected values (labels) $\hat{y}$ there are different possibilities to compute the loss.

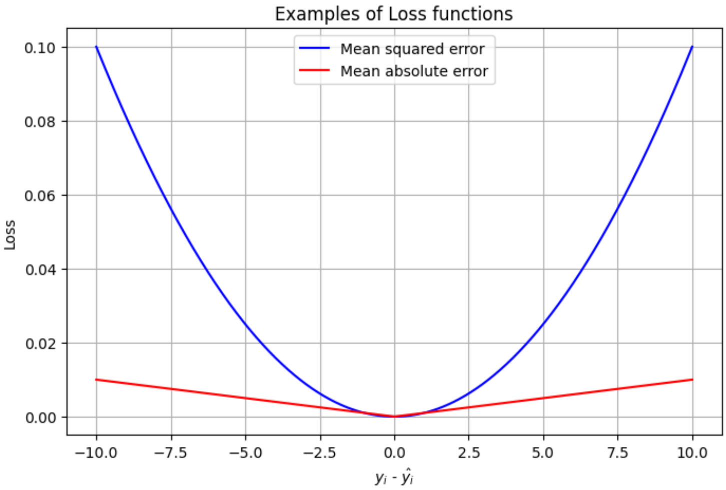
Regression:

- Mean Squared Error (MSE): $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ with $\hat{y} \in \mathbb{R}^n$
- Mean Absolute Error (MAE): $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ with $\hat{y} \in \mathbb{R}^n$

Classification:

- Binary Cross-Entropy (CE): $-(\hat{y}_i log(y_i) + (1 - \hat{y}_i) log(1 - y_i))$ with $\hat{y}_i \in \{0, 1\}$
- Hinge Loss (SVM): $max(0, 1 - \hat{y}_i y_i)$ with $\hat{y}_i \in \{-1, 1\}$

Example: Loss functions



Live Voting

Gradient Descent

2.2

Optimization

- Given: a *model* as a function f with parameters $\theta : f_\theta$
- Data $\mathcal{D}$, e.g. points $\{(x_1, y_1), \ldots (x_N, y_N)\}$
- Goal: Find optimal model $\rightarrow$ parameters θ^* for which the error c is minimal:

$$\theta^* = \arg\min_\theta c(f_\theta, \mathcal{D})$$

- If c is convex (exactly one minimum), there exists an analytic solution:

$$\frac{d}{d\theta} c \stackrel{!}{=} 0 \Rightarrow \theta^* = \ldots$$

Gradient Descent

Otherwise, gradient-based algorithms can be used:

- We know: the gradient points into the direction of steepest ascent
- $\rightarrow$ We can improve the parameters θ by following the negative gradient for a value of c.
- Therefore, given our data, we can
  - compute the loss function c for our current model
  - compute the derivative of the loss: $\frac{dc}{d\theta}$
  - update the parameters θ of our model by shifting them a bit into the direction of the negative gradient

$$\theta = \theta - \eta \cdot \nabla f^{\top}$$

where η is the step size

Gradient Descent

- We repeat this process until convergence, i.e., until model parameters θ (almost) do not change anymore
- Note that we care about the direction of the gradient, but its magnitude is arbitrary, and hence one might want to normalize the gradient:

$$\theta = \theta - \eta \cdot \frac{\nabla f^\top}{|\nabla f|}$$

Gradient Descent

The step size is important:

- Too small $\rightarrow$ slow, might get stuck in local minimum.
- Too large $\rightarrow$ can miss global minimum, diverge.
- Convergence depends on the cost function, the optimizer, the initial parameters and the data.
- Optimizing neural networks is difficult since the loss function can become very complex.

Examples

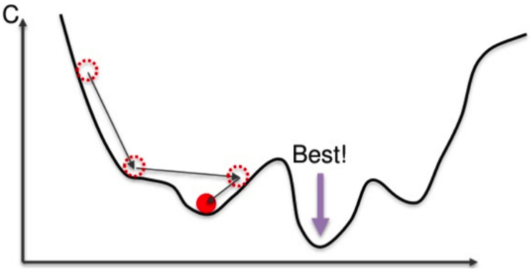


Figure: Gradient descent stuck in local minimum

(source: https://www.superdatascience.com/blogs/artificial-neural-networks-stochastic-gradient-descent)

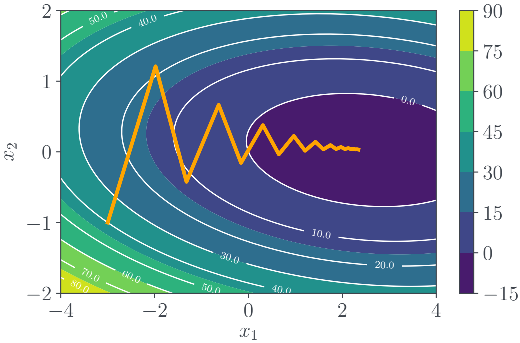


Figure: Gradient descent on 2D function (Deisenroth, Faisal, and Ong 2020)

Optimizing Neural Networks

Neural networks can be become arbitrarily complex, so can the loss function:

- Ideally, we integrate all training data into the batched gradient:

$$\nabla f = \sum_{n=1}^{N} \nabla f_n$$

- This is often infeasible, and we instead approximate the gradient using a subset of our training data; this is called stochastic mini-batch gradient descent
- There exist many other optimizers, e.g., using momentum or adaptive learning rate

Live Voting

Thanks for your attention - Questions?

References

Deisenroth, Marc Peter, A. Aldo Faisal, and Cheng Soon Ong (2020). *Mathematics for Machine Learning*. Cambridge University Press. DOI: 10.1017/9781108679930.