Mathematics for Machine Learning

Vazgen Mikayelyan

August 27, 2020

FAST DISCOVERING THE FUTURE

Differential

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^2$ and (x_0, y_0) is an interior point of X.

Differential

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^2$ and (x_0, y_0) is an interior point of X.

Definition

f is called differentiable at the point (x_0, y_0) if there exists $A, B \in \mathbb{R}$ such that

$$f(x_0 + \Delta x, y_0 + \Delta y) = f(x_0, y_0) + A\Delta x + B\Delta y + o(\rho), \rho \to 0,$$

where $\rho = \sqrt{\Delta x^2 + \Delta y^2}$.

Differential

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^2$ and (x_0, y_0) is an interior point of X.

Definition

f is called differentiable at the point (x_0, y_0) if there exists $A, B \in \mathbb{R}$ such that

$$f(x_0 + \Delta x, y_0 + \Delta y) = f(x_0, y_0) + A\Delta x + B\Delta y + o(\rho), \rho \to 0,$$

where $\rho = \sqrt{\Delta x^2 + \Delta y^2}$.

Theorem

If partial derivatives of the first degree of f are continuous at (x_0, y_0) then it is differentiable at (x_0, y_0). The inverse is not true.

Differential

Example

$$f(x,y) = \begin{cases} x^2 \sin \dfrac{1}{x}, \ \textit{if } (x,y) \neq (0,0), \\ 0, \ \textit{if } (x,y) = (0,0) \end{cases}$$

Differential

Example

$$f(x,y) = \begin{cases} x^2 \sin \dfrac{1}{x}, \ \textit{if } (x,y) \neq (0,0), \\ 0, \ \textit{if } (x,y) = (0,0) \end{cases}$$

Definition

$$df(x_0,y_0) = \frac{\partial f}{\partial x}(x_0,y_0)\,\Delta x + \frac{\partial f}{\partial y}(x_0,y_0)\,\Delta y \ \textit{is called differential of } f.$$

Directional Derivative

Theorem

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$ and x_0 is an interior point of X. If f is differentiable at x_0, then for all $v \in \mathbb{R}^n$ such that $\|v\| = 1$ we have

$$\lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t} = \frac{\partial f}{\partial x_1}(x_0)\, v_1 + \ldots + \frac{\partial f}{\partial x_n}(x_0)\, v_n$$

Directional Derivative

Theorem

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$ and x_0 is an interior point of X. If f is differentiable at x_0, then for all $v \in \mathbb{R}^n$ such that $\|v\| = 1$ we have

$$\lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t} = \frac{\partial f}{\partial x_1}(x_0)\, v_1 + \ldots + \frac{\partial f}{\partial x_n}(x_0)\, v_n = \nabla f(x_0) \cdot v.$$

Directional Derivative

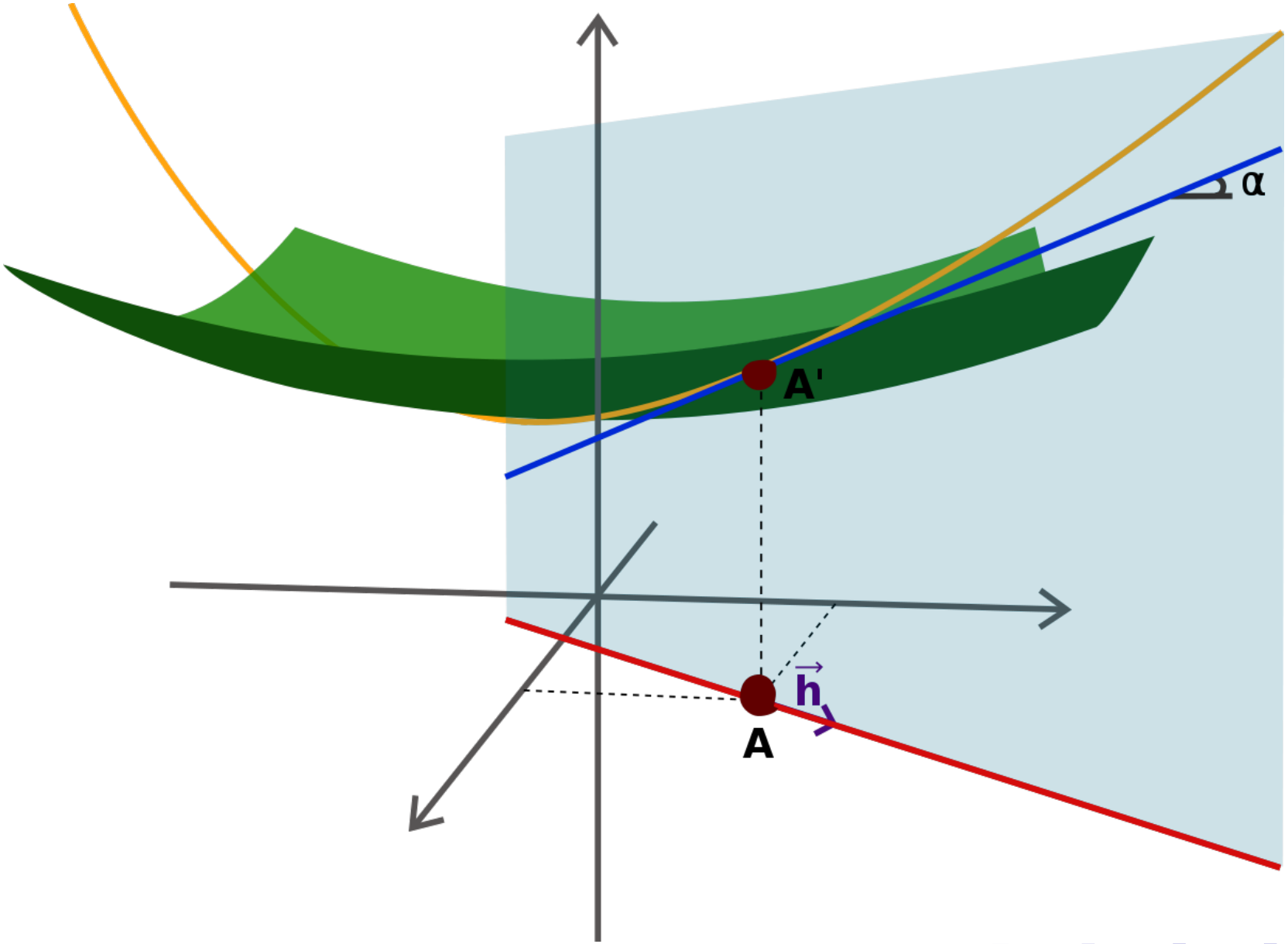
Theorem

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$ and x_0 is an interior point of X. If f is differentiable at x_0, then for all $v \in \mathbb{R}^n$ such that $\|v\| = 1$ we have

$$\lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t} = \frac{\partial f}{\partial x_1}(x_0)\, v_1 + \ldots + \frac{\partial f}{\partial x_n}(x_0)\, v_n = \nabla f(x_0) \cdot v.$$

If the limit of right hand side exists, it is called directional derivative of f and is denoted by $\frac{\partial f}{\partial v}$.

Directional Derivative



Extrema of Functions of Many Variables

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$ and x_0 is an interior point of X.

Extrema of Functions of Many Variables

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$ and x_0 is an interior point of X.

Definition

x_0 is called a local maximum (minimum) point of f if there exists a ball $B(x_0, \delta)$ such that $f(x) \leq f(x_0)$ $(f(x) \geq f(x_0))$ for all $x \in B(x_0, \delta)$.

Extrema of Functions of Many Variables

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$ and x_0 is an interior point of X.

Definition

x_0 is called a local maximum (minimum) point of f if there exists a ball $B(x_0, \delta)$ such that $f(x) \leq f(x_0)$ $(f(x) \geq f(x_0))$ for all $x \in B(x_0, \delta)$.

Theorem

If x_0 is a local extremum point of f and there exists $\nabla f(x_0)$, then $\nabla f(x_0) = 0$.

Extrema of Functions of Many Variables

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$ and x_0 is an interior point of X.

Definition

x_0 is called a local maximum (minimum) point of f if there exists a ball $B(x_0, \delta)$ such that $f(x) \leq f(x_0)$ $(f(x) \geq f(x_0))$ for all $x \in B(x_0, \delta)$.

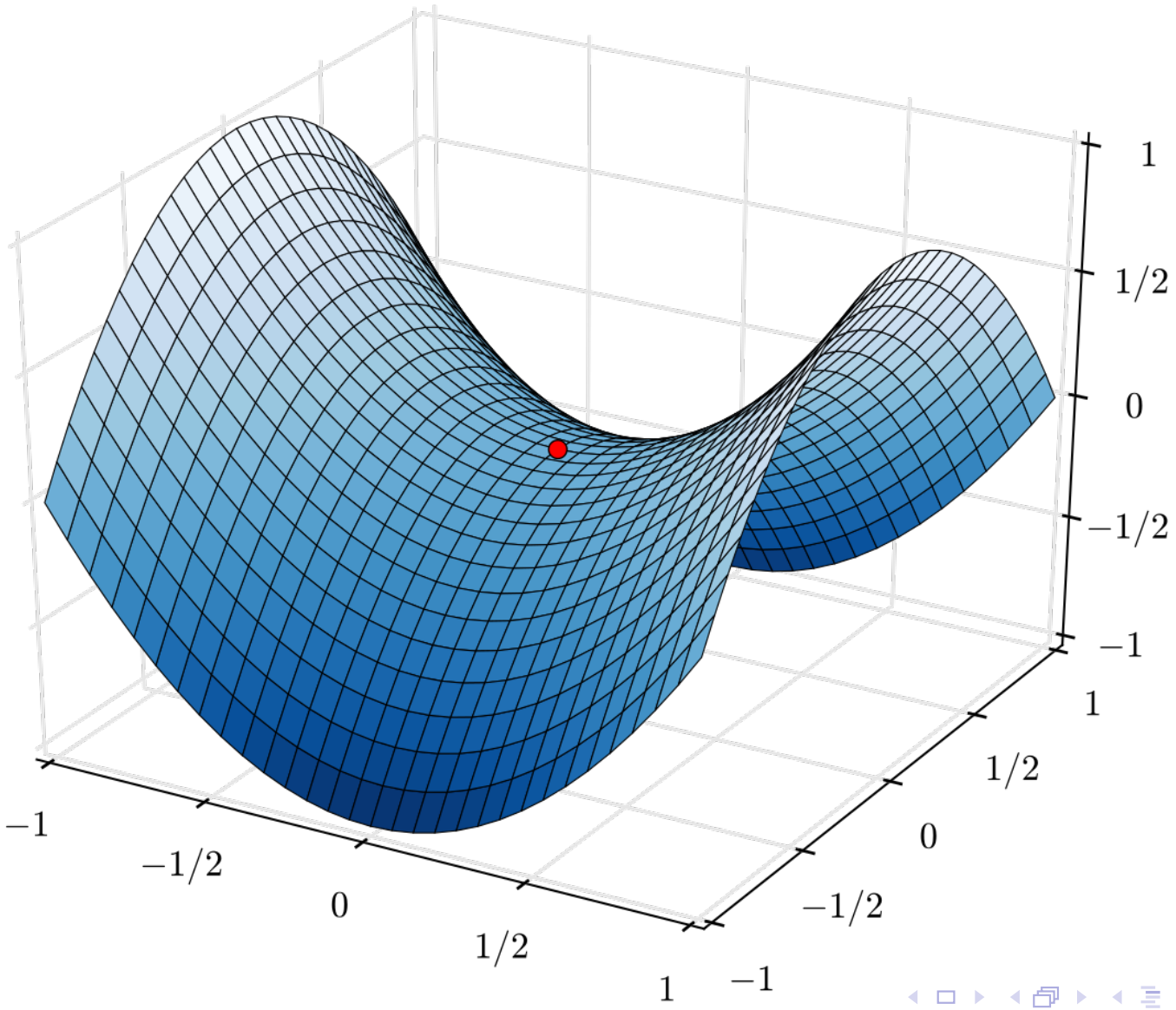
Theorem

If x_0 is a local extremum point of f and there exists $\nabla f(x_0)$, then $\nabla f(x_0) = 0$.

Definition

x_0 is called a saddle point of f if $\nabla f(x_0) = 0$ but x_0 is not an local extremum point of f.

Extrema of Functions of Many Variables



Extrema of Functions of Many Variables

Definition

Let $f : X \to \mathbb{R}$ and $X \subset \mathbb{R}^n$. If all second partial derivatives of f exist and are continuous over the domain of the function, then the matrix $H_{i,j} = \dfrac{\partial^2 f}{\partial x_i \partial x_j}$, $i, j = 1, \ldots, n$ is called the Hesian matrix of f.

Extrema of Functions of Many Variables

Definition

Let $f : X \to \mathbb{R}$ and $X \subset \mathbb{R}^n$. If all second partial derivatives of f exist and are continuous over the domain of the function, then the matrix $H_{i,j} = \dfrac{\partial^2 f}{\partial x_i \partial x_j}$, $i, j = 1, \ldots, n$ is called the Hesian matrix of f.

Theorem

Hesian matrix is symmetric.

Extrema of Functions of Many Variables

Definition

Let $f : X \to \mathbb{R}$ and $X \subset \mathbb{R}^n$. If all second partial derivatives of f exist and are continuous over the domain of the function, then the matrix $H_{i,j} = \dfrac{\partial^2 f}{\partial x_i \partial x_j}$, $i, j = 1, \ldots, n$ is called the Hesian matrix of f.

Theorem

Hesian matrix is symmetric.

Theorem

If f is convex, then its Hesian matrix is positive semi-definite.

Extrema of Functions of Many Variables

Theorem

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$ and x_0 is a critical point of f. If all second partial derivatives of f exist and are continuous at x_0 then

1. *if $H(x_0)$ is positive definite, then f attains a local minimum at x_0,*

Extrema of Functions of Many Variables

Theorem

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$ and x_0 is a critical point of f. If all second partial derivatives of f exist and are continuous at x_0 then

1. *if $H(x_0)$ is positive definite, then f attains a local minimum at x_0,*
2. *if $H(x_0)$ is negative definite, then f attains a local maximum at x_0,*

Extrema of Functions of Many Variables

Theorem

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$ and x_0 is a critical point of f. If all second partial derivatives of f exist and are continuous at x_0 then

1. *if $H(x_0)$ is positive definite, then f attains a local minimum at x_0,*
2. *if $H(x_0)$ is negative definite, then f attains a local maximum at x_0,*
3. *if $H(x_0)$ has both positive and negative eigenvalues then x_0 is a saddle point for f.*

Gradient Descent

Let $f : \mathbb{R}^k \to \mathbb{R}$ be a convex function and we want to find its global minimum.

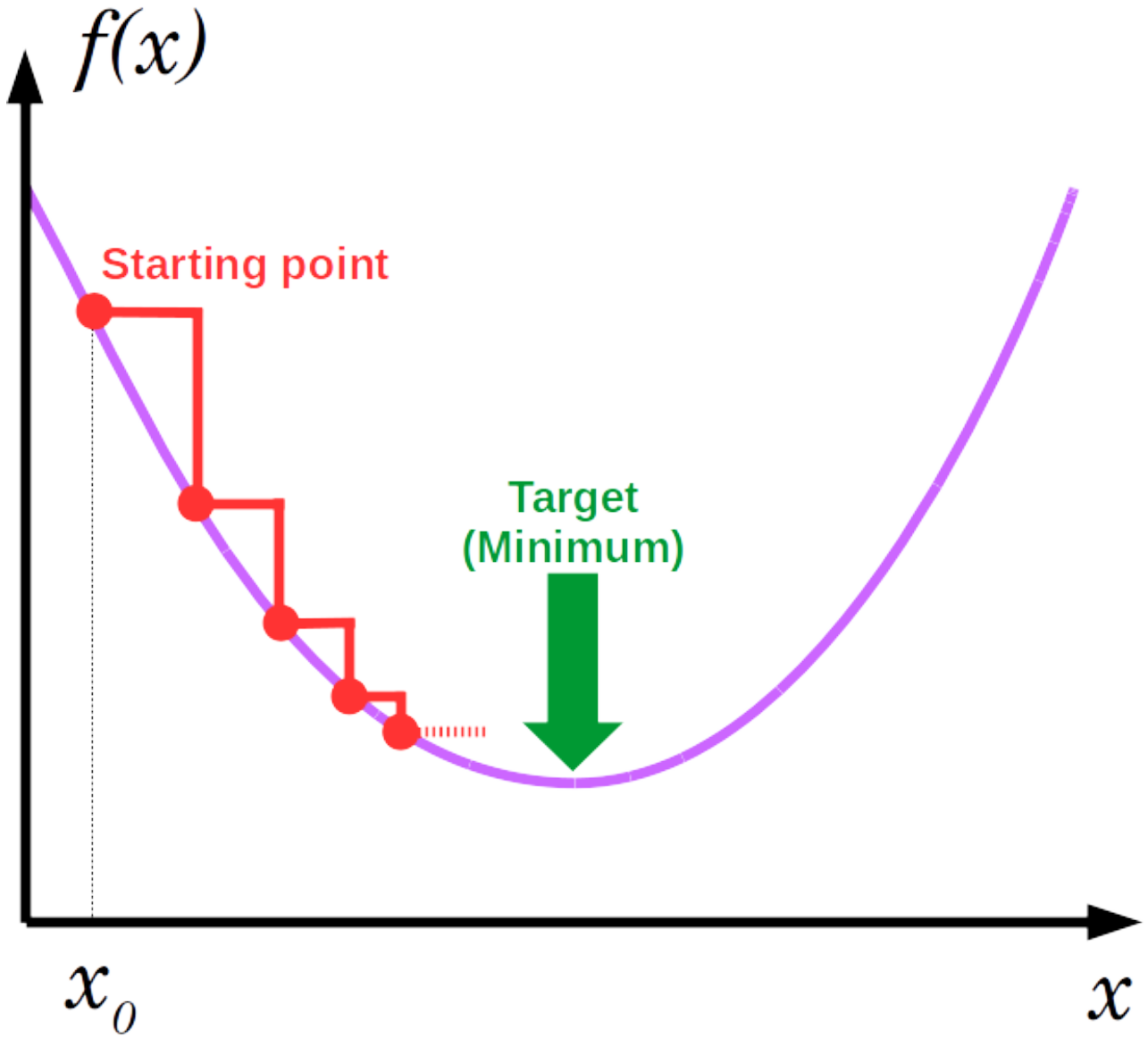
Gradient Descent

Let $f : \mathbb{R}^k \to \mathbb{R}$ be a convex function and we want to find its global minimum. This optimization algorithm is based on the fact that the fastest decreasing direction of the function is the opposite direction of gradient:

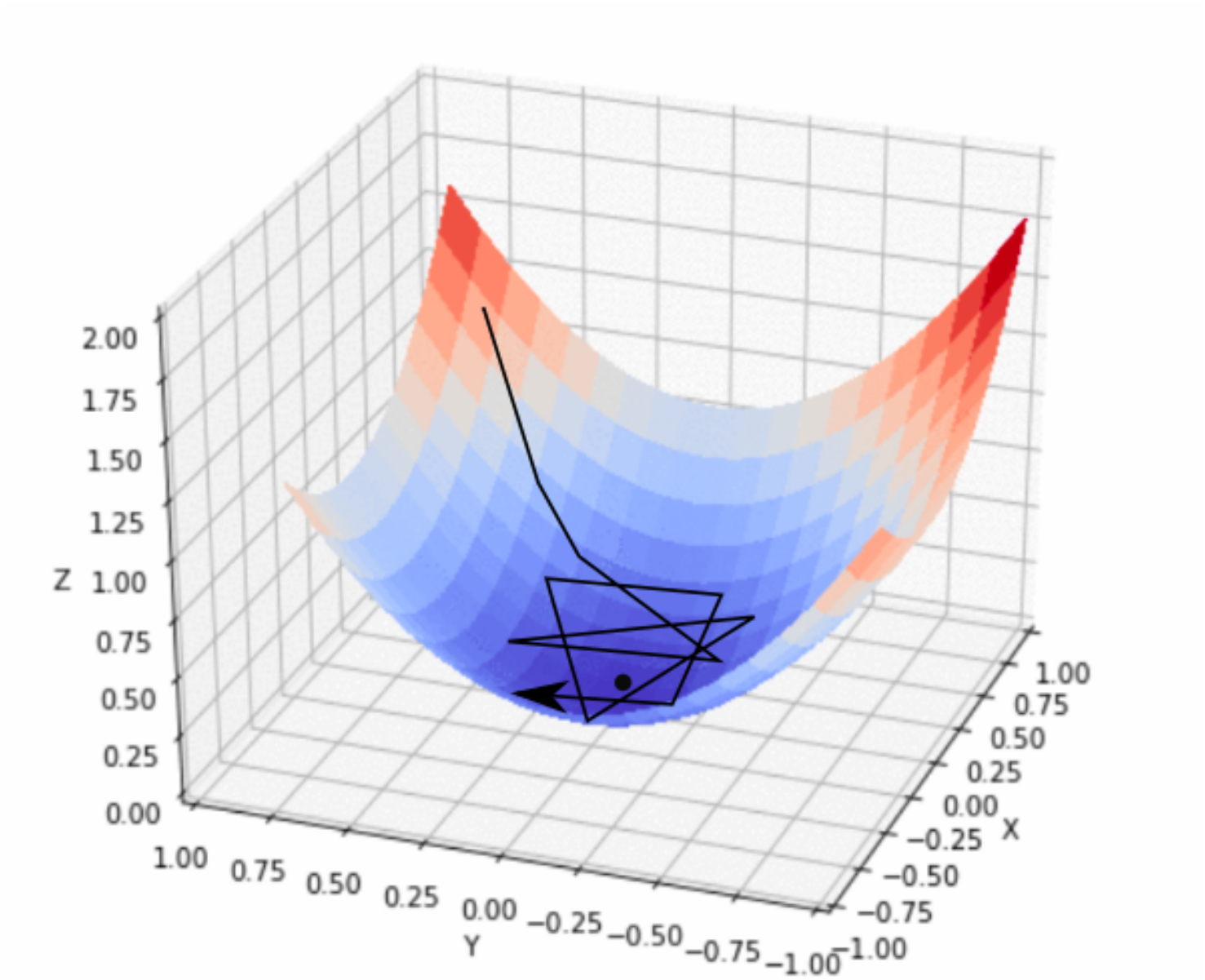
$$x_{n+1} = x_n - \alpha \nabla f\left(x_n\right)$$

and $x_0 \in \mathbb{R}^k$ is a arbitrary point and $\alpha > 0$.

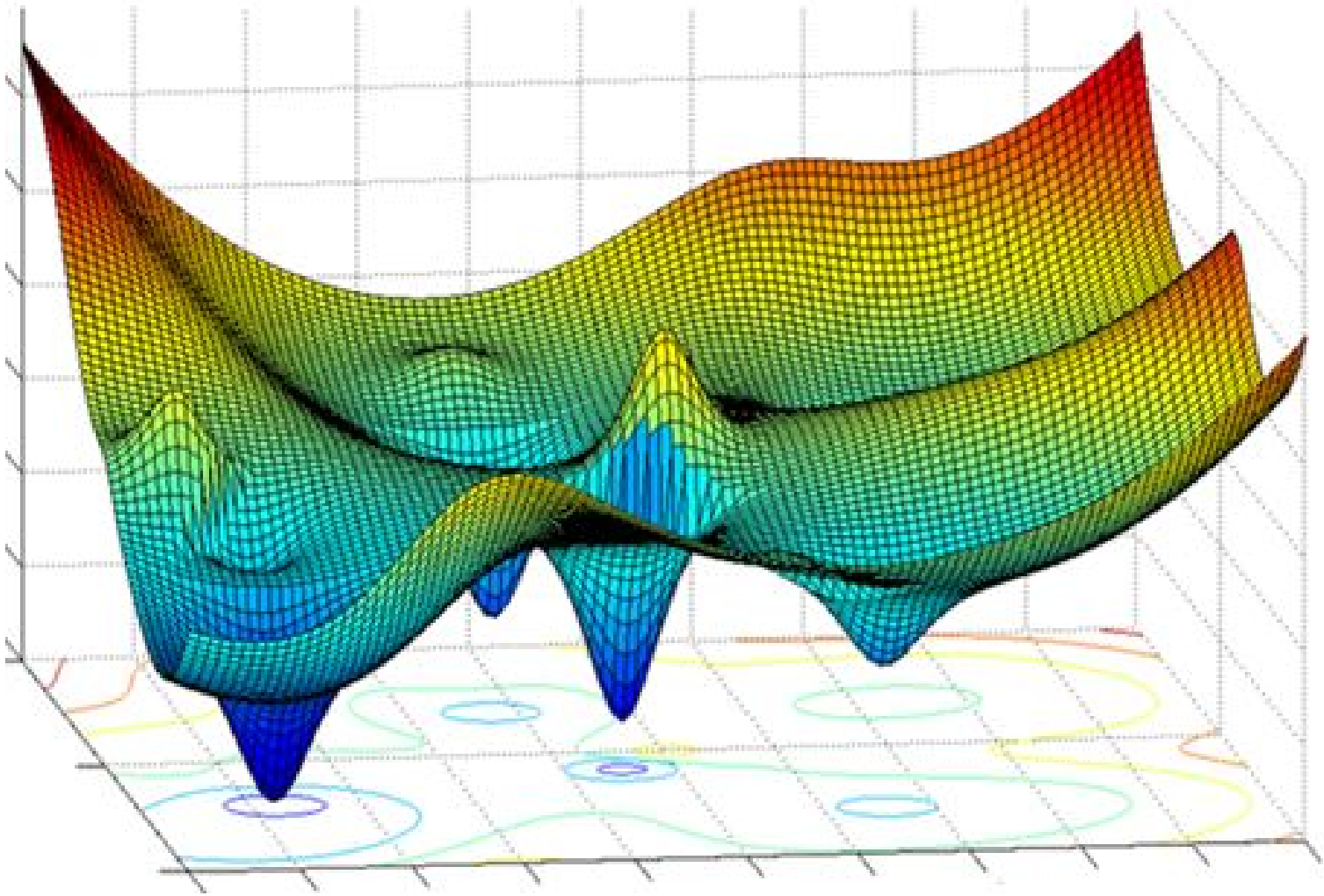
Gradient Descent



Gradient Descent



Gradient Descent



Probability

Experiment, Outcomes and the Sample Space

Experiment, Outcomes and the Sample Space

- A **random (or probabilistic) Experiment** is a situation, where we are uncertain about the result.

Experiment, Outcomes and the Sample Space

- A **random (or probabilistic) Experiment** is a situation, where we are uncertain about the result.
- An **Outcome** is a possible result of an Experiment.

Experiment, Outcomes and the Sample Space

- A **random (or probabilistic) Experiment** is a situation, where we are uncertain about the result.
- An **Outcome** is a possible result of an Experiment.
- The set of all Outcomes of an Experiment is called the **Sample Space** of that Experiment:

Experiment, Outcomes and the Sample Space

- A **random (or probabilistic) Experiment** is a situation, where we are uncertain about the result.
- An **Outcome** is a possible result of an Experiment.
- The set of all Outcomes of an Experiment is called the **Sample Space** of that Experiment:

$$\Omega = \text{the Sample Space of the Experiment} =$$

$$= \text{the set of all outcomes of our Experiment}$$

Examples

- Our Experiment: we are tossing a (fair) coin.

Examples

- Our Experiment: we are tossing a (fair) coin.
- **Heads** is one of the outcomes.

Examples

- Our Experiment: we are tossing a (fair) coin.
- **Heads** is one of the outcomes.
- The Sample Space in this Example is:

$$\Omega = \text{Sample Space} = \{\text{Heads, Tails}\} = \{H, T\}$$

Examples

- Experiment: we are rolling a (fair) die.

Examples

- Experiment: we are rolling a (fair) die.
- One of the outcomes is 3.

Examples

- Experiment: we are rolling a (fair) die.
- One of the outcomes is 3.
- The Sample Space in this Example is: $\{1, 2, 3, 4, 5, 6\}$

Examples

- Experiment: we are interested in the remaining lifetime (in years) of a person (for insurance reasons, say).

Examples

- Experiment: we are interested in the remaining lifetime (in years) of a person (for insurance reasons, say).
- One of the outcomes is 30.1.

Examples

- Experiment: we are interested in the remaining lifetime (in years) of a person (for insurance reasons, say).
- One of the outcomes is 30.1.
- The Sample Space in this Example is: $[0, 150]$

Events Examples

- Experiment: Rolling a die

Events Examples

- Experiment: Rolling a die
- Sample Space $= \Omega = \{1, 2, 3, 4, 5, 6\}$

Events Examples

- Experiment: Rolling a die
- Sample Space $= \Omega = \{1, 2, 3, 4, 5, 6\}$
- Some Events:
  - The Result is Odd $= \{1, 3, 5\}$

Events Examples

- Experiment: Rolling a die
- Sample Space $= \Omega = \{1, 2, 3, 4, 5, 6\}$
- Some Events:
  - The Result is Odd $= \{1, 3, 5\}$
  - The Result is larger than 2 $= \{3, 4, 5, 6\}$

Events Examples

- Experiment: Rolling a die
- Sample Space $= \Omega = \{1, 2, 3, 4, 5, 6\}$
- Some Events:
  - The Result is Odd $= \{1, 3, 5\}$
  - The Result is larger than 2 $= \{3, 4, 5, 6\}$
  - Any Result $= \Omega$

Events Examples

- Experiment: Rolling a die
- Sample Space $= \Omega = \{1, 2, 3, 4, 5, 6\}$
- Some Events:
  - The Result is Odd $= \{1, 3, 5\}$
  - The Result is larger than 2 $= \{3, 4, 5, 6\}$
  - Any Result $= \Omega$
  - No Result $= \varnothing$

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train
- An example of an outcome: 3.24.

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train
- An example of an outcome: 3.24.
- Sample Space $= \Omega = [0, 20]$

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train
- An example of an outcome: 3.24.
- Sample Space $= \Omega = [0, 20]$
- It is not interesting to have the probability of one outcome: say, what is the probability that the waiting time will be 3.24312456231?

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train
- An example of an outcome: 3.24.
- Sample Space $= \Omega = [0, 20]$
- It is not interesting to have the probability of one outcome: say, what is the probability that the waiting time will be 3.24312456231? **Exactly**, the answer is 0.

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train
- An example of an outcome: 3.24.
- Sample Space $= \Omega = [0, 20]$
- It is not interesting to have the probability of one outcome: say, what is the probability that the waiting time will be 3.24312456231? **Exactly**, the answer is 0.
- So in this case we are interested in events' probabilities rather than in particular outcome probability.

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train
- An example of an outcome: 3.24.
- Sample Space $= \Omega = [0, 20]$
- It is not interesting to have the probability of one outcome: say, what is the probability that the waiting time will be 3.24312456231? **Exactly**, the answer is 0.
- So in this case we are interested in events' probabilities rather than in particular outcome probability.
- Some Events:
  - The WT is larger than $3 = (3, 20]$

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train
- An example of an outcome: 3.24.
- Sample Space $= \Omega = [0, 20]$
- It is not interesting to have the probability of one outcome: say, what is the probability that the waiting time will be 3.24312456231? **Exactly**, the answer is 0.
- So in this case we are interested in events' probabilities rather than in particular outcome probability.
- Some Events:
  - The WT is larger than $3 = (3, 20]$
  - The WT is between 2 and 5, included $= [2, 5]$

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train
- An example of an outcome: 3.24.
- Sample Space $= \Omega = [0, 20]$
- It is not interesting to have the probability of one outcome: say, what is the probability that the waiting time will be 3.24312456231? **Exactly**, the answer is 0.
- So in this case we are interested in events' probabilities rather than in particular outcome probability.
- Some Events:
  - The WT is larger than $3 = (3, 20]$
  - The WT is between 2 and 5, included $= [2, 5]$
  - The WT is anything $= \Omega$

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train
- An example of an outcome: 3.24.
- Sample Space $= \Omega = [0, 20]$
- It is not interesting to have the probability of one outcome: say, what is the probability that the waiting time will be 3.24312456231? **Exactly**, the answer is 0.
- So in this case we are interested in events' probabilities rather than in particular outcome probability.
- Some Events:
  - The WT is larger than $3 = (3, 20]$
  - The WT is between 2 and 5, included $= [2, 5]$
  - The WT is anything $= \Omega$
  - No Result $= \varnothing$

σ-algebra

Definition

Let Ω be some set and $\mathcal{F}$ be a set of some subsets of Ω.

σ-algebra

Definition

Let Ω be some set and $\mathcal{F}$ be a set of some subsets of Ω. $\mathcal{F}$ is called a σ-algebra if it satisfies the following three properties:

σ-algebra

Definition

Let Ω be some set and $\mathcal{F}$ be a set of some subsets of Ω. $\mathcal{F}$ is called a σ-algebra if it satisfies the following three properties:

- $\Omega \in \mathcal{F}$,

σ-algebra

Definition

Let Ω be some set and $\mathcal{F}$ be a set of some subsets of Ω. $\mathcal{F}$ is called a σ-algebra if it satisfies the following three properties:

- $\Omega \in \mathcal{F}$,
- *if* $A \in \mathcal{F}$, *then* $\Omega \setminus A \in \mathcal{F}$,

σ-algebra

Definition

Let Ω be some set and $\mathcal{F}$ be a set of some subsets of Ω. $\mathcal{F}$ is called a σ-algebra if it satisfies the following three properties:

- $\Omega \in \mathcal{F}$,
- *if* $A \in \mathcal{F}$, *then* $\Omega \setminus A \in \mathcal{F}$,
- *if* $A_n \in \mathcal{F}$, $n \in \mathbb{N}$, *then* $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Probability (Measure) Definition

Probability (Measure) Definition

Probability Measure Definition

A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called a **Probability Measure** on $(\Omega, \mathcal{F})$, if it satisfies the following axioms:

Probability (Measure) Definition

Probability Measure Definition

A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called a **Probability Measure** on $(\Omega, \mathcal{F})$, if it satisfies the following axioms:

P1. For any $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$;

Probability (Measure) Definition

Probability Measure Definition

A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called a **Probability Measure** on $(\Omega, \mathcal{F})$, if it satisfies the following axioms:

P1. For any $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$;

P2. $\mathbb{P}(\Omega) = 1$;

Probability (Measure) Definition

Probability Measure Definition

A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called a **Probability Measure** on $(\Omega, \mathcal{F})$, if it satisfies the following axioms:

P1. For any $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$;

P2. $\mathbb{P}(\Omega) = 1$;

P3. For any sequence of pairwise mutually exclusive (disjoint) events $A_n \in \mathcal{F}$, i.e., for any sequence $A_n \in \mathcal{F}$ with $A_i \cap A_j = \varnothing$ for $i \neq j$, we have

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Probability (Measure) Definition

Probability Measure is very similar (and shares the properties of) any other Measure -

- Cardinality (no. of elements),
- Length (in 1D),
- Area (in 2D),
- Volume (in 3D and moreD).

Probability (Measure) Definition

Probability Measure is very similar (and shares the properties of) any other Measure -

- Cardinality (no. of elements),
- Length (in 1D),
- Area (in 2D),
- Volume (in 3D and moreD).

The difference is only that the Probability of the Sample Space is 1, $\mathbb{P}(\Omega) = 1$.

Properties of the Probability Measure

1. $\mathbb{P}(\varnothing) = 0$;

Properties of the Probability Measure

1. $\mathbb{P}(\varnothing) = 0$;
2. if $A, B \in \mathcal{F}$ are mutually exclusive events, i.e., if $A \cap B = \varnothing$, then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B);$$

Properties of the Probability Measure

1. $\mathbb{P}(\varnothing) = 0$;
2. if $A, B \in \mathcal{F}$ are mutually exclusive events, i.e., if $A \cap B = \varnothing$, then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B);$$

3. for any event $A \in \mathcal{F}$,

$$\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A);$$

Here $\overline{A} = A^c = \Omega \setminus A$.

Properties of the Probability Measure

4. If $A_1, A_2, ..., A_n \in \mathcal{F}$ are pairwise disjoint (mutually exclusive), i.e., if $A_i \cap A_j = \varnothing$ for $i \neq j$, then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i);$$

Properties of the Probability Measure

4. If $A_1, A_2, ..., A_n \in \mathcal{F}$ are pairwise disjoint (mutually exclusive), i.e., if $A_i \cap A_j = \varnothing$ for $i \neq j$, then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i);$$

5. for any events $A, B \in \mathcal{F}$ (not necessarily disjoint),

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B);$$