Mathematics Basics

Optimization Problems

Contents

- Mathematics and AI
- Linear Algebra
- Probability and Statistics
- Optimization Problems
    - Classification of Optimization Problems
    - Gradient Descent Method
    - Newton's Method and Conjugate Gradient

Optimization Problems

- **Optimization problem**: a problem of changing the values of parameters(decision variables) x to minimize or maximize objective function $f(x)$, It can be represented by

$$x^* = arg\min_{x} f(x) \quad , x = (x_1, x_2, \cdots, x_n)^T \in R^n$$

$$s.t. \quad c_i(x) \geq 0, i = 1, 2, \cdots, m, \qquad \text{inequality constraint}$$

$$c_j(x) = 0, j = 1, 2, \cdots, p, \qquad \text{equality constraint}$$

- Constraints define a feasible region, which is nonempty.
- If we seek a maximum of $f(x)$ it is equivalent to seeking to a min of $-f(x)$.
- In an optimization problem, if there are no other constraints for each variable except for the objective function, then it is called an unconstrained optimization problem. Otherwise, it is called a constrained optimization problem.

Solutions to Optimization Problems

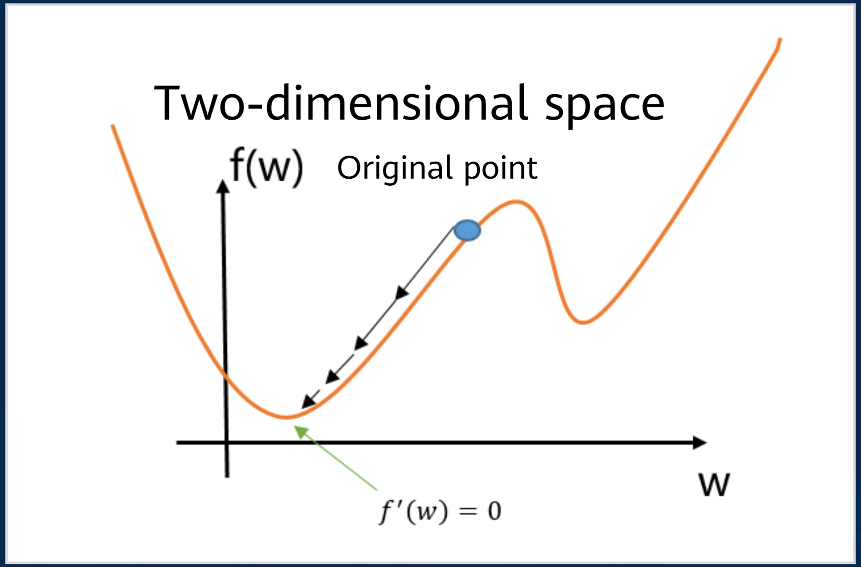
- **Solutions to unconstrained optimization**: mainly include analytical methods and direct methods.
  - Direct methods are usually used when the representation of an objective function is complicated or cannot be specified. Through numerical calculation in a series of iterative processes, a range of points will be generated for searching for an optimal point.
  - The analytical methods, also known as indirect methods, obtain the optimal solution based on the analytical expression of the objective function that an unconstrained optimization problem focuses on. The analytical methods mainly include gradient descent method, Newton's method, Quasi-Newton method, conjugate direction method, and conjugate gradient method.

Solutions to Optimization Problems

- **Solutions to constrained optimization:** The method of Lagrange multiplier is usually used in solving optimization problems subject to equality constraints, while the **Karush–Kuhn–Tucker (KKT)** approach is used in solving problems subject to inequality constraints. These methods turn constrained optimization problems involving n variables and k constraints into unconstrained optimization problems involving (n+k) variables.
- In this course, we focus on the most common solution to unconstrained optimization problems in deep learning, that , the gradient descent method and Newton method.

Extension to N dimensions

- How big N can be?
  - Problem sizes can vary from a handful of parameters to many thousands.
- We will consider examples for N=2, so that cost function surfaces can be visualized.

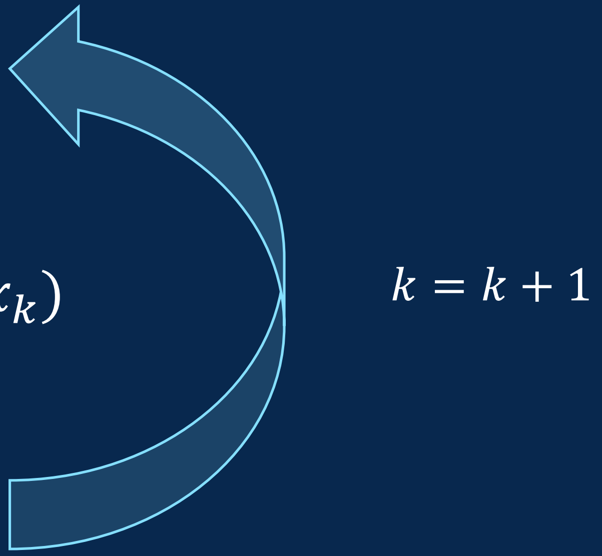


An optimization Algorithm

- Start at x_0, $\mathrm{k} = 0$.

1. Compute a search direction p_k.
2. Compute a step size α_k, such that $f(x_k + \alpha_k p_k) < f(x_k)$
3. Update $x_k = x_k + \alpha_k p_k$
4. Check for convergence(stopping criteria)

   e.g. $\nabla f(x) = 0$



$k = k + 1$

Contents

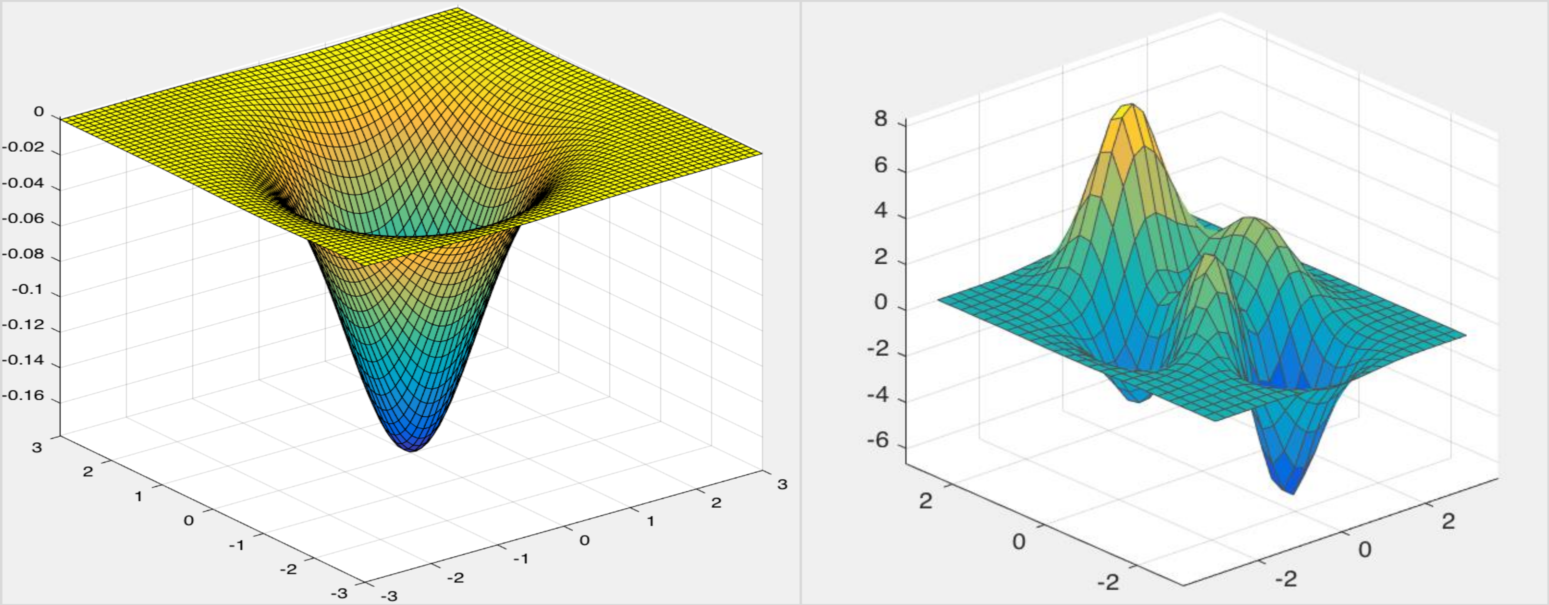
- ◆ Mathematics and AI
- ◆ Linear Algebra
- ◆ Probability and Statistics
- ◆ Optimization Problems
  - Classification of Optimization Problems
  - Gradient Descent Method
  - Newton's Method and Conjugate Gradient

Gradient Descent

- Convex function: If $\lambda \in (0,1)$ and any x_1, $x_2 \in R$ satisfy

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2),$$

$f(x)$ is called a convex function. The minima of a convex function appears at the stationary points.



Gradient Descent

- Basic principle is to minimize the N-dimensional function by a series of 1D line-minimizations:

$$x_{k+1} = x_k + \alpha_k p_k$$

- The gradient descent method chooses p_k to be parallel to the gradient

$$p_k = -\nabla f(x_k)$$

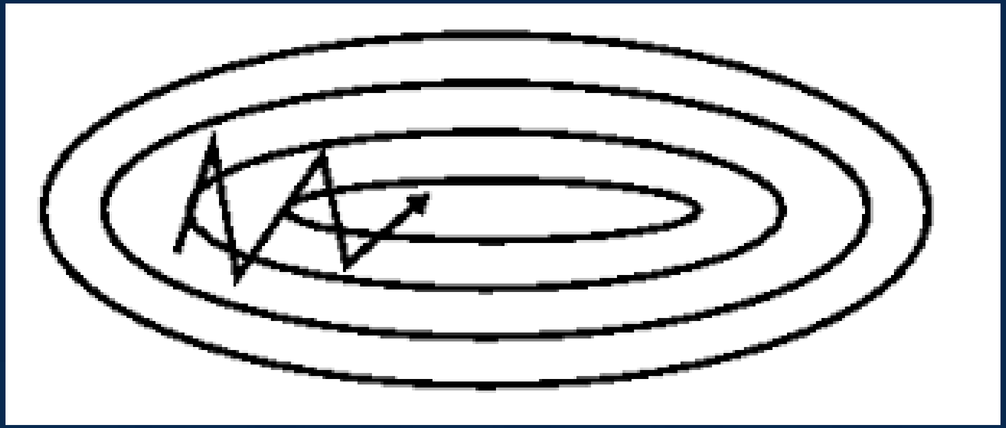
- Step size α_k is the **learning rate**, a positive scalar determining step chosen to minimize $f(x_k + \alpha_k p_k)$.

$$\alpha_k = \frac{p_k^T p_k}{p_k^T H p_k}$$

- Gradient descent converges when every element of the gradient is zero or close to zero.

Gradient Descent

- The gradient is everywhere perpendicular to the contour lines.
- After each line minimization the new gradient orthogonal to the previous step direction. Therefore the iterates tend to zig-zag down the valley.



Contents

◆ Mathematics and AI

◆ Linear Algebra

◆ Probability and Statistics

◆ Optimization Problems

- Classification of Optimization Problems
- Gradient Descent Method
- Newton's Method and Conjugate Gradient

Newton’s Method-1D

- Fit a quadratic approximation to $f(x)$ using both gradient and curvature information at x.
- Expand $f(x)$ locally using a Taylor series.

$$f(x+\delta x)=f(x)+f(\acute{x})\delta x+\frac{1}{2}f(\ddot{x})\delta x^2+o(\delta x^2)$$

- Find the δx which minimizes this local quadratic approximation.

$$\delta x=-\frac{f(\acute{x})}{f(\ddot{x})}$$

- Update x. $x_{n+1}=x_n-\delta x=x_n-\frac{f(\acute{x})}{f(\ddot{x})}$

Newton’s Method-N Dimension

- Expand $f(x)$ locally using a Taylor series x_k.

$$f(x_k + \delta x) = f(x_k) + g_{\mathrm{k}}^T \delta x + \frac{1}{2} \delta x^T H_k \delta x$$

Where the gradient is the vector

$$g_k = \nabla f(x_k) = \left[\frac{\partial f}{x_1} \dots \frac{\partial f}{x_N}\right]^T$$

And the Hessian is the symmetric matrix

$$H_k = H(x_k) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_N \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_N^2} \end{bmatrix}$$

Newton's Method-N Dimension

- For a minima we require that $\nabla f(x) = 0$, and so $\nabla f(x) = g_k + H_k \delta x = 0$
- With the solution $\delta x = -H_k^{-1} g_k$, this gives the iterative update

$$x_{k+1} = x_k - H_k^{-1} g_k$$

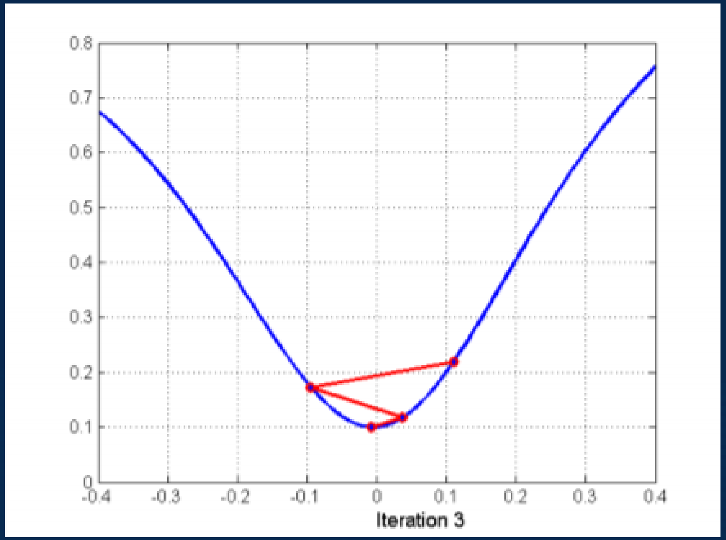
  - If $f(x)$ is quadratic, then the solution is found in one step.
  - The method has quadratic convergence(as in the 1D case).
  - The solution $\delta x = -H_k^{-1} g_k$ is guaranteed to be a downhill direction.
  - Rather than jump straight to the minimum, better to perform a line minimization which ensures global convergence

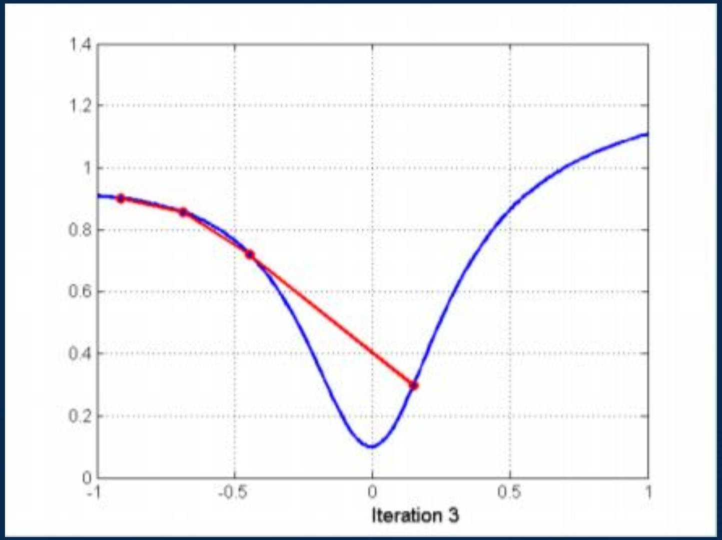
$$x_{k+1} = x_k - \alpha_k H_k^{-1} g_k$$

  - If $H = I$ then this reduces to gradient descent.

Newton's Method

- Quadratic convergence(decimal accuracy doubles at each iteration)
- Global convergence of Newton's method is poor if the starting point is too far from the minima.
- In practice, combined with a globalization strategy which reduces the step size until the function decrease is assured.





Conjugate Gradient

- Each direction p_k is chosen to be conjugate to all previous directions with respect to Hessian H:

$$p_i^T H p_j = 0, \mathrm{i} \neq j; \;\; p_k = \nabla f_k + \left(\frac{\nabla f_k^T \nabla f_k}{\nabla f_{k-1}^T \nabla f_{k-1}} \right) p_{k-1}$$

- Compute step size α_k for x_k at Hessian H_k. Set $x_{k+1} = x_k + \alpha_k d_k$ and calculate $f_{k+1} = f(x_{k+1})$.

$$\alpha_k = \frac{g_k^T g_k}{d_k^T H d_k}$$

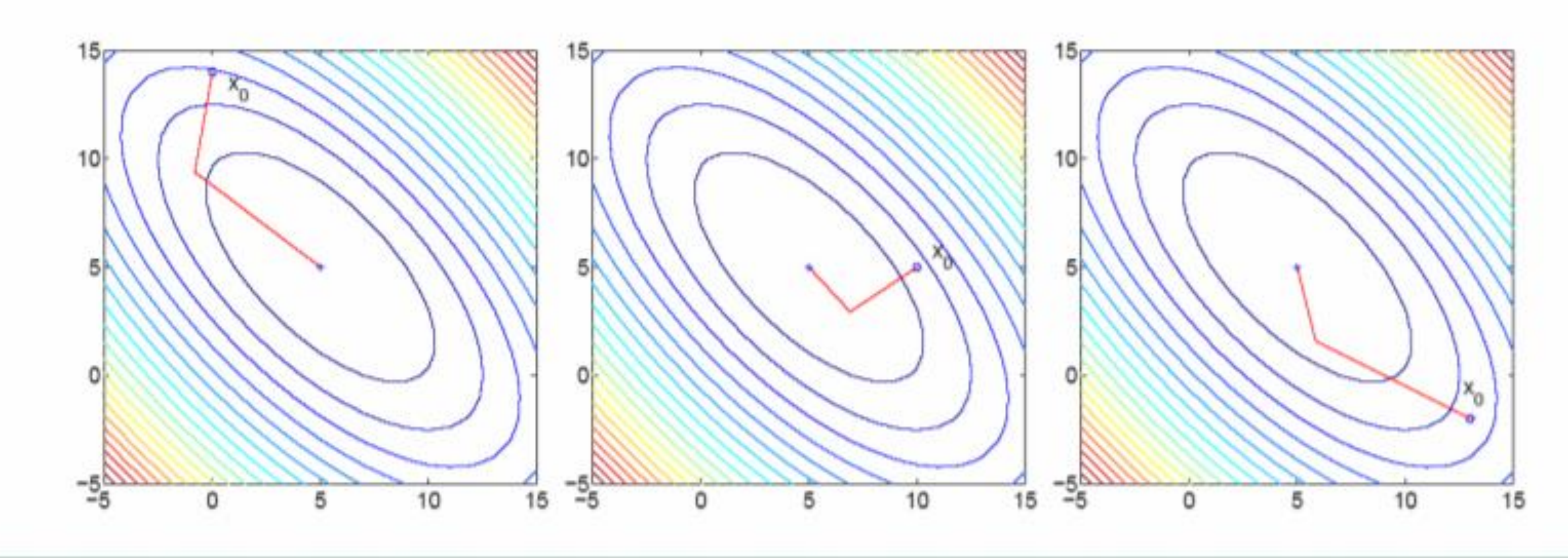
- If $\|\alpha_k d_k\| < \varepsilon$, output $x^* = x_{k+1}$ and $f(x^*) = f_{k+1}$and stop.

- Compute g_{k+1}. Compute $\beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$

- Generate new direction $d_{k+1} = -g_{k+1} + \beta_k d_k$

Conjugate Gradient

- An N-dimensional quadratic form can be minimized in at most N conjugate descent steps.



Summary

- This chapter mainly introduces the essential mathematics topics used in AI, including linear algebra, probability and statistics, as well as the optimization problems. It lays a foundation for other learning materials.

More Information

Huawei e-Learning website

- https://support.huawei.com/learning/en/newindex.html

Huawei support case library

- https://support.huawei.com/enterprise/en/index.html

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

HUAWEI