# مبانی یادگیری ماشین

# Intro to Machine Learning


**بابک نجار اعرابی**

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران


نیم سال اول سال تحصیلی 1403-04

موضوع این جلسه

# مروری بر روش های بهینه سازی

## جلسه چهارم

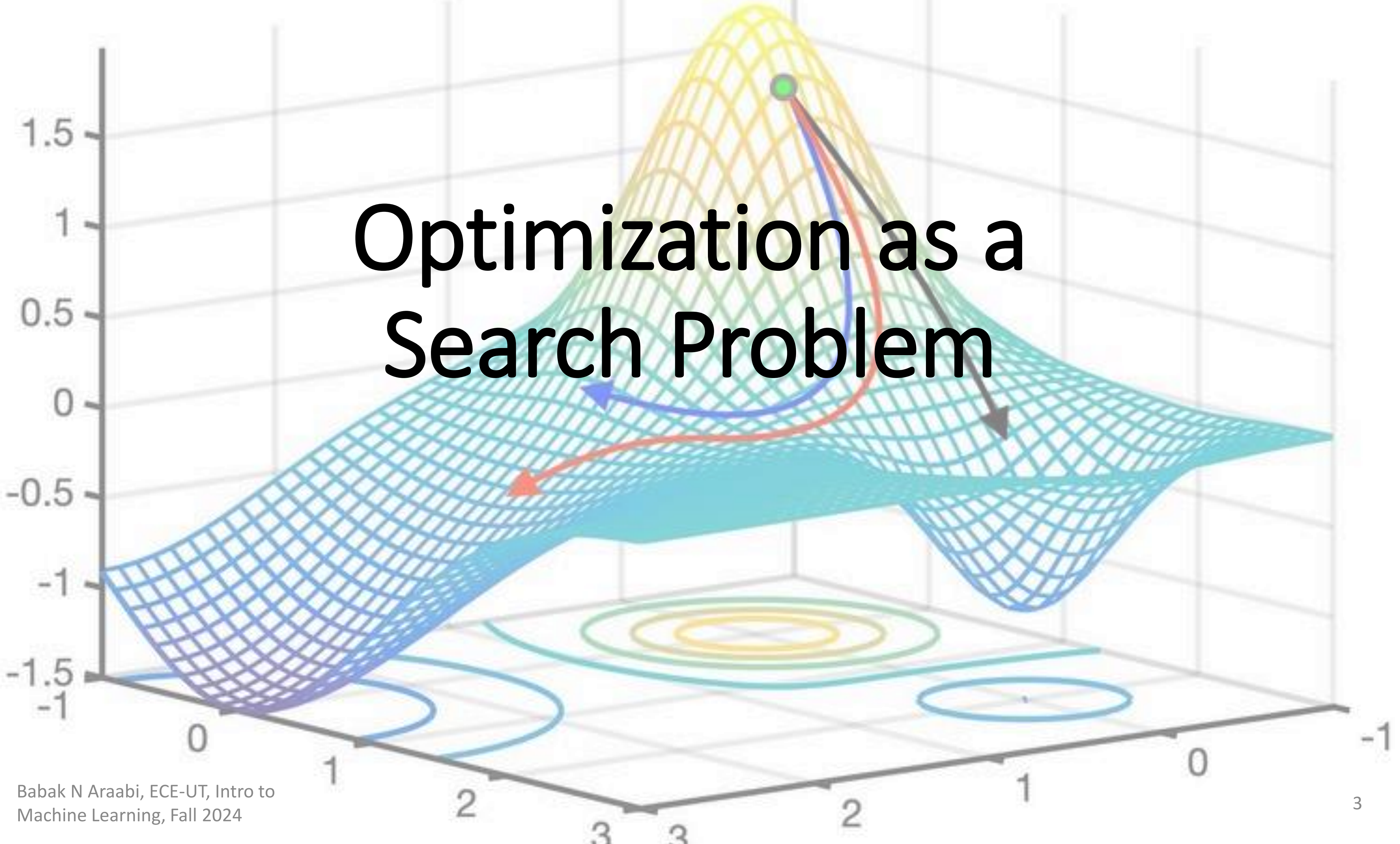
**Babak Nadjar Araabi**

School of Electrical & Computer Eng

University of Tehran

ECE-UT - Fall 2024

# Optimization as a Search Problem



# Nonlinear Programing Gradient Based Methods

for Local Optimization

# Gradient Descent Method (Steepest Descent)

for Local Optimization



https://en.wikipedia.org/wiki/Gradient_descent

# Newton Method

for Local Optimization



https://en.wikipedia.org/wiki/Newton%27s_method_in_optimization

# Quasi-Newton Method

for Local Optimization



https://en.wikipedia.org/wiki/Quasi-Newton_method

# Conjugate Gradient Method

for Local Optimization



https://en.wikipedia.org/wiki/Gauss%E2%80%93Newton_algorithm

# Nonlinear Least Squares

## Nonlinear Programing
## Gradient Based Methods

for Local Optimization

# Gauss-Newton Method

for Local Optimization

# Levenberg–Marquardt Method

for Local Optimization



https://en.wikipedia.org/wiki/Levenberg%E2%80%93Marquardt_algorithm

# Stochastic Gradient Descent

- **Stochastic approximation of gradient descent optimization**
- **Actual gradient:** calculated from the entire data set
- **Estimate gradient:** calculated from a randomly selected subset of the data (a batch of data)
- Especially in **high-dimensional optimization** problems this reduces the very high computational burden, achieving **faster iterations in exchange for a lower convergence rate**

https://en.wikipedia.org/wiki/Stochastic_gradient_descent

# SGD

- Minimizing an objective function that has the form of a sum:

$$Q(w) = \frac{1}{n}\sum_{i=1}^{n} Q_i(w),$$

where the parameter $w$ that minimizes $Q(w)$ is to be estimated. Each summand function $Q_i$ is typically associated with the $i$-th observation in the data set (used for training).

The sum-minimization problem also arises for empirical risk minimization. There, $Q_i(w)$ is the value of the loss function at $i$-th example, and $Q(w)$ is the empirical risk.

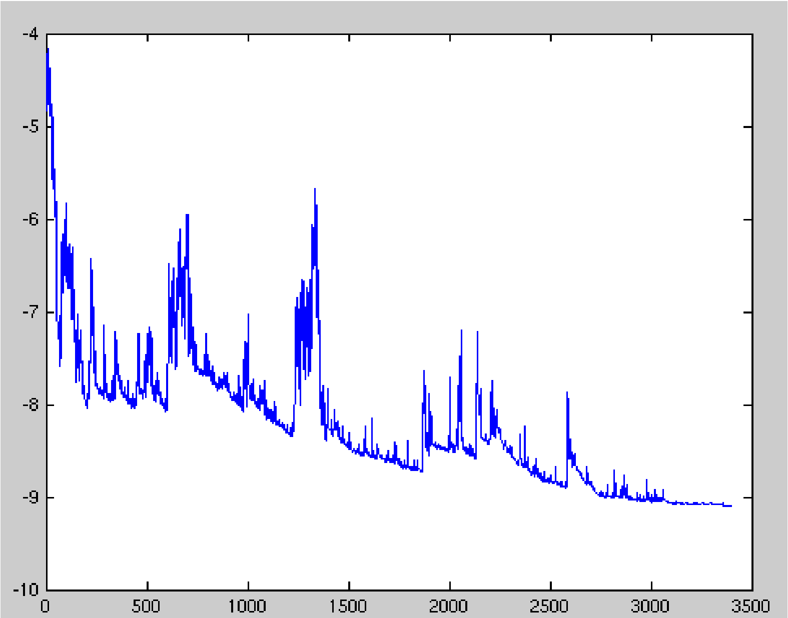
When used to minimize the above function, a standard (or "batch") gradient descent method would perform the following iterations:

$$w := w - \eta\nabla Q(w) = w - \frac{\eta}{n}\sum_{i=1}^{n} \nabla Q_i(w).$$

In stochastic (or "on-line") gradient descent, the true gradient of $Q(w)$ is approximated by a gradient at a single sample:

$$w := w - \eta \nabla Q_i(w).$$

Fluctuations in the total objective function as gradient steps with respect to mini-batches are taken.



- As the algorithm sweeps through the training set, it performs the above update for each training sample. Several passes (epochs) can be made over the training set until the algorithm converges. If this is done, the data can be shuffled for each pass to prevent cycles. Typical implementations may use an adaptive learning rate so that the algorithm converges

# Extensions and variants of SGD

- **Need to set a learning rate (step size)**
- **Implicit updates (ISGD)**
- **Momentum method (Heavy ball method)**
- **Averaging**
- **AdaGrad**
- **RMSProp**
- **Adam (Adaptive Moment Estimation)**

https://en.wikipedia.org/wiki/Stochastic_gradient_descent