

# Assignment Project Exam Help

Machine learning lecture slides

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

Optimization I: Convex optimization

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

- ▶ Convex sets and convex functions
- ▶ Local minimizers and global minimizers
- ▶ Gradient descent
- ▶ Analysis for smooth objective functions
- ▶
- ▶

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

# Convex sets

- Convex set: a set that contains every line segment between pairs of points in the set.

- Examples:

- All of  $\mathbb{R}^d$

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pro

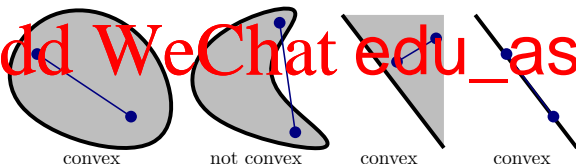


Figure 1: Which of these sets are convex?

## Convex functions (1)

- Convex function: a function satisfying the two-point version of Jensen's inequality:

$$f((1-\alpha)w + \alpha w') \leq (1-\alpha)f(w) + \alpha f(w'), \quad w, w' \in \mathbb{R}^d, \alpha \in [0, 1].$$

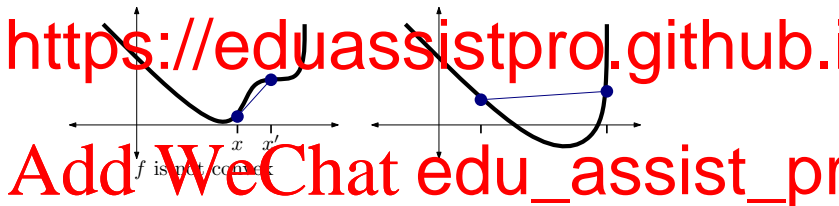


Figure 2: Which of these functions are c

## Convex functions (2)

### ► Examples:

►  $f(w) = c$  for  $c \in \mathbb{R}$

►  $f(w) = \exp(w)$  (on  $\mathbb{R}$ )

►  $f(w) = |w|^c$  for  $c \geq 1$  (on  $\mathbb{R}$ )

►  $f(w) = b^\top w$  for  $b \in \mathbb{R}^d$

<https://eduassistpro.github.io>

►  $f(w) = \text{logsumexp}(w) = \ln \left( \sum_i \exp(w_i) \right)$

►  $v \mapsto f(g(v))$  for convex function

Add WeChat [edu\\_assist\\_pro](#)

- Verify  $f(w) = \|w\|$  is convex

# Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

# Convexity of differentiable functions (1)

- Differentiable function  $f$  is convex iff

$$f(w) \geq f(w_0) + \nabla f(w_0)^\top (w - w_0) \quad \text{for all } w, w_0 \in \mathbb{R}^d.$$

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pro

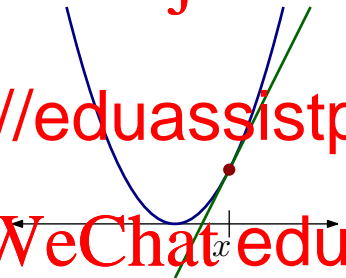


Figure 3: Affine approximation

- Twice-differentiable function  $f$  is convex iff  $\nabla^2 f(w)$  is positive semidefinite for all  $w \in \mathbb{R}^d$ .



## Convexity of differentiable functions (2)

- ▶ Example: Verify  $f(w) = w^4$  is convex
- ▶ Use second-order condition

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

## Convexity of differentiable functions (3)

- ▶ Example: Verify  $f(w) = e^{b^T w}$  for  $b \in \mathbb{R}^d$  is convex
- ▶ Use first-order condition

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

## Verifying convexity of least squares linear regression

- Verify  $f(w) = \|Aw - b\|_2^2$  is convex

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

## Verifying convexity of logistic regression MLE problem

- Verify  $f(w) = \frac{1}{n} \sum_{i=1}^n \ln(1 + e^{-y_i x_i^T w})$  is convex

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

## Local minimizers

- Say  $w^* \in \mathbb{R}^d$  is a local minimizer of  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  if there is an “open ball”  $U = \{w \in \mathbb{R}^d : \|w - w^*\|_2 < r\}$  of positive radius  $r > 0$  such that  $f(w^*) \leq f(u)$  for all  $u \in U$ .
- i.e., nothing looks better in the immediate vicinity of  $w^*$ .

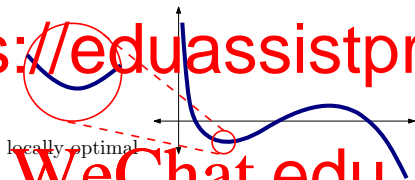


Figure 4: Local minimiz

## Local minimizers of convex problems

- ▶ If  $f$  is convex, and  $w^*$  is a local minimizer, then it is also a global minimizer.

- ▶ “Local-to-global” phenomenon

- ▶ Local search is well-motivated for convex optimization problems

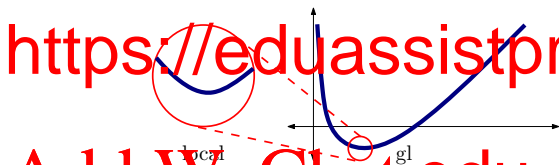


Figure 5: Local-to-global phe

- ▶ Consider (unconstrained) convex optimization problem

Assignment  $\min_{w \in \mathbb{R}^d} f(w)$ . Project Exam Help

- ▶ \_\_\_\_\_

- ▶ <https://eduassistpro.github.io>

- ▶  $w^{(t)} \leftarrow w^{(t-1)}$   
Add WeChat edu\_assist\_pro  
(Lots of things unspecified here ...)

## Motivation for gradient descent

- ▶ Why move in direction of (negative) gradient?
- ▶ Affine approximation of  $f(w + \delta)$  around  $w$ :

$$f(w + \delta) \approx f(w) + \nabla f(w)^T \delta.$$



- ▶ <https://eduassistpro.github.io>

$$\nabla f(w) - \eta \nabla f(w) \quad \eta \quad f(w) <$$

as long as  $\nabla f(w) \neq 0$ .

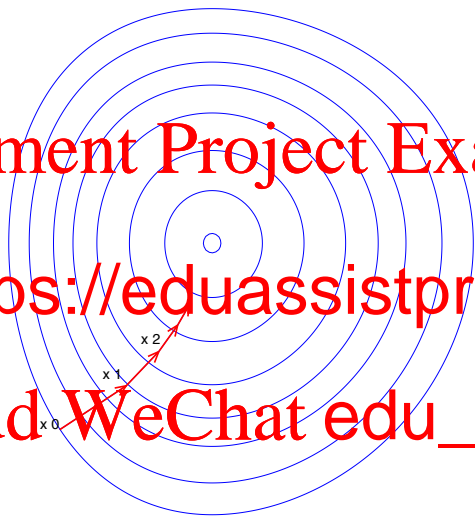
- ▶ Need  $\eta$  to be small enough so still have small error of affine approximation.



Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pro



The figure shows a series of concentric blue ellipses centered at the origin, representing the level sets of a cost function. A red line with arrows indicates the path of gradient descent, starting from an initial point  $x_0$  and moving towards the center. The path is labeled with  $x_0$ ,  $x_1$ , and  $x_2$  at successive points along the trajectory.

Figure 6: Trajectory of gradient descent

## Example: Gradient of logistic loss

- ▶ Negative gradient of logistic loss on  $i$ -th training example:  
using chain rule,

$$-\nabla \{\ell_{\text{logistic}}(y_i x_i^\top w)\} = -\ell_{\text{logistic}}(y_i x_i^\top w) y_i x_i$$

<https://eduassistpro.github.io>

where  $\sigma$  is the sigmoid function.

- ▶ Recall,  $\Pr_w(Y = y | X = x) = \sigma(y x^\top w)$  using the logistic regression model.

## Example: Gradient descent for logistic regression

- ▶ Objective function:

Assignment Project Exam Help

$$J(w) = \frac{1}{n} \sum_{i=1}^n \ell_{\text{logistic}}(y_i; x_i^\top w)$$

- ▶ <https://eduassistpro.github.io> <sup>step size</sup>

Add WeChat edu\_assist\_pr

$$\begin{aligned} w^{(t)} &:= w^{(t-1)} + \eta \nabla f(w^{(t-1)}) \\ &= w^{(t-1)} + \eta \frac{1}{n} \sum_{i=1}^n (1 - y_i x_i^\top w^{(t-1)}) x_i \end{aligned}$$

- ▶ Interpretation of update:
  - ▶ How much of  $y_i x_i$  to add to  $w^{(t-1)}$  is scaled by how far  $\sigma(y_i x_i^\top w^{(t-1)})$  currently is from 1.

# Convergence of gradient descent on smooth objectives

- **Theorem:** Assume  $f$  is twice-differentiable and convex, and  $\lambda_{\max}(\nabla^2 f(w)) \leq \beta$  for all  $w \in \mathbb{R}^d$  (" $f$  is  $\beta$ -smooth"). Then gradient descent with step size  $\eta := 1/\beta$  satisfies

$$\|w^{(0)} - w^*\|^2$$

<https://eduassistpro.github.io>

Add WeChat [edu\\_assist\\_pro](https://eduassistpro.github.io)

- Note: it is possible to have convergence even in some cases; should really treat  $\eta$  as a hyperparameter.

## Example: smoothness of empirical risk with squared loss

- Empirical risk with squared loss

Assignment Project Exam Help

So objective function is  $\beta$ -smooth with  $\beta = \lambda (A^T A)$ .

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

## Example: smoothness of empirical risk with logistic loss

- Empirical risk with logistic loss

Assignment Project Exam Help

$$\nabla^2 \frac{1}{n} \sum_{i=1}^n \ln(1 + \exp(-y_i x_i^\top w))$$

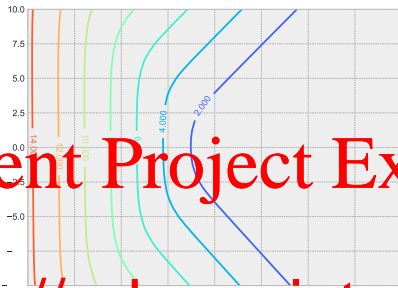
<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

# Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr



# Analysis of gradient descent for smooth objectives (1)

- By Taylor's theorem, can upper-bound  $f(w + \delta)$  by quadratic:

Assignment Project Exam Help



<https://eduassistpro.github.io>

$$f(w + \delta) \leq f(w) + \nabla f(w)^\top \delta + \frac{\beta}{2} \|\delta\|_2^2.$$

Add WeChat edu\_assist\_pro

- Minimized by  $\delta^* = -\frac{1}{\beta} \nabla f(w)$
- Plug-in this value of  $\delta$  into above ine

$$f\left(w - \frac{1}{\beta} \nabla f(w)\right) - f(w) \leq -\frac{1}{2\beta} \|\nabla f(w)\|_2^2.$$



## Analysis of gradient descent for smooth objectives (2)

- If  $f$  is convex (in addition to  $\beta$ -smooth), then repeatedly making such local changes is sufficient to approximately minimize  $f$ .

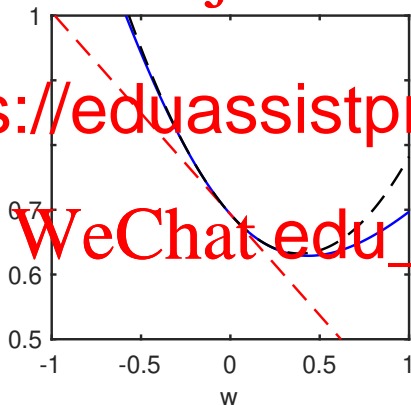


Figure 8: Linear and quadratic approximations to a convex function

## Example: Text classification (1)

- ▶ Data: articles posted to various internet message boards
- ▶ Label: -1 for articles from "religion", +1 for articles from "politics"
- ▶ Features:

Assignment Project Exam Help

- ▶ <https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pr

## Example: Text classification (2)

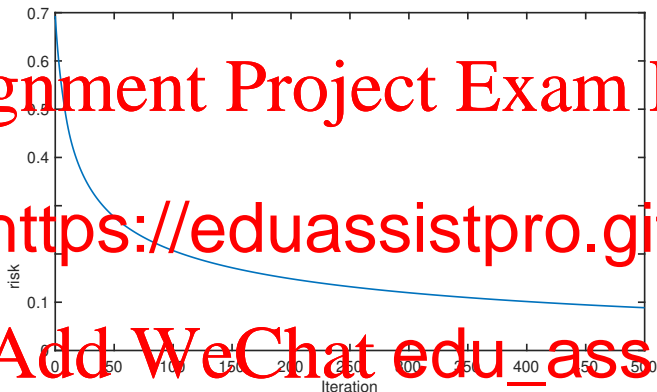


Figure 9: Objective value as a function of number of gradient descent iterations

## Example: Text classification (3)

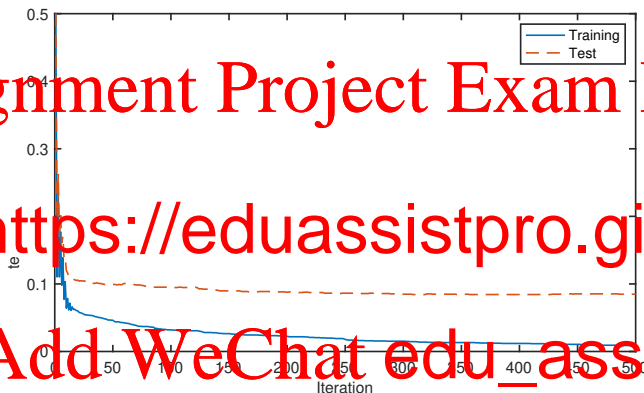


Figure 10: Error rate as a function of number of gradient descent iterations

# Stochastic gradient method (1)

- ▶ Every iteration of gradient descent takes  $\Theta(nd)$  time.
  - ▶ Pass through all training examples to make a single update.
  - ▶ If  $n$  is enormous, too expensive to make many passes.
- ▶ Alternative: Stochastic gradient descent (SGD)

Assignment Project Exam Help

<https://eduassistpro.github.io>

$$-\frac{1}{n} \sum_{j=1}^n \nabla \ell(\mathbf{x}^{\top} \mathbf{w}^{(t)})$$
  
Add WeChat edu\_assist\_pro  
(A.k.a. full batch gradient.)

- ▶ Pick term  $J$  uniformly at random:

$$\nabla \ell(y_J x_J^{\top} \mathbf{w}^{(t)}).$$

- ▶ What is expected value of this random vector?

## Stochastic gradient method (2)

- ▶ Minibatch

- ▶ To reduce variance of estimate, use several random examples

$J_1, \dots, J_B$  and average — called minibatch gradient.

$$\frac{1}{B}$$

<https://eduassistpro.github.io>

- ▶ Alternative: instead of picking example uniformly at random, shuffle order of training examples, and take this order.

- ▶ Verify that expected value is same!
  - ▶ Seems to reduce variance as well, but not fully understood.

## Example: SGD for logistic regression

- ▶ Logistic regression MLE for data  $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$ .
- ▶ Start with  $w^{(0)} \in \mathbb{R}^d, \|w\| > 0$ ,  $t = 1$
- ▶ For epoch  $p = 1, 2, \dots$ :

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

# Optimization for linear regression

- ▶ Back to considering ordinary least squares.
- ▶ Gaussian elimination to solve normal equations can be slow when  $d$  is large (time is  $O(n d^2)$ ).
- ▶ Alternative: find approximate solution using gradient descent
- ▶

<https://eduassistpro.github.io>

- ▶ Time to multiply matrix by vector is linear in  $d$
- ▶ So each iteration takes time  $O(d)$
- ▶ Can describe behavior of gradient descent for (empirical risk) objective very precisely.



# Behavior of gradient descent for linear regression

- **Theorem:** Let  $\hat{w}$  be the minimum Euclidean norm solution to normal equations. Assume  $w^{(0)} = 0$ . Write eigendecomposition

$A^T A = \sum_{i=1}^r \lambda_i v_i v_i^T$  with  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$ . Then  $w^{(t)} \in \text{range}(A^T)$  and

► <https://eduassistpro.github.io>

- If we choose  $\eta$  such that  $2\eta\lambda_i$

$$2\eta\lambda_i \sum_{k=0}^{t-1} (1 - 2\eta\lambda_i)^k$$

which converges to 1 as  $t \rightarrow \infty$ .

- So, when  $2\eta\lambda_1 < 1$ , we have  $w^{(t)} \rightarrow \hat{w}$  as  $t \rightarrow \infty$ .
- Rate of convergence is geometric, i.e., “exponentially fast convergence”.
- Algorithmic inductive bias!

- ▶ There are many optimization algorithms for convex optimization
  - ▶ Gradient descent, Newton's method, BFGS, coordinate descent, mirror descent, etc.
  - ▶ Stochastic variants thereof



<https://eduassistpro.github.io>

- ▶ E.g., want coordinates of  $w$  to lie i



The algorithmic inductive bias not always it is there!