Exercises
**Introduction to Machine Learning**
Spring 2023

Institute for Machine learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause, Prof. Dr. Fanny Yang

For questions, please refer to Moodle.
Released on **8 March 2023**

# Homework 1
# (Linear Regression, Optimization, Regularization)

General Instructions

- Submission of solutions is not mandatory but solving the exercises is highly recommended.
- The master solution will be released next week.

## Exercise 1: Convex functions

In the following exercise we consider real-valued functions $f : \text{dom}(f) \to \mathbb{R}$, with $\text{dom}(f) \subseteq \mathbb{R}^d$.

A function $f : \text{dom}(f) \to \mathbb{R}$ is *convex* if (i) $\text{dom}(f)$ is a convex set and (ii) for all $\boldsymbol{x}, \boldsymbol{y} \in \text{dom}(f)$ and all $\lambda \in [0, 1]$, we have

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(x) + (1-\lambda) f(\boldsymbol{y}) \tag{1}$$

Geometrically, the condition means that the line segment connecting the points $(\boldsymbol{x}, f(\boldsymbol{x})), (\boldsymbol{y}, f(\boldsymbol{y})) \in \mathbb{R}^{d+1}$ lies point-wise above the graph of $f$ ; see Figure 1 below.

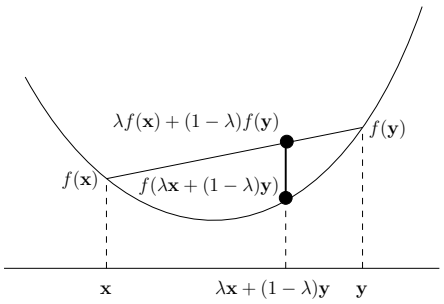


Figure 1: A convex function

### Questions

(a) A norm is a function $f : \mathbb{R}^d \to \mathbb{R}$ which satisfies the following three properties:

- $f(x) > 0$ for all $x \neq 0$
- $f(\theta x) = |\theta| f(x)$ for all $\theta \in \mathbb{R}$ and $x \in \mathbb{R}^d$
- $f(x + y) \leq f(x) + f(y)$ for all $x, y \in \mathbb{R}^d$

Show that any valid norm $f$ is a convex function using the definition of convexity in Equation 1.

Consider now the function $f(x, y) = x^2 + y^2$. The graph of $f$ is the unit paraboloid in $\mathbb{R}^3$ which looks convex. However, verifying the condition in (1) directly is somewhat cumbersome.

To address this problem, we develop better ways to check convexity if the function under consideration is differentiable. In particular, suppose that $\mathrm{dom}(f)$ is open and that $f$ is differentiable, i.e. the gradient

$$\nabla f(\boldsymbol{x}) \coloneqq \left( \frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_p}(x) \right)$$

exists at every point $\boldsymbol{x} \in \mathrm{dom}(f)$. Then, we will prove that an easier condition to verify convexity is the following:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \tag{3}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$. Geometrically, this means that for all $\boldsymbol{x} \in \mathrm{dom}(f)$, the graph of $f$ lies above its tangent hyperplane at the point $(\boldsymbol{x}, f(\boldsymbol{x}))$ ; see Figure 2 below.

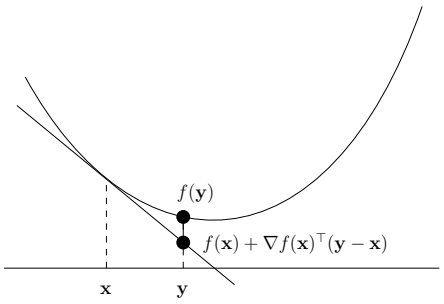


Figure 2: First-order characterization of convexity

## Questions

(b) Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Prove that $f$ is convex if and only if $\mathrm{dom}(f)$ is convex and

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \tag{4}$$

holds for all $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$.

(c) Suppose that $f$ is differentiable and convex. Prove that $\boldsymbol{x}^\star$ is a global minimum of $f$ if and only if $\nabla f(\boldsymbol{x}^\star) = 0$.

(d) Show that $f(x, y) = x^2 + y^2$ is a convex function and that the point $(0, 0)$ is a global minimum.

A third way to verify convexity is through:

$$\nabla^2 f(x) \succeq 0 \tag{10}$$

Geometrically, this means that the graph of $f$ has non-negative curvature everywhere and hence looks like a bowl. We will come back to this when discussing ridge regression in Exercise 3.

## Questions

(e) A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called $\alpha$-strongly convex for some $\alpha > 0$, if for any points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ one has

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\alpha}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

If $f$ is twice differentiable, an equivalent condition is that for any point $x \in \mathbb{R}^d$, one has

$$\nabla^2 f(x) \succeq \alpha I$$

which means $\nabla^2 f(\boldsymbol{x}) - \alpha I$ is positive semi-definite for all $\boldsymbol{x} \in \mathbb{R}^d$. Prove that a strongly convex function admits a unique minimizer in $\mathbb{R}^d$.
*Hint: This is not an easy exercise. First prove that $f(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$ to show that there is some minimizer.*

## Exercise 2: Linear regression

In the lecture we have learned how to fit an affine function to data by performing linear regression. In the tutorial on Friday we will discuss how to fit more general nonlinear functions to data. The goal of this exercise is to solidify our understanding of some of the concepts that have been touched upon.

Consider a dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R} \times \mathbb{R}$ and the hypothesis space of affine functions $H = \{w_0 + w_1 x : w_0, w_1 \in \mathbb{R}\}$. The error of a solution $f \in H$, i.e., $f(x) = w_0 + w_1 x$ for some $w_0, w_1 \in \mathbb{R}$ is given by

$$L(w_0, w_1) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2. \tag{11}$$

### Questions

(a) Let us for a moment consider the simpler case where we fix $w_0 = 0$. Compute the optimal linear fit to the data by computing $w_1^* = \arg\min_{w_1 \in \mathbb{R}} L(0, w_1)$.

(b) Prove that, for $n \geq 2$ and $x_i \neq x_j$ for $i \neq j$, (11) is a strictly convex function with respect to $\boldsymbol{w} = (w_0, w_1)$.

(c) The unique global minimum of a strictly convex function can be computed by setting its gradient to zero. Compute the gradient

$$\nabla L(w_0, w_1) = \begin{pmatrix} \frac{\partial L(w_0, w_1)}{\partial w_0} \\ \frac{\partial L(w_0, w_1)}{\partial w_1} \end{pmatrix}. \tag{14}$$

(d) Compute the optimal parameters $(w_0^*, w_1^*) = \arg\min_{w_0, w_1 \in \mathbb{R}} L(w_0, w_1)$ by solving the linear system of equations obtained by setting (14) to zero, i.e., $\nabla L(w_0, w_1) = 0$.

Another way of writing (11) is in matrix notation

$$L(\boldsymbol{w}) = \frac{1}{2n}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|^2, \tag{15}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is the vector of target values, $\boldsymbol{w} = (w_0, w_1)^T$ is our weight vector and

$$\Phi = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{16}$$

is the data matrix.

For $n \geq 2$ different observations (15) is a strictly convex function and can be minimized by setting its gradient

$$\nabla L(\boldsymbol{w}) = \frac{1}{n}(\Phi^T\Phi\boldsymbol{w} - \Phi^T\boldsymbol{y}) \tag{17}$$

to zero.

The benefit of (15) is that it straightforwardly generalizes to multiple inputs $x_i \in \mathbb{R}^d$ using

$$\Phi = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1d} \\ 1 & x_{21} & \ldots & x_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nd} \end{pmatrix}, \tag{18}$$

where we have one row per data point and one column per input.

## Questions

(e) Provide necessary conditions for $\Phi$ such that $\Phi^T\Phi$ is invertible.

(f) Show that if $\Phi^T\Phi$ is invertible, then $\boldsymbol{w}^* = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{y}$ is the unique minimum of $L(\boldsymbol{w})$ in (15).

(g) Show that for $n < d + 1$ the regression problem

$$\min_{\boldsymbol{w}\in\boldsymbol{R}^{d+1}} \|\boldsymbol{y} - \Phi\boldsymbol{w}\|^2 \tag{23}$$

does not admit a unique solution.

Next, recall the gradient descent update as discussed in lecture: $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta\nabla_{\boldsymbol{w}}L(\boldsymbol{w}^t)$. We would like to compare the computational complexity of the closed-form solution $\boldsymbol{w}^* = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{y}$ against the one of the gradient descent algorithm. For the next exercises you may use the contraction inequality as discussed during lecture,

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|_2 \leq \overbrace{\|I - \eta\Phi^\top\Phi\|_{op}}^{:=\rho}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|_2$$

## Questions

(h) Prove that if $\Phi^\top\Phi$ is full rank and the learning rate satisfies $\eta < \frac{2}{\lambda_{max}(\Phi^\top\Phi)}$, where $\lambda_{max}$ is the largest eigenvalue, then $\rho < 1$. Conclude that gradient descent converges to the optimal solution $w^*$ as $t \to \infty$.

Hint: remember that an equivalent characterization of the operator norm in terms of the eigenvalues of $A$ is given by $\|A\|_{op} = \max\{|\lambda| : \lambda \text{ eigenvalue of } A\}$

(i) Assume that the stepsize $\eta$ is such that $\|I - \eta\Phi^T\Phi\|_{op} < 1$. Compute the number of gradient steps $\tau$ and the overall complexity required to obtain a solution $\boldsymbol{w}^\tau$ that satisfies $\|\boldsymbol{w}^\tau - \boldsymbol{w}^*\| < \varepsilon$, where $\boldsymbol{w}^\tau$ is the parameter vector computed by gradient descent after $\tau$ steps.

(j) For the linear regression loss function defined in Equation 15, prove that $L(w^{t+1}) < L(w^t)$ for small enough stepsize $\eta$.

(k) Now, say you are free to choose a constant stepsize. What is the minimum number of iterations $\tau$ required to obtain a solution $\boldsymbol{w}^\tau$ that satisfies $\|\boldsymbol{w}^\tau - \boldsymbol{w}^*\| < \varepsilon$? How does it depend on the maximum and minimum eigenvalues $\lambda_{max}$, $\lambda_{min}$ of the matrix $\Phi^T\Phi$?

(l) Compare the computational complexity of gradient descent to the one required to solve the linear system of equations $\Phi^T\Phi\boldsymbol{w}^* = \Phi^T\boldsymbol{y}$ in closed form.

## Exercise 3: Bias variance tradeoff

In the lecture we have seen the bias variance decomposition of the expected squared error. In this exercise, we will first derive this decomposition in the case of linear regression and we will then use it to study ridge regression. Consider a dataset with fixed data matrix $\Phi \in \mathbb{R}^{n\times(d+1)}$ as shown in Equation 18 and noisy labels $\boldsymbol{y} \in \mathbb{R}^n$ with $y_i = \boldsymbol{w}^\top\phi(\boldsymbol{x}_i) + \epsilon_i$. We define the noise $\epsilon_i$ such that $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. We assume that we have used this training data to train a linear regressor with estimator $\hat{\boldsymbol{w}}$. We now consider the expected squared error of this estimate defined as $\mathbb{E}_\epsilon\left[\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|^2\right]$. In this exercise, $\|\cdot\|$ always refers to the Euclidean norm. Remember the definitions of bias and variance as given in the lecture:

$$\mathrm{Bias}(\hat{\boldsymbol{w}})^2 = \|\boldsymbol{w} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}]\|^2 \tag{25}$$

$$\mathrm{Var}(\hat{\boldsymbol{w}}) = \mathbb{E}_\epsilon\left[\|\hat{\boldsymbol{w}} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}]\|^2\right] \tag{26}$$

### Questions

(a) Prove that $\mathbb{E}_\epsilon\left[\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|^2\right] = \mathrm{Bias}(\hat{\boldsymbol{w}})^2 + \mathrm{Var}(\hat{\boldsymbol{w}})$.

Remember that while bias drives underfitting, variance is related to overfitting. During training, we aim to find the model with the lowest expected generalization error which requires joint optimization over these two components. As seen in the lecture, regularization is frequently used to control model complexity which reduces variance at the expense of increased bias. Ridge regression is a regularized version of linear regression. The optimization problem with parameter $\lambda > 0$ is given by Equation 27 which is optimizing over the loss we have seen in Exercise 2 but with an added regularization term $\lambda\|\boldsymbol{w}\|^2$.

$$\boldsymbol{w}^*_{\text{ridge}} = \arg\min_{\boldsymbol{w}} L_{\text{ridge}}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}}\left[\sum_{i=1}^{n}(y_i - \boldsymbol{w}^\top\phi(\boldsymbol{x}_i))^2 + \lambda\|\boldsymbol{w}\|^2\right] \tag{27}$$

### Questions

(b) Prove that the solution to Equation 27 is unique for any matrix $\Phi$ by showing that $L_{\text{ridge}}$ is strictly convex.

In the following, we want to develop a better understanding of the bias variance tradeoff by comparing ridge regression to unregularized linear regression.

### Questions

(c) Derive the closed-form solution $\boldsymbol{w}^*_{\text{ridge}}$ by computing the unique minimizer defined in Equation 27.

(d) Show that its bias is given by $\text{Bias}(\boldsymbol{w}^*_{\text{ridge}}) = \|\lambda \left(\Phi^\top \Phi + \lambda I_d\right)^{-1} \boldsymbol{w}\|$. How does it compare to the bias of unregularized linear regression?

(e) We next focus on the variance of the unregularized estimator as derived in Exercise 2. Remember that this estimator is defined as $\boldsymbol{w}^* = (\Phi^\top \Phi)^{-1} \Phi^\top \boldsymbol{y}$ assuming that $\Phi$ is full rank. Show that its variance is given by

$$\text{Var}(\boldsymbol{w}^*) = \sum_{i=1}^{m} \frac{\sigma^2}{\sigma_i^2}$$

where the $\sigma_i$ are the singular values of $\Phi$ and $m = \min(n, d+1)$.

*Hint: Remember that* $Var(\hat{\boldsymbol{w}}) = \mathbb{E}_\epsilon \left[\|\hat{\boldsymbol{w}} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}]\|^2\right]$ *and use the fact that* $A = Tr(A)$ *if* $A$ *is a real number and that the trace is invariant under cyclic permutations* $Tr(ABC) = Tr(BCA) = Tr(CAB)$.

(f) Similarly, it can be shown that the ridge estimator has variance

$$\text{Var}(\boldsymbol{w}^*_{\text{ridge}}) = \sigma^2 \sum_{i=1}^{m} \frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2}$$

where the $\sigma_i$ are the singular values of $\Phi$ and $m = \min(n, d+1)$. Compare the variance of the unregularized estimator to that of the ridge estimator.

(g) How does the choice of $\lambda$ affect bias and variance? How do bias and variance behave as $\lambda \to 0$ and $\lambda \to \infty$?

## Exercise 4: Gradient Descent for Linear Regression (Bonus)

In this exercise, we are going to prove that under mild conditions the gradient descent algorithm for ordinary linear regression problems converges to the solution with minimum norm. This is a very good exercise to practice your linear algebra skills. To help you prove this argument, we have divided the complete proof into smaller chunks. As in the lecture, suppose $X \in \mathbb{R}^{n \times d}$ is the data matrix and $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector. The goal is to find a vector $\boldsymbol{w} \in \mathbb{R}^d$ such that $L(\boldsymbol{w}) := \frac{1}{2}\|X\boldsymbol{w} - \boldsymbol{y}\|^2$ is minimized. For this, we use the gradient descent algorithm: starting from an initial vector $\boldsymbol{w}^0$, the iterates of the gradient descent algorithm for a step size $\eta$ are

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \eta \nabla L(\boldsymbol{w}^k), \quad k = 0, 1, \ldots$$

(a) By computing the gradient of $L$, confirm that

$$\boldsymbol{w}^{k+1} = (I - \eta X^\top X)\boldsymbol{w}^k + \eta X^\top \boldsymbol{y}.$$

(b) By using induction on $k$, prove that

$$\boldsymbol{w}^k = \underbrace{(I - \eta X^\top X)^k \boldsymbol{w}^0}_{\text{(A)}} + \underbrace{\eta \left( \sum_{j=0}^{k-1} (I - \eta X^\top X)^j \right) X^\top \boldsymbol{y}}_{\text{(B)}} \tag{30}$$

From (b) it is clear that powers of the matrix $I - \eta X^\top X$ play an important role in understanding what happens to $\boldsymbol{w}^k$ when $k$ is large. It is usual to look at the eigenvalues of a matrix when studying its powers. Hence, we start by the SVD of $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{d \times d}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{n \times d}$ is a rectangular diagonal matrix with non-negative real numbers $\sigma_1, \ldots, \sigma_n$ on its diagonal. *From this part onwards, we focus on the over-parameterized case where* $n < d$.

(c) Verify that the eigenvalue decomposition of $I - \eta X^\top X$ is $V(I - \eta\Lambda)V^\top$, where $\Lambda \in \mathbb{R}^{d \times d}$ is a diagonal matrix whose first $n$ diagonal entries are $\sigma_i^2$ and the rest are zero.

(d) Denote by $\sigma_{\max} := \max \sigma_i$. Observe that if $\eta \leq 1/\sigma_{\max}^2$, all eigenvalues of $I - \eta X^\top X$ will be non-negative.

(e) Compute $(I - \eta X^\top X)^k$ in closed form for any $k \geq 1$ based on $V$, $\eta$ and $\Lambda$.

We now compute parts (A) and (B) in Equation (30) separately. *From now on, we assume that* $\eta \leq 1/\sigma_{\max}^2$.

(f) If $\boldsymbol{v}^i$ is an eigenvector of $X^\top X$ corresponding to the eigenvalue $\sigma_i^2$, compute $(I - \eta X^\top X)^k \boldsymbol{v}^i$. Describe what happens when $k \to \infty$. (*Hint: you have to consider two cases: when* $\sigma_i = 0$ *and* $\sigma_i > 0$)

(g) Based on the last step, compute part (A) when $k \to \infty$. (*Hint: decompose* $\boldsymbol{w}^0 = \boldsymbol{v} + \boldsymbol{u}$, *where* $\boldsymbol{u} \in \ker(X)$ *and* $\boldsymbol{v} \in \ker(X)^\perp$)

(h) Show that part (B) is equal to

$$V\left(\eta \sum_{j=0}^{k-1} (I - \eta\Lambda)^j\right)\Sigma^\top U^\top \boldsymbol{y}$$

(i) Compute (B) when $k \to \infty$. (*Hint: treat zero and positive singular values separately.*)

(j) Prove that the limit computed above equals $X^+\boldsymbol{y}$, where $X^+$ is the Moore-Penrose pseudo-inverse.

(k) Notice that the above argument also works for the case where $n \geq d$. Make the necessary adjustments and prove that gradient descent initialized at $\mathbf{0}$ and with a small enough step size converges to the correct solution in under-parameterized setting.