Bowen Weng
*Department of Computer Science*
*Iowa State University*

**HW3 — Optimization**

# 1 Definitions

1. Are the following sets convex?
   (a) $\{x \mid x \in \mathbb{R}^2, x^T x \leq 2\}$ Yes
   (b) $\{x \mid x \in \mathbb{R}^2, x^T x \geq 2\}$ No
2. Are the following sets strictly convex?
   (a) $\{x \mid x \in \mathbb{R}^2, x^T x \leq 2\}$ Yes
   (b) $\{x \mid Ax \leq 0\}$ $(A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix})$ No
3. Are the following functions convex? in general, you can just plot the graph of the function and check it yourself, some of you misunderstood the graph of $x^T x + 4$, that is basically a "bowl" with the minimum as $(0, 4)$, and the last question "cuts off" its bottom with all points $(x)$ having a radius of 2 or shorter, making it non-convex (e.g., pick two points $(3, 3)$ and $(-3, -3)$ and connect them.)
   (a) $f(x) = x^2, x \in \mathbb{R}$ yes
   (b) $f(x) = x^2, x \in [0, 1]$ yes
   (c) $f(x) = x^T x + 4, x \in \mathbb{R}^2$ yes
   (d) $f(x) = x^T x + 4, x \in \mathbb{R}^2, x^T x \geq 2$ no
4. Are the following matrices Positive Definite (PD), Positive Semi-Definite (PSD), Negative Definite (ND), or Negative Semi-Definite (NSD)? Please explain your answers. Correct answer is one point. Explanation is one point. I appreciate those who attempted to calculate eigenvalues. You will get the point (if you made the calculations correctly). In this course, the suggested method is to use the given definitions (see slides and the following answers).
   (a) $\begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}$ $x^T Q x = 4x_2^2 \geq 0$ and the zero equality is valid when $x_1 \in \mathbb{R}, x_2 = 0$ (i.e., not only $x = 0$), hence it is PSD
   (b) $\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$ $x^T Q x = x_1^2 + 4x_2^2 + 2x_1 x_2 = (x_1 + x_2)^2 + 3x_2^2 \geq 0$ and the zero equality is only valid when $x = 0$ $(x_1 = x_2 = 0)$, hence it is PD
   (c) $\begin{bmatrix} -5 & 2 & 0 \\ 2 & -3 & 1 \\ 0 & 1 & -2 \end{bmatrix}$ $x^T Q x = -(2x_1 - x_2)^2 - 2x_2^2 - (x_1 - x_3)^2 - x_3^2 \leq 0$ and the zero equality is only valid when $x = 0$, hence it is ND
5. Are the following functions Lipschitz continuous? Why? First check if the function is differentiable. If yes, your explanation should check $|\nabla f(x)|$ and see if there exists a constant positive value $L > 0$ such that $|\nabla f(x)| < L, \forall x \in X$, where $X$ denotes the set of respective admissible values of $x$.
   (a) $f(x) = \sqrt{x}, x \in [0, 1]$ No
   (b) $f(x) = x^2, x \in \mathbb{R}$ No
   (c) $f(x) = x^2, x \in [0, 1]$ Yes

(d) $f(x) = |x|, x \in \mathbb{R}$ Yes,. I originally intended to confine our discussion to differentiable function only, so I wanted the student to identify that it is not differentiable (at $x = 0$). But I just rechecked my notes, it was not given in the definition. As a result, one can indeed find an $L > 0$.

(e) $f(x) = \sqrt{7x^2 + 4}, x \in \mathbb{R}$ Yes

(f) $f(x) = \sin x, x \in \mathbb{R}$ Yes

(g) $f(x) = x^5, x \in \mathbb{R}$ No

6. Are the following functions Lipschitz smooth? Why? First check if the function is twice differentiable. If yes, your explanation should check $|H(x) = \nabla_{xx} f(x)|$ and see if there exists a constant positive value $L > 0$ such that $|H(x)| < L, \forall x \in X$, where $X$ denotes the set of respective admissible values of $x$.

(a) $f(x) = \sqrt{x}, x \in [0, 1]$ No

(b) $f(x) = x^2, x \in \mathbb{R}$ Yes

(c) $f(x) = x^2, x \in [0, 1]$ Yes

(d) $f(x) = |x|, x \in \mathbb{R}$ No, the differentiable function is a prior in the definition.

(e) $f(x) = \sqrt{7x^2 + 4}, x \in \mathbb{R}$ Yes

(f) $f(x) = \sin x, x \in \mathbb{R}$ Yes

(g) $f(x) = x^5, x \in \mathbb{R}$ No

In the future, when it is allowed, there are many symbolic tools that can help you derive, sometimes very complex, derivatives, such as www.wolframalpha.com. Feel free to try it out if you are still confused about the gradient and Hessian calculations for the above two sets of questions.

7. Recall the (second) definition of the convex function given in the lecture, consider the following lemma (note $\langle a, b \rangle$ denotes $a^T b$ for some vectors $a$ and $b$):

**Lemma.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and convex, then*

$$\forall x, y \in \mathbb{R}^n, f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle. \tag{1-1}$$

(a) What does the lemma mean (hint: in the perspective of the first definition of the convex function given in the lecture)? Every tangent line to the graph of $f(x)$ is a lower bound of all function values.

(b) Prove the lemma (hint: use the convexity definition and the notion of limits).

*Proof.* By the convexity definition, for all $t \in (0, 1]$:

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y). \tag{1-2}$$

Let both sides be divided by $t$ and with some rearrangements,

$$\frac{f(tx + (1 - t)y) - f(y)}{t} \leq f(x) - f(y) \tag{1-3}$$

Take the limit of both sides as $t \to 0$:

$$\langle \nabla f(y), x - y \rangle \leq f(x) - f(y) \tag{1-4}$$

(divided by $t$ is one point, taking limit is one point, there are other correct solutions that do not follow this approach, TA can direct those answers to be for grading.) Since the selections of $x, y$ are arbitrary, swap the selections (i.e., let $x$ take the value of $y$ and $y$ take the value of $x$), we have

$$\langle \nabla f(x), y - x \rangle \leq f(y) - f(x). \tag{1-5}$$

This completes the proof. □

8. Recall the smooth function definition, let $f : \mathbb{R}^n \to \mathbb{R}$ be $L$-smooth, convex, with the minimum value at $x = x^*$, prove

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*)) \tag{1-6}$$

The proof is straightforward if you know the Taylor expansion of $f(x)$, and I somehow had this impression that it was given in Calculus III and various versions of differential equations. That's being said, if you have taken any of the courses before, you should be good to go. If you are not aware of Taylor expansion, I'm sorry I'm not aware of other alternatives. One may have to lose one point here.

*Proof.* Given the definition of $L$-smoothness, let $y = x - \frac{1}{L}\nabla f(x)$

$$f(x - \frac{1}{L}\nabla f(x)) \leq f(x) - \frac{1}{L}\langle \nabla f(x), \nabla f(x)\rangle + \frac{L}{2}\left\|\frac{1}{L}\nabla f(x)\right\|^2 = f(x) - \frac{1}{2L}\|\nabla f(x)\|^2 \tag{1-7}$$

Given the definition of the global minimum $x^*$ ($f(x^*) \leq f(y), \forall y$) and the above, we have

$$f(x^*) \leq f(y) \Rightarrow f(x^*) - f(x) \leq f(y) - f(x) \leq f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|^2 \tag{1-8}$$

Some rearrangements of the above inequality complete the proof. □

9. How many feasible solutions do the following optimization problems have? Why? Two rules: (i) solution means $x$, not f(x). (ii) a feasible solution is a global optimal solution, a global optimal solution is a feasible solution (necessary and sufficient). You need to use the definitions to explain your answer, not the graph. But plotting the graph can always help you form your answers with definitions.

   (a) $\min_x x^4, x \in \mathbb{R}$ One. By definition, $0^4 \leq x^4, \forall x \in \mathbb{R}$ and the equality is only valid at $x = 0$.

   (b) $\min_x x^T x, x \in \{x \mid x \in \mathbb{R}^2, x^T x \geq 2\}$ Infinitely many. By definition, $\forall x'$ satisfying $x'^T x' = 2$, $x'^T x' \leq x^T x, \forall x \in \{x \mid x \in \mathbb{R}^2, x^T x \geq 2\}$ and the equality is only valid when $x^T x = 2$. Recall the "bottom cut-off bowl" mentioned above.

   (c) $\min_x f(x), f(x) = \begin{cases} x^2, \text{if } x \geq 0 \\ 0, \text{otherwise} \end{cases}$ Infinitely many, all $x < 0$ are feasible solutions

   (d) $\min_x x^4 - 4x^2 + 4, x \in \mathbb{R}$ Two. This one is an non-constrained problem. Zero gradient gives three solutions and two of them are minimums ($\sqrt{2}$ and $-\sqrt{2}$). By definition, for $f(x^*) \leq f(x), forall x \in \mathbb{R}$ with $f(x) = x^4 - 4x^2 + 4$. The equality is only valid when $f(x^*) = f(\sqrt{2}) = f(-\sqrt{2})$.

   (e) $\min_x \sin(x), x \in \mathbb{R}$ Infinitely many. All $x$ with $\sin(x) = -1$ make global minimum solutions.

## 2 Stochastic Gradient Descent (or Life)

1. Consider the function

$$f(x) = \frac{x^2}{4} + 1 - \cos(2\pi x), x \in [-4.5, 4.5] \tag{1-9}$$

   (a) Draw $f(x)$ versus $x$. You can use a picture of hand-drawing (no need to be perfect) or with the help of the computer program. See Figure 1

   (b) Find an initial point $x_0$ where gradient descent could end up converging to a local minimum that is not the global minimum. There is no unique solution, TA can justify.

   (c) Find a subset of $[-4.5, 4.5]$ such that f(x) has a non-unique global minimum. There is no unique solution, TA can justify.

   (d) Find a subset of $[-4.5, 4.5]$ such that f(x) has a unique global minimum. There is no unique solution, TA can justify.

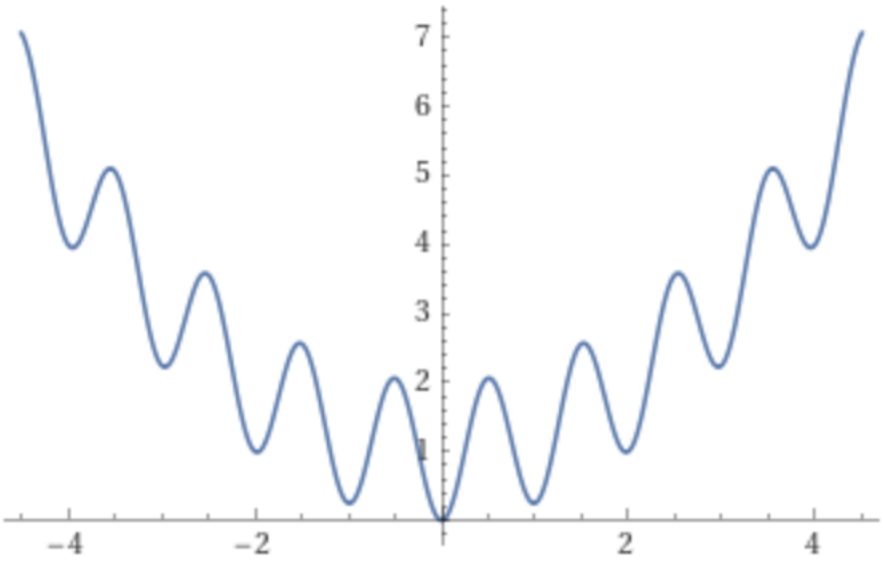


Figure 1: The plot is generated with wolframalpha.com

2. Consider the linear regression problem in the previous lectures. Recall the "real estate" data set (see code on Canvas and more details in the slides). Consider the data set as it stands (i.e., incorporating all six $x$ features and the target $y$ being the house price of unit area). Given the candidate function

$$y = \omega^T x + b, \omega \in \mathbb{R}^6, b \in \mathbb{R}, \tag{1-10}$$

and the cost function defined as RSS (unless noted otherwise), consider the following questions. Note the gradient formulation (if used) is expected to be derived analytically (then implemented in code), but the calculation of gradients can be done numerically in a variety of ways. As a result, your Python code is expected to use only standard Python libraries and `numpy` only.

   (a) Starting from the initial point at $\omega = \begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1\end{bmatrix}^T$ and $b = 10$, execute the gradient descent algorithm for 4140 steps at a constant learning rate $\gamma = 0.001$, what is the updated solution (you are extremely likely to need some help from your computer here)? As I mentioned in-class, normalization is important to get a correct answer (or technically speaking any numerical answer). So, the first point gives to normalization (through code, or explanations). The second point gives to the correct implementation of the gradient calculation (through code, or analytical formulations). Solution itself does not have any point.

   (b) If you follow the configuration described in the previous question (i.e., same initial point, same learning rate), for a sufficiently large running steps, will Gradient Descent lead you to the same optimal solution you have derived from the regression lecture (in other words, will it converge?)? Why? Yes, it will. By the convergence theorem of GD provided in the lecture, the RSS function is $L$-smooth ($L \geq 2$) and convex. As a result, the convergence is guaranteed as steps tend to infinity and a step size less than 0.5 (could be smaller for sub-linear convergence, but this is not required here).

   (c) Starting from the initial point at at $\omega = \begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1\end{bmatrix}^T$ and $b = 10$, execute the stochastic gradient descent algorithm for 4140 steps at a constant learning rate $\gamma = 0.001$ (let's say an invisible magical force is controlling the generation of all randomness on our planet, and your "stochastic" samples end up being the enumeration of the 414 points in the real estate data set following the exact order as it stands in the .csv file, thus to have the 4140 steps, you will have to repeat the enumeration at the same order for 10 times), what is the updated solution? Is your solution better than, the same with, or worth than the solution from 2(a)? Why? Normalization

is obviously also required in this one, but it is not graded for points. The first point gives to the correct implementation of SGD (based on one sample rather than the complete data set, through code implementation or analytical formulations). The second point gives to the correct performance judgement using $R^2$ (and $R^2$ only!). Students may have different $R^2$ values. To relief the TA's grading burdens, the one point is given assuming the student's calculation is correct as long as one demonstrates the correct metric ($R^2$) is used and the relative comparison is given.

(d) Let's consider a cost that is not RSS, but takes the following form:

$$J(X, y; \omega, b) = \sum_{i=1}^{n} |y_i - \omega^T x_i - b|. \tag{1-11}$$

i. Is the optimal solution unique? If the optimal solution is not unique, why? If it is unique, is it still the same with the original one derived from the RSS cost? The correct answer is Yes and Yes, but any answer get two points. See the next question for why.

ii. For sufficiently large running steps with learning rate $\gamma = 0.001$, can you prove the convergence of Gradient Descent using the updated cost function (1-11)? Why? The key point here is to recognize that this updated new RSS is no longer differentiable (1 pt)! If you say Yes, you need to prove the convergence of GD with such a piece-wise cost function (feasible, but challenging). If you say No, you just need to explain that the theorem and definition given in our lecture do not apply in this case. The correct justification (regardless of yes or no) gets the other one point.

3. A manufacturing company produces a product using different materials. The goal is to minimize the production cost while meeting certain quality constraints. The company uses three materials, A, B, and C, in varying quantities to produce one unit of the product. The cost per unit of each material is as follows:

- Material A: \$4 per unit
- Material B: \$7 per unit
- Material C: \$3 per unit

Additionally, the company must meet the following quality constraints for each unit produced:

- The product must contain at least 20 units of Material A.
- The product must contain at least 15 units of Material B.
- The product must contain at least 10 units of Material C.

The company wants to determine the optimal quantities of each material to use in production to minimize the total cost while satisfying the quality constraints.

(a) Formulate the above problem as an optimization problem.

(b) Solve it with Gradient Descent.

Everyone gets 4 pts. **And for those who submitted an answer for this question, consider this a last warning!** I don't count or enforce attendance. Yet you still need to have an approach (a scientific one or through some superhuman magic) to make sure you receive all information from this course (including the slides, the instructions in the lectures, and other notifications on Canvas or through emails, to name a few). If you ignore my repeated instructions, and get caught (or revealed) through your future HW submissions or exam submissions, you will lose the points. Thanks!

## 3 A “Bonus" Question, This is the Last Time (1 pt)

From a scale of 1 to 5, how difficult is HW3? 1 is “I can do it in my sleep". 5 is “Bowen is ridiculous". 0 is “I refuse to answer this question". This is for my own reference to improve the quality of future assignments. Thanks!

All, as I have highlighted repeatedly, you are officially at the global maximum point of this course in terms of the difficulty level. I sincerely appreciate your efforts and enthusiasm for making to this point and still maintaining your enrollment status. You should find yourself in a more relaxed status moving forward with this course as we re-start our journey after the Spring break.

This is a cross-level course with both undergraduate and graduate students. It means this is the last machine learning course for some of you, but also the first machine learning course for others, including those who might continue their ML journey and make to a wonderland that I have never been to or even dreamed of. As a result, the context is prepared to maintain a balance in between the diverged groups. It is inevitable that some of you would find some of the points ridiculous and unnecessary. The special rubric of counting the best 5 out of 7 HW assignments towards the final grade was thus established for this potential problem.

Finally, as I have mentioned in the first lecture of this semester: “This is my first course as a faculty. All suggestions, complaints, warnings are welcome and encouraged!" Thank you all, and good luck!

*Submitted by Bowen Weng on .*