STA 243: Homework 2

- Homework due in Canvas: 05/15/2024 at 9:00PM. Please follow the instructions provided in Canvas about homeworks, carefully.

1. (**5 Points**) Prove that a differentiable function $f(\theta) : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$f(\theta_2) \geq f(\theta_1) + \nabla f(\theta_1)^\top (\theta_2 - \theta_1) \tag{1}$$

Hint: Think of 1-dimensional case and extend the intuition to d-dimensional case.

Definition 1. *Convex function from $\mathbb{R}^d \to \mathbb{R}^1$ is defined as: $\forall \theta_1, \theta_2 \in \mathbb{R}^d$, $\forall t \in [0,1]$*

$$tf(\theta_1) + (1-t)f(\theta_2) \geq f(t\theta_1 + (1-t)\theta_2) \tag{2}$$

Given Assumption that $f \subset C^1$ (which means that f is first-order differentiable)

Proof. $1 \implies 2$
let $\theta_0 = t\theta_1 + (1-t)\theta_2$, according to 1, we have

$$f(\theta_1) \geq f(\theta_0) + \nabla f(\theta_0)^\top (\theta_1 - \theta_0) \tag{3}$$

$$f(\theta_2) \geq f(\theta_0) + \nabla f(\theta_0)^\top (\theta_2 - \theta_0) \tag{4}$$

$3 \times t + 4 \times (1-t)$ is known as:

$$\begin{aligned} tf(\theta_1) + (1-t)f(\theta_2) &\geq f(\theta_0) + \nabla f(\theta_0)^\top (t\theta_1 - t\theta_2 + \theta_2 - \theta_0 - t\theta_2 + t\theta_0) \\ &= f(\theta_0) + \nabla f(\theta_0)^\top (t\theta_1 + (1-t)\theta_2 - \theta_0) \\ &= f(t\theta_1 + (1-t)\theta_2) + \nabla f(\theta_0)^\top (\theta_0 - \theta_0) \\ &= f(t\theta_1 + (1-t)\theta_2) \end{aligned}$$

$2 \implies 1$
Note that:

$$tf(\theta_1) + (1-t)f(\theta_2) \geq f(t\theta_1 + (1-t)\theta_2)$$

So that:

$$f(\theta_1) \geq \frac{f(t\theta_1 + (1-t)\theta_2) - (1-t)f(\theta_2)}{t}$$

That is

$$f(\theta_1) \geq \frac{f(\theta_2 + t(\theta_1 - \theta_2)) - f(\theta_2)}{t} + f(\theta_2)$$

Note the Taylor Expansion of f at θ_2 can be writtern as:

$$f(\theta_2 + t(\theta_1 - \theta_2)) = f(\theta_2) + \nabla f(\theta_2)^\top t(\theta_1 - \theta_2) + o(t(\theta_1 - \theta_2)), \quad t \to 0$$

So that when $t \to 0$:

$$\begin{aligned} f(\theta_1) &\geq \frac{f(\theta_2 + t(\theta_1 - \theta_2)) - f(\theta_2)}{t} + f(\theta_2) \\ &= \frac{f(\theta_2) + \nabla f(\theta_2)^\top t(\theta_1 - \theta_2) + o(t(\theta_1 - \theta_2)) - f(\theta_2)}{t} + f(\theta_2) \\ &= \frac{\nabla f(\theta_2)^\top t(\theta_1 - \theta_2) + o(t(\theta_1 - \theta_2))}{t} + f(\theta_2) \\ &= \nabla f(\theta_2)^\top (\theta_1 - \theta_2) + f(\theta_2) \end{aligned}$$

$\square$

2. (**20 Points**) The origin of the dataset `housingprice.csv` we will use in this question is from the Coursera open course Machine Learning Foundations: A Case Study Approach by Prof. Carlos Guestrin and Prof. Emily Fox. Load the training data `train.data.csv` and testing data `test.data.csv`. We'll build our regression model on the training data and evaluate the model on the testing data.

   (a) Build a linear model (you are free to use any Python package for this) on the training data by regressing the housing price on these variables: `bedrooms`, `bathrooms`, `sqft_living`, and `sqft_lot`. What's the R^2 of the model on training data? What's the R^2 on testing data?

Dataset	R^2 **Value**
Train	0.51011
Test	0.50499

Table 1: R-squared values for Train and Test datasets

   (b) The image below is Bill Gates' house. Load the file `fancyhouse.csv` to obtain the features of the house. Guess the price of his house using your linear model. Do you think the predicted price is reasonable?

   Predicted price for Bill Gats's house is $15,436,769.538, which is reasonable.

   (c) Let's continue to improve the linear model we have. Instead of throwing only the raw data into the statistical model, we might want to use our intuition and domain expertise to extract more meaningful features from the raw data. This step is called feature engineering. Using meaningful features in the model is often crucial for successful data analysis. Add another variable by multiplying the number of bedrooms by the number of bathrooms, which describes the combined benefit of having more bedrooms and bathrooms. Add this variable to the linear model we have in Part (a). What's the R^2 of the new model on the training data and testing data respectively?

Dataset	R^2 **Value**
Train	0.51735
Test	0.51054

Table 2: R-squared values for Train and Test datasets

   (d) Perform parts (a), (b) and (c) above without using any in-built function in Python (i.e., any packages that are related directly to linear regression), but by using `gradient descent algorithm` on the sample-based least-squares objective function, to estimate the OLS regression parameter vector. How does your result compare to the result from previous part ? Note that you have to set the step-size parameter appropriately for this method.

   We choose **Gradient Descent Algorithm**

   - **Input**: Initial vector $\boldsymbol{\theta}^{(0)} \in \mathbb{R}^d$
   - **Do**:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta_t \nabla f\left(\boldsymbol{\theta}^{(t)}\right)$$

- **While** $\left\|\nabla f\left(\boldsymbol{\theta}^{(t)}\right)\right\| \geq \tau$
- **Return** $\boldsymbol{\theta}^{(t)}$

Optimization Task
We use gradient descent to solve the following optimization task:

$$\begin{aligned}\beta^* &= \arg\min_{\beta\in\mathbb{R}^4} (y - X\beta)^\top (y - X\beta) \\ &= \arg\min_{\beta\in\mathbb{R}^4} (\beta - \overline{\beta})^\top (X^\top X)(\beta - \overline{\beta}) \\ &= \arg\min_{\beta\in\mathbb{R}^4} \frac{1}{2}(\beta - \overline{\beta})^\top A(\beta - \overline{\beta}) \\ &= \arg\min_{\beta\in\mathbb{R}^4} f(\beta)\end{aligned}$$

where $\overline{\beta} = (X^\top X)^{-1} X^\top y$, and $A = 2X^\top X$.

Gradient and Stepsize
In this problem:

- The gradient is given by $X^\top (X\beta - y)$.
- We firstly choose the stepsize $\eta_t = \eta = \frac{2}{\lambda_{\max}(A)+\lambda_{\min}(A)}$, and then we play around this numerical number and let $\eta_t = 10^-5$ to speed up our training method.

Model	X1	X2	X3	X4	X1X2
OLS	-0.15147017	0.00773629	0.7918328	-0.04514359	
OLS, GD	-0.15147017	0.00773629	0.7918328	-0.04514359	
Interaction	-0.32877205	-0.23326364	0.7719275	-0.04491517	0.38964848
Interaction, GD	-0.32877205	-0.23326364	0.7719275	-0.04491517	0.38964848

Table 3: $\hat{\beta}$ in Standardized Form for GD

Model	Train	Test
OLS	0.51011	0.50499
OLS, GD	0.51011	0.50499
Interaction	0.51735	0.51054
Interaction, GD	0.51735	0.51054

Table 4: R^2 for GD

GD got exactly the optimal solution in both with and without interaction term.

(e) Perform arts (a), (b) and (c) above now using `stochastic gradient descent` (with one sample in each iteration). How does your result compare to the result from previous parts ? Note: while running `stochastic gradient descent`, you can sample without replacement and when you run out of samples, just start over. Note that you have to set the step-size parameter appropriately for this method.

In the section, we are implementing the stochastic gradient algorithm to solve the stochastic optimization problem in the form of finite sum.

$$\beta^* = \underset{\beta \in \mathbb{R}^4}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 = \underset{\beta \in \mathbb{R}^4}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} f_i(\beta)$$

For each step, we simply select one data point (x_i, y_i) without replacement to estimate the gradient.

$$\beta^{(t+1)} = \beta^{(t)} - \eta_t \nabla f_i(\beta^{(t)}) = \beta^{(t)} - \eta_t (2x_i^\top x_i \beta^{(t)} - 2x_i^\top y_i)$$

Model	X1	X2	X3	X4	X1X2
OLS	-0.15147017	0.00773629	0.7918328	-0.04514359	
OLS, SGD	-0.15070764	0.00874859	0.7927156	-0.04433098	
Interaction	-0.32877205	-0.23326364	0.7719275	-0.04491517	0.38964848
Interaction, SGD	-0.32410012	-0.22732823	0.77209111	-0.04496272	0.3810138

Table 5: $\hat{\beta}$ in Standardized Form for SGD

Model	Train	Test
OLS	0.51011	0.50499
OLS, SGD	0.51011	0.50501
Interaction	0.51735	0.51054
Interaction, SGD	0.51735	0.51055

Table 6: R^2 for SGD

The final results are pretty close to the optimal solution and GD (since GD is exactly the optimal solution).Although several estimated coefficients are a little bit different. The most tricky thing here is the batch size for SGD here is only 1, which may leads to high variance of the gradient.

3. (**15 Points**) Prove the Fact in Page 91 of `Opt.pdf` and solve the recursion in Page 92 to obtain the final result of the Theorem in Page 86. (**Hint:** You can use induction)

Lemma 1. *If the function is μ-strongly convex, then* $(\nabla f(\theta_1) - \nabla f(\theta_2))^T (\theta_1 - \theta_2) \geq \mu\|\theta_1 - \theta_2\|^2$

Proof. Same as Problem(1), we use Taylor expansion here, which expand $f(\theta_1)$ at θ_2 and $f(\theta_2)$ at θ_1:

$$f(\theta_1) = f(\theta_2) + \nabla f(\theta_2)^T(\theta_1 - \theta_2) + \frac{1}{2}\langle \nabla^2 f(\theta_2 + t(\theta_1 - \theta_2))(\theta_1 - \theta_2), (\theta_1 - \theta_2)\rangle \tag{5}$$

$$f(\theta_2) = f(\theta_1) + \nabla f(\theta_1)^T(\theta_2 - \theta_1) + \frac{1}{2}\langle \nabla^2 f(\theta_1 + p(\theta_2 - \theta_1))(\theta_2 - \theta_1), (\theta_2 - \theta_1)\rangle \tag{6}$$

where $t, p \in [0, 1]$. Adding 5 and 6, we get:

$$\begin{aligned}
(\nabla f(\theta_1) - \nabla f(\theta_2))^T(\theta_1 - \theta_2) &= \frac{1}{2}\langle \nabla^2 f(\theta_2 + t(\theta_1 - \theta_2))(\theta_1 - \theta_2), (\theta_1 - \theta_2)\rangle \\
&\quad + \frac{1}{2}\langle \nabla^2 f(\theta_1 + p(\theta_2 - \theta_1))(\theta_2 - \theta_1), (\theta_2 - \theta_1)\rangle \\
&\geq \frac{1}{2}\min(\nabla^2 f(\theta_2 + t(\theta_1 - \theta_2)))\|\theta_1 - \theta_2\|^2 \\
&\quad + \frac{1}{2}\min(\nabla^2 f(\theta_1 + p(\theta_2 - \theta_1)))\|\theta_1 - \theta_2\|^2 \\
&\text{(Note that } \min(*) \geq \mu \text{ due to the strong convexity assumption)} \\
&\geq \frac{1}{2}\mu\|\theta_1 - \theta_2\|^2 + \frac{1}{2}\mu\|\theta_1 - \theta_2\|^2 \\
&= \mu\|\theta_1 - \theta_2\|^2
\end{aligned}$$

□

Part II

Theorem 0.1.

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta_t \mathbf{g}\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{\xi}^{(t)}\right)$$

where $\mathbf{g}\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{\xi}^{(t)}\right) = \nabla_{\boldsymbol{\theta}} F\left(\boldsymbol{\theta}^{(t)}; \boldsymbol{\xi}^{(t)}\right)$ *for solving the optimization problem. If*

$$\eta_t = \frac{c}{t+1}$$

for some $c > 0$, then we have

$$\mathbb{E}\left[\left\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\right\|_2^2\right] \leq \frac{c_0}{t+1}$$

where c_0 is a numerical constant.

Proof. Note that the relationship between $\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*$:

$$\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^*\|_2^2 = \|\boldsymbol{\theta}^{(t)} - \eta_t \mathbf{g}(\boldsymbol{\theta}^{(t)}; \xi^{(t)}) - \theta^*\|_2^2 \tag{7}$$

$$= \|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2^2 + \eta_t^2\|\mathbf{g}(\boldsymbol{\theta}^{(t)}; \xi^{(t)})\|_2^2 - 2\eta_t(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*)^\top \mathbf{g}(\boldsymbol{\theta}^{(t)}; \xi^{(t)}) \tag{8}$$

Analysis the Expectation of $\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2^2$ is just analying the Expectation of three parts above.

- $\|\mathbf{g}(\boldsymbol{\theta}^{(t)};\xi^{(t)})\|_2^2$ **part**
  By assumption, we can easily get:

$$\mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t)}}[\|\mathbf{g}(\boldsymbol{\theta}^{(t)};\xi^{(t)})\|_2^2] \leq \sigma_g^2 + M_g \mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}\|\nabla f(\boldsymbol{\theta}^{(t)})\|_2^2$$

  At this point, we bound $\mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t)}}[\|\mathbf{g}(\boldsymbol{\theta}^{(t)};\xi^{(t)})\|_2^2]$ in terms of $\mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}\|\nabla f(\boldsymbol{\theta}^{(t)})\|_2^2$. Note that using the property of L-smooth functions:

$$\|\nabla f(\boldsymbol{\theta}^{(t)})\|_2^2 = \|\nabla f(\boldsymbol{\theta}^{(t)}) - \nabla f(\boldsymbol{\theta}^*)\|_2^2 \leq L^2\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2^2$$

  We have:

$$\mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t)}}[\|\mathbf{g}(\boldsymbol{\theta}^{(t)};\xi^{(t)})\|_2^2] \leq \sigma_g^2 + M_g L^2 \mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2^2$$

- $(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*)^\top \mathbf{g}(\boldsymbol{\theta}^{(t)};\xi^{(t)})$ **part**
  And by assumptions and lemma 1, we can get

$$\begin{aligned}
\mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t)}}[(\theta(t) - \theta^*)^T\mathbf{g}(\boldsymbol{\theta}^{(t)};\xi^{(t)})] &= \mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}\left[\mathop{\mathbb{E}}_{\xi^{(t)}}[(\theta^t - \theta^*)^T\mathbf{g}(\boldsymbol{\theta}^{(t)};\xi^{(t)})|\xi^{(1)},\cdots,\xi^{(t-1)}]\right] \\
&= \mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}\left[(\theta^t - \theta^*)^T\mathop{\mathbb{E}}_{\xi^{(t)}}[(\mathbf{g}(\boldsymbol{\theta}^{(t)};\xi^{(t)}))|\xi^{(1)},\cdots,\xi^{(t-1)}]\right] \\
&= \mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}[(\boldsymbol{\theta}^{(t)} - \theta^*)^T\nabla f(\boldsymbol{\theta}^{(t)})] \quad \text{Using lemma 1 to get below} \\
&\leq \mu \mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}[\|\boldsymbol{\theta}^{(t)} - \theta^*\|^2]
\end{aligned}$$

Bring those to parts back to 8 and taking expectation, we got a induction inequity;

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^*\right\|_2^2\right] \leq& \mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}[\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2^2] + \eta_t^2(\sigma_g^2 + M_g L^2 \mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}[\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2^2)] \\
&- 2\eta_t\mu \mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}[\|\boldsymbol{\theta}^{(t)} - \theta^*\|^2] \\
&= (1 - 2\mu\eta_t + \eta_t^2 M_g L^2)\mathop{\mathbb{E}}_{\xi^{(1)},\cdots,\xi^{(t-1)}}[\|\boldsymbol{\theta}^{(t)} - \theta^*\|^2] + \eta_t^2\sigma_g^2
\end{aligned}$$

Set $c = \frac{1}{\mu}, c_0 = \frac{c^2\sigma_g^2}{2\mu c - 1}$. Using the fact our discussion set $M_g = 0$, So that $\eta_t = \frac{1}{\mu(t+1)}$ By induction over t, we can get the results:

(a) When $t = 0$, so that $\eta_0 = \frac{1}{\mu}$

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^*\right\|_2^2\right] &\leq (1 - 2\mu\eta_0)\mathbb{E}[\|\boldsymbol{\theta}^{(0)} - \theta^*\|^2] + \eta_0^2\sigma_g^2 \\
&= (1 - 2\mu\frac{1}{\mu})c_0 + \frac{\sigma_g^2}{\mu^2} \\
&= 0 \\
&\leq \frac{c_0}{0+1}
\end{aligned}$$

(b) Assume when $t = k - 1$ our consequence holds, which means($\mathbb{E}[\|\boldsymbol{\theta}^{(k)} - \theta^*\|^2] \leq \frac{c_0}{k+1}$), consider $t = k$. Note that $\eta_k = \frac{1}{\mu(k+1)}$

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^*\right\|_2^2\right] &\leq (1 - 2\mu\eta_k)\mathbb{E}[\|\boldsymbol{\theta}^{(k)} - \theta^*\|^2] + \eta_k^2\sigma_g^2 \\
&\leq (1 - 2\mu \times \frac{1}{\mu(k+1)})\frac{c_0}{k+1} + \frac{\sigma_g^2}{\mu^2(k+1)^2} \\
&= \frac{k-1}{(k+1)^2}c_0 + \frac{\sigma_g^2}{\mu^2(k+1)^2} \\
&= \frac{k}{(k+1)^2}c_0 \\
&\leq \frac{c_0}{k+2}
\end{aligned}$$

Combine (a) and (b), we used induction to get the proof of Theorem 0.1.
Q.E.D. □

4. (**10 Points**) In class, we defined the notion of a sub-gradient and worked out the example of the absolute function ($f(\theta) = |\theta|$).

   Consider the function from $\mathbb{R}^d \to \mathbb{R}$ which is the Euclidean norm for a vector, i.e., $f(\theta) := \|\theta\|_2$. Compute the sub-gradient of this function.

   For $x \neq \mathbf{0}$,

$$\nabla \|x\|_2 = \frac{x}{\|x\|_2}$$

   At $x = \mathbf{0}$, we know that $u \in \partial \|x\|_2$ if

$$\|y\|_2 \geq \|\mathbf{0}\|_2 + \langle y - \mathbf{0}, u \rangle = \langle y, u \rangle \quad \text{for all } y \in \mathbb{R}^n \tag{9}$$

   We can find u that meet these conditions using the Cauchy-Schwarz inequality. Note that

$$\langle y, u \rangle \leq \|y\|_2 \|u\|_2,$$

   so 9 will hold when $\|u\|_2 \leq 1$.

   On the other hand, if $\|u\|_2 > 1$, then for $y = u$, we have

$$\langle y, u \rangle = \|y\|_2^2 > \|y\|_2$$

   and 9 does not hold.
   Therefore

$$\partial \|x\|_2 = \begin{cases} \{u : \langle u, x \rangle = \|x\|_2, \|u\|_2 \leq 1\}, & x = \mathbf{0} \\ \frac{x}{\|x\|_2}, & x \neq \mathbf{0} \end{cases}$$

Pledge:
Please sign below (print full name) after checking (✓) the following. If you cannot honestly check each of these responses, please email me at kbala@ucdavis.edu to explain your situation.

- I pledge that I am a honest student with academic integrity and I have not cheated on this homework.
- These answers are my own work.
- I did not give any other students assistance on this homework (beyond what is allowed as per syllabus).
- I understand that to submit work that is not my own and pretend that it is mine is a violation of the UC Davis code of conduct and will be reported to Student Judicial Affairs.
- I understand that suspected misconduct on this homework will be reported to the Office of Student Support and Judicial Affairs and, if established, will result in disciplinary sanctions up through Dismissal from the University and a grade penalty up to a grade of "F" for the course.

Signature: Jingzhi Sun