Problem 1.
By definition, a differentiable function $f(\theta) : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$f(\lambda\theta_2 + (1-\lambda)\theta_1) \leq (\theta_2) + (1-\lambda)f(\theta_1),\ \forall\lambda \in (0,1)$$

Simplifying and rearranging we obtain

$$f(\theta_1 + \lambda(\theta_2 - \theta_1)) \leq f(\theta_1) + \lambda f(\theta_2) - \lambda f(\theta_1)$$

$$\Rightarrow \frac{f(\theta_1 + \lambda(\theta_2 - \theta_1)) - f(\theta_1)}{\lambda} \leq f(\theta_2) - f(\theta_1)$$

Noting the definition of a directional derivative, as $\lambda \to 0$:

$$f(\theta_2) - f(\theta_1) \geq \nabla f^T(\theta_1)(\theta_2 - \theta_1)$$

Rearranging, we obtain:

$$f(\theta_2) \geq f(\theta_1) + \nabla f^T(\theta_1)(\theta_2 - \theta_1)\blacksquare$$

Problem 2.

(a) To build our linear model, we utilized the *scikit-learn* package in Python and more specifically the **LinearRegression()** method. We obtained an R^2 of 0.51 on the training data and an R^2 of 0.5 on the test data.

(b) To predict the price of Bill Gates's house, we used the **predict()** method from *scikit-learn*. Our model predicted Bill Gates's house to be worth \$15,436,769.54.

(c) After fitting the linear model with the interaction effect between `bedrooms` and `bathrooms`, in addition to the four original variables, the R^2 for the training data remained the same at 0.51 and the R^2 for the test data slightly increased to 0.51. Therefore, it seems that adding this interaction effect improved the performance of the model.

(d) To implement the gradient descent algorithm, we began by initializing the OLS regression parameter vector $\vec{\beta}$ as a vector of ones. Also, to help the algorithm converge, we scaled the data by subtracting the mean and dividing by the standard deviation by using **StandardScaler()** from *scikit-learn*. Then, we defined two conditions that would cause the algorithm to terminate: 1) if the ℓ_2 norm of the gradient vector was less than τ or 2) if the algorithm reached the maximum number of iterations. Once one of those conditions is satisfied, the algorithm is terminated and the resulting $\vec{\beta}$ is returned. To calculate $\vec{\beta}$, we performed gradient descent using the following equation

$$\vec{\beta}^{(t+1)} = \vec{\beta}^{(t)} - \eta\nabla f(\vec{\beta}^{(t)})$$

where $\vec{\beta}^{(t+1)}$ is the OLS parameter vector at the current iteration, $\vec{\beta}^{(t)}$ is the OLS parameter vector at the previous iteration, η is the step size, and $\nabla f(\vec{\beta}^{(t)}) = X^T(X\vec{\beta}^{(t)} - y)$. We let $\tau = 0.01$ and set the maximum number of iterations to be 10,000. We achieved the most accurate results when using a step size of 0.00005 for the data without the interaction and a step size of 0.00009 for the data with the interaction.

Using our gradient descent algorithm, we obtained an R^2 of 0.51 on the training data and an R^2 of 0.5 on the testing data, which matches our R^2 values from part (a). The algorithm was terminated at iteration 167 once the ℓ_2 norm of the gradient was less than τ. Our gradient descent algorithm predicted the price of Bill Gates's house to be \$14,920,927.7, which is not too far off from the prediction in part (b). When including the interaction term between `bedrooms` and `bathrooms`, our gradient descent algorithm had an R^2 of 0.51 on both the train and test sets, which matches our result from part (c). Adding the interaction caused the algorithm to run for a big longer since it was terminated at iteration 1,278 once the ℓ_2 norm of the gradient was less than τ.

(e) To implement the stochastic gradient descent algorithm, we began by initializing the OLS regression parameter vector $\vec{\beta}$ as a vector of ones. Then, we set the number of iterations and begin running the algorithm. We first obtained a random sample of size m, which is our batch size. We then selected the observations from the dataset at the index(es) of the random sample and then performed SGD to compute $\vec{\beta}$. To calculate $\vec{\beta}$, we performed stochastic gradient descent using the following equation

$$\vec{\beta}^{(t+1)} = \vec{\beta}^{(t)} - \eta \frac{1}{m} \nabla F(\vec{\beta}^{(t)}, \xi^{(t)})$$

where $\vec{\beta}^{(t+1)}$ is the OLS parameter vector at the current iteration, $\vec{\beta}^{(t)}$ is the OLS parameter vector at the previous iteration, η is the step size, m is the batch size, and $\nabla F(\vec{\beta}^{(t)}, \xi^{(t)}) = X^T(X\vec{\beta}^{(t)} - y)$. It is important to note that $\vec{\beta}^{(t+1)}$ and $\nabla F(\vec{\beta}^{(t)}, \xi^{(t)})$ are random vectors. We achieved the most accurate results using a step size of 0.0008 with 18,000 iterations for the data without the interaction effect and a step size of 0.0009 with 15,000 iterations for the data with the interaction.

Using our stochastic gradient descent algorithm, we obtained an R^2 that ranged from 0.49-0.51 on the training data and an R^2 of 0.48-0.5 on the testing data. Our SGD algorithm predicted the price of Bill Gates's house to be somewhere between \$14.8M to \$15.7M, which is generally closer to the prediction in part (b) than the prediction from gradient descent was. When including the interaction term between `bedrooms` and `bathrooms`, our SGD algorithm had an R^2 in the range of 0.49-0.51 on both the train and test sets. It is worth noting that the fluctuation in R^2 values is due to the randomness in the SGD algorithm.

Problem 3.
First, We prove that if the function is μ-strongly convex, then $(\nabla f(\theta_1) - \nabla(f\theta_2)^T(\theta_1 - \theta_2) \geq \mu\|\theta_1 - \theta_2\|^2$:
It is given that f is differentiable, in which case, being *strictly* convex is equivalent to

$$= f(\theta_2) \geq f(\theta_1) + \nabla f(\theta_1)^T(\theta_2 - \theta_1)$$

For differentiable and strongly convex functions, We also have that

$$g(x) = f(x) - \frac{\mu}{2}||x||^2$$

We also have that the gradient of a convex function is monotone, s.t.

$$\nabla g(\theta_1)^T(\theta_1 - \theta_2) - \nabla g(\theta_2)^T(\theta_1 - \theta_2) \geq 0$$

With this information we can now show what we need to show:

$$= (\nabla f(\theta_1) - \mu\theta_2)^T(\theta_1 - \theta_2) - (\nabla f(\theta_2) - \mu\theta_2)^T(\theta_1 - \theta_2) \geq 0 \tag{1}$$

$$= (\nabla f(\theta_1)^T - \mu\theta_2^T)(\theta_1 - \theta_2) - (\nabla f(\theta_2)^T - \mu\theta_2^T)(\theta_1 - \theta_2) \geq 0 \tag{2}$$

$$= \nabla f(\theta_1)^T(\theta_1 - \theta_2) - \mu\theta_1^T(\theta_1 - \theta_2) - \nabla f(\theta_2)^T(\theta_1 - \theta_2) + \mu\theta_2^T(\theta_1 - \theta_2) \geq 0 \tag{3}$$

$$= (\nabla f(\theta_1) - \nabla f(\theta_2))^T(\theta_1 - \theta_2) \geq \mu\theta_1^T(\theta_1 - \theta_2) - \mu\theta_2^T(\theta_1 - \theta_2) \tag{4}$$

$$= (\nabla f(\theta_1) - \nabla f(\theta_2))^T(\theta_1 - \theta_2) \geq \mu(\theta_1 - \theta_2)^T(\theta_1 - \theta_2) \tag{5}$$

$$= (\nabla f(\theta_1) - \nabla f(\theta_2))^T(\theta_1 - \theta_2) \geq \mu||\theta_1 - \theta_2||^2 \tag{6}$$

Now, for the proof of Theorem 6.1:
Assuming $M_g = 0$, $\eta_t = \frac{c}{t+1}$ we have

$$\mathbb{E}\big[||\theta^{t+1} - \theta^*||^2\big] \leq (1 - 2\mu\eta_t)\mathbb{E}\big[||\theta^t - \theta^*||^2\big] + \eta_t^2\sigma_g^2 \tag{7}$$

To prove the convergence rate by induction, first consider the base case where $t = 0$:

$$\mathbb{E}\big[||\theta^1 - \theta^*||^2\big] \leq \left(1 - 2\mu\frac{c}{1}\right)\mathbb{E}\big[||\theta^0 - \theta^*||^2\big] + c^2\sigma_g^2 \tag{8}$$

Letting $c = \frac{1}{\mu}$ This simplifies to:

$$[||\theta^1 - \theta^*||^2] \leq \frac{\max\left(\mathbb{E}\big[||\theta^0 - \theta^*||^2\big], \frac{\sigma^2}{\mu^2}\right)}{1} \tag{9}$$

To prove the convergence rate, we assume $t = k - 1$ is true, and then must show the inequality holds for $t = k$: Let $c_0 := \frac{\max\left(\mathbb{E}[||\theta^k - \theta^*||^2], \frac{\sigma^2}{\mu^2}\right)}{k+1}$

$$\mathbb{E}\big[||\theta^{k+1} - \theta^*||^2\big] \leq \left(1 - 2\mu\frac{c}{k+1}\right)\mathbb{E}\big[||\theta^k - \theta^*||^2\big] + \frac{c^2}{(k+1)^2}\sigma^2 \tag{10}$$

Plugging in c_0 for $\mathbb{E}[||\theta^k - \theta^*||^2]$ and letting $c = \frac{1}{\mu}$:

$$\mathbb{E}\big[||\theta^{k+1} - \theta^*||^2\big] \leq \left(1 - \frac{2}{k+1}\right)\frac{c_0}{k+1} + \frac{1}{\mu^2(k+1)^2}\sigma^2 \tag{11}$$

Now Plugging in c_0 for $\frac{\sigma^2}{\mu^2}$:

$$\mathbb{E}\big[||\theta^{k+1} - \theta^*||^2\big] \leq \left(1 - \frac{2}{k+1}\right)\frac{c_0}{k+1} + \frac{c_0}{(k+1)^2} \tag{12}$$

$$\mathbb{E}\big[||\theta^{k+1} - \theta^*||^2\big] \leq \left(\frac{1}{k+1} - \frac{1}{(k+1)^2}\right)c_0 \tag{13}$$

$$\mathbb{E}\big[||\theta^{k+1} - \theta^*||^2\big] \leq \left(\frac{k}{(k+1)^2}\right)c_0 \tag{14}$$

Note that $\left(\frac{k}{(k+1)^2}\right)c_0 \leq \left(\frac{k+1}{(k+1)^2}\right)c_0$, so if the expectation on the left is bounded by $\left(\frac{k}{(k+1)^2}\right)c_0$, then it is also bounded by the slightly larger $\left(\frac{k+1}{(k+1)^2}\right)c_0$. Simplifying,

$$\mathbb{E}[||\theta^{k+1} - \theta^*||^2] \leq \frac{c_0}{(k+1)} \tag{15}$$

This completes the proof. ■

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

- We pledge that we are honest students with academic integrity and we have not cheated on this homework. ✓
- These answers are our own work. ✓
- We did not give any other students assistance on this homework. ✓
- We understand that to submit work that is not our own and pretend that it is our is a violation of the UC Davis code of conduct and will be reported to Student Judicial Affairs. ✓
- We 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. ✓

Team Member 1
Collin Kennedy

Team Member 2
Grant Gambetta