6.867 Machine Learning

Homework 2

1 IMPLEMENTING GRADIENT DESCENT

Gradient descent is an iterative procedure for finding a vector that minimizes an objective function. Typically, the vector is our parameters and the objective function is the cost function. In each iteration, we find the gradient of the objective function evaluated at the vector and update the vector in the direction of the negative gradient.

1.1 Basic Gradient Descent

To demonstrate the gradient descent algorithm, we began by finding the minimum of two well defined functions with closed form derivatives shown below.

Negative Gaussian:

$$f(x) = -\frac{1}{\sqrt{(2\pi)^n|\Sigma|}} exp[-\frac{1}{2}(x-u)^T\Sigma^{-1}(x-u)] \tag{1}$$

$$\frac{\partial f(x)}{\partial x} = -f(x)\Sigma^{-1}(x-u) \tag{2}$$

Quadratic Bowl:

$$f(x) = \frac{1}{2}x^TAx - x^Tb \tag{3}$$

$$\frac{\partial f(x)}{\partial x} = Ax - b \tag{4}$$

In each iteration, update x according to

$$x_{t+1} = x_t - \eta \bigtriangledown f(x) \tag{5}$$

where η is the step size. Use the gradient function above to directly compute the gradient at x.

The gradient descent algorithm also takes three structural parameters, the starting guess, step size, and convergence criterion. Each affects the end result of the algorithm. The starting guess is important because gradient descent iteratively follows the gradient. Thus it can get stuck in local minimum. By running the algorithm repeated with random initialization, we increase our chance of finding a global minimum. The step size affects convergence behavior of the algorithm. As shown in figures 1-3, if the step size is too small, the algorithm will converge slowly. If the step size is slightly too big, the algorithm will oscillate. If the step size is much too big, the algorithm will actually diverge. The convergence criterion is used to determine whether the algorithm has converged or not. It is a threshold for the difference in the cost function in two successive iterations. If the difference of the cost function of two successive iterations is less than the convergence criterion, the algorithm is determined to have converged. Decreasing this gives greater accuracy but increases runtime.

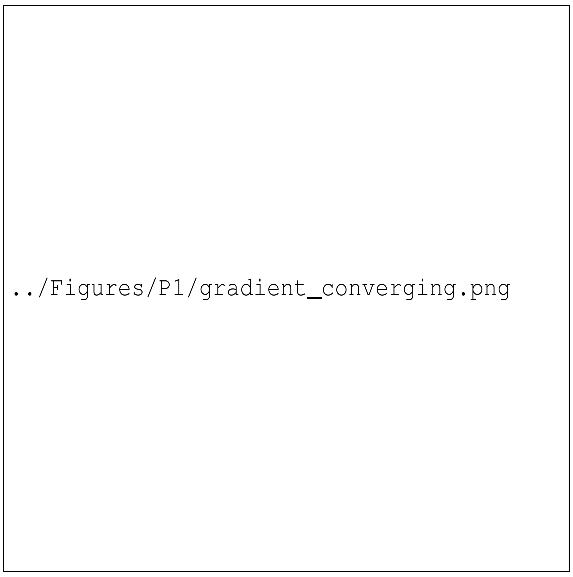

Figure 1: Gradient descent on negative gaussian with step size η = 10^7. The norm of the gradient is converging to zero.

1.2 Central Difference Approximation

For many objective functions, it is impossible to write a closed form gradient function. Thus, use central difference approxiation to approximate the gradient. For each dimension i, estimate its partial derivative by

$$\bigtriangledown_i f(x) = \frac{f(x+d*\hat{\imath}) - f(x-d*\hat{\imath})}{2d} \tag{6}$$

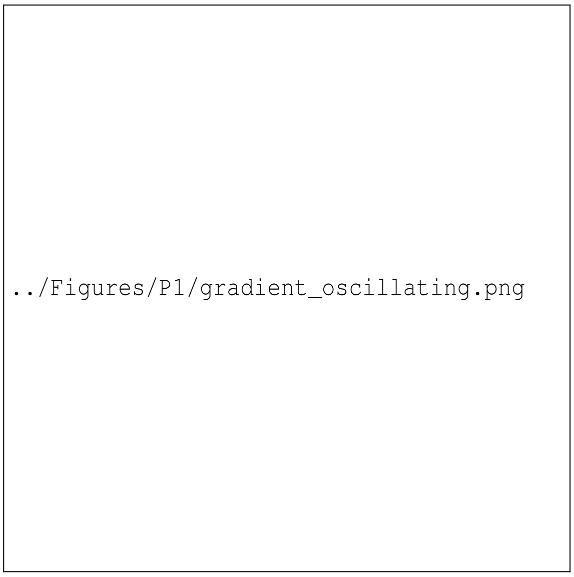

Figure 2: Gradient descent on negative gaussian with step size $\eta = 10^7$. The norm of the gradient is oscillating and never converging.

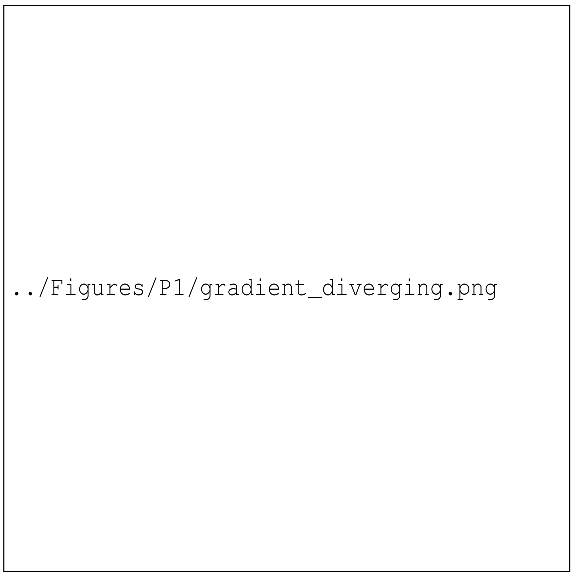

Figure 3: Gradient descent on negative gaussian with step size $\eta = 10^7$. The norm of the gradient is diverging.

for some small d. In general, the larger d is, the more inaccurate the gradient approximation is. To check the correctness of our gradient functions, we compared the computed gradient to the numerical approximation of the gradient evaluated with $d = 10^{-3}$. Interestingly, for the quadratic bowl function, the numerical approximation is always accurate regardless of how large d is. This is because the derivative of the quadratic bowl is linear in x.

1.3 Batch vs. Stochastic Gradient Descent

We used gradient descent to find the parameters θ that minimized the least squared error objective function

$$J(\theta) = ||X\theta - y||^2 \tag{7}$$

where each row of X and y is a data sample pair.

In batch gradient descent, θ is updated with the gradient of the cost function for the entire training dataset. We used the gradient function

$$\bigtriangledown_\theta J(\theta) = 2X(X\theta - y) \tag{8}$$

Using a step size of 0.01 and a convergence criterion of $10^{-}8$, the quadratic bowl function converges in 179 iterations. The dataset had 100 points so it took 17900 point wise gradient evaluations.

In contrast, stochastic gradient descent updates θ with the gradient of the cost function for each data point. θ is updated according to

$$\theta_{t+1} = \theta_t - \eta_t \bigtriangledown_\theta J(\theta_t; x^{(i)}, y^{(i)}) \tag{9}$$

where η is the learning rate and

$$J(\theta_t; x^{(i)}, y^{(i)}) = (x^{(i)T}\theta_t - y^{(i)})^2 \tag{10}$$

$$\bigtriangledown_\theta J(\theta_t; x^{(i)}, y^{(i)}) = 2(x^{(i)T}\theta_t - y^{(i)})x^{(i)}. \tag{11}$$

We iterate through the entire dataset for n rounds and for each round, we iterate through the data points in a different order. We converge when the cost between rounds decreases by less than a certain threshold. The stochastic gradient descent took 200000 pointwise gradient evaluations. The batch gradient descent performed much faster to achieve the same level of convergence but in practice, it is sometimes infeasible to compute gradients on large datasets. For such datasets, stochastic gradient descent can be the best option.

2 LINEAR BASIS FUNCTION REGRESSION

In the supervised learning problem of regression, we are given a set of n data points and n target values and the goal is to find a function that relates x to y. We have to do this in such a way that this function generalizes well to unseen values of x. Linear Basis Function Regression aims to find the optimal linear combination of basis functions to create a function mapping x to y. This linear combination is expressed as a vector of weights for each of these basis functions. These basis functions take the form

$$\phi_i(x) = x^i, \forall i \in [0, M] \tag{12}$$

We evaluated different methods of finding the optimal weight vector, w. To do this, we generated 11 points from $y(x) = \cos(\pi x) + 1.5\cos(2\pi x)$ with some added noise.

2.1 Closed Form Solution

From the textbook, we know the closed form solution for the maximum likelihood weight vector given data values X, target values Y, and the value of M, where M is the highest order polynomial in our polynomial basis. The closed form solution for the maximum likelihood weight vector is

$$(\phi^T \phi)^{-1} \phi^T Y \tag{13}$$

where ϕ is the design matrix. We tested our solution by plotting the polynomials our models generated against the 11 data points. As can be seen in figures 4-7, when M is larger the higher polynomial model overfits to the data. We found the optimal M to be 4.

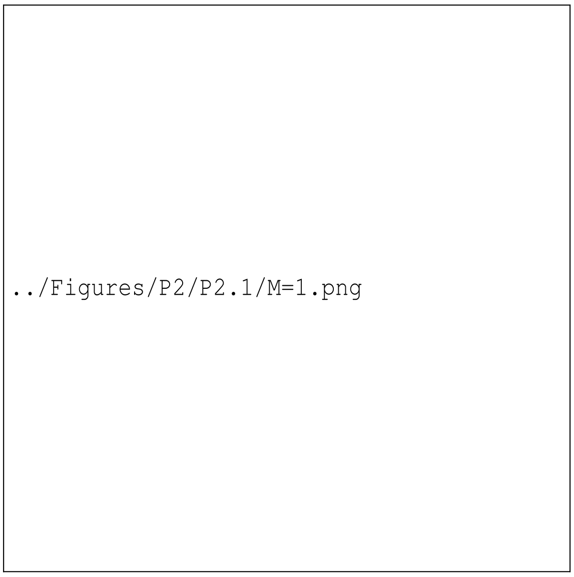

Figure 4

2.2 Objective Function

The objective function we were trying to minimize was Sum Squared Error. The function and its derivative are

$$f(x) = \frac{1}{2} \sum_{n=1}^{N} (y_n - w^T \phi(x_n))^2 \tag{14}$$

$$\bigtriangledown f(x) = \sum_{n=1}^{N} (y_n - w^T \phi(x_n)) \phi(x_n)^T \tag{15}$$

. We then used our gradient descent algorithms to find the weight vector that minimized the Sum Squared error.

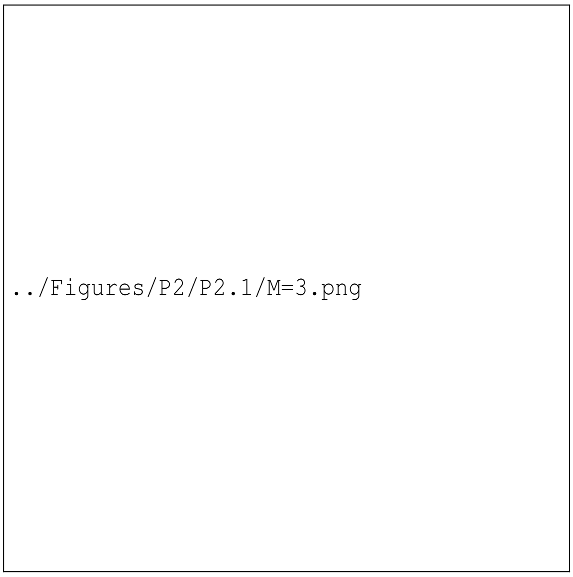

Figure 5

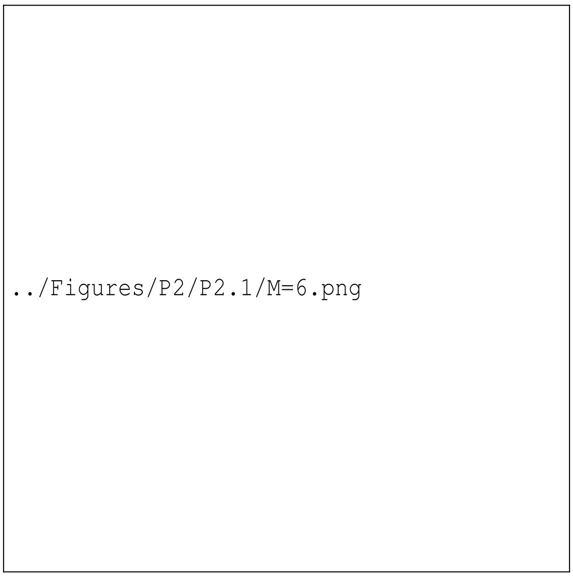

Figure 6

2.3 Batch vs Stochastic Gradient Descent

We tested different combinations of convergence criterion, step size, initial guess, and M. We found that with batch gradient descent, a step size of 0.1

../Figures/P2/P2.1/M=8.png

Figure 7

achieved the best results. The algorithm achieved lower costs on the cost function and required less iterations than smaller step sizes. Values lower than that would cause the algorithm to converge to a suboptimal value, and values higher than that caused the algorithm to oscillate and fail. We believe that this is because the loss function is less sensitive to changes, allowing for larger step sizes without oscillation. Additionally, we believe that this insensitivity to changes makes it easy for the function to have nearly identical values on successive iterations when the step size is small, causing the algorithm to converge to suboptimal weights. A lower convergence criterion led to better results, but came with the tradeoff of longer convergence times. For example, with $M = 4$ and a step size of .01, lowering the convergence criterion from 10^{-6} to 10^{-7} quintupled the number of iterations needed and improved the cost by .02. We initialized our parameter vector at random and it didn't affect the converged values of the algorithm. We believe this is because the function we are minimizing is convex, thus the algorithm always moves towards the minimum and eventually converges to it.

With Stochastic Gradient Descent, we found similar trends. Lowering the convergence criterion led to lower costs but more iterations. Increasing the step size (with a max value of 1) generally led to lower costs. A step size of 1 would sometimes give the best results, but other times it caused the algorithm to oscillate and fail. We initialized our parameter vector at random several times and it didn't affect the converged values of the algorithm.

Batch gradient descent achieved better results, yielding lower costs in nearly every configuration we tested. However, it took a significant number of more iterations than stochastic gradient descent. It often needed more than 10 times the number of iterations.

2.4 Cosine Basis Functions

We then tested our function for computing the maximum likelihood weight vector by using our prior knowledge about the origin of the data. We used cosine basis functions instead of polynomial ones. The real values of the function that generated the data are $\cos(\pi x) + 1.5\cos(2\pi x)$, which means its weight vector would be $[1, 1.5]$. Our calculations yielded $[0.7789928, 1.17413213]$. The plot of our model can be seen in Figure 8.

../Figures/P2/P2.3/cosine_graph.png

Figure 8: Regression using cosine basis.

3 RIDGE REGRESSION

Linear regressions, especially those with sum squared error cost functions, are prone to highly overfit the data. One way to ameliorate this is to introduce a regularization term on your weight vector. Ridge regression punishes weight vectors with large sizes. The parameter λ controls how much the size of a vector is penalized.

3.1 Closed Form Solution

We tested different values of λ on ridge regressions on a small data set. The tested values of λ ranging from 1000 to .00001. With very large λ values, the weight vector became the zero vector. In general, as λ got smaller, the best fit line produced by the regression was able to fit the data better. This is because when $\lambda = 0$, then it is a normal regression problem with a least squares solution. λ effectively works as a bias term, the larger it is, the lower the variance of the weight vector.

3.2 Results

We decided to evaluate the effect λ and M have on unseen data. To evaluate this, we trained several models on a training set, A, and then used a validation set to determine what models performed the best. We then calculated the error that each of these selected models had on an unseen test set. The models that performed the best on the test set had very low λ values (between .000001 and .0001) and had $M = 6$. Large λ values tended to make the function flatter, even if it had a high order polynomial basis. Low λ values allowed the model to fit the data well. Models with different combinations of λ and M values are shown in figures 9 - 11.

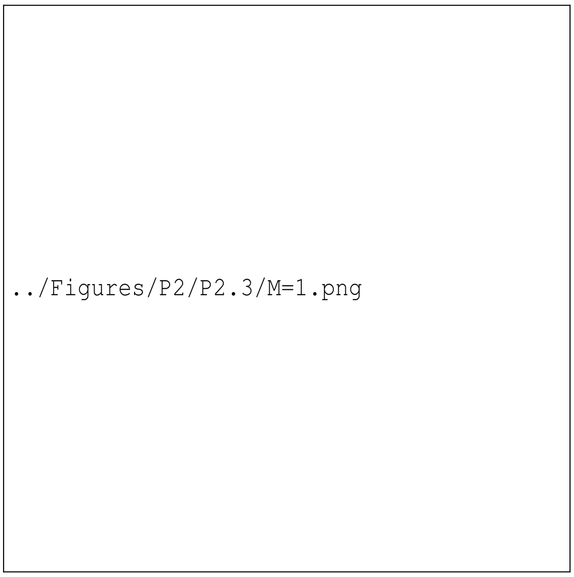

Figure 9: $M = 1$ with different values for λ

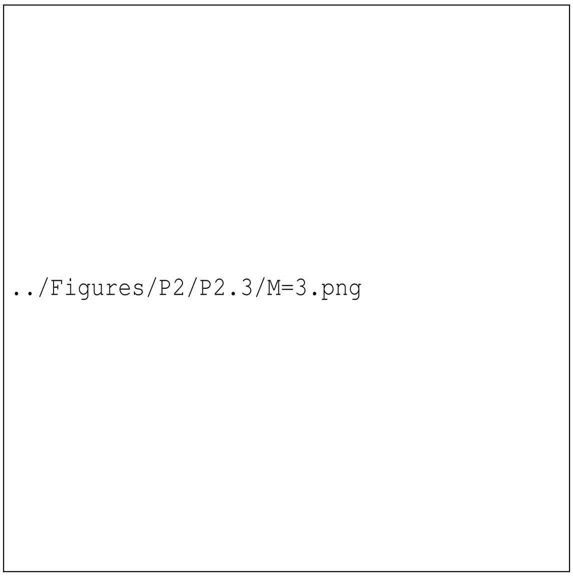

Figure 10: $M = 3$ with different values for λ

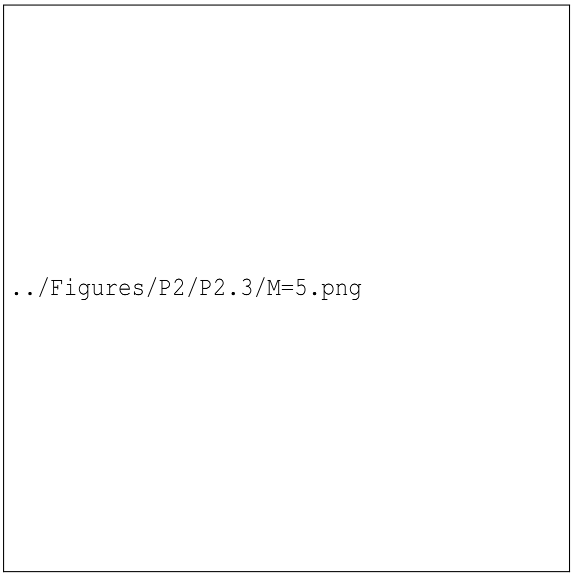

Figure 11: Higher order polynomial basis ($M = 5$) with different λ.

4 LASSO REGRESSION

In some applications, we have reason to believe the underlying function generating the data has sparse weights. In this case, using the l_1 norm for regularization, LASSO regression, can perform better than using the l_2 norm in ridge regression. To demonstrate this, we have a dataset generated according to

$$y = w_{true}^T \phi(x) + \varepsilon \tag{16}$$

ε is some error and $\phi(x)$ is a basis vector defined as

$$\phi(x) = (x, \sin(0.4\pi x * 1), ..., \sin(0.4\pi x * 12)) \tag{17}$$

w_{true}^T is a vector with sparse weights. We will try to estimate them from the training data by performing batch gradient descent to minimize the following objective function for LASSO:

$$\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - w^T\phi(x^{(i)}))^2 + \lambda\sum_{j=1}^{M}|w_j| \qquad (18)$$

Different values of λ give different weight estimates. The larger λ is, the more sparse the weight estimates will be. Unsurprisingly, when λ is extremely large, the weights are all zero. These weight estimates can then be plotted by $y = w_{est}^T\phi(x)$ to see how well they performed against the training data. As you can see in Figure 12, the lower λ is, the more closely the function matches the training data.

Then, from all possible λ, w_{est} pairs, we used to validation dataset to choose the pair that performed best. The best λ was $10^{-2.55}$ and it yielded an average squared error from the test dataset of 0.0484489. It is depicted in Figure 13.

Using a same procedure with ridge regression, we minimize the objective function

$$\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - w^T\phi(x^{(i)}))^2 + \lambda\sum_{j=1}^{M}(w_j)^2 \qquad (19)$$

generate weight estimates, and plot them against the sample data. In figure 14, we see weight estimates from different lambdas plotted against the training data.

Figure 15 plots the best ridge regression as selected from the validation data set. It is clear the ridge regression performs much worse than LASSO. In fact, the average squared error is 0.635523, 50% more than LASSO. LASSO was much better at guessing the sparse true weights.

../Figures/P4/lasso_varying.png

Figure 12: LASSO weight estimates from varying λs graphed against training data.

../Figures/P4/lasso_best.png

Figure 13: Best LASSO weight estimate.

../Figures/P4/ridge_varying.png

Figure 14: Ridge regression weight estimates from varying λs graphed against training data.

../Figures/P4/ridge_best.png

Figure 15: Best ridge regression weight estimate.