STAT 154 Notes

Raaz Dwivedi

UC Berkeley

February 21, 2019

Least Squares

- We consider the following optimization problem:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{X}\theta - \mathbf{y}\|_2^2 \qquad \text{(Least Squares)}$$

  where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the feature matrix and $\mathbf{y} \in \mathbb{R}^n$ is the set of observations.
- Note that the normalization factor $\frac{1}{2}$ is used for convenience later on–it does not change the optimization problem.
- When $n \geq d$ and $\mathbf{X}$ is full column rank, one way to compute the solution to this problem is using the closed form expression:

$$\theta^{\mathrm{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X} \mathbf{y}.$$

Convex optimization

- We can also study least squares in the framing of convex optimization.
- Consider the following optimization problem

$$\min_{\theta} f(\theta) \quad \text{where } f \text{ is convex.} \tag{1}$$

- Note that because of convexity all minimizers $\theta^\star$ satisfy

$$\nabla_\theta f(\theta^\star) = 0.$$

  In fact the closed form for θ^{OLS} is obtained by solving for this equation.

Gradient Descent

- A popular algorithm for finding these stationary points is gradient descent

$$\theta^{t+1} = \theta^t - \eta \nabla_\theta f(\theta^t).$$

- Note that these updates will not move when $\theta^t = \theta^\star$.
- Note that the direction of $-\nabla_\theta f$ is also the direction of steepest descent so for small enough η, the gradient step will reduce the function value.
- We now see some details of gradient descent for least-squares.

Gradient Descent on Least Squares

- We now study how the gradient descent method behaves when applied to the Least squares problem.
- First we expand the objective so that gradient computation is easy:

$$f(\theta) = \frac{1}{2}\|\mathbf{X}\theta - \mathbf{y}\|_2^2 = \frac{1}{2}\theta^\top(\mathbf{X}^\top\mathbf{X})\theta + \frac{1}{2}\mathbf{y}^\top\mathbf{y} - \theta^\top\mathbf{X}^\top\mathbf{y}.$$

- Then we have

$$\nabla_\theta f(\theta) = \mathbf{X}^\top\mathbf{X}\theta - \mathbf{X}^\top\mathbf{y}.$$

And hence gradient descent updates with step size η become:

$$\begin{aligned}\theta^{t+1} &= \theta^t - \eta\nabla_\theta f(\theta^t) \\ &= \theta^t - \eta(\mathbf{X}^\top\mathbf{X}\theta^t - \mathbf{X}^\top\mathbf{y}).\end{aligned}$$

How do we choose the step size?

Gradient Descent on Least Squares

- We now derive the recursion in the error. To simplify our calculations, we assume that there exists $\theta^\star$ such that $\mathbf{y} = \mathbf{X}\theta^\star$.
- Under that assumption, the gradient descent updates simplify to

$$\begin{aligned}
\theta^{t+1} &= \theta^t - \eta((\mathbf{X}^\top\mathbf{X})\theta^t - \mathbf{X}^\top\mathbf{y}) \\
&= \theta^t - \eta((\mathbf{X}^\top\mathbf{X})\theta^t - \mathbf{X}^\top\mathbf{X}\theta^\star) \\
&= \theta^t - \eta\mathbf{X}^\top\mathbf{X}(\theta^t - \theta^\star).
\end{aligned}$$

- As a result, we have

$$\begin{aligned}
\theta^{t+1} - \theta^\star &= \theta^t - \eta\mathbf{X}^\top\mathbf{X}(\theta^t - \theta^\star) - \theta^\star \\
&= (\mathbf{I} - \eta\mathbf{X}^\top\mathbf{X})(\theta^t - \theta^\star) \\
&\vdots \\
&= (\mathbf{I} - \eta\mathbf{X}^\top\mathbf{X})^{t+1}(\theta^0 - \theta^\star).
\end{aligned}$$

Gradient Descent on Least Squares

- Now to choose the step size, we see that

$$\begin{aligned}\|\theta^t - \theta^\star\|_2 &= \|(\mathbf{I} - \eta \mathbf{X}^\top \mathbf{X})^t (\theta^0 - \theta^\star)\|_2 \\ &= \|(\mathbf{I} - \eta \mathbf{X}^\top \mathbf{X})\|_2^t \|\theta^0 - \theta^\star\|_2\end{aligned}$$

  where we use $\|\mathbf{A}\|_2$ to denote the operator norm of the matrix.
- If the eigenvalues of the matrix $\mathbf{X}^\top \mathbf{X}$ lie between m and L, then we have

$$\lambda(\mathbf{I} - \eta \mathbf{X}^\top \mathbf{X}) \in [1 - \eta L, 1 - \eta m].$$

Gradient Descent on Least Squares

- Note that the operator norm in this case is bounded as

$$\begin{aligned}\|(\mathbf{I}-\eta\mathbf{X}^\top\mathbf{X})\|_2 &= \max\left\{\left|\lambda_{\min}(\mathbf{I}-\eta\mathbf{X}^\top\mathbf{X})\right|, \left|\lambda_{\max}(\mathbf{I}-\eta\mathbf{X}^\top\mathbf{X})\right|\right\} \\ &\leq \underbrace{\max\left|1-\eta L\right|, \left|1-\eta m\right|}_{=:\alpha}.\end{aligned}$$

- Thus we have

$$\begin{aligned}\|\theta^t-\theta^\star\|_2 &= \|(\mathbf{I}-\eta\mathbf{X}^\top\mathbf{X})\|_2^t\|\theta^0-\theta^\star\|_2 \\ &\leq \alpha^t\|\theta^0-\theta^\star\|_2\end{aligned}$$

- So as long as

$$\alpha := \max\left|1-\eta L\right|, \left|1-\eta m\right| < 1$$

  we have a geometric rate of convergence.
- Verify that this holds for any $\eta \in (0, 2/L)$.

Gradient Descent on Least Squares

- That is

$$\|\theta^t - \theta^\star\|_2 \leq \epsilon \quad \text{for } t \geq \frac{\log(\|\theta^0 - \theta^\star\|_2/\epsilon)}{\log(1/\alpha)}.$$

- Verify that α is guaranteed to lie in between 0 and 1 for any step size that satisfies $\eta \in (0, 2/L)$.