STAT 154 Notes: Gradient Descent (GD) for LS

Raaz Dwivedi

UC Berkeley

February 26, 2019

Least Squares

- We consider the following optimization problem:

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

  where $\mathbf{X} \in \mathbb{R}^{n \times p}$ 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 $n \geq p$ and $\mathbf{X}$ is full column rank, one way to compute the solution to this problem is using the closed form expression:

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

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

$$\theta_{t+1} = \theta_t - \gamma \nabla_\theta f(\theta_t).$$

- 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 γ, 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 γ become:

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

  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 $\mathbf{y} = \mathbf{X}\theta^\star$ (in general, we can use $H\mathbf{y} = X^T\theta^{OLS}$ or $\theta^* = \theta^{OLS}$).
- Under that assumption, the gradient descent updates simplify to

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

- As a result, we have

$$
\begin{aligned}
\theta_{t+1} - \theta^\star &= \theta_t - \gamma\mathbf{X}^\top \mathbf{X}(\theta_t - \theta^\star) - \theta^\star \\
&= (\mathbf{I} - \gamma\mathbf{X}^\top \mathbf{X})(\theta_t - \theta^\star) \\
&\vdots \\
&= (\mathbf{I} - \gamma\mathbf{X}^\top \mathbf{X})^{t+1}(\theta_0 - \theta^\star).
\end{aligned}
$$

Gradient Descent on Least Squares

- Now to choose the step size, we see that

$$\begin{aligned}\|\theta_t - \theta^\star\|_2 &= \|(\mathbf{I} - \gamma \mathbf{X}^\top \mathbf{X})^t (\theta_0 - \theta^\star)\|_2 \\ &\leq ||(\mathbf{I} - \gamma \mathbf{X}^\top \mathbf{X}||_{op}^t \|\theta_0 - \theta^\star\|_2\end{aligned}$$

  where we use $||\mathbf{A}||_{op}$ to denote the operator norm of the matrix (also called the spectral norm).
- If the eigenvalues of the matrix $\mathbf{X}^\top \mathbf{X}$ lie between m and L, then we have the eigen values of $\mathbf{I} - \gamma \mathbf{X}^\top \mathbf{X}$ lie in

$$[1 - \gamma L, 1 - \gamma m].$$

Gradient Descent on Least Squares

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

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

- Thus we have

$$\begin{aligned} \|\theta_t - \theta^\star\|_2 &= ||\mathbf{I} - \gamma \mathbf{X}^\top \mathbf{X}||_{op}^t \|\theta_0 - \theta^\star\|_2 \\ &\leq \alpha^t \|\theta_0 - \theta^\star\|_2 \end{aligned}$$

- So as long as

$$\alpha := \max |1 - \gamma L|, |1 - \gamma m| < 1$$

we have a geometric rate of convergence to the LS solution θ^{OLS} in the general case.

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 $\gamma \in (0, 2/L)$.