EECS-553 Homework 3

Vipransh Sinha

March 6, 2025

Exercise 1: Linear Regression with Gradient Descent

Given $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times d}$, we consider the linear regression problem $J(w) = 0.5\|y - Xw\|_{l2}^2$ where $w \in \mathbb{R}^d$ is the weight vector. In this problem, we will consider the gradient descent algorithm for minimizing $J(w)$. With step size $\eta > 0$, the algorithm is given by

$$w_{t+1} = w_t - \eta \nabla J(w_t)$$

(a) Recall that Hessian matrix is $H = X^T X$. Show that the gradient of $J(w)$ is $\nabla J(w) = Hw - X^T y$.

(b) Suppose $H \succ 0$. Prove that, unique global minima is $w^* = (X^T X)^{-1} X^T y$.

(c) Define residual $e_t = w_t - w^*$. Prove the iterations obey the following recursion

$$e_{t+1} = e_t - \eta H e_t.$$

(d) Prove that $\|e_t\|_2 \leq \text{rate}_\eta^t \|e_0\|_2$, where $\text{rate}_\eta = \|I - \eta H\|$ where $\|\cdot\|$ denoting the spectral norm of a matrix.

(e) Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues of H. Prove that $\text{rate}_\eta = \max(1 - \eta\lambda_{\min}, \eta\lambda_{\max} - 1)$.

(f) What is the optimal step size η^* that ensures fastest convergence of gradient descent by minimizing rate_η? What is the corresponding convergence rate rate_{η^*}?

Solution

(a) To find $\nabla J(w)$ we must first rewrite the problem, expand, and the simplify:

$$\begin{aligned}
\nabla J(w) &= \nabla \frac{1}{2}(y - Xw)^T(y - Xw) \\
&= \nabla \frac{1}{2}(y^T y - w^T X^T y - y^T Xw + w^T X^T Xw) \\
&= \frac{1}{2}(-2y^T X + 2X^T Xw) \\
\nabla J(w) &= Hw - X^T y
\end{aligned}$$

(b) To show that the unique global minima is $w^* = (X^T X)^{-1} X^T y = H^{-1} X^T y$, we must set the gradient to 0, and solve for w^*:

$$\begin{aligned}
&0 = Hw^* - X^Y y \\
\Rightarrow\ & X^T y = Hw^* \\
\Rightarrow\ & w^* = H^{-1} X^T y \\
\Rightarrow\ & w^* = (X^T X)^{-1} X^T y
\end{aligned}$$

(c) Substituting we get:

$$\begin{aligned}
w_{t+1} - w^* &= (w_t - w^*) - \eta H(w_t - w^*) \\
w_{t+1} &= w_t - \eta H(w^* - w_t)
\end{aligned}$$

To show that the iterations obey the recursion, we must show that $\eta \nabla J(w_t) = Hw_t - X^T y = \eta H(w_t - w^*)$, from the initial given gradient descent algorithm.

$$\begin{aligned}
Hw_t - X^T y &= H(w_t - w^*) \\
&= Hw_t - Hw^* \\
&= Hw_t - X^T Xw^* \\
&= Hw_t - X^T X((X^T X)^{-1} X^T y) \\
&= Hw^t - X^T y
\end{aligned}$$

Thus we have shown the iterations obey $e_{t+1} = e_t - \eta He_t$.

(d) Rewriting the problem, we have to show $\|e_t\|_2 \leq \|I - \eta H\|^t \|e_0\|_2$. We can derive this from the previous part's given of:

$$e_{t+1} = e_t - \eta He_t$$

Rewriting this using recursion, we get:

$$e_t = (I - \eta H)^t e_0$$

Taking the $l2$ norm of this on both sides we get

$$\|e_t\|_2 = \|(I - \eta H)^t e_0\|_2$$

Finally, using the submulitiplicative of norms, we arrive at our result of

$$\|e_t\|_2 \leq \|I - \eta H\|^t \|e_0\|$$

(e) We can rewrite the problem into: $\|I - \eta H\| = \max(1 - \eta\lambda_{\min}, \eta\lambda_{\max} - 1)$. Since the spectral norm can be thought of as the maximum eigenvalue of $I - \eta H$, we get

$$\text{rate}_\eta = \max |1 - \eta\lambda_i|$$

Taking the two extremes of eigenvalues of H, we get cases of

$$\begin{aligned} \text{For } \lambda_{\min} &\Rightarrow |1 - \eta\lambda_{min}| \Rightarrow 1 - \eta\lambda_{min} \\ \text{For } \lambda_{\max} &\Rightarrow |1 - \eta\lambda_{max}| \Rightarrow \eta\lambda_{min} - 1 \end{aligned}$$

Thus we get: $\text{rate}_\eta = \max(1 - \eta\lambda_{\min}, \eta\lambda_{\max} - 1)$.

(f) We need to solve:

$$\min_\eta \max(1 - \eta\lambda_{\min}, \eta\lambda_{\max} - 1)$$

We notice that the terms inside of the maximum are minimized when they are equal, giving:

$$\begin{aligned} & 1 - \eta\lambda_{\min} = \eta\lambda_{\max} - 1 \\ \Rightarrow \quad & 2 = \eta(\lambda_{\max} + \lambda_{\min}) \\ \Rightarrow \quad & \eta^* = \frac{2}{\lambda_{\min} + \lambda_{\max}} \end{aligned}$$

Using this we can find the optimal rate by substituting into either expression to be:

$$\begin{aligned} \text{rate}_{\eta^*} &= 1 - \eta^*\lambda_{\min} \\ \text{rate}_{\eta^*} &= \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \end{aligned}$$

Exercise 3: Optimal soft-margin hyperplane

Consider a variation of the optimal soft-margin linear classifier defined by

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}[(1 - \alpha) \sum_{i:y_i=1}^{n} \xi_i + \alpha \sum_{i:y_i=-1} \xi_i] \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \forall i \\ & \xi_i \geq 0, \forall i \end{aligned}$$

where $\alpha \in (0,1)$ is a parameter that captures a desire to penalize either false positives or false negatives more than the other. Show that the resulting linear classifier can also be derived by regularized empirical risk minimization with a particular loss. Determine the loss (which will depend on α), and state the regularization parameter λ in terms of C.

Solution

The form for Empirical Risk Minimization is:

$$\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \Omega(f)$$

where $L(y_i, f(x_i))$ is the loss function, and the $\lambda\Omega(f)$ is the regularizar. Letting the loss function be chosen as hinge loss:

$$\begin{aligned} L(y_i, f(x_i)) &= \max(0, 1 - y_i f(x_i)) \\ &= \max(0, 1 - y_i(w^T x_i + b)) \end{aligned}$$

This gives us

$$\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b)) + \lambda \Omega(f)$$

We consider the two cases of y_i to remove that term in the equation and turn into weighted hinge loss. Doing this we get:

$$L(y_i, w^T x + b) = \begin{cases} \max(0, 1 - w^T x_i - b) & , y_i = 1 \\ \max(0, 1 + w^T x_i + b) & , y_i = -1 \end{cases}$$

Doing this, and letting our $\Omega(f) = 1/2\|w\|^2$, and C be arbitrary scalar, we get:

$$\min_{f \in F} \frac{C\lambda}{2n} \|w\|^2 + \frac{C}{n} \sum_{i=1} (1-\alpha) \max(0, 1 - w^T x_i - b) + \alpha \max(0, 1 + w^T x_i + b)$$

To cancel out λ, we let $\lambda = \frac{n}{C}$, thus giving:

$$\min_{f \in F} \frac{1}{2} \|w\|^2 + \frac{C}{n} ((1-\alpha) \sum_{i=1} \max(0, 1 - w^T x_i - b) + \alpha \sum_{i} \max(0, 1 + w^T x_i + b))$$

With some substitution we arrive at the optimal soft-margin linear classifier above. The loss is:

$$L(y_i, w^T x + b) = \begin{cases} (1-\alpha) \max(0, 1 - w^T x_i - b) & , y_i = 1 \\ \alpha \max(0, 1 + w^T x_i + b) & , y_i = -1 \end{cases}$$

and we find $\lambda = \frac{n}{C}$.

Exercise 5: Inner Product Kernels

Let $k(u,v) = (u^T v + 1)^2$ where $u, v \in \mathbb{R}^3$. Find Φ such that $k(u,v) = \langle \Phi(u), \Phi(v) \rangle$.

Solution

Expanding we get:

$$\begin{aligned} k(u,v) &= (u^T v + 1)^2 \\ &= (u^T v)^2 + 2u^T v + 1 \end{aligned}$$

Converting into a summation form we get

$$k(u,v) = 1 + 2\sum_{i=1}^{3} u_i v_i + \sum_{i=1}^{3} u_i^2 v_i^2 + 2\sum_{i<j} u_i u_j v_i v_j$$

Doing this we can rewrite this into inner product form of:

$$k(u,v) = \left\langle \begin{bmatrix} 1 \\ \sqrt{2}u_1 \\ \sqrt{2}u_2 \\ \sqrt{2}u_3 \\ u_1^2 \\ u_2^2 \\ u_3^2 \\ \sqrt{2}u_1u_2 \\ \sqrt{2}u_1u_3 \\ \sqrt{2}u_2u_3 \end{bmatrix}, \begin{bmatrix} 1 \\ \sqrt{2}v_1 \\ \sqrt{2}v_2 \\ \sqrt{2}v_3 \\ v_1^2 \\ v_2^2 \\ v_3^2 \\ \sqrt{2}v_1v_2 \\ \sqrt{2}v_1v_3 \\ \sqrt{2}v_2v_3 \end{bmatrix} \right\rangle$$

Thus we find $\Phi(\vec{x}) = [1 \;\; \sqrt{2}x_1 \;\; \sqrt{2}x_2 \;\; \sqrt{2}x_3 \;\; x_1^2 \;\; x_2^2 \;\; x_3^2 \;\; \sqrt{2}x_1x_2 \;\; \sqrt{2}x_1x_3 \;\; \sqrt{2}x_2x_3]^T$.