Geet Sethi
Course: Attacking LLMs Using Projected Gradient Descent

Assignment - 2 | Part - 1
Date: Jun 09, 2024

1 Theory

1.1 Lagrange's Duality

The *Lagrange's multiplier* along with *KKT* conditions can be used to find the *argmin*. It results in the following:

Definition 1.1: Lagrange's Duality

If $\Omega : g(x) \leq c$ is a convex domain and we want to find $z = \underset{x \in \Omega}{\operatorname{argmin}} f(x)$, then the following conditions must hold for optimality:

- If $x \in \Omega$, then $z = x$.
- If $x \notin \Omega$, then
  - $g(x) = c$
  - $\nabla f(x) = \lambda \nabla g(x)$

  Solving these two equations simultaneously will yield the result!

1.2 Overview of the Projection Step

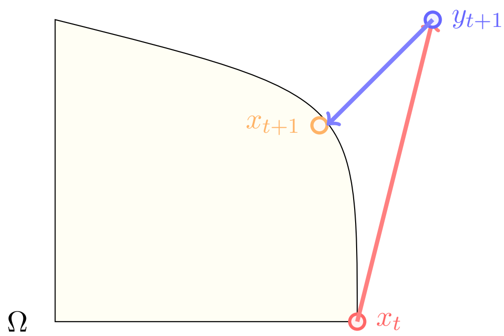

Figure 1: Projection Step

$$y_{t+1} = x_t - \gamma \nabla f(x_t)$$
$$x_{t+1} = \operatorname{proj}_\Omega (y_{t+1})$$

After carrying out the usual gradient descent, we apply the projection operator on the result(y_{t+1}) to make sure that the final result (x_{t+1}) is contained inside the domain (Ω)

1.3 The Projection Operator

Definition 1.2: Projection Operator

The projection operator projects the given vector (x) onto the domain (Ω) which is a convex set. It is mathematically defined as:

$$\operatorname{proj}_\Omega(x) = \underset{u \in \Omega}{\operatorname{argmin}} \frac{1}{2}\|u - x\|^2$$

2 Problems

Problem 1

Compute the projection step

$$P_{\mathbb{C}}(z) = \underset{x \in \mathbb{C}}{\operatorname{argmin}} \ \frac{1}{2\gamma} \|x - z\|^2$$

where $\mathbb{C} = \{x \in \ \mathbb{R} \mid a^T x \leq c\}$ for some fixed $a \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

Solution: Let $\Omega = \{x \in \mathbb{R} \mid a^T x = c\}$ be the domain onto which we want to project z. (This Ω is the hyperplane with $n-1$ dimensions where $a^T \in \mathbb{R}^n$)
Therefore, taking all possible orientation (i.e. leaving one dimension at a time), we can construct a matrix A with n columns and $n-1$ rows.
$\Longrightarrow$ Our domain: $\Omega : Ax = c$

We know that:

$$\operatorname{proj}_{\mathbb{C}}(z) = \underset{x \in \mathbb{C}}{\operatorname{argmin}} \frac{1}{2\gamma} \|x - z\|^2$$

Let $f(x, z) = \frac{1}{2}\|z - x\|^2$ and $g(x) = Ax - c$.
Therefore from Lagrange's duality, we have the following conditions:

$$\begin{aligned}
g(x) = 0 &\Rightarrow Ax - c = 0 && (1) \\
\nabla f(x) = \lambda \nabla g(x) &\Rightarrow \frac{1}{\gamma}\ (z - x) + \lambda \nabla g(x) = 0 && \left(\text{since } \nabla f(x) = \frac{1}{\gamma}\ (x - z)\right) \\
&\Rightarrow (z - x) + \lambda\gamma\ \nabla g(x) = 0 \\
&\Rightarrow (z - x) + \beta\ \nabla g(x) = 0 && (\text{let } \lambda\gamma = \beta) \\
&\Rightarrow (z - x) + \beta\ A^T = 0 && (\text{since } \nabla g(x) = A^T) \quad (2)
\end{aligned}$$

Solving further, we get:

$$\begin{aligned}
& z - x + \beta A^T = 0 \\
& \Rightarrow x = z + \beta A^T \\
& \Rightarrow Ax = Az + \beta A A^T && (\text{pre-multiplying by } A) \\
& \Rightarrow c = Az + \beta A A^T && (\text{since } g(x) = Ax - b = 0) \\
& \Rightarrow \beta A A^T = Az - c \\
& \Rightarrow \beta = (AA^T)^{-1}(Az - c) \\
& \Rightarrow x = z + (AA^T)^{-1}(Az - c)A^T
\end{aligned}$$

Therefore, the answer is: $\boxed{x = z + (AA^T)^{-1}(Az - c)A^T}$

Note 2.1 Similarity to Linear Algebra

If you look closely, this result is very similar to the result derived in *Linear Algebra* courses for finding the projection of a vector onto a hyperplane where we find the projection matrix!