Homework 1: Adversarial Attack

Byung Jae Bae
2023952231

November 22, 2023

1 Question 1

To use KKT conditions to show that $x\prime = x + r^*$ can be expressed by $x\prime = x + \epsilon \cdot \text{sgn}(\nabla_x L(x, y))$, we first define the optimization problem as:

$$r^* \in argmax_{r \in \mathbb{R}^p} L(x, y) + r^\intercal \nabla_x L(x, y) \text{ subject to } ||r||_\infty \leq \epsilon \tag{1}$$

Since we need to maximize the r but with a contraint of the infinity norm of r less than or equal to ϵ, those are our two conditions we need to account for.

We can simplify the optimization problem to:

$$r^* \in argmax_{r \in \mathbb{R}^p} r^\intercal \nabla_x L(x, y) \text{ subject to } ||r||_\infty \leq \epsilon \tag{2}$$

since the lone term $L(x, y)$ is irrelevant to the optimization problem of r.

We break down $||r||_\infty \leq \epsilon$ to $max_{i=0,1,2,...,p}|r_i| \leq \epsilon$. This means that $-\epsilon \leq r_i \leq \epsilon$. We can break this down into two equations $-\epsilon \leq r_i$ and $r_i \leq \epsilon$. To simplify this further we have $r_i + \epsilon \geq 0$ and $r_i - \epsilon \geq 0$, respectively, where $i = 0, 1, 2, ..., p$. These two equations make up our inequality set.

To prove the KKT conditions, we first set the Lagragian, $\mathcal{L}(x, \lambda)$. The Lagragian Function is given as $\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x)$. we can substitute our simplified objective function in place of the $f(x)$ in our Lagragian.

$$\mathcal{L}(r, \lambda) = r^\intercal \nabla_x L(x, y) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x) \tag{3}$$

Since we have two constraints, $c(x)$, in the form of $r_i + \epsilon \geq 0$ and $r_i - \epsilon \geq 0$ we write the two different λ and replace them $\alpha_i \geq 0$ and $\beta_i \geq 0$ (Dual Feasibility) with $r_i + \epsilon \geq 0$ and $r_i - \epsilon \geq 0$, respectively,

$$\mathcal{L}(r, \alpha, \beta) = r^\intercal \nabla_x L(x, y) - \sum_{i=1}^{n} \alpha_i (r_i + \epsilon) - \sum_{i=1}^{n} \beta_i (r_i - \epsilon) \tag{4}$$

To use KKT, We must satisfy 4 conditions:

Stationarity The gradient of the Lagragian is equal to zero.

Primal Feasibility All constraints are satisfied, either from the equality or inequality set.

Dual Feasibility All Lagrage multipliers are non-negative for inequality constraints.

Complimentary Slackness For each inequality constraint, the constraint is either active or inactive or the Lagrage multiplier is zero

First, we set the gradient of the Lagragian equal to zero. We find the gradient w.r.t r, our input, which results in this equation set to zero.

$$\nabla_r \mathcal{L}(r, \alpha, \beta) = \nabla_x L(x, y) + \alpha_i - \beta_i = 0 \tag{5}$$

For Primal Feasibility and Dual Feasibility, we will make sure the inequality constraints satisfied in the Complimentary Slackness part of the KKT conditions. We already stated that the Primal Feasibility is that $r_i + \epsilon \geq 0$ and $r_i - \epsilon \geq 0$, where $i = 0, 1, 2, ..., p$. The Dual Feasibility was also established by defining $\alpha_i \geq 0$ and $\beta_i \geq 0$. As long as we do not violate any of these conditions in the Complimentary Slackness, the proof is valid.

Complimentary Slackness demands that $\alpha_i(r_i + \epsilon) = 0$ and $\beta_i(r_i - \epsilon) = 0$ There are only possible 3 scenarios: $|r_i| < \epsilon$, $r_i = -\epsilon$, or $r_i = \epsilon$. The other case, $|r_i| > \epsilon$, violates the Primal Feasibility.

If $|r_i| < \epsilon$, then $\alpha_i = 0$ and $\beta_i = 0$ since $(r_i + \epsilon)$ and $(r_i - \epsilon)$ terms in $\alpha_i(r_i + \epsilon) = 0$ and $\beta_i(r_i - \epsilon) = 0$ are not equal to 0 This partially satisfies Complimentary Slackness, but Stationarity still needs to acheived. Since α_i and β_i are 0, $\nabla_r \mathcal{L}(r, \alpha, \beta) = \nabla_x L(x, y) + \alpha_i - \beta_i = 0$, this makes $\nabla_x L(x, y) = 0$.

If $r_i = -\epsilon$, then $\alpha_i \geq 0$, since $(r_i + \epsilon) = 0$ and $\beta_i = 0$, since $(r_i - \epsilon) \neq 0$. To complete Stationarity, $-\nabla_x L(x, y) = \alpha_i \geq 0$.

If $r_i = \epsilon$, then $\_i \geq 0$, since $(r_i - \epsilon) = 0$ and $\alpha = 0$, since $(r_i + \epsilon) \neq 0$. To complete Stationarity, $\nabla_x L(x, y) = \beta_i \geq 0$.

KKT Conditions give us that when $r_i = \epsilon$ then the constraint is active for α_i, when $r_i = -\epsilon$ then the constraint is active for β_i, and when $|r_i| < \epsilon$ then the constraints are not active and can be ignored, since $\lambda_i = 0$.

This means that when $\nabla_x L(x, y) > 0$, then $r_i = \epsilon$, and if $\nabla_x L(x, y) < 0$, then $r_i = -\epsilon$. And the last case $\nabla_x L(x, y) = 0$ can be ignored setting ϵ to 0 as no constraints are active.

This is representative of the sign function and $x\prime = x + r^*$ can be expressed by $x\prime = x + \epsilon \cdot \mathrm{sgn}(\nabla_x L(x, y))$