Comparison of Frank-Wolfe Varients for White-Box Adversarial Attacks

Tanner Aaron Graves - 2073559 Alessandro Pala - 2107800

June 2024

1 Introduction

What are Adversarial attacks
Problem statement
Type of Norms

2 Algorithims

2.1 Frank-Wolfe

2.2 Pairwise Frank-Wolfe

2.3 Away-Step Frank-Wolfe

3 Results

Introduce Datasets

3.1 Momentum

3.2 LMO vs. Linesearch

I really don't want to implement linesearch. As it is a waste of time when LMO is cheap but if we need easy content we can do this section.

3.3 ϵ Choice

Create plot showing how accurate attacks are with different ϵ constraints.

4 Convergence Analysis

The constrained nature of the Adversarial Attack problem means that the norm of the gradient $||\nabla_x f(x)||$ is not a sutible convergence criterion as boundry points need not have 0 gradient. The Frank-Wolfe gap provides provides measure of both optimality and point feasibility. It is a measure of the maximum improvement over the current iteration x_t within the constraints C and defined

$$g(x_t) = \max_{x \in C} \langle x - x_t, -\nabla f(x_t) \rangle$$

We always have $g(x_t) \geq 0$ and its usefullness as a convergence criterion comes from $g(x_t) = 0$ iff x_t is a stationary point. For convex problems, we would have that the linear approximation $f(x_t) + \langle x_t - x, -\nabla f(x_t) \rangle \geq f(x)$. However, the loss of DNNs as commnly the subject of adversarial attacks, are highly non-convex, making this only true locally. This complicate the convergence of Frank-Wolfe in this application, but it is still gaurenteed.

4.1 Frank-Wolfe

4.2 Pairwise Frank-Wolfe

4.3 Away-Step Frank-Wolfe