

Comparison of Frank-Wolfe Variants 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 Algorithms

2.1 Frank-Wolfe

2.2 Pairwise Frank-Wolfe

2.3 Away-Step Frank-Wolfe

3 Results

Introduce Datasets

3.1 Momentum

3.2 Early-Stopping (Convergence Criterion)

It is worth noting that Convergence Criterion For Frank-Wolfe is a somewhat imprecise surrogate for success in the context of adversarial attacks. For many examples, we find that Frank-Wolfe methods create successful attacks several iterations before convergence. We attribute this to an incorrect class probability being greater than the correct class being sufficient for success where convergence is reached when the new output class probability is maximized. We observe the convergence of the Frank-Wolfe gap

3.3 Stepsize

The methods for stepsize were implemented: fixed, where $\gamma_t = 1$ for all t , linesearching which solves $\arg \min_{\gamma} f(x + \gamma d_t)$ where $\gamma \in (0, 1]$, and decaying stepsize $\gamma_t = \frac{2}{t+2}$.

3.4 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 suitable convergence criterion as boundary points need not have 0 gradient. The Frank-Wolfe gap provides a 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 usefulness 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 commonly the subject of adversarial attacks, are highly non-convex, making this only true locally. This complicates the convergence of Frank-Wolfe in this application, but it is still guaranteed.

4.1 Frank-Wolfe

4.2 Pairwise Frank-Wolfe

4.3 Away-Step Frank-Wolfe