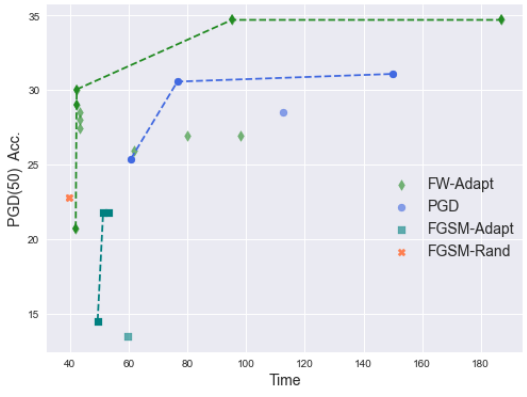
Appendices

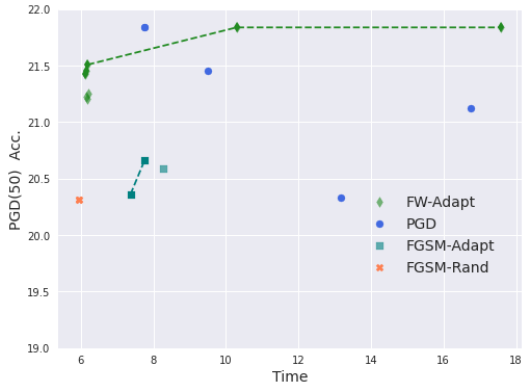
A. Results on Additional Datasets

We further validate our method on two higher resolution datasets. The first is a dataset of 7 types of skin lesions at a resolution of 224×224, ISIC-2018 [9], and the second is a subset of ImageNet where the task is to classify 10 breeds of dogs resized to 64×64, ImageWoof. Both datasets have $\sim$10K/1K train/val images. Due to computation constraints, we show only results against PGD(50) attack. Figure 6 shows the results of these experiments.

For the ISIC-2018 dataset we see an even stronger trend than with CIFAR-10/100. Here FW-AT-ADAPT almost uniformly outperforms competing methods with respect to optimal tradeoffs. A similar trend to CIFAR-10/100 holds for ImageWoof; however, the results are much less pronounced. In particular PGD struggled at higher steps. We suspect the lower performance is due to the lower resolution increasing the difficulty of differentiating dog breeds which share many semantically similar features, although we note FW-AT-ADAPT seems to maintain performance across parameters.



(a) ISIC-2018



(b) ImageWoof

Figure 6. Adversarial accuracy against $\epsilon = 8/255$ attacks with PGD(50). Dashed line spans optimal parameters. (PGD steps 2,3,5,7)

B. Effects of Minimum Distortion Ratio Bound on FW-Adapt

To better understand the impact of the minimum distortion ratio, r, on FW-ADAPT we run the same distortion ratios as in our main set of experiments over 5 independent runs, and analyzed various performance and training metrics. In Figure 7 plots both training time and adversarial accuracy as a function of the minimum distortion ratio bound, r. Adversarial accuracy is computed against a PGD(10) attack with step size $2.5\epsilon/10$. Both time and adversarial accuracy are reported as the mean of 5 independent training runs. We see for both $\epsilon = 8/255$ and $16/255$ the adversarial accuracy increases with training time. In both cases, there seems to be an optimal r in terms of training time vs robustness tradeoffs, around 0.9 and 0.88 for $\epsilon = 8/255$ and $16/255$ respectively.

In Figure 8 we show how the number of steps used by FW-ADAPT evolves during training for our different values of the minimum distortion ratio bound r. We do not consider the first batch as this is always a two-step attack to monitor the distortion. Steps are averaged across 5 independent runs.

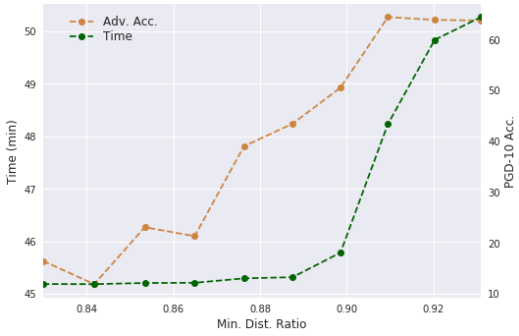
Higher values of r result in a linear increase towards the maximum number of steps, 15, and very low values of r result in primarily, although importantly not exclusively, single steps of attacks during training. As expected, the optimal value of r based on Figure 7 corresponds to training strategies which used a small number of steps initially and then modestly increase during training.

Figure 9 is in a sense dual to Figure 8 in that we plot the value of FW(2) distortion used in the adaptive step check. Again, we averaged the values over five independent runs.

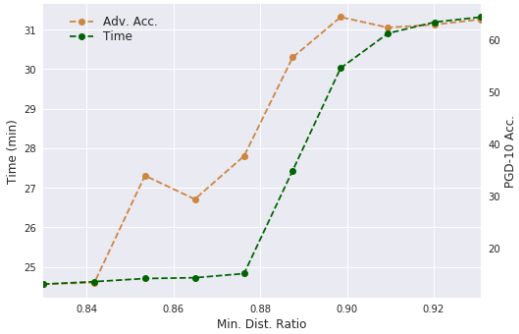
The high values of r which quickly increased their training steps have a smooth gradual decay in their distortion check; whereas, lower values had much more variation in their checks. The overall trend of decaying distortion is interesting and reinforces the fact that as AT progresses, stronger multi-step attacks are needed to more effectively increase the loss, but early in training such steps are not necessary. FW-ADAPT is able to capitalize on this to achieve faster training times. In future work, we hope to better understand the decaying trend of distortion, and perhaps develop more sophisticated adaptive criterion and step modifications to further improve performance.

C. FW-AT is As Good As PGD-AT

Although we focus on the novel FW-ADAPT algorithm here, we note that using FW optimization (Algorithm 1) in place of PGD with no other alterations performs as well as PGD in terms of robustness and training times. Figure 10 shows the training times and accuracy against PGD(5)0



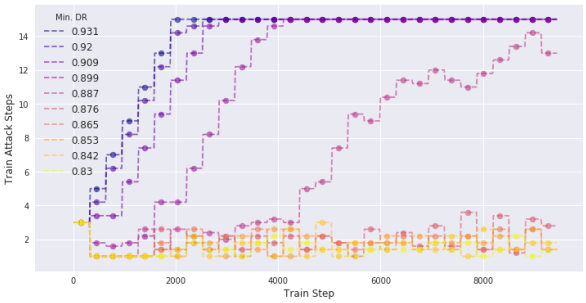
(a) $\epsilon = 8/255$



(b) $\epsilon = 16/255$

Figure 7. Average Training times and adversarial accuracy against PGD(10) as a function of minimum distortion ratio over five independent runs.

(a) $\epsilon = 8/255$



(b) $\epsilon = 16/255$

Figure 8. Number of attack steps during training for varying minimum distortion ratios. Plot ignores the first batch which is always done with 2 steps. Results averaged over 5 runs.

attacks for models trained with PGD-K-AT and FW-K-AT for $K \in \{1, 2, 3, 5, 7, 10\}$. The training parameters are the same as those above except accuracy and training time are averaged over three independent runs and we train for 40 epochs. We see that FW-AT performs comparably to PGD-AT. We hope this will encourage further study of FW for deep learning and AT.

D. Distortion and Gradient Alignment as Catastrophic Overfitting Signals

The basis of FW-ADAPT is that a small number of batches going through low-compute adversarial training (FW(2)) can provide a strong signal as to how many steps are needed for the rest of the epoch. As a particular example of this we showed empirically that the distortion of FW(2) attacks is a strong signal of catastrophic overfitting (CO), the phenomena where a model is trained with a single step attack and is achieving high accuracy against strong multi-step attacks, but then suddenly loses robustness against strong attacks while still being robust to the single step attack. In [2] authors note that gradient misalignment is also strongly associated with CO and they use it to regularize single step methods (FGSM-GA).

Here we compare FW(2) distortion and gradient alignment (GA) as signals for CO. In figure 11 shows the gradual (top) and overall transition of model into CO. Both the distributions of GA and FW(2) distortion are able to distinguish the CO model from the non-CO model. During the transition we see the GA score distribution becomes more diffuse during the transition; whereas, the FW(2) distortion has a more gradual peak shift during transition.

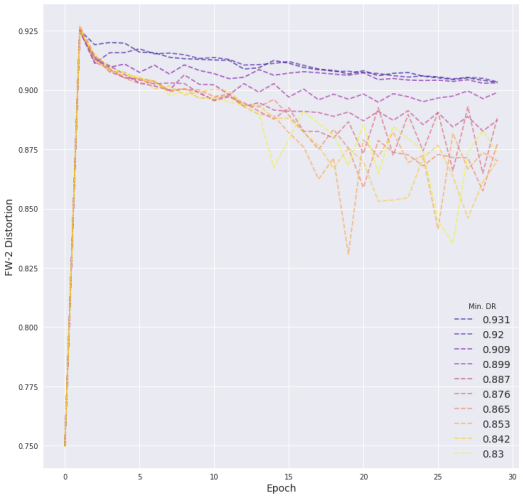
Interestingly, the GA signal may be slightly clearer than the FW(2) distortion for CO detection. Although, as we see above, using the GA as a regularizer for single step methods is not able to achieve the same level of robustness as multi-step methods. This suggest that there is more to the gap between single and multi-step methods than merely fixing CO. Building upon the theoretical foundation for exactly what is missed by single step methods is an interesting direction of further research.

E. Proofs

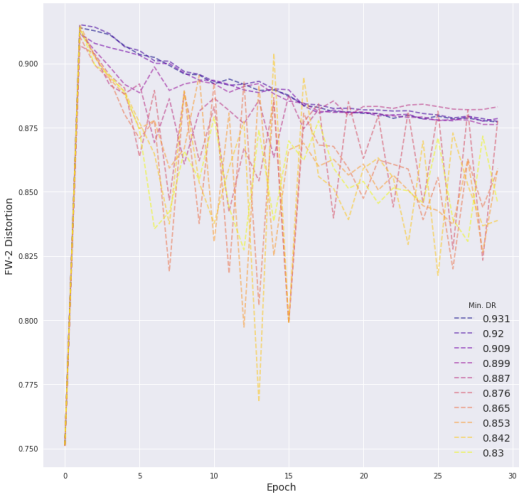
E.1. Proof of Proposition 1

Proof. The LMO solution is given by $\bar{\delta}_k = \epsilon\, \phi_p(\nabla_\delta \mathcal{L}(x + \delta_k, y))$ and the update becomes

$$\begin{aligned}\delta_{k+1} &= \delta_k + \gamma_k(\bar{\delta}_k - \delta_k) \\ &= (1 - \gamma_k)\delta_k + \gamma_k\, \epsilon\, \phi_p(\nabla_\delta \mathcal{L}(x + \delta_k, y))\end{aligned}$$



(a) $\epsilon = 8/255$



(b) $\epsilon = 16/255$

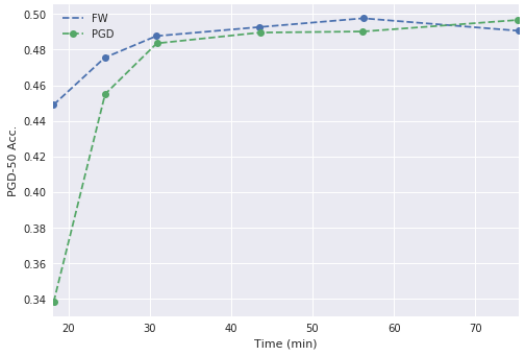
Figure 9. Average distortion check value during training for varying minimum distortion ratios.Results averaged over 5 runs.

Using induction on this relation yields after K steps:

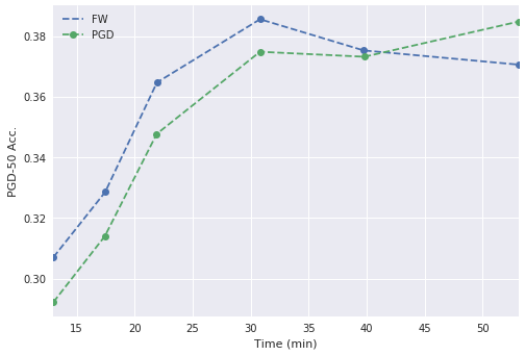
$$\delta_K = \delta_0 \prod_{l=0}^{K-1}(1-\gamma_l) + \epsilon \sum_{l=0}^{K-1} \gamma_l \prod_{i=l+1}^{K-1}(1-\gamma_i)\phi_p(\nabla_\delta \mathcal{L}(x+\delta_k, y)) \quad (11)$$

where δ_0 is the initial point which affects both terms in (11) and $\gamma_k = c/(c+k)$ for $k \geq 0$. Since $\gamma_0 = 1$, the first term vanishes and (11) simplifies to

$$\delta_K = \epsilon \sum_{l=0}^{K-1} \alpha_l \phi_p(\nabla_\delta \mathcal{L}(x+\delta_l, y)) \quad (12)$$



(a) $\epsilon = 8/255$



(b) $\epsilon = 16/255$

Figure 10. Accuracy against PGD(5)0 attacks on CIFAR-10 validation images for FW-AT and PGD-AT at various ϵ values.

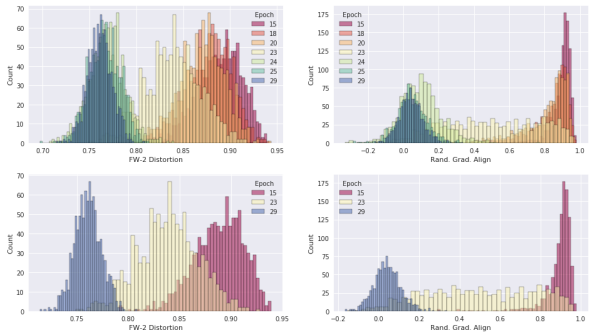


Figure 11. Distribution of FW(2) Distortion and Gradient direction at the point and and random direction (Grad. Align signal) for 1024 randomly sampled CIFAR-10 validation images during FGSM-AT training where catastrophic overfitting occurs as in Figure 4. (Top) The transition across 7 epochs into CO, (Bottom) focusing on epochs clearly before, during, and clearly after CO occurs.

where the coefficients are

$$\alpha_l = \gamma_l \prod_{i=l+1}^{K-1}(1-\gamma_i) \quad (13)$$

Since $\gamma_l \in [0,1]$, it follows that $\alpha_l \in [0,1]$. Induction on (13) yields that $\sum_{l=0}^{K-1} \alpha_l = 1$. Furthermore, $\alpha_l \leq \alpha_{l+1}$

follows from:

$$\begin{aligned} \alpha_l &\leq \alpha_{l+1} \\ \Leftrightarrow \gamma_l\,(1-\gamma_{l+1}) &\leq \gamma_{l+1} \\ \Leftrightarrow \frac{c}{c+l}\Big(1-\frac{c}{c+l+1}\Big) &\leq \frac{c}{c+l+1} \\ \Leftrightarrow \frac{l+1}{c+l} &\leq 1 \\ \Leftrightarrow 1 &\leq c \end{aligned}$$

Thus, the sequence α_l is non-decreasing in l. Since the coefficients sum to unity, (12) is in the convex hull of the generated LMO sequence $\{\phi_p(\nabla_\delta \mathcal{L}(x+\delta_l)) : l = 0, \ldots, K-1\}$. □

E.2. Proof of Theorem 1

Proof. From Proposition 1, we obtain the following decomposition of the adversarial perturbation:

$$\delta_K = \epsilon \sum_{l=0}^{K-1} \alpha_l \text{sgn}(\nabla_\delta \mathcal{L}(x+\delta_l, y))$$

To bound the magnitude of the adversarial perturbation, we have

$$\|\delta_K\|_2 = \sqrt{\|\delta_K\|_2^2} = \epsilon \sqrt{\Big\|\sum_l \alpha_l s_l\Big\|_2^2}$$

where we use the shorthand notation $s_l = \text{sgn}(\nabla_\delta \mathcal{L}(x+\delta_l, y))$. The squared ℓ_2 norm in the above is bounded as:

$$\begin{aligned} \Big\|\sum_l \alpha_l s_l\Big\|_2^2 &= \sum_l \sum_j \alpha_l \alpha_j \langle s_l, s_j \rangle \\ &= \sum_l (\alpha_l)^2 \|s_l\|_2^2 + \sum_{l\neq j} \alpha_l \alpha_j \|s_l\|_2 \|s_j\|_2 \cos\beta_{lj} \\ &= d\Big(\sum_l (\alpha_l)^2 + \sum_{l\neq j} \alpha_l \alpha_j \cos\beta_{lj}\Big) \\ &= d\Big(\sum_l (\alpha_l)^2 + \sum_{l\neq j} \alpha_l \alpha_j - \sum_{l\neq j} \alpha_l \alpha_j (1-\cos\beta_{lj})\Big) \\ &= d\Big(1 - \sum_{l\neq j} \alpha_l \alpha_j (1-\cos\beta_{lj})\Big) \\ &= d\Big(1 - 2\sum_{l<j} \alpha_l \alpha_j (1-\cos\beta_{lj})\Big) \end{aligned}$$

where we used $\|s_l\|_2 = \sqrt{d}$ and from Proposition 1 $(\sum_l \alpha_l)^2 = 1$. The final step follows from symmetry. This concludes the proof. □

E.3. Proof of Theorem 2

Proof. From Theorem 1 and the lower bound on the distortion, it follows that:

$$\sum_{l<j} \alpha_l \alpha_j (1-\cos\beta_{lj}) \leq \eta/2 \tag{14}$$

Letting $s_i = \text{sgn}(\nabla \mathcal{L}(x+\delta_i, y))$ and expanding the squared difference of signed gradients:

$$\begin{aligned} \|s_l - s_j\|_2^2 &= \|s_l\|_2^2 + \|s_j\|_2^2 - 2\langle s_j, s_l \rangle \\ &= \|s_l\|_2^2 + \|s_j\|_2^2 - 2\|s_j\|_2 \|s_l\|_2 \cos\beta_{lj} \\ &= d + d - 2d\cos\beta_{lj} \\ &= 2d(1-\cos\beta_{lj}) \end{aligned} \tag{15}$$

Using (15) into (14),

$$\sum_{l<j} \alpha_l \alpha_j \|s_l - s_j\|_2^2 \leq \eta d \tag{16}$$

For the FGSM deviation bound, i.e., $k_0 = 1$, we have by the triangle inequality:

$$\begin{aligned} \|\delta_K - \epsilon\,\text{sgn}(\nabla_x \mathcal{L}(x,y))\|_2 &= \|\epsilon \sum_{l=0}^{K-1} \alpha_l s_l - \epsilon s_0\|_2 \\ &= \|\epsilon \sum_l \alpha_l s_l - \sum_l \alpha_l \epsilon s_0\|_2 \\ &= \epsilon \|\sum_l \alpha_l (s_l - s_0)\|_2 \\ &\leq \epsilon \sum_{l>0} \alpha_l \|s_l - s_0\|_2 \end{aligned} \tag{17}$$

Using Cauchy-Schwarz inequality, we obtain:

$$\begin{aligned} \sum_{l>0} \alpha_l \|s_l - s_0\|_2 &\leq \sqrt{K-1} \sqrt{\sum_{l>0} (\alpha_l)^2 \|s_l - s_0\|_2^2} \\ &\leq \sqrt{K-1} \sqrt{\sum_{l<j} (\alpha_l)^2 \|s_l - s_j\|_2^2} \\ &\leq \sqrt{K-1} \sqrt{\sum_{l<j} \alpha_l \alpha_j \|s_l - s_j\|_2^2} \\ &\leq \sqrt{K-1} \cdot \sqrt{\eta d} \end{aligned} \tag{18}$$

where we used the non-decreasing property of the sequence $\{\alpha_l\}_l$ and the bound (16). This concludes the first part.

Given $1 \leq k_0 \leq K$, we have via using Proposition 1

twice:

$$\begin{aligned}
\delta_K - \delta_{k_0} &= \epsilon \sum_{l=0}^{K-1} \alpha_l s_l - \delta_{k_0} \\
&= \epsilon \sum_{l=0}^{K-1} \alpha_l s_l - \sum_{l} \alpha_l \delta_{k_0} \\
&= \epsilon \sum_{l=0}^{K-1} \alpha_l (s_l - \delta_{k_0}/\epsilon) \\
&= \epsilon \sum_{l=0}^{K-1} \alpha_l (s_l - \sum_{j=0}^{k_0-1} \tilde{\alpha}_j s_j) \\
&= \epsilon \sum_{l=0}^{K-1} \alpha_l \sum_{j=0}^{k_0-1} \tilde{\alpha}_j (s_l - s_j) \\
&= \epsilon \sum_{l=0}^{K-1} \sum_{j=0}^{k_0-1} \alpha_l \tilde{\alpha}_j (s_l - s_j) \qquad (19)
\end{aligned}$$

where $\alpha_l = \gamma_l \prod_{i=l+1}^{K-1}(1-\gamma_i), 0 \le l \le K-1$ and $\tilde{\alpha}_j = \gamma_j \prod_{i=l+1}^{k_0-1}(1-\gamma_i), 0 \le j \le k_0 - 1$.

Taking the ℓ_2 norm of both sides of (19) and using the triangle inequality, we obtain:

$$\| \delta_K - \delta_{k_0} \|_2 \le \epsilon \sum_{l=0}^{K-1} \sum_{j=0}^{k_0-1} \alpha_l \tilde{\alpha}_j \|s_l - s_j\|_2$$

Using the Cauchy-Schwarz inequality yields:

$$\begin{aligned}
&\sum_{l=0}^{K-1} \sum_{j=0}^{k_0-1} \alpha_l \tilde{\alpha}_j \|s_l - s_j\|_2 \\
&\le \sqrt{\sum_{l=0}^{K-1} \sum_{j=0}^{k_0-1} (\alpha_l \tilde{\alpha}_j)^2} \sqrt{\sum_{l=0}^{K-1} \sum_{j=0}^{k_0-1} \|s_l - s_j\|_2^2} \\
&\le \sqrt{\sum_{l=0}^{K-1} (\alpha_l)^2 \sum_{j=0}^{k_0-1} (\tilde{\alpha}_j)^2} \sqrt{\sum_{l=0}^{K-1} \sum_{j=0}^{K-1} \|s_l - s_j\|_2^2} \\
&\le \sqrt{\sum_{l=0}^{K-1} (\alpha_l)^2 \sum_{j=0}^{k_0-1} (\tilde{\alpha}_j)^2} \sqrt{\frac{2\eta d}{\min_{l<j}\{\alpha_l \alpha_j\}}} \\
&= \sqrt{\frac{2 \sum_{l=0}^{K-1} (\alpha_l)^2 \sum_{j=0}^{k_0-1} (\tilde{\alpha}_j)^2}{\alpha_0 \alpha_1}} \sqrt{\eta d}
\end{aligned}$$

where we used (16) and the nondecreasing sequence $\{\alpha_l\}$ implies $\min_{l<j}\{\alpha_l \alpha_j\} = \alpha_0 \alpha_1$. This concludes the proof of the second part. □

E.4. Proof of Theorem 3

Proof. Using the triangle inequality and the L-Lipschitz continuous loss gradient assumption:

$$\begin{aligned}
&\|g(\theta, \delta(K)) - g(\theta, \delta(k_0))\|_2 \\
&= \Big\| \frac{1}{|B|} \sum_{i \in B} (\nabla_\theta \mathcal{L}(f_\theta(x_i + \delta_i(K)), y_i) \\
&\qquad - \nabla_\theta \mathcal{L}(f_\theta(x_i + \delta_i(k_0)), y_i)) \Big\|_2 \\
&\le \frac{1}{|B|} \sum_{i \in B} \|\nabla_\theta \mathcal{L}(f_\theta(x_i + \delta_i(K)), y_i) \\
&\qquad - \nabla_\theta \mathcal{L}(f_\theta(x_i + \delta_i(k_0)), y_i))\|_2 \\
&\le \frac{L}{|B|} \sum_{i \in B} \|\delta_i(K) - \delta_i(k_0)\|_2 \qquad (20)
\end{aligned}$$

The average distortion condition yields via Proposition 1 (with the superscript (i) denoting the i-th example variables):

$$\frac{1}{|B|} \sum_{i \in B} \sqrt{1 - 2 \sum_{l<j} \alpha_l \alpha_j (1 - \cos \beta_{lj}^{(i)})} \ge \sqrt{1 - \eta}$$

Using Jensen's inequality (and the concavity of the square root function) further yields after some algebra:

$$\frac{1}{|B|} \sum_{i \in B} \sum_{l<j} \alpha_l \alpha_j (1 - \cos \beta_{lj}^{(i)}) \le \frac{\eta}{2}$$

Borrowing the relation (15) from the proof of Theorem 2, we further obtain:

$$\frac{1}{|B|} \sum_{i \in B} \sum_{l<j} \alpha_l \alpha_j \|s_l^{(i)} - s_j^{(i)}\|_2^2 \le \eta d \qquad (21)$$

Using the relation (19), it follows:

$$\begin{aligned}
&\frac{1}{|B|} \sum_{i \in B} \|\delta_i(K) - \delta_i(k_0)\|_2 \\
&\overset{(a)}{\le} \frac{1}{|B|} \sum_{i \in B} \epsilon \sum_{l=0}^{K-1} \sum_{j=0}^{k_0-1} \alpha_l \tilde{\alpha}_j \|s_l^{(i)} - s_j^{(i)}\|_2 \\
&\overset{(b)}{\le} \epsilon \sqrt{\sum_{l=0}^{K-1} \alpha_l^2 \sum_{j=0}^{k_0-1} \tilde{\alpha}_j^2} \cdot \frac{1}{|B|} \sum_{i \in B} \sqrt{\sum_{l=0}^{K-1} \sum_{j=0}^{k_0-1} \|s_l^{(i)} - s_j^{(i)}\|_2} \\
&\overset{(c)}{\le} \epsilon \sqrt{\sum_{l=0}^{K-1} \alpha_l^2 \sum_{j=0}^{k_0-1} \tilde{\alpha}_j^2} \cdot \sqrt{\frac{1}{|B|} \sum_{i \in B} \sum_{l=0}^{K-1} \sum_{j=0}^{k_0-1} \|s_l^{(i)} - s_j^{(i)}\|_2} \qquad (22)
\end{aligned}$$

where we used (a) triangle inequality, (b) Cauchy-Schwarz, and (c) Jensen's inequality.

From (21), it follows that:

$$\frac{1}{|B|}\sum_{i\in B}\sum_{l=0}^{K-1}\sum_{j=0}^{k_0-1}\|s_l^{(i)}-s_j^{(i)}\|_2^2 \leq \frac{2\eta d}{\alpha_0\alpha_1} \tag{23}$$

Combining (23) with (22) yields:

$$\frac{1}{|B|}\sum_{i\in B}\|\delta_i(\theta,K)-\delta_i(\theta,k_0)\|_2 \leq \epsilon\sqrt{d}\sqrt{\eta}C_{k_0,K}$$

where $C_{k_0,K}=\sqrt{\frac{2\sum_{l=0}^{K-1}\alpha_l^2\sum_{j=0}^{k_0-1}\tilde{\alpha}_j^2}{\alpha_0\alpha_1}}$. Using this bound in (20) concludes the proof. □

E.5. Convergence Analysis

Loss functions $\mathcal{L}(x+\delta,y)$ in deep neural networks are nonconvex in general. For a targeted attack that aims to fool the classifier to predict a specific label, without loss of generality, we seek to minimize the loss $f(\delta)=\mathcal{L}(x+\delta,y')$ over a ℓ_p constraint set. The untargeted case follows similarly. [1] For general nonconvex constrained optimization, the Frank-Wolfe gap given by [11]:

$$G(\delta_k)=\max_{\delta\in B_p(\epsilon)}\langle\delta-\delta_k,\nabla_\delta\mathcal{L}(x+\delta_k,y)\rangle \tag{24}$$

is nonnegative in general and zero at stationary points. The convergence of FW on non-convex functions has been studied in [17] and recently improved for strongly convex constraints in [26].

Assumption 2. *The function f has L-Lipschitz continuous gradients on $B_p(\epsilon)$, i.e., $\|\nabla f(u)-\nabla f(v)\|\leq L\|u-v\|,\forall u,v\in B_p(\epsilon)$.*

Assumption 2 is a standard assumption for the nonconvex setting and has been made in several works [8, 17]. A recent study [29] shows that the batch normalization layer used in modern neural networks makes the loss much smoother. Furthermore, the process of adversarial training smooths the loss landscape in comparison to standard models significantly [22, 25].

Given Assumption 2 and the compactness of the constraint sets, all limit points of FW are stationary points [4]. The convergence rate of FW to a stationary point for optimization over arbitrary convex sets was first shown in [17] given by

$$\min_{1\leq s\leq t}G(\delta_s)\leq\frac{\max\{2h_0,L\,\mathrm{diam}(B)\}}{\sqrt{t+1}}$$

where $h_0=f(\delta_0)-\min_{\delta\in B(\epsilon)}f(\delta)$ is the initial global suboptimality. It follows that larger ϵ imply a larger diameter and more iterations may be needed to converge [2]. This result implies that an approximate stationary point can be found with gap less than ϵ_0 in at most $O(1/\epsilon_0^2)$ iterations. Theorem 4 in [26] shows that for smooth non-convex functions over strongly convex constraint sets, FW yields an improved convergence rate $O\left(\frac{1}{t}\right)$, which importantly does not hold for the ℓ_∞ constraint.

[1] For untargeted attacks, $\min_{\delta\in B(\epsilon)}-\mathcal{L}(x+\delta,y)$ is considered and the FW gap becomes (24).

[2] The diameter of ℓ_2 ball is 2ϵ and for the ℓ_∞ ball $2\epsilon\sqrt{d}$.