Supplementary Material: Revisiting Outer Optimization in Adversarial Training

Anonymous ECCV submission

Paper ID 7022

1 Theoretical Analysis

1.1 Bound on the variance of Gradients

Let $\mathcal{P}$ be an arbitrary distribution of random vectors, and consider the vector transformation $T(\mathbf{a}) = \min(\frac{\alpha}{||\mathbf{a}||}, 1)\mathbf{a}$ with $\alpha > 0$.

Definition 1. *The variance of vectors with distribution $\mathcal{P}$ is defined as: [1, 2]:*

$$\sigma^2 := \mathbb{E}_{\mathbf{a}\sim\mathcal{P}}\big[||\mathbf{a} - \mathbb{E}_{\mathbf{b}\sim\mathcal{P}}[\mathbf{b}]||^2\big]. \tag{1}$$

Lemma 1. *Applying the vector transformation T bounds the norm of the mean vector to α,* i.e., *:*

$$||\mathbb{E}_{\mathbf{a}\sim\mathcal{P}}[T(\mathbf{a})]|| \leq \alpha. \tag{2}$$

Proof. Proof is straightforward by considering that $||T(\mathbf{a})|| \leq \alpha,\ \forall \mathbf{a}$.

Theorem 1. *Applying the vector transformation T bounds the variance of the vectors to $4\alpha^2$,* i.e., *:*

$$\mathbb{E}_{\mathbf{a}\sim\mathcal{P}}\big[||T(\mathbf{a}) - \mathbb{E}_{\mathbf{b}\sim\mathcal{P}}[T(\mathbf{b})]||^2\big] \leq 4\alpha^2. \tag{3}$$

Proof. Note that for any $\mathbf{a}$ we have:

$$||T(\mathbf{a}) - \mathbb{E}_{\mathbf{b}\sim\mathcal{P}}[T(\mathbf{b})]|| \leq ||T(\mathbf{a})|| + ||\mathbb{E}_{\mathbf{b}\sim\mathcal{P}}[T(\mathbf{b})]||. \tag{4}$$

Using $||T(\mathbf{a})|| \leq \alpha$ and Lemma 1 concludes the proof.

1.2 Convergence Analysis

To analyze the convergence of ENGM, we first define the total empirical risk as $A(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} L_i(\boldsymbol{\theta})$, where $L_i(\boldsymbol{\theta}) = L(F_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$, and n is the total number of examples in the dataset. Based on the previous works [1, 2], we make Assumptions 1 and 2 to analyze the convergence of A.

Assumption 1 *(bounded variance) For any $\boldsymbol{\theta}$ the variance of the gradients is bounded by σ^2 as:*

$$\mathbb{E}\big[||\nabla A(\boldsymbol{\theta}) - \nabla L_i(\boldsymbol{\theta})||^2\big] \leq \sigma^2. \tag{5}$$

Assumption 2 *(Smoothness) $A(\boldsymbol{\theta})$ is smooth with modulus $p > 0$ if for any θ_0, θ_1:*

$$A(\boldsymbol{\theta}_1) \leq A(\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}_0)^\top (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) + \tfrac{p}{2}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0||^2. \tag{6}$$

Lemma 2. *For $\mathbf{g} = \frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{v}_i$, where $w_i = \min\big(\frac{\alpha}{||\mathbf{v}_i||}, 1\big)$, we have $||\mathbf{g}|| \leq \alpha$.*

Proof. Proof is straightforward by considering that $||w_i \mathbf{v}_i|| \leq \alpha, \ \forall i$.

Lemma 3. *(Extension of Lemma 2 in [2]) For $\mathbf{v}_{t+1} = \beta \mathbf{v}_t + \mathbf{g}_t$ where $\beta < 1$ and $||\mathbf{g}_t|| \leq \alpha$, $\forall t \geq 0$ we have:*

$$||\mathbf{v}_t|| \leq \tfrac{\alpha}{1-\beta}. \tag{7}$$

Proof.

$$\begin{aligned} ||\mathbf{v}_{t+1}|| &\leq \beta||\mathbf{v}_t|| + \alpha \\ &\leq \beta^2||\mathbf{v}_{t-1}|| + \alpha(\beta+1) \\ &\leq \beta^{t+1}||\mathbf{v}_0|| + \alpha(\beta^t + \beta^{t-1} + \cdots + 1) \\ &\leq \tfrac{\alpha}{1-\beta}. \end{aligned} \tag{8}$$

Theorem 2. *Consider ENGM for optimizing $A(\boldsymbol{\theta})$ with the following update rule:*

$$\mathbf{v}_{t+1} = \beta \mathbf{v}_t + \mathbf{g}_t, \tag{9}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \mathbf{v}_{t+1}, \tag{10}$$

where $\mathbf{g}_t = \frac{1}{|\mathcal{I}_t|}\sum_{i \in \mathcal{I}_t} w_i \nabla_{\boldsymbol{\theta}} L_i(\boldsymbol{\theta}_t)$, and $w_i = \min\big(\frac{\alpha}{||\nabla_{\boldsymbol{\theta}} L_i(\boldsymbol{\theta}_t)||}, 1\big)$, for any $\alpha > 0$ the convergence is $O(\sigma)$ and is given as:

$$\begin{aligned} \tfrac{1}{t1}\sum_{i=0}^{t1-1} \mathbb{E}[||\nabla A(\boldsymbol{\theta}_t)||] \leq \tfrac{1-\beta}{\eta\alpha t_1}(A(\boldsymbol{\theta}_0) - A(\boldsymbol{\theta}^*)) \\ + (\tfrac{p\eta}{2(1-\beta)})\alpha. \end{aligned} \tag{11}$$

Proof. We drop $\mathbf{x}$ from $L(\mathbf{x}, \boldsymbol{\theta})$ and $\boldsymbol{\theta}$ from $\nabla_{\boldsymbol{\theta}}$ for brevity. Let $\boldsymbol{\varphi}_t = \boldsymbol{\theta}_t + \frac{\beta}{1-\beta}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})$, then we have:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \mathbf{g}_t + \beta(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}), \tag{12}$$

and

$$\boldsymbol{\varphi}_{t+1} = \boldsymbol{\varphi}_t - \tfrac{\eta}{1-\beta}\mathbf{g}_t. \tag{13}$$

Using Assumption 2 we have:

$$\begin{aligned} A(\boldsymbol{\varphi}_{t+1}) &\leq A(\boldsymbol{\varphi}_t) - \tfrac{\eta}{1-\beta}\nabla A(\boldsymbol{\varphi}_t)^\top \mathbf{g}_t + \tfrac{p\eta^2}{2(1-\beta)^2}||\mathbf{g}_t||^2 \\ &\leq A(\boldsymbol{\varphi}_t) - \tfrac{\eta}{1-\beta}||\mathbf{g}_t|| + \tfrac{p\eta^2}{2(1-\beta)^2}||\mathbf{g}_t||^2 \\ &\quad - \tfrac{\eta}{1-\beta}\Big[\big(\nabla A(\boldsymbol{\varphi}_t) - \nabla A(\boldsymbol{\theta}_t)\big)^\top \mathbf{g}_t \\ &\quad + \big(\nabla A(\boldsymbol{\theta}_t) - \mathbf{g}_t\big)^\top \mathbf{g}_t\Big]. \end{aligned} \tag{14}$$

Using Lemma 2, we have:

$$\begin{aligned} A(\boldsymbol{\varphi}_{t+1}) \leq A(\boldsymbol{\varphi}_t) - \tfrac{\eta\alpha}{1-\beta} + \tfrac{p\eta^2\alpha^2}{2(1-\beta)^2} \\ - \tfrac{\eta\alpha}{1-\beta}\Big[p||\boldsymbol{\varphi}_t - \boldsymbol{\theta}_t|| + ||\nabla A(\boldsymbol{\theta}_t) - \mathbf{g}_t||\Big]. \end{aligned} \tag{15}$$

Combining Lemma 3 with Equation 12, we obtain:

$$||\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t|| \leq \beta||\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}|| + \eta\alpha \leq \tfrac{\eta\alpha}{1-\beta}. \tag{16}$$

Consequently, we have:

$$||\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}|| \leq \tfrac{\eta\alpha}{1-\beta}, \tag{17}$$

and:

$$||\boldsymbol{\varphi}_t - \boldsymbol{\theta}_t|| = \tfrac{\beta}{1-\beta}||\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}|| \leq \tfrac{\eta\beta\alpha}{(1-\beta)^2}. \tag{18}$$

Inserting Equation 18 in Equation 15, we obtain:

$$\begin{aligned} ||\nabla A(\boldsymbol{\theta}_t) - \mathbf{g}_t|| \leq \tfrac{1-\beta}{\eta\alpha}(A(\boldsymbol{\varphi}_t) - A(\boldsymbol{\varphi}_{t+1})) \\ + \tfrac{p\eta\alpha}{2(1-\beta)} - \tfrac{p\eta\beta\alpha}{(1-\beta)^2}. \end{aligned} \tag{19}$$

Combining Equation 19 with the fact that $||\nabla A(\boldsymbol{\theta}_t)|| \leq ||\nabla A(\boldsymbol{\theta}_t) - \mathbf{g}_t|| + ||\mathbf{g}_t|| \leq ||\nabla A(\boldsymbol{\theta}_t) - \mathbf{g}_t|| + \alpha$, we obtain:

$$\begin{aligned} ||\nabla A(\boldsymbol{\theta}_t)|| \leq \tfrac{1-\beta}{\eta\alpha}(A(\boldsymbol{\varphi}_t) - A(\boldsymbol{\varphi}_{t+1})) \\ + \tfrac{p\eta\alpha}{2(1-\beta)} - \tfrac{p\eta\beta\alpha}{(1-\beta)^2} + \alpha. \end{aligned} \tag{20}$$

Finally, taking expectation from both sides, considering that $\boldsymbol{\theta}_0 = \boldsymbol{\varphi}_0$, and summing up the above inequality from $t = 0$ to $t1$ concludes the proof.

2 Evaluations on ℓ_2-norm Threat Model

Here, we evaluate the performance of different AT methods combined with ENGM on CIFAR-10/100 datasets. PGD with 10 steps (PGD^{10}), $\epsilon = 128/255$, and step size 15/255 is used as the attack to maximize the adversarial loss during the training. Table 1 presents the results for these evaluations. We observe that similar to the ℓ_∞-norm threat model, ENGM provides better robustness than MSGD in ℓ_2-norm threat model. Furthermore, it reduces the robust overfitting across all evaluations.

	AT Method	Optim. Method	Accuracy (%) Natural	PGD^{20}	AA	Overfit. (%)
CIFAR-10	Vanilla	MSGD	**89.64**	67.12	65.20	10.6
		ENGM	89.15	69.50	66.58	4.3
	TRADES	MSGD	87.93	68.33	67.01	6.2
		ENGM	87.37	69.95	68.11	3.5
	MART	MSGD	88.10	68.42	67.32	6.1
		ENGM	87.78	70.14	68.47	5.3
	AWP	MSGD	88.08	70.30	68.91	3.9
		ENGM	88.52	**71.16**	**69.96**	**3.3**
CIFAR-100	Vanilla	MSGD	**63.46**	41.31	38.54	14.6
		ENGM	62.63	43.53	39.92	6.8
	TRADES	MSGD	61.25	43.40	39.33	9.29
		ENGM	61.11	44.64	40.18	6.7
	MART	MSGD	61.90	43.75	39.20	10.8
		ENGM	61.43	44.19	40.68	8.8
	AWP	MSGD	61.83	45.26	40.28	8.0
		ENGM	62.00	**45.31**	**41.24**	**6.4**

Table 1: Comparison of methods for outer optimization on different AT approaches. Note that ENGM consistently outperforms MSGD.

References

1. Yu, H., Jin, R., Yang, S.: On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In: International Conference on Machine Learning. pp. 7184–7193. PMLR (2019)
2. Zhao, S.Y., Xie, Y.P., Li, W.J.: Stochastic normalized gradient descent with momentum for large batch training. arXiv preprint arXiv:2007.13985 (2020)