Random Noise Defense Against Query-based Black-Box Attacks

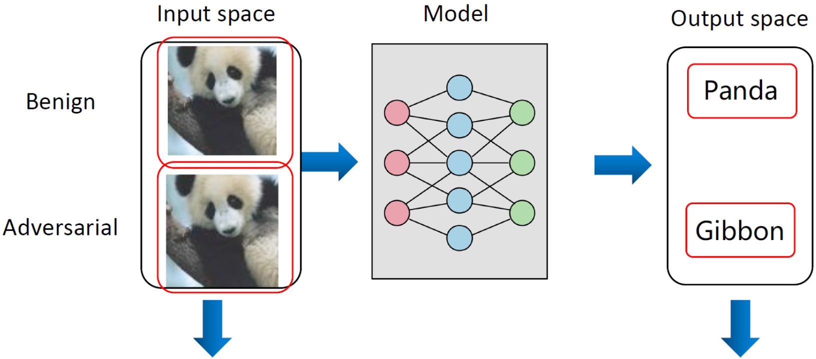
Zeyu Qin[1], **Yanbo Fan**[2], **Hongyuan Zha**[1], **Baoyuan Wu**[1]

[1]School of Data Science, Shenzhen Research Institute of Big Data,

The Chinese University of Hong Kong, Shenzhen

[2]Tencent AI Lab

Adversarial Examples

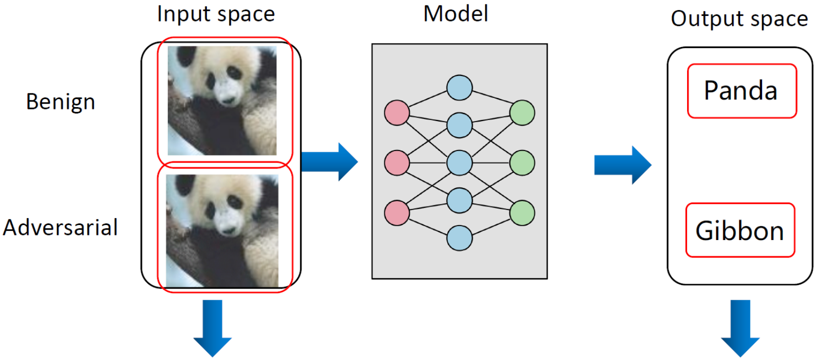


Small difference in input space

Large difference in output space

- Imperceptible: $\|x_{adv} - x\|_p \leq \epsilon$
- Misclassified: $y \neq \arg\max \mathcal{F}(x_{adv})$

Adversarial Examples



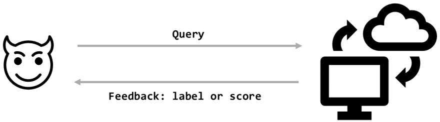
Small difference in input space

Large difference in output space

- Imperceptible: $\|\boldsymbol{x}_{adv} - \boldsymbol{x}\|_p \leq \epsilon$
- Misclassified: $\boldsymbol{y} \neq \arg\max \mathcal{F}(\boldsymbol{x}_{adv})$

Query-based Black-Box attacks

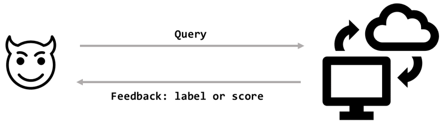
- However, in real scenarios such as autonomous driving, face recognition and verification, the DNN model as well as the training dataset, are often hidden from users.
- Only the model feedback for each query (labels or confidence scores) are accessible.
- By iteratively querying the targeted model, the attackers generate adversarial examples x_{adv} based on exact feedback of each query.



reference face vs target face, similarity 89%

Query-based Black-Box attacks

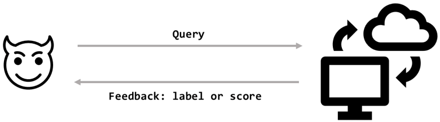
- However, in real scenarios such as autonomous driving, face recognition and verification, the DNN model as well as the training dataset, are often hidden from users.
- Only the model feedback for each query (labels or confidence scores) are accessible.
- By iteratively querying the targeted model, the attackers generate adversarial examples x_{adv} based on exact feedback of each query.



reference face vs target face, similarity 89%

Query-based Black-Box attacks

- However, in real scenarios such as autonomous driving, face recognition and verification, the DNN model as well as the training dataset, are often hidden from users.
- Only the model feedback for each query (labels or confidence scores) are accessible.
- By iteratively querying the targeted model, the attackers generate adversarial examples x_{adv} based on exact feedback of each query.



reference face vs target face, similarity 89%

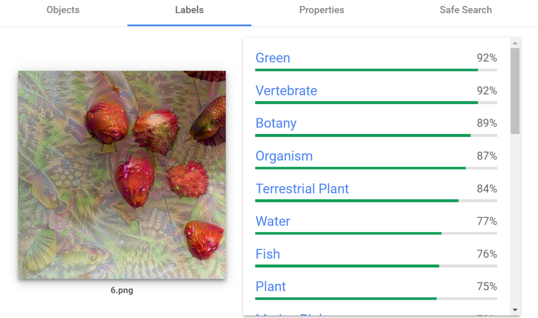
Score-based attacks

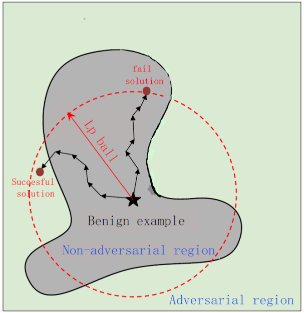
- Score-based : confidence score returned
  - untargeted attack:

$$\min_{\boldsymbol{x}_{adv}} f(\boldsymbol{x}_{adv}) = \max(0, \mathcal{F}(\boldsymbol{x}_{adv}, \boldsymbol{y}) - \max_{j \neq y} \mathcal{F}(\boldsymbol{x}_{adv}, \boldsymbol{j})), \quad s.t. \ \|\boldsymbol{x}_{adv} - \boldsymbol{x}\|_p \leq \epsilon \tag{1}$$

  - targeted attack:

$$\min_{\boldsymbol{x}_{adv}} f(\boldsymbol{x}_{adv}) = \max(0, \max_{j \neq t} \mathcal{F}(\boldsymbol{x}_{adv}, \boldsymbol{j}) - \mathcal{F}(\boldsymbol{x}_{adv}, \boldsymbol{t})), \quad s.t. \ \|\boldsymbol{x}_{adv} - \boldsymbol{x}\|_p \leq \epsilon \tag{2}$$





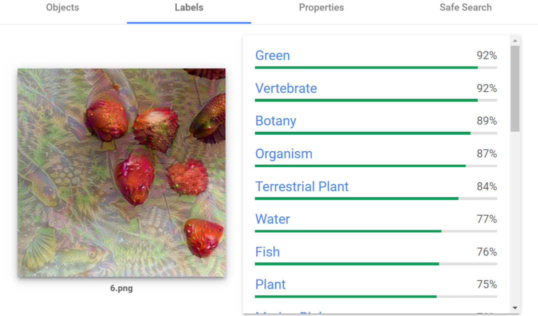
Score-based attacks

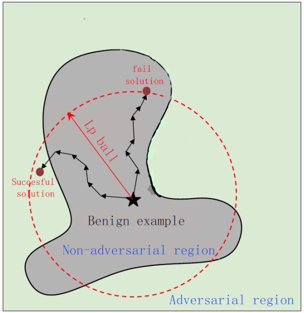
- Score-based : confidence score returned
  - untargeted attack:

$$\min_{\boldsymbol{x}_{adv}} f(\boldsymbol{x}_{adv}) = \max(0, \mathcal{F}(\boldsymbol{x}_{adv}, \boldsymbol{y}) - \max_{j \neq y} \mathcal{F}(\boldsymbol{x}_{adv}, \boldsymbol{j})), \quad s.t. \ \|\boldsymbol{x}_{adv} - \boldsymbol{x}\|_p \leq \epsilon \tag{1}$$

  - targeted attack:

$$\min_{\boldsymbol{x}_{adv}} f(\boldsymbol{x}_{adv}) = \max(0, \max_{j \neq t} \mathcal{F}(\boldsymbol{x}_{adv}, \boldsymbol{j}) - \mathcal{F}(\boldsymbol{x}_{adv}, \boldsymbol{t})), \quad s.t. \ \|\boldsymbol{x}_{adv} - \boldsymbol{x}\|_p \leq \epsilon \tag{2}$$





How to find the adversarial directions

- Zero Order (ZO) Attacks:
  - Randomized Gradient-Free (RGF) method (ZO Optimization) [1,2]:

$$g_{\mu}(\boldsymbol{x}) = \frac{f(\boldsymbol{x} + \mu\boldsymbol{u}) - f(\boldsymbol{x})}{\mu}\boldsymbol{u}, \tag{3}$$

    where f represents $f(\boldsymbol{x}_{adv})$.
  - Conducting projection gradient descent:

$$\boldsymbol{x}_{t+1} = \mathrm{Proj}_{\mathcal{N}_R(\boldsymbol{x})}\left(\boldsymbol{x}_t - \eta_t g_{\mu}\left(\boldsymbol{x}_t\right)\right). \tag{4}$$

- Search-based Attacks:
  - Random Search:

$$s_{\mu}(\boldsymbol{x}) = \mathbb{I}\{h_{\mu}(\boldsymbol{x}) < 0\} \cdot \mu\boldsymbol{u} \quad \text{where } h_{\mu}(\boldsymbol{x}) = f(\boldsymbol{x} + \mu\boldsymbol{u}) - f(\boldsymbol{x}), \tag{5}$$

  - Conducting projection gradient descent:

$$\boldsymbol{x}_{t+1} = \mathrm{Proj}_{\mathcal{N}_R(\boldsymbol{x})}\left(\boldsymbol{x}_t + \eta_t s\left(\boldsymbol{x}_t\right)\right). \tag{6}$$

[1] Yurii Nesterov et al., Random gradient-free minimization of convex functions, 2017
[2] John Duchi et al., Optimal rates for zero-order convex optimization: The power of two function evaluations, 2015

How to find the adversarial directions

- Zero Order (ZO) Attacks:
  - Randomized Gradient-Free (RGF) method (ZO Optimization) [1,2]:

$$g_\mu(\boldsymbol{x}) = \frac{f(\boldsymbol{x} + \mu \boldsymbol{u}) - f(\boldsymbol{x})}{\mu} \boldsymbol{u}, \tag{3}$$

    where f represents $f(\boldsymbol{x}_{adv})$.
  - Conducting projection gradient descent:

$$\boldsymbol{x}_{t+1} = \mathrm{Proj}_{\mathcal{N}_R(\boldsymbol{x})} \left(\boldsymbol{x}_t - \eta_t g_\mu \left(\boldsymbol{x}_t \right) \right). \tag{4}$$

- Search-based Attacks:
  - Random Search:

$$s_\mu(\boldsymbol{x}) = \mathbb{I}\{h_\mu(\boldsymbol{x}) < 0\} \cdot \mu \boldsymbol{u} \quad \text{where } h_\mu(\boldsymbol{x}) = f(\boldsymbol{x} + \mu \boldsymbol{u}) - f(\boldsymbol{x}), \tag{5}$$

  - Conducting projection gradient descent:

$$\boldsymbol{x}_{t+1} = \mathrm{Proj}_{\mathcal{N}_R(\boldsymbol{x})} \left(\boldsymbol{x}_t + \eta_t s \left(\boldsymbol{x}_t \right) \right). \tag{6}$$

[1] Yurii Nesterov et al., Random gradient-free minimization of convex functions, 2017
[2] John Duchi et al., Optimal rates for zero-order convex optimization: The power of two function evaluations, 2015

Black-Box Defense

- Main challenges in real scenarios,
  - the defender should not significantly influence the model's feedback to normal queries, but it is difficult to know whether a query is normal or malicious;
  - the defender has no information about what kinds of black-box attack strategies adopted by the attacker.
- We define defense task to address the above two challenges as **Black-Box Defense**. For product providers, the Black-Box defense should satisfy the below requirements:
  - well keeping clean accuracy
  - being robust against all kinds of black-box attacks
- However, the SOTA white-box defense, Adversarial Training (AT), is not suitable choice:
  - significant degradation of the clean accuracy
  - poor generalization for new data and adversarial attacks

Black-Box Defense

- Main challenges in real scenarios,
  - the defender should not significantly influence the model's feedback to normal queries, but it is difficult to know whether a query is normal or malicious;
  - the defender has no information about what kinds of black-box attack strategies adopted by the attacker.
- We define defense task to address the above two challenges as **Black-Box Defense**. For product providers, the Black-Box defense should satisfy the below requirements:
  - well keeping clean accuracy
  - being robust against all kinds of black-box attacks
- However, the SOTA white-box defense, Adversarial Training (AT), is not suitable choice:
  - significant degradation of the clean accuracy
  - poor generalization for new data and adversarial attacks

Black-Box Defense

- Main challenges in real scenarios,
  - the defender should not significantly influence the model's feedback to normal queries, but it is difficult to know whether a query is normal or malicious;
  - the defender has no information about what kinds of black-box attack strategies adopted by the attacker.
- We define defense task to address the above two challenges as **Black-Box Defense**. For product providers, the Black-Box defense should satisfy the below requirements:
  - well keeping clean accuracy
  - being robust against all kinds of black-box attacks
- However, the SOTA white-box defense, Adversarial Training (AT), is not suitable choice:
  - significant degradation of the clean accuracy
  - poor generalization for new data and adversarial attacks

Random Noise Defense

- The core of query-based attack: find an attack direction by **gradient estimation or random search based on the exact feedback** of consecutive queries.

$$g_{\mu}(\boldsymbol{x}) = \frac{f(\boldsymbol{x}+\mu\boldsymbol{u}) - f(\boldsymbol{x})}{\mu}\boldsymbol{u},$$

$$s_{\mu}(\boldsymbol{x}) = \mathbb{I}\{h_{\mu}(\boldsymbol{x}) < 0\} \cdot \mu\boldsymbol{u} \quad \text{where } h_{\mu}(\boldsymbol{x}) = f(\boldsymbol{x}+\mu\boldsymbol{u}) - f(\boldsymbol{x}).$$

- *Random Noise Defense* (RND) is realized by **adding a random noise to each query at the inference time**. There the gradient estimator and searching direction become

$$g_{\mu,\nu}(\boldsymbol{x}) = \frac{f(\boldsymbol{x}+\mu\boldsymbol{u}+\nu\boldsymbol{v}_1) - f(\boldsymbol{x}+\nu\boldsymbol{v}_2)}{\mu}\boldsymbol{u} \tag{7}$$

$$s_{\nu}(\boldsymbol{x}) = \mathbb{I}(h_{\nu}(\boldsymbol{x}) < 0) \cdot \mu\boldsymbol{u} \quad \text{where } h_{\nu}(\boldsymbol{x}) = f(\boldsymbol{x}+\mu\boldsymbol{u}+\nu\boldsymbol{v}_1) - f(\boldsymbol{x}+\nu\boldsymbol{v}_2) \tag{8}$$

Random Noise Defense

- The core of query-based attack: find an attack direction by **gradient estimation or random search based on the exact feedback** of consecutive queries.

$$g_\mu(\boldsymbol{x}) = \frac{f(\boldsymbol{x}+\mu\boldsymbol{u}) - f(\boldsymbol{x})}{\mu}\boldsymbol{u},$$

$$s_\mu(\boldsymbol{x}) = \mathbb{I}\{h_\mu(\boldsymbol{x}) < 0\} \cdot \mu\boldsymbol{u} \quad \text{where } h_\mu(\boldsymbol{x}) = f(\boldsymbol{x}+\mu\boldsymbol{u}) - f(\boldsymbol{x}).$$

- *Random Noise Defense* (RND) is realized by **adding a random noise to each query at the inference time**. There the gradient estimator and searching direction become

$$g_{\mu,\nu}(\boldsymbol{x}) = \frac{f\left(\boldsymbol{x}+\mu\boldsymbol{u}+\nu\boldsymbol{v}_1\right) - f\left(\boldsymbol{x}+\nu\boldsymbol{v}_2\right)}{\mu}\boldsymbol{u} \tag{7}$$

$$s_\nu(\boldsymbol{x}) = \mathbb{I}\left(h_\nu(\boldsymbol{x}) < 0\right) \cdot \mu\boldsymbol{u} \ \text{ where } h_\nu(\boldsymbol{x}) = f\left(\boldsymbol{x}+\mu\boldsymbol{u}+\nu\boldsymbol{v}_1\right) - f\left(\boldsymbol{x}+\nu\boldsymbol{v}_2\right) \tag{8}$$

Random Noise Defense

- For RND, the feedback for one query is $\mathcal{F}(\boldsymbol{x} + \nu \boldsymbol{v})$, with $\boldsymbol{v} \sim \mathcal{N}(0, \mathrm{I})$. And, ν controls magnitude of random noise.
- RND should satisfy two conditions
  - prediction of each query will not be changed significantly.
  - the estimated gradient or direction searching should be perturbed as large as possible.
- In the following, we provide the theoretical analysis of RND, which can shed light on the setting of ν.

Random Noise Defense

- For RND, the feedback for one query is $\mathcal{F}(\boldsymbol{x} + \nu\boldsymbol{v})$, with $\boldsymbol{v} \sim \mathcal{N}(0, \mathrm{I})$. And, ν controls magnitude of random noise.
- RND should satisfy two conditions
  - prediction of each query will not be changed significantly.
  - the estimated gradient or direction searching should be perturbed as large as possible.
- In the following, we provide the theoretical analysis of RND, which can shed light on the setting of ν.

Random Noise Defense

- For RND, the feedback for one query is $\mathcal{F}(\boldsymbol{x} + \nu \boldsymbol{v})$, with $\boldsymbol{v} \sim \mathcal{N}(0, \mathrm{I})$. And, ν controls magnitude of random noise.
- RND should satisfy two conditions
  - prediction of each query will not be changed significantly.
  - the estimated gradient or direction searching should be perturbed as large as possible.
- In the following, we provide the theoretical analysis of RND, which can shed light on the setting of ν.

Theoretical Analysis of RND Against ZO Attacks

To facilitate subsequent analyses, we first introduce some assumptions, definitions, and notations.

Assumption 1.

$f(\boldsymbol{x})$ is Lipschitz-continuous, i.e., $|f(\boldsymbol{y}) - f(\boldsymbol{x})| \leq L_0(f)\|\boldsymbol{y} - \boldsymbol{x}\|$.

Assumption 2.

$f(\boldsymbol{x})$ is continuous and differentiable, and $\nabla f(\boldsymbol{x})$ is Lipschitz-continuous, i.e., $\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \leq L_1(f)\|\boldsymbol{y} - \boldsymbol{x}\|$.

Definition 1.

The Gaussian-Smoothing function corresponding to $f(\boldsymbol{x})$ with $\nu > 0, \boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is

$$f_\nu(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \int f(\boldsymbol{x} + \nu\boldsymbol{v}) \cdot \mathrm{e}^{-\frac{1}{2}\|\boldsymbol{v}\|_2^2} \, \mathrm{d}\boldsymbol{v}. \tag{9}$$

Theoretical Analysis of RND Against ZO Attacks

Notations.

- The perturbation measure is specified as ℓ_2 norm, $\mathcal{N}_R(\boldsymbol{x}) = \{\boldsymbol{x}' \| \|\boldsymbol{x}' - \boldsymbol{x}\|_2 \leq R\}$.
- $d = |\boldsymbol{X}|$ denotes the input dimension.
- $\boldsymbol{\mathcal{U}}_t = \{\boldsymbol{u}_0, \boldsymbol{u}_1, ..., \boldsymbol{u}_t\}$, $\boldsymbol{\mathcal{V}}_t = \{\boldsymbol{v}_{01}, \boldsymbol{v}_{02}, ..., \boldsymbol{v}_{t1}, \boldsymbol{v}_{t2}\}$, represent the noise added by attacker or defenders. t is the iteration index.
- The benign example $\boldsymbol{x}$ is used as the initial solution, *i.e.*, $\boldsymbol{x}_0 = \boldsymbol{x}$.
- The generated sequential solutions are denoted as $\{\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_Q\}$.
- We define $S_Q = \sum_{t=0}^{Q} \eta_t$.

We study the convergence property of ZO attacks in Eq.(11) with $g_{\mu,\nu}(x)$ in Eq.(10) being the gradient estimator.

$$g_{\mu,\nu}(\boldsymbol{x}) = \frac{f(\boldsymbol{x} + \mu\boldsymbol{u} + \nu\boldsymbol{v}_1) - f(\boldsymbol{x} + \nu\boldsymbol{v}_2)}{\mu}\boldsymbol{u} \tag{10}$$

$$\boldsymbol{x}_{t+1} = \operatorname{Proj}_{\mathcal{N}_R(\boldsymbol{x}_0)}\left(\boldsymbol{x}_t - \eta_t g_{\mu,\nu}(\boldsymbol{x}_t)\right). \tag{11}$$

Theoretical Analysis of RND Against ZO Attacks

Notations.

- The perturbation measure is specified as ℓ_2 norm, $\mathcal{N}_R(\boldsymbol{x}) = \{\boldsymbol{x}' | \|\boldsymbol{x}' - \boldsymbol{x}\|_2 \le R\}$.
- $d = |\boldsymbol{X}|$ denotes the input dimension.
- $\boldsymbol{\mathcal{U}}_t = \{\boldsymbol{u}_0, \boldsymbol{u}_1, ..., \boldsymbol{u}_t\}$, $\boldsymbol{\mathcal{V}}_t = \{\boldsymbol{v}_{01}, \boldsymbol{v}_{02}, ..., \boldsymbol{v}_{t1}, \boldsymbol{v}_{t2}\}$, represent the noise added by attacker or defenders. t is the iteration index.
- The benign example $\boldsymbol{x}$ is used as the initial solution, *i.e.*, $\boldsymbol{x}_0 = \boldsymbol{x}$.
- The generated sequential solutions are denoted as $\{\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_Q\}$.
- We define $S_Q = \sum_{t=0}^{Q} \eta_t$.

We study the convergence property of ZO attacks in Eq.(11) with $g_{\mu,\nu}(\boldsymbol{x})$ in Eq.(10) being the gradient estimator.

$$g_{\mu,\nu}(\boldsymbol{x}) = \frac{f\left(\boldsymbol{x} + \mu \boldsymbol{u} + \nu \boldsymbol{v}_1\right) - f\left(\boldsymbol{x} + \nu \boldsymbol{v}_2\right)}{\mu} \boldsymbol{u} \tag{10}$$

$$\boldsymbol{x}_{t+1} = \mathrm{Proj}_{\mathcal{N}_R(\boldsymbol{x}_0)}\left(\boldsymbol{x}_t - \eta_t g_{\mu,\nu}\left(\boldsymbol{x}_t\right)\right). \tag{11}$$

Theoretical Analysis of RND Against ZO Attacks

Theorem 1.

Under Assumption 1, for any $Q \geq 0$*, consider a sequence* $\{\boldsymbol{x}_t\}_{t=0}^{Q}$ *generated according to the descent update Eq.(11) using the gradient estimator* $g_{\mu,\nu}(\boldsymbol{x})$ *Then, we have*

$$\frac{1}{S_Q}\sum_{t=0}^{Q}\eta_t\mathbb{E}_{\mathcal{U}_t,\mathcal{V}_t}(\|\nabla f_{\mu,\nu}(\boldsymbol{x}_t)\|^2) \leq \frac{f_{\mu,\nu}(\boldsymbol{x}_0)-f_\nu^*}{S_Q}+\frac{1}{S_Q}\sum_{t=0}^{Q}\eta_t^2 L_0(f)^3 d^{\frac{5}{2}}(\frac{1}{2\mu}+\frac{\sqrt{2}\nu}{\mu^2}+\frac{\nu^2}{\mu^3}).$$

We have $|f_{\mu,\nu}(\boldsymbol{x})-f_\nu(\boldsymbol{x})| \leq \mu L_0(f)d^{1/2}$*. To ensure* $|f_{\mu,\nu}(\boldsymbol{x}_t)-f_\nu(\boldsymbol{x}_t)| \leq \epsilon$*, We choose* $\mu \leq \frac{\epsilon}{d^{1/2}L_0(f)}$ *and set* $\alpha = \frac{\nu}{\mu}$*. With constant stepsize,* $\eta = \left[\frac{R\epsilon}{(\alpha+\frac{\sqrt{2}}{2})^2 d^3 L_0^3(f)(Q+1)}\right]^{1/2}$*, we have*

$$\frac{1}{Q+1}\sum_{t=0}^{Q}\mathbb{E}_{\mathcal{U}_t,\mathcal{V}_t}(\|\nabla f_{\mu,\nu}(\boldsymbol{x}_t)\|^2) \leq \frac{2L_0(f)^{\frac{5}{2}}R^{\frac{1}{2}}d^{\frac{3}{2}}}{(Q+1)^{\frac{1}{2}}\epsilon^{\frac{1}{2}}}(\alpha+\frac{\sqrt{2}}{2}). \tag{12}$$

In order to ensure that expected squared norm of $\nabla f_{\mu,\nu}$ *can reach* δ*,* **the query complexity is** $O\left((\alpha+\frac{\sqrt{2}}{2})^2\frac{d^3L_0^5(f)R}{\epsilon\delta^2}\right)$.

Theoretical Analysis of RND Against ZO Attacks

Remark 1.

- *Due to the non-convexity assumption, we only guarantee the convergence to a stationary point of the function $f_{\mu,\nu}(\boldsymbol{x})$, which is a smoothing approximation of f_ν.*
- *To make sure $|f_{\mu,\nu}(\boldsymbol{x}_t) - f_\nu(\boldsymbol{x}_t)| \leq \epsilon$, $\forall\ \boldsymbol{x}_t \in \mathcal{N}_R(\boldsymbol{x}_0)$, we utilize the Theorem 1 in [1], $|f_{\mu,\nu}(\boldsymbol{x}) - f_\nu(\boldsymbol{x})| \leq \mu L_0(f) d^{1/2}$. So, we could choose $\mu \leq \frac{\epsilon}{d^{1/2} L_0(f)}$.*
- *In order to ensure that expected squared norm of $\nabla f_{\mu,\nu}$ can reach δ, we set $\frac{2L_0(f)^{\frac{5}{2}} R^{\frac{1}{2}} d^{\frac{3}{2}}}{(Q+1)^{\frac{1}{2}} \epsilon^{\frac{1}{2}}}(\alpha + \frac{\sqrt{2}}{2}) = \delta$. Therefore, the expected number of queries is $O\left((\alpha + \frac{\sqrt{2}}{2})^2 \frac{d^3 L_0^5(f) R}{\epsilon \delta^2}\right)$.*

[1] Yurii Nesterov et al., Random gradient-free minimization of convex functions, 2017

Theoretical Analysis of RND Against ZO Attacks

Remark 1.

- *Due to the non-convexity assumption, we only guarantee the convergence to a stationary point of the function $f_{\mu,\nu}(\boldsymbol{x})$, which is a smoothing approximation of f_ν.*
- *To make sure $|f_{\mu,\nu}(\boldsymbol{x}_t) - f_\nu(\boldsymbol{x}_t)| \leq \epsilon,\ \forall\ \boldsymbol{x}_t \in \mathcal{N}_R(\boldsymbol{x}_0)$, we utilize the Theorem 1 in [1], $|f_{\mu,\nu}(\boldsymbol{x}) - f_\nu(\boldsymbol{x})| \leq \mu L_0(f) d^{1/2}$. So, we could choose $\mu \leq \frac{\epsilon}{d^{1/2} L_0(f)}$.*
- *In order to ensure that expected squared norm of $\nabla f_{\mu,\nu}$ can reach δ, we set $\frac{2L_0(f)^{\frac{5}{2}} R^{\frac{1}{2}} d^{\frac{3}{2}}}{(Q+1)^{\frac{1}{2}} \epsilon^{\frac{1}{2}}}(\alpha + \frac{\sqrt{2}}{2}) = \delta$. Therefore, the expected number of queries is $O\left((\alpha + \frac{\sqrt{2}}{2})^2 \frac{d^3 L_0^5(f) R}{\epsilon \delta^2}\right)$.*

[1] Yurii Nesterov et al., Random gradient-free minimization of convex functions, 2017

Theoretical Analysis of RND Against ZO Attacks

Remark 1.

- *Due to the non-convexity assumption, we only guarantee the convergence to a stationary point of the function* $f_{\mu,\nu}(\boldsymbol{x})$*, which is a smoothing approximation of* f_ν*.*
- *To make sure* $|f_{\mu,\nu}(\boldsymbol{x}_t) - f_\nu(\boldsymbol{x}_t)| \leq \epsilon,\ \forall\ \boldsymbol{x}_t \in \mathcal{N}_R(\boldsymbol{x}_0)$*, we utilize the Theorem 1 in [1],* $|f_{\mu,\nu}(\boldsymbol{x}) - f_\nu(\boldsymbol{x})| \leq \mu L_0(f) d^{1/2}$*. So, we could choose* $\mu \leq \frac{\epsilon}{d^{1/2} L_0(f)}$*.*
- *In order to ensure that expected squared norm of* $\nabla f_{\mu,\nu}$ *can reach* δ*, we set* $\frac{2L_0(f)^{\frac{5}{2}} R^{\frac{1}{2}} d^{\frac{3}{2}}}{(Q+1)^{\frac{1}{2}} \epsilon^{\frac{1}{2}}}(\alpha + \frac{\sqrt{2}}{2}) = \delta$*. Therefore, the expected number of queries is* $O\left((\alpha + \frac{\sqrt{2}}{2})^2 \frac{d^3 L_0^5(f) R}{\epsilon \delta^2}\right)$*.*

[1] Yurii Nesterov et al., Random gradient-free minimization of convex functions, 2017

Theoretical Analysis of RND Against ZO Attacks

- Theorem 1 shows the convergence rate is positive related to the ratio $\frac{\nu}{\mu}$. **The larger ratio $\frac{\nu}{\mu}$ will lead to the higher upper bound of convergence error and slower convergence rate.**
- Under the queries limited setting, the attack efficiency will be significantly reduced, leading to **failed attacks or a much larger number of queries for successful attacks**.
- The larger ratio $\frac{\nu}{\mu}$ leads the effectiveness of RND.

Theoretical Analysis of RND Against ZO Attacks

- Theorem 1 shows the convergence rate is positive related to the ratio $\frac{\nu}{\mu}$. **The larger ratio $\frac{\nu}{\mu}$ will lead to the higher upper bound of convergence error and slower convergence rate.**
- Under the queries limited setting, the attack efficiency will be significantly reduced, leading to **failed attacks or a much larger number of queries for successful attacks**.
- The larger ratio $\frac{\nu}{\mu}$ leads the effectiveness of RND.

Theoretical Analysis of RND Against ZO Attacks

- Theorem 1 shows the convergence rate is positive related to the ratio $\frac{\nu}{\mu}$. **The larger ratio $\frac{\nu}{\mu}$ will lead to the higher upper bound of convergence error and slower convergence rate.**
- Under the queries limited setting, the attack efficiency will be significantly reduced, leading to **failed attacks or a much larger number of queries for successful attacks**.
- The larger ratio $\frac{\nu}{\mu}$ leads the effectiveness of RND.

Theoretical Analysis of RND Against ZO Attacks

- Trade-off of Larger ν and Clean Accuracy:
  If $f(\boldsymbol{x})$ is Lipschitz-continuous, then $|f_\nu(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \nu L_0(f) d^{1/2}$. The larger ν is, the larger the gap between $f_\nu(\boldsymbol{x})$ and $f(\boldsymbol{x})$. So the clean accuracy of model with adding larger noise will also decrease. This forms a **trade-off between defense performance of RND and clean accuracy.**

- Larger Noise Size μ Adopted by Attackers:
  The attacker may be aware of the defense mechanism, so they can also increase the adopted noise size μ. As shown in figure in next page, for NES attack, the attack failure rate is almost 0, when $\nu = \mu = 0.01$.
  However, increasing the noise size μ will also lead less accurate gradient estimation and random search in Eq.(3) and Eq.(5), **leading to a significant decrease in attack performance**.

Theoretical Analysis of RND Against ZO Attacks

- Trade-off of Larger ν and Clean Accuracy:
  If $f(\boldsymbol{x})$ is Lipschitz-continuous, then $|f_\nu(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \nu L_0(f) d^{1/2}$. The larger ν is, the larger the gap between $f_\nu(\boldsymbol{x})$ and $f(\boldsymbol{x})$. So the clean accuracy of model with adding larger noise will also decrease. This forms a **trade-off between defense performance of RND and clean accuracy.**

- Larger Noise Size μ Adopted by Attackers:
  The attacker may be aware of the defense mechanism, so they can also increase the adopted noise size μ. As shown in figure in next page, for NES attack, the attack failure rate is almost 0, when $\nu = \mu = 0.01$.
  However, increasing the noise size μ will also lead less accurate gradient estimation and random search in Eq.(3) and Eq.(5), **leading to a significant decrease in attack performance**.

Theoretical Analysis of RND Against ZO Attacks

Experimental results verify our theoretical findings.

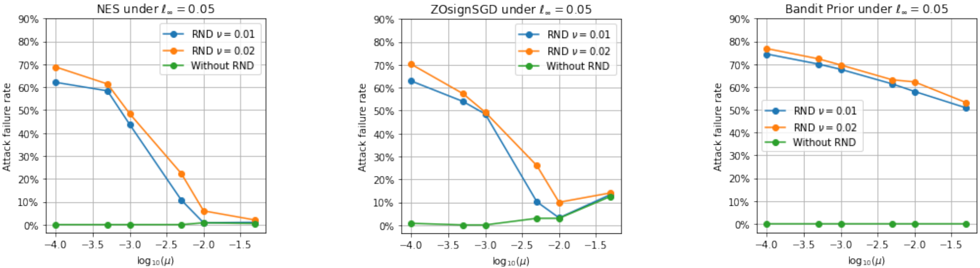


Figure: Attack failure rate (%) of query-based attacks on VGG-16 and CIFAR-10 under different values of μ and ν. We adopt logarithm scale for better illustration.

Theoretical Analysis of RND Against Adaptive Attacks

- As suggested in recent studies of robust defense [1, 2], the defender should take a robust evaluation against the **corresponding adaptive attack**, in which case **the attacker is aware of the defense mechanism**.
- Since the idea of RND is to insert random noise, an adaptive attacker could utilize Expectation Over Transformation (EOT) [1] to obtain an more accurate estimation, *i.e.*, querying one sample multiple times to obtain the average.
- Then, the original gradient estimator used in ZO attacks Eq.(11) is

$$g_{\mu,\nu}(\boldsymbol{x}) = \frac{f\left(\boldsymbol{x} + \mu\boldsymbol{u} + \nu\boldsymbol{v}_1\right) - f\left(\boldsymbol{x} + \nu\boldsymbol{v}_2\right)}{\mu}\boldsymbol{u}$$

Now, it becomes

$$\tilde{g}_{\mu,\nu}(\boldsymbol{x}) = \frac{1}{M}\sum_{j=1}^{M}\frac{f(\boldsymbol{x} + \mu\boldsymbol{u} + \nu\boldsymbol{v}_{j1}) - f(\boldsymbol{x} + \nu\boldsymbol{v}_{j2})}{\mu}\boldsymbol{\mu}, \tag{13}$$

[1] Anish Athalye et al., Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, ICML 2018
[2] Florian Tramer et al., On Adaptive Attacks to Adversarial Example Defenses, NeurIPS 2020

Theoretical Analysis of RND Against Adaptive Attacks

- As suggested in recent studies of robust defense [1, 2], the defender should take a robust evaluation against the **corresponding adaptive attack**, in which case **the attacker is aware of the defense mechanism**.
- Since the idea of RND is to insert random noise, an adaptive attacker could utilize Expectation Over Transformation (EOT) [1] to obtain an more accurate estimation, *i.e.*, querying one sample multiple times to obtain the average.
- Then, the original gradient estimator used in ZO attacks Eq.(11) is

$$g_{\mu,\nu}(\boldsymbol{x}) = \frac{f\left(\boldsymbol{x}+\mu\boldsymbol{u}+\nu\boldsymbol{v}_1\right)-f\left(\boldsymbol{x}+\nu\boldsymbol{v}_2\right)}{\mu}\boldsymbol{u}$$

Now, it becomes

$$\tilde{g}_{\mu,\nu}(\boldsymbol{x}) = \frac{1}{M}\sum_{j=1}^{M}\frac{f(\boldsymbol{x}+\mu\boldsymbol{u}+\nu\boldsymbol{v}_{j1})-f(\boldsymbol{x}+\nu\boldsymbol{v}_{j2})}{\mu}\boldsymbol{\mu}, \tag{13}$$

[1] Anish Athalye et al., Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, ICML 2018
[2] Florian Tramer et al., On Adaptive Attacks to Adversarial Example Defenses, NeurIPS 2020

Theoretical Analysis of RND Against Adaptive Attacks

- As suggested in recent studies of robust defense [1, 2], the defender should take a robust evaluation against the **corresponding adaptive attack**, in which case **the attacker is aware of the defense mechanism**.
- Since the idea of RND is to insert random noise, an adaptive attacker could utilize Expectation Over Transformation (EOT) [1] to obtain an more accurate estimation, *i.e.*, querying one sample multiple times to obtain the average.
- Then, the original gradient estimator used in ZO attacks Eq.(11) is

$$g_{\mu,\nu}(\boldsymbol{x}) = \frac{f\left(\boldsymbol{x} + \mu \boldsymbol{u} + \nu \boldsymbol{v}_1\right) - f\left(\boldsymbol{x} + \nu \boldsymbol{v}_2\right)}{\mu} \boldsymbol{u}$$

Now, it becomes

$$\tilde{g}_{\mu,\nu}(\boldsymbol{x}) = \frac{1}{M} \sum_{j=1}^{M} \frac{f(\boldsymbol{x} + \mu \boldsymbol{u} + \nu \boldsymbol{v}_{j1}) - f(\boldsymbol{x} + \nu \boldsymbol{v}_{j2})}{\mu} \boldsymbol{\mu}, \tag{13}$$

[1] Anish Athalye et al., Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, ICML 2018

[2] Florian Tramer et al., On Adaptive Attacks to Adversarial Example Defenses, NeurIPS 2020

Theoretical Analysis of RND Against Adaptive Attacks

The convergence analysis of ZO attack with Eq.(13) against RND is presented in Theorem 2.

Theorem 2.

Under Assumption 1 and 2, for any $Q \geq 0$, consider a sequence $\{\boldsymbol{x}_t\}_{t=0}^{Q}$ generated according to the descent update Eq.(11) using the gradient estimator $\tilde{g}_{\mu,\nu}(\boldsymbol{x})$ Eq.(13), we have

$$\begin{aligned}\frac{1}{S_Q}\sum_{t=0}^{Q}\eta_t\mathbb{E}_{\mathcal{U}_t,\mathcal{V}_t}(\|\nabla f_{\mu,\nu}(\boldsymbol{x}_t)\|^2) \leq & \frac{L_0(f)R}{S_Q} + \frac{1}{S_Q}\sum_{t=0}^{Q}\eta_t^2(L_0(f)^2L_1(f)d^2(\frac{1}{2}+\frac{2\nu^2}{\mu^2 M}) \\ & + \frac{\nu^2 L_0(f)L_1(f)^2}{\mu}d^{\frac{5}{2}} + \frac{\nu^4 L_1(f)^3(M+1)}{2\mu^2 M}d^3)\end{aligned} \tag{14}$$

Theoretical Analysis of RND Against Adaptive Attacks

- The larger M for EOT:
  Theorem 2 shows that with larger M, the upper bound will decrease. Therefore, EOT can mitigate the defense effect caused by the randomness of RND.
  However, with $M \longrightarrow \infty$, the upper bound of expected convergence error (*i.e.*, Eq. (14)) becomes

$$\frac{1}{S_Q}\sum_{t=0}^{Q}\eta_t\mathbb{E}_{\mathcal{U}_t,\mathcal{V}_t}(\|\nabla f_{\mu,\nu}(\boldsymbol{x}_t)\|^2) \leq \frac{L_0(f)R}{S_Q} + \frac{1}{S_Q}\sum_{t=0}^{Q}\eta_t^2(\frac{1}{2}L_0(f)^2L_1(f)d^2 + \frac{\nu^2 L_0(f)L_1(f)^2}{\mu}d^{\frac{5}{2}} + \frac{\nu^4 L_1(f)^3}{2\mu^2}d^3)$$

  which is still dominated by the max term $\frac{\nu^4}{\mu^2}d^3$. **It implies that the attack improvement from EOT is limited, especially with the larger ratio $\frac{\nu}{\mu}$.**

Theoretical Analysis of RND Against Adaptive Attacks

- The larger M for EOT:
  Theorem 2 shows that with larger M, the upper bound will decrease. Therefore, EOT can mitigate the defense effect caused by the randomness of RND.
  However, with $M \longrightarrow \infty$, the upper bound of expected convergence error (*i.e.*, Eq. (14)) becomes

$$\frac{1}{S_Q}\sum_{t=0}^{Q}\eta_t\mathbb{E}_{\mathcal{U}_t,\mathcal{V}_t}(\|\nabla f_{\mu,\nu}(\boldsymbol{x}_t)\|^2) \leq \frac{L_0(f)R}{S_Q} + \frac{1}{S_Q}\sum_{t=0}^{Q}\eta_t^2(\frac{1}{2}L_0(f)^2L_1(f)d^2 + \frac{\nu^2 L_0(f)L_1(f)^2}{\mu}d^{\frac{5}{2}} + \frac{\nu^4 L_1(f)^3}{2\mu^2}d^3)$$

which is still dominated by the max term $\frac{\nu^4}{\mu^2}d^3$. **It implies that the attack improvement from EOT is limited, especially with the larger ratio $\frac{\nu}{\mu}$.**

Theoretical Analysis of RND Against Adaptive Attacks

Experimental results verify our theoretical findings. **The relative performance improvements induced by EOT generally decrease as M increases.**

Theoretical Analysis of RND Against Adaptive Attacks

Experimental results verify our theoretical findings. **The relative performance improvements induced by EOT generally decrease as M increases.**

settings	Methods	M=1	M=5	M= 10	Methods	M=1	M=5	M= 10
adaptive	NES	1448/0.484	4078/0.361	5763/0.342	NES	2532/0.762	5364/0.705	7582/0.691
	ZS	1489/0.493	3189/0.374	5912/0.349	ZS	2824/0.825	5735/0.761	7662/0.740
fixed	NES	1448/0.484	2528/0.452	3246/0.443	NES	2533/0.762	5240/0.775	5658/0.781
	ZS	1489/0.493	2765/0.448	3123/0.421	ZS	2824/0.825	4023/0.842	4652/0.861
	Bandit	436/0.696	276/0.582	314/0.543	Bandit	305/0.604	759/0.523	946/0.49
	Square	380/0.301	181/0.162	223/0.121	Square	93/0.353	145/0.20	328/0.171
	SignHunter	459/0.367	559/0.224	759/0.191	SignHunter	173/0.532	336/0.456	659/0.431
	ECO	904/0.720	1681/0.761	2560/0.793	ECO	1237/0.666	3065/0.678	3091/0.692
	SimBA	1353/0.650	3852/0.467	4103/0.396	SimBA	274/0.891	468/0.878	517/0.869

Figure: The evaluation of EOT with ℓ_∞ attack on CIFAR-10 and ImageNet under the *adaptive and fixed query setting*. The left part is the results on **CIFAR-10** and the right part is on **ImageNet. The average number of query** of successful attack as well as **the attack failure rate** are reported.

Theoretical Analysis of RND Against Search-based Attacks

Recall the original searching direction is

$$s_{\mu}(\boldsymbol{x}) = \mathbb{I}\{h_{\mu}(\boldsymbol{x}) < 0\} \cdot \mu\boldsymbol{u} \quad \text{where } h_{\mu}(\boldsymbol{x}) = f(\boldsymbol{x} + \mu\boldsymbol{u}) - f(\boldsymbol{x}).$$

Therefore, the searching direction under RND becomes

$$s_{\nu}(\boldsymbol{x}) = \mathbb{I}\left(h_{\nu}(\boldsymbol{x}) < 0\right) \cdot \mu\boldsymbol{u} \;\; \text{where } h_{\nu}(\boldsymbol{x}) = f\left(\boldsymbol{x} + \mu\boldsymbol{u} + \nu\boldsymbol{v}_1\right) - f\left(\boldsymbol{x} + \nu\boldsymbol{v}_2\right) \tag{15}$$

- By adding noise $\nu\boldsymbol{v}$, the value of $h_{\nu}(\boldsymbol{x})$ will be different from that of $h_{\mu}(\boldsymbol{x})$, and there is certain probability that $\text{Sign}(h_{\nu}(\boldsymbol{x}))$ be different from $\text{Sign}(h_{\mu}(\boldsymbol{x}))$.
- When the random noise $\nu\boldsymbol{v}$ causes inconsistence between $\text{Sign}(h_{\nu}(\boldsymbol{x}))$ and $\text{Sign}(h_{\mu}(\boldsymbol{x}))$, RND will mislead the attackers to select the incorrect attack directions (*i.e.*, abandoning the descent direction *w.r.t.* f or selecting the ascent direction), so as to decrease the attack performance.

Theoretical Analysis of RND Against Search-based Attacks

Recall the original searching direction is

$$s_{\mu}(\boldsymbol{x}) = \mathbb{I}\{h_{\mu}(\boldsymbol{x}) < 0\} \cdot \mu\boldsymbol{u} \quad \text{where } h_{\mu}(\boldsymbol{x}) = f(\boldsymbol{x} + \mu\boldsymbol{u}) - f(\boldsymbol{x}).$$

Therefore, the searching direction under RND becomes

$$s_{\nu}(\boldsymbol{x}) = \mathbb{I}\left(h_{\nu}(\boldsymbol{x}) < 0\right) \cdot \mu\boldsymbol{u} \;\; \text{where } h_{\nu}(\boldsymbol{x}) = f\left(\boldsymbol{x} + \mu\boldsymbol{u} + \nu\boldsymbol{v}_1\right) - f\left(\boldsymbol{x} + \nu\boldsymbol{v}_2\right) \tag{15}$$

- By adding noise $\nu\boldsymbol{v}$, the value of $h_{\nu}(\boldsymbol{x})$ will be different from that of $h_{\mu}(\boldsymbol{x})$, and there is certain probability that $\text{Sign}(h_{\nu}(\boldsymbol{x}))$ be different from $\text{Sign}(h_{\mu}(\boldsymbol{x}))$.
- When the random noise $\nu\boldsymbol{v}$ causes inconsistence between $\text{Sign}(h_{\nu}(\boldsymbol{x}))$ and $\text{Sign}(h_{\mu}(\boldsymbol{x}))$, RND will mislead the attackers to select the incorrect attack directions (*i.e.*, abandoning the descent direction *w.r.t.* f or selecting the ascent direction), so as to decrease the attack performance.

Theoretical Analysis of RND Against Search-based Attacks

Recall the original searching direction is

$$s_{\mu}(\boldsymbol{x}) = \mathbb{I}\{h_{\mu}(\boldsymbol{x}) < 0\} \cdot \mu\boldsymbol{u} \quad \text{where } h_{\mu}(\boldsymbol{x}) = f(\boldsymbol{x} + \mu\boldsymbol{u}) - f(\boldsymbol{x}).$$

Therefore, the searching direction under RND becomes

$$s_{\nu}(\boldsymbol{x}) = \mathbb{I}\left(h_{\nu}(\boldsymbol{x}) < 0\right) \cdot \mu\boldsymbol{u} \ \text{ where } h_{\nu}(\boldsymbol{x}) = f\left(\boldsymbol{x} + \mu\boldsymbol{u} + \nu\boldsymbol{v}_1\right) - f\left(\boldsymbol{x} + \nu\boldsymbol{v}_2\right) \tag{15}$$

- By adding noise $\nu\boldsymbol{v}$, the value of $h_{\nu}(\boldsymbol{x})$ will be different from that of $h_{\mu}(\boldsymbol{x})$, and there is certain probability that Sign$(h_{\nu}(\boldsymbol{x}))$ be different from Sign$(h_{\mu}(\boldsymbol{x}))$.
- When the random noise $\nu\boldsymbol{v}$ causes inconsistence between Sign$(h_{\nu}(\boldsymbol{x}))$ and Sign$(h_{\mu}(\boldsymbol{x}))$, RND will mislead the attackers to select the incorrect attack directions (*i.e.*, abandoning the descent direction *w.r.t.* f or selecting the ascent direction), so as to decrease the attack performance.

Theoretical Analysis of RND Against Search-based Attacks

Recall the original searching direction is

$$s_\mu(\boldsymbol{x}) = \mathbb{I}\{h_\mu(\boldsymbol{x}) < 0\} \cdot \mu\boldsymbol{u} \quad \text{where } h_\mu(\boldsymbol{x}) = f(\boldsymbol{x} + \mu\boldsymbol{u}) - f(\boldsymbol{x}).$$

Therefore, the searching direction under RND becomes

$$s_\nu(\boldsymbol{x}) = \mathbb{I}\left(h_\nu(\boldsymbol{x}) < 0\right) \cdot \mu\boldsymbol{u} \ \text{ where } h_\nu(\boldsymbol{x}) = f\left(\boldsymbol{x} + \mu\boldsymbol{u} + \nu\boldsymbol{v}_1\right) - f\left(\boldsymbol{x} + \nu\boldsymbol{v}_2\right) \tag{15}$$

- By adding noise $\nu\boldsymbol{v}$, the value of $h_\nu(\boldsymbol{x})$ will be different from that of $h_\mu(\boldsymbol{x})$, and there is certain probability that Sign$(h_\nu(\boldsymbol{x}))$ be different from Sign$(h_\mu(\boldsymbol{x}))$.
- When the random noise $\nu\boldsymbol{v}$ causes inconsistence between Sign$(h_\nu(\boldsymbol{x}))$ and Sign$(h_\mu(\boldsymbol{x}))$, RND will mislead the attackers to select the incorrect attack directions (*i.e.*, abandoning the descent direction *w.r.t.* f or selecting the ascent direction), so as to decrease the attack performance.

Theoretical Analysis of RND Against Search-based Attacks

Theorem 3.

Under Assumption 1, considering the direction update Eq.(6) with Eq.(15) in search-based attacks, we have,

$$P(\mathit{Sign}(h_\mu(\boldsymbol{x})) \neq \mathit{Sign}(h_\nu(\boldsymbol{x})) \leq \frac{2L_0(f)\nu\sqrt{d}}{|h_\mu(\boldsymbol{x})|} \tag{16}$$

Remark 2.

- *Theorem 3 shows the probability of misleading attacker is positive correlated with* $\frac{\nu}{|h_\mu(\boldsymbol{x})|}$.
- *Due to the small value* μ *and local linearity of smooth function, we* $|h_\mu(x)| = |f(x+\mu\boldsymbol{u}) - f(x)| \approx C\mu\|\boldsymbol{u}\|$. *The* $|h_\mu(x)|$ *is also positive correlated with the stepsize* μ *within the small neighborhoods.*
- **So the probability of changing the sign is positive correlated with** $\frac{\nu}{\mu}$.

Theoretical Analysis of RND Against Search-based Attacks

Theorem 3.

Under Assumption 1, considering the direction update Eq.(6) with Eq.(15) in search-based attacks, we have,

$$P(\mathit{Sign}(h_\mu(\boldsymbol{x})) \neq \mathit{Sign}(h_\nu(\boldsymbol{x})) \leq \frac{2L_0(f)\nu\sqrt{d}}{|h_\mu(\boldsymbol{x})|} \tag{16}$$

Remark 2.

- *Theorem 3 shows the probability of misleading attacker is positive correlated with* $\frac{\nu}{|h_\mu(\boldsymbol{x})|}$.
- *Due to the small value* μ *and local linearity of smooth function, we* $|h_\mu(x)| = |f(x+\mu\boldsymbol{u}) - f(x)| \approx C\mu\|\boldsymbol{u}\|$. *The* $|h_\mu(x)|$ *is also positive correlated with the stepsize* μ *within the small neighborhoods.*
- **So the probability of changing the sign is positive correlated with** $\frac{\nu}{\mu}$.

Theoretical Analysis of RND Against Search-based Attacks

Theorem 3.

Under Assumption 1, considering the direction update Eq.(6) with Eq.(15) in search-based attacks, we have,

$$P(\mathit{Sign}(h_\mu(\boldsymbol{x})) \neq \mathit{Sign}(h_\nu(\boldsymbol{x})) \leq \frac{2L_0(f)\nu\sqrt{d}}{|h_\mu(\boldsymbol{x})|} \tag{16}$$

Remark 2.

- *Theorem 3 shows the probability of misleading attacker is positive correlated with* $\frac{\nu}{|h_\mu(\boldsymbol{x})|}$.
- *Due to the small value* μ *and local linearity of smooth function, we* $|h_\mu(\boldsymbol{x})| = |f(\boldsymbol{x}+\mu\boldsymbol{u}) - f(\boldsymbol{x})| \approx C\mu\|\boldsymbol{u}\|$. *The* $|h_\mu(\boldsymbol{x})|$ *is also positive correlated with the stepsize* μ *within the small neighborhoods.*
- So the probability of changing the sign is positive correlated with $\frac{\nu}{\mu}$.

Theoretical Analysis of RND Against Search-based Attacks

Theorem 3.

Under Assumption 1, considering the direction update Eq.(6) with Eq.(15) in search-based attacks, we have,

$$P(\mathit{Sign}(h_\mu(\boldsymbol{x})) \neq \mathit{Sign}(h_\nu(\boldsymbol{x})) \leq \frac{2L_0(f)\nu\sqrt{d}}{|h_\mu(\boldsymbol{x})|} \tag{16}$$

Remark 2.

- *Theorem 3 shows the probability of misleading attacker is positive correlated with* $\frac{\nu}{|h_\mu(\boldsymbol{x})|}$.
- *Due to the small value* μ *and local linearity of smooth function, we* $|h_\mu(\boldsymbol{x})| = |f(\boldsymbol{x} + \mu\boldsymbol{u}) - f(\boldsymbol{x})| \approx C\mu\|\boldsymbol{u}\|$. *The* $|h_\mu(\boldsymbol{x})|$ *is also positive correlated with the stepsize* μ *within the small neighborhoods.*
- **So the probability of changing the sign is positive correlated with** $\frac{\nu}{\mu}$.

Theoretical Analysis of RND Against Search-based Attacks

Experimental results verify our theoretical findings.

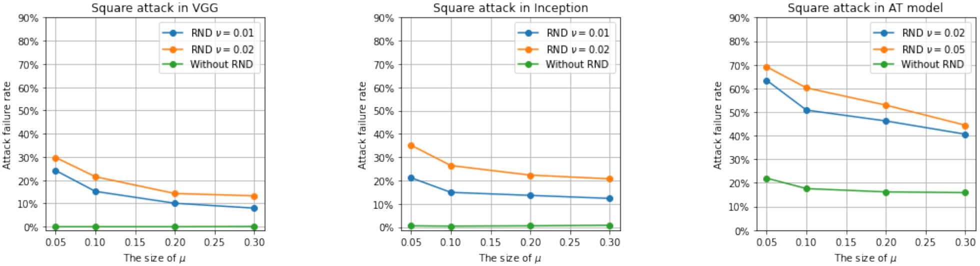


Figure: Attack failure rate (%) of Square ℓ_∞ attacks on VGG-16(CIFAR-10), Inception v3(ImageNet) and AT model (ImageNet) under different values of μ and ν, where μ is the square size in Square attacks.

Better Trade-off Between Defense Effect and Clean Accuracy

- To achieve a high-quality balance, we could reduce the sensitivity of the target model to random noises.
- We propose to utilize **Gaussian Augmentation Fine-tuning (GF)**, the loss function (CE loss) is

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\in D} - \boldsymbol{y}^T \log \left(\underset{\delta\sim\mathcal{N}(0,\sigma^2 I)}{\mathbb{E}} \left[\left(F\left(\boldsymbol{x} + \delta \right) \right) \right] \right)$$

Better Trade-off Between Defense Effect and Clean Accuracy

- To achieve a high-quality balance, we could reduce the sensitivity of the target model to random noises.
- We propose to utilize **Gaussian Augmentation Fine-tuning (GF)**, the loss function (CE loss) is

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\in D} - \boldsymbol{y}^T \log \left(\mathop{\mathbb{E}}_{\delta\sim\mathcal{N}(0,\sigma^2 I)} \left[\left(F\left(\boldsymbol{x} + \delta \right) \right) \right] \right)$$

Better Trade-off Between Defense Effect and Clean Accuracy

- To achieve a high-quality balance, we could reduce the sensitivity of the target model to random noises.
- We propose to utilize **Gaussian Augmentation Fine-tuning (GF)**, the loss function (CE loss) is

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\in D} - \boldsymbol{y}^T \log \left(\underset{\delta \sim \mathcal{N}(0,\sigma^2 I)}{\mathbb{E}} \left[\left(F\left(\boldsymbol{x} + \delta\right)\right)\right] \right)$$

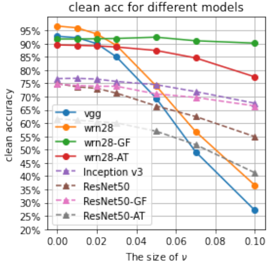


Figure: Clean accuracy for different models on CIFAR-10 and ImageNet. The **circle lines** and **triangle lines** represent models on **CIFAR-10** and **ImageNet** respectively.

Better Trade-off Between Defense Effect and Clean Accuracy

- Compared with RND, **RNG-GF** significantly improves the defense performance under all attack methods while maintaining the good clean accuracy.
- Combining AT with RND, RND-AT significantly improves the robustness against all attacks and achieves best performance among all methods.

Table 2: The comparison of RND ($\nu = 0.02$), GF, RND-GF ($\nu = 0.05$), AT, RND-AT ($\nu = 0.05$), PNI, RSE, and FD on CIFAR-10 and Imagenet. **The average number of queries** of successful attack and **the attack failure rates** are reported. The best and second best attack failure rate under each attack are highlighted in bold and underlined, respectively. The evaluation under ℓ_2 attack is shown in Section B.6 of supplementary materials.

Datasets	Methods	Clean Acc	NES(ℓ_∞)	ZS(ℓ_∞)	Bandit(ℓ_∞)	Sign(ℓ_∞)	Square(ℓ_∞)	SimBA(ℓ_2)	ECO(ℓ_∞)
CIFAR-10 (WideNet-28)	Clean Model	**96.60%**	465.5/0.01	581.8/0.06	210.2/0.03	167.6/0.03	137.1/0.02	457.2/0.04	457.8/0.0
	GF	91.72%	999.0/0.407	759.9/0.544	744.5/0.116	348.3/0.027	581.0/0.061	1146.8/0.395	883.9/0.067
	RSE[30]	84.12%	1246.3/0.396	1327.8/0.422	281.7/0.372	243.7/0.221	413.3/0.243	498.3/0.337	578.3/0.534
	PNI[23]	87.20%	1071.4/0.725	1310.7/0.823	324.9/0.824	267.0/0.708	295.3/0.612	945.0/0.857	2342.2/0.623
	AT[20]	89.48%	821.6/0.807	614.9/0.862	1451.5/0.623	766.3/0.476	1135.4/0.499	1523.2/0.635	1180.4/0.484
	RND	93.60%	842.5/0.05	941.8/0.143	273.1/0.478	977.2/0.226	762.4/0.116	2112.6/0.549	912.8/0.688
	RND-GF	92.40%	2805.7/0.516	2966.3/0.730	1223.5/0.841	1017.1/0.407	1207.3/0.378	1220.2/0.863	687.2/0.872
	RND-AT	87.40%	2499.2/**0.842**	2625.7/**0.923**	891.5/**0.891**	767.9/**0.737**	1170.7/**0.730**	1787.4/**0.912**	687.4/ **0.911**
ImageNet (ResNet-50)	Clean Model	**74.90%**	1031.9/0.0	2013.0/0.235	329.2/0.02	264.1/0.03	76.5/0.0	1234.5/0.281	347.7/0.0
	GF[37]	74.70%	1685.5/0.03	1712.1/0.347	601.4/0.02	329.0/0.0	97.28/0.0	1417.4/0.112	362.4/0.0
	FD[49]	54.20%	1997.2/0.679	1555.5/0.775	1579.2/0.426	1633.1/0.332	1092.4/0.242	2607.9/0.613	1501.0/0.240
	AT[17]	61.60%	2113.4/0.724	1688.7/0.815	1091.5/0.416	1522.7/0.289	1109.0/0.159	2638.2/0.651	1440.6/0.200
	RND	73.00%	3041.5/0.245	2266.2/0.330	390.6/0.536	661.0/0.314	81.5/0.101	825.3/0.612	2435.5/0.540
	RND-GF	71.15%	2489.3/0.421	2053.5/0.563	495.9/0.603	514.0/0.348	1009.9/0.146	777.2/ 0.762	994.8/0.702
	RND-AT	58.15%	2556.6/**0.864**	2596.6/**0.870**	448.0/**0.810**	724.2/**0.632**	1306.3/**0.386**	1210.5/**0.953**	631.1/**0.865**

Better Trade-off Between Defense Effect and Clean Accuracy

- Compared with RND, **RNG-GF** significantly improves the defense performance under all attack methods while maintaining the good clean accuracy.
- Combining AT with RND, RND-AT significantly improves the robustness against all attacks and achieves best performance among all methods.

Table 2: The comparison of RND ($\nu = 0.02$), GF, RND-GF ($\nu = 0.05$), AT, RND-AT ($\nu = 0.05$), PNI, RSE, and FD on CIFAR-10 and Imagenet. **The average number of queries** of successful attack and **the attack failure rates** are reported. The best and second best attack failure rate under each attack are highlighted in bold and underlined, respectively. The evaluation under ℓ_2 attack is shown in Section B.6 of supplementary materials.

Datasets	Methods	Clean Acc	NES(ℓ_∞)	ZS(ℓ_∞)	Bandit(ℓ_∞)	Sign(ℓ_∞)	Square(ℓ_∞)	SimBA(ℓ_2)	ECO(ℓ_∞)
CIFAR-10 (WideNet-28)	Clean Model	**96.60%**	465.5/0.01	581.8/0.06	210.2/0.03	167.6/0.03	137.1/0.02	457.2/0.04	457.8/0.0
	GF	91.72%	999.0/0.407	759.9/0.544	744.5/0.116	348.3/0.027	581.0/0.061	1146.8/0.395	883.9/0.067
	RSE[30]	84.12%	1246.3/0.396	1327.8/0.422	281.7/0.372	243.7/0.221	413.3/0.243	498.3/0.337	578.3/0.534
	PNI[23]	87.20%	1071.4/0.725	1310.7/0.823	324.9/0.824	267.0/<u>0.708</u>	295.3/<u>0.612</u>	945.0/0.857	2342.2/0.623
	AT[20]	89.48%	821.6/<u>0.807</u>	614.9/<u>0.862</u>	1451.5/0.623	766.3/0.476	1135.4/0.499	1523.2/0.635	1180.4/0.484
	RND	<u>93.60%</u>	842.5/0.05	941.8/0.143	273.1/0.478	977.2/0.226	762.4/0.116	2112.6/0.549	912.8/0.688
	RND-GF	92.40%	2805.7/0.516	2966.3/0.730	1223.5/<u>0.841</u>	1017.1/0.407	1207.3/0.378	1220.2/<u>0.863</u>	687.2/<u>0.872</u>
	RND-AT	87.40%	2499.2/**0.842**	2625.7/**0.923**	891.5/**0.891**	767.9/**0.737**	1170.7/**0.730**	1787.4/**0.912**	687.4/ **0.911**
ImageNet (ResNet-50)	Clean Model	**74.90%**	1031.9/0.0	2013.0/0.235	329.2/0.02	264.1/0.03	76.5/0.0	1234.5/0.281	347.7/0.0
	GF[37]	<u>74.70%</u>	1685.5/0.03	1712.1/0.347	601.4/0.02	329.0/0.0	97.28/0.0	1417.4/0.112	362.4/0.0
	FD[49]	54.20%	1997.2/0.679	1555.5/0.775	1579.2/0.426	1633.1/0.332	1092.4/<u>0.242</u>	2607.9/0.613	1501.0/0.240
	AT[17]	61.60%	2113.4/<u>0.724</u>	1688.7/<u>0.815</u>	1091.5/0.416	1522.7/0.289	1109.0/0.159	2638.2/0.651	1440.6/0.200
	RND	73.00%	3041.5/0.245	2266.2/0.330	390.6/0.536	661.0/0.314	81.5/0.101	825.3/0.612	2435.5/0.540
	RND-GF	71.15%	2489.3/0.421	2053.5/0.563	495.9/<u>0.603</u>	514.0/<u>0.348</u>	1009.9/0.146	777.2/ <u>0.762</u>	994.8/<u>0.702</u>
	RND-AT	58.15%	2556.6/**0.864**	2596.6/**0.870**	448.0/**0.810**	724.2/**0.632**	1306.3/**0.386**	1210.5/**0.953**	631.1/**0.865**