Diversified Adversarial Attacks based on Conjugate Gradient Method

Keiichiro Yamamura[1], Haruki Sato[1], Nariaki Tateiwa[1,3], Nozomi Hata[1], Toru Mitsutake[1], Issa Oe[1], Hiroki Ishikura[1], Katsuki Fujisawa[2]

[1] Graduate School of Mathematics, Kyushu University
[2] Institute of Mathematics for Industry, Kyushu University
[3] Present affiliation is NTT Software Innovation Center, NTT Corporation

International Conference on Machine Learning, 2022

Background

- Deep learning models are vulnerable to adversarial examples.
  - *Adversarial examples*：Perturbed images that mislead the prediction of a model.
  - *Adversarial attack*： The way to generate adversarial examples.
  - *Adversarial training*： One of the famous defenses against adversarial attacks, which uses adversarial examples during training phase.

→To make the model robust to strong attacks, adversarial examples created by the strong attack method are required.

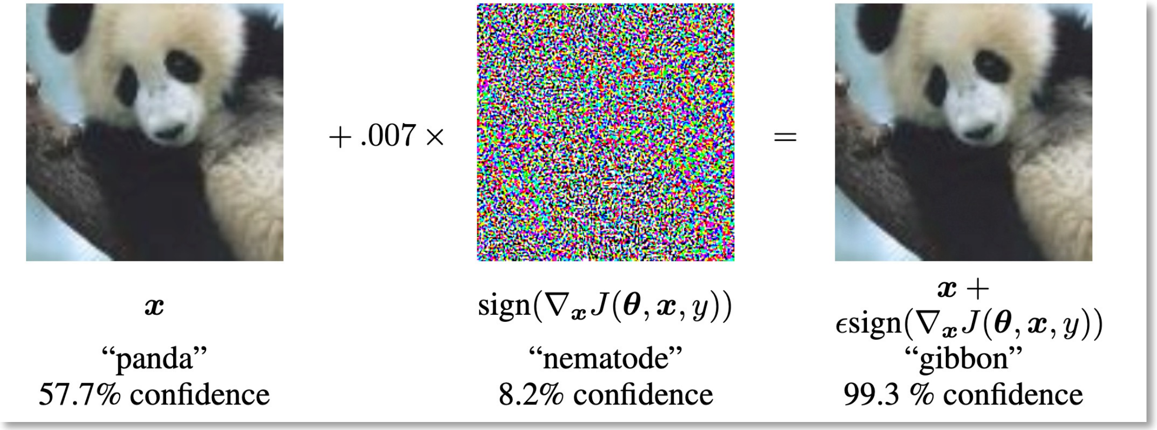

$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure:（Ian Goodfellow, et al, 2014)

Background

- Deep learning models are vulnerable to adversarial examples.
  - *Adversarial examples*：Perturbed images that mislead the prediction of a model.
  - *Adversarial attack*： The way to generate adversarial examples.
  - *Adversarial training*： One of the famous defenses against adversarial attacks, which uses adversarial examples during training phase.

→To make the model robust to strong attacks, adversarial examples created by the strong attack method are required.

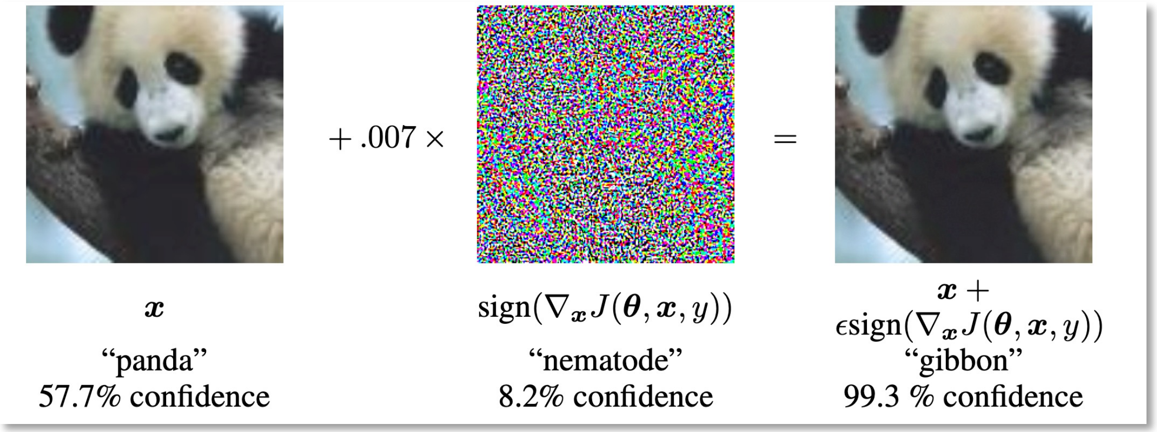

+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure:（Ian Goodfellow, et al, 2014)

Contributions

1. Propose a more diversified algorithm, **Auto-Conjugate Gradient Attack (ACG)**
   - ✓ Based on Conjugate Gradient Method
   - ✓ Higher ASR than APGD for 63 models out of 64
   - ✓ Higher ASR than APGD for 49 models out of 64 with fewer iterations and only deterministic operations

2. Propose Diversity Index (DI) which quantify the degree of diversification of the attacks
   - ✓ First attempt to quantify the degree of diversification of adversarial attacks
   - ✓ Based on the global clustering coefficients of the graph whose nodes are latest K search points
   - ✓ Capture the difference among algorithms by the transitions of DI.

Contributions

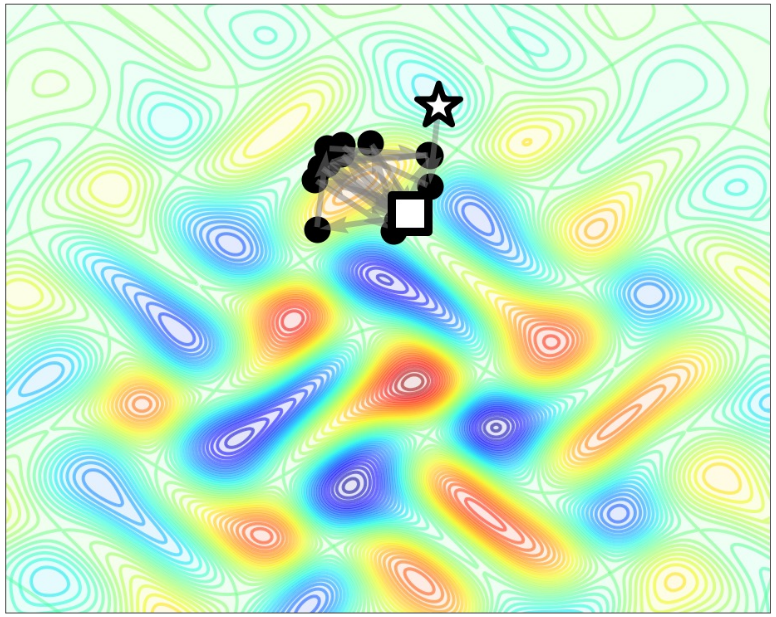
1. Propose a more diversified algorithm, Auto-Conjugate Gradient Attack (ACG)
   - ✓ Based on Conjugate Gradient Method
   - ✓ Higher ASR than APGD for 63 models out of 64
   - ✓ Higher ASR than APGD for 49 models out of 64 with fewer iterations and only deterministic operations

2. Propose **Diversity Index (DI)** to quantify the degree of diversification of the attacks
   - ✓ First attempt to quantify the degree of diversification of adversarial attacks
   - ✓ Based on the global clustering coefficients of the graph whose nodes are latest K search points
   - ✓ Capture the difference among algorithms by the transitions of DI

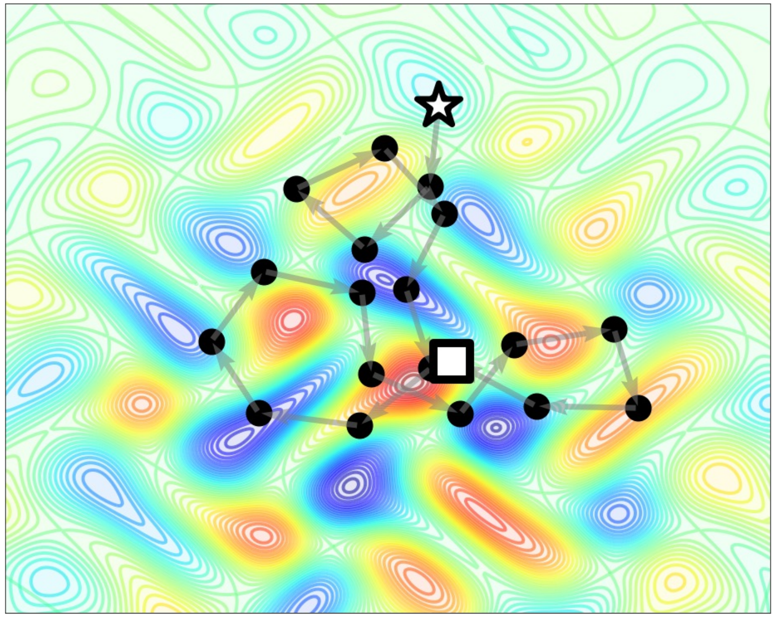
Section 3: Auto-Conjugate Gradient Attack (ACG)

Motivation

The first method considered for optimization problems with gradients is the steepest descent method.
However, when the steepest descent method does not work well, the CG method is used as the second option.
To avoid stacking in local optima (left figure), the diversification of the search is important (right figure).
Then, we applied CG-inspired method to this problem in the hope of achieving search diversification.



Bad: stack in local optima



Good: almost find the global optima.

Section 3: Comparison with ACG and APGD

APGD (Croce & Hein, 2020)

State-of-the-art steepest based method.

P_S: projection onto S

$$\boldsymbol{z}^{(k+1)} = P_S\left(\boldsymbol{x}^{(k)} + \eta^{(k)} \cdot \mathrm{Sign}\left(\nabla f\left(\boldsymbol{x}^{k}\right)\right)\right)$$

$$\boldsymbol{x}^{(k+1)} = P_S\left(\boldsymbol{x}^{(k)} + \alpha\left(\boldsymbol{z}^{(k+1)} - \boldsymbol{x}^{(k)}\right) + (1-\alpha)\left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right)\right)$$

ACG (proposed method)

$$\boldsymbol{y}^{(k-1)} = \nabla f\left(\boldsymbol{x}^{(k-1)}\right) - \nabla f\left(\boldsymbol{x}^{(k)}\right),$$

$$\beta_{HS}^{(k)} = \frac{\left\langle -\nabla f\left(\boldsymbol{x}^{(k)}\right), \boldsymbol{y}^{(k-1)} \right\rangle}{\left\langle \boldsymbol{s}^{(k-1)}, \boldsymbol{y}^{(k-1)} \right\rangle},$$

$$\boldsymbol{s}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right) + \beta_{HS}^{(k)} \boldsymbol{s}^{(k-1)},$$

$$\boldsymbol{x}^{(k+1)} = P_S\left(\boldsymbol{x}^{(k)} + \eta^{(k)} \cdot \mathrm{Sign}\left(\boldsymbol{s}^{(k)}\right)\right)$$

As describe above, there are two major difference between APGD and ACG.

1. Their update direction; ACG moves to CG direction while APGD moves to gradient and momentum direction.
2. Whether to use the momentum term.

While APGD tries to diversify its search by controlling step size, we try to perform a more diversified search using signed CG direction in the update.

Section 3: Comparison with ACG and APGD

APGD (Croce & Hein, 2020)

State-of-the-art steepest based method.

P_S: projection onto S

$$\boldsymbol{z}^{(k+1)} = P_S\left(\boldsymbol{x}^{(k)} + \eta^{(k)} \cdot \mathrm{Sign}\left(\nabla f\left(\boldsymbol{x}^{k}\right)\right)\right)$$

$$\boldsymbol{x}^{(k+1)} = P_S\left(\boldsymbol{x}^{(k)} + \alpha\left(\boldsymbol{z}^{(k+1)} - \boldsymbol{x}^{(k)}\right) + (1-\alpha)\left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right)\right)$$

ACG (proposed method)

$$\boldsymbol{y}^{(k-1)} = \nabla f\left(\boldsymbol{x}^{(k-1)}\right) - \nabla f\left(\boldsymbol{x}^{(k)}\right),$$

$$\beta_{HS}^{(k)} = \frac{\left\langle -\nabla f\left(\boldsymbol{x}^{(k)}\right), \boldsymbol{y}^{(k-1)}\right\rangle}{\left\langle \boldsymbol{s}^{(k-1)}, \boldsymbol{y}^{(k-1)}\right\rangle},$$

$$\boldsymbol{s}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right) + \beta_{HS}^{(k)} \boldsymbol{s}^{(k-1)},$$

$$\boldsymbol{x}^{(k+1)} = P_S\left(\boldsymbol{x}^{(k)} + \eta^{(k)} \cdot \mathrm{Sign}\left(\boldsymbol{s}^{(k)}\right)\right)$$

As describe above, there are two major difference between APGD and ACG.

1. Their update direction; ACG moves to CG direction while APGD moves to gradient and momentum direction.
2. Whether to use the momentum term.

While APGD tries to diversify its search by controlling step size, we try to perform a more diversified search using signed CG direction in the update.

Section 3: Comparison with ACG and APGD

APGD (Croce & Hein, 2020)

State-of-the-art steepest based method.
P_S: projection onto S

$$\boldsymbol{z}^{(k+1)} = P_S\left(\boldsymbol{x}^{(k)} + \eta^{(k)} \cdot \mathrm{Sign}\left(\nabla f\left(\boldsymbol{x}^{k}\right)\right)\right)$$

$$\boldsymbol{x}^{(k+1)} = P_S\left(\boldsymbol{x}^{(k)} + \alpha\left(\boldsymbol{z}^{(k+1)} - \boldsymbol{x}^{(k)}\right) + (1-\alpha)\left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right)\right)$$

ACG (proposed method)

$$\boldsymbol{y}^{(k-1)} = \nabla f\left(\boldsymbol{x}^{(k-1)}\right) - \nabla f\left(\boldsymbol{x}^{(k)}\right),$$

$$\beta_{HS}^{(k)} = \frac{\langle -\nabla f\left(\boldsymbol{x}^{(k)}\right), \boldsymbol{y}^{(k-1)} \rangle}{\langle \boldsymbol{s}^{(k-1)}, \boldsymbol{y}^{(k-1)} \rangle},$$

$$\boldsymbol{s}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right) + \beta_{HS}^{(k)} \boldsymbol{s}^{(k-1)},$$

$$\boldsymbol{x}^{(k+1)} = P_S\left(\boldsymbol{x}^{(k)} + \eta^{(k)} \cdot \mathrm{Sign}\left(\boldsymbol{s}^{(k)}\right)\right)$$

As describe above, there are two major difference between APGD and ACG.

1. Their update direction; ACG moves to CG direction while APGD moves to gradient and momentum direction.
2. Whether to use the momentum term.

While APGD tries to diversify its search by controlling step size, ACG tries to perform a more diversified search using signed CG direction in the update.

Section 4: Numerical Experiment (untargeted attack)

- #Models: 64 models in RobustBench
- Dataset: CIFAR-10, CIFAR-100, ImageNet
- Evaluation: Untargeted attacks
  (targeted attacks require more iterations than untargeted attacks)

The objective function:CW loss*

$$L(x, c) = -g_c(x) + \max_{i \neq c} g_i(x)$$

Summary of the results using CW loss

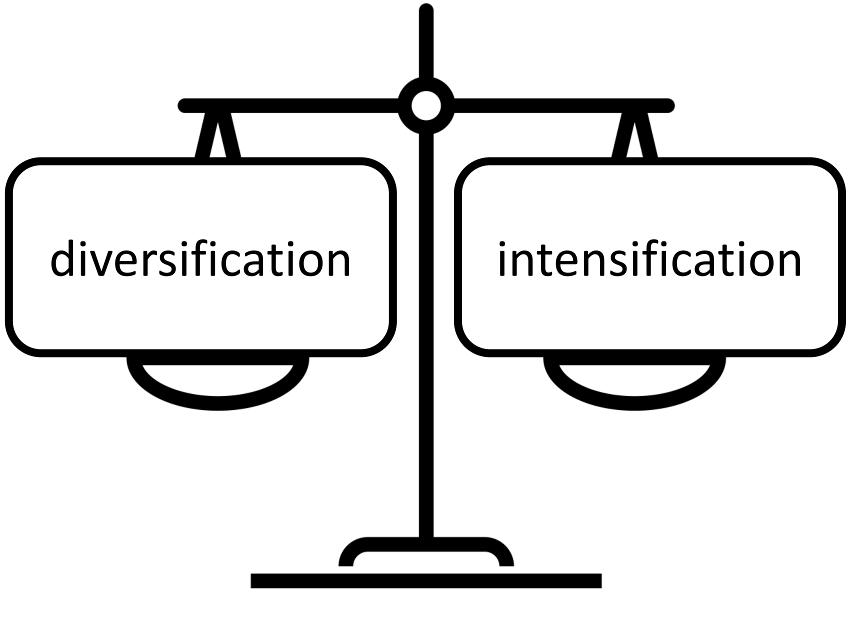
	APGD(1)	ACG(1)	APGD(5)	ACG(5)
#best	0	0	1	**63**
#2nd best	1	**49**	14	0

- ACG got higher ASR for **63/64** models
- **Only with a single initial point,**
  ACG got higher ASR for **49/64** models
  (In this case, ACG only uses **deterministic** operations.)

*We also evaluated the performance using DLR loss (Croce & Hein, 2020). Details are in Appendix F.

Section 5: Diversity Index

Motivation



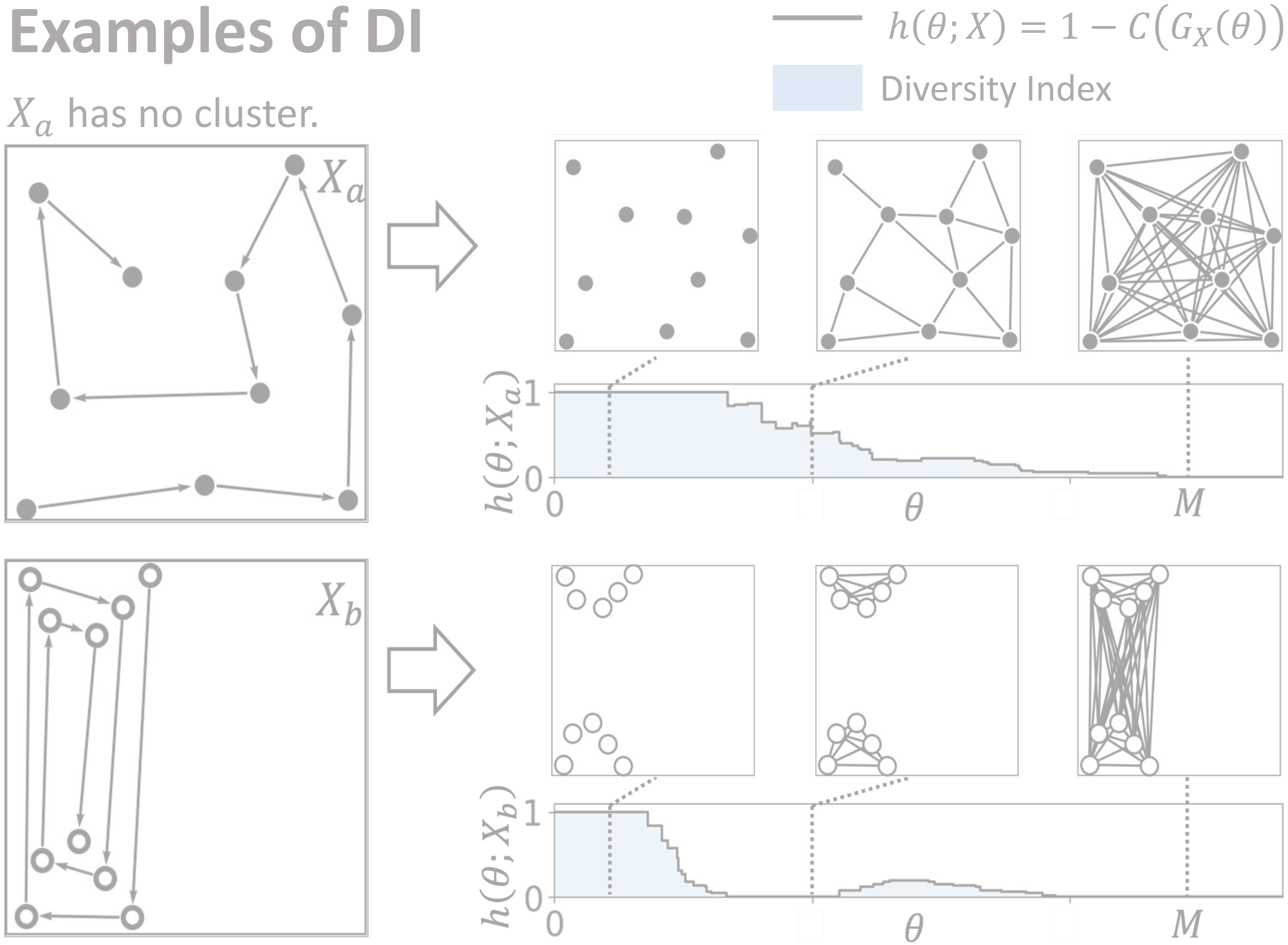
diversification

intensification

quantification

Examples of DI

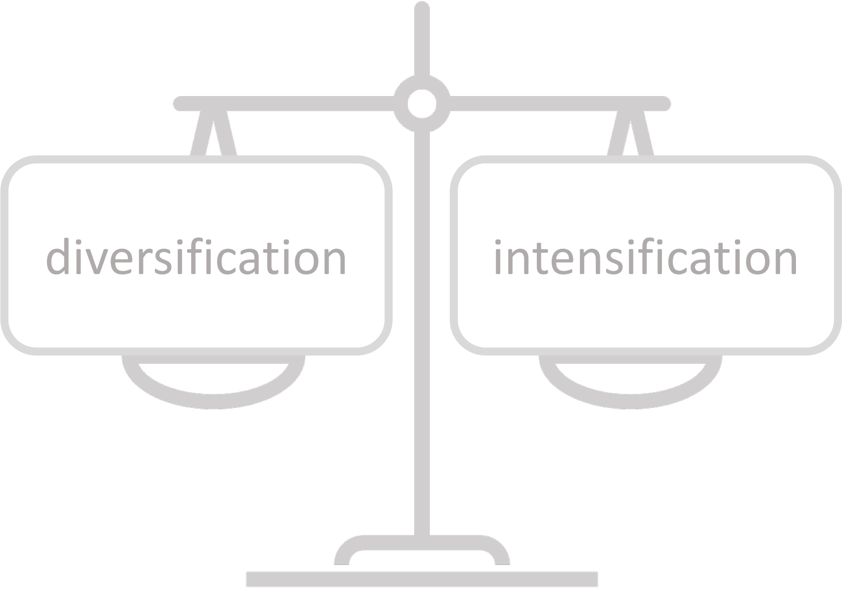
X_a has no cluster.



X_b has two clusters.

Section 5: Diversity Index

Motivation



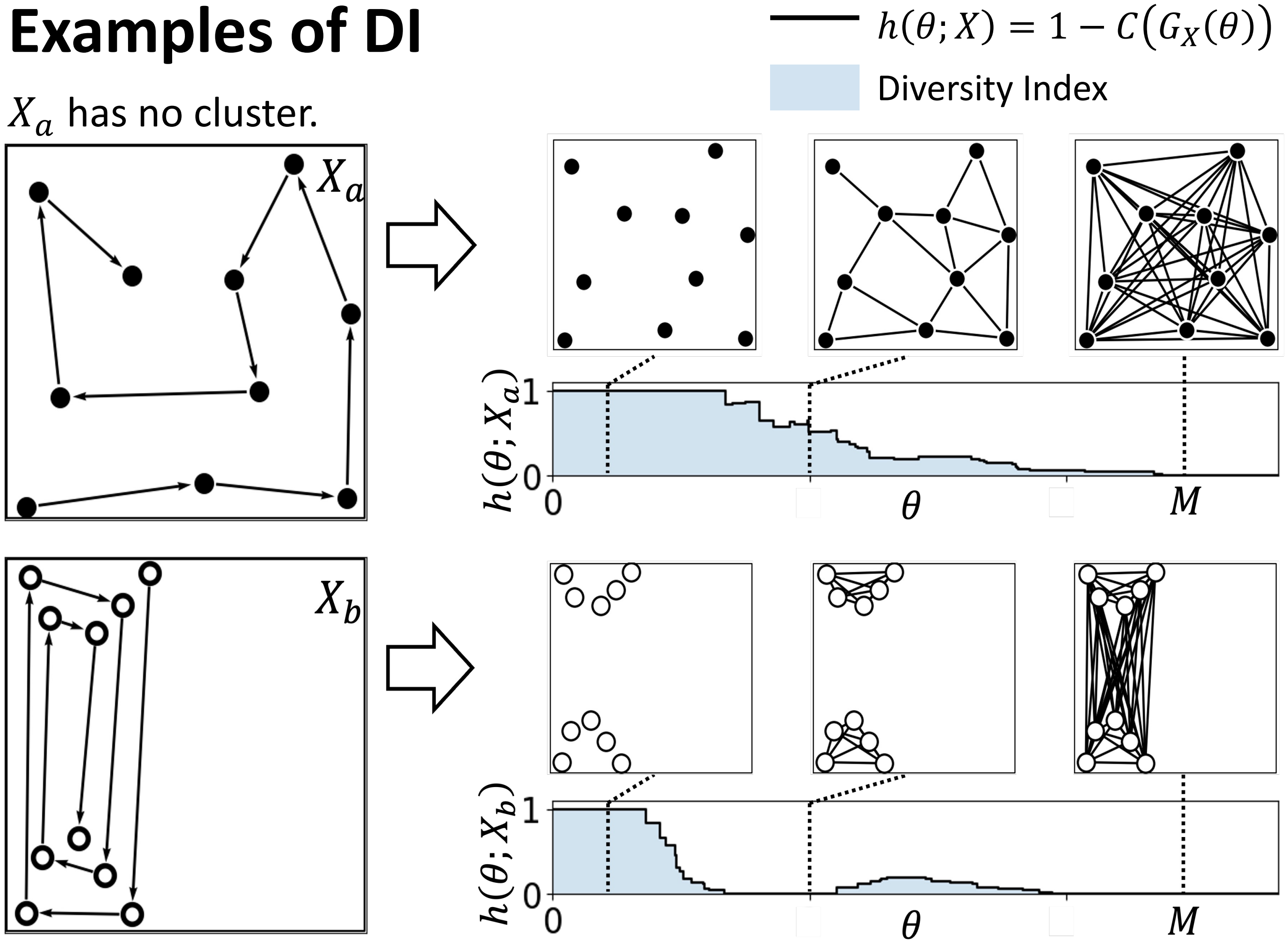
diversification

intensification

quantification

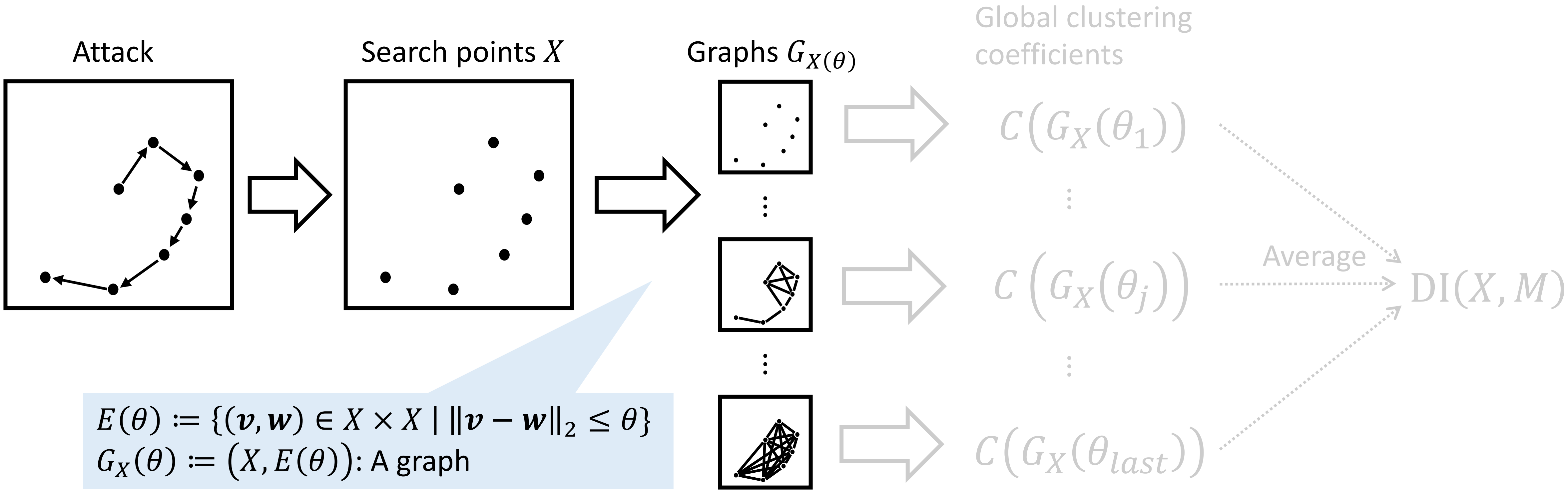
Examples of DI

X_a has no cluster.



X_b has two clusters.

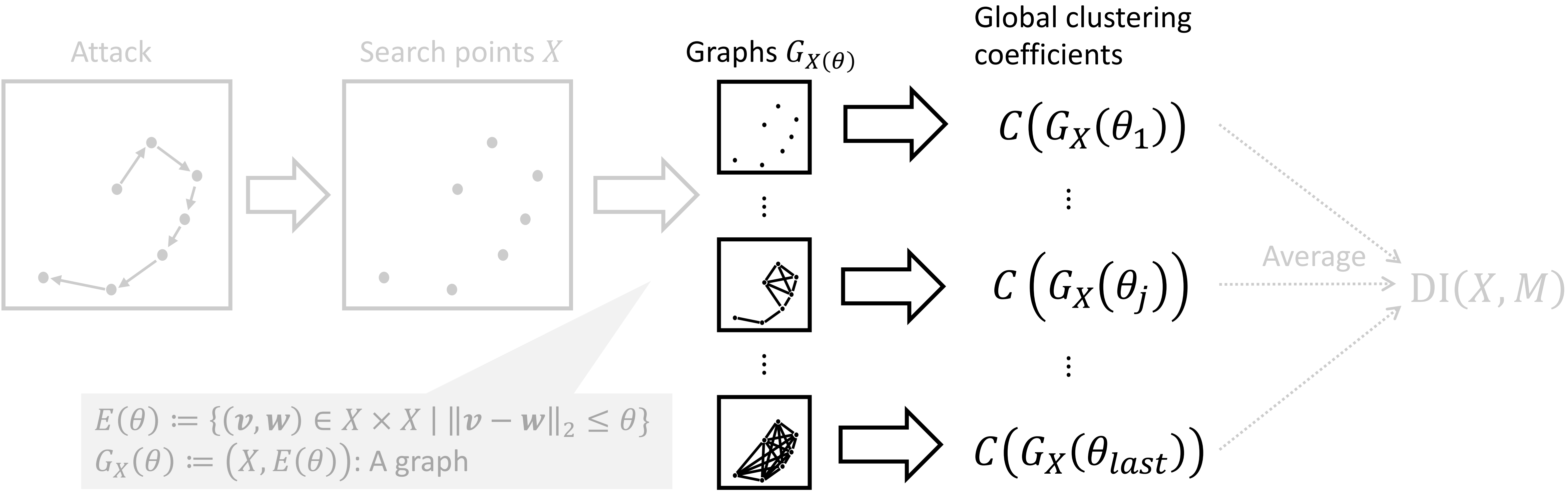
Section 5: Diversity Index



$E(\theta) \coloneqq \{(\boldsymbol{v}, \boldsymbol{w}) \in X \times X \mid \|\boldsymbol{v} - \boldsymbol{w}\|_2 \leq \theta\}$

$G_X(\theta) \coloneqq \left(X, E(\theta)\right)$: A graph

Section 5: Diversity Index

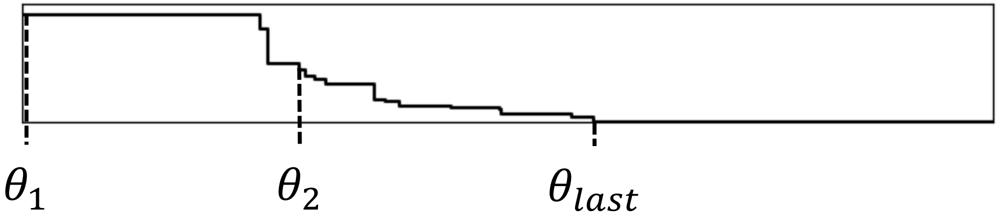


$E(\theta) \coloneqq \{(\boldsymbol{v}, \boldsymbol{w}) \in X \times X \mid \|\boldsymbol{v} - \boldsymbol{w}\|_2 \leq \theta\}$
$G_X(\theta) \coloneqq (X, E(\theta))$: A graph

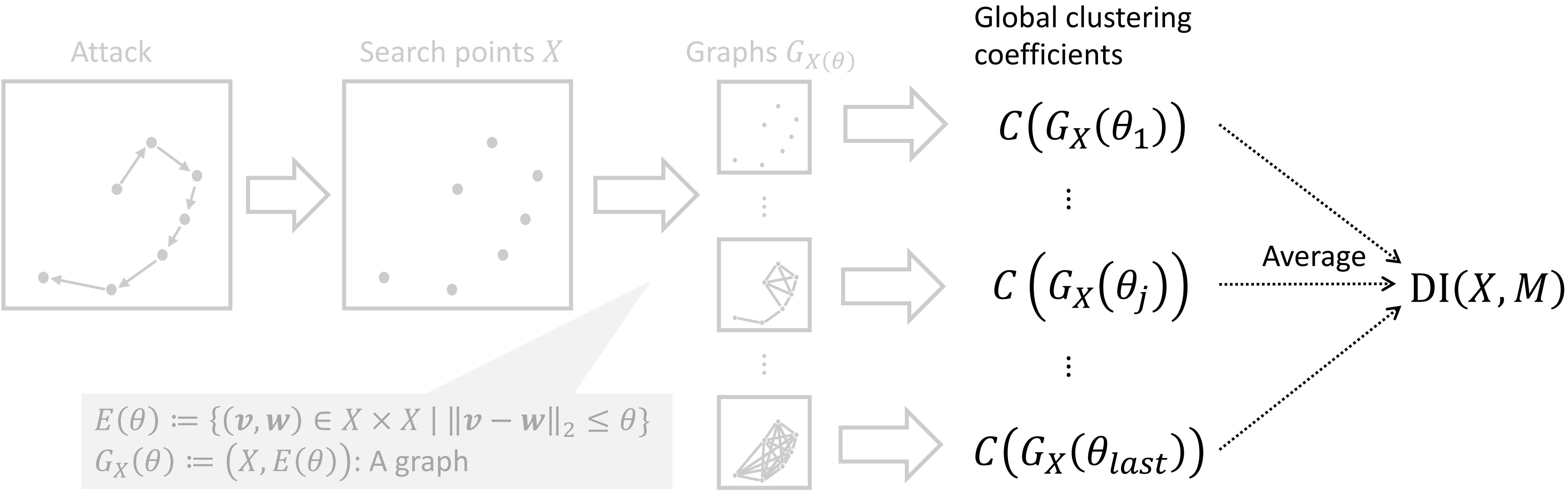
$$C(G_X(\theta)) = \mathrm{average}_{v \in X}(C_v),$$

$$C_v = \frac{2\times\#\{\, e_{uw} \in E(\theta) \mid u, w \in N_v \,\}}{\#N_v\times(\#N_v - 1)},$$

$$N_v = \{\, w \in X \mid e_{vw} \in E(\theta) \,\}$$



Section 5: Diversity Index


Attack
Search points X
Graphs $G_{X(\theta)}$
Global clustering coefficients
$C\left(G_X(\theta_1)\right)$
⋮
$C\left(G_X(\theta_j)\right)$
⋮
$C\left(G_X(\theta_{last})\right)$
Average
$\mathrm{DI}(X, M)$

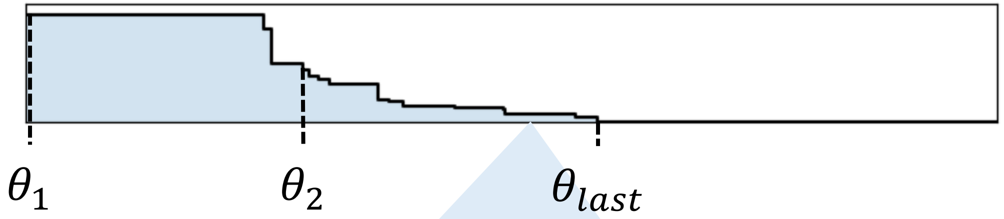
$E(\theta) \coloneqq \{(\boldsymbol{v}, \boldsymbol{w}) \in X \times X \mid \|\boldsymbol{v} - \boldsymbol{w}\|_2 \leq \theta\}$

$G_X(\theta) \coloneqq \left(X, E(\theta)\right)$: A graph

$$C\left(G_X(\theta)\right) = \mathrm{average}_{v \in X}(C_v),$$

$$C_v = \frac{2 \times \#\{\, e_{uw} \in E(\theta) \mid u, w \in N_v \,\}}{\#N_v \times (\#N_v - 1)},$$

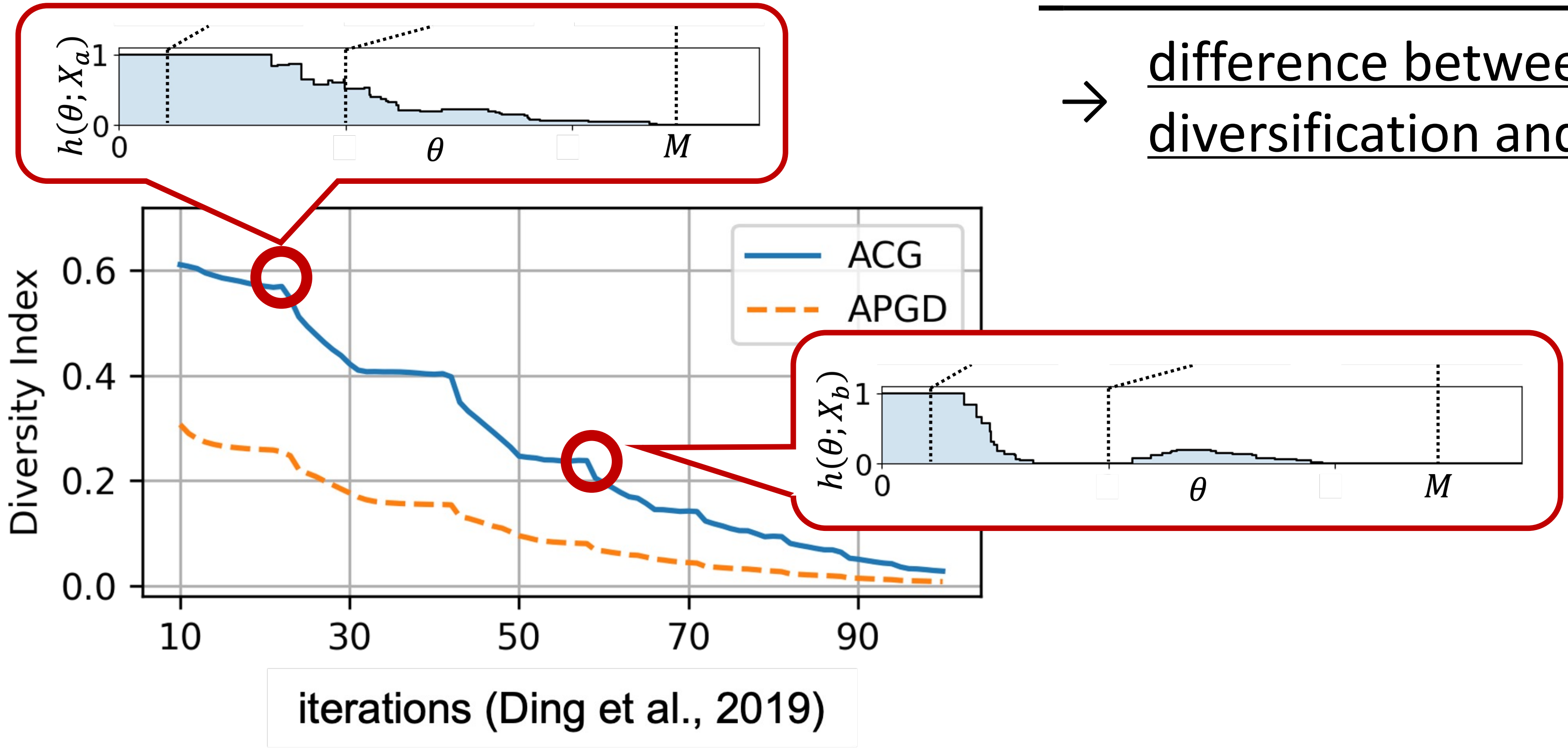
$$N_v = \{\, w \in X \mid e_{vw} \in E(\theta) \,\}$$



$$DI(X, M) = \frac{1}{M} \int_0^M (1 - C\left(G_X(\theta)\right) d\theta$$

Section 5: Diversity Index – Transition diagram

An example of the transition of DI.



Transition of DI for APGD and ACG

→ difference between the balance of diversification and intensification

Each data points correspond to $h(\theta; X)$.
We use 10 search points to compute DI.

Section 5: Diversity Index – Analysis of ACG

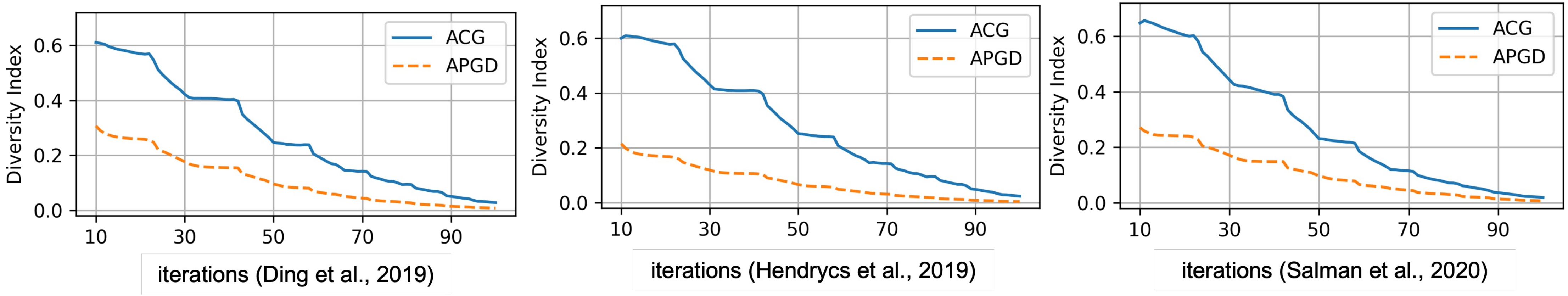


Figure: Transitions of Diversity Index on three models.

Both algorithms : diversification→ intensification

ACG : better diversification and enough intensification

APGD : better intensification

⇩

ACG’s **diversified search at early iterations** contributes to its higher ASR.

Conclusion

1. Propose a more diversified algorithm, Auto-Conjugate Gradient Attack (ACG)
   - ✓ Based on Conjugate Gradient Method
   - ✓ Higher ASR than APGD for 63 models out of 64
   - ✓ Higher ASR than APGD for 49 models out of 64 with fewer iterations and only deterministic operations

2. Propose Diversity Index (DI) which quantify the degree of diversification of the attacks
   - ✓ First attempt to quantify the degree of diversification of adversarial attacks
   - ✓ Based on the global clustering coefficients of the graph whose nodes are latest K search points
   - ✓ Capture the difference among algorithms by the transitions of DI

According to our thorough experiments, we believe that the diversified search of ACG contributes to achieving higher ASR than APGD.