Curvature Dynamic Black-box Attack

Revisiting Adversarial Robustness via Dynamic Curvature Estimation

Peiran Sun
Lanzhou University

Introduction

Adversarial attacks expose the vulnerability of deep networks. Most prior curvature studies rely on the Hessian of the loss or outputs, which reflects internal model curvature but not the geometry of the **decision boundary**—the object that directly governs adversarial robustness.

Decision boundary curvature is difficult to obtain in a hard-label black-box setting. **Dynamic Curvature Estimation (DCE)** provides a query-efficient way to estimate local curvature along the attack trajectory, revealing a clear relation between curvature and robustness and motivating a curvature-aware attack, CDBA.

Dynamic Curvature Estimation (DCE)

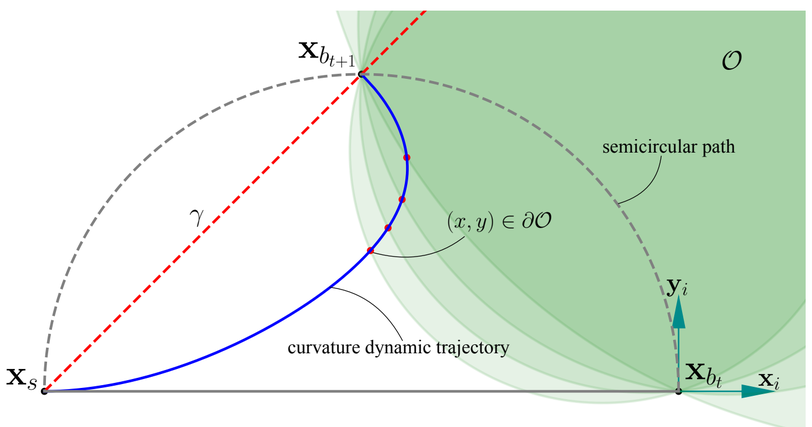
DCE works on the 2D plane spanned by $(x_t - x_s)$ and the estimated normal vector. While CGBA performs semicircular search, DCE models the boundary between iterations using a family of circles whose closest points form the **curvature dynamic trajectory**:

$$(\tan\gamma\cos\theta - \sin\theta)r^2 - r\tan\gamma + \sin\theta = 0.$$

Curvature is then inferred by:

$$\hat{\kappa} = \left[\left(\frac{0.5\tan\gamma}{\tan\gamma - \tan\theta} - 1\right)^2 + \left(\frac{0.5\tan\gamma\tan\theta}{\tan\gamma - \tan\theta}\right)^2\right]^{-1/2}.$$

DCE requires minimal additional queries and remains stable across datasets and attack types.

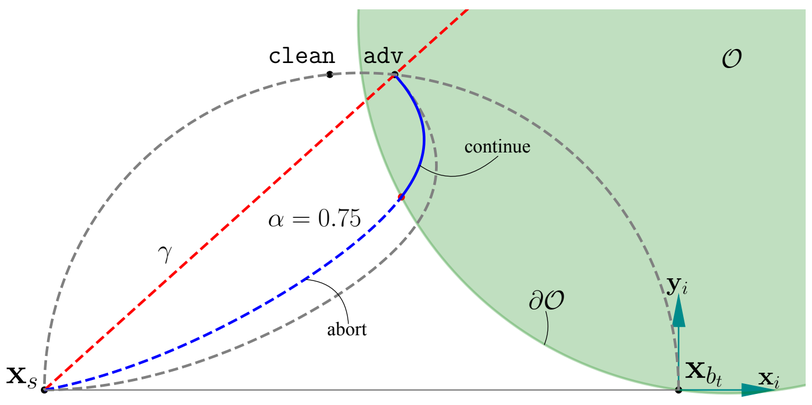


Curvature Dynamic Black-box Attack (CDBA)

CDBA integrates curvature-guided refinement into CGBA:

- **Abort rule** — skips curvature search when curvature is very small.
- **Step parameter** α — query at $(\rho^\star, \gamma - \alpha * (\gamma - \theta^\star))$, reduces local minima and stabilizes late iterations.

When curvature is high, the curvature trajectory accelerates descent; when curvature becomes small, α maintains robustness. CGBA-H appears as the special case α=0.



Results: CDBA Performance

We run CGBA, CGBA-H, CDBA ($\alpha = 1, 0.75$) separately for 15 iterations of targeted attack on standard ResNet-50 classifier to show the average ℓ_2-norm versus query number results in different attack methods. The least ℓ_2 perturbation is marked in bold.

Query	250	500	750	1000
CGBA	96.93	94.11	91.63	89.05
CGBA-H	82.78	74.13	68.60	63.91
CDBA($\alpha = 1$)	**81.72**	73.86	68.37	64.23
CDBA($\alpha = 0.75$)	82.25	**73.85**	**67.69**	**62.78**

CDBA($\alpha = 0.75$) achieves the least ℓ_2-norm perturbation under a limited query budget 1000. Moreover, standard CDBA outperforms other attacks during the initial descent, with relatively high curvature of decision boundary.

Results: Curvature vs Adversarial Robustness

For ℓ_p robust classifiers, we selected seven WideResNet-28-10 classifiers on CIFAR-10 including three ℓ_2 robust ones and four ℓ_∞ robust ones. For certified robust classifiers, we chose seven classifiers including four ResNet-101 and three ResNet-50 classifiers.

Instead of taking the average $\hat{\kappa}$ over the entire iteration, the results are first divided into different bins according to their ℓ_2-norm. We also take the logarithm of $\hat{\kappa}$ for better visualization.

Datasets	Structure	Metric	Classifier	$\overline{log(\hat{\kappa})}$				
CIFAR-10	WideResNet	-	**Standard**	**-0.63**	**0.18**	**1.17**	**1.58**	**1.43**
		ℓ_2	Rony et al. (2019)	-1.75	-1.64	-0.93	-0.97	-1.41
			Rebuffi et al. (2021)	**-3.32**	**-3.41**	**-3.18**	**-2.78**	**-2.39**
			Wang et al. (2023)	-3.12	-3.33	-2.65	-2.34	-1.37
		ℓ_∞	Hendrycks et al. (2019)	-1.90	-1.82	-1.10	-1.01	-0.74
			Sehwag et al. (2020)	-1.64	-1.48	-1.59	-1.46	-1.34
			Zhang et al. (2021)	-1.70	-1.55	-1.64	-1.37	-1.20
			Cui et al. (2021)	**-3.03**	**-2.57**	**-2.10**	**-1.74**	**-1.53**
	ResNet-101	-	**Standard**	**-0.26**	**0.44**	**0.75**	**1.08**	**1.54**
		Certified	$\sigma = 0.12$	-0.72	-0.26	-0.32	0.03	0.65
			$\sigma = 0.25$	-0.65	-0.73	-0.49	-0.15	-0.96
			$\sigma = 0.50$	-1.47	**-1.17**	-0.92	-1.36	**-1.35**
			$\sigma = 1.00$	**-2.32**	-1.10	**-1.18**	**-1.70**	-0.59
ImageNet	ResNet-50	-	**Standard**	**3.24**	**3.34**	**3.39**	**3.42**	**3.44**
		ℓ_∞	Wong et al. (2020)	0.59	1.24	1.93	1.93	2.14
			Engstrom et al. (2019)	0.80	1.60	1.79	2.36	2.36
			Salman et al. (2020)	**0.17**	**1.07**	**1.45**	**1.91**	**2.10**
		Certified	$\sigma = 0.25$	0.91	0.77	1.12	1.52	1.70
			$\sigma = 0.50$	0.34	**0.44**	0.73	1.33	1.04
			$\sigma = 1.00$	**0.20**	0.48	**0.19**	**0.01**	**-0.07**

The clean and robust accuracy of classifiers as reported in RobustBench:

	Classifier	Clean acc.	Robust acc.
CIFAR(ℓ_2)	Standard	94.78%	0
	Rony et al. (2019)	89.05%	66.44%
	Rebuffi et al. (2021)	91.79%	78.80%
	Wang et al. (2023)	95.16%	83.68%
CIFAR(ℓ_∞)	Standard	94.78%	0
	Hendrycks et al. (2019)	87.11%	54.92%
	Sehwag et al. (2020)	88.98%	57.14%
	Zhang et al. (2021)	89.36%	59.64%
	Cui et al. (2021)	92.16%	67.73%
ImageNet	Standard	76.52%	0
	Wong et al. (2020)	55.62%	26.24%
	Engstrom et al. (2019)	62.56%	29.22%
	Salman et al. (2020)	64.02%	34.96%

- **Standard models** show high curvature and irregular boundaries.
- **Adversarially robust models** (ℓ_2, ℓ_∞) show clearly **flatter boundaries**.
- **Certified (smoothed) models** also reduce curvature.

Curvature ordering aligns with robustness accuracy, especially at small perturbation radii, indicating that adversarial robustness manifests as local boundary flattening.

Results: Curvature vs Common Corruption Robustness

Common corruption robustness behaves differently:

- Curvature decreases slightly for robust models.
- Correlates weakly with corruption robustness.

Classifier	$\overline{log(\hat{\kappa})}$				
Standard	**3.24**	**3.34**	**3.39**	**3.42**	**3.44**
Geirhos et al. (2018)	2.75	2.99	3.08	3.05	3.20
Erichson et al. (2022)	2.71	2.85	3.03	3.17	3.29
Hendrycks et al. (2021)	2.87	2.73	2.68	2.77	2.85

This might be because the perturbation of common corruptions is generally larger, which doesn't necessarily result in flattened local boundary.

Conclusion

DCE provides the first practical, query-efficient estimation of decision boundary curvature in black-box attacks. Across diverse robustness settings, flatter boundaries consistently correspond to stronger robustness.

CDBA leverages this curvature information to achieve lower distortion and more reliable convergence, showing that decision-boundary geometry is a powerful signal for improving black-box adversarial attacks.