Finding Optimal Tangent Points for Reducing Distortions of Hard-label Attacks

Chen Ma[1], Xiangyu Guo[2], Li Chen[1],

Jun-Hai Yong[1], Yiseng Wang[3]

[1] School of Software, BNRist, Tsinghua University

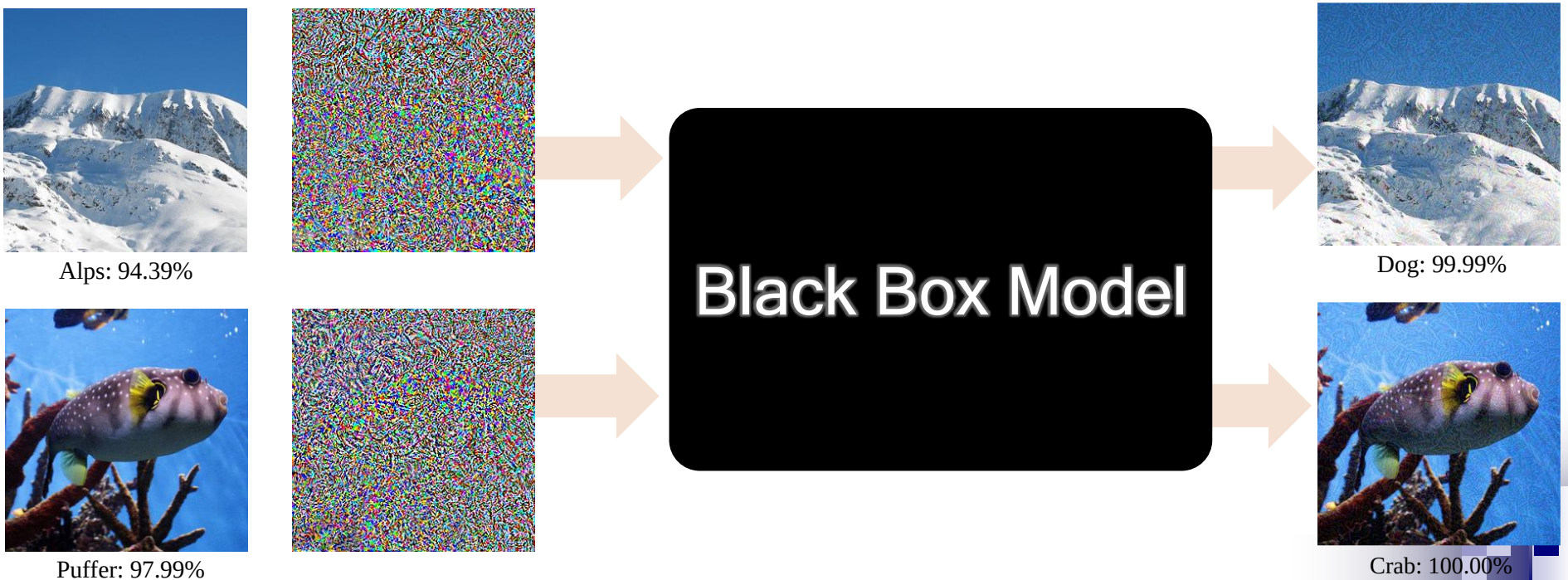
[2] Department of Computer Science and Engineering of SUNY Buffalo

[3] School of EECS, Peking University

[1] machenstar@163.com [2] xiangyug@buffalo.edu

Background

- An adversarial example should be visually indistinguishable from the corresponding normal one, but yet are misclassified by the target model.
- Adversarial attacks are the algorithm to find such examples. Hard-label attacks belong to black-box attacks.



Black-box Attacks

- Transfer-based
  - Generate adversarial examples against white-box models, and leverage transferability for attacks
  - Require no knowledge of the target model, no queries
  - Issue: require white-box surrogate models (datasets), it assumes this model and the target model are similar.
- Query-based
  - Get some information from the target model directly, through queries
    - Score-based
    - **Decision-based**
  - Goal: save queries and reduce the distortions of examples
  - In addition, our method does not need any surrogate model!

Hard-label Attacks

- Goal: Given classifier $f(x)$: $\mathbb{R}^d \rightarrow \mathbb{R}^K$ and input-label pair (x, y), the hard-label attack needs to generate the adversarial example x_{adv} with only the top-1 predicted label of the classifier:

$$\hat{y} = \text{argmax}_i f(x_{adv})_i \ , i \in \{1, \dots, K\}$$

Class	Prob
Dog:	0.9
Cat:	0.04
...	
Bird:	0.03

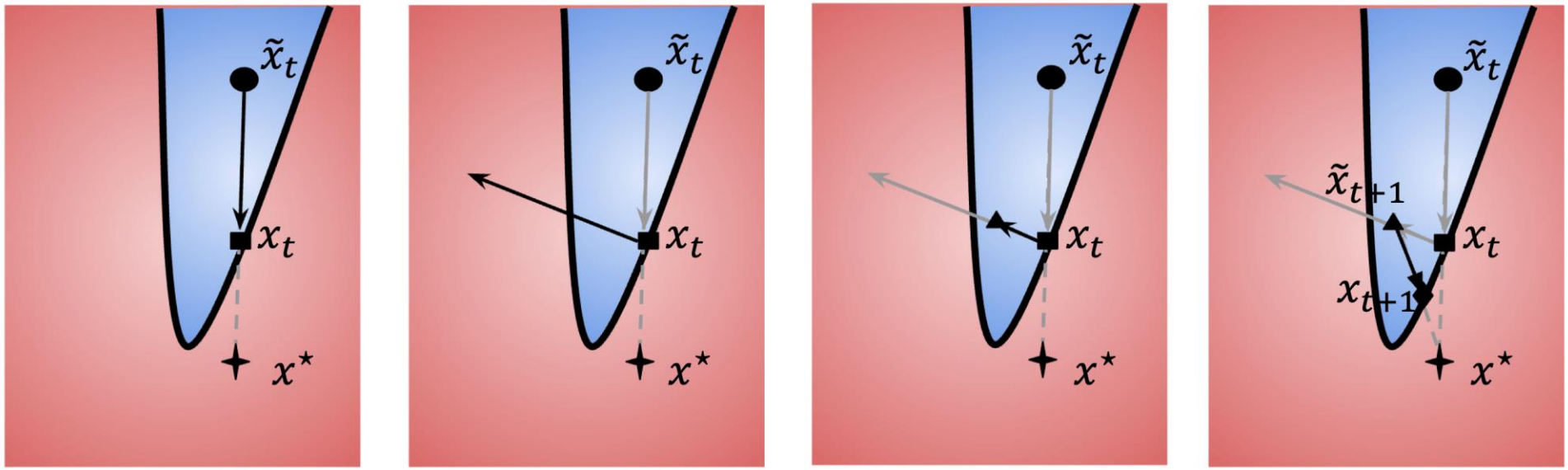
- x^{adv} can be generated by solving

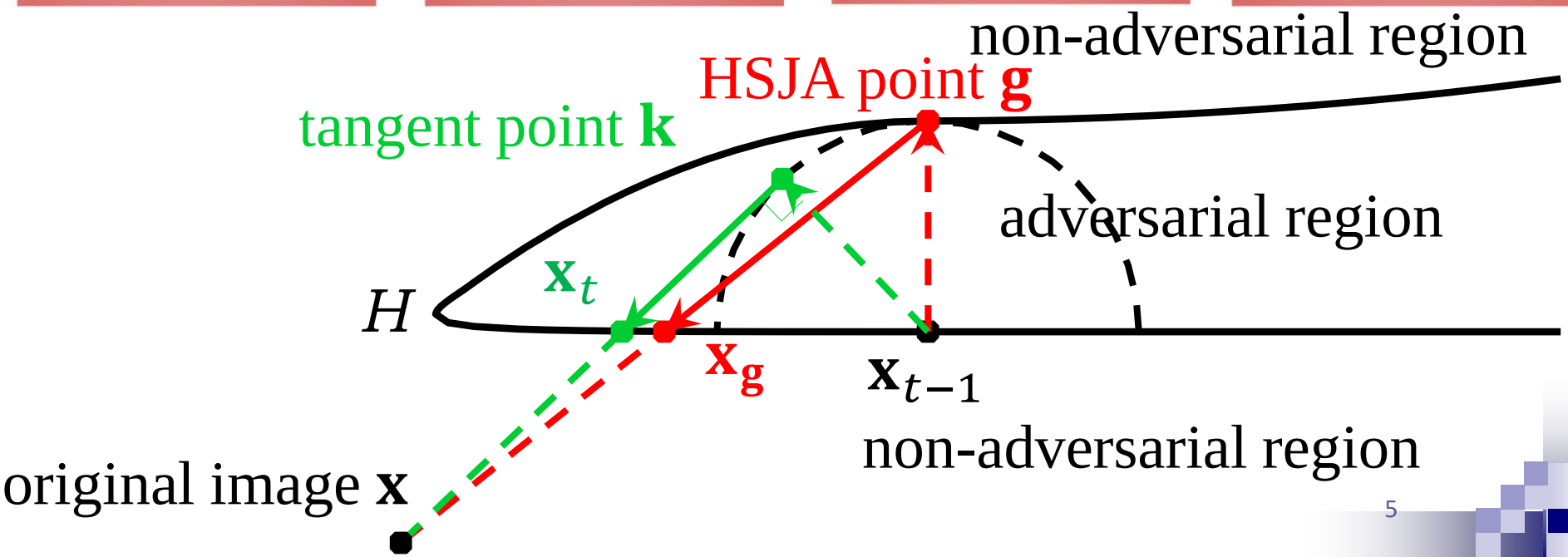
$$x^{\text{adv}} = \underset{x^{\text{adv}}}{\text{argmin}}\, d(x^{\text{adv}}, x) \quad \text{s.t.} \quad \phi(x_{\text{adv}}) = 1$$

where $\phi(x_{\text{adv}}) = \begin{cases} 1 \ if \ \hat{y} = y_{adv} \text{ in the targeted attack} \\ \text{or } \hat{y} \neq y \text{ in the untargeted attack} \\ 0 \ \text{ otherwise} \end{cases}$

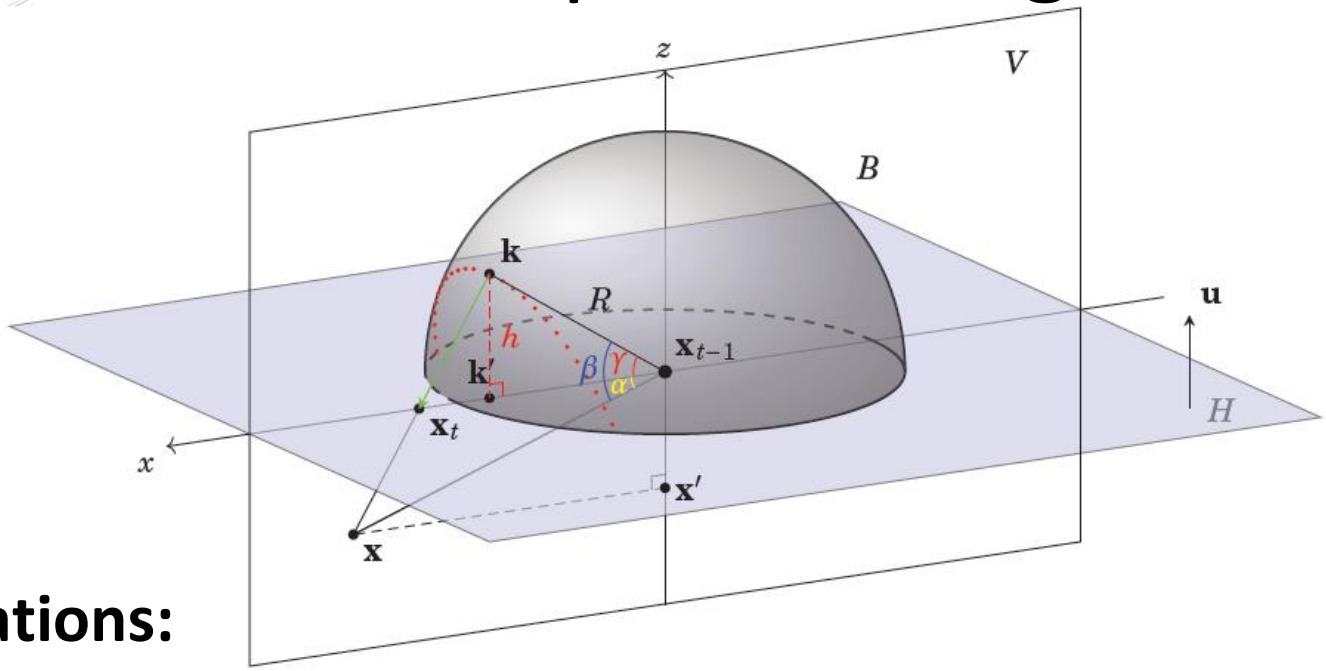
i.e., $\phi(x_{\text{adv}})$ indicates an successful attack.

A typical hard-label attack: HSJA



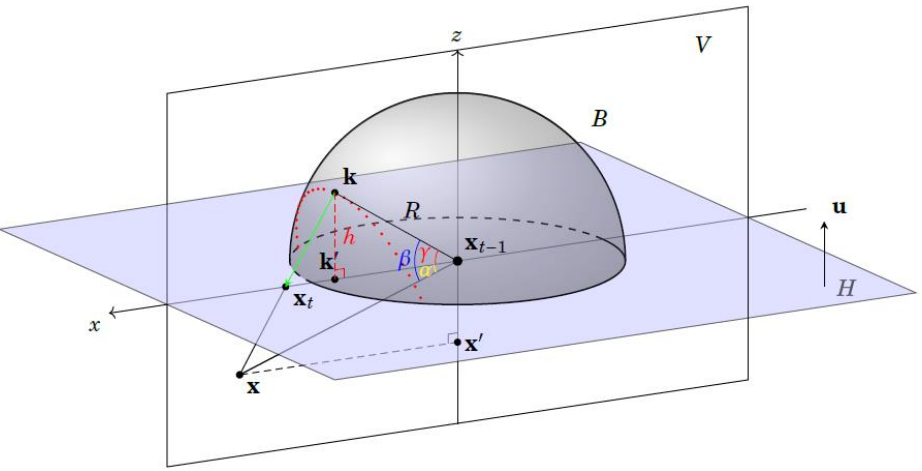


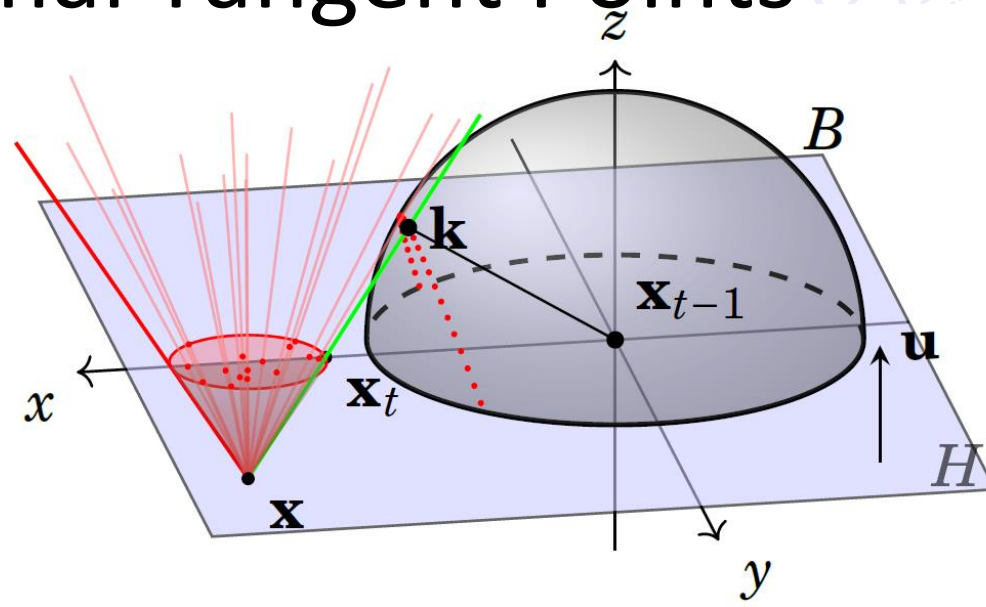
Definition of Optimal Tangent Points



- **Notations:**
- H denotes the decision hyperplane,
- **u** denotes the unit gradient vector,
- **k** denotes any point in the hemisphere B, which centered at $\mathbf{x}_{t-1}$
- R denotes the radius of hemisphere B,
- $\mathbf{x}_t$ denotes the intersection of H and the line passing through **x** and **k**.

Definition of Optimal Tangent Points





Theorem 1: Let H, $\mathbf{u}$, $\mathbf{x}$ and $\mathbf{x}_{t-1}$ be defined above, then the distance $||\mathbf{x} - \mathbf{x}_{t-1}||_2$ is the shortest if $\mathbf{k}$ is the optimal solution of the following constrained optimization problem:

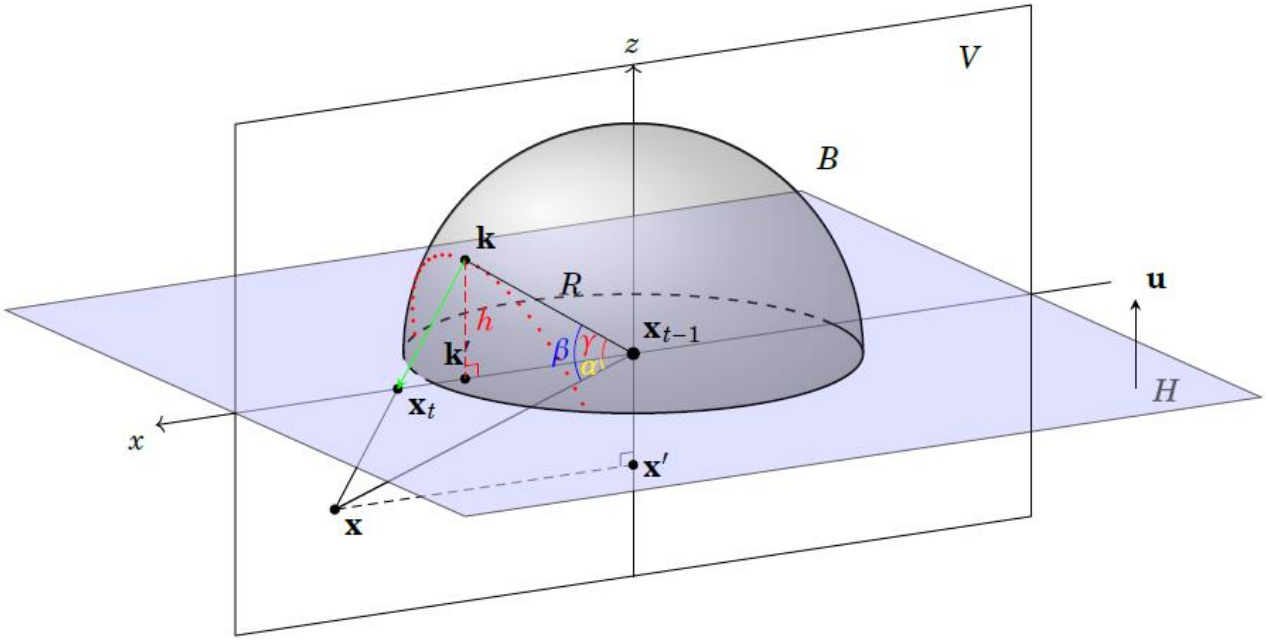
$$\text{argmax}_{\mathbf{k}} \langle \mathbf{k} - \mathbf{x}_{t-1}, \mathbf{u} \rangle$$

$$\text{s.t.} \quad \langle \mathbf{k} - \mathbf{x}_{t-1}, \mathbf{x} - \mathbf{k} \rangle = 0$$

$$||\mathbf{k} - \mathbf{x}_{t-1}||_2 = 0$$

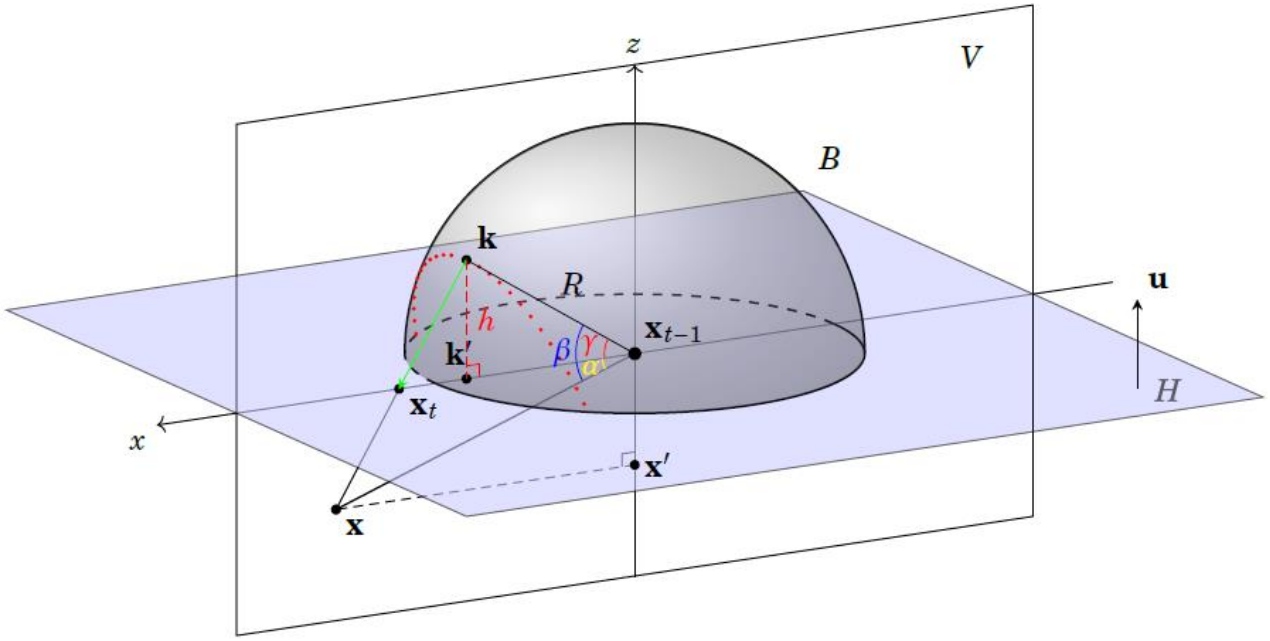
$$\langle \mathbf{k} - \mathbf{x}_{t-1}, \mathbf{u} \rangle > 0$$

Closed-form Solution of Optimal Tangent Points



- α, β, γ are on the same plane V, so $\beta = \alpha + \gamma$。
- the angle between vector **x** and **u** is $\frac{\pi}{2} + \alpha$, thus
- $\langle \mathbf{x}, \mathbf{u} \rangle = ||\mathbf{x}||_2 \cdot ||\mathbf{u}||_2 \cdot \cos\left(\frac{\pi}{2} + \alpha\right) = ||\mathbf{x}||_2 \cdot ||\mathbf{u}||_2 \cdot (-\sin\alpha)$
- Thus, $\sin\alpha = -\frac{\langle \mathbf{x}, \mathbf{u} \rangle}{||\mathbf{x}||_2 \cdot ||\mathbf{u}||_2}$

Closed-form Solution of Optimal Tangent Points

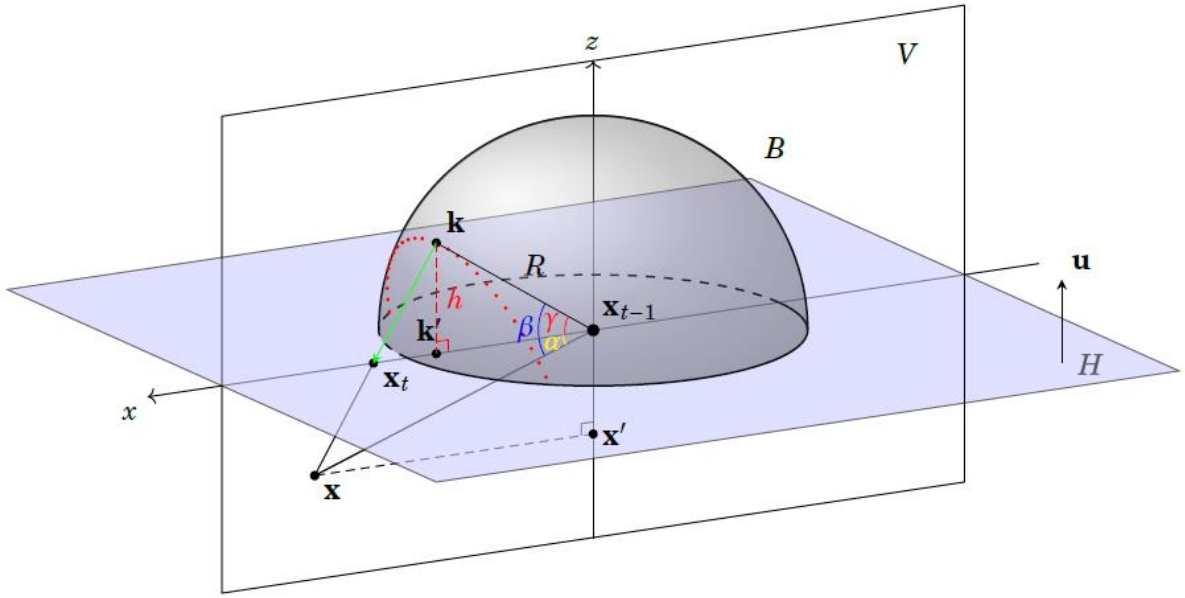


$$\cos\alpha = \sqrt{1-\sin^2\alpha} = \frac{\sqrt{||\mathbf{x}||_2^2 - (\langle \mathbf{x}, \mathbf{u} \rangle)^2}}{||\mathbf{x}||_2}$$

$$\cos\beta = \frac{R}{||\mathbf{x}||_2}$$

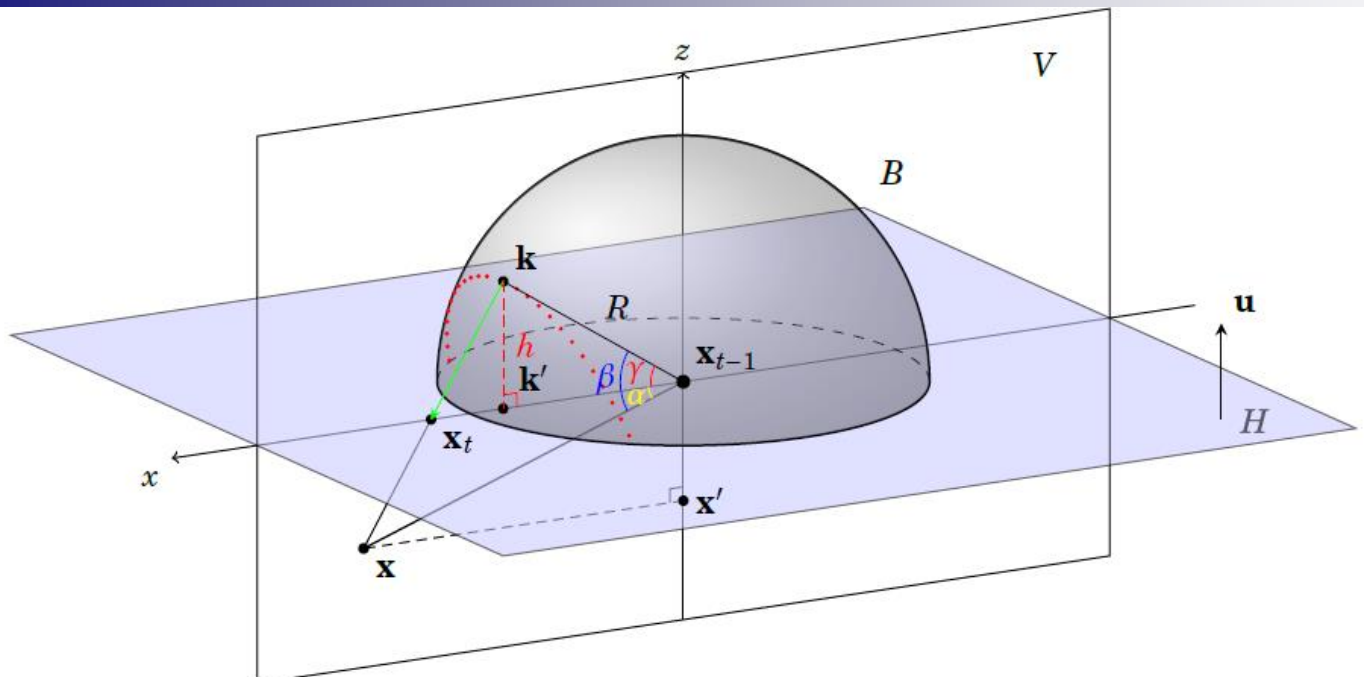
$$\sin\beta = \sqrt{1-\cos^2\beta} = \frac{\sqrt{||\mathbf{x}||_2^2 - R^2}}{||\mathbf{x}||_2}$$

Closed-form Solution of Optimal Tangent Points



- $\sin\gamma = \sin(\beta - \alpha) = \sin\beta\cos\alpha - \cos\beta\sin\alpha$
- $\cos\gamma = \cos(\beta - \alpha) = \cos\beta\cos\alpha + \sin\beta\sin\alpha$
- Make the project point $\mathbf{k}'$ of $\mathbf{k}$ onto the hyperplane H,
- and $||\mathbf{k} - \mathbf{k}'||_2 = h$.
- With $\sin\gamma$ and h, we have $h = R \cdot \sin\gamma = R \cdot (\sin\beta\cos\alpha - \cos\beta\sin\alpha$
- Make the project point $\mathbf{x}'$ of $\mathbf{x}$ onto the z-axis, then$\mathbf{x}' = \langle \mathbf{x}, \mathbf{u} \rangle \cdot \mathbf{u}$ we have $\frac{\mathbf{k}'}{||\mathbf{k}'||} = \frac{\mathbf{x}-\mathbf{x}'}{||\mathbf{x}-\mathbf{x}'||}$

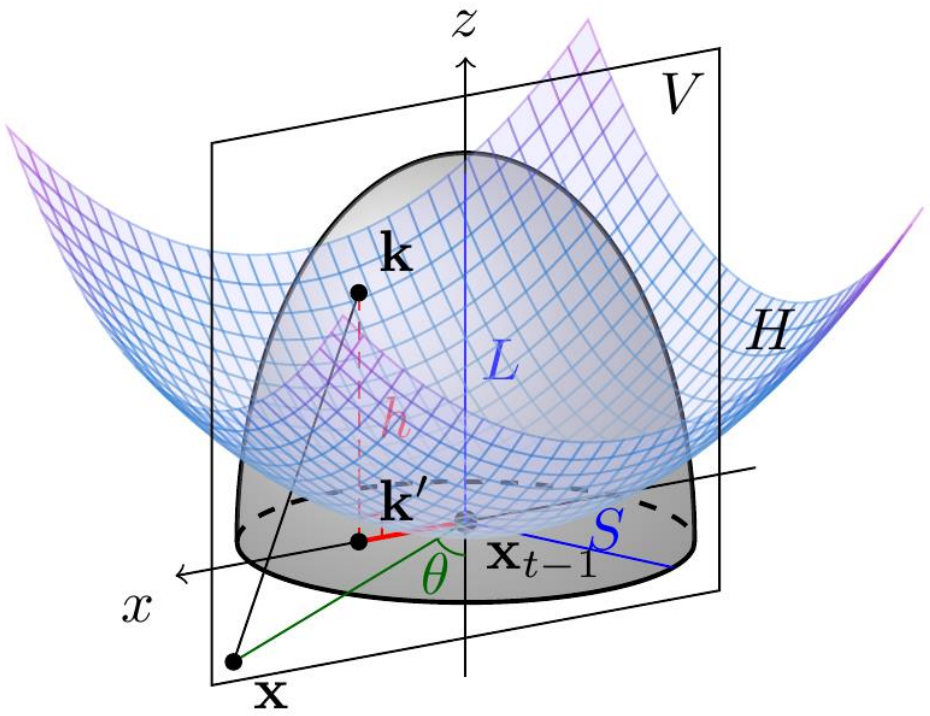
Closed-form Solution of Optimal Tangent Points



$$\frac{\mathbf{k}'}{||\mathbf{k}'||_2} = \frac{\mathbf{x}-\mathbf{x}'}{||\mathbf{x}-\mathbf{x}'||_2}$$

- Substitute $||\mathbf{k}'||_2 = R \cdot \cos\gamma$ and $\mathbf{x}' = \langle \mathbf{x}, (-\mathbf{u}) \rangle \cdot (-\mathbf{u}) = \langle \mathbf{x}, \mathbf{u} \rangle \cdot \mathbf{u}$
- We have $$\mathbf{k}' = \frac{\mathbf{x}-\mathbf{x}'}{||\mathbf{x}-\mathbf{x}'||_2} \cdot R \cdot \cos\gamma = \frac{\mathbf{x}-\langle \mathbf{x}, \mathbf{u} \rangle \cdot \mathbf{u}}{||\mathbf{x}-\langle \mathbf{x}, \mathbf{u} \rangle \cdot \mathbf{u}||_2} \cdot R \cdot \cos\gamma$$
- Finally $$\mathbf{k} = \mathbf{k}' + h \cdot \mathbf{u} = \frac{\mathbf{x}-\langle \mathbf{x}, \mathbf{u} \rangle \cdot \mathbf{u}}{||\mathbf{x}-\langle \mathbf{x}, \mathbf{u} \rangle \cdot \mathbf{u}||_2} \cdot R \cdot \cos\gamma + h \cdot \mathbf{u}$$

Generalized Tangent Attack



$$x_k = \frac{S^2\left(L^2 - z_0 \cdot \frac{L^2S^2z_0 + L^2x_0\sqrt{-L^2S^2 + L^2x_0^2 + S^2z_0^2}}{L^2x_0^2 + S^2z_0^2}\right)}{L^2 \cdot x_0}$$

$$z_k = \frac{L^2S^2z_0 + L^2x_0\sqrt{-L^2S^2 + L^2x_0^2 + S^2z_0^2}}{L^2x_0^2 + S^2z_0^2}$$

Then, $\mathbf{k} = |x_k| \cdot \mathbf{v} + z_k\mathbf{u}$

The complete algorithm

Algorithm 1 Tangent Attack

Input: benign image $\mathbf{x}$, attack success indicator function $\phi(\cdot)$ defined in Eq. (1), initial batch size B_0, iteration T, mode $m \in$ {semi-ellipsoid, hemisphere}, radius ratio r.
Initialize $\tilde{\mathbf{x}}_0$ that satisfies $\phi(\tilde{\mathbf{x}}_0) = 1$.
$\mathbf{x}_0 \leftarrow \text{BinarySearch}(\tilde{\mathbf{x}}_0, \mathbf{x}, \phi)$. {boundary search}
$d_0 = ||\mathbf{x}_0 - \mathbf{x}||_2$.
for t **in** $1, 2, \ldots, T-1$ **do**
  Sample $B_t \leftarrow B_0\sqrt{t}$ vectors to estimate gradient $\mathbf{u}$.
  Initialize $R \leftarrow d_{t-1}/\sqrt{t}$. {the initial radius}
  while true do
    Compute $\mathbf{k}$ based on Eq (11) **if** $m =$ hemisphere **else** Eq (12)
    $R \leftarrow \frac{R}{2}$ {search the radius, and we set $L = R, S = \frac{L}{r}$ if $m =$ semi-ellipsoid}
    if $\phi(\mathbf{k}) = 1$ **then**
      break.
    end if
  end while
  $\mathbf{k} \leftarrow \text{Clip}(\mathbf{k}, 0, 1)$
  $\mathbf{x}_t \leftarrow \text{BinarySearch}(\mathbf{k}, \mathbf{x}, \phi)$. {boundary search}
  $d_t = ||\mathbf{x}_t - \mathbf{x}||_2$.
end for

Experiment Setting

- **Dataset**
- CIFAR-10: 1000 images with the 32×32 resolution
- ImageNet:1000 images with 299×299 , 224×224 resolutions,

- **Target Model**
- CIFAR-10: PyramidNet-272, GDAS, WRN-28, WRN-40
- ImageNet: Inception-v3, Inception-v4, SENet-154, ResNet-101

- **Defensive Model:**
- Adversarial Training (AT), TRADES, JPEG, Feature Distillation

Experimental Results: ImageNet

Table 1: Mean ℓ_2 distortions of different query budgets on ImageNet dataset, where $r = 1.1$.

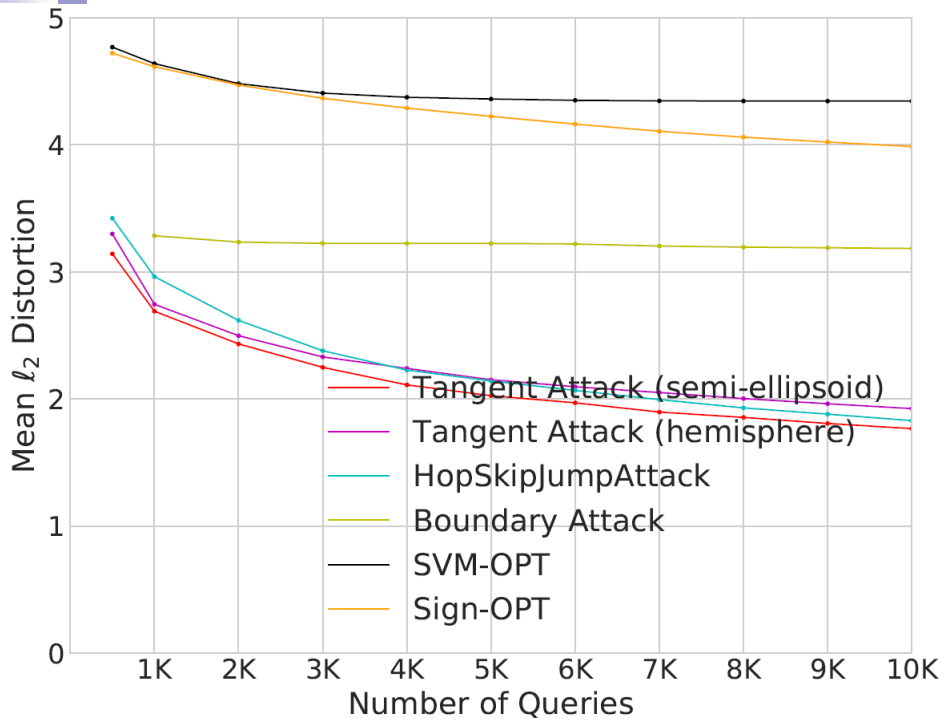
Target Model	Method	Targeted Attack						Untargeted Attack					
		@300	@1K	@2K	@5K	@8K	@10K	@300	@1K	@2K	@5K	@8K	@10K
Inception-v3	BA	111.798	108.044	106.283	102.715	86.931	78.326	-	107.558	102.309	95.776	78.668	60.296
	Sign-OPT	103.939	87.706	71.291	46.744	34.640	29.414	121.085	79.158	43.642	16.625	10.557	8.680
	SVM-OPT	**101.630**	82.950	67.965	46.275	35.694	31.106	121.135	66.027	36.763	15.736	10.501	8.789
	HSJA	111.562	95.295	82.111	52.544	37.395	30.425	103.605	57.295	37.185	15.484	9.989	7.967
	Ours (hemisphere)	103.781	**80.327**	**66.708**	**42.121**	**30.846**	**25.566**	94.752	52.523	35.229	15.040	9.748	7.793
	Ours (ellipsoid)	103.724	81.089	67.168	42.434	31.011	25.587	**94.668**	**52.037**	**34.540**	**14.643**	**9.540**	**7.618**
Inception-v4	BA	110.343	106.616	104.586	100.321	84.058	75.507	-	116.075	111.474	104.451	86.572	66.283
	Sign-OPT	101.620	85.731	69.719	46.416	34.957	30.004	132.991	86.431	48.292	18.678	11.567	9.262
	SVM-OPT	**99.856**	81.342	66.982	45.667	35.477	31.152	132.227	72.920	41.095	17.611	11.418	9.372
	HSJA	109.670	93.916	80.937	52.358	37.773	30.958	110.727	63.731	42.290	17.936	11.367	8.911
	Ours (hemisphere)	101.666	**78.683**	**65.304**	**41.629**	30.993	25.958	101.207	58.616	40.314	17.639	11.304	8.907
	Ours (ellipsoid)	101.495	79.210	65.888	42.002	**30.965**	**25.847**	**101.173**	**58.225**	**39.788**	**17.265**	**11.008**	**8.677**
SENet-154	BA	81.090	77.723	76.122	71.967	55.953	47.652	-	75.998	71.671	66.983	53.917	40.725
	Sign-OPT	75.722	62.876	49.191	30.155	21.333	17.672	70.035	47.705	27.314	10.890	6.643	5.245
	SVM-OPT	74.658	58.677	46.827	30.264	22.461	19.186	69.854	40.291	23.692	10.494	6.666	5.409
	HSJA	77.035	63.488	51.802	30.138	19.680	16.261	71.248	38.035	24.895	10.218	5.855	4.842
	Ours (hemisphere)	70.739	55.256	**43.694**	**24.961**	**16.756**	**13.876**	**65.589**	**35.689**	24.037	10.039	5.774	4.766
	Ours (ellipsoid)	**70.591**	**55.224**	44.047	25.041	16.854	14.047	65.871	35.768	**23.954**	**9.959**	**5.733**	**4.734**
ResNet-101	BA	81.565	77.903	76.366	72.392	58.746	51.679	-	64.007	60.389	56.544	44.175	31.371
	Sign-OPT	76.732	63.939	51.231	32.439	23.160	19.248	56.244	38.282	21.985	10.048	7.050	6.050
	SVM-OPT	77.031	61.417	49.842	32.806	24.553	20.964	55.894	32.638	19.409	9.830	7.185	6.281
	HSJA	76.121	63.091	52.301	31.018	20.472	16.911	56.264	27.443	17.717	7.649	4.723	**4.019**
	Ours (hemisphere)	**72.434**	**57.969**	**47.142**	**27.699**	**18.788**	**15.414**	53.197	26.777	17.651	7.730	4.822	4.107
	Ours (ellipsoid)	72.459	58.320	47.297	27.905	19.045	15.633	**53.058**	**26.631**	**17.384**	**7.602**	**4.720**	4.026

Experimental Results: CIFAR-10

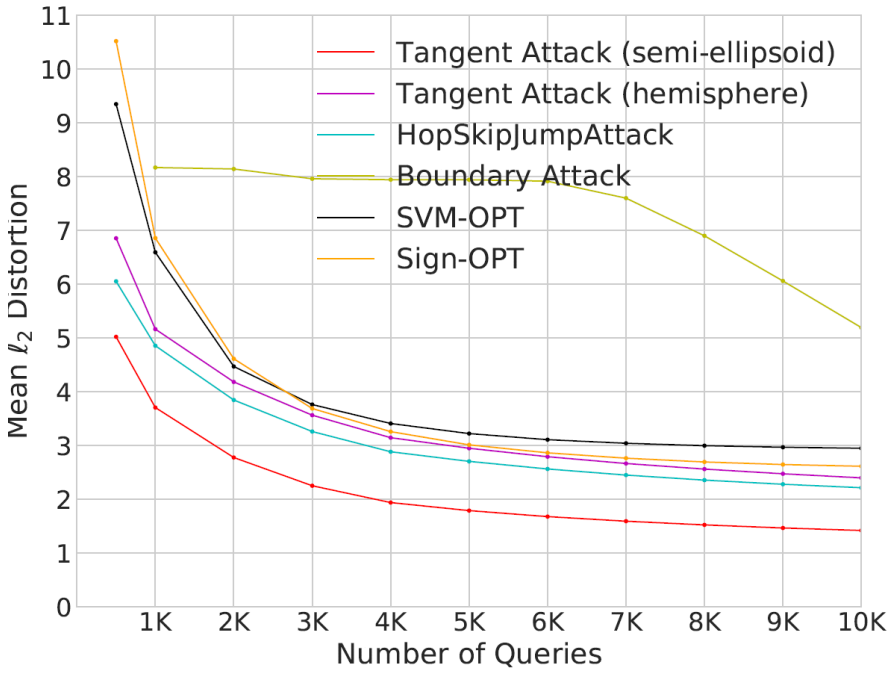
Table 2: Mean ℓ_2 distortions with different query budgets on CIFAR-10 dataset, where $r = 1.5$.

Target Model	Method	Targeted Attack						Untargeted Attack					
-		@300	@1K	@2K	@5K	@8K	@10K	@300	@1K	@2K	@5K	@8K	@10K
PyramidNet-272	BA	8.651	8.073	8.013	6.387	4.189	3.333	-	5.636	4.725	4.414	2.750	1.696
	Sign-OPT	8.279	6.331	4.250	1.718	0.960	0.718	4.387	2.334	1.178	0.403	0.267	0.226
	SVM-OPT	9.207	6.801	4.530	2.010	1.207	0.947	4.481	2.318	1.093	0.414	0.276	0.236
	HSJA	7.917	4.329	2.523	**0.793**	**0.489**	**0.397**	4.505	1.279	0.713	0.333	0.255	0.227
	Ours (hemisphere)	7.943	**4.267**	2.488	0.809	0.503	0.406	**4.256**	1.275	0.710	0.329	0.253	0.226
	Ours (ellipsoid)	**7.816**	4.277	**2.469**	0.803	0.505	0.412	4.432	**1.270**	**0.702**	**0.329**	**0.252**	**0.225**
GDAS	BA	8.487	7.885	7.821	6.034	3.632	2.703	-	2.717	2.514	2.373	1.642	1.106
	Sign-OPT	8.372	6.514	4.351	1.827	0.987	0.711	4.917	4.159	3.260	1.352	0.452	0.250
	SVM-OPT	9.529	7.243	5.092	2.347	1.317	0.958	4.909	3.950	2.736	1.082	0.371	0.234
	HSJA	7.714	3.566	1.966	0.591	0.365	0.301	2.188	0.756	0.483	0.261	0.208	0.189
	Ours (hemisphere)	**7.674**	**3.529**	**1.946**	0.585	0.366	0.302	2.190	0.774	0.485	0.257	0.206	0.187
	Ours (ellipsoid)	7.697	3.558	1.959	**0.583**	**0.361**	**0.298**	**2.161**	**0.745**	**0.476**	**0.255**	**0.204**	**0.185**
WRN-28	BA	8.688	8.046	7.984	5.786	2.486	1.555	-	4.425	3.648	3.435	1.543	0.832
	Sign-OPT	8.258	5.576	3.260	1.087	0.593	0.459	3.093	1.494	0.828	0.319	**0.239**	**0.213**
	SVM-OPT	9.516	5.968	3.744	1.367	0.728	0.553	2.977	1.466	0.723	0.325	0.245	0.221
	HSJA	6.810	2.603	1.326	0.518	0.389	0.347	3.052	0.797	0.508	0.299	0.250	0.232
	Ours (hemisphere)	6.802	2.556	1.311	0.519	0.394	0.353	**2.974**	0.785	**0.496**	**0.293**	0.249	0.233
	Ours (ellipsoid)	**6.755**	**2.543**	**1.281**	**0.513**	**0.387**	**0.345**	2.995	**0.782**	0.502	0.298	0.250	0.232
WRN-40	BA	8.658	8.014	7.953	5.738	2.484	1.566	-	4.377	3.586	3.367	1.487	0.821
	Sign-OPT	8.156	5.579	3.300	1.186	0.646	0.501	4.754	3.239	1.885	0.311	**0.226**	**0.201**
	SVM-OPT	9.339	6.061	3.840	1.445	0.800	0.605	4.457	2.756	0.739	0.310	0.229	0.206
	HSJA	6.909	2.648	1.330	0.528	0.400	**0.357**	2.992	0.777	0.498	0.290	0.242	0.225
	Ours (hemisphere)	6.944	**2.579**	**1.295**	**0.523**	**0.398**	0.358	**2.926**	**0.770**	**0.490**	0.288	0.243	0.227
	Ours (ellipsoid)	**6.783**	2.605	1.320	0.535	0.403	0.361	2.952	0.772	0.492	**0.288**	0.241	0.223

Experimental Results: defensive model



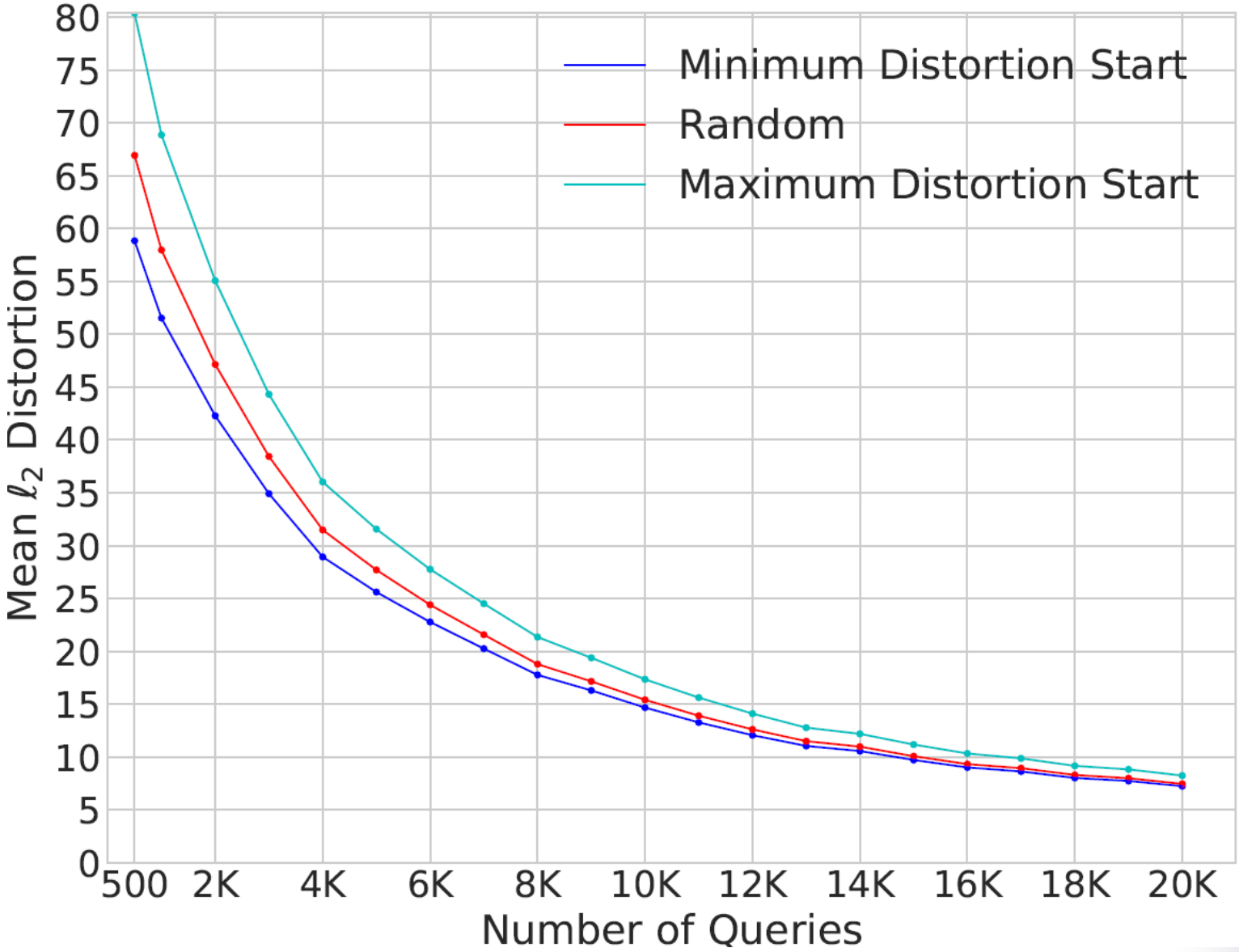
JPEG



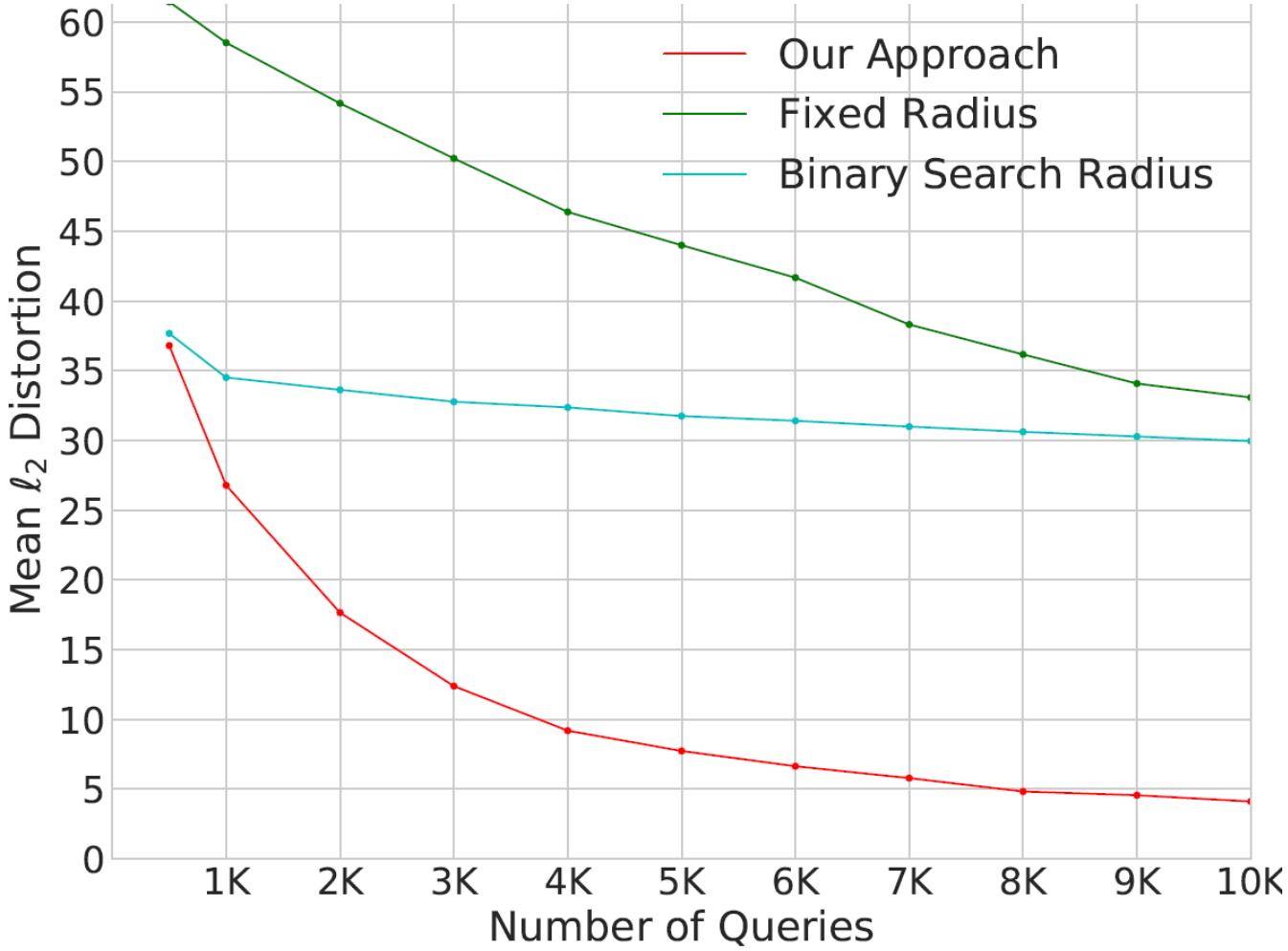
TRADES

Dataset :CIFAR-10

Ablation study: initialization

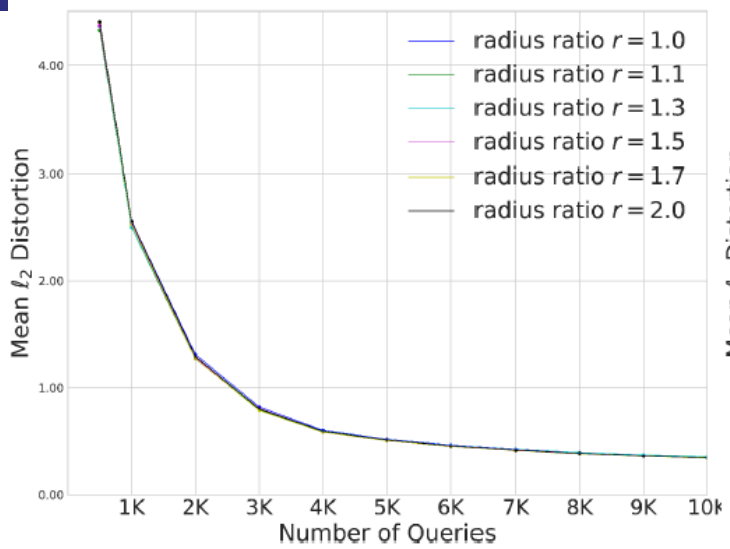


Ablation study: how to determining radius

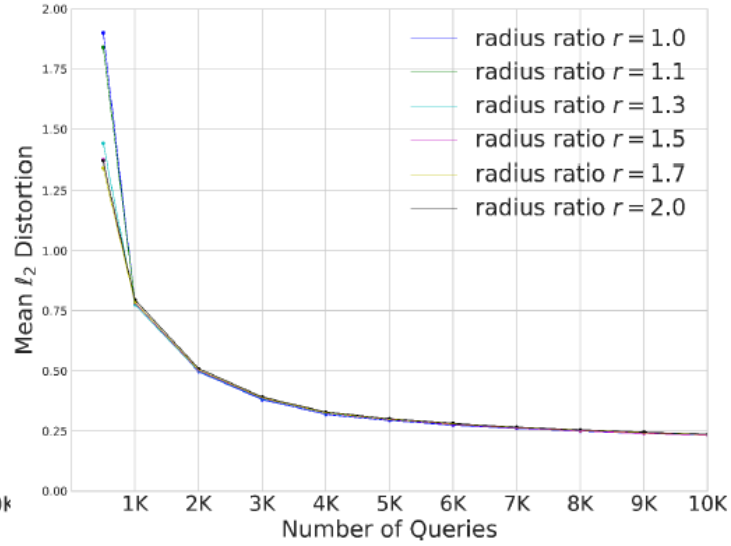


- Radius schedule

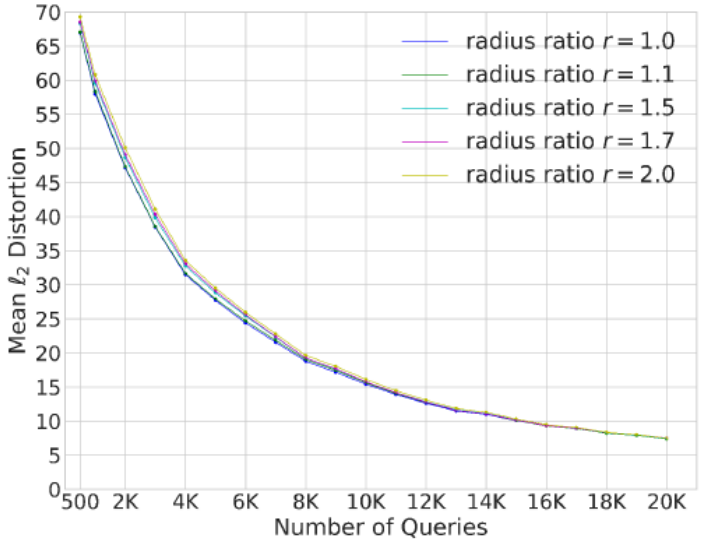
Ablation Study: radius ratio of G-TA



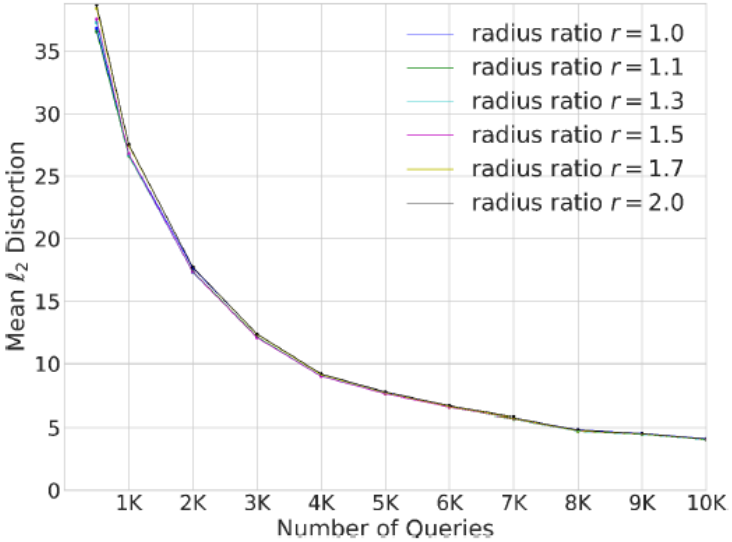
Targeted Attack WRN-28 on CIFAR-10



Untargeted Attack WRN-28 on CIFAR-10



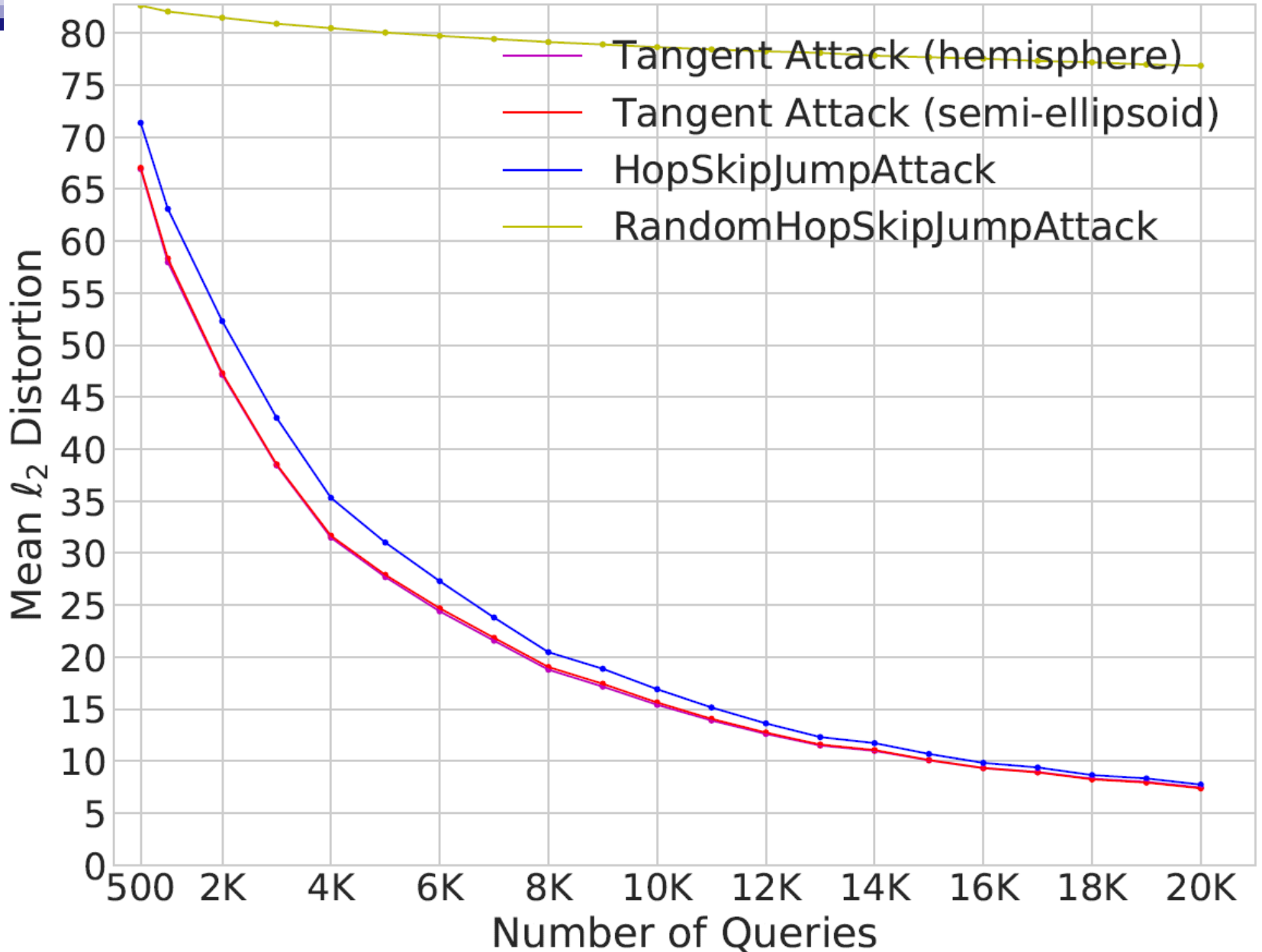
Targeted Attack ResNet-101 on ImageNet



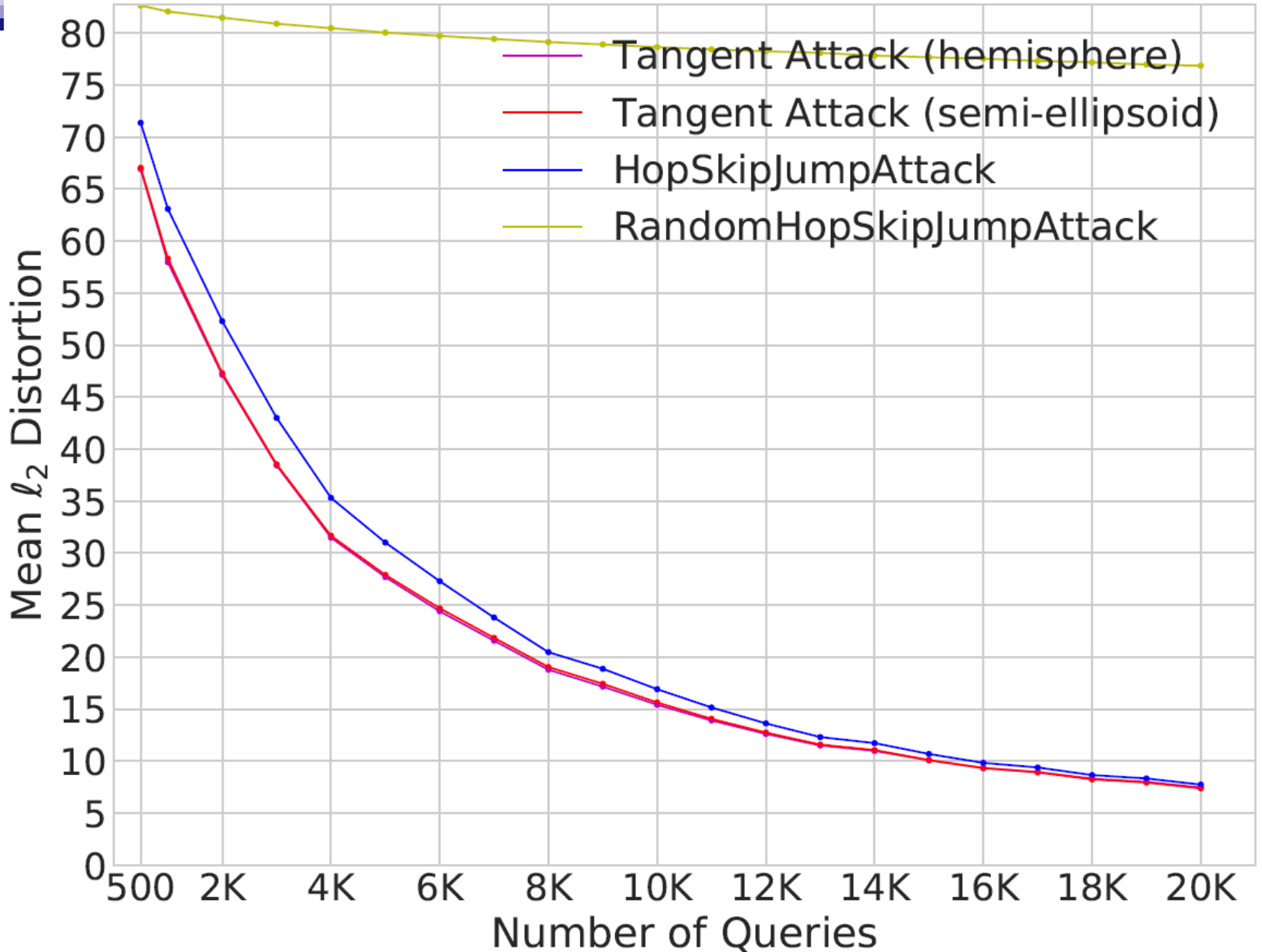
Untargeted Attack ResNet-101 on ImageNet

The ablation study of radius ratio r on CIFAR-10 and ImageNet datasets

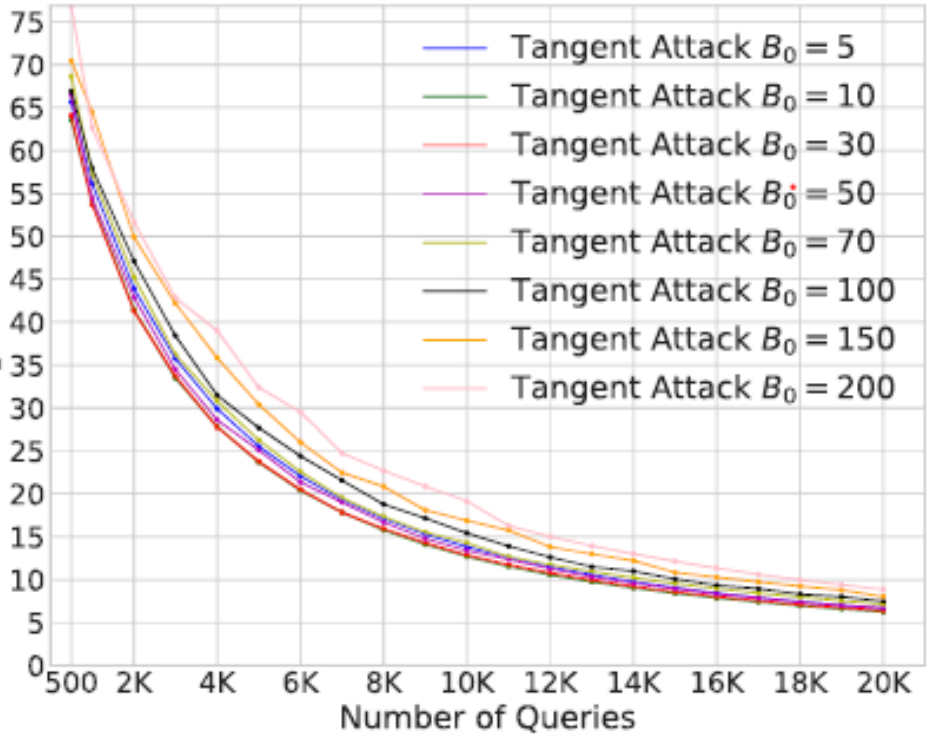
Ablation Study: jump direction



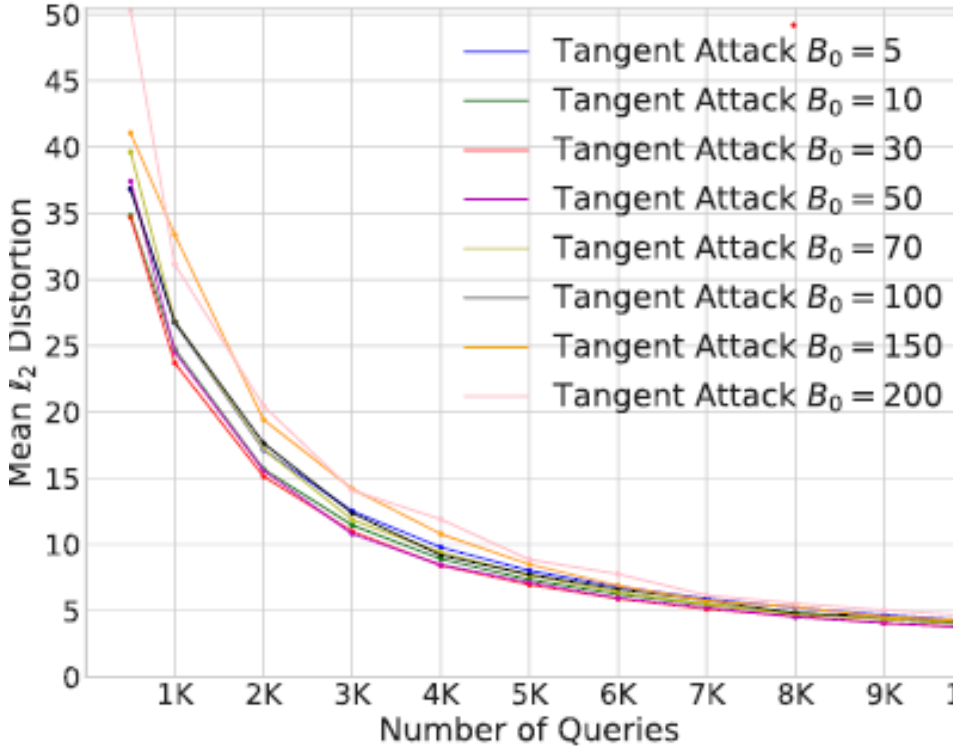
Ablation Study: jump direction



Initial batch size for estimating gradient



Targeted Attack ResNet-101 on ImageNet dataset



Untargeted Attack ResNet-101 on ImageNet datase

Thanks for your listening!