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

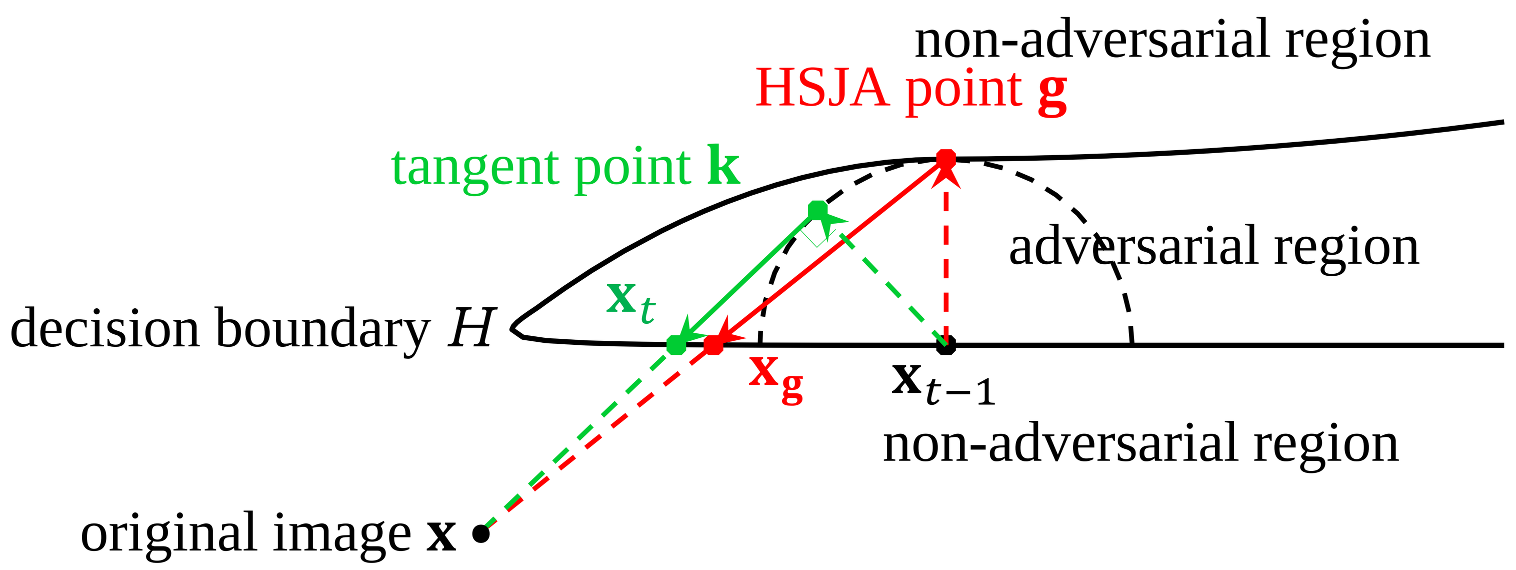
Chen Ma [1], Xiangyu Guo [2], Li Chen [1], Jun-Hai Yong [1], Yisen Wang [3]

[1] School of Software, BNRist, Tsinghua University, Beijing, China

[2] Department of Computer Science and Engineering, University at Buffalo, Buffalo NY, USA

[3] Key Lab. of Machine Perception, School of EECS, Peking University, Beijing, China

Introduction



The hard-label attack is a challenging adversarial attack, in which only the top-1 predicted label is available. Existing hard-label attacks do not thoroughly investigate the geometric properties of the decision boundary to accelerate the attack. Let us take HSJA (HopSkipJumpAttack) for example. As shown in the above figure, $\mathbf{x}_{t-1}$ is the current adversarial example mapped onto the decision boundary at the $(t-1)$-th iteration of the attack. HSJA updates $\mathbf{x}_{t-1}$ along the gradient direction to reach $\mathbf{g}$ and then maps it to H at $\mathbf{x}_\mathbf{g}$ along the line through $\mathbf{x}$ and $\mathbf{g}$. However, the optimal update is not $\mathbf{g}$. It is easy to observe that moving along the tangent line can reach the nearest location of the decision boundary to the original image $\mathbf{x}$, thereby producing the adversarial example with the minimum distortion. In real attack scenarios, the image data reside in a high-dimensional space, and the semicircle becomes a hemisphere B. In n-dimensional space where $n \geq 3$, there are infinitely many tangent lines from $\mathbf{x}$ to B which produce infinitely many tangent points on B. Still, we will show that **exactly one tangent point** can lead to the minimum distortion when mapping it onto H along the tangent line, which is defined by the following theorem.

Definition of Optimal Tangent Points

Theorem 1: Let u be the unit normal vector of H, $\mathbf{k}$ be any point on the surface of B, then the distance $||\mathbf{x}-\mathbf{x}_t||_2$ is the shortest if $\mathbf{k}$ is the optimal solution of the following constrained optimization problem:

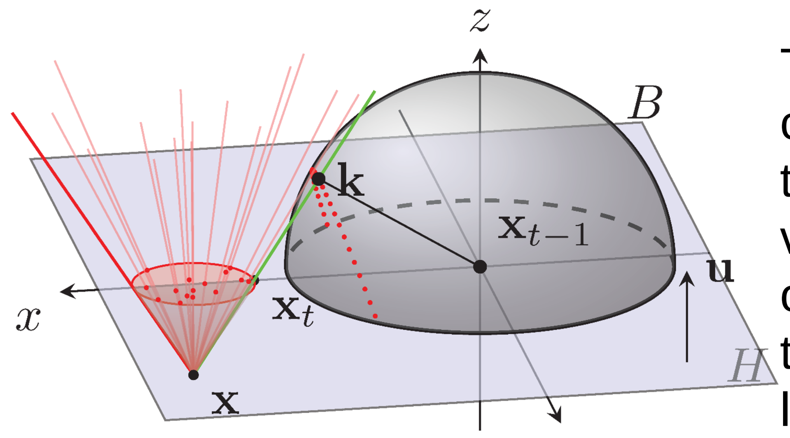
$$\operatorname*{argmax}_{\mathbf{k}} \langle \mathbf{k}-\mathbf{x}_{t-1}, \mathbf{u}\rangle, \longrightarrow \text{maximizes the projection of } \mathbf{k}-\mathbf{x}_{t-1} \text{ onto } \mathbf{u}.$$

$$\text{s.t. } \langle \mathbf{k}-\mathbf{x}_{t-1}, \mathbf{x}-\mathbf{k}\rangle = 0, \longrightarrow \text{ensures that } \mathbf{k} \text{ is a tangent point.}$$

$$||\mathbf{k}-\mathbf{x}_{t-1}|| = 0, \longrightarrow \text{indicates } \mathbf{k} \text{ is on the surface of } B.$$

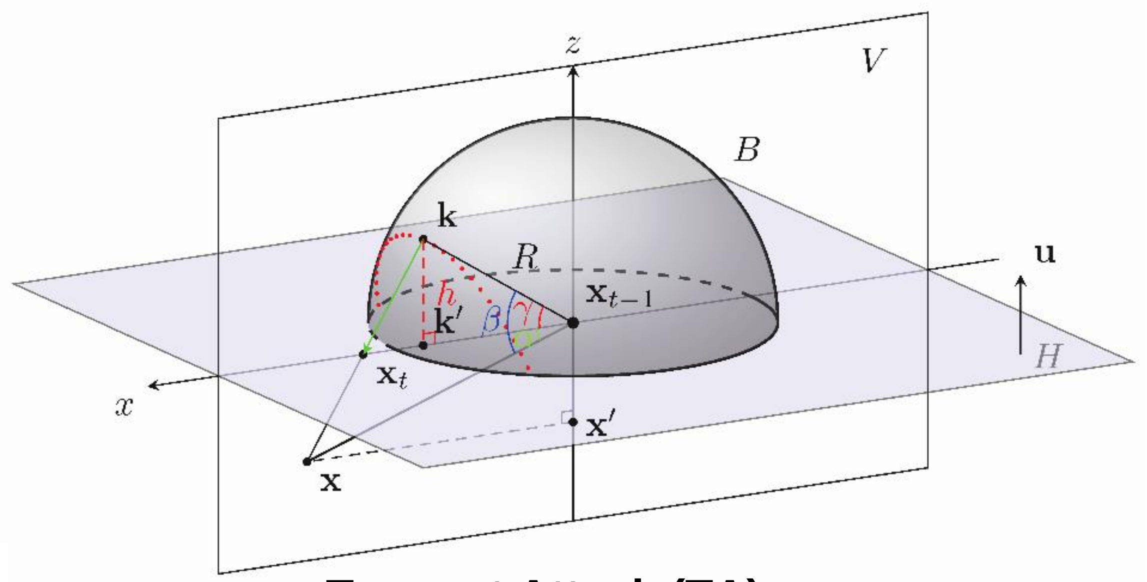
$$\langle \mathbf{k}-\mathbf{x}_{t-1}, \mathbf{u}\rangle > 0. \longrightarrow \text{states that } \mathbf{k} \text{ cannot appear on the same side of } H \text{ as } \mathbf{x}\text{, which is always satisfied in our algorithm.}$$

Geometrical Explanation of Theorem 1



The left figure show that all points on H that are closer to $\mathbf{x}$ than $\mathbf{x}_t$ are within a red disk, which is the intersection of H and the red cone whose vertex is $\mathbf{x}$. Clearly $\mathbf{k}$ is the only intersection point of the cone and the hemisphere B. Thus, of all the lines intersecting B, only the tangent line leads to the shortest distance from $\mathbf{x}$ to H.

Computation of Optimal Tangent Points



Tangent Attack (TA)

$$\beta = \alpha + \gamma$$

$$\sin\alpha = -\frac{\langle \mathbf{x},\mathbf{u}\rangle}{||\mathbf{x}|| \cdot ||\mathbf{u}||}$$

$$\cos\alpha = \frac{\sqrt{||\mathbf{x}||^2 \cdot ||\mathbf{u}||^2 - \langle \mathbf{x},\mathbf{u}\rangle^2}}{||\mathbf{x}|| \cdot ||\mathbf{u}||}$$

$$\sin\beta = \sqrt{1-\cos^2\beta}$$

$$\cos\gamma = \cos\beta\cos\alpha + \sin\beta\sin\alpha$$

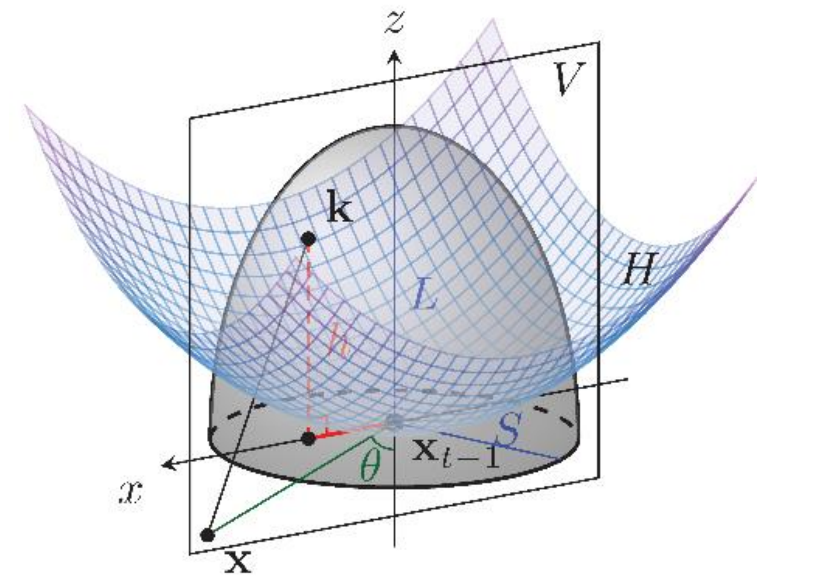
$$\sin\gamma = \sin\beta\cos\alpha - \cos\beta\sin\alpha$$

$$h = R\cdot\sin\gamma = R\cdot(\sin\beta\cos\alpha - \cos\beta\sin\alpha)$$

$$\mathbf{x}' = \langle \mathbf{x},\mathbf{u}\rangle \cdot \mathbf{u} \,/\, ||\mathbf{u}||^2$$

$$\mathbf{k} = \mathbf{k}' + h\cdot\mathbf{u}$$

$$= R\cdot\cos\gamma\cdot\frac{\mathbf{x}-\langle \mathbf{x},\mathbf{u}\rangle\cdot\mathbf{u} \,/\, ||\mathbf{u}||^2}{||\mathbf{x}-\langle \mathbf{x},\mathbf{u}\rangle\cdot\mathbf{u} \,/\, ||\mathbf{u}||^2||} + h\cdot\mathbf{u}$$



Generalized Tangent Attack (G-TA)

$$\theta = \arccos\left(\frac{\langle \mathbf{x}, -\mathbf{u}\rangle}{||\mathbf{x}|| \cdot ||\mathbf{u}||}\right)$$

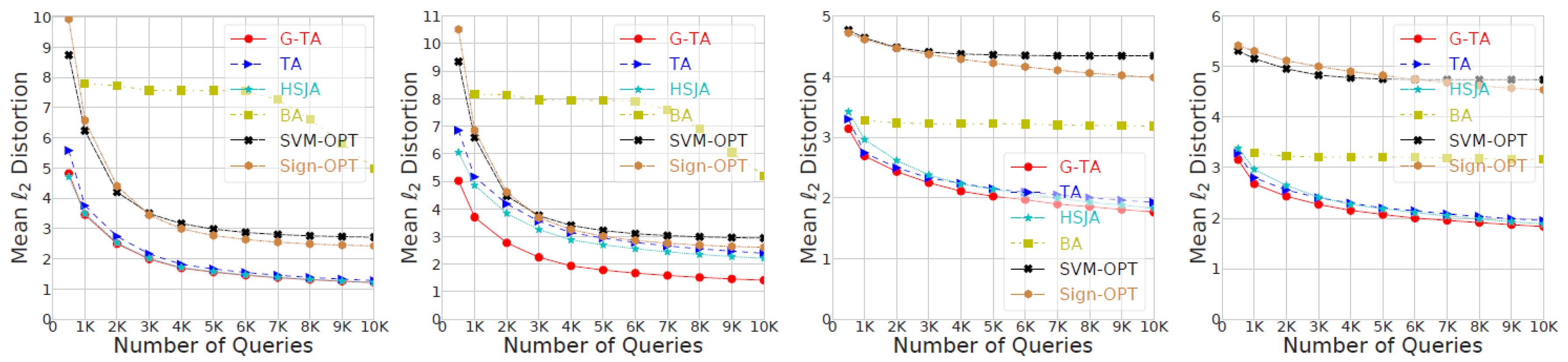
$$(x_0, z_0) = (||\mathbf{x}||\cdot\sin\theta, -||\mathbf{x}||\cdot\cos\theta)$$

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

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

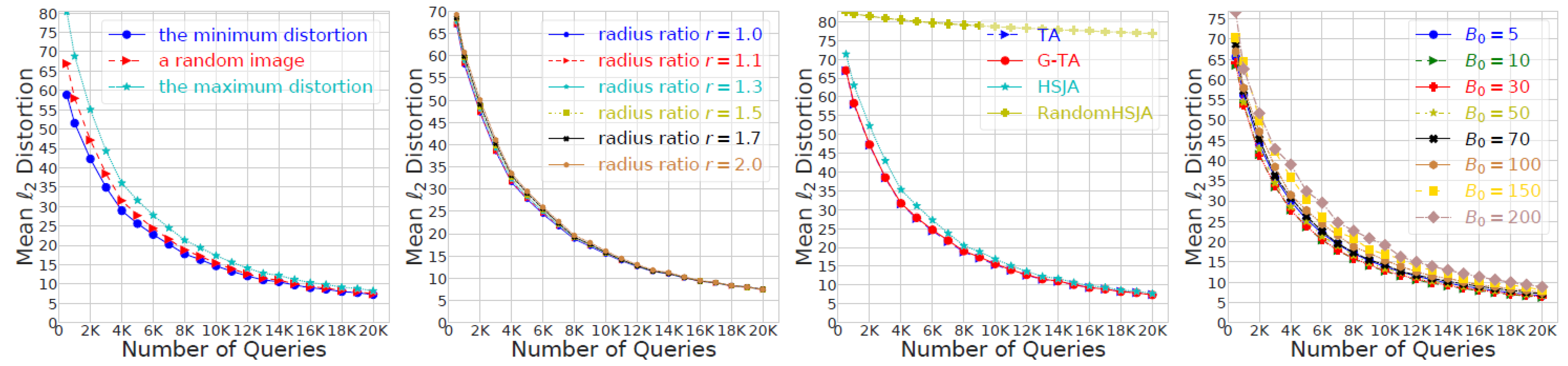
$$\mathbf{k} = |x_k| \cdot \frac{\mathbf{x}-\langle \mathbf{x},\mathbf{u}\rangle\cdot\mathbf{u} \,/\, ||\mathbf{u}||^2}{||\mathbf{x}-\langle \mathbf{x},\mathbf{u}\rangle\cdot\mathbf{u} \,/\, ||\mathbf{u}||^2||} + z_k\cdot\mathbf{u}$$

Experimental Results of Attacks against Defense Models



(a) AT ($\epsilon = 8/255$) (b) TRADES ($\epsilon = 8/255$) (c) JPEG (d) Feature Distillation

Comprehensive Understanding of Tangent Attack



(a) initial $\tilde{x}_0$ selection (b) radius ratio (c) jump directions (d) initial batch size B_0

Experimental Results

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

Target Model	Method	Targeted Attack @300	@1K	@2K	@5K	@8K	@10K	Untargeted Attack @300	@1K	@2K	@5K	@8K	@10K
Inception-v3	BA [4]	111.798	108.044	106.283	102.715	86.931	78.326	-	107.558	102.309	95.776	78.668	60.296
	Sign-OPT [11]	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 [11]	**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 [7]	111.562	95.295	82.111	52.544	37.395	30.425	103.605	57.295	37.185	15.484	9.989	7.967
	TA	103.781	**80.327**	**66.708**	**42.121**	**30.846**	**25.566**	94.752	52.523	35.229	15.040	9.748	7.793
	G-TA	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 [4]	110.343	106.616	104.586	100.321	84.058	75.507	-	116.075	111.474	104.451	86.572	66.283
	Sign-OPT [11]	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 [11]	**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 [7]	109.670	93.916	80.937	52.358	37.773	30.958	110.727	63.731	42.290	17.936	11.367	8.911
	TA	101.666	**78.683**	**65.304**	**41.629**	30.993	25.958	101.207	58.616	40.314	17.639	11.304	8.907
	G-TA	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 [4]	81.090	77.723	76.122	71.967	55.953	47.652	-	75.998	71.671	66.983	53.917	40.725
	Sign-OPT [11]	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 [11]	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 [7]	77.035	63.488	51.802	30.138	19.680	16.261	71.248	38.035	24.895	10.218	5.855	4.842
	TA	70.739	55.256	**43.694**	**24.961**	**16.756**	**13.876**	**65.589**	**35.689**	24.037	10.039	5.774	4.766
	G-TA	**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 [4]	81.565	77.903	76.366	72.392	58.746	51.679	-	64.007	60.389	56.544	44.175	31.371
	Sign-OPT [11]	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 [11]	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 [7]	76.121	63.091	52.301	31.018	20.472	16.911	56.264	27.443	17.717	7.649	4.723	**4.019**
	TA	**72.434**	**57.969**	**47.142**	**27.699**	**18.788**	**15.414**	53.197	26.777	17.651	7.730	4.822	4.107
	G-TA	72.459	58.320	47.297	27.905	19.045	15.633	**53.058**	**26.631**	**17.384**	**7.602**	**4.720**	4.026

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

Target Model	Method	Targeted Attack @300	@1K	@2K	@5K	@8K	@10K	Untargeted Attack @300	@1K	@2K	@5K	@8K	@10K
PyramidNet-272	BA [4]	8.651	8.073	8.013	6.387	4.189	3.333	-	5.636	4.725	4.414	2.750	1.696
	Sign-OPT [11]	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 [11]	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]	7.917	4.329	2.523	**0.793**	**0.489**	**0.397**	4.505	1.279	0.713	0.333	0.255	0.227
	TA	7.943	**4.267**	2.488	0.809	0.503	0.406	**4.256**	1.275	0.710	**0.329**	0.253	0.226
	G-TA	**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 [4]	8.487	7.885	7.821	6.034	3.632	2.703	-	2.717	2.514	2.373	1.642	1.106
	Sign-OPT [11]	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 [11]	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]	7.714	3.566	1.966	0.591	0.365	0.301	2.188	0.756	0.483	0.261	0.208	0.189
	TA	**7.674**	**3.529**	**1.946**	0.585	0.366	0.302	2.190	0.774	0.485	0.257	0.206	0.187
	G-TA	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 [4]	8.688	8.046	7.984	5.786	2.486	1.555	-	4.425	3.648	3.435	1.543	0.832
	Sign-OPT [11]	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 [11]	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 [7]	6.810	2.603	1.326	0.518	0.389	0.347	3.052	0.797	0.508	0.299	0.250	0.232
	TA	6.802	2.556	1.311	0.519	0.394	0.353	**2.974**	0.785	**0.496**	**0.293**	0.249	0.233
	G-TA	**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 [4]	8.658	8.014	7.953	5.738	2.484	1.566	-	4.377	3.586	3.367	1.487	0.821
	Sign-OPT [11]	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 [11]	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 [7]	6.909	2.648	1.330	0.528	0.400	**0.357**	2.992	0.777	0.498	0.290	0.242	0.225
	TA	6.944	**2.579**	**1.295**	**0.523**	**0.398**	0.358	**2.926**	**0.770**	**0.490**	**0.288**	0.243	0.227
	G-TA	**6.783**	2.605	1.320	0.535	0.403	0.361	2.952	0.772	0.492	**0.288**	0.241	0.223

Table 3: Experimental results of the combined method of QEBA-S and TA.

Target Model	Method	Targeted Attack @300	@1K	@2K	@5K	@8K	@10K
Inception-v3	QEBA-S [25]	100.295	79.604	63.621	35.194	22.773	18.414
	QEBA-S + TA	104.490	**75.622**	**59.836**	**33.112**	**22.329**	**17.799**
Inception-v4	QEBA-S [25]	**97.772**	77.347	62.451	35.275	23.204	19.002
	QEBA-S + TA	101.845	**73.838**	**58.554**	**33.288**	**23.160**	**18.736**
SENet-154	QEBA-S [25]	**72.831**	55.367	42.674	21.988	13.888	11.210
	QEBA-S + TA	76.547	**52.269**	**39.740**	**20.608**	**13.873**	**11.016**
ResNet-101	QEBA-S [25]	**75.567**	57.929	44.983	23.209	14.402	11.467
	QEBA-S + TA	78.709	**53.917**	**41.245**	**21.198**	**13.856**	**10.773**