Boosting Ray Search Procedure of Hard-label Attacks with Transfer-based Priors

Chen Ma[1,2], Xinjie Xu[1], Shuyu Cheng[3], Qi Xuan[1,2]

[1] Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou, China

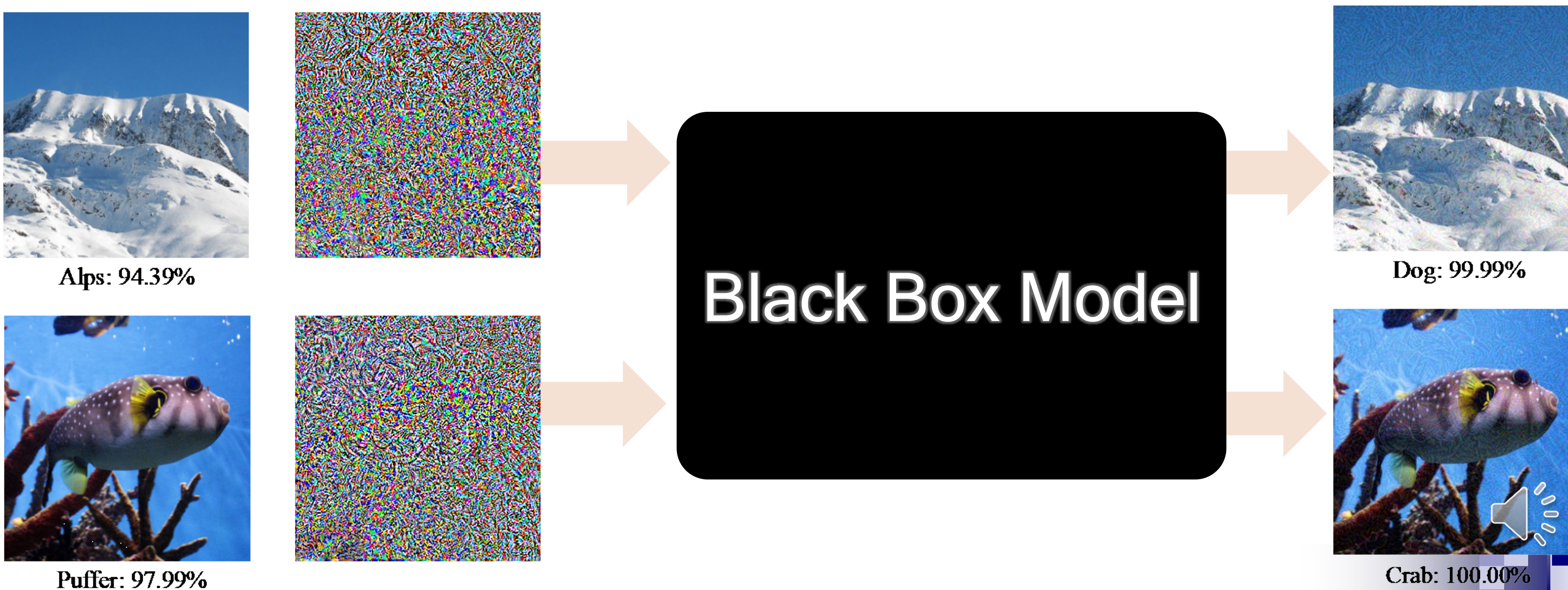
[2] Binjiang Institute of Artificial Intelligence, ZJUT, Hangzhou, China

[3] JQ Investments, Shanghai, China

machenstar@163.com

Background: black-box adversarial attack

- An adversarial example should be visually indistinguishable from the corresponding normal one, yet it is misclassified by the target model.
- Hard-label adversarial attacks aim to generate adversarial examples using only the top-1 predicted label.



Black-box Attackss

- 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: the adversary knows the output logits of the target model
    - **Decision-based: the adversary only knows the top-1 predicted labels**
  - Goal: save queries and reduce the distortions of examples

Hard-label Attacks

- Goal: For a classifier $f(x)$: $\mathbb{R}^d \rightarrow \mathbb{R}^K$ and input-label pair (x, y), the hard-label black-box attack generates an adversarial example x_{adv} using only the classifier's top-1 predicted label:

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

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

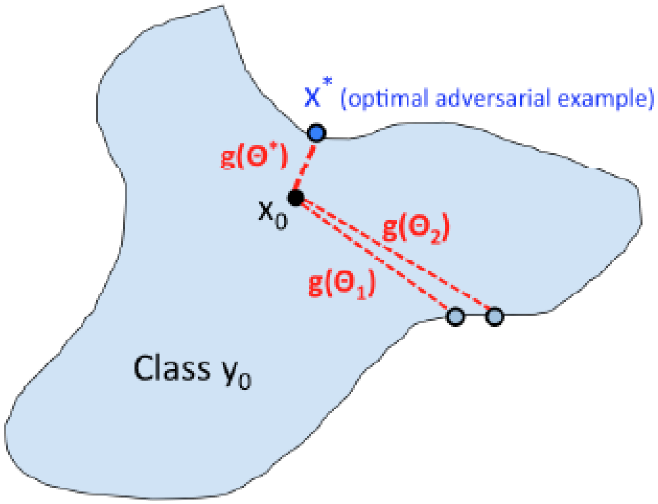
- x^{adv} can be generated by solving

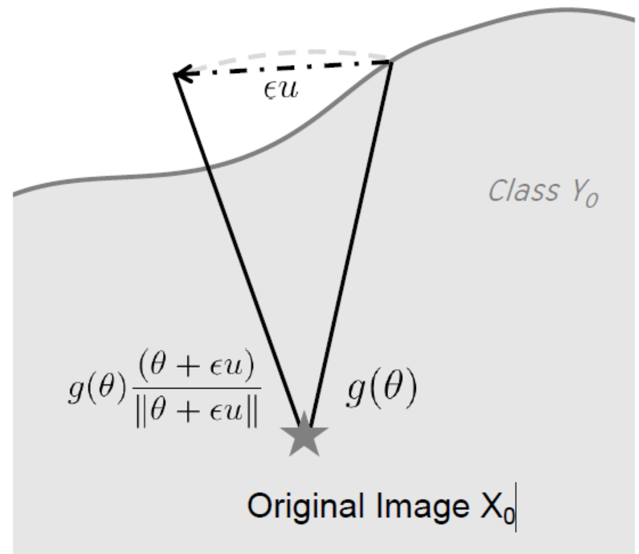
$$x^{\mathrm{adv}} = \underset{x^{\mathrm{adv}}}{\mathrm{argmin}}\, d(x^{\mathrm{adv}}, x) \ \ \text{s.t.} \ \ \phi(x_{\mathrm{adv}}) = 1$$

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

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

A typical hard-label attack: Sign-OPT



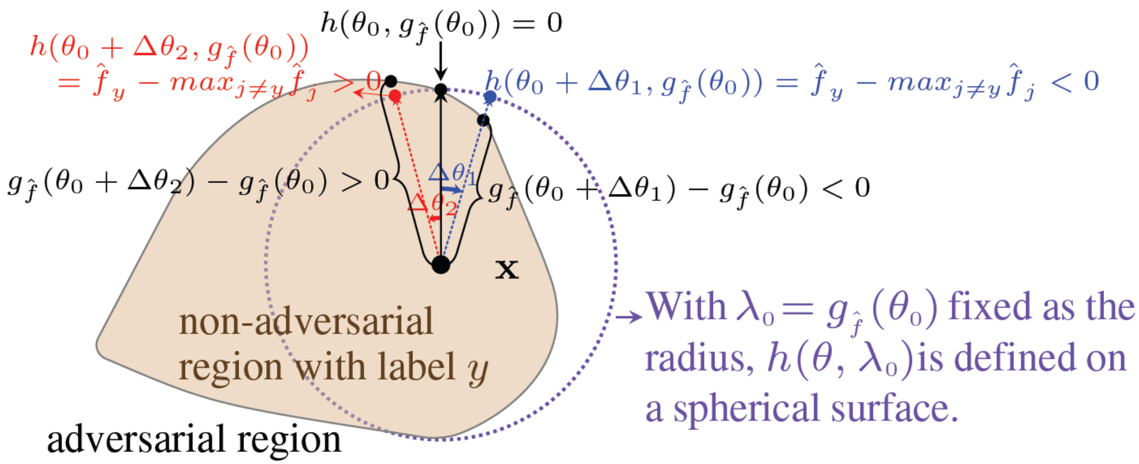


Untargeted attack: $g(\boldsymbol{\theta}) = \min_{\lambda>0} \lambda \quad \texttt{s.t} \quad f(\boldsymbol{x}_0 + \lambda \frac{\boldsymbol{\theta}}{||\boldsymbol{\theta}||}) \neq y_0$

Targeted attack (given target t): $g(\boldsymbol{\theta}) = \min_{\lambda>0} \lambda \quad \texttt{s.t} \quad f(\boldsymbol{x}_0 + \lambda \frac{\boldsymbol{\theta}}{||\boldsymbol{\theta}||}) = t$

$$\text{sign}(g(\boldsymbol{\theta} + \epsilon \mathbf{u}) - g(\boldsymbol{\theta})) = \begin{cases} +1, & f(x_0 + g(\boldsymbol{\theta})\frac{(\boldsymbol{\theta}+\epsilon\mathbf{u})}{\|\boldsymbol{\theta}+\epsilon\mathbf{u}\|}) = y_0, \\ -1, & \text{Otherwise.} \end{cases}$$

Acquisition of Transfer-based Priors



(1) When $g_{\hat{f}}(\theta) \downarrow$ with $\Delta\theta_1$, $h(\theta, \lambda) \downarrow$ as well.

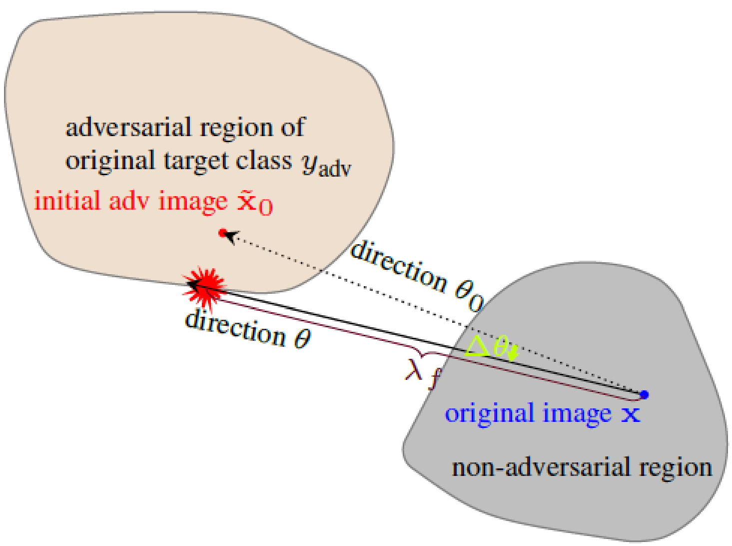
(2) When $g_{\hat{f}}(\theta) \uparrow$ with $\Delta\theta_2$, $h(\theta, \lambda) \uparrow$ as well.

To approximate the gradient of the non-differentiable function $g_{\hat{f}}(\theta_0)$, we introduce a surrogate loss $h(\theta, \lambda)$ by fixing $\lambda_0 = g_{\hat{f}}(\theta_0)$, where $\hat{f}$ denotes the surrogate model. This yields a transfer-based prior via $\nabla g_{\hat{f}}(\theta_0) \propto \nabla_\theta h(\theta_0, \lambda_0)$. The surrogate loss $h(\theta, \lambda)$ is defined as

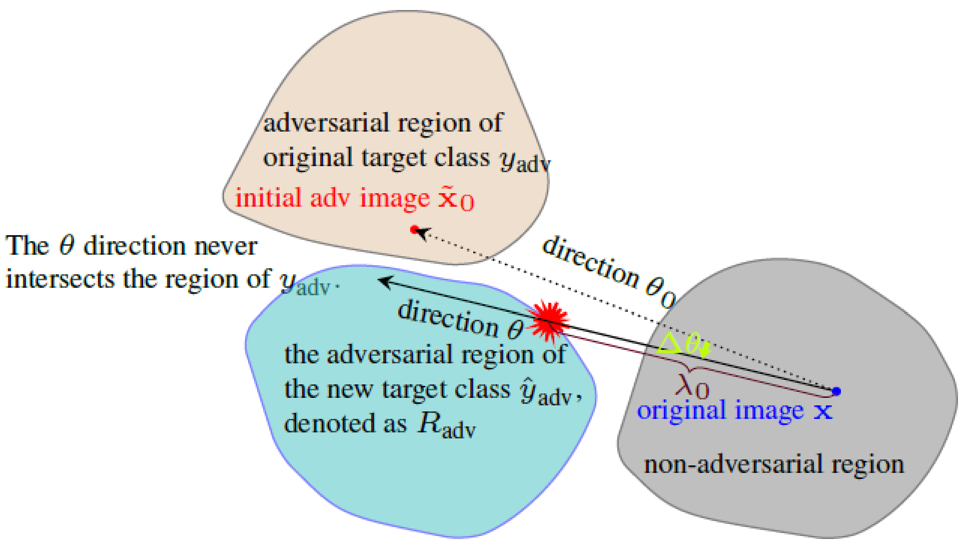
$$h(\theta, \lambda) := \begin{cases} \hat{f}_y - \max_{j \neq y} \hat{f}_j, & \text{if untargeted attack,} \\ \max_{j \neq \hat{y}_{\text{adv}}} \hat{f}_j - \hat{f}_{\hat{y}_{\text{adv}}}, & \text{if targeted attack,} \end{cases}$$

Acquisition of Transfer-based Priors

- For targeted attacks, the transfer-based priors are more difficult to obtain.

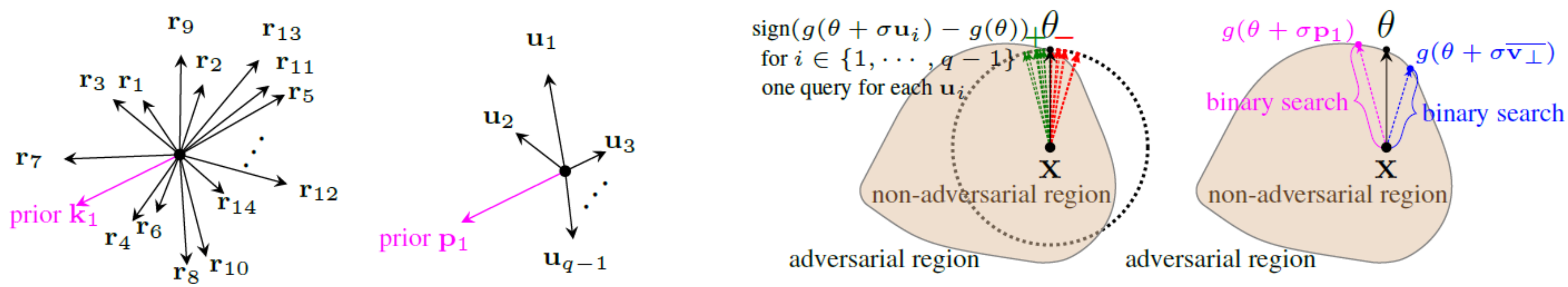


(a) The θ direction in the target model f.



(b) The θ direction in the surrogate model $\hat{f}$.

The Optimization of Ray Directions



(a) Get prior $\mathbf{k}_1$ and sample $\mathbf{r}_i$. (b) Orthonormal basis. (c) Estimate a sign-based $\mathbf{v}_\perp$. (d) Estimate $\mathbf{v}^*$ with $\mathbf{v}_\perp$ and $\mathbf{p}_1$.

1. Prior-Sign-OPT

$$\mathbf{v}^* = \sum_{i=1}^{s} \text{sign}(g(\theta + \sigma \mathbf{p}_i) - g(\theta)) \cdot \mathbf{p}_i + \sum_{i=1}^{q-s} \text{sign}(g(\theta + \sigma \mathbf{u}_i) - g(\theta)) \cdot \mathbf{u}_i.$$

2. Prior-OPT

$$\mathbf{v}^* = \sum_{i=1}^{s} \frac{g(\theta + \sigma \mathbf{p}_i) - g(\theta)}{\sigma} \cdot \mathbf{p}_i + \frac{g(\theta + \sigma \overline{\mathbf{v}_\perp}) - g(\theta)}{\sigma} \cdot \overline{\mathbf{v}_\perp},$$

where $\overline{\mathbf{v}_\perp}$ is the ℓ_2 normalization of $\mathbf{v}_\perp$, and $\mathbf{v}_\perp$ is obtained by:

$$\mathbf{v}_\perp = \sum_{i=1}^{q-s} \text{sign}(g(\theta + \sigma \mathbf{u}_i) - g(\theta)) \cdot \mathbf{u}_i.$$

Algorithm 1 Prior-Sign-OPT and Prior-OPT attack

Input: benign image $\mathbf{x}$, objective function $g(\cdot)$, attack success indicator $\Phi(\cdot)$ defined in Eq. (2), iteration T, method $m \in \{$ `Prior-OPT`, `Prior-Sign-OPT`$\}$, the initialization strategy of untargeted attacks $\in \{\theta_0^{\text{PGD}}, \theta_0^{\text{RND}}\}$, the maximum gradient norm $\mathbf{g}_{\max}$, attack norm $p \in \{2, \infty\}$, surrogate models $\mathbb{S} = \{\hat{f}_1, \cdots, \hat{f}_s\}$.

Output: adversarial example $\mathbf{x}^*$ that satisfies $\Phi(\mathbf{x}^*) = 1$.

$\tilde{\mathbf{x}}_0 \leftarrow \text{PGD}(\mathbf{x}, \hat{f}_1)$ if initialization $= \theta_0^{\text{PGD}}$, otherwise a random $\tilde{\mathbf{x}}_0$ that satisfies $\Phi(\tilde{\mathbf{x}}_0) = 1$ is selected, which is θ_0^{RND} strategy; ▷ the targeted attack selects an image from the target class as $\tilde{\mathbf{x}}_0$.

$\theta_0 \leftarrow \frac{\tilde{\mathbf{x}}_0 - \mathbf{x}}{\|\tilde{\mathbf{x}}_0 - \mathbf{x}\|}, \quad d_0 \leftarrow \|\tilde{\mathbf{x}}_0 - \mathbf{x}\|_p;$

for t **in** $1, \ldots, T$ **do**

  for $\hat{f}_i$ **in** $\mathbb{S}$ **do**

    $\lambda_{t-1} \leftarrow \text{BinarySearch}(\mathbf{x}, \theta_{t-1}, \hat{f}_i, \Phi);$

    $\mathbf{k}_i \leftarrow \nabla_\theta h(\theta_{t-1}, \lambda_{t-1})$ on $\hat{f}_i$ with λ_{t-1} treated as a constant in differentiation; ▷ obtain s transfer-based priors

  end for

  $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $i = 1, \cdots, q - s;$

  $\mathbf{p}_1, \cdots, \mathbf{p}_s, \mathbf{u}_1, \cdots, \mathbf{u}_{q-s} \leftarrow$ Gram-Schmidt Orthogonalization($\{\mathbf{k}_1, \cdots, \mathbf{k}_s, \mathbf{r}_1, \cdots, \mathbf{r}_{q-s}\}$);

  Estimate a gradient $\mathbf{v}^*$ using Eq. (7) if $m =$ `Prior-Sign-OPT`, otherwise using Eq. (13);

  $\mathbf{v}^* \leftarrow \text{ClipGradNorm}(\mathbf{v}^*, \mathbf{g}_{\max});$

  $\eta^* \leftarrow \text{LineSearch}(\mathbf{x}, \mathbf{v}^*, \Phi, d_{t-1}, \theta_{t-1});$ ▷ search step size.

  $\theta_t \leftarrow \theta_{t-1} - \eta^* \mathbf{v}^*, \quad \theta_t \leftarrow \frac{\theta_t}{\|\theta_t\|};$

  $d_t \leftarrow \|g(\theta_t) \cdot \theta_t\|_p;$

end for

return $\mathbf{x}^* \leftarrow \mathbf{x} + g(\theta_T)\frac{\theta_T}{\|\theta_T\|};$

Theory (part I)

Theorem 3.2. *For the Sign-OPT estimator approximated by Eq. (6) (defined as Eq. (44)), we let* $\gamma := \overline{\mathbf{v}}^\top \overline{\nabla g(\theta)}$ *be its cosine similarity to the true gradient, where* $\overline{\mathbf{v}} := \frac{\mathbf{v}}{\|\mathbf{v}\|}$*, then*

$$\mathbb{E}[\gamma] = \sqrt{q}\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\sqrt{\pi}}, \tag{9}$$

$$\mathbb{E}[\gamma^2] = \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right). \tag{10}$$

The proof of Theorem 3.2 is included in Appendix C.1. For Prior-Sign-OPT, we have Theorem 3.3.

Theorem 3.3. *For the Prior-Sign-OPT estimator approximated by Eq. (6) (defined as Eq. (82)), we let* $\gamma := \overline{\mathbf{v}^*}^\top \overline{\nabla g(\theta)}$ *be its cosine similarity to the true gradient, where* $\overline{\mathbf{v}^*} := \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|}$*, then*

$$\mathbb{E}[\gamma] = \frac{1}{\sqrt{q}}\left[\sum_{i=1}^{s}|\alpha_i| + (q-s)\sqrt{1-\sum_{i=1}^{s}\alpha_i^2}\cdot\frac{\Gamma(\frac{d-s}{2})}{\Gamma(\frac{d-s+1}{2})\sqrt{\pi}}\right], \tag{11}$$

$$\begin{aligned}\mathbb{E}[\gamma^2] = \frac{1}{q}\Bigg[&\left(\sum_{i=1}^{s}|\alpha_i|\right)^2 + \frac{q-s}{d-s}\left(\frac{2}{\pi}(q-s-1)+1\right)\left(1-\sum_{i=1}^{s}\alpha_i^2\right)\\ &+2\left(\sum_{i=1}^{s}|\alpha_i|\right)(q-s)\sqrt{1-\sum_{i=1}^{s}\alpha_i^2}\cdot\frac{\Gamma(\frac{d-s}{2})}{\Gamma(\frac{d-s+1}{2})\sqrt{\pi}}\Bigg],\end{aligned} \tag{12}$$

where $\alpha_i := \mathbf{p}_i^\top \overline{\nabla g(\theta)}$ *is the cosine similarity between the* i*-th prior and the true gradient.*

Theory (part II)

Theorem 3.4. *For the Prior-OPT estimator approximated by Eq. (6) (defined as Eq. (114)), we le* $\gamma := \overline{\mathbf{v}^*}^\top \overline{\nabla g(\theta)}$ *be its cosine similarity to the true gradient, where* $\overline{\mathbf{v}^*} := \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|}$*, then*

$$\mathbb{E}[\gamma] \geq \sqrt{\sum_{i=1}^{s} \alpha_i^2 + \frac{(q-s)(1-\sum_{i=1}^{s}\alpha_i^2)}{\pi}\left(\frac{\Gamma(\frac{d-s}{2})}{\Gamma(\frac{d-s+1}{2})}\right)^2}, \quad (15$$

$$\mathbb{E}[\gamma] \leq \sqrt{\sum_{i=1}^{s} \alpha_i^2 + \frac{1}{d-s}\left(\frac{2}{\pi}(q-s-1)+1\right)\left(1-\sum_{i=1}^{s}\alpha_i^2\right)}, \quad (16$$

$$\mathbb{E}[\gamma^2] = \sum_{i=1}^{s} \alpha_i^2 + \frac{1}{d-s}\left(\frac{2}{\pi}(q-s-1)+1\right)\left(1-\sum_{i=1}^{s}\alpha_i^2\right), \quad (17$$

where $\alpha_i := \mathbf{p}_i^\top \overline{\nabla g(\theta)}$ *is the cosine similarity between the* i*-th prior and the true gradient.*

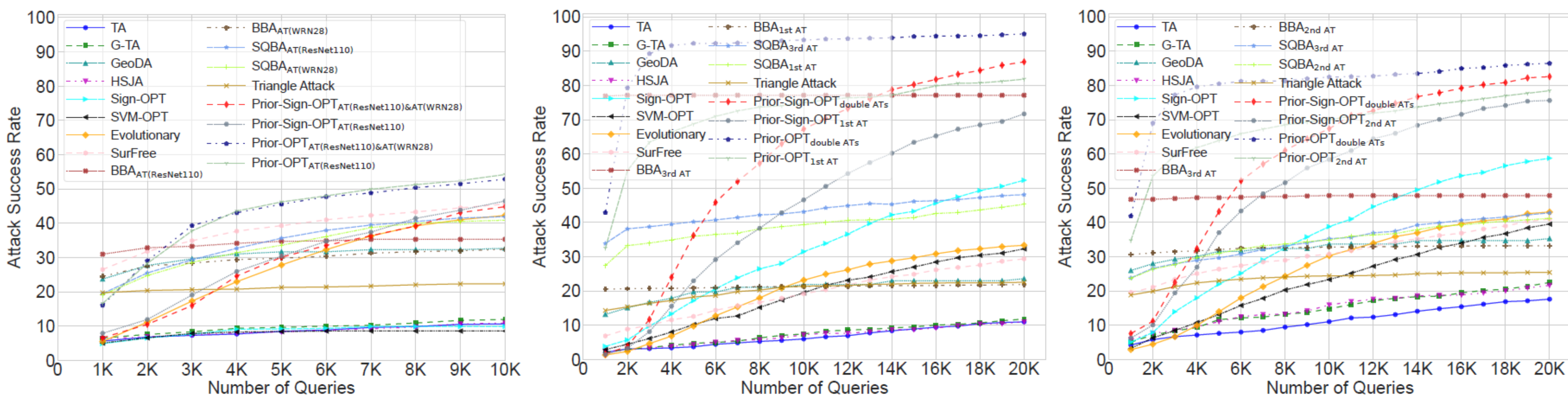
Experimental Results

Target Model	Method	Untargeted Attack @1K	@2K	@5K	@8K	@10K	Targeted Attack @1K	@2K	@5K	@8K	@10K	@15K	@20K
Inception-v4	HSJA (Chen et al., 2020)	75.392	44.530	20.567	14.194	11.645	95.876	79.001	52.176	39.190	32.951	24.546	19.522
	TA (Ma et al., 2021b)	67.496	42.233	20.352	14.175	11.694	78.883	61.990	40.669	31.506	27.111	21.079	17.319
	G-TA (Ma et al., 2021b)	67.842	41.946	19.962	13.865	11.448	79.297	62.291	40.529	30.941	26.427	20.268	16.569
	Sign-OPT (Cheng et al., 2020)	86.716	48.233	18.258	11.067	8.786	80.366	65.200	42.866	32.104	27.526	20.394	16.281
	SVM-OPT (Cheng et al., 2020)	89.863	47.914	18.297	11.091	8.839	79.807	65.590	43.426	33.090	28.797	22.354	18.795
	GeoDA (Rahmati et al., 2020)	29.157	20.119	12.487	11.010	9.688	-	-	-	-	-	-	-
	Evolutionary (Dong et al., 2019)	61.966	42.665	20.815	13.382	10.839	81.761	65.060	43.021	32.120	27.385	19.942	15.610
	SurFree (Maho et al., 2021)	51.685	38.482	22.845	16.374	13.818	84.925	74.887	55.991	44.475	39.004	29.354	23.153
	Triangle Attack (Wang et al., 2022)	27.217	25.853	23.743	22.581	22.132	-	-	-	-	-	-	-
	$\text{SQBA}_{\text{IncResV2}}$ (Park et al., 2024)	26.134	19.035	11.189	8.432	7.417	-	-	-	-	-	-	-
	$\text{SQBA}_{\text{Xception}}$ (Park et al., 2024)	23.672	17.424	10.502	8.036	7.115	-	-	-	-	-	-	-
	$\text{BBA}_{\text{IncResV2}}$ (Brunner et al., 2019)	38.782	28.437	18.757	15.474	14.191	66.746	56.283	41.324	34.066	30.942	25.757	22.630
	$\text{BBA}_{\text{Xception}}$ (Brunner et al., 2019)	43.317	31.519	20.504	16.712	15.282	63.069	53.363	39.740	33.166	30.221	25.438	22.561
	$\text{Prior-Sign-OPT}_{\text{IncResV2}}$	81.991	42.403	12.835	7.365	5.842	74.597	55.421	31.856	22.958	19.513	14.361	11.665
	$\text{Prior-Sign-OPT}_{\text{IncResV2\&Xception}}$	77.683	37.099	9.058	5.195	4.199	69.526	49.368	**26.882**	**19.324**	**16.697**	**12.821**	**10.769**
	$\text{Prior-Sign-OPT}_{\theta_0^{\text{PGD}}+\text{IncResV2}}$	23.596	15.347	8.074	5.729	4.863	-	-	-	-	-	-	-
	$\text{Prior-OPT}_{\text{IncResV2}}$	49.279	18.135	5.718	4.451	4.027	67.300	49.842	33.477	27.602	25.281	21.837	19.800
	$\text{Prior-OPT}_{\text{IncResV2\&Xception}}$	42.541	13.418	**3.919**	**3.321**	**3.119**	**60.211**	**42.631**	27.547	23.011	21.441	19.193	17.983
	$\text{Prior-OPT}_{\theta_0^{\text{PGD}}+\text{IncResV2}}$	**22.852**	**12.194**	6.568	5.114	4.548	-	-	-	-	-	-	-
ViT	HSJA (Chen et al., 2020)	37.813	19.386	9.031	6.604	5.637	61.491	44.853	23.947	16.926	14.152	10.791	8.922
	TA (Ma et al., 2021b)	37.923	19.867	9.078	6.636	5.674	52.110	36.455	**20.536**	15.145	12.885	10.158	8.609
	G-TA (Ma et al., 2021b)	37.425	19.347	8.948	6.496	5.643	52.550	36.720	20.857	15.436	13.255	10.490	8.933
	Sign-OPT (Cheng et al., 2020)	51.120	25.290	8.559	5.482	4.572	55.941	41.867	23.784	16.541	13.873	10.129	8.267
	SVM-OPT (Cheng et al., 2020)	55.802	26.580	9.242	5.988	5.070	56.002	41.899	23.909	17.273	14.848	11.739	10.320
	GeoDA (Rahmati et al., 2020)	18.880	12.904	8.039	7.153	6.313	-	-	-	-	-	-	-
	Evolutionary (Dong et al., 2019)	40.382	25.709	11.925	7.974	6.719	57.141	40.187	21.782	15.191	12.795	9.677	8.311
	SurFree (Maho et al., 2021)	28.228	19.016	10.194	7.321	6.303	70.337	53.129	30.054	20.595	16.908	11.794	9.204
	Triangle Attack (Wang et al., 2022)	**12.789**	12.144	11.064	10.411	10.097	-	-	-	-	-	-	-
	$\text{SQBA}_{\text{ResNet50}}$ (Park et al., 2024)	21.741	14.004	7.738	5.861	5.201	-	-	-	-	-	-	-
	$\text{SQBA}_{\text{ConViT}}$ (Park et al., 2024)	12.886	9.762	6.240	4.947	4.452	-	-	-	-	-	-	-
	$\text{BBA}_{\text{ResNet50}}$ (Brunner et al., 2019)	29.755	20.053	12.580	10.375	9.567	**43.231**	**33.365**	21.889	17.635	16.046	13.726	12.463
	$\text{BBA}_{\text{ConViT}}$ (Brunner et al., 2019)	22.716	16.153	10.893	9.193	8.595	45.588	35.227	22.865	18.325	16.614	14.028	12.623
	$\text{Prior-Sign-OPT}_{\text{ResNet50}}$	50.161	27.953	9.474	5.872	4.850	55.095	40.480	22.354	15.626	13.201	9.789	8.048
	$\text{Prior-Sign-OPT}_{\text{ResNet50\&ConViT}}$	46.196	23.869	7.327	4.694	3.967	53.925	38.418	20.673	**14.422**	**12.153**	**9.090**	**7.544**
	$\text{Prior-Sign-OPT}_{\theta_0^{\text{PGD}}+\text{ResNet50}}$	29.912	18.425	7.848	5.175	4.331	-	-	-	-	-	-	-
	$\text{Prior-OPT}_{\text{ResNet50}}$	42.838	22.704	8.848	6.024	5.195	54.348	40.930	24.408	18.117	15.803	12.638	11.070
	$\text{Prior-OPT}_{\text{ResNet50\&ConViT}}$	26.495	**11.287**	**4.929**	**3.937**	**3.609**	53.369	40.002	24.706	19.148	17.116	14.114	12.650
	$\text{Prior-OPT}_{\theta_0^{\text{PGD}}+\text{ResNet50}}$	29.099	17.754	8.208	5.782	5.009	-	-	-	-	-	-	-

Table 2: Mean ℓ_2 distortions of the different numbers of priors on the ImageNet dataset.

Method	Priors	Target Model: ResNet-101[1]					Target Model: Swin Transformer[2]					Target Model: GC ViT[2]				
		@1K	@2K	@5K	@8K	@10K	@1K	@2K	@5K	@8K	@10K	@1K	@2K	@5K	@8K	@10K
Sign-OPT	no prior	37.248	21.235	8.982	5.811	4.754	86.373	53.399	20.686	12.406	9.899	57.903	35.762	14.763	9.047	7.185
Prior-Sign-OPT	1 prior	34.150	18.733	6.111	3.718	3.019	84.124	52.882	20.344	11.880	9.254	57.171	36.949	14.963	8.931	6.899
	2 priors	32.848	17.548	5.121	3.136	2.593	77.459	43.062	13.614	7.903	6.331	54.896	32.418	11.012	6.651	5.342
	3 priors	31.156	15.455	4.074	2.527	2.122	73.110	37.852	10.264	5.939	4.778	52.744	28.939	8.707	5.245	4.215
	4 priors	29.984	14.707	3.698	2.333	1.989	70.246	34.470	8.526	5.066	4.169	50.256	26.027	6.435	3.804	3.212
	5 priors	**29.601**	**14.195**	**3.573**	**2.275**	**1.951**	**67.616**	**32.225**	**7.321**	**4.219**	**3.467**	**48.935**	**24.821**	**6.123**	**3.601**	**2.893**
Prior-OPT	1 prior	18.355	7.100	2.840	2.324	2.158	69.432	39.447	16.536	11.241	9.625	50.467	29.091	11.537	7.311	5.948
	2 priors	17.373	6.465	2.454	2.096	1.979	41.152	17.977	7.289	5.453	4.896	36.055	16.176	6.094	4.413	3.747
	3 priors	15.373	5.350	1.919	1.714	1.653	**36.636**	13.877	5.166	4.008	3.687	**33.181**	13.005	4.702	3.644	3.264
	4 priors	**15.422**	**5.220**	**1.849**	**1.654**	**1.596**	38.343	12.650	3.784	3.027	2.850	34.396	10.994	3.047	2.356	2.171
	5 priors	15.556	5.395	1.881	1.672	1.605	37.712	**12.070**	**3.488**	**2.747**	**2.577**	33.351	**10.369**	**2.921**	**2.329**	**2.159**

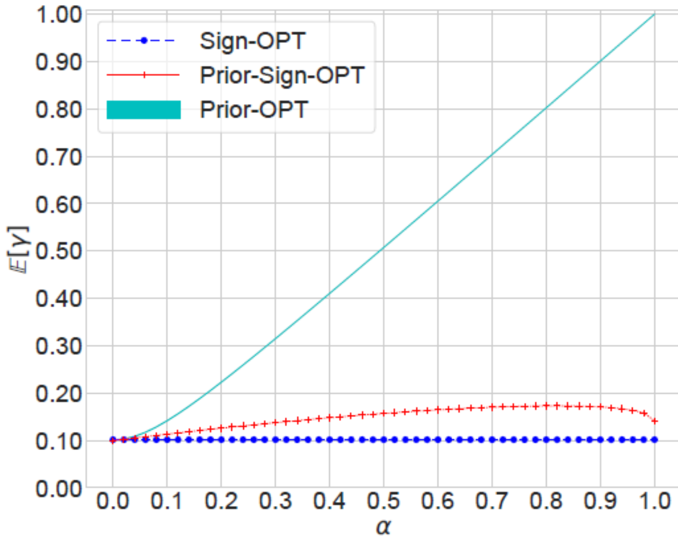
[1] Five surrogate models: ResNet-50, SENet-154, ResNeXt-101 (64 × 4d), VGG-13, SqueezeNet v1.1
[2] Five surrogate models: ResNet-50, ConViT, CrossViT, MaxViT, ViT



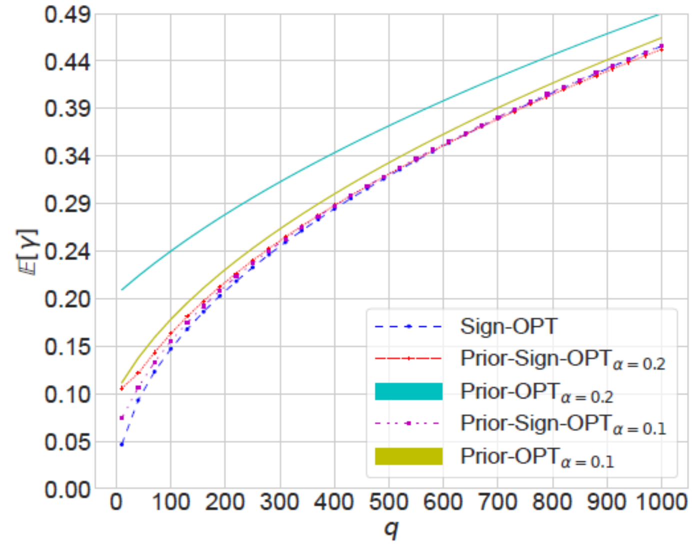
(a) TRADES (CIFAR-10) (b) $\mathrm{AT}_{\epsilon_\infty=4/255}$ (ImageNet)[1] (c) $\mathrm{AT}_{\epsilon_\infty=8/255}$ (ImageNet)[2]

Figure 4: Attack success rates of untargeted attacks with ℓ_2 norm constraint against defense models.

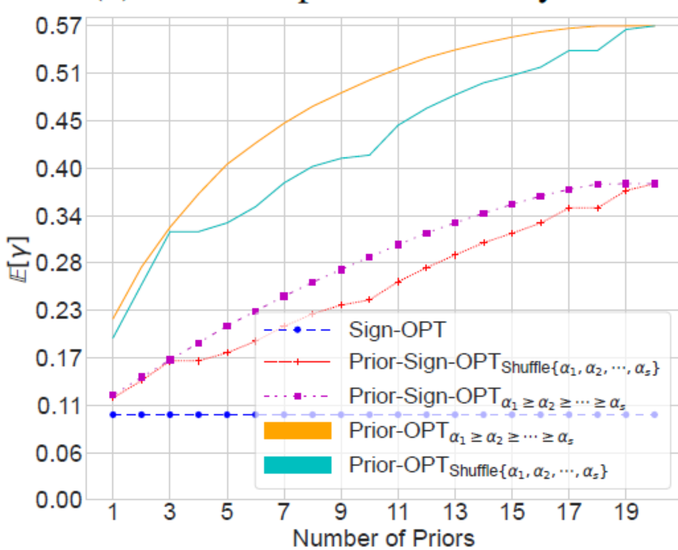
Comprehensive Understanding



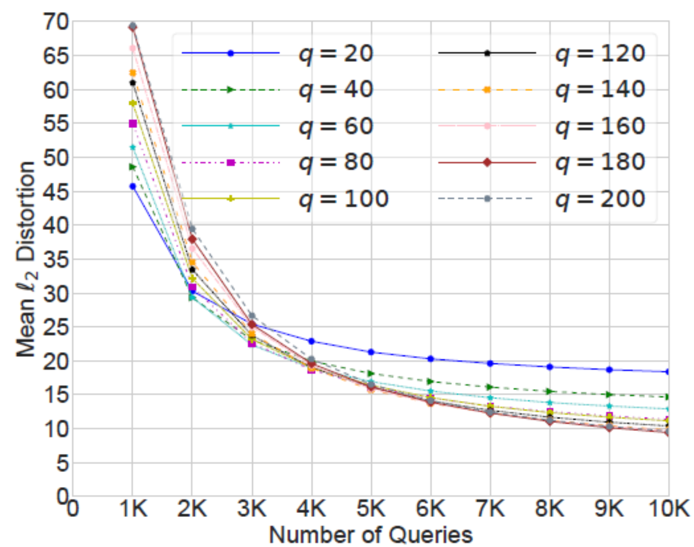
(a) Effect of prior's accuracy α



(b) Effect of number of vectors q



(c) Effect of number of priors



(d) Prior-OPT with different q

Thanks for your listening!