Supplementary Material for Parsimonious Black-Box Adversarial Attacks via Efficient Combinatorial Optimization

Seungyong Moon [*] [1] [2] Gaon An [*] [1] [2] Hyun Oh Song [1] [2]

A. Proofs

A.1. Proof of Lemma 1

Lemma 1. *Let $\mathcal{S}$ be the solution obtained by performing the local search algorithm. Then $\mathcal{S}$ is a local optima.*

Proof. Suppose $\mathcal{S}$ is not a local optima, then there exists an element x that satisfies one of the followings: $x \in \mathcal{S}$ and $F(\mathcal{S}\backslash\{x\}) \geq F(\mathcal{S})$ or $x \in \mathcal{V}\backslash\mathcal{S}$ and $F(\mathcal{S}\cup\{x\}) \geq F(\mathcal{S})$. This means the algorithm must not terminate with $\mathcal{S}$. Contradiction. □

A.2. Proof of Theorem 1

Before proving Theorem 1, we introduce submodularity index (SmI) which is a measure of the degree of submodularity (Zhou & Spanos, 2016).

Definition 4. *The submodularity index (Zhou & Spanos, 2016) for a set function $F : 2^{\mathcal{V}} \rightarrow \mathbb{R}$, a set $\mathcal{L}$, and a cardinality k is defined as*

$$\lambda_F(\mathcal{L},k) = \min_{\substack{\mathcal{A}\subseteq\mathcal{L}\\ \mathcal{S}\cap\mathcal{A}=\emptyset\\ |\mathcal{S}|\leq k}} \left\{ \phi_F(\mathcal{S},\mathcal{A}) \triangleq \sum_{x\in\mathcal{S}} F_x(\mathcal{A}) - F_{\mathcal{S}}(\mathcal{A}) \right\},$$

where $F_{\mathcal{S}}(\mathcal{A}) = F(\mathcal{A}\cup\mathcal{S}) - F(\mathcal{A})$.

It is easy to verify $\forall\mathcal{I} \subseteq \mathcal{J}$, SmI satisfies $\lambda_F(\mathcal{I},k) \geq \lambda_F(\mathcal{J},k)$ and for the optimal solution $\mathcal{C}$, $-2F(\mathcal{C}) \leq \lambda_F(\mathcal{V},2) \leq 2F(\mathcal{C})$. Following lemma bounds the degradation in submodularity with SmI.

Lemma 2. *Let $\mathcal{A}$ be an arbitrary set, $\mathcal{B} = \mathcal{A}\cup\{y_1, ..., y_M\}$ and $x \in \overline{\mathcal{B}}$. Then, $F_x(\mathcal{A}) - F_x(\mathcal{B}) \geq M\lambda_F(\mathcal{B},2)$*

Proof. See Zhou & Spanos (2016) Lemma 3. □

Lemma 3. *Let $\mathcal{Y}$ be an arbitrary set and $\mathcal{A} \subseteq \mathcal{B}$, Then*

$$\begin{aligned} &F(\mathcal{A}\cup\mathcal{Y}) - F(\mathcal{A}) \\ &\quad\geq F(\mathcal{B}\cup\mathcal{Y}) - F(\mathcal{B}) + |\mathcal{B}\setminus\mathcal{A}|\cdot|\mathcal{Y}|\cdot\lambda_F(\mathcal{B}\cup\mathcal{Y},2) \end{aligned}$$

Proof. Let $\mathcal{Y} = \{a_1, ..., a_n\}$. Then,

$$\begin{aligned} &F(\mathcal{A}\cup\{a_1\}) - F(\mathcal{A}) \\ &\qquad\geq F(\mathcal{B}\cup\{a_1\}) - F(\mathcal{B}) + |\mathcal{B}\setminus\mathcal{A}|\lambda_F(\mathcal{B},2) \\ &F(\mathcal{A}\cup\{a_1,a_2\}) - F(A\cup\{a_1\}) \\ &\qquad\geq F(\mathcal{B}\cup\{a_1,a_2\}) - F(\mathcal{B}\cup\{a_1\}) \\ &\qquad\quad + |\mathcal{B}\setminus\mathcal{A}|\lambda_F(\mathcal{B}\cup\{a_1\},2) \\ &\qquad\qquad\vdots \\ &F(\mathcal{A}\cup\mathcal{Y}) - F(\mathcal{A}\cup\mathcal{Y}\setminus\{a_n\}) \\ &\qquad\geq F(\mathcal{B}\cup\mathcal{Y}) - F(\mathcal{B}\cup\mathcal{Y}\setminus\{a_n\}) \\ &\qquad\quad + |\mathcal{B}\setminus\mathcal{A}|\lambda_F(\mathcal{B}\cup\mathcal{Y}\setminus\{a_n\},2) \end{aligned}$$

By telescoping sum,

$$\begin{aligned} &F(\mathcal{A}\cup\mathcal{Y}) - F(\mathcal{A}) \\ &\geq F(\mathcal{B}\cup\mathcal{Y}) - F(\mathcal{B}) + |\mathcal{B}\setminus\mathcal{A}|\sum_{i=1}^{n}\lambda_F(\mathcal{B}\cup\{a_1, ...a_{i-1}\},2) \\ &\geq F(\mathcal{B}\cup\mathcal{Y}) - F(\mathcal{B}) + |\mathcal{B}\setminus\mathcal{A}|\cdot|\mathcal{Y}|\cdot\lambda_F(\mathcal{B}\cup\mathcal{Y},2) \\ &\hspace{16em}\text{(By property of SmI)} \end{aligned}$$

□

Next lemma relates the local optima solution from local search with submodularity index.

Lemma 4. *If $\mathcal{S}$ is a local optima for a function F, then for any subsets $\mathcal{I} \subseteq \mathcal{S} \subseteq \mathcal{J}$, the following holds.*

$$F(\mathcal{I}) \leq F(\mathcal{S}) - \binom{|\mathcal{S}\setminus\mathcal{I}|}{2}\lambda_F(\mathcal{S},2)$$

$$F(\mathcal{J}) \leq F(\mathcal{S}) - \binom{|\mathcal{J}\setminus\mathcal{S}|}{2}\lambda_F(\mathcal{J},2)$$

Proof. Let $\mathcal{I} = \mathcal{T}_0 \subseteq \mathcal{T}_1 \subseteq \cdots \subseteq \mathcal{T}_k = \mathcal{S}$ be a chain of sets where $\mathcal{T}_i \setminus \mathcal{T}_{i-1} = \{a_i\}$. For each $1 \leq i \leq k$, the following holds.

[*]Equal contribution [1]Department of Computer Science and Engineering, Seoul National University, Seoul, Korea [2]Neural Processing Research Center. Correspondence to: Hyun Oh Song <hyunoh@snu.ac.kr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

$$\begin{aligned}
F(\mathcal{T}_i) - F(\mathcal{T}_{i-1}) &\geq F(\mathcal{S}) - F(\mathcal{S} \setminus \{a_i\}) + (k-i)\lambda_F(\mathcal{S} \setminus \{a_i\}, 2) \\
&\qquad \text{(By Lemma 3)} \\
&\geq (k-i)\lambda_F(\mathcal{S} \setminus \{a_i\}, 2) \\
&\qquad \text{(By the definition of local optima)} \\
&\geq (k-i)\lambda_F(\mathcal{S}, 2) \\
&\qquad \text{(By the property of SmI)}
\end{aligned}$$

By telescoping sum,

$$F(\mathcal{S}) - F(\mathcal{I}) \geq \sum_{i=1}^{k} (k-i)\lambda_F(\mathcal{S}, 2) = \binom{|\mathcal{S} \setminus \mathcal{I}|}{2} \lambda_F(\mathcal{S}, 2)$$

Similarly, Let $\mathcal{S} = \mathcal{T}_0 \subseteq \mathcal{T}_1 \subseteq \cdots \subseteq \mathcal{T}_k = \mathcal{J}$ be a chain of sets where $\mathcal{T}_i \setminus \mathcal{T}_{i-1} = \{a_i\}$. For each $1 \leq i \leq k$, the following holds.

$$\begin{aligned}
F(\mathcal{T}_i) - F(\mathcal{T}_{i-1}) &\leq F(\mathcal{S} \cup \{a_i\}) - F(\mathcal{S}) - (i-1)\lambda_F(\mathcal{T}_{i-1}, 2) \\
&\qquad \text{(By Lemma 3)} \\
&\leq -(i-1)\lambda_F(\mathcal{T}_{i-1}, 2) \\
&\qquad \text{(By the definition of local optima)} \\
&\leq -(i-1)\lambda_F(\mathcal{J}, 2) \\
&\qquad \text{(By the property of SmI)}
\end{aligned}$$

By telescoping sum,

$$\begin{aligned}
F(\mathcal{J}) - F(\mathcal{S}) &\leq -\sum_{i=1}^{k} (i-1)\lambda_F(\mathcal{J}, 2) \\
&= -\binom{|\mathcal{J} \setminus \mathcal{S}|}{2} \lambda_F(\mathcal{J}, 2)
\end{aligned}$$

□

Now, we prove Theorem 1.

Theorem 1. *Let $\mathcal{C}$ be an optimal solution for a function F and $\mathcal{S}$ be the solution obtained by the local search algorithm. Then,*

$$2F(\mathcal{S}) + F(\mathcal{V} \setminus \mathcal{S}) \geq F(\mathcal{C}) + \xi \lambda_F(\mathcal{V}, 2),$$

where

$$\xi = \binom{|\mathcal{S} \setminus \mathcal{C}|}{2} + \binom{|\mathcal{C} \setminus \mathcal{S}|}{2} + |\overline{\mathcal{S} \cup \mathcal{C}}| \cdot |\mathcal{S}| + |\mathcal{C} \setminus \mathcal{S}| \cdot |\mathcal{S} \cap \mathcal{C}|$$

Proof. Since $\mathcal{S}$ is a local optimum, The following holds by Lemma 4.

$$F(\mathcal{S}) \geq F(\mathcal{S} \cap \mathcal{C}) + \binom{|\mathcal{S} \setminus \mathcal{C}|}{2} \lambda_F(\mathcal{S}, 2)$$
$$F(\mathcal{S}) \geq F(\mathcal{S} \cup \mathcal{C}) + \binom{|\mathcal{C} \setminus \mathcal{S}|}{2} \lambda_F(\mathcal{S} \cup \mathcal{C}, 2)$$

Also from Lemma 3, we have,

$$\begin{aligned}
&F(\mathcal{S} \cup \mathcal{C}) + F(\mathcal{V} \setminus \mathcal{S}) \\
&\quad \geq F(\mathcal{C} \setminus \mathcal{S}) + F(\mathcal{V}) + |\overline{\mathcal{S} \cup \mathcal{C}}| \cdot |\mathcal{S}| \cdot \lambda_F(\mathcal{V}, 2) \\
&\quad \geq F(\mathcal{C} \setminus \mathcal{S}) + |\overline{\mathcal{S} \cup \mathcal{C}}| \cdot |\mathcal{S}| \cdot \lambda_F(\mathcal{V}, 2) \\
&\qquad \text{(By non-negativity)}
\end{aligned}$$

Also,

$$\begin{aligned}
&F(\mathcal{S} \cap \mathcal{C}) + F(\mathcal{C} \setminus \mathcal{S}) \\
&\quad \geq F(\mathcal{C}) + F(\emptyset) + |\mathcal{C} \setminus \mathcal{S}| \cdot |\mathcal{S} \cap \mathcal{C}| \cdot \lambda_F(\mathcal{C}, 2) \\
&\quad \geq F(\mathcal{C}) + |\mathcal{C} \setminus \mathcal{S}| \cdot |\mathcal{S} \cap \mathcal{C}| \cdot \lambda_F(\mathcal{C}, 2) \\
&\qquad \text{(By non-negativity)}
\end{aligned}$$

Summing the inequalities, we get

$$\begin{aligned}
2F(\mathcal{S}) + F(\mathcal{V} \setminus \mathcal{S}) \geq F(\mathcal{C}) &+ \binom{|\mathcal{S} \setminus \mathcal{C}|}{2} \lambda_F(\mathcal{S}, 2) \\
&+ \binom{|\mathcal{C} \setminus \mathcal{S}|}{2} \lambda_F(\mathcal{S} \cup \mathcal{C}, 2) \\
&+ |\overline{\mathcal{S} \cup \mathcal{C}}| \cdot |\mathcal{S}| \cdot \lambda_F(\mathcal{V}, 2) \\
&+ |\mathcal{C} \setminus \mathcal{S}| \cdot |\mathcal{S} \cap \mathcal{C}| \cdot \lambda_F(\mathcal{C}, 2)
\end{aligned}$$

Since all $\lambda_F(\cdot, 2)$'s are greater than or equal to $\lambda_F(\mathcal{V}, 2)$ by the property of SmI, we get

$$\begin{aligned}
2F(\mathcal{S}) + F(\mathcal{V} \setminus \mathcal{S}) &\geq F(\mathcal{C}) + \Bigg[\binom{|\mathcal{S} \setminus \mathcal{C}|}{2} + \binom{|\mathcal{C} \setminus \mathcal{S}|}{2} \\
&\quad + |\overline{\mathcal{S} \cup \mathcal{C}}| \cdot |\mathcal{S}| + |\mathcal{C} \setminus \mathcal{S}| \cdot |\mathcal{S} \cap \mathcal{C}| \Bigg] \lambda_F(\mathcal{V}, 2) \\
&= F(\mathcal{C}) + \xi \lambda_F(\mathcal{V}, 2)
\end{aligned}$$

□

B. Hyperparameters

B.1. Experiments on Cifar-10

Hyperparameters for NES and Bandits on Cifar-10 dataset in untargeted setting are shown in Table 1 and Table 2 respectively. Note that the hyperparameters are tuned in a setting where images are normalized in a scale of $[0, 1]$ to maintain consistency with the experiments on ImageNet dataset.

Hyperparameter	Value
σ for NES	0.001
n, size of each NES population	100
η, learning rate	0.01
β, momentum	0.9

Table 1. Hyperparameters for NES untargeted attack on Cifar-10.

Hyperparameter	Value
η, OCO learning rate	0.1
h, image learning rate	0.01
δ, bandit exploration	0.1
η, finite difference probe	0.1
tile size	16

Table 2. Hyperparameters for Bandits untargeted attack on Cifar-10.

B.2. Untargeted attacks on ImageNet

Hyperparameters for NES and Bandits on ImageNet dataset in untargeted setting are listed in Table 3 and Table 4. We use NES implementation from Ilyas et al. (2018b), since Ilyas et al. (2018a) conducted experiments only in the targeted setting.

Hyperparameter	Value
n, sample per step	100
η, finite difference probe	0.1
h, image learning rate	0.01

Table 3. Hyperparameters for NES untargeted attack on ImageNet.

Hyperparameter	Value
η, OCO learning rate	100
h, image learning rate	0.01
δ, bandit exploration	1.0
η, finite difference probe	0.1
tile size	50

Table 4. Hyperparameters for Bandits untargeted attack on ImageNet.

B.3. Targeted attacks on ImageNet

Hyperparameters for NES targeted attack on ImageNet dataset are shown in Table 5. All the hyperparameters except for momentum are referred from the original paper. For momentum, we tuned with range $\beta \in \{0.5, 0.7, 0.9\}$. The result of tuning momentum is in Table 6. We choose $\beta = 0.7$ which records the lowest average queries.

Hyperparameter	Value
σ for NES	0.001
n, size of each NES population	50
η, learning rate	0.01
β, momentum	0.7

Table 5. Hyperparameters for NES targeted attack on ImageNet.

Momentum	Success rate	Avg. queries	Med. queries
0.5	99.2%	16977	13375
0.7	99.7%	**16284**	**12650**
0.9	**99.8%**	16725	13525

Table 6. Result of tuning momentum for NES.

B.4. Untargeted attacks on ImageNet with smaller ϵ

Hyperparameters for NES and Bandit with smaller maximum perturbation are given in Table 7 and Table 8. Since we run the experiments in untargeted setting, we use NES implementation from Ilyas et al. (2018b).

Hyperparameter	Value	
	$\epsilon = 0.01$	$\epsilon = 0.03$
n, samples per step	100	100
η, finite difference probe	1	1
h, image learning rate	0.001	0.005

Table 7. Hyperparameters for NES untargeted attack on ImageNet with smaller ϵ.

Hyperparameter	Value	
	$\epsilon = 0.01$	$\epsilon = 0.03$
η, OCO learning rate	100	100
h, image learning rate	0.001	0.005
δ, bandit exploration	0.1	1
η, finite difference probe	1	1
tile size	50	50

Table 8. Hyperparameters for Bandits untargeted attack on ImageNet with smaller ϵ.

C. Tuning Bandits for targeted attack

In applying Bandits to targeted attack, we tuned for image learning rate h and OCO learning rate η. Other hyperparameters were set as the untargeted setting given by the authors. We performed grid search on the two hyperparameters, with range $h \in \{0.0001, 0.001, 0.005, 0.01, 0.05\}$ and $\eta \in \{1, 10, 100, 1000\}$. This sweep range covers the method's original untargeted setting, which is $h = 0.01$ and $\eta = 100$. Evaluation metrics were attack success rate and average queries. Results can be found below.

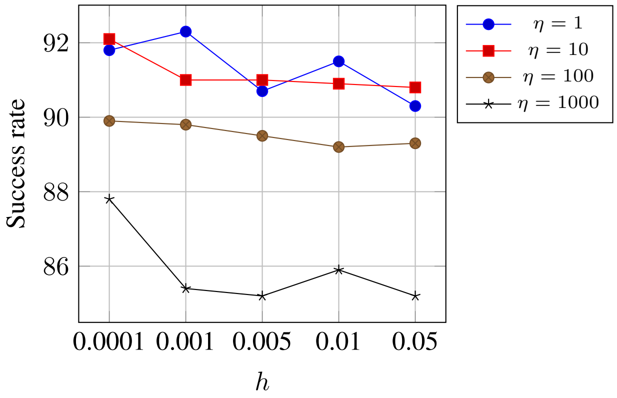

Figure 1. Success rate with given hyperparameters.

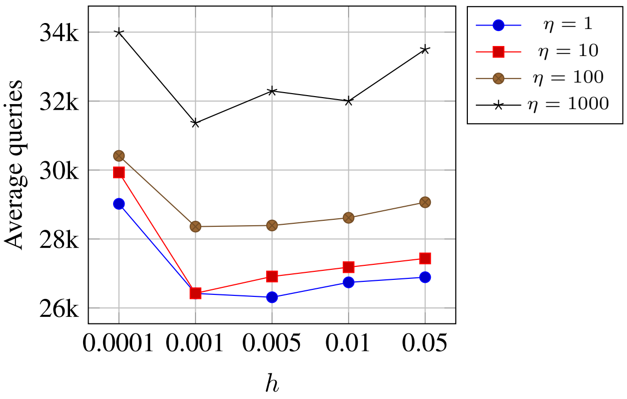

Figure 2. Average queries with given hyperparameters.

On the paper's Table 3 we used $h = 0.001$ and $\eta = 1$, which shows the best result on success rate with low average queries.

D. Additional plot on hyperparameter sensitivity analysis

To show the robustness of our method to hyperparameters more explicitly, we draw a mean and standard deviation plot of success rate against the number of queries across different hyperparameter settings for each attack method. The experimental protocol is the same as in Section 5.5 in the main text. The results are shown in Figure 3. The figure shows that our method is less sensitive to the hyperparameters than Bandits at every query limit.

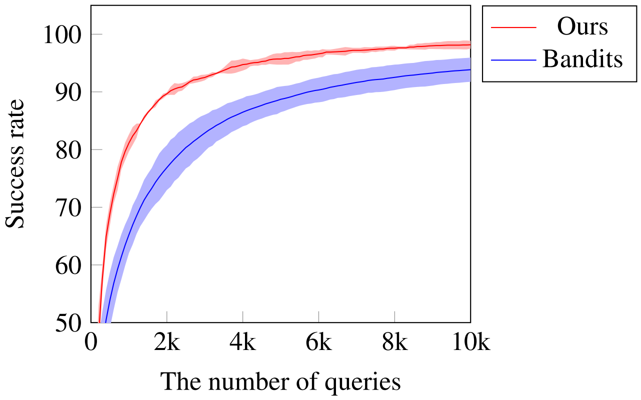

Figure 3. Mean and standard deviation plots of success rate against the number of queries across different hyperparamters. The solid lines show the average success rate (y-axis) at each query limit (x-axis).

References

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018a.

Ilyas, A., Engstrom, L., and Madry, A. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018b.

Zhou, Y. and Spanos, C. J. Causal meets submodular: Subset selection with directed information. In *NIPS*, 2016.