Appendix A. Solutions for different ℓ_p norms in z-step

In problem (11), we set $D(\boldsymbol{z}) = \|\boldsymbol{z}\|_2^2$ to measure the similarity between the legitimate image and the adversarial example. But $D(\boldsymbol{z})$ can also take other ℓ_p norms and the solutions in z-step can be obtained with minor modifications. In the following, we show the z-step solutions[3] for $D(\boldsymbol{z}) = \|\boldsymbol{z}\|_0$, $D(\boldsymbol{z}) = \|\boldsymbol{z}\|_1$, and $D(\boldsymbol{z}) = \|\boldsymbol{z}\|_1 + \frac{\beta}{2}\|\boldsymbol{z}\|_2^2$, derived from proximal operators which are applicable and well-suited to problems of substantial recent interest involving large or high-dimensional datasets.

A.1. Solutions for ℓ_0 norm

If $D(\boldsymbol{z}) = \|\boldsymbol{z}\|_0$, the solution to problem (11) can be obtained as follows,

$$[\mathbf{z}^{k+1}]_i = \begin{cases} \min\{1 - [\mathbf{x}_0]_i, \epsilon\} & \text{if } c_i > \min\{1 - [\mathbf{x}_0]_i, \epsilon\} \\ \max\{-[\mathbf{x}_0]_i, -\epsilon\} & \text{if } c_i < \max\{-[\mathbf{x}_0]_i, -\epsilon\} \\ c_i & \text{otherwise,} \end{cases} \tag{28}$$

where

$$c_i = \begin{cases} a_i & \text{if } a_i^2 > \frac{2\gamma}{\rho} \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

A.2. Solutions for ℓ_1 norm

If $D(\boldsymbol{z}) = \|\boldsymbol{z}\|_1$, the solution to problem (11) can be obtained as below,

$$[\mathbf{z}^{k+1}]_i = \begin{cases} \min\{1 - [\mathbf{x}_0]_i, \epsilon\} & \text{if } (a_i - \frac{\gamma}{\rho})_+ - (-a_i - \frac{\gamma}{\rho})_+ \\ & > \min\{1 - [\mathbf{x}_0]_i, \epsilon\} \\ \max\{-[\mathbf{x}_0]_i, -\epsilon\} & \text{if } (a_i - \frac{\gamma}{\rho})_+ - (-a_i - \frac{\gamma}{\rho})_+ \\ & < \max\{-[\mathbf{x}_0]_i, -\epsilon\} \\ (a_i - \frac{\gamma}{\rho})_+ & \\ -(-a_i - \frac{\gamma}{\rho})_+ & \text{otherwise,} \end{cases} \tag{30}$$

where $(x)_+ = x$ if $x \geq 0$ and 0 otherwise.

A.3. Solutions for combination of ℓ_1 and ℓ_2 norm

If $D(\boldsymbol{z}) = \|\boldsymbol{z}\|_1 + \frac{\beta}{2}\|\boldsymbol{z}\|_2^2$, which is also know as elastic net regularization, the solution to problem (11) can be obtained through,

$$[\mathbf{z}^{k+1}]_i = \begin{cases} \min\{1 - [\mathbf{x}_0]_i, \epsilon\} & \text{if } \frac{1}{1+\frac{\gamma\beta}{\rho}}((a_i - \frac{\gamma}{\rho})_+ \\ & -(-a_i - \frac{\gamma}{\rho})_+) \\ & > \min\{1 - [\mathbf{x}_0]_i, \epsilon\} \\ \max\{-[\mathbf{x}_0]_i, -\epsilon\} & \text{if } \frac{1}{1+\frac{\gamma\beta}{\rho}}((a_i - \frac{\gamma}{\rho})_+ \\ & -(-a_i - \frac{\gamma}{\rho})_+) \\ & < \max\{-[\mathbf{x}_0]_i, -\epsilon\} \\ \frac{1}{1+\frac{\gamma\beta}{\rho}}((a_i - \frac{\gamma}{\rho})_+ & \\ -(-a_i - \frac{\gamma}{\rho})_+) & \text{otherwise,} \end{cases} \tag{31}$$

Appendix B. Derivation for maximizing EI

EI can be transformed as follows,

$$\begin{aligned} \mathrm{EI}(\boldsymbol{\delta}) &\overset{l'=\frac{l(\boldsymbol{\delta})-\mu}{\sigma}}{=} \mathbb{E}_{l'}\left[(l^+ - l'\sigma - \mu)\mathcal{I}\left(l' \leq \frac{l^+ - \mu}{\sigma}\right)\right] \\ &= (l^+ - \mu)\Phi\left(\frac{l^+ - \mu}{\sigma}\right) - \sigma E_{l'}\left[l'\mathcal{I}\left(l' \leq \frac{l^+ - \mu}{\sigma}\right)\right] \\ &= (l^+ - \mu)\Phi\left(\frac{l^+ - \mu}{\sigma}\right) - \sigma\int_{-\infty}^{\frac{l^+ - \mu}{\sigma}} l'\phi(l')dl' \\ &= (l^+ - \mu)\Phi\left(\frac{l^+ - \mu}{\sigma}\right) + \sigma\phi\left(\frac{l^+ - \mu}{\sigma}\right), \end{aligned} \tag{32}$$

Appendix C. BO-ZO-ADMM

In BO-ZO-ADMM, BO is used to obtain a query-efficient attack solution (at early ADMM iterations) for initializing the ZO method, which can further minimize the adversarial distortion (at later ADMM iterations). Additional experiments showed that when reaching the same ℓ_2 distortion as ZO-ADMM, BO-ZO-ADMM requires 380 queries on MNIST and 320 queries on CIFAR-10, outperforming 493 and 421 queries in Table 1.

Appendix D. Comparison with AutoZoom and Boundary method

For the comparison with AutoZoom [39], we report the averaged number of queries for attacking 500images at the same ℓ_2 distortion level for MNIST, CIFAR-10,and ImageNet in Table A1. As we can see, the proposed ZO-ADMM method is more query-efficient, while it is worth noting that AutoZOOM produces adversarial perturbation in low-dimensional latent space,and thus saves more computation cost.

Table A1. Comparison to AutoZOOM in attack success rate (ASR) and query #.

	MNIST		CIFAR-10		ImageNet	
	ASR	# of Query	ASR	# of Query	ASR	# of Query
AutoZoom	100%	1821	99.2%	1639	98.3%	43547
ZO-ADMM	100%	562	99%	492	99%	16390

[3] We do not investigate the case of ℓ_∞ norm since the constraint $\|\boldsymbol{z}\|_\infty \leq \epsilon$ on the ℓ_∞ norm is already taken into consideration.

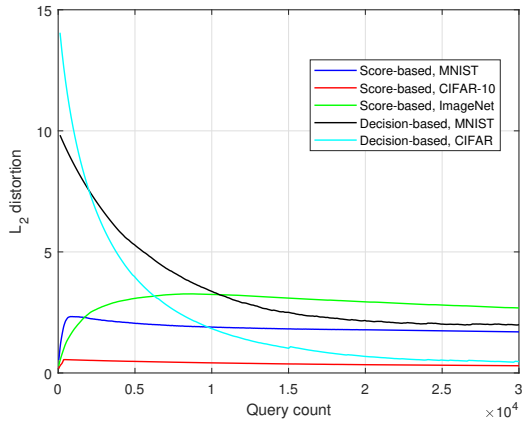
In Table A2, we show the comparison of ZO-ADMM method with the Query-limited [18] and Boundary methods [5] in terms of query number and ℓ_p norms on ImageNet.

Table A2. Experimental results on ImageNet

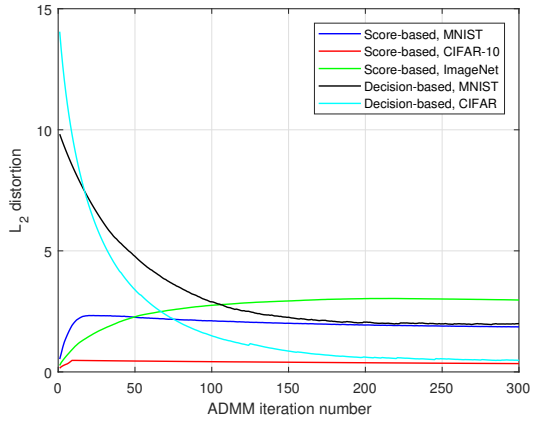
Settings	Methods	ASR	ℓ_1	ℓ_2	ℓ_∞	Query #
Score-based	Query-limited [18]	98%	1251	4.8	0.049	3.4×10^5
	ZO-ADMM	97%	785	3.5	0.039	1.6×10^5
Decision-based	Boundary [20]	85%	1120	3.99	0.045	2.2×10^6
	ZO-ADMM	93%	962	3.92	0.042	1.5×10^6

Appendix E. Convergence of the ZO-ADMM attack

Figure A1 shows the convergence of the ZO-ADMM attack v.s. query number or ADMM iteration number. Figure A2 shows the convergence comparison of the ZO-ADMM method and the Boundary method.



(a) ℓ_2 norm v.s. query number



(b) ℓ_2 norm v.s. ZO-ADMM iteration number]

Figure A1. Convergence of the ZO-ADMM attack.

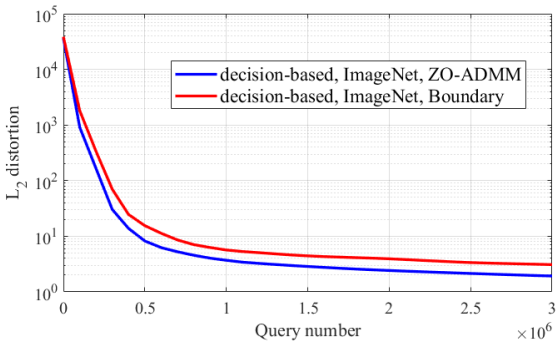


Figure A2. ℓ_2 distortion of decision-based attack vs queries on ImageNet.

Appendix F. Examples for the decision-based ZO-ADMM attack

In the following, we provide more adversarial examples generated by the proposed ZO-ADMM decision-based black-box attack.

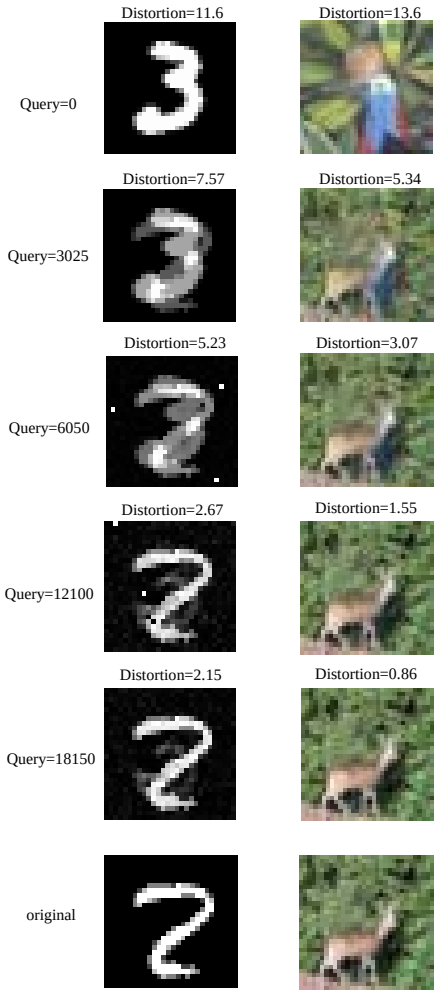


Figure A3. Adversarial examples generated by the proposed decision-based black-box attack with ZO-ADMM on MNIST and CIFAR-10.

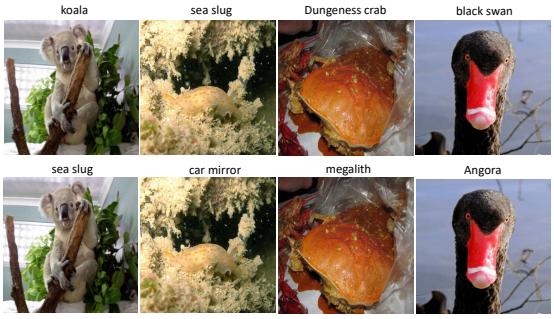


Figure A4. Adversarial examples on ImageNet. The original images are on the top row and their corresponding adversarial examples are shown on the bottom row with target labels.