GeoDA: a geometric framework for black-box adversarial attacks Supplementary Material

Ali Rahmati*, Seyed-Mohsen Moosavi-Dezfooli†, Pascal Frossard‡, and Huaiyu Dai*

*Department of ECE, North Carolina State University
†Institue for Machine Learning, ETH Zurich
‡Ecole Polytechnique Federale de Lausanne
arahmat@ncsu.edu, seyed.moosavi@inf.ethz.ch, pascal.frossard@epfl.ch, hdai@ncsu.edu

A. Proof of Lemma 2

Proof. Let X_i be a random vector taking values in $\mathbb{R}^d$ with mean $\boldsymbol{\mu} = \mathbb{E}[X]$ and covariance matrix $\boldsymbol{R} = \mathbb{E}(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^T$. Given the $X_1, \ldots, X_n$, the goal is to estimate $\boldsymbol{\mu}$. If X has a multivariate Gaussian or sub-Gaussian distribution, the sample mean $\bar{\boldsymbol{\mu}}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ is the result of MLE estimation, which satisfies, with probability at least $1 - \delta$

$$||\bar{\boldsymbol{\mu}}_N - \boldsymbol{\mu}|| \leq \sqrt{\frac{\text{Tr}(\boldsymbol{R})}{N}} + \sqrt{\frac{2\lambda_{\max}\log(1/\delta)}{N}} \qquad (1)$$

where $\text{Tr}(\boldsymbol{R})$ and $\lambda_{\max}$ denote the trace and largest eigenvalue of the covariance matrix $\boldsymbol{R}$, respectively [3]. We already know the truncated normal distribution mean and variance. Although, the truncated distribution is similar to Gaussian, we need to prove that it satisfies the sub-Gaussian distribution property so that we can use the bound in (1).

The truncated distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{R}$ is a sub-Gaussian distribution. A given distribution is sub-Gaussian if for all unit vectors $\{\boldsymbol{v} \in \mathbb{R}^d : ||\boldsymbol{v}|| = 1\}$ [5], the following condition holds

$$\mathbb{E}\left[\exp(\lambda\langle \boldsymbol{v}, X - \boldsymbol{\mu}\rangle)\right] \leq \exp(c\lambda^2\langle \boldsymbol{v}, \boldsymbol{\Sigma}\boldsymbol{v}\rangle). \qquad (2)$$

Assuming the hyperplane $\boldsymbol{w}^T X \geq 0$ truncated Normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$, the left hand side of the (13) can be computed as:

$$\mathbb{E}\left[\exp(\lambda\langle \boldsymbol{v}, X - \boldsymbol{\mu}\rangle)\right] = \int_{\mathcal{H}^+} \exp(\lambda \boldsymbol{v}^T X)\phi_d(X|\boldsymbol{\Sigma})dX \qquad (3)$$

where $\mathcal{H}^+ = \{X \in \mathbb{R}^d : \boldsymbol{w}^T X \geq 0\}$. Since $\boldsymbol{R}$ is a symmetric, positive definite matrix, using Cholesky decomposition we can have $\boldsymbol{R}^{-1} = \boldsymbol{\Psi}^T\boldsymbol{\Psi}$ where $\boldsymbol{\Psi}$ is a non-singular, upper triangular matrix [4]. By transforming the variables, we have $Y = \boldsymbol{\Psi}X$. Using Y, with some manipulation as in [7], one can get

$$\mathbb{E}\left[\exp(\lambda\langle \boldsymbol{v}, X - \boldsymbol{\mu}\rangle)\right] = \exp\left(\frac{1}{2}\lambda^2 \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}\right)\Phi\left[\frac{\lambda \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}}{\sigma}\right] \qquad (4)$$

and $\sigma^2 = \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}$, and $\Phi[.]$ is the cumulative distribution function of the univariate normal distribution. Plugging $\boldsymbol{\Sigma} = \mathcal{I}$, one can get

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda\langle \boldsymbol{v}, X\rangle)\right] &= \exp\left(\frac{1}{2}\lambda^2\right)\Phi\left[\lambda \boldsymbol{w}^T\boldsymbol{v}\right] \\ &\leq \exp\left(\frac{1}{2}\lambda^2\right), \end{aligned} \qquad (5)$$

where the inequality is valid due to the fact that the CDF function is equal to 1 in the maximum. Comparing with the right hand side of the (13):

$$\exp\left(\frac{1}{2}\lambda^2\right) \leq \exp\left(\frac{1}{2}c\lambda^2\right), \qquad (6)$$

one can see that it is valid for any $c \geq 1$. Thus, the truncated Normal distribution is a sub-Gaussian distribution. □

The above proof is consistent with our intuition as the truncated Gaussian has the tails approaching zero at least as fast as exponential distribution. The truncated part of the Gaussian is already equal to zero so there is no chance for being a heavy tailed distribution. Thus, the bound provided in (1) can be valid for our problem [5].

Since the covariance matrix $\boldsymbol{R}$ is unknown, we need to find bounds for $\text{Tr}(\boldsymbol{R})$ and $\lambda_{\max}$ as well. It can easily be obtained that

$$\text{Tr}(\boldsymbol{R}) = d + c_2\boldsymbol{w}^T\boldsymbol{w} = d + c_2 \qquad (7)$$

In order to obtain the maximum eigenvalue of the $\boldsymbol{R}$, we use Weyl's inequality to have an upper bound for largest

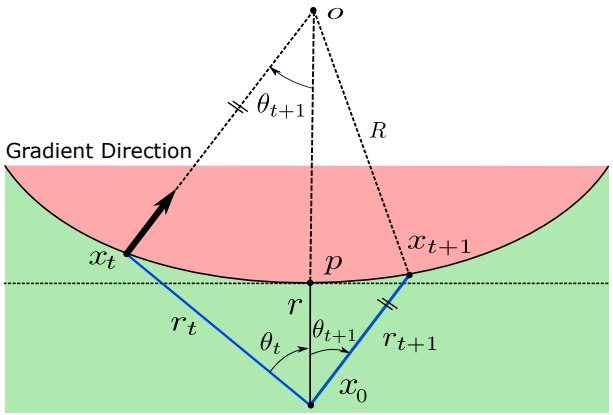

Figure 1: Convex decision boundary with bounded curvature.

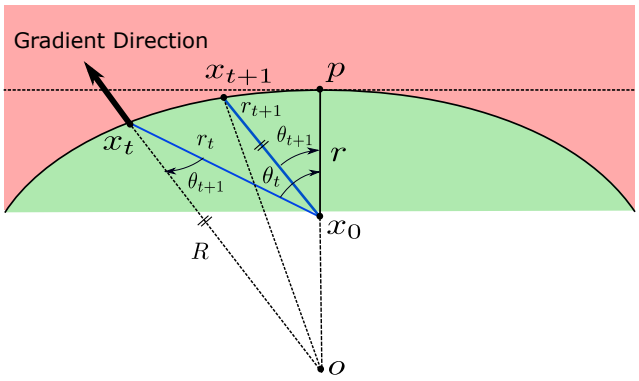

Figure 2: Concave decision boundary with bounded curvature.

eigenvalue of the covariance matrix as [6]:

$$\lambda_{\max}(\boldsymbol{A}+\boldsymbol{B}) \leq \lambda_{\max}(\boldsymbol{A}) + \lambda_{\max}(\boldsymbol{B}) \tag{8}$$

The largest eigenvalue for the identity matrix $\mathcal{I}$ is 1. For the rank-1 matrix $c_2\boldsymbol{w}\boldsymbol{w}^T$ which is the outer product of the normal vector is given by:

$$\lambda_{\max}(c_2\boldsymbol{w}\boldsymbol{w}^T) = c_2\mathrm{Tr}(\boldsymbol{w}\boldsymbol{w}^T) = c_2\boldsymbol{w}^T\boldsymbol{w} = c_2 \tag{9}$$

which immediately results in $\lambda_{\max}(\boldsymbol{R}) \leq 1 + c_2$. Substituting the above values to the (12), the sample mean $\bar{\boldsymbol{\mu}}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ is the result of MLE estimation, which satisfies, with probability at least $1-\delta$

$$\|\bar{\boldsymbol{\mu}}_N - \boldsymbol{\mu}\| \leq \sqrt{\frac{d+c_2}{N}} + \sqrt{\frac{2(1+c_2)\log(1/\delta)}{N}} \tag{10}$$

This bound can provide an upper bound with probability at least $1-\delta$ for the error of the sample mean while getting N queries from the neural network.

$$\frac{R}{\sin(\theta_t)} = \frac{r_t}{\sin(\theta_{t+1})} \tag{11}$$

B. Proof of Theorem 1

In the following subsections, we consider two cases for the curvature of the boundary.

Convex Curved Bounded Boundary

We assume that the curvature of the boundary is convex as given in Fig. 1. As given in [2], if θ_t satisfies the two assumptions $\tan^2(\theta_t) \leq 0.2R/r$ and $r/R < 1$, the value for $\|\boldsymbol{x}_t - \boldsymbol{x}_0\|_2 = r_t$ is given as follows:

$$r_t = -(R-r)\cos(\theta_t) + \sqrt{(R-r)^2\cos^2(\theta_t) + 2Rr - r^2} \tag{12}$$

where $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_0\|_2 = r_{t+1}$ can be obtained in a similar way. It can be observed that the value of the r_t is an increasing function of the θ_t because:

$$\frac{\partial r_t}{\partial \theta_t} = (R-r)\sin(\theta_t) - \frac{(R-r)^2\cos(\theta_t)\sin(\theta_t)}{\sqrt{(R-r)^2\cos^2(\theta_t) + 2Rr - r^2}}, \tag{13}$$

Setting $\frac{\partial r_t}{\partial \theta_t} > 0$, with some manipulations one can get $2R > r$ which shows that r_t is an increasing function of the θ_t. Thus, if we can show that $\theta_t > \theta_{t+1}$, it means that $r_t > r_{t+1}$ which means that r_t can converge to r. Here, we assume that the given image is in the vicinity of the boundary $r/R < 1$. The line connecting point o to x_0 intersects the two parallel lines. Based on the law of sines, one can get Since $r_t < R$, one can conclude that $\theta_t > \theta_{t+1}$ using the sines law. Thus, as r_t is an increasing function of θ_t, we can get $r_{t+1} < r_t$. Thus, after several iterations, the following update rule

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + r_t\hat{\boldsymbol{w}}_{N_t} \tag{14}$$

converges to the minimum perturbation r.

Applying the sine law for k iterations, one can get the following equation using (11):

$$\sin(\theta_t) = \frac{\prod_{i=0}^{t} r_i}{R^t}\sin(\theta_0) \tag{15}$$

We know that $r_t < R$ and in each iteration, it gets smaller and smaller. Thus, for the convergence, we consider the worst case. We know that $\max_{i=0,1,\ldots,t}\{r_i\} = r_t$. Thus, To bound this, we can have:

$$\sin(\theta_{t+K}) = \frac{\prod_{k=0}^{K} r_{t+k}}{R^K}\sin(\theta_t) \leq (\frac{r_t}{R})^K\sin(\theta_t) \tag{16}$$

where can be reduced to

$$\sin(\theta_{t+K}) \leq (\frac{r_t}{R})^K\sin(\theta_t) \tag{17}$$

This shows that $\sin(\theta_{t+K})$ converges to zero exponentially since $r_t < R$. Thus, θ_{t+k} goes to zero which results that the in coinciding the r_t and r in the same magnitude. Thus, we have

$$\lim_{k\to\infty} r_{t+k} = r \tag{18}$$

We already know that

$$\begin{aligned} r_{t+1} =& -(R-r)\cos(\theta_{t+1}) \\ &+ \sqrt{(R-r)^2\cos^2(\theta_{t+1}) + 2Rr - r^2}, \end{aligned} \tag{19}$$

Considering the cosine law, based on the figure, we can see that

$$r_t^2 = (R+r)^2 + R^2 - 2R(R+r)\cos(\theta_{t+1}) \tag{20}$$

By combining the above equations and eliminating the $\cos(\theta_{t+1})$, one can get:

$$\begin{aligned} r_{t+1} =& -(R-r)\frac{(R+r)^2 + R^2 - r_t^2}{2R(R+r)} \\ &+ \sqrt{\frac{(R-r)^2((R+r)^2+R^2-r_t^2)^2}{4R^2(R+r)^2} + 2Rr - r^2}, \end{aligned} \tag{21}$$

Plugging (21) into the following limit,

$$\lim_{t\to\infty} \frac{r_{t+1} - r}{r_t - r} \tag{22}$$

for $t \to \infty$, we get $\frac{0}{0}$. Thus, using the L'Hospital's Rule, we take the derivative of the numerator and the denominator as:

$$\begin{aligned} \frac{\partial r_{t+1}}{\partial r_t} =& -\frac{r_t(R-r)}{R(R+r)} \\ &+ \frac{((R+r)^2 + R^2 - r_t^2)r_t}{2\sqrt{\frac{(R-r)^2((R+r)^2+R^2-r_t^2)^2}{4R^2(R+r)^2} + 2Rr - r^2}} \frac{(R-r)^2}{R^2(R+r)^2} \end{aligned} \tag{23}$$

Having $t \to \infty$, we can get $r_t \to r$, since we have $\hat{r}_t \to r_t$, thus:

$$\lim_{t\to\infty} \frac{\hat{r}_{t+1} - r}{\hat{r}_t - r} = \frac{r^2(R-r)}{R^2(R+r)} = \lambda < 1 \tag{24}$$

As $r < R$, the rate of convergence $\lambda \in (0, 1)$ which completes the proof.

Concave Curved Bounded Boundary

As in [2], the value for $\|\boldsymbol{x}_t - \boldsymbol{x}_0\|_2 = r_t$ is given as follows:

$$r_t = (R+r)\cos(\theta_t) - \sqrt{(R+r)^2\cos^2(\theta_t) - 2Rr - r^2} \tag{25}$$

where $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_0\|_2 = r_{t+1}$ can be obtained in a similar way. It can easily be seen that the $\theta_t > \theta_{t+1}$. Assuming $r/R < 1$, r_t is a decreasing function with respect to θ_t which results in $r_t < r_{t+1}$. Similar proof of convergence can be obtained for this case as well.

C. Proof of Theorem 2

Given the point r_{t-1}, the goal is to find the estimate of the $\hat{r}_t$ with limited query. Assuming the normalized version of the true gradient $\boldsymbol{w}_t = \boldsymbol{\mu}_t/\|\boldsymbol{\mu}_t\|_2$, we have

$$\|\hat{\boldsymbol{w}}_{N_t} - \boldsymbol{w_t}\| \leq \frac{\gamma}{\sqrt{N_t}} \tag{26}$$

where $\gamma = \sqrt{\mathrm{Tr}(\boldsymbol{R})} + \sqrt{2\lambda_{\max}\log(1/\delta)}$, $\hat{\boldsymbol{w}}_{N_t}$ is the estimated gradient at iteration t and N_t is the number of queries to estimate the gradient at point $\boldsymbol{x}_{t-1}$. Based on the reverse triangle inequality $\|\boldsymbol{x}\| - \|\boldsymbol{y}\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$, we can have

$$\|\hat{\boldsymbol{w}}_{N_t}\| - 1 \leq \|\hat{\boldsymbol{w}}_{N_t} - \boldsymbol{w_t}\| \leq \frac{\gamma}{\sqrt{N_t}}. \tag{27}$$

Multiplying by r_t, we have:

$$r_t - \frac{\gamma r_t}{\sqrt{N_t}} \leq \hat{r}_t \leq r_t + \frac{\gamma r_t}{\sqrt{N_t}}. \tag{28}$$

where $\hat{r}_t = r_t\|\hat{\boldsymbol{w}}_{N_t}\|$. Here, we conduct the analysis in the limit sense and we observe in the simulations that it is valid in limited iterations as well. Given r_{t-1}, for large t, we have:

$$r_t - r \approx \lambda(r_{t-1} - r) \tag{29}$$

Considering the best and worst case for the estimated gradient, we can find the following bound. In particular, the best case is the case in which all the gradient errors are constructive and make the $\hat{r}_t$ in each iteration smaller than r_t. In contrast, the worst case happens when all the gradients directions are destructive and make the $\hat{r}_t$ greater than r_t. In practice, however, what is happening is something in between. Substituting r_t from (28) in (29), one can obtain:

$$\lambda(r_{t-1} - r) - \frac{\gamma r_t}{\sqrt{N_t}} \leq \hat{r}_t - r \leq \lambda(r_{t-1} - r) + \frac{\gamma r_t}{\sqrt{N_t}} \tag{30}$$

By using the iterative equation, one can get the following bound:

$$\lambda^t(r_0 - r) - e(\boldsymbol{N}) \leq \hat{r}_t - r \leq \lambda^t(r_0 - r) + e(\boldsymbol{N}) \tag{31}$$

where $e(\boldsymbol{N}) = \gamma\sum_{i=1}^{t} \frac{\lambda^{t-i} r_i}{\sqrt{N_i}}$ is the error due to limited number of queries.

D. Proof of Theorem 3

It can easily be observed that the optimization problem is convex. Thus, the duality gap between this problem and its dual optimization problem is zero. Therefore, we can solve the given problem by solving its dual problem. The Lagrangian is given by:

$$\mathcal{L}(\boldsymbol{N}, \alpha) = \sum_{i=1}^{T} \frac{\lambda^{-i} r_i}{\sqrt{N_i}} + \alpha\left(\sum_{i=1}^{T} N_i - N\right) \tag{32}$$

where α is the non-negative dual variable associated with the budget constraint. The KKT conditions are given as follows [1]:

$$\frac{\partial \mathcal{L}(\boldsymbol{N}, \alpha)}{\partial N_t} = 0, \ \forall i \tag{33}$$

$$\alpha \left(\sum_{i=0}^{T} N_i - N \right) = 0 \tag{34}$$

$$\sum_{i=1}^{T} N_i \leq N \tag{35}$$

Based on (33), taking the derivative and setting equal to zero, we can have

$$N_t = \left(\frac{\lambda^{-t} r_t}{2\alpha} \right)^{\frac{2}{3}} \tag{36}$$

We see that the constraint holds with equality. Assume that $\sum_{i=0}^{t} N_i \neq N$, then based on (34), $\alpha = 0$. If $\alpha = 0$ then based on (36), we have $N_i = \infty$, $\forall i$ which contradicts with (35). Substituting (36) in $\sum_{i=0}^{t} N_i = N$, the Lagrangian multiplier can be obtained as

$$\alpha^{\frac{2}{3}} = \frac{1}{2^{\frac{2}{3}}} \frac{\sum_{i=1}^{T} (\lambda^{-i} r_i)^{\frac{2}{3}}}{N} \tag{37}$$

Substituting α in (36), one can get the optimal number of queries as:

$$N_t^* = \frac{(\lambda^{-t} r_t)^{\frac{2}{3}}}{\sum_{i=1}^{T} (\lambda^{-i} r_i)^{\frac{2}{3}}} N \tag{38}$$

For $t \to \infty$, we have $r_t \to r$. Based on this, the ratio of the optimal number if queries for each iteration is given by:

$$N_t^* \approx \frac{\lambda^{-\frac{2}{3}t}}{\sum_{i=1}^{T} \lambda^{-\frac{2}{3}i}} N \tag{39}$$

This equation shows that the distribution of the queries should be increased by a factor of $\lambda^{-\frac{2}{3}}$ where $0 < \lambda < 1$. By approximation, we have

$$\frac{N_{t+1}^*}{N_t^*} \approx \lambda^{-\frac{2}{3}} \tag{40}$$

which completes the proof.

1. Additional experiment results

Here we show more experiments on the performance of GeoDA on different ℓ_p norms. In Figs. 3, Figs. 4, Figs. **??**, and Figs. **??**, we have generated adversarial examples using GeoDA. For each image, the first row consists of (from left to right) original image, ℓ_2 fullspace adversarial example, ℓ_2 subspace adversarial example, ℓ_∞ fullspace adversarial example, ℓ_∞ subspace adversarial example, and ℓ_1 adversarial example, respectively. However, as can be seen the perturbations are not quite visible in the actual adversarial examples in the first row. In the second row, we show the magnified version of perturbations for ℓ_2 and ℓ_∞. To do so, the norm of all the perturbations is magnified to 100 given that the images coordinate normalized to the 0 to 1 scale. For the sparse case, we do not magnify the perturbations as they are visible and equal to their maximum (minimum) values. Finally, in the third row, we added a magnified version of the perturbation with norm of 30 to have a better visualization.

	Queries	**ResNet-50**	**ResNet-101**
GeoDA (ℓ_2)	500	11.76	17.91
	2000	3.35	6.38
	10000	1.06	1.87

Table 1: The performance comparison of GeoDA on different ResNet image classifiers.

In Table 1, we have compared the performance of GeoDA with different deep network image classifiers. The proposed algorithm GeoDA follows almost the same trend on a wide variety of deep networks. The reason is that the core assumption of GeoDA, i.e. boundary has a low mean curvature in vicinity of the datapoints, is verified empirically for a wide variety of deep networks. We can provide the experimental results on different networks.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 4

[2] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pages 1632–1640, 2016. 2, 3

[3] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971. 1

[4] Nicholas J Higham. *Analysis of the Cholesky decomposition of a semi-definite matrix*. Oxford University Press, 1990. 1

[5] Gábor Lugosi, Shahar Mendelson, et al. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019. 1

[6] Rachid Marsli. Bounds for the smallest and largest eigenvalues of hermitian matrices. *International Journal of Algebra*, 9(8):379–394, 2015. 2

[7] GM Tallis. Plane truncation in normal populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 27(2):301–307, 1965. 1

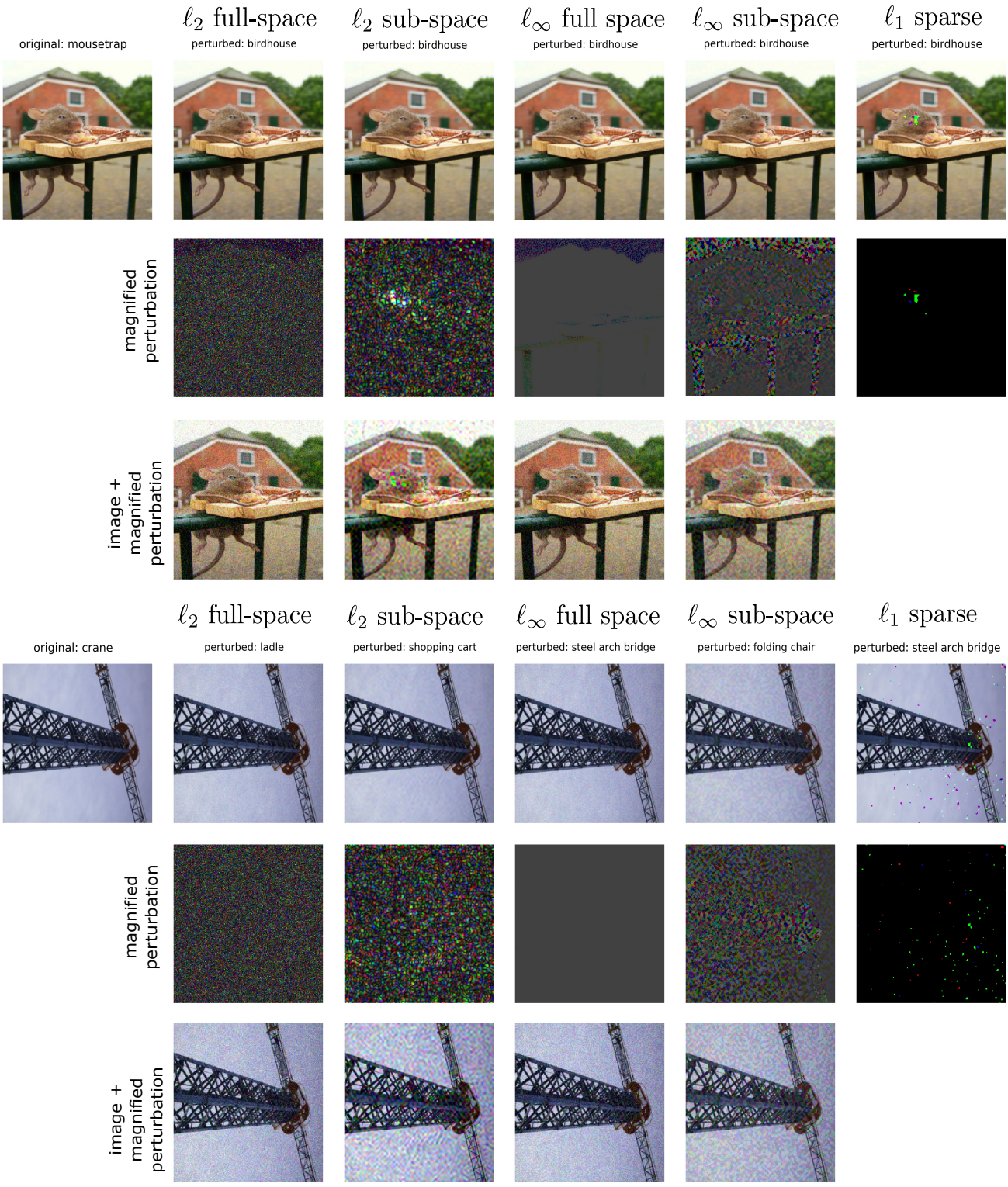

Figure 3: Original images and adversarial perturbations generated by GeoDA for ℓ_2 fullspace, ℓ_2 subspace, ℓ_∞ fullspace, ℓ_∞ subspace, and ℓ_1 sparse with $N = 10000$ queries.

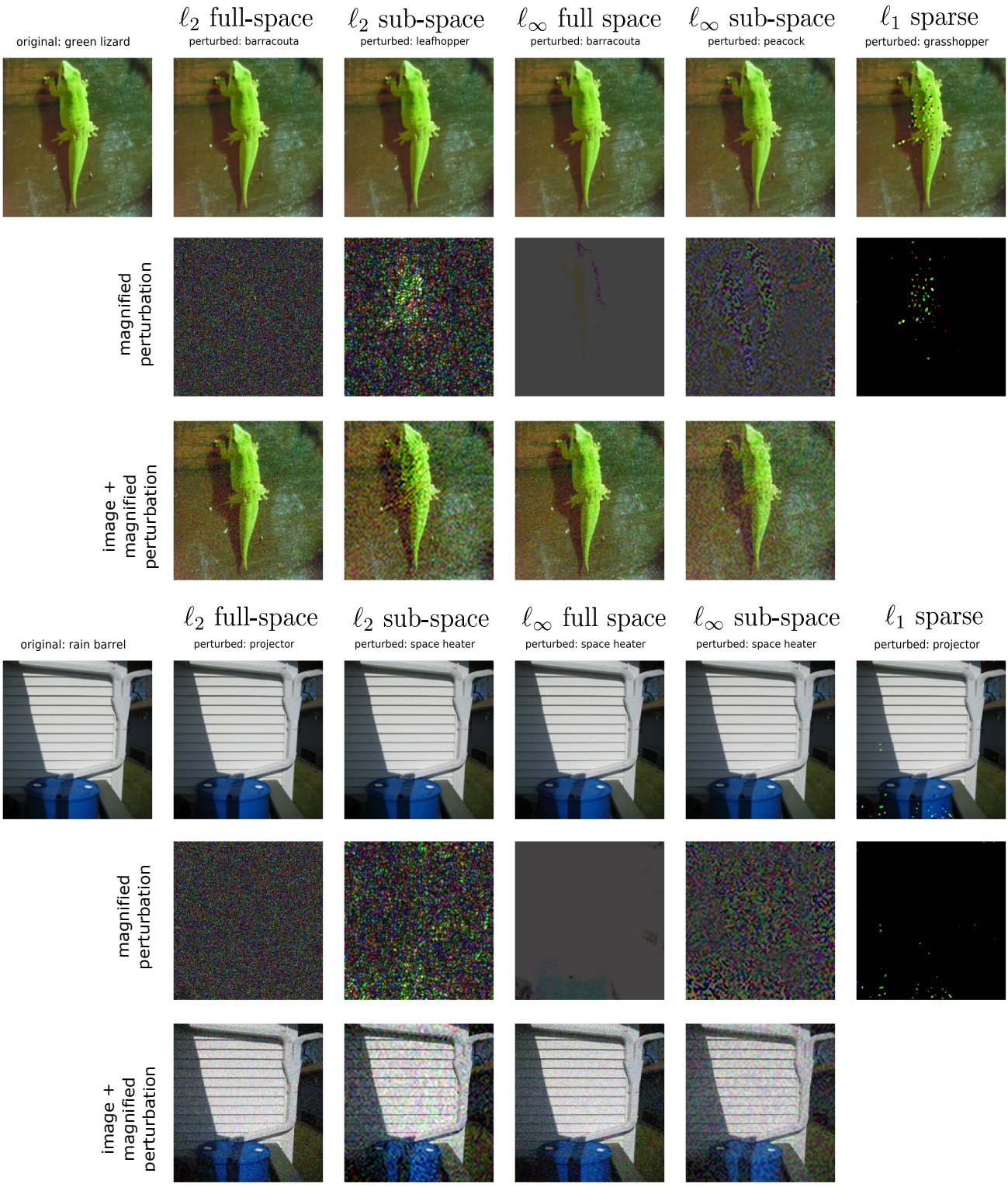

Figure 4: Original images and adversarial perturbations generated by GeoDA for ℓ_2 fullspace, ℓ_2 subspace, ℓ_∞ fullspace, ℓ_∞ subspace, and ℓ_1 sparse with $N = 10000$ queries.