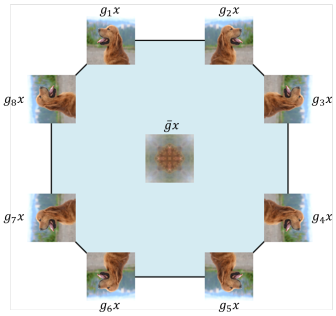
Journal of Machine Learning Research 1–13

On the Reconstruction of Training Data from Group Invariant Networks

Abstract

Reconstructing training data from trained neural networks is an active area of research with significant implications for privacy and explainability. Recent advances have demonstrated the feasibility of this process for several data types. However, reconstructing data from group-invariant neural networks poses distinct challenges that remain largely unexplored. This paper addresses this gap by first formulating the problem and discussing some of its basic properties. We then provide an experimental evaluation demonstrating that conventional reconstruction techniques are inadequate in this scenario. Specifically, we observe that the resulting data reconstructions gravitate toward symmetric inputs on which the group acts trivially, leading to poor-quality results. Finally, we propose two novel methods aiming to improve reconstruction in this setup and present promising preliminary experimental results. Our work sheds light on the complexities of reconstructing data from group invariant neural networks and offers potential avenues for future research in this domain.

1. Introduction



Recent works (Haim et al., 2022; Oz et al., 2024; Loo et al., 2023) have shown that it is possible to reconstruct training data from simple neural networks. However, the reconstruction from group invariant neural networks, such as networks applied to point clouds (Zaheer et al., 2017; Qi et al., 2017), graph data (Gilmer et al., 2017) or images with rotation and reflection symmetries (Cohen and Welling, 2016), remains largely unexplored.

Unlike the standard case explored in previous works, reconstructing data from group invariant models faces the unique challenge of multiple distinct inputs representing the same data point (namely, the orbit of a data point). This paper addresses the task of reconstructing data from group invariant networks, focusing on the limitations of conventional methods and proposing novel solutions. Our key contributions include:

1. A formal definition of the reconstruction problem for invariant networks, accompanied by a discussion of its fundamental properties and invariance characteristics.

2. Empirical evidence demonstrating that conventional reconstruction methods converge to symmetric inputs (i.e., inputs on which the group acts trivially), producing low-quality reconstructions. The inset illustrates this phenomenon for $G = D_4$: rather than recovering an element from a training example's orbit, conventional methods often converge to the (symmetric) orbit average.

3. Introduction of two novel techniques that enhance standard methods, enabling them to go beyond symmetric reconstructions with encouraging initial results.

4. A discussion of potential future research directions.

2. Preliminaries

Invariance and equivariance. Let (V, ρ),(V', ρ') be group representations of a finite group G. We denote $orb(\boldsymbol{x})$ as the orbit of a vector $\boldsymbol{x} \in V$ under the group action and $Stab(\boldsymbol{x})$ as its stabilizer group. The orbitope of a point $\boldsymbol{x}$ (Sanyal et al. (2011)) is defined as the convex hull of $orb(\boldsymbol{x})$ as illustrated in Figure 1. We denote the vectors on which the group acts trivially as V^G, formally $V^G = \{\boldsymbol{v} \in V | \rho(g)\boldsymbol{v} = \boldsymbol{v}, \forall g \in G\}$. The projection of a vector $\boldsymbol{x}$ on V^G, denoted as $\bar{g}x$, is the average of its orbit, $\bar{g} \cdot \boldsymbol{x} = \frac{1}{|G|} \sum_{g \in G} \rho(g) \cdot \boldsymbol{x}$. A function $f : V \to V'$ is G-invariant if $f \circ \rho(g) = f$ for any $g \in G$ and a function $F : V \to V'$ is G-equivariant if $F \circ \rho(g) = \rho'(g) \circ F$ for any $g \in G$. For simplicity, we denote $\rho(g)\boldsymbol{x} = g\boldsymbol{x}$.

Data Reconstruction. There are several methods to reconstruct training data from trained neural networks $\phi(\boldsymbol{x}; \theta)$, where $\boldsymbol{x} \in \mathbb{R}^d$ is the network's input, and θ is a vectorization of its parameters. Here, we focus on two methods: (1) Activation Maximization (AM) (Fredrikson et al., 2015; Yang et al., 2019), where the goal is to look for the input, which maximizes the model output for the desired target class. Namely, the objective for class i is defined as $\mathcal{L}_{rec} = \max_{\boldsymbol{x} \in \mathbb{R}^d} (\phi(\boldsymbol{x}; \theta))_i$. (2) KKT-based reconstruction (Haim et al., 2022; Buzaglo et al., 2023). This method uses the implicit bias of homogeneous neural networks trained with gradient methods toward margin maximization. Here, the following objective is optimized: $\mathcal{L}_{rec}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m, \lambda_1, \ldots, \lambda_m) = \|\theta - \sum_{i=1}^m \lambda_i y_i \phi(\boldsymbol{x}_i; \theta)\|$, (The $y_i's$ denote the labels.). For a detailed description of the reconstruction methods, see Appendix B.

3. Reconstruction from Invariant networks

3.1. Problem definition

Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \{\pm 1\}$ be a training dataset and let $(\mathbb{R}^d, \rho)$ is an orthogonal representation of a finite group G. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a pre-trained G-invariant neural network with weights $\theta \in \mathbb{R}^p$. We aim to find a set of reconstructions $\mathcal{S} = \{\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_m\} \subset \mathbb{R}^d$ such that $\mathcal{S}$ is the closest set to $\mathcal{D}$ with respect to some evaluation metric.

Evaluation under group symmetries. Since the reconstruction of any element from the orbit is equally valid, the evaluation metric for invariant data reconstruction problems should be invariant to these group symmetries to ensure a fair assessment of model performance. This can be done, for example, by defining the distance between a reconstruction x and a training example x' using the following metric $d(x, x') = \min_{g \in G} \|x - gx'\|$. Note that invariant metrics may be computationally challenging, e.g. when the input is a graph.

Applying AM and KKT to the invariant case. In most cases the training of (homogeneous) invariant neural networks is conducted in a way that the conditions of both methods (AM and KKT-based) are met, so we can apply them to the invariant case.

3.2. Challenges and theoretical observations

Here we present basic theoretical findings and challenges in reconstructing data from invariant models, applicable to any orthogonal representation of a finite group. Full proofs are in the appendix.

(1) Multiple equivalent solutions. When reconstructing data from invariant models, a critical factor to consider is that each training sample can have multiple equivalent representations. These representations form what is known as an orbit under the group action.

Proposition 1 *If the model is G-invariant then the objective functions of the methods mentioned in Section 2 are G-invariant. Formally,*

$$\mathcal{L}_{rec}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m) = \mathcal{L}_{rec}(g_1 \cdot \boldsymbol{x}_1, \dots, g_m x_m),\ \forall (x_1, g_1), \dots, (x_m, g_m) \in \mathbb{R}^d \times G \tag{1}$$

(2) Optimizing invariant reconstruction objectives using GD. As invariant reconstruction objectives have multiple optima, both initialization and optimization methods are crucial in determining the final solution. The following proposition sheds light on GD's behavior in this context:

Proposition 2 *If the reconstruction objective function $\mathcal{L}_{rec}(\boldsymbol{x}; \theta)$ is G-invariant function then: (i) the GD step $x_t = x_{t-1} - \eta_t \nabla_x \mathcal{L}_{rec}(x_{t-1})$ is G-equivariant function of x_{t-1}; and (ii) $Stab_G(x_{t-1}) \subseteq Stab_G(x_t)$*

First, the above part (i) implies that GD is an equivariant function of the initialization, as it is a composition of equivariant functions (GD iterations). Therefore, the initialization determines which element the method converges to. Moreover, it implies that if we use invariant distribution for the initialization ($P(x_0)$ is a G-invariant function) the algorithm induces an invariant distribution over the reconstructions. Moreover, the nesting of the stabilizers mentioned in part (ii) of Proposition 2 indicates that as optimization progresses, the stabilizers of the iterates may become more restrictive, thereby narrowing the exploration of the solution space. As we will see in the following sections, we believe that this property plays a significant role in the dynamics of optimization and can influence the final outcomes.

3.3. Ineffectiveness of standard methods in Invariant Reconstruction

This subsection presents experimental evidence demonstrating the ineffectiveness of AM and KKT-based methods in solving the invariant reconstruction problem. The full experimental results and description are on the appendix.

Setup and evaluation. We focus on the reconstruction of image data from invariant models of different groups of reflections and rotations (see Section E). To evaluate the results we used DSSIM proposed in Baker et al. (2023) for measuring differences between images (high DSSIM implies high structural dissimilarity).To ensure the invariance of our metric, we match reconstructions to training samples across all group transformations.

Results. As illustrated in Figures 3(a) and 3(b), the optimization often converges to invariant reconstructions, resulting in a significant loss of information and low-quality reconstructions. Moreover, our observations reveal that many reconstructions lie on the convex hull of sample orbits, or orbitopes. To understand the distribution of reconstructions on the orbitopes we sampled various points on the ground truth orbitopes and identified the nearest neighbors for each reconstruction. as depicted in Figure 4 the observed distribution aligns with our predictions in Section 3.2: the reconstructions are concentrated around the group average that lies in V^G with the largest possible stabilizer. Notably, the KKT-based method outperforms activation maximization, as shown in Figure 2. Furthermore, we observe that increasing either the group size or the training set size leads to poorer results.

4. Symmetry-aware reconstruction

We propose two methods to improve reconstruction: mitigating GD's bias towards symmetry and imposing meaningful input space priors.

Symmetry-Aware Memory-Enhanced Gradient Descent (SAME-GD). Empirically, we tend to converge to points with nontrivial stabilizers, in particular to points on or close to V^G. We suggest aggregating the current query point with previous points in the optimization trajectory in a way that breaks the nesting property proved in Proposition 2. For simplicity, we used convex aggregation in the form of $\boldsymbol{x}_t \leftarrow \alpha_t \boldsymbol{x}_t + (1-\alpha_t)(\boldsymbol{x}_{prev} - \bar{g}\boldsymbol{x}_{prev})$, see Algorithm 1.

Incorporating Deep Image Prior (DIP). As proposed in Ulyanov et al. (2020), convolutional neural networks can be used as an implicit prior when it comes to inverse problems. We propose to use the same objective functions of the existing methods, but parameterizing the reconstruction variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ as the output of a randomly initialized CNN instead of optimizing them directly. The motivation is that the natural image prior could potentially break the symmetry and enhance the quality of the reconstructions.

Preliminary experimental results. We investigated the reconstruction abilities of the proposed methods under different configurations as listed on Table 4. We extended our experimental results to include CIFAR-10 images, with binary labels indicating animals and vehicles. SAME-GD and DIP, when combined with the KKT objective, yield notably improved reconstructions. These methods show a reduced tendency to converge to group averages, thus preserving more meaningful data characteristics. Particularly noteworthy is the KKT with DIP approach (Figure 1), which excels in producing piece-wise smooth asymmetric reconstructions by exploiting the implicit prior induced by DIP.

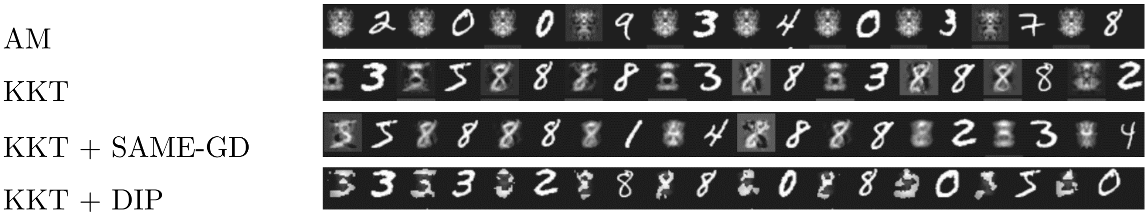


Figure 1: Pairs of training samples and their corresponding nearest neighbors reconstructions on their left, where $n = 50, |G| = 2$ is the group of right-left reflections.

Dataset	Group size	Train set size	AM	KKT	KKT + SAME-GD	KKT + DIP
MNIST	2	50	0.482 ± 0.000	0.464 ± 0.000	0.429 ± 0.004	$\mathbf{0.285 \pm 0.018}$
MNIST	2	100	0.484 ± 0.000	0.467 ± 0.000	0.450 ± 0.004	$\mathbf{0.271 \pm 0.000}$
CIFAR-10	2	50	0.446 ± 0.000	$\mathbf{0.346 \pm 0.001}$	0.369 ± 0.000	–
CIFAR-10	2	100	0.463 ± 0.000	0.371 ± 0.001	$\mathbf{0.370 \pm 0.007}$	–
MNIST	8	50	0.490 ± 0.002	0.471 ± 0.000	0.465 ± 0.000	$\mathbf{0.314 \pm 0.021}$
MNIST	8	100	0.494 ± 0	0.471 ± 0.001	0.469 ± 0.000	$\mathbf{0.323 \pm 0.025}$

Table 1: The Mean DSSIM value for different methods across datasets, group sizes, and training set sizes.

Discussion. Our work highlights the challenges in reconstructing training data from group-invariant neural networks. The theoretical and experimental foundations laid here raise questions about the behavior of reconstruction methods applied to these networks. Although we provide some insights and novel approaches, It is still unclear why standard reconstruction methods fail for invariant models. There are many future directions to be explored, some of them are discussed in more details in Appendix G.

References

Allison H. Baker, Alexander Pinard, and Dorit M. Hammerling. Dssim: a structural similarity index for floating-point data, 2023. URL https://arxiv.org/abs/2202.02616.

Gon Buzaglo, Niv Haim, Gilad Yehudai, Gal Vardi, Yakir Oz, Yaniv Nikankin, and Michal Irani. Deconstructing data reconstruction: Multiclass, weight decay and general losses. In *Advances in Neural Information Processing Systems*, volume 36, pages 51515–51535, 2023.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *Advances in Neural Information Processing Systems*, 35:22911–22924, 2022.

Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 603–618, 2017.

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems*, 34:7232–7241, 2021.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Noel Loo, Ramin Hasani, Mathias Lechner, and Daniela Rus. Dataset distillation fixes dataset reconstruction attacks. *arXiv preprint arXiv:2302.01428*, 2023.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Yakir Oz, Gilad Yehudai, Gal Vardi, Itai Antebi, Michal Irani, and Niv Haim. Reconstructing training data from real world models trained with transfer learning. *arXiv preprint arXiv:2407.15845*, 2024.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

Raman Sanyal, Frank Sottile, and Bernd Sturmfels. Orbitopes. *Mathematika*, 57(2): 275–314, June 2011. ISSN 2041-7942. doi: 10.1112/s002557931100132x. URL http://dx.doi.org/10.1112/S002557931100132X.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *International Journal of Computer Vision*, 128(7):1867–1888, March 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01303-4. URL http://dx.doi.org/10.1007/s11263-020-01303-4.

Bang Wu, Xiangwen Yang, Shirui Pan, and Xingliang Yuan. Adapting membership inference attacks to gnn for graph classification: Approaches and implications, 2021.

Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 225–240, 2019.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Zaixi Zhang, Qi Liu, Zhenya Huang, Hao Wang, Chengqiang Lu, Chuanren Liu, and Enhong Chen. Graphmi: Extracting private graph data from graph neural networks, 2021.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32, 2019.

Appendix A. Previous work

Several methods have been developed to reconstruct training samples from neural networks under different settings. Activation-maximization attacks Fredrikson et al. (2015); Yang et al. (2019) optimize the target output class over the input. Another method is reconstruction in a federated learning setup Zhu et al. (2019); Hitaj et al. (2017); Geiping et al. (2020); Huang et al. (2021) where the attacker is assumed to have knowledge of the sample's gradient. Several works use the implicit bias of neural networks towards margin maximization Lyu and Li (2020); Ji and Telgarsky (2020) to devise reconstruction losses, and thus reconstruct samples that are on the margin Haim et al. (2022); Buzaglo et al. (2023); Loo et al. (2023); Oz et al. (2024). Some prior studies have explored reconstructing graph structures from trained networks. The majority of existing research has concentrated on single-graph learning scenarios. In these cases, known node feature matrices effectively break symmetries, which simplifies the problem Zhang et al. (2021); Wu et al. (2021).

Appendix B. Current reconstruction methods

This is elaboration of Section 2 in the main text. There are several methods to reconstruct training data from trained neural networks. In this work we focused on two methods which allow data reconstruction in a general setting with minimal assumptions on the model's architecture.

Activation-Maximization (AM). We are given a trained multi-class classifier $\phi : \mathbb{R}^d \to \mathbb{R}^C$ with C classes. The predicted class of the classifier is defined as $\max_{i\in[C]}(\phi(\boldsymbol{x}))_i$, namely, the class with the maximal output. In this reconstruction method, to reconstruct a sample in class i, we randomly initialize an input $\boldsymbol{x} \sim \mathcal{N}\left(0, \frac{1}{d}I\right)$ and maximize the loss objective $\mathcal{L}_{rec} = \max_{\boldsymbol{x}\in\mathbb{R}^d}(\phi(\boldsymbol{x}))_i$. This is done by applying a first-order optimization method such as (Gradient Descent) GD.

KKT-based reconstruction. Lyu and Li (2020); Ji and Telgarsky (2020) show that given a homogeneous[1] neural network $\phi(\cdot, \theta)$, trained with gradient flow using an exponentially tailed loss (e.g., binary cross entropy) on a binary classification dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \{\pm 1\}$, its parameters θ converge to a KKT point of the following margin maximization problem

$$\min \|\theta\|^2 \quad \text{s.t.} \;\; \forall i = 1, \dots, n, \;\; y_i\phi(\boldsymbol{x}_i, \theta) \geq 1. \tag{2}$$

In particular, the KKT stationary condition is satisfied, namely there exist $\lambda_i \geq 0$ for $i = 1, \dots, n$ such that $\theta = \sum_{i=1}^n \lambda_i y_i \phi(\boldsymbol{x}_i, \theta)$. In Haim et al. (2022) the authors use the stationary condition to construct an objective that reconstructs the training data $\boldsymbol{x}_i$ given the trained weights θ. Namely, they propose optimizing the following loss objective

$$\mathcal{L}_{rec}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m, \lambda_1, \dots, \lambda_m) = \left\| \theta - \sum_{i=1}^m \lambda_i y_i \phi(\boldsymbol{x}_i, \theta) \right\| \tag{3}$$

where m, the number of reconstruction candidates is chosen to be $m \gg n$. This optimization problem in practice is solved by GD or similar optimization methods.

Appendix C. Proof of proposition 1

The activation maximization objective function uses the model output directly. Since the model is invariant it is trivially implying that the objective loss is also invariant. As the KKT based method involves first-order derivatives we would start by proving the following lemma:

Lemma 3 *Let $f(x; \theta) : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ be G-invariant function w.r.t x. Assume f has a partial gradient by θ at (x_0, θ_0) . Then $f(x; \theta)$ also has partial gradient by θ at $\{(g \cdot x_0, \theta_0)\}_{g\in G}$. Moreover, $\nabla_\theta f(x_0; \theta) = \nabla_\theta f(g \cdot x_0; \theta), \forall g \in G$*

Proof Denote $\{e_1, e_2, ..., e_p\}$ to be the standard basis of $\mathbb{R}^p$. f is G-invariant, therefore $\forall i = 1, \dots, p, \forall \epsilon \in \mathbb{R}, \forall g \in G$,

$$\frac{f(x; \theta + \epsilon e_i) - f(x; \theta)}{\epsilon} = \frac{f(g \cdot x; \theta + \epsilon e_i) - f(x; \theta)}{\epsilon} \tag{4}$$

[1] A function f is L homogeneous if for every $\alpha > 0$ we have $f(\alpha \boldsymbol{x}) = \alpha^L f(\boldsymbol{x})$

Since it is given the limit of $\epsilon \to 0$ exists for the left side of the equation then the limit of the right side also exists and is equal to it. If we take the limit of both side, by definition we get:

$$\frac{\partial f(x;\theta)}{\partial \theta_i} = \frac{\partial f(g \cdot x;\theta)}{\partial \theta_i} \tag{5}$$

■

In other words $\nabla_\theta f(\cdot, \theta)$ is G-invariant. As the trained model ϕ is invariant, we can say that $\nabla_\theta \phi$ is also invariant and therefore the objective loss in the KKT based method is also invariant.

Appendix D. Proof of proposition 2

We prove here the extended version of Proposition 2.

Proposition 4 *Let $\mathcal{L}(\boldsymbol{x};\theta) : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ be G-invariant function w.r.t $\boldsymbol{x}$. Consider the following optimization problem*

$$min_{x \in \mathbb{R}^d} \mathcal{L}(x, \theta) \tag{6}$$

solved by the following iterates of GD with some learning rates η_t

$$x_t = x_{t-1} - \eta_t \nabla_x \mathcal{L}(x_{t-1}, \theta), \quad t = 1, 2, \ldots, T \tag{7}$$

Then

1. *The gradient step is G-equivariant function of x_{t-1}.*
2. *$Stab_G(x_{t-1}) \subseteq Stab_G(x_t)$*

We first start by addressing the equivariance of the gradient by x.

Lemma 5 *Let $\mathcal{L}(x;\theta) : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ be G-invariant function w.r.t x and assume $\rho(g)$ is an orthogonal matrix for all $g \in G$. If $\mathcal{L}$ is derivable by x at (x_0, θ_0) then $\mathcal{L}(x;\theta)$ is also derivable by x at $\{(g \cdot x_0, \theta_0)\}_{g \in G}$.*
Moreover, $\nabla_x \mathcal{L}(g \cdot x;\theta) = g \nabla_x \mathcal{L}(x;\theta), \forall g \in G$.

Proof For convenience, we would write $\mathcal{L}(\cdot)$ instead of $\mathcal{L}(\cdot, \theta_0)$
By definition, for any direction $h \in \mathbb{R}^d, ||h||_2 = 1$,

$$D_h f(x) =< \nabla f(x), h > \tag{8}$$

Where $D_h f(x)$ is the directional derivative of f at x.
f is G-invariant, therefore:

$$D_{g^{-1} \cdot h} \mathcal{L}(x) = \lim_{\epsilon \to 0} \frac{\mathcal{L}(x + \epsilon g^{-1} \cdot h) - \mathcal{L}(x)}{\epsilon} \tag{9}$$

$$= \lim_{\epsilon \to 0} \frac{\mathcal{L}(g \cdot x + \epsilon g \cdot g^{-1} \cdot h) - \mathcal{L}(g \cdot x)}{\epsilon} \tag{10}$$

$$= lim_{\epsilon \to 0} \frac{\mathcal{L}(g \cdot x + \epsilon \cdot h) - \mathcal{L}(g \cdot x)}{\epsilon} \tag{11}$$

$$= D_h \mathcal{L}(g \cdot x) \tag{12}$$

On one hand,

$$D_{g^{-1}\cdot h}\mathcal{L}(x) =< \nabla\mathcal{L}(x), g^{-1}h > \tag{13}$$

On the other hand g^{-1} is orthogonal, then

$$\begin{aligned} D_{g^{-1}\cdot h}\mathcal{L}(x) &= D_h\mathcal{L}(g\cdot x) && (14)\\ &=< \nabla\mathcal{L}(g\cdot x), h > && (15)\\ &=< g^{-1}\nabla\mathcal{L}(g\cdot x), \cdot g^{-1}\cdot h > && (16)\\ & && (17) \end{aligned}$$

Therefore,

$$\nabla\mathcal{L}(x) = g^{-1}\nabla\mathcal{L}(g\cdot x) \tag{18}$$

■

In other words $\nabla_\theta\mathcal{L}(\cdot,\theta)$ is G-equivariant.
$\mathcal{L}$ is G-invariant, then by Lemma 5, for every $g \in G$ and for any $x_{t-1} \in \mathbb{R}^d$:

$$gx_{t-1} - \eta_t\nabla_x\mathcal{L}(gx_{t-1},\theta) = gx_{t-1} - g\eta_t\nabla_x\mathcal{L}(x_{t-1}),\theta) = g\cdot x_t \tag{19}$$

Therefore the gradient step is G-equivariant.
if $g \in Stab(x_{t-1})$, then by definition $gx_{t-1} = x_{t-1}$. Therefore:

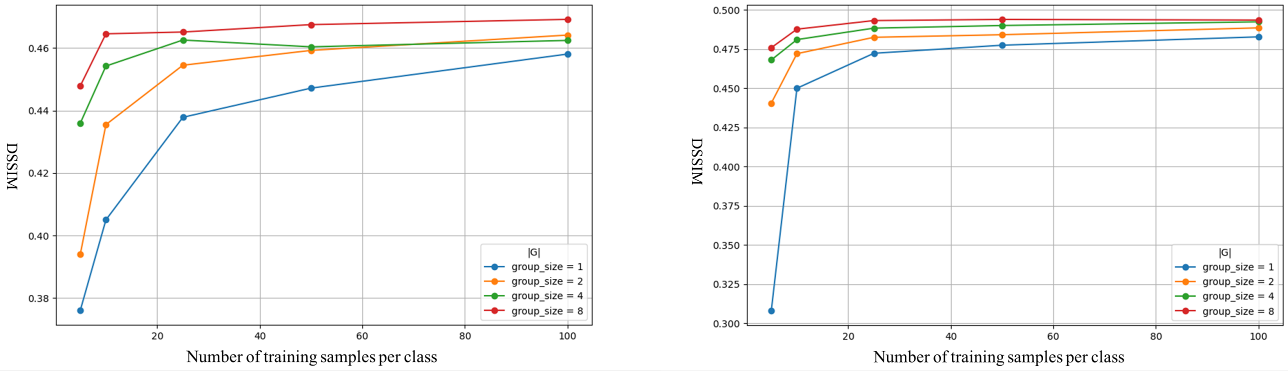
$$\begin{aligned} g\cdot x_t &= g\cdot(x_{t-1} - \eta_t\nabla\mathcal{L}(x_{t-1},\theta)) && (20)\\ &= g\cdot x_{t-1} - \eta_t g\cdot\nabla\mathcal{L}(x_{t-1},\theta) && (21)\\ &= x_{t-1} - \eta_t\nabla\mathcal{L}(g\cdot x_{t-1},\theta) && (22)\\ &= x_{t-1} - \eta_t\nabla\mathcal{L}(x_{t-1},\theta) && (23)\\ &= x_t && (24) \end{aligned}$$

Therefore $g \in Stab(x_t)$

Appendix E. Experimental setting

Setting. We focus on image data and considered 4 groups for our experiments - the trivial group, group of 2 elements acting as horizontal reflection, the group of 4 elements acting as horizontal and vertical reflection G_4 (Klein four-group), and the Dihedral group D_4 (rotations and reflections). To construct the invariant model we used a ReLU neural network with two hidden layers of width 1000 each and applied symmetrization [2]. Initially we trained the models on MNIST images with binary labels for odd or even digits. For each group we trained a neural network with different training set sizes $n = \{10, 20, 50, 100, 200\}$ for 100K epochs. All training ended with $\sim 1e-6$ training error and 100% accuracy. For each method and configuration, we ran 5 experiments of reconstruction with different seeds and $m = 1000$ candidates (500 per class).

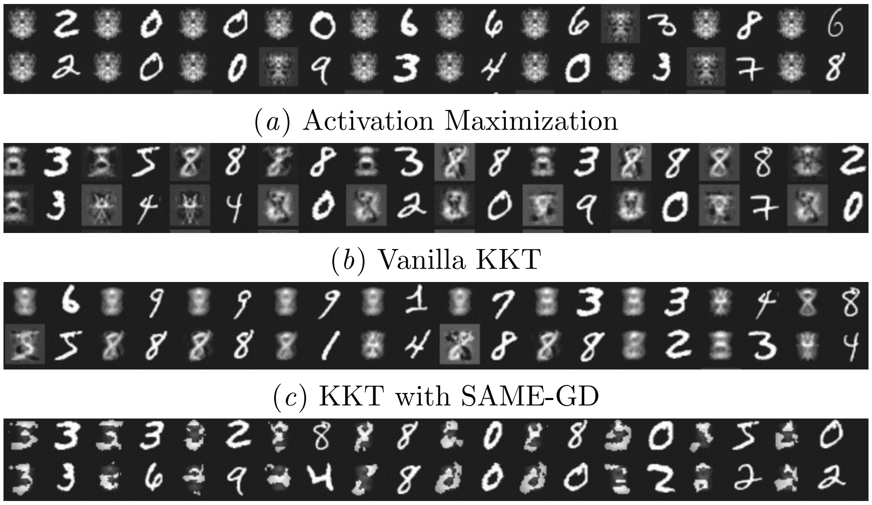
[2] symmetrization is a common practice to project functions on the invariant function space using Reynolds operator $\phi(x;\theta) = \frac{1}{|G|}\sum_{g\in G}\tilde{\phi}(gx;\theta)$



(a) KKT

(b) AM

Figure 2: The mean DSSIM over MNIST training subsets with varying size and different groups.



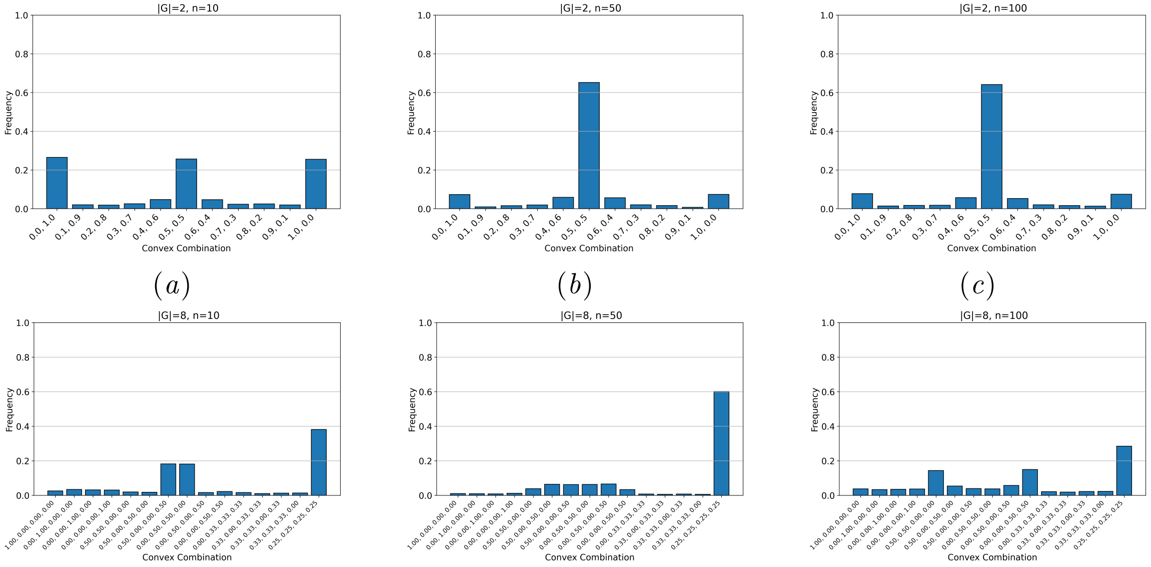
(a) Activation Maximization

(b) Vanilla KKT

(c) KKT with SAME-GD

(d) KKT with Deep Image Prior

Figure 3: Pairs of training samples and their corresponding nearest neighbors reconstructions on their right , where $n = 50, |G| = 2$.



(a) (b) (c)

(d) (e) (f)

Figure 4: The empirical distribution of reconstructions across orbitopes using KKT-based method on MNIST-trained invariant networks. Orbitopes are discretized into bins, each representing a convex combination of orbit elements. Reconstructions are assigned to bins based on their nearest neighbor in the discretized orbitope.

Appendix F. SAME-GD

```
Algorithm 1: SAME-GD
Input: $\{\eta_t, \alpha_t, \beta_t\}_{t=1}^{T}, T_{save}, T_{update}$
Output: $x_T$
Draw $x_0$ ;
$x_{prev} \leftarrow x_0$;
for $i \leftarrow 1$ to $T$ do
    if $t\%T_{update} \neq 0$ then
        $x_t = x_{t-1} - \eta_t \nabla \mathcal{L}(x_{t-1})$;
    end
    else
        $x_t \leftarrow \alpha_t x_t + (1 - \alpha_t)(x_{prev} - \bar{g} x_{prev})$ ;
    end
    if $t\%T_{save} == 0$ then
        $x_{prev} \leftarrow \beta_t x_t + (1 - \beta_t)\nabla \mathcal{L}(x_t)^2$ ;
    end
end
```

Appendix G. Discussion and future directions

In this section we discuss in details some challenges and future directions that arise from this work:

- Our work focuses on relatively small groups, containing at most 8 elements, which contain only rotation and reflection transformations. It would be interesting to further study reconstruction from models that are invariant to much larger groups.

- Our results indicate that reconstructions from invariant models lie close to the orbitope, mostly to the average over group elements. This finding only scratches the surface regarding how the reconstructions are distributed inside the orbitope, which may be effected by different factors such as initialization, architecture of the network, and structure of the group.

- We propose several methods to guide the reconstructions towards specific elements in the group, and thus to reconstruct the actual training samples (up to group action). However, it is not clear how well these methods generalize to larger datasets or larger groups. It will also be interesting to find new methods for this task, for example methods that take advantage of the geometrical properties of the orbitope, and may push the reconstructions towards extreme points of the orbitope.

- Our work focuses on image datasets, namely MNIST and CIFAR. It would be interesting to use these methods to reconstruct training samples from different data modalities, such as graphs or point clouds.

- In this work the models are constructed to be invariant using symmetrization of the feed-forward model. There are other methods to construct invariant networks, such as parameter sharing-based techniques, and it would be interesting to study data reconstruction attacks on such networks.