Data Augmentation: A Fourier Analysis Perspective

Behrooz Tahmasebi BEHROOZ_TAHMASEBI@SEAS.HARVARD.EDU
Harvard University

Melanie Weber MWEBER@SEAS.HARVARD.EDU
Harvard University

Stefanie Jegelka STEFANIE.JEGELKA@TUM.DE
TUM and MIT

Editors: List of editors' names

Abstract

Data augmentation, which supplements a dataset with transformed copies of each datum according to a known symmetry group, provides a model-agnostic approach to enforcing invariances, in contrast to methods that encode symmetries directly into the model. Although data augmentation has proven effective in theory and practice, full group-sized augmentation is often computationally infeasible, prompting the question: *Can partial augmentation still achieve the same performance as full augmentation in terms of generalization bounds and sample complexity?* In this paper, we develop a theoretical framework based on Fourier analysis, showing that partial data augmentation can achieve the full statistical benefits of full data augmentation. To our knowledge, this is the first proof of the efficacy of partial augmentation, highlighting an underexplored aspect of why augmentation remains a powerful and widely applicable strategy, even when performed only partially.

Keywords: data augmentation, Fourier analysis, invariance, equivariance, symmetry

1. Introduction

One of the most widely used model-agnostic techniques in machine learning and artificial intelligence for leveraging task structure is data augmentation. In data augmentation, the dataset is supplemented with transformed copies of each datum according to a known structure inherent in the underlying task. For instance, in learning with invariances, data augmentation leverages the group of symmetries associated with the task to improve model performance, including generalization to unseen data.

Due to its simplicity and model-agnostic nature, data augmentation for learning under invariances is widely used in practice. Its applications span a broad range of domains, including physics, materials science, drug discovery, molecular machine learning, computer vision and image processing, among many others.

However, despite its flexibility and wide applicability, this approach becomes challenging to apply in full when the symmetry groups are prohibitively large, a situation that frequently arises in practice. For instance, the permutation group and the sign-flip group are both exponentially large in the data dimension, making *full* data augmentation computationally infeasible. In such cases, one can still use *partial* data augmentation, as is often done in practice via heuristic methods, where only a subset of group elements is used. Yet, the theoretical understanding of the quality of partial data augmentation in downstream tasks remains underexplored.

© 2025 B. Tahmasebi, M. Weber & S. Jegelka.

In this paper, we initiate a rigorous study of this problem by asking the following question: *Can partial augmentation with substantially smaller subsets of the group still achieve statistical performance comparable to that of full group augmentation?* As an instance of this general question, in this extended abstract, we focus on the classical problem of density estimation. Somewhat surprisingly, we prove that even very small subsets of the group suffice to recover the statistical benefits of full data augmentation. Our proof builds on tools from Fourier analysis on groups together with standard facts from group and representation theory.

In short, this extended abstract makes the following contributions:

- We investigate the problem of efficient data augmentation for learning over symmetric data (with respect to a given group of invariances), and demonstrate that small subsets of group elements are sufficient to attain the full statistical benefits of augmentation.
- We introduce techniques from Fourier analysis on groups to analyze data augmentation, providing tools and perspectives that may be of independent interest.

Note. A detailed review of related work is provided in the appendices.

2. Problem Formulation and Main Results

We consider the classical setting of density estimation, where we are given n i.i.d. samples x_i, $i \in [n]$, from an unknown distribution with density $f^\star(x)$ over a domain $\mathcal{X} \subseteq \mathbb{R}^d$. For simplicity, we assume $\mathcal{X} = \mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$, but the results extend to other domains under mild conditions. Moreover, we assume $f^\star \in L^2(\mathbb{S}^{d-1})$.

An estimator takes the observed samples and returns $\hat{f} \in \mathcal{F}$, where $\mathcal{F}$ denotes the set of admissible estimators. The quality of this estimation is measured by the excess population risk (sometimes referred to as the generalization error), defined as

$$\mathcal{R}(\hat{f}) := \mathbb{E}\Big[\|f^\star - \hat{f}\|^2_{L^2(\mathbb{S}^{d-1})}\Big] - \min_{f \in \mathcal{F}} \|f^\star - f\|^2_{L^2(\mathbb{S}^{d-1})}, \tag{1}$$

where the expectation is taken over the randomness of the data.

In this paper, for simplicity, we focus on *low-degree* estimators, where $\mathcal{F} = \mathcal{F}_k$ denotes the space of multivariable polynomials of total degree at most k, for some $k \in \mathbb{N}$. The parameter k controls the model capacity. Moreover, we are particularly interested in the case where the unknown distribution $f^\star$ is invariant under the action of a given group G on the domain: $f^\star(gx) = f^\star(x), \forall g \in G$, where the equality is understood in the L^2 sense.

Generic algorithm under data augmentation. Assume we are given the optimal[1] (i.e., minimax) density estimator, i.e., the optimal algorithm that maps any dataset to an estimated distribution, and we want to combine it with data augmentation.

We begin by selecting a subset of the group $S \subseteq G$ and augmenting the dataset accordingly. The augmented dataset is $\{sx_i : s \in S,\, i \in [n]\}$. Full augmentation corresponds to $S = G$, while partial augmentation uses substantially smaller subsets. Since data augmentation is model-agnostic, we do not modify the optimal algorithm and apply it to the augmented dataset. We denote the output as $\hat{f}_S$.

1. Details about the optimal algorithms are provided in the proof sketch. Minimax optimality here should be interpreted with respect to generic distributions, without restriction to invariant distributions.

Intuitively, larger sets S provide greater statistical benefits (i.e., lower risk). However, for many groups commonly arising in applications (e.g., permutations, sign-flips), performing full augmentation with $S = G$ is computationally infeasible, since the size of G is typically exponential in the dimension. This motivates the following question:

> Can partial data augmentation with small subsets $S \subseteq G$ achieve statistical benefits (i.e., low risk) comparable to those of full augmentation?

The main result of this paper is summarized in the following theorem.

Theorem 1 *There exists a subset $S \subseteq G$ of size $|S| = \mathcal{O}(nk \log d)$ such that*

$$c_1 \, \mathcal{R}(\hat{f}_G) \;\leq\; \mathcal{R}(\hat{f}_S) \;\leq\; c_2 \, \mathcal{R}(\hat{f}_G), \tag{2}$$

for some universal positive constants c_1, c_2. Moreover, such a subset S can be efficiently constructed via i.i.d. sampling from the group G.

The above result is significant, as it shows that only logarithmic-sized subsets S (with respect to the data dimension d) suffice to achieve the full benefits of data augmentation. For comparison, consider the case of permutation invariances where $|G| = d!$. The theorem implies that one can obtain the same statistical gain with only $|S| = \mathcal{O}(\log d)$, which constitutes a double-exponential improvement over the baseline, assuming $n, k = \mathcal{O}(1)$. Even if $n = \text{poly}(d)$, the improvement remains exponential over the baseline. Please check Remark 4 for further explanation.

> Partial augmentation with only logarithmic-sized subsets $|S| = \mathcal{O}_{n,k}(\log d)$ achieves the full statistical benefits of data augmentation. For large groups $|G| = \exp(\Omega(d))$, this yields a *double-exponential* improvement in the dependence on dimension.

Proof sketch. We now outline the key ideas underlying the proof of our main result. First, let us introduce the following algorithm.

Optimal algorithm. For this classical problem (without invariances, in the minimax setting), the optimal algorithm is given by the following spectral estimator:

$$\hat{f}(x) = \sum_{\ell=0}^{k} \sum_{j=1}^{d_\ell} \hat{f}_{\ell,j} \, \phi_{\ell,j}(x), \quad \hat{f}_{\ell,j} = \frac{1}{n} \sum_{i=1}^{n} \phi_{\ell,j}(x_i), \tag{3}$$

where $\phi_{\ell,j}$ denotes the spherical harmonic of degree ℓ indexed by $j \in [d_\ell]$, and d_ℓ is the dimension of the space of spherical harmonics of degree ℓ. For additional background on spherical harmonics, see the appendices.

Consider the spectral algorithm applied to a dataset augmented with a subset $S \subseteq G$. We introduce the following notation: let $\boldsymbol{f}^\star \in \mathbb{R}^r$ denote the vector obtained by concatenating all spectral coefficients $f^\star_{\ell,j}$ for $\ell \leq k$, where $r \in \mathbb{N}$ is the appropriate dimension. Similarly, define $\hat{\boldsymbol{f}}_S \in \mathbb{R}^r$ for the estimator obtained from the augmented dataset, and let

$\hat{\boldsymbol{f}} \in \mathbb{R}^r$ denote the naive estimator without augmentation. Using standard spectral analysis, one can show that

$$\mathcal{R}(\hat{f}_S) = \|\hat{\boldsymbol{f}}_S - \boldsymbol{f}^\star\|_2^2 = \|D_S \hat{\boldsymbol{f}} - \boldsymbol{f}^\star\|_2^2, \quad D_S := \frac{1}{|S|}\sum_{s\in S} \rho(s) \in \mathbb{R}^{r\times r}, \tag{4}$$

where ρ denotes the linear representation of G induced on the space of polynomials of degree at most k on the unit sphere. For a trivial set S (with no augmentation), we have $D_S = I_r$. At the other extreme, if $S = G$, then

$$\exists P \in U(r): \quad D_G \;=\; P \begin{bmatrix} I_{r_{\text{inv}}} & 0 \\ 0 & 0 \end{bmatrix} P^\dagger \;\in \mathbb{R}^{r\times r},$$

where r_{inv} is the dimension of the space of G-invariant polynomials of degree at most k. More generally, for any subset $S \subseteq G$, one can show that the following representation exists:

$$D_S \;=\; P \begin{bmatrix} I_{r_{\text{inv}}} & 0 \\ 0 & \widetilde{D}_S \end{bmatrix} P^\dagger. \tag{5}$$

From this decomposition, a straightforward calculation yields

$$\mathcal{R}(\hat{f}_S) \;\lesssim\; \frac{r_{\text{inv}}}{n} + \|\widetilde{D}_S\|_{\text{op}}^2, \tag{6}$$

and this upper bound is tight in the case $S = G$, where $\|\widetilde{D}_S\|_{\text{op}} = 0$. If instead S is chosen uniformly at random, standard concentration bounds imply that with high probability,

$$\|\widetilde{D}_S\|_{\text{op}}^2 = \mathcal{O}\left(\frac{\log(r)}{|S|}\right).$$

Since $r = \mathcal{O}(d^k)$, this yields

$$\mathcal{R}(\hat{f}_S) \;\lesssim\; \frac{r_{\text{inv}}}{n} + \frac{k\log(d)}{|S|}. \tag{7}$$

Therefore, it suffices to take $|S| = \mathcal{O}(nk\log d)$ to achieve the desired risk bound. This completes the proof of Theorem 1.

Remark 2 *A more refined analysis incorporating r_{inv} yields sharper bounds on $|S|$. However, in this paper we report the conservative bound assuming $r_{\text{inv}} = \Omega(1)$ in order to keep the presentation streamlined. See Remark 4 for further discussion.*

Remark 3 *A key ingredient in our analysis is the concentration of the matrix D_S around zero. Note that D_S is given by the direct sum of averages over S of the irreducible components in the decomposition of ρ, which from the Fourier-analytic perspective corresponds to the* Fourier transform *of $\frac{1}{|S|}\sum_{s\in S}\delta_s$, where δ_s is the Dirac delta at s. Our result relies on the existence of functions on G with small support (i.e., small $|S|$) and simultaneously small Fourier coefficients. This connects naturally to uncertainty principles in Fourier analysis on groups, and to the best of our knowledge, this is the first work to highlight such a link in the context of data augmentation.*

Note. The framework extends beyond the present setting to Sobolev spaces, RKHSs, and nonparametric regression. With appropriate tools, it also extends to neural networks (e.g., two-layer networks under the neural tangent kernel). We leave a detailed study of these directions to future work.

Acknowledgments

BT and MW were partially supported by NSF Award CBET-2112085. MW acknowledges partial funding from NSF award DMS-2406905 and a Sloan Research Fellowship. Stefanie Jegelka acknowledges support from NSF AI Institute TILOS and the Alexander von Humboldt Foundation.

References

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Simon Batzner, Albert Musaelian, and Boris Kozinsky. Advancing molecular simulation with equivariant interatomic potentials. *Nature Reviews Physics*, 5(8):437–438, 2023.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Yunhao Chen, Zihui Yan, and Yunjie Zhu. A comprehensive survey for generative data augmentation. *Neurocomputing*, 600:128167, 2024.

Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Chris De Sa, and Christopher Ré. A kernel theory of modern data augmentation. In *Int. Conference on Machine Learning (ICML)*, 2019.

Yijun Dong, Yuege Xie, and Rachel Ward. Adaptively weighted data augmentation consistency regularization for robust optimization under concept shift. In *Int. Conference on Machine Learning (ICML)*, 2023.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *Int. Conference on Machine Learning (ICML)*, 2023.

Polina Kirichenko, Mark Ibrahim, Randall Balestriero, Diane Bouchacourt, Shanmukha Ramakrishna Vedantam, Hamed Firooz, and Andrew G Wilson. Understanding the detrimental class-level effects of data augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. *Ai Open*, 3:71–90, 2022.

Chi-Heng Lin, Chiraag Kaushik, Eva L Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *Journal of Machine Learning Research*, 25(91):1–85, 2024.

George Ma, Yifei Wang, Derek Lim, Stefanie Jegelka, and Yisen Wang. A canonicalization perspective on invariant and equivariant learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Guozheng Ma, Zhen Wang, Zhecheng Yuan, Xueqian Wang, Bo Yuan, and Dacheng Tao. A comprehensive survey of data augmentation in visual reinforcement learning. *International Journal of Computer Vision*, pages 1–38, 2025.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory (COLT)*, 2021.

Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE, 2018.

Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022.

Pratik Patil and Jin-Hong Du. Generalized equivalences between subsampling and ridge regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132:109803, 2023.

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *Int. Conference on Learning Representations (ICLR)*, 2022.

Ruoqi Shen, Sébastien Bubeck, and Suriya Gunasekar. Data augmentation as feature manipulation. In *Int. Conference on Machine Learning (ICML)*, 2022.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Tess E Smidt. Euclidean symmetry and equivariance in machine learning. *Trends in Chemistry*, 3(2):82–85, 2021.

Ashkan Soleymani, Behrooz Tahmasebi, Patrick Jaillet, and Stefanie Jegelka. From finite to infinite groups: A polynomial-time algorithm for learning with exact invariances. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025a.

Ashkan Soleymani, Behrooz Tahmasebi, Stefanie Jegelka, and Patrick Jaillet. Learning with exact invariances in polynomial time. In *Int. Conference on Machine Learning (ICML)*, 2025b.

Ashkan Soleymani, Behrooz Tahmasebi, Stefanie Jegelka, and Patrick Jaillet. A robust kernel statistical test of invariance: Detecting subtle asymmetries. In *Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025c.

Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Behrooz Tahmasebi and Stefanie Jegelka. Sample complexity bounds for estimating probability divergences under invariances. In *Int. Conference on Machine Learning (ICML)*, 2024.

Behrooz Tahmasebi and Stefanie Jegelka. Regularity in canonicalized models: A theoretical perspective. In *Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025a.

Behrooz Tahmasebi and Stefanie Jegelka. Generalization bounds for canonicalization: A comparative study with group averaging. In *Int. Conference on Learning Representations (ICLR)*, 2025b.

Shuo Yang, Yijun Dong, Rachel Ward, Inderjit S Dhillon, Sujay Sanghavi, and Qi Lei. Sample efficiency of data augmentation consistency regularization. In *Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *Foundations and Trends® in Machine Learning*, 18 (4):385–912, 2025.

Tong Zhao, Wei Jin, Yozen Liu, Yingheng Wang, Gang Liu, Stephan Günnemann, Neil Shah, and Meng Jiang. Graph data augmentation for graph machine learning: A survey. *arXiv preprint arXiv:2202.08871*, 2022.

Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. Toward understanding generative data augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Appendix A. Related Work

The field of geometric machine learning has found applications across a wide range of domains, including quantum, atomistic, and continuum systems, among others (Zhang et al., 2025; Batzner et al., 2022; Bronstein et al., 2017; Smidt, 2021; Batzner et al., 2023). On the theoretical side, the sample complexity benefits of exploiting symmetries have been studied for kernel methods through group averaging (Tahmasebi and Jegelka, 2023, 2024) and for canonicalization (Tahmasebi and Jegelka, 2025b); see also (Tahmasebi and Jegelka, 2025a) for regularization, and (Soleymani et al., 2025c) for testing. On the other hand, the time complexity of learning under invariances has also been a subject of recent study (Soleymani et al., 2025a,b). Approximation guarantees for equivariant learning have also been established (Petrache and Trivedi, 2023). By contrast, the theoretical understanding

of data augmentation remains relatively limited. Note that canonicalization (Kaba et al., 2023; Ma et al., 2024) and frame averaging (Puny et al., 2022) represent two alternative approaches for introducing symmetries directly into learning algorithms, rather than through augmentation.

Several works have studied the role of data augmentation from different perspectives: its effect on training dynamics in neural networks (Shen et al., 2022), its interpretation as a form of regularization (Lin et al., 2024; Yang et al., 2023), and its group-theoretic foundations (Chen et al., 2020). Closely related to our setting, Dao et al. (2019) developed a kernel-based theory of augmentation, while further analyses are given in (Patil and Du, 2023; Mei et al., 2021).

Beyond passive augmentation, several recent works have explored generative, active, and adaptive augmentation strategies (Zheng et al., 2023; Dong et al., 2023; Chen et al., 2024), though these aspects are not addressed in the present paper. For surveys, see Shorten and Khoshgoftaar (2019) for image augmentation in deep learning, Ma et al. (2025) for reinforcement learning, and Li et al. (2022); Pellicer et al. (2023) for natural language tasks. Additional studies examine applications in image classification (Mikołajczyk and Grochowski, 2018), graph learning (Zhao et al., 2022), and other domains (Mumuni and Mumuni, 2022). Finally, it is worth noting that augmentation can also have drawbacks, as highlighted in Kirichenko et al. (2023).

Appendix B. Background

B.1. Spherical Harmonics

Let us briefly review the theory of spherical harmonics. The space $L^2(\mathbb{S}^{d-1})$ admits a decomposition into an orthogonal direct sum of the spaces of homogeneous harmonic polynomials:

$$L^2(\mathbb{S}^{d-1}) = V_0 \oplus V_1 \oplus V_2 \oplus \cdots , \tag{8}$$

where V_ℓ denotes the space of all polynomials $p(x)$ of total degree $\ell = 0, 1, \ldots$ satisfying $\Delta p(x) = 0$ for all $x \in \mathbb{R}^d$. Let $d_\ell = \dim(V_\ell) = \binom{d+\ell-1}{\ell} - \binom{d+\ell-3}{\ell-2}$, and denote by $\{\phi_{\ell,j}\}_{j=1}^{d_\ell}$ an orthonormal basis of harmonic polynomials for V_ℓ. Then, any function $f \in L^2(\mathbb{S}^{d-1})$ can then be expressed as a convergent expansion in spherical harmonics:

$$f = \sum_{\ell=0}^{\infty} \sum_{j=1}^{d_\ell} f_{\ell,j}\, \phi_{\ell,j}. \tag{9}$$

Truncating the expansion to $\ell \leq k$ yields the best approximation of f by polynomials of degree at most k.

B.2. Spectral Estimators

By the orthonormality of spherical harmonics, we can write

$$f = \sum_{\ell=0}^{\infty} \sum_{j=1}^{d_\ell} f_{\ell,j}\, \phi_{\ell,j} \qquad \Longrightarrow \qquad f_{\ell,j} = \int_{\mathbb{S}^{d-1}} f(x)\, \phi_{\ell,j}(x)\, dx = \mathbb{E}_{x\sim\mu}\big[\phi_{\ell,j}(x)\big], \tag{10}$$

where the expectation is taken over random points on the unit sphere distributed according to $f(x)$.

This observation implies that, given data, one can approximate the coefficients $f_{\ell,j}$ by empirical averages of $\phi_{\ell,j}(x_i)$, $i \in [n]$, and thereby construct an estimator of the density function. Such spectral estimators are known to achieve minimax optimal rates in the classical density estimation problem.

B.3. Miscellaneous Facts about Representation Theory

For an orthogonal group representation $D(g) \in \mathbb{R}^{r\times r}$, one can show that

$$\mathbb{E}_g[D(g)] \;=\; P \begin{bmatrix} I_{r_{\text{inv}}} & 0 \\ 0 & 0 \end{bmatrix} P^{\dagger} \tag{11}$$

for some unitary matrix $P \in U(r)$.

To see this, note that if $D(g)$ is irreducible, then the average is either the zero matrix or the identity matrix, depending on whether $D(g)$ is trivial or not. In the general case, complete reducibility implies that $D(g)$ decomposes into irreducible components, from which the result follows.

Appendix C. Proof of Theorem 1

Proof We provide a complementary explanation to the proof sketch presented in the main body of the paper. By complete reducibility, we may assume without loss of generality that $D(g)$ is an irreducible non-trivial representation. In this case, we have $\mathbb{E}_g[D(g)] = 0$. Our goal is to measure how close

$$\frac{1}{|S|}\sum_{g\in S} D(g)$$

is to zero in operator norm, for a random subset $S \subset G$. Note that classical concentration results, such as the Hanson–Wright inequality, imply that with probability at least $1-\delta$,

$$\Big\|\frac{1}{|S|}\sum_{g\in S} D(g)\Big\|_{\text{op}} \;\le\; \mathcal{O}_\delta\!\left(\frac{\log(r)}{|S|}\right),$$

where r is the dimension of the representation. Thus, it remains to relate this operator norm bound to the generalization gap. Note that

$$\begin{aligned}
\mathcal{R}(\hat{f}_S) &= \|\hat{\boldsymbol{f}}_S - \boldsymbol{f}^\star\|_2^2 && (12)\\
&= \|D_S\hat{\boldsymbol{f}} - \boldsymbol{f}^\star\|_2^2 && (13)\\
&\lesssim \|D_S\hat{\boldsymbol{f}} - D_G\hat{\boldsymbol{f}}\|_2^2 + \|D_G\hat{\boldsymbol{f}} - \boldsymbol{f}^\star\|_2^2. && (14)
\end{aligned}$$

The second term is bounded, via a variance calculation, by $\frac{r_{\text{inv}}}{n}$. For the first term, we have

$$\|D_S\hat{\boldsymbol{f}} - D_G\hat{\boldsymbol{f}}\|_2^2 \;\lesssim\; \|\widetilde{D}_S\|_{\text{op}}^2, \tag{15}$$

which establishes the claim. ■

Remark 4 *Consider the case of permutation invariances. For polynomial regression under such invariances, it is known that* $r_{\text{inv}} = \exp(\Theta(\sqrt{k}))$*, which is independent of the ambient dimension* d*. Consequently, although the group size is* $|G| = d!$*, a subset of size* $|S| = \mathcal{O}_k(\log d)$ *suffices to achieve the full benefits of data augmentation.*

In particular, to guarantee a generalization gap of ϵ*, full augmentation requires at least* $n = \mathcal{O}_k\left(\frac{1}{\epsilon}\right)$ *samples, while partial augmentation attains the same gap with only* $|S| = \mathcal{O}_k\left(\frac{\log(d)}{\epsilon}\right)$*, representing a double-exponential improvement in the size of the augmentation scheme compared to the baseline.*