Dimension-Free Scaling Laws for Invariant Score Matching

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Modern machine learning models increasingly operate on variable-size data, such as sets and graphs, where the input dimension may differ between training and testing. Classical estimation theory is largely tied to fixed domains, with guarantees that often deteriorate as dimension grows. We ask whether invariant score matching obeys dimension-free scaling laws: can a score estimator learned on one domain transfer reliably to another, possibly much larger, domain? We answer this question by developing a general theory of invariant M-estimation across dimensions, with score matching as a central instance. Our analysis introduces a new spectral complexity measure that governs the scaling law of the estimator by quantifying how efficiently the target score is captured by low-complexity invariant components. For score matching on sets and graphs, we prove dimension-uniform convergence guarantees and show that low-complexity invariant scores can be estimated at fast, dimension-free rates, nearly matching the parametric rate. Our theory identifies a sharp scaling transition: below an explicit spectral threshold, invariant score matching transfers across dimensions with fast rates independent of ambient dimension; above this threshold, the sample complexity can become exponentially larger, making dimension transfer statistically impossible without additional structure. Beyond score matching, our framework applies to general invariant M-estimators under group actions, yielding dimension-free guarantees for a broad class of statistical and machine learning procedures. These results provide, to our knowledge, the first general theoretical framework for dimension-free scaling laws in invariant score matching, and more broadly for invariant M-estimation on variable domains.

1. Introduction

In many geometric learning problems, the domain is not fixed once and for all. Sets vary in cardinality, graphs vary in the number of nodes, and structured objects at test time may be larger than those observed during training. The goal is therefore not only to generalize to new samples, but to generalize across domain sizes. This phenomenon is central in geometric machine learning [6, 54] and is often studied as *size generalization* for sets and graphs [3, 56].

Despite its empirical importance, the statistical scaling laws of size generalization remain poorly understood. Classical estimation theory is usually formulated on a fixed domain, with rates that often depend explicitly on dimension. It therefore does not explain when learning from small objects can transfer to larger ones with stable sample complexity. This gap is especially important for invariant data, such as sets and graphs, where permutation symmetries reduce effective complexity. The key question is how this reduction changes the scaling law of estimation.

We study this question for invariant parameter estimation across dimensions, with *score matching* [16] as the central application. Score matching is a fundamental method for estimating score functions and a core tool in modern score-based generative modeling. For invariant distributions

Table 1: Dimension-free score-matching rates from optimized bias–variance tradeoffs (Theorem 2).

Data type	Source condition	Rate	Scaling regime
Sets in $(\mathbb{R}^r)^d$	$\mathsf{Exp}(a,\nu),\ \nu > \frac{r}{r+1}$	$n^{-1+o(1)}$	Nearly parametric
Sets in $(\mathbb{R}^r)^d$	$\mathsf{Exp}(a,\nu),\ \nu = \frac{r}{r+1}$	n^{-c}	Polynomial
Sets in $(\mathbb{R}^r)^d$	$\mathsf{Exp}(a,\nu),\ 0 < \nu < \frac{r}{r+1}$	$\exp\left(-c(\log n)^{\frac{\nu(r+1)}{r}}\right)$	Stretched logarithmic
Sets in $(\mathbb{R}^r)^d$	$\mathsf{Alg}(\tau)$	$(\log n)^{-\frac{\tau(r+1)}{r}}$	Logarithmic
Graphs with edge attributes	$\mathsf{Exp}(a,\nu),\ \nu > 1$	$n^{-1+o(1)}$	Nearly parametric
Graphs with edge attributes	$\mathsf{Exp}(a,\nu),\ \nu = 1$	$\exp\left(-c\frac{\log n}{\log\log n}\right)$	Subpolynomial
Graphs with edge attributes	$\mathsf{Exp}(a,\nu),\ 0 < \nu < 1$	$\exp\left(-c\left(\frac{\log n}{\log\log n}\right)^{\nu}\right)$	Stretched logarithmic
Graphs with edge attributes	$\mathsf{Alg}(\tau)$	$\left(\frac{\log n}{\log\log n}\right)^{-\tau}$	Logarithmic

on sets or graphs, the score inherits the corresponding equivariance structure. Thus, the central question becomes: when does invariant score matching obey dimension-free scaling laws, so that a score learned at one size can transfer to larger sizes without rates deteriorating with dimension?

We answer this question by developing a general theory of invariant M-estimation across dimensions. The framework applies beyond score matching to moment estimation, maximum likelihood-type objectives, and other invariant estimation problems under group actions. It shows that statistical rates can be governed by invariant complexity rather than ambient dimension.

For score matching on sets and graphs, we introduce a spectral complexity measure that quantifies how efficiently the true score is approximated by low-degree invariant features. This measure determines the optimized bias–variance scaling law. Low spectral complexity yields dimension-free, nearly parametric rates, while slower invariant coefficient decay leads to polynomial, stretched-logarithmic, or logarithmic regimes. In particular, we identify explicit source thresholds separating fast and slow scaling regimes for sets and graphs; see Table 1.

Our work is also connected to the recent program of *any-dimensional machine learning*, which studies architectures and guarantees that remain meaningful across changing domains. Building on ideas from representation stability [13], this line of work has obtained general transfer guarantees for structured data across dimensions [25]. Our contribution is complementary: we derive statistical scaling laws for such transfer, showing that representation stability must be paired with a quantitative bias–variance analysis governed by spectral complexity.

In short, this paper makes the following contributions:

- We develop a framework for invariant M-estimation across dimensions, showing that statistical rates can be governed by invariant complexity rather than ambient dimension. This gives dimension-free convergence guarantees for general invariant estimation problems under group actions.
- We establish dimension-free scaling laws for invariant score matching on sets and graphs. A new spectral complexity measure captures how efficiently the target score is approximated by invariant components, and yields sharp source thresholds separating nearly parametric, polynomial, stretched-logarithmic, and logarithmic regimes.

2. Problem Statement

We study estimation problems indexed by a dimension or size parameter $d \in \mathbb{N}$. For each d, let $\mathcal{X}_d$ be a measurable space with data distribution μ_d, let $\Theta_d \subseteq \mathbb{R}^{p_d}$ be a parameter space, and let $\ell_d : \mathcal{X}_d \times \Theta_d \to \mathbb{R}$ be a measurable loss. The population target is the M-estimation parameter

$$\theta_d^\star \in \arg\min_{\theta \in \Theta_d} \{L_d(\theta) \coloneqq \mathbb{E}_{X \sim \mu_d}[\ell_d(X, \theta)]\}.$$

Given i.i.d. samples $X_1, \ldots, X_n \sim \mu_d$, the empirical estimator is

$$\hat{\theta}_{n,d} \in \arg\min_{\theta \in \Theta_d} \left\{ \hat{L}_{n,d}(\theta) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell_d(X_i, \theta) \right\}.$$

This framework includes mean, covariance, moment, maximum likelihood, empirical risk minimization, and score-matching estimators as special cases. Our framework provides dimension-free bounds for general invariant M-estimators, with score matching serving as the main application throughout the paper. Additional definitions and examples are given in Appendix B and Appendix E.

Invariant estimation. For each d, let G_d be a group acting measurably on $\mathcal{X}_d$ and on Θ_d. We assume that the data distribution is invariant, meaning $g \cdot X \sim \mu_d$ for every $g \in G_d$ and $X \sim \mu_d$, and that the loss is compatible with the two actions, meaning $\ell_d(g \cdot x, g \cdot \theta) = \ell_d(x, \theta)$ for all $g \in G_d$, $x \in \mathcal{X}_d$, and $\theta \in \Theta_d$. Consequently, the population risk is invariant: $L_d(g \cdot \theta) = L_d(\theta)$. We denote the invariant parameter space by $\Theta_d^{G_d} \coloneqq \{\theta \in \Theta_d : g \cdot \theta = \theta \text{ for all } g \in G_d\}$. When the population minimizer is unique, invariance of the risk implies $\theta_d^\star \in \Theta_d^{G_d}$. Thus, under invariant data and losses, the target lies in a lower-complexity invariant subspace.

Any-dimensional setting. Classical estimation asks whether $\hat{\theta}_{n,d}$ approximates $\theta_d^\star$ for a fixed d. We ask a stronger question: can an estimator learned from samples at one dimension be transferred to larger dimensions with guarantees that do not deteriorate with the ambient size? For every pair $d \leq D$, suppose we are given a transfer map $\Phi_{d \to D} : \Theta_d^{G_d} \to \Theta_D^{G_D}$, which embeds invariant parameters at size d into invariant parameters at size D. Given samples from μ_d, we compute an invariant estimator $\hat{\theta}_{n,d} \in \Theta_d^{G_d}$ and evaluate its transferred version $\Phi_{d \to D} \hat{\theta}_{n,d}$ against the target $\theta_D^\star$ at the larger dimension D.

Invariant any-dimensional estimation. Given samples $X_1, \ldots, X_n \sim \mu_d$, construct an invariant estimator $\hat{\theta}_{n,d} \in \Theta_d^{G_d}$ such that, after transfer to any larger dimension $D \geq d$,

$$\sup_{D \geq d} \left\| \Phi_{d \to D} \hat{\theta}_{n,d} - \theta_D^\star \right\| \leq \varepsilon, \qquad \text{with probability at least } 1 - \delta.$$

using a sample size $n = n(\varepsilon, \delta)$ that is independent of d and D.

This formulation captures the central paper's question: when do invariant estimators obey dimension-free scaling laws? Our results show that the answer is governed by the spectral complexity of the invariant components needed to approximate the target.

Motivating examples: sets and graphs. The canonical examples are sets and graphs, where $G_d = S_d$ acts by relabeling elements or nodes. For set-valued data $X = (x_1, \dots, x_d) \in (\mathbb{R}^r)^d$, the action of $\pi \in S_d$ is $\pi \cdot X = (x_{\pi^{-1}(1)}, \dots, x_{\pi^{-1}(d)})$, so S_d-invariance means that the distribution is exchangeable. For graph-valued data represented by an attributed adjacency tensor $A \in \mathbb{R}^{d \times d \times r'}$, node relabeling acts by $(\pi \cdot A)_{ij,b} = A_{\pi^{-1}(i),\pi^{-1}(j),b}$. Thus, invariance means that the graph distribution is independent of the node ordering. If μ_d admits an invariant density p_d, then the score $s_d^\star(x) = \nabla_x \log p_d(x)$ is equivariant under the same action. In the graph case, for example, $s_d^\star(\pi \cdot A) = \pi \cdot s_d^\star(A)$. Therefore, score matching on sets and graphs is naturally an invariant M-estimation problem across changing dimensions. The rest of the paper develops convergence guarantees for this setting and identifies when these rates become fast and dimension-free.

3. Main Results: Invariant M-Estimation

We first give a general statistical principle for invariant M-estimation. The key point is that the relevant complexity is not the ambient parameter dimension p_d, but the dimension of the invariant model. Recall that $\Theta_d^{G_d} \subseteq \Theta_d$ denotes the invariant parameter space. We define the invariant M-estimator by $\hat{\theta}_{n,d}^G \in \arg\min_{\theta \in \Theta_d^{G_d}} \hat{L}_{n,d}(\theta)$. Let p_d^G denote the local dimension of $\Theta_d^{G_d}$ around $\theta_d^\star$. The following informal statement summarizes our general result; the full theorem, including the regularity assumptions, is given in Appendix C.

Theorem 1 (Invariant M-estimation, informal) *Suppose that $\theta_d^\star \in \Theta_d^{G_d}$, the loss is uniformly strongly convex on the invariant model, and the empirical gradient at $\theta_d^\star$, restricted to $\Theta_d^{G_d}$, is uniformly sub-Gaussian. Then, with probability at least $1-\delta$, $\left\|\hat{\theta}_{n,d}^G - \theta_d^\star\right\|_2^2 \lesssim \frac{\sigma^2}{\alpha^2} \frac{p_d^G + \log(1/\delta)}{n}$. In particular, if $p^G := \sup_d p_d^G < \infty$, then the same bound holds with p_d^G replaced by p^G, uniformly over d. Moreover, if the transfer maps $\Phi_{d \to D}$ preserve the targets and distances on the invariant model, then with probability at least $1-\delta$*

$$\sup_{D \geq d} \left\|\Phi_{d \to D} \hat{\theta}_{n,d}^G - \theta_D^\star\right\|_2^2 \lesssim \frac{\sigma^2}{\alpha^2} \frac{p^G + \log(1/\delta)}{n}.$$

Theorem 1 shows that invariant estimation can obey dimension-free scaling laws whenever the invariant dimension p_d^G remains bounded with the ambient size d. Thus, invariance changes the effective sample complexity from one controlled by the raw dimension p_d to one controlled by the intrinsic invariant dimension p_d^G.

Low-degree moments on sets and graphs. This principle is especially transparent for moment estimation using low-degree invariant polynomial features. For sets, degree-k invariant features are multisymmetric polynomials whose dimension stabilizes once $d \geq k$; denote the stable dimension by $N_r(k)$. For graphs with node and edge attributes, degree-k invariant features are indexed by unlabeled attributed multigraph patterns and stabilize once $d \geq 2k$; denote the stable dimension by $M_{r,r'}(k)$. Applying Theorem 1 gives, with probability at least $1-\delta$,

$$\left\|\hat{\theta}_{n,d}^G - \theta_d^\star\right\|_2^2 \lesssim \frac{\sigma^2}{\alpha^2} \frac{N_r(k) + \log(1/\delta)}{n}, \quad \left\|\hat{\theta}_{n,d}^G - \theta_d^\star\right\|_2^2 \lesssim \frac{\sigma^2}{\alpha^2} \frac{M_{r,r'}(k) + \log(1/\delta)}{n}.$$

Thus, for fixed degree k, invariant moment estimation on sets and graphs has sample complexity independent of the number of elements or nodes. The growth of $N_r(k)$ and $M_{r,r'}(k)$ is recalled in the score-matching section, where it determines the dimension-free scaling laws.

4. Main Results: Invariant Score Matching

We now specialize the general invariant M-estimation framework to score matching on variable-size sets and graphs. The main finding is that invariant score matching obeys dimension-free scaling laws: once the invariant feature hierarchy stabilizes, the statistical rate no longer depends directly on the number of elements or nodes; it is determined by the complexity of the invariant score itself.

Invariant score features. For set-valued data $X = (x_1, \dots, x_d) \in (\mathbb{R}^r)^d$ and graph-valued data with node attributes in $\mathbb{R}^r$ and edge attributes in $\mathbb{R}^{r'}$, assume that μ_d admits an invariant density p_d. Then the score $s_d^\star = \nabla \log p_d$ is equivariant under the corresponding S_d-action. We model $s_d^\star$ using gradients of invariant Hermite features, $s_d^\star = \sum_{j \geq 1} \theta_{d,j}^\star \psi_{d,j}$, ordered by degree. For cutoff k, the number of equivariant score features stabilizes once $d \geq k$ for sets and $d \geq 2k$ for graphs, and satisfies

$$N_r(k) = \exp\Big(C_r k^{\frac{r}{r+1}} + o\Big(k^{\frac{r}{r+1}}\Big)\Big), \quad M_{r,r'}(k) = \exp\Big(k \log k + O_{r,r'}(k \log\log k)\Big).$$

Here $N_r(k)$ and $M_{r,r'}(k)$ are the stable dimensions of the set and graph score-feature spaces, respectively; the latter are indexed by unlabeled attributed graph patterns. Their different growth laws determine the different scaling thresholds for sets and graphs.

Theorem 2 (Dimension-free scaling laws for invariant score matching, informal) *Let $\hat{s}_{n,d}$ be the optimally truncated invariant score-matching estimator learned at size d, and let $\Phi_{d\to D}$ transfer it to any $D \geq d$. An exponential source condition with exponent $\nu > 0$ means that the invariant coefficient tail decays as $\exp(-ak^\nu)$, while an algebraic source condition with exponent $\tau > 0$ means decay as $k^{-\tau}$. Under the finite-dimensional invariant M-estimation conditions, the following bounds hold with probability at least $1-\delta$, uniformly over source and target sizes. Full statements are given in Appendix D.1 and Appendix D.2:*

$$\|\Phi_{d\to D}\hat{s}_{n,d} - s_D^\star\|^2 \lesssim \begin{cases} n^{-1+o(1)}, & \textit{Exp., } \nu > \frac{r}{r+1}, \\ n^{-c}, & \textit{Exp., } \nu = \frac{r}{r+1}, \\ e^{-c(\log n)^{\frac{\nu(r+1)}{r}}}, & \textit{Exp., } 0 < \nu < \frac{r}{r+1}, \\ (\log n)^{-\frac{\tau(r+1)}{r}}, & \textit{Alg., } \tau, \end{cases} \qquad \begin{cases} n^{-1+o(1)}, & \textit{Exp., } \nu > 1, \\ e^{-c\frac{\log n}{\log\log n}}, & \textit{Exp., } \nu = 1, \\ e^{-c(\frac{\log n}{\log\log n})^\nu}, & \textit{Exp., } 0 < \nu < 1, \\ (\frac{\log n}{\log\log n})^{-\tau}, & \textit{Alg., } \tau. \end{cases}$$

Sets *Graphs*

The hidden constants depend on the source and feature-growth parameters, but not on d or D.

Theorem 2 gives the scaling laws for invariant score matching. For exponential source conditions, there is a critical threshold: $\nu = r/(r+1)$ for sets and $\nu = 1$ for graphs. Above the threshold, the rate is nearly parametric; at or below it, the rate slows down, and under algebraic source conditions, convergence is still dimension-free but only logarithmic, with different logarithmic exponents for sets and graphs. Thus, the rate is controlled by invariant coefficient decay rather than by the number of elements or nodes.

Conclusion. Invariant score matching admits dimension-free scaling laws, but the exponent is governed by invariant complexity. The critical threshold is determined by the growth rate of the invariant feature hierarchy: above this threshold, the rate is nearly parametric; below it, achieving the same accuracy may require exponentially more samples.

References

[1] Matan Atzmon, Koki Nagano, Sanja Fidler, Sameh Khamis, and Yaron Lipman. Frame averaging for equivariant shape space learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Umberto Benedetto, Stuart J Head, Gianni D Angelini, and Eugene H Blackstone. Statistical primer: propensity score matching and its alternatives. *European Journal of Cardio-Thoracic Surgery*, 53(6):1112–1117, 2018.

[3] Beatrice Bevilacqua, Yangze Zhou, and Bruno Ribeiro. Size-invariant graph representations for graph classification extrapolations. In *Int. Conference on Machine Learning (ICML)*, 2021.

[4] Alexander Bogatskiy, Brandon Anderson, Jan Offermann, Marwah Roussi, David Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics. In *Int. Conference on Machine Learning (ICML)*, 2020.

[5] Pietro Bongini, Monica Bianchini, and Franco Scarselli. Molecular generative graph neural networks for drug discovery. *Neurocomputing*, 450:242–252, 2021.

[6] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[7] Chen Cai and Yusu Wang. Convergence of invariant graph networks. In *Int. Conference on Machine Learning (ICML)*, pages 2457–2484, 2022.

[8] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

[9] Ziyu Chen, Markos Katsoulakis, Luc Rey-Bellet, and Wei Zhu. Sample complexity of probability divergences under group symmetry. In *Int. Conference on Machine Learning (ICML)*, 2023.

[10] Mateo Díaz, Dmitriy Drusvyatskiy, Jack Kendrick, and Rekha R Thomas. Invariant kernels: Rank stabilization and generalization across dimensions. *arXiv preprint arXiv:2502.01886*, 2025.

[11] Nadav Dym, Hannah Lawrence, and Jonathan W Siegel. Equivariant frames and the impossibility of continuous canonicalization, 2024.

[12] Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[13] Benson Farb. Representation stability. *arXiv preprint arXiv:1404.4065*, 2014.

[14] Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *Int. Conference on Machine Learning (ICML)*, 2020.

[15] Daniel Herbst and Stefanie Jegelka. Higher-order graphon neural networks: Approximation and cut distance. In *Int. Conference on Learning Representations (ICLR)*, 2025.

[16] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[17] Haotian Ju, Dongyue Li, Aneesh Sharma, and Hongyang R Zhang. Generalization in graph neural networks: Improved pac-bayesian bounds on graph diffusion. In *Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[18] 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.

[19] Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. Convergence and stability of graph convolutional networks on large random graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21512–21523, 2020.

[20] Ron Levie. A graphon-signal analysis of graph neural networks. *Advances in Neural Information Processing Systems*, 36:64482–64525, 2023.

[21] Ron Levie, Wei Huang, Lorenzo Bucci, Michael Bronstein, and Gitta Kutyniok. Transferability of spectral graph convolutional neural networks. *Journal of Machine Learning Research*, 22(272):1–59, 2021.

[22] Eitan Levin and Venkat Chandrasekaran. Free descriptions of convex sets. *arXiv preprint arXiv:2307.04230*, 2023.

[23] Eitan Levin and Venkat Chandrasekaran. Limits of weighted graphs via random quotients. *arXiv:2512.23149*, 2026.

[24] Eitan Levin and Mateo Díaz. Any-dimensional equivariant neural networks. In *Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

[25] Eitan Levin, Yuxin Ma, Mateo Díaz, and Soledad Villar. On transferring transferability: Towards a theory for size generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[26] Xiao Li, Li Sun, Mengjie Ling, and Yan Peng. A survey of graph neural network based recommendation in social networks. *Neurocomputing*, 549:126441, 2023.

[27] Renjie Liao, Raquel Urtasun, and Richard Zemel. A pac-bayesian approach to generalization bounds for graph neural networks. In *Int. Conference on Learning Representations (ICLR)*, 2020.

[28] Yuchao Lin, Jacob Helwig, Shurui Gui, and Shuiwang Ji. Equivariance via minimal frame averaging for more symmetries and efficiency. In *Int. Conference on Machine Learning (ICML)*, 2024.

[29] Siwei Lyu. Interpretation and generalization of score matching. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

[30] 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.

[31] Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *Int. Conference on Learning Representations (ICLR)*, 2019.

[32] Sohir Maskey, Ron Levie, Yunseok Lee, and Gitta Kutyniok. Generalization analysis of message passing neural networks on large random graphs. *Advances in neural information processing systems*, 35:4805–4817, 2022.

[33] Sohir Maskey, Ron Levie, and Gitta Kutyniok. Transferability of graph neural networks: an extended graphon approach. *Applied and Computational Harmonic Analysis*, 63:48–83, 2023.

[34] Sohir Maskey, Gitta Kutyniok, and Ron Levie. Generalization bounds for message passing networks on mixture of graphons. *SIAM Journal on Mathematics of Data Science*, 7(2):802–825, 2025.

[35] Christopher Morris, Floris Geerts, Jan Tönshoff, and Martin Grohe. Wl meet vc. In *Int. Conference on Machine Learning (ICML)*, 2023.

[36] Quynh T Nguyen, Louis Schatzki, Paolo Braccia, Michael Ragone, Patrick J Coles, Frederic Sauvage, Martin Larocca, and Marco Cerezo. Theory for equivariant quantum neural networks. *PRX Quantum*, 5(2):020328, 2024.

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

[38] Trung V Phan, George A Kevrekidis, Soledad Villar, Yannis G Kevrekidis, and Juan M Bello-Rivas. Towards coordinate-and dimension-agnostic machine learning for partial differential equations. *arXiv preprint arXiv:2505.16549*, 2025.

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

[40] Levi Rauchwerger, Stefanie Jegelka, and Ron Levie. Generalization, expressivity, and universality of graph neural networks on attributed graphs. In *Int. Conference on Learning Representations (ICLR)*, 2025.

[41] Luana Ruiz, Luiz Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1702–1712, 2020.

[42] Luana Ruiz, Luiz FO Chamon, and Alejandro Ribeiro. Transferability properties of graph neural networks. *IEEE Transactions on Signal Processing*, 71:3474–3489, 2023.

[43] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Int. Conference on Machine Learning (ICML)*, 2021.

[44] Franco Scarselli, Ah Chung Tsoi, and Markus Hagenbuchner. The Vapnik–Chervonenkis dimension of graph and recursive neural networks. *Neural Networks*, 108:248–259, 2018.

[45] Zakhar Shumaylov, Peter Zaika, James Rowbottom, Ferdia Sherry, Melanie Weber, and Carola-Bibiane Schönlieb. Lie algebra canonicalization: Equivariant neural operators under arbitrary Lie groups. In *Int. Conference on Learning Representations (ICLR)*, 2025.

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

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

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

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

[50] Antonis Vasileiou, Ben Finkelshtein, Floris Geerts, Ron Levie, and Christopher Morris. Covered forest: Fine-grained generalization analysis of graph neural networks. In *Int. Conference on Machine Learning (ICML)*, 2025.

[51] Antonis Vasileiou, Stefanie Jegelka, Ron Levie, and Christopher Morris. Survey on generalization theory for graph neural networks. *arXiv preprint arXiv:2503.15650*, 2025.

[52] Dian Wang, Robin Walters, and Robert Platt. $\mathrm{SO}(2)$-equivariant reinforcement learning. In *Int. Conference on Learning Representations (ICLR)*, 2022.

[53] Rui Wang, Robin Walters, and Rose Yu. Incorporating symmetry into deep dynamics models for improved generalization. In *Int. Conference on Learning Representations (ICLR)*, 2021.

[54] Melanie Weber. Geometric machine learning. *Wiley Online Library*, 2025.

[55] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *Int. Conference on Learning Representations (ICLR)*, 2019.

[56] Gilad Yehudai, Ethan Fetaya, Eli Meirom, Gal Chechik, and Haggai Maron. From local structures to size generalization in graph neural networks. In *Int. Conference on Machine Learning (ICML)*, 2021.

[57] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Appendix A. Related Work

Transferability and size generalization. Transferability across graph sizes has been studied extensively in the GNN literature, often through graph limits and graphon-based analysis [19, 21, 41]. Related work further develops transferability guarantees for graph neural networks and graph-based learning methods using graphon and higher-order limit perspectives [7, 15, 33, 42]. These works primarily study architectural transferability or stability of learned graph models. In contrast, we study transferability as a statistical estimation problem, asking when an estimator learned on one domain can achieve dimension-uniform convergence on another.

Representation stability and any-dimensional learning. Our work builds on representation stability [13], which describes how structured families of representations stabilize across dimensions. Representation stability has recently been used in machine learning to design and analyze architectures for learning under symmetries [24], optimization across dimensions [22], graph limits [23], scientific computing and PDEs [38], and finite-dimensional kernel learning under invariances [10]. Most closely related is Levin et al. [25], which proves size-generalization bounds in a broad setting that includes sets and graphs. While their framework establishes transfer across dimensions, their rates still depend on dimension-dependent quantities, such as the covering numbers of the domains. Our score-matching results instead identify regimes where the rates become fully independent of dimension and characterize the spectral thresholds that separate fast, nearly parametric transfer from regimes requiring exponentially larger sample complexity. To our knowledge, this sharp transition has not appeared in prior work and is specific to the spectral complexity viewpoint developed in this paper.

Learning under invariances. Symmetry and invariance are central principles in geometric machine learning. For sets, a classical architecture is Deep Sets [57]; for graphs, graph neural networks and higher-order invariant/equivariant architectures provide widely used models for permutation-equivariant learning [31, 43, 55]. Beyond architectures that hard-code invariance, several architecture-agnostic approaches have been developed, including canonicalization [11, 18, 30, 45] and frame averaging [1, 28, 39]. Applications of invariant learning span reinforcement learning [52], physics [4, 36], drug discovery [5], and social networks [26]. Our work is complementary to these modeling approaches: rather than proposing a new architecture, we study when invariant statistical estimation can transfer across dimensions.

Generalization and sample complexity under symmetry. The statistical benefits of symmetry have been widely studied in learning theory. Prior work has shown that invariance can reduce sample complexity in supervised learning and related estimation problems [12, 37, 47, 49, 53], with extensions to estimating probability divergences [9, 48]. Other work studies how such gains can be made computationally efficient [46]. Generalization bounds for graph neural networks have also been developed using tools such as Rademacher complexity [14], VC dimension [35, 44], and PAC-Bayesian analysis [17, 27]. Further results on GNN generalization and transfer include [20, 32, 34, 40, 50]; see also the survey of Vasileiou et al. [51]. Our contribution differs from these works by focusing on invariant estimation across changing dimensions, with explicit dimension-uniform rates and phase transitions for score matching.

Score matching. Score matching was introduced by Hyvärinen and Dayan [16] as a method for estimating unnormalized statistical models through their score functions, and has since become a

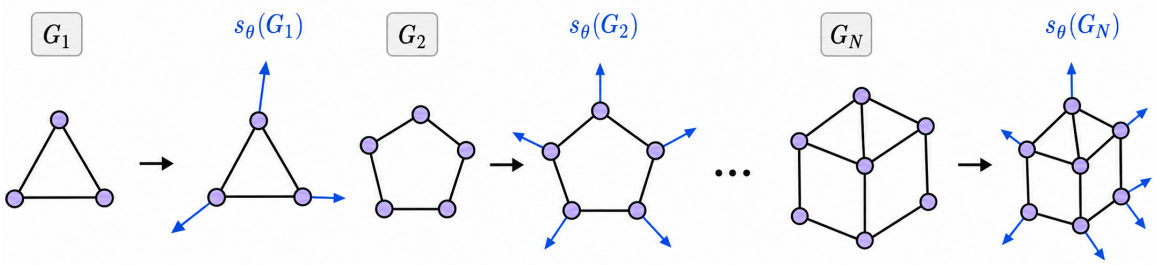

Figure 1: A learned score function maps any graph to a vector field on its nodes. The graphs can have different numbers of nodes (variable-size), the model shares parameters across all sizes.

foundational tool in modern score-based generative modeling. Classical estimation and statistical properties of score matching have been studied in [2, 8, 16, 29]. Our work studies score matching in a different regime: invariant data whose domain size may change between training and testing. We show that transfer across dimensions is possible precisely in low spectral-complexity regimes, while high-complexity scores may require exponentially larger sample complexity.

Appendix B. Detailed Problem Statement

This section gives a concise overview of the problem formulation. A detailed review of the definitions, background, and preliminaries used throughout the paper is provided in Appendix E.

Preliminaries on M-estimators. We adopt the M-estimation framework to formalize the problem, using $d \in \mathbb{N}$ as an index, allowing us to study collections of related problems jointly across different settings (i.e., dimensions). For each d, let $\mathcal{X}_d$ be a measurable space (e.g., $\mathbb{R}^d$) and let μ_d be a probability distribution on $\mathcal{X}_d$. Let $\Theta_d \subseteq \mathbb{R}^{p_d}$ denote the parameter space, $p_d \in \mathbb{N}$, and let $\ell_d : \mathcal{X}_d \times \Theta_d \to \mathbb{R}$ be a measurable loss function. The target parameter is defined as

$$\theta_d^\star \in \arg\min_{\theta \in \Theta_d} \Big\{ L_d(\theta) := \mathbb{E}_{X \sim \mu_d}[\ell_d(X, \theta)] \Big\},$$

and represents the quantity of interest in the estimation problem considered in this paper.

In a classical estimation setting, the distribution μ_d is unknown, and only independent and identically distributed samples $X_1, \ldots, X_n \sim \mu_d$ are available. The corresponding M-estimator is defined as

$$\hat{\theta}_{n,d} \in \arg\min_{\theta \in \Theta_d} \Big\{ \hat{L}_{n,d}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell_d(X_i, \theta) \Big\}.$$

Note that the parameter $\theta_d^\star$ represents the population minimizer, while $\hat{\theta}_{n,d}$ is its empirical counterpart. The main objective in this setting is to characterize the conditions and sample complexity under which $\hat{\theta}_{n,d}$ converges to and accurately approximates $\theta_d^\star$.

The M-estimation framework recovers several classical estimators as special cases:

- **Mean estimation:** Let $\mathcal{X}_d = \mathbb{R}^d$, $\Theta_d = \mathbb{R}^d$, and $\ell_d(x, \theta) = \|x - \theta\|_2^2$. Then,

$$\theta_d^\star = \mathbb{E}_{X \sim \mu_d}[X], \quad \hat{\theta}_{n,d} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- **Covariance estimation:** Let $\mathcal{X}_d = \mathbb{R}^d$, $\Theta_d = \mathbb{R}^{d \times d}$, and $\ell_d(x, \theta) = \|xx^\top - \theta\|_F^2$. Then

$$\theta_d^\star = \mathbb{E}[XX^\top], \quad \hat{\theta}_{n,d} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top.$$

- **Higher-order moment (tensor) estimation:** For a fixed order $k \geq 1$, let $\Theta_d = (\mathbb{R}^d)^{\otimes k}$ and define $\ell_d(x, \theta) = \|x^{\otimes k} - \theta\|_F^2$. Then,

$$\theta_d^\star = \mathbb{E}[X^{\otimes k}], \quad \hat{\theta}_{n,d} = \frac{1}{n} \sum_{i=1}^n X_i^{\otimes k}.$$

- **Maximum likelihood estimation.** Let $\{p_\theta : \theta \in \Theta_d\}$ be a parametric family of densities on $\mathcal{X}_d$, and define $\ell_d(x, \theta) = -\log p_\theta(x)$. Then,

$$\theta_d^\star \in \arg\min_\theta \mathbb{E}[-\log p_\theta(X)], \quad \hat{\theta}_{n,d} \in \arg\min_\theta \frac{1}{n} \sum_{i=1}^n -\log p_\theta(X_i).$$

- **Empirical risk minimization (ERM):** Let $\mathcal{X}_d = \mathfrak{X}_d \times \mathcal{Y}$ denote input-output pairs. For $(x, y) \in \mathcal{X}_d$, define $\ell_d((x, y), \theta) = \ell\left(f_\theta(x), y\right)$. Then

$$\theta_d^\star \in \arg\min_\theta \mathbb{E}[\ell(f_\theta(X), Y)], \quad \hat{\theta}_{n,d} \in \arg\min_\theta \frac{1}{n} \sum_{i=1}^n \ell(f_\theta(X_i), Y_i).$$

- **Score matching and denoising.** Let $\mathcal{X}_d = \mathbb{R}^d$, and let Θ_d parameterize vector fields $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$. Score estimation can be formulated within the M-estimation framework via suitable loss functions. In particular, score matching defines $\ell_d(x, \theta) = \frac{1}{2}\|s_\theta(x)\|_2^2 + \nabla \cdot s_\theta(x)$, leading to $\theta_d^\star \in \arg\min_\theta \mathbb{E}[\ell_d(X, \theta)]$. Alternatively, denoising score matching introduces noisy observations $\tilde{X} = X + \sigma\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I)$, and defines $\ell_d((x, \tilde{x}), \theta) = \|s_\theta(\tilde{x}) + \frac{1}{\sigma^2}(\tilde{x} - x)\|_2^2$, yielding $\theta_d^\star \in \arg\min_\theta \mathbb{E}[\ell_d((X, \tilde{X}), \theta)]$. Assuming μ_d admits a density p_d, these formulations estimate the score function $\nabla_x \log p_d(x)$ and are central to modern diffusion-based generative models.

Groups and symmetries. We next introduce the invariance structure used throughout the paper. For each $d \in \mathbb{N}$, let G_d be a group acting on $\mathcal{X}_d$. That is, each $g \in G_d$ defines a measurable bijection $x \mapsto g \cdot x$ on $\mathcal{X}_d$, satisfying $e \cdot x = x$ and $g \cdot (h \cdot x) = (gh) \cdot x$ for all $g, h \in G_d$, where $e \in G_d$ denotes the identity element of the group. We assume that the data distribution is invariant under this action, meaning that $g \cdot X \sim \mu_d$ whenever $X \sim \mu_d$, for every $g \in G_d$.

We further assume that G_d acts on the parameter space Θ_d, and that the loss is compatible with the two actions in the sense that $\ell_d(g \cdot x, g \cdot \theta) = \ell_d(x, \theta)$, for all $g \in G_d,\ x \in \mathcal{X}_d,\ \theta \in \Theta_d$.

Under this compatibility condition, the population risk is invariant: $L_d(g \cdot \theta) = L_d(\theta)$, for all $g \in G_d$, $\theta \in \Theta_d$. Define the invariant parameter space $\Theta_d^{G_d} := \{\theta \in \Theta_d : g \cdot \theta = \theta \text{ for all } g \in G_d\}$. If the population minimizer $\theta_d^\star$ is unique, then $\theta_d^\star \in \Theta_d^{G_d}$. Indeed, since $L_d(g \cdot \theta_d^\star) = L_d(\theta_d^\star)$, the parameter $g \cdot \theta_d^\star$ is also a minimizer of L_d, and uniqueness therefore implies $g \cdot \theta_d^\star = \theta_d^\star$ for all $g \in G_d$.

Any-dimensional estimation. In classical estimation, the goal is to ensure that an empirical estimator $\hat{\theta}_{n,d}$ accurately approximates the population target $\theta_d^\star$ for a fixed problem indexed by d. In the any-dimensional setting, we ask for a stronger guarantee: an estimator constructed from samples at one dimension should remain meaningful after being transferred to larger dimensions.

We consider a family of invariant estimation problems indexed by $d \in \mathbb{N}$. For each dimension d, the parameter space Θ_d is equipped with a group action, and the target parameter $\theta_d^\star$ is constrained by the corresponding invariance structure. For every pair $d \le D$, suppose we are given a transfer map

$$\Phi_{d\to D} : \Theta_d \to \Theta_D,$$

which embeds parameters from dimension d into the parameter space at dimension D in a way that is compatible with the invariant structure. Given samples $X_1, \dots, X_n \sim \mu_d$, we compute an invariant estimator $\hat{\theta}_{n,d} \in \Theta_d^{G_d}$ and then transfer it to $\Theta_D^{G_D}$ via $\Phi_{d\to D}\hat{\theta}_{n,d}$. The goal is to control its error relative to the invariant target $\theta_D^\star \in \Theta_D^{G_D}$ uniformly over all larger dimensions $D \ge d$.

Invariant any-dimensional estimation. Given samples $X_1, \dots, X_n \sim \mu_d$, compute an invariant estimator $\hat{\theta}_{n,d}$ and transfer it to any larger dimension $D \ge d$ through a map $\Phi_{d\to D} : \Theta_d \to \Theta_D$. The goal is to have a sample size $n = n(\varepsilon, \delta)$, independent of d and D, such that

$$\sup_{D \ge d} \left\| \Phi_{d\to D}\hat{\theta}_{n,d} - \theta_D^\star \right\|_2 \le \varepsilon, \quad \text{with probability at least } 1-\delta.$$

Canonical examples: sets and graphs. We conclude with two motivating examples; a more detailed discussion is provided in Appendix F. First, consider set-valued data with d elements in $\mathbb{R}^r$, represented as $X = (x_1, \dots, x_d) \in (\mathbb{R}^r)^d$, where the ordering is arbitrary. The symmetric group S_d acts by permuting elements, $\pi \cdot X = (x_{\pi^{-1}(1)}, \dots, x_{\pi^{-1}(d)})$. The assumption that μ_d is S_d-invariant means that the distribution does not depend on the chosen ordering, i.e., the represented sequence is exchangeable. To see a nontrivial parameter action, let $\Theta_d = (\mathbb{R}^r)^d$ and consider the loss

$$\ell_d(X, \theta) = \frac{1}{d}\sum_{i=1}^{d} \|x_i - \theta_i\|_2^2, \qquad \theta = (\theta_1, \dots, \theta_d).$$

The group acts on Θ_d by $\pi \cdot \theta = (\theta_{\pi^{-1}(1)}, \dots, \theta_{\pi^{-1}(d)})$, and the loss satisfies $\ell_d(\pi \cdot X, \pi \cdot \theta) = \ell_d(X, \theta)$. Exchangeability implies that the population minimizer lies in the invariant subspace, so $\theta_1^\star = \cdots = \theta_d^\star$, recovering the usual set mean as the invariant representative.

Second, consider graph-valued data represented by an edge-attributed adjacency tensor $A \in \mathbb{R}^{d\times d\times r'}$, where $A_{ij,b}$ denotes the b-th attribute of the edge from node i to node j. Node relabeling acts as $(\pi \cdot A)_{ij,b} = A_{\pi^{-1}(i),\pi^{-1}(j),b}$. In the scalar edge case $r' = 1$, this reduces to the usual action $A \mapsto P_\pi A P_\pi^\top$, where P_π is the permutation matrix associated with $\pi \in S_d$. Thus, S_d-invariance of μ_d means that the graph distribution is independent of node labels. Classical graph

statistics, including edge statistics, triangle counts, subgraph counts, and spectral moments in the scalar edge case, are invariant under this action and fit the M-estimation framework through losses $\ell_d(A,\theta) = \|\varphi_d(A) - \theta\|_2^2$, where φ_d is node-relabeling invariant. If μ_d admits an invariant density p_d, then its score $s_d^\star(A) = \nabla_A \log p_d(A)$ is node-permutation equivariant:

$$s_d^\star(\pi \cdot A) = \pi \cdot s_d^\star(A), \qquad (\pi \cdot s_d^\star(A))_{ij,b} = s_d^\star(A)_{\pi^{-1}(i),\pi^{-1}(j),b}.$$

The same principle applies to attributed graphs (A, H), with node features $H \in \mathbb{R}^{d\times r}$, under the diagonal action $(A, H) \mapsto (\pi \cdot A, \pi \cdot H)$, where $(\pi \cdot H)_{i,a} = H_{\pi^{-1}(i),a}$.

Appendix C. Main Results: Invariant M-Estimation

We first provide a general statistical guarantee for invariant M-estimators and then instantiate it for moment parameters on sets and graphs.

C.1. General invariant M-estimation

For each d, consider the invariant parameter space $\Theta_d^{G_d} \subseteq \Theta_d$ introduced in Section B. We define the invariant M-estimator by

$$\hat\theta_{n,d}^G \in \arg\min_{\theta\in\Theta_d^{G_d}} \hat L_{n,d}(\theta).$$

To make the dimension precise, we assume that $\Theta_d^{G_d}$ is a smooth finite-dimensional model in a neighborhood of $\theta_d^\star$, and denote its local dimension by $p_d^G := \dim_{\theta_d^\star}(\Theta_d^{G_d})$. When the invariant dimensions stabilize, we write $p^G := \sup_d p_d^G < \infty$.

The following result gives the sharp parametric rate on the invariant model.

Theorem 3 (Invariant M-estimation) *Assume that $\theta_d^\star \in \Theta_d^{G_d}$ for every d, and suppose that $\Theta_d^{G_d}$ is a finite-dimensional invariant model as defined above. Assume that, for every $x \in \mathcal{X}_d$, the loss $\ell_d(x,\cdot)$ is α-strongly convex on $\Theta_d^{G_d}$, uniformly over d. Suppose also that the empirical gradient at $\theta_d^\star$, restricted to $\Theta_d^{G_d}$, is sub-Gaussian with variance proxy at most σ^2, uniformly over d. Then there exists an absolute constant $C > 0$ such that, for every d and every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left\|\hat\theta_{n,d}^G - \theta_d^\star\right\|_2^2 \le C\frac{\sigma^2}{\alpha^2}\frac{p_d^G + \log(1/\delta)}{n}.$$

In particular, if $p^G := \sup_d p_d^G < \infty$, then, for every d, with probability at least $1-\delta$,

$$\left\|\hat\theta_{n,d}^G - \theta_d^\star\right\|_2^2 \le C\frac{\sigma^2}{\alpha^2}\frac{p^G + \log(1/\delta)}{n}.$$

Moreover, suppose that for every $D \ge d$, the transfer maps $\Phi_{d\to D}$ satisfy $\Phi_{d\to D}\theta_d^\star = \theta_D^\star$ and are distance preserving on the invariant model, i.e.,

$$\left\|\Phi_{d\to D}\theta - \Phi_{d\to D}\theta'\right\|_2 = \left\|\theta - \theta'\right\|_2 \qquad \forall \theta,\theta' \in \Theta_d^{G_d}.$$

Then, with probability $1-\delta$,

$$\sup_{D\ge d}\left\|\Phi_{d\to D}\hat\theta_{n,d}^G - \theta_D^\star\right\|_2^2 \le C\frac{\sigma^2}{\alpha^2}\frac{p^G + \log(1/\delta)}{n}.$$

Theorem 3 shows that the statistical complexity of invariant estimation is controlled by the quantity p_d^G, not by the raw parameter dimension p_d. Therefore, if p_d^G remains bounded or grows slowly with d, the sample complexity of estimating and transferring invariant parameters can be independent of the ambient problem size. This is the mechanism behind any-dimensional estimation in our framework.

C.2. Application: moment estimation on sets and graphs

Theorem 3 becomes particularly informative when the invariant dimension p_d^G stabilizes with the size parameter d. This occurs naturally for moment parameters defined by low-degree invariant polynomial features. In this case, $\theta_d^\star$ is the expectation of a vector of invariant degree-k features, and the corresponding squared loss is strongly convex in the coefficient parameter. Thus the theorem gives a rate controlled by the number of invariant features rather than the raw number of degree-k monomials.

For set-valued data $X = (x_1, \dots, x_d) \in (\mathbb{R}^r)^d$, with S_d acting by permuting elements, invariant homogeneous degree-k polynomial features are multisymmetric polynomials. As shown in Appendix E.4, their dimension stabilizes once $d \geq k$. Denoting this stable dimension by $N_r(k)$, we have, for fixed r,

$$N_r(k) = \exp\left(C_r k^{\frac{r}{r+1}} + o\left(k^{\frac{r}{r+1}}\right)\right),$$

where $C_r > 0$ is a constant depending only on r. Applying Theorem 3 gives, with probability at least $1 - \delta$,

$$\left\|\hat{\theta}_{n,d}^G - \theta_d^\star\right\|_2^2 \leq C\frac{\sigma^2}{\alpha^2}\frac{N_r(k) + \log(1/\delta)}{n}, \qquad d \geq k.$$

Thus, for fixed r and k, low-degree invariant moment estimation on sets has sample complexity independent of the number of set elements.

For graph-valued data, consider node attributes in $\mathbb{R}^r$ and edge attributes in $\mathbb{R}^{r'}$, with S_d acting by node relabeling. Invariant homogeneous degree-k polynomial features are indexed by unlabeled attributed multigraph patterns of total degree k. Let $M_{r,r'}(k)$ denote this stable dimension. As shown in Appendix E.6, $p_d^G = M_{r,r'}(k)$ once $d \geq 2k$. Consequently, Theorem 3 yields

$$\left\|\hat{\theta}_{n,d}^G - \theta_d^\star\right\|_2^2 \leq C\frac{\sigma^2}{\alpha^2}\frac{M_{r,r'}(k) + \log(1/\delta)}{n}, \qquad d \geq 2k,$$

with probability at least $1 - \delta$. If $r' \geq 1$, then

$$M_{r,r'}(k) = \exp(k \log k + O(k \log\log k)).$$

Appendix D. Main Results: Score Matching

We now turn to convergence guarantees for score matching on data represented as sets and graphs.

D.1. Score matching on set-valued data

We next consider score matching on set-valued data. Let $X = (x_1, \dots, x_d) \in (\mathbb{R}^r)^d$, and assume that μ_d admits an S_d-invariant density. Then the score $s_d^\star = \nabla \log p_d$ is permutation equivariant. We model $s_d^\star$ using gradients of invariant Hermite features. Let $\{\psi_{d,j}\}_{j \geq 1}$ be an ordered basis of

permutation-equivariant score features, ordered by the degree of the underlying invariant Hermite polynomial, and write

$$s_d^\star = \sum_{j\geq 1} \theta_{d,j}^\star \psi_{d,j}.$$

We measure score estimation error in the coefficient norm

$$\|s_d - s_d^\star\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} := \sum_{j\geq 1} (\theta_{d,j} - \theta_{d,j}^\star)^2.$$

For a degree cutoff k, let $\mathcal{I}_{d,k}$ be the set of invariant score features of degree at most k, and define

$$m_r(k) := |\mathcal{I}_{d,k}|.$$

For $d \geq k$, this quantity is independent of d, and by Appendix E.4,

$$m_r(k) = \exp\left(C_r k^{\frac{r}{r+1}} + o\left(k^{\frac{r}{r+1}}\right)\right),$$

where $C_r > 0$ depends only on r.

We now define two source norms for scores. For $a, \nu > 0$, define the exponential source norm

$$\|s_d^\star\|^2_{\mathsf{Exp}(a,\nu)} := \sup_{k\geq 1} \left\{ e^{ak^\nu} \sum_{j\notin\mathcal{I}_{d,k}} (\theta_{d,j}^\star)^2 \right\}.$$

For $\tau > 0$, define the algebraic source norm

$$\|s_d^\star\|^2_{\mathsf{Alg}(\tau)} := \sup_{k\geq 1} \left\{ k^\tau \sum_{j\notin\mathcal{I}_{d,k}} (\theta_{d,j}^\star)^2 \right\}.$$

These norms quantify how well the true score is approximated by low-degree score features.

Theorem 4 (Invariant score matching on sets) *Assume the truncated score matching estimator over $\mathcal{I}_{d,k}$ satisfies the finite-dimensional invariant M-estimation conditions of Theorem 3, uniformly over d and k, with constants $\alpha, \sigma > 0$. Let $\hat{s}_{n,d}^{(k)}$ be the score estimator using invariant score features of degree at most k. Then there exists an absolute constant $C > 0$ such that, for every $d \geq k$ and every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\hat{s}_{n,d}^{(k)} - s_d^\star\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} \leq \sum_{j\notin\mathcal{I}_{d,k}} (\theta_{d,j}^\star)^2 + C\frac{\sigma^2}{\alpha^2}\frac{m_r(k) + \log(1/\delta)}{n}.$$

Consequently,

$$\inf_{k\geq 1} \|\hat{s}_{n,d}^{(k)} - s_d^\star\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} \leq \inf_{k\geq 1} \left\{ \sum_{j\notin\mathcal{I}_{d,k}} (\theta_{d,j}^\star)^2 + C\frac{\sigma^2}{\alpha^2}\frac{m_r(k) + \log(1/\delta)}{n} \right\}.$$

Moreover, the following optimized rates hold:

(i) ***Exponential source condition.*** *If* $\sup_d \|s_d^\star\|_{\mathsf{Exp}(a,\nu)} < \infty$*, then*

$$\inf_{k\geq 1} \|\hat{s}_{n,d}^{(k)} - s_d^\star\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} \leq \begin{cases} n^{-1+o(1)}, & \nu > \frac{r}{r+1}, \\ n^{-c_1}, & \nu = \frac{r}{r+1}, \\ \exp\left(-c_2(\log n)^{\frac{\nu(r+1)}{r}}\right), & \nu < \frac{r}{r+1}, \end{cases}$$

where $c_1, c_2 > 0$ *are constants depending on the source condition and the invariant feature growth. Indeed,* $c_1 = \frac{a}{a+C_r} \in (0,1)$.

(ii) ***Algebraic source condition.*** *If* $\sup_d \|s_d^\star\|_{\mathsf{Alg}(\tau)} < \infty$*, then*

$$\inf_{k\geq 1} \|\hat{s}_{n,d}^{(k)} - s_d^\star\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} \leq (\log n)^{-\frac{\tau(r+1)}{r}}.$$

Theorem 4 shows that score matching across set sizes is governed by a bias-variance tradeoff over the degree cutoff k. The variance is controlled by the number of invariant score features, which grows as $\exp(C_r k^{r/(r+1)} + o(k^{r/(r+1)}))$, while the bias is controlled by the source norm of the score. The exponential source norm yields a sharp phase transition at $\nu = r/(r+1)$. Above this threshold, the estimator is nearly parametric; at the threshold, the rate is polynomial; below it, the rate is stretched-logarithmic. By contrast, the algebraic source norm is much weaker and yields only logarithmic convergence. Thus the source norm determines whether invariant score matching admits fast sample complexity or only slow consistency.

D.2. Score matching on graph-valued data

We finally consider score matching on graph-valued data. Let the graph have node attributes $H \in \mathbb{R}^{d\times r}$ and edge attributes $A \in \mathbb{R}^{d\times d\times r'}$, with $r' \geq 1$. The symmetric group S_d acts by node relabeling:

$$(\pi \cdot H)_{i,a} = H_{\pi^{-1}(i),a}, \qquad (\pi \cdot A)_{ij,b} = A_{\pi^{-1}(i),\pi^{-1}(j),b}.$$

If μ_d admits an S_d-invariant density, then its score is node-permutation equivariant. We model this score using gradients of node-relabeling invariant Hermite features. Let $\{\psi_{d,j}\}_{j\geq 1}$ denote an ordered basis of equivariant graph score features, ordered by the degree of the underlying invariant Hermite polynomial, and similar to the previous case write

$$s_d^\star = \sum_{j\geq 1} \theta_{d,j}^\star \psi_{d,j}.$$

We measure score error in the norm

$$\|s_d - s_d^\star\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} := \sum_{j\geq 1} (\theta_{d,j} - \theta_{d,j}^\star)^2.$$

For a degree cutoff k, let $\mathcal{J}_{d,k}$ be the set of equivariant graph score features of degree at most k, and define

$$M_{r,r'}(k) := |\mathcal{J}_{d,k}|.$$

As shown in Appendix E.6, for $d \geq 2k$, this quantity is independent of the number of nodes d, and for $r' \geq 1$,

$$M_{r,r'}(k) = \exp(k \log k + O(k \log \log k)) .$$

We define graph score source norms analogously to the set case.

Theorem 5 (Invariant score matching on graphs) *Assume $r' \geq 1$. Suppose the truncated graph score matching estimator over $\mathcal{J}_{d,k}$ satisfies the finite-dimensional invariant M-estimation conditions of Theorem 3, uniformly over d and k, with constants $\alpha, \sigma > 0$. Let $\hat{s}^{(k)}_{n,d}$ be the graph score estimator using equivariant score features of degree at most k. Then there exists an absolute constant $C > 0$ such that, for every $d \geq 2k$ and every $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\|\hat{s}^{(k)}_{n,d} - s^\star_d\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} \leq \sum_{j \notin \mathcal{J}_{d,k}} (\theta^\star_{d,j})^2 + C \frac{\sigma^2}{\alpha^2} \frac{M_{r,r'}(k) + \log(1/\delta)}{n}.$$

Consequently, optimizing over deterministic cutoffs k gives the bias-variance tradeoff

$$\inf_{k \geq 1} \left\{ \sum_{j \notin \mathcal{J}_{d,k}} (\theta^\star_{d,j})^2 + C \frac{\sigma^2}{\alpha^2} \frac{M_{r,r'}(k) + \log(1/\delta)}{n} \right\}.$$

Moreover, the following optimized rates hold:

(i) ***Exponential source condition.*** *If $\sup_d \|s^\star_d\|_{\mathsf{Exp}(a,\nu)} < \infty$, then*

$$\inf_{k \geq 1} \|\hat{s}^{(k)}_{n,d} - s^\star_d\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} \leq \begin{cases} n^{-1+o(1)}, & \nu > 1, \\ \exp\left(-c_1 \frac{\log n}{\log \log n}\right), & \nu = 1, \\ \exp\left(-c_2 \left(\frac{\log n}{\log \log n}\right)^\nu\right), & 0 < \nu < 1, \end{cases}$$

where $c_1, c_2 > 0$ depend on the source condition and the graph feature growth.

(ii) ***Algebraic source condition.*** *If $\sup_d \|s^\star_d\|_{\mathsf{Alg}(\tau)} < \infty$, then*

$$\inf_{k \geq 1} \|\hat{s}^{(k)}_{n,d} - s^\star_d\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} \leq \left(\frac{\log n}{\log \log n}\right)^{-\tau}.$$

Theorem 5 is the graph analogue of Theorem 4. The main difference is the growth of the invariant feature dimension. For sets, the number of invariant degree-k features grows like $\exp(C_r k^{r/(r+1)})$. For graphs with edge attributes, it grows like $\exp(k \log k + O(k \log \log k))$. Consequently, the phase transition for exponential source conditions occurs at $\nu = 1$. If the graph score has coefficient tail faster than $\exp(-ak)$, then graph score matching is nearly parametric. At the critical scale $\nu = 1$, the rate is subpolynomial but faster than any logarithmic power. For $0 < \nu < 1$, the rate becomes stretched logarithmic in $\log n / \log \log n$. Algebraic source conditions yield consistency, but only at logarithmic speed. Thus graph score matching remains dimension stable in the number of nodes, but its degree complexity is substantially richer than in the set case.

D.3. Discussion: the main conclusion for score matching

The score matching results above reveal a simple principle. Invariant score estimation can be made dimension stable, in the sense that the number of elements or nodes does not directly determine the statistical rate once the invariant feature spaces stabilize. However, dimension stability does not mean that all scores are equally easy to estimate. The rate is governed by a bias and variance tradeoff: low-degree invariant features reduce variance, while the approximation quality of these features controls the bias.

Our source norms make this tradeoff explicit. Scores with rapidly decaying invariant coefficients can be estimated at nearly parametric rates, while scores with slower coefficient decay lead to polynomial, stretched logarithmic, or logarithmic rates. Thus convergence is possible under broad source conditions, but the speed of convergence depends on how efficiently the score can be approximated by stable invariant features.

The distinction between sets and graphs comes from the growth of their invariant feature spaces. Sets have a smaller invariant degree complexity, while graphs with edge attributes have richer graph-pattern complexity. As a result, the phase transition for fast score matching occurs at a different source scale in the two cases. In short, symmetry removes the dependence on raw size, but the approximation structure of the invariant score determines the statistical difficulty.

> Invariant score matching across dimensions is possible when the true score admits a low-degree invariant approximation. Fast rates occur when the score coefficients decay faster than the growth of the invariant feature dimension. For sets, the critical threshold is $\nu = r/(r+1)$; for graphs with edge attributes, the critical threshold is $\nu = 1$. Above the threshold, the optimized rate is nearly parametric; at or below the threshold, the rates become substantially slower, even requiring exponentially more samples in extreme cases.

Appendix E. Preliminaries on Symmetries and Invariant Polynomial Spaces

This section collects all background and notation used throughout the paper. We review group actions, invariant distributions, compatible losses, any-dimensional models, and the invariant polynomial spaces that arise for set-valued and graph-valued data. We also briefly discuss dimension formulas for these spaces and the stabilization as the size parameter d grows.

E.1. Group actions and invariant estimation

A group G is a set equipped with an associative binary operation $G \times G \to G$, denoted by $(g, h) \mapsto gh$, an identity element $e \in G$ satisfying $eg = ge = g$ for all $g \in G$, and an inverse $g^{-1} \in G$ for every $g \in G$. A left action of G on a measurable space $\mathcal{X}$ is a map $G \times \mathcal{X} \to \mathcal{X}$, denoted by $(g, x) \mapsto g \cdot x$, such that

$$e \cdot x = x, \qquad g \cdot (h \cdot x) = (gh) \cdot x,$$

for all $g, h \in G$ and $x \in \mathcal{X}$. Similarly, an action of G on a parameter space Θ is a map $G \times \Theta \to \Theta$, denoted by $(g, \theta) \mapsto g \cdot \theta$, satisfying

$$e \cdot \theta = \theta, \qquad g \cdot (h \cdot \theta) = (gh) \cdot \theta,$$

for all $g, h \in G$ and $\theta \in \Theta$.

A probability distribution μ on $\mathcal{X}$ is G-invariant if

$$g \cdot X \sim \mu \quad \text{whenever} \quad X \sim \mu,\ g \in G.$$

Equivalently, $\mu(g^{-1}A) = \mu(A)$ for all measurable sets $A \subseteq \mathcal{X}$.

A loss $\ell : \mathcal{X} \times \Theta \to \mathbb{R}$ is compatible with the two actions if

$$\ell(g \cdot x, g \cdot \theta) = \ell(x, \theta), \qquad \forall g \in G,\ x \in \mathcal{X},\ \theta \in \Theta.$$

If μ is G-invariant, then the population risk $L(\theta) = \mathbb{E}_{X\sim\mu}[\ell(X, \theta)]$ satisfies

$$L(g \cdot \theta) = L(\theta), \qquad \forall g \in G,\ \theta \in \Theta.$$

Indeed, by compatibility of the loss,

$$L(g \cdot \theta) = \mathbb{E}_{X\sim\mu}[\ell(X, g \cdot \theta)] = \mathbb{E}_{X\sim\mu}[\ell(g^{-1} \cdot X, \theta)].$$

Since μ is G-invariant, $g^{-1} \cdot X \sim \mu$, and therefore

$$\mathbb{E}_{X\sim\mu}[\ell(g^{-1} \cdot X, \theta)] = \mathbb{E}_{X\sim\mu}[\ell(X, \theta)] = L(\theta).$$

Thus the population objective is invariant under the induced action on parameters.

The invariant parameter space is

$$\Theta^G := \{\theta \in \Theta : g \cdot \theta = \theta \text{ for all } g \in G\}.$$

If L has a unique minimizer $\theta^\star$, then $\theta^\star \in \Theta^G$. Indeed, for every $g \in G$, $g \cdot \theta^\star$ is also a minimizer by invariance of L, and uniqueness implies $g \cdot \theta^\star = \theta^\star$.

If the minimizer is not unique, invariant minimizers still exist under standard convexity assumptions. Suppose Θ is convex, the action of G on Θ is linear, G is finite or compact, and L is convex and G-invariant. If $\theta^\star$ is any minimizer, then the group average

$$\bar{\theta}^\star := \mathbb{E}_{g\in G}[g \cdot \theta^\star]$$

is invariant and also minimizes L, since

$$L(\bar{\theta}^\star) \le \mathbb{E}_{g\in G} L(g \cdot \theta^\star) = L(\theta^\star).$$

For finite groups, the expectation over G denotes the uniform average while for compact groups, it is the left Haar probability measure.

E.2. Any-dimensional models

The paper considers families of estimation problems indexed by $d \in \mathbb{N}$. For each d, we have a data space $\mathcal{X}_d$, distribution μ_d, parameter space Θ_d, loss ℓ_d, and population target $\theta_d^\star$. The index d may represent dimension, sequence length, number of nodes, or another size parameter.

An any-dimensional estimation problem additionally specifies transfer maps between parameter spaces. For $d \le D$, let

$$\Phi_{d\to D} : \Theta_d \to \Theta_D$$

be a parameter transfer map. The goal is to estimate $\theta_d^\star$ from samples at size d, transfer the estimator to size D, and compare it to $\theta_D^\star$. A typical dimension-uniform guarantee takes the form

$$\Pr\left(\sup_{D\geq d}\left\|\Phi_{d\to D}\hat{\theta}_{n,d}-\theta_D^\star\right\|\leq\varepsilon\right)\geq 1-\delta,$$

with a sample size $n=n(\varepsilon,\delta)$ that does not depend on d or D.

When each problem has a symmetry group G_d, the transfer maps should be compatible with the invariant structure. Informally, invariant parameters at size d should transfer to invariant parameters at size D:

$$\Phi_{d\to D}(\Theta_d^{G_d})\subseteq\Theta_D^{G_D}.$$

This compatibility is what makes dimension-stable estimation possible: if the effective dimension of $\Theta_d^{G_d}$ stabilizes or grows slowly with d, then the sample size required for estimation can be independent of the dimension.

E.3. Invariant polynomial parameters

A central class of examples in the paper arises when the parameter θ consists of coefficients of a polynomial. Let V_d be a finite-dimensional vector space on which G_d acts linearly (i.e., via invertible matrices). Let $\mathcal{P}_{d,k}$ denote the space of homogeneous degree k polynomial functions on V_d. The action of G_d on V_d induces an action on polynomials by

$$(g\cdot f)(x)\coloneqq f(g^{-1}\cdot x).$$

The invariant homogeneous polynomial space is

$$\mathcal{P}_{d,k}^{G_d}\coloneqq\{f\in\mathcal{P}_{d,k}:g\cdot f=f\text{ for all }g\in G_d\}.$$

When θ denotes the coefficient vector of a polynomial $f_\theta\in\mathcal{P}_{d,k}$, the invariant parameter space $\Theta_d^{G_d}$ is naturally identified with $\mathcal{P}_{d,k}^{G_d}$.

For finite groups, the dimension of the invariant subspace can be expressed using characters. If $\rho_{d,k}$ is the representation of G_d on $\mathcal{P}_{d,k}$ (i.e., the collection of matrices corresponding to the action) and $\chi_{d,k}(g)=\operatorname{tr}(\rho_{d,k}(g))$ is its character, then

$$\dim\mathcal{P}_{d,k}^{G_d}=\frac{1}{|G_d|}\sum_{g\in G_d}\chi_{d,k}(g).$$

Equivalently, if P_{G_d} denotes the group-averaging projection

$$P_{G_d}f=\frac{1}{|G_d|}\sum_{g\in G_d}g\cdot f,$$

then

$$\dim\mathcal{P}_{d,k}^{G_d}=\operatorname{tr}(P_{G_d}).$$

For compact groups, the finite average is replaced by Haar integration.

These formulas make explicit that invariant estimation reduces the number of effective parameters from $\dim\mathcal{P}_{d,k}$ to $\dim\mathcal{P}_{d,k}^{G_d}$. Representation stability refers to the phenomenon that, for natural sequences of groups and representations, these invariant dimensions stabilize as d grows. This stabilization is the algebraic mechanism behind dimension-uniform estimation in our setting.

E.4. Set-valued data and multisymmetric polynomials

Consider set-valued data with d elements in $\mathbb{R}^r$. We represent the set as

$$X = (x_1, \dots, x_d) \in (\mathbb{R}^r)^d, \qquad x_i = (x_{i,1}, \dots, x_{i,r}) \in \mathbb{R}^r.$$

The symmetric group S_d acts by permuting the d elements:

$$\pi \cdot X = (x_{\pi^{-1}(1)}, \dots, x_{\pi^{-1}(d)}).$$

The corresponding invariant polynomials are called multisymmetric polynomials: they are polynomials in the dr variables $x_{i,a}$ that are invariant under permutations of the block index i.

Let $\mathcal{P}_{d,r,k}^{S_d}$ denote the space of homogeneous degree k multisymmetric polynomials. A monomial is described by a collection of multi-indices

$$\alpha_i = (\alpha_{i,1}, \dots, \alpha_{i,r}) \in \mathbb{N}^r, \qquad 1 \le i \le d,$$

with total degree $\sum_i |\alpha_i| = k$. The group S_d permutes the multi-indices α_i. Therefore an invariant monomial orbit is determined by a multiset of nonzero multi-indices

$$\{\alpha^{(1)}, \dots, \alpha^{(q)}\}, \qquad \alpha^{(j)} \in \mathbb{N}^r \setminus \{0\}, \qquad \sum_{j=1}^{q} |\alpha^{(j)}| = k.$$

For $d \ge k$, every such multiset can be realized using at most k nonzero blocks, so the dimension stabilizes in d.

For each $m \ge 1$, the number of nonzero block types of degree m is

$$a_r(m) = \#\{\alpha \in \mathbb{N}^r : |\alpha| = m\} = \binom{r+m-1}{m}.$$

Hence the stable dimension of homogeneous degree k invariants is

$$N_r(k) \coloneqq \dim \mathcal{P}_{d,r,k}^{S_d} = [t^k] \prod_{m \ge 1} (1 - t^m)^{-a_r(m)} \qquad \text{for } d \ge k,$$

where $[t^k]F(t)$ denotes the coefficient of t^k in the formal power series $F(t)$. Equivalently,

$$N_r(k) = [t^k] \prod_{m \ge 1} (1 - t^m)^{-\binom{r+m-1}{m}}.$$

For $r = 1$, this reduces to the ordinary partition number:

$$N_1(k) = p(k).$$

Thus the invariant degree-k dimension is independent of d once $d \ge k$, whereas the full homogeneous degree-k polynomial space in dr variables has dimension

$$\binom{dr+k-1}{k},$$

which grows polynomially in d of order d^k for fixed k.

The asymptotic behavior of $N_r(k)$ for fixed r and large k is subexponential:

$$\log N_r(k) \sim C_r k^{\frac{r}{r+1}},$$

where

$$C_r = \frac{r+1}{r}\big(r\zeta(r+1)\big)^{\frac{1}{r+1}},$$

and $\zeta(s) \coloneqq \sum_{n=1}^{\infty} n^{-s}$ denotes the Riemann zeta function. Thus

$$N_r(k) = \exp\left(C_r k^{\frac{r}{r+1}} + o\left(k^{\frac{r}{r+1}}\right)\right).$$

For example, $r = 1$ recovers the Hardy–Ramanujan exponent

$$\log p(k) \sim \pi\sqrt{\frac{2k}{3}}.$$

For $r = 2$,

$$\log N_2(k) \sim \frac{3}{2}(2\zeta(3))^{1/3}k^{2/3},$$

and for $r = 3$,

$$\log N_3(k) \sim \frac{4}{3}(3\zeta(4))^{1/4}k^{3/4}.$$

These formulas show that for fixed element dimension r and polynomial degree k, the invariant polynomial parameter dimension stabilizes as the set size d grows. This is the representation-stability mechanism underlying the dimension-independent rates for set-valued data.

E.5. Hermite features for set-valued data

For data on $(\mathbb{R}^r)^d$, a natural basis is given by multivariate Hermite polynomials relative to a Gaussian reference measure. Let H_α denote an orthonormal Hermite polynomial indexed by $\alpha \in \mathbb{N}^{dr}$. For block-structured data, we may write

$$H_{\alpha_1,\dots,\alpha_d}(X) = \prod_{i=1}^{d} H_{\alpha_i}(x_i), \qquad \alpha_i \in \mathbb{N}^r.$$

The group S_d permutes the block indices α_i. Therefore invariant Hermite features are obtained by symmetrizing Hermite monomials over S_d. The dimension count for homogeneous degree k invariant Hermite features is the same as for multisymmetric polynomials:

$$\dim \mathcal{H}^{S_d}_{d,r,k} = N_r(k) \qquad \text{for } d \geq k.$$

This basis is useful for score matching. If p_d is a density relative to a Gaussian reference γ_d, write

$$\log \frac{p_d}{\gamma_d}(X) = \sum_j \theta^\star_j \phi_j(X),$$

where ϕ_j are invariant Hermite features. The score has the form

$$\nabla_X \log p_d(X) = -X + \sum_j \theta_j^\star \nabla_X \phi_j(X).$$

Gradients of invariant scalar features are equivariant vector fields:

$$\nabla_X \phi_j(\pi \cdot X) = \pi \cdot \nabla_X \phi_j(X).$$

Thus truncating the Hermite expansion at degree k gives a finite-dimensional equivariant score class whose dimension is $N_r(\leq k) = \sum_{q=1}^{k} N_r(q)$, independent of d for $d \geq k$.

If the parameter sequence has a controlled tail, e.g.

$$\sum_{j>m} (\theta_j^\star)^2 \leq b_m^2,$$

then truncating to m invariant Hermite features incurs coefficient bias at most b_m. In score norm, a weighted tail appears because taking gradients increases the effective degree and for Hermite features, the corresponding weight is proportional to the polynomial degree.

E.6. Attributed graphs

For attributed graphs, the data consist of node attributes and edge attributes:

$$H \in \mathbb{R}^{d \times r}, \qquad A \in \mathbb{R}^{d \times d \times r'}.$$

Here $H_{i,a}$ denotes the a-th attribute of node i, and $A_{ij,b}$ denotes the b-th attribute of the edge from node i to node j. The symmetric group S_d acts diagonally by relabeling nodes:

$$(\pi \cdot H)_{i,a} = H_{\pi^{-1}(i),a}, \qquad (\pi \cdot A)_{ij,b} = A_{\pi^{-1}(i),\pi^{-1}(j),b}.$$

For example, in the scalar edge case $r' = 1$, this action is

$$(A, H) \mapsto (P_\pi A P_\pi^\top, P_\pi H).$$

A homogeneous degree k polynomial in the entries of (A, H) is a linear combination of monomials

$$\prod_{i,a} H_{i,a}^{\alpha_{i,a}} \prod_{i,j,b} A_{ij,b}^{m_{ij,b}}, \qquad \sum_{i,a} \alpha_{i,a} + \sum_{i,j,b} m_{ij,b} = k.$$

Under the action of S_d, such a monomial is identified with all monomials obtained by relabeling node indices. Hence invariant homogeneous degree k polynomials are indexed by orbits of these monomials, equivalently by unlabeled attributed multigraph patterns of total degree k. In this interpretation, factors of $H_{i,a}$ are vertex decorations, while factors of $A_{ij,b}$ are edge decorations, both with multiplicity.

Let $\mathcal{R}_{d,r,r',k}^{S_d}$ denote the space of homogeneous degree k invariant polynomials in the entries of (A, H). Then, we have

$$\dim \mathcal{R}_{d,r,r',k}^{S_d} = \#\{\text{unlabeled attributed multigraph patterns of total degree } k \text{ realizable on } d \text{ nodes}\}.$$

Any degree k monomial can involve at most $2k$ distinct nodes through edge variables and at most k distinct nodes through node variables. Therefore this count stabilizes once $d \geq 2k$. Thus, for $d \geq 2k$,

$$\dim \mathcal{R}^{S_d}_{d,r,r',k} = M_{r,r'}(k),$$

where $M_{r,r'}(k)$ denotes the number of unlabeled attributed multigraph patterns of total degree k. The stable dimension depends on k, the node-attribute dimension r, the edge-attribute dimension r', and graph conventions, but not on the number of nodes d.

Equivalently, this dimension can be computed by character theory. Let

$$W_d \coloneqq \mathbb{R}^{d\times r} \oplus \mathbb{R}^{d\times d\times r'}$$

denote the vector space of attributed graph variables, with the diagonal S_d-action above. The homogeneous degree k polynomial space is $\mathrm{Sym}^k(W_d^*)$, and

$$\mathcal{R}^{S_d}_{d,r,r',k} = \left(\mathrm{Sym}^k(W_d^*)\right)^{S_d}.$$

Thus

$$\dim \mathcal{R}^{S_d}_{d,r,r',k} = \frac{1}{d!}\sum_{\pi\in S_d} \chi_{\mathrm{Sym}^k(W_d^*)}(\pi).$$

Equivalently, the generating function is given by the Molien series

$$\sum_{k\geq 0} \dim \mathcal{R}^{S_d}_{d,r,r',k} t^k = \frac{1}{d!}\sum_{\pi\in S_d} \frac{1}{\det(I - t\,\rho_d(\pi))},$$

where ρ_d is the representation of S_d on W_d^*.

The asymptotic dependence on k separates two regimes. If $r' = 0$, so that only node attributes are present, then the stable dimension reduces to the multisymmetric count

$$N_r(k) = [t^k] \prod_{m\geq 1} (1 - t^m)^{-\binom{r+m-1}{m}},$$

and therefore

$$N_r(k) = \exp\left(C_r k^{\frac{r}{r+1}} + o\left(k^{\frac{r}{r+1}}\right)\right).$$

In contrast, if $r' \geq 1$, edge variables generate graph-pattern complexity and

$$M_{r,r'}(k) = \exp(k\log k + O(k\log\log k)), \qquad k\to\infty.$$

Thus attributed graph invariant spaces remain stable in the number of nodes d, but their degree complexity is larger than that of set-valued data.

Appendix F. Additional Details on Sets and Graphs

In this appendix, we provide further details for the two canonical examples discussed in the main text: set-valued data and graph-valued data. These examples illustrate how the abstract invariance assumptions specialize to common variable-size data structures, and how both classical M-estimators and score matching fit naturally into the same framework.

Set-valued data. Consider a collection of d elements, each belonging to $\mathbb{R}^r$. We represent such an object as

$$X = (x_1, \dots, x_d) \in (\mathbb{R}^r)^d, \qquad x_i \in \mathbb{R}^r.$$

The ordering of the elements is arbitrary. The symmetric group S_d acts by permuting the elements:

$$\pi \cdot X = (x_{\pi^{-1}(1)}, \dots, x_{\pi^{-1}(d)}), \qquad \pi \in S_d.$$

A distribution μ_d on $(\mathbb{R}^r)^d$ is invariant under this action if $\pi \cdot X \sim \mu_d$ whenever $X \sim \mu_d$. This is precisely the finite-dimensional exchangeability assumption: the joint distribution of the represented sequence $x_1, \dots, x_d$ is unchanged by any permutation of the indices.

To illustrate the parameter action, consider the parameter space $\Theta_d = (\mathbb{R}^r)^d$, whose elements are $\theta = (\theta_1, \dots, \theta_d)$ with $\theta_i \in \mathbb{R}^r$. The group S_d acts on Θ_d by

$$\pi \cdot \theta = (\theta_{\pi^{-1}(1)}, \dots, \theta_{\pi^{-1}(d)}).$$

Consider the squared loss

$$\ell_d(X, \theta) = \frac{1}{d} \sum_{i=1}^{d} \|x_i - \theta_i\|_2^2.$$

This loss is compatible with the two actions:

$$\ell_d(\pi \cdot X, \pi \cdot \theta) = \ell_d(X, \theta), \qquad \forall \pi \in S_d.$$

By the general invariance argument in the main text, since the population minimizer is unique, it must belong to the invariant parameter subspace

$$\Theta_d^{S_d} = \{\theta \in (\mathbb{R}^r)^d : \pi \cdot \theta = \theta \text{ for all } \pi \in S_d\}.$$

This subspace consists exactly of parameters with identical components,

$$\theta_1 = \theta_2 = \cdots = \theta_d.$$

Thus, exchangeability forces the population minimizer to have the form $\theta^\star = (m_d, \dots, m_d)$, where

$$m_d = \mathbb{E}_{X \sim \mu_d} \left[\frac{1}{d} \sum_{i=1}^{d} x_i \right].$$

Thus, the usual set mean arises as the invariant representative of an equivariant estimation problem.

More general invariant estimators can be constructed from permutation-invariant features of X. Examples include symmetric polynomial features, invariant Hermite features, averages of single-element features, pairwise interactions, and higher-order symmetric interactions. Such features form natural finite-dimensional truncation classes for invariant estimation on sets.

The same symmetry is also natural for score matching. Suppose μ_d admits a density p_d with respect to a reference measure on $(\mathbb{R}^r)^d$, and suppose p_d is permutation invariant:

$$p_d(\pi \cdot X) = p_d(X), \qquad \pi \in S_d.$$

Then the score

$$s_d^\star(X) = \nabla_X \log p_d(X)$$

is permutation equivariant:

$$s_d^\star(\pi \cdot X) = \pi \cdot s_d^\star(X).$$

Therefore, when estimating the score, it is natural to restrict to permutation-equivariant vector fields. One way to construct such fields is to take gradients of permutation-invariant scalar features. In particular, gradients of invariant Hermite or polynomial features produce equivariant vector fields, giving a symmetry-compatible finite-dimensional class for score matching.

Graph-valued data. We next consider graph-valued data with edge attributes. Let

$$A \in \mathbb{R}^{d \times d \times r'}$$

denote the edge-attributed adjacency tensor of a graph on d nodes, where $A_{ij,b}$ is the b-th attribute of the edge from node i to node j. The case $r' = 1$ recovers ordinary weighted adjacency matrices. The symmetric group S_d acts by relabeling nodes:

$$(\pi \cdot A)_{ij,b} = A_{\pi^{-1}(i),\pi^{-1}(j),b}, \qquad \pi \in S_d.$$

Equivalently, when $r' = 1$, this action is $A \mapsto P_\pi A P_\pi^\top$, where P_π is the permutation matrix associated with π. A distribution μ_d over such graphs is invariant under node relabeling if

$$\pi \cdot A \sim \mu_d \quad \text{whenever} \quad A \sim \mu_d.$$

This formalizes the usual idea that graph distributions should not depend on arbitrary node labels.

Many graph-level statistics are invariant under this action. Examples include total edge attributes, triangle and subgraph counts with edge labels or weights, and spectral moments such as $\mathrm{tr}(A^k)$ in the scalar edge case $r' = 1$. Let $\varphi_d(A)$ be any such node-relabeling invariant statistic, and consider the target parameter

$$\theta_d^\star = \mathbb{E}_{A \sim \mu_d}[\varphi_d(A)].$$

Then the squared loss

$$\ell_d(A, \theta) = \|\varphi_d(A) - \theta\|_2^2$$

defines an M-estimator compatible with the graph permutation symmetry.

The same principle applies to score estimation on edge-attributed graphs. If μ_d admits a density p_d over $\mathbb{R}^{d \times d \times r'}$ and p_d is invariant under node relabeling, then

$$p_d(\pi \cdot A) = p_d(A).$$

Consequently, the score

$$s_d^\star(A) = \nabla_A \log p_d(A)$$

is equivariant under node relabeling:

$$s_d^\star(\pi \cdot A) = \pi \cdot s_d^\star(A),$$

where the action on the score tensor is given by

$$(\pi \cdot s_d^\star(A))_{ij,b} = s_d^\star(A)_{\pi^{-1}(i),\pi^{-1}(j),b}.$$

Thus, score matching for graph-valued data naturally leads to node-permutation equivariant vector fields on edge-attributed adjacency tensors.

Finally, the same formulation extends to graphs with both node and edge attributes. Let

$$H \in \mathbb{R}^{d\times r}, \qquad A \in \mathbb{R}^{d\times d\times r'}$$

denote node and edge attributes, respectively. The graph is represented as (A, H), and node relabeling acts diagonally:

$$(A, H) \mapsto (\pi \cdot A, \pi \cdot H),$$

where

$$(\pi \cdot H)_{i,a} = H_{\pi^{-1}(i),a}, \qquad (\pi \cdot A)_{ij,b} = A_{\pi^{-1}(i),\pi^{-1}(j),b}.$$

If μ_d admits an invariant density $p_d(A, H)$, the score decomposes as

$$s^\star_d(A, H) = \big(s^\star_{A,d}(A, H), s^\star_{H,d}(A, H)\big),$$

where

$$s^\star_{A,d}(A, H) = \nabla_A \log p_d(A, H), \qquad s^\star_{H,d}(A, H) = \nabla_H \log p_d(A, H).$$

These components transform equivariantly:

$$s^\star_d(\pi \cdot A, \pi \cdot H) = \big(\pi \cdot s^\star_{A,d}(A, H),\ \pi \cdot s^\star_{H,d}(A, H)\big),$$

where we have

$$(\pi \cdot s^\star_{A,d}(A, H))_{ij,b} = s^\star_{A,d}(A, H)_{\pi^{-1}(i),\pi^{-1}(j),b}, \qquad (\pi \cdot s^\star_{H,d}(A, H))_{i,a} = s^\star_{H,d}(A, H)_{\pi^{-1}(i),a}.$$

Invariant estimators depend only on the isomorphism class of the attributed graph, while score estimators are equivariant under this diagonal action. This provides a unified symmetry model for sets, graphs, and attributed graphs.

Appendix G. Proof of Theorem 3

We prove the invariant M-estimation guarantee stated in Theorem 3. Under the strong convexity assumption on the loss, the proof follows: empirical optimality and strong convexity reduce the parameter error to the size of the empirical gradient at the population target, and the latter is controlled by sub-Gaussian concentration on the invariant parameter space.

G.1. Restricted gradients on the invariant model

Fix $d \in \mathbb{N}$. Let $\Theta_d^{G_d}$ be the invariant parameter space and suppose $\theta_d^\star \in \Theta_d^{G_d}$. We write

$$p_d^G = \dim(\Theta_d^{G_d}).$$

For simplicity of notation, we identify $\Theta_d^{G_d}$ locally with a Euclidean space of dimension p_d^G. Equivalently, if $\Theta_d^{G_d}$ is a smooth model, the gradient below is understood in any normalized local coordinate chart around $\theta_d^\star$. Thus, $\nabla_G \hat{L}_{n,d}(\theta_d^\star)$ denotes the gradient of the empirical risk restricted to the invariant model.

The assumption that $\ell_d(x,\cdot)$ is α-strongly convex on $\Theta_d^{G_d}$ for every $x \in \mathcal{X}_d$ implies that the empirical risk

$$\hat{L}_{n,d}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_d(X_i,\theta)$$

is also α-strongly convex on $\Theta_d^{G_d}$. Indeed, an average of α-strongly convex functions is again α-strongly convex.

We will use the following standard concentration bound for the restricted empirical gradient.

Lemma 6 (Restricted gradient concentration) *Suppose the restricted empirical gradient at $\theta_d^\star$ is sub-Gaussian with variance proxy at most σ^2. That is, writing*

$$Y_i := \nabla_G \ell_d(X_i,\theta_d^\star) \in \mathbb{R}^{p_d^G},$$

we have $\mathbb{E}Y_i = 0$ and, for every unit vector $v \in \mathbb{R}^{p_d^G}$,

$$\mathbb{E}\exp(\lambda\langle v, Y_i\rangle) \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \qquad \forall \lambda \in \mathbb{R}.$$

Then there exists an absolute constant $C_0 > 0$ such that, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left\|\nabla_G \hat{L}_{n,d}(\theta_d^\star)\right\|_2^2 = \left\|\frac{1}{n}\sum_{i=1}^{n} Y_i\right\|_2^2 \le C_0\sigma^2\frac{p_d^G + \log(1/\delta)}{n}.$$

Proof Let $p = p_d^G$ and define

$$\bar{Y}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Since the Y_i's are independent, mean-zero, and σ^2-sub-Gaussian, the average $\bar{Y}_n$ is sub-Gaussian with variance proxy σ^2/n. Hence for every unit vector $v \in \mathbb{R}^p$,

$$\mathbb{E}\exp(\lambda\langle v, \bar{Y}_n\rangle) \le \exp\left(\frac{\lambda^2\sigma^2}{2n}\right).$$

Let $\mathcal{N}$ be a $1/2$-net of the unit sphere $\mathbb{S}^{p-1}$ with $|\mathcal{N}| \le 5^p$. For every $u \in \mathbb{R}^p$,

$$\|u\|_2 \le 2\max_{v\in\mathcal{N}}\langle v, u\rangle.$$

Therefore, for any $t > 0$,

$$\Pr(\|\bar{Y}_n\|_2 \ge 2t) \le \Pr\left(\max_{v\in\mathcal{N}}\langle v, \bar{Y}_n\rangle \ge t\right).$$

By a union bound and the one-dimensional sub-Gaussian tail bound,

$$\Pr\left(\max_{v\in\mathcal{N}}\langle v, \bar{Y}_n\rangle \ge t\right) \le |\mathcal{N}|\exp\left(-\frac{nt^2}{2\sigma^2}\right) \le \exp\left(p\log 5 - \frac{nt^2}{2\sigma^2}\right).$$

Taking

$$t = \sigma\sqrt{\frac{2(p\log 5 + \log(1/\delta))}{n}},$$

we obtain, with probability at least $1-\delta$,

$$\|\bar{Y}_n\|_2 \le 2\sigma\sqrt{\frac{2(p\log 5 + \log(1/\delta))}{n}}.$$

Squaring both sides yields

$$\|\bar{Y}_n\|_2^2 \le 8\sigma^2\frac{p\log 5 + \log(1/\delta)}{n}.$$

Thus the claim holds with an absolute constant C_0. ∎

G.2. Proof of the estimation bound

Let

$$\hat{\theta}^G_{n,d} \in \arg\min_{\theta\in\Theta_d^{G_d}} \hat{L}_{n,d}(\theta).$$

Since $\theta^\star_d \in \Theta_d^{G_d}$, empirical optimality gives

$$\hat{L}_{n,d}(\hat{\theta}^G_{n,d}) \le \hat{L}_{n,d}(\theta^\star_d).$$

Because $\hat{L}_{n,d}$ is α-strongly convex on $\Theta_d^{G_d}$, we have

$$\hat{L}_{n,d}(\hat{\theta}^G_{n,d}) \ge \hat{L}_{n,d}(\theta^\star_d) + \left\langle \nabla_G\hat{L}_{n,d}(\theta^\star_d), \hat{\theta}^G_{n,d} - \theta^\star_d\right\rangle + \frac{\alpha}{2}\left\|\hat{\theta}^G_{n,d} - \theta^\star_d\right\|_2^2.$$

Combining the previous two displays gives

$$0 \ge \left\langle \nabla_G\hat{L}_{n,d}(\theta^\star_d), \hat{\theta}^G_{n,d} - \theta^\star_d\right\rangle + \frac{\alpha}{2}\left\|\hat{\theta}^G_{n,d} - \theta^\star_d\right\|_2^2.$$

By Cauchy–Schwarz,

$$0 \ge -\left\|\nabla_G\hat{L}_{n,d}(\theta^\star_d)\right\|_2\left\|\hat{\theta}^G_{n,d} - \theta^\star_d\right\|_2 + \frac{\alpha}{2}\left\|\hat{\theta}^G_{n,d} - \theta^\star_d\right\|_2^2.$$

If $\hat{\theta}^G_{n,d} = \theta^\star_d$, the desired bound is immediate. Otherwise, dividing by $\|\hat{\theta}^G_{n,d} - \theta^\star_d\|_2$ gives

$$\left\|\hat{\theta}^G_{n,d} - \theta^\star_d\right\|_2 \le \frac{2}{\alpha}\left\|\nabla_G\hat{L}_{n,d}(\theta^\star_d)\right\|_2.$$

Squaring and applying Lemma 6, we obtain that with probability at least $1-\delta$,

$$\left\|\hat{\theta}^G_{n,d} - \theta^\star_d\right\|_2^2 \le \frac{4}{\alpha^2}\left\|\nabla_G\hat{L}_{n,d}(\theta^\star_d)\right\|_2^2 \le C\frac{\sigma^2}{\alpha^2}\frac{p^G_d + \log(1/\delta)}{n},$$

for an absolute constant $C > 0$. This proves the first claim of Theorem 3.

If

$$p^G := \sup_d p^G_d < \infty,$$

then $p^G_d \le p^G$ for every d, and therefore

$$\left\|\hat{\theta}^G_{n,d} - \theta^\star_d\right\|_2^2 \le C\frac{\sigma^2}{\alpha^2}\frac{p^G + \log(1/\delta)}{n}.$$

G.3. Proof of the transfer bound

Fix a source dimension d. Suppose that for every $D \geq d$,

$$\Phi_{d\to D}\theta_d^\star = \theta_D^\star,$$

and that $\Phi_{d\to D}$ is distance preserving on the invariant model:

$$\left\|\Phi_{d\to D}\theta - \Phi_{d\to D}\theta'\right\|_2 = \left\|\theta - \theta'\right\|_2 \qquad \forall \theta, \theta' \in \Theta_d^{G_d}.$$

Then, for every $D \geq d$,

$$\begin{aligned}\left\|\Phi_{d\to D}\hat{\theta}_{n,d}^G - \theta_D^\star\right\|_2 &= \left\|\Phi_{d\to D}\hat{\theta}_{n,d}^G - \Phi_{d\to D}\theta_d^\star\right\|_2 \\ &= \left\|\hat{\theta}_{n,d}^G - \theta_d^\star\right\|_2.\end{aligned}$$

Taking the supremum over $D \geq d$, we get

$$\sup_{D\geq d}\left\|\Phi_{d\to D}\hat{\theta}_{n,d}^G - \theta_D^\star\right\|_2^2 = \left\|\hat{\theta}_{n,d}^G - \theta_d^\star\right\|_2^2.$$

Combining this identity with the estimation bound above yields

$$\sup_{D\geq d}\left\|\Phi_{d\to D}\hat{\theta}_{n,d}^G - \theta_D^\star\right\|_2^2 \leq C\frac{\sigma^2}{\alpha^2}\frac{p_d^G + \log(1/\delta)}{n}$$

with probability at least $1-\delta$. If $p^G = \sup_d p_d^G < \infty$, the right-hand side is bounded by

$$C\frac{\sigma^2}{\alpha^2}\frac{p^G + \log(1/\delta)}{n}.$$

This proves the transfer statement and completes the proof of Theorem 3.

Appendix H. Proof of Theorem 4

We prove the score matching result for set-valued data. The proof consists of a bias-variance decomposition in coefficient space and an optimization over the degree cutoff.

H.1. Coefficient representation and source norms

Let $\{\psi_{d,j}\}_{j\geq 1}$ be the ordered basis of permutation-equivariant score features used in the main text. We write the target score as

$$s_d^\star = \sum_{j\geq 1}\theta_{d,j}^\star\psi_{d,j}, \qquad \theta_d^\star = (\theta_{d,1}^\star, \theta_{d,2}^\star, \ldots) \in \ell_2(\mathbb{N}).$$

For an estimator

$$\hat{s}_d = \sum_{j\geq 1}\hat{\theta}_{d,j}\psi_{d,j},$$

we measure error in the coefficient norm

$$\|\hat{s}_d - s_d^\star\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} = \sum_{j\geq 1}(\hat{\theta}_{d,j} - \theta^\star_{d,j})^2.$$

For a degree cutoff k, let $\mathcal{I}_{d,k}$ denote the set of invariant score features of degree at most k, and let

$$m_r(k) = |\mathcal{I}_{d,k}|.$$

For $d \geq k$, this dimension is independent of d, and satisfies

$$m_r(k) = \exp\left(C_r k^{\frac{r}{r+1}} + o\left(k^{\frac{r}{r+1}}\right)\right).$$

Let $\theta^\star_{d,k}$ be the truncation of $\theta^\star_d$ to the coordinates in $\mathcal{I}_{d,k}$:

$$(\theta^\star_{d,k})_j = \begin{cases} \theta^\star_{d,j}, & j \in \mathcal{I}_{d,k}, \\ 0, & j \notin \mathcal{I}_{d,k}. \end{cases}$$

The truncation bias is

$$B_d(k) \coloneqq \|\theta^\star_d - \theta^\star_{d,k}\|^2_{\ell_2} = \sum_{j\notin\mathcal{I}_{d,k}} (\theta^\star_{d,j})^2.$$

The exponential source norm is

$$\|s_d^\star\|^2_{\mathsf{Exp}(a,\nu)} = \sup_{k\geq 1} e^{ak^\nu} B_d(k),$$

so that

$$B_d(k) \leq \|s_d^\star\|^2_{\mathsf{Exp}(a,\nu)} e^{-ak^\nu}.$$

Similarly, the algebraic source norm is

$$\|s_d^\star\|^2_{\mathsf{Alg}(\tau)} = \sup_{k\geq 1} k^\tau B_d(k),$$

so that

$$B_d(k) \leq \|s_d^\star\|^2_{\mathsf{Alg}(\tau)} k^{-\tau}.$$

H.2. Bias-variance decomposition

Let $\hat{s}^{(k)}_{n,d}$ be the score matching estimator restricted to $\mathcal{I}_{d,k}$, and write its coefficient vector as $\hat{\theta}^{(k)}_{n,d}$, extended by zero outside $\mathcal{I}_{d,k}$. Since $\hat{\theta}^{(k)}_{n,d} - \theta^\star_{d,k}$ is supported on $\mathcal{I}_{d,k}$, while $\theta^\star_d - \theta^\star_{d,k}$ is supported on the complement, the two components are orthogonal in ℓ_2. Therefore,

$$\|\hat{s}^{(k)}_{n,d} - s_d^\star\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} = \|\hat{\theta}^{(k)}_{n,d} - \theta^\star_{d,k}\|^2_{\ell_2} + B_d(k).$$

The finite-dimensional score matching problem over $\mathcal{I}_{d,k}$ is an invariant M-estimation problem with parameter dimension $m_r(k)$. By Theorem 3, with probability at least $1-\delta$,

$$\|\hat{\theta}^{(k)}_{n,d} - \theta^\star_{d,k}\|^2_{\ell_2} \leq C\frac{\sigma^2}{\alpha^2}\frac{m_r(k) + \log(1/\delta)}{n}.$$

Thus

$$\|\hat{s}^{(k)}_{n,d} - s^\star_d\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} \le B_d(k) + C\frac{\sigma^2}{\alpha^2}\frac{m_r(k) + \log(1/\delta)}{n}.$$

Taking the infimum over k gives

$$\inf_{k\ge 1} \|\hat{s}^{(k)}_{n,d} - s^\star_d\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} \le \inf_{k\ge 1}\left\{B_d(k) + C\frac{\sigma^2}{\alpha^2}\frac{m_r(k) + \log(1/\delta)}{n}\right\}.$$

H.3. Exponential source condition

Assume

$$\sup_d \|s^\star_d\|_{\mathsf{Exp}(a,\nu)} < \infty.$$

Then, uniformly over d,

$$B_d(k) \le R^2 e^{-ak^\nu}$$

for some $R < \infty$. Also, for every fixed $\epsilon > 0$ and all sufficiently large k,

$$m_r(k) \le \exp\left((C_r + \epsilon)k^{\frac{r}{r+1}}\right).$$

Thus, up to constants and lower-order confidence terms, it suffices to optimize

$$R_n(k) = e^{-ak^\nu} + \frac{1}{n}\exp\left(bk^{\frac{r}{r+1}}\right),$$

where $b > C_r$ is fixed. The optimal cutoff balances the two exponents:

$$ak^\nu + bk^{\frac{r}{r+1}} \asymp \log n.$$

Case $\nu > \frac{r}{r+1}$. Here ak^ν dominates $bk^{r/(r+1)}$, so the balancing cutoff satisfies

$$k_n \asymp \left(\frac{\log n}{a}\right)^{1/\nu}.$$

Then

$$bk_n^{\frac{r}{r+1}} = O\left((\log n)^{\frac{r}{\nu(r+1)}}\right) = o(\log n).$$

Therefore,

$$R_n(k_n) = n^{-1+o(1)}.$$

Case $\nu = \frac{r}{r+1}$. The two exponents have the same order. The balancing equation becomes

$$(a+b)k^{\frac{r}{r+1}} \asymp \log n.$$

Thus,

$$k_n \asymp \left(\frac{\log n}{a+b}\right)^{\frac{r+1}{r}}.$$

The optimized rate is polynomial:

$$R_n(k_n) = n^{-c}$$

for constant $c = a/(a+b) \in (0,1)$.

Case $\nu < \frac{r}{r+1}$. Here $bk^{r/(r+1)}$ dominates the variance exponent. The balancing cutoff satisfies

$$k_n \asymp \left(\frac{\log n}{b}\right)^{\frac{r+1}{r}}.$$

Then the bias term gives

$$e^{-ak_n^\nu} = \exp\left(-c(\log n)^{\frac{\nu(r+1)}{r}}\right)$$

for some $c > 0$. Therefore,

$$R_n(k_n) = \exp\left(-c(\log n)^{\frac{\nu(r+1)}{r}}\right).$$

This proves the three exponential-source regimes.

H.4. Algebraic source condition

Assume

$$\sup_d \|s_d^\star\|_{\mathsf{Alg}(\tau)} < \infty.$$

Then

$$B_d(k) \le R^2 k^{-\tau}$$

uniformly over d. We optimize

$$R_n(k) = k^{-\tau} + \frac{1}{n}\exp\left(bk^{\frac{r}{r+1}}\right).$$

Choose

$$k_n \asymp \left(\frac{\log n}{b}\right)^{\frac{r+1}{r}}.$$

Then

$$k_n^{-\tau} \asymp (\log n)^{-\frac{\tau(r+1)}{r}}.$$

Hence,

$$\inf_k R_n(k) \le (\log n)^{-\frac{\tau(r+1)}{r}}.$$

This proves the algebraic-source rate and completes the proof of Theorem 4.

Appendix I. Proof of Theorem 5

We prove the graph score matching result. The proof parallels the set-valued case, with the graph invariant dimension $M_{r,r'}(k)$ replacing the set invariant dimension $N_r(k)$.

I.1. Coefficient representation and graph source norms

Let $\{\psi_{d,j}\}_{j\geq 1}$ be the ordered basis of node-permutation equivariant graph score features. We write the target graph score as

$$s_d^\star = \sum_{j\geq 1} \theta_{d,j}^\star \psi_{d,j}, \qquad \theta_d^\star = (\theta_{d,1}^\star, \theta_{d,2}^\star, \ldots) \in \ell_2(\mathbb{N}).$$

For an estimator

$$\hat{s}_d = \sum_{j\geq 1} \hat{\theta}_{d,j} \psi_{d,j},$$

we measure error in the coefficient norm

$$\|\hat{s}_d - s_d^\star\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} = \sum_{j\geq 1} (\hat{\theta}_{d,j} - \theta_{d,j}^\star)^2.$$

For a degree cutoff k, let $\mathcal{J}_{d,k}$ be the set of equivariant graph score features of degree at most k, and let

$$M_{r,r'}(k) = |\mathcal{J}_{d,k}|.$$

For $d \geq 2k$, this dimension is independent of d, and for $r' \geq 1$,

$$M_{r,r'}(k) = \exp(k \log k + O(k \log\log k)).$$

Let $\theta_{d,k}^\star$ be the truncation of $\theta_d^\star$ to the coordinates in $\mathcal{J}_{d,k}$:

$$(\theta_{d,k}^\star)_j = \begin{cases} \theta_{d,j}^\star, & j \in \mathcal{J}_{d,k}, \\ 0, & j \notin \mathcal{J}_{d,k}. \end{cases}$$

The truncation bias is

$$B_d^{\mathrm{graph}}(k) \coloneqq \|\theta_d^\star - \theta_{d,k}^\star\|_{\ell_2}^2 = \sum_{j \notin \mathcal{J}_{d,k}} (\theta_{d,j}^\star)^2.$$

The graph exponential source norm is

$$\|s_d^\star\|^2_{\mathsf{Exp}(a,\nu)} = \sup_{k\geq 1} e^{ak^\nu} B_d^{\mathrm{graph}}(k),$$

so that

$$B_d^{\mathrm{graph}}(k) \leq \|s_d^\star\|^2_{\mathsf{Exp}(a,\nu)} e^{-ak^\nu}.$$

Similarly, the graph algebraic source norm is

$$\|s_d^\star\|^2_{\mathsf{Alg}(\tau)} = \sup_{k\geq 1} k^\tau B_d^{\mathrm{graph}}(k),$$

so that

$$B_d^{\mathrm{graph}}(k) \leq \|s_d^\star\|^2_{\mathsf{Alg}(\tau)} k^{-\tau}.$$

I.2. Bias-variance decomposition

Let $\hat{s}^{(k)}_{n,d}$ be the score matching estimator restricted to $\mathcal{J}_{d,k}$, and write its coefficient vector as $\hat{\theta}^{(k)}_{n,d}$, extended by zero outside $\mathcal{J}_{d,k}$. Since $\hat{\theta}^{(k)}_{n,d} - \theta^\star_{d,k}$ is supported on $\mathcal{J}_{d,k}$, while $\theta^\star_d - \theta^\star_{d,k}$ is supported on the complement, the two components are orthogonal in ℓ_2. Therefore,

$$\|\hat{s}^{(k)}_{n,d} - s^\star_d\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} = \|\hat{\theta}^{(k)}_{n,d} - \theta^\star_{d,k}\|^2_{\ell_2} + B^{\mathrm{graph}}_d(k).$$

The finite-dimensional graph score matching problem over $\mathcal{J}_{d,k}$ is an invariant M-estimation problem with parameter dimension $M_{r,r'}(k)$. By Theorem 3, with probability at least $1-\delta$,

$$\|\hat{\theta}^{(k)}_{n,d} - \theta^\star_{d,k}\|^2_{\ell_2} \le C\frac{\sigma^2}{\alpha^2}\frac{M_{r,r'}(k) + \log(1/\delta)}{n}.$$

Thus

$$\|\hat{s}^{(k)}_{n,d} - s^\star_d\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} \le B^{\mathrm{graph}}_d(k) + C\frac{\sigma^2}{\alpha^2}\frac{M_{r,r'}(k) + \log(1/\delta)}{n}.$$

This proves the bias-variance bound.

I.3. Rates under exponential graph source conditions

Assume

$$\sup_d \|s^\star_d\|_{\mathsf{Exp}(a,\nu)} < \infty.$$

Then, uniformly over d,

$$B^{\mathrm{graph}}_d(k) \le R^2 e^{-ak^\nu}$$

for some $R < \infty$. Since

$$M_{r,r'}(k) = \exp(k\log k + O(k\log\log k)),$$

for every fixed $\epsilon > 0$ and all sufficiently large k,

$$M_{r,r'}(k) \le \exp\left((1+\epsilon)k\log k\right).$$

Up to constants and lower-order confidence terms, it is enough to optimize

$$R_n(k) = e^{-ak^\nu} + \frac{1}{n}\exp\left((1+\epsilon)k\log k\right).$$

Case $\nu > 1$. The bias exponent ak^ν grows faster than $k\log k$. Choose

$$k_n \asymp \left(\frac{\log n}{a}\right)^{1/\nu}.$$

Then

$$k_n \log k_n = o(\log n),$$

and hence

$$\frac{1}{n}\exp\left((1+\epsilon)k_n\log k_n\right) = n^{-1+o(1)}.$$

The bias term is also $n^{-1+o(1)}$. Therefore the optimized rate is

$$n^{-1+o(1)}.$$

Case $\nu = 1$. The variance exponent $k \log k$ dominates the linear bias exponent ak. Choose

$$k_n \asymp \frac{\log n}{\log \log n}.$$

Then

$$k_n \log k_n \asymp \log n,$$

so the variance term is controlled, and the bias term gives

$$e^{-ak_n} = \exp\left(-a \frac{\log n}{\log \log n}\right).$$

Thus the optimized rate is bounded by

$$\exp\left(-c_1 \frac{\log n}{\log \log n}\right)$$

for some $c_1 > 0$.

Case $0 < \nu < 1$. Again choose

$$k_n \asymp \frac{\log n}{\log \log n}.$$

The variance term is controlled as above, while the bias term becomes

$$e^{-ak_n^\nu} = \exp\left(-a \left(\frac{\log n}{\log \log n}\right)^\nu\right).$$

Therefore, the optimized rate is

$$\exp\left(-c_2 \left(\frac{\log n}{\log \log n}\right)^\nu\right)$$

for some $c_2 > 0$. This proves the exponential-source regimes.

I.4. Rates under algebraic graph source conditions

Assume

$$\sup_d \|s_d^\star\|_{\mathsf{Alg}(\tau)} < \infty.$$

Then

$$B_d^{\mathrm{graph}}(k) \le R^2 k^{-\tau}$$

uniformly over d. We optimize

$$R_n(k) = k^{-\tau} + \frac{1}{n} \exp\left(k \log k\right).$$

Choose

$$k_n \asymp \frac{\log n}{\log \log n}.$$

Then

$$k_n^{-\tau} = \left(\frac{\log n}{\log\log n}\right)^{-\tau}.$$

Therefore,

$$\inf_k R_n(k) \le \left(\frac{\log n}{\log\log n}\right)^{-\tau}.$$

This proves the algebraic-source rate and completes the proof of Theorem 5.

Appendix J. Experiments

J.1. Subgraph counting

We include a small synthetic experiment illustrating the role of invariant low-degree structure in transfer across graph sizes. The purpose of the experiment is not to benchmark a new architecture, but to visualize the basic mechanism behind our theory: a low-degree invariant statistic has a stable meaning across graph sizes, while a non-invariant coordinate polynomial does not transfer canonically.

We generate graphs from an Erdős–Rényi model with a graph-dependent edge probability. For each sample, we draw

$$p \sim \mathrm{Unif}(0.15, 0.85), \qquad A \sim G(d, p),$$

where $A \in \{0,1\}^{d\times d}$ is the adjacency matrix of an undirected graph with no self-loops. The response is the population triangle density $y \approx p^3$. Thus, the target is a low-degree graph moment, but the learner observes only the sampled graph A, not the latent parameter p. We train all models only on graphs with $d = 10$ nodes and evaluate them on larger graphs with

$$D \in \{25, 30, 35, 40, 50, 60, 75, 100, 125, 150, 175, 200\}.$$

We compare two predictors. The first is an invariant polynomial model using normalized graph statistics

$$\phi_{\mathrm{inv},d}(A) = \big(E_d(A), W_d(A), T_d(A)\big),$$

where $E_d(A)$ is the edge density, $W_d(A)$ is the normalized wedge density, and $T_d(A)$ is the normalized triangle density. A linear model is fit on these features at $d = 10$, and transfer to size D is performed by evaluating the same normalized invariant statistics on the larger graph:

$$\widehat{f}_{\mathrm{inv},D}(A) = \widehat{\beta}_0 + \widehat{\beta}_1 E_D(A) + \widehat{\beta}_2 W_D(A) + \widehat{\beta}_3 T_D(A).$$

This gives a direct size-transfer rule because the features are invariant graph moments whose interpretation is stable across graph sizes.

The second predictor is a non-invariant degree-three polynomial baseline. At the training size $d = 10$, we use the raw triangle indicators

$$A_{ij}A_{ik}A_{jk}, \qquad 1 \le i < j < k \le 10,$$

as features and fit a linear model. Such a model assigns separate coefficients to labeled triangles and therefore has no canonical extension to graphs with more than ten nodes. To give this baseline a

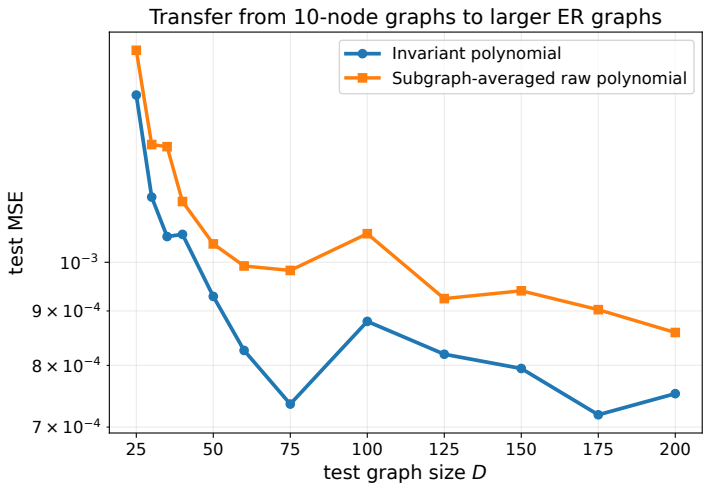

Figure 2: Transfer from 10-node graphs to larger Erdős–Rényi graphs. Both methods are trained only at size $d = 10$. The invariant polynomial model transfers by evaluating the same normalized graph moments at the larger size. The raw polynomial model is trained on labeled triangle indicators and is made transferable by averaging over random 10-node induced subgraphs. The results illustrate that stable size transfer comes from invariant low-degree structure; the raw baseline improves only after an explicit symmetrization step.

favorable transfer rule, we average it over random induced subgraphs: for a test graph on D nodes, we sample subsets $S_1, \dots, S_M \subseteq [D]$ of size ten and define

$$\widehat{f}^{\mathrm{avg}}_{\mathrm{raw},D}(A) = \frac{1}{M} \sum_{m=1}^{M} \widehat{f}_{\mathrm{raw},10}(A_{S_m \times S_m}).$$

This is a Monte Carlo symmetrization of the raw non-invariant predictor.

Figure 2 shows the test mean-squared error as the test graph size increases. The invariant polynomial model transfers directly and improves with graph size, since the normalized graph moments are computed using all available nodes and become more accurate estimates of the underlying population quantities. The subgraph-averaged raw model also benefits from averaging, but this improvement comes from an explicit symmetrization step at test time. The experiment therefore illustrates the main message of our theory: dimension transfer is enabled not merely by low degree, but by low-degree features whose invariant meaning is stable across sizes.

J.2. Score transfer on weighted graphs

We next include a small synthetic experiment illustrating the score-estimation mechanism behind the theory. The goal is to isolate the effect of invariant spectral complexity in a setting where the true score is known analytically. We consider continuous weighted graphs represented by symmetric matrices

$$A \in \mathbb{R}^{d \times d},$$

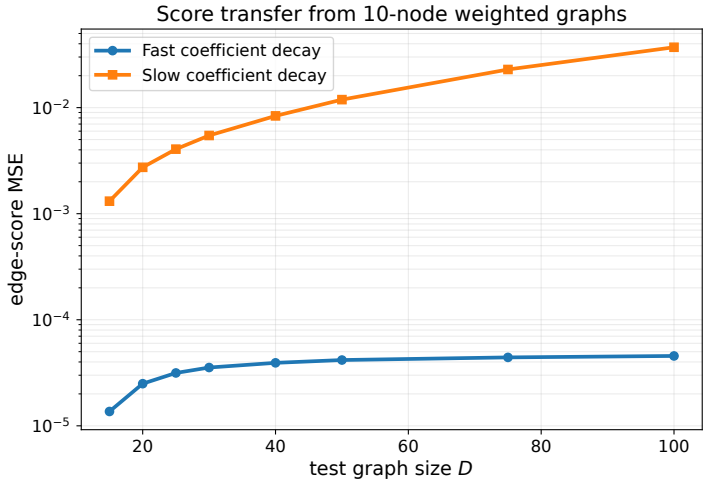

Figure 3: Toy score transfer on weighted graphs. A degree-three equivariant score model is trained on 10-node graphs and evaluated on larger graphs. Fast coefficient decay yields accurate transfer, while slow decay leaves a larger high-degree tail and leads to higher error.

with zero diagonal. The entries A_{ij}, $i < j$, are sampled independently from a standard Gaussian distribution. We define graph distributions through invariant potentials of the form

$$p_d(A) \propto \exp(-U_d(A))\gamma_d(A),$$

where γ_d is the standard Gaussian reference measure on the edge variables. The corresponding score is

$$s_d^\star(A) = \nabla_A \log p_d(A) = -A - \nabla_A U_d(A).$$

Since the potentials U_d used below are invariant under node relabeling, the score $s_d^\star$ is permutation-equivariant.

We use invariant spectral potentials

$$U_d(A) = \sum_{q=2}^{K} b_q d^{-q} \operatorname{tr}(A^q),$$

for which the score is available in closed form:

$$s_d^\star(A) = -A - \sum_{q=2}^{K} b_q q d^{-q} A^{q-1}.$$

This construction lets us control the invariant complexity of the score through the decay of the coefficients b_q. We compare two regimes. In the first regime, the coefficients decay rapidly,

$$b_q = 2\exp(-0.45q^2),$$

so the score is well approximated by low-degree spectral components. In the second regime, the coefficients decay slowly,

$$b_q = 2q^{-1/2},$$

so higher-degree components remain important.

For both regimes, we train a linear equivariant score model using only graphs with $d = 10$ nodes. The model has the form

$$\widehat{s}_d(A) = \widehat{\theta}_0 \mathbf{1} + \widehat{\theta}_1 A + \widehat{\theta}_2 A^2 + \widehat{\theta}_3 A^3,$$

where the diagonal entries are ignored and the loss is the mean squared error over off-diagonal edge scores. We then transfer the same learned coefficients to larger graphs by evaluating the same equivariant features on graphs of size

$$D \in \{15, 20, 25, 30, 40, 50, 75, 100\}.$$

The test error is

$$\frac{1}{\binom{D}{2}} \sum_{i<j} \left(\widehat{s}_D(A)_{ij} - s^\star_D(A)_{ij}\right)^2,$$

averaged over test graphs.

Figure 3 shows the resulting score-estimation error. The fast-decay regime transfers accurately across graph sizes, since the score is well captured by low-degree equivariant spectral features. In contrast, slow decay leaves a larger high-degree tail, leading to higher error. This supports the main message that score transfer is governed by low-complexity invariant approximability, not by graph size alone.