Which Physical Symmetries Can Be Imposed by Averaging?

Raunav Mendiratta
Westview High School
San Diego, CA, USA

Aiden Seungchan Lee
Westview High School
San Diego, CA, USA

Palaash Gang
Indus International School Pune
Pune, India

Abstract

Averaging a trained model over a compact symmetry group makes it invariant, and when the data distribution shares the symmetry the averaged model has no larger risk than the original under any convex loss. Many symmetries of physical systems, among them translations, Lorentz boosts and changes of units, form non-compact groups, for which the average has to be restricted to a bounded set of group elements, usually a ball around the identity. In this paper we examine which of the two properties survive this truncation. For connected Lie groups acting freely on the inputs, we show that ball averages become invariant for every bounded model if and only if the volume of balls grows polynomially. Under exponential growth, as for the Lorentz group, they do not, and the average of a model is instead governed by its values at the cutoff. The guarantee on the risk is lost for a different reason, since no data distribution is invariant under a non-compact group that moves inputs arbitrarily far, as translations and boosts do, and we illustrate this with an exact counterexample and with a top-quark tagger trained on a standard benchmark. Although neither result depends on amenability, amenability determines whether some sequence of sets other than balls yields invariant averages, and we construct such sets for the affine and similarity groups.

1 Introduction

Averaging a model over a symmetry group is a simple way to make it invariant or equivariant. One transforms the input by each group element, applies the model, maps the output back and takes the mean. For a compact group the average is a projection onto the equivariant models [12, 47], it applies to a trained model without the retraining that an equivariant architecture [9, 25] would need, and if the data distribution shares the symmetry it does not increase the risk under any convex loss [8, 12, 28]. Foit et al. [13] applied it at test time to surrogate models trained on The Well [32] and reported lower evaluation loss at no training cost. Data augmentation rests on the same average [8], so the question of which groups admit a useful average bears on augmentation as well.

Many symmetries used in physics, such as translations of an unbounded domain, Lorentz boosts of particle-physics data and changes of origin and units of dimensioned quantities, form non-compact groups, which Foit et al. list as out of scope. A non-compact group has no uniform distribution [46], so the average has to be restricted to a bounded set of group elements, usually a ball of radius H around the identity, and divided by its volume. In this paper we study whether the truncated average becomes invariant as H grows, and whether it can increase the risk. The first answer depends on how fast the volume of balls grows and the second on whether the group moves inputs arbitrarily far, and neither depends on amenability, the property usually associated with averaging over infinite groups.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Representations for the Physical Sciences Workshop.

Our contributions. In Section 3 we show that, for a connected Lie group acting freely, ball averages become invariant for every bounded model if and only if the group has polynomial volume growth, and that under exponential growth the worst-case change of the averaged prediction under a fixed group element tends to a positive limit, which we compute in closed form for groups acting on hyperbolic space. On the Lorentz group, the ball average of a model that settles at large rapidity tends instead to an average of the model over rotations at the cutoff. Amenability, which decides whether some sequence of sets gives invariant averages, does not decide whether balls do, and for the affine and similarity groups, which are amenable but whose balls do not give invariant averages, we construct such sets explicitly. In Section 4 we show with an exact counterexample that averaging over a non-compact group that moves inputs arbitrarily far can increase the risk, whatever the growth of the group, and in Section 5 we measure this effect on a trained top-quark tagger. Section 6 gives a recommendation for each class of groups.

Related work. Under polynomial growth balls are Følner sets [7, 42], and Milnes [30] showed that they are not in the affine group, which is amenable. Følner averages appear in statistics [1] and learning [33], and amenability plays a similar role in the Hunt–Stein theory of invariant procedures [5]. Villar et al. [44] and Spinner et al. [39] noted that no invariant data distribution exists for scalings and for boosts, respectively, and Lawrence and Harris [26] showed that augmentation over a non-compact group shifts the data distribution. Building on these results, we separate the two questions for averaging a trained model and answer both for the main non-compact groups of physics, listed in Table 1. Appendix F discusses frames, canonicalization and other related work.

2 Preliminaries

We assume familiarity with group actions, and refer the reader to Weil [46] for Haar measure, to Greenleaf [19] for amenability and to Helgason [21] for the structure of the Lorentz group. We define the main objects here and introduce others where they are used.

Groups and models. A group G acts on inputs by $x \mapsto gx$. A model f is *invariant* if $f(gx) = f(x)$ for all g and x, and *equivariant* if $f(gx) = \rho(g)f(x)$ for a representation ρ of G on the outputs. We state results for invariance, and they extend to equivariance through the factor $\rho(g)$ in (1).

Haar measure and averages. The *Haar measure* μ of G assigns a volume $\mu(E)$ to each set E of group elements and does not change when the set is shifted, $\mu(gE) = \mu(E)$. Its total volume is finite if and only if G is compact [46], and then we normalize it to the uniform distribution on G. For a set F of group elements with finite positive volume, the *average* of f over F is

$$A_F f(x) = \frac{1}{\mu(F)} \int_F \rho(g)\, f(g^{-1}x)\, d\mu(g), \tag{1}$$

the mean prediction over the inputs $g^{-1}x$ with g in F. For compact G the average over the whole group, $A_G f$, also known as the Reynolds operator, is invariant. We use the form $f(g^{-1}x)$ with left Haar measure, for which the results hold for all the groups we consider (Appendix B).

Truncation. For non-compact G we measure distances between group elements with a left-invariant metric (Appendix A) and write B_H for the ball of radius H, the *cutoff*, around the identity. The *truncated average* $A_H = A_{B_H}$ is the mean of $f(g^{-1}x)$ over g drawn from μ_H, the uniform distribution on B_H. For translations of the line, $B_H = [-H, H]$ and $A_H f(x)$ is the mean of f over the interval $[x - H, x + H]$.

Equivariance defect. The defect of a model φ at $h \in G$, $\mathrm{def}_h(\varphi) = \sup_x |\varphi(hx) - \varphi(x)|$, is the largest change of its prediction when the input is moved by h. A model is invariant if and only if its defect vanishes for every h, so the defect of $A_H f$ measures how far it is from invariant.

Følner sets and amenability. A sequence of sets $F_n \subset G$ of finite positive volume is a *Følner sequence* if $\mu(hF_n \triangle F_n)/\mu(F_n) \to 0$ for every h, where $\triangle$ is the symmetric difference, so that shifting such a set by a fixed element changes a vanishing fraction of it. A group is *amenable* if it has a Følner sequence [19]. Compact, abelian and solvable groups are amenable, and the Lorentz group is not.

Volume growth. G has *polynomial growth* if the volume $\mu(B_H)$ of balls grows like a power of H, as for translations and rigid motions. It has *exponential growth* if $\mu(B_H)$ grows like e^{pH} up to

Table 1: Groups from the classes that Foit et al. [13] leave out of scope, with the compact rotation group for comparison. Volume growth decides whether balls are Følner and hence whether ball averages become invariant (Proposition 1), while amenability decides whether any Følner sequence exists. Only the compact group keeps the risk guarantee, and only if the data are invariant (Section 4). Appendix A gives the degrees and rates of growth. *Of an unbounded domain, since translations of a periodic domain form a compact group.

group	symmetry	volume growth	amenable	balls Følner	average over
$\mathrm{SO}(3)$	rotations	compact	yes	–	whole group
$\mathbb{R}^d$	translations*	polynomial	yes	yes	balls
$\mathrm{SE}(3)$	rigid motions	polynomial	yes	yes	balls
Heis_3	phase-space shifts	polynomial	yes	yes	balls
$\mathbb{R}^k_{>0}$	changes of units	polynomial	yes	yes	balls
$\mathrm{Aff}(1)$	changes of origin and unit	exponential	yes	no	sets F_n
$\mathrm{SIM}(2)$	… and of orientation	exponential	yes	no	sets F_n
$\mathrm{SO}(1,3)^+$	Lorentz transformations	exponential	no	no	rotations only
$\mathrm{SO}(1,4)^+$	de Sitter symmetries	exponential	no	no	rotations only

polynomial factors, as for the Lorentz group, which acts on the hyperbolic space of four-velocities. A connected Lie group has one or the other [20, 22], and which one does not depend on the metric.

Risk and the risk guarantee. Data (x, y) follow a distribution P, and the risk of a model is its expected loss, $\mathcal{R}(f) = \mathbb{E}_P\, \ell(f(x), y)$, for a loss ℓ that is convex in the prediction, such as the squared error or the cross-entropy of a predicted probability. P is *G-invariant* if (gx, y) has the same distribution as (x, y) for every g, so that a transformed data point is as likely as the original and has the same label. The *risk guarantee* is the statement that for compact G and G-invariant P, $\mathcal{R}(A_G f) \le \mathcal{R}(f)$ for every model f [8, 12, 28], so that under these conditions a trained model can be averaged without checking its risk again.

3 Invariance of truncated averages

The first question, whether $A_H f$ becomes invariant as the cutoff grows, reduces by Proposition 1 to a property of the balls alone.

Proposition 1 (Følner sets give invariant averages)**.** *(a) For every set F of finite positive volume, every bounded model f and every $h \in G$, $\mathrm{def}_h(A_F f) \le \|f\|_\infty\, \mu(h^{-1}F \triangle F)/\mu(F)$. Equality holds for some f with $\|f\|_\infty \le 1$ when the action is free, that is, when no input is fixed by an element other than the identity. (b) For a connected Lie group, the balls B_H form a Følner sequence if and only if G has polynomial growth.*

Part (a) says that, for a free action, the worst-case defect of an average over F, the largest defect over models with values in $[-1, 1]$, is the volume in which F and its shift differ, relative to the volume of F. Part (b) combines the dichotomy between the two types of growth with results of Breuillard [7] and Tessera [42] on the volume of balls (proofs in Appendix A). Together, the two parts say that, for a free action, ball averages become invariant for every bounded model if and only if growth is polynomial. Under exponential growth the worst-case defect has a positive limit, which can be computed when the group acts on a hyperbolic space.

Proposition 2 (Defect under exponential growth)**.** *Let G be $\mathrm{Aff}(1)$, $\mathrm{SIM}(2)$ or $\mathrm{SO}(1,n)^+$, each acting on a hyperbolic space, with B_H the elements that move a base point by at most H, and let h move the base point by a distance $t > 0$. Then the worst-case defect $\mu(h^{-1}B_H \triangle B_H)/\mu(B_H)$ converges, as $H \to \infty$, to a positive limit given in Appendix A. For the Lorentz group $\mathrm{SO}(1,3)^+$ and for $\mathrm{SIM}(2)$, the limit is $2\tanh(t/2)$.*

By Proposition 2, enlarging the ball does not make the average invariant, because most of the volume of a large ball lies near its boundary, so that a ball and its shift by a fixed element keep differing in a fixed fraction of their volume (Figure 1a).

Amenability does not decide whether balls are Følner, as two groups from Table 1 show. The affine group $\mathrm{Aff}(1)$ of maps $x \mapsto e^s x + b$, which change the origin and unit of a quantity, and the similarity

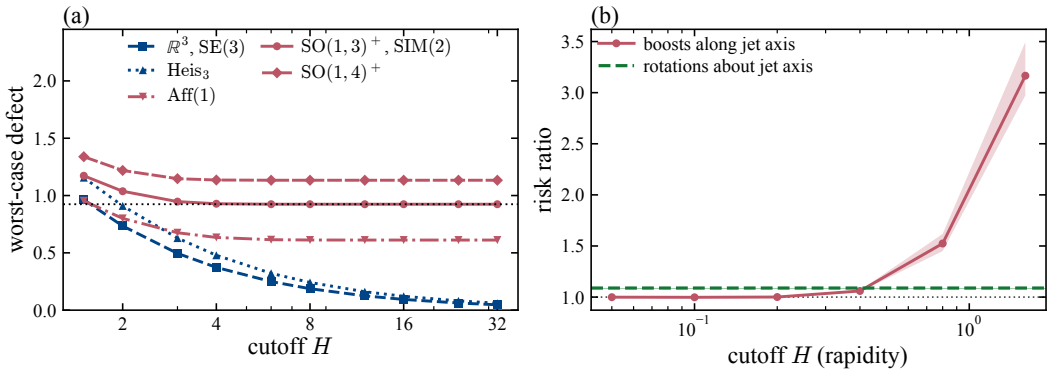


Figure 1: **Truncated averages on non-compact groups.** (a) Worst-case defect of the ball average for an element h that moves the base point by unit distance, $\mu(h^{-1}B_H \triangle B_H)/\mu(B_H)$. It tends to zero under polynomial growth (blue) and not under exponential growth (red), with the Lorentz limit $2\tanh(1/2)$ dotted. (b) Ratio $\mathcal{R}(A_H f)/\mathcal{R}(f)$ of the risk of the averaged top-quark tagger to that of the tagger, for boosts along the jet axis against the cutoff and for rotations about the jet axis (dashed). Lines are means over three independently trained models, and the shaded band spans them.

group $\mathrm{SIM}(2)$ of the plane are solvable, hence amenable, yet their balls are not Følner, as Milnes showed for the affine group [30]. Being amenable, they have other Følner sequences, and these can be written down explicitly.

Proposition 3 (Følner sets for the affine and similarity groups)**.** *In the coordinates (s, b) of $\mathrm{Aff}(1)$, the sets $F_n = \{(s, b) : 0 \le s \le n,\ 0 \le b \le e^s\}$ form a Følner sequence, with $\mu(hF_n \triangle F_n)/\mu(F_n) = O(1/n)$ for every h. For $\mathrm{SIM}(2)$, the sets with $0 \le s \le n$ and translation part $|v| \le e^s$ do the same.*

The sets F_n allow translations over a range that widens at the rate at which the Haar measure thins out, so a shifted set differs from F_n in a set of bounded volume while the volume of F_n grows like n (Appendix B). By Proposition 1(a), averages over F_n become invariant for every bounded model.

The Lorentz group. Although the balls of the Lorentz group are not Følner, the ball average of a particular model can still converge [18, 31]. Each Lorentz transformation is a rotation, followed by a boost of rapidity η along a fixed axis and another rotation, and B_H consists of the transformations with $\eta \le H$, most of whose volume lies at rapidities close to H. We say that a model *settles* if its prediction approaches a limit as the boost applied to the input grows. For such a model the ball average therefore approaches the average of the model over rotations applied before and after a boost of rapidity H (Proposition 5 in Appendix C, proved for rank-one groups such as the Lorentz group). This limit is invariant under rotations for every cutoff, but under boosts at most as $H \to \infty$, and then only under further conditions on the model (Appendix C). Its value is then determined by the model at infinite rapidity, far from any data, so ball averages are not a practical way to make a model boost-invariant.

4 Risk of truncated averages

The risk guarantee rests on two hypotheses, that the averaging distribution is a probability distribution and that the data distribution is invariant, and truncation supplies only the first.

Lemma 1. *Let ℓ be convex in its first argument and let ν be a probability distribution on G. Then the average $A_\nu f(x) = \mathbb{E}_{g\sim\nu} f(g^{-1}x)$ satisfies $\mathcal{R}(A_\nu f) \le \mathbb{E}_{g\sim\nu}\mathcal{R}_g(f)$, where $\mathcal{R}_g(f) = \mathbb{E}_P\, \ell(f(g^{-1}x), y)$ is the risk of the model applied to transformed inputs.*

Lemma 1 applies to the truncated average with $\nu = \mu_H$. If P is G-invariant, every transformed model has the same risk as f, and the risk guarantee follows. A non-compact group that moves inputs arbitrarily far, however, admits no invariant data distribution. To see this, call an action *proper* if group elements far from the identity move every bounded set of inputs off itself. Under a proper action, a bounded set holding more than half of the probability could be moved off itself,

and its image would hold as much again, so the total probability would exceed one (Appendix D). Translations of points and Lorentz transformations of massive four-momenta, or of pairs of them, are proper actions. Actions that are not proper, such as translations of a stationary random field, can have invariant distributions, and then the risk guarantee holds for truncated averages too. The failure does not depend on growth, so it already appears for translations of the line.

Proposition 4 (A counterexample for translations)**.** *Let $\mathbb{R}$ act on $\mathbb{R}$ by translation, with target $y \equiv 0$, squared loss, model $f(x) = \mathbf{1}\{|x| \geq 1\}$ and data $x \sim \mathcal{N}(0, \varepsilon^2)$. Then $\mathcal{R}(f) = 2\Phi(-1/\varepsilon)$, where Φ is the standard normal distribution function, while $A_H f(x) = 1 - 1/H$ whenever $|x| \leq H - 1$, so that $\mathcal{R}(A_H f) \geq (1 - 1/H)^2 \, \mathbb{P}(|x| \leq H - 1)$.*

As H grows, the risk of the averaged model tends to one, the largest risk of any model with values in $[0, 1]$, for every data scale ε, while the risk of f is below one and tends to zero as $\varepsilon \to 0$ (Appendix D). The average replaces the prediction at each input by its mean over a long stretch of the line, most of which carries no data.

5 Averaging a trained jet tagger

We now test whether the risk also increases for a trained model on a public benchmark. We use ParticleNet [36], a standard tagger that separates top-quark jets from jets of light quarks and gluons, trained with its authors' code on the reference top-tagging dataset [24] (Appendix E). We average its predicted probability over two subgroups of the Lorentz group, both acting on all particles of a jet. Rotations about the jet axis form a compact group, while boosts along the jet axis form a one-parameter group whose balls, the boosts up to rapidity H, are Følner, so the boost average becomes invariant as H grows. Neither transformation changes the label, but the jets in the dataset are selected in a window of transverse momentum, and these boosts move them out of it.

Figure 1(b) shows the ratio of the risk of each average to the risk of the tagger. The boost average barely changes the risk for small cutoffs and raises it steeply for the larger cutoffs that invariance requires. The rotation average also raises the risk, by less than the boost average at large cutoffs. By Lemma 1 this can only happen if the jet distribution is not rotation-invariant either. This is expected, since the beam axis singles out directions transverse to the jet, and the Lorentz-equivariant taggers evaluated on this dataset take the beam direction as an additional input [17, 39].

6 Discussion

Of the four classes of groups in Table 1, the compact groups are the simplest, since the average can be taken over the whole group and the risk guarantee holds when the data are invariant. This is the setting of Foit et al. [13], whose groups are compact (dihedral rotations and reflections, and translations of periodic domains), and the one dataset on which they find no gain has initial conditions that break the symmetry, as our results would lead one to expect. Under polynomial growth the ball average becomes invariant, but nothing guarantees a lower risk, so the risk should be checked on held-out data and the cutoff reported. The same holds for the affine and similarity groups, with the sets F_n in place of balls. For the Lorentz group no Følner sequence exists, so for a free action no average over sets of group elements becomes invariant for every bounded model (Proposition 1a). Averaging over rotations keeps the guarantee when the data are rotation-invariant, while boost invariance is better imposed by invariant architectures [4, 17, 39] or by frames and canonicalization [27, 40], which may be discontinuous [11].

In neural solvers for partial differential equations, non-compact symmetries such as Galilean boosts and scalings are used as augmentation with small ranges tuned on validation data [6, 29], or imposed by canonicalization [38]. Our results support this practice, since without the risk guarantee no range is safe a priori and the useful range depends on how far the data are from invariant. The same holds for the bounded non-compact transformations sampled in self-supervised pipelines for jets [10].

Limitations. The tagger experiment covers one architecture and one dataset, and the defect results are worst cases over bounded models. Proposition 5 covers rank-one groups such as $\mathrm{SO}(1, n)^+$, while the ball average on higher-rank or non-semisimple groups such as $GL(d)$ remains to be characterized.

References

[1] Morgane Austern and Peter Orbanz. Limit theorems for distributions invariant under groups of transformations. *The Annals of Statistics*, 50(4):1960–1991, 2022.

[2] Armin Beck and Peter Ochs. Symmetries in PAC-Bayesian learning. In *International Conference on Machine Learning*, 2026. arXiv:2510.17303.

[3] Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. In *Advances in Neural Information Processing Systems*, pages 18673–18684, 2021.

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

[5] James V. Bondar and Paul Milnes. Amenability: A survey for statistical applications of Hunt–Stein and related conditions on groups. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(1): 103–128, 1981.

[6] Johannes Brandstetter, Max Welling, and Daniel E. Worrall. Lie point symmetry data augmentation for neural PDE solvers. In *International Conference on Machine Learning*, 2022.

[7] Emmanuel Breuillard. Geometry of locally compact groups of polynomial growth and shape of large balls. *Groups, Geometry, and Dynamics*, 8(3):669–732, 2014.

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

[9] Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, 2016.

[10] Barry M. Dillon, Gregor Kasieczka, Hans Olischläger, Tilman Plehn, Peter Sorrenson, and Lorenz Vogel. Symmetries, safety, and self-supervision. *SciPost Physics*, 12:188, 2022.

[11] Nadav Dym, Hannah Lawrence, and Jonathan W. Siegel. Equivariant frames and the impossibility of continuous canonicalization. In *International Conference on Machine Learning*, 2024.

[12] Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *International Conference on Machine Learning*, pages 2959–2969, 2021.

[13] Valentino F. Foit, David W. Hogg, and Soledad Villar. Group averaging for physics applications: Accuracy improvements at zero training cost. In *Machine Learning and the Physical Sciences Workshop, NeurIPS*, 2025. arXiv:2511.09573.

[14] Jan Gerken, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Equivariance versus augmentation for spherical images. In *International Conference on Machine Learning*, 2022.

[15] Madison Giacofci, Anouar Meynaoui, and Alex Podgorny. A kernel two-sample test invariant under group action with applications to functional data. *arXiv preprint arXiv:2603.16294*, 2026.

[16] Andrés Gomberoff and Donald Marolf. On group averaging for SO(n,1). *International Journal of Modern Physics D*, 8:519–535, 1999.

[17] Shiqi Gong, Qi Meng, Jue Zhang, Huilin Qu, Congqiao Li, Sitian Qian, Weitao Du, Zhi-Ming Ma, and Tie-Yan Liu. An efficient Lorentz equivariant graph neural network for jet tagging. *Journal of High Energy Physics*, 2022(7):30, 2022.

[18] Alexander Gorodnik and Amos Nevo. *The Ergodic Theory of Lattice Subgroups*, volume 172 of *Annals of Mathematics Studies*. Princeton University Press, 2010.

[19] Frederick P. Greenleaf. *Invariant Means on Topological Groups and Their Applications*. Van Nostrand Reinhold, 1969.

[20] Yves Guivarc'h. Croissance polynomiale et périodes des fonctions harmoniques. *Bulletin de la Société Mathématique de France*, 101:333–379, 1973.

[21] Sigurdur Helgason. *Groups and Geometric Analysis*. Academic Press, 1984.

[22] Joe W. Jenkins. Growth of connected locally compact groups. *Journal of Functional Analysis*, 12(1): 113–127, 1973.

[23] Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*, 2023.

[24] Gregor Kasieczka, Tilman Plehn, et al. The machine learning landscape of top taggers. *SciPost Physics*, 7:014, 2019.

[25] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, 2018.

[26] Hannah Lawrence and Mitchell Tong Harris. Learning polynomial problems with $\mathrm{SL}(2, \mathbb{R})$-equivariance. In *International Conference on Learning Representations*, 2024.

[27] Yuchao Lin, Jacob Helwig, Shurui Gui, and Shuiwang Ji. Equivariance via minimal frame averaging for more symmetries and efficiency. In *International Conference on Machine Learning*, pages 30042–30079, 2024.

[28] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020.

[29] Grégoire Mialon, Quentin Garrido, Hannah Lawrence, Danyal Rehman, Yann LeCun, and Bobak T. Kiani. Self-supervised learning with Lie symmetries for partial differential equations. In *Advances in Neural Information Processing Systems*, 2023.

[30] Paul Milnes. Counterexample to a conjecture of Greenleaf. *Canadian Mathematical Bulletin*, 13(4): 497–499, 1970.

[31] Amos Nevo. Pointwise ergodic theorems for actions of groups. In *Handbook of Dynamical Systems*, volume 1B, pages 871–982. Elsevier, 2006.

[32] Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Keaton Burns, Stuart B. Dalziel, Drummond B. Fielding, et al. The Well: a large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

[33] Peter Orbanz. Global optimality under amenable symmetry constraints. *arXiv preprint arXiv:2402.07613*, 2024.

[34] Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Advances in Neural Information Processing Systems*, pages 61936–61959, 2023.

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

[36] Huilin Qu and Loukas Gouskos. ParticleNet: Jet tagging via particle clouds. *Physical Review D*, 101: 056019, 2020.

[37] Huilin Qu, Congqiao Li, and Sitian Qian. Particle transformer for jet tagging. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18281–18292, 2022.

[38] 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 *International Conference on Learning Representations*, 2025.

[39] Jonas Spinner, Victor Bresó, Pim de Haan, Tilman Plehn, Jesse Thaler, and Johann Brehmer. Lorentz-equivariant geometric algebra transformers for high-energy physics. In *Advances in Neural Information Processing Systems*, 2024.

[40] Jonas Spinner, Luigi Favaro, Peter Lippmann, Sebastian Pitz, Gerrit Gerhartz, Tilman Plehn, and Fred A. Hamprecht. Lorentz local canonicalization: How to make any network Lorentz-equivariant. In *Advances in Neural Information Processing Systems*, pages 28779–28807, 2025.

[41] Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. In *Advances in Neural Information Processing Systems*, pages 55616–55646, 2023.

[42] Romain Tessera. Volume of spheres in doubling metric measured spaces and in groups of polynomial growth. *Bulletin de la Société Mathématique de France*, 135(1):47–64, 2007.

[43] Soledad Villar, David W. Hogg, Kate Storey-Fisher, Weichi Yao, and Ben Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics. In *Advances in Neural Information Processing Systems*, 2021.

[44] Soledad Villar, Weichi Yao, David W. Hogg, Ben Blum-Smith, and Bianca Dumitrascu. Dimensionless machine learning: Imposing exact units equivariance. *Journal of Machine Learning Research*, 24(109): 1–32, 2023.

[45] Soledad Villar, David W. Hogg, Weichi Yao, George A. Kevrekidis, and Bernhard Schölkopf. Towards fully covariant machine learning. *Transactions on Machine Learning Research*, 2024.

[46] André Weil. *L'intégration dans les groupes topologiques et ses applications*. Hermann, 1940.

[47] Dmitry Yarotsky. Universal approximations of invariant maps by neural networks. *Constructive Approximation*, 55(1):407–474, 2022.

A Volume growth and the defect

Metrics and growth. Balls are taken in the geodesic radial coordinate of the non-compact part of G. Compact factors, such as the rotations in $\mathrm{SE}(3)$ and $\mathrm{SIM}(2)$, contribute bounded volume and change no limit. Writing $J(r)$ for the radial Haar density, so that $\mu(B_H) = \int_0^H J$, the groups of Table 1 have the following densities and growth.

group	radial density $J(r)$	growth
$\mathbb{R}^d$; $\mathbb{R}^k_{>0}$ in logarithmic coordinates	$\propto r^{d-1}$; $\propto r^{k-1}$	polynomial, degree d; k
$\mathrm{SE}(3)$	$4\pi r^2$	polynomial, degree 3
Heis_3, Korányi gauge	$4r^3$	polynomial, degree 4
$\mathrm{Aff}(1)$	$2\pi \sinh r$	exponential, rate 1
$\mathrm{SIM}(2)$	$4\pi \sinh^2 r$	exponential, rate 2
$\mathrm{SO}(1,3)^+$	$\sinh^2 r$	exponential, rate 2
$\mathrm{SO}(1,4)^+$	$\sinh^3 r$	exponential, rate 3

For $\mathrm{Aff}(1)$, $\mathrm{SIM}(2)$ and $\mathrm{SO}(1,n)^+$ the metric is normalized so that the associated hyperbolic space $\mathbb{H}^m$ has curvature -1, which fixes the rate $p = m - 1$. Another normalization rescales p, while the type of growth and the degree do not depend on the metric. The density $\sinh^{n-1}\eta$ of $\mathrm{SO}(1,n)^+$ is the Cartan form of its Haar measure [21, Ch. I, §5] and agrees with the volume form induced on the mass shell. The Heisenberg group has degree 4 by the Bass–Guivarc'h formula [7, 20], and in $\mathrm{SIM}(n)$ the dilation scales all n coordinates, so $p = n$.

Volume near the boundary. If $\mu(B_H) \sim Ce^{pH}$, the share of the volume of B_H within unit distance of its boundary, $1 - \mu(B_{H-1})/\mu(B_H)$, tends to $1 - e^{-p}$, while under polynomial growth it tends to zero. This is why, under exponential growth, a ball and its shift keep differing in a fixed fraction of their volume (Proposition 2).

Proof of Proposition 1. (a) Substituting $g \mapsto hg$ in (1) and using left invariance of μ gives $A_F f(hx) = \mu(F)^{-1}\int_{h^{-1}F} f(g^{-1}x)\,d\mu(g)$, so $A_F f(hx) - A_F f(x) = \mu(F)^{-1}\big(\int_{h^{-1}F} - \int_F\big) f(g^{-1}x)\,d\mu(g)$, whose absolute value is at most $\|f\|_\infty \mu(h^{-1}F \triangle F)/\mu(F)$ because $\mu(h^{-1}F) = \mu(F)$. If G acts freely, the map $g \mapsto g^{-1}x_0$ is injective for any fixed x_0, so f can be chosen equal to $+1$ at the points $g^{-1}x_0$ with $g \in h^{-1}F \setminus F$ and -1 at those with $g \in F \setminus h^{-1}F$, which attains the bound.

(b) If G has polynomial growth, $\mu(B_H) \sim cH^d$ for left-invariant Riemannian metrics [7, Thm. 1.1]. Let r be the distance from the identity to h. Then $hB_H \subset B_{H+r}$, so $\mu(hB_H \triangle B_H) = 2\mu(hB_H \setminus B_H) \le 2(\mu(B_{H+r}) - \mu(B_H))$, and the ratio to $\mu(B_H)$ tends to zero (see also [42] and [7, Cor. 1.10]). Conversely, suppose the balls are Følner and fix $0 < \delta < 1$. Cover B_1 by finitely many sets $h_i B_\delta$. Since the metric is geodesic, $B_{H+1} = B_1 B_H \subset \bigcup_i h_i B_{H+\delta}$, so $\mu(B_{H+1}) \le \mu(B_{H+\delta}) + \sum_i \mu(h_i B_{H+\delta} \setminus B_{H+\delta}) = \mu(B_{H+\delta})(1 + o(1))$. Hence $\mu(B_{H+c})/\mu(B_H) \to 1$ for every $c > 0$, so

$\log \mu(B_H)$ grows sublinearly in H, the growth is not exponential, and by the dichotomy of Guivarc'h and Jenkins [20, 22] it is polynomial. □

Proof of Proposition 2. We show that the limit is $\int_{S^{m-1}} |1 - P_t(\xi)|\, d\sigma(\xi)$, where $m = 2$, 3 and n for $\mathrm{Aff}(1)$, $\mathrm{SIM}(2)$ and $\mathrm{SO}(1,n)^+$, σ is the uniform distribution on the sphere of directions and P_t is the Poisson kernel of $\mathbb{H}^m$ for a point at distance t. Each of these groups acts transitively on $\mathbb{H}^m$ with compact stabilizer, B_H is the preimage of the hyperbolic ball of radius H about o, and $h^{-1}B_H$ is the preimage of the ball about $o' = h^{-1}o$, at distance t from o. The ratio in the proposition is therefore the same ratio for hyperbolic balls. In polar coordinates (r, ξ) about o, a point far from o is at distance $r + \beta_t(\xi) + o(1)$ from o', where $\beta_t(\xi) = \log(\cosh t - \sinh t \cos\theta)$ is the Busemann function and θ is the angle between ξ and the direction of o'. The point lies in the first ball and not in the second when $H - \beta_t(\xi) < r \le H$. Since the volume element is $\sinh^{m-1} r\, dr\, d\xi \approx 2^{1-m} e^{(m-1)r}\, dr\, d\xi$, the share of $\mu(B_H)$ in this set tends to $\int (1 - e^{-(m-1)\beta_t})_+\, d\sigma$, and $e^{-(m-1)\beta_t(\xi)} = (\cosh t - \sinh t\cos\theta)^{1-m} = P_t(\xi)$ is the Poisson kernel. The symmetric difference has twice this volume, and $2\int (1 - P_t)_+\, d\sigma = \int |1 - P_t|\, d\sigma$ because $\int P_t\, d\sigma = 1$. For $m = 3$, $\cos\theta$ is uniform on $[-1, 1]$ and the integral is $\int_1^{e^t} (1 - u^{-2})\, du / \sinh t = 2\tanh(t/2)$. For $m = 2$ the same computation gives $(4/\pi)\,\mathrm{gd}(t/2)$, with gd the Gudermannian function. □

Numerical values. Table 2 gives the worst-case defect at unit displacement, computed by exact quadrature in polar coordinates for the hyperbolic balls, in closed form for Euclidean balls, and by Monte Carlo over the Korányi ball for Heis_3, with h a unit horizontal translation. For $\mathrm{Aff}(1)$ the limit is $(4/\pi)\,\mathrm{gd}(1/2) = 0.612$, and for $\mathrm{SIM}(2)$ and $\mathrm{SO}(1,3)^+$ it is $2\tanh(1/2) = 0.924$. The maximum possible value is 2.

Table 2: Worst-case equivariance defect $\mu(h^{-1}B_H \triangle B_H)/\mu(B_H)$ for a group element h that moves the base point by unit distance. These are the curves of Figure 1(a).

group	H=4	H=8	H=16	H=40
$\mathbb{R}$	0.250	0.125	0.062	0.025
$\mathbb{R}^3$, $\mathrm{SE}(3)$	0.373	0.187	0.094	0.037
Heis_3	0.478	0.242	0.121	0.048
$\mathrm{Aff}(1)$	0.634	0.612	0.612	0.612
$\mathrm{SIM}(2)$, $\mathrm{SO}(1,3)^+$	0.928	0.924	0.924	0.924
$\mathrm{SO}(1,4)^+$	1.135	1.133	1.133	1.133

B Følner sets for $\mathrm{Aff}(1)$ and $\mathrm{SIM}(2)$

Write $\mathrm{Aff}(1) = \{(a, b) : a > 0\}$ acting by $x \mapsto ax + b$ and put $a = e^s$. Left Haar measure is $e^{-s}\, ds\, db$, and left translation by (s_0, b_0) acts as $(s, b) \mapsto (s + s_0,\ e^{s_0} b + b_0)$. For F_n as in Proposition 3, $\mu(F_n) = \int_0^n e^{-s} e^s\, ds = n$. The translate is $hF_n = \{(s, b) : s_0 \le s \le n + s_0,\ b_0 \le b \le e^s + b_0\}$. On the two end slabs of width $|s_0|$ the symmetric difference has measure $|s_0|$ each. On the overlap the two b-intervals differ by $2\min(|b_0|, e^s)$, contributing at most $2|b_0| \int_{\max(0,s_0)}^n e^{-s}\, ds \le 2|b_0|$. Hence $\mu(hF_n \triangle F_n) \le 2|s_0| + 2|b_0|$, which is bounded in n, and the ratio is $O(1/n)$.

For $\mathrm{SIM}(2) = \mathbb{R}^2 \rtimes (\mathbb{R}_{>0} \times \mathrm{SO}(2))$ with dilation e^s, left Haar measure is $e^{-2s}\, dv\, ds\, dr$, and left translation by (v_0, s_0, r_0) sends $(v, s, r) \mapsto (v_0 + e^{s_0} r_0 v,\ s + s_0,\ r_0 r)$. With $F_n = \{(v, s, r) : 0 \le s \le n,\ |v| \le e^s\}$ each slab is a disc, which is invariant under r_0, so the translate is the disc of the same radius centred at v_0. The slab symmetric difference is $2(\pi R^2 - \mathrm{lens}(R, |v_0|))$ with $R = e^s$. Since $\pi R^2 - \mathrm{lens}(R, d) = R^2[\pi - 2\arccos\varrho + 2\varrho\sqrt{1 - \varrho^2}]$ with $\varrho = d/2R$, the weight e^{-2s} cancels R^2, and since the bracket is at most $4\varrho = 2|v_0|e^{-s}$, the overlap contributes at most $4|v_0|$. The end slabs contribute $\pi|s_0|$ each, and with $\mu(F_n) = \pi n$ the ratio is again $O(1/n)$.

Følner averages over amenable groups were used in a learning context by Orbanz [33], and general constructions of Følner sequences for amenable Lie groups are surveyed by Bondar and Milnes [5]. The sets above are closed-form instances for the two groups of Table 1 whose balls are not Følner.

Two forms of the average. The bounds above concern left translates hF_n, which control the average (1) through Proposition 1(a). The other common form, $\mu(F)^{-1}\int_F f(gx)\,d\mu(g)$ with left Haar measure, is controlled instead by right translates F_nh, which need not shrink in a group that is not unimodular. Table 3 compares the two forms on $\mathrm{Aff}(1)$ acting on $\mathbb{R}$, for $f=\tanh$, taking the largest defect over a few inputs and group elements. Only the form used in this paper becomes invariant, and the other approaches the largest possible defect. For the unimodular groups of Table 1 with symmetric balls, the two forms coincide.

Table 3: Largest defect of the average over F_n on $\mathrm{Aff}(1)$, for the two forms of the average.

n	$\int_{F_n} f(gx)\,d\mu(g)$	$\int_{F_n} f(g^{-1}x)\,d\mu(g)$, Eq. (1)
10	1.971	0.1630
30	1.990	0.0543
100	1.997	0.0163
300	1.999	0.0054

C The Lorentz ball average

Cartan coordinates. Every proper orthochronous Lorentz transformation can be written as $g=k_1a_\eta k_2$, with rotations $k_1,k_2\in K=\mathrm{SO}(3)$ and a boost a_η of rapidity $\eta\ge 0$ along a fixed axis. In these coordinates the Haar measure is $\sinh^2\eta\,d\eta\,dk_1\,dk_2$, so $p=2$, and B_H consists of the elements with $\eta\le H$ [21]. The weight concentrates near the cutoff, which suggests comparing A_Hf with the spherical average at the cutoff,

$$D_Hf(x)=\int_K\int_K f(k_1a_Hk_2\,x)\,dk_1\,dk_2. \tag{2}$$

That is, D_Hf rotates the input, boosts it by the cutoff rapidity H, rotates it again, applies the model and averages over both rotations [31].

Proposition 5 (Limit of the Lorentz ball average)**.** *Let $G=KA^+K$ be a rank-one semisimple Lie group, such as $\mathrm{SO}(1,n)^+$, whose radial Haar density grows like $e^{p\eta}$, and let f be bounded. If $|f(k_1a_\eta k_2x)-f_\infty(k_1,k_2,x)|\le Ce^{-\alpha\eta}$ for all k_1, k_2 and x, then*

$$\sup_x|A_Hf(x)-D_Hf(x)|\le C'\big(1+H\,\mathbf{1}[\alpha=p]\big)\,e^{-\min(p,\alpha)H}.$$

Moreover D_Hf is invariant under K for every H, and D_Hf converges to $D_\infty f=\int_K\int_K f_\infty\,dk_1\,dk_2$.

The proof below integrates the deviation $e^{-\alpha\eta}$ of the model against the weight, which is proportional to $e^{p(\eta-H)}$. The ball average therefore forgets the model below the cutoff at the smaller of two rates, the rate p at which the weight concentrates near the cutoff and the rate α at which the model settles along a boost. Figure 2 shows the gap to D_Hf closing at the predicted rate for a model that is not rotation-invariant, while the gap to the model at a single boost of rapidity H does not close, so the limit cannot be replaced by the model at a single group element.

Invariance of the limit. D_Hf is invariant under rotations for every cutoff, but under boosts only in the limit, and only if the model settles. If f_∞ depends on the rotations only through the direction k_1 of the boost and on the input only through its Lorentz invariants, as for any model that depends continuously on the invariants and on the directions of the four-vectors, then $D_\infty f$ is Lorentz-invariant. Numerically, the defect of A_Hf decreases towards zero for a model that settles and stays of order one for a bounded model that oscillates along boosts (Table 4). This invariance is of little use for prediction, because $D_\infty f$ is determined by the values of the model at infinite rapidity, far from any data, and for a single four-vector it is a constant. A similar truncation for $\mathrm{SO}(1,n)$ arises in refined algebraic quantization [16].

Proof of Proposition 5. The Haar measure of $G=KA^+K$ is $\delta(\eta)\,d\eta\,dk_1\,dk_2$ [21, Ch. I, §5]. Writing $g^{-1}=k_2^{-1}a_\eta^{-1}k_1^{-1}$ and $a_\eta^{-1}=wa_\eta w^{-1}$ for a Weyl element $w\in K$, invariance of the Haar measure of K gives

$$A_Hf(x)=\int_0^H w_H(\eta)\,g_x(\eta)\,d\eta,\qquad w_H(\eta)=\frac{\delta(\eta)}{\int_0^H\delta},$$

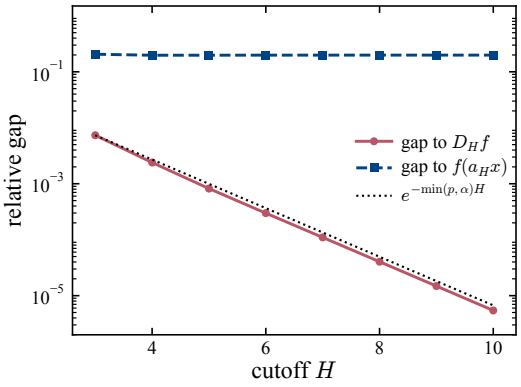


Figure 2: Gap between the ball average $A_H f$ on $\mathrm{SO}(1,3)^+$ and two candidate limits, for a model that is not rotation-invariant. Only the gap to the rotation average $D_H f$ closes, at the rate $\min(p,\alpha)$ of Proposition 5.

with $g_x(\eta) = \int_K \int_K f(k_1 a_\eta k_2 x)\, dk_1\, dk_2$, and $D_H f(x) = g_x(H)$. By hypothesis $|g_x(\eta) - g_x(\infty)| \le Ce^{-\alpha\eta}$, with $g_x(\infty) = D_\infty f(x)$, so

$$
\begin{aligned}
|A_H f(x) - D_H f(x)| &\le \int_0^H w_H(\eta)\big(|g_x(\eta) - g_x(\infty)| + |g_x(H) - g_x(\infty)|\big)\, d\eta \\
&\le Ce^{-\alpha H} + C\int_0^H w_H(\eta) e^{-\alpha\eta}\, d\eta.
\end{aligned}
$$

Since $\delta(\eta) \le c_2 e^{p\eta}$ and $\int_0^H \delta \ge c_1 e^{pH}$ for $H \ge 1$, the last integral is at most $(c_2/c_1)e^{-pH}\int_0^H e^{(p-\alpha)\eta}\, d\eta$, which is at most of order $e^{-\alpha H}$ if $\alpha < p$, at most of order e^{-pH} if $\alpha > p$, and equal to $(c_2/c_1)He^{-pH}$ if $\alpha = p$. For $k \in K$, $D_H f(kx) = \int\int f(k_1 a_H k_2 k\, x)\, dk_1\, dk_2 = D_H f(x)$ by right invariance of the Haar measure of K, and $D_H f \to D_\infty f$ by dominated convergence. □

The argument gives an upper bound. For the model below, whose deviation along a boost is of order $e^{-\alpha\eta}$ with $\alpha = 1$ and no smaller, the measured rate agrees with $\min(p,\alpha) = 1$. Along a single boost axis with the uniform weight, the Haar measure of that one-parameter group, $A_H f(x) - f(a_H x) = H^{-1}\int_0^H (f(a_\eta x) - f(a_H x))\, d\eta$ is of order $1/H$ for any model that settles, because the rapidities at which the model has not yet settled keep weight of order $1/H$.

Measurements on $\mathrm{SO}(1,3)^+$. The data are 2000 four-momenta of unit mass with isotropic three-momenta. The average uses 200 quadrature nodes in η and 240 Monte Carlo pairs (k_1, k_2), held fixed across cutoffs so that the two candidate limits are compared with common random numbers. The model is $f(Z) = \tanh(Z_0/\|Z\| + 1.5\, Z_1/\|Z\|)$, where $\|Z\|$ is the Euclidean norm of the four-vector, and its second term makes it not rotation-invariant and settles like $e^{-\eta}$, so that $\alpha = 1 < p = 2$. Over $H \in [5, 10]$ the gap $\|A_H f - D_H f\|/\|A_H f\|$ decays at fitted rate 1.00 on the full group. The gap to the single-boost value $f(a_H x)$ stays between 0.198 and 0.199 for H from 4 to 10, and at $H = 10$ the spread of $f(k_1 a_H k_2 x)$ over rotations is 0.76 of its mean (Figure 2).

Defect of the averaged model. Table 4 gives the defect $\|A_H f(h\,\cdot) - A_H f\|/\|A_H f\|$ of the full-group average under a boost h of unit rapidity along an axis orthogonal to the one used for a_η, with the same data and 120×120 rotations. For the model above that settles, the defect decreases towards zero as H grows. For the bounded model $\sin(3 \log Z_0)$, which oscillates along boosts, the defect stays of order one. With a finite sample of rotations, $D_H f$ is rotation-invariant up to Monte Carlo error.

Table 4: Relative defect of the averaged model under a unit boost.

model	$H{=}2$	$H{=}4$	$H{=}6$	$H{=}8$
settles, $\tanh(Z_0/\lVert Z\rVert + 1.5\,Z_1/\lVert Z\rVert)$	0.020	0.0085	0.0014	0.00019
does not settle, $\sin(3\log Z_0)$	1.389	1.464	1.472	1.446

D Proofs for Section 4

Proof of Lemma 1. For each x, $\ell(A_\nu f(x), y) = \ell(\mathbb{E}_{g\sim\nu} f(g^{-1}x), y) \le \mathbb{E}_{g\sim\nu}\ell(f(g^{-1}x), y)$ by Jensen's inequality, since ℓ is convex in its first argument and ν is a probability distribution. Taking the expectation over P and exchanging the order of integration gives the claim. □

Proper actions. An action is proper if, for every compact set C, the set of g with $gC \cap C \neq \emptyset$ is relatively compact. Suppose P is a G-invariant probability distribution on the inputs and G is not compact. Choose a compact C with $P(C) > 1/2$. Because G is not compact, some h lies outside the closure of $\{g : gC \cap C \neq \emptyset\}$, so hC and C are disjoint, and $P(hC \cup C) = 2P(C) > 1$, a contradiction. The hypothesis is needed, since translations of a stationary random field, or a non-compact group acting on a finite-volume quotient, preserve probability distributions. Villar et al. [44] noted the same obstruction for scalings, Spinner et al. [39] for boosts in Lorentz-equivariant generative models, and Lawrence and Harris [26] for augmentation. Footnotes in Lyle et al. [28] and Chen et al. [8] extend their results to more general groups or measures. For a proper action of a non-compact group the invariance hypothesis they keep cannot be satisfied, so these extensions do not yield a risk guarantee for truncated averages.

Proof of Proposition 4. Take $G = \mathbb{R}$ acting on $\mathbb{R}$ by translation, target $y \equiv 0$, squared loss, model $f(x) = \mathbf{1}\{|x| \ge 1\}$ and data $x \sim \mathcal{N}(0, \varepsilon^2)$. The averaging distribution μ_H is a probability distribution, the target is invariant and the loss is convex, so Lemma 1 applies and only the invariance of the data distribution is missing. Then $\mathcal{R}(f) = \mathbb{E}f(x)^2 = 2\Phi(-1/\varepsilon)$ and

$$A_H f(x) = 1 - \frac{\big|[x-1, x+1] \cap [-H, H]\big|}{2H} = 1 - \frac{1}{H} \quad \text{for } |x| \le H - 1,$$

so $\mathcal{R}(A_H f) \ge (1 - 1/H)^2\,\mathbb{P}(|x| \le H - 1)$, which tends to one as $H \to \infty$ for every ε. Since $0 \le A_H f \le 1$, one is also the largest possible risk. □

The risk of f is below one for every ε and tends to zero as $\varepsilon \to 0$, so the ratio of the two risks mostly reflects how small $\mathcal{R}(f)$ is. Table 5 therefore lists both risks, computed by quadrature, together with the lower bound, which is tight once $H - 1$ is large compared with ε. For small cutoffs averaging can still help, as in the row with $\varepsilon = 1$ and $H = 2$.

Table 5: Risk of the model and of its truncated average in the translation counterexample.

ε	H	$\mathcal{R}(f)$	$\mathcal{R}(A_H f)$	lower bound
0.25	2	6.33×10^{-5}	0.2500	0.2500
0.25	4	6.33×10^{-5}	0.5625	0.5625
0.25	8	6.33×10^{-5}	0.7656	0.7656
0.25	32	6.33×10^{-5}	0.9385	0.9385
0.5	2	0.046	0.2523	0.2386
0.5	4	0.046	0.5625	0.5625
0.5	8	0.046	0.7656	0.7656
0.5	32	0.046	0.9385	0.9385
1.0	2	0.317	0.3007	0.1707
1.0	4	0.317	0.5626	0.5610
1.0	8	0.317	0.7656	0.7656
1.0	32	0.317	0.9385	0.9385

E The top-quark tagging experiment

Data and model. We use the top-quark tagging reference dataset [24], which contains jets with transverse momentum between 550 and 650 GeV and $|\eta| < 2$, represented by the four-momenta of

up to 200 constituents, with 1.2M jets for training, 400k for validation and 400k for testing, in the converted form distributed with the Particle Transformer code [37]. The model is ParticleNet [36], trained with the reference implementation and the training command for this dataset released with that code (`train_TopLandscape.sh PN kin`, with weaver-core 0.4.17), without changes. Its inputs are seven kinematic features of up to 128 particles, namely the logarithms of the transverse momentum and energy, their logarithmic ratios to the jet values, and the pseudorapidity, azimuth and angular distance relative to the jet axis, with the pseudorapidity and azimuth also serving as the coordinates for its nearest-neighbour graphs. Training runs for 20 epochs with the Ranger optimizer, and the epoch with the highest validation accuracy is kept. We train three models independently, which on the full test set reach AUCs of 0.9857 to 0.9858 and an accuracy of 0.940, and reject background at 30% signal efficiency by factors of 1518 to 1741. The authors of ParticleNet report an AUC of 0.9858, an accuracy of 0.940 and a rejection factor of 1615 ± 93 on this dataset [36].

Averages. Both groups act on all constituents of a jet, and a rotation by an angle θ about the jet axis is a Lorentz transformation that leaves the jet momentum unchanged. A boost of rapidity b along the jet axis keeps the direction of the jet but changes its energy and momentum, and so moves jets out of the transverse-momentum window of the dataset. The rotation average uses 16 equally spaced angles, and the boost average uses 16 equally spaced rapidities in $[-H, H]$, which approximate the uniform distribution on a ball of the one-parameter boost group. The transformed jets pass through the same conversion and preprocessing code as the original test jets. We average the predicted probability of the top class, so the binary cross-entropy is convex in the averaged prediction and Lemma 1 applies. The averages are evaluated on the first 100k test jets, and intervals are bootstrap percentiles over these jets.

Results. Table 6 gives, for each of the three models, the ratio of the binary cross-entropy of the averaged tagger to that of the tagger on the same jets. Over rotations the risk rises by about 9% for every model, with intervals far above one, and the AUC falls, so the increase is more than a change of calibration. Over boosts the risk changes by less than half a percent, in either direction, up to a cutoff of 0.2, and rises steeply beyond it, to about three times its value at a cutoff of 1.6.

Table 6: Averages of ParticleNet over rotations about the jet axis and over boosts along it, on the first 100k test jets. The risk ratio is the binary cross-entropy of the averaged tagger divided by that of the tagger, for each of the three models. The half-width is the largest half-width of the bootstrap 95% intervals of the three ratios, and the AUC is the mean over the models.

average	cutoff H	risk ratio, three models	half-width	AUC
none	–	1, 1, 1	–	0.9854
rotations	–	1.090, 1.089, 1.087	0.008	0.9828
boosts	0.05	0.999, 1.000, 0.999	0.001	0.9854
boosts	0.1	0.998, 0.999, 0.997	0.001	0.9854
boosts	0.2	1.000, 1.004, 0.997	0.002	0.9853
boosts	0.4	1.062, 1.065, 1.052	0.007	0.9843
boosts	0.8	1.505, 1.450, 1.618	0.024	0.9822
boosts	1.6	3.029, 2.973, 3.495	0.071	0.9793

Numerical sensitivity. Converting the untransformed test jets with the repository's code reproduces the official test files up to float32 rounding, with differences below 10^{-6} in the particle coordinates. ParticleNet builds nearest-neighbour graphs from these coordinates, and rounding at this level can change which particles are neighbours. On the first 2000 test jets it changes the predicted probability by more than 0.01 for 1.4% of the jets, and running the same files on a CPU instead of a GPU does so for 2.9%. Over 100k jets the largest change is 0.36, but such changes average out in the risk, and computed from the converted jets the risk of each model differs from that of the reference predictions by less than 0.1%, and the AUC by less than 10^{-4}. Changes of the risk ratio of a few tenths of a percent, as for the smallest cutoffs, are therefore of the same order as this numerical effect.

F Further related work

Averages, frames and canonicalization. Frame averaging [35] averages over an input-dependent subset of the group rather than over the group itself, works with any architecture, and includes centering the inputs as a frame for translations. Minimal frames exist for the Lorentz group and

for $GL(d)$ [27], and local canonicalization makes any network Lorentz-equivariant [40], although canonicalization [23] can be discontinuous, as shown for several compact groups [11]. For the non-compact Lie point symmetries of partial differential equations, canonicalization has been used to impose equivariance on pretrained neural operators [38]. These constructions do not need an invariant measure on the group and are therefore not limited by the results of this paper. Averaging against the conditional distribution of the data on each orbit keeps the risk from increasing for free actions of any group [2]. Invariant two-sample tests for non-compact groups have been built by integrating an orbit-weighted kernel against Haar measure, which avoids normalizing the measure [15].

Benefits of symmetry under compactness. The benefit of imposing a symmetry has been quantified as a reduction in sample complexity [3, 41], and as an approximation-generalization trade-off when the symmetry holds only approximately [34], in both cases for a finite or compact group. On spherical images, rotation augmentation matches an equivariant network on classification only with much more data and larger models, and falls short on segmentation [14]. For physics data, invariant scalars [43] and units equivariance [44, 45] impose symmetries, including non-compact ones, by construction rather than by averaging.

Lorentz symmetry in jet tagging. Lorentz-equivariant taggers evaluated on the top-tagging dataset take the beam directions as additional inputs [17, 39], so they are invariant only under the transformations that fix the beam. Spinner et al. [39] name the beam direction as the main source of symmetry breaking in measured data, followed by the resolution of the detector. Gong et al. [17] boost the test jets along an axis transverse to the beam and find that taggers without Lorentz symmetry lose accuracy as the boost grows.

Amenability in statistics. In decision theory, the best invariant procedure is minimax when the group is amenable, through averages over asymptotically invariant sequences of measures, and the conclusion can fail for non-amenable groups [5]. Orbanz connects this Hunt–Stein theory to symmetry constraints in learning and uses Følner averages for amenable groups [33]. The risk there is a worst case over an orbit of distributions, whereas the risk guarantee of this paper concerns a single data distribution and needs it to be invariant, which for a proper action is possible only if the group is compact, whether or not it is amenable.

The Weyl unitarian trick. Foit et al. [13] remark that some averaging techniques can be extended to non-compact groups using the Weyl unitarian trick, citing Villar et al. [44]. There the trick is carried out for the scaling group, where a real-analytic model is continued analytically and averaged over a compact real form. This produces an invariant model for real-analytic f, but not an average over the group itself, and Villar et al. show that the resulting projection differs from the orthogonal projection for every data distribution, because no scaling-invariant probability distribution exists. The trick therefore avoids truncation altogether and does not restore the risk guarantee.