On Uncertainty Calibration for Invariant Functions

Edward Berman BERMAN.ED@NORTHEASTERN.EDU
Northeastern University, Boston, USA

Jacob Ginesin GINESIN@CMU.EDU
Carnegie Mellon University, Pittsburgh, USA

Marco Pacini MPACINI@FBK.EU
University of Trento, Trento, Italy

Robin Walters R.WALTERS@NORTHEASTERN.EDU
Northeastern University, Boston, USA

Editors: List of editors' names

Abstract

Data-sparse settings such as robotic manipulation, and molecular physics, and galaxy morphology classification are some of the hardest domains for deep learning. For these problems, equivariant networks can help improve modeling across undersampled parts of the input space, and uncertainty estimation can guard against overconfidence. However, until now, the relationships between equivariance and model confidence, and more generally equivariance and model calibration, has yet to be studied. In this work, we present the first theory relating invariance to uncertainty estimation. By proving lower and upper bounds on uncertainty calibration errors under various invariance conditions, we elucidate the generalization limits of invariant models.

Keywords: Approximation Error Bound, Group Invariance, Uncertainty Estimation

1. Introduction

Equivariant neural networks are a class of neural networks that encode group symmetries into the structure of the network architecture such that the symmetries do not need to be learned from data. Understanding both model calibration and confidence is particularly useful in data-sparse settings where equivariant neural networks tend to thrive, such as pick-and-place robotics tasks (Kalashnikov et al., 2018; Wang et al., 2022b,a; Fu et al., 2023; Huang et al., 2023, 2024b,a; Wang, 2025), galaxy morphology classification (Pandya et al., 2023, 2025), and molecular physics (Zou et al., 2023; Ramakrishnan et al., 2014). While equivariance has proved invaluable in these scenarios, it does have some drawbacks, including fairly limited benefits at scale (Wang et al., 2023b; Klee et al., 2023; Gruver et al., 2023; Brehmer et al., 2024; Abramson et al., 2024), provable degradation on model performance in cases of symmetry mismatch (Wang et al., 2024), more complex architectures, and higher compute costs. Despite these drawbacks, a surprising result of Wang et al. (2023a) is that equivariant neural networks can still be effective even in cases of mismatch between the model and the data symmetry. This finding motivated the work of Wang et al. (2024), which explored how equivariance can affect model *accuracy*, both positively and negatively. However, it is not yet understood how equivariance impacts model *calibration*, loosely defined as the disagreement between a model's accuracy and predicted *confidence*.

© 2025 E. Berman, J. Ginesin, M. Pacini & R. Walters.

To help understand the tradeoffs associated with the inductive bias of equivariance, we seek to quantify the impact of equivariance on expected calibration error. While no previous works are using the generalization limits of equivariant models, it is apparent that previous results on the subject can be applied to the study of calibration error. This is because calibration errors contain expressions corresponding to classification or regression errors themselves. In this work, we extend the error bounds given by Wang et al. (2024) to a broader class of calibration losses. In this way, we can quantify the effect of invariance not just on accuracy, but also on calibration. In particular, we show that calibration error is related to typical classification errors over the fibers of each confidence prediction. These classification errors have known bounds for invariant functions, which we use to provide lower and upper calibration error bounds. This study illustrates that the effect of invariance on model calibration is dependent on where we are in the equivariance taxonomy (i.e. correct, incorrect, or extrinsic equivariance), supporting the previous line of work in Wang et al. (2023a, 2024).

2. Background

2.1. Invariance and Equivariance

Here, we give precise definitions of equivariance and invariance. Let G be a group with representations $\rho^{\mathcal{X}}$ and $\rho^{\mathcal{Y}}$ which transform vectors in the vector spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. Representations map group elements to invertible linear transformations. When clear, we omit the representation map and write gx for $\rho^{\mathcal{X}}(g)x$. A map $\phi : \mathcal{X} \to \mathcal{Y}$ is *equivariant* if $\rho^{\mathcal{Y}}(g)[\phi(x)] = \phi(\rho^{\mathcal{X}}(g)[x])$ for all $g \in G, x \in \mathcal{X}$. Invariance is a special case of equivariance in which $\rho^{\mathcal{Y}} = \mathrm{Id}^{\mathcal{Y}}$ for all $g \in G$. That is, a map $\phi : \mathcal{X} \to \mathcal{Y}$ is *invariant* if it satisfies $\phi(x) = \phi(\rho^{\mathcal{X}}(g)[x])$ for all $g \in G, x \in \mathcal{X}$.

Fundamental Domain. This paper will use iterated integration over orbits and a set of orbit representatives. We call the set of orbit representatives the fundamental domain, denoted F. A precise definition is given in Appendix A.

2.2. Equivariant Learning

Consider a function $f : X \to Y$. Let $p : X \to \mathbb{R}$ be the probability density function of the domain X. We assume there is no distribution shift during testing – that is, p is always the underlying distribution during training and testing. The goal for a model class $\{h : X \to Y\}$ is to fit the function f by minimizing an error function $\mathrm{err}(h)$. We assume the model class $\{h\}$ is arbitrarily expressive except that it is constrained to be equivariant with respect to a group G. Let $\mathbb{1}(\mathcal{I}(x))$ be an indicator function that equals to 1 if the condition $\mathcal{I}$ is satisfied and 0 otherwise. The classification error is given by $\mathrm{err}_{\mathrm{cls}}(h) = \mathbb{E}_{x\sim p}\left[\mathbb{1}(f(x) \neq h(x))\right]$.

2.3. Error Bounds for Invariant Classifiers

Our goal is to generalize the bounds from Wang et al. (2024) to a calibration objective. We briefly review the main classification result from Wang et al. (2024). Given that equivariance is not always correct, the following definition and theorem detail how symmetry mismatch can harm model fitting for invariant classification.

Definition 1 (Majority Label Total Dissent) *For the orbit Gx of $x \in X$, the total dissent $k(Gx)$ is the integrated probability density of the elements in the orbit Gx having a different label than the majority label: $k(Gx) = \min_{y \in Y} \int_{Gx} p(z)\mathbb{1}(f(z) \neq y)dz$.*

Theorem 2 (Theorem 4.3 in Wang et al. (2024)) *The error $err_{cls}(h)$ is bounded below by $\int_F k(Gx)dx$.*

Additional background on misspecified equivariance is given in Appendix B.

3. Invariant Classification Bounds

We first note that the expected calibration error from Guo et al. (2017) is bounded in the interval $[0, 1]$, and that under the hypothesis of an invariant model class both the lower and upper bounds can be tightened. We present complete proofs in Appendix F.

Classification Problem. Consider a function $f : X \to Y$ where Y is a finite set of labels. Let $q : X \to \mathbb{R}$ be a probability density on the domain X. We define a model class $\{h : X \to Y \times [0, 1]\}$. Let $h(x) = (h_Y, h_P)$ where h_P represents the probability estimate associated with the predicted label h_Y.The goal is for h to fit the function f *and* to properly predict its own confidence by minimizing the expected calibration error (Equation 1, and Equation 2 in Guo et al. (2017)). Following Wang et al. (2024), we assume that the class $\{h\}$ is arbitrarily expressive except that it is constrained to be equivariant with respect to a group G. In the classification setting, we specifically assume h to be $G-$invariant, which, although not strictly necessary, is the case in most classification problems considered in the literature. Let $r(p)$ be the probability density such that $\mathbb{P}(p_1 \leq h_P(x) \leq p_2) = \int_{p_1}^{p_2} r(p)dp$. This is the push-forward of q over h_P. The expected calibration error is nominally defined

$$\mathrm{ECE}(h) = \mathbb{E}_{h_P}\left[\left|\, \mathbb{P}\left(f = h_Y | h_P = p\right) - p \,\right|\right] \tag{1}$$

as in Guo et al. (2017). Intuitively, if a model has confidence p, then it should be accurate with probability p. This metric penalizes the discrepancy between accuracy and confidence averaged over all of confidences weighted by the push-forward density r. We abbreviate $\mathbb{P}(f(x) = h_Y(x)|h_P(x) = p)$ with $\mathrm{Acc}_p(h)$, which denotes the true accuracy of the model when the predicted confidence is p. Hence h is well calibrated at confidence p when $\mathrm{Acc}_p(h) = p$, underconfident when $\mathrm{Acc}_p(h) > p$, and overconfident when $\mathrm{Acc}_p(h) < p$.

We briefly comment on the well-definedness of Equation 1 in Appendix C.

ECE Upper Bounds. We first note that ECE is a bounded between 0 and 1, since it is bounded below by 0 and bounded above by 1. The upper bound of 1 is generally loose without any further assumptions. Therefore, we now show that the assumption of invariance on the model class allows us to tighten the upper bound. We start with the following assumption on a fundamental domain F and orbit Gx that ensures iterated integration on the fibers of h_P is well defined.

Assumption 3 *For a group G acting on a domain X, we assume the union of all pairwise intersections $\cup_{g_1 \neq g_2}(g_1 F \cap g_2 F)$ have measure 0 and that F and Gx are differentiable manifolds for all $x \in X$. This holds for domains $\mathcal{F}_p = h_P^{-1}(p) \subseteq X$ on which G also acts.*

Proposition 4 *Denote the fiber $\mathcal{F}_p = h_P^{-1}(p)$. Denote the total density on a fiber $\mathcal{F}_p$ by $q(\mathcal{F}_p) = \int_{\mathcal{F}_p} q(x)dx$ and the renormalized density by $q_p(x) = q(x)/q(\mathcal{F}_p)$. Let $k(Gx, p)$ be the total dissent of an orbit on $\mathcal{F}_p$ with the renormalized probability $q_p(x)$. Denote a fundamental domain of G in $\mathcal{F}_p$ as F_p. Let $P_1 = \{p : \text{Acc}_p(h) \leq \frac{1}{2}\}$ and $P_2 = \{p : \text{Acc}_p(h) \geq \frac{1}{2}\}$. Let Gx^* be the orbit with the smallest nonzero total dissent $k(Gx^*)$, i.e. $x^* = \arg\min_{x \in X}\{k(Gx)\}$. ECE is bounded above by $\frac{1}{2} + \int_0^1 r(p)|\frac{1}{2} - p|dp - k(Gx^*)\int_{P_2} r(p)dp$.*

Proof Sketch 1 We observe that $|\text{Acc}_p(h) - p| \leq |\text{Acc}_p(h) - 1/2| + |1/2 - p|$. The upper bound on ECE is determined by the upper bound of $|\text{Acc}_p(h) - 1/2|$, and we observe that the accuracy on each fiber is constrained by invariance. By considering the orbit with the lowest nonzero total dissent, we can compute an upper bound that is tighter than 1 even without knowing the error lower bound on each fiber, i.e. $\int_{F_p} k(Gx, p)dx$, or the fibers themselves $\mathcal{F}_p$.

This upper bound is tighter than 1 since it accounts for the error caused by incorrect invariance along the subset of fibers where accuracy is at least 50%. In other words, if we consider all of the orbits with incorrect invariance, then ECE is only as bad as the best of those orbits. One tradeoff this bound makes is that it is in terms of $k(Gx^*)$, which only considers error along one orbit. If we know which data points are in each fiber of h_P, then we can tighten the bound. In Appendix D, we introduce further assumptions that allow us to tighten the bound and provide examples on how the bound may be computed.

We can similarly derive an ECE lower bound using invariance. We start by defining the minority label total dissent $\kappa(Gx)$, which is the integrated density of the elements in the orbit Gx having a different label than the minority label (see Appendix E for a formal definition). In Appendix E, we prove that $\text{err}_{\text{cls}}(h)$ is bounded above by $\int_F \kappa(Gx)dx$. Leveraging this, we can now prove the ECE invariant lower bound.

Theorem 5 *We will denote the fundamental domain of G in $\mathcal{F}_p$ as F_p, where $\mathcal{F}_p$ is as defined in Proposition 4. As in Proposition 4, the total dissent on an orbit in a fiber $\mathcal{F}_p$ is denoted $\kappa(Gx, p)$ and is defined in terms of the renormalized density $q_p(x) = q(x)/q(\mathcal{F}_p)$. Define the minimum fiberwise classification accuracy as $m = \min_{p \in [0,1]} 1 - \int_{F_p} \kappa(Gx, p)dx$. ECE is bounded below by $\int_0^m r(p)(m - p)dp$.*

Proof Sketch 2 If an orbit contains each label in $|Y|$, then an invariant model will be correct at least once (i.e. a broken clock is right twice a day). We find the accuracy lower bound for each orbit and then bound using the smallest one, m. We integrate the discrepancy between accuracy and confidence in the region where the confidence is less than the accuracy lower bound.

Special cases that tighten and exemplify the bound are provided in Appendix E.

4. Conclusion

This work proves upper and lower bounds for ECE for invariant functions, elucidating the generalization limits of invariant functions in cases of symmetry mismatch between the model and the data.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale? *arXiv preprint arXiv:2410.23179*, 2024.

Jiahui Fu, Yilun Du, Kurran Singh, Joshua B Tenenbaum, and John J Leonard. Neuse: Neural se (3)-equivariant embedding for consistent spatial understanding with objects. *arXiv preprint arXiv:2303.07308*, 2023.

Nate Gruver, Marc Anton Finzi, Micah Goldblum, and Andrew Gordon Wilson. The lie derivative for measuring learned equivariance. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JL7Va5Vy15J.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

L. Hörmander. *The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis*. Classics in Mathematics. Springer Berlin Heidelberg, 2015. ISBN 9783642614972. URL https://books.google.com/books?id=aaLrCAAAQBAJ.

Haojie Huang, Dian Wang, Xupeng Zhu, Robin Walters, and Robert Platt. Edge grasp network: A graph-based se (3)-invariant approach to grasp detection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3882–3888. IEEE, 2023.

Haojie Huang, Owen Howell, Dian Wang, Xupeng Zhu, Robin Walters, and Robert Platt. Fourier transporter: Bi-equivariant robotic manipulation in 3d. *arXiv preprint arXiv:2401.12046*, 2024a.

Haojie Huang, Dian Wang, Arsh Tangri, Robin Walters, and Robert Platt. Leveraging symmetries in pick and place. *The International Journal of Robotics Research*, 43(4): 550–571, 2024b.

E. T. Jaynes. *Probability theory*. Cambridge University Press, Cambridge, 2003. ISBN 0-521-59271-2. doi: 10.1017/CBO9780511790423. URL https://doi.org/10.1017/CBO9780511790423. The logic of science, Edited and with a foreword by G. Larry Bretthorst.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-opt: scalable deep reinforcement learning for vision-based robotic manipulation. corr abs/1806.10293 (2018). *arXiv preprint arXiv:1806.10293*, 2018.

David Klee, Jung Yeon Park, Robert Platt, and Robin Walters. A comparison of equivariant vision models with imagenet pre-training. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023.

Andrey Kolmogorov. Grundbegriffe der wahrscheinlichkeitsrechnung (in german), berlin: Julius springer. 1933.

Sneh Pandya, Purvik Patel, Jonathan Blazek, et al. E (2) equivariant neural networks for robust galaxy morphology classification. *arXiv preprint arXiv:2311.01500*, 2023.

Sneh Pandya, Purvik Patel, Brian D Nord, Mike Walmsley, and Aleksandra Ciprijanovic. Sidda: Sinkhorn dynamic domain adaptation for image classification with equivariant neural networks. *Machine Learning: Science and Technology*, 2025. URL http://iopscience.iop.org/article/10.1088/2632-2153/adf701.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1): 1–7, 2014.

Dian Wang. *Equivariant Policy Learning for Robotic Manipulation*. PhD thesis, Northeastern University, 2025. URL https://link.ezproxy.neu.edu/login?url=https://www.proquest.com/dissertations-theses/equivariant-policy-learning-robotic-manipulation/docview/3224180630/se-2. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2025-07-16.

Dian Wang, Mingxi Jia, Xupeng Zhu, Robin Walters, and Robert Platt. On-robot learning with equivariant models. *arXiv preprint arXiv:2203.04923*, 2022a.

Dian Wang, Robin Walters, Xupeng Zhu, and Robert Platt. Equivariant q learning in spatial action spaces. In *Conference on Robot Learning*, pages 1713–1723. PMLR, 2022b.

Dian Wang, Jung Yeon Park, Neel Sortur, Lawson L.S. Wong, Robin Walters, and Robert Platt. The surprising effectiveness of equivariant models in domains with latent symmetry. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=P4MUGRM4Acu.

Dian Wang, Xupeng Zhu, Jung Yeon Park, Mingxi Jia, Guanang Su, Robert Platt, and Robin Walters. A general theory of correct, incorrect, and extrinsic equivariance. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuyang Wang, Ahmed AA Elhag, Navdeep Jaitly, Joshua M Susskind, and Miguel Ángel Bautista. Swallowing the bitter pill: Simplified scalable conformer generation. In *Forty-first International Conference on Machine Learning*, 2023b.

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

Zihan Zou, Yujin Zhang, Lijun Liang, Mingzhi Wei, Jiancai Leng, Jun Jiang, Yi Luo, and Wei Hu. A deep learning model for predicting selected organic molecular spectra. *Nature Computational Science*, 3(11):957–964, 2023.

Appendix A. Fundamental Domain

Here, we give a precise definition for the fundamental domain.

Definition 6 (Definition 4.1 in Wang et al. (2024)) *Let d be the dimension of a generic orbit of G in X and n the dimension of X. Let ν be the $(n-d)$ dimensional Hausdorff measure in X. A closed subset F of X is called a fundamental domain of G in X if X is the union of conjugates of F, i.e., $X = \cup_{g\in G} gF$, and the intersection of any two conjugates has 0 measure under ν.*

Appendix B. Equivariance Taxonomy: Correct, Incorrect, and Extrinsic.

Wang et al. (2024) establish a taxonomy which describes the relationship of the symmetry of the model class to the symmetry in the data. We review the definitions of correct, incorrect, and extrinsic equivariance from Wang et al. (2024) . These definitions help us understand the ability of equivariant functions to approximate datasets that may or may not have the same symmetries. A key inclusion here is extrinsic symmetry, which describes the case where the action of the group moves data points out of the support of their original distribution.

Definition 7 (Equivariance Taxonomoy, Definitions 3.1-3.3 in Wang et al. (2024)) *For all $x \in X$, $g \in G$ where $p(x) > 0$, if $p(gx) > 0$ and $f(gx) = gf(x)$, h has correct equivariance with respect to f. For all $x \in X$, $g \in G$ where $p(x) > 0$, if $p(gx) > 0$ and $f(gx) \neq gf(x)$, h has incorrect equivariance with respect to f. For all $x \in X$, $g \in G$ where $p(x) > 0$, if $p(gx) = 0$, h has extrinsic equivariance with respect to f.*

An important nuance is that different subsets of a dataset can belong to different classes of the taxonomy, e.g. the data set can be $\frac{1}{3}$ correct, $\frac{1}{3}$ incorrect, and $\frac{1}{3}$ extrinsic. We can add specificity to Definition 7 by considering the type of equivariance at *each point* in the dataset.

Definition 8 (Pointwise Equivariance, Definitions 3.5-3.7 in Wang et al. (2024)) *For $g \in G$ and $x \in X$ where $p(x) \neq 0$, if $p(gx) \neq 0$ and $f(gx) = gf(x)$, h has correct equivariance with respect to f at x under transformation g. For $g \in G$ and $x \in X$ where $p(x) \neq 0$, if $p(gx) \neq 0$ and $f(gx) \neq gf(x)$, h has incorrect equivariance with respect to f at x under transformation g. For $g \in G$ and $x \in X$ where $p(x) \neq 0$, if $p(gx) = 0$, h has extrinsic equivariance with respect to f at x under transformation g.*

Appendix C. Well-definedness of ECE

In this section, we commented on the requirements of Equation 2 in order for it to be well defined. The definition in Guo et al. (2017) abuses notation slightly, in that the probability of any event drawn from a continuous random variable has probability zero, i.e. $p(h_P = p) = 0$ for all p. We can rectify this by defining ECE as

$$\text{ECE}(h) = \lim_{\varepsilon \to 0} \mathbb{E}_{p \sim r(p)} \left[\left| \mathbb{P}\left(f = h_Y | p - \varepsilon < h_P < p + \varepsilon\right) - p \right| \right], \tag{2}$$

however, we drop the limits for brevity throughout. First, $\mathbb{P}\left(f = h_Y | p - \varepsilon < h_P < p + \varepsilon\right)$ is well defined when $r(p) \neq 0$ for all $p \in [0, 1]$. Moreover, we note that in general, if $B = \{C = c\}$, it is not always permissible to define $P(A|B) = \lim_{\varepsilon \to 0} P(A | c - \varepsilon < C < c + \varepsilon)$. This is because we face contradictions when $\{D = d\} = B = \{C = c\}$, but the random variables C and D have different *densities* defined with respect to different *measures*. This results in contradictions where $P(A|B) = \lim_{\varepsilon \to 0} P(A | c - \varepsilon < C < c + \varepsilon)$ and $P(A|B) = \lim_{\varepsilon \to 0} P(A | d - \varepsilon < D < d + \varepsilon)$ but $\lim_{\varepsilon \to 0} P(A | c - \varepsilon < C < c + \varepsilon) \neq \lim_{\varepsilon \to 0} P(A | d - \varepsilon < D < d + \varepsilon)$, see for example the Borel-Kolmogorov Paradox (Kolmogorov, 1933; Jaynes, 2003)[1]. In other words, the probability density conditioned on an event with zero probability can only be specified with respect to a given reference measure that determines the probability density function being conditioned on. Therefore, we specify a measure on X so that the *random variable* h_p has *push-forward density* $r(p)$ defined with respect to the *push-forward measure*. In particular, let $\mathcal{H}$ be the $|X|$ dimensional Hausdorff measure in X that defines $q(x)$. $r(p)$ is the push-forward density of q over h_P, meaning it is defined with respect to the accompanying push-forward measure $h_P \# \mathcal{H}$ on $[0, 1]$. This is sufficient for Equation 2 to be uniquely defined. We also note that we don't need these well-definedness properties to hold in the special case where $r(p)$ is discrete or when we are computing approximations that treat h_P as discrete. In each case, we average over the confidences (or confidence bins) with non-zero probability.

Appendix D. Computing the Upper Bound

Corollary 9 *Define* $m = \min_{p \in [0,1]} \int_{F_p} k(Gx, p)$. *Then* $\text{ECE} \leq \frac{1}{2} + \int_0^1 r(p) |\frac{1}{2} - p| dp - m \int_{P_2} r(p) dp$.

A key subtlety in the proof of Corollary 9 is that m is a minimum over error lower bounds defined on fibers of $[0, 1]$ and not orbits. This is stated formally in Remark 10.

Remark 10 *By assumption of invariance on* h_P, *the fibers of* $[0, 1]$ *contain entire orbits. The integrated total dissent* $\int_{F_p} k(Gx, p) dx$ *is defined on the collection of orbits where the confidence is always given by* $h_p(x_p) = p$, *but the label* $h_Y(x_p)$ *itself may very. This is possible because points* x_{p_1} *and* x_{p_2} *may belong to distinct orbits which map to different distinct labels* y_1 *and* y_2 *under* h_Y *but map to the same confidence* p *under* h_P.

1. This paradox is most easily exemplified with the Great Circle Puzzle.

We note that for the special case of binary classification, an accuracy of 50% is essentially the minimum accuracy on each fiber. If the accuracy of a classifier is less than 50% on each fiber, we can construct a classifier that simply chooses the opposite label to improve its accuracy so that it is accurate over 50% of the time.

Corollary 11 *Assume $|Y| = 2$. ECE is bounded above by $1 - k(Gx^*)$, or $1 - m$ in the special case of Corollary 9.*

Improving the Unconstrained ECE Upper Bound. In our proof of Proposition 4, we used the fact that we could bound $|\text{Acc}_p(h) - \frac{1}{2}|$ on P_2 using invariance. However, we also could have made a simplication without that assumption, noting that $|\text{Acc}_p(h) - \frac{1}{2}| \leq \frac{1}{2}$. We refer to this as the upper bound in the unconstrained case.

Proposition 12 *ECE is bounded from above by $\frac{1}{2} + \int_0^1 r(p)|\frac{1}{2} - p|dp$.*

Comparing with Proposition 4, we see that the assumption of invariance decreases the upper bound by $k(Gx^*) \int_{P_2} r(p)dp$.

Reparameterizing the Distributions. Notice that both of the upper bounds in Propositions 4 and 12 are expressed in terms of $r(p)$. The density $r(p)$ is not in general easily derivable from $q(x)$. In order to express each bound in terms of $q(x)$, we introduce extra assumptions on h_P, which we will now examine.

From Hörmander (2015), we have that

$$r(p) = \int_X q(x)\delta(p - h_P(x))dx = \int_{\mathcal{F}_p} \frac{1}{|\nabla h_P(x)|} q(x)dx_p$$

where δ is the Dirac-Delta distribution. This is valid if we assume that $h_P(x)$ is continuously differentiable and has gradient nowhere 0. Now, to attain our upper bound on ECE independent of h, we must find the upper bound on $\frac{1}{|\nabla h_P(x)|}$. This is achievable if $h_P(x)$ is bi-Lipschitz. Let us define.

Definition 1 *Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is Lipschitz continuous if there exists a constant $K \geq 0$ such that for all x_1, $x_2 \in X$ we have*

$$d_Y(f(x_1), f(x_2)) \leq K d_X(x_1, x_2).$$

Furthermore, a function is (K_1, K_2)-Bi-Lipschitz continuous if

$$\frac{1}{K_2} d_X(x_1, x_2) \leq d_Y(f(x_1), f(x_2)) \leq K_1 d_X(x_1, x_2).$$

The Lipschitz constant K further serves as a bound for the gradient of f, since we can consider the limit as $x_1 \to x_2$. If the function is bi-lipschitz, then K_1 bounds the gradient and K_2 bounds the reciprocal of the gradient.

Proposition 13 *We assume $h_P(x)$ is differentiable, has gradient nowhere 0, and is (K_1, K_2) bi-Lipschitz continuous. Let $G\widetilde{x}$ be the orbit with the least integrated probability density. The upper bound on ECE in the incorrectly invariant case becomes $\frac{2+K_2 \int_{G\widetilde{x}} q(x)dx}{4} - k(Gx^*)K_2 \int_{G\widetilde{x}} q(x)dx$ and in the unconstrained case $\frac{2+\int_{G\widetilde{x}} q(x)dx K_2}{4}$.*

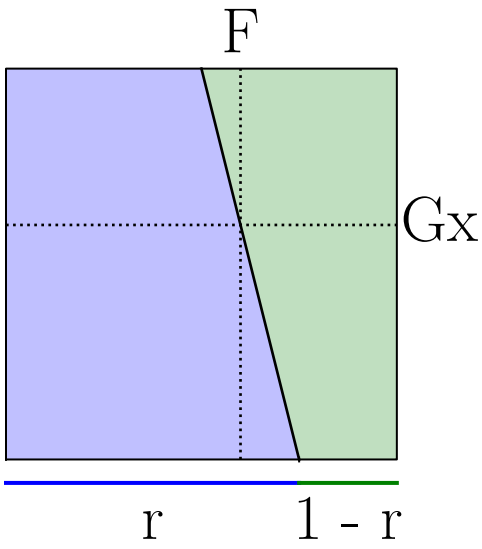

Figure 1: Binary Classification on a Unit Square with Translation Invariance. Blue and green represent true labels.

Comparing the Unconstrained and Invariant Upper Bounds. A natural question now arises, how much tighter is the upper bound on ECE than the bound in the unconstrained case? In the special case where $h_P(x)$ is bi-Lipschitz with known lipschitz constants K_1 and K_2, we are able to compute an exact number for each upper bound. Still, being bi-Lipschitz is a massive constraint, and so we would like to compare the bounds without this assumption. We do this by using test functions for $r(p)$. Specifically, Example 1 considers a binary classification task where $r(p)$ is a Truncated Normal Distribution with various means. We choose means μ that roughly correspond to the model has wavering confidence, the model is not confident, and the model is very confident.

Example 1 (Binary Classification on the Unit Square) *This example takes place on a unit square in $\mathbb{R}^2$ with a uniform density $p(x, y) = 1$ if $0 \leq x \leq 1$ and $0 \leq y \leq 1$ and $p(x, y) = 0$ otherwise. The function h is invariant to translations in the $x-$direction. Let us now consider the unconstrained bound for three different test functions $r(p)$.*

Unconstrained Bound with Truncated Normal Density: *Recall that the Truncated Normal Distribution with mean μ, variance σ^2, and bounds (a, b) has a probability density*

$$f(x; \mu, \sigma, a, b) = \frac{1}{\sigma} \frac{\varphi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

for $a \leq x \leq b$ and $f(x) = 0$ otherwise and where

$$\varphi(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2}, \qquad \Phi(z) = \frac{1}{2}\left(1 + erf\left(\frac{z}{\sqrt{2}}\right)\right).$$

In all cases, we consider $\sigma = 0.1$ *and of course* $a = 0$, $b = 1$. *For* $\mu = 0.5$ *the upper bound is*

$$\frac{1}{2} + \int_0^1 \frac{1}{0.1} \frac{\varphi\left(\frac{p-0.5}{0.1}\right)}{\Phi(5) - \Phi(-5)} |0.5 - p| dp \approx 0.58.$$

Similarly, for $\mu = 0.25$ *or for* $\mu = 0.75$ *the upper bound is* ≈ 0.75.

Invariant Bound with Truncated Normal Density: *As seen in Figure 1, the orbit with the smallest integrated total dissent is the one on the* $x-$*axis,* $k(Gx^*) = 1 - r$. *Since this task is binary classification, we have* $P_2 = [0, 1]$ *and* $\int_{p_2} r(p)dp = 1$. *The upper bound on ECE, using the assumption of incorrect invariance, decreases by* $(1 - r)$.

Conclusion: *In the unconstrained case, the upper bound is tighter when the confidence is concentrated around* 50%, *which can be interpreted as the model "hedging its bet." The same is true in the invariant case, and the bound is always tightened according to accuracy of the best performing orbit, regardless of the distribution* $r(p)$.

The following example illustrates the utility of our bound in the special case when we know the fibers $\mathcal{F}_p$.

Example 2 (Reflection Invariance Upper Bound) *Consider the unit circle* S^1 *embedded in* $\mathbb{R}^2$. *Along* S^1 *we have* 20 *points that are each assigned either a blue or orange label. The cyclic group* C_2 *acts on elements of* S^1 *by reflecting them over the* $x-$*axis and trivially on the labels. As indicated in Figure 2, we have pointwise incorrect invariance on each half of the circle, though the model is still able to correctly classify* 90% *of its labels on right half . We assume that on the right half where* $x > 0$, *each of the* 5 *orbits map to the same confidence* p_1 *under* h_P *but may map to different labels under* h_Y. *Similarly, each of the* 5 *orbits where* $x < 0$ *map to the same confidence* p_2.

If $p_1 \neq p_2$, *or the confidence of the model* h *on the left half is not equal to its confidence on the right half, then we have two fibers to consider. This is indicated by the right half of the circle having diagonal lines in the first panel of Figure 2. On the left half, the error* $\int_{F_p} k(Gx, p)dx \geq 0.5$, *and on the right half,* $\int_{F_p} k(Gx, p)dx \geq 0.1$. *Taking the minimum, we find that* $m = 0.1$ *and ECE is bounded above by* $1 - 0.1 = 0.9$.

If instead $p_1 = p_2$, *or the confidence of the model is the same on each orbit, then we have one fiber to consider. In the second panel of Figure 2, this is indicated by the circle having no shaded regions. The error over the whole dataset is bounded from below by* $0.5(0.1 + 0.5) = 0.3$, *and ECE is bounded above by* $1 - 0.3 = 0.7$. *This concludes the first example.*

Having two fibers may be reasonable in a real world setting if there is a prevailing noise (e.g., shadow, camera artifacts, background patterns) on just one side of the field of view of a camera, leading to approximately two different confidence regions in our model output. This example serves as the ECE analogue to Figure 2 in Wang et al. (2023a).

Example 3 (Rotation Invariance Upper Bound) *The setup is as before with our dataset. However, now we consider rotation invariance instead of reflection invariance. In this case, we have only one orbit where we will predict one label and one confidence. Accordingly there*

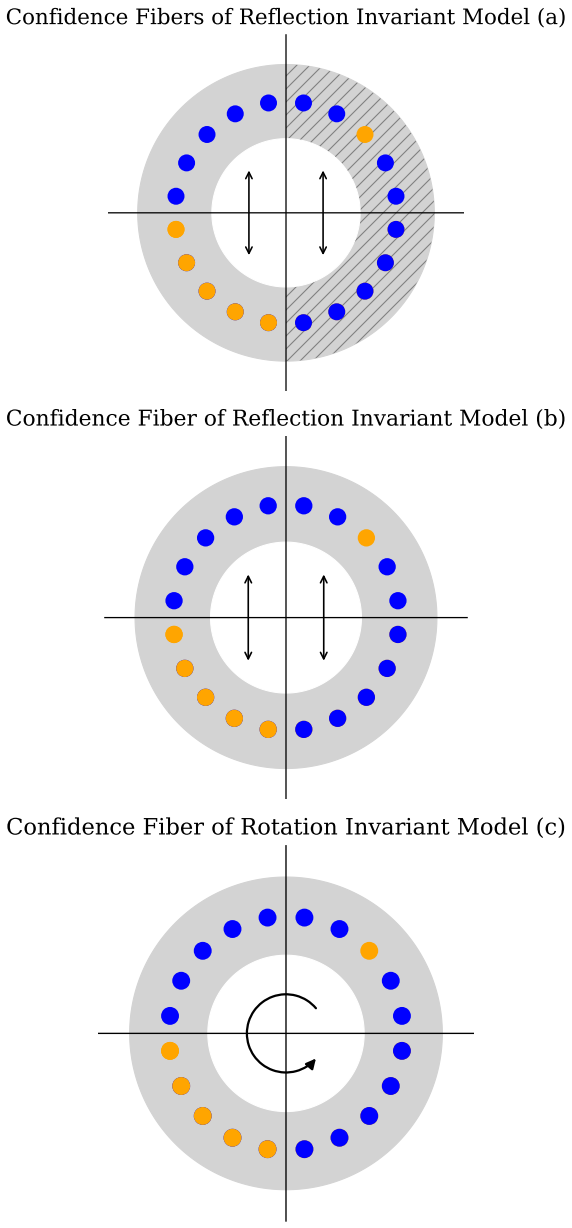

Figure 2: The first two panels indicate that the model is invariant C_2 reflections over the x-axis. The presence of diagonal lines in the first panel depicts that the model has 2 confidence fibers. The last panel indicates that the model has rotation invariance. Colors represent labels in all panels.

is only one fiber to consider, which is the entire dataset, as illustrated in the last panel of Figure 2. We minimize error when we predict each circle to be blue, since there are more blue labels then orange labels in our dataset. This gives us an error lower bound of 0.3*, and our ECE upper bound is* 0.7*. We see that this is the same result for reflection invariance in the special case where the confidences are the same across each fiber. This concludes the example.*

Appendix E. ECE Invariant Lower Bound

ECE Invariant Model Lower Bound. The assumption of invariance can also be used to obtain an ECE lower bound. In the naïve case, the lower bound is 0 because we can have that $\text{Acc}_p(h) = p$ for all p. However, if we have an accuracy lower bound m, then $\text{Acc}_p(h) \neq p$ for $p < m$. Let us now derive this accuracy lower bound.

To start, we define the minority label, the label that causes the maximal error on a given orbit Gx (analogous to the majority label that minimizes error on a given orbit in Wang et al. (2024)). We define the error on this orbit as the *minority label total dissent.*

Definition 2 (Minority Label Total Dissent) *For an orbit Gx of $x \in X$, the minority label total dissent $\kappa(Gx)$ is the integrated probability density of the elements in the orbit Gx having a different label than the minority label:*

$$\kappa(Gx) = \max_{y \in Y} \int_{Gx} q(z) \mathbb{1}(f(z) \neq y)) dz.$$

We prove in Proposition 14 that the total classification error is upper bounded by the integrated minority label total dissent.

Proposition 14 *$err_{cls}(h)$ is upper bounded by $\int_F \kappa(Gx)dx$.*

Accordingly, the accuracy is lower bounded by $1 - \int_F \kappa(Gx)dx$. We now use this accuracy lower bound derived from the minority label total dissent to lower bound ECE. We restate the theorem for convenience here.

Theorem 15 *We will denote the fundamental domain of G in $\mathcal{F}_p$ as F_p, where $\mathcal{F}_p$ is as defined in Proposition 4. As in Proposition 4, the total dissent on an orbit in a fiber $\mathcal{F}_p$ is denoted $\kappa(Gx, p)$ and is defined in terms of the renormalized density $q_p(x) = q(x)/\int_{\mathcal{F}_p} q(x)dx$. Define the minimum fiberwise classification accuracy as $m = \min_{p \in [0,1]} 1 - \int_{F_p} \kappa(Gx, p)dx$. ECE is lower bounded by $\int_0^m r(p)(m - p)dp$.*

Recall that for ECE to be well defined we assumed that $r(p)$ is supported everywhere on $[0, 1]$. Therefore, $r(p) \neq 0$ anywhere on $[0, m]$ and the lower bound is strictly greater than 0.

Again, it is ideal for our lower bound to be stated as using $q(x)$ instead of $r(p)$, since $q(x)$ is more likely to be known. We would also like to find an accuracy lower bound m' that does not depend on on the fibers $\mathcal{F}_p$, which implicitly depend on h_P. The following Propositions build up the necessary conditions needed to express the lower bound independently of h_P. We start by finding the accuracy lower bound m'.

Proposition 16 *Assume that h_P is a continuously differentiable function and that its gradient is nowhere 0. Define $x^* = \arg\min_{x \in X} \int_{z \in Gx} q(z)dz$ so that the orbit with the smallest integrated density is Gx^*. Let $p^* = \arg\min_{p \in [0,1]}(1 - \int_{F_p} \kappa(Gx, p)dx)$ and let $m' = 1 - \frac{1}{\int_{Gx^*} q(z)dz} \int_F \kappa(Gx)dx$. m is as defined in Theorem 15. ECE is lower bounded by*

$$\int_0^{m'} \int_{\mathcal{F}_p} \frac{1}{|\nabla h_P(x)|} q(x)dx_p(m' - p)dp.$$

This proof indicates that the accuracy lower bound on the entire dataset, inversely weighted by the integrated probability of the least likely orbit, is less than the accuracy lower bound on any given fiber. This allows us to derive an accuracy lower bound m' independent of the fibers $\mathcal{F}_p$.

As with the upper bound, the lower bound on ECE is related to how quickly the function h_P changes as a function of x. We can get a precise lower bound independent of $|\nabla h_P|$ if we have knowledge of the Lipschitz constant.

Proposition 17 *Assume h_P is differentiable, has gradient nowhere 0, and has Lipschitz constant of K. Gx^* and m' are the same as in Proposition 16. Then, ECE is lower bounded by*

$$\int_0^{m'} \int_{Gx^*} \frac{1}{K} q(x) dx (m' - p) dp.$$

Proof The ECE lower bound from Proposition 16 is minimized when $\frac{1}{|\nabla h_P(x)|}$ is minimized, so we are interested in when $|\nabla h_P(x)|$ is maximized. This upper bound is given by the Lipschitz constant K. The integral $\int_{\mathcal{F}_p} \frac{1}{K} q(x) dx_P = \int_{F_p} \int_{z \in Gx} q(z) \frac{1}{K} dz_p$. If Gx_p^* is the orbit with the smallest integrated probability density in F_p, then $\int_{F_p} \int_{z \in Gx} q(z) \frac{1}{K} dx_P \geq \int_{z \in Gx_p^*} q(z) \frac{1}{K} dx_P$. Now, we remove the dependence on p by considering the orbit with the smallest integrated probability density, including orbits not on F_p. This gives us $\int_{z \in Gx_p^*} q(z) \frac{1}{K} dz_p \geq \int_{Gx^*} q(x) \frac{1}{K} dx$. Therefore, our ECE lower bound becomes

$$\text{ECE}(h) \geq \int_0^{m'} \int_{z \in Gx^*} q(z) \frac{1}{K} dx (m' - p) dp.$$

■

With this in hand, we now give an example of a Lipschitz and invariant network where a precise lower bound independent of h_P can be obtained.

Example 4 (ECE lower bound for a $2-$layer $\mathbb{S}_n$ invariant network) *In this example, we consider a $2-$layer network that is permutation invariant for h. For this, we use a modified version of Deep Sets linear layers (Zaheer et al., 2017) of the form $W = \tanh(\lambda_1) I + \tanh(\lambda_2) 1 1^T$ where $\lambda_1, \lambda_2 \in \mathbb{R}$ are learnable parameters, I is the $m \times m$ identity matrix, and $1 = [1, \ldots, 1]^T$ is a vector in $\mathbb{R}^m$. The layer acts on $m \times n$ matrices, where m is the "set dimension." Following W, we apply the ReLU nonlinearity. These layers are permutation equivariant, meaning they satisfy $f([x_{\pi(1)}, \ldots, x_{\pi(M)}] = [f_{\pi(1)}(x), \ldots, f_{\pi(M)}(x)]$.*

To get strict invariance, we use a final readout layer of $\tanh(\lambda_3) 1^T$ where $\lambda_3 \in \mathbb{R}$ is also a learnable parameter. $\lambda_3 1^T$ performs mean pooling over the set dimension. All together,

$$h_P(x) = \tanh(\lambda_3) 1^T ReLU((\tanh(\lambda_1) I + \tanh(\lambda_2) 1 1^T)) x. \tag{3}$$

For Lipschitz functions f and g with Lipschitz constants L_1 and L_2, it is known that the Lipschitz constant for the composition $f(g(x))$ has Lipschitz constant $L_2 L_1$ and that the Lipschitz constant for the sum $f(x) + g(x)$ is the average Lipschitz constant $\frac{1}{2}(L_1 + L_2)$.

Moreover, the Lipschitz constant for a linear map is given by its maximum singular value $\sigma_{\max}$. *Finally, note that ReLU is* $1-$*Lipschitz. Thus, the Lipschitz constant for Equation 3 is* $\sigma_{\max}(1^T)\frac{1}{2}(\sigma_{\max}(I)+\sigma_{\max}(11^T))$ *which simplifies to* $\sigma_{\max}(1^T)(\frac{1}{2})(1+\sigma_{\max}(11^T))$.

Applying Proposition 17 the lower bound on ECE becomes

$$\int_0^{m'}\int_{Gx^*}\frac{1}{\sigma_{\max}(1^T)(\frac{1}{2})(1+\sigma_{\max}(11^T))}q(x)dx(m'-p)dp.$$

Now, the lower bound doesn't depend on the size of the fibers, and we have eliminated all dependence on $|\nabla h_P|$ *in the lower bound. This concludes the example.*

Appendix F. Proofs

Proof [Theorem **??**] Since ECE is the expectation of a random variable bounded between 0 and 1, ECE is bounded by 0 and 1. ■

Proof [Proposition 4]

We start by the revisiting the bound on ECE from Theorem **??**. Observe that

$$\begin{aligned}\left|\operatorname{Acc}_p(h)-p\right| &= \left|\operatorname{Acc}_p(h)-\frac{1}{2}+\frac{1}{2}-p\right| \\ &\leq \left|\operatorname{Acc}_p(h)-\frac{1}{2}\right|+\left|\frac{1}{2}-p\right|\end{aligned}$$

Integrating over $[0,1]$,

$$\begin{aligned}\int_{p=0}^{p=1} r(p)\left|\operatorname{Acc}_p(h)-p\right|dp &\leq \int_{p=0}^{p=1} r(p)\left(\left|\operatorname{Acc}_p(h)-\frac{1}{2}\right|+\left|\frac{1}{2}-p\right|\right)dp \\ &= \int_{p=0}^{p=1} r(p)\left(\left|\operatorname{Acc}_p(h)-\frac{1}{2}\right|\right)dp+\int_{p=0}^{p=1} r(p)\left|\frac{1}{2}-p\right|dp.\end{aligned}$$

Note P_1 and P_2 partition $[0,1]$. By definition of P_1 and P_2,

$$\int_{p=0}^{p=1} r(p)\left(\left|\operatorname{Acc}_p(h)-\frac{1}{2}\right|\right)dp = \int_{P_1} r(p)\left(\frac{1}{2}-\operatorname{Acc}_p(h)\right)dp+\int_{P_2} r(p)\left(\operatorname{Acc}_p(h)-\frac{1}{2}\right)dp \tag{4}$$

$$= \frac{1}{2}\left(\int_{P_1} r(p)dp-\int_{P_2} r(p)dp\right)-\int_{P_1} r(p)\operatorname{Acc}_p(h)dp+\int_{P_2} r(p)\operatorname{Acc}_p(h)dp. \tag{5}$$

By Theorem 2, the accuracy $\operatorname{Acc}_p(h)$ on any fiber of p is bounded above by $1-\int_{F_p}k(Gx,p)dx$. Combining this bound with the bounds defining P_1 and P_2 yields,

$$0<\operatorname{Acc}_p(h)<\min\left(1-\int_{F_p}k(Gx,p)dx,\frac{1}{2}\right)\qquad\forall p\in P_1.$$

$$\frac{1}{2}<\operatorname{Acc}_p(h)<\left(1-\int_{F_p}k(Gx,p)dx\right)\qquad\forall p\in P_2.$$

Observe that the upper bound for ECE is determined by the upper bound of Equations 4 and 5 and is related to the accuracy of h by the last two integrals in Equation 5. In particular, the model h that maximizes ECE satisfies $\text{Acc}_p(h) = 0$ on P_1 and $\text{Acc}_p(h) = 1 - \int_{F_p} k(Gx, p)dx$ on P_2. Substituting these values into Equation 5 gives upper bound

$$\left[\frac{1}{2}\left(\int_{P_1} r(p)dp - \int_{P_2} r(p)dp\right) + \int_{P_2} r(p)dp - \int_{P_2} r(p)\int_{F_p} k(Gx,p)dxdp\right] + \int_{p=0}^{p=1} r(p)|\frac{1}{2} - p|dp$$

which simplifies to

$$\frac{1}{2} + \int_{p=0}^{p=1} r(p)|\frac{1}{2} - p|dp - \int_{P_2} r(p)\int_{F_p} k(Gx,p)dxdp.$$

Finally,

$$\begin{aligned}
-\int_{P_2} r(p)\int_{F_p} k(Gx,p)dxdp &= -\int_{P_2} r(p)\int_{F_p} \min_{y\in Y}\int_{Gx} q_p(z)\mathbb{1}(f(z)\neq y)dzdxdp \\
&\leq -\int_{P_2} r(p)\int_{F_p} \min_{y\in Y}\int_{Gx} q(z)\mathbb{1}(f(z)\neq y)dzdxdp \\
&\leq -\int_{P_2} r(p)\min_{y\in Y}\int_{Gx^*} q(z)\mathbb{1}(f(z)\neq y)dzdp \\
&= -k(Gx^*)\int_{P_2} r(p)dp
\end{aligned}$$

and so ECE $\leq \frac{1}{2} + \int_0^1 r(p)|\frac{1}{2} - p|dp - k(Gx^*)\int_{P_2} r(p)dp$. This completes the proof. ■

Proof [Corollary 9] We compute $\frac{1}{2} + \int_0^1 r(p)|\frac{1}{2} - p|dp - \int_{p_2} r(p)\int_{F_p} k(Gx,p)dxdp \leq \frac{1}{2} + \int_0^1 r(p)|\frac{1}{2} - p|dp - m\int_{p_2} r(p)dp$. ■

Proof [Corollary 11] Note that $\int_0^1 r(p)|\frac{1}{2} - p|dp$ is bounded above by $\frac{1}{2}$. We have $\int_{P_2} r(p)dp = \int_0^1 r(p)dp = 1$ by assumption. Substituting these values into Proposition 4 and Corollary 9 completes the proof. ■

Proof [Proposition 12] See that $|\text{Acc}_p(h) - p| = |\text{Acc}_p(h) - p + \frac{1}{2} - \frac{1}{2}| \leq |\text{Acc}_p(h) - \frac{1}{2}| + |\frac{1}{2} - p| \leq |\frac{1}{2} - p| + \frac{1}{2}$. Therefore, $\int_0^1 r(p)(|\text{Acc}_p(h) - p|dp \leq \int_0^1 r(p)\left(|\frac{1}{2} - p| + \frac{1}{2}\right)dp = \frac{1}{2} + \int_0^1 r(p)|\frac{1}{2} - p|dp$. ■

Proof [Proposition 13] Substituting our expression for $r(p)$, we have that the upper bound on ECE in the incorrectly invariant case becomes $\frac{1}{2} + \int_0^1 \int_{\mathcal{F}_p} K_2 q(x)dx_p|\frac{1}{2} - p|dp - k(Gx^*)\int_{p_2}\int_{\mathcal{F}_p} K_2 q(x)dx_p dp$ and in the unconstrained case $\frac{1}{2} + \int_0^1 \int_{\mathcal{F}_p} K_2 q(x)dx_p|\frac{1}{2} - p|dp$. For the unconstrained case, note that

$$\begin{aligned}\frac{1}{2}+\int_0^1\int_{\mathcal{F}_p} K_2 q(x)dx_p|\frac{1}{2}-p|dp &\leq \frac{1}{2}+\int_0^1\int_{G\widetilde{x}} K_2 q(x)dx|\frac{1}{2}-p|dp\\ &= \frac{1}{2}+K_2\int_{G\widetilde{x}} q(x)dx\frac{1}{4}\\ &= \frac{2+K_2\int_{G\widetilde{x}} q(x)dx}{4}.\end{aligned}$$

For the invariant case, see that

$$-k(Gx^*)\int_{P_2}\int_{\mathcal{F}_p} K_2 q(x)dx_p dp \leq -k(Gx^*)\int_{P_2}\int_{G\widetilde{x}} K_2 q(x)dxdp \leq -k(Gx^*)K_2\int_{G\widetilde{x}} q(x)dx$$

so the upper bound becomes $\frac{2+K_2\int_{G\widetilde{x}} q(x)dx}{4} - k(Gx^*)K_2\int_{G\widetilde{x}} q(x)dx$.

■

Proof [Proposition 14] See that

$$\begin{aligned}\mathrm{err}_{\mathrm{cls}}(h) &= \int_X q(x)\mathbb{1}(f(x)\neq h(x))dx\\ &= \int_F\int_{Gx} q(z)\mathbb{1}(f(z)\neq h(z))dzdx\\ &\leq \int_F \max_{y\in Y}\int_{Gx} q(z)\mathbb{1}(f(z)\neq y))dx = \int_F \kappa(Gx)dx.\end{aligned}$$

■

Proof [Theorem 15] By Proposition 14, the classification accuracy on each fiber is lower bounded by $1-\int_{F_p}\kappa(Gx,p)dx$. $\mathrm{Acc}_p(h)$ is therefore lower bounded by m. See that $\int_0^1 r(p)|\mathrm{Acc}_p(h)-p|dp \geq \int_0^m r(p)|\mathrm{Acc}_p(h)-p|dp \geq \int_0^m r(p)(m-p)dp$ since $\mathrm{Acc}_p(h)\geq m > p$ when integrating p on $[0,m]$. ■

Proof [Proposition 16] As before, we have that $r(p)=\int_X q(x)\delta(p-h_P(x))dx = \int_{\mathcal{F}_p}\frac{1}{|\nabla h_P(x)|}q(x)dx_p$. As in Proposition 4, we note that the accuracy lower bound for each fiber must be computed in terms of the renormalized probabilities. See that

$$\begin{aligned}m=\min_{p\in[0,1]} 1-\int_{F_p}\kappa(Gx,p)dx = 1-\int_{F_p^*}\kappa(Gx,p)dx &= 1-\int_{F_p^*}\max_{y\in Y}\int_{Gx}\frac{q(z)}{\int_{\mathcal{F}_{p^*}}q(z)dz}\mathbb{1}(f(z)\neq y)dzdx\\ &\geq 1-\int_{F_p^*}\max_{y\in Y}\int_{Gx}\frac{q(z)}{\int_{Gx^*}q(z)dz}\mathbb{1}(f(z)\neq y)dzdx\\ &= 1-\frac{1}{\int_{Gx^*}q(z)dz}\int_{F_{p^*}}\max_{y\in Y}\int_{Gx} q(z)\mathbb{1}(f(z)\neq y)dzdx\\ &\geq 1-\frac{1}{\int_{Gx^*}q(z)dz}\int_F\max_{y\in Y}\int_{Gx} q(z)\mathbb{1}(f(z)\neq y)dzdx\\ &= 1-\frac{1}{\int_{Gx^*}q(z)dz}\int_F\kappa(Gx)dx = m'.\end{aligned}$$

So, ECE is lower bounded by $\int_0^{m'} \int_{\mathcal{F}_p} \frac{1}{|\nabla h_P(x)|} q(x) dx_p(m' - p) dp$.

■