Journal of Machine Learning Research 21 (2020) 1-71 Submitted 2/20; Published 9/20

A Group-Theoretic Framework for Data Augmentation

Shuxiao Chen SHUXIAOC@WHARTON.UPENN.EDU
Edgar Dobriban DOBRIBAN@WHARTON.UPENN.EDU
Department of Statistics
The Wharton School of the University of Pennsylvania
Philadelphia, PA, 19104-6340, USA

Jane H. Lee JANEHLEE@SAS.UPENN.EDU
Department of Mathematics, and Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104-6309, USA

Editor: Rob Fergus

Abstract

Data augmentation is a widely used trick when training deep neural networks: in addition to the original data, properly transformed data are also added to the training set. However, to the best of our knowledge, a clear mathematical framework to explain the performance benefits of data augmentation is not available. In this paper, we develop such a theoretical framework. We show data augmentation is equivalent to an averaging operation over the orbits of a certain group that keeps the data distribution approximately invariant. We prove that it leads to variance reduction. We study empirical risk minimization, and the examples of exponential families, linear regression, and certain two-layer neural networks. We also discuss how data augmentation could be used in problems with symmetry where other approaches are prevalent, such as in cryo-electron microscopy (cryo-EM).

Keywords: Data Augmentation, Deep Learning, Empirical Risk Minimization, Invariance, Variance Reduction

1. Introduction

Deep learning algorithms such as convolutional neural networks (CNNs) are successful in part because they exploit natural symmetry in the data. For instance, image identity is roughly invariant to translations and rotations: so a slightly translated cat is still a cat. Such invariances are present in many datasets, including image, text and speech data. Standard architectures are invariant to some, but not all transforms. For instance, CNNs induce an approximate equivariance to translations, but not to rotations. This is an *inductive bias* of CNNs, and the idea dates back at least to the neocognitron (Fukushima, 1980).

To make models invariant to arbitrary transforms beyond the ones built into the architecture, *data augmentation* is commonly used. Roughly speaking, the model is trained not just with the original data, but also with transformed data. Data augmentation is a crucial component of modern deep learning pipelines, and it is typically needed to achieve state of the art performance. It has been used, e.g., in AlexNet (Krizhevsky et al., 2012), and other pioneering works (Cireşan et al., 2010). The state-of-the-art results on aggregator websites

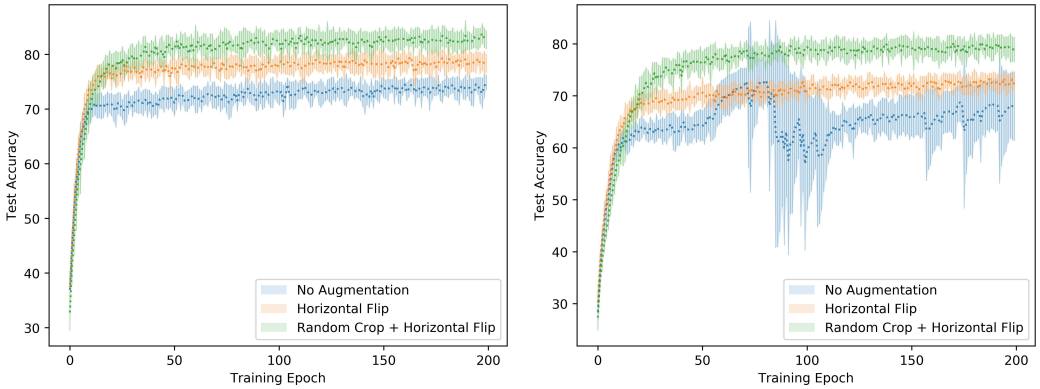

Figure 1: Benefits of data augmentation: A comparison of test accuracy across training epochs of ResNet18 (He et al., 2016) (1) without data augmentation, (2) horizontally flipping the image with 0.5 probability, and (3) a composition of randomly cropping a 32×32 portion of the image and random horizontal flip. The experiment is repeated 15 times, with the dotted lines showing the average test accuracy and the shaded regions representing 1 standard deviation around the mean. The left graph shows results from training on the full CIFAR10 training data and the right uses half of the training data as that of the left. Data augmentation leads to increased performance, especially with limited data.

such as `https://paperswithcode.com/sota` often crucially rely on effective new ways of data augmentation. See Figure 1 for a small experiment (see Appendix D for details).

Equivariant or invariant architectures (such as CNNs) are attractive for tackling invariance. However, in many cases, datasets have symmetries that are naturally described in a *generative* form: we can specify generators of the group of symmetries (e.g., rotations and scalings). In contrast, computing the equivariant features requires designing new architectures. Thus, data augmentation is a universally applicable, generative, and algorithmic way to exploit invariances.

However, a general framework for understanding data augmentation is missing. Such a framework would enable us to reason clearly about the benefits offered by augmentation, in comparison to invariant features. Moreover, such a framework could also shed light on questions such as: How can we improve the performance of our models by simply adding transformed versions of the training data? Under what conditions can we see benefits? Developing such a framework is challenging for several reasons: first, it is unclear what mathematical approach to use, and second, it is unclear how to demonstrate that data augmentation "helps".

In this paper, we propose such a general framework. We use group theory as a mathematical language, and model invariances as "approximate equality" in distribution under a group action. We show that data augmentation can be viewed as invariant learning by averaging over the group action. We then demonstrate that data augmentation leads to sample efficient learning, both in the non-asymptotic setting (relying on results from stochastic

convex optimization and Rademacher complexity), as well as in the asymptotic setting (by using asymptotic statistical theory for empirical risk minimizers/M-estimators).

We show how to apply data augmentation beyond deep learning, to other problems in statistics and machine learning that have invariances. In addition, we explain the connections to several other important notions from statistics and machine learning, including sufficiency, invariant representations, equivariance, variance reduction in Monte Carlo methods, and regularization.

We can summarize our main contributions as follows:

1. We study data augmentation in a group-theoretic formulation, where there is a group acting on the data, and the distribution of the data is equal (which we refer to as *exact invariance*), or does not change too much (which we refer to as *approximate invariance*) under the action. We explain that in empirical risk minimization (ERM), this leads to minimizing an augmented loss, which is the average of the original loss under the group action (Section 3.1). In the special case of maximum likelihood estimation (MLE), we discuss several variants of MLE that may potentially exploit invariance (Section 3.2). We also propose to extend data augmentation beyond ERM, using the "augmentation distribution" (Section 3.3).

2. We provide several theoretical results to support the benefits of data augmentation. When the data is exactly invariant in distribution (exact invariance), we show that averaging over the group orbit (e.g., all rotations) reduces the variance of *any function*. We can immediately conclude that estimators based on the "augmentation distribution" gain efficiency and augmentation reduces the mean squared error (MSE) of general estimators (Section 4.1).

3. Specializing the variance reduction to the loss and the gradients of the loss, we show that the empirical risk minimizer with data augmentation enjoys favorable properties in a non-asymptotic setup. Specifically, "loss-averaging" implies that data augmentation reduces the Rademacher complexity of a loss class (Section 4.2.1), which further suggests that the augmented model may generalize better. On the other hand, we show that "gradient-averaging" reduces the variance of the ERM when the loss is strongly-convex, using a recent result from stochastic convex optimization (Section 4.2.2).

4. Moving to the asymptotic case, we characterize the *precise* variance reduction obtained by data augmentation under exact invariance. We show that this depends on the covariance of the loss gradients along the group orbit (Section 4.2.3). This implies that data augmentation can improve upon the Fisher information of the un-augmented maximum likelihood estimator (MLE) (Section 4.2.4). For MLE, we further study the special case when the subspace of parameters with invariance is a low-dimensional manifold. We connect this to geometry showing that the projection of the gradient onto the tangent space is always invariant; however, it does not always capture all invariance. As a result, the augmented MLE cannot always be as efficient as the "constrained MLE", where the invariance is achieved by constrained optimization.

5. We work out several examples of our theory under exact invariance: exponential families (Section 5.1), least-squares regression (Section 5.2) and least squares classification

(Section 5.3). As a notable example, we work out the efficiency gain for two-layer neural networks with circular shift data augmentation in the heavily underparametrized regime (where most of our results concern quadratic activations). We also provide an example of "augmentation distribution", in the context of linear regression with general linear group actions (Section 5.4).

6. We extend most of the results to the approximate invariance case, where the distribution of the data is close, but not exactly equal to its transformed copy (Section 6). Using optimal transport theory, we characterize an intriguing bias-variance tradeoff: while the orbit averaging operation reduces variance, a certain level of bias is created because of the non-exact-invariance. This tradeoff suggests that in practice where exact invariance does not hold, data augmentation may not always be beneficial, and its performance is governed by both the variability of the group and a specific Wasserstein distance between the data and its transformed copy.

7. We illustrate the bias-variance tradeoff by studying the generalization error of overparameterized two-layer networks trained by gradient descent (Section 7), using recent results on Neural Tangent Kernel (see, e.g., Jacot et al. 2018; Arora et al. 2019; Ji and Telgarsky 2019; Chen et al. 2019).

8. We also describe a few important problems where symmetries occur, but where other approaches—not data augmentation—are currently used (Section 8): cryo-electron microscopy (cryo-EM), spherically invariant data, and random effects models. These problems may be especially promising for using data augmentation.

2. Some related work

In this section, we discuss some related works, in addition to those that are mentioned in other places.

Data augmentation methodology in deep learning. There is a great deal of work in developing efficient methods for data augmentation in deep learning. Here we briefly mention a few works. Data augmentation has a long history, and related ideas date back at least to Baird (1992), who built a "pseudo-random image generator", that "*given an image and a set of model parameters, computes a degraded image*". This is recounted in Schmidhuber (2015).

Conditional generative adversarial networks (cGAN) are a method for learning to generate from the class-conditional distributions in a classification problem (Mirza and Osindero, 2014). This has direct applications to data augmentation. Data augmentation GANs (DAGAN) (Antoniou et al., 2017) are a related approach that train a GAN to discriminate between x, x_g, and x, x', where x_g is generated as $x_g = f(x, z)$, and x' is an independent copy. This learns the conditional distribution $x'|x$, where x', x are sampled "independently" from the training set. Here the training data is viewed as non-independent and they learn the dependence structure.

Hauberg et al. (2016) construct class-dependent distributions over diffeomorphisms for learned data augmentation. Ratner et al. (2017) learn data augmentation policies using

reinforcement learning, starting with a known set of valid transforms. Tran et al. (2017) propose a Bayesian approach. RenderGAN (Sixt et al., 2018) combines a 3D model with GANs for image generation. DeVries and Taylor (2017a) propose to perform data augmentation in feature space. AutoAugment (Cubuk et al., 2018) is another approach for learning augmentation policies based on reinforcement learning, which is one of the state of the art approaches. RandAugment (Cubuk et al., 2019) proposes a strategy to reduce the search space of augmentation policies, which also achieves state of the art performance on ImageNet classification tasks. Hoffer et al. (2019) argues that replicating instances of samples within the same batch with different data augmentations can act as an accelerator and a regularizer when appropriately tuning the hyperparameters, increasing both the speed of training and the generalization performance.

Neural net architecture design. There is a parallel line of work designing invariant and equivariant neural net architectures. A key celebrated example is convolutions, dating back at least to the neocognitron (Fukushima, 1980), see also LeCun et al. (1989). More recently, group equivariant Convolutional Neural Networks (G-CNNs), have been proposed, using G-convolutions to exploit symmetry (Cohen and Welling, 2016a). That work designs concrete architectures for groups of translations, rotations by 90 degrees around any center of rotation in a square grid, and reflections. Dieleman et al. (2016) designed architectures for cyclic symmetry.

Worrall et al. (2017) introduces Harmonic Networks or H-Nets, which induce equivariance to patch-wise translation and 360 degree rotation. They rely on circular harmonics as invariant features. Cohen and Welling (2016b) propose steerable CNNs and Cohen et al. (2018a) develop a more general approach. There are several works on SO(3) equivariance, see Cohen et al. (2018b); Esteves et al. (2018a,b, 2019).

Gens and Domingos (2014) introduces deep symmetry networks (symnets), that form feature maps over arbitrary symmetry groups via kernel-based interpolation to pool over symmetry spaces. See also Ravanbakhsh et al. (2017); Kondor and Trivedi (2018); Weiler et al. (2018); Kondor et al. (2018). There are also many examples of data augmentation methods developed in various application areas, e.g., Jaitly and Hinton (2013); Xie et al. (2019); Park et al. (2019); Ho et al. (2019), etc.

Data augmentation as a form of regularization. There is also a line of work proposing to add random or adversarial noise to the data when training neural networks. The heuristic behind this approach is that the addition of noise-perturbed data should produce a classifier that is robust to random or adversarial corruption.

For example, DeVries and Taylor (2017b) proposes to randomly mask out square regions of input images and fill the regions with pure gray color; Zhong et al. (2017) and Lopes et al. (2019) propose to randomly select a patch within an image and replace its pixels with random values; Bendory et al. (2018) proposes to add Perlin noise (Perlin, 1985) to medical images; Zhang et al. (2017) proposes to train with convex combinations of two images *as well as their labels*. The experiments done by those papers show that augmenting with noise-perturbed data can lead to lower generalization error and better robustness against corruption.

Szegedy et al. (2013) and Cohen et al. (2019) demonstrate that training with adversarial examples can lead to some form of regularization. Hernández-García and König (2018a) has argued that data augmentation can sometimes even replace other regularization mechanisms such as weight decay, while Hernández-García and König (2018b) has argued that this can be less sensitive to hyperparameter choices than other forms of regularization. While this approach is also called data augmentation in the literature, it is *fundamentally different* from what we consider. We study a way to exploit invariance in the data, while those works focus on smoothing effects (adding noise cannot possibly lead to exactly invariant distributions).

More related to our approach, Maaten et al. (2013) propose to train the model by minimizing the expected value of the loss function under the "corrupting distribution" and show such a strategy improves generalization. Chao et al. (2017); Mazaheri et al. (2019) aim to reduce the variance and enhance stability by "functional integration" inspired averaging over a large number of different representations of the same data. However, our quantitative approaches differ. Lyle et al. (2019) study the effect of invariance on generalization in neural networks. They study feature averaging, which agrees with our augmentation distribution defined later. They mention variance reduction, but we can, in fact, prove it.

Other works connected to data augmentation. There is a tremendous amount of other work connected to data augmentation. On the empirical end, Bengio et al. (2011) has argued that the benefit of data augmentation goes up with depth. On the theoretical end, Rajput et al. (2019) investigate if gaussian data augmentation leads to a positive margin, with some negative results. Connecting to adversarial examples, Engstrom et al. (2017) shows that adversarially chosen group transforms such as rotations can already be enough to fool neural network classifiers. Javadi et al. (2019) shows that data augmentation can reduce a certain Hessian-based complexity of neural networks. Liu et al. (2019) show that data augmentation can significantly improve the optimization landscape of neural networks, so that SGD avoids bad local minima and leads to much more accurate trained networks. Hernández-García et al. (2018) has shown that it also leads to better biological plausibility in some cases.

Dao et al. (2019) also seek to establish a theoretical framework for understanding data augmentation, but focus on the connection between data augmentation and kernel classifiers. Dao et al. (2019) study k-NN and kernel methods. They show how data augmentation with a kernel classifier yields approximations which look like feature averaging and variance regularization, but do not explicitly quantify how this *improves* classification. One of the most related works by Bloem-Reddy and Teh (2019) use similar probabilistic models, but without focusing on data augmentation.

We also note that data augmentation has another meaning in Bayesian statistics, namely the introduction of auxiliary variables to help compute the posterior (see e.g., Tanner and Wong, 1987). The naming clash is unfortunate. However, since the term "data augmentation" is well established in deep learning, we decided to keep it in our current work.

Group invariance in statistical inference. There has been significant work on group invariance in statistical inference (e.g., Giri, 1996). However, the questions investigated there are different from the ones that we study. Among other contributions, Helland

(2004) argues that group invariance can form natural non-informative priors for the parameters of the model, and introduces permissible sub-parameters as those upon which group actions can be defined.

Physical invariances. There is a long history of studying invariance and symmetry in physics. Invariances lead to conservation laws, such as conservation of mass and momentum. In addition, invariances in Hamiltonians of physical systems lead to reductions in the number of parameters of probability distributions governing the systems. This has been among the explanations proposed of why deep learning works (Lin et al., 2017).

3. Methodology for data augmentation

3.1 ERM

We start by explaining our framework in the context of empirical risk minimization (ERM).

Consider observations $X_1, \ldots, X_n \in \mathcal{X}$ (e.g., images along with their labels) sampled i.i.d. from a probability distribution $\mathbb{P}$ on the sample space $\mathcal{X}$. Since data augmentation is a way of "teaching invariance to the model", we need to assume our data is invariant to certain transformations. Consider thus a group G of transforms (e.g., the set of all rotations of images), which *acts* on the sample space: there is a function $\phi : G \times \mathcal{X} \to \mathcal{X}, (g, x) \mapsto \phi(g, x)$, such that $\phi(e, x) = x$ for the identity element $e \in G$, and $\phi(gh, x) = \phi(g, \phi(h, x))$ for any $g, h \in G$. For notational simplicity, we write $\phi(g, x) \equiv gx$ then there is no ambiguity. To model invariance, we assume that for any group element $g \in G$ and almost any $X \sim \mathbb{P}$, we have an "approximate equality" in distribution:

$$X \approx_d gX. \tag{1}$$

For ease of exposition, we start with *exact invariance* $X =_d gX$ in Section 4, and we handle *approximate invariance* in Section 6.

For supervised learning applications, $X_i = (Z_i, Y_i)$ contains both the features Z_i and the outcome Y_i. The assumption (1) means that the probability of an image being a bird is (either exactly or approximately) the same as the probability for a rotated image.

As an aside, the invariance in distribution (1) arises naturally in many familiar statistical settings, including permutation tests, time series analysis. See Section 3.4.

In the current context, data augmentation corresponds to "adding all datapoints gX_i, $g \in G$, $i = 1, \ldots, n$" to the dataset. When the group is finite, this can effectively be implemented by enlarging the data size. However, many important groups are infinite, and to understand data augmentation in that setting, it is most clear if we argue from first principles.

In practice, data augmentation is performed via the following approach. To start, we consider a loss function $L(\theta, X)$, and attempt to minimize the empirical risk

$$R_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} L(\theta, X_i). \tag{2}$$

We iterative over time $t = 1, 2, \ldots$ using stochastic gradient descent (SGD) or variants (see Algorithm 1). At each step t, a minibatch of X_i's (say with indices S_t) is chosen. In data

Algorithm 1: Augmented SGD

Input : Data X_i, $i = 1, \dots, n$; Method to compute gradients $\nabla L(\theta, X)$ of the loss; Method to sample augmentations $g \in G$, $g \sim \mathbb{Q}$; Learning rates η_t; Batch sizes $|S_t|$; Initial parameters θ_0; Stopping criterion.

Output: Final parameters.

Set $t = 0$

While stopping criterion is not met

Sample random minibatch $S_t \subset \{1, \dots, n\}$

Sample random augmentation $g_{i,t} \sim \mathbb{Q}$ for each batch element

Update parameters

$$\theta_{t+1} \leftarrow \theta_t - \frac{\eta_t}{|S_t|} \sum_{i \in S_t} \nabla L(\theta, g_{i,t} X_i)$$

$t \leftarrow t + 1$

return θ

augmentation, a random transform $g_{i,t} \in G$ is sampled and applied to each data point in the minibatch. Then, the parameter is updated as

$$\theta_{t+1} = \theta_t - \frac{\eta_t}{|S_t|} \sum_{i \in S_t} \nabla L(\theta, g_{i,t} X_i). \tag{3}$$

Here we need a probability distribution $\mathbb{Q}$ on the group G, from which $g_{i,t}$ is sampled. For a finite G, one usually takes $\mathbb{Q}$ to be the uniform distribution. However, care must be taken if G is infinite. We assume G is a compact topological group, and we take $\mathbb{Q}$ to be the *Haar probability measure*.[1] Hence, for any $g \in G$ and measurable $S \subseteq G$, the following *translation invariant* property holds: $\mathbb{Q}(gS) = \mathbb{Q}(S), \mathbb{Q}(Sg) = \mathbb{Q}(S)$.

A key observation is that the update rule (3) corresponds to SGD on an *augmented* empirical risk, where we take an average over all augmentations according to the measure $\mathbb{Q}$:

$$\min_\theta \bar{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \int_G L(\theta, gX_i) d\mathbb{Q}(g). \tag{4}$$

To be precise, $\nabla L(\theta, g_{i,t} X_i)$ is an unbiased stochastic gradient for the *augmented* loss function

$$\bar{L}(\theta, X) := \int_G L(\theta, gX) d\mathbb{Q}(g), \tag{5}$$

and hence we can view the resulting estimator as an empirical risk minimizer of $\bar{R}_n = n^{-1} \sum_{i=1}^{n} \bar{L}(\theta, X_i)$.

This can be viewed as Rao-Blackwellizing the loss, meaning taking a conditional expectation of the loss over the conditional distribution of x belonging to a certain group orbit. See Lemma 1 for a precise statement and Section 3.4 for more discussion.

1. Haar measures are used for convenience. Most of our results hold for more general measures with slightly more lengthy proofs.

3.2 Variants of MLE under invariance

The maximum likelihood estimation (MLE) is a special case of ERM when there is an underlying parametric model. The usual approach goes as follows. One starts with a parametric statistical model (i.e., a collection of probability measures) $\{\mathbb{P}_\theta : \theta \in \Theta\}$, where Θ is some parameter space. Let p_θ be the density of $\mathbb{P}_\theta$, and we define the log-likelihood function as $\ell_\theta(x) = \log p_\theta(x)$. We then define the MLE as any solution of the following problem:

$$\widehat{\theta}_{\mathrm{MLE},n} \in \arg\max_{\theta\in\Theta} \frac{1}{n} \sum_{i\in[n]} \ell_\theta(X_i). \tag{6}$$

The invariance assumption (1) is only imposed on the "true" parameter θ_0, so that for $\mathbb{P}_{\theta_0}$-a.e. x and $\mathbb{Q}$-a.e. g, we have

$$p_{\theta_0}(gx) \cdot |\det Jac(x \to gx)| = p_{\theta_0}(x), \tag{7}$$

where $Jac(x \to gx)$ is the Jacobian of the transform $x \mapsto gx$.

There are multiple ways to adapt the likelihood function to the invariance structure (with data augmentation being one of them), and we detail some of them below.

Constrained MLE. The invariance structure is a constraint on the density function. Hence we obtain a natural *constrained* (or restricted) maximum likelihood estimation problem. Define the *invariant subspace* as

$$\Theta_G = \{\theta \in \Theta : X =_d gX, \forall g \in G\}. \tag{8}$$

Then the constrained MLE is

$$\widehat{\theta}_{\mathrm{cMLE},n} \in \arg\max_{\theta\in\Theta_G} \sum_{i\in[n]} \ell_\theta(X_i). \tag{9}$$

In general, this can be much more sample-efficient than the original MLE. For instance, suppose we are trying to estimate a normal mean based on one sample: $X \sim \mathcal{N}(\theta, 1)$. Let the group be negation, i.e., $G = (\{\pm 1\}, \cdot) = \mathbb{Z}_2$. Then $\Theta_G = \{0\}$, because the only normal density symmetric around zero is the one with zero mean. Hence, the invariance condition uniquely identifies the parameter, showing that the constrained MLE perfectly recovers the parameter.

However, in general, optimizing over the restricted parameter set may be computationally more difficult. This is indeed the case in the applications we have in mind. For instance, in deep learning, G may correspond to the set of all translations, rotations, scalings, shearings, color shifts, etc. And it's not clear how to obtain information on Θ_G (for example, how to compute the projection operator onto Θ_G).

Augmented MLE. A particular example of the augmented ERM is augmented maximum likelihood estimation. Here the loss is the negative log-likelihood, $L(\theta, x) = -\ell_\theta(x)$. Then the augmented ERM program (4) becomes

$$\widehat{\theta}_{\mathrm{aMLE},n} \in \arg\max_{\theta} \sum_{i\in[n]} \int_G \ell_\theta(gX_i) d\mathbb{Q}(g). \tag{10}$$

Table 1: Optimization objectives

Method	Sample Objective	Population Objective
ERM/MLE	$\min_{\theta\in\Theta}\frac{1}{n}\sum_{i\in[n]}L(\theta,X_i)$	$\min_{\theta\in\Theta}\mathbb{E}_{\theta_0}L(\theta,X)$
Constrained ERM/MLE	$\min_{\theta\in\Theta_G}\frac{1}{n}\sum_{i\in[n]}L(\theta,X_i)$	$\min_{\theta\in\Theta_G}\mathbb{E}_{\theta_0}L(\theta,X)$
Augmented ERM/MLE	$\min_{\theta\in\Theta}\frac{1}{n}\sum_{i\in[n]}\int L(\theta,gX_i)d\mathbb{Q}(g)$	$\min_{\theta\in\Theta}\mathbb{E}_{\theta_0}\int L(\theta,gX)d\mathbb{Q}(g)$
Invariant ERM/MLE	$\min_{\theta\in\Theta}\frac{1}{n}\sum_{i\in[n]}L(\theta,T(X_i)),\quad T(x)=T(gx)$	$\min_{\theta\in\Theta}\mathbb{E}_{\theta_0}L(\theta,T(X))$
Marginal MLE	$\max_{\theta\in\Theta}\frac{1}{n}\sum_{i\in[n]}\log\left(\int p_\theta(gX_i)d\mathbb{Q}(g)\right)$	$\max_{\theta\in\Theta}\mathbb{E}_{\theta_0}\log\left(\int p_\theta(gX)d\mathbb{Q}(g)\right)$

Invariant MLE. Another perspective to exploit invariance is that of invariant representations, i.e., learning over representations $T(x)$ of the data such that $T(gx)=T(x)$ (see also Section 3.4). However, it turns out that in some natural examples, the invariant MLE does not gain over the usual MLE (see Appendix C for a specific example). Thus we will not consider this in much detail.

Marginal MLE. There is a natural additional method to estimate the parameters, the marginal MLE. Under exact invariance, our original local invariance assumption (7) is equivalent to the following *latent variable* model. Under the true parameter θ_0, we sample a random group element $g\sim\mathbb{Q}$, and a random datapoint $\tilde{X}\sim\mathbb{P}_{\theta_0}$, i.e.,

$$\tilde{X}\sim\mathbb{P}_{\theta_0},\ g\sim\mathbb{Q}. \tag{11}$$

Then, we observe $X=g\tilde{X}$. We repeat this independently over all datapoints to obtain all X_i. Since $gX=_d X$ under θ_0, this sampling process is exactly equivalent to the original model, under θ_0.

Suppose that instead of fitting the model (11), we attempt to fit the relaxed model $\tilde{X}\sim\mathbb{P}_\theta$, $g\sim\mathbb{Q}$, observing $X=g\tilde{X}$. The only change is that we assume that the invariance holds for all parameter values. This model is mis-specified, nonetheless, it may be easier to fit computationally. Moreover, its MLE may still retain consistency for the original true parameter. Now the maximum marginal likelihood estimator (ignoring terms constant with respect to θ) can be written as:

$$\widehat{\theta}_{\mathrm{mMLE},n}=\arg\max_\theta\sum_{i\in[n]}\log\left(\int_G p_\theta(gX_i)d\mathbb{Q}(g)\right). \tag{12}$$

We emphasize that this is not the same as the augmented MLE estimator considered above. This estimator has the $\log(\cdot)$ terms outside the G-integral, while the augmented one effectively has the $\log(\cdot)$ terms inside.

Summary of methods. Thus we now have several estimators for the original problem: MLE, constrained MLE, augmented MLE, invariant MLE, and marginal MLE. We note that the former four methods are general in the ERM context, whereas the last one is specific

to likelihood-based models. See Table 1 for a summary. *Can we understand them?* In the next few sections, we will develop theoretical results to address this question.

3.3 Beyond ERM

The above ideas apply to empirical risk minimization. However, there are many popular algorithms and methods that are not most naturally expressed as plain ERM. For instance:

1. Regularized ERM, e.g., Ridge regression and Lasso
2. Shrinkage estimators, e.g., Stein shrinkage
3. Nearest neighbors, e.g., k-NN classification and regression
4. Heuristic algorithms like Forward stepwise (stagewise, streamwise) regression, and backward stepwise. While these may be associated with an objective function, there may be no known computationally efficient methods for finding global solutions.

To apply data augmentation for those methods, let us consider a general estimator $\widehat{\theta}(x)$ based on data x. The simplest idea would be to try to compute the estimator on *all the data*, including the actual and transformed sets. Following the previous logic, if we have the invariance $X \approx_d gX$ for all $g \in G$, then after observing data x, we should "augment" our data with gx, for all $g \in G$. Finally, we should run our method on this data. However, this idea can be impractical, as the entire data can be too large to work with directly. Therefore, we will take a more principled approach and work through the logic step by step, considering all possibilities. We will eventually recover the above estimator as well.

Augmentation distribution & General augmentations. We define the *augmentation distribution* as the set of values

$$\widehat{\theta}(gx),\ g \in G. \tag{13}$$

We think of $x = (x_1, \ldots, x_n)$ as the entire dataset, and of $g \in G$ being a group element summarizing the action on every data point. The augmentation distribution is simply the collection of values of the estimator we would get if we were to apply all group transforms to the data. It has a special role, because we think of each transform as equally informative, and thus each value of the augmentation distribution is an equally valid estimator.

We can also make (13) a proper probability distribution by taking a random $g \sim \mathbb{Q}$. Then we can construct a final estimator by computing a summary statistic on this distribution, for instance, the mean

$$\widehat{\theta}_G(x) = \mathbb{E}_{g\sim\mathbb{Q}}\widehat{\theta}(gx). \tag{14}$$

It is worth noticing that this summary statistic is exactly invariant, so that $\widehat{\theta}_G(x) = \widehat{\theta}_G(gx)$. Moreover, this estimator can be approximated in the natural way in practice via sampling $g_i \sim \mathbb{Q}$ independently:

$$\widehat{\theta}_k(x) = \frac{1}{k}\sum_{i=1}^{k}\widehat{\theta}(g_i x). \tag{15}$$

Connection to previous approach. To see how this connects to the previous ideas, let us consider $X = (X_1, \ldots, X_n)$, and let $\widehat{\theta}$ be the ERM with loss function $L(\theta, \cdot)$. Consider the group $G^n = G \times \ldots \times G$ acting on X elementwise. Then, the augmentation distribution is the set of values

$$\widehat{\theta}(X_1, \ldots, X_n; g_1, \ldots, g_n) := \arg\min_\theta \frac{1}{n} \sum_{i \in [n]} L(\theta, g_i X_i),$$

where we range $g \in G$. Then, the final estimator, according to (14), would be

$$\widehat{\theta}_G(X_1, \ldots, X_n) := \mathbb{E}_{g_1, \ldots, g_n \sim Q} \arg\min_\theta \frac{1}{n} \sum_{i \in [n]} L(\theta, g_i X_i).$$

Compared to the previous augmented ERM estimator, the current one changes the order of averaging over G and minimization. Specifically, the previous one is $\arg\min \mathbb{E}_g R_n(\theta; gX)$, while the current one is $\mathbb{E}_g \arg\min R_n(\theta; gX)$. If we know that the estimator is obtained from minimizing a loss function, then we can average that loss function; but in general we can only average the estimator, which justifies the current approach.

The two approaches above are closer than one may think. We can view the SGD iteration (3) as an online optimization approach to minimize a randomized objective of the form $\sum_{i \in [n]} L(\theta, g_i X_i)$. This holds exactly if we take one pass over the data in a deterministic way, which is known as a type of incremental gradient method. In this case, minimizing the augmented ERM has a resemblance to minimizing the mean of the augmentation distribution. However, in practice, people take multiple passes over the data, so this interpretation is not exact.

Augmentation in sequence of estimators. In the above generalization of data augmentation beyond ERM, we assumed only the bare minimum, meaning that the estimator $\widehat{\theta}(x)$ exists. Suppose now that we have slightly more structure, and the estimator is part of a sequence of estimators $\widehat{\theta}_n$, defined for all n. This is a mild assumption, as in general estimators can be embedded into sequences defined for all data sizes. Then we can directly augment our dataset $X_1, \ldots, X_n$ by adding new datapoints. We can define augmented estimators in several ways. For instance, for any fixed m, we can compute the estimator on a uniformly resampled set of size m from the data, applying uniform random transforms:

$$\begin{aligned} &\widehat{\theta}_m(g_1 X_{i_1}, g_2 X_{i_2}, \ldots, , g_m X_{i_m}) \\ &i_k \sim \text{Unif}([n]),\ g_k \sim \mathbb{Q}. \end{aligned}$$

This implicitly assumes a form of symmetry of the estimator with respect to its arguments. There are many variations: e.g., we may insist that m should be a multiple of n, or we can include all datapoints. This leads us to a "completely augmented" estimator which includes all data and all transforms (assuming $|G|$ is finite)

$$\widehat{\theta}_{n|G|}(\{g_j X_i\}_{i \in [n], j \in [|G|]}).$$

The advantage of the above reasoning is that it allows us to design augmented/invariant learning procedures extremely generally.

3.4 Connections and perspectives

Our approach has connections to many important and well-known concepts in statistics and machine learning. Here we elaborate on those connections, which should help deepen our understanding of the problems under study.

Sufficiency. The notion of sufficiency, due to Ronald A Fisher, is a fundamental concept in statistics. Given an observation (datapoint) X from a statistical model $X \sim \mathbb{P}$, a statistic $T := T(X)$ is said to be sufficient for a parameter $\theta := \theta(\mathbb{P})$ if the conditional distribution of $X|T = t$ does not depend on θ for almost any t. Effectively, we can reduce the data X to the statistic T without any loss of information about θ. A statistic T is said to be minimal sufficient if any other sufficient statistic is a function of it.

In our setup, assuming the invariance $X =_d gX$, on the invariant subspace Θ_G, the *orbits* $G \cdot x := \{gx | g \in G\}$ are minimal sufficient for θ. More generally, the local invariance condition (where invariance only holds for a subset of the parameter space) implies that the group orbits are a *locally sufficient* statistic for our model. From the perspective of statistical theory, this suggests that we should work with the orbits. However, this can only be practical under the following conditions:

1. We can conveniently compute the orbits, or we can conveniently find representatives;
2. We can compute the distribution induced by the model on the orbits;
3. We can compute estimators/learning rules defined on the orbits in a convenient way.

This is possible in many cases (Lehmann and Casella, 1998; Lehmann and Romano, 2005), but in complex cases such as deep learning, some or all of these steps can be impractical. For instance, the set of transforms may include translations, rotations, scalings, shearings, color shifts, etc. How can we compute the orbit of an image? It appears that an explicit description would be hard to find.

Invariant representations. The notions of invariant representations and features are closely connected to our approach. Given an observation x, and a group of transforms G acting on $\mathcal{X}$, a feature $F : \mathcal{X} \to \mathcal{Y}$ is invariant if $F(x) = F(gx)$ for all x, g. This definition does not require a probabilistic model. By design, convolutional filters are trained to look for spatially localized features, such as edges, whose pixel-wise representation is invariant to location. In our setup, we have a group acting on the data. In that case, it is again easy to see that the orbits $G \cdot x := \{gx | g \in G\}$ are the maximal invariant representations.

Related work by Mallat, Bölcskei and others (e.g., Mallat, 2012; Bruna and Mallat, 2013; Wiatowski and Bölcskei, 2018; Anselmi et al., 2019) tries to explain how CNNs extract features, using ideas from harmonic analysis. They show that the features of certain models of neural networks (Deep Scattering Networks for Mallat) are increasingly invariant with respect to depth.

Equivariance. The notion of equivariance is also key in statistics (e.g., Lehmann and Casella, 1998). A statistical model is called equivariant with respect to a group G acting on the sample space if there is an *induced group* G^* acting on the parameter space Θ such

that for any $X \sim \mathbb{P}_\theta$, and any $g \in G$, there is a $g^* \in G^*$ such that $gX \sim \mathbb{P}_{g^*\theta}$. Under equivariance, it is customary to restrict to equivariant estimators, i.e., those that satisfy $\widehat{\theta}(gx) = g^*\widehat{\theta}(x)$. Under some conditions, there are Uniformly Minimum Risk Equivariant (UMRE) estimators.

Our invariance condition can be viewed as having a "trivial" induced group G^*, which always acts as the identity. Then the equivariant estimators are those for which $\widehat{\theta}(gx) = \widehat{\theta}(x)$. Thus, equivariant estimators are invariant on the orbits.

The above mentioned UMRE results crucially use that the induced group has large orbits on the parameter space (or in the extreme case, is transitive), so that many parameter values are equivalent. In contrast, we have the complete opposite setting, where the orbits are singletons. Thus our setting is very different from classical equivariance.

Exact invariance in some statistical models. The exact invariance in distribution $X =_d gX$ arises naturally in many familiar statistical settings. In permutation tests, we are testing a null hypothesis H_0, under which the distribution of the observed data $X_1, \ldots, X_n$ is invariant to some permutations. For instance, in two-sample tests against a mean shift alternative, we are interested to test the null of no significant difference between two samples. Under the null, the data is invariant to all permutations, and this can be the basis of statistical inference (e.g., Lehmann and Romano, 2005). In this case, the group is the symmetric group of all permutations, and this falls under our invariance assumption. However, the goals in permutation testing are very different from our ones.

In time series analysis, we have observations $X = (\ldots, X_1, X_2, \ldots, X_t, \ldots)$ measured over time. Here, the classical notion of strict stationarity is the same as invariance under shifts, i.e., $(\ldots, X_t, X_{t+1}, \ldots) =_d (\ldots, X_{t+1}, X_{t+2}, \ldots)$ (e.g., Brockwell and Davis 2009). Thus, our invariance in distribution can capture an important existing notion in time series analysis. Moreover, data augmentation corresponds to adding all shifts of the time series to the data. Hence, it is connected to auto-regressive methods.

We discuss other connections, e.g., to nonparametric density estimation under symmetry constraints, and U-statistics, in other places of this paper.

Variance reduction in Monte Carlo. Variance reduction techniques are widely used in Monte Carlo methods (Robert and Casella, 2013). Data augmentation can be viewed as a type of variance reduction, and is connected to other known techniques. For instance, $f(gX)$ can be viewed as *control* variates for the random variable $f(X)$. The reason is that $f(gX)$ has the same marginal distribution, and hence the same mean as $f(X)$. Taking averages can be viewed as a suboptimal, but universal way to combine control variates.

We briefly mention that under a reflection symmetry assumption, data augmentation can also be viewed as a special case of the method of *antithetic variates*.

Connection to data augmentation. We can summarize the connections to the areas mentioned above. Data augmentation is *computationally feasible approach* to *approximately* learn on the orbits (which are both the minimal sufficient statistics and maximal invariant features). The computational efficiency is partly because we never explicitly compute or store the orbits and invariant features.

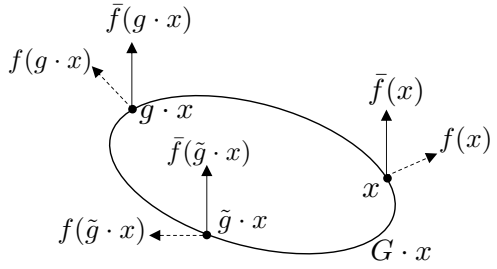

Figure 2: A pictorial illustration of orbit averaging. The circle represents the orbit $G \cdot x = \{gx : g \in G\}$, and $x, gx, \tilde{g}x$ are three different points on this orbit. Although $f(x), f(gx), f(\tilde{g}x)$ may be different, after orbit averaging, $\bar{f}(x), \bar{f}(gx), \bar{f}(\tilde{g}x)$ all take the same value.

4. Theoretical results under exact invariance

In this section, we present our theoretical results for data augmentation under the assumption of *exact invariance*: $gX =_d X$ for almost all $g \sim \mathbb{Q}$ and $x \sim \mathbb{P}$.

4.1 General estimators

We start with some general results on variance reduction. The following lemma characterizes the bias and variance of a general estimator under augmentation:

Lemma 1 (Invariance lemma) *Let f be an arbitrary function s.t. the map $(X, g) \mapsto f(gX)$ is in $L^2(\mathbb{P} \times \mathbb{Q})$. Assume exact invariance holds for $\mathbb{Q}$-almost all $g \in G$. Let $\bar{f}(x) := \mathbb{E}_{g \sim \mathbb{Q}} f(gx)$ be the "orbit average" of f. Then:*

1. *For any x, $\bar{f}(x)$ is the conditional expectation of $f(X)$, conditional on the orbit: $\bar{f}(x) = \mathbb{E}[f(X)|X \in Gx]$, where $Gx := \{gx : g \in G\}$;*
2. *Therefore, by the law of total expectation, the mean of $\bar{f}(X)$ and $f(X)$ coincide: $\mathbb{E}_{X \sim \mathbb{P}} f(X) = \mathbb{E}_{X \sim \mathbb{P}} \bar{f}(X)$;*
3. *By the law of total covariance, the covariance of $f(X)$ can be decomposed as*

$$\mathrm{Cov}_{X \sim \mathbb{P}} f(X) = \mathrm{Cov}_{X \sim \mathbb{P}} \bar{f}(X) + \mathbb{E}_{X \sim \mathbb{P}} \mathrm{Cov}_{g \sim \mathbb{Q}} f(gX);$$

4. *Let φ be any real-valued convex function. Then $\mathbb{E}_{X \sim \mathbb{P}}[\varphi(f(X))] \geq \mathbb{E}_{X \sim \mathbb{P}}[\varphi(\bar{f}(X))]$.*

Proof See Appendix A.1. ■

Figure 2 gives an illustration of orbit averaging.

Augmentation leads to variance reduction. This lemma immediately implies that data augmentation has favorable variance reduction properties. For general estimators, we obtain a direct consequence:

Proposition 2 (Augmentation decreases MSE of general estimators) *Assume exact invariance holds. Consider an estimator $\widehat{\theta}(X)$ of θ_0, and its augmented version $\widehat{\theta}_G(X) = \mathbb{E}_{g\sim\mathbb{Q}}\widehat{\theta}(gX)$. The bias of the augmented estimator is the same as the bias of the original estimator, and the covariance matrix decreases in the Loewner order:*

$$\mathrm{Cov}\left[\widehat{\theta}_G(X)\right] \preceq \mathrm{Cov}\left[\widehat{\theta}(X)\right].$$

Hence, the mean-squared erro (MSE) decreases by augmentation. Moreover, for any convex loss function $L(\theta_0, \cdot)$, we have

$$\mathbb{E}L(\theta_0, \widehat{\theta}(X)) \geq \mathbb{E}L(\theta_0, \widehat{\theta}_G(X)).$$

The estimator $\widehat{\theta}_G(X)$ is an instance of data augmentation via *augmentation distribution* (14). We see that the augmented estimator is no worse than the original estimator according to any convex loss. In some cases, using such a strategy gives an unexpectedly large efficiency gain (See Section 5.4 for an example).

For ERM and MLE, the claim implies that the variance of the augmented loss, log-likelihood, and score functions all decreases. For other estimators based on the augmentation distribution, such as the median, one can show that other measures of error, such as the mean absolute error decrease. This shows how data augmentation can be viewed as a form of *algorithmic regularization*. Indeed, the mean behavior of loss/log-likelihood, etc are all unchanged, but the variance decreases. This indeed shows that augmentation is a natural form of regularization.

Connection to Rao-Blackwell theorem. Lemma 1 has the same flavor as the celebrated Rao-Blackwell theorem, where a "better" estimator is constructed from a preliminary estimator by conditioning on a sufficient statistic. In fact, the orbit GX is a sufficient statistic in many cases and we provide a classical example below to illustrate this point:

Example 1 (U-statistic as an augmented estimator) *Consider data $X_1, ..., X_n$ sampled i.i.d. from some distribution $\mathbb{P}$. We are interested in estimating some functional θ of $\mathbb{P}$. Suppose we have a crude preliminary estimator $\widehat{\theta}(X_1, ..., X_r)$, which takes $r \leq n$ arguments. Let G be the group acting on $(X_1, ..., X_n)$ by randomly selecting r samples without replacement. Then the augmented estimator is*

$$\mathbb{E}_{g\sim\mathbb{Q}}[\widehat{\theta}(g(X_1, ..., X_n))] = \frac{1}{\binom{n}{r}} \sum_{i_1 \neq i_2 \neq \cdots \neq i_r} \widehat{\theta}(X_{i_1}, ..., X_{i_r}),$$

where the summation is taken over the set of all unordered subsets $\{i_1, ..., i_r\}$ of r different integers chosen from $[n]$. This is the U-statistic of order r with kernel $\widehat{\theta}$. The statistical properties of the U-statistic are better than its non-augmented counterpart, which does not use all the data. There are well-known explicit formulas for the variance reduction (e.g., Van der Vaart, 1998).

Beyond groups. Some of our conclusions hold without requiring a group structure on the set of transforms. Instead, it is enough to consider a set (i.e., a semigroup, because the identity always makes sense to include) of transforms $T : \mathcal{X} \to \mathcal{X}$, with a probability

measure $\mathbb{Q}$ on them. This is more realistic in some applications, e.g., in deep learning where we also consider transforms such as cropping images, which are not invertible. Then we still get the variance reduction as in Lemma 1 (specifically, part 2, 3, and 4 still hold). Therefore, under appropriate regularity conditions, we also get improved performance for augmented ERM, as stated previously, as well as in the next sections. However, some of the results and interpretations do not hold in this more general setting. Specifically, the orbits are not necessarily defined anymore, and so we cannot view the augmented estimators as conditional expectations over orbits. The semigroup extension may allow us to handle methods such as subsampling or the bootstrap in a unified framework.

4.2 ERM / M-estimators

We now move on to present our results on the behavior of ERM. Recall our setup from Section 3.1, namely that

$$\theta_0 \in \arg\min_{\theta\in\Theta} \mathbb{E}L(\theta, X), \qquad \theta_G \in \arg\min_{\theta\in\Theta} \mathbb{E}\bar{L}(\theta, X),$$
$$\widehat{\theta}_n \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(\theta, X_i), \qquad \widehat{\theta}_{n,G} \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \bar{L}(\theta, X_i),$$

where Θ is some parameter space. Under exact invariance, it is easy to see that $\mathbb{E}L(\theta, X) = \mathbb{E}\bar{L}(\theta, X)$, and hence there is a one-to-one correspondence between θ_0 and θ_G. To evaluate an estimator $\widehat{\theta}$, we may want

1. small generalization error: we want $\mathbb{E}L(\widehat{\theta}, X) - \mathbb{E}L(\theta_0, X)$ to be small;
2. small parameter estimation error: we want $\|\widehat{\theta} - \theta_0\|_2$ to be small.

Comparing $\widehat{\theta}_n$ with $\widehat{\theta}_{n,G}$, we will see in Section 4.2.1 that the reduction in generalization error can be quantified by averaging the *loss function* over the orbit, whereas in Section 4.2.2, it is shown that the reduction in parameter estimation error can be quantified by averaging the *gradient* over the orbit.

4.2.1 Effect of loss-averaging

To obtain a bound on the generalization error, we first present what can be deduced quite directly from known results on Rademacher complexity. The point of this section is mainly pedagogical, i.e., to remind readers of some basic definitions, and to set a baseline for the more sophisticated results to follow. We recall the classical approach of decomposing the generalization error into terms that can be bounded via concentration and Rademacher complexity (Bartlett and Mendelson, 2002; Shalev-Shwartz and Ben-David, 2014):

$$\mathbb{E}L(\widehat{\theta}_n, X) - \mathbb{E}L(\theta_0, X) = \mathbb{E}L(\widehat{\theta}_n, X) - \frac{1}{n}\sum_{i=1}^{n} L(\widehat{\theta}_n, X_i) + \frac{1}{n}\sum_{i=1}^{n} L(\widehat{\theta}_n, X_i) - \mathbb{E}L(\theta_0, X).$$

The second half of the RHS above can be bounded as

$$\frac{1}{n}\sum_{i=1}^{n} L(\widehat{\theta}_n, X_i) - \mathbb{E}L(\theta_0, X) \le \frac{1}{n}\sum_{i=1}^{n} L(\theta_0, X_i) - \mathbb{E}L(\theta_0, X),$$

because $\widehat{\theta}_n$ is a minimizer of the empirical risk. Hence we arrive at

$$\mathbb{E}L(\widehat{\theta}_n, X) - \mathbb{E}L(\theta_0, X) \leq \mathbb{E}L(\widehat{\theta}_n, X) - \frac{1}{n}\sum_{i=1}^n L(\widehat{\theta}_n, X_i) + \frac{1}{n}\sum_{i=1}^n L(\theta_0, X_i) - \mathbb{E}L(\theta_0, X)$$
$$\leq \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n L(\theta, X_i) - \mathbb{E}L(\theta, X)\right| + \left(\frac{1}{n}\sum_{i=1}^n L(\theta_0, X_i) - \mathbb{E}L(\theta_0, X)\right).$$

Using exact invariance, a similar computation gives

$$\mathbb{E}L(\widehat{\theta}_{n,G}, X) - \mathbb{E}L(\theta_0, X) \leq \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \bar{L}(\theta, X_i) - \mathbb{E}\bar{L}(\theta, X)\right| + \left(\frac{1}{n}\sum_{i=1}^n \bar{L}(\theta_0, X_i) - \mathbb{E}\bar{L}(\theta_0, X)\right).$$

To control the two terms in the RHS, we need to show that both $n^{-1}\sum_{i=1}^n L(\theta, X_i)$ and $n^{-1}\sum_{i=1}^n \bar{L}(\theta, X_i)$ concentrate around their means, respectively, uniformly for all $\theta \in \Theta$. Intuitively, as $\bar{L}$ is an averaged version of L, we would expect that the concentration of $\bar{L}$ *happens at a faster rate* than that of L, hence giving a tighter generalization bound. We will make this intuition rigorous in the following theorem:

Theorem 3 (Rademacher bounds under exact invariance) *Assume the loss $L(\cdot,\cdot) \in [0,1]$ and assume exact invariance holds. Then with probability at least $1-\delta$ over the draw of $X_1, ..., X_n$, we have*

$$\mathbb{E}L(\widehat{\theta}_n, X) - \mathbb{E}L(\theta_0, X) \leq 2\mathcal{R}_n(L\circ\Theta) + \sqrt{\frac{2\log 2/\delta}{n}} \qquad \text{(Classical Rademacher bound)}$$
$$\mathbb{E}L(\widehat{\theta}_{n,G}, X) - \mathbb{E}L(\theta_0, X) \leq 2\mathcal{R}_n(\bar{L}\circ\Theta) + \sqrt{\frac{2\log 2/\delta}{n}}, \qquad \text{(Implication for data augmentation)}$$

where $\mathcal{R}_n(L\circ\Theta)$ and $\mathcal{R}_n(\bar{L}\circ\Theta)$ are the Rademacher complexity of the original loss class and the augmented loss class, repectively, defined as

$$\mathcal{R}_n(L\circ\Theta) = \mathbb{E}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i L(\theta, X_i)\right|, \qquad \mathcal{R}_n(\bar{L}\circ\Theta) = \mathbb{E}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \bar{L}(\theta, X_i)\right|,$$
$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} \text{Rademacher}, \ \{\varepsilon_i\}_{i=1}^n \perp\!\!\!\perp \{X_i\}_{i=1}^n,$$

where the expectation is taken over both $\{\varepsilon_i\}_1^n$ and $\{X_i\}_1^n$. Moreover, augmentation decreases *the Rademacher complexity of the loss class:*

$$\mathcal{R}_n(\bar{L}\circ\Theta) \leq \mathcal{R}_n(L\circ\Theta).$$

Proof This is a special case of Theorem 17. We refer the readers to Appendix B.3 for a proof. ■

The first generalization bound is a standard result. It can be viewed as an intermediate between Theorems 26.3 and Theorem 26.5 of Shalev-Shwartz and Ben-David (2014), part 3, because it is a high-probability bound (like 26.5) for expected generalization error (like

26.3). The second bound simply applies the first one to data augmentation, and the point is that we obtain a sharper result.

We also note that the decrease of the Rademacher complexity under augmentation fits into the "Rademacher calculus" Shalev-Shwartz and Ben-David (2014), along with results such as bounds on the Rademacher complexity under convex combinations and Lipschitz transforms. The current claim holds by inspection.

4.2.2 Effect of gradient-averaging

Just as the averaged loss function concentrates faster around its mean, the averaged gradient also concentrates faster. Specifically, an application of Lemma 1 gives

$$\mathrm{Cov}\nabla L(\theta, X) = \mathrm{Cov}\nabla \bar{L}(\theta, X) + \mathbb{E}_{X\sim\mathbb{P}}\mathrm{Cov}_{g\sim\mathbb{Q}}\nabla L(\theta, gX).$$

If we assume the loss function is strongly convex w.r.t. the parameter, then this observation, along with a recent result from Foster et al. (2019), gives a bound of the parameter estimation error for both $\widehat{\theta}_n$ and $\widehat{\theta}_{n,G}$:

Theorem 4 (Data augmentation reduces variance in ERM) *Assume the loss $L(\cdot, x)$ is λ-strongly convex and assume exact invariance holds. Then*

$$\mathbb{E}\|\widehat{\theta}_n - \theta_0\|_2^2 \leq \frac{4}{\lambda^2 n}\,\mathrm{tr}\left(\mathrm{Cov}\nabla L(\theta_0, X)\right) \qquad \text{(Classical ERM bound)}$$

$$\mathbb{E}\|\widehat{\theta}_{n,G} - \theta_0\|_2^2 \leq \frac{4}{\lambda^2 n}\,\mathrm{tr}\left(\mathrm{Cov}\nabla L(\theta_0, X) - \mathbb{E}_X\mathrm{Cov}_g\nabla L(\theta, gX)\right)$$

(Implication for data augmentation).

Proof The proof is simply a matter of checking the conditions required by Theorem 7 of Foster et al. (2019). It is not hard to see that the strong convexity is preserved under averaging, and so the strong convexity constant of $\theta \to \bar{L}(\theta, x)$ is at least as large as that of $\theta \to L(\theta, x)$. Then invoking part 3 of Lemma 1 gives the desired result. ■

The above result is similar to Theorem 3 in that it gives a sharper upper bound on the estimation error.

4.2.3 Asymptotic results

The finite sample results in Theorem 3 and 4 do not precisely tell us how much we gain. This motivates us to consider an asymptotic regime where we can characterize precisely how much we gain by data augmentation.

We assume the sample space $\mathcal{X} \subseteq \mathbb{R}^d$ and the parameter space $\Theta \subseteq \mathbb{R}^p$. We will consider the classical under-parameterized regime, where both d and p are fixed and $n \to \infty$. The over-parameterized regime will be handled in Section 7. Under weak regularity conditions, it is a well-known result that $\widehat{\theta}_n$ is asymptotically normal with covariance given by the inverse Fisher information (see, e.g, Van der Vaart 1998). We will see that a similar asymptotic normality result holds for $\widehat{\theta}_{n,G}$ and its asymptotic covariance is explicitly computable. This requires some standard assumptions, and we state them for completeness below.

Assumption A (Regularity of the population risk minimizer) *The minimizer θ_0 of the population risk is well separated: for any $\varepsilon > 0$, we have*

$$\sup_{\theta:\|\theta-\theta_0\|\geq\varepsilon} \mathbb{E}L(\theta, X) > \mathbb{E}L(\theta_0, X).$$

Assumption B (Regularity of the loss function) *For the loss function $L(\theta, x)$, we assume that*

1. *uniform weak law of large number holds:*

$$\sup_{\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^{n} L(\theta, X_i) - \mathbb{E}L(\theta, X)| \xrightarrow{p} 0;$$

2. *for each $\theta \in \Theta$, the map $x \mapsto L(\theta, x)$ is measurable;*
3. *the map $\theta \mapsto L(\theta, x)$ is differentiable at θ_0 for almost every x;*
4. *there exists a $L^2(\mathbb{P})$ function $\dot{L}$ s.t. for almost every x and for every θ_1, θ_2 in a neighborhood of θ_0, we have*

$$|L(\theta_1, x) - L(\theta_2, x)| \leq \dot{L}(x)\|\theta_1 - \theta_2\|;$$

5. *the map $\theta \mapsto \mathbb{E}L(\theta, X)$ admits a second-order Taylor expansion at θ_0 with non-singular second derivative matrix V_{θ_0}.*

Note that the two assumptions above can be defined in a similar fashion for the pair $(\theta_G, \bar{L})$. The lemma below says that such a re-definition is unnecessary under exact invariance.

Lemma 5 (Regularity of the augmented loss) *For each $\theta \in \Theta$, assume the map $(X, g) \mapsto L(\theta, gX)$ is in $L^1(\mathbb{P} \times \mathbb{Q})$. Assume exact invariance holds. If the pair (θ_0, L) satisfies Assumption A and B, then the two assumptions also hold for the pair $(\theta_G, \bar{L})$.*

Proof See Appendix A.2. ■

By the above lemma, we conclude that $\widehat{\theta}_{n,G}$ is also asymptotically normal, and a covariance calculation gives the following theorem. We present the classical results for M-estimators in parallel for clarity.

Theorem 6 (Asymptotic normality for the augmented estimator) *Assume Θ is open and the conditions in Lemma 5 hold. Then $\widehat{\theta}_n$ and $\widehat{\theta}_{n,G}$ admit the following Bahadur representation:*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} V_{\theta_0}^{-1} \sum_{i=1}^{n} \nabla L(\theta_0, X_i) + o_p(1) \quad \text{(Classical representation for M-estimators)}$$

$$\sqrt{n}(\widehat{\theta}_{n,G} - \theta_0) = \frac{1}{\sqrt{n}} V_{\theta_0}^{-1} \sum_{i=1}^{n} \nabla \bar{L}(\theta_0, X_i) + o_p(1). \quad \text{(Representation for augmented M-estimators)}$$

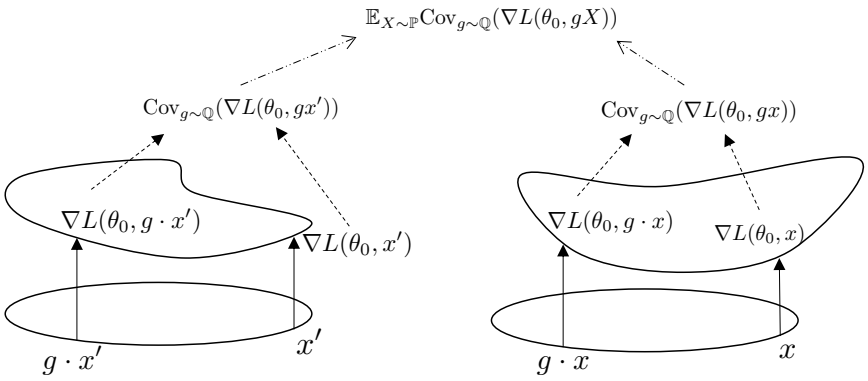

Figure 3: A pictorial illustration on computing the average covariance $\mathbb{E}_X \mathrm{Cov}_g \nabla L(\theta_0, gX)$ of the gradient of the loss over an orbit.

Therefore, both $\widehat{\theta}_n$ and $\widehat{\theta}_{n,G}$ are asymptotically normal:

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_0), \qquad \sqrt{n}(\widehat{\theta}_{n,G} - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_G),$$

where the covariance is given by

$$\begin{aligned} \Sigma_0 &= V_{\theta_0}^{-1} \mathbb{E}_X[\nabla L(\theta_0, X) \nabla L(\theta_0, X)^\top] V_{\theta_0}^{-1} \qquad \text{(Classical covariance)} \\ \Sigma_G &= \Sigma_0 - V_{\theta_0}^{-1} \mathbb{E}_X[\mathrm{Cov}_g \nabla L(\theta_0, gX)] V_{\theta_0}^{-1}. \qquad \text{(Augmented covariance)} \end{aligned}$$

As a consequence, the asymptotic relative efficiency of $\widehat{\theta}_{n,G}$ compared to $\widehat{\theta}_n$ is $\mathrm{RE} = \frac{\mathrm{tr}(\Sigma_0)}{\mathrm{tr}(\Sigma_G)} \geq 1$.

Proof See Appendix A.3. ■

Interpretation. The reduction in covariance is governed by $\mathbb{E}_X \mathrm{Cov}_g \nabla L(\theta_0, gX)$. This is the average covariance of the gradient ∇L along the orbits Gx. If the gradient varies a lot along the orbits, then augmentation gains a lot of efficiency. This makes sense, because this procedure effectively denoises the gradient, making it more stable. See Figure 3 for an illustration.

Alternatively, the covariance of the augmented estimator can also be written as $V_{\theta_0}^{-1}$ $\mathrm{Cov}_X \mathbb{E}_g \nabla L(\theta_0, gX)$ $V_{\theta_0}^{-1}$. The inner term is the covariance matrix of the orbit-average gradient, and similar to above, augmentation effectively denoises and reduces it.

Understanding the Bahadur representation. It is worthwhile to understand the form of the Bahadur representations. For the ERM, we sum the terms $f(X_i) := \nabla L(\theta_0, X_i)$, while for the augmented ERM, we sum the terms $\bar{f}(X_i) = \nabla \bar{L}(\theta_0, gX_i)$. Thus, even in the Bahadur representation, we can see clearly that the effect of data augmentation is to *average the gradients over the orbits.*

Statistical inference. It follows automatically from our theory that statistical inference for θ can be performed in the usual way. Specifically, assuming that $\theta \mapsto L(\theta, x)$ is twice differentiable at θ_0 on a set of full measure, we will have that $V_{\theta_0} = \mathbb{E}\nabla^2 L_{\theta_0}(X)$. We can then compute the plug-in estimator of V_{θ_0}:

$$\widehat{V}_{\theta_0} = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 L(\widehat{\theta}_{n,G}, X_i).$$

Let us also define the plug-in estimator of the gradient covariance:

$$\widehat{\widetilde{I}_{\theta_0}} = \frac{1}{n}\sum_{i=1}^{n} \nabla L(\widehat{\theta}_{n,G}, X_i)\nabla L(\widehat{\theta}_{n,G}, X_i)^\top.$$

We can then define the plug-in covariance estimator

$$\widehat{\Sigma}_G = \widehat{V}_{\theta_0}^{-1}\widehat{\widetilde{I}_{\theta_0}}\widehat{V}_{\theta_0}^{-1}.$$

This leads to the following corollary, whose proof is quite direct and we omit it for brevity.

Corollary 7 (Statistical inference with the augmented estimator) *Assume the same conditions as Theorem 6. Assume in addition that $\theta \to L(\theta, x)$ is twice differentiable at θ_0 on a set of full x-measure, and that each entry of the Hessian $\nabla^2 L(\theta_0, \cdot)$ is in $L^1(\mathbb{P})$. Assume that for both of the functions F_i, $i = 1, 2$ $F_1(\theta, x) = \nabla^2 L(\theta, x)$ and $F_1(\theta, x) = \nabla L(\theta, x)\nabla L(\theta, x)^\top$ there exist $L^1(\mathbb{P})$ functions L_i such that for every θ_1, θ_2 in a neighborhood of θ_0, we have*

$$\|F_i(\theta_1, x) - F_i(\theta_2, x)\| \le L_i(x)\|\theta_1 - \theta_2\|.$$

Then we have:

$$\widehat{\Sigma}_G \xrightarrow{p} \Sigma_G.$$

Therefore, we have the asymptotic normality

$$\sqrt{n}\widehat{\Sigma}_G^{-1/2}(\widehat{\theta}_{n,G} - \theta_0) \Rightarrow \mathcal{N}\Big(0, I_p\Big).$$

Hence, statistical inference for θ_0 can be performed in the usual way, constructing normal confidence intervals and tests based on the asymptotic pivot.

ERM in high dimensions. We conclude this subsection by noting that asymptotic covariance formula can hold in some high-dimensional ERM problems, where the dimension of the parameter space can diverge as $n \to \infty$. Specifically, if we assume θ_0 is sparse and we are able to perform variable selection in a consistent way, then the original (high-dimensional) problem can be reduced to a low-dimensional one, and the same asymptotic variance formulas hold (see, e.g., Wainwright 2019). Then augmentation will lead to benefits as above.

4.2.4 Implications for MLE

Now we consider the special case of maximum likelihood estimation in a model $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Recall the setup from Section 3.2:

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i), \qquad \widehat{\theta}_{\mathrm{aMLE}} \in \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \bar{\ell}_\theta(X_i),$$

where $\bar{\ell}_\theta(x) = \mathbb{E}_{g\sim\mathbb{Q}}\ell_\theta(gx)$ is the averaged version of the log-likelihood.

In the context of ERM, we have $L(\theta, x) = -\ell_\theta(x)$. Hence, if the conditions in Theorem 6 hold, we have

$$\begin{aligned}
\sqrt{n}(\widehat{\theta}_{\mathrm{MLE}} - \theta_0) &\Rightarrow \mathcal{N}(0, I_{\theta_0}^{-1}) \\
\sqrt{n}(\widehat{\theta}_{\mathrm{aMLE}} - \theta_0) &\Rightarrow \mathcal{N}\bigg(0, I_{\theta_0}^{-1}\Big(I_{\theta_0} - \mathbb{E}_{\theta_0}\mathrm{Cov}_g \nabla\ell_{\theta_0}(gX)\Big)I_{\theta_0}^{-1}\bigg),
\end{aligned}$$

where $I_{\theta_0} = \mathbb{E}_{\theta_0}\nabla\ell_{\theta_0}\nabla\ell_{\theta_0}^\top$ is the Fisher information at θ_0.

Invariance on a submanifold. Note that the gain in efficiency is again determined by the term

$$\mathbb{E}_{\theta_0}\mathrm{Cov}_g \nabla\ell_{\theta_0}(gX).$$

We will see that the magnitude of this quantity is closely related to the geometric structure of the parameter space. To be clear, let us recall the invariant subspace Θ_G defined in Equation (8). If this subspace is "small", we expect augmentation to be very efficient. Continuing to assume that the whole parameter set Θ is an open subset of $\mathbb{R}^p$, we additionally suppose that we can identify $\Theta_G \subseteq \Theta$ with a smooth submanifold of $\mathbb{R}^p$. If we assume G *acts linearly* on $\mathcal{X}$, then every $g \in G$ can be represented as a matrix, and thus the Jacobian of $x \mapsto gx$ is simply the matrix representation of g. As a result, for every $\theta \in \Theta_G$, Equation (7) becomes

$$p_\theta(x) = p_\theta(g^{-1}x)/|\det(g)|.$$

As a result, we have

$$\ell_\theta(gx) + \log|\det(g)| = \ell_\theta(x), \ \ \forall\theta \in \Theta_G, g \in G. \tag{16}$$

It is temping to differentiate both sides at θ_0 and obtain an identity. However, it is important to note that in general, for the score function,

$$\nabla\ell_{\theta_0}(gx) \neq \nabla\ell_{\theta_0}(x).$$

To be clear, this is because the differentiation operation at θ_0 may not be valid. Indeed, Equation (16) does not necessarily hold in an open neighborhood of θ_0. If it holds in an open neighborhood, we can differentiate and get the identity for the score. However, it may only hold in a lower dimensional submanifold locally around θ_0, in which case the desired equation for the score function only holds in the *tangential direction* at θ_0. Figure 4 provides an illustration of this intuition, and this intuition is made rigorous by the following proposition:

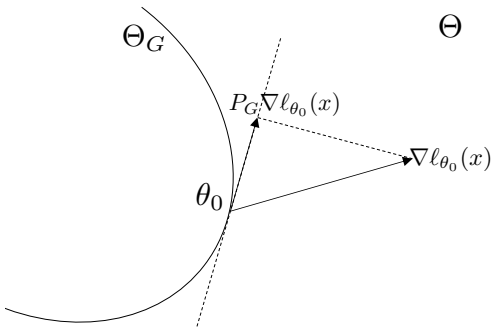

Figure 4: A pictorial illustration of projection onto local tangent space of Θ_G at the point θ_0.

Proposition 8 (MLE under invariance on a submanifold) *Let the conditions of Theorem 6 hold with $L(\theta, x) = -\ell_\theta(x)$. Assume G acts linearly on the sample space $\mathcal{X}$. Let P_G be the orthogonal projection operator onto the tangent space of the manifold Θ_G at the point θ_0, and let $P_G^\perp = Id - P_G$. Then we can decompose*

$$\nabla\ell_{\theta_0}(gx) = P_G\nabla\ell_{\theta_0}(gx) + P_G^\perp\nabla\ell_{\theta_0}(gx),$$

and the tangential part is invariant:

$$P_G\nabla\ell_{\theta_0}(gx) = P_G\nabla\ell_{\theta_0}(x), \ \ \forall g \in G.$$

Moreover, we have that the covariance of the gradient is the covariance of the projection "out" of the tangent space:

$$\mathbb{E}_{\theta_0}\mathrm{Cov}_g(\nabla\ell_{\theta_0}(gX)) = \mathbb{E}_{\theta_0}\mathrm{Cov}_g(P_G^\perp\nabla\ell_{\theta_0}(gX)).$$

Proof See Appendix A.4. ■

The above proposition sends a clear message on the magnitude of $\mathbb{E}_{\theta_0}\mathrm{Cov}_g(\nabla\ell_{\theta_0}(gX))$:

The larger the tangential part is, the less we gain from augmentation.

At one extreme, if Θ_G contains an open neighborhood of θ_0, then $P_G = Id$, so that $\mathbb{E}_{\theta_0}\mathrm{Cov}_G(\nabla\ell_{\theta_0}(gX)) = 0$. This means augmentation does not gain us anything. At another extreme, if Θ_G is a singleton, then $P_G^\perp = Id$, and so augmentation leads to a great variance reduction.

Orbit averaging and tangential projection. We now have two operators that "capture the invariance". The first one is the orbit averaging operator, defined by $\mathbb{E}_G : \nabla\ell_{\theta_0}(x) \mapsto \mathbb{E}_{g\sim\mathbb{Q}}\nabla\ell_{\theta_0}(gx)$. The second one is the tangential projection operator P_G, which projects $\nabla\ell_{\theta_0}(x)$ to the tangent space of the manifold Θ_G at the point θ_0. How does $\mathbb{E}_G$ relate to P_G? In the current notation, recall that the inner term of the asymptotic

covariance Σ_G can be written as $\mathrm{Cov}_{\theta_0}\mathbb{E}_G\nabla\ell_{\theta_0}$. Proposition 8 allows us to do the following decomposition:

$$\mathrm{Cov}_{\theta_0}\mathbb{E}_G\nabla\ell_{\theta_0} = \mathrm{Cov}_{\theta_0}P_G\nabla\ell_{\theta_0}(X) + \mathrm{Cov}_{\theta_0}\mathbb{E}_G P_G^{\perp}\nabla\ell_{\theta_0}, \tag{17}$$

which relates the two operators. An interesting question now is whether the two operators are equivalent. A careful analysis shows that this is not true in general.

Let us consider a special case, where $\theta = (\theta_1, \theta_2)$, and the invariant set Θ_G is characterized by $\theta_2 = 0$. In this special case, the tangent space is exactly $\{(\theta_1, \theta_2) : \theta_2 = 0\}$, and so

$$P_G\nabla\ell_{\theta_0} = \begin{bmatrix}\nabla_1\ell_{\theta_0}\\ 0\end{bmatrix}, \qquad P_G^{\perp}\nabla\ell_{\theta_0} = \begin{bmatrix}0\\ \nabla_2\ell_{\theta_0}\end{bmatrix}, \tag{18}$$

where ∇_1 and ∇_2 denote the gradient w.r.t. the first and the second component of the parameter, respectively. Now, since $\mathbb{E}_G$ is a conditional expectation (see Lemma 1), we know $\mathbb{E}_G\nabla\ell_\theta$ and $\nabla\ell_\theta - \mathbb{E}_G\nabla\ell_\theta$ are uncorrelated. Hence, if $\mathbb{E}_G = P_G$, then $\nabla_1\ell_{\theta_0}$ and $\nabla_2\ell_{\theta_0}$ are uncorrelated. If we write the Fisher information in a block matrix:

$$I_{\theta_0} = \mathbb{E}_{\theta_0}\nabla\ell_{\theta_0}\nabla\ell_{\theta_0}^\top = \mathbb{E}_{\theta_0}\begin{bmatrix}\nabla_1\ell_{\theta_0}\cdot\nabla_1\ell_{\theta_0}^\top & \nabla_1\ell_{\theta_0}\cdot\nabla_2\ell_{\theta_0}^\top\\ \nabla_2\ell_{\theta_0}\cdot\nabla_1\ell_{\theta_0}^\top & \nabla_2\ell_{\theta_0}\cdot\nabla_2\ell_{\theta_0}^\top\end{bmatrix} = \begin{bmatrix}I_{11}(\theta_0) & I_{12}(\theta_0)\\ I_{21}(\theta_0) & I_{22}(\theta_0)\end{bmatrix},$$

then we would have $I_{12}(\theta_0) = 0$. This shows that if $\mathbb{E}_G = P_G$, then the two parameter blocks are orthogonal, i.e., $I_{12}(\theta_0) = 0$, which does not hold in general.

aMLE vs cMLE. Recall that another method that exploits the invariance structure is the constrained MLE, defined by

$$\widehat{\theta}_{\mathrm{cMLE}} \in \arg\max_{\theta\in\Theta_G}\frac{1}{n}\sum_{i=1}^n \ell_\theta(X_i),$$

where we seek the minimizer over the invariant subspace Θ_G. How does it compare to the augmented MLE? If Θ_G is open, and if the true parameter belongs to the interior of the invariant subspace, $\theta_0 \in \mathrm{int}\Theta_G$, then by an application of Theorem 6 (with Θ replaced by Θ_G) and Proposition 8, it is clear that all three estimators – $\widehat{\theta}_{\mathrm{MLE}}, \widehat{\theta}_{\mathrm{aMLE}}$, and $\widehat{\theta}_{\mathrm{cMLE}}$ – share the same asymptotic behavior. At the other extreme, if Θ_G is a singleton, then the constrained MLE, provided it can be solved, gives the exact answer $\widehat{\theta}_{\mathrm{cMLE}} = \theta_0$. In comparison, the augmented MLE gains some efficiency but in general will not recover θ_0 exactly.

What happens to constrained MLE when the manifold dimension of Θ_G is somewhere between 0 and p? We provide an answer to this question in an simplified setup, where we again assume that $\theta = (\theta_1, \theta_2)$ and that Θ_G can be characterized by $\{(\theta_1, \theta_2) : \theta_2 = 0\}$. We first recall the known behavior of cMLE in this case, which has asymptotic covariance matrix $(I_{11}(\theta_0))^{-1}$ (see e.g., Van der Vaart, 1998). By an application of the Schur complement formula, we have

$$(I_{11}(\theta_0))^{-1} \preceq [I_{\theta_0}^{-1}]_{11} = [I_{11}(\theta_0) - I_{12}(\theta_0)(I_{22}(\theta_0))^{-1}I_{21}(\theta_0)]^{-1},$$

which says that the cMLE is more efficient than the MLE–which has has asymptotic covariance matrix $I_{\theta_0}^{-1}$. A further computation gives the following result:

Proposition 9 (Relation between aMLE and cMLE in parametric models that decompose) *Suppose that the parameter has two blocks, $\theta = (\theta_1, \theta_2)$, and the constrained set Θ_G is characterized by $\theta_2 = 0$. Denote by $\bar{I}(\theta) := \mathrm{Cov}\mathbb{E}_g \nabla \ell_\theta(gX)$ the covariance of the average gradient. Define*

$$M_\theta = \begin{bmatrix} (I_{11}(\theta))^{-1} & (I_\theta^{-1})_{1\bullet} \\ (I_\theta^{-1})_{\bullet 1} & (\bar{I}(\theta))^{-1} \end{bmatrix}$$

where the notation $(I^{-1})_{1\bullet}$ refers to the submatrix of I^{-1} corresponding to the rows indexed by the coordinates of θ_1. Then the aMLE is asymptotically more efficient than the cMLE in estimating θ_1 if and only if M_{θ_0} is p.s.d.

Proof See Appendix A.5. ■

In general, it seems that the two are not easy to compare, and the required condition would have to be checked on a case by case basis.

4.3 Finite augmentations

The augmented estimator considered in above sections involves integrals over G, which are in general intractable if G is infinite. As we have seen in Section 3, in practice, one usually randomly samples some finite number of elements in G at each iteration. Therefore, it is of interest to understand the behavior of the augmented estimator in the presence of such a "finite approximation". It turns out that we can repeat the above arguments with minimal changes. Specifically, given a function $f(x)$, we can define $\bar{f}_k(x) = k^{-1} \sum_{j=1}^k f(g_j x)$, where g_j are arbitrary elements of G. Similarly to before, we find that the mean is preserved, while the variance is reduced, in the following way:

1. $\mathbb{E} f(X) = \mathbb{E} \bar{f}_k(X)$;
2. $\mathrm{Cov} f(X) = \mathrm{Cov} \bar{f}_k(X) + \mathbb{E}_{X \sim \mathbb{P}} \mathrm{Var}_k[f(g_1 X), \ldots, f(g_k X)]$, where $\mathrm{Var}_k(a_1, \ldots, a_k)$ is the variance of the k numbers a_i.

The above arguments suggest that for ERM problems, the efficiency gain is governed by the expected variance of k vectors $\nabla L(\theta_0, g_i X)$, where $i = 1, ..., k$. The general principle for practitioners is that the augmented estimator performs better if we choose g_i to "vary a lot" in the appropriate variance metric.

5. Examples

In this section, we give several examples of models where exact invariance occurs. We characterize how much efficiency we can gain by doing data augmentation and compare it with various other estimators. Some examples are simple enough to give a finite-sample characterization, whereas others are calculated according to the asymptotic theory developed in the previous section.

5.1 Exponential families

We start with exponential families, which are a fundamental class of models in statistics (e.g., Lehmann and Casella, 1998; Lehmann and Romano, 2005). Suppose $X \sim \mathbb{P}_\theta$ is

distributed according to an exponential family, so that the log-likelihood can be written as

$$\ell_\theta(X) = \theta^\top T(X) - A(\theta),$$

where $T(X)$ is the sufficient statistic, θ is the natural parameter, $A(\theta)$ is the log-partition function. The densities of $\mathbb{P}_\theta$ are assumed to exist with respect to some common dominating σ-finite measure. Then the score function and the Fisher information is given by

$$\nabla \ell_\theta(X) = T(X) - \nabla A(\theta), \qquad I_\theta = \mathrm{Cov}\left[T(X)\right] = \nabla^2 A(\theta).$$

Given invariance with respect to a group G, by Theorem 6, the asymptotic covariance matrix of the aMLE equals $I_\theta^{-1} J_\theta I_\theta^{-1}$, where J_θ is the covariance of the orbit-averaged sufficient statistic $J_\theta = \mathrm{Cov}_X \mathbb{E}_g T(gX)$.

Assuming a linear action by the group G, by Equation (16), the invariant parameter space Θ_G consists of those parameters θ for which

$$\theta^\top [T(gx) - T(x)] + v(g) = 0, \quad \forall g, x,$$

where $v(g) = \log|\det g|$ is the log-determinant. This is a set of linear equations in θ. Moreover, the log-likelihood is concave, and hence the constrained MLE estimator can in principle be computed as the solution to the following convex optimization problem:

$$\begin{aligned} \widehat{\theta}_{\mathrm{cMLE}} \in \arg\max_\theta \;\; & \theta^\top T(X) - A(\theta) \\ s.t. \;\; & \theta^\top [T(gx) - T(x)] + v(g) = 0, \quad \forall g \in G, x \in \mathcal{X}. \end{aligned}$$

Assume that $\Theta = \mathbb{R}^p$, so that the exponential family is well defined for all natural parameters, and that ∇A is invertible on the range of $\mathbb{E}_g T(gX)$. The KKT conditions of the above convex program is given by

$$\begin{aligned} \widehat{\theta}_{\mathrm{cMLE}} \in [\nabla A]^{-1}(T(X) + \mathrm{span}\{T(gz) - T(z) : z \in \mathbb{R}^d, g \in G\}) \\ s.t. \;\; \theta^\top [T(gx) - T(x)] + v(g) = 0, \quad \forall g, x. \end{aligned}$$

Meanwhile, augmented MLE is the solution of the optimization problem where we replace the sufficient statistic $T(x)$ by $\bar{T}(x) = \mathbb{E}_g T(gx)$:

$$\widehat{\theta}_{\mathrm{aMLE}} \in \arg\max_\theta \;\; \theta^\top \mathbb{E}_g T(gX) - A(\theta).$$

We then have $\widehat{\theta}_{\mathrm{aMLE}} = [\nabla A]^{-1} \mathbb{E}_g T(gX)$. Therefore, for exponential families we were able to give more concrete expressions for the augmented and constrained MLEs.

Gaussian mean. Consider now the important special case of Gaussian mean estimation. Suppose that X is a standard Gaussian random variable, so that $A(\theta) = \|\theta\|^2/2$, and $T(x) = x$. Assume for simplicity that G acts orthogonally. Then we have $\Theta_G = \{v : g^\top v = v, \forall g \in G\}$. Recall that maximizing the Gaussian likelihood is equivalent to minimizing the distance $\|\theta - X\|_2^2$. Hence, the constrained MLE, by definition of the projection, takes the following form:

$$\widehat{\theta}_{\mathrm{cMLE}} = P_G X,$$

where we recall that P_G is the orthogonal projection operator onto the tangent space of Θ_G at θ. However, since Θ_G is a linear space in our case, P_G is simply the orthogonal projection operator onto Θ_G. On the other hand, we have

$$\widehat{\theta}_{\text{aMLE}} = \mathbb{E}_{g\sim\mathbb{Q}}[g]X.$$

In fact, under the current setup, the augmented MLE equals the constrained MLE:

Proposition 10 *Assume G acts linearly and orthogonally. If X is d-dimensional standard Gaussian, then $P_G = \mathbb{E}_{g\sim\mathbb{Q}}[g]$, so that both the aMLE and cMLE are equal to the projection onto the invariant subspace Θ_G. In particular, their risk equals $\dim \Theta_G$.*

Proof See Appendix A.6. ■

For instance, suppose $G = (\{\pm 1\}, \cdot)$ acting by flipping the signs. Then it is clear that $\Theta_G = \{0\}$, and so both the cMLE and aMLE are identically equal to zero.

On the other hand, the marginal MLE (12) is a different object, even in the one-dimensional case. Suppose that $X \sim \mathcal{N}(\theta, 1)$, and we consider the reflection group $G = \{1, -1\}$. The marginal distribution of the data is a Gaussian mixture model

$$Z \sim \frac{1}{2}[\mathcal{N}(\theta, 1) + \mathcal{N}(-\theta, 1)].$$

So the mMLE fits a mixture model, solving

$$\begin{aligned}
\widehat{\theta}_{\text{mMLE}} &\in \arg\max_{\theta} \sum_{i\in[n]} \log\left(\int_G p_\theta(gX_i) d\mathbb{Q}(g)\right) \\
&= \arg\max_{\theta} \sum_{i\in[n]} \log\left(\frac{1}{2}[p_\theta(X_i) + p_\theta(-X_i)]\right) \\
&= \arg\max_{\theta} \sum_{i\in[n]} \log\left(\exp[-(X_i-\theta)^2/2] + \exp[-(-X_i-\theta)^2/2]\right)
\end{aligned}$$

The solution to this is not necessarily identically equal to zero, and in particular it does not agree with the cMLE and aMLE.

Numerical results. We present some numerical results to support our theory. We consider $X \sim \mathcal{N}(\mu, I_d)$, and invariance to the reflection group *that reverses the order of the vector*. This is a stylized model of invariance, which occurs for instance in objects like faces.

In Figure 5, we show the results of two experiments. On the left figure, we show the histograms of the mean squared errors (normalized by dimension) of the MLE and the augmented MLE on a $d = 100$ dimensional Gaussian problem. We repeat the experiment $n_{MC} = 100$ times. We see that the MLE has average MSE roughly equal to unity, while the augmented MLE has average MSE roughly equal to one half. Thus, data augmentation reduces the MSE two-fold. This confirms our theory.

On the right figure, we change the model to each coordinate X_i of X being sampled independently as $X_i \sim Poisson(\lambda)$. We show that the relative efficiency (the relative decrease in MSE) of the MLE and the augmented MLE is roughly equal to two regardless of λ. This again confirms our theory.

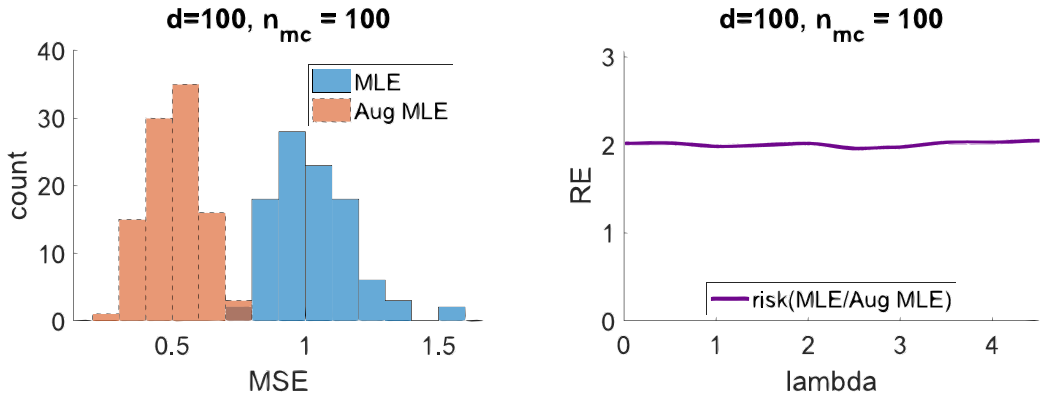

Figure 5: Plots of the increase in efficiency achieved by data augmentation in a *flip symmetry* model.

5.2 Parametric regression models

We consider a regression problem where we observe an iid random sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \subseteq \mathbb{R}^d \times \mathbb{R}$ from the law of a random vector (X, Y). This follows the model:

$$Y = f(\theta_0, X) + \varepsilon, \qquad \varepsilon \perp\!\!\!\perp X, \qquad \mathbb{E}\varepsilon = 0,$$

where $\theta_0 \in \mathbb{R}^p$. We have a group G acting on $\mathbb{R}^d \times \mathbb{R}$ *only through* X:

$$g(X, Y) = (gX, Y),$$

and the invariance is characterized by

$$(gX, Y) =_d (X, Y).$$

In regression or classification problems in deep learning, we typically apply the augmentations conditionally on the outcome or class label. This corresponds to the conditional invariance

$$(gX|Y = y) =_d (X|Y = y).$$

The conditional invariance can be deduced from $(gX, Y) =_d (X, Y)$ by conditioning on $Y = y$. Conversely, if it holds conditionally for each y, we deduce that it also holds jointly, i.e., $(gX, Y) =_d (X, Y)$. Thus, conditional and joint invariance are equivalent. Moreover, each of them clearly implies marginal invariance of the features, i.e., $gX =_d X$. By conditional invariance, $\mathbb{E}(Y|X = x) = \mathbb{E}(Y|X = gx)$, hence

$$f(\theta_0, gx) = f(\theta_0, x) \tag{19}$$

for every non-random x and any $g \in G$. This is what we would expect in the applications we have in mind: the label should be preserved provided there is no random error.

5.2.1 Non-linear least squares regression

We now focus on the least squares loss:

$$L(\theta, X, Y) = (Y - f(\theta, X))^2. \tag{20}$$

The population risk is

$$\mathbb{E}L_\theta(X,Y) = \mathbb{E}(Y - f(\theta_0, X) + f(\theta_0, X) - f(\theta, X))^2 = \mathbb{E}(f(\theta_0, X) - f(\theta, X))^2 + \gamma^2,$$

where γ^2 is the variance of the error ε. Under standard assumptions, the minimizer $\widehat{\theta}_{ERM}$ of $\theta \mapsto \sum_{i=1}^n L(\theta, X_i, Y_i)$ is consistent (see e.g., Example 5.27 of Van der Vaart 1998). Similarly, under standard smoothness conditions, we have

$$\mathbb{E}L(\theta, X, Y) = \underbrace{\gamma^2}_{=\mathbb{E}L(\theta_0, X, Y)} + \frac{1}{2}\mathbb{E}\left[(\theta - \theta_0)^\top \left(2\nabla f(\theta_0, X)\nabla f(\theta_0, X)^\top\right)(\theta - \theta_0)\right] + o(\|\theta - \theta_0\|^2),$$

where $\nabla f(\theta, X)$ is the gradient w.r.t. θ. This suggests that we can apply Theorem 6 with $V_{\theta_0} = 2\mathbb{E}\nabla f(\theta_0, X)\nabla f(\theta_0, X)^\top$ and $\nabla L(\theta_0, X, Y) = -2(Y - f(\theta_0, X))\nabla f(\theta_0, X) = -2\varepsilon\nabla f(\theta_0, X)$, which gives (with the Fisher information $I_\theta = \mathbb{E}\nabla f(\theta, X)\nabla f(\theta, X)^\top$)

$$\sqrt{n}(\widehat{\theta}_{\mathrm{ERM}} - \theta_0) \Rightarrow \mathcal{N}\left(0, V_{\theta_0}^{-1}\mathbb{E}\left[4\varepsilon^2\nabla f(\theta_0, X)\nabla f(\theta_0, X)^\top\right]V_{\theta_0}^{-1}\right) =_d \mathcal{N}\left(0, \gamma^2 I_{\theta_0}^{-1}\right). \tag{21}$$

On the other hand, the augmented ERM estimator is the minimizer $\widehat{\theta}_{\mathrm{aERM}}$ of $\theta \mapsto \sum_{i=1}^n \mathbb{E}_g L(\theta, gX_i, Y_i)$. Now applying Theorem 6 gives

$$\sqrt{n}(\widehat{\theta}_{\mathrm{aERM}} - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_{\mathrm{aERM}}), \tag{22}$$

with the asymptotic covariance being

$$\begin{aligned}
\Sigma_{\mathrm{aERM}} &= \gamma^2 I_{\theta_0}^{-1} - V_{\theta_0}^{-1}\mathbb{E}\left[\mathrm{Cov}_g \nabla L(\theta_0, gX, Y)\right]V_{\theta_0}^{-1} \\
&= \gamma^2 I_{\theta_0}^{-1} - I_{\theta_0}^{-1}\mathbb{E}\left[\mathrm{Cov}_g (Y - f(\theta_0, gX))\nabla f(\theta_0, gX)\right]I_{\theta_0}^{-1} \\
&= \gamma^2 I_{\theta_0}^{-1} - I_{\theta_0}^{-1}\mathbb{E}\left[\varepsilon^2 \mathrm{Cov}_g \nabla f(\theta_0, gX)\right]I_{\theta_0}^{-1} \\
&= \gamma^2 \cdot \left(I_{\theta_0}^{-1} - I_{\theta_0}^{-1}\mathbb{E}\left[\mathrm{Cov}_g \nabla f(\theta_0, gX)\right]I_{\theta_0}^{-1}\right),
\end{aligned}$$

where we used $f(\theta_0, gx) = f(\theta_0, x)$ in the second to last line. Using Lemma 1, we can also write

$$\begin{aligned}
\Sigma_{\mathrm{aERM}} &= \gamma^2 I_{\theta_0}^{-1}\left(I_{\theta_0} - \mathbb{E}\left[\mathrm{Cov}_G \nabla f(\theta_0, gX)\right]\right)I_{\theta_0}^{-1} \\
&= \gamma^2 I_{\theta_0}^{-1}\left(\mathrm{Cov}_X \nabla f(\theta_0, X) - \mathbb{E}\left[\mathrm{Cov}_G \nabla f(\theta_0, gX)\right]\right)I_{\theta_0}^{-1} \\
&= \gamma^2 I_{\theta_0}^{-1}\bar{I}_{\theta_0} I_{\theta_0}^{-1},
\end{aligned}$$

where $\bar{I}_{\theta_0}$ is the "averaged Fisher information", defined as

$$\bar{I}_{\theta_0} = \mathrm{Cov}_X[\mathbb{E}_g \nabla f(\theta_0, gX)] = \mathbb{E}_X\left[\left(\mathbb{E}_g \nabla f(\theta_0, gX)\right)\left(\mathbb{E}_g \nabla f(\theta_0, gX)\right)^\top\right].$$

5.2.2 Two-layer neural network for regression tasks

As an example, consider a two-layer neural network

$$f(\theta, x) = a^\top \sigma(Wx).$$

Here x is a d-dimensional input, W is a $p \times d$ weight matrix, σ is a nonlinearity applied elementwise to the preactivations Wx. The overall parameters are $\theta = (a, W)$. For simplicity, let us focus on the case where $a = 1_p$ is the all ones vector. This will simplify the expressions for the gradient. We can then write

$$f(W, x) = 1^\top \sigma(Wx). \tag{23}$$

Let $I_W = \mathbb{E}\nabla f(W, X)\nabla f(W, X)^\top$ be the fisher information, and denote its averaged version as

$$\bar{I}_W = \mathbb{E}_X\left[\left(\mathbb{E}_g \nabla f(W, gX)\right)\left(\mathbb{E}_g \nabla f(W, gX)\right)^\top\right].$$

Recall Equation (21) and (22):

$$\sqrt{n}(\widehat{W}_{\text{ERM}} - W) \Rightarrow \mathcal{N}\left(0, \sigma^2 I_W^{-1}\right)$$
$$\sqrt{n}(\widehat{W}_{\text{aERM}} - W) \Rightarrow \mathcal{N}\left(0, \sigma^2 I_W^{-1}\bar{I}_W I_W^{-1}\right).$$

Therefore, the efficiency of the ERM and aERM estimators is determined by the magnitude of I_W and $\bar{I}_W$. In general, those are difficult to calculate. However, an exact calculation is possible under a natural example of *translation invariance*. Let the group $G = \{g_0, g_1, ..., g_{p-1}\}$, where g_i acts by shifting a vector circularly by i units:

$$(g_i x)_{j+i \bmod p} = x_j. \tag{24}$$

We equip G with a uniform distribution. Under such a setup, we have the following result:

Theorem 11 (Circular shift data augmentation in two-layer networks) *Consider the two-layer neural network model* (23) *trained using the least squared loss* (20)*. Then:*

1. *The Fisher information matrix I_W can be viewed as a tensor*

$$I_W = \mathbb{E}(\sigma'(WX) \otimes \sigma'(WX)) \cdot (X \otimes X)^\top.$$

If, furthermore, the activation function is quadratic, $\eta(x) = x^2/2$, then I_W can be written as the product of a $p^2 \times d^2$ tensor and a $d^2 \times d^2$ tensor:

$$I_W = (W \otimes W) \cdot \mathbb{E}(XX^\top \otimes XX^\top).$$

2. *Assume the activation function is quadratic. Let C_v be the circulant matrix associated with the vector v, with entries $C_v(i, j) = v_{i-j+1}$. Then $\bar{I}_W$ can be written as*

$$\bar{I}_W = (W \otimes W) \cdot d^{-2}\mathbb{E}(C_X C_X^\top \otimes C_X C_X^\top).$$

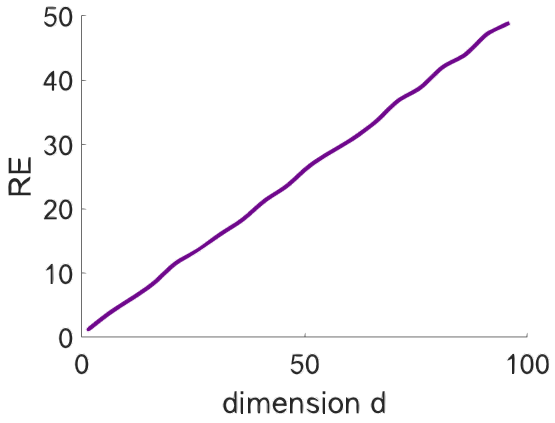

Figure 6: Plot of the increase in efficiency achieved by data augmentation in a *circular symmetry* model.

If furthermore the distribution of X is normal, $X \sim \mathcal{N}(0, I_d)$, then we have

$$\bar{I}_W = (W \otimes W) \cdot F_2^* \cdot (F_2^2 \odot M) \cdot F_2^*.$$

Here $F_2 = F \otimes F$, where F is the $d \times d$ DFT matrix, M is the $d^2 \times d^2$ tensor with entries

$$M_{iji'j'} = F_i^\top F_j \cdot F_{i'}^\top F_{j'} + F_i^\top F_{j'} \cdot F_{i'}^\top F_j + F_i^\top F_{i'} \cdot F_i^\top F_{j'},$$

and F_2^* is the complex conjugate of F_2.

Proof See Appedix A.7. ■

This theorem shows in a precise quantitative sense how much we gain from data augmentation in a two-layer neural network. To get a sense of the magnitude of improvement, we will attempt to understand how much "smaller" $\bar{I}_W$ is compared to I_W by calculating their MSEs. For simplicity, we will impose a prior $W \sim \mathcal{N}(0, I_p \otimes I_d)$, and the MSE is calculated by taking expectation w.r.t. this prior. First we have $\mathbb{E}\,\mathrm{tr}\, I_W = \mathbb{E}\|WXX^\top\|_{Fr}^2$. Let $S = XX^\top$. Now $W \sim \mathcal{N}(0, I_p \otimes I_d)$. So conditional on X, we have $WS \sim \mathcal{N}(0, I_p \otimes S^2)$. Hence, $\mathbb{E}\|WS\|_{Fr}^2 = p\mathbb{E}\,\mathrm{tr}\, S^2 = p\mathbb{E}\,\mathrm{tr}(XX^\top)^2$. Similarly, we find $\mathbb{E}\,\mathrm{tr}\,\bar{I}_W = p\mathbb{E}\,\mathrm{tr}(C_X C_X^\top)^2/d^2$.

Numerical results. In Figure 6, we show the results of an experiment where we randomly generate the input as $X \sim \mathcal{N}(0, I_d)$. For a fixed X, we compute the values of $\mathbb{E}\,\mathrm{tr}\, I_W = p\,\mathrm{tr}(XX^\top)^2$ and $\mathbb{E}\,\mathrm{tr}\,\bar{I}_W = p\,\mathrm{tr}(C_X C_X^\top)^2/d^2$, and record their ratio. We repeat the experiment $n_{MC} = 100$ times. We then show the relative efficiency of aMLE with respect to MLE as a function of the input dimension d. We find that the relative efficiency scales as $RE(d) \sim d/2$. Thus, for the efficiency gain increases as a function of the input dimension. However, the efficiency gain does not depend on the output dimension p. This makes sense, as circular invariance affects and reduces only the input dimension.

5.3 Parametric classification models

Similar calculations as in the previous subsection carry over to the classification problems. We now have a random sample $\{(X_1, Y_1), ..., (X_n, Y_n)\} \subseteq \mathbb{R}^d \times \{0, 1\}$ from the law of a random vector (X, Y), which follows the model:

$$\mathbb{P}(Y = 1 \mid X) = \eta(f(\theta_0, X)),$$

where $\theta_0 \in \mathbb{R}^p$, $\eta : \mathbb{R} \to [0, 1]$ is an increasing activation function, and $f(\theta_0, \cdot)$ is a real-valued function. For example, the sigmoid $\eta(x) = 1/(1 + e^{-x})$ gives the logistic regression model, using features extracted by $f(\theta_0, \cdot)$. As in the regression case, we have a group G acting on $\mathbb{R}^d \times \{0, 1\}$ via

$$g(X, Y) = (gX, Y),$$

and the invariance is

$$(gX, Y) =_d (X, Y).$$

The interpretation of the invariance relation is again two-fold. On the one hand, we have $gX =_d X$. On the other hand, for almost every (w.r.t. the law of X) x, we have

$$\mathbb{P}(Y = 1 \mid gX = x) = \mathbb{P}(Y = 1 \mid X = x).$$

The LHS is $\eta(f(\theta_0, g^{-1}x))$, whereas the RHS is $\eta(f(\theta_0, x))$. This shows that for any (non-random) $g \in G$ and x, we have

$$\eta(f(\theta_0, gx)) = \eta(f(\theta_0, x)).$$

For image classification, the invariance relation says that the class probabilities stay the same if we transform the image by the group action. Moreover, since we assume η is monotonically strictly increasing, applying its inverse gives

$$f(\theta_0, gx) = f(\theta_0, x).$$

5.3.1 Non-linear least squares classification

We consider using the least square loss to train the classifier:

$$L(\theta, X, Y) = (Y - \eta(f(\theta, X)))^2. \tag{25}$$

Though this is not the most popular loss, in some cases it can be empirically superior to the default choices, e.g., logistic loss and hinge loss (Wu and Liu, 2007; Nguyen and Sanner, 2013). The loss function has a bias-variance decomposition:

$$\begin{aligned}\mathbb{E}L(\theta, X, Y) &= \mathbb{E}[Y - \eta(f(\theta_0, X)) + \eta(f(\theta_0, X)) - \eta(f(\theta, X))]^2 \\ &= \underbrace{\mathbb{E}[Y - \eta(f(\theta_0, X))]^2}_{\mathbb{E}L(\theta_0, X, Y)} + \mathbb{E}[\eta(f(\theta_0, X)) - \eta(f(\theta, X))]^2,\end{aligned}$$

where the cross-term vanishes because $\eta(f(\theta_0, X)) = \mathbb{E}[Y|X]$. Note that

$$\begin{aligned}&\mathbb{E}[Y - \eta(f(\theta_0, X))]^2 = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] = \mathbb{E}\Big[\mathbb{E}[(Y - \mathbb{E}[Y|X])^2 \mid X]\Big] \\ &= \mathbb{E}\mathrm{Var}(Y|X) = \mathbb{E}\mathrm{Var}[\mathrm{Bernoulli}(\eta(f(\theta_0, X)))] = \mathbb{E}\eta(f(\theta_0, X))(1 - \eta(f(\theta_0, X))).\end{aligned}$$

Meanwhile, since $\nabla\eta(f(\theta,X)) = \eta'(f(\theta,X))\nabla f(\theta,X)$, for sufficiently smooth η, we have a second-order expansion of the population risk:

$$\mathbb{E}L(\theta,X,Y) = \mathbb{E}L(\theta_0,X,Y)+\frac{1}{2}(\theta-\theta_0)^\top\mathbb{E}[2\eta'(f(\theta_0,X))^2\nabla f(\theta_0,X)\nabla f(\theta_0,X)^\top](\theta-\theta_0)+o(\|\theta-\theta_0\|^2).$$

This suggests that we can apply Theorem 6 with $V_{\theta_0} = \mathbb{E}[2\eta'(f(\theta_0,X))^2\nabla f(\theta_0,X)\nabla f(\theta_0,X)^\top]$ and $\nabla L(\theta,X,Y) = -2(Y-\eta(f(\theta,X)))\eta'(f(\theta,X))\nabla f(\theta,X)$, which gives

$$\sqrt{n}(\widehat{\theta}_{\mathrm{ERM}}-\theta_0) \Rightarrow \mathcal{N}(0,\Sigma_{\mathrm{ERM}}), \tag{26}$$

where the asymptotic covariance is

$$\begin{aligned}\Sigma_{\mathrm{ERM}} &= \mathbb{E}[U_{\theta_0}(X)]^{-1}\mathbb{E}[v_{\theta_0}(X)U_{\theta_0}(X)]\mathbb{E}[U_{\theta_0}(X)]^{-1} \\ v_{\theta_0}(X) &= \eta(f(\theta_0,X))\cdot(1-\eta(f(\theta_0,X))) \\ U_{\theta_0}(X) &= \eta'(f(\theta_0,X))^2\nabla f(\theta_0,X)\nabla f(\theta_0,X)^\top.\end{aligned} \tag{27}$$

Here $v_{\theta_0}(X)$ can be viewed as the noise level, which corresponds $\mathbb{E}\varepsilon^2$ in the regression case. Also, $U_{\theta_0}(X)$ is the information, which corresponds to $\mathbb{E}\nabla f(\theta_0,X)\nabla f(\theta_0,X)^\top$ in the regression case. The classification problem is a bit more involved, because the noise and the information do not decouple (they both depend on X). In a sense, the asymptotics of classification correspond to a regression problem with heteroskedastic noise, whose variance depends on the mean signal level.

In contrast, applying Theorem 6 for the augmented loss gives

$$\sqrt{n}(\widehat{\theta}_{\mathrm{aERM}}-\theta_0) \Rightarrow \mathcal{N}(0,\Sigma_{\mathrm{aERM}}), \tag{28}$$

where

$$\Sigma_{\mathrm{ERM}} - \Sigma_{\mathrm{ERM}} = V_{\theta_0}^{-1}\mathbb{E}\mathrm{Cov}_g\nabla L(\theta_0,gX)V_{\theta_0}^{-1}.$$

We now compute the gain in efficiency:

$$\begin{aligned}\mathbb{E}\mathrm{Cov}_g\nabla L(\theta_0,gX) &= \mathbb{E}\mathrm{Cov}_g\bigg(2(Y-\eta(f(\theta_0,gX)))\eta'(f(\theta_0,gX))\nabla f(\theta_0,gX)\bigg) \\ &= 4\mathbb{E}\bigg[(Y-\eta(f(\theta_0,X)))^2\mathrm{Cov}_g\bigg(\eta'(f(\theta_0,gX))\nabla f(\theta_0,gX)\bigg)\bigg] \\ &= 4\mathbb{E}\bigg[v_{\theta_0}(X)\mathrm{Cov}_g\bigg(\eta'(f(\theta_0,gX))\nabla f(\theta_0,gX)\bigg)\bigg].\end{aligned}$$

In summary, the covariance of ERM is larger than the covariance of augmented ERM by

$$\Sigma_{\mathrm{ERM}} - \Sigma_{\mathrm{aERM}} = \mathbb{E}[U_{\theta_0}(X)]^{-1}\mathbb{E}\bigg[v_{\theta_0}(X)\mathrm{Cov}_g\bigg(\eta'(f(\theta_0,gX))\nabla f(\theta_0,gX)\bigg)\bigg]\mathbb{E}[U_{\theta_0}(X)]^{-1}. \tag{29}$$

5.3.2 Two-layer neural network for classification tasks

We consider the two-layer neural network (23) here. Most of the computations for the regression case carry over to the classification case. Recall that in the current setup, the model (23) becomes

$$\mathbb{P}(Y = 1 \mid X) = \eta(f(W, X)) := \eta(1^\top \sigma(WX)), \tag{30}$$

where $W \in \mathbb{R}^{p\times d}$ and η is a nonlinearity applied elementwise. We then have an analog of Theorem 11:

Corollary 12 *Consider the two-layer neural network model* (23) *trained using the least squares loss* (25)*. Assume the activation function is quadratic: $\sigma(x) = x^2/2$. Let U_W, v_W be defined as in* (27)*. Then*

$$\Sigma_{\text{ERM}} = \mathbb{E}[U_W(X)]^{-1}\mathbb{E}[v_W(X)U_X(X)]\mathbb{E}[U_W(X)]^{-1},$$

where

$$\begin{aligned}\mathbb{E}U_W(X) &= (W \otimes W) \cdot \mathbb{E}[\eta'(f(W, X))^2 \cdot XX^\top \otimes XX^\top] \\ \mathbb{E}[v_W(X)U_W(X)] &= (W \otimes W)\mathbb{E}[v_W(X)\eta'(f(W, X))^2 \cdot XX^\top \otimes XX^\top].\end{aligned}$$

If we further assume that the group G acts by cirlular shift (24) *and $\mathbb{Q}$ is the uniform distribution on G, then*

$$\mathbb{E}U_W(X)\Sigma_{\text{aERM}}\mathbb{E}U_W(X) = (W \otimes W) \cdot d^{-2}\mathbb{E}[v_W(X)\eta'(f(W, X))^2 \cdot C_X C_X^\top \otimes C_X C_X^\top].$$

Thus, the gain by augmentation is characterized by

$$\begin{aligned}&\mathbb{E}[U_W(X)](\Sigma_{\text{ERM}} - \Sigma_{\text{aERM}})\mathbb{E}[U_W(X)] \\ &= (W \otimes W)\mathbb{E}\left[v_W(X)\eta'(f(W, X))^2\left(XX^\top \otimes XX^\top - C_X C_X^\top \otimes C_X C_X^\top\right)\right].\end{aligned}$$

Proof See Appendix A.8. ■

5.4 Improving linear regression by augmentation distribution

In this section, we provide an example on how to use the "augmentation distribution", described in Section 3.3, to improve the performance of ordinary least squares estimator in linear regression. We will see that, perhaps unexpectedly, the idea of augmentation distribution gives an estimator nearly as efficient as the constrained ERM in many cases.

We consider the classical linear regression model

$$Y = X^\top\beta + \varepsilon, \qquad \beta \in \mathbb{R}^p.$$

We again let $\gamma^2 = \mathbb{E}\varepsilon^2$. We will assume that the action is linear, so that g can be represented as a $p \times p$ matrix. If we augment by a single fixed g, we get

$$\arg\min \|y - Xg^\top\beta\|_2^2 = ((Xg^\top)^\top Xg^\top)^{-1}(Xg^\top)^\top y.$$

Following the ideas on augmentation distribution, we can then average the above estimator over $g \sim \mathbb{Q}$ to obtain

$$\widehat{\beta}_{\text{aDIST}} = \mathbb{E}_{g\sim\mathbb{Q}}\left[((Xg^\top)^\top Xg^\top)^{-1}(Xg^\top)^\top y\right].$$

On the other hand, let us consider the estimator arising from constrained ERM. By Equation (19), we have

$$x^\top\beta = (gx)^\top\beta$$

for $\mathbb{P}_X$-a.e. x and $\mathbb{Q}$-a.e. g. This is a set of *linear constraints* on the regression coefficient β. Formally, supposing that x can take any value (i.e., $\mathbb{P}_X$ has mass on the entire $\mathbb{R}^p$), we conclude that β is constrained to be in the invariant parameter subspace Θ_G, which is a linear subspace defined by

$$\Theta_G = \{v : g^\top v = v, \forall g \in G\}.$$

If x can only take values in a smaller subset of $\mathbb{R}^p$, then we get fewer constraints. So the constrained ERM, defined as

$$\widehat{\beta}_{\text{cERM}} = \arg\min_\beta \|y - X\beta\|_2^2 \qquad s.t.\ (g^\top - I_p)\beta = 0\ \forall g \in G,$$

can in principle, be solved via convex optimization.

Intuitively, we expect both $\widehat{\beta}_{\text{aDIST}}$ and $\widehat{\beta}_{\text{cERM}}$ to be better than the vanilla ERM

$$\widehat{\beta}_{\text{ERM}} = \arg\min_\beta \|y - X\beta\|_2^2,$$

Let $r_{\text{ERM}}, r_{\text{aDIST}}, r_{\text{cERM}}$ be the mean squared errors of the three estimators. We summarize the relationship between the three estimators in the following proposition:

Proposition 13 (Comparison between ERM, aDIST and cERM in linear regression) *Let the action of G be linear. Then:*

1. *Denote $v_j \in \mathbb{R}^p$ as the j-th eigenvector of X^TX and d_j^2 as the corresponding eigenvalue. We have*

$$r_{\text{ERM}} = \gamma^2 \operatorname{tr}[X^\top X]^{-1} = \gamma^2 \sum_{j=1}^p d_j^{-2}, \qquad r_{\text{aDIST}} = \gamma^2 \sum_{j=1}^p d_j^{-2}\|\mathcal{G}^\top v_j\|_2^2,$$

   where $\mathcal{G} = \mathbb{E}_{g\sim\mathbb{Q}}[g]$.

2. *If G acts orthogonally, then $r_{\text{aDIST}} \le r_{\text{ERM}}$.*

3. *If G is the permutation group over $\{1, ..., p\}$, then*

$$r_{\text{aDIST}} = \gamma^2 p^{-1} 1_p^\top (X^\top X)^{-1} 1_p, \qquad r_{\text{cERM}} = \gamma^2 p (1_p^\top X^\top X 1_p)^{-1}.$$

   Furthermore, if X is an orthogonal design so that $X^\top X = I_p$, we have

$$r_{\text{ERM}} = p\gamma^2, \qquad r_{\text{aDIST}} = r_{\text{cERM}} = \gamma^2.$$

Proof See Appendix A.9. ■

In general, constrained ERM can be even more efficient than the estimator obtained by the augmention distribution. However, by the third point in the above proposition, in the special case where G is the permutation group, we have $r_{\text{aDIST}} = r_{\text{cERM}} \ll r_{\text{ERM}}$ when the dimension p is large. A direct extension of the above proposition shows that such a phenomenon occurs when G is the permutation group on a subset of $\{1, \dots, p\}$. There are several other subgroups of interest of the permutation group, including the group of cyclic permutations and the group that contains the identity and the operation that "flips" or reverses each vector.

We note briefly that the above results apply *mutatis mutandis* to logistic regression. There, the outcome $Y \in \{-1, 1\}$ is binary, and $P(Y = 1|X = x) = \sigma(x^\top \beta)$, where $\sigma(z) = 1/(1 + \exp(-x))$ is the sigmoid function. The invariance condition reduces to the same as for linear regression. We omit the details.

6. Extension to approximate invariance

In this section, we develop extensions of the results in Section 4 without assuming exact invariance. We only require *approximate invariance* $gX \approx_d X$ in an appropriate sense. We start by recalling the notion of distance between probability distributions based on transporting the mass from one distribution to another (see, e.g., Villani 2003 and references therein):

Definition 14 (Wasserstein metric) *Let $\mathcal{X}$ be a Polish space. Let d be a lower semi-continuous metric on $\mathcal{X}$. For two probability distributions μ, ν on $\mathcal{X}$, we define*

$$\mathcal{W}_d(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x, y) d\pi(x, y),$$

where $\Pi(\mu, \nu)$ are all couplings whose marginals agree with μ and ν. When $\mathcal{X}$ is a Euclidean space and d is the Euclidean distance, we denote $\mathcal{W}_{\ell_2} \equiv \mathcal{W}_1$ and refer to it as the Wasserstein-1 distance.

6.1 General estimators

Since we no longer have $gX =_d X$, the equation $\mathbb{E}f = \mathbb{E}\bar{f}$ cannot hold in general. However, we still expect some variance reduction by averaging a function over the orbit. Hence, we should see a bias-variance tradeoff, which is made clear in the following lemma. For symmetric matrices A, B, C, we will use the notation $A \in [B, C]$ to mean that $B \preceq A \preceq C$ in the Loewner order.

Lemma 15 (Approximate invariance lemma) *Assume the conditions in Lemma 1 hold, but $gX \neq_d X$. Let $\|f\|_\infty = \sup \|f(x)\|_2$ (which can be ∞). Then:*

1. *The expectations satisfy $\|\mathbb{E}_X \bar{f}(X) - \mathbb{E}_X f(X)\|_2 \leq \mathbb{E}_g \mathcal{W}_1(f(gX), f(X))$;*

2. *The covariances satisfy* $\mathrm{Cov}_X \bar{f}(X) = \mathrm{Cov}_{(X,g)} f(gX) - \mathbb{E}_X \mathrm{Cov}_g f(gX)$, *and according to the Loewner order, we have*

$$\mathrm{Cov}_X \bar{f}(X) - \mathrm{Cov}_X f(X) \in \Big[- \mathbb{E}_X \mathrm{Cov}_g f(gX) \pm 4\|f\|_\infty \mathbb{E}_g \mathcal{W}_1(f(gX), f(X)) \cdot I \Big],$$

where I *is the identity matrix;*

3. *Let* φ *be any real-valued convex function, and let* $\overline{\varphi \circ f}(x) = \mathbb{E}_g \varphi \circ f(gx)$. *Then*

$$\mathbb{E}_X \varphi(\bar{f}(X)) - \mathbb{E}_X \varphi(f(X)) \in \Big[\Big(\mathbb{E}_X \varphi(\bar{f}(X)) - \mathbb{E}_X \overline{\varphi \circ f}(X) \Big) \pm \|\varphi\|_{\mathrm{Lip}} \mathbb{E}_g \mathcal{W}_1(f(gX), f(X)) \Big],$$

where $\|\varphi\|_{\mathrm{Lip}}$ *is the (possibly infinite) Lipschitz constant of* φ. *We recall from our prior results that* $\mathbb{E}_X \varphi(\bar{f}(X)) - \mathbb{E}_X \overline{\varphi \circ f}(X) \le 0$.

Proof See Appendix B.1. ■

The above lemma reduces to Lemma 1 under exact invariance. With this lemma at hand, we prove an analog of Proposition 2 below, which characterizes the mean squared error if we think of f as a general estimator:

Proposition 16 (MSE of general estimators under approximate invariance) *Under the setup of Lemma 15, consider the estimator* $\widehat{\theta}(X)$ *of* θ_0, *and its augmented version* $\widehat{\theta}_G(X) = \mathbb{E}_{g \sim \mathbb{Q}} \widehat{\theta}(gX)$. *Then we have*

$$\mathrm{MSE}(\widehat{\theta}_G) - \mathrm{MSE}(\widehat{\theta}) \in \Big[- \mathbb{E}_X \operatorname{tr}(\mathrm{Cov}_g \widehat{\theta}(gX)) \pm \Delta \Big],$$

where

$$\Delta = \mathbb{E}_g \mathcal{W}_1(\widehat{\theta}(gX), \widehat{\theta}(X)) \cdot \Big[\mathbb{E}_g \mathcal{W}_1(\widehat{\theta}(gX), \widehat{\theta}(X)) + 2\|\mathrm{Bias}(\widehat{\theta}(X))\|_2 + 4\|\widehat{\theta}\|_\infty \Big].$$

Proof See Appendix B.2. ■

Compared with exact invariance , apart from the variance reduction term $\mathbb{E}_X \operatorname{tr}(\mathrm{Cov}_g \widehat{\theta}(gX))$, we have an additional "bias" term Δ as a result of approximate invariance. This bias has three components in the present form. We need the following to be small: (1) the Wasserstein-1 distance $\mathbb{E}_g \mathcal{W}_1(\widehat{\theta}(gX), \widehat{\theta}(X))$ (which is small if the invariance is close to being exact), (2) the bias of the original estimator $\mathrm{Bias}(\widehat{\theta}(X))$ is small (which is a minor requirement, because otherwise there is no point in using this estimator), and (3) the sup-norm $\|\widehat{\theta}\|_\infty$ is small (which is reasonable, because we can always standardize the estimating target θ_0).

6.2 ERM / M-estimators

We now extend the results on the behavior of ERM. Recall that θ_0 and θ_G are minimizers of the population risks $\mathbb{E}L(\theta, X)$ and $\mathbb{E}\bar{L}(\theta, X)$, respectively. Meanwhile, $\widehat{\theta}_n$ and $\widehat{\theta}_{n,G}$ are minimizers of the empirical risks $n^{-1} \sum_{i=1}^n L(\theta, X_i)$ and $n^{-1} \sum_{i=1}^n \bar{L}(\theta, X_i)$, respectively.

Under regularity conditions, it is clear that $\widehat{\theta}_{n,G}$ is an asymptotically unbiased estimator for θ_G. However, under approximate invariance, some quantities, for example, θ_0 and θ_G, may not coincide exactly, introducing an additional bias. As a result, there is a bias-variance tradeoff, similar to that observed in Proposition 16.

We first illustrate this tradeoff for the generalization error, in an extension of Theorem 3:

Theorem 17 (Rademacher bounds under approximate invariance) *Let $L(\theta, \cdot)$ be Lipschitz uniformly over $\theta \in \Theta$, with a (potentially infinite) Lipschitz constant $\|L\|_{\text{Lip}}$. Assume $L(\cdot, \cdot) \in [0, 1]$. Then with probability at least $1 - \delta$ over the draw of $X_1, \ldots, X_n$, we have*

$$\mathbb{E}L(\widehat{\theta}_n, X) - \mathbb{E}L(\theta_0, X) \le 2\mathcal{R}_n(L \circ \Theta) + \sqrt{\frac{2\log 2/\delta}{n}} \qquad \text{(Classical Rademacher bound)}$$

$$\mathbb{E}L(\widehat{\theta}_{n,G}, X) - \mathbb{E}L(\theta_0, X) \le 2\mathcal{R}_n(\bar{L} \circ \Theta) + \sqrt{\frac{2\log 2/\delta}{n}} + 2\|L\|_{\text{Lip}} \cdot \mathbb{E}_{g\sim\mathbb{Q}}\mathcal{W}_1(X, gX).$$

(Bound for augmentation under approx invariance)

Moreover, the Rademacher complexity of the augmented loss class can further be bounded as

$$\mathcal{R}_n(\bar{L} \circ \Theta) - \mathcal{R}_n(L \circ \Theta) \le \Delta + \|L\|_{\text{Lip}} \cdot \mathbb{E}_{g\sim\mathbb{Q}}\mathcal{W}_1(X, gX),$$

where

$$\Delta = \mathbb{E}\sup_\theta |\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{E}_g L(\theta, gX_i)| - \mathbb{E}\mathbb{E}_g \sup_{\theta\in\Theta} |\frac{1}{n}\varepsilon_i L(\theta, gX_i)| \le 0$$

is the "variance reduction" term.

Proof See Appendix B.3. ■

Compared to Theorem 3, we see that there is an additional bias term $2\|L\|_{\text{Lip}} \cdot \mathbb{E}_g\mathcal{W}_d(X, gX)$ due to approximate invariance. The bound $\Delta \le 0$ can be loose, and this variance reduction term can be much smaller than zero. Consequently, as long as $\mathbb{E}_g\mathcal{W}_1(X, gX)$ is small, the above theorem can potentially yield a tighter bound for the augmented estimator $\widehat{\theta}_{n,G}$.

We now illustrate the bias-variance tradeoff in the asymptotic regime. The next theorem is analogous to Theorem 6:

Theorem 18 (Asymptotic normality under approximate invariance) *Assume Θ is open. Let Assumption A and B hold for both pairs (θ_0, L) and $(\theta_G, \bar{L})$. Let V_0, V_G be the Hessian of $\theta \mapsto \mathbb{E}L(\theta, X)$ and $\theta \mapsto \mathbb{E}\bar{L}(\theta, X)$, respectively. Let $M_0(X) = \nabla L(\theta_0, X)\nabla L(\theta_0, X)^\top$ and $M_G(X) = \nabla L(\theta_G, X)\nabla L(\theta_G, X)^\top$. Then we have*

$$\begin{aligned} n(\text{MSE}(\widehat{\theta}_{n,G}) - \text{MSE}(\widehat{\theta}_n)) \to\ & n\|\theta_G - \theta_0\|_2^2 + \mathbb{E}_g\mathbb{E}_X\left\langle M_G(gX) - M_G(X), V_G^{-2}\right\rangle \\ & + \mathbb{E}_X\left\langle M_G(X) - M_0(X), V_G^{-2}\right\rangle + \langle \text{Cov}_X \nabla L(\theta_0, X), V_G^{-2} - V_0^{-2}\rangle \\ & - \langle \mathbb{E}_X \text{Cov}_g(\nabla L(\theta_G, gX)), V_G^{-2}\rangle. \end{aligned}$$

Proof See Appendix B.4. ■

From the above theorem, we see that the the variance reduction is determined by the term

$$\mathbb{E}_X \mathrm{Cov}_g(\nabla L(\theta_G, gX)),$$

whereas we have three extra bias terms on the RHS due to approximate invariance. Indeed, if gX and X are close in the $\mathcal{W}_1$ metric, one can check that the three additional bias terms are small. For example, by the dual representation of Wasserstein metric, we have

$$\mathbb{E}_g \mathbb{E}_X \left\langle M_G(gX) - M_G(X), V_G^{-2} \right\rangle \le \mathbb{E}_g \mathcal{W}_1 \left(\langle M_G(gX), V_G^{-2} \rangle, \langle M_G(X), V_G^{-2} \rangle \right)$$
$$\mathbb{E}_X \left\langle M_G(X) - M_0(X), V_G^{-2} \right\rangle \le \mathcal{W}_1 \left(\langle M_G(X), V_G^{-2} \rangle, \langle M_0(X), V_G^{-2} \rangle \right),$$

and the upper bound goes to zero as we approach exact invariance. Hence we can see an efficiency gain by data augmentation.

7. A Case Study on Over-Parameterized Two-Layer Nets

The result on two-layer nets in Section 5 has three major limitations. First, we only considered quadratic activations, whereas in practice rectified linear units (ReLU) are usually used. Also, the entire result concerns the under-parameterized regime, whereas in practice, neural nets are often over-parameterized. Finally, we assumed that we can optimize the weights of the neural network to consistently estimate the true weights, which can be a hard nonconvex optimization problem with only partially known solutions.

In this section, we present an example that addresses the above three limitations: we consider a ReLU network; we allow the input dimension and the number of neurons to scale with n; we use gradient descent for training.

Consider a binary classification problem, where the data points $\{X_i, Y_i\}_1^n \subseteq \mathbb{S}^{n-1} \times \{\pm 1\}$ are sampled i.i.d. from some data distribution. For technical convenience, we assume $|G| < \infty$, and $gX \in \mathbb{S}^{d-1}$ for all $g \in G$ and almost every X from the feature distribution. We again consider a two-layer net $f(x; W, a) = \frac{1}{\sqrt{m}} a^\top \sigma(Wx)$, where $W \in \mathbb{R}^{m \times d}, a \in \mathbb{R}^m$ are the weights and $\sigma(x) = \max(x, 0)$ is the ReLU activation. We consider the following initialization: $W_{ij} \sim \mathcal{N}(0, 1), a_s \sim \mathrm{unif}(\{\pm 1\})$. Such a setup is common in recent literature on Neural Tangent Kernel.

In the training process, we *fix a and only train W*. We use the logistic loss $\ell(z) = \log(1 + e^{-z})$. In our previous notation, we have $L(\theta, X, Y) = \ell(Y f(X; W, a))$. The weight is then trained by gradient descent: $W_{t+1} = W_t - \eta_t \nabla \bar{R}_n(W_t)$, where $\bar{R}_n$ is the augmented empirical risk (4).

To facilitate the analysis, we impose a margin condition below, similar to that in Ji and Telgarsky (2019).

Assumption C (Margin condition) *Let $\mathcal{H}$ be the space of functions $v : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $\int \|v(z)\|_2^2 d\mu(z) < \infty$, where μ is the d-dimensional standard Gaussian probability measure. Assume there exists $\bar{v} \in \mathcal{H}$ and $\gamma > 0$, s.t. the Euclidean norm satisfies $\|\bar{v}(z)\|_2 \le 1$ for any $z \in \mathbb{R}^d$, and that*

$$Y \int \langle \bar{v}(z), gX \rangle \mathbf{1}\{\langle z, gX \rangle > 0\} d\mu(z) \ge \gamma$$

for all $g \in G$ and amost all (X, Y) from the data distribution.

This says that there is a classifier that can distinguish the (augmented) data with positive margin.

We need a few notations. For $\rho > 0$, we define $\mathscr{W}_\rho := \{W \in \mathbb{R}^{m\times d} : \|w_s - w_{s,0}\|_2 \leq \rho \text{ for any } s \in [m]\}$, where $w_s, w_{s,0}$ is the s-th row of W, W_0, respectively. We let

$$\mathcal{R}_n := \mathbb{E} \sup_{W\in\mathscr{W}_\rho} |\frac{1}{n}\varepsilon_i[-\ell'(y_i f(X_i; W, a))]|$$

be the Rademacher complexity of the non-augmented gradient, where the expectation is taken over both $\{\varepsilon_i\}_1^n$ and $\{X_i, Y_i\}_1^n$. Similarly, writing $f_{i,g}(W) = f(gX_i; W, a)$, we define

$$\bar{\mathcal{R}}_n := \mathbb{E} \sup_{W\in\mathscr{W}_\rho} |\frac{1}{n}\varepsilon_i \mathbb{E}_g[-\ell'(Y_i f_{i,g}(W))]|$$

to be the Rademacher complexity of the augmented gradient. The following theorem characterizes the performance gain by data augmentation in this example:

Theorem 19 (Benefits of data augmentation for two-layer ReLU nets) *Under Assumption C, take any $\varepsilon \in (0, 1)$ and $\delta \in (0, 1/5)$. Let*

$$\lambda = \frac{\sqrt{2\log(4n|G|/\delta)} + \log(4/\varepsilon)}{\gamma/4}, M = \frac{4096\lambda^2}{\gamma^6},$$

and let $\rho = 4\lambda/(\gamma\sqrt{m})$. Let k be the best iteration (with the lowest empirical risk) in the first $\lceil 2\lambda^2/n\varepsilon\rceil$ steps. Let $\alpha = 16[\sqrt{2\log(4n|G|/\delta)} + \log(4/\varepsilon)]/\gamma^2 + \sqrt{md} + \sqrt{2\log(1/\delta)}$. For any $m \geq M$ and any constant step size $\eta \leq 1$, with probability at least $1 - 5\delta$ over the random initialization and i.i.d. draws of the data points, we have

$$\mathbb{P}(Y f(X; W_k, a) \leq 0) \leq 2\varepsilon + [\sqrt{\frac{2\log 2/\delta}{n}} + 4\bar{\mathcal{R}}_n] + \frac{1}{2}\mathbb{E}_Y \mathbb{E}_g \mathcal{W}_1(X|Y, gX|Y) \cdot \alpha.$$

The three terms bound the optimization error, generalization error, and the bias due to approximate invariance. Moreover, with probability at least $1 - \delta$ over the random initialization, we have

$$\bar{\mathcal{R}}_n - \mathcal{R}_n \leq \Delta + \frac{1}{4}\mathbb{E}_Y \mathbb{E}_g \mathcal{W}_1(X|Y, gX|Y) \cdot \alpha,$$

where

$$\Delta = \mathbb{E} \sup_{W\in\mathscr{W}_\rho} |\frac{1}{n}\varepsilon_i \mathbb{E}_g[-\ell'(y_i f_{i,g}(W))]| - \mathbb{E}\mathbb{E}_g \sup_{W\in\mathscr{W}_\rho} |\frac{1}{n}\varepsilon_i[-\ell'(y_i f_{i,g}(W))]| \leq 0$$

is the "variance reduction" term.

Proof See Appendix B.5. ■

The proof idea is largely based on results in Ji and Telgarsky (2019). We decompose the

overall error into optimization error and generalization error. The optimization error is taken care of by a corollary of Theorem 2.2 in Ji and Telgarsky (2019). The generalization error is dealt with by adapting several arguments in Theorem 3.2 of Ji and Telgarsky (2019) and using some arguments in the proof of Theorem 17.

Again, we see a bias-variance tradeoff in this example due to approximate invariance. We conclude this section by noting that the term $\mathcal{R}_n$ can be further bounded by invoking Rademacher calculus and the Rademacher complexity bound for linear classifiers. We refer readers to Section 3 of Ji and Telgarsky (2019) for details.

8. Potential applications

8.1 Cryo-EM and related problems

In this section, we describe several important problems in the biological and chemical sciences, and how data augmentation may be useful. Cryo-Electron Microscopy (Cryo-EM) is a revolutionary technique in structural biology, allowing us to determine the structure of molecules to an unprecedented resolution (e.g., Frank, 2006). The technique was awarded the Nobel Prize in Chemistry in 2017.

The data generated by Cryo-EM poses significant data analytic (mathematical, statistical, and computational) challenges (Singer, 2018). In particular, the data possesses several invariance properties that can be exploited to improve the accuracy of molecular structure determination. However, exploiting these invariance properties is highly nontrivial, because of the massive volume of the data, and due to the high levels of noise. In particular, exploiting the invariance is an active area of research (Kam, 1980; Frank, 2006; Zhao et al., 2016; Bandeira et al., 2017; Bendory et al., 2018). Classical and recent approaches involve mainly (1) latent variable models for the unknown symmetries, and (2) invariant feature approaches. Here we will explain the problem, and how data augmentation may help.

In the imaging process, many copies of the molecule of interest are frozen in a thin layer of ice, and then 2D images are taken via an electron beam. A 3D molecule is represented by an electron density map $\phi : \mathbb{R}^3 \to \mathbb{R}$. Each molecule is randomly rotated, via a rotation that can be represented by a 3D orthogonal rotation matrix $R_i \in O(3)$. Then we observe the noisy line integral

$$Y_i = \int_z \phi(R_i[x, y, z]^\top)dz + \varepsilon_i.$$

We observe several iid copies, and the goal is to estimate the density map ϕ. Clearly the model is invariant under rotations of ϕ. Existing approaches mainly work by fitting statistical methods for latent variable models, such as the expectation maximization (EM) algorithm. Data augmentation is a different approach, where we add the data transformed according to the symmetries. It is interesting, but beyond our scope, to investigate if this can improve the estimation accuracy.

Invariant denoising. A related problem is invariant denoising, where we want to denoise images subject to an invariance of their distribution, say according to rotations (see e.g., Vonesch et al., 2015; Zhao et al., 2016, 2018). This area is well studied, and popular approaches rely on invariant features. It is known how to do it for rotations. However, capturing translation-invariance poses complications to the invariant features approach. In

principle, data augmentation could be used as a more general approach.

XFEL. Another related technique, X-ray free electron lasers (XFEL), is a rapidly developing and increasingly popular experimental method for understanding the three-dimensional structure of molecules (e.g., Favre-Nicolin et al., 2015; Maia and Hajdu, 2016; Bergmann et al., 2017). Single molecule XFEL imaging collects two-dimensional diffraction patterns of single particles at random orientations. A key advantage is that XFEL uses extremely short femtosecond X-ray pulses, during which the molecule does not change its structure. On the other hand, we only capture one diffraction pattern per particle and the particle orientations are unknown, so it is challenging to reconstruct the 3D structure at a low signal-to-noise ratio. The images obtained are very noisy due to the low number of photons that are typical for single particles (Pande et al., 2015).

A promising approach for 3-D structure reconstruction is Kam's method (Kam, 1977, 1980; Saldin et al., 2009), which requires estimating the covariance matrix of the noiseless 2-D images. This is extremely difficult due to low photon counts, and motivated prior work to develop improved methods for PCA and covariance estimation such as ePCA (Liu et al., 2018), as well as the steerable ePCA method Zhao et al. (2018) that builds in invariances. As above, it would be interesting to investigate if we can use data augmentation as another approach for rotation-invariance.

8.2 Spherically invariant data

Here we discuss models for spherically invariant data, and how data augmentation may be used. See for instance Fisher et al. (1993) for more general models of spherical data. In the invariant model, the data $X \in \mathbb{R}^p$ is such that $X =_d OX$ for any orthogonal matrix O. One can see that the Euclidean norms $\|X\|$ are sufficient statistics. There are several problems of interest:

- Estimating the radial density. By taking the norms of the data, this reduces to estimating their 1D density.
- Estimating the marginal density f of a single coordinate. Here it is less obvious how to exploit spherical invariance. However, data augmentation provides an approach.

A naive estimator for the marginal density is any density estimator applied to the first coordinates of the data, $X_1(1), \ldots, X_n(1)$. Since $X_i(1) \sim_{iid} f$, we can use any estimator, $\widehat{f}(X) = \widehat{f}(X_1(1), \ldots, X_n(1))$ for instance a kernel density estimator. However, this is inefficient, because it does not use information in all coordinates.

In data augmentation we rotate our data uniformly, leading to

$$\widehat{f}_a(X) = \int \widehat{f}(gX) d\mathbb{Q}(g) = \mathbb{E}_{O_1,\ldots,O_p \sim O(p)} f([O_1 X_1](1), \ldots, [O_p X_p](1)).$$

Note that if $O \sim O(p)$, then for any vector x, $[Ox](1) =_d \|x\| Z(1)/\|Z\|$, where $Z \sim \mathcal{N}(0, I_p)$. Hence the expectation can be rewritten in terms of Gaussian integrals. It is also possible to write it as a 2-dimensional integral, in terms of $Z(1), \|Z(2:p)\|^2$, which have independent normal and Chi-squared distributions. However, in general, it may be hard to compute exactly.

When the density estimator decouples into a sum of terms over the datapoints, then this expression simplifies. This is the case for kernel density estimators: $\widehat{f}(x) = (nh^p)^{-1} \sum_{i=1}^n k([x - x_i(1)]/h)$. More generally, if $\widehat{f}(x) = \sum_{i=1}^n T(x - x_i(1))$, then we only need to calculate

$$\begin{aligned}\tilde{T}(x) &= \mathbb{E}_O T\left(x - [Oy](1)\right) \\ &= \mathbb{E}_{Z\sim\mathcal{N}(0,I_p)} T\left(x - \|y\|\frac{Z(1)}{\|Z\|}\right).\end{aligned}$$

This is significantly simpler than the previous expression. It can also be viewed as a form of convolution of a kernel with T, which is already a kernel typically. Therefore, we have shown how data augmentation can be used to estimate the marginal density of coordinates for spherically uniform data.

8.3 Random effects models

Data augmentation may have applications to certain random effect models (Searle et al., 2009). Consider the one-way layout $X_{ij} = \mu + A_i + B_{ij}$, $i = 1\ldots, s$, $j = 1, \ldots, n_i$, where $A_i \sim \mathcal{N}(0, \sigma_A^2)$, $B_{ij} \sim \mathcal{N}(0, \sigma_B^2)$ independently. We want to estimate the global mean μ and the variance components σ_A^2, σ_B^2. If the layout is unbalanced, that is the number of replications is not equal, this can be somewhat challenging. Two general approaches for estimation are the restricted maximum likelihood (REML), and minimum norm quadratic estimation (MINQUE) methods.

Here is how one may use data augmentation. Consider a simple estimator of σ_B^2 such as $\widehat{\sigma}_B^2 = s^{-1} \sum_{i=1}^n \mathbb{E}(X_{i1} - X_{i2})^2/2$ (we assume that $n_i \geq 2$ for all i). This is a heuristic plug-in estimator, which is convenient to write down and compute in a closed form. It is also unbiased. However, it clearly does not use all samples, and therefore, it should be possible to improve it.

Now let us denote by X_i the block of i-th observations. These have a joint normal distribution $X_i \sim \mathcal{N}(\mu 1_{n_i}, \sigma_A^2 1_{n_i} 1_{n_i}^\top + \sigma_B^2 I_{n_i})$. The model is invariant under the operations

$$X_i \to O_i X_i,$$

for any orthogonal matrix O_i of size n_i for which $O_i 1_{n_i} = 1_{n_i}$, i.e., a matrix that has the vector of all ones 1_{n_i} as an eigenvector. Let G_i be the group of such matrices. Then the overall model is invariant under the action of the direct product $G_1 \times G_2 \times \ldots \times G_s$. Therefore, any estimator that is not invariant with respect to this group can be improved by data augmentation.

Going back to the estimator $\widehat{\sigma}_B^2$, to find its augmented version, we need to compute the quantity $\mathbb{E}([Ox]_1 - [Ox]_2)^2$, where $x \in \mathbb{R}^k$ is fixed and O is uniformly random from the group of orthogonal matrices such that $O1_k = 1_k$. Write $x = \bar{x}1_k + r$, where $\bar{x}$ is the mean of the entries of x. Then $Ox = \bar{x}1_k + Or$, and $[Ox]_j = \bar{x} + [Or]_j$. Thus we need $\mathbb{E}([Or]_1 - [Or]_2)^2$. This can be done by using that Or is uniformly distributed on the $k-1$ dimensional orthocomplement of the 1_k vector, and we omit the details.

Acknowledgments

We are grateful to Zongming Ma for proposing the topic, and for participating in our early meetings. We thank Jialin Mao for participating in several meetings and for helpful references. We thank numerous people for valuable discussions, including Peter Bickel, Kostas Daniilidis, Jason Lee, William Leeb, Po-Ling Loh, Tengyu Ma, Jonathan Rosenblatt, Yee Whye Teh, Tengyao Wang, the members in Konrad Kording's lab, with special thanks to David Rolnick. We thank Dylan Foster for telling us about the results from Foster et al. (2019). This work was supported in part by NSF BIGDATA grant IIS 1837992 and NSF TRIPODS award 1934960. We thank the Simons Institute for the Theory of Computing for providing AWS credits.

Appendix A. Proofs for results under exact invariance

A.1 Proof of Lemma 1

We first prove part 1. Let x be fixed. Let $A = \{X \in Gx\}$. It suffices to show

$$\int_A \mathbb{E}_g f(gx) d\mathbb{P}(X) = \int_A f(X) d\mathbb{P}(X).$$

For an arbitrary $g \in G$, the RHS above is equal to

$$\begin{aligned}\int_A f(X) d\mathbb{P}(X) &= \int f(gX) \mathbb{1}\{gX \in Gx\} d\mathbb{P}(X) \\ &= \int f(gX) \mathbb{1}\{X \in Gx\} d\mathbb{P}(X),\end{aligned}$$

where the first equality is by the exact invariance, and the second equality is by the definition of the orbit. Taking expectation w.r.t. $\mathbb{Q}$, we get

$$\int_A f(X) d\mathbb{P}(X) = \int_G \int f(gX) \mathbb{1}\{X \in Gx\} d\mathbb{P}(X) d\mathbb{Q}(g).$$

On the event A, there exists g_X^*, potentially depending on X, s.t. $X = g_X^* x$. Hence, we have

$$\begin{aligned}\int_A f(X) d\mathbb{P}(X) &= \int_G \int f(g \circ g_X^* x) \mathbb{1}\{X \in Gx\} d\mathbb{P}(X) d\mathbb{Q}(g) \\ &= \int \int_G f(g \circ g_X^* x) d\mathbb{Q}(g) \mathbb{1}\{X \in Gx\} d\mathbb{P}(X) \\ &= \int \int_G f(gx) d\mathbb{Q}(g) \mathbb{1}\{X \in Gx\} d\mathbb{P}(X) \\ &= \int_A \mathbb{E}_g f(gx) d\mathbb{P}(X),\end{aligned}$$

where the second equality is by Fubini's theorem, and the third inequality is due to the translation invariant property of the Haar measure.

Part 2 follows by law of total expectation along with the above point.

Part 3 follows directly from part 1 and the law of the total covariance applied to the random variable $f(gX)$, where $g \sim \mathbb{Q}, X \sim \mathbb{P}$.

Part 4 follows from Jensen's inequality.

A.2 Proof of Lemma 5

By exact invariance and Fubuni's theorem, it is clear that $\mathbb{E}\bar{L}(\theta, X) = \mathbb{E}L(\theta, X)$ for any $\theta \in \Theta$. Hence $\theta_G = \theta_0$ and Assumption A is verified for $(\theta_G, \bar{L})$. We now verify the five parts of Assumption B.

For part 1, we have

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\bar{L}(\theta, X_i) - \mathbb{E}\bar{L}(\theta, X)\right| &= \sup_{\theta\in\Theta}\left|\mathbb{E}_g[\frac{1}{n}\sum_{i=1}^{n}L(\theta, gX_i) - \mathbb{E}L(\theta, gX)]\right| \\ &\le \sup_{\theta\in\Theta}\mathbb{E}_g\left|\frac{1}{n}\sum_{i=1}^{n}L(\theta, gX_i) - \mathbb{E}L(\theta, gX)\right| \\ &\le \mathbb{E}_g\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}L(\theta, gX_i) - \mathbb{E}L(\theta, gX)\right| \\ &= o_p(1),\end{aligned}$$

where the two inequalities is by Jensen's inequality, and the convergence statement is true because of the exact invariance and the fact that the original loss satisfies part 1 of Assumption B.

Part 2 is true because we have assumed that the action $x \mapsto gx$ is continuous.

For part 3, since (θ_0, L) satisfies this assumption, we know that on an event with full probability, we have

$$\lim_{\delta\to 0}\frac{\left|L(\theta_0+\delta, gX) - L(\theta_0, gX) - \delta^\top\nabla L(\theta_0, gX)\right|}{\|\delta\|} = 0.$$

Now we have

$$\begin{aligned}&\frac{\left|\mathbb{E}_g L(\theta_0+\delta, gX) - \mathbb{E}_g L(\theta_0, gX) - \delta^\top\mathbb{E}_g\nabla L(\theta_0, gX)\right|}{\|\delta\|} \\ &\le \frac{\mathbb{E}_g\left|L(\theta_0+\delta, gX) - L(\theta_0, gX) - \delta^\top\nabla L(\theta_0, gX)\right|}{\|\delta\|} \\ &\le \mathbb{E}_g[\dot{L}(gX) + \|\nabla L(\theta_0, gX)\|],\end{aligned}$$

where the first inequality is by Jensen's inequality, and the second inequality is by part 4 applied to (θ_0, L). We have assumed that $\mathbb{E}[\dot{L}(X)^2] < \infty$, and hence so does $\mathbb{E}[\dot{L}(X)]$. By exact invariance, we have

$$\mathbb{E}_X\mathbb{E}_g[\dot{L}(gX)] = \mathbb{E}_g\mathbb{E}_X[\dot{L}(gX)] = \mathbb{E}[\dot{L}(X)] < \infty,$$

which implies $\mathbb{E}_g[\dot{L}(gx)] < \infty$ for $\mathbb{P}$-a.e. x. On the other hand, part 5 applied to (θ_0, L) implies the existence of $\mathbb{E}\nabla L(\theta_0, X)\nabla L(\theta_0, X)^\top$, and hence $\mathbb{E}\|\nabla L(\theta_0, X)\|^2 < \infty$ and so does $\mathbb{E}\|\nabla L(\theta_0, X)\|$. Now a similar argument shows that under exact invariance, we have

$$\mathbb{E}_g\|\nabla L(\theta_0, gx)\| < \infty.$$

for $\mathbb{P}$-a.e. x. Hence we can apply dominated convergence theorem to conclude that

$$\begin{aligned}
&\lim_{\delta\to 0} \frac{\left|\mathbb{E}_g L(\theta_0+\delta, gX) - \mathbb{E}_g L(\theta_0, gX) - \delta^\top \mathbb{E}_g \nabla L(\theta_0, gX)\right|}{\|\delta\|} \\
&\leq \mathbb{E}_g \lim_{\delta\to 0} \frac{\left|L(\theta_0+\delta, gX) - L(\theta_0, gX) - \delta^\top \nabla L(\theta_0, gX)\right|}{\|\delta\|} \\
&= 0,
\end{aligned}$$

which implies that $\theta \mapsto \bar{L}(\theta, x)$ is indeed differentiable at θ_0.

For part 4, by assumption, we have

$$|L(\theta_1, gx) - L(\theta_2, gx)| \leq \dot{L}(gx)\|\theta_1 - \theta_2\|$$

for almost every x and every θ_1, θ_2 in a neighborhood of $\theta_0 = \theta_G$. Taking expectation w.r.t. $g \sim \mathbb{Q}$, we get

$$\begin{aligned}
|\bar{L}(\theta_1, x) - \bar{L}(\theta_2, x)| &\leq \mathbb{E}_g |L(\theta_1, gx) - L(\theta_2, gx)| \\
&\leq \mathbb{E}_g \dot{L}(gx)\|\theta_1 - \theta_2\|.
\end{aligned}$$

Now it suffices to show $x \mapsto \mathbb{E}_g \dot{L}(gx)$ is in $L^2(\mathbb{P})$. This is true by an application of Jensen's inequality and exact invariance:

$$\begin{aligned}
\mathbb{E}_X[(\mathbb{E}_g \dot{L}(gX))^2] &\leq \mathbb{E}_X \mathbb{E}_g[(\dot{L}(gX))^2] \\
&= \mathbb{E}_g \mathbb{E}_X[(\dot{L}(gX))^2] \\
&= \mathbb{E}_g \mathbb{E}\dot{L}^2 \\
&= \mathbb{E}\dot{L}^2 \\
&< \infty.
\end{aligned}$$

Part 5 is true because under exact invariance, we have $\mathbb{E}L(\theta, X) = \mathbb{E}\bar{L}(\theta, X)$.

A.3 Proof of Theorem 6

The results concerning $\widehat{\theta}_n$ is classical (see, e.g., Theorem 5.23 of Van der Vaart 1998). By Lemma 3.4, we can apply Theorem 5.23 of Van der Vaart (1998) to the pair $(\theta_G = \theta_0, \bar{L})$ to conclude the Bahadur representation. Hence the asymptotic normality of $\widehat{\theta}_{n,G}$ follows and we have

$$\Sigma_G = V_{\theta_0}^{-1} \mathbb{E}[\bar{L}(\theta_0, X)\bar{L}(\theta_0, X)^\top] V_{\theta_0}^{-1}.$$

The final representation of Σ_G follows from part 3 of Lemma 1.

A.4 Proof of Proposition 8

Let T_pM be the tangent space of M at the point p. The inclusion map from Θ_G to $\mathbb{R}^p$ is an immersion. So we can decompose $T_{\theta_0}\mathbb{R}^p = T_{\theta_0}\Theta_G \oplus (T_{\theta_0}\Theta_G)^\perp$, i.e., the direct sum of

the tangent space of Θ_G at θ_0 and its orthogonal complement. Hence the decomposition is valid. Now as $\ell_\theta(gx) + \log\det g = \ell_\theta(x)$ for any $\theta \in \Theta_G$, it is clear that the gradient of ℓ_{θ_0} w.r.t. Θ_G is invariant.

Using $P_G \nabla \ell_{\theta_0}(gX) = P_G \nabla \ell_{\theta_0}(X)$, we have

$$\begin{aligned}\mathbb{E}_{\theta_0}\mathrm{Cov}_g(\nabla \ell_{\theta_0}(gX)) &= \mathbb{E}_{\theta_0}\mathrm{Cov}_g(P_G\nabla \ell_{\theta_0}(gX) + P_G^\perp \nabla \ell_{\theta_0}(gX)) \\ &= \mathbb{E}_{\theta_0}\mathrm{Cov}_g(P_G\nabla \ell_{\theta_0}(X) + P_G^\perp \nabla \ell_{\theta_0}(gX)) \\ &= \mathbb{E}_{\theta_0}\mathrm{Cov}_g(P_G^\perp \nabla \ell_{\theta_0}(gX)),\end{aligned}$$

which concludes the proof.

A.5 Proof of Proposition 9

From our theory we know that the asymptotic covariance matrix of cMLE for estimating parameters $\theta = (\theta_1, \theta_2)$ is $I^{-1}\bar{I}I^{-1}$, while that of aMLE for estimating θ_1 is I_{11}^{-1}. Thus aMLE is asymptotically more efficient than the cMLE if and only if

$$I_{11}^{-1} \succeq [I^{-1}\bar{I}I^{-1}]_{11}.$$

Denoting $K := (I^{-1})_{1\bullet}$, this is equivalent to

$$I_{11}^{-1} \succeq K\bar{I}K^\top,$$

by Equation (17) and (18), Using the Schur complement formula, this is equivalent to the statement that the Shur complement of the matrix I_G^{-1} in the matrix M_θ is PSD, where M_θ equals

$$M_\theta = \begin{bmatrix} I_{11}^{-1} & K \\ K^\top & \bar{I}^{-1} \end{bmatrix}.$$

Now, the top left block of this matrix, I_{11}^{-1}, is PSD. Thus, from the properties of Schur complements, the entire matrix M_θ is also PSD. Therefore, the condition is equivalent to the matrix M_θ being PSD.

A.6 Proof of Proposition 10

Let $C = \mathbb{E}_{g\sim\mathbb{Q}}[g]$. By orthogonality, for each g we have that $g^\top = g^{-1}$ is also in G. Along with the fact that $\mathbb{Q}$ is Haar, we conclude that the matrix C is symmetric. Then for any $v \in \Theta_G$, we have $Cv = \mathbb{E}_{g\sim\mathbb{Q}}[gv] = \mathbb{E}_{g\sim\mathbb{Q}}[g^\top v] = \mathbb{E}_{g\sim\mathbb{Q}}[v] = v$. Moreover, for any $w \in \Theta_G^\perp$, we have $Cw = \mathbb{E}_{g\sim\mathbb{Q}}[gw] = \mathbb{E}_{g\sim\mathbb{Q}}[g^\top w] = \mathbb{E}_{g\sim\mathbb{Q}}[0] = 0$. Hence, C is exactly the orthogonal projection into the subspace Θ_G, which finishes the proof.

A.7 Proof of Theorem 11

In the following proof, we will use θ and W interchangeably when there is no ambiguity.

For part 1 of the theorem, we have

$$\nabla f(W, x) = \sigma'(Wx) \cdot x^\top \in \mathbb{R}^{p\times d}.$$

We then have

$$\nabla f(W,x) = \sigma'(Wx)\cdot x^\top \in \mathbb{R}^{p\times d}.$$

We can think of the Fisher information matrix $I_\theta = \mathbb{E}\nabla f(\theta,X)\nabla f(\theta,X)^\top$ as a tensor, i.e,

$$\begin{aligned} I_W &= \mathbb{E}(\sigma'(WX)\cdot X^\top)\otimes(\sigma'(WX)\cdot X^\top) \\ &= \mathbb{E}(\sigma'(WX)\otimes\sigma'(WX))\cdot(X\otimes X)^\top. \end{aligned}$$

The i,j,i',j'-th entries of this tensor are

$$I_W(i,j,i',j') = \mathbb{E}\sigma'(W_i^\top X)\sigma'(W_{i'}^\top X)\cdot X_j X_{j'}.$$

For a quadratic activation function, $\sigma(x) = x^2/2$, we can get more detailed results. We then have

$$\begin{aligned} I_W &= \mathbb{E}(WXX^\top)\otimes(WXX^\top) \\ &= (W\otimes W)\cdot\mathbb{E}(XX^\top\otimes XX^\top). \end{aligned}$$

The group acts by $gx = T_g x$, where T_g is an operator that shifts a vector circularly by g units. We can then write the neural network $f(W,x) = \sum_{i=1}^p h(W_i;x)$ as a sum, where $h(a,x) = \sigma(a^\top x)$. Therefore, the invariant function corresponding to f_W can also be written in terms of the corresponding invariant functions corresponding to the h-s:

$$\bar f(W,x) = \frac{1}{d}\sum_{g=1}^d f(W,T_g x) = \sum_{i=1}^p \bar h(W_i;x).$$

where $\bar h(a;x) = \frac{1}{d}\sum_{g=1}^d h(a;T_g x)$. We can use this representation to calculate the gradient. We first notice $\nabla h(a;x) = \sigma'(a^\top x)x$. Thus,

$$\begin{aligned} \nabla\bar h(a;x) &= \frac{1}{d}\sum_{g=1}^d \nabla h(a;T_g x) = \frac{1}{d}\sum_{g=1}^d \sigma'(a^\top T_g x)T_g x \\ &= \frac{1}{d}C_x\cdot\sigma'(C_x^\top a). \end{aligned}$$

Here C_x is the circulant matrix

$$C_x = [x, T_1 x, \dots, T_{d-1}x] = \begin{bmatrix} x_1, & x_d, & \dots, & x_{d-1} \\ x_2, & x_1, & \dots, & x_d \\ & & \cdots & \\ x_d, & x_{d-1}, & \dots, & x_1 \end{bmatrix}.$$

Hence the gradient of the invariant neural network $\bar f(W,x)$ as a matrix-vector product

$$\nabla\bar f(W,x) = \begin{bmatrix} \nabla\bar h(W_1;x)^\top \\ \cdots \\ \nabla\bar h(W_p;x)^\top \end{bmatrix} = \frac{1}{d}\begin{bmatrix} \sigma'(W_1^\top C_x)\cdot C_x^\top \\ \cdots \\ \sigma'(W_p^\top C_x)\cdot C_x^\top \end{bmatrix} = \frac{1}{d}\sigma'(WC_x)\cdot C_x^\top.$$

So the Fisher information can also expressed in terms of matrix products

$$\begin{aligned}\bar{I}_W &= \mathbb{E}(\sigma'(WC_X)\cdot C_X^\top)\otimes(\sigma'(WC_X)\cdot C_X^\top)\\ &= \mathbb{E}(\sigma'(WC_X)\otimes\sigma'(WC_X))\cdot(C_X\otimes C_X)^\top.\end{aligned}$$

For quadratic activation functions, we have

$$\begin{aligned}\bar{I}_W &= \frac{1}{d^2}\mathbb{E}(WC_XC_X^\top)\otimes(WC_XC_X^\top)\\ &= (W\otimes W)\cdot\frac{1}{d^2}\mathbb{E}(C_XC_X^\top\otimes C_XC_X^\top)\\ &= (W\otimes W)\cdot\frac{1}{d^2}\mathbb{E}(C_X\otimes C_X)\cdot(C_X\otimes C_X)^\top.\end{aligned}$$

Therefore, the efficiency gain can be characterized by the move from the 4th moment tensor of X to that of $\frac{1}{\sqrt{d}}C_X$.

Finally, if we assume $X\sim\mathcal{N}(0,I_d)$, then we can express our results in a simpler form using the discrete Fourier transform. Let F be the $d\times d$ Discrete Fourier Transform (DFT) matrix, with entries $F_{j,k}=d^{-1/2}\exp(-2\pi i/d\cdot(j-1)(k-1))$. Then Fx is called the DFT of the vector x, and $F^{-1}y=F^*y$ is called the inverse DFT. The DFT matrix is a unitary matrix with $FF^*=F^*F=I_d$. Thus $F^{-1}=F^*$. It is also a symmetric matrix with $F^\top=F$. Then the circular matrix can be diagonalized as

$$\frac{1}{\sqrt{d}}C_x=F^*\operatorname{diag}(Fx)F.$$

The eigenvalues of $d^{-1/2}C_x$ are the entries of Fx, with eigenvectors the corresponding columns of F.

So we can write, with $D:=\operatorname{diag}(FX)$,

$$\begin{aligned}d^{-1}C_X\otimes C_X &= F^*DF\otimes F^*DF\\ &= (F\otimes F)^*\cdot(D\otimes D)\cdot(F\otimes F)=F_2^*D_2F_2,\end{aligned}$$

where $F_2=F\otimes F$, and $D_2=D\otimes D$ is a diagonal matrix. So

$$\begin{aligned}d^{-2}\mathbb{E}(C_X\otimes C_X)\cdot(C_X\otimes C_X)^\top &= \mathbb{E}F_2^*D_2F_2\cdot(F_2^*D_2F_2)^\top\\ &= \mathbb{E}F_2^*D_2F_2\cdot F_2^\top D_2F_2^{*,T}\\ &= F_2^*\cdot\mathbb{E}D_2F_2^2D_2\cdot F_2^*.\end{aligned}$$

Here we used that $F=F^\top$, hence $F_2^\top=(F\otimes F)^\top=F^\top\otimes F^\top=F_2$.

Now, D_2 can be viewed as a $d^2\times d^2$ matrix, with diagonal entries $D_2(i,j,i,j)=D_iD_j=F_i^\top X\cdot F_j^\top X$, where F_i are the rows (which are also equal to the columns) of the DFT. Thus the inner expectation can be written as an elementwise product (also known as Hadamard or odot product)

$$\mathbb{E}D_2F_2^2D_2=F_2^2\odot\mathbb{E}D_2D_2^\top.$$

So we only need to calculate the 4th order moment tensor M of the Fourier transform FX,

$$M_{iji'j'}=\mathbb{E}F_i^\top X\cdot F_j^\top X\cdot F_{i'}^\top X\cdot F_{j'}^\top X.$$

Let us write $r := FX$. Then by Wick's formula,

$$\mathbb{E} f_i f_j f_{i'} f_{j'} = \mathbb{E} f_i f_j \cdot \mathbb{E} f_{i'} f_{j'} + \mathbb{E} f_i f_{j'} \cdot \mathbb{E} f_{i'} f_j + \mathbb{E} f_i f_{i'} \cdot \mathbb{E} f_i f_{j'}.$$

Now

$$\mathbb{E} f_i f_j = \mathbb{E} F_i^\top X \cdot F_j^\top X = F_i^\top \cdot \mathbb{E} X X^\top \cdot F_j = F_i^\top F_j.$$

Hence

$$M_{ij i' j'} = F_i^\top F_j \cdot F_{i'}^\top F_{j'} + F_i^\top F_{j'} \cdot F_{i'}^\top F_j + F_i^\top F_{i'} \cdot F_i^\top F_{j'}.$$

This leads to a completely explicit expression for the average information. Recall $F_2 = F \otimes F$, and M is the $d^2 \times d^2$ tensor with entries given above. Then

$$\bar{I}_W = (W \otimes W) \cdot F_2^* \cdot (F_2^2 \odot M) \cdot F_2^*.$$

A.8 Proof of Corollary 12

For notational simplicity, for a generic matrix A, we will use $A^{\otimes 2}$ to denote the tensor product $A \otimes A$. Recalling the definitions in (27), we have

$$\mathbb{E} U_W(X) = \mathbb{E}[\eta'(f(W,X))^2 (\sigma'(WX) \otimes \sigma'(WX)) \cdot (X \otimes X)^\top],$$

and similarly,

$$\begin{aligned}\mathbb{E}[U_W(X)] \Sigma_{\mathrm{ERM}} \mathbb{E}[U_W(X)] &= \mathbb{E}[v_W(X) U_W(X)] \\ &= \mathbb{E}[v_W(X) \eta'(f(W,X))^2 (\sigma'(WX) \otimes \sigma'(WX)) \cdot (X \otimes X)^\top].\end{aligned}$$

On the other hand, under the circular symmetry model (24), for the augmented estimator, by Theorem 6, we can directly compute

$$\begin{aligned}\mathbb{E}[U_W(X)] \Sigma_{\mathrm{aERM}} \mathbb{E}[U_W(X)] &= \mathbb{E}\left[\left(\mathbb{E}_G[(Y - \eta(f(W,gX))) \eta'(f(W,gX)) \nabla f(W,gX)]\right)^{\otimes 2}\right] \\ &= \mathbb{E}\left[(Y - \eta(f(W,X)))^2 \eta'(f(W,X))^2 (\mathbb{E}_G \nabla f(W,gX))^{\otimes 2}\right] \\ &= \mathbb{E}\left[v_W(X) \eta'(f(W,X))^2 (\sigma'(WC_X))^{\otimes 2} (C_X^{\otimes 2})^\top\right],\end{aligned}$$

where in the second line we used $f(W,x) = f(W,gx)$, and the last equality is by a similar computation as in the proof of Theorem 11. Then it is clear that

$$\begin{aligned}&\mathbb{E}[U_W(X)](\Sigma_{\mathrm{ERM}} - \Sigma_{\mathrm{aERM}})\mathbb{E}[U_W(X)] \\ &= \mathbb{E}\left[v_W(X) \eta'(f(W,X))^2 \times \left((\sigma'(WX))^{\otimes 2} (X^{\otimes 2})^\top - (\sigma'(WC_X))^{\otimes 2} (C_X^{\otimes 2})^\top\right)\right].\end{aligned}$$

Assuming $\sigma(x) = x^2/2$, we have

$$\begin{aligned}&\mathbb{E}[U_W(X)](\Sigma_{\mathrm{ERM}} - \Sigma_{\mathrm{aERM}})\mathbb{E}[U_W(X)] \\ &= W^{\otimes 2} \mathbb{E}\left[v_W(X) \eta'(f(W,X))^2 \left((XX^\top)^{\otimes 2} - (C_X C_X^\top)^{\otimes 2}\right)\right],\end{aligned}$$

which concludes the proof.

A.9 Proof of Proposition 13

For part 1, we have

$$\begin{aligned}
\widehat{\beta}_{\mathrm{aDIST}} &= \mathbb{E}_g\left[((Xg^\top)^\top Xg^\top)^{-1}(Xg^\top)^\top y\right] \\
&= \mathbb{E}_g\left[(gX^\top Xg^\top)^{-1}gX^\top(Xg^\top\beta+\varepsilon)\right] \\
&= \beta + \mathbb{E}_g\left[(gX^\top Xg^\top)^{-1}gX^\top\varepsilon\right] \\
&= \beta + \mathbb{E}_g\left[g^{-1}(X^\top X)^{-1}g^{-1}gX^\top\varepsilon\right] \\
&= \beta + \mathbb{E}_g[g^{-1}](X^\top X)^{-1}X^\top\varepsilon \\
&= \beta + \mathcal{G}^\top(X^\top X)^{-1}X^\top\varepsilon.
\end{aligned}$$

Let $X = UDV^\top$ be a SVD of X, where $V \in \mathbb{R}^{p\times p}$ is unitary. Note that $\widehat{\beta}_{\mathrm{aDIST}}$ is unbiased, so its ℓ_2 risk is

$$\begin{aligned}
r_{\mathrm{aDIST}} &= \gamma^2\,\mathrm{tr}(\mathrm{Var}(\widehat{\beta}_{\mathrm{aDIST}})) = \gamma^2\,\mathrm{tr}(\mathcal{G}^\top(X^\top X)^{-1}\mathcal{G}) \\
&= \gamma^2\,\mathrm{tr}(\mathcal{G}^\top VD^{-2}V^\top\mathcal{G}) = \gamma^2\,\mathrm{tr}(D^{-2}V^\top\mathcal{G}\mathcal{G}^\top V) \\
&= \gamma^2\sum_{j=1}^p d_j^{-2}e_j^\top V^\top\mathcal{G}\mathcal{G}^\top Ve_j = \gamma^2\sum_{j=1}^p d_j^{-2}\|\mathcal{G}^\top v_j\|_2^2,
\end{aligned}$$

where $v_j \in \mathbb{R}^p$ is j-th eigenvector of $X^\top X$ and d_j^2 is j-th eigenvalue of $X^\top X$. As a comparison, for the usual ERM, we have

$$\widehat{\beta}_{\mathrm{ERM}} = (X^\top X)^{-1}X^\top y = \beta + (X^\top X)^{-1}X^\top\varepsilon,$$

so its ℓ_2 risk is

$$r_{\mathrm{ERM}} = \gamma^2\,\mathrm{tr}((X^\top X)^{-1}) = \gamma^2\sum_{j=1}^p d_j^{-2}.$$

So part 1 is proved.

We now prove part 2. For $r_{\mathrm{aDIST}} \le r_{\mathrm{ERM}}$ we need to show

$$\mathrm{tr}((X^\top X)^{-1}\mathcal{G}\mathcal{G}^\top) \le \mathrm{tr}((X^\top X)^{-1}).$$

A sufficient condition is that, in the partial ordering of positive semidefinite matrices,

$$\mathcal{G}\mathcal{G}^\top \le I_p.$$

This is equivalent to the claim that for all v $\|\mathbb{E}_g g^\top v\|^2 \le \|v\|^2$. However, by Jensen's inequality, $\|\mathbb{E}_g g^\top v\|^2 \le \mathbb{E}_G\|g^\top v\|^2$. Since G is a subgroup of the orthogonal group, we have $\|g^\top v\|^2 = \|v\|^2$, which finishes the proof for part 2.

We finally prove part 3. We assume G is the permutation group. This group is clearly a subgroup of the orthogonal group. Note that invariance w.r.t. G implies that the true parameter is a multiple of the all ones vector: $\beta = 1_p b$. So we have

$$\widehat{\beta}_{\text{cERM}} = 1_p \widehat{b}, \qquad \widehat{b} = \arg\min \|y - X 1_p b\|_2^2.$$

Solving the least-squares equation gives

$$\widehat{b} = \frac{1^\top X^\top y}{1_p^\top X^\top X 1_p}.$$

The risk of estimating b is then $\gamma^2 (1_p^\top X^\top X 1_p)^{-1}$, so that the risk of estimating β by $1_p \widehat{b}$ is

$$r_{\text{cERM}} = \gamma^2 p (1_p^\top X^\top X 1_p)^{-1}.$$

Finally, we have

$$r_{\text{aDIST}} = \frac{\gamma^2}{p^2} \operatorname{tr}(1_p 1_p^\top (X^\top X)^{-1} 1_p 1_p^\top) = \frac{\gamma^2}{p} 1_p^\top (X^\top X)^{-1} 1_p,$$

which is equal to r_{cERM} if $X^\top X = I_p$.

Appendix B. Proofs for results under approximate invariance

B.1 Proof of Lemma 15

For part 1, we have

$$\begin{aligned}
\|\mathbb{E}_X \mathbb{E}_g f(gX) - \mathbb{E}_X f(X)\|_2 &= \sup_{\|v\|_2 \le 1} \mathbb{E}_g \mathbb{E}_X \langle v, f(gX) - \mathbb{E}_X f(X) \rangle \\
&\le \sup_{\|v\|_2 \le 1} \mathbb{E}_g \|v\|_2 \mathcal{W}_1(f(gX), f(X)) \\
&= \mathbb{E}_g \mathcal{W}_1(f(gX), f(X)),
\end{aligned}$$

where the inequality is due to Kantorovich-Rubinstein theorem, i.e., the dual representation of the W_1 metric (see. e.g., Villani 2003).

For part 2, by law of total variance, we have

$$\operatorname{Cov}_X \bar{f}(X) - \operatorname{Cov}_X f(X) = -\mathbb{E}_X \operatorname{Cov}_g f(gX) + \Delta_1 + \Delta_2,$$

where

$$\begin{aligned}
\Delta_1 &= \mathbb{E}_{(X,g)} f(gX) f(gX)^\top - \mathbb{E}_X f(X) f(X)^\top \\
\Delta_2 &= \mathbb{E}_X f(X) \mathbb{E}_X f(X)^\top - \mathbb{E}_{(X,g)} f(gX) \mathbb{E}_{(X,g)} f(gX)^\top
\end{aligned}$$

For any non-zero vector v, we have

$$\begin{aligned}
|v^\top \Delta_1 v| &= \left| \mathbb{E}_g \mathbb{E}_X \left[\langle v, f(gX) \rangle^2 - \langle v, f(X) \rangle^2 \right] \right| \\
&\le \mathbb{E}_g \left| \mathbb{E}_X \left[\langle v, f(gX) \rangle^2 - \langle v, f(X) \rangle^2 \right] \right|.
\end{aligned}$$

The Lipschitz constant for the function $w \mapsto \langle v, w\rangle^2$, where $w \in \text{Range}(f)$, is bounded above by $2\|v\|_2^2\|f\|_\infty$. Invoking Kantorovich-Rubinstein theorem again, we have

$$|v^\top \Delta_1 v| \le 2\|v\|_2^2\|f\|_\infty \mathbb{E}_g \mathcal{W}_1(f(gX), f(X)).$$

Similarly, we have

$$\begin{aligned}
|v^\top \Delta_2 v| &= \left| \left(\mathbb{E}_X \langle v, f(X)\rangle \right)^2 - \left(\mathbb{E}_{(X,g)} \langle v, f(gX)\rangle \right)^2 \right| \\
&\le 2\|f\|_\infty \left| \mathbb{E}_g \mathbb{E}_X \left[\langle v, f(X)\rangle - \langle v, f(gX)\rangle \right] \right| \\
&\le 2\|f\|_\infty \mathbb{E}_g \left| \mathbb{E}_X \left[\langle v, f(X)\rangle - \langle v, f(gX)\rangle \right] \right| \\
&\le 2\|v\|_2^2\|f\|_\infty \mathbb{E}_g \mathcal{W}_1(f(gX), f(X)).
\end{aligned}$$

Part 2 is then proved by recalling the definition of the Loewner order.

For part 3, we have

$$\mathbb{E}_X \varphi(\bar{f}(X)) - \mathbb{E}_X \varphi(f(X)) = \mathbb{E}_X \varphi(\bar{f}(X)) - \mathbb{E}_X \varphi \,\bar{\circ}\, f(X) + \mathbb{E}_g \mathbb{E}_X \varphi \circ f(gX) - \mathbb{E}_X \varphi(f(X)).$$

We finish the proof by noting that

$$\left| \mathbb{E}_g \mathbb{E}_X \varphi \circ f(gX) - \mathbb{E}_X \varphi(f(X)) \right| \le \|\varphi\|_{\text{Lip}} \mathcal{W}_1(f(gX), f(X)).$$

B.2 Proof of Proposition 16

For notational simplicity, we let $\bar{f} = \widehat{\theta}_G$ and $f = \widehat{\theta}$. Then using bias-variance decomposition, we have

$$\text{MSE}(\bar{f}) - \text{MSE}(f) = B + V,$$

where

$$\begin{aligned}
B &= \|\text{Bias}(\bar{f})\|_2^2 - \|\text{Bias}(f)\|_2^2 \\
V &= \text{tr}(\text{Cov}_X \bar{f}(X)) - \text{tr}(\text{Cov}_X f(X)).
\end{aligned}$$

We first analyze the bias term. Note that by Lemma 15, we have

$$\begin{aligned}
\left| \|\text{Bias}(\bar{f})\|_2 - \|\text{Bias}(f)\|_2 \right| &\le \|\mathbb{E}_X \bar{f}(X) - \mathbb{E}_X f(X)\|_2 \\
&\le \mathbb{E}_g \mathcal{W}_1(f(gX), f(X)).
\end{aligned}$$

Hence

$$\begin{aligned}
|B| &\le \left(\|\text{Bias}(\bar{f})\|_2 + \|\text{Bias}(f)\|_2 \right) \left| \|\text{Bias}(\bar{f})\|_2 - \|\text{Bias}(f)\|_2 \right| \\
&\le \left(\|\mathbb{E}_X \bar{f}(X) - \mathbb{E}_X f(X)\|_2 + 2\|\text{Bias}(f)\|_2 \right) \cdot \|\mathbb{E}_X \bar{f}(X) - \mathbb{E}_X f(X)\|_2 \\
&\le \left(\mathbb{E}_G \mathcal{W}_1(f(gX), f(X)) + 2\|\text{Bias}(f)\|_2 \right) \cdot \mathbb{E}_g \mathcal{W}_1(f(gX), f(X)).
\end{aligned}$$

The variance term V can be bounded by the following arguments. We have

$$\begin{aligned}
|\operatorname{tr}(\Delta_1)| &= \left|\mathbb{E}_g \mathbb{E}_X \left[\|f(gX)\|_2^2 - \|f(X)\|_2^2\right]\right| \\
&\leq 2\|f\|_\infty \mathbb{E}_g \left|\mathbb{E}_X \left[\|f(gX)\|_2 - \|f(X)\|_2\right]\right| \\
&\leq 2\|f\|_\infty \mathbb{E}_g \mathcal{W}_1(f(gX), f(X)).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
|\operatorname{tr}(\Delta_2)| &= \left|\|\mathbb{E}_X f(X)\|_2^2 - \|\mathbb{E}_{X,g} f(gX)\|_2^2\right| \\
&\leq 2\|f\|_\infty \left|\|\mathbb{E}_X f(X)\|_2 - \|\mathbb{E}_{X,g} f(gX)\|_2\right| \\
&\leq 2\|f\|_\infty \|\mathbb{E}_X f(X) - \mathbb{E}_X \bar{f}(X)\|_2 \\
&\leq 2\|f\|_\infty \mathbb{E}_g \mathcal{W}_1(f(gX), f(X)),
\end{aligned}$$

where the last inequality is due to Lemma 15.

Combining the bound for B and V gives the desired result.

B.3 Proof of Theorem 17

We first prove a useful lemma.

Lemma 20 (Triangle inequality/Tensorization) *For two random vectors $(X_1, ..., X_n)$, $(Y_1, ..., Y_n) \in \mathcal{X}^n$, we denote the joint laws as μ^n, ν^n respectively, and the marginal laws as $\{\mu_i\}_1^n, \{\nu_i\}_1^n$ respectively. We have*

$$\mathcal{W}_{d^n}(\mu^n, \nu^n) \leq \sum_i \mathcal{W}_d(\mu_i, \nu_i).$$

Proof By Kantorovich duality (see, e.g., Villani 2003), for each coordinate, we can choose optimal couplings $(X_i^*, Y_i^*) \in \Pi(\mu_i, \nu_i)$ s.t. $\mathcal{W}_d(X_i, Y_i) = \mathbb{E}d(X_i^*, Y_i^*)$. We then conclude that proof by noting that $(\{X_i^*\}_1^n, \{Y_i^*\}_1^n) \in \Pi(\mu^n, \nu^n)$. ■

Now we are ready to give the proof. We will do the proof for a general metric d. The desired result is a special case when d is the Euclidean metric.

The results concerning $\widehat{\theta}_n$ is classical. We present a proof here for completeness. We recall the classical approach of decomposing the generalization error into terms that can be bounded via concentration and Rademacher complexity (Bartlett and Mendelson, 2002; Shalev-Shwartz and Ben-David, 2014):

$$\mathbb{E}L(\widehat{\theta}_n, X) - \mathbb{E}L(\theta_0, X) = \mathbb{E}L(\widehat{\theta}_n, X) - \frac{1}{n}\sum_{i=1}^n L(\widehat{\theta}_n, X_i) + \frac{1}{n}\sum_{i=1}^n L(\widehat{\theta}_n, X_i) - \mathbb{E}L(\theta_0, X).$$

Hence we arrive at

$$\mathbb{E}L(\widehat{\theta}_n, X) - \mathbb{E}L(\theta_0, X) \leq \mathbb{E}L(\widehat{\theta}_n, X) - \frac{1}{n}\sum_{i=1}^n L(\widehat{\theta}_n, X_i) + \frac{1}{n}\sum_{i=1}^n L(\theta_0, X_i) - \mathbb{E}L(\theta_0, X)$$
$$\leq \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n L(\theta, X_i) - \mathbb{E}L(\theta, X)\right| + \left(\frac{1}{n}\sum_{i=1}^n L(\theta_0, X_i) - \mathbb{E}L(\theta_0, X)\right),$$

where the first inequality is because $\widehat{\theta}_n$ is a minimizer of the empirical risk. By McDiarmid's inequality, we have

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n L(\theta_0, X_i) - \mathbb{E}L(\theta_0, X) > t) \leq \exp\{-2nt^2\}.$$

So w.p. at least $1 - \delta/2$, we have

$$\frac{1}{n}\sum_{i=1}^n L(\theta_0, X_i) - \mathbb{E}L(\theta_0, X) \leq \sqrt{\frac{\log 2/\delta}{2n}}.$$

It remains to control

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n L(\theta, X_i) - \mathbb{E}L(\theta, X)\right|.$$

We bound the above quantity using Rademacher complexity. The arguments are standard and can be found in many textbooks (see, e.g., Shalev-Shwartz and Ben-David 2014). Since we've assumed $L(\theta, x) \in [0, 1]$, for two data sets $\{X_i\}_1^n$ and $\{\tilde{X}_i\}_1^n$ which only differ in the i-th coordinate, we have

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n L(\theta, X_i) - \mathbb{E}L(\theta, X)\right| - \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n L(\theta, \tilde{X}_i) - \mathbb{E}L(\theta, X)\right|$$
$$\leq \frac{1}{n}\sup_{\theta\in\Theta}|L(\theta, X_i) - L(\theta, \tilde{X}_i)| \leq \frac{1}{n}.$$

By McDiarmid's inequality, we have

$$\mathbb{P}\bigg(\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n L(\theta, X_i) - \mathbb{E}L(\theta, X)\right| - \mathbb{E}\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n L(\theta, X_i) - \mathbb{E}L(\theta, X)\right|\right] \geq t\bigg)$$
$$\leq \exp\{-2nt^2\}.$$

It follows that w.p. $1 - \delta/2$, we have

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n L(\theta, X_i) - \mathbb{E}L(\theta, X)\right| - \mathbb{E}\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n L(\theta, X_i) - \mathbb{E}L(\theta, X)\right|\right] \leq \sqrt{\frac{\log 2/\delta}{2n}}.$$

A standard symmetrization argument then shows that

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n L(\theta, X_i) - \mathbb{E}L(\theta, X)\right|\right] \leq 2\mathcal{R}_n(L \circ \Theta),$$

where the Rademacher complexity of the function class $L \circ \Theta = \{x \mapsto L(\theta, x) : \theta \in \Theta\}$ is defined as

$$\mathcal{R}_n(L \circ \Theta) = \mathbb{E} \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i L(\theta, X_i) \right|,$$

where the expectation is taken over both the data and IID Rademacher random variables ε_i, which are independent of the data. Summarizing the above computations (along with a union bound) finishes the proof for $\widehat{\theta}_n$.

Now we consider the results for $\widehat{\theta}_{n,G}$ under approximate invariance. We start by doing a similar decomposition

$$\mathbb{E}L(\widehat{\theta}_G, X) - \mathbb{E}L(\theta_0, X) = \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + \mathrm{V},$$

where

$$\begin{aligned}
\mathrm{I} &= \mathbb{E}L(\widehat{\theta}_G, X) - \mathbb{E}\mathbb{E}_G L(\widehat{\theta}_G, gX) \\
\mathrm{II} &= \mathbb{E}\mathbb{E}_G L(\widehat{\theta}_G, gX) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_G L(\widehat{\theta}_G, gX_i) \\
\mathrm{III} &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_G L(\widehat{\theta}_G, gX_i) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_G L(\theta_0, gX_i) \\
\mathrm{IV} &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_G L(\theta_0, gX_i) - \mathbb{E}\mathbb{E}_G L(\theta_0, gX) \\
\mathrm{V} &= \mathbb{E}\mathbb{E}_G L(\theta_0, gX) - \mathbb{E}L(\theta_0, X).
\end{aligned}$$

By construction, we have $\mathrm{III} \leq 0$ and

$$\mathrm{II} \leq \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_G L(\theta, gX_i) - \mathbb{E}\mathbb{E}_G L(\theta, gX) \right|.$$

Moreover, we have

$$\mathrm{I} + \mathrm{V} \leq 2 \sup_{\theta \in \Theta} \left| \mathbb{E}L(\theta, X) - \mathbb{E}\mathbb{E}_G L(\theta, gX) \right|,$$

which is equal to zero under exact invariance $gX =_d X$.

The term $\mathrm{II} + \mathrm{IV}$ is taken care of by essentially the same arguments as the proof for $\widehat{\theta}_n$. One uses concentration to bound IV and uses Rademacher complexity to bound II. These arguments give

$$\mathrm{II} + \mathrm{IV} \leq 2\mathcal{R}_n(\bar{L} \circ \Theta) + \sqrt{\frac{2 \log 2/\delta}{n}}$$

w.p. at least $1 - \delta$, where

$$\mathcal{R}_n(\bar{L} \circ \Theta) = \mathbb{E} \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mathbb{E}_G L(\theta, gX_i) \right|.$$

Now we have

$$\mathcal{R}_n(\bar{L}\circ\Theta)-\mathcal{R}_n(L\circ\Theta)\le\Delta+\mathbb{E}_g\left[\mathbb{E}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_iL(\theta,gX_i)\right|-\mathbb{E}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_iL(\theta,X_i)\right|\right],$$

where we recall

$$\Delta=\mathbb{E}\sup_\theta|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{E}_gL(\theta,gX_i)|-\mathbb{E}\mathbb{E}_g\sup_{\theta\in\Theta}|\frac{1}{n}\varepsilon_iL(\theta,gX_i)|\le 0$$

by Jensen's inequality.

By our assumption, for any $x,\tilde{x}\in\mathcal{X},\theta\in\Theta$, we have

$$L(\theta,x)-L(\theta,\tilde{x})\le\|L\|_{\mathrm{Lip}}\cdot d(x,\tilde{x})$$

for some constant $\|L\|_{\mathrm{Lip}}$. For a fixed vector $(\varepsilon_1,...,\varepsilon_n)$, consider the function

$$h:(x_1,...,x_n)\mapsto\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_iL(\theta,x_i)\right|.$$

We have

$$\begin{aligned}|h(x_1,...,x_n)-h(y_1,...,y_n)|&\le\frac{1}{n}\sup_{\theta\in\Theta}\left|\sum_{i=1}^n\varepsilon_iL(\theta,x_i)-\varepsilon_iL(\theta,y_i)\right|\\&\le\frac{1}{n}\|L\|_{\mathrm{Lip}}\cdot\sum_i d(x_i,y_i).\end{aligned}$$

That is, the function $h:\mathcal{X}^n\to\mathbb{R}$ is $(\|L\|_{\mathrm{Lip}}/n)$-Lipschitz w.r.t. the l.s.c. metric d_n, defined by $d_n(\{x_i\}_1^n,\{y_i\}_1^n)=\sum_i d(x_i,y_i)$. Applying the tensorization lemma and Kantorovich-Rubinstein theorem, for arbitrary random vectors $(X_1,...,X_n)$ and $(Y_1,...,Y_n)$, we have

$$\begin{aligned}|\mathbb{E}h(X_1,...,X_n)-h(Y_1,...,Y_n)|&\le\frac{1}{n}\|L\|_{\mathrm{Lip}}\cdot\mathcal{W}_{d^n}(\mu^n,\nu^n)\\&\le\frac{1}{n}\|L\|_{\mathrm{Lip}}\cdot\sum_{i=1}^n\mathcal{W}_d(X_i,Y_i).\end{aligned}$$

Hence we arrive at

$$\begin{aligned}\mathcal{R}_n(\bar{L}\circ\Theta)-\mathcal{R}_n(L\circ\Theta)&\le\Delta+\|L\|_{\mathrm{Lip}}\cdot\frac{1}{n}\sum_i\mathbb{E}_g\mathcal{W}_d(X_i,gX_i)\\&=\|L\|_{\mathrm{Lip}}\cdot\mathbb{E}_g\mathcal{W}_d(X,gX).\end{aligned}$$

Summarizing the above computations, we have

$$\mathrm{II}+\mathrm{IV}\le 2\mathcal{R}_n(L\circ\Theta)+2\|L\|_{\mathrm{Lip}}\cdot\mathbb{E}_G\mathcal{W}_d(X,gX)+\sqrt{\frac{2\log 2/\delta}{n}}$$

w.p. at least $1-\delta$.

We now bound I + V. We have

$$\begin{aligned}\mathrm{I}+\mathrm{V} &\le 2\sup_{\theta\in\Theta}\left|\mathbb{E}L(\theta,X)-\mathbb{E}\mathbb{E}_G L(\theta,gX)\right| \\ &\le 2\sup_{\theta\in\Theta}\mathbb{E}_G\left|\mathbb{E}L(\theta,X)-\mathbb{E}L(\theta,gX)\right| \\ &\le 2\|L\|_{\mathrm{Lip}}\cdot\mathcal{W}_d(X,gX).\end{aligned}$$

Combining the bounds for the five terms gives the desired result.

B.4 Proof of Theorem 18

Recall that

$$\theta_G=\arg\min_{\theta\in\Theta}\mathbb{E}\mathbb{E}_g L(\theta,gX).$$

By our assumptions, we can apply Theorem 5.23 of Van der Vaart (1998) to obtain the Bahadur representation:

$$\sqrt{n}(\widehat{\theta}_G-\theta_G)=\frac{1}{\sqrt{n}}V_G^{-1}\sum_{i=1}^{n}\nabla\mathbb{E}_G L(\theta_0,gX_i)+o_p(1),$$

so that we get

$$\sqrt{n}(\widehat{\theta}_G-\theta_G)\Rightarrow\mathcal{N}\left(0,V_G^{-1}\mathbb{E}\left[\nabla\mathbb{E}_G L(\theta_G,gX)(\mathbb{E}_G\nabla L(\theta_G,gX))^\top\right]V_G^{-1}\right).$$

To simplify notations, we let

$$C_0=\mathrm{Cov}_X(\nabla L(\theta_0,X)),\quad C_G=\mathrm{Cov}_X(\nabla\mathbb{E}_g L(\theta_G,gX)).$$

By bias-variance decomposition, we have

$$\begin{aligned}\mathrm{MSE}_0&=n^{-1}\,\mathrm{tr}(V_0^{-1}C_0V_0^{-1})\\ \mathrm{MSE}_G&=n^{-1}\,\mathrm{tr}(V_G^{-1}C_GV_G^{-1})+\|\theta_G-\theta_0\|^2.\end{aligned}$$

We have

$$\begin{aligned}\mathrm{tr}(V_G^{-1}C_GV_G^{-1})-\mathrm{tr}(V_0^{-1}C_0V_0^{-1})&=\langle C_G,V_G^{-2}\rangle-\langle C_0,V_0^{-2}\rangle\\ &=\langle C_G-C_0+C_0,V_G^{-2}\rangle-\langle C_0,V_0^{-2}\rangle\\ &=\langle C_G-C_0,V_G^{-2}\rangle+\langle C_0,V_G^{-2}-V_0^{-2}\rangle.\end{aligned}$$

We let $M_0(X)=\nabla L(\theta_0,X)\nabla L(\theta_0,X)^\top$ and $M_G(X)=\nabla L(\theta_G,X)\nabla L(\theta_G,X)^\top$. Then we have

$$\begin{aligned}C_G-C_0&=C_G-\mathbb{E}_X M_G(X)+\mathbb{E}_X M_G(X)-C_0\\ &=\left(C_G-\mathbb{E}_X\mathbb{E}_g M_G(gX)\right)+\mathbb{E}_g\mathbb{E}_X\left[M_G(gX)-M_G(X)\right]+\mathbb{E}_X\left[M_G(X)-M_0(X)\right]\\ &=-\mathbb{E}_X\mathrm{Cov}_g(\nabla L(\theta_G,gX))+\mathbb{E}_g\mathbb{E}_X\left[M_G(gX)-M_G(X)\right]+\mathbb{E}_X\left[M_G(X)-M_0(X)\right].\end{aligned}$$

Hence we arrive at

$$n(\mathrm{MSE}_G - \mathrm{MSE}_0) = -\left\langle \mathbb{E}_X \mathrm{Cov}_G(\nabla L(\theta_G, gX)), V_G^{-2} \right\rangle + \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV},$$

where

$$\begin{aligned}
\mathrm{I} &= n\|\theta_G - \theta_0\|^2 \\
\mathrm{II} &= \mathbb{E}_g \mathbb{E}_X \left\langle M_G(gX) - M_G(X), V_G^{-2} \right\rangle \\
\mathrm{III} &= \mathbb{E}_X \left\langle M_G(X) - M_0(X), V_G^{-2} \right\rangle \\
\mathrm{IV} &= \langle C_0, V_G^{-2} - V_0^{-2} \rangle,
\end{aligned}$$

and this is the desired result.

B.5 Proof of Theorem 19

We seek to control the misclassification error of the two-layer net at step k. By Markov's inequality, for a new sample (X, Y) from the data distribution, we have

$$\begin{aligned}
\mathbb{P}(Yf(x; W_k, a) \le 0) &= \mathbb{P}\left(\frac{1}{1 + e^{Yf(X; W_k, a)}} \ge \frac{1}{2} \right) \\
&\le 2\mathbb{E}\Big[-\ell'(Yf(X; W_k, a)) \Big].
\end{aligned}$$

The population quantity in the RHS is decomposed by

$$\mathbb{E}\Big[-\ell'(Yf(X; W_k, a)) \Big] = \mathrm{I} + \mathrm{II},$$

where

$$\mathrm{I} = \frac{1}{n} \sum_{i \in [n]} \mathbb{E}_g \Big[-\ell'(Y_i f_{i,g}(W_k)) \Big]$$

and

$$\mathrm{II} = \mathbb{E}\Big[-\ell'(Yf(X; W_k, a)) \Big] - \frac{1}{n} \sum_{i \in [n]} \mathbb{E}_g \Big[-\ell'(Y_i f_{i,g}(W_k)) \Big]$$

The first term (optimization error) is controlled by calculations based on the Neural Tangent Kernel. The second term (generalization error) is controlled via Rademacher complexity.

We first control the first term (optimization error). In fact, everything is set up so that we can directly invoke Theorem 2.2 of Ji and Telgarsky (2019). We note that their result holds for any fixed dataset and there is no independence assumption. This gives the following result:

Proposition 21 *Given $\varepsilon \in (0,1), \delta \in (0,1/3)$. Let*

$$\lambda = \frac{\sqrt{2\log(4n|G|/\delta)} + \log(4/\varepsilon)}{\gamma/4}, \qquad M = \frac{4096\lambda^2}{\gamma^6}.$$

For any $m \geq M$ and any constant step size $\eta \leq 1$, w.p. $1-3\delta$ over the random initialization, we have

$$\frac{1}{T}\sum_{t<T} \bar{R}_n(W_t) \leq \varepsilon, \qquad T = \lceil 2\lambda^2/(n\varepsilon) \rceil.$$

Moreover, for any $0 \leq t < T$ and any $1 \leq s \leq m$, we have

$$\|w_{s,t} - w_{s,0}\|_2 \leq \frac{4\lambda}{\gamma\sqrt{m}},$$

where $w_{s,t}$ is the s-th row of the weight matrix at step t.

Proof This is a direct corollary of Theorem 2.2 in Ji and Telgarsky (2019). ■

Assume the above event happens. Since we've chosen k to be the best iteration (with the lowest empirical loss) in the first T steps. Then with the same probability as above, we have $\bar{R}_n(W_k) \leq \varepsilon$. Now, let us note that the logistic loss satisfies the following fundamental self-consistency bound: $-\ell' \leq \ell$. This shows that if the loss is small, then the magnitude of the derivative is also small. Thus on the same event, we have that the term I is also bounded,

$$\mathrm{I} \leq \bar{R}_n(W_k) \leq \varepsilon.$$

We then control the second term (generalization error). The calculations below are similar to the proof of Theorem 4.4. We begin by decomposing

$$\mathrm{II} = \mathrm{II.1} + \mathrm{II.2},$$

where

$$\mathrm{II.1} = \mathbb{E}\Big[-\ell'(Yf(X;W_k,a))\Big] - \mathbb{E}\mathbb{E}_g\Big[-\ell'(Yf(gX;W_k,a))\Big]$$

and

$$\mathrm{II.2} = \mathbb{E}\mathbb{E}_g\Big[-\ell'(Yf(gX;W_k,a))\Big] - \frac{1}{n}\sum_{i\in[n]} \mathbb{E}_g\Big[-\ell'(Y_i f(gX_i;W_k,a))\Big].$$

We control term II.1 by exploiting the closedness between the distribution of (X,Y) and that of (gX,Y). Note that the Lipschitz constant of the map $x \mapsto -\ell'(yf(x;W_k,a))$ (w.r.t.

the Euclidean metric on $\mathbb{R}^d$) can be computed by:

$$\begin{aligned}
|\ell'(yf(x;W_k,a)) - \ell'(yf(\tilde{x};W_k,a))| &\le \frac{1}{4}|f(x;W_k,a) - f(\tilde{x};W_k,a)| \\
&= \frac{1}{4}\left|\frac{1}{\sqrt{m}}\sum_{s\in[m]}\sigma(w_{s,k}^\top x) - \frac{1}{\sqrt{m}}\sum_{s\in[m]}\sigma(w_{s,k}^\top \tilde{x})\right| \\
&\le \frac{1}{4}\frac{1}{\sqrt{m}}\sum_{s\in[m]}|w_{s,k}^\top(x-\tilde{x})| \\
&\le \frac{1}{4}\frac{1}{\sqrt{m}}\sum_{s\in[m]}\|w_{s,k}\|_2\|x-\tilde{x}\|_2 \\
&\le \frac{1}{4}\frac{1}{\sqrt{m}}\sum_{s\in[m]}(\|w_{s,0}\|_2 + \|w_{s,0}-w_{s,k}\|_2)\|x-\tilde{x}\|_2 \\
&\le \frac{1}{4}\left(\rho\sqrt{m} + \frac{1}{\sqrt{m}}\sum_{s\in[m]}\|w_{s,0}\|_2\right)\|x-\tilde{x}\|_2,
\end{aligned}$$

where $\rho = \frac{4\lambda}{\gamma\sqrt{m}}$ and the last inequality is by Proposition 21. Note that each $\|w_{s,0}\|_2$ is 1-subgaussian as a 1-Lipschitz function of a Gaussian random vector (for example, by Theorem 2.1.12 of Tao 2012), so that

$$\mathbb{P}\left(\frac{1}{\sqrt{m}}\sum_{s\in[m]}\|w_{s,0}\|_2 - \mathbb{E}\left[\frac{1}{\sqrt{m}}\sum_{s\in[m]}\|w_{s,0}\|_2\right] \ge t\right) \le e^{-t^2/2}.$$

Hence w.p. at least $1-\delta$, we have

$$\begin{aligned}
\frac{1}{\sqrt{m}}\sum_{s\in[m]}\|w_{s,0}\|_2 &\le \mathbb{E}\left[\frac{1}{\sqrt{m}}\sum_{s\in[m]}\|w_{s,0}\|_2\right] + \sqrt{2\log\frac{1}{\delta}} \\
&\le \frac{1}{\sqrt{m}}\sum_{s\in[m]}\sqrt{\mathbb{E}\|w_{s,0}\|_2^2} + \sqrt{2\log\frac{1}{\delta}} \\
&= \sqrt{md} + \sqrt{2\log 1/\delta}.
\end{aligned}$$

So w.p. at least $1-\delta$, the Lipschitz constant of the map $x \mapsto -\ell'(yf(x;W_k,a))$ is bounded above by

$$\frac{1}{4}\left(\frac{4\lambda}{\gamma} + \sqrt{md} + \sqrt{2\log 1/\delta}\right).$$

Assume the above event happens (along with the previous event, the overall event happens w.p. at least $1-4\delta$). This information allows us to exploit the closeness between $X|Y$ and

$gX|Y$. We have

$$\begin{aligned}\text{II.1} &= \mathbb{E}_Y \mathbb{E}_{g\sim\mathbb{Q}}\left[\mathbb{E}_{X|Y}[-\ell'(Yf(X;W_k,a))] - \mathbb{E}_{gX|Y}[-\ell'(Yf(gX;W_k,a))]\right] \\ &\leq \mathbb{E}_Y \mathbb{E}_{g\sim\mathbb{Q}}\left[\frac{1}{4}\left(\frac{4\lambda}{\gamma} + \sqrt{md} + \sqrt{2\log 1/\delta}\right)\cdot \mathcal{W}_1(X|Y, gX|Y)\right] \\ &= \frac{1}{4}\left(\frac{4\lambda}{\gamma} + \sqrt{md} + \sqrt{2\log 1/\delta}\right)\cdot \mathbb{E}_Y \mathbb{E}_{g\sim\mathbb{Q}} \mathcal{W}_1(X|Y, gX|Y),\end{aligned}$$

where we let $X|Y$ to denote the conditional distribution of X given Y, and the inequality is by the dual representation of the Wasserstein distance. Note that under exact invariance, II.1 $= 0$.

The term II.2 is controlled by standard results on Rademacher complexity. Indeed, by the same arguments as in the proof of Theorem 6.4 of the main manuscript, w.p. at least $1-\delta$, we have

$$\text{II.2} \leq 2\bar{\mathcal{R}}_n + \sqrt{\frac{\log 2/\delta}{2n}}.$$

Taking a union bound (now w.p. at least $1-5\delta$), we have proved the generalization error bound.

Finally, we prove the bound on $\bar{\mathcal{R}}_n - \mathcal{R}_n$. Under exact invariance, Jensen's inequality gives $\bar{\mathcal{R}}_n \leq \mathcal{R}_n$. However, under approximate invariance, we have an extra bias term. We have

$$\bar{\mathcal{R}}_n - \mathcal{R}_n = \Delta + \mathbb{E}\mathbb{E}_g \sup_{W\in\mathscr{W}_\rho}\left|\frac{1}{n}\varepsilon_i\left[-\ell'(Y_i f_{i,g}(W))\right]\right| - \mathcal{R}_n.$$

where

$$\Delta = \mathbb{E}\sup_{W\in\mathscr{W}_\rho}\left|\frac{1}{n}\varepsilon_i \mathbb{E}_g\left[-\ell'(Y_i f_{i,g}(W))\right]\right| - \mathbb{E}\mathbb{E}_g \sup_{W\in\mathscr{W}_\rho}\left|\frac{1}{n}\varepsilon_i\left[-\ell'(Y_i f_{i,g}(W))\right]\right| \leq 0$$

by Jensen's inequality. Now by the computations when bounding term II.1 and the arguments in the proof of Theorem 4.4, we have

$$\begin{aligned}\mathbb{E}\mathbb{E}_g \sup_{W\in\mathscr{W}_\rho}\left|\frac{1}{n}\varepsilon_i\left[-\ell'(Y_i f_{i,g}(W))\right]\right| - \mathcal{R}_n \leq \frac{1}{4}\left(\frac{4\lambda}{\gamma} + \sqrt{md} + \sqrt{2\log 1/\delta}\right) \\ \cdot \mathbb{E}_Y \mathbb{E}_{g\sim\mathbb{Q}} \mathcal{W}_1(X|Y, gX|Y)\end{aligned}$$

w.p. at least $1-\delta$. Combining the above bounds finishes the proof.

Appendix C. Invariant MLE

Another perspective to exploit invariance is that of invariant representations. The natural question is, how can we work with invariant representations, and what are the limits of information we can extract from them?

Suppose therefore that in our model it is possible to choose a representation $T(x)$ such that $(T(x), 0_m) \in G\cdot x$ for all x(where 0_m is the zero vector with m entries). Thus, T

chooses a representative from each orbit. This is equivalent to $(T(x), 0_m) = g_0(x) \cdot x$, for some specific $g_0(x) \in G$. Suppose $T(\cdot), g(\cdot)$ satisfy sufficient regularity conditions, such as smoothness. For example, when G is the orthogonal rotation group $O(d)$, we can take $T(x) := \|x\|_2$, and g any orthogonal rotation such that $g_0(x) = (\|x\|_2, 0_{d-1})$.

How can we estimate the parameters θ based on this representation? A natural approach is to construct the MLE based on the data $T(X_1), \ldots, T(X_n)$. We can also construct invariant ERM using the same principle, but we will focus on MLE first. Let therefore Q_θ be the induced distribution of $T(X)$, when $X \sim P_\theta$, and assume it has a density q_θ with respect to Lebesgue measure on a potentially lower dimensional Euclidean subspace (say d' dimensional, where d is original dimension and $m = d - d'$). We can construct the invariant MLE (iMLE):

$$\widehat{\theta}_{iMLE,n} = \arg\max_\theta \sum_{i \in [n]} \log q_\theta(T(X_i)).$$

How does this compare to the previous approaches? It turns out that in general this is not better than the un-augmented MLE. Suppose that the group G is discrete. Then we have

$$q_T(t) = \sum_{g \in G} p_X(g \cdot (t, 0)) = |G| \cdot p_X((t, 0)).$$

Therefore, in this case the iMLE equals the MLE. Therefore, the invariant MLE does not actually gain anything over the usual MLE, and in particular augmented MLE is better.

Appendix D. Experiment details

Our code is available at `https://github.com/dobriban/data_aug`. Our experiment to generate Figure 1 is standard: We train ResNet18 (He et al., 2016) on CIFAR10 (Krizhevsky, 2009) for 200 epochs, based on the code of `https://github.com/kuangliu/pytorch-cifar`. The CIFAR10 dataset is standard and can be downloaded from `https://www.cs.toronto.edu/~kriz/cifar.html`. We use the default settings from that code, including the SGD optimizer with a learning rate of 0.1, momentum 0.9, weight decay $5 \cdot 10^{-4}$, and batch size of 128. We train three models: (1) without data augmentation, (2) horizontally flipping the image with 0.5 probability, and (3) a composition of randomly cropping a 32×32 portion of the image and random horizontal flip; besides the data augmentation, all other hyperparameters and settings are kept the same. We repeat this experiment 15 times and plot the average test accuracy for each number of training epochs. The shaded regions represent 1 standard deviation around the average test accuracy. We train both on the full CIFAR10 training data, as well as a randomly chosen half of the training data. We do this to evaluate the behavior of data augmentation in the limited data regime, because there it may to lead to higher benefits. This experiment was done on a p3.2xlarge (GPU) instance on Amazon Web Services (AWS).

References

F. Anselmi, G. Evangelopoulos, L. Rosasco, and T. Poggio. Symmetry-adapted representation learning. *Pattern Recognition*, 86:201–208, 2019.

A. Antoniou, A. Storkey, and H. Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019.

H. S. Baird. Document image defect models. In *Structured Document Image Analysis*, pages 546–556. Springer, 1992.

A. S. Bandeira, B. Blum-Smith, J. Kileel, A. Perry, J. Weed, and A. S. Wein. Estimation under group actions: recovering orbits from invariants. *arXiv preprint arXiv:1712.10163*, 2017.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

T. Bendory, N. Boumal, C. Ma, Z. Zhao, and A. Singer. Bispectrum inversion with application to multireference alignment. *IEEE Transactions on Signal Processing*, 66(4): 1037–1050, 2018.

Y. Bengio, F. Bastien, A. Bergeron, N. Boulanger-Lewandowski, T. Breuel, Y. Chherawala, M. Cisse, M. Côté, D. Erhan, J. Eustache, et al. Deep learners benefit more from out-of-distribution examples. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 164–172, 2011.

U. Bergmann, V. Yachandra, and J. Yano, editors. *X-Ray Free Electron Lasers*. Energy and Environment Series. The Royal Society of Chemistry, 2017.

B. Bloem-Reddy and Y. W. Teh. Probabilistic symmetry and invariant neural networks. *arXiv preprint arXiv:1901.06082*, 2019.

P. J. Brockwell and R. A. Davis. *Time series: theory and methods*. Springer, 2009.

J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

P. Chao, T. Mazaheri, B. Sun, N. B. Weingartner, and Z. Nussinov. The stochastic replica approach to machine learning: Stability and parameter optimization. *arXiv preprint arXiv:1708.05715*, 2017.

Z. Chen, Y. Cao, D. Zou, and Q. Gu. How much over-parameterization is sufficient to learn deep relu networks? *arXiv preprint arXiv:1911.12360*, 2019.

D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber. Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, 22(12):3207–3220, 2010.

J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

T. Cohen and M. Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016a.

T. Cohen, M. Geiger, and M. Weiler. A general theory of equivariant cnns on homogeneous spaces. *arXiv preprint arXiv:1811.02017*, 2018a.

T. S. Cohen and M. Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016b.

T. S. Cohen, M. Geiger, J. Köhler, and M. Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018b.

E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2019.

T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, and C. Re. A kernel theory of modern data augmentation. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

T. DeVries and G. W. Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017a.

T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017b.

S. Dieleman, J. De Fauw, and K. Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. *arXiv preprint arXiv:1602.02660*, 2016.

L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis. Learning so(3) equivariant representations with spherical cnns. In *The European Conference on Computer Vision (ECCV)*, September 2018a.

C. Esteves, A. Sud, Z. Luo, K. Daniilidis, and A. Makadia. Cross-domain 3d equivariant image embeddings. *arXiv preprint arXiv:1812.02716*, 2018b.

C. Esteves, Y. Xu, C. Allen-Blanchette, and K. Daniilidis. Equivariant multi-view networks. *arXiv preprint arXiv:1904.00993*, 2019.

V. Favre-Nicolin, J. Baruchel, H. Renevier, J. Eymery, and A. Borbély. XTOP: high-resolution X-ray diffraction and imaging. *Journal of Applied Crystallography*, 48(3):620–620, 2015.

N. I. Fisher, T. Lewis, and B. J. Embleton. *Statistical analysis of spherical data*. Cambridge university press, 1993.

D. Foster, A. Sekhari, O. Shamir, N. Srebro, K. Sridharan, and B. Woodworth. The complexity of making the gradient small in stochastic convex optimization. *arXiv preprint arXiv:1902.04686*, 2019.

J. Frank. *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. Oxford University Press, 2006.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

R. Gens and P. M. Domingos. Deep symmetry networks. In *Advances in neural information processing systems*, pages 2537–2545, 2014.

N. C. Giri. *Group invariance in statistical inference*. World Scientific, 1996.

S. Hauberg, O. Freifeld, A. B. L. Larsen, J. Fisher, and L. Hansen. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *Artificial Intelligence and Statistics*, pages 342–350, 2016.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90.

I. S. Helland. Statistical inference under symmetry. *International Statistical Review*, 72(3): 409–422, 2004.

A. Hernández-García and P. König. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*, 2018a.

A. Hernández-García and P. König. Further advantages of data augmentation on convolutional neural networks. In *International Conference on Artificial Neural Networks*, pages 95–103. Springer, 2018b.

A. Hernández-García, J. Mehrer, N. Kriegeskorte, P. König, and T. C. Kietzmann. Deep neural networks trained with heavier data augmentation learn features closer to representations in hit. In *Conference on Cognitive Computational Neuroscience*, 2018.

D. Ho, E. Liang, I. Stoica, P. Abbeel, and X. Chen. Population based augmentation: Efficient learning of augmentation policy schedules. *arXiv preprint arXiv:1905.05393*, 2019.

E. Hoffer, T. Ben-Nun, I. Hubara, N. Giladi, T. Hoefler, and D. Soudry. Augment your batch: better training with larger batches. *arXiv preprint arXiv:1901.09335*, 2019.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

N. Jaitly and G. E. Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, volume 117, 2013.

H. Javadi, R. Balestriero, and R. Baraniuk. A hessian based complexity measure for deep networks. *arXiv preprint arXiv:1905.11639*, 2019.

Z. Ji and M. Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Z. Kam. Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations. *Macromolecules*, 10(5):927–934, 1977.

Z. Kam. The reconstruction of structure from electron micrographs of randomly oriented particles. *Journal of Theoretical Biology*, 82(1):15–39, 1980.

R. Kondor and S. Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *arXiv preprint arXiv:1802.03690*, 2018.

R. Kondor, Z. Lin, and S. Trivedi. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. In *Advances in Neural Information Processing Systems*, pages 10117–10126, 2018.

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

E. Lehmann and G. Casella. Theory of point estimation. *Springer Texts in Statistics*, 1998.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2005.

H. W. Lin, M. Tegmark, and D. Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, 2017.

L. T. Liu, E. Dobriban, and A. Singer. e pca: High dimensional exponential family pca. *The Annals of Applied Statistics*, 12(4):2121–2150, 2018.

S. Liu, D. Papailiopoulos, and D. Achlioptas. Bad global minima exist and sgd can reach them. *arXiv preprint arXiv:1906.02613*, 2019.

R. G. Lopes, D. Yin, B. Poole, J. Gilmer, and E. D. Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019.

C. Lyle, M. Kwiatkowksa, and Y. Gal. An analysis of the effect of invariance on generalization in neural networks. In *International conference on machine learning Workshop on Understanding and Improving Generalization in Deep Learning*, 2019.

L. Maaten, M. Chen, S. Tyree, and K. Weinberger. Learning with marginalized corrupted features. In *International Conference on Machine Learning*, pages 410–418, 2013.

F. R. Maia and J. Hajdu. The trickle before the torrent—diffraction data from X-ray lasers. *Scientific Data*, 3, 2016.

S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

T. Mazaheri, B. Sun, J. Scher-Zagier, A. Thind, D. Magee, P. Ronhovde, T. Lookman, R. Mishra, and Z. Nussinov. Stochastic replica voting machine prediction of stable cubic and double perovskite materials and binary alloys. *Physical Review Materials*, 3(6): 063802, 2019.

M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

T. Nguyen and S. Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In *International Conference on Machine Learning*, pages 1085–1093, 2013.

K. Pande, M. Schmidt, P. Schwander, and D. Saldin. Simulations on time-resolved structure determination of uncrystallized biomolecules in the presence of shot noise. *Structural Dynamics*, 2(2):024103, 2015.

D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

K. Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985.

S. Rajput, Z. Feng, Z. Charles, P.-L. Loh, and D. Papailiopoulos. Does data augmentation lead to positive margin? *arXiv preprint arXiv:1905.03177*, 2019.

A. J. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré. Learning to compose domain-specific transformations for data augmentation. In *Advances in neural information processing systems*, pages 3236–3246, 2017.

S. Ravanbakhsh, J. Schneider, and B. Poczos. Equivariance through parameter-sharing. 2017.

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

D. K. Saldin, V. L. Shneerson, R. Fung, and A. Ourmazd. Structure of isolated biomolecules obtained from ultrashort x-ray pulses: exploiting the symmetry of random orientations. *Journal of Physics: Condensed Matter*, 21(13), 2009.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117, 2015.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*, volume 391. John Wiley & Sons, 2009.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

A. Singer. Mathematics for cryo-electron microscopy. *arXiv preprint arXiv:1803.06714*, 2018.

L. Sixt, B. Wild, and T. Landgraf. Rendergan: Generating realistic labeled data. *Frontiers in Robotics and AI*, 5:66, 2018.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

T. Tao. *Topics in random matrix theory*. American Mathematical Society, 2012.

T. Tran, T. Pham, G. Carneiro, L. Palmer, and I. Reid. A bayesian data augmentation approach for learning deep models. In *Advances in Neural Information Processing Systems*, pages 2797–2806, 2017.

A. W. Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

C. Vonesch, F. Stauber, and M. Unser. Steerable pca for rotation-invariant image recognition. *SIAM Journal on Imaging Sciences*, 8(3):1857–1873, 2015.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

M. Weiler, M. Geiger, M. Welling, W. Boomsma, and T. Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In *Advances in Neural Information Processing Systems*, pages 10381–10392, 2018.

T. Wiatowski and H. Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. *IEEE Transactions on Information Theory*, 64(3):1845–1866, 2018.

D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017.

Y. Wu and Y. Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le. Unsupervised data augmentation. *arXiv preprint arXiv:1904.12848*, 2019.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Z. Zhao, Y. Shkolnisky, and A. Singer. Fast steerable principal component analysis. *IEEE Transactions on Computational Imaging*, 2(1):1–12, 2016.

Z. Zhao, L. T. Liu, and A. Singer. Steerable e pca. *arXiv preprint arXiv:1812.08789*, 2018.

Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.